Language Representations for Computational Argumentation

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von
Anne Ruth Lauscher
aus Trier

Mannheim, 2021

Dekan	Dr. Bernd Lübcke, Universität Mannheim
Referent	Prof. Dr. Goran Glavaš, Universität Mannheim
Korreferent	Prof. Dr. Simone Paolo Ponzetto, Universität Mannheim
Korreferent	Prof. Dr. Kai Eckert, Hochschule der Medien Stuttgart
Korreferent	Prof. Dr. Marie-Francine Moens, Katholieke Universiteit te Leuven

Tag der mündlichen Prüfung: 02. August 2021

Acknowledgements

During this journey, I have received tremendous support, advice, and inspiration.

I first and foremost thank my primary supervisor Prof. Goran Glavaš. You supported me all the way: providing close guidance and even mental support when needed and letting me fly when I wanted to work independently and with other people on other projects. You heard my ideas and supported my passion for research. You are a great person showing an extreme devotion to being a good researcher and supervisor.

Similarly, I like to thank my secondary and third supervisors Prof. Simone Paolo Ponzetto and Prof. Kai Eckert. They shared their research visions and advice and provided me with inspiring academic environments throughout the years. In addition, I would like to thank the other two members of my Ph.D. committee: Prof. Heiko Paulheim and Prof. Marie-Francine Moens – without you, it would not have been possible to close this chapter. Thank you also for the exciting and challenging questions during my defense!

Now, I like to thank my colleagues, who (at least before this pandemic) always have been available for fun chats when I needed some distraction. They have also built a super great Ph.D. hat, which I promise not to throw away for a very long time. I cannot name all of you, and therefore I will not try to do so except for Dr. Federico Nanni, who initially supervised my master thesis, which made me want to conduct a Ph.D.! You have been a great mentor, office buddy, and friend – and I think you are a person of great integrity.

Further, I would like to thank all my collaborators throughout the years of my Ph.D. (here I list only the ones from works that relate to the thesis): Dr. Ivan Vulić, Dr. Edoardo Maria Ponti, Prof. Anna Korhonen, Lily Ng, Dr. Courtney Napoles, Dr. Joel Tetreault, Vinit Ravishankar, Olga Majewska, Leonardo Ribeiro, Prof. Iryna Gurevych, and Nikolai Rozanov. I learned a lot from working with you!

Then I would like to thank my family (Walter, Edith, and Dr. Dirk Lauscher) and my many friends. I cannot say how glad I am to have you in my life. Here, I can also not name all, which is why I will restrict myself to the ones who have been most closely following my Ph.D. journey: Kilian Theil, Arne Lange, Dr. Johanna Fernández Castro, Theresa Bührle, Julia Wagner, Noah Jacobi, Katharina Pfeil, and Ophelia Sullivan.

And finally,

thank you for the music.

Abstract

Argumentation is an essential feature and, arguably, one of the most exciting phenomena of natural language use. Accordingly, it has fascinated scholars and researchers in various fields, such as linguistics and philosophy, for long. Its computational analysis, falling under the notion of *computational argumentation*, is useful in a variety of domains of text for a range of applications. For instance, it can help to understand users' stances in online discussion forums towards certain controversies, to provide targeted feedback to users for argumentative writing support, and to automatically summarize scientific publications. As in all natural language processing pipelines, the text we would like to analyze has to be introduced to computational argumentation models in the form of numeric features. Choosing such suitable semantic representations is considered a core challenge in natural language processing. In this context, research employing static and contextualized pretrained text embedding models has recently shown to reach state-of-the-art performances for a range of natural language processing tasks. However, previous work has noted the specific difficulty of computational argumentation scenarios with language representations as one of the main bottlenecks and called for targeted research on the intersection of the two fields. Still, the efforts focusing on the interplay between computational argumentation and representation learning have been few and far apart. This is despite (a) the fast-growing body of work in both computational argumentation and representation learning in general and (b) the fact that some of the open challenges are well known in the natural language processing community.

In this thesis, we address this research gap and acknowledge the specific importance of research on the intersection of representation learning and computational argumentation. To this end, we (1) identify a series of challenges driven by inherent characteristics of argumentation in natural language and (2) present new analyses, corpora, and methods to address and mitigate each of the identified issues. Concretely, we focus on five main challenges pertaining to the current state-of-the-art in computational argumentation:

(C1) External knowledge: static and contextualized language representations encode distributional knowledge only. We propose two approaches to complement this knowledge with knowledge from external resources. First, we inject lexico-semantic knowledge through an additional prediction objective in the pretraining stage. In a second study, we demonstrate how to inject conceptual knowledge post hoc employing the adapter framework. We show the effectiveness of these approaches on general natural language understanding and argumentative reasoning tasks.

(C2) Domain knowledge: pretrained language representations are typically trained

on big and general-domain corpora. We study the trade-off between employing such large and general-domain corpora versus smaller and domain-specific corpora for training static word embeddings which we evaluate in the analysis of scientific arguments.

(C3) Complementarity of knowledge across tasks: many computational argumentation tasks are interrelated but are typically studied in isolation. In two case studies, we show the effectiveness of sharing knowledge across tasks. First, based on a corpus of scientific texts, which we extend with a new annotation layer reflecting fine-grained argumentative structures, we show that coupling the argumentative analysis with other rhetorical analysis tasks leads to performance improvements for the higher-level tasks. In the second case study, we focus on assessing the argumentative quality of texts. To this end, we present a new multi-domain corpus annotated with ratings reflecting different dimensions of argument quality. We then demonstrate the effectiveness of sharing knowledge across the different quality dimensions in multi-task learning setups.

(C4) Multilinguality: argumentation arguably exists in all cultures and languages around the globe. To foster inclusive computational argumentation technologies, we dissect the current state-of-the-art in zero-shot cross-lingual transfer. We show big drops in performance when it comes to resource-lean and typologically distant target languages. Based on this finding, we analyze the reasons for these losses and propose to move to inexpensive few-shot target-language transfer, leading to consistent performance improvements in higher-level semantic tasks, e.g., argumentative reasoning.

(C5) Ethical considerations: envisioned computational argumentation applications, e.g., systems for self-determined opinion formation, are highly sensitive. We first discuss which ethical aspects should be considered when representing natural language for computational argumentation tasks. Focusing on the issue of unfair stereotypical bias, we then conduct a multi-dimensional analysis of the amount of bias in monolingual and cross-lingual embedding spaces. In the next step, we devise a general framework for implicit and explicit bias evaluation and debiasing. Employing intrinsic bias measures and benchmarks reflecting the semantic quality of the embeddings, we demonstrate the effectiveness of new debiasing methods, which we propose. Finally, we complement this analysis by testing the original as well as the debiased language representations for stereotypically unfair bias in argumentative inferences.

We hope that our contributions in language representations for computational argumentation fuel more research on the intersection of the two fields and contribute to fair, efficient, and effective natural language processing technologies.

Zusammenfassung

Argumentation ist eine essentielle Eigenschaft und eines der wohl aufregendsten Phänomene in der Benutzung natürlicher Sprache. Entsprechend sind Forscher*innen verschiedenster Disziplinen, wie beispielsweise der Linguistik oder der Philosophie, seit langer Zeit fasziniert von ihrem Studium. Die computergestütze Analyse von Argumentation, die unter den Begriff *Computational Argumentation* fällt, ist in einer Vielfalt von Textdomänen und Anwendungen nützlich. So kann sie z.B. dabei helfen, Haltungen von Benutzern von Online-Foren in Bezug auf unterschiedlichste Kontroversen zu verstehen, gezieltes Feedback zur Qualität argumentativer Texte zu geben und automatisch wissenschaftliche Publikationen zusammenzufassen. Wie in allen Pipelines in Natural Language Processing, muss der Text, der analysiert werden soll, den Computational Argumentation-Modellen in Form numerischer Features eingegeben werden. Repräsentationen zu finden, die die Semantik eines Texts adäquat reflektieren, wird als eine der Kernfragestellungen in Natural Language Processing betrachtet. In diesem Kontext erzielte kürzlich Forschung, die vortrainierte statische und kontextualisierte Embedding-Methoden einsetzt, state-of-the-art Ergebnisse in einer Reihe von Textverstehensaufgaben. Vorhergegangene Arbeit hat jedoch bereits die spezifische Schwierigkeit von Szenarien in Computational Argumentation erkannt und dabei Sprachrepräsentationen als einen Hauptengpass identifiziert. Dennoch gibt es nur wenige Anstrengungen, die sich gezielt auf die Schnittstelle von Sprachrepräsentationen und Computational Argumentation beziehen und das trotz (a) einer schnell wachsenden Anzahl von Arbeiten in beiden Forschungsbereichen und (b) des Fakts, dass manche der Probleme der Natural Language Processing-Gemeinschaft wohlbekannt sind.

In der vorliegenden Thesis adressieren wir diese Forschungslücke und erkennen die spezifische Wichtigkeit von Forschung am Zusammenspiel zwischen Computational Argumentation und Repräsentationslernen an. Dazu (1) identifizieren wir zunächst eine Serie von Herausforderungen basierend auf inhärenten Charakteristika von Argumentation und (2) präsentieren neue Analysen, Maßzahlen, Textkorpora und Methoden, um jedes der zuvor identifizierten Probleme zu adressieren. Konkret fokussieren wir uns dabei auf die folgenden fünf Herausforderungen:

(**C1**) Externes Wissen: Aktuelle Sprachrepräsentationen kodieren ausschließlich distributionelles Wissen. Wir schlagen zwei neue Ansätze vor, um dieses mit Wissen aus externen Ressourcen zu komplementieren. Als erstes fügen wir lexiko-semantisches Wissen in der Vortrainingsphase über ein zusätzliches Vorhersageziel hinzu. In einer zweiten Studie demonstrieren wir wie konzeptuelles Wissen post hoc über das Adapter-

Framework injeziert werden kann. Wir zeigen die Effektivität dieser Ansätze in generellen Textverstehensaufgaben und im argumentativen Schlussfolgern.

(**C2**) Domänen-spezifisches Wissen: Vortrainierte Sprachrepräsentationen werden typischerweise auf großen und allgemeinen Textkorpora trainiert. Wir studieren den Trade-off zwischen dem Einsatz großer und allgemeiner vs. kleiner und domänen-spezifischer Korpora, welche wir in der Analyse wissenschaftlicher Argumente evaluieren.

(**C3**) Geteiltes Wissen zwischen Aufgaben: Viele der Natural Language Processing-Aufgaben in Computational Argumentation sind miteinander verknüpft, werden aber oft in Isolation betrachtet. In zwei Fallstudien demonstrieren wir die Effektivität dessen, Wissen zwischen solchen Aufgaben zu teilen. Zuerst zeigen wir, dass es zu Performanzverbesserungen führt, die feingranulare argumentative Strukturanalyse mit anderen Aufgaben in der rhetorischen Analyse wissenschaftlicher Texte zu verknüpfen. Dazu erstellen wir außerdem neue Annotationen, welche diese argumentative Struktur in einem Korpus wissenschaftlicher Texte ausweisen. In der zweiten Fallstudie fokussieren wir uns auf das Bewerten von Argumentationsqualität. Hierzu präsentieren wir ein neues multi-domänen Korpus, welches mit Bewertungen verschiedener Dimensionen von Argumentationsqualität annotiert ist. Wir demonstrieren dann, dass es zu Verbesserungen führt, wenn Wissen zwischen diesen verschiedenen Dimensionen geteilt wird.

(**C4**) Multilingualität: Um inklusive Computational Argumentation-Technologien zu gewährleisten, sezieren wir den aktuellen State-of-the-Art in Zero-Shot Cross-Lingual Transfer. Wir zeigen hier, dass große Performanzverluste für ressourcenarme und typologisch weit von der Quellsprache entfernte Zielsprachen entstehen. Basierend darauf analysieren wir die Gründe dafür und schlagen im Anschluss alternativ dazu den effizienten Few-Shot Target-Language Transfer vor, welcher zu konsistenten Performanzverbesserungen in z.B. argumentativem Schlussfolgern führt.

(**C5**) Ethische Überlegungen: Manche der angestrebten Computational Argumentation-Anwendungen sind hochgradig sensitiv. Daher diskutieren wir zunächst, welche ethischen Aspekte berücksichtigt werden müssen. Im Anschluss fokussieren wir uns auf das Problem unfairer stereotypischer Verzerrungen in statischen Sprachrepräsentationen. Hierzu analysieren wir zunächst das Ausmaß dieser Verzerrungen. Im nächsten Schritt entwickeln wir ein generelles Framework für implizite und explizite Verzerrungsevaluation und zum Entzerren solcher Repräsentationsräume. In einer intrinsischen Evaluation demonstrieren wir die Effektivität neuer Entzerrungsmethoden, die wir vorschlagen. Zuletzt vervollständigen wir diese Analyse extrinsisch, in dem wir die Sprachrepräsentationen auf unfaire Verzerrung in argumentativem Schlussfolgern testen.

Wir hoffen, dass unsere Forschung zu Sprachrepräsentationen für Computational Argumentation weitere Forschung zu diesem Thema antreibt und wir zu fairen, effizienten und effektiven Sprachverarbeitungstechnologien beitragen.

Contents

List of Publications

Most of the work presented in this thesis has been previously published in proceedings of top-tier international conferences and workshops. This includes text, figures, and tables. We reference the corresponding publications in the respective Chapters and list them here in inverse chronological order. Conference ranking denotations (A* and A) are taken from CORE Conference Ranking, a widely adopted ranking of conferences in Computer Science. Denotation A* indicates a leading conference in a discipline area and A indicates a venue highly respected in a discipline area.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING, A)*, pages 1371–1383, Barcelona, Spain (Online), December 2020, International Committee on Computational Linguistics.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. Rhetoric, Logic, and Dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING, A)*, pages 4563–4574, Barcelona, Spain (Online), December 2020, International Committee on Computational Linguistics.

Lily Ng,[*] **Anne Lauscher**,[*] Joel Tetreault, and Courtney Napoles. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining (ArgMining)*, pages 117–126, Online, December 2020, Association for Computational Linguistics. *Awarded outstanding paper.*

Anne Lauscher,[*] Vinit Ravishankar,[*] Ivan Vulić, and Goran Glavaš. From Zero to Hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP, A)*, pages 4483–4499, Online, November 2020, Association for Computational Linguistics.

[*] Equal contribution.

Anne Lauscher, Olga Majewska, Leonardo FR Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. Common Sense or World Knowledge? Investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online, November 2020, Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI, A*)*, pages 8131–8138, New York, New York, January 2020, AAAI Press.

Anne Lauscher and Goran Glavaš. Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 85–91, Minneapolis, Minnesota, June 2019, Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. Investigating the role of argumentation in the rhetorical analysis of scientific publications with neural multi-task learning models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP, A)*, pages 3326–3338, Brussels, Belgium, October–November 2018, Association for Computational Linguistics

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining (ArgMining)*, pages 40–46, Brussels, Belgium, October-November 2018, Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. In *Proceedings of the 6th International Workshop on Mining Scientific Publications*, pages 24–28, Toronto, ON, Canada, December 2017, ACM Press.

In addition to the publications listed before, which are directly related to the content of this thesis, the author contributed to and published other research work during the course of her doctoral studies. We list them here in inverse chronological order:

Soumya Barikeri, **Anne Lauscher**, Ivan Vulić, Goran Glavaš. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers, ACL, A*)*, pages 1941—1955, Online, August 2021, Association for Computational Linguistics.

Niklas Friedrich, **Anne Lauscher**, Simone Paolo Ponzetto, Goran Glavaš. DebIE: A Platform for Implicit and Explicit Debiasing of Word Embedding Spaces. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL, A)*, pages 91–98, Online, April 2021, Association for Computational Linguistics.

Shintaro Yamamoto, **Anne Lauscher**, Simone Paolo Ponzetto, Goran Glavaš, Shigeo Morishima. Self-Supervised Learning for Visual Summary Identification in Scientific Publications. In *Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval*, pages 5–19, Lucca, Italy (online), April 2021, CEUR Workshop Proceedings.

Marilena Daquino, Silvio Peroni, David Shotton, Giovanni Colavizza, Behnam Ghavimi, **Anne Lauscher**, Philipp Mayr, Matteo Romanello, and Philipp Zumstein. The OpenCitations data model. In *The Semantic Web – ISWC 2020. Lecture Notes in Computer Science, vol 12507., (ISWC, A)*, pages, 447–463, November 2020, Springer, Cham.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. AraWEAT: Multidimensional analysis of biases in arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP)*, pages 192–199, December 2020, Barcelona, Spain (Online), Association for Computational Linguistics.

Anne Lauscher, Yide Song, and Kiril Gashteovski. MinScIE: Citation-centered open information extraction. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL, A*)*, pages 386–387, June 2019, Champaign, IL, IEEE.

Anne Lauscher, Goran Glavaš, and Kai Eckert. ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining (ArgMining)*, pages 22–28, Brussels, Belgium, October–November 2018, Association for Computational Linguistics.

Anne Lauscher, Kai Eckert, Lukas Galke, Ansgar Scherp, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, Philipp Zumstein, and Annette Klein. Linked Open Citation Database: Enabling libraries to contribute to an open and interconnected citation graph. In *Proceedings of the Joint Conference on Digital Libraries (JCDL, A*)*, pages 109–118, Fort Worth, TX, USA, May 2018, ACM Press.

Thorsten Keiper, Zhonghao Lyu, Sara Pooladzadeh, Yuan Xu, Jingyi Zhang, **Anne Lauscher**, and Simone Paolo Ponzetto. UniMa at SemEval-2018 Task 7: Semantic relation extraction and classification from scientific publications. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 826–830, New Orleans, Louisiana, June 2018, Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, and Kai Eckert. Citation-based summarization of scientific articles using semantic textual similarity. In *Proceedings of the Computational Linguistics Scientific Summarization Shared Task*, pages 33–42, Tokyo, Japan, August 2017, CEUR Workshop Proceedings.

List of Figures

List of Tables

List of Acronyms

AC	Argument Components. 92
Acc	Accuracy. 62, 65, 70
ACI	Argument Component Identification. 93–98
ACL	Association for Computational Linguistics. 78–80
AI	Artificial Intelligence. 35, 36
AM	Argument Mining. 12, 20, 21, 86
AQ	Argument Quality. 4, 6, 7, 12, 17, 22, 24, 32, 42, 47–49, 60, 82, 99–106, 109–115, 117, 118, 157, 159, 208
AX	Diagnostics. 61, 62, 70
B–I–O	Begin–Inside–Outside. 21, 93
BAM	Bias Alignment Method. 6, 130, 143, 144, 147–151, 153
BAT	Bias Analogy Test. 5, 130, 145, 147–149
BERT	Bidirectional Encoder Representations from Transformers. 28, 29, 32, 33, 42, 46, 51, 54–72, 102, 112–117, 121, 122, 156
CA	Computational Argumentation. 1–7, 12, 17, 19, 20, 22–25, 30, 34, 38–42, 44, 45, 47–49, 51–54, 66, 68, 69, 73, 74, 77, 82, 100, 118, 119, 129–132, 135, 139, 140, 156–159
CBOW	Continuous Bag of Words. 25–27, 134, 136, 147–149, 153
CC	Citation Contexts. 92
CC-100	CommonCrawl-100. 33, 121
CCC	Citation Context Identification. 93–98
CL	Computational Linguistics. 41, 74, 77, 78, 100, 101
CLWE	Cross-lingual Word Embeddings. 120–122, 135, 143
CM	ConvinceMe. 107
CMV	Change My View. 107
CNN	Convolutional Neural Network. 75–77, 79–81
CoLA	Corpus of Linguistic Acceptability. 60–62, 70
CQA	Community Questions and Answers. 101, 106–109, 111–114, 116–118, 159
CRF	Conditional Random Fields. 85, 97
CS	Common Sense. 71
CV	Cross Validation. 79

DBN DebiasNet. 143, 148–151
DebiasNet Explicit Neural Debiasing. 6, 130, 141, 143, 144, 147–149, 151, 153
DebIE Debiasing Embeddings Implicitly and Explicitly. 5, 140, 155
DR Discourse Roles. 92
DRC Discourse Role Classification. 93, 94, 96–98

ECT Embedding Coherence Test. 5, 145, 148–150
EM Exact Match. 124
EQT Embedding Quality Test. 145

F1 F1-Measure. 62, 64, 65, 70, 80, 97
Fa Factivity. 63
FN Fraction Neutral. 152, 153
FT fastText. 147–150, 153

GBDD General Bias Direction Debiasing. 130, 142, 144, 147–151, 153
GloVe Global Vectors. 25, 27, 38, 78, 97, 131, 134, 136, 137, 140, 147–150
GLUE General Language Understanding Evaluation. 55, 56, 60–62, 66–70
GNLU General Natural Language Understanding. 20, 54
GPU Graphics Processing Unit. 51, 129

HMM Hidden Markov Model. 97

IAA Inter-Annotator Agreement. 89, 90, 99, 103, 106, 109
IAC Internet Argument Corpus V2. 104, 107
IAT Implicit Association Test. 132

KB Knowledge Base. 43, 57, 67, 71, 121
KCS Knowledge and Common Sense. 63, 70, 71
KM KMeans++. 148–151

LE Lexical Entailment. 63
LeS Lexical Semantics. 63, 70, 71
LIBERT Lexically Informed BERT. 54–66, 156
LM Language Model. 55, 66–68, 120, 121
Lo Logic. 63, 71
LRC Lexical Relation Classification. 56, 57, 59–61
LS Lexical Simplification. 56, 65, 66
LSTM Long Short-Term Memory. 94–96, 98

MAE Mean Absolute Error. 125
mBERT Multilingual BERT. 33, 50, 119, 121–128
MCC Matthews Correlation Coefficient. 62, 70, 71
MLM Masked Language Modeling. 28, 55, 57, 59, 61–66, 68, 121
MMT Massively Multilingual Transformer. 5, 6, 33, 50, 51, 119–129, 158
MN Morphological Negation. 63

MNLI Multi-Genre Natural Language Inference. 61, 123, 152
MNLI-m MNLI-matched. 61, 62, 70, 152, 153
MNLI-mm MNLI-mismatched. 61, 62, 70, 152, 153
MRPC Microsoft Research Paraphrase Corpus. 60–63, 70
MT Machine Translation. 121, 123
MTL Multi-Task Learning. 6, 32, 34, 48, 49, 56, 61, 62, 66, 82–86, 93–99, 102, 112, 113, 115, 157

NE Named Entities. 63, 71
NER Named Entity Recognition. 21, 122
NLI Natural Language Inference. 2, 22, 23, 29, 39, 40, 42–44, 56, 61, 63, 66, 68, 70–72, 119, 122, 123, 128, 139, 140, 151, 153, 156
NLP Natural Language Processing. 1, 2, 10, 20, 22–25, 30–33, 37, 38, 41, 45–47, 49, 51, 52, 54, 55, 66, 74, 77, 78, 85, 100–102, 119, 129, 131, 135, 144
NLU Natural Language Understanding. 2, 4, 7, 19, 41–44, 48, 54, 56, 123, 125, 156
NMI Normalized Mutual Information. 92
NSP Next Sentence Prediction. 28, 29, 55, 57, 59, 61–64, 66, 121

OMCS Open Mind Common Sense. 67–72

P Precision. 64, 65, 80, 97
PAS Predicate-Argument Structure. 63, 71
PCA Principal Component Analysis. 151
Pears Pearson's Correlation Coefficient. 62, 125
POS Part-of-Speech. 122

QA Question Answering. 1, 61, 119, 123, 128
QNLI Question NLI. 61, 62, 70
QQP Quora Question Pairs. 60–62, 70
Qu Quantifiers. 63

R Recall. 64, 65, 80, 97
RBF Radial Basis Function. 75, 77, 97, 146
Re Redundancy. 63
RoBERTa Robustly Optimized BERT Approach. 28, 29, 42, 121
RTE Recognizing Textual Entailment. 61, 62, 70

SA Subjective Aspects. 92
SAC Subjective Aspect Classification. 93, 94, 96–99
SemQ Semantic Quality. 148
SL SimLex-999. 148–151, 153
SNLI Stanford Natural Language Inference. 22, 152, 153
Spear Spearman's Rank Correlation Coefficient. 70, 125
SQuAD Stanford Question Answering Dataset. 123
SR Summary Relevances. 92

SRC Summary Relevance Classification. 94, 96–98
SST-2 Stanford Sentiment Treebank v2. 60–63, 70
STILT Supplementary Training on Intermediate Labeled-Data Tasks. 32, 48, 49, 82, 117, 118
STL Single-Task Learning. 96, 97
STS-B Semantic Textual Similarity Benchmark. 60–62, 70
SVM Support Vector Machines. 75–77, 79–81, 97, 98, 146, 148, 149
SVR Support Vector Regression. 112–115, 125

TF–IDF Term Frequency–Inverse Document Frequency. 2, 77, 79, 80, 97, 113

W WEAT. 150, 151
WCBOW Weighted Continuous Bag of Words. 77
WEAT Word Embedding Association Test. 4, 52, 131–138, 141, 145–152, 154, 155, 158
World World Knowledge. 71
WS WordSim-353. 148, 149, 153

XLM-R XLM-RoBERTa. 33, 50, 121–128
XNLI Cross-Lingual Natural Language Inference. 123–128
XQuAD Cross-lingual Question Answering Dataset. 123–128
XWEAT Cross-lingual WEAT. 4, 5, 130, 132, 137, 138, 146, 150, 151, 154

For referring to specific languages, we use ISO 639-1 codes, e.g., EN for English.

Mathematical Notation

Throughout this work we adhere to the following notation.

Numbers, Arrays, Sets, and Tuples

a	A scalar
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
a_i	Element at position i of vector $\mathbf{a}$
a_{ij}	Element at position i, j of matrix $\mathbf{A}$
$a_{i,:}$	Row i of matrix $\mathbf{A}$
$a_{:,j}$	Column j of matrix $\mathbf{A}$
A	A set
$\mathbb{R}$	The set of real numbers
$A \subset B$	Set A is a subset of set B
$a^{(i)}$	An element i in the set
$(x_1, x_2, ..., x_n)$	An ordered n-tuple

Linear Algebra Operations

$\mathbf{A}^\top$	Transpose of a matrix $\mathbf{A}$
$\mathbf{a} \cdot \mathbf{b}$	The dot product between vectors $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{a} \frown \mathbf{b}$	The concatenation of two vectors $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{a} \odot \mathbf{b}$	The element-wise product of two vectors $\mathbf{a}$ and $\mathbf{b}$ (Hadamard product)

Set Operations

$A \times B$	The cartesian product of the sets A and B
$A \setminus B$	The difference of the sets A and B
$A \cup B$	The union of the sets A and B

Probabilities

A	The random variable X
$P(A)$	The probabilty distribution over the random variable A
$P(A\|B)$	The conditional probabilty distribution over A given B

Functions

$f(\cdot)$	A function f
$f \circ g$	Composition of functions f and g
$\log x$	Natural logarithm of x

Learning

$\mathcal{X}$	A feature space
$\mathcal{Y}$	A label space
$f : \mathcal{X} \to \mathcal{Y}$	A function f that maps from $\mathcal{X}$ to $\mathcal{Y}$
Ω	A search space
$\mathcal{T}$	A machine learning task
$\mathcal{D}$	A machine learning domain
θ	The set of model parameters

(This page is intentionally left blank.)

Chapter 1

Introduction

1.1 Motivation and Problem Statement

> *Rien n'est stupide comme vaincre; la vraie gloire est convaincre.*
> *(Nothing is so stupid as to vanquish; the real glory is to convince.)*
>
> Victor Hugo, Les Misérables

Argumentation, as a direct reflection of human reasoning in natural language, has fascinated scholars and researchers in various disciplines, such as philosophy, logic, and linguistics, for long. Being tied to the development of democracy and public discourse in Europe, argumentation-theoretic literature teaching the art to convince the other can be traced back to the origins of the city-states in ancient Greece. But argumentation does not only occur in the political and public discourse – it plays an important role in solving internal controversies with ourselves as well as in any "social arena" (Atkinson et al., 2017). Accordingly, the theory of argumentation has been studied in a variety of textual domains, such as web debates (e.g., Habernal and Gurevych, 2016), business reviews (e.g., Wachsmuth et al., 2015), and scientific writing (e.g., Green, 2015b).

Computational argumentation (CA), which covers the computational (a) mining of arguments, (b) assessment of arguments, and (c) reasoning over arguments, requires deep language understanding capabilities (Moens, 2018). Much like other semantically challenging natural language processing (NLP) tasks, such as question answering (QA; Rajpurkar et al., 2016) and reading comprehension (Saha et al., 2018), CA tasks have received more and more attention with the growing amount of publicly available textual data and the increased amount of computational processing power. Atkinson et al. (2017) acknowledge the importance of the field as follows: *"[...] argumentation pervades our intelligent behavior and the challenge of developing artificial argumentation systems appears to be as diverse and exciting as the challenge of artificial intelligence itself."* But there is not only this inherent interest, which is tied to the fundamental challenges of artificial intelligence research with the "holy grail" of creating a general artificial intelligence – the output of CA systems and especially of argumentative understanding models applied to natural language texts are useful in many practical scenarios and have an impact on

other NLP tasks. For instance, given a debate thread on vaccination on social media, a CA system can enable us to extract and understand not only the stances people have but also their particular premises and conclusions they base their positions on, including the underlying reasoning processes. All this information can next be employed in downstream applications, for instance, for efficiently and effectively summarizing the whole controversy and for automatically retrieving good arguments for a particular topic and stance (Wachsmuth et al., 2017c). Similarly, based on the output of automatically analyzing citations, which are argumentative tools in scientific writing (Gilbert, 1977), and assigning sentential argument roles, i.e., *argumentative zoning* (Teufel et al., 1999), we can anticipate future trends in scientific research (e.g., McKeown et al., 2016).

Here, like in any other NLP task, the input, i.e., the text, has to be provided in a numeric format to allow for computational processing. How to optimally represent textual data numerically is, however, an ongoing research topic which has been focused on in NLP since the genesis of the field (see, e.g., Luhn, 1957). While researchers first adhered to sparse lexical document representations, such as the term frequency–inverse document frequency (TF–IDF) vectors (Sparck Jones, 1972), dense semantic representations are employed in most state-of-the-art natural language understanding (NLU) models in the field (see Wang et al., 2019b,a). Here, we can distinguish between static and contextualized embedding models. While the former provide a single vector representation for a span of text, such as a word or a subword, the latter consist of multi-layered architectures and dynamically compute the representations of spans of text based on the context provided.

However, when employing those representations in CA scenarios, a series of challenges tied to inherent characteristics of argumentation arises. Consequently, previous work indicated language representations as one of the main bottlenecks in argumentative understanding models (Moens, 2018). For instance, though static and contextualized embedding models operate fundamentally differently from each other, they are both grounded in the *distributional hypothesis* (Harris, 1954), and as such have the tendency to conflate together true lexical similarity with broader topical relatedness (Hill et al., 2015; Schwartz et al., 2015). This poses a problem, as distinguishing between similarity and relatedness can be crucial in many argumentative reasoning scenarios, such as natural language inference (NLI; Wang et al., 2019b). As a second example, previous research has shown that dense semantic representations encode biases, which reflect many human stereotypes. This is not particularly surprising as humans exhibit (a) a series of cognitive biases and (b) are socialized in certain cultural and institutional contexts, which often leads to unfair decisions, stereotypes, and prejudices about individuals in minoritized groups, e.g., due to their gender, sexuality, nationality, or religion. These prejudices, in turn, are reflected in language and consequently projected in human-produced texts. For instance, the term *man* typically occurs more often in the context of career-related terms, while the term *woman* occurs more often in the context of family terms. As these texts serve as input for inducing semantic embedding models, the numeric representation of the term *man* will be more similar to the induced representations of career-related terms than to family terms. Vice versa, the embedding of *woman* will be less similar to career-related terms and more similar to family-related terms. Employing such biased representations in NLP systems is *stereotyping*, a representational harm (Blodgett et al., 2020), and depending

on the socio-technical scenario, it might lead to bias amplification, systematically unfair system decisions, and decreased performance for minoritized classes (Sun et al., 2016). Recently, the issue of bias has been identified as a critical concern for CA (Spliethöver and Wachsmuth, 2020), given the high sensitivity of envisioned CA systems, as in the case of support systems for self-determined opinion formation (Wachsmuth et al., 2017c). As a final example, given that argumentation is supposed to exist in all of the world's 7,000 languages (Eberhard et al., 2020), we need to ensure truly multilingual CA systems to foster inclusion and democratization of language technology. However, at the moment, this is only ensured for resource-rich languages, e.g., English. Moreover, those languages are currently supported with ever-larger language representations (Bender et al., 2021), with training costs for single models, which are exceeding the ecological damage produced by taking a trans-American flight (Strubell et al., 2019). In the long run, this trend is clearly not sustainable. As those models are specifically employed in tasks requiring deep semantic understanding, as it is the case for most CA tasks, this is an additional issue with current language representations for CA. Given these three examples alone, it is evident that further research on semantic language representations for CA is required in order to ensure effective, efficient, inclusive, sustainable, and fair CA systems.

While (1) the fields of CA and representation learning are both active research fields and (2) the specific need for advanced language representations for CA has been recognized in previous research, it is surprising that to date, no work has systematically studied the ties and interrelations between those two fields.

In this work, we aim towards closing this research gap by identifying and systematically addressing a set of five prominent challenges when employing dense semantic language representations in CA research. In particular, we study the following challenges:

(C1) *External knowledge:* static and contextualized language representations encode distributional knowledge only. How can we complement this knowledge by injecting external knowledge into language representation models?

(C2) *Domain knowledge:* pretrained language representations are typically trained on big and general-domain corpora. How can we adapt language representations to encode knowledge relevant to specific domains?

(C3) *Complementarity of knowledge across tasks:* many CA tasks are interrelated, but are most often studied in isolation. How can we improve our language representations by sharing knowledge across multiple tasks?

(C4) *Mulilinguality:* argumentation arguably exists in all cultures. How can we foster inclusion in CA by accounting for multilinguality in language representations?

(C5) *Ethical considerations:* envisioned CA applications are highly sensitive. Which ethical aspects should be considered when representing natural language for CA and how can we adjust to those? How can we ensure fairness?

For each of these challenges, which we derive from inherent characteristics of argumentation and from envisioned CA systems, we conduct one or two case studies relating to the

problem. In those case studies, we provide either an extensive analysis on certain aspects of the challenge and/or propose new measures and/or approaches for mitigating the issue.

1.2 Contributions

After identifying issues with commonly employed language representations used in CA, we build on top of these insights and present contributions that can be attributed to the field of CA as well as to the representation learning area. We demonstrate the effectiveness of newly proposed techniques in representation learning by evaluating them on CA problems and general NLU tasks, which are, in turn, fundamental for mastering the area of CA. We present new approaches and resources as well as analytical insights into the challenges identified in these areas. Concretely, we make the following contributions:

Corpora. We create new annotation layers and textual resources for training and evaluating computational CA and language representation models.

1. *Argument-augmented Dr. Inventor Corpus*: in order to advance research on scientific argumentation and to allow for a better understanding of the role of fine-grained argumentative structures within the multi-layered argumentative nature of scientific writing, we present an additional annotation layer for the Dr. Inventor corpus (Fisas et al., 2016) capturing fine-grained argumentative components and relationships. This effort results in the first corpus of English scientific texts annotated with fine-grained argumentative structures and enables us to study the complementarity of knowledge (**C3**) in language representations employed for the rhetorical analysis of scientific text (see Section 6.1). We hereby also introduce the notion of *scitorics*, the rhetorical aspects of scientific argumentation, which correspond to a domain-specific group of argumentative analysis tasks (**C2**).

2. *GAQCorpus*: secondly, aiming to advance theory-based argument quality (AQ) assessment (Wachsmuth et al., 2017b), which treats overall AQ as being composed of rhetorical, logical, and dialectical aspects, we present the largest English multi-domain corpus annotated with theory-based AQ scores. This corpus enables us to study the complementarity of knowledge (**C3**) across these theory-based AQ dimensions (see Section 6.2). In this context, we also present initial results on domain-specificic aspects of language representations for AQ assessment (**C2**).

3. *XWEAT*: furthermore, to be able to measure potentially problematic stereotypical biases (**C5**) in multilingual and cross-lingual language representations (**C4**), we present cross-lingual WEAT, a translation of the Word Embedding Association Test term sets (WEAT; Caliskan et al., 2017) from English to six languages.[1] To date, XWEAT is the bias resource covering most languages. We employ the test sets in, what is to date, the largest study on bias in distributional word vector spaces (see Section 8.1) and as specifications supporting our proposed framework for implicit and explicit bias evaluation and debiasing (see Section 8.2).

[1] In addition to those six, in a recent study, we presented AraWEAT, an Arabic extension to XWEAT.

Measures. Related to measures, we present the following two contributions.

1. To take a more holistic perspective on biases in static language representations (**C5**), in this work, we assemble a framework for the implicit and explicit evaluation of stereotypical biases in distributional word vector spaces, dubbed DebIE (see Section 8.2). The framework is based on bias test specifications consisting of sets of *stimuli* among which the bias is expected to exist and integrates XWEAT, thereby allowing to measure bias in multilingual and cross-lingual scenarios. We then adapt existing measures to operate within the unified notion of these specifications, such as the Embedding Coherence Test (ECT; Dev and Phillips, 2019).

2. In addition to unifying and adapting existing bias tests, in Section 8.2, we introduce the Bias Analogy Test (BAT). BAT is a new measure testing for the existence of an explicit bias in static word embedding spaces based on the idea of biased analogies as originally introduced by Bolukbasi et al. (2016).

Analyses. Based on the newly introduced resources and measures outlined above, we conduct a series of analyses towards obtaining a better understanding of the individual challenges of language representations for CA identified.

1. We are the first to quantify unfair stereotypical bias (**C5**) in distributional word vector spaces across a variety of languages and in cross-lingual embedding spaces, including other relevant factors, such as the domain of the text corpus and embedding models in our study. This effort results in the most extensive analysis of bias in static language representations to date (see Section 8.1).

2. Relating to **C2**, domain-specific knowledge, we examine the trade-off between larger and noisier vs. smaller and more homogeneous pretraining corpora for static word embeddings. We study this trade-off within the task of semantically classifying citations as main argumentative tools in scientific writing (see Chapter 5).

3. In the context of multilinguality (**C4**), we dissect the current state-of-the-art zero-shot cross-lingual transfer approach based on massively multilingual transformer (MMT) models by quantifying the loss in performance in the transfer. To this end, we employ two tasks requiring deep semantic knowledge, including argumentative reasoning. We analyze the factors contributing to the transfer performance, such as the size of the monolingual corpora employed in pretraining (see Chapter 7). Note that ensuring multilinguality is also vital for enabling democratization of CA technologies (**C5**).

Methods. We explore and propose several new approaches with respect to the challenges identified by exploiting state-of-the-art transfer learning paradigms.

1. We are the first to employ convolutional neural networks and domain-specific word embeddings for the semantic classification of citations (**C2**, see Chapter 5).

2. Concerning the identified limitations of zero-shot cross-lingual transfer with MMT models (**C4**), we propose to move to inexpensive annotation cycles and few-shot target-language fine-tuning. We demonstrate consistent performance improvements in argumentative reasoning (see Chapter 7).

3. For tackling the issue of underrepresented external knowledge in language representations (**C1**), a bottleneck for argumentative understanding, we present two new methods: (a) the injection of lexico-semantic knowledge leading to a specialization for true semantic similarity of large language models via an extension to the pretraining procedure (see Section 4.1); and (b) the efficient injection of conceptual knowledge post hoc via adapter layers (see Section 4.2).

4. Further, we investigate (a) the role of argumentation in scientific writing and (b) the complementarity of knowledge across theory-based AQ dimensions with neural multi-task learning (MTL) models. In the case of (a), we are also the first to employ a joint loss function based on homoscedastic uncertainty in the MTL setup. For (b), we also propose a hierarchical combination of the different objectives as well as a sequential task transfer setup (**C3**, see Sections 6.1 and 6.2).

5. Last, we propose two new techniques for mitigating unfair stereotypical bias in static language representations: (1) Bias Alignment Method (BAM), which is inspired by projection-based cross-lingual word embedding spaces, and (2) Explicit Neural Debiasing (DebiasNet), inspired by previous work on semantic specialization of distributional word vector spaces (**C5**, see Section 8.2).

An overview of all resources published in the context of this thesis can be found in Part A of the supplementary material. We hope that our work fuels future research on the intersection between language representations and CA and beyond.

1.3 Outline

We first discuss this thesis's theoretical background (Chapter 2), consisting of fundamental knowledge relating to computational argumentation and representation learning. Based on inherent characteristics of argumentation and findings of previous research, we then identify shortcomings and challenges when representing text for CA applications (Chapter 3). The subsequent Chapters describe our efforts to systematically address the previously identified challenges starting with (**C1**) external knowledge (Chapter 4) to (**C5**) ethical considerations (Chapter 8). In each of those Chapters, we deal with one or two case studies related to CA, for which we provide motivation and briefly survey the related work before the actual discussion of the methodology and the results. Finally, in Chapter 9, we conclude our work and provide directions for future research.

Chapter 2

Theoretical Background

In this Chapter, we introduce fundamental concepts pertaining to the two main topics of this thesis: (1) computational argumentation, and (2) representation learning (i.e., machine learning methods for acquiring semantic representations of text).

2.1 Computational Argumentation

Acknowledging argumentation as a direct reflection of human reasoning manifested in natural language, we start by outlining the history of argumentative studies from the ancient Greeks to computational argumentation (CA). Building upon this, we then introduce argumentation-theoretic concepts, such as argument models and the notion of argument quality (AQ). As a highly stylized, domain-specific example, we discuss the special case of scientific argumentation. Finally, we investigate the link between general NLU and CA and discuss prominent CA tasks.

2.1.1 From Ancient Greeks to Computational Argumentation

The study of argumentation has a long-lasting tradition. In the western world, it can be traced back to the fifth century B.C.E. with the emerging concept of democracy in Athens and the emergence of the city-state, the *polis* (πόλις), as a political space (Vernant, 1965). Based on the idea that (*male*!) citizens[1] could participate in governing the polis, it became more important to be able to speak in public and convince the audience of a certain idea or policy, and, consequently, actively shape the future of the polis. Accordingly, a culture around the art of publicly speaking emerged and so-called *sophists* (σοφιστής) offered their services in teaching, among other skills, how to choose and combine the right structures and words into compelling arguments. As such, the study of argumentation has always been goal-oriented. One of the most influential works from ancient Greece is Aristotle's *On Rhetoric* (Aristotle, ca. 350 B.C.E./ translated 2006), which was later not only referred to by other members of the peripatetic school[2] but also by famous

[1] We use the term "citizen" to refer to individuals with full political and judicial rights. In this sense, women were not considered to be citizens in Athens (Loraux, 1994).

[2] The school of philosophy founded by Aristotle himself.

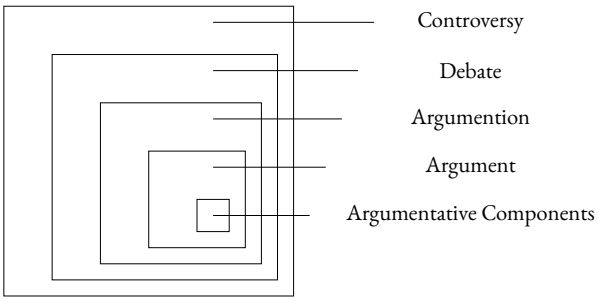


Figure 2.1: The hierarchical relationship between controversies, debates, argumentation, arguments, and their argumentative components.

Roman rhetoricians such as Cicero (Fortenbaugh, 2005). In his work, Aristotle (384–322 B.C.E.) focuses on rhetoric, which – according to him – is the art of speaking in public as opposed to dialectic as defined by his teacher Plato, which concerns academic or private matters, and is, moreover, characterized by a sequence of questions and answers. Aristotle further defines two types of speeches, which are highly argumentative in nature:[3] (1) the *deliberative* speech, which advises on a course of future action, e.g., a new policy in the polis, and (2) the *judicial* speech, which accuses or defends someone, thereby corresponding to legal argumentation, an argumentation over conclusions of past events. Both of these situations have in common that they start from some *controversy*: while in (1) the controversy is about a course of future action, and the audience, who needs to be convinced, is the public, in (2) the controversy lies in the judicial question, and the audience corresponds to a judge (Kennedy, 2009). The idea of a controversy as starting point for argumentation is even more explicitly expressed by other authors: "*[c]ontroversy is an essential prerequisite of debate.* Where there is no *clash* of ideas, proposals, interests, or expressed positions on issues, there is no debate" (Freeley and Steinberg, 2013). This clash of ideas likely results in different standpoints on the issue, may it be relating to a past event, or a course of future action, or certain beliefs, which then encourage people to *argue* for their stance in the form of a *debate*. In debates, two or more arguers present their argumentation with regard to a certain issue related to which the controversy occurred. This can be highly formalized, as in a British Union-style debate situation, in which two "houses" argue for their stance (see Haapala, 2012), and, similarly, in a very informal context among friends or relatives. Accordingly, debates are dialogical, while argumentation itself can also occur in a monological or hybrid way, for instance, in the case of scientific publications: here, scientists present their argumentation predominantly as a monologue, but link their arguments to the overall scientific discourse by using references to previous work, thereby adding a dialogical component. But what exactly is argumentation? Stede and Schneider (2018) who recently reviewed the field adopt the prominent definition of van Eemeren and Grootendorst (2010). We follow Stede and Schneider (2018) and adopt

[3] The third type, *epideictic* rhetoric, corresponds to ceremonial discourse and does not aim at persuasion directly (Lockwood, 1996), which is why we consider it non-argumentative.

this definition throughout this thesis, as the authors beautifully managed to incorporate the most relevant aspects of argumentation in a single concise sentence:

Definition 1 (Argumentation)**.** *"Argumentation is a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint."*

Stede and Schneider (2018) dissect this definition into its eight essential characteristics, which we discuss next. As a running example, we employ the case of scientific publications as we will later discuss and experiment with arguments from the scientific domain in more detail (see Section 2.1.3, Chapter 5, and Section 6.1).

Verbal Activity. Argumentation is and always has been an "(...) inherently linguistic activity" and can be either expressed in writing or in speech (Stede and Schneider, 2018). Whether it is textually or orally expressed depends on the debate situation. For instance, in the case of a scientific publication, the argumentation is mostly expressed in textual form. However, we also want to remind the reader that there are, ultimately, more forms argumentation can take, e.g., when we consider scientific publications as more complex multi-modal documents, in which information is also conveyed in visual form, helping the reader to better understand the scientific argument (Nelson et al., 1976). Still, in its core, we agree with argumentation as a "verbal activity" as highlighted by Stede and Schneider (2018), because even in such a multi-modal document, the main part of the argumentation is expressed verbally while visual parts act rather supportively. Accordingly, the present thesis focuses on argumentation expressed in natural language only.

Social Activity. According to Stede and Schneider (2018), argumentation is an interaction, usually performed between two or more people. There always has to be someone to argue with, even in monological argumentation. Here, the authors also mention the possibility of mentally arguing with one-self, but they conclude that for a real argument, there always has to be *the other*, i.e., someone to argue with. We do not necessarily agree with the authors' opinion, as for solving internal controversies, individuals can build proper arguments for each of the possible stances (even in textual form) to finally arrive at a well-founded opinion. However, in most cases, there clearly is *the other*. In scientific writing, they usually correspond to members of the scientific community, for instance, peer reviewers, or researchers working on the same or similar topics.

Rational Activity. As Stede and Schneider (2018) point out "[...] argumentation targets specifically the dimension of reason." While this is definitely an important aspect and has already been expressed by Aristotle in terms of the concept of *logos* (Aristotle, ca. 350 B.C.E./ translated 2006), it is interesting to note that this rationality, i.e., logic, corresponds to a single qualitative dimension of argumentation (see also Section 2.1.2). As a consequence, we also have to be aware that the rationality of an argument can be expressed in varying degrees, and other aspects, e.g., the emotional appeal of an argument,

can be similarly important in order to convince. In a scientific publication, however, the rational aspects should be predominant, as science is, per se, considered a rational activity.

Standpoint. Argumentation relates to a particular stance regarding a topic of discussion. Often, the set of possible standpoints in a debate is expressed in dichotomies, for instance, *pro* vs. *contra* gay marriage. In scientific publications, the set of stances is often not clear in advance. For instance, in NLP, a publication can argue for the superiority of a certain method, the superiority of a certain method in certain cases only, or the reason why a certain method works or does not work. What all these cases have in common is that the authors argue for the validity of their work and, consequently, for the validity of their opinions, ultimately aiming to be accepted by the respective scientific community.[4]

Convincement of Acceptability. On the one hand, the general idea of argumentation is to convince *the other* of the arguer's own standpoint relating to a certain topic, which typically amounts to changing the beliefs of the audience. On the other hand, in addition to changing the stance of the opponent as mentioned by Stede and Schneider (2018), argumention can also be successful if it does not change the stance of the audience as also pointed out by Al-Khatib et al. (2016): depending on one's prior belief, an argument is also successful if it empowers the audience and enables one to better defend one's standpoint. Tindale (2007) further provides five main intentions associated with argumentation: (i) persuasion of an audience, (ii) resolution of a dispute, (iii) achieving agreement in a negotiation, (iv) recommending, and (v) completing and inquiry. In scientific writing, it is typically the first intent, the persuasion of the scientific community of the described work as a valid contribution to science (Teufel, 2014).

Constellation of Propositions. Sometimes, an argument can consist of a single proposition only and can, accordingly, be expressed as a simple single sentence. However, often it is more complex and corresponds to a constellation of propositions, which, in sum, support one's overall stance. For instance, as outlined by Teufel (2014) the overall intent to persuade the scientific community of the work as a valid contribution to science (see above) is, in turn, divisible in subintents related to scientific argumentation, e.g., convincing the audience of the novelty and soundness of the work.

Justification of the Proposition(s). In an ideal argument, a speaker does not convince due to the fact that they are louder or funnier, but because they provide justifications for their propositions, which they link to their stance in the debate. This aspect clearly relates to the notion of rationality discussed above. However, we want to remind the reader that there might be a varying amount of justifications provided and that in the wilderness of real-world arguments, justifications might sometimes be rare. Furthermore, the importance of particular properties of the arguer, e.g., their estimated credibility, plays a non-negligible role in terms of convincingness (see Subsection 2.1.2). In scientific argumentation, the justification of a proposition is often tied to experiments. In case those

[4] This holds even for surveys, opinion works, etc.

experiments were properly conducted according to standards of the specific scientific community, the results obtained are believed to be facts. Based on these facts, certain conclusions can be drawn, which, in turn, are considered tentative knowledge. Another popular way of providing justifications in science is mathematical proofs. Based on certain theorems or axioms, proofs are logical arguments, which show that a certain conclusion is entailed by the assumptions and a (scientifically accepted) set of inference rules. The complexity of providing justifications for claims is also illustrated by different models of argumentation discussed in Subsection 2.1.2.

Reasonable Critic. Stede and Schneider (2018) further highlight the aspect of "a reasonable critic", which relates to two aspects already mentioned before: argumentation is (1) a rational, and (2) a social activity, relating to *the other*. They further highlight the fact that the notion of the reasonable critic is very much dependent on the context of the argumentative situation, e.g., in scientific argumentation, there is typically a scientific audience involved, which, ideally, has specific prior scientific knowledge in the field.

All of these eight aspects highlight the complexity of argumentation, and accordingly, the difficulty of composing and selecting the "right" arguments in a debate. Tindale (2007) characterizes arguments by the fact that they have a "[...] particular structure, where one or more statements (premises) are given in support of a conclusion [...]." This micro-structure of an argument is also reflected more globally: a controversy can be seen as the starting point for a debate, in which opponents present their argumentation, which itself is composed of individual arguments, and finally, argumentative components. The hierarchical relationship between controversies, debates, argumentation, and argumentative components is illustrated in Figure 2.1.

As discussed, controversies can appear in many situations, and as a result, argumentation seems to be almost omnipresent: it occurs from the more formalized Oxford Union-style debates and legal argumentation over political debates broadcast via TV and scientific publications to daily situations in which we like to convince our romantic partners of which place to choose for vacation. Particularly the rise of the Web 2.0 with its online social media platforms leads to an increase of readily available argumentative text: there exist platforms specifically targeted to debating, e.g., the Reddit subforum *Change My View*,[5] and *CreateDebate*,[6] and other platforms more generally designed for exchanging opinions about products and businesses, i.e., online review forums,[7] such as *Yelp*,[8] which all allow users to exchange arguments (we deal with data from three different domains of online argumentative writing in Section 6.2). In consequence, the rising amount of textual argumentative data increases the need for effective and efficient computational analysis of argumentative text, aligned with the overall goal of efficient computational processing in data science. The collection of techniques, which can be applied to this end, fall under the notion of *argument mining* (Lawrence and Reed, 2019).

[5] `https://www.reddit.com/r/changemyview/`

[6] `https://www.createdebate.com/`

[7] Note that Tindale (2007)'s five intentions associated with argumentation include "recommending".

[8] `https://www.yelp.com`

In line with other definitions of the field (e.g., Peldszus and Stede, 2013; Cabrio and Villata, 2018, *inter alia*), Lawrence and Reed (2019) define argument mining (sometimes also referred to as *argumentation mining*) as follows:

Definition 2 (Argument Mining)**.** *"Argument mining is the automatic identification and extraction of the structure of inference and reasoning expressed as arguments presented in natural language."*

Argument mining (AM) can, therefore, be seen as an analysis task, which can help to understand (a) the *stance* of the natural language text, and (b) the *reason(s)* for this particular stance, with the reason(s) being argumentative components combined with an argumentative structure to form an overall argument. Accordingly, AM relates to the field of opinion mining, which deals, more generally, with understanding opinions, i.e., stances. However, as Habernal et al. (2014) point out, "[t]he key question which brings argumentation on the scene is *why do they think so?*", with which the authors highlight the additional dimension that distinguished AM from more general opinion mining or sentiment analysis. This added dimension, however, yields an increase in complexity, requiring advanced language representation techniques.[9] Even more complex, while all of the fields mentioned, i.e., AM, opinion mining, and sentiment analysis, focus on the analytical process of dissecting a stance and its associated justifications, they are part of the more general field of computational argumentation (CA). We define CA as follows:

Definition 3 (Computational Argumentation)**.** *The computational analysis and synthesis of natural language argumentation, based on argumentative reasoning.*

Accordingly, CA includes not only aspects of the pure computational analysis as in "classic" AM, but reaches beyond this subfield by additionally covering other subfields, such as argumentative reasoning as well as argument generation. Next, we introduce argumentation-theoretic aspects dealing with argument structure and AQ. Afterward, the subfields of CA we already touched upon here will be discussed in detail.

2.1.2 The Theory of Arguments

The field of CA is largely based on theoretical studies of argumentation. For this reason, we will now discuss theories related to structural and qualitative aspects, which underpin the computational tasks and models we will later present and evaluate our approaches on (specifically Section 2.1.4, and Chapters 5–7).

Argument Models

Models of argumentation reflect the internal or external structure of arguments, depending on the perspective or granularity (as illustrated in Figure 2.1) applied. Originally described by Aristotle (ca. 350 B.C.E./ translated 1989), the so-called *syllogism* (συλλογισμός) can be seen as the "classic" form of logical arguments in natural language and the prototype of deductive reasoning. It is defined by a major premise, which corresponds to a

[9] We will discuss the challenges in more detail in Chapter 3.

more general statement, a minor premise, which is a specific statement, and a conclusion, which can be deduced from the combination of the major premise and the minor premise. As an example, consider the following famous syllogism:

> All men are mortal (*major premise*)
> Socrates is a man (*minor premise*).
> ---
> Therefore, Socrates is mortal (*conclusion*) $\therefore$

The syllogism, as a traditional argument model, consists of exactly three argumentative components, which have to be composed in a certain scheme in order to form a logical argument. As Corcoran (2003) points out, Aristotle presented the "world's first extant logical system", in which he, crucially, assumes a limited domain of propositions. In combination with a method of deduction, the limited domain of propositions allows him to gaplessly deduce conclusions and assess their validity. While this relates to a *closed* assumption about the world, its states, its actors, and events, modern theories of argumentation extend upon the idea of formal logic by taking an *open* world assumption, thereby emphasizing the uncertainty of real-life situations. Here, arguments are framed as tentative proofs: at any given point in time, an argument can turn out to be invalid in case new information relevant to the argument comes up. For instance, in a scientific argument, it is generally valid to build hypotheses based on observations (inductive reasoning), and new observations can lead to new hypotheses and invalidate former ones.[10] Generally, natural language argumentation is considered to be fuzzy and imprecise: often, the arguments presented are highly dependent on the context, the speaker, the audience, and their relationship. They do not adhere to a clear structure, and certain parts of an argument (compared to the perfect syllogism) are left implicit.[11] As Blair and Johnson (1987) point out, "[...] in most cases, arguments as products of communication in such natural language practices as rational persuasion or rational inquiry are simply not chains of deductive inferences." As a consequence, formal frameworks are difficult to apply to natural language argumentation, which led to the study of *informal logic*. Informal logic "[...] seeks to develop standards, criteria and procedures for the interpretation, evaluation and construction of arguments and argumentation used in natural language" (Blair and Johnson, 1987). Within this field, new models of argumentation emerged. Taking a practical perspective, we highlight that all of the aforementioned points make the computational understanding of natural language arguments extremely challenging.

Bentahar et al. (2010) surveyed and classified existing argument models according to their (a) structure, (b) foundation, and (c) linkage properties and distinguish between (1) rhetorical, (2) dialogical, and (3) monological models. Their conceptual framework is characterized in Table 2.1. Rhetorical models deal with the perception of the arguments by the audience. They therefore focus on how to connect arguments to an overall

[10] This relates to the principle of *falsifiability* in science (Popper, 1935, edition 2002), which we briefly discuss in Section 2.1.3 when we introduce the special case of scientific argumentation.

[11] Arguments with implicit premises fall under the notion of the *enthymeme*, forms of the syllogism initially described by Aristotle (Aristotle, ca. 350 B.C.E./ translated 1989,c).

Type	Structure	Foundation	Linkage
Rhetorical	Rhetorical structure of arguments	Audience's perception of arguments	Connecting arguments in a persuasion structure
Dialogical	Macro-structure of arguments	Defeasible reasoning	Connecting a set of arguments in a dialogical structure
Monological	Micro-structure of arguments	Arguments as tentative proofs	Connecting a set of premises to a claim at the level of each argument

Table 2.1: Conceptual framework of argumentation models, consisting of (1) rhetorical, (2) dialogical, and (3) monological models according to Bentahar et al. (2010).

argumentative structure and highlight persuasive aspects. Dialogical models focus on the interactive aspect of argumentation, i.e., the macro-structure of arguments in the context of debates. In contrast, monological models work on the level of argumentative components to understand and model the micro-structure of arguments.

The Toulmin Model. The Toulmin model of argumentation (Toulmin, 1958, 2003 edition), originally developed for the legal domain, focuses on the notion of practical arguments and the process of justification in contrast to a theoretical and formal view on argumentation. As a monological model of argumentation (Bentahar et al., 2010), it dissects the micro-structure of arguments and defines an argument to consist of six parts: (1) claim, (2) data, (3) warrant, (4) qualifier, (5) rebuttal, and (6) backing (see Figure 2.2).

(1) Claim. A claim corresponds to an argumentative statement in question – an assertion which is put in front of the audience for establishing its merit. It reflects therefore the author's opinion to a controversy. *Example:* (So,) Harry is a British subject.

(2) Data. Data is a fact or evidence, which can serve as a foundation for the claim. It is often also called premise or ground. *Example:* Harry was born in Bermuda.

(3) Warrant. A warrant is a statement that provides the justification for the inference procedure from the data to the claim component. *Example:* (Since,) A man born in Bermuda will generally be a British subject.

(4) Backing. Backing provides additional support for the warrant, for instance, in terms of a reference to a legal document. *Example:* (On account of) The following statutes: ... .

(5) Rebuttal. The rebuttal presents restrictions to the claim, e.g., exceptions in which the argumentative statement does not hold. *Example:* (Unless) Both his parents were aliens.

(6) Qualifier. The qualifier corresponds to the degree to which the arguer believes that the claim holds, e.g., *certainly*, *presumably*, or *most probably*.

We employ an adapted version of the Toulmin Model when studying fine-grained argumentative structures in scientific publications in the context of the complementarity of knowledge across rhetorical analysis tasks in Section 6.1.

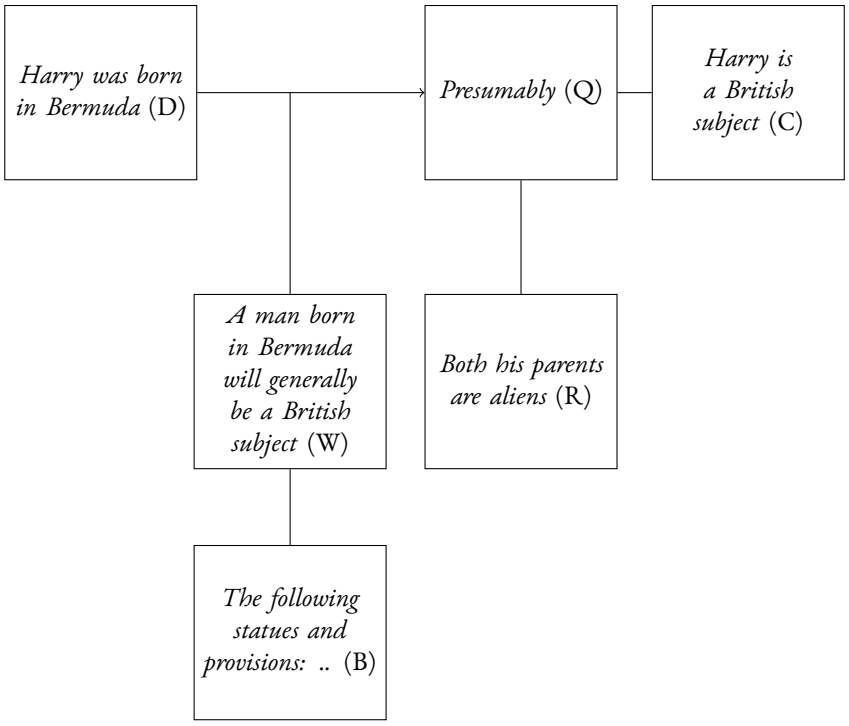


Figure 2.2: The Toulmin model of argumentation with the six argument components (1) claim (C), (2) data (D), (3) warrant (W), (4) qualifier (Q), (5) rebuttal (R), and (6) backing (B), illustrated with Toulmin's original argument example concerning the British citizenship of a human subject (Toulmin, 1958, 2003 edition).

Dung's Model. Contrary to Toulmin's Model (Toulmin, 1958, 2003 edition), Dung's Model (Dung, 1995) belongs to the class of dialogical argumentation models (Bentahar et al., 2010) and focuses on the logical acceptability of arguments, which is, as he outlines, dependent on whether the arguments can be successfully defended against attacking arguments. As such, it is based on the relationship between an agent's own arguments and external arguments, in particular, their *attack* structure. His model allows for the evaluation of the acceptability of arguments based on the notion of defeasible reasoning. The framework is a pair consisting of a set of arguments AR and a binary relation, *attacks*, on AR, i.e.,

$$AF = (AR, attacks)\,, \tag{2.1}$$

with $attacks \subset AR \times AR$. Given two arguments A and B, $attacks(A, B)$ means that A attacks B. For a set of arguments S, if there are no two arguments A and B such that $attacks(A, B)$, S is called *conflict-free*. And finally, based on this notion, Dung (1995) defines an acceptable argument $A \in AR$ with respect to S as *acceptable* iff for each $B \in AR$: if $attacks(B, A)$ then $attacks(S, B)$. In other words, an argument A is acceptable if it can be defended against all attacks. Later, in Section 6.1, we draw

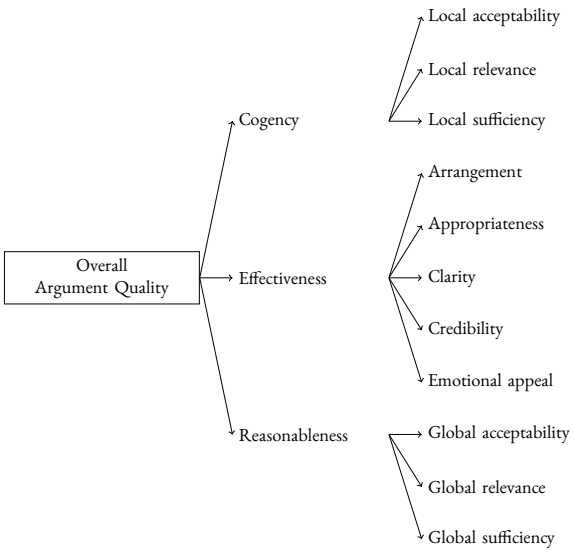


Figure 2.3: Taxonomy of theory-based argument quality (Wachsmuth et al., 2017b).

some inspiration from Dung's model in order to reflect the hybrid nature of scientific publications: being monological documents, they still exhibit dialogical aspects as they engage with the overall scientific discourse. As such, they might restate, and *attack* arguments from other authors (linked via citations). We account for this by adding relationships between individual claims in our adapted Toulmin model when studying fine-grained argumentative structures in scientific literature.

Argument Quality

The quality of argumentation can be assessed according to different perspectives. For instance, in his *On Rhetoric*, Aristotle (ca. 350 B.C.E./ translated 2006) describes three technical means of persuasion, which characterize the quality of an argument:

Logos (λόγος). The argument itself has to be logical in order to be reasonable. A logical argument can, according to Aristotle, take two forms: it can either be *deductive* or *inductive*, which still builds the basis for scientific reasoning.

Ethos (ἦθος). In contrast, ethos is grounded in the arguer: the speaker, as a person, has to display (a) practical intelligence, (b) a virtuous character, and (c) goodwill in order to appear credible to the audience, especially when there is room for doubt.

Pathos (πάθος). Pathos deals with the emotional state of the audience in relation to the argument. The argument should be presented in a way that evokes emotions in the audience, which are beneficial for making it judged in the desired way.

In CA, researchers have often focused on specific practical conceptualizations of argumentative quality, for instance on *clarity* (Persing and Ng, 2015). Later on, Wachsmuth et al. (2017a) proposed a taxonomy of argumentation quality (AQ) by synthesizing AQ theories and mapping those to approaches discussed in computational argumentation. The taxonomy is depicted in Figure 2.3. It defines *overall AQ* as being composed of three sub-dimensions (Cogency, Effectiveness, and Reasonableness), each of which is, in turn, composed of a series of quality-related subaspects:

Cogency. Cogency relates to the logical aspects of argument quality. High cogency indicates that an argument's premises are acceptable, as well as relevant, and sufficient with respect to the argument's conclusion.

Effectiveness. Effectiveness reflects the persuasive power of how an argument is stated and is thereby tied to the rhetorical aspects of argumentative quality. Important aspects of an effective argument include its arrangement, clarity, appropriateness in a given context, emotional appeal, and the author's credibility.

Reasonableness. Reasonableness indicates the quality of an argument in the context of a debate and thereby relates to dialectical AQ, i.e., its relevance, its acceptability, and the way it is stated as a whole, and its sufficiency toward the resolution of the issue.

Mapping this taxonomy to the technical means of persuasion defined by Aristotle, Cogency represents *logos*, and the Effectiveness, as the rhetorical dimension, reflects aspects of *ethos* and *pathos*. While each of the dimensions represents a separate series of aspects of the argumentative quality of texts, they are interrelated, and all contribute to overall AQ. We employ the taxonomy in Section 6.2 when we study how to exploit the complementarity of knowledge across CA tasks within language representations.

The outlined qualitative dimensions are generally present in and can be assessed across all argumentative domains of text, but they can be pronounced with a varying degree depending on the argumentative context. For instance, an argument presented in a business review forum might describe rather subjective experiences and put an emphasis on the emotional appeal of the argument (see Section 6.2). Similarly, some structural argument models are more directly applicable in specific domains of text than in others. For instance, recall that Toulmin's model was specifically developed for legal arguments. As an interesting case of argumentation, we next discuss the special case of scientific arguments. We choose this domain in order to study domain-specificity of language representations for CA (see Section 3.2 and Chapter 5), and in the context of the complementarity of knowledge across tasks (given that scientific argumentation can be seen as being composed of multiple interrelated rhetorical layers, see Sections 3.3 and 6.1).

2.1.3 The Special Case of Scientific Argumentation

As already discussed, argumentation is nearly ubiquitous in our lives and can be found in many domains of text, such as news editorials, student essays, and online debate forums. In particular, in significant portions of this thesis, we will be focusing on scientific argumentation. According to Weinstein (1990) "[...] almost all in science includes argumentation [...]", with which he is referring to the epistemological nature of scientific work and the central role of the falsifiability of scientific claims (Popper, 1935, edition 2002), which is, in turn, in line with the idea of defeasible reasoning as a central notion in modern argumentation theories (see also Section 2.1.2). Popper argues that scientific knowledge is provisional: scientific hypotheses can be seen as tentative proofs at a certain point in time and with a certain amount of information available, which should be testable, and, ultimately, falsifiable. Consequently, this allows for controversies to arise in the scientific community, the starting point for debates (as discussed in Section 2.1.1). Weinstein acknowledges that "[...] much in science includes explicit argumentation". This means, instead of only being tacitly inherent to scientific reasoning processes, in order to resolve arising controversies, part of the epistemic process is to externalize, i.e., verbalize, scientific argumentation. Here, the, arguably, most prominent externalization form is the scientific paper. In these publications, we try to convince the scientific audience of the validity and merit of our work, of accepting our findings, and, ultimately, of our work as a valid contribution to science (Teufel, 2014). Accordingly, argumentation can be seen as a key feature in scientific writing (Green, 2015a). In Section 6.1, we analyze the role of fine-grained argumentative structures in the rhetorical analysis of scientific literature, and in Chapter 5, we analyze citations as a central argumentative tool in scientific writing.

The motivation for focusing on scientific writing as one particular domain of argumentation in this thesis is twofold: (1) efficient computational analysis of scientific publications is needed for ensuring efficient access to scientific knowledge, and (2) scientific argumentation is particularly challenging to analyze.

(1) First, the exponential growth in the number of scientific publications (Bornmann and Mutz, 2015) raises the need for effective and efficient computational analysis tools of the large body of research work. As we are experiencing now in the face of the COVID-19 pandemic, scientific information access is, especially in situations which require fast governmental decisions, crucial to crisis response and societal welfare.[12] Here, as outlined by Green (2015a), it is important to understand argumentation for three main reasons: (a) argumentation provides the critical context within which we should interpret the text, (b) it can be beneficial for downstream applications, as shown for summarization (Cohan and Goharian, 2015; Abu-Jbara and Radev, 2011, *inter alia*), and (c) arguments within scientific literature are tied to the scientific discourse; therefore, the global relationship of scientific claims which reflects the state of knowledge in a field of research can be understood via the analysis of argumentation in scientific writing.

(2) Secondly, the challenging nature (Green, 2017) and the unique features of scientific in contrast to "ordinary" argumentation make the task particularly interesting:

[12] This is reflected within the scientific community: in the first 4 months after the first confirmed COVID-19 case, 16,000 related scientific papers were published (Fraser et al., 2021).

scientific reasoning and scientific argumentation are generally recognized as complex processes (Kuhn et al., 2000), which require demanding epistemic reasoning, such as hypothesizing and evaluating evidence, and is therefore acknowledged to be difficult to acquire (Klahr and Dunbar, 1988; Osborne, 2010). Furthermore, scientific argumentation is framed within a complex network of previous results and commitments, such as already accepted claims in the field, as well as stylized practices (Weinstein, 1990), and is highly ritualized (Latour and Woolgar, 1987). The conditions upon which scientific argumentation is placed are therefore more convoluted than in other fields of argumentation (Weinstein, 1990). These conditions manifest as well in scientific writing: for instance, scientific publications typically follow a community-established structure and use a certain terminology, as well as certain rhetorical moves, which underpin the higher-level argumentative intentions (Teufel, 2014). Further, in order to get their work accepted within a peer review process, researchers need to show that they are aware of the latest developments in their field with sufficient and up-to-date citations, and each scientific field has certain accepted ways of referring to those works, for instance, by providing a "related work" section and adhering to a certain citation style. As Gilbert (1977) notes, referencing can be seen as persuasion. By referencing other works and, afterward, being referenced by others, researchers respond to previous scientific claims and thereby also connect their own work to and position it within the overall scientific discourse (see Chapter 5). Due to the use of citations, scientific publications exhibit a hybrid nature: they are monological arguments, which are placed within and connected to a dialogical debate. As a result, in order to be able to present a scientific argument sufficiently and position it with respect to previous works and concerning a certain research field and problem, scientific publications are typically long and complex documents, which makes understanding argumentation difficult (Kirschner et al., 2015). Looking at the micro-level, i.e., the level of individual arguments within the course of the document, we note that argumentative components are not necessarily expressed in adjacent phrasal units (Green, 2017) and the content of several arguments may be interleaved at the text level (Green, 2016). Furthermore, some of the argumentative components may be left implicit, resulting in enthymemes, i.e., arguments with implicit components (Green, 2017). Finally, all the aspects mentioned above, e.g., adhering to overall community-established styles, such as the structure, referencing others, and positioning the work within the discourse, as well as building up a finer-grained argumentative structure work together in forming a convincing argument and persuade other members of the scientific community of the proposed contribution to science. This makes CA for scientific documents extremely challenging. Due to all these reasons, this thesis focuses on the computational understanding of scientific text as a challenging and interesting case study of argumentation.

In the next Subsection, we introduce argumentation from the perspective of NLU, and, accordingly, highlight the most prominent CA tasks. We hereby also discuss specific tasks related to the argumentative analysis of scientific text.

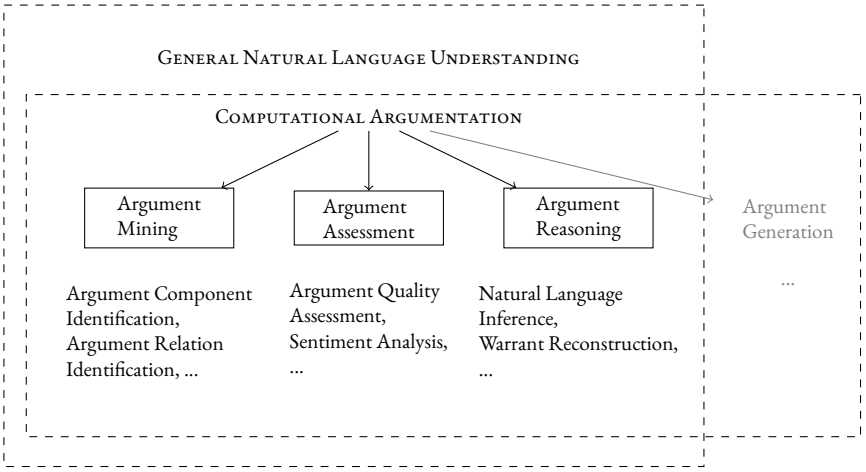


Figure 2.4: Computational Argumentation with its four subfields (1) Argument Mining, (2) Argument Assessment, (3) Argument Reasoning, and (4) Argument Generation and its relation to General Natural Language Understanding.

2.1.4 Argumentation and Natural Language Understanding

The field of computational argumentation can be subdivided into four subfields, each of which corresponds to a collection of concrete NLP tasks: (1) argument mining, (2) argument assessment, (3) argument reasoning, and (4) argument generation. While the first three groups of tasks cover the *analysis* and *understanding* of argumentation, the last subfield, argument generation, relates to the *synthesis* of arguments. As the present thesis focuses on the understanding of arguments, we will not further cover argument generation and instead discuss the other three aspects, which are grounded within general natural language understanding (GNLU). Interestingly, GNLU, a model's general ability to understand natural language, can be seen as a necessary prerequisite for, arguably, most of the CA tasks. For some, there is even a direct correspondence, with, for instance, natural language inference directly reflecting argumentative reasoning capabilities (Moens, 2018).[13] But, as Moens (2018) states, argumentative understanding "[...] puts an extra dimension to the language understanding process." While some of the tasks are domain-independent in the sense that they are relevant and similarly formulated tasks for many domains of argumentative text, some relate to certain domains only. Here, specifically the tasks relating to scientific argumentation stand out (as explained before, see Section 2.1.3), which we will cover at the end of this Subsection.

Argument Mining (AM). As discussed before (see Definition 2), AM deals with the identification and extraction of structural aspects of arguments with, as Lippi and Torroni

[13] An overview of correspondences between argumentative understanding tasks and more established NLP tasks is also given by Lippi and Torroni (2016b).

(2016b) argue, the main goal to provide structured input data for other systems, such as reasoning engines. As discussed in Section 2.1.2, there exist a range of theoretical argumentation models in the literature, e.g., Toulmin's Model (Toulmin, 1958, 2003 edition), and depending on the model chosen, the structure of an argument is considered to consist of a different set of argumentative components and relationships. In consequence, the prediction space of the task is defined according to the argument model chosen. For instance, we can simply distinguish between *claims* and *premises* with premises *supporting* the claims, and those component and relation types then correspond to the labels we can assign to extracted portions of text and their relationships. The full structural analysis, i.e., the AM pipeline, can be broken down into the following subtasks (Lippi and Torroni, 2015): (1) argument detection (e.g., Moens et al., 2007), (2) argument component identification (e.g., Morio and Fujita, 2019), and (3) argument structure prediction (e.g., Galassi et al., 2018). We will now explain each subtask by means of the example: *"Our method is superior to previously proposed ones, because it requires less data to perform similarly well."*

(1) Argument Detection. The idea is to identify argumentative portions of text, which is typically handled as a text classification task. Given a span of text x, the task is then to assign one of the labels $y^{(i)}$ out of the set of labels $Y = \{\text{argumentative}, \text{non-argumentative}\}$.

"Our method is superior to previously proposed ones, because it requires less data to perform similarly well." $\longrightarrow$ argumentative

The granularity of the text can differ, e.g., Lippi and Torroni (2016b) describe a sentence-level variant, but it can also be cast as a classification of larger or smaller text portions.

(2) Argument Component Identification. In argument component identification, which is also known as argument component boundary detection (Lippi and Torroni, 2016a), the task is to identify the different argumentative components according to the argument model chosen, e.g., claims and premises, in an argumentative text.

"Our method is superior to previously proposed ones [claim], because it requires less data to perform similarly well [premise]."

The task is typically cast as a sequence labeling task. More formally, given a sequence of tokens $t_1, ..., t_n$ assign to each token a label $y^{(i)}$ out of the set of argumentative component labels Y usually based on a begin–inside–outside (B–I–O) labeling scheme, e.g., $Y = \{\text{begin\_claim}, \text{inside\_claim}, ...\}$. The B–I–O format is also common in other sequence labeling tasks, such as named entity recognition (NER).

(3) Argument Structure Prediction. In argument structure prediction, the task is to predict the overall argument structure, which typically amounts to predicting the relationships between two or more (1) arguments or (2) argument components (as discussed before). The set of possible relations is again grounded in the argument model applied. For instance, it can consist of *supports* and *contradicts/ attacks* relationships.

"Our method is superior to previously proposed ones [claim]" $\xleftarrow{supports}$ *"it requires less data to perform similarly well [premise]."*

We will now turn our attention to the next subfield, argument assessment.

Argument Assessment. Argument assessment refers to the assessment of certain properties of an argument, such as its stance (Wojatzki and Zesch, 2016), sentiment (Wachsmuth et al., 2015), and quality (Wachsmuth et al., 2017b). The range of possible properties one might be interested in is large and depends, naturally, on the goal of the final application. Here, we focus on AQ prediction, as (a) it encompasses a big pool of concrete argument assessment tasks that have been tackled in the CA community (e.g., Wachsmuth et al., 2017a,b; Habernal and Gurevych, 2016, *inter alia*), and (b) later on, we study the complementarity of knowledge across argumentative quality dimensions (see Section 6.2).

Argument Quality Assessment. The task of scoring an argument according to its quality was tackled in many different conceptualizations, e.g., as *clarity* (e.g., Persing and Ng, 2013), and *prompt adherence* (Persing and Ng, 2014). Most often, it is casted as a regression task (e.g., Persing and Ng, 2015; Persing et al., 2010), in which given an argumentative text $x^{(i)}$, the task is to predict a score $y \in \mathbb{R}$, which reflects the quality of the argument according to some quality aspect chosen. Sometimes, it has also been casted as a pairwise classification task: given a pair of arguments $(x_1, x_2)^{(i)}$, decide whether x_1 is preferrable over x_2 or vice versa (e.g., Gretz et al., 2020). Similar to the argument mining tasks, the prediction of the argumentative quality of a text can be based on an underlying theoretical framework, for instance, the taxonomy of AQ presented by Wachsmuth et al. (2017b). In this case, the underlying theory determines the dimensions and concrete properties for the manual or computational annotation of the texts, e.g., the assessment of an argumentative text according to logical, rhetorical, and dialectical aspects (see Section 2.1.2).

Argument Reasoning. Argument reasoning is the task of reasoning over arguments. In NLP, there are currently two popular flavors of this task: (1) natural language inference and (2) argument reasoning comprehension.

(1) Natural Language Inference (NLI). Also known as recognizing textual entailment (Giampiccolo et al., 2007), NLI reflects general argumentative reasoning capabilities (Moens, 2018). Given a *premise* p and a *hypothesis* h, the task is to identify whether p entails h, i.e., whether h can be inferred from p. The set of possible labels depends on the data set. In many cases there are three classes: *entailment*, *contradiction*, and *neutral*. Example:

Premise	*A man in an orange vest leans over a pickup truck.*
Hypothesis	*A man is touching a truck.*
Label	*Entailment*

This example, which we have drawn from the Stanford Natural Language Inference (SNLI; Bowman et al., 2015) corpus, illustrates that though plain NLI does not deal with canonical arguments, the task is designed for testing precise reasoning capabilities and advanced knowledge as required in argumentation. For instance, the models have to know that "*leaning over a truck*" implies "*touching the truck*". We explain and address this challenge in Sections 3.1 and 4, and further employ the task for our evaluation in the context of multilinguality (see Chapter 7). As a more challenging extension of the plain NLI task, Camburu et al. (2018) proposed the e-SNLI task based on the SNLI corpus, in which the models additionally have to argue for their inference decision by providing explanations.

(2) Argument Reasoning Comprehension. Proposed by Habernal et al. (2018), argument reasoning comprehension can be seen as another variant of NLI, in which the task is to explain why a claim follows from its premises, similar to e-SNLI (Camburu et al., 2018) discussed above. However, in contrast to e-SNLI, argument reasoning comprehension relates specifically to the Toulmin Model (Toulmin, 1958, 2003 edition): the task is to reconstruct and analyze *warrants*, which are often left implicit (see Section 2.1.2).

Relations between CA Subfields. The four CA fields are related and may depend on each other depending on the application scenario. As a simple example, consider the task of identifying "good" arguments given a topic and a stance. The first step could consist of extracting arguments in a collection of argumentative texts relating to this topic (*argument mining*). As a next step, one could assess the stance of these arguments as well as their quality (*argument assessment*). And as a final step, one could filter the arguments according to the given stance and retrieve the top k arguments ranked based on the quality score assigned. The final result is then dependent on the output of each of the pipelined models. This example falls under *argument retrieval* (e.g., Wachsmuth et al., 2017c) an adaptation of standard information retrieval for the case of arguments.

In this Subsection, we have so far focused on the most important "standard" CA tasks, but as in the example above, we also note that there are more, rather application-specific variants, which we do not explicitly cover. We now turn our attention to domain-specific tasks dealing with the argumentative analysis of scientific text.

Analyzing Scientific Argumentation: Scitorics. As outlined in Section 2.1.3, scientific argumentation is interesting but also challenging to analyze. While, theoretically, all tasks discussed before can be transferred to the case of scientific text, directly transferring standard task formulations is difficult due to the convoluted and highly stylized nature of scientific argumentation. Accordingly, tasks tailored to analyzing the *rhetorical aspects* of scientific writing, dubbed *scitorics*, emerged. Here, we briefly overview the most traditional analysis tasks: (1) argumentative zoning and (2) citation context analysis.

(1) Argumentative Zoning. As the first task in NLP, which relates to the argumentative structure of scientific publications, Teufel et al. (1999) proposed argumentative zoning. Based on the idea that a scientific paper follows a certain, community-established discourse structure, the task is, given a sentence $x^{(i)}$, to assign a discourse role out the set of possible sentential discourse roles to it, such as *motivation*, *method*, and *result*. Example:

"Our work results in a new corpus for argumentative discourse analysis." $\longrightarrow$ result

The sentence clearly states the outcome of the authors' work, and is labeled as *result*, accordingly. The set of argumentative discourse roles, i.e., labels, varies across the works (e.g., Teufel et al., 1999; Ronzano and Saggion, 2015).

(2) Citation Context Analysis. Based on the role of citations as "tools of persuasion" (Gilbert, 1977), the analysis of citation contexts, which are, in the case of sentential contexts, known as *citances* (Nakov et al., 2004), is seen as an important task in the argumentative analysis of scientific texts. Accordingly, different notions of the task have been proposed. For

instance, previous research deals with the *extraction of citation contexts* (Jha et al., 2016), with predicting the *citation polarity* or *sentiment* (Athar, 2011; Abu-Jbara et al., 2013), and with classifying the *citation purpose* or *citation intent* (Cohan et al., 2019). All these notions relate to the idea of understanding the citer's motivation. Example:

"We use the tool of Author (Year), as it has shown to perform best in our prestudy." $\longrightarrow$ Polarity: positive, Intent: use

In this example sentence, the citation marker *Author (Year)* references a previous work, which is cited in a *positive* way, and the citing publication *uses* an artifact of this research. As with the task of argumentative zoning, different labeling schemes have been proposed. We deal with citation polarity and citation purpose classification in Chapter 5.

After having discussed the field of Computational Argumentation, we now introduce the second main topic of this thesis: representation learning.

2.2 Representation Learning

The second main topic of this work, *representation learning*, is a fundamental area in NLP. After providing a brief introduction to machine learning basics (Subsection 2.2.1), we outline methods for inducing semantic representations of text. Next, we discuss transfer learning (Subsection 2.2.3), a learning paradigm underlying the introduced representation models. We exploit several types of transfer learning, e.g., cross-lingual learning, in the course of this thesis. Finally, we look at the topic of bias, a fundamental concept in both human and machine learning, with a focus on the ethical issue of encoding unfair stereotypical bias in language representations (Subsection 2.2.4).

2.2.1 Machine Learning

In machine learning, the general idea is to learn a computational model from data. According to Mitchell (1997), *learning* in this context means that for given a task $\mathcal{T}$ with an associated performance measure P, the performance of a machine on $\mathcal{T}$ measured by P improves with experience, i.e., with the number of examples it has seen.

We broadly distinguish *unsupervised*, *supervised*, and *self-supervised* machine learning. In (1) unsupervised machine learning, no supervision signal is given. For example, the task $\mathcal{T}$ can be to cluster a set of text documents, and the performance measure P can be some measure of the intra-cluster similarity. In (2) supervised machine learning, human supervision is part of the training process. This means that during the training process, for each input example $\mathbf{x} \in X$, where X is the set of training examples, a label y is given. For instance, in AQ assessment (see Subsection 2.1.4), the task can be to associate a real number $y \in \mathbb{R}$ with a vectorized input text $\mathbf{x}$ indicating the overall AQ and the performance measure P can be the squared difference $(\hat{y} - y)^2$ between the predicted scores $\hat{y}$ and the true scores y. More precisely, the goal of supervised machine learning is the following: given an input domain $\mathcal{D}$, consisting of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ with $\mathbf{x} = \{x_1, ..., x_n\} \in \mathcal{X}$, as well as a *task* $\mathcal{T}$ with a label space $\mathcal{Y}$, prior distribution over the labels $P(Y)$, and a

conditional probability distribution $P(Y|X)$, learn a function $f(\cdot)$, which maps $\mathcal{X}$ to $\mathcal{Y}$ approximating $P(Y|X)$, i.e., $f(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$. We search for $f(\cdot)$ in the search space Ω. Related to this, the notion of a domain is discussed in Subsections 2.2.3 and 3.2 in more detail. In this thesis, we evaluate our approaches on many supervised CA tasks. Finally, in (3) self-supervised learning, the model is exploiting supervision signals that are not explicitly given by humans but are inherently part of the input data. For instance, in some of the language representation models which we will introduce next, the models' goal is to learn to predict a word given its context words.

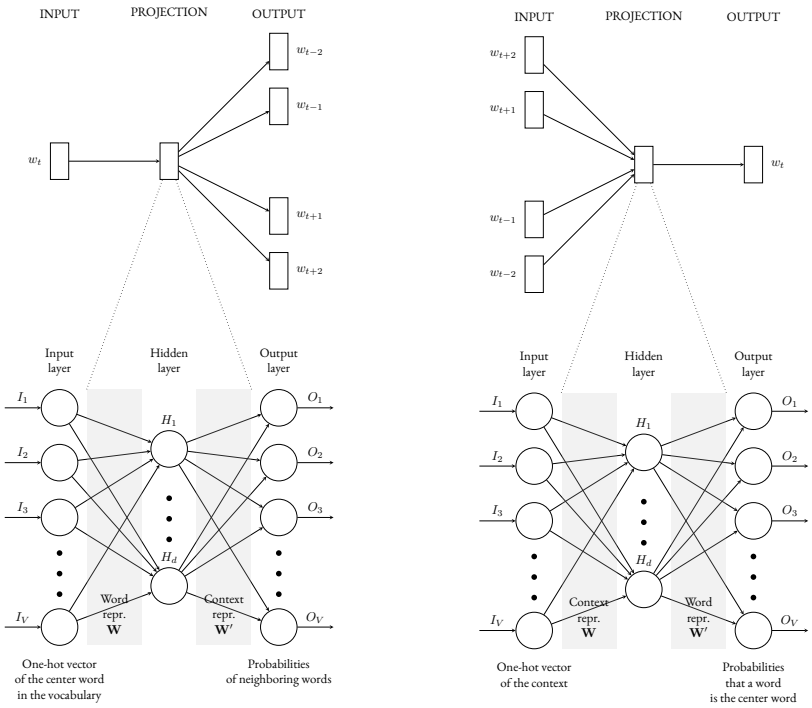
2.2.2 Language Representation Methods

As for all tasks in NLP, also for CA, the textual input has to be represented numerically in order to be processed by computational models. This Section introduces dense representations of words or subwords, i.e., *embeddings*, which are used in the research work covered in this thesis. Generally, the techniques can be broken down into (1) static word embeddings and (2) contextualized word embeddings. Both embedding techniques are based on the so-called *distributional hypothesis* (Harris, 1954), which underpins the field of distributional semantics and was captured by Firth (1957) in the popular quote: *"You shall know a word by the company it keeps"*. The general idea is that the meaning of a term can be explained by other terms it typically appears with, i.e., its typical context. This idea can then be exploited in unsupervised or self-supervised learning scenarios (as discussed in the previous Section) to obtain numeric representations. Most often, the representations are induced on large collections of text, e.g., Wikipedia, and afterward adjusted to specific tasks. This procedure is referred to as *pretrain then fine-tune* paradigm, an example of *transfer learning* (see Subsection 2.2.3). The intuition behind this is that the models can acquire general language understanding capabilities before being fine-tuned to accentuate specific phenomena that are important features for a particular downstream task.

Static Word Embeddings

Static embeddings assign to each token t, which in most cases corresponds to a single word, a static numerical representation, i.e., a vector of real numbers $\mathbf{e}$, which represents its meaning. In static embedding spaces, often also referred to as *distributional word vector spaces*, the vector representation does not change depending on the context in which the token appears. Popular algorithms for inducing these representations inlcude Word2Vec (Mikolov et al., 2013c), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017), which will be discussed in the following paragraphs.

Word2Vec. With Word2Vec, Mikolov et al. (2013a) presented two word embedding methods, which are both based on a simple single-layer feed-forward neural network (see Figures 2.5a and 2.5b): skipGram, and continuous bag of words (CBOW), which can be distinguished by their model architectures and respective training objectives.



(a) The SKIPGRAM model architecture. (b) The CBOW model architecture.

Figure 2.5: The difference between the SKIPGRAM and CBOW model architectures (upper parts from (Mikolov et al., 2013a)).

SKIPGRAM. Given a sequence of words $w_{t-n}, ..., w_t, ...w_{t+n}$, with a center word w_t, the task of the model is to predict its surrounding tokens, i.e., the context words $W_t = \{w_{t-n}, ..., w_t, ...w_{t+n}\} \setminus \{w_t\}$. This can be expressed by maximizing the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-n<j<n;j!=0} \log P(w_{t+j}|w_t)\,, \tag{2.2}$$

with T as the total number of terms in the sequence and n as the context size before and after the center word. While the probability $P(w_{t+j}|w_t)$ was originally defined using the *softmax* function, i.e.,

$$P(w_{t+j}|w_t) = \frac{\exp(\mathbf{x'_{t+1}}^{\top}\mathbf{x_t})}{\sum_{i=1}^{|V|}\exp(\mathbf{x'_i}^{\top}\mathbf{x_t})}\,, \tag{2.3}$$

with the vocabulary V, and $\mathbf{x_i}$, and $\mathbf{x'_i}$, as the word and context embeddings, i.e., rows and columns of weight matrices $\mathbf{W}$ and $\mathbf{W}'$, respectively (see Figure 2.5a), Mikolov et al. (2013c) proposed a more efficient softmax approximation based on *negative sampling*, a simplification of noise contrastive estimation (Gutmann and Hyvärinen, 2012). Instead

of predicting the probabilities over the whole vocabulary, we only predict over a subset of the vocabulary, which includes the true context word, as well as randomly sampled negative examples. This approximation reduces the computational complexity originally arising from computing the probabilities over the whole vocabulary.

CBOW. In contrast to SKIPGRAM, given a sequence of words $w_{t-n}, ..., w_t, ...w_{t+n}$, with a center word w_t and its set of context words $W_t = \{w_{t-n}, ..., w_t, ...w_{t+n}\} \setminus \{w_t\}$, the goal of the CBOW architecture is to learn representations, which are optimized for predicting the center word w_t based on its context W_t. This can be expressed via the following loss function:

$$L_{\text{CBOW}} = -\log P(w_t|W_t)\,, \tag{2.4}$$

with $P(w_t|W_t)$ as the probabability of w_t being the center word, conditioned on W_t.

GLOVE. While the algorithms in WORD2VEC induce dense word representations based on local contexts and backpropagation, Pennington et al. (2014) proposed an analytical approach to obtain word vectors based on global corpus statistics, similar to simpler co-occurence-based representations. However, instead of resorting to direct co-occurence probabilities, given two words w_i and w_j, the main intuition is that their ratio of co-occurrence probabilities with a third term, the probe word w_k, expressed as $\frac{P(w_i,w_k)}{P(w_j,w_k)}$ is for their semantic relationship more indicative than the direct co-occurence probability.

FASTTEXT. Based on the observation that word embedding models, which associate each word with a single vector, ignore the internal (morphological) structure of words, Bojanowski et al. (2017) proposed FASTTEXT. The method is based on the original SKIPGRAM architecture but designed to capture subword information by representing words as bags of character n-grams. Each character n-gram is linked to a distinct vector representation, and a word vector, in turn, is defined as the sum of its character n-gram vectors. As a result, even for rare words, a reliable representation can be learned.

We employ static word embedding spaces for the semantic characterization of citations in Chapter 5, and for the rhetorical analysis of scientific argumentation in Section 6.1. Furthermore, we analyze unfair stereotypical biases encoded in those representations in Chapter 8. We next discuss their successors: contextualized word embeddings.

Contextualized Word Embeddings

While static word embeddings associate each token with a single vector, i.e., a *static* representation, contextualized embedding models assign a representation to a token based on its context. The state-of-the-art in nowadays pretrained language models is based on the so-called *transformer* architecture, in particular, on its encoder, proposed by Vaswani et al. (2017). The transformer encoder consists of n identical layers, which each consist of a *self-attention* and *feed-forward network* sublayer. The *self-attention* typically corresponds to the scaled dot product-attention. Here, given a matrix of input representations, three attention matrices are created, in which each row corresponds to

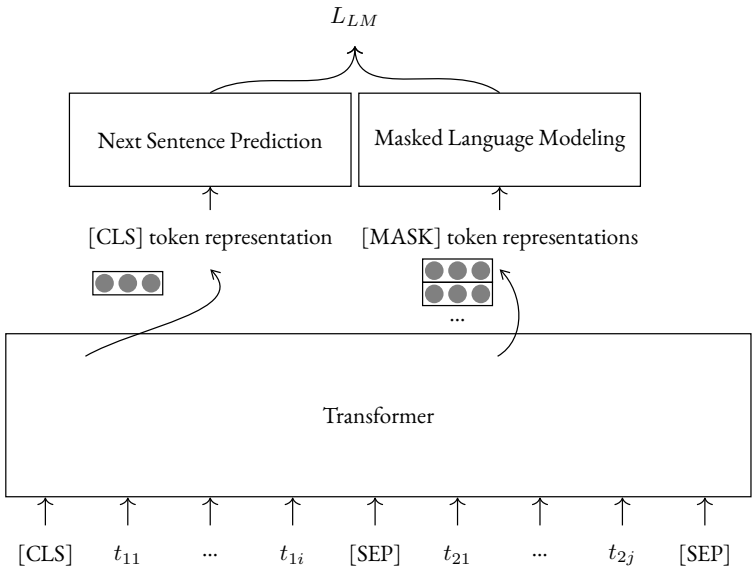


Figure 2.6: The BERT model architecture with its two pretraining objectives: (1) Next sentence prediction, and (2) Masked language modeling.

an input token: the query matrix $\mathbf{Q}$, the key matrix $\mathbf{K}$, and the value matrix $\mathbf{V}$. The self-attention is then computed as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\mathbf{V}) \,, \tag{2.5}$$

with the scaling factor $\frac{1}{\sqrt{d_k}}$, which impedes gradient underflow. To allow to jointly attend to information from different representation subspaces the model employs *multi-head attention*, i.e., h attention layers (= heads) are run in parallel. The outputs of each of the heads are concatenated and projected using the output weight matrix $\mathbf{W^O}$:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\mathbf{W^O} \,, \tag{2.6}$$

with $\text{head}_i = \text{Attention}(\mathbf{QW_i^Q}, \mathbf{KW_i^K}, \mathbf{VW_i^V})$.

Pretrained language models based on this encoder architecture are, among others, BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019).

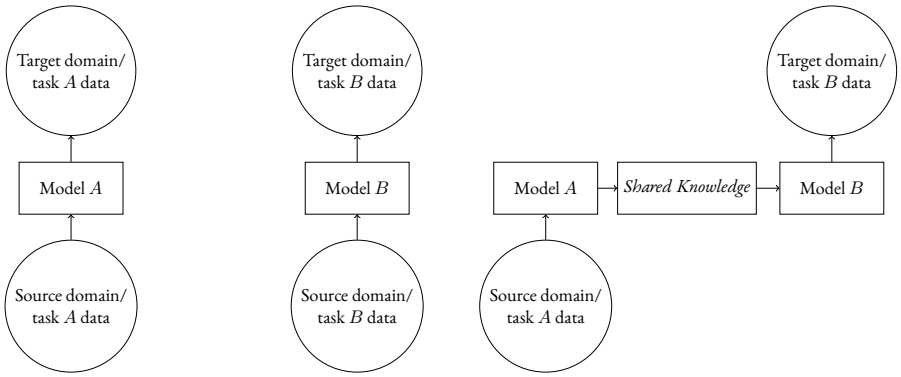
Bidirectional Encoder Representations from Transformers (BERT). The core of the BERT model (Devlin et al., 2019) is a multi-layer bidirectional transformer encoder (Vaswani et al., 2017) as explained above. It is pretrained using two objectives: masked language modeling (MLM) and next sentence prediction (NSP). (1) MLM is a token-level prediction task, also referred to as *Cloze* task (Taylor, 1953): among the input

data, a certain percentage of tokens is masked out and needs to be recovered. (2) In contrast, NSP operates on the sentence-level and can be seen as a higher-level sequence modeling task that captures information across sentences. NSP predicts if two given sentences are adjacent in text (negative examples are created by randomly pairing sentences). For representing the input, BERT uses WordPiece embeddings (Wu et al., 2016) with a 30,000 token vocabulary, as well as position and segment embeddings. As different sentences are concatenated together to a single input sequence (e.g., for the NSP task in the pretraining), the segment embeddings help the model to differentiate the different parts of the input. For each token, the input representation in the first layer of the model corresponds to the sum over its token, segment, and position embeddings. Furthermore, BERT uses two types of special tokens: the separator token (`[SEP]`) and the sequence start token (`[CLS]`). The separator tokens are an additional way of indicating different parts of the input sequence. The sequence start token is used as a representation of the whole input sequence and, accordingly, its final hidden state is used as input for sequence classification tasks, e.g., NSP. The input data used to pretrain the original model consists of a concatenation of the BooksCorpus (800M words; Zhu et al., 2015) and the English Wikipedia (2,500M words). Figure 2.6 illustrates BERT's pretraining framework.

In contrast to static word embeddings, where only a single matrix consisting of word vectors needs to be transferred in order to fine-tune these representations on a downstream task, the authors propose to fine-tune all of BERT's encoder layers. To this end, the input needs to be prepared in a "BERT-compatible" format and a prediction head, which is appropriate for the particular fine-tuning task at hand, needs to be placed on top of the encoder. For example, for NLI (see Subsection 2.1.4), the premise and the hypothesis are first tokenized. Then, the tokens are concatenated, and the special separator and sequence start tokens are added in between and in front of the sequence, respectively. After piping the input through the model, the transformed representation of the sequence start token can then be fed into a simple softmax classifier for predicting the entailment relationship.

Robustly Optimized BERT Approach (RoBERTa). RoBERTa (Liu et al., 2019) is a robustly optimized BERT model, for which the authors reevaluated different configurations and design choices. In particular, two main findings are incorporated in RoBERTa: dynamic masking, and removal of the NSP loss. (1) In contrast to masking a certain percentage of tokens of the input data for the Cloze task in a preprocessing stage only once, thereby producing a single static mask across all training epochs, for each input sequence, a mask is dynamically generated. This is especially effective when pretraining on larger data sets or for more epochs. (2) Furthermore, the model does not rely on the NSP loss, i.e., it does not explicitly model relationships between sentences. For both changes, i.e., using dynamic instead of static masking, and removing the NSP objective, the authors empirically demonstrate performance improvements. Additionally, RoBERTa is trained in larger batches and with a larger vocabulary on more data.

We employ contextualized embeddings in Chapters 4, and 7, and in Section 6.2.



(a) Traditional machine learning. (b) Transfer Learning.

Figure 2.7: The difference between (a) the traditional machine learning setup and (b) the transfer learning scenario: in transfer learning, a portion of knowledge from source domain or source task A is reused for the target domain or target task B.

2.2.3 Transfer Learning

In an ideal supervised machine learning setup (discussed in Subsection 2.2.1), the training data and the test data originate from the same feature space and the same distribution. Similarly, the prediction task, defined by its label space and its objective predictive function, stays the same between the training and test scenario. However, in reality, across all NLP and CA tasks, this is often not the case: not for all of the world's around 7,000 languages (Eberhard et al., 2020), such as Urdu or Swahili, not for all possible domains of text, such as scientific writing or online debates, and not for all imaginable tasks annotated data is available. In fact, it seems impossible to ever reach complete coverage. To alleviate this problem termed *data scarcity* or framed as *low-resource scenario*, researchers in NLP have been working on making effective use of the data that is already available, even in the case of mismatches in data distribution and mismatches in the nature of the task between training and inference time. The general idea is to provide mechanisms, which allow for transferring previously acquired knowledge such that new problems can be solved faster or better. This is aligned with the human way of problem-solving: someone, who learns to play the guitar from scratch, typically also acquires knowledge about music theory, which then can be transferred to learning a new instrument, e.g., piano, in which they then might exhibit a steeper learning curve. This is because there is a portion of shared knowledge involved, which does not need to be learned from scratch but can be transferred. Before, we have already discussed examples of this paradigm, which we call *transfer learning*: the semantic language representations introduced in Subsection 2.2.2 acquire general knowledge on large corpora in the pretraining stage, which is then reused on a particular task in the fine-tuning stage. The difference between traditional machine learning scenarios and transfer learning is depicted in Figure 2.7.

Transfer learning has been researched in artificial intelligence since the 70s (Bozi-

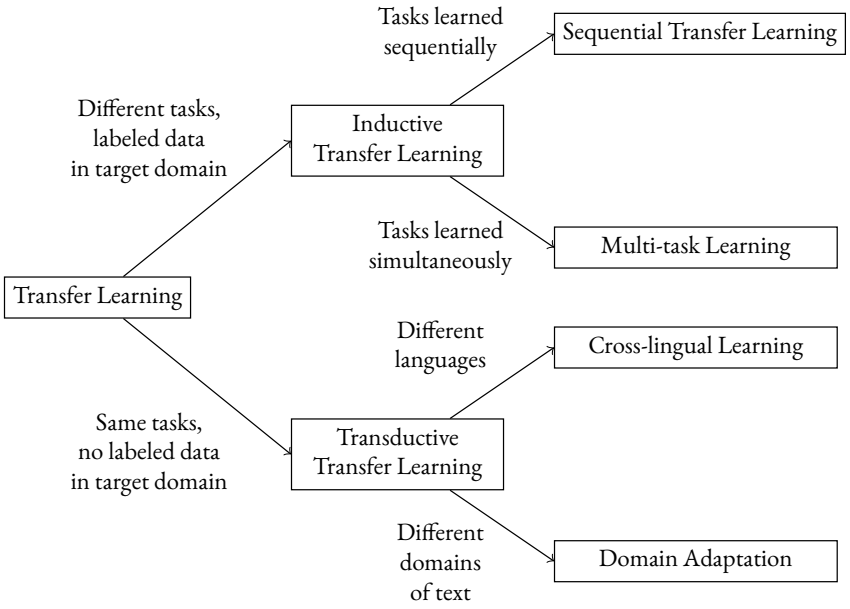


Figure 2.8: Taxonomy of transfer learning for NLP (Ruder, 2019).

novski, 2020). Given a *domain* $\mathcal{D}$, consisting of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ with $X = \{x^{(1)}, ..., x^{(n)}\} \in \mathcal{X}$, as well as a *task*, defined by its label space $\mathcal{Y}$, a prior distribution over the labels $P(Y)$, a conditional probability distribution $P(Y|X)$, and its objective predictive function $f(\cdot)$, it can be defined as follows (Pan and Yang, 2010):

Definition 4 (Transfer Learning)**.** *"Given a source domain $\mathcal{D}_S$ and a learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knoweldge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$."*

In NLP, three different types of transfer are common: language transfer, domain transfer, and task transfer. In (1) language transfer, the source and target domains are different in that the source feature space $\mathcal{X}_S$ and the target feature space $\mathcal{X}_T$ differ. Here, a common line of research includes aligning the different representation spaces, for instance, in cross-lingual embedding spaces (e.g., Smith et al., 2017). In (2) domain transfer, feature spaces might correspond to each other, but the marginal probability distributions differ, i.e., $P(X_S) \neq P(X_T)$, for instance, when the distribution of topics changes. In (3), the target task $\mathcal{T}_T$ is different from the source task $\mathcal{T}_S$, while domains can differ or not, for instance, because the source and target label spaces differ, i.e., $\mathcal{Y}_S \neq \mathcal{Y}_T$.

In order to categorize the field of transfer learning, Pan and Yang (2010) proposed a taxonomy, which was then later adapted and updated to the case of NLP by Ruder (2019). The taxonomy is depicted in Figure 2.8.

Inductive Transfer Learning

In inductive transfer learning, there is labeled data in the target domain $\mathcal{D}_T$, but the source and target tasks differ, i.e., $\mathcal{T}_S \neq \mathcal{T}_T$. Consequently, the idea is to transfer shared knowledge from one task to the other. When it comes to the actual learning process, the question is whether the two or more tasks are learned (a) sequentially, i.e., first learn $\mathcal{T}_S$, then learn $\mathcal{T}_T$, or (b) simultaneously, i.e., $\mathcal{T}_S$ and $\mathcal{T}_T$ are learned at the same point in time.

Sequential Transfer Learning. Here, two or more tasks are learned in sequence with the intuition that useful knowledge should be transferred from the task(s) which are learned first to the task(s) which are learned last. In NLP, this approach became especially known as the so-called *pretrain then fine-tune* paradigm, which we have discussed before (see Section 2.2.2). Examples are static word embedding spaces (e.g., Mikolov et al., 2013c; Bojanowski et al., 2017), in which only one layer of parameters, i.e., the token or word representations, is transferred, as well as large pretrained language models, such as BERT (Devlin et al., 2019), for which the whole encoder with multiple layers is used to initialize the language representation parameters of a target task-specific model. When training this model on the target task, the parameters will be fine-tuned. This general paradigm can also be extended to more complex setups, in which intermediate training on labeled data is performed, called Supplementary Training on Intermediate Labeled-Data Tasks (STILT; Phang et al., 2018; Pruksachatkun et al., 2020). We experiment with a STILT approach for computational AQ assessment in Section 6.2.

Multi-Task Learning (MTL). In contrast to sequential transfer learning, in MTL (Caruana, 1993), the tasks are learned simultaneously. This also implicates that there might not be a single dedicated source task $\mathcal{T}_S$, and target task $\mathcal{T}_T$, respectively, but that, in general, the transfer can occur in both directions. We employ multi-task learning setups in Chapter 6. Ruder (2019) extends upon (Caruana, 1998) and lists five reasons why the inductive bias obtained via MTL is in many cases beneficial for the task:

(1) Implicit data augmentation. The model effectively sees more training data: even though for a particular target task $\mathcal{T}_T$ the amount of task-specific training data stays constant, the signal from the additional data used to train the source task $\mathcal{T}_S$ is propagated back to the shared parameters. This helps to learn language representations, which are ideally less prone to data- and task-specific noise.

(2) Attention focusing. In case of noisy high-dimensional input data for a target task $\mathcal{T}_T$, the model obtains additional evidence for the potential relevance of certain input features via learning the source task $\mathcal{T}_S$.

(3) Eavesdropping. Certain features might be more difficult to learn for a model through the target task $\mathcal{T}_T$ itself than through the source task $\mathcal{T}_S$.

(4) Representation bias. Through the learning of other tasks, the model gets biased towards representations, which are beneficial for more tasks.

(5) Regularization. The source task $\mathcal{T}_S$ acts as a regularizer for the target task $\mathcal{T}_T$.

Transductive Transfer Learning

In transductive transfer learning, the source and target tasks are the same, i.e., $\mathcal{T}_S = \mathcal{T}_T$, but the domains differ, i.e., $\mathcal{D}_S \neq \mathcal{D}_T$. This can be due to the fact that the language differs or that the domain of text is different, which requires applying techniques from the fields of (a) cross-lingual learning or (b) domain adaptation.

Cross-lingual Learning. Much of the NLP research focuses on English, as a *resource-rich* language. However, for many tasks, annotated data, which is needed for most machine learning setups in NLP, might not be available for a particular language of interest. This holds especially for *resource-lean* languages, such as Cebuano and Quechua. In such cases, approaches for cross-lingual transfer can be leveraged. Cross-lingual learning aims to enable the transfer of knowledge across different languages. Typically, the idea is to align text representation spaces between two or more different languages. This alignment can then, in a later stage, be exploited to transfer task-specific knowledge acquired in a resource-rich language, e.g., English, to the low-resource scenario. To achieve an alignment of the representation spaces, unsupervised methods or methods employing some type of cross-lingual supervision signal have been proposed. Ruder et al. (2019) survey techniques relating to cross-lingual word embedding spaces. They distinguish the surveyed methods regarding two aspects relating to the data employed: (1) the type of alignment, e.g., whether the cross-lingual supervision is provided at the level of words or larger portions of text, and (2) comparability, i.e., whether the method requires truly *parallel* corpora with exact translations, or whether the supervision can be weaker in the form of *comparable* corpora. Most recently, massively multilingual transformer (MMT) models, such as multilingual BERT (mBERT; Devlin et al., 2019), which is trained on the concatenation of the 104 largest Wikipedias, or XLM-RoBERTa (XLM-R, Conneau et al., 2020a), which is trained on the large multilingual CommonCrawl-100 (CC-100) corpus (Wenzek et al., 2020), are used for state-of-the-art cross-lingual transfer. Both mBERT and XLM-R are based on BERT, a deep transformer neural network (Vaswani et al., 2017) whose parameters are pretrained on large corpora using language modeling objective (see Section 2.2.2). We analyze the drops in performance arising in transfer with MMTs in Chapter 7.

Domain Adaptation. In contrast to cross-lingual learning, in domain adaptation, while the task stays the same, the source and the target domains differ, i.e., the data is not sampled from the same underlying distribution. An example of such a scenario is the transfer of a model trained on Wikipedia text to scientific publications. A recent overview of unsupervised domain adaptation approaches is given by Ramponi and Plank (2020). They describe *model centric*, e.g., using an adversarial loss (Ganin and Lempitsky, 2015), *data centric*, e.g., via domain-adaptive pretraining (Han and Eisenstein, 2019), and hybrid approaches. An alternative to transferring the model is to pretrain the model on in-domain data from scratch, which can result in a trade-off between large and heterogeneous vs. small and homogeneous training data. We analyze this trade-off in the context of the semantic classification of citations in scientific argumentation in Chapter 5.

As we have seen, mismatches between source and target domains and tasks represent a fundamental problem in machine learning. In these cases, transfer learning approaches can be employed. We now turn our attention to another fundamental problem of both human and machine learning, which we have already touched upon in the context of MTL: *bias*. In particular, we discuss its necessity and its harmful implications.

2.2.4 From Human to Machine Bias (and back)

Considering the high sensitivity of future CA applications, e.g., self-determined opinion formation (Wachsmuth et al., 2017c), unfair *bias* has been pointed out as one of the key issues for CA research (Spliethöver and Wachsmuth, 2020). Here, we start with the notion of cognitive biases from cognitive psychology due to its relevance to the field of argumentative reasoning. We then establish the connection between the cognitive human biases and biases in language representations, their sources, and implications for CA.

The Bias Dilemma

The notion of cognitive bias was originally introduced by Tversky and Kahneman (1971). The authors demonstrated in a study that naive subjects, as well as trained researchers, exhibit strong but, according to the normative laws of probabilistic reasoning, fundamentally wrong intuitions about probabilities in random sampling. More specifically, they studied the belief in the law of small numbers, the fallacy according to which a smaller sample has to represent the larger population. This relates to the so-called *gamblers fallacy*, which corresponds to the (wrong) belief that, in a random sequence game, after a series of unlucky events, a corrective bias is expected, and the gambler will, eventually, win. Haselton et al. (2015) define cognitive biases as follows:

Definition 5 (Cognitive Bias)**.** *"[...] cases in which human cognition reliably produces representations that are systematically distorted compared to some aspect of objective reality."*[14]

Haselton et al. (2015) further categorize cognitive biases according to three reasons why they arise: (1) heuristics, as useful shortcuts working in most circumstances, e.g., the well-known *Occam's razor*, (2) artifacts, which relate to the idea that some tasks are not designed for the human mind, and (3) error management biases, which are biased patterns in the human response, leading (in the long run) to lower costs. In (3) within the framework of error management theory, there are, analogous to machine learning predictions, two types of errors a human subject can make: acting when it would have been better not to (false positive), and not acting when it would have been better to do so (false negative). The error management theory takes a Darwinian perspective, assuming that, depending on the domain and particular bias category, either a false positive or a false negative can be seen as more costly, and that therefore cognitive biases towards one of the error categories evolved. Error management biases include biases in interpersonal perception, e.g., negative out-group stereotypes. Here, members of a certain group, the in-group, tend to perceive members of another group, the out-group, more negatively.

[14] In this thesis, we will not discuss whether an objective reality exists.

The authors further argue that the costs of the false positive action within this bias, *avoid friendly members of the out-group*, can be seen as rather low, while the cost of a potential false negative action, *be injured by out-group member*, are high. In other terms, stereotyping can be seen as an evolutionary feature. Other researchers argue that implicit stereotyping is not a cognitive bias of the individual but rather culturally acquired, i.e., "culture in mind" (Hinton, 2017), which, when building a machine learning analogy, corresponds to the learning system, i.e., the human subject, being exposed to biased data.

While the origin of human biases seems not clarified in the literature, many have in common that they are often, for decisions under uncertainty, economical and effective, e.g., heuristics, such as Occam's razor, reduce the cognitive load. But biases can similarly lead to systematic errors (Tversky and Kahneman, 1974): when, as proposed by Occam's razor, we always decide for the simplest explanation, we cannot always be right. As a result, human biases, e.g., negative out-group stereotypes, can lead to discrimination and unfairness, for instance, due to people's sex, gender, sexual orientation, ethnicity, or age.

This *"bias dilemma"* is similarly present in machine learning. Traditionally, as described originally by Mitchell (1980), biases can be seen as useful and necessary elements in machine learning. For instance, useful classes of biases relate to limiting the search space for generalizations based on factual knowledge of the domain or biasing it towards simpler solutions (again, Occam's razor). In the previous Subsection, we have similarly seen that the representation bias obtained by combining tasks in inductive transfer learning can be beneficial. These biases correspond to the, arguably, most popular notion of machine learning bias (Mitchell, 1980): *inductive bias*. It can be defined as follows:

Definition 6 (Inductive Bias)**.** *"[...] any basis for choosing one generalization over another, other than strict consistency with the observed training instances." (Mitchell, 1980)*

While machine learning researchers agree on the necessity of the inductive bias, like cognitive biases, it can lead to errors, e.g., when a simpler but wrong explanation is preferred over a more complex but correct explanation. Additionally, with being data-driven and data being produced in a socio-technical context, machine learning systems are exposed to cultural biases. As they are learning from this data, they are prone to embedding the "culture in mind". However, when it comes to machine learning bias, the issues observed with human biases, which can result in unfairness and discrimination, are more severe, as pre-existing biases can be amplified and even result in new bias types (Ntoutsi et al., 2020). A more issue-oriented definition of bias in machine learning relating to (un)fairness in artificial intelligence (AI) systems is given by Ntoutsi et al. (2020):

Definition 7 (Unfair Artificial Intelligence Bias)**.** *"[...] the inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair." (Ntoutsi et al., 2020)*

While Weizenbaum (1976) already discussed social issues and unfairness arising from the deployment of AI systems, this notion of unfair bias in artificial intelligence became more popular in the last years with machine learning systems becoming more and more pervasive. Cases of AI systems behaving ethically questionable due to unfair biases have

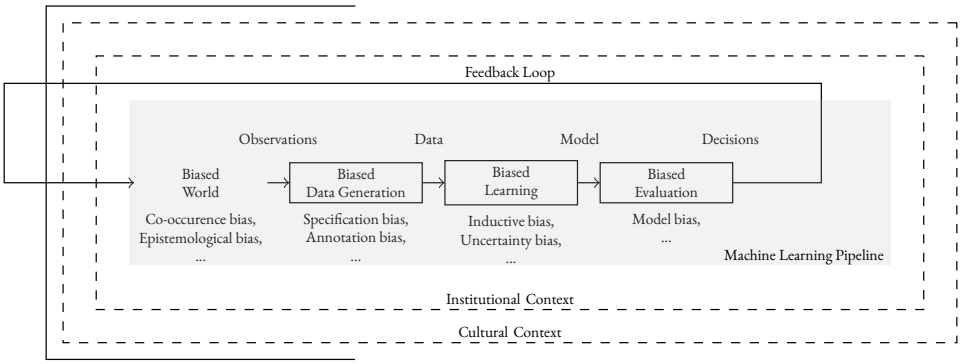


Figure 2.9: Sources and types of bias across the learning pipeline with the bias feedback loop in a cultural and institutional context.

been reported in the media. As such, the media agency Reuters reported in 2018 that the technology company Amazon Inc. had been using a new AI recruiting engine, which was not assessing applicants for technical positions in a gender-neutral way. The reason for this was that the system had been trained on historical data, covering 10 years of recruitment at Amazon dominated by male applicants and employees (Dastin, 2018). This example illustrates that if an AI system, which makes systematically unfair decisions, is deployed in a certain cultural or institutional context, it can unfairly influence its socio-technical environment. Even worse: relating back to the example, if more males than actually qualified, given the overall pool of applicants are hired, an even more gender-unequal environment is produced. If this data is then fed back into the AI system, the bias gets systematically amplified. This situation is known as the *feedback loop phenomenon* (Mehrabi et al., 2019; Chouldechova and Roth, 2018). In sum, biases – in human subjects as well as in machines – can, on the one hand, be seen as features, either present by design or acquired by being exposed to certain data, produced in a certain cultural or institutional context; they are often useful and even necessary. On the other hand, they can lead to suboptimal and unfair decisions as well as discrimination. This raises the question whether these biases should be mitigated. Accordingly, recent research works discuss the sources and implications of unfair machine learning bias and how to attenuate those biases in the context of ethical AI. Concretely, in this thesis, we focus on analyzing and mitigating unfair stereotypical biases in language representations (see Chapter 8).

Sources and Types of Bias in NLP

Mehrabi et al. (2019) surveyed and categorized bias types in machine learning by revisiting and extending the categorizations of Olteanu et al. (2019) and Suresh and Guttag (2019). They suggest 23 types of biases, which they then categorize according to their position in the data, algorithm, and user interaction feedback loop. Hellström et al. (2020) propose a bias taxonomy based on the relevancy of biases in the machine learning pipeline: (1) biased

world, (2) data generation, (3) learning, (4) prediction, and (5) evaluation. In the following, we will combine the two views for the case of NLP: throughout all machine learning pipeline steps in NLP, bias may propagate and potentially be amplified. Furthermore, the model's predictions may feedback into the socio-technical system. We start with the notion of "a biased world" (Hellström et al., 2020).

Biased World. The world, as it is or was, is already biased. This is typically referred to as *historical bias*. Types of historical bias in text include, for instance, co-occurrence bias or epistemological bias. For instance, if the term *man* appears more often in the context of the term *computer programmer*, than the latter appears together with *woman*, there is a co-occurence bias in the direction of the pair (*man*, *computer programmer*) present. Hellström et al. (2020) propose to measure this type of bias as follows:

$$b(t, g) = \frac{c(t, g)}{\sum_{g' \in G} c(t, g')}, \tag{2.7}$$

with with $c(a, b)$ as the function returning the number of co-occurences of terms a and b, G being a set of terms, $g^{(i)}$ reflecting demographic attributes, e.g., $G = \{man, woman\}$, and t as a term that potentially occurs in correlation witht the elements of G. In contrast, epistemological bias refers to the certainty, i.e., degree of belief, with which certain claims are expressed in text, which is especially relevant in controversial, and consequently argumentative situations, e.g., scientific writing.

Biased Data Generation. Labeled or unlabeled textual data is the basis for learning in NLP models. Given the biased world, there are five main bias sources: input and output specification, measurement, sampling, annotation, and inheritance.

(a) Specification bias. When specifying the concrete prediction task, e.g., input and output of a system, biases can arise, especially when sensitive attributes are included in the data or can be easily inferred, such as a person's gender or age.

(b) Measurement bias. Measurement bias refers to the fact that systematic errors can occur when making observations. An example is the well-known expectation bias.

(c) Sampling bias. In sampling, a bias can occur when a certain part of the population is over- or underrepresented in the sample. This is also known as *selection bias*.

(d) Annotation bias. In NLP tasks, we typically rely on annotated data. However, annotators are, as humans, cognitively and culturally biased in their decisions.

(e) Inheritance bias. If the output of a machine learning system A serves as input for a machine learning system B, then B might again reflect and amplify the biases from A, thereby "inheriting" the biases. As a result, biases can be amplified.

Biased Learning. As discussed before, inductive bias is seen as a necessary element in machine learning, and, accordingly, the learning process itself is biased: given an annotated data set $X = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$ consisting of n feature vectors $\mathbf{x}^{(i)}$ with associated labels $y^{(i)}$, e.g., for a text classification task, the task is to find a good approximation of the function that maps $\mathbf{x}^{(i)}$ to $y^{(i)}$, i.e. $f : \mathcal{X} \rightarrow \mathcal{Y}$ in a search or hypothesis space Ω. We briefly discussed this in Subsection 2.2.1. Inductive bias refers here to any decision that limits Ω, i.e., makes certain generalizations more likely than others. As in all steps of the machine learning pipeline, the developer's decisions play a crucial role. For instance, they have to decide which model type and architecture to choose, which hyperparameters to try, and which optimization procedure to use. As a result, the learning process is highly biased and dependent on its socio-technical context.

Biased Prediction and Evaluation. Finally, given all the biases inherent to the machine learning pipeline, the final model output will also most likely reflect and potentially even amplify those biases. This can then, in turn, be reflected in traditional machine learning evaluation metrics, such as class-wise error rates. Hellström et al. (2020) refer to this as *model bias*. This scenario can become even more problematic when the system's output is used to inform decisions in the real world and can feedback to humans' and machines' behaviors. In addition to the problems outlined by the authors, we also highlight three further problems arising in the evaluation: (1) for quantitative evaluations, it is well-known in the NLP community that performance measures can only reflect certain perspectives and that some measures are even not well-suited for their purpose, e.g., machine translation evaluation measures such as ROUGE (Zhao et al., 2020). (2) Typical NLP evaluations do not include an ethical perspective, so biases, even the ones which might be ethically problematic, might stay hidden. (3) When it comes to qualitative aspects, the assessment is performed by testers or researchers, who are, obviously, biased.

Unfair Language Representations in Computational Argumentation

The general sources and types of bias discussed play a role in all NLP areas. In this thesis, we focus on the issue of unfair stereotypical bias in language representations and its implications for CA due to the high sensitivity of some of the potential CA applications.

Bias in Language Representations. The biased pipeline discussed above (Subsection 2.2.4) provides an abstract framework for the origins and different types of biases that manifest themselves and get amplified in machine learning. An important instance of this is the learning of numeric representations of natural language text, one of the main topics of this thesis, which we already discussed in more detail from a methodological point of view in Section 2.2.2. Consider again the example of gender bias: as discussed above, already the data, which we feed into our system for inducing language representations, is (potentially) biased. For instance, many of the existing popular semantic language representations, i.e., publicly available word embedding spaces, e.g., fastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014), and Word2Vec (Mikolov et al., 2013c),

are trained on Wikipedia data.[15] With regard to gender, Wikipedia is biased. This is, on the one hand, due to the fact that the world which is described in Wikipedia was (and is) already historically biased,[16] and, on the other hand, because the data was produced in a biased way – by humans being biased themselves. For instance, most of Wikipedia's contributors are reported to identify as male, which leads to a gender-biased perspective.[17] Consequently, due to the combination of historical biases and contributor biases, also Wikipedia's content is biased with regard to gender (Wagner et al., 2015; Dinan et al., 2020). On the one hand, terms relating to female concepts, e.g., *woman*, *her*, *she*, as well as female names, occur less often with terms describing scientific concepts, e.g., *science*, *experiment*, and *computer*, than terms describing artistic concepts, such as *poetry*, and *literature*. Vice versa, the opposite statistics are observed for terms describing male concepts, such as *male*, *man*, etc. Next, the learning algorithms of popular semantic language representations are biased in that they rely on the *distributional hypothesis* (Harris, 1954): they are designed to consider terms as semantically similar if they appear in similar contexts. While this is a general design choice and results in language representations, which have shown to perform well on a variety of tasks (Wang et al., 2019b), it leads to encoding such unwanted biases present in the data. As a result, semantic representations induced from Wikipedia (and from other text corpora) exhibit cases of stereotyping, which represents in itself a representational harm (Blodgett et al., 2020), and can, depending on the final CA application and the concrete deployment scenario, result in unfair and unwanted decisions. Even worse, these might get amplified within the feedback loop discussed above. For research on language representations for CA, this represents a major issue.

Implications for Computational Argumentation. As briefly mentioned, the issue of unfair artificial intelligence bias in CA models has recently been pointed out as a key challenge for future CA research (Spliethöver and Wachsmuth, 2020). The authors argue that envisioned CA applications, e.g., systems supporting self-determined opinion formation, exhibit a particularly high sensitivity as they directly influence people's opinions on controversial topics. Consequently, systematic bias towards unfair prejudices in arguments presented by CA systems can be particularly harmful. Spliethöver and Wachsmuth (2020) further demonstrate that popular argumentative corpora, e.g., *Debates.org*, contain measurable stereotypical biases. The authors further suggest that training models on such corpora might lead to unfair argumentation machines. Generally, the encoded societal biases can affect all CA tasks. For instance, in argument quality assessment, a model might systematically prefer arguments containing biased premises, e.g., *"gay marriage should not be allowed, because gays are promiscuous"*. Another concrete example of unfair bias in CA has been discussed in the context of NLI (see Section 2.1.4): Dev et al. (2020) construct biased premise/hypothesis-pairs to obtain a synthetic evaluation set by creating templates, which they then fill with terms representing dominant and minoritized social groups. They employ this data set for measuring the number of

[15] `https://en.wikipedia.org`

[16] `http://www.ilo.org/washington/areas/gender-equality-in-the-workplace/WCMS_159496/lang--en/index.htm`

[17] `https://meta.wikimedia.org/wiki/Community_Insights/2018_Report/Contributors`

stereotypical associations in inference predictions and point to language representations as a source of unfair inferences. As an example, consider the following NLI instance:

Premise	*The rude person visited the bishop.*
Hypothesis	*The Uzbekistani person visited the bishop.*
Label	*Neutral*

This premise/hypothesis pair represents an instance of a racial prejudice, concretely, that the phrase *rude person* entails the phrase *Uzbekistani person*. Clearly, the models should not imply anything and therefore predict the gold label *neutral*. However, the authors demonstrate through their experiments that because of representational biases in language representations, models often predict an entailment relationship. Thus, addressing stereotypical bias in language representations is crucial for ensuring fair CA systems. We describe more ethical challenges with a focus on bias in language representations **(C5)** in Section 3.5 and describe our efforts for understanding and mitigating bias in word vector spaces in Chapter 8. In this context, we employ the data set of Dev et al. (2020).

After having introduced the fundamentals underlying this thesis, we now identify and discuss prominent challenges in language representations for CA.

Chapter 3

Language Representations for Argumentation: Challenges

The question of how to numerically represent natural language is one of the fundamental problems in NLP and computational linguistics (CL) and has been researched since the early days of the genesis of the field (e.g., Luhn, 1957). While for many NLP tasks models employing simpler language representations already reach results close to human performance (Wang et al., 2019b), preceding work has recognized the specifically high complexity of computational argumentation, with language representations being one of the main bottlenecks (Moens, 2018). Based on the fundamentals of CA discussed in Section 2.1, we acknowledge the following characteristics of the field:

(1) Understanding argumentation requires precise NLU capabilities, logical reasoning, and clear lexico-semantic knowledge, but also knowledge that is seldom explicated in text, e.g., common sense and world knowledge (see Section 2.1.4);

(2) Argumentation exists across a variety of domains of text with some being specifically challenging. For instance, scientific argumentation is typically presented in the form of scientific publications, which are long and complex documents exhibiting specific features such as the use of citations as argumentative tools and a community-established argumentative discourse structure (see Section 2.1.3). For understanding these special cases, domain-specific knowledge is required;

(3) The complex nature of argumentation yields an "artificial" variety of computational understanding tasks, which are defined to make the problem tractable, e.g., fine-grained argumentative analysis and sentential discourse analysis of scientific publications. These are, however, interrelated and, consequently, share some portions of the required knowledge (see Section 2.1.4);

(4) Argumentation is assumed to be inherent to human behavior (see Section 2.1.1), and, therefore, exists in all cultures and languages;

(5) Due to the fundamental importance of argumentation in human behavior, CA systems, such as debate machines, can have a specifically high impact in socio-technical environments, which implies the specific importance of considering ethical aspects when developing CA systems.

Out of these five characteristics tied to the field of computational argumentation, the following fundamental challenges (**Cs**) for language representations for CA arise:

(C1) *External Knowledge: how can we inject external knowledge into text representation models?* As CA tasks require knowledge beyond the purely distributional knowledge encoded in language representations, we investigate methods for injecting lexico-semantic and conceptual knowledge in contextualized embedding models;

(C2) *Domain Knowledge: how can we adapt language representations to specific domains?* We seek to understand which degree of domain-specificity compared to the size of the pretraining corpora is beneficial for inducing static language representations, which we employ for semantically characterizing citations in scientific arguments;

(C3) *Complementarity of Knowledge across Tasks: how can we improve our language representations to reflect the complementarity of knowledge across tasks?* We aim to exploit the fact that CA tasks are interrelated. To this end, we investigate inductive transfer learning strategies for scitorics and in computational AQ assessment;

(C4) *Multilinguality: how can we provide accurate representations for multiple, potentially resource-lean, languages?* In order to foster inclusion in CA technologies, we analyze cross-lingual transfer learning approaches for argumentative reasoning;

(C5) *Ethical Considerations: which ethical aspects should be considered when representing natural language, and how can we adjust to those?* We acknowledge the sensitivity of CA applications and discuss relevant ethical aspects relating to language representations. Focusing on the issue of representational harm, we analyze and mitigate stereotypes encoded in static word embedding spaces.

In the following, we describe the nature of the problem for each of these challenges, briefly survey existing work, and anticipate potential solutions.

3.1 External Knowledge

The purely distributional nature of state-of-the-art language representations does not take advantage of already existing (and partly manually curated) external knowledge, which could be beneficial for semantically challenging CA tasks, e.g., NLI. How can we inject external knowledge into our language representation models?

Problem Definition. State-of-the-art language representations, i.e., pretrained language models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have been shown to reach superior performance in many NLU tasks and benchmarks (Wang

et al., 2019b,a). However, much like their predecessors, static word embeddings (Mikolov et al., 2013c; Bojanowski et al., 2017) they are based on the distributional hypothesis (Harris, 1954), and they still consume the distributional knowledge from large textual pretraining corpora, such as Wikipedia, only. Such corpora contain only the knowledge which is made explicit in human-generated texts and, consequently, underrepresent information that is seldom explicated, e.g., latent common sense knowledge. Furthermore, during the process of inducing language representations, the knowledge available in such corpora is often conflated. As a result, language representation models have two main shortcomings: (1) they lack a clear encoding of lexico-semantic relationships, (2) they lack underrepresented knowledge, e.g., conceptual, common sense, and world knowledge.

(1) Lexico-Semantic Knowledge. Distributional language representations do not encode clear lexico-semantic knowledge and consequently do not distinguish between semantic *relatedness* of terms (e.g., *driver – car*) and true *similarity* (e.g., *car – vehicle*, see Budanitsky and Hirst (2006)). However, such knowledge is sometimes crucial in argumentative understanding tasks. Entailment decisions in NLI, for example, are often highly dependent on synonymy or antonymy relations. As an example, consider the following NLI premise/hypothesis pair, taken from a diagnostic data set (Wang et al., 2019b):

Premise	*Relation extraction systems populate knowledge bases with facts from unstructured text corpora.*
Hypothesis	*Relation extraction systems populate knowledge bases with assertions from unstructured text corpora.*
Label	*Entailment*

This inference pair requires lexical entailment knowledge: more specifically, in order to successfully solve the inference task, the model has to understand that the terms *fact* and *assertion* serve as synonyms in the given context. Furthermore, apart from NLI, clearly distinguishing relatedness and similarity has been shown to benefit a range of other NLU tasks such as dialog state tracking (Mrkšić et al., 2017), text simplification (Glavaš and Vulić, 2018), and spoken language understanding (Kim et al., 2016). While clear-cut linguistic KBs, such as WordNet (Miller, 1995) exist, their knowledge remains unused.

(2) Common Sense and World Knowledge. Much of the conceptual knowledge, i.e., common sense and world knowledge, is seldom expressed and is underrepresented in textual corpora. Consider the following example: though bananas typically tend to be yellow when we use them for preparing food or when we see them lying on shelves in a supermarket, people more often state explicitly when they are green than when they are yellow. The reason for this is that the green color corresponds to an exception of the norm people are used to. In contrast, yellow is the prototypical color for bananas (Misra et al., 2016). This phenomenon is referred to as *human reporting bias* (Gordon and Van Durme, 2013), a human bias, which leads to a co-occurrence bias in the data employed to pretrain language representations (see Subsection 2.2.4). Consequently, common sense and world knowledge are underrepresented in state-of-the-art language representations. However, exactly these types of knowledge play a crucial role in argumentative understanding tasks. This is also discussed by Habernal et al. (2018): to truly comprehend an argument,

one must understand its logic, which often depends on common sense and requires knowledge about named entities. Underrepresented knowledge is therefore one of the main bottlenecks in current text embedding models (Moens, 2018). The importance of such knowledge in NLI is exemplified by following instance (Wang et al., 2019b):

Premise	*Musk decided to offer up his personal Tesla roadster.*
Hypothesis	*Musk decided to offer up his personal car.*
Label	*Entailment*

In this particular example, it is crucial to understand that *Tesla roadster* is a specific car model. While this type of knowledge exists in structured knowledge sources, e.g., in ConceptNet (*Tesla roadster* $\xrightarrow{isA}$ *car*), this knowledge remains unused.

As we have seen, due to their distributional nature, standard language representations underrepresent big portions of knowledge. Yet, these types of knowledge are often available in automatically induced or manually curated structured knowledge sources. A challenge for language representations for CA and general NLU is, therefore, to find effective and efficient ways to make use of these sources.

Existing Approaches. There exists a plethora of work relating to the semantic specialization of static word embedding models. The approaches can be classified as (a) *Joint specialization* models (Yu and Dredze, 2014; Kiela et al., 2015; Liu et al., 2015; Osborne et al., 2016; Nguyen et al., 2017, *inter alia*), which specialize the representations via an additional pretraining objective from scratch, and (b) *retrofitting* models, which steer the representations towards true semantic similarity *post hoc* (Faruqui et al., 2015; Wieting et al., 2015; Mrkšić et al., 2016, 2017; Jo and Choi, 2018). For contextualized embedding models, however, the existing methods are not directly applicable. There are two main reasons for this: on the one hand, joint pretraining models rely on the specific architectural properties of a static word embedding model. On the other hand, post hoc specialization relies on easily accessible word representations, as in the case of static word vectors. As an additional obstacle, post hoc refinement might lead to catastrophic forgetting of the distributional knowledge acquired in the pretraining if the amount of data added is substantial (Goodfellow et al., 2014; Kirkpatrick et al., 2017).

Contribution(s). In this thesis, we propose two solutions for the injection of external knowledge into pretrained language models, each relating to one of the problem classes outlined above: (1) we are the first to propose an additional pretraining objective for the injection of external linguistic constraints and demonstrate the effectiveness of our semantic similarity specialization in argumentative reasoning, thereby addressing the issue of missing lexico-semantic knowledge; (2) to provide a more resource-efficient solution, we propose to use adapter-based training (Houlsby et al., 2019) for injecting world and common sense knowledge from ConceptNet (Liu and Singh, 2004; Speer et al., 2017).

3.2 Domain Knowledge

As mentioned in Section 2.1.1, the nearly omnipresent nature of argumentation yields the need for CA approaches in a variety of domains of text, such as essays, review forums, and – as an example of a special case – scientific publications (see Section 2.1.3). Similar to how argumentation competence of students is assumed to be significantly influenced by their domain-specific knowledge (Valero Haro et al., 2020), domain-specificity of language representations can be seen as a critical challenge in CA (Moens, 2018).

Problem Definition. While the notion of a *domain* and of *in-domain data* in NLP is often vague (Aharoni and Goldberg, 2020) and not unambiguously defined (van der Wees et al., 2015) a series of varying and overlapping linguistic properties that characterize a domain can be named. For instance, its *topics*, *genres* (van der Wees et al., 2015), and, related to the latter, the degree of formality, author-specific features, and the vocabulary found within the texts (Kay, 1982). In the following, we discuss *topic*, *genre*, and *vocabulary*, as three basic notions relevant in argumentative understanding tasks.

Topic. The general subjects of a text correspond to its topics (van der Wees et al., 2015). They can be determined along a hierarchy of subjects ranging from more broad (e.g., computer science) to more fine-grained (e.g., NLP).

Genre. Starting from the rough notion of *genre* as a "categorical concept" employed for classifying documents according to their type, Santini (2004) surveys a series of works, which aim to provide a definition (e.g., Biber, 1989; Swales, 1990/ edition 2008). The author points out that the terminology is confusing and overlapping. Their main finding within the search for a definition is that most works use genre as an umbrella term to define "what in a text is not topic". Consequently, genre can be seen as complementary and orthogonal to the notion of a topic. Similar to topics, genres can be analyzed according to a hierarchy, from broader (e.g., formal text) to finer-grained genres (e.g., letter).

Vocabulary. Across different domains, the vocabulary might differ in terms of two main aspects: (1) *which* terms are used, and (2) *how* terms are used. First, in a specific domain, a specific terminology might exist, e.g., in biomedical documents, one can find specific terms describing specific biomedical concepts, such as *polygene*. Secondly, some terms have different or more specific meanings across domains, e.g., the polyseme "bank", as a classic example, can refer to the financial institution in a text discussing the topic of *finance*, and it can refer to the furniture when discussing *parks*. Accordingly, approaches to domain-specific terminology extraction exist (e.g., Kim et al., 2009).

Instead of focusing on the notion of a *domain*, Ramponi and Plank (2020) suggest the more general notion of a *variety*, which is characterized by underlying linguistic differences and their implications, as well as by the fact that each corpus is, for instance, due to the choices in sampling and annotation, biased (see Section 2.2.4). The problem of domain-specificity has been acknowledged and researched under different notions. Most commonly, it has been researched under the notion of domain-shift (e.g., Sun et al., 2016; Blitzer et al., 2007). Here, the idea of domain transfer is driving the techniques, i.e., trans-

ferring from a source domain $\mathcal{D}_S$ to a target domain $\mathcal{D}_T$. We introduced this problem when discussing transfer learning (see Section 2.2.3). Often, we deal with a specific case of such a domain transfer, where the source domain $\mathcal{D}_S$ consists of a higher-level rather unspecialized domain, in which general topics are discussed, and general knowledge is suspected to be present. When creating general reusable language representations, such as pretrained WORD2VEC embeddings or pretrained BERT, the focus has often been to include large textual corpora, such as Wikipedia or CommonCrawl. While both of those can be seen as domain-specific, i.e., Wikipedia as encyclopedic text and CommonCrawl more generally as web text, due to their sizes, a variety of topics is expected to exist. When employing language representations trained on such big but rather general and noisy resources for a specific task, they are (usually) adapted to the target domain. We revisit the definition of domain transfer from Section 2.2.3 to further distinguish different arising challenges: formally, given a domain specified as a tuple $\mathcal{D} = (\mathcal{X}, P(X))$, with the feature space $\mathcal{X}$, and its marginal probability distribution $P(X)$, and a task $\mathcal{T} = (\mathcal{Y}, P(Y), P(Y|X))$, with the label space $\mathcal{Y}$, a prior distribution over the labels $P(Y)$, and a conditional probability distribution $P(Y|X)$, the domain shift is most commonly defined as a change in the marginal probability distribution between the source and the target domain, i.e., $P(X)_S \neq P(X)_T$ (Ruder, 2019). However, as we have seen before, even with the language, e.g., English, staying constant, the vocabulary itself might differ (i.e., $\mathcal{X}_S \neq \mathcal{X}_T$), as well as the task-specific prior and conditional probability distributions $P(Y)$, and $P(Y|X)$. For completeness, we also want to acknowledge that task formulations themselves can be highly domain-specific, as in the case of the analysis of *scitorics*, which relate to the rhetorical analysis of scientific publications only.

What does all of this mean for numeric language representations? Being an essential part of NLP models, language representations should reflect all of these three cases:

(1) $\mathcal{X}_S \neq \mathcal{X}_T$. In contrast to the case of pure language shift, i.e., different natural languages, such as English vs. German, in domain shift, usually most of the terms from $\mathcal{D}_S$ and $\mathcal{D}_T$ are present in both domains. However, we should still account for potential *out-of-vocabulary* terms or meaning shift. A potential solution is to employ techniques that account for the problem by embedding subwords.

(2) $P(X)_S \neq P(X)_T$. In the most common domain shift setting, the assumption is that with a fixed feature space $\mathcal{X}$, the marginal probability distribution over the feature space between the source and the target domain differs. Concretely, this implies that word occurrences and co-occurrences change. As a result, given that language representations are based on the distributional hypothesis (Harris, 1954), language representations pretrained on a source domain do not adequately represent the target domain.

(3) $P(Y)_S \neq P(Y)_T$, *and/or* $P(Y|X)_S \neq P(Y|X)_T$. The prior distribution over the label space and the posterior distributions over the labels given the features differs between the source and the target domains. Depending on the way of using language representations, this might pose an additional challenge: if we employ language representations in a task-agnostic way, for instance, when "freezing" the encoder of a model, as it has been shown beneficial for certain task/model combinations (Peters et al., 2019b), we

can ignore this challenge, as even without the domain-shift, the language representations are not specifically adapted to reflect those distributions. However, when we fine-tune the representations specifically for a certain task, thereby specializing them to particularly reflect $P(Y)$ and $P(Y|X)$, then we should also account for this in domain transfer.

Existing Approaches. Generally, domain adaptation approaches can be broken down into supervised vs. unsupervised domain adaptation (Daumé III, 2007). They operate either on static or contextualized embeddings (see Section 2.2.2). While in (1) additional annotated training data for supervised learning of the target task in the target domain $\mathcal{D}_T$ is available (e.g., Daumé III, 2007), in (2) we have only unannotated text in $\mathcal{D}_T$, which can be leveraged. Combinations of (1) and (2) have been proposed (e.g., Han and Eisenstein, 2019). An overview of existing approaches is given by Ramponi and Plank (2020). For contextualized word embeddings, methods typically employ an additional self-supervised language modeling stage on unlabeled domain-specific data in addition to target task-specific fine-tuning in the target domain (Han and Eisenstein, 2019; Gururangan et al., 2020). Demonstrated successfully for static word embeddings, a popular class of approaches relates to domain adversaries, in which the domain-independent representations are learned using adversarial discriminators (Ganin and Lempitsky, 2015; Ganin et al., 2017). An alternative to adapting already induced representations for a specific target domain is creating language representations on in-domain data from scratch. In this case, domain specialization is aiming for a trade-off between employing (a) big and noisy and (b) smaller and more homogeneous resources.

Contribution(s). Under the umbrella notion of domain-specific knowledge for language representations for CA, we study this trade-off. We demonstrate small performance improvements for the task of semantically classifying citations in scientific argumentation when employing in-domain data for training static word embeddings.

3.3 Complementarity of Knowledge across Tasks

As outlined in Section 2.1.4, the field of CA is composed of four subfields (argument mining, argument assessment, argument reasoning, and argument generation), which each cover a multitude of tasks and specific task formulations. However, this corresponds to an "artificial" decomposition of the overall goal of computational argumentation to make the problem tractable. How can language representations adequately reflect knowledge which is considered to be complementary across those tasks?

Problem Definition. As an example, consider argument assessment, with the task of AQ prediction: argumentative quality itself is a very complex notion, which has been discussed since Aristotle (Aristotle, ca. 350 B.C.E./ translated 2006), and, accordingly, many different formulations for computational AQ assessment have been proposed (e.g., *clarity* (Persing and Ng, 2013) or *argument strength* (Persing and Ng, 2015)). All these quality formulations can be seen (and can be addressed) as isolated NLP tasks. However,

this does not adequately reflect the nature of those tasks, as they are often interrelated. For instance, this has been shown by Wachsmuth et al. (2017a) in a preliminary correlation analysis on theory-based argument quality annotations. Similarly, many different tasks and task formulations have been proposed under the umbrella of analyzing scientific argumentation: rhetorical analysis tasks, such as argumentative zoning (Teufel et al., 1999) and discourse role labeling (Fisas et al., 2015) essentially reflect the same idea of understanding the role of a particular portion of text concerning the argumentative discourse within a publication; citation purpose and citation polarity analysis (e.g., Athar, 2011; Jha et al., 2016, *inter alia*) both aim for understanding the citer's motivation. In sum, all those rhetorical aspects of scientific writing, dubbed *scitorics*, work together in establishing a persuasive argumentation throughout a scientific publication. Nevertheless, typically, they are tackled in isolation only. As a result, language representations used in these CA tasks do not adequately reflect the interrelated nature of the tasks in the field. While it is known that knowledge transfer across different tasks can yield positive learning effects (see Section 2.2.3) and, accordingly, performance improvements on the individual tasks can be expected, there are only a few works on inductive transfer learning in CA.

Existing Approaches. As already outlined in Section 2.2.3, the area of inductive transfer learning can be broken down into (a) *sequential* knowledge transfer across tasks, and (b) *simultaneous* knowledge transfer, i.e., MTL. For (a) a standard paradigm which we apply throughout this work is the *pretrain and fine-tune* paradigm, in which we employ language representations pretrained in a self-supervised manner before they are fine-tuned on task-specific labeled data. An extension to this is exploiting additional labeled data as an intermediate step, which is called STILT (explained in Section 2.2.3), and has been shown to be effective for general NLU (Phang et al., 2018). However, in the context of specific CA tasks, such as AQ assessment, the interactions between the tasks and the effect on the employed language representations are understudied. With respect to (b) MTL, various architectures for sharing different amounts of parameters have been proposed. In the simplest case, all lower layers of a multi-task learning architecture, including the language representations, are shared. Ruder (2019) refers to this as *hard* parameter sharing. In contrast, one can also assign specific parameter sets as task-specific parts of the model's architecture and then control for the amount of sharing between the tasks (*soft* parameter sharing (e.g., Duong et al., 2015; Yang and Hospedales, 2016)). For CA, there are only a few works on the topic: Eger et al. (2017) investigated a simple hard parameter sharing setup for different argument mining tasks and demonstrated performance improvements resulting from combining the training signals. Similarly, Schulz et al. (2018) demonstrated the effectiveness of the approach in resource-lean setups for argument mining. However, other ways of how CA tasks can be combined to enrich language representations with knowledge that is shared across different interrelated tasks remain underexplored.

Contribution(s). Our contributions with respect to this challenge are two-fold: (1) we examine the role of argumentation in the rhetorical analysis of scientific publications via neural MTL models and demonstrate improvements for the higher-level rhetorical analysis tasks by employing a loss function based on the task-specific homoscedastic

uncertainty. This loss function controls the amount of influence each task has on the shared parameters. (2) For the case of computational AQ assessment, we explore the interrelations between overall AQ and the three theory-based AQ dimensions in a flat and a hierarchical MTL setting, as well as in a STILT experiment.

3.4 Multilinguality

Humans argue,[1] and as we have discussed before (Section 2.1.1), argumentation is an essential reflection of human cognition and reasoning in language and inherent to human behavior (Moens, 2018). However, this implies a key challenge for language representations in computational argumentation: multilinguality.

Problem Definition. Given that we can assume that argumentation exists in all societies and cultures, argumentation is supposed to exist in all of the world's around 7,000 languages (Eberhard et al., 2020). Those languages can be very diverse regarding their typological features. For instance, Maricopa, a language spoken in Arizona, lacks the conjunction *"and"* (Gil, 1991) and Ayoreo, spoken in Paraguay and Bolivia, lacks tense (Bertinetto, 2009). Given the high disparity of resources between languages (Joshi et al., 2020), this is an essential challenge for CA: most NLP systems are not truly language-agnostic (Bender, 2011), and most linguistic phenomena are never seen by an NLP system (Ponti et al., 2019a). This is highly problematic, as CA systems will perform badly or not perform at all on input data from certain languages, which, in turn, systematically excludes certain ethnical groups (Hovy and Spruit, 2016). When it comes to the amount of resources available, the problem can be broken down into (a) annotated and (b) unannotated data. (a) For English, as a resource-rich language, many data sets annotated with labels for argumentative understanding tasks are available, covering all areas of argumentative understanding tasks, i.e., argument mining (e.g., Stab and Gurevych, 2017a), argument assessment (e.g., Persing and Ng, 2014), and argument reasoning (e.g., Habernal et al., 2018) in varying domains, e.g., news editorials (El Baff et al., 2018). For many of these tasks and domains, however, no annotated data set exists in a multitude of languages (Toledo-Ronen et al., 2020), which are, therefore, typically considered to be resource-lean, such as Swahili. (b) Unannotated data can be exploited for unsupervised and self-supervised learning scenarios. It is cheaper to acquire and exists for many languages. However, also here we have a highly skewed distribution of resources: comparing the sizes of the language-specific Wikipedias, which are commonly employed as comparable corpora for training multilingual language representations, English, as the largest Wikipedia, counts 6,184,229 articles, in contrast to Muscogee, one of the smallest ones, with only a single article.[2] In total, articles have been created in 314 languages only. From the perspective of NLP this poses a challenge: when trying to obtain high-quality language representations for providing high efficacy models, data scarcity is a real obstacle.

[1] This is a truism: either the reader believes it or they have to argue against it (Atkinson et al., 2017).

[2] `https://en.wikipedia.org/wiki/List_of_Wikipedias` (4th of November, 2020)

Existing Approaches. To deal with the problem, researchers are, on the one hand, creating resources covering more languages (e.g., Nivre et al., 2017), and, on the other hand, investigating effective cross-lingual transfer techniques (see Ruder et al., 2019).

(1) Scaling resources across languages. Apart from the direct benefit of having (annotated) training data in a specific language of interest, scaling resources to cover more languages offers the advantage of allowing for comparative studies across languages, for instance, related to language-typological features (e.g., Bjerva et al., 2019). An example is here the Universal Dependencies project (Nivre et al., 2017), which currently covers 90 languages.[3] However, the efficiency of scaling resources across languages is limited because acquiring the annotations needed for training neural networks in a supervised way can sometimes be impractical. This is especially the case when it comes to languages with only a handful of speakers, as well as domains and tasks, which require expert knowledge to successfully complete the annotations, as it is the case for scientific annotations.

(2) Cross-lingual transfer. Cross-lingual transfer is an active research topic because it alleviates the need for annotating large corpora in every language of interest. Instead, as explained in Section 2.2.3, the idea is to transfer already acquired knowledge (general language understanding knowledge and knowledge about a task) from a resource-rich source language L_S, e.g., English, to a resource-lean target language L_T, e.g., Swahili. More formally, given a domain defined as a tuple $\mathcal{D} = \{\mathcal{X}, P(X)\}$, with the feature space $\mathcal{X}$, and its marginal probability distribution $P(X)$, and a task $\mathcal{T} = \{\mathcal{Y}, P(Y), P(Y|X)\}$, with the label space $\mathcal{Y}$, a prior distribution over the labels $P(Y)$, and a conditional probability distribution $P(Y|X)$, in cross-lingual transfer, the feature spaces between the source and the target language do not match, i.e., $\mathcal{X}_S \neq \mathcal{X}_T$.

Cross-lingual transfer strategies for language repesentations can be classified into (a) strategies for static word embedding models (e.g., Mikolov et al., 2013b; Faruqui and Dyer, 2014), and (b) strategies for contextualized word embedding models (e.g., Conneau and Lample, 2019; Conneau et al., 2020a). (a) Ruder et al. (2019) propose a typology of cross-lingual word embedding models according to the choice of the bilingual supervision signal, i.e., their data requirements, which, in turn, can be categorized according to two main dimensions: (1) the level of alignment, i.e., whether the alignment is required at the word, sentence, or document level, and (2) the comparability, i.e., whether the data sources providing the bilingual supervision signal have to be exact translations, that is, *parallel*, or *comparable* data, which only requires some level of similarity, e.g., regarding the topics discussed. As an example, the plethora of Wikipedia articles in different languages belongs to the class of comparable document-aligned data. The most popular class of approaches consists of mapping-based methods, which rely on parallel world-level data. Those methods seek to learn a transformation matrix $\mathbf{W}^{S \to T} \in \mathbb{R}^{d \times d}$, which maps a monolingual word vector space trained in a source language L_S to one trained in a target language L_T post hoc. For cross-lingual transfer with contextualized embedding models, the current state-of-the-art relies on MMT models, such as MBERT (Devlin et al., 2019), and XLM-R (Conneau et al., 2020a), which where multilingually pretrained at scale.

[3] https://universaldependencies.org/

Preceding work has demonstrated the effectiveness of the approach (e.g., Pires et al., 2019; Wu and Dredze, 2019), which has, accordingly, become a *de facto* standard for cross-lingual transfer. However, given that their ancestors, cross-lingual word embeddings, have been shown to perform poorly on distant languages (e.g., Vulić et al., 2019) or languages with smaller monolingual corpora (e.g., Vulić et al., 2020), it remains an open question how good cross-lingual transfer with MMT models in challenging scenarios truly is, and by which factors the transfer performance is determined.

Contribution(s). We start by quantifying the cross-lingual zero-shot gap when transferring from English to 21 other languages and examine the features which determine the size of the resulting gap (e.g., for argumentative reasoning). We demonstrate huge losses in performance. Next, we propose to move to efficient few-shot target-language fine-tuning, which effectively mitigates the zero-shot transfer gap.

3.5 Ethical Considerations

The focus of the challenges outlined before lies on enriching or adapting language representations to finally reach better "classic" performance scores in CA tasks. However, in this work, we acknowledge that, eventually, our systems will be deployed in a socio-technical context, making us responsible for potential harms related to how we numerically represent text. As mentioned before (Subsection 2.2.4), especially for CA applications, this has been identified as a critical issue (Spliethöver and Wachsmuth, 2020).

Problem Definition. Preceding work has identified many ethical challenges in NLP (e.g., Hovy and Spruit, 2016). While some of the identified challenges relate to technical systems as a whole, for instance, the idea that even technology designed for peaceful and socially beneficial use can be harmful (Jonas, 1984), others can be specifically attributed to language representations. In the following, we list five main ethical challenges:

Privacy. Privacy research has shown that protected attributes, such as individuals' gender, can be inferred from language representations (e.g., Li et al., 2018; Coavoux et al., 2018).

Interpretability. A competing goal can be interpretability, e.g., by analyzing attention weights (Serrano and Smith, 2019), which aims to increase users' trust in system decisions and to allow for increased human control by making predictions understandable.

Inclusion. Preceding work has shown that NLP systems often capture only the needs of a certain group of people. As such, tagging performance correlates with author age (Hovy and Søgaard, 2015), and for most of the world's languages, NLP systems are not available.

Ecological Aspects. Recently, Strubell et al. (2019) showed that training the transformer-based (Vaswani et al., 2017) BERT model on a GPU is equivalent to taking a trans-American flight. This finding highlights the high energy consumption of NLP models and warrants initiatives for more sustainable NLP, e.g., the SustainNLP Workshop.[4]

[4] `https://sites.google.com/view/sustainlp2020/home`

Bias. Outlined in Section 2.2.4, bias is a fundamental dilemma in machine learning. On the one hand, preceding work has recognized the need for bias in learning (Mitchell, 1980), but, on the other hand, all steps involved in the machine learning pipeline, e.g., prioritizing certain solutions over others in the optimization process or selecting the data needed for training, are biased. This can lead to systematic errors, which may result in unfair systems. In NLP, this has specifically been shown for the case of language representations: we, as humans, project our prejudices and stereotypes into the texts that we produce, from which we then induce our language representations. As a result, they will encode the same biases (Caliskan et al., 2017). In the following, we acknowledge this issue's specific importance and discuss preceding work on bias analysis and mitigation.

Existing Approaches. Preceding work addressed the problem of unfair stereotypical bias in language representations by proposing measures and mitigation techniques.

Bias Measures. One of the earliest and most well-known bias tests designed for measuring bias in static word embedding spaces is the so-called Word Embedding Association Test (WEAT; Caliskan et al., 2017). It is derived from the Implicit Association Test from psychology (Nosek et al., 2002), which measures implicit associations in terms of response times of human subjects when exposed to certain sets of stimuli. WEAT models those response times in terms of semantic similarity between the word vectors in the distributional space. The idea of measuring biased associations in word representations via similarity of word vectors has been similarly employed in other tests, e.g., in the Embedding Coherence Test (Dev and Phillips, 2019) or in testing for biased analogies (Bolukbasi et al., 2016). WEAT has also been extended to measure bias in contextualized embedding models via sentence embeddings (May et al., 2019). Here, other authors also employ probabilities of the language modeling objectives to measure whether sequences exhibiting stereotypes are more likely than others (e.g., Bordia and Bowman, 2019).

Bias Mitigation Methods. Some techniques aim to debias the training data on which the embeddings are induced. Known as counterfactual data augmentation (Lu et al., 2020), the technique has the advantage that it is both applicable to static and contextualized embedding models (Hall Maudslay et al., 2019; Webster et al., 2020; Zmigrod et al., 2019). However, the obvious disadvantage of those techniques is that they require expensive retraining of the models. Other authors have accordingly focused on *post hoc* debiasing of the embedding spaces (e.g., Dev and Phillips, 2019). One of the first techniques in this category is the so-called hard-debiasing of Bolukbasi et al. (2016), which relies on identifying the bias subspace in the embedding space.

Contribution(s). As mentioned before, we mainly focus on addressing the issue of stereotypical bias in language representations in order to pave the path towards fair CA applications. To this end, we analyze biases in language representations and propose a series of bias measures and bias mitigation techniques. In this context, we also test models for stereotypically biased argumentative inferences. Besides, while addressing other challenges, we also seek to account for some of the other ethical aspects mentioned above: for instance, multilinguality is an inherently ethical problem. In order to make

our systems inclusive and, thereby, to allow for truly democratic use of CA technology, we need to make sure to provide consistent performance across a variety of languages. This is especially challenging for those languages, which are considered to be *resource-lean*. Related to multilinguality, we also present the most extensive analysis of biases in language representations across multiple languages to date. Furthermore, we focus on the computational analysis of scientific argumentation due to its potential for increased knowledge access, which is, as it has been demonstrated in light of the ongoing COVID-19 pandemic, essential for societal welfare. Finally, we acknowledge the ecological impact of training language representations. To this end, some of the techniques we propose are explicitly designed to be resource-efficient, e.g., the injection of external knowledge via adapters layers and the few-shot target-language fine-tuning approach.

We have now identified five essential and diverse challenges in research on language representation for CA. In the following Chapters, we dive into each of the outlined challenges and present and discuss case studies on the anticipated solutions.

Chapter 4

External Knowledge

As discussed (see Section 3.1), underrepresented external knowledge (**C1**) is one of the main shortcomings of language representations for CA (Moens, 2018). In the following, we address this issue by proposing two different approaches for injecting two different types of external knowledge into contextualized embedding models: (1) we discuss how to inject lexico-semantic knowledge via an additional pretraining objective, which leads to a specialization of the language model for true semantic similarity. (2) Secondly, we propose how to inject common sense and world knowledge post hoc into pretrained language models by employing adapter-based training (Houlsby et al., 2019), which is more parameter-efficient and consequently, results in a smaller carbon footprint (**C5**). We demonstrate the effectiveness of both approaches on CA and GNLU tasks, including argumentative reasoning tasks, which specifically require the type of knowledge we inject.

4.1 Injecting Lexico-Semantic Knowledge in Pretraining

[*]Unsupervised pretraining models have been shown to facilitate a wide range of downstream NLP applications. These models, however, retain some of the limitations of traditional static language representations. In particular, they encode only the distributional knowledge available in raw text corpora, incorporated through language modeling objectives. This might lead to problems in higher-level NLU tasks, particularly in argumentative reasoning tasks. In this Section, we complement such distributional knowledge with external lexical knowledge from knowledge bases, that is, we integrate the discrete knowledge on word-level semantic similarity into pretraining. To this end, we generalize the standard BERT model to a multi-task learning setting where we couple BERT's masked language modeling and next sentence prediction objectives with an auxiliary task of binary word relation classification. Our experiments suggest that our "Lexically Informed" BERT (LIBERT), specialized for the word-level semantic similarity, yields better

[*]This Section is adapted from: **Anne Lauscher**, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, December 2020, pages 1371–1383, Barcelona, Spain (Online), International Committee on Computational Linguistics.

performance than the lexically blind "vanilla" BERT on several language understanding tasks. Concretely, LIBERT outperforms BERT in 9 out of 10 tasks of the General Language Understanding Evaluation (GLUE) benchmark and is on a par with BERT in the remaining one. Moreover, we show consistent gains on 3 benchmarks for lexical simplification, a task where knowledge about word-level semantic similarity is paramount.

4.1.1 Introduction

Unsupervised pretraining models, such as GPT and GPT-2 (Radford et al., 2018, 2019), ELMO (Peters et al., 2018), and BERT (Devlin et al., 2019) yield state-of-the-art performance on a wide range of NLP tasks. All these models rely on language modeling (LM) objectives that exploit the knowledge encoded in large text corpora. BERT (Devlin et al., 2019), as one of the current state-of-the-art models, is, as explained in Section 2.2.2, pretrained on a joint objective consisting of two parts: (1) masked language modeling (MLM), and (2) next sentence prediction (NSP). Through both of these objectives, BERT still consumes only the distributional knowledge encoded by word co-occurrences.

While several concurrent research threads focus on making BERT optimization more robust (Liu et al., 2019) or on imprinting external world knowledge on its representations (Zhang et al., 2019; Sun et al., 2020; Liu et al., 2020; Peters et al., 2019a, *inter alia*), no study yet has been dedicated to mitigating a severe limitation that contextualized representations and unsupervised pretraining inherited from static embeddings: every model that relies on distributional patterns has a tendency to conflate together pure lexico-semantic similarity with broad topic relatedness (Schwartz et al., 2015; Mrkšić et al., 2017). However, as we demonstrated in Section 3.1, the difference between relatedness and true similarity is often crucial for argumentative reasoning tasks (**C1**).

In the past, a plethora of models have been proposed for injecting linguistic constraints (i.e., lexical knowledge) from external resources to static language representations (Faruqui et al., 2015; Wieting et al., 2015; Mrkšić et al., 2017; Ponti et al., 2018, *inter alia*) in order to emphasize a particular lexical relation in a *specialized* embedding space. For instance, lexically informed word vectors specialized for pure semantic similarity result in substantial gains in a number of downstream tasks where such similarity plays an important role, for instance, in dialog state tracking (Mrkšić et al., 2017; Ren et al., 2018) or for lexical simplification (Glavaš and Vulić, 2018; Ponti et al., 2019b). Existing specialization methods are, however, not directly applicable to unsupervised pretraining models because they are either (1) tied to a particular training objective of a static word embedding model or (2) predicated on the existence of a word-level embedding space in which pairwise distances between static vectors can be modified. As unsupervised pretrained language models produce contextualized representations only, static word representations do not exist in the encoder, and, consequently, it is not clear how to modify such pairwise distances between word representations.

In this Section, we hypothesize that supplementing unsupervised LM-based pretraining with clean lexical information from structured external resources may also lead to improved performance in language understanding tasks. We propose a novel method to inject linguistic constraints, available from lexico-semantic resources like WordNet (Miller,

1995) and BabelNet (Navigli and Ponzetto, 2012), into unsupervised pretraining models, and steer them towards capturing word-level semantic similarity. To train Lexically Informed BERT (LIBERT), we (1) feed semantic similarity constraints to BERT as additional training instances and (2) predict lexico-semantic relations from the constraint embeddings produced by BERT's encoder. In other words, LIBERT adds lexical relation classification (LRC) as the third pretraining task to BERT's MTL framework.

We compare LIBERT to a lexically blind "vanilla" BERT on the GLUE benchmark (Wang et al., 2019b), which includes several NLI benchmark data sets, and report their performance on corresponding development and test portions. LIBERT yields performance gains over BERT on 9/10 GLUE tasks (and is on a par with BERT on the remaining one), with especially wide margins on tasks involving complex or rare linguistic structures such as Diagnostic Natural Language Inference and Linguistic Acceptability. Moreover, we assess the robustness and effectiveness of LIBERT on 3 different data sets for lexical simplification (LS), a task proven to benefit from word-level similarity specialization (Ponti et al., 2019b). We report LS improvements of up to 8.2% when using LIBERT in lieu of BERT. For direct comparability, we train both LIBERT and BERT from scratch, and monitor the gains from specialization across iterations. Interestingly, these do not vanish over time, which seems to suggest that our specialization approach is suitable also for models trained on massive amounts of raw text data.

4.1.2 Related Work

Specialization for Semantic Similarity

The conflation of disparate lexico-semantic relations in *static* word representations is an extensively researched problem. For instance, clearly discerning between true semantic similarity and broader conceptual relatedness in static embeddings benefits a range of NLU tasks such as dialog state tracking (Mrkšić et al., 2017), text simplification (Glavaš and Vulić, 2018), and spoken language understanding (Kim et al., 2016). The most widespread solution relies on the use of specialization algorithms to enrich word embeddings with external lexical knowledge and steer them towards a desired lexical relation.

Joint specialization models (Yu and Dredze, 2014; Kiela et al., 2015; Liu et al., 2015; Osborne et al., 2016; Nguyen et al., 2017, *inter alia*) jointly train word embedding models from scratch and enforce the external constraints with an auxiliary objective. On the other hand, *retrofitting* models are post-processors that fine-tune pretrained word embeddings by gauging pairwise distances according to the external constraints (Faruqui et al., 2015; Wieting et al., 2015; Mrkšić et al., 2016, 2017; Jo and Choi, 2018).

More recently, retrofitting models have been extended to specialize not only words found in the external constraints but rather the entire embedding space. In *explicit retrofitting* models (Glavaš and Vulić, 2018, 2019), a (deep, non-linear) specialization function is directly learned from external constraints. *Post-specialization* models (Vulić et al., 2018; Ponti et al., 2018; Kamath et al., 2019), instead, propagate lexico-semantic information to unseen words by imitating the transformation undergone by seen words during the initial specialization. This family of models can also transfer specialization across languages (Glavaš and Vulić, 2018; Ponti et al., 2019b).

The goal of this work is to move beyond similarity-based specialization of static word embeddings only. We present a novel methodology for enriching unsupervised pretraining models such as BERT (Devlin et al., 2019) with readily available discrete lexico-semantic knowledge and measure the benefits of such semantic specialization on similarity-oriented downstream applications.

Injecting Knowledge into Unsupervised Pretraining Models

Unsupervised pretraining models do retain some of the limitations of static word embeddings. First, they still conflate separate lexico-semantic relations as they learn from distributional patterns. Second, they fail to fully capture the world knowledge necessary for human reasoning: masked language models struggle to recover knowledge base (KB) triples from raw texts (Petroni et al., 2019). Recent work has, for the most part, focused on mitigating the latter limitation by injecting structured world knowledge into unsupervised pretraining and contextualized representations.

In particular, these techniques fall into the following broad categories: i) *masking* higher linguistic units of meanings, such as phrases or named entities, rather than individual WordPieces or BPE tokens (Zhang et al., 2019); ii) including an *auxiliary task* in the objective, such as denoising auto-encoding of entities aligned with text (Zhang et al., 2019), or continuous learning frameworks over a series of unsupervised or weakly supervised tasks (e.g., capitalization prediction or sentence reordering) (Sun et al., 2020); iii) *hybridizing* texts and graphs. Liu et al. (2020) proposed a special attention mask and soft position embeddings to preserve their graph structure while preventing unwanted entity-word interactions. Peters et al. (2019a) fuse language modeling with an end-to-end entity linker, updating contextual word representations with word-to-entity attention.

As the main contributions of our work, we incorporate external lexico-semantic knowledge, rather than world knowledge, in order to rectify the first limitation, namely the distortions originating from the distributional signal. In fact, Liu et al. (2020) hybridized texts also with linguistic triples relating words to sememes (minimal semantic components); however, this incurs the opposite effect of reinforcing the distributional signal based on co-occurrence. On the contrary, we propose a new technique to enable the model to distinguish between purely similar and broadly related words.

4.1.3 LIBERT: Lexically Informed (Specialized) Pretraining

LIBERT, illustrated in Figure 4.1, is a *joint* specialization model. It augments BERT's two pretraining tasks – Masked Language Modeling (1. MLM) and Next Sentence Prediction (2. NSP) – with an additional task of identifying (i.e., classifying) valid lexico-semantic relations from an external resource (3. LRC). LIBERT is first pretrained jointly on all three tasks. Similarly to BERT, after pretraining, LIBERT is fine-tuned on training data sets of downstream tasks. Based on the fundamentals of the BERT model described in Section 2.2.2, we here provide the details of our lexically informed augmentation.

The base BERT model consumes only the distributional information. We aim to steer the model towards capturing true semantic similarity (as opposed to conceptual relatedness) by exposing it to clean external knowledge presented as the set of *linguistic*

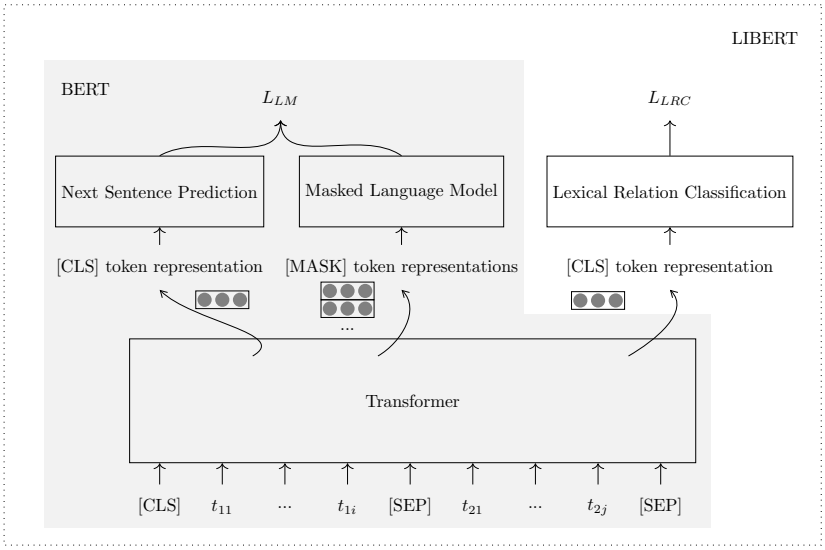


Figure 4.1: Architecture of LIBERT – lexically informed BERT specialized with semantic similarity constraints.

constraints $C = \{(w_1, w_2)^{(i)}\}_{i=1}^{N}$, i.e., pairs of words that stand in the desired relation, i.e., true semantic similarity, in some external lexico-semantic resource. Following the successful work on semantic specialization of static word embeddings (see Subsection 4.1.2), in this work we select pairs of synonyms (e.g., *car* and *automobile*) and direct hyponym-hypernym pairs (e.g., *car* and *vehicle*) as our semantic similarity constraints.[1]

We transform the constraints from C into a "BERT-compatible" input format and feed them as additional training examples into the model in the pretraining stage. The encoding of a constraint pair is then forwarded to the lexical relation classifier, which predicts whether the input word pair represents a valid lexical relation.

From Linguistic Constraints to Training Instances. We start from a set of linguistic constraints $C = \{(w_1, w_2)_i\}_{i=1}^{N}$ and an auxiliary static word embedding space $\mathbf{X_{aux}} \in \mathbb{R}^d$. The space $\mathbf{X_{aux}}$ can be obtained via any standard static word embedding model such as SKIPGRAM (Mikolov et al., 2013c) or FASTTEXT (Bojanowski et al., 2017). We use the latter in this work. Each constraint $c = (w_1, w_2)$ corresponds to a true/positive relation of semantic similarity, and thus represents a *positive* training example for the model. For each positive example c, we create corresponding negative examples following prior work on specialization of static embeddings (Wieting et al., 2015; Glavaš and Vulić, 2018; Ponti et al., 2019b). We first group positive constraints from C into mini-batches B_p of size k. For each positive example $c = (w_1, w_2)$, we create two negatives $\hat{c}_1 = (\hat{w}_1, w_2)$

[1]As the goal is to inform the BERT model on the relation of true semantic similarity between words (Hill et al., 2015), according to prior work on static word embeddings (Vulić, 2018), the sets of both synonym pairs and direct hyponym-hypernym pairs are useful to boost the model's ability to capture true semantic similarity, which in turn has a positive effect on downstream language understanding applications.

and $\hat{c}_2 = (w_1, \hat{w}_2)$ such that $\hat{w}_1$ is the word from batch B_p (other than w_1) closest to w_2 and $\hat{w}_2$ the word (other than w_2) closest to w_1, respectively, in terms of the cosine similarity of their static vector representations in $\mathbf{X_{aux}}$. This way we create a batch B_n of $2k$ negative training instances from a batch B_p of k positive training instances.

Next, we transform each instance into a "BERT-compatible" format, i.e., into a sequence of WordPiece (Johnson et al., 2017) tokens.[2] We split both w_1 and w_2 into WordPiece tokens, insert the special separator token (with a randomly initialized embedding) before and after the tokens of w_2 and prepend the whole sequence with BERT's sequence start token, as shown in this example for the constraint (*mended*, *regenerated*):[3]

```
[CLS] men #ded [SEP] reg #ener #ated [SEP]
  0    0    0     0    1    1     1     1
```

As in the original work (Devlin et al., 2019), we sum the WordPiece embedding of each token with the embeddings of the segment and position of the token. We assign the segment ID of `0` to the `[CLS]` token, all w_1 tokens, and the first `[SEP]` token; segment ID `1` is assigned to all tokens of w_2 and the final `[SEP]` token.

Lexical Relation Classifier. Original BERT feeds transformer-encoded token representations to two classifiers: MLM classifier (predicting the masked tokens), and the NSP classifier (predicting whether two sentences are adjacent). LIBERT introduces the third pretraining classifier: it predicts whether an encoded constraint pair represents a desired lexico-semantic relation (i.e., a positive example where two words stand in the relation of true semantic similarity – synonyms or hypernym-hyponym pairs) or not. Let $\mathbf{x}_{CLS} \in \mathbb{R}^H$ be the transformed vector representation of the sequence start token `[CLS]` that encodes the whole constraint (w_1, w_2). Our lexical relation predictor (LRC) is a simple softmax classifier formulated as follows:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{x}_{CLS}\mathbf{W}_{LRC}^{\top} + \mathbf{b}_{LRC}) \,, \tag{4.1}$$

with $\mathbf{W}_{LRC} \in \mathbb{R}^{H \times 2}$ and $\mathbf{b}_{LRC} \in \mathbb{R}^2$ as the classifier's trainable parameters. The relation classification loss L_{LRC} is then simply the negative log-likelihood over k instances in the training batch consisting of our lexical constraints:

$$L_{LRC} = -\sum_{k} \ln \hat{\mathbf{y}}_k \cdot \mathbf{y}_k \,, \tag{4.2}$$

where $\mathbf{y} \in \{[0, 1], [1, 0]\}$ is the true relation label for a word-pair training instance.

4.1.4 Language Understanding Evaluation

To isolate the effects of injecting external linguistic knowledge into BERT, we train base BERT and LIBERT in the same setting: the only difference is that we additionally update

[2] We use the same 30K WordPiece vocabulary as Devlin et al. (2019). Sharing WordPieces helps our word-level task as lexico-semantic relationships are similar for words composed of the same morphemes.

[3] The sign # denotes split WordPiece tokens.

the parameters of LIBERT's transformer encoder based on the gradients of the LRC loss L_{LRC} from Equation (4.2). In the first set of experiments, we probe the usefulness of injecting semantic similarity knowledge on the well-known suite of GLUE tasks (Wang et al., 2019b). Later, in Subsection 4.1.5, we additionally present an evaluation on lexical simplification, another task that has been shown to specifically benefit from semantic similarity specialization (Glavaš and Vulić, 2018).

Experimental Setup

Pretraining Data. We minimize BERT's original objective $L_{MLM} + L_{NSP}$ on training examples that we obtain from the English Wikipedia.[4] We collect the set of constraints C for the L_{LRC} term from the body of previous work on semantic specialization of static language representations (Zhang et al., 2014; Vulić et al., 2018; Ponti et al., 2018). In particular, we collect 1,023,082 synonymy pairs from WordNet (Miller, 1995) and from Roget's Thesaurus (Kipfer, 2005) and combine them with 326,187 direct hyponym-hypernym pairs (Vulić and Mrkšić, 2018) from WordNet.[5]

Fine-Tuning (Downstream) Tasks. We evaluate BERT and LIBERT on the the following tasks from the GLUE benchmark (Wang et al., 2019b), where sizes of training, development, and test data sets for each task are provided in Table 4.1:

Corpus of Linguistic Acceptability (CoLA). A binary sentence classification task, in which the model is asked to predict if sentences from linguistic publications are grammatically acceptable (Warstadt et al., 2019); note that grammaticality is related to to the AQ quality aspect of *clarity* (see Subsection 2.1.2);

Stanford Sentiment Treebank v2 (SST-2). A binary sentence classification, in which the task is to predict sentiment (positive or negative) for movie review sentences (Socher et al., 2013); note that movie reviews are argumentative texts with the argumentative intent of *recommending* (Tindale, 2007) and that sentiment analysis in movies corresponds to understanding users' stances in argument assessment;

Microsoft Research Paraphrase Corpus (MRPC). A binary sentence-pair classification, predicting whether two sentences are mutual paraphrases (Dolan and Brockett, 2005); being able to understand paraphrases, is beneficial for argument recognition;

Semantic Textual Similarity Benchmark (STS-B). A sentence-pair regression task; the task is predicting the degree of semantic similarity for a pair of sentences (Cer et al., 2017); again, this relates to the ability of recognizing similar arguments;

Quora Question Pairs (QQP). A binary classification task, in which the models are tested for their ability to recognize question paraphrases (Chen et al., 2018); as before, we need similar capabilities in argument recognition;

[4] We acknowledge that training the models on larger corpora would likely lead to better absolute downstream scores; however, the main goal of this work is not to achieve state-of-the-art downstream performance but to compare the base model against its lexically informed counterpart.

[5] Note again that similar to the work of Vulić (2018), both WordNet synonyms and direct hyponym-hypernym pairs are treated exactly the same: as positive examples for the relation of true semantic similarity.

	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	AX
# Train	8,551	67,349	3,668	5,749	363,870	392,702	392,702	104,743	2,490	–
# Dev	1,042	872	408	1,501	40,431	9,815	9,832	5,463	278	–
# Test	1,063	1,821	1,725	1,379	390,964	9,796	9,847	5,463	3,000	1,104

Table 4.1: Data set sizes for tasks in the GLUE benchmark (Wang et al., 2019b).

Multi-Genre Natural Language Inference (MNLI). Ternary NLI classification of sentence pairs (Williams et al., 2018). Two test sets are given: a matched version (MNLI-matched (MNLI-m)) in which the test domains match with training data domains, and a mismatched version (MNLI-mismatched (MNLI-mm)) with different test domains;

Question NLI (QNLI). A binary classification version of the Stanford question answering (QA) data set (Rajpurkar et al., 2016); inference capabilities are not only important in argumentative reasoning but also relate to other semantically challenging tasks, e.g., QA;

Recognizing Textual Entailment (RTE). Another NLI data set, ternary entailment classification for sentence pairs (Giampiccolo et al., 2007);

Diagnostics (AX). A small, manually curated NLI data set (i.e., a ternary classification task), with examples encompassing different linguistic phenomena relevant for entailment (Wang et al., 2019b); we have already seen two example instances from this data set in Section 3.1, when we discussed relevant types of knowledge in argumentative reasoning.[6]

Training and Evaluation. We train both BERT and LIBERT from scratch, with the configuration of the BERT$_{BASE}$ model (Devlin et al., 2019): $L = 12$ transformer layers with the hidden state size of $H = 768$, and $A = 12$ self-attention heads. We train in batches of $k = 16$ instances;[7] the input sequence length is 128. The learning rate for both models is $2 \cdot 10^{-5}$ with a warm-up over the first $1,000$ training steps. Other hyperparameters are set to the values reported by Devlin et al. (2019).

LIBERT combines BERT's MLM and NSP objectives with our LRC objective in a MTL setup. We update its parameters in a balanced alternating regime: (1) we first minimize BERT's $L_{MLM} + L_{NSP}$ objective on one batch of masked sentence pairs and then (2) minimize the LRC objective L_{LRC} on one batch of linguistic constraints.

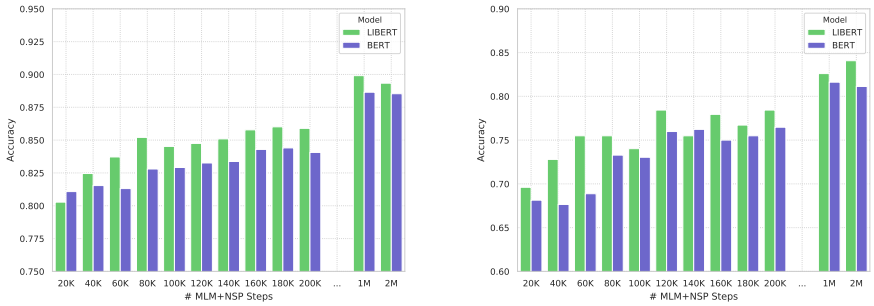
During fine-tuning, for each task, we independently find the optimal hyperparameter configurations of the downstream classifiers for the pretrained BERT and LIBERT: this implies that it is valid to compare their performances on the downstream development sets. Finally, we evaluate fine-tuned BERT and LIBERT on all 10 test sets.

[6] Following Devlin et al. (2019), we do not evaluate on Winograd NLI, given its well-documented issues.

[7] Due to hardware restrictions, we train in smaller batches than in the the original work (Devlin et al., 2019) ($k = 256$). This means that for the same number of update steps, our models will have observed less training data than the original BERT model of Devlin et al. (2019).

			CoLA MCC	SST-2 Acc	MRPC F1/Acc	STS-B Pears	QQP F1/Acc	MNLI-m Acc	MNLI-mm Acc	QNLI Acc	RTE Acc	AX MCC
1M	Dev	BERT	29.4	88.7	87.1/81.6	86.4	85.9/89.5	78.2	**78.8**	86.2	63.9	–
		LIBERT	**35.3**	**89.9**	**87.9/82.6**	**87.2**	**86.3/89.8**	**78.5**	78.7	**86.5**	**65.3**	–
		Δ	+5.9	+1.2	+0.8/+1.0	+0.8	+0.4/+0.3	+0.3	-0.1	+0.3	+1.4	–
	Test	BERT	21.5	87.9	84.8/78.8	**80.8**	68.6/87.9	78.2	**77.6**	85.8	61.3	26.8
		LIBERT	**31.4**	**89.6**	**86.1/80.4**	80.5	**69.0/88.1**	**78.4**	77.4	**86.2**	**62.6**	**32.8**
		Δ	+9.9	+1.7	+1.3/+1.6	-0.3	+0.4/+0.2	+0.2	-0.2	+0.4	+1.3	+6.0
2M	Dev	BERT	30.0	88.5	86.4/81.1	87.0	86.3/89.8	78.8	79.3	86.6	64.3	–
		LIBERT	**37.2**	**89.3**	**88.7/84.1**	**88.3**	**86.5/90.0**	**79.6**	**80.0**	**87.7**	**66.4**	–
		Δ	+7.2	+0.8	+2.3/+3.0	+1.3	+0.2/+0.2	+0.8	+0.7	+1.1	+2.1	–
	Test	BERT	28.8	89.7	84.9/79.1	81.1	69.0/88.0	78.6	78.1	**87.2**	63.4	30.8
		LIBERT	**35.3**	**90.8**	**86.6/81.7**	**82.6**	**69.3/88.2**	**79.8**	**78.8**	**87.2**	**63.6**	**33.3**
		Δ	+6.5	+1.1	+1.7/+2.6	+1.5	+0.3/+0.2	+1.2	+0.7	+0.0	+0.2	+2.5

Table 4.2: Results on the 10 GLUE tasks after 1M and 2M MLM+NSP steps with BERT and LIBERT, our lexically informed extension.



(a) SST-2 (b) MRPC

Figure 4.2: Accuracy over time for BERT (*blue*) and LIBERT (*green*) on (a) SST-2 and (b) MRPC on the corresponding development sets.

Results and Discussion

Main Results. The main results are summarized in Table 4.2: we report both development set and test set performance. After 1M MLM+NSP steps, LIBERT outperforms BERT on 8 out of 9 tasks (dev) and 8 out of 10 tasks (test). After 2M MLM+NSP steps, LIBERT is superior in all 9 tasks (dev) and 9 out of 10 tasks (test). For the test set of the tenth task (QNLI), LIBERT is on a par with BERT. While large gains are reported on CoLA, AX, and visible gains appear on SST-2 and MRPC, it is encouraging to see that slight and consistent gains are observed on almost all other tasks. These results suggest that available external lexical knowledge can be used to supplement unsupervised pretraining models with useful information which cannot be fully captured solely through large text data and their distributional signal. The results indicate that LIBERT, our lexically informed MTL method, successfully blends such curated linguistic knowledge with distributional learning signals. It also further validates intuitions from relevant work on specializing static word embeddings (Wieting et al., 2015; Mrkšić et al., 2017) that steering distributional models towards capturing true semantic similarity (as also done here) has a positive impact on language understanding applications in general.

			Coarse-grained				Fine-grained					
	Model	All	LeS	PAS	Lo	KCS	LE	MN	Fa	Re	NE	Qu
1M	BERT	26.8	24.5	38.8	19.6	12.8	17.5	29.3	04.9	22.5	15.6	**57.2**
	LIBERT	**32.8**	**35.2**	**39.7**	**25.3**	**19.4**	**28.5**	**51.4**	**18.7**	**59.2**	**18.0**	56.9
	Δ	6.0	10.7	0.9	5.7	6.6	11.0	22.2	13.8	36.7	2.4	-0.3
2M	BERT	30.8	31.3	40.0	21.7	**19.7**	21.2	51.3	09.1	59.2	**21.0**	60.5
	LIBERT	**33.3**	**40.6**	39.9	**24.5**	18.3	**33.2**	**72.0**	**21.0**	59.2	18.3	**68.4**
	Δ	2.5	9.3	-0.1	2.8	-1.4	12.0	20.7	11.9	0.0	-2.7	7.9

Table 4.3: Linguistic analysis of LIBERT's and BERT's predictions on the Diagnostic data set. The scores are R_3 coefficients between gold and predicted labels, scaled by 100, for sentences containing linguistic phenomena of interest. We report all the coarse-grained categories: *Lexical Semantics* (**LeS**), *Predicate-Argument Structure* (**PAS**), *Logic* (**Lo**), and *Knowledge and Common Sense* (**KCS**). Moreover, we report fine-grained categories for Lexical Semantics: *Lexical Entailment* (**LE**), *Morphological Negation* (**MN**), *Factivity* (**Fa**), *Redundancy* (**Re**), *Named Entities* (**NE**), and *Quantifiers* (**Qu**).

Fine-grained Analysis. To understand how lexical information corroborates the model predictions, we perform a fine-grained analysis on the Diagnostic data set (Wang et al., 2019b), measuring the performance of LIBERT on specific sets of NLI instances annotated for the linguistic phenomena they contain. We report the results in Table 4.3. As expected, *Lexical Semantics* is the category of phenomena that benefits the most (+43.7% for 1M iterations, +29.7% for 2M), but with significant gains also in phenomena related to *Logic* (+29.1% for 1M and +29.1% for 2M) and *Knowledge & Common Sense* (+51.7% for 1M). Interestingly, these results seem to suggest that knowledge about semantic similarity and lexical relations also partially encompasses factual knowledge about the world.

By inspecting even finer-grained phenomena related to *Lexical Semantics*, LIBERT outdistances its baseline by a large margin in: i) *Lexical Entailment* (+62.9% for 1M, +56.6% for 2M), as expected from the guidance of hypernym-hyponym pairs; ii) *Morphological Negation* (+75.8% for 1M, +40.4% for 2M). Crucially, the lower performance of BERT cannot be explained by the low frequency of morphologically derived words (prevented by the WordPiece tokenization), but exactly because of the distributional bias and the resulting conflation of lexico-semantic relationships. iii) *Factivity* (+281.7% for 1M, +130.8% for 2M), which is a lexical entailment between a clause and the entire sentence it is embedded in. Since it depends on specific lexical triggers (usually verbs or adverbs), it is clear that lexico-semantic knowledge better characterizes the trigger meanings. The improvement margin for *Redundancy* and *Quantifiers* fluctuate across different iterations; hence no conclusions can be drawn from the current evidence.

Performance over Time. Further, an analysis of the models' performances over time (in terms of MLM+NSP training steps for BERT and LIBERT) for one single-sentence classification task (SST-2) and one sentence-pair classification task (MRPC) is reported in Figures 4.2a and 4.2b. The scores clearly suggest that the impact of the external linguistic knowledge does not vanish over time: the gains with the lexically informed LIBERT

# Steps		BenchLS P	BenchLS R	BenchLS F1	LexMTurk P	LexMTurk R	LexMTurk F1	NNSeval P	NNSeval R	NNSeval F1
1M	BERT	.2167	.1765	.1945	.3043	.1420	.1937	.1499	.1200	.1333
	LIBERT	**.2348**	**.1912**	**.2108**	**.3253**	**.1518**	**.2072**	**.1646**	**.1318**	**.1464**
	Δ	.0181	.0147	.0163	.0210	.0098	.0135	.0147	.0118	.0131
2M	BERT	.2408	.1960	.2161	.3267	.1524	.2079	.1583	.1267	.1408
	LIBERT	**.2766**	**.2252**	**.2483**	**.3700**	**.1727**	**.2354**	**.1925**	**.1541**	**.1712**
	Δ	.0358	.0292	.0322	.0433	.0203	.0275	.0342	.0274	.0304

Table 4.4: Results on the lexical simplification candidate generation task on three data sets: BenchLS, LexMTurk, and NNSeval. For each data set we report the performance after 1M and 2M MLM+NSP steps (# Steps) with BERT and LIBERT in terms of Precision (P), Recall (R), and F1-Measure (F1).

# Steps		BenchLS Accuracy	LexMTurk Accuracy	NNSeval Accuracy
1M	BERT	.3854	.5260	.2469
	LIBERT	**.4338**	**.6080**	**.2678**
	Δ	.0484	.0820	.0209
2M	BERT	.4241	.5920	.2594
	LIBERT	**.4887**	**.6540**	**.2803**
	Δ	.0646	.0620	.0209

Table 4.5: Results on the full lexical simplification pipeline on three data sets: BenchLS, LexMTurk, and NNSeval. For each data set we report the performance after 1M and 2M MLM+NSP steps (# Steps) with BERT and LIBERT in terms of accuracy.

persist at different time steps. This finding again indicates the complementarity of useful signals encoded in large text data versus lexical resources (Faruqui, 2016; Mrkšić et al., 2017), which should be investigated more in future work.

4.1.5 Downstream Evaluation: Lexical Simplification

Task Description. The goal of lexical simplification is to replace a target word w in a context sentence S with simpler alternatives of equivalent meaning. Generally, the task can be divided into two main parts: (1) generation of substitute candidates and (2) candidate ranking, in which the simplest candidate is selected (Paetzold and Specia, 2017). Unsupervised approaches to candidate generation seem to be predominant lately (e.g., Glavaš and Štajner, 2015; Ponti et al., 2019b, *inter alia*). In this task, discerning between pure semantic similarity and broad topical relatedness (as well as from other lexical relations such as antonymy) is crucial. Consider the example: *Einstein unlocked the door to the atomic age*, where *unlocked* is the target word. In this context, the model should avoid confusion both with related words (e.g., *repaired*) and opposite words (e.g., *closed*) that fit in the context but alter the original meaning of the sentence.

Experimental Setup. In order to evaluate the simplification capabilities of LIBERT versus BERT, we adopt a standard BERT-based approach to lexical simplification (LS), BERT-LS (Qiang et al., 2020). It exploits the BERT MLM pretraining task objective for candidate generation. Given the complex word w and a context sentence S, we mask w in a new sequence S'. Next, we concatenate S and S' as a sentence pair and create the BERT-style input by running WordPiece tokenization on the sentences, adding the `[CLS]` and `[SEP]` tokens before, in-between, and after the sequence, and setting segment IDs accordingly. We then feed the input either to BERT or LIBERT, and obtain the probability distribution over the vocabulary outputted by the MLM predictor based on the masked token $p(\cdot|S, S'\backslash\{w\})$. Based on this, we select the candidates as top k words according to their probabilities, excluding morphological variations of the masked word.

For the substitution ranking component, we also follow Qiang et al. (2020). Given the set of candidate tokens C, we compute for each c_i in C a set of features: (1) BERT prediction probability, (2) loss of the likelihood of the whole sequence according to the MLM when choosing c_i instead of w, (3) semantic similarity between the fastText vectors (Bojanowski et al., 2017) of the original word w and the candidate c_i, and (4) word frequency of c_i in the top 12 million texts of Wikipedia and in the Children's Book Test corpus.[8] Based on the individual features, we rank the candidates in C and consequently, obtain a set of ranks for each c_i. The best candidate is chosen according to its average rank across all features. In our experiments, we fix the number of candidates k to 6.

Evaluation Data. We run the evaluation on three standard data sets for LS:

(1) LexMTurk (Horn et al., 2014). The data set consists of 500 English instances, which are collected from Wikipedia. The complex word and the simpler substitutions were annotated by 50 crowd workers on Amazon Mechanical Turk.

(2) BenchLS (Paetzold and Specia, 2016) is a merge of LexMTurk and LSeval (De Belder and Moens, 2010) containing 929 sentences. The latter data set focuses on text simplification for children. The authors of BenchLS applied additional corrections over the instances of the two data set in order to provide a high-quality data set.

(3) NNSeval (Paetzold and Specia, 2017) is an English data set specifically focused on text simplification for non-native speakers and consists in total of 239 prediction instances. Similar to BenchLS, the data set is based on LexMTurk, but filtered for (a) instances that contain a complex target word for non-native speakers and (b) lexical simplification candidates that were found to be non-complex by non-native speakers.

We report the scores on all three data sets in terms of Precision (P), Recall (R) and F1-Measure (F1) for the candidate generation sub-task, and in terms of the standard lexical simplification metric of Accuracy (Horn et al., 2014; Glavaš and Štajner, 2015) for the full simplification pipeline. This metric computes the number of correct simplifications (i.e., when the replacement made by the system is found in the list of gold standard replacements) divided by the total number of target complex words.

[8] A detailed description of these features can be found in the original work.

Results and Discussion. The results for BERT and LIBERT for the simplification candidate generation task and for the full lexical simplification pipeline evaluation are provided in Table 4.4 and Table 4.5, respectively. We report the performance of both models after 1M and 2M MLM+NSP pretraining steps. We observe that LIBERT consistently outperforms BERT by at least 0.9 percentage points across all evaluation setups, measures, and for all three evaluation sets. Same as in the GLUE evaluation, the gains do not vanish as we train both models for a longer period of time (i.e., compare the differences between the two models after 1M vs. 2M training steps). On the contrary, for the candidate generation task, the gains of LIBERT over BERT are even more pronounced after 2M steps. The gains achieved by LIBERT are also visible in the full simplification pipeline: for instance, on LexMTurk, replacing BERT with LIBERT yields a gain of 8.2 percentage points. In sum, these results confirm the importance of specialization for true semantic similarity for a similarity-oriented downstream task such as lexical simplification.

4.1.6 Conclusion

Given the need for lexico-semantic knowledge in argumentative reasoning (see Section 3.1), in this Section, we have presented LIBERT, a lexically informed extension of the state-of-the-art unsupervised pretraining model BERT. Our model is based on a MTL framework that allows us to steer (i.e., specialize) the purely distributional BERT model to accentuate a lexico-semantic relation of true semantic similarity (as opposed to broader semantic relatedness), which is crucial in many argumentative reasoning tasks. The framework combines standard BERT objectives with a third pretraining objective formulated as a lexical relation classification task. We evaluated the approach on CA tasks, e.g., NLI, and other NLP tasks. The gains stemming from such explicit injection of lexical knowledge from external knowledge sources into pretraining were observed for 9 out of 10 language understanding tasks from the GLUE benchmark, as well as for 3 LS benchmarks.

As shown, injecting knowledge in the pretraining stage is effective, but pretraining large transformer-based architectures is computationally expensive and therefore poses the jeopardy of ecological damage (Strubell et al., 2019). This is why, in the next Section, we focus on the injection of knowledge using an efficient adapter-based approach.

4.2 Injecting Conceptual Knowledge via Adapters

[*]Following the major success of neural language models, such as BERT or GPT-2 on a variety of language understanding tasks, recent work focused on injecting (structured) knowledge from external resources into these models, which is crucial for argumentative reasoning tasks (see Section 3.1). While on the one hand, joint pre-training (i.e., training from scratch, adding objectives based on external knowledge to the primary LM objective)

[*]This Section is adapted from: **Anne Lauscher**, Olga Majewska, Leonardo FR Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. Common Sense or World Knowledge? Investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online, November 2020, Association for Computational Linguistics.

as presented in the previous Section may be prohibitively computationally expensive, post hoc fine-tuning on external knowledge, on the other hand, may lead to the catastrophic forgetting of distributional knowledge. In this Section, we investigate models for complementing the distributional knowledge of BERT with conceptual knowledge from ConceptNet and its corresponding Open Mind Common Sense (OMCS) corpus using *adapter training*. While overall results on the GLUE benchmark paint an inconclusive picture, a deeper analysis reveals that our adapter-based models substantially outperform BERT (up to 15–20 performance points) on argumentative inference tasks that require the type of conceptual knowledge explicitly present in ConceptNet and OMCS.

4.2.1 Introduction

Self-supervised neural models like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019; Liu et al., 2019), GPT (Radford et al., 2018, 2019), or XLNet (Yang et al., 2019) have rendered language modeling a very suitable pretraining task for learning language representations that are useful for a wide range of language understanding tasks (Wang et al., 2019b,a). Although shown versatile w.r.t. the types of knowledge (Rogers et al., 2020) they encode, much like their predecessors – static word embedding models (Mikolov et al., 2013c; Pennington et al., 2014) – neural language models still only "consume" the distributional information from large corpora. Yet, a number of structured knowledge sources exist – general purpose KBs (Suchanek et al., 2007; Auer et al., 2007) and lexico-semantic networks (Miller, 1995; Liu and Singh, 2004; Navigli and Ponzetto, 2010) – encoding many types of knowledge that are underrepresented in text corpora and play an important role in argumentative reasoning (see Section 3.1, **C1**).

Starting from this observation, most recent efforts focused on injecting factual (Zhang et al., 2019; Liu et al., 2020; Peters et al., 2019a) and, as also in the previous Section, linguistic knowledge (Peters et al., 2019a) into pretrained language models and demonstrated the usefulness of such knowledge in language understanding tasks (Wang et al., 2019b,a). *Joint pretraining models*, on the one hand, augment distributional LM objectives with additional objectives based on external resources (Yu and Dredze, 2014; Nguyen et al., 2016) and train the extended model from scratch. We proposed such a procedure in Section 4.1. For models like BERT, however, this implies computationally expensive retraining from scratch of the encoding transformer network. *post hoc fine-tuning* models (Zhang et al., 2019; Liu et al., 2020; Peters et al., 2019a), on the other hand, use the objectives based on external resources to fine-tune the encoder's parameters, pretrained via distributional LM objectives. If the amount of fine-tuning data is substantial, however, this approach may lead to (catastrophic) forgetting of distributional knowledge obtained in pretraining (Goodfellow et al., 2014; Kirkpatrick et al., 2017).

In this Section, similar to the concurrent work of Wang et al. (2020), we resort to the recently proposed *adapter-based fine-tuning* paradigm (Rebuffi et al., 2018; Houlsby et al., 2019), which remedies for shortcomings of both joint pretraining and standard post hoc fine-tuning. Adapter-based training injects additional parameters into the encoder and only tunes their values: the original transformer parameters are kept fixed. Because of freezing these layers, adapter training preserves the distributional information obtained

in LM pretraining, without the need for any distributional (re-)training. While (Wang et al., 2020) inject factual knowledge from Wikidata (Vrandečić and Krötzsch, 2014) into BERT, in this work, we investigate two resources that are commonly assumed to contain *general-purpose* and *common sense* knowledge,[9] types of knowledge that are useful for argumentative reasoning tasks: ConceptNet (Liu and Singh, 2004; Speer et al., 2017) and the Open Mind Common Sense (OMCS) corpus (Singh et al., 2002), from which the ConceptNet graph was (semi-)automatically induced. For our first model, dubbed CN-ADAPT, we first create a synthetic text corpus by randomly traversing the ConceptNet graph and then learn adapter parameters with MLM training (Devlin et al., 2019) on that synthetic corpus. For our second model, named OM-ADAPT, we learn the adapter parameters via MLM training directly on the OMCS corpus.

As in Section 4.1, we evaluate both models on the GLUE benchmark, which contains a variety of tasks relevant for CA, where we observe limited improvements over BERT on a subset of GLUE tasks. However, a more detailed inspection reveals large improvements over the base BERT model (up to 20 Matthews correlation points) on language inference (NLI) subsets labeled as requiring World Knowledge or knowledge about Named Entities. Investigating further, we relate this result to the fact that ConceptNet and OMCS contain much more of what in downstream is considered to be factual world knowledge than what is judged as common sense knowledge. Our findings pinpoint the need for more detailed analyses of the compatibility between (1) the types of knowledge contained by external resources; and (2) the types of knowledge that benefit concrete downstream tasks; within the emerging body of work on injecting knowledge into pretrained transformers.

4.2.2 Knowledge Injection Models

In this work, we are primarily set to investigate if injecting specific types of knowledge (given in the external resource) benefits downstream argumentative inference that clearly requires those exact types of knowledge. Because of this, we resort to arguably the most straightforward mechanisms for injecting the ConceptNet and OMCS information into BERT and leave the exploration of potentially more effective knowledge injection objectives for future work. We inject the external information into adapter parameters of the adapter-augmented BERT (Houlsby et al., 2019) via BERT's natural objective – MLM, explained in Section 2.2.2. OMCS, already a corpus in natural language, is directly subjectable to MLM training – we filtered out non-English sentences. To subject ConceptNet to MLM training, we need to transform it into a (synthetic) corpus.

Unwrapping ConceptNet. Following established previous work (Perozzi et al., 2014; Ristoski and Paulheim, 2016), we induce a synthetic corpus from ConceptNet by randomly traversing its graph. We then convert the relation strings, which are part of the obtained walks, into natural language phrases (e.g., `synonyms` to *is a synonym of*) and duplicate the object node of a triple, using it as the subject for the next sentence. For example, from the path "*alcoholism* $\xrightarrow{causes}$ *stigma* $\xrightarrow{hasContext}$ *christianity* $\xrightarrow{partOf}$ *religion*"

[9] Our results in Subsection 4.2.3 scrutinize this assumption.

we create the text "*alcoholism causes stigma. stigma is used in the context of christianity. christianity is part of religion.*". We set the walk lengths to 30 relations and sample the starting and neighboring nodes from uniform distributions. In total, we performed 2,268,485 walks, resulting in a corpus of 34,560,307 synthetic sentences.

Adapter-Based Training. We follow Houlsby et al. (2019) and adopt the adapter-based architecture for which they report solid performance across the board. We inject *bottleneck adapters* into BERT's transformer layers. In each transformer layer, we insert two bottleneck adapters: one after the multi-head attention sub-layer and another after the feed-forward sub-layer. Let $\mathbf{X} \in \mathbb{R}^{T \times H}$ be the sequence of contextualized vectors (of size H) for the input of T tokens in some transformer layer, input to a bottleneck adapter. The bottleneck adapter, consisting of two feed-forward layers and a residual connection, yields an output defined as follows:

$$\text{Adapter}(\mathbf{X}) = \mathbf{X} + f\left(\mathbf{X}\mathbf{W}_d + \mathbf{b}_d\right)\mathbf{W}_u + \mathbf{b}_u\,, \tag{4.3}$$

where the matrices $\mathbf{W}_d$ (with the bias $\mathbf{b}_d$) and $\mathbf{W}_u$ (with the bias $\mathbf{b}_u$) are the adapter's parameters, that is, the weights of the linear down-projection and up-projection sub-layers and f is the non-linear activation function. Matrix $\mathbf{W}_d \in \mathbb{R}^{H \times m}$ compresses the vectors in $\mathbf{X}$ to the *adapter size* $m << H$, and the matrix $\mathbf{W}_u \in \mathbb{R}^{m \times H}$ projects the activated down-projections back to the transformer's original hidden size H.

4.2.3 Evaluation

We first briefly describe the downstream tasks and training details and then proceed with the discussion of results obtained with our adapter models.

Experimental Setup

Downstream Tasks. We evaluate BERT and our two adapter-based models, CN-ADAPT and OM-ADAPT, with injected knowledge from ConceptNet and OMCS, respectively, on the tasks from the GLUE benchmark (Wang et al., 2019b), described in Section 4.1.4. The benchmark contains a large variety of tasks, which are relevant to CA.

Training Details. We inject our adapter layers into a BERT$_{BASE}$ model (12 transformer layers with 12 attention heads each; $H = 768$) pretrained on lowercased corpora. Following (Houlsby et al., 2019), we set the size of all adapters to $m = 64$ and use gaussian error linear unit (Hendrycks and Gimpel, 2016) as the adapter activation function f. We train the adapter parameters with the Adam algorithm (Kingma and Ba, 2015) and set the initial learning rate to $1 \cdot 10^{-4}$, with 10000 warm-up steps and the weight decay factor of 0.01. In the downstream fine-tuning, we train in batches of size 16 and limit the input sequences to $T = 128$ WordPiece tokens. For each task, we find the optimal hyperparameter configuration by searching in the following grid: learning rate $l \in \{2 \cdot 10^{-5}, 3 \cdot 10^{-5}\}$ and epochs in $n \in \{3, 4\}$.

Model	CoLA MCC	SST-2 Acc	MRPC F1	STS-B Spear	QQP F1	MNLI-m Acc	MNLI-mm Acc	QNLI Acc	RTE Acc	AX MCC	Avg –
BERT Base	52.1	93.5	**88.9**	85.8	71.2	**84.6**	83.4	90.5	66.4	34.2	75.1
OM-ADAPT (25K)	49.5	93.5	88.8	85.1	71.4	84.4	83.5	**90.9**	67.5	35.7	75.0
OM-ADAPT (100K)	**53.5**	93.4	87.9	**85.9**	71.1	84.2	**83.7**	90.6	68.2	34.8	75.3
CN-ADAPT (50K)	49.8	**93.9**	**88.9**	85.8	**71.6**	84.2	83.3	90.6	**69.7**	37.0	**75.5**
CN-ADAPT (100K)	48.8	92.8	87.1	85.7	71.5	83.9	83.2	90.8	64.1	**37.8**	74.6

Table 4.6: Results on test portions of GLUE benchmark tasks. Numbers in parentheses next to adapter-based models (25K, 50K, 100K) indicate the number of update steps of adapter training on the synthetic ConceptNet corpus (for CN-ADAPT) or on the original OMCS corpus (for OM-ADAPT). **Bold**: the best score in each column.

Results and Analysis

GLUE Results. Table 4.6 reveals the performance of CN-ADAPT and OM-ADAPT in comparison with BERT$_{BASE}$ on the GLUE evaluation tasks.[10] We show the results for two snapshots of OM-ADAPT, after 25K and 100K update steps, and for two snapshots of CN-ADAPT, after 50K and 100K steps of adapter training. Overall, none of our adapter-based models with injected external knowledge from ConceptNet or OMCS yields significant improvements over BERT Base on GLUE. However, we observe substantial improvements (of around 3 points) on RTE and on the Diagnostics NLI data set (AX), which encompasses inference instances that require a specific type of knowledge.

Since our adapter models draw specifically on the conceptual knowledge encoded in ConceptNet and OMCS, we expect the positive impact of injected external knowledge – assuming effective injection – to be most observable on test instances that target the same types of conceptual knowledge. To investigate this further, we measure the model performances across different categories of the Diagnostic NLI data set (as in Section 4.1). This allows us to tease apart inference instances which truly test the efficacy of our knowledge injection methods. We show the results obtained on different categories of the Diagnostic NLI data set in Table 4.7. The improvements of our adapter-based models over BERT Base on these phenomenon-specific subsections of the Diagnostics NLI data set are generally much more pronounced: e.g., OM-ADAPT (25K) yields a 7% improvement on inference that requires factual or common sense knowledge (KCS), whereas CN-ADAPT (100K) yields a 6% boost for inference that depends on lexico-semantic knowledge (LeS). These results suggest that (1) ConceptNet and OMCS do contain the specific types of knowledge required for these inference categories and that (2) we managed to inject that knowledge into BERT by training adapters on these resources.

[10] Note that these results are not comparable with Table 4.2, as the original BERT checkpoint from which we start here, has seen more and slightly different data than the model we trained from scratch in Section 4.1.

Model	LeS	KCS	Lo	PAS	All
BERT Base	38.5	20.2	26.7	39.6	34.2
OM-Adapt (25K)	39.1	**27.1**	26.1	39.5	35.7
OM-Adapt (100K)	37.5	21.2	27.4	41.0	34.8
CN-Adapt (50K)	40.2	24.3	30.1	**42.7**	37.0
CN-Adapt (100K)	**44.2**	25.2	**30.4**	41.9	37.8

Table 4.7: Breakdown of Diagnostics NLI performance (Matthews correlation), according to information type needed for inference (coarse-grained categories): *Lexical Semantics* (**LeS**), *Knowledge and Common Sense* (**KCS**), *Logic* (**Lo**), and *Predicate-Argument Structure* (**PAS**).

Model	CS	World	NE
BERT Base	**29.0**	10.3	15.1
OM-Adapt (25K)	28.5	25.3	31.4
OM-Adapt (100K)	24.5	17.3	22.3
CN-Adapt (50K)	25.6	21.1	26.0
CN-Adapt (100K)	24.4	**25.6**	**36.5**

Table 4.8: Results (Matthews correlation) on *Common Sense* (**CS**), *World Knowledge* (**World**), and *Named Entities* (**NE**) categories of the Diagnostic NLI data set. Our models outperform BERT on World and NE knowledge.

Fine-Grained Knowledge Type Analysis. In our final analysis, we "zoom in" our models' performances on three fine-grained categories of the Diagnostics NLI data set – inference instances that require Common Sense Knowledge (CS), World Knowledge (World), and knowledge about named entities (NE), respectively. The results for these fine-grained categories are given in Table 4.8. These results show an interesting pattern: our adapter-based knowledge-injection models massively outperform BERT Base (up to 15 and 21 MCC points, respectively) for NLI instances labeled as requiring World Knowledge or knowledge about Named Entities. In contrast, we see drops in performance on instances labeled as requiring common sense knowledge. This initially came as a surprise, given the common belief that OMCS and ConcepNet contain the so-called *common sense* knowledge. A manual follow-up analysis of the diagnostic test instances from both CS and World categories uncovers a noticeable mismatch between the kind of information that is considered common sense in KBs like ConceptNet and what is considered common sense knowledge in the downstream. In fact, the majority of information present in ConceptNet and OMCS falls under the World Knowledge definition of the Diagnostic NLI data set, including factual geographic information (`stockholm [partOf] sweden`), domain knowledge (`roadster [isA] car`) and specialized terminology (`indigenous [synonymOf] aboriginal`). Diagnostic NLI examples from the World Knowledge and Common Sense categories are depicted in Table 4.9. In contrast, many of the common sense inference instances require complex, high-level reasoning, understanding metaphorical and idiomatic meaning, and making far-reaching connections. In such cases, explicit conceptual links often do not suffice for a correct inference and much of the required knowledge is not explicitly encoded in the external resources. Consider, e.g., the following common sense NLI instance: [`premise:` *My jokes fully reveal my character* `; hypothesis:` *If everyone believed my jokes, they'd know exactly who I was* `; entailment`]. While ConceptNet and OMCS may associate *character* with *personality* or *personality* with *identity*, the knowledge that the phrase *who I was* may refer to *identity* is beyond these resources.

Knowledge	Premise	Hypothesis	ConceptNet?
World	*The sides came to an agreement after their meeting in **Stockholm**.*	*The sides came to an agreement after their meeting in **Sweden**.*	`stockholm [partOf] sweden`
	*Musk decided to offer up his personal Tesla **roadster**.*	*Musk decided to offer up his personal **car**.*	`roadster [isA] car`
	*The Sydney area has been inhabited by **indigenous** Australians for at least 30,000 years.*	*The Sydney area has been inhabited by **Aboriginal** people for at least 30,000 years.*	`indigenous [synonymOf] aboriginal`
Common Sense	*My jokes fully reveal my character.*	*If everyone believed my jokes, they'd know exactly who I was.*	
	The systems thus produced are incremental: dialogues are processed word-by-word, shown previously to be essential in supporting natural, spontaneous dialogue.	*The systems thus produced support the capability to interrupt an interlocutor mid-sentence.*	
	He deceitfully proclaimed: "This is all I ever really wanted."	*He was satisfied.*	

Table 4.9: Premise-hypothesis examples from the diagnostic NLI data set tagged for common sense and world knowledge, and relevant ConceptNet relations, where available.

4.2.4 Conclusion

In this Section, we presented two simple strategies for injecting knowledge from ConceptNet and OMCS, respectively, into BERT via bottleneck adapters. Additional adapter parameters store the external knowledge and allow for the preservation of the corpus knowledge obtained in the pretraining of the original transformer parameters. We demonstrated the effectiveness of these models in language understanding settings that require precisely the type of knowledge one finds in ConceptNet and OMCS, in which our adapter-based models outperform BERT up to 20 performance points. Our findings stress the importance of detailed analyses comparing the types of knowledge found in external sources and the types of knowledge needed in concrete reasoning tasks.

To address the challenge of underrepresented external knowledge in distributional language representation models for CA (**C1**), in this Chapter, we presented two case studies grounded for knowledge injection from external sources: (1) injection of lexico-semantic constraints via an additional pretraining objective, and (2) injection of conceptual knowledge via adapter layers. Next, we address (**C2**), domain-specific knowledge.

Chapter 5

Domain Knowledge

[*]Given that argumentation exists in a large variety of domains, e.g., scientific writing, a challenge for language representations for CA is their suitability for domain-specific scenarios (**C2**, see Section 3.2). As discussed before, one possibility is to identify domain-specific data from which, in turn, domain-specific language representations can be induced. However, this might not always result in improved performance in downstream tasks, as the degree of specificity of the particular domain along the hierarchy of topics and genres can correlate with the amount of data available. This can imply a trade-off between bigger and more noisy vs. smaller and more homogeneous data, affecting the quality of the resulting embeddings. In order to address the challenge of domain-specificity in language representations for computational argumentation, we focus on analyzing this trade-off for the case of scientific argumentation. In particular, we study the impact of employing general vs. general scientific vs. CL-specific corpora in order to induce word embeddings for semantically characterizing citations in NLP and CL publications in terms of polarity and purpose, tasks which fall under the category of *scitorics* (see Sections 2.1.3 and 2.1.4, respectively). To this end, we frame polarity and purpose detection as classification tasks and investigate the performance of convolutional networks with general and domain-specific word embeddings on these tasks. Our best-performing model outperforms previously reported results on a benchmark data set by a wide margin.

5.1 Introduction

Citations play a vital role in scientific argumentation as they connect the authors' monological argument to the overall scientific discourse (see Section 2.1.3). Acknowledging the importance of these references, citation graphs and citation indices have long been supporting various analyses in the sociology of science (Garfield, 1955; Garfield et al., 1984). As such, citation graphs are used to detect research communities and retrace the evolution

[*]This Chapter is adapted from: **Anne Lauscher**, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. In *Proceedings of the 6th International Workshop on Mining Scientific Publications*, pages 24–28, Toronto, ON, Canada, December 2017, ACM Press.

of ideas within the scientific discourse over time. Various measures reflecting the impact of a publication, journal, or author exploit only raw citation counts. For example, the *h-index* (Hirsch, 2005) is commonly used to assess the impact of a researcher.

Purely quantitative measures alone, however, may often be misleading regarding the positive impact of some research. For example, a publication on widely-criticized work will still have a large number of citations. Being based on simple counts, quantitative scientometric measures reflect quantitative rather than qualitative aspects of research – we are not only interested in how often a work is cited, but also why it is being cited, and accordingly, which argumentative intent caused the citer to refer to another work. Knoth and Herrmannova (2014) recently introduced the term *semantometrics* to describe a new category of scientometric measures that account for qualitative aspects of citations. Automated qualitative analysis of publications is challenging, as it requires processing the textual content of all citing publications. Historically, models for qualitative analysis of citations employ a range of heavily manually-engineered features.

In this Chapter, we evaluate models that require virtually no feature engineering on tasks of citation polarity and purpose classification while, at the same time, we seek to understand the effect of domain-specificity of our employed language representations (see Section 3.2). Citation polarity (also known as citation sentiment classification) assigns a polarity (positive, negative, or neutral) to a citation, considering the citation context (Athar, 2011). Citation purpose classification (also known as citation function and citation intent classification, see Section 2.1.4) is a more fine-grained type of analysis that aims to provide a functional characterization of a citation (Teufel et al., 2006).

The contributions of this work are twofold. First, following a series of successful applications of convolutional neural networks (CNNs; LeCun and Bengio, 1998) in short text classification (Kim, 2014; Kalchbrenner et al., 2014), we present the first CNN application in the area of qualitative citation analysis. Using CNNs allows us to avoid extensive feature-engineering present in existing semantometric models. Secondly, we investigate the impact of using domain-specific word embeddings.[1]

Experimental results on a benchmark data set show that our best performing models outperform previously reported results for both classification tasks by a wide margin.

5.2 Related Work

A significant body of work exists both for citation polarity classification (Athar, 2011; Jochim and Schütze, 2012; Abu-Jbara et al., 2013; Kim and Thoma, 2015) and citation purpose classification (Teufel et al., 2006; Dong and Schäfer, 2011).

Athar (2011) first worked on citation polarity classification, combining a range of lexical, dictionary-based, and syntactic features with a linear support vector machines (SVM) classifier. Similarly, Jochim and Schütze (2012) fed a range of features for citation polarity classification to a maximum entropy classifier, whereas Kim and Thoma (2015) trained an SVM model with radial basis function (RBF) kernel using occurrence statistics of

[1] The domain-specific word vector representations produced in this research are available for download at: `https://github.com/anlausch/scientific-domain-embeddings`.

n-grams in an annotated corpus as features. Teufel et al. (2006) classified function of citations into one of 12 categories. They employed a k-NN classifier using cue phrases, self-citation, and the position of the citing sentence as features. Dong and Schäfer (2011) analyzed the effectiveness of different feature groups (e.g., positional, lexical, syntactic) for function classification over a range of classifiers, pointing to syntactic features as being most useful. Xu et al. (2013) focused on discerning functional from perfunctory citations, using a combination of textual and external features. Abu-Jbara et al. (2013) and Jha et al. (2016) addressed both polarity and purpose classification with an SVM employing an extended set of features such as speculation cues and self-citation indicators. All of the above models rely on heavy manual feature design and feature engineering.

Jochim and Schütze (2014) were the first to apply a deep learning model to the citation polarity classification. In a domain-adaption setting, they trained a marginalized stacked denoising autoencoders (mSDA) model on product reviews and used it to predict the polarity of citations. To the best of our knowledge, there have been no attempts to apply convolutional neural networks, achieving state-of-the-art performance on a range of text classification tasks (Kim, 2014; Kalchbrenner et al., 2014; Severyn and Moschitti, 2015; Shrestha et al., 2017, *inter alia*), to citation context analysis.

5.3 Classification Models

Our primary goals are to avoid tedious feature engineering for citation classification and to understand the trade-off between larger, more heterogeneous vs. smaller, more homogeneous corpora for inducing word embeddings. Here, we describe two models that satisfy the first criterion, and which we will employ in our experiments towards understanding the degree of domain-specificity that is beneficial for the models' performances.

5.3.1 Convolutional Neural Network

CNNs (LeCun and Bengio, 1998), introduced to the NLP community by Collobert and Weston (2008), exhibit state-of-the-art performance on a range of text classification tasks (Kim, 2014; Severyn and Moschitti, 2015; Shrestha et al., 2017). A CNN is a feed-forward neural network consisting of one or more convolution layers. Each convolution layer consists of a set of filters. When applied to textual data, convolutions of filters and text slices – matrices produced by sequentially sliding a window of size k over the embedding-based representation of text – are computed. Each convolution layer is followed by a pooling layer, which subsamples the output of the convolution layer (e.g., by taking N maximal values). This architecture allows the network to capture local aspects, i.e., the most informative k-grams in text for the task. We use a CNN with a single convolution and single max-pooling layer. We use rectified linear unit activation and optimize the network parameters with the RMSprop algorithm (Tieleman and Hinton, 2012) to minimize the cross-entropy loss. To be subdued to a CNN, texts must be represented as numerical vectors, which can be achieved by using word embeddings (Mikolov et al., 2013c; Pennington et al., 2014, *inter alia*). More precisely, each text is represented as a matrix of size $N \times L$, where N is the length of the text (in number of tokens), and L

is the length of the word embeddings. Because the CNN expects the same number of features for all texts, all instances must be of equal length. In our experiments, we set N to the length of the longest text in the data set and pad all other sentences with a special padding token to which we assign a random embedding vector.

5.3.2 SVM with Embedding Features

Having in mind (1) that SVM has been widely used for citation polarity and purpose classification and (2) that by employing word embeddings, we may still avoid manual feature engineering and study the domain-specificity of those, we decided to compare CNN's performance to that of an SVM model using the semantic embedding of the text. We compute the embedding of the text as weighted continuous bag of words (WCBOW) aggregation of word embeddings (Mikolov et al., 2013c):

$$\text{WCBOW}(t_1, \ldots, t_k) = \frac{1}{\sum_{i=1}^{k} a_i} \sum_{i=1}^{k} a_i \mathbf{t_i}\,, \tag{5.1}$$

where t_i is the i-th token of a k-token-long text, $\mathbf{t_i}$ is the word embedding of the token t_i, and a_i is the TF–IDF weight of the token. We compute the TF–IDF weight on the training set and use it in order to reflect the relative informativeness of words. This results in a single aggregate embedding vector for each text, which we then feed to the SVM classifier with a radial basis function (RBF) kernel.

5.3.3 General vs. Domain-Specific Word Embeddings

Both above models use static language representations – semantic vectors that capture the meaning of words (see Section 2.2.2). As discussed before, those representations are generally trained in a self-supervised manner on large general-domain corpora, e.g., Wikipedia. However, in all our experiments, we classify argumentative texts involving citations from a specific subdomain of scientific publications: from the area of NLP and CL (see Section 5.4). A research question that naturally arises and which relates to one of the five main challenges in the context of language representations for CA (**C2**, see Section 3.2) is whether domain-specific word embeddings, i.e., static word embeddings trained on an in-domain corpus, would lead to better classification performance than word embeddings trained on general-domain corpora. To investigate the effects of using domain-specific embeddings, we evaluate three different variants of the above two models, employing (1) general word embeddings, (2) embeddings trained on domain corpora consisting of scientific publications from various research fields, and (3) embeddings trained on a narrowly in-domain corpus of publications from the area of NLP and CL.

5.4 Data

We briefly describe the corpora used to train different language representation spaces and the classification data set used in our experiments.

Data set	Size (in tokens)
Wikipedia + GigaWord	6,000,000,000
CORE corpus	2,530,738,678
ACL Reference Corpus	81,365,802

Table 5.1: Corpora used to train word embeddings. The corpus with the highest degree of domain-specificity with respect to our target argumentative domain is the smallest (ACL Reference Corpus), while our largest corpus is the least domain-specific (Wikipedia + Giga Word).

Classification	Label	Proportion
Polarity	*positive*	32.6%
	negative	12.4%
	neutral	55.0%
Purpose	*criticizing*	16.3%
	comparison	8.1%
	use	18.0%
	substantiating	8.0%
	basis	5.3%
	neutral	44.3%

Table 5.2: Citation label distributions.

5.4.1 Word Embeddings Corpora

We experimented with 50-dimensional GloVe embeddings (Pennington et al., 2014) trained on three different corpora: (1) general domain (Wikipedia + GigaWord corpus),[2] (2) the CORE corpus of scientific publications aggregated from Open Access repositories and journals (Knoth and Zdrahal, 2012), and (3) the Association for Computational Linguistics (ACL) Reference Corpus[3] (Bird et al., 2008). We compare the sizes of these three corpora in Table 5.1. The CORE corpus is significantly larger than the ACL Reference Corpus, as it aggregates publications over various disciplines, whereas the ACL Reference Corpus only contains publications related to CL and NLP. Accordingly, the sizes of these corpora are inversely correlated with their homogeneity.

5.4.2 Citation Classification Corpus

We use the data set from Abu-Jbara et al. (2013) and Jha et al. (2016) in our experiments. In total, it contains 3,271 citation context instances, each consisting of four sentences: the sentence citing a given target reference, one preceding sentence, and two following sentences. All of these contexts have been annotated with citation polarity and citation purpose information. Citation polarity was annotated with one of three labels –*positive*, *negative*, and *neutral*. Furthermore, one of six categories has to be chosen as a label for the citation purpose: *criticism*, *comparison*, *use*, *substantiation*, *basis*, and *neutral*. The distribution of instances over the different categories for both polarity and purpose are shown in Table 5.2. In addition to assigning polarity and purpose labels to citation contexts, annotators labeled each sentence of the context as being informative for the polarity and polarity classification or not. We observe that the data set is heavily skewed towards the least informative *neutral* class for both classification dimensions.

[2] `http://nlp.stanford.edu/data/glove.6B.zip`

[3] Version 20160301, ParsCit structured XML.

5.5 Evaluation

We describe the experimental setting, the model variants and baselines we evaluate, and the performance levels they reach for citation polarity and purpose classification.

5.5.1 Models and Baselines

We evaluate the two models from Section 5.3: CNN and SVM with aggregate text embeddings. For each of these two models we evaluate three variants, using static language representations trained on different corpora: General, CORE, and ACL (see Section 5.4). We compare our models with the following baselines:

(1) Given the heavily skewed label distributions for both tasks, we use the majority class baseline predicting the most frequent class in the training set (*neutral* in both cases);

(2) We also evaluate a linear SVM with discrete TF–IDF-weighted bag-of-words features. Comparing this baseline with the embedding-based SVM model provides insights into usefulness of word embeddings for citation classification tasks;

(3) Last, we report the performance of the SVM model with a rich set of features from Jha et al. (2016), as they evaluate their model on the same data set (Abu-Jbara et al., 2013).

5.5.2 Experimental Setting

In order to make our results comparable to those reported by Jha et al. (2016), we evaluate the models in 10-fold cross validation (CV) setting. More precisely, for each model, we execute a nested CV evaluation, where for each fold of the outer CV loop, we optimize the model's hyperparameters via grid search in the inner CV. The reported performance is macro-averaged over the folds of the outer CV loop.

5.6 Results

Polarity classification results are shown in Table 5.3 and purpose classification results in Table 5.4. Surprisingly, the linear SVM with bag-of-words features is a very competitive baseline on both classification tasks. More surprisingly, it performs 8 percentage points (polarity) and 14 percentage points (purpose) better than the SVM model from Jha et al. (2016), which uses a much richer set of features. This is probably because Jha et al. (2016), reportedly, do not optimize their model's hyperparameters. Also, the SVM models with embedding features do not outperform the linear SVM baseline, regardless of the corpus used to train the embeddings. All this suggests that citation polarity and purpose are strongly indicated by a particular set of lexical clues.

The CNN model has a slight edge over all SVM-based models, but the performance gains are much lower than reported for other text classification tasks (Kim, 2014; Shrestha et al., 2017). The in-domain specialization of the language representations does not seem to play a significantly positive role. The best results are obtained using the super-domain CORE embeddings. The in-domain ACL embeddings are probably of lower quality due to the much smaller size of the training corpus. This confirms the expected trade-off.

Model	P	R	F1
Majority	18.3	33.3	23.6
Jha et al. (2016)	67.1	70.6	68.8
SVM TF–IDF	77.9	76.3	77.1
SVM General emb.	79.1	74.0	76.5
SVM CORE emb.	83.2	72.1	75.3
SVM ACL emb.	81.3	75.4	77.3
CNN General emb.	82.0	75.9	**78.8**
CNN CORE emb.	81.8	76.1	**78.8**
CNN ACL emb.	81.2	75.4	78.2

Table 5.3: Polarity classification results.

Model	P	R	F1
Majority	7.4	16.7	10.3
Jha et al. (2016)	54.9	62.5	58.4
SVM TF–IDF	74.3	70.9	72.6
SVMGeneral emb.	86.8	64.7	74.1
SVM CORE emb.	81.7	66.2	73.1
SVM ACL emb.	81.7	66.0	73.0
CNN General emb.	79.9	68.2	73.6
CNN CORE emb.	80.8	68.8	**74.3**
CNN ACL emb.	76.7	68.4	72.3

Table 5.4: Purpose classification results.

Classification	Model	Context	P	R	F1
Polarity	CNN CORE emb.	Citing Sentence	81.8	76.1	78.8
	CNN CORE emb.	Gold Standard	85.8	78.7	**82.1**
	SVM CORE emb	Citing Sentence	83.2	72.1	75.3
	SVM CORE emb.	Gold Standard	84.1	75.6	79.6
Purpose	CNN CORE emb.	Citing Sentence	80.8	68.8	74.3
	CNN CORE emb.	Gold Standard	85.2	73.3	**78.9**
	SVM CORE emb.	Citing Sentence	81.7	66.2	73.1
	SVM CORE emb.	Gold Standard	84.8	69.2	76.2

Table 5.5: Impact of the choice of the citation context on the classification results.

Table 5.5 shows the classification results of the SVM and CNN models with CORE embedding features when using different citation context sizes. As it can be seen, for all models, the performance improves by around 3 to 4 percentage points when the gold standard citation context is taken into account instead of only the directly citing sentence. This suggests that a fine-grained identification of the citation context is an important step that needs to precede the citation classification tasks at hand.

When analyzing the results in depth, we notice that for both classification tasks, most errors that happened correspond to a misclassification of a context into the category *neutral*. This type of error occurred in 61% of all the misclassifications that happened in the purpose classification and in 59% of the errors which occurred when classifying polarity. We hypothesize that this may be due to the skewness of the benchmark data set we used. Another frequent error that happened in the purpose classification is the misclassification of an instance of the category *basis* as *use*, which is probably due to the high interrelation of those two purposes. Similarly, all purpose classifiers often confuse the instances of the class *comparison* with instances of the class *criticizing*.

5.7 Conclusion

Understanding citations plays an important role in the argumentative analysis of scientific publications: they connect the authors' argumentation to the overall scientific debate and act as central tools in building a convincing scientific argument (see Section 2.1.3). Existing models for the semantic classification of citations rely on extensive feature engineering. In this Chapter, we investigated two models that do not require any manual feature design – CNN and SVM with aggregate text embeddings – on citation polarity and citation purpose classification tasks. The investigated models outperform previously reported results on a benchmark data set by a wide margin. However, only CNN models slightly outperform a simple linear SVM with lexical features. This suggests that lexical clues alone quite strongly indicate citation polarity and purpose.[4] We also find that using highly domain-specific word embeddings provides no observable performance boost, confirming the expected trade-off between larger and more general vs. smaller and more domain-specific corpora. In the next Chapter, we investigate the complementarity of knowledge across a variety of argumentative analysis tasks (**C3**).

[4]Note that this work was performed in 2017, before the era of contextualized embedding models. We expect that employing such language representations, e.g., BERT and the domain-specific SciBERT (Beltagy et al., 2019), will yield better results compared to employing static representations.

Chapter 6

Complementarity of Knowledge across Tasks

As outlined in Section 3.3, the complexity of the computational argumentation field with its variety of interrelated and interdependent problems (see Section 2.1.4) naturally lends itself to sharing knowledge encoded in language representations (**C3**). Within this frame, preceding work has shown the effectiveness of multi-task learning (MTL) on argumentative tasks for low-resource scenarios (Schulz et al., 2018). In this Chapter, we employ such inductive transfer learning techniques (see Section 2.2.3) for addressing two specific problems in CA: (1) we acknowledge the argumentative, multi-layered nature of scientific text (discussed in Section 2.1.3) and study the role of argumentation with respect to other scitorics[1] with neural MTL models. To this end, we extend a corpus of scientific literature with an additional fine-grained argumentation annotation layer. We then demonstrate performance improvements when coupling argumentation with the other rhetorical analysis problems in a joint MTL setup, thereby sharing knowledge in the language representations. (2) We move from the special case of scientific argumentation to AQ in multiple domains of online writing. Here, especially the theory-based perspective (Wachsmuth et al., 2017b), as explained in Subsection 2.1.2, remains underexplored. So far, no large-scale corpus annotated with theory-based AQ dimensions (logic, rhetoric, and dialectic) allowing for training computational models which exploit the complementarity of knowledge across tasks is in place. We close this research gap by presenting GAQCorpus, the first English multi-domain theory-based argument quality corpus. We further demonstrate performance improvements in two settings exploiting complementarity of knowledge in contextualized embedding models: (a) in a flat and a hierarchical multi-task learning setup, and (b) in a sequential task transfer setup (STILT).

[1] The rhetorical aspects of scientific writing which we discussed in Section 2.1.4.

6.1 Complementarity of Knowledge across Scitorics

[*]The exponential growth in the number of scientific publications yields the need for the effective automatic analysis of the rhetorical aspects of scientific writing, which we collectively dub *scitorics* (see Section 2.1.4). Acknowledging the argumentative nature of scientific text, in this Section, we investigate the link between the argumentative structure of scientific publications and other rhetorical aspects such as discourse categories or citation contexts. To this end, we firstly (1) augment a corpus of scientific publications annotated with four layers of rhetoric annotations with argumentation annotations. Concretely, we add the argumentative annotations to the existing Dr. Inventor Corpus (Fisas et al., 2015, 2016), already annotated for four other rhetorical aspects. We analyze the annotated argumentative structures and investigate the relations between argumentation and other rhetorical aspects of scientific writing, such as discourse roles and citation contexts. Secondly, (2) we investigate the complementarity of knowledge in language representations (**C3**, see Section 3.3) using neural multi-task learning (MTL) architectures (discussed in Section 2.2.3) combining argument extraction with a set of rhetorical classification tasks. By coupling the rhetorical classifiers with the extraction of argumentative components in a joint MTL setting, we obtain statistically significant performance gains for the different rhetorical analysis tasks.[2]

6.1.1 Introduction

Scientific publications, as highly argumentative texts in research (Gilbert, 1977), are carefully composed documents written to convince the reader of the validity and merit of the researchers' work (see Section 2.1.3). As such, they are inherently argumentative and often adhere to well-trodden rhetorical patterns and follow established structures and practices of the respective research field. As demonstrated during the COVID-19 pandemic, knowledge access is vital when it comes to societal crises. However, the accelerated growth of scientific literature (Bornmann and Mutz, 2015) makes the exploration and analysis of relevant publications increasingly difficult. This yields the need for automatic analyses of these documents, including their argumentative and rhetorical structure.

Accordingly, as discussed in Section 2.1.4 and as dealt with in the previous Chapter, computational models already support a series of publication analysis tasks, e.g., classification of citation purpose and polarity (Athar, 2011; Jha et al., 2016, *inter alia*) and classification of (sentential) discourse roles (Teufel et al., 1999; Liakata et al., 2010, *inter alia*). Further, rhetorical predictions at the (sub-)sentence level obtained using these models have been shown useful in higher-level downstream tasks such as publication

[*]Adapted from: (1) **Anne Lauscher**, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. Investigating the role of argumentation in the rhetorical analysis of scientific publications with neural multi-task learning models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3326–3338, Brussels, Belgium, October–November 2018, Association for Computational Linguistics. (2) **Anne Lauscher**, Goran Glavaš, and Simone Paolo Ponzetto. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining (ArgMining)*, pages 40–46, Brussels, Belgium, October–November 2018, Association for Computational Linguistics.

[2]Code and data are available here: `https://github.com/anlausch/multitask_sciarg`.

classification (Teufel et al., 1999), (extractive) publication summarization (Cohan and Goharian, 2015), and research trend prediction (McKeown et al., 2016).

To allow for the holistic analysis of scientific publications with respect to the interactions between different rhetorical aspects of scientific text (**C3**), which we collectively dub *scitorics*, Fisas et al. (2016) created a corpus of scientific publications with manual annotations of several high-level rhetorical aspects of scientific writing (e.g., sentence-level discourse roles), but without annotations of the argumentative structure of publications. Despite (1) scientific texts being inherently argumentative (Gilbert, 1976), (2) the existence of theoretical argumentative frameworks (Toulmin, 1958, 2003 edition; Kirschner et al., 2015), and (3) a wide range of argument extraction models in other domains (e.g., debates or essays, see Palau and Moens (2009); Habernal and Gurevych (2017), *inter alia*), there is still very little work on automatic argumentation mining from scientific literature. Consequently, to the best of our knowledge, there has been no work analyzing associations between argumentation and other rhetorical constructs in scientific writing, although such dependencies clearly exist. Consider the following example:

> *"In general, our OMR preserves the high frequency content of the motion quite well [claim], since inverse rate control is directed by Jacobian values [data]."*

Here, the authors make a *claim* (underlined text) about their approach and support it with a technical fact (*data*) about the method (wave-underlined text). At the same time, regarding other rhetorical constructs, this sentence is stating the subjective aspect of *advantage* (of the proposed method), belongs to the discourse category of *outcome* (of the authors' work), and may be considered *relevant* for the summary of the publication. We argue that these rhetorical dimensions are interconnected and that fine-grained argumentation underpins other rhetorical layers in scientific text. For example, sentences stating an *advantage* of a method are likely to be argumentative and may contain *claims* that should be included in the summary. In contrast, purely descriptive, non-argumentative sentences often describe low-level technical details (e.g., belong to discourse class *approach*) and, lacking any claims, should not be included in the summary.

Assuming that argumentation guides rhetorics in scientific text, we investigate neural MTL models, which couple argument extraction with several other rhetorical analysis tasks. To this end, we augment the existing corpus of scientific publications (Fisas et al., 2016), containing several layers of rhetorical annotations, with an additional layer of argumentative components and relations. We then explore two neural MTL architectures based on shared recurrent encoders, intra-sentence attention, and private task-specific classifiers and couple the neural architectures with a joint MTL objective with uncertainty-based weighting of task-specific losses (Kendall et al., 2018). We validate our approach by testing that it outperforms traditional machine learning models in single-task settings. We finally show that coupling rhetorical analysis tasks with argument extraction using MTL models significantly improves the results for the rhetorical analysis tasks.

Contributions. We propose a general argument annotation scheme for scientific text that can cover various research domains. We next extend the Dr. Inventor corpus (Fisas et al., 2015, 2016) with an annotation layer containing fine-grained argumentative com-

ponents and relations. These efforts result in the first argument-annotated corpus of scientific publications in English, which allows for joint analyses of argumentation and other rhetorical dimensions of scientific writing. We make the argument-annotated corpus publicly available. Next, we offer an extensive statistical and information-theoretic analysis of the corpus. We then carry out the first study on dependencies between different rhetorical dimensions in the computational analysis of scientific writing. Using MTL models, we show that argumentation informs other rhetorical analysis tasks. Finally, in the context of MTL research, our results indicate that the dynamic uncertainty-based loss weighting (Kendall et al., 2018) is beneficial for high-level NLP tasks.

6.1.2 Related Work

We provide an overview of (1) studies analyzing rhetorical aspects in scientific publications and (2) a large body of work on argumentation mining.

Rhetorical Analysis of Scientific Texts

Previous work has analyzed a number of rhetorical aspects of scientific publications. Pioneering annotation efforts of Teufel and Moens (1999a,b); Teufel et al. (1999) focused on discourse-level argumentation (dubbed *argumentative zones*), denoting more the rhetorical structure of the publications than fine-grained argumentation, i.e., there are no (1) fine-grained argumentative components (at sub-sentence level) and no (2) relations between components, giving rise to an argumentation graph. Liakata et al. (2010) proposed a more general discourse scheme dubbed *core scientific concepts* and in subsequent work (Liakata et al., 2012) trained a conditional random fields (CRF) model to assign discourse labels to text spans. Blake (2010) distinguishes between explicit and implicit claims, correlations, comparisons, and observations in biomedical publications. In contrast, we are not interested in how the claim is made, but rather in what are the claims (and what is not a claim) and how they are mutually connected. Several authors focused on tasks relating to citations: extraction of citation context (e.g., Abu-Jbara et al., 2013; Jha et al., 2016), classification of citation polarity (e.g., Athar, 2011) and purpose (e.g., Teufel et al., 2006; Jochim and Schütze, 2012), and the automatic detection of referenced parts of the cited publication (Jaidka et al., 2016). Both discourse and citation information have been exploited for summarizing scientific publications (Cohan and Goharian, 2015; Teufel and Moens, 2002; Abu-Jbara and Radev, 2011; Chen and Zhuge, 2014). Intuitively, citation contexts may contain information relevant to the summary. Similarly, summaries commonly contain sentences with diversified discourse properties.

Fisas et al. (2016) provided different layers of rhetorical annotations on the same corpus of scientific text. Their Dr. Inventor Corpus is annotated with a combination of existing discourse annotation schemes (Teufel et al., 2009; Liakata et al., 2010) and citation-based annotations. Despite the argumentative nature of scientific texts, the Dr. Inventor Corpus contains no annotations of argumentative components such as claims. Several computational studies followed, addressing the rhetorical tasks corresponding to the layers of the Dr. Inventor Corpus (Ronzano and Saggion, 2015, 2016; Accuosto et al., 2017), but none of them investigated dependencies between different

tasks. Green (2014a,b, 2015b, 2016) proposed methods for identifying and annotating argumentative structures in scientific publications, but released no publicly available annotated corpus, and consequently, no computational models.

The work of Kirschner et al. (2015) is the closest to ours since they annotated scientific publications with fine-grained argumentation. However, their corpus is in German and contains no annotations of other rhetorical dimensions. Moreover, their corpus is significantly smaller than the Dr. Inventor Corpus (Fisas et al., 2016). In contrast, we augment the Dr. Inventor Corpus with an argumentation layer, allowing for combinations of argumentation extraction and other rhetorical analysis tasks in MTL settings.

Argumentation Mining

In their pioneering work on automatic AM, Palau and Moens (2009) discriminated argumentative from non-argumentative sentences and proposed a rule-based approach for extracting argumentative structures in documents. Habernal and Gurevych (2016, 2017) extracted argumentative components from online discussions. They framed the argumentative component extraction as a sequence labeling task and applied structured SVMs as a learning model. Recent work started exploiting dependencies between AM tasks using global optimization (Peldszus and Stede, 2015; Persing and Ng, 2016; Stab et al., 2014) and MTL models (Eger et al., 2017; Niculae et al., 2017). Peldszus and Stede (2015) used decoding based on minimum spanning trees to jointly predict argumentative segments and their types as well as argumentative relations, to generate an argumentation graph from text. Persing and Ng (2016) and Stab and Gurevych (2017a) similarly produced argumentative structures by globally optimizing local predictions of argumentative components and relations. Potash et al. (2017) proposed a neural architecture based on a pointer network for jointly predicting types of argumentative components and identifying argumentative relations. In a similar effort, Eger et al. (2017) combined the AM tasks using the MTL framework of Søgaard and Goldberg (2016). Remedying for data sparsity, Schulz et al. (2018) treated different argumentation formalisms as different tasks and combined respective extraction tasks and data sets in a MTL setting. In contrast to these efforts that combine several AM subtasks or formalisms with joint optimization and MTL models, in this work, we examine the dependencies between argumentative components and other rhetorical aspects of scientific writing.

6.1.3 Data Annotation

We first briefly describe the original Dr. Inventor Corpus (Fisas et al., 2016), which we augment with fine-grained argumentative annotations. We then explain in more detail our argumentation annotation scheme and the annotation process.

Dr. Inventor Corpus

We chose the Dr. Inventor Corpus (Fisas et al., 2015, 2016) as a starting point for our study of associations between argumentative structure and rhetorical aspects of scientific publications for two reasons. First, containing 40 publications with a total of 10,789

Annotation Layer	Labels	%
Discourse Role	*Background*	20
	Challenge	5
	Approach	57
	Outcome	16
	Future Work	2
Subjective Aspect	*Advantage*	33
	Disadvantage	16
	Adv.-Disadv.	3
	Disadv.-Adv.	1
	Novelty	13
	Common Practice	32
	Limitation	2

Annotation Layer	Labels	%
Summarization Rel.	*Totally irrelevant*	66
	Should not appear	6
	May appear	14
	Relevant	6
	Very relevant	8
Citation Purpose	*Criticism*	23
	Comparison	9
	Use	11
	Substantiation	1
	Basis	5
	Neutral	53

Table 6.1: Annotation layers of the Dr. Inventor Corpus (Fisas et al., 2016).

sentences, it is one of the largest corpora of scientific arguments, which is manually labeled with rhetorical information. Secondly, it contains *four* different layers of rhetorical annotations which allow for studying complementarity of knowledge across tasks: (1) a *discourse* layer, specifying discourse roles of sentences, (2) a *citation context* layer, specifying the textual context of citations, (3) a layer with *subjective aspect* categories assigned to sentences, and (4) a *summarization relevance* layer, indicating how relevant sentences are for the summary. The overview of labels for all annotation layers with the distribution of instances across labels is shown in Table 6.1. For more details on the original Dr. Inventor Corpus we refer the reader to (Fisas et al., 2015, 2016).

Argumentation Annotation Scheme

We considered several existing argumentation frameworks (e.g., Anscombre and Ducrot, 1997; Walton et al., 2008, *inter alia*) and selected Toulmin's model (Toulmin, 1958, 2003 edition), explained in Section 2.1.2, as basis for our annotation scheme. We chose Toulmin's model because: (1) it is a well-established in philosophy as well as in computer science (e.g, Freeman, 1991; Bench-Capon, 1998; Verheij, 2009, *inter alia*) and (2) it contains different types of argumentative components and takes the relations between them into account, which is useful for fine-grained analyses. To test the applicability of the framework for our purposes, we first carried out a preliminary annotation round with two expert annotators and adjusted the annotation scheme according to their observations.

Argumentative components. We devised an adapted version of the Toulmin model,[3] containing the following argumentative components:

- *Background claim*: an argumentative statement related to the work of other authors, state-of-the-art methods, or common practices;

[3]We omitted some of Toulmin's component types due to very rare occurrences in the corpus.

"The range of breathtaking realistic 3D models is only limited by the creativity of artists and resolution of devices."

- *Own claim*: an argumentative statement about the authors' own work;

  "Using our method, character authors may use any tool they like to author characters."

- *Data*: a fact that the authors state as evidence that supports or contradicts a claim.

  "SSD is widely adopted in games, virtual reality, and other realtime applications due to its ease of implementation and low cost of computing."

Argumentative components are annotated as arbitrary spans of text (in terms of length, annotated components ranged from a single token to multiple sentences). Annotators were instructed to annotate the shortest possible span of text that completely captures the argumentative component. Thus, we do not bind arguments to sentences, i.e., we allow for fine-grained argumentative components.

Argumentative relations. Authors connect argumentative components in order to form convincing reasoning chains. We also annotated relations between argumentative components. Following proposals from previous work (Dung, 1995; Bench-Capon, 1998), we distinguish between three relation types:

- *Supports*: indicates that a *claim* component is supported by a data component or another claim. The (assumed) validity of the *supporting* component (data or claim) contributes to the validity of the *supported* claim.
- *Contradicts*: indicates that the validity of a claim decreases with the validity of another argumentative component. If an argumentative component is assumed to be true, the claim it contradicts is assumed to be false, and vice versa.
- *Same claim*: connects different mentions of what is essentially the same claim. It is common to repeat important claims (e.g., the central claim) of the work several times in the publication, e.g., in the inroduction and in the conclusion.

Further details about the annotation scheme can be found in the annotation guidelines.[4]

Annotation Procedure

We hired four annotators for the task, one of whom we considered to be an *expert* annotator[5] and executed the process in two phases. In the first phase, we calibrated the annotators for the task in five iterations on five publications from the Dr. Inventor Corpus. After all annotators labeled one of the five documents, we met with them, discussed the disagreements, identified erroneous annotations, and, when required, revised the annotation guidelines. At the end of the calibration phase, the annotators re-annotated

[4] `http://data.dws.informatik.uni-mannheim.de/sci-arg/annotation_guidelines.pdf`

[5] A researcher in computer science, albeit not in computer graphics, which is the domain of the corpus.

the five calibration publications and resolved the remaining disagreements by consensus. In Figure 6.1 we show the inter-annotator agreement (IAA) for both component identification and relation classification, in terms of averaged pairwise F_1 score,[6] after each of the five calibration iterations. The evolution of IAA over the five calibration iterations is depicted in two variants: (1) a *strict* version in which components have to match exactly in span and type and relations have to match exactly in both components, direction and type of the link and (2) a *relaxed* version in which components only have to match in type and overlap in span (by at least half of the length of the shorter of them). Expectedly, we observe higher agreements with more calibration as the discussions helped to get a common understanding of the task among the annotators. The agreement on argumentative relations is 23% lower than on the components, which we think is due to the high ambiguity of argumentation structures, as it was also previously noted by Stab et al. (2014). That is, given an argumentative text with pre-identified argumentative components, there are often multiple valid interpretations of an argumentative relation between them, i.e., it is "[...] hard or even impossible to identify one correct interpretation" (Stab et al., 2014). Additionally, disagreements in component identification are propagated to relations as well, since the agreement on a relation implies the agreement on annotated components at both ends of the relation. Interestingly, the average agreement of our *expert* annotator with the *non-expert* annotators was similar to the average agreement between non-expert annotators. This is encouraging because it suggests that annotating argumentative structures in scientific text does not require expert knowledge of the domain. In the second phase, we evenly split the remaining 35 documents of the Dr. Inventor Corpus among the four annotators, without any overlaps.

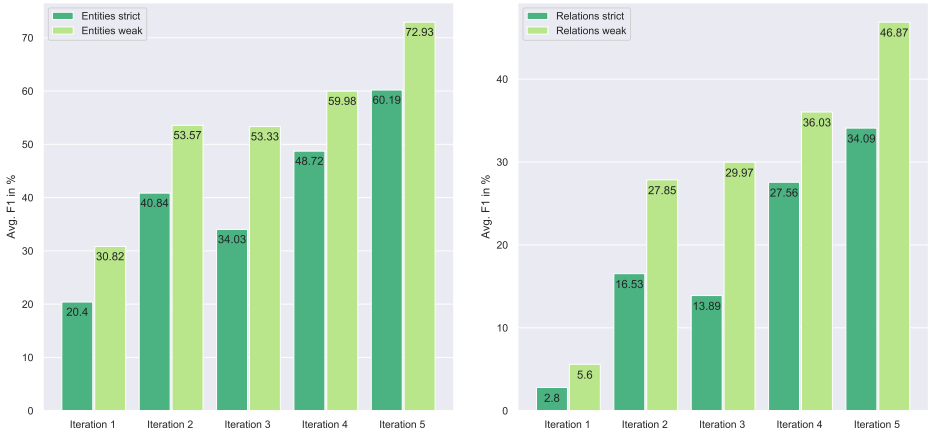
6.1.4 Corpus Analysis

We make the Dr. Inventor Corpus augmented with argumentation annotations (together with the annotation guidelines) publicly available.[7] The final corpus contains 12,289 annotations of argumentative components and 6,530 relation annotations. We next study the argumentation layer we annotated in isolation. Afterwards, we focus on the interrelations with other rhetorical annotation layers.

Analysis of Argumentation Annotations. Table 6.2 lists the number of components and relations in total and on average per publication. The number of *own claims* roughly doubles the amount of *background claims*, as the corpus consists only of original research papers, in which the authors mainly emphasize their own contributions. Interestingly, there are only half as many *data* components as claims. We can see two reasons for this – first, not all claims are supported and secondly, claims can be supported by other claims. There are many more *supports* than *contradicts* relations. This is intuitive,

[6] We measured the agreement in terms of the F_1 measure because (1) it is straight-forward to compute, (2) it is directly interpretable, and (3) it can account for spans of varying length, allowing for computing relaxed agreements in terms of partial overlaps, and (4) the chance-corrected measures, e.g., Cohen's Kappa, approach F_1-measure when the number of negative instances grows (Hripcsak, 2005).

[7] `http://data.dws.informatik.uni-mannheim.de/sci-arg/compiled_corpus.zip`



(a) Argumentative Components (b) Argumentative Relations

Figure 6.1: IAA evolution over calibration phases: (a) argumentative components; (b) relations. We report both *strict* (annotated components match in span and type; relations match in type and components at both ends match strictly) and *relaxed* agreement scores (components match in type and overlap in span; relations match in type and their components at both ends match according to the relaxed criterion).

Category	Label	Total	Per Publication
Component	*Background claim*	2,751	68.8 ± 25.2
	Own claim	5,445	136.1 ± 46.0
	Data	4,093	102.3 ± 32.1
Relation	*Supports*	5,790	144.8 ± 43.1
	Contradicts	696	17.4 ± 9.1
	Semantically same	44	1.1 ± 1.81

Table 6.2: Total and per-publication distributions of labels of argumentative components and relations in the augmented Dr. Inventor Corpus.

Label	Min	Max	Avg (μ)	Std (σ)
Background claim	5	340	87.46	43.74
Own claim	3	500	85.70	44.03
Data	1	244	25.80	27.59

Table 6.3: Statistics on the length of argumentative components (in number of characters) identified in the augmented Dr. Inventor Corpus.

as authors mainly argue by providing *supporting* evidence for their own claims.

Table 6.3 shows the statistics on length of argumentative components. While the *background claims* and *own claims* are on average of similar length (85 and 87 characters, respectively), they are much longer than *data* components (average of 25 characters). This

Criterion	Min	Max	Avg (μ)	Std (σ)
Diameter	2	5	3.05	0.71
Max In-Degree	3	11	6.33	1.97
# standalone claims	27	127	63.00	21.40
# unsupported claims	39	180	94.38	29.14
# unconnected subgraphs	78	231	147.23	35.78
# components per subgraph	1	17	2.09	1.5

Table 6.4: Graph-based analysis of the argumentative structures identified in the augmented Dr. Inventor Corpus. We report per publication statistics.

Type	Pub.	Claim with maximal PageRank score
background claim	A13	'physical validity is often sacrificed for performance'
	A21	'a tremendous variety of materials exhibit this type of behavior'
own claim	A39	'the solution to the problem of asymmetry is to modify the CG method so that it can operate on equation (15), while procedurally applying the constraints inherent in the matrix W at each iteration'

Table 6.5: Claims with maximum PageRank score in a publication.

is intuitive given the domain of the corpus, as facts in computer science often require less explanation than claims. For example, we noticed that authors often refer to tables and figures as evidence for their claims. Similarly, when claiming weaknesses or strengths of related work, authors commonly provide references as evidence.

The argumentative structure of an individual publication corresponds to a forest of directed acyclic graphs (DAG) with annotated argumentative components as nodes and argumentative relations as edges. Thus, to obtain further insight into structural properties of argumentation in scientific publications, in Table 6.4 we provide graph-based measures like the number of connected components (i.e., subgraphs), the diameter, and the number of standalone claims (i.e., nodes without incoming or outgoing edges) and unsupported claims (i.e., nodes with no incoming *supports* edges). Our annotators identified an average of 141 connected components per publication, with an average diameter of 3. This indicates that either authors write very short argumentative chains or that our annotators had difficulties noticing long-range argumentative dependencies.

On the one hand, there are at least 27 standalone claims in each publication, i.e., claims not connected with any other components. On the other hand, the maximum in-degree of a claim in a publication, on average, is 6, indicating that there are claims with a lot of evidence given. Intuitively, the claims for which more evidence is given should be more prominent. We next run PageRank (Page et al., 1999) on argumentation graphs of individual publications to identify most prominent claims. We list a couple of examples of claims with the highest PageRank scores in Table 6.5. Somewhat unexpectedly, in 30 out of 40 publications, the highest-ranked claim was a *background claim*. This suggests that in computer graphics, authors emphasize more research gaps and motivation for their work than they justify its impact (for which empirical results often suffice).

	AC	DR	SA	SR
AC	–	–	–	–
DR	**0.22**	–	–	–
SA	0.08	0.11	–	–
SR	0.04	0.10	0.13	–
CC	**0.18**	0.10	0.04	0.01

Table 6.6: Normalized mutual information between the label sets of the annotation layers indicating Argument Components (AC), Discourse Roles (DR), Subjective Aspects (SA), and Citation Contexts (CC) in the extended Dr. Inventor Corpus.

Links to Other Rhetorical Aspects. We next investigate the interdependencies between the newly added argumentative annotations and the existing rhetorical annotations of the Dr. Inventor Corpus. An inspection of dependencies between different annotation layers in the corpus may indicate the usefulness of computational approaches that aim to exploit such interrelations. E.g., Bjerva (2017) recently showed that the measure of mutual information strongly correlates with performance gains obtained by multi-task learning models. Accordingly, We employ the measure of normalized mutual information (NMI) (Strehl and Ghosh, 2003) to assess the amount of information shared between the five annotation layers. NMI is a variant of mutual information scaled to the interval $[0, 1]$ through normalization with the entropy of each of the two label sets. For our analysis, we port all token-level annotations to the sentence-level, and then compute pairwise NMI. In Table 6.6 we show the NMI scores for all pairs of annotations layers: Argument Components (AC), Discourse Roles (DR), Citation Contexts (CC), Subjective Aspects (SA), and Summary Relevances (SR). The strongest association is found between AC and DR. Looking at the labels of these two annotation layers, this seems plausible – *background claim* (AC) is likely to appear in a sentence of discourse role *background* (DR). Similarly, *own claims* more frequently appear in sections describing the *outcomes* of the work. To confirm this intuition, we computed co-occurrence matrices for pairs of label sets – indeed, the AC label *own claim* most frequently appears together with the discourse role *approach* and *outcome*, and the *background claim* with discourse roles *background* and *challenge*. Consider the following sentence:

> "*With the help of modeling tools or capture devices, complicated 3D character models are widely used in the fields of entertainment, virtual reality, medicine, etc.*"

It contains a general claim about the research area (i.e., it is a *background claim*) and it also offers *background* information in terms of the overall scientific discourse of the publication. A similar set of intuitive label alignments justifies the higher NMI score between argumentative components (AC) and citation contexts (CC): *citation contexts* often appear in sentences with a *background claim*. Again, this is not surprising, as authors need to reference other publications and in order to motivate their work and to position their work within their respective research field.

This is exemplified by the following two sentences:

> *"An improvement based on addition of auxiliary joints has been also proposed in [Weber 2000]. Although this reduces the artifacts, the skin to joints relationship must be re-designed after joint addition."*

In the above example, the wave-underlined text, i.e., the citation, serves as the *data* for the underlined text, which is the *background claim* stating a research gap in the referenced work. Simultaneously, the underlined text acts as the *citation context* of the reference.

6.1.5 Multi-task Learning for Analyzing Scientific Argumentation

We next exploit the augmented corpus to study the dependencies between fine-grained argumentation and other scitorics. To this end, we adopt neural MTL.

Tasks

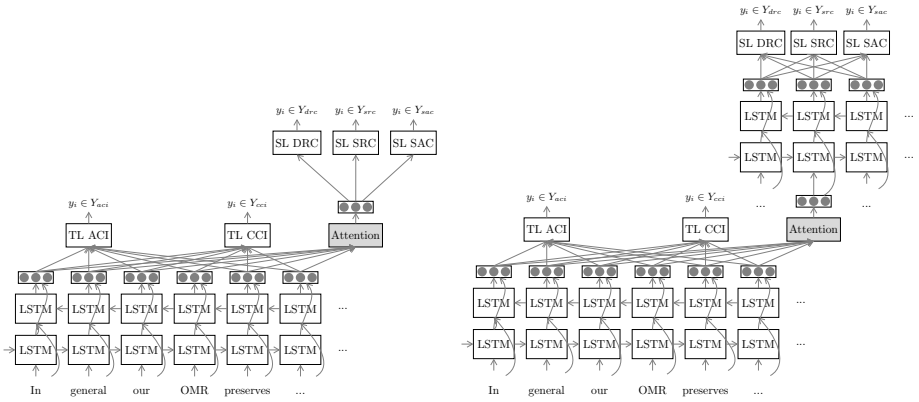
The following are the rhetorical analysis and argument extraction tasks we investigate. We discussed those from a general perspective in Section 2.1.4) and introduce here the concrete task formalization we are dealing with in our study.

Argument Component Identification (ACI). The task is to extract and classify argumentative components. We frame ACI as a token-level sequence labeling task: given a sequence of tokens $\mathbf{x} = (x_1, .., x_n)$ of length n, the task is to assign a sequence of tags $\mathbf{y}_{aci} = (y_1, .., y_n), y_i \in Y_{aci}$. The tagset Y_{aci} contains seven token-level tags, obtained by combining the standard B–I–O annotation scheme with three types of argumentative components: *Own claim*, *Background claim*, and *Data*.

Discourse Role Classification (DRC). The multi-class classification task in which each sentence needs to be assigned one out of the set of discourse roles $Y_{drc} = \{$ *Background*, *Unspecified*, *Challenge*, *FutureWork*, *Approach*, *Outcome*$\}$.

Citation Context Identification (CCC). The task is to identify the span of the publication text that introduces or explains a reference. It is also a token-level sequence-labeling task – a sequence of tags $\mathbf{y}_{cci} = (y_1, .., y_n)$ with $y_i \in Y_{cci} = \{B_{CC}, I_{CC}, O\}$ is assigned to a sequence of tokens $\mathbf{x} = (x_1, .., x_n)$.

Subjective Aspect Classification (SAC). Another sentence-level classification task in which the model has to assign one of the subjective aspect labels, $Y_{sac} = \{$*None*, *Limitation*, *Advantage*, *Disadvantage-Advantage*, *Disadvantage*, *Common Practice*, *Novelty*, *Advantage-Disadvantage*$\}$, to each sentence.



(a) Simple model.

(b) Hierarchical model.

Figure 6.2: Neural MTL architectures for the rhetorical and argumentative analysis of scientific publications: (a) the *Simple model* addresses sentence-level tasks (DRC, SAC, SRC) as plain classification tasks, whereas (b) the *Hierarchical model* treats sentence-level tasks as sequence labeling tasks thereby considering surrounding context. Both models address ACI and CCC as token-level sequence labeling tasks.

Summary Relevance Classification (SRC). The task is to predict the relevance of sentences for the summary of the publication. Each sentence needs to be assigned a label with $Y_{src} = \{$*Very relevant*, *Relevant*, *May appear*, *Should not appear*, *Totally irrelevant*$\}$.

ACI and CCC are token-level sequence labeling tasks. The remaining three tasks can be cast as either (1) plain sentence classification tasks or (2) sentence-level sequence labeling tasks (assuming that there are regularities in sequences of sentence-level labels that can be captured). We propose one MTL architecture for each of the two possibilities.

Multi-Task Learning Models

We propose two different MTL architectures for the rhetorical and argumentative analysis of scientific publications. The *Simple model* treats sentence-level tasks (DRC, SAC, and SRC) as plain classification tasks (i.e., the prediction for each sentence ignores the content and labels of other, neighboring sentences). The *Hierarchical model* addresses sentence-level tasks as sequence labeling tasks. This model can be seen as a hierarchical sequence labeling model, in which the sentence-level recurrent network is stacked on top of the token-level sequence labeling network. Both architectures are illustrated in Figure 6.2.

Token-level Predictions. Given a sentence $\mathbf{s}_i = (x_{i1}, .., x_{in})$ out of a sequence of sentences $\mathbf{D} = (\mathbf{s}_1, .., \mathbf{s}_m)$ we first retrieve the pre-trained embedding vector for each token $\mathbf{x}_{ij}$.We then obtain context-aware token representations $\mathbf{h}_{ij}$ by applying a bidirectional recurrent network with long short-term memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) on the sequence of pre-trained word embeddings:

$$\mathbf{h}_{ij} = [\overrightarrow{\text{LSTM}}(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ij}); \overleftarrow{\text{LSTM}}(\mathbf{x}_{in}, \ldots, \mathbf{x}_{ij})] \,. \tag{6.1}$$

This token-level Bi-LSTM encoder is shared between the tasks combined by the MTL models. Next, we define a separate classifier for each of the token-level (TL) tasks (i.e., ACI and CCC) and feed the contextualized token representations $\mathbf{h}_{ij}$ to these classifiers. Each of the classifiers is defined as a feed-forward network with a single hidden layer. The label probability distribution is obtained by applying the *softmax* function on its output.

$$y_{ijt} = \text{softmax}(\mathbf{W}_t \mathbf{h}_{ij} + \mathbf{b}_t) \,, \tag{6.2}$$

where $\mathbf{W}_t \in \mathbb{R}^{2K \times |Y_t|}$ and $\mathbf{b}_t \in \mathbb{R}^{|Y_t|}$ are the task-specific classification parameters for the task t, with K being the size of the LSTM state and $|Y_t|$ the number of labels of t.

Sentence-level Predictions. We learn to aggregate a sentence representation $\mathbf{s}_i$ from contextualized vectors of its tokens, $\mathbf{h}_{ij}$ (produced by the token-level Bi-LSTM), using the intra-sentence attention mechanism (Yang et al., 2016):

$$\mathbf{s}_i = \sum_j \alpha_{ij} \mathbf{h}_{ij} \,, \tag{6.3}$$

with the weights α_i computed dynamically as:

$$\alpha_i = \text{softmax}(\mathbf{U}_i \, \mathbf{u}_{att}) \,, \tag{6.4}$$

where $\mathbf{u}_{att}$ is the trainable attention head vector and $\mathbf{U}_i$ is a matrix with non-linearly transformed token representations ($\mathbf{h}_{ij}$) as rows:

$$\mathbf{U}_{ij} = \tanh(\mathbf{W}_{att} \mathbf{h}_{ij} + \mathbf{b}_{att}) \,. \tag{6.5}$$

In the *Simple* architecture, sentence representations $\mathbf{s}_i$ are fed directly to the sentence-level task-specific classifiers, which are also single-layer feed-forward networks:

$$y_{it} = \text{softmax}(\mathbf{W}_t \mathbf{s}_i + \mathbf{b}_t) \,. \tag{6.6}$$

Within the *Hierarchical* architecture, sentence representations are first contextualized with representations of other sentences via the sentence-level Bi-LSTM layer (denoted with the function Bi-LSTM$_S$) and then forwarded to the classifier:

$$y_{it} = \text{softmax}(\mathbf{W}_t \text{Bi-LSTM}_S(\mathbf{s}_i) + \mathbf{b}_t) \,. \tag{6.7}$$

Joint optimization and loss functions. All of the tasks we consider are framed as multi-class classification tasks. Thus, we simply specify all task-specific losses to be L2-regularized cross-entropy errors. Let y_{to} be the one-hot ground truth label vector for the prediction instance o[8] of the task t, and let y'_{to} be the predicted probability distribution over the task labels for the same instance. With Y_t as the set of classification labels for the task t, the task-specific loss L_t is computed as follows:

$$L_t = \lambda \|\theta_t\|_2 - \sum_o \sum_{k=1}^{|Y_t|} y_{to}^{(k)} \cdot \ln \left(y_{to}^{\prime(k)} \right), \tag{6.8}$$

where θ_t is the set of model's parameters relevant for the task t[9] and λ is the regularization factor. We train the MTL model jointly on different tasks by defining and minimizing the joint loss function L that combines task-specific losses L_t. Instead of using constant weights, we opt for dynamic weighting of task-specific losses during the training process, based on the homoscedastic uncertainty of tasks, as proposed by Kendall et al. (2018):

$$L = \sum_t \frac{1}{2\sigma_t^2} L_t + \ln \sigma_t^2 \,, \tag{6.9}$$

where σ_t is the variance of the task-specific loss over training instances used to quantify the uncertainty of task t. Kendall et al. (2018) show that better MTL results can be obtained by dynamically assigning less weight to the more uncertain tasks, as opposed to constant task weights throughout the whole training process.[10]

6.1.6 Evaluation

We run two sets of experiments. First, we evaluate the performance of the *Simple* and the *Hierarchical* neural models on individual tasks (i.e., in single-task learning (STL) scenarios). We then evaluate the impact of the argumentative signal on other dimensions of rhetorical analysis by combining them in joint MTL settings.

Experimental Setup.

We randomly split the corpus on the document level into train (roughly 70%, 28 documents containing 6,697 sentences) and test portions (roughly 30%; 12 documents with 2,874 sentences). We used roughly 20% of the train portion as the validation set.

Model Configuration and Training. We ran an initial grid search on the validation set with values for the hyperparameters learning rate $\nu \in \{10^{-4}, 10^{-5}\}$, L2 regularization factor $\lambda \in \{0.001, 0.0001\}$, and LSTM states $K \in \{64, 128, 256\}$ and found

[8] The prediction instance is a token for ACI and CCC, and a sentence for DRC, SAC, and SRC.

[9] The set of relevant parameters differs across tasks: for token-level tasks (e.g., ARI) θ_t denotes token-level Bi-LSTM parameters and the parameters $\mathbf{W}_t$ and $\mathbf{b}_t$ of task t's classifier; for a sentence-level task (e.g., DRC) within the *Hierarchical* architecture, θ_t includes all parameters of both token- and sentence-level Bi-LSTMs, intra-sentence attention parameters, and parameters of the task-specific classifier.

[10] Later, we experiment with constant weights and confirm this observation.

	ACI			CCC		
Model	P	R	F1	P	R	F1
HMM	30.8	17.2	20.8	18.3	13.1	15.0
$\text{CRF}_{lexical}$	38.8	29.1	31.7	15.3	17.8	16.4
$\text{CRF}_{embeddings}$	37.9	23.3	26.1	12.8	1.4	2.5
Neural: *Simple*	**47.0**	**44.5**	**44.7**	**48.7**	**43.8**	**46.1**

Table 6.7: Single-task results for the token-level classification tasks (Precision (P), Recall (R), and F1 performances macro-averaged over the classes).

the configuration $\nu = 10^{-4}$, $\lambda = 0.001$, and $K = 128$ to be optimal for the majority of the STL and MTL models. In all experiments, we represent tokens with pretrained 300-dimensional GLOVE embeddings (Pennington et al., 2014)[11] and optimize the model parameters using the Adam algorithm (Kingma and Ba, 2015). We initialize all model parameters using Xavier initialization (Glorot and Bengio, 2010), train the models in batches of $N = 16$ sentences and apply early stopping based on the validation set performance.

Baselines. As a type of "sanity check", we first compare the performance of the two neural architectures against traditional supervised machine learning algorithms on each of the tasks separately. For the token-level sequence labeling tasks (ACI and CCC) we use a hidden markov model (HMM) and CRF (Lafferty et al., 2001) as baselines. The HMM works directly on the tokens, while we feed either the lexical representation or the embedding representation of the tokens as features for the CRF. For the sentence classification tasks (DRC, SAC, and SRC), we evaluate as baselines (1) the linear SVM with TF–IDF feature vectors and (2) SVM with RBF kernel and embedding features. In the latter case, we obtain a sentence representation by averaging the pretrained embeddings of sentence words. We tune the hyperparameter values of the SVM by conducting a grid search with possible penality parameter values $c \in \{0.1, 1.0, 10.0\}$ (linear SVM and SVM with RBF kernel) and the parameter of the radial basis function $\gamma \in \{0.01, 0.1, 1.0\}$ (SVM with RBF kernel). The possible hyperparameter values for the L1 regularization coefficient $c1$ and for L2 regularization coefficient $c2$ of the CRF are $c1, c2 \in \{0.1, 0.2, 0.001, 0.0001\}$.

In MTL the experiments, we consider the respective task performances from the STL experiments as well as MTL with a joint loss function with fixed equal weighting of the task losses, i.e., weights set to 0.5 when coupling two tasks, as baselines.

Single-Task Experiments. We first report the model performances for individual tasks in STL settings. Results for token-level tasks are shown in Table 6.7, whereas Table 6.8 displays results for sentence-level tasks. The scores (Precision, Recall, and F1 score) are reported as macro-averages over all task labels. Expectedly, our neural architectures substantially outperform the traditional machine learning baselines on all tasks. For the three sentence-level tasks, the *Hierarchical* architecture outperforms the

[11] `http://nlp.stanford.edu/data/glove.840B.300d.zip`

Model	DRC	SAC	SRC
SVM_{tfidf}	34.0	10.3	22.2
$SVM_{embeddings}$	25.7	08.5	19.3
Neural: *Simple*	**44.1**	**20.5**	31.5
Neural: *Hierarchical*	42.6	19.1	**33.2**

Table 6.8: Single-task results for sentence-level tasks (macro-averaged F1 scores).

	CCC	DRC	SAC	SRC
Single Task				
Simple	46.1	44.1	20.5	31.5
Hierarchical	–	42.6	19.1	33.2
Multi Task (w. ACI)				
$Simple_{0.5}$	43.8 (44.2)	43.5 (41.6)	18.0 (42.0)	32.2 (41.9)
$Simple_{\text{uncert}}$	**49.9** (40.5)	**45.2** (38.6)	**22.1** (39.4)	**34.8** (41.0)
$Hierarchical_{0.5}$	–	41.6 (42.1)	17.8 (42.9)	30.3 (43.4)
$Hierarchical_{\text{uncert}}$	–	**43.9** (40.8)	18.9 (41.6)	**34.8** (40.8)

Table 6.9: MTL results: rhetorical analysis tasks coupled with argumentative component identification. We report the F1 score macro-averaged over the classes. The scores achieved for argumentative component identification are shown in parentheses.[12]

Simple model only when classifying sentences by summary relevance (SRC). This result seems intuitive – a *Very relevant* sentence is likely to be surrounded with *Relevant* and *May appear* sentences (and an *Irrelevant* sentence with other *Irrelevant* and *Should not appear* sentences). The fact that we observe no gains from the additional sentence-level Bi-LSTM encoder for DRC and SAC suggests that the content of the sentence informs its discourse role and subjective aspect much more strongly than neighboring sentences. In other words, DRC and SAC seem to be more localized classification tasks than SRC.

Multi-Task Learning Results. Our core research question relates to the effect that recognizing fine-grained argumentative components has on other rhetorical analysis tasks, thereby addressing the issue of complementarity of knowledge in language representations across tasks (**C3**). This is why, in our central set of experiments, we evaluate MTL models with homoscedastic uncertainty weighting which combine the ACI (as an auxiliary task) with each of the four other tasks. In each MTL model, the token-level Bi-LSTM encoder is shared between the two tasks. For sentence-level tasks (DRC, SAC, SRC), we evaluate both the *Simple* and *Hierarchical* architecture. In Table 6.9 we show the performances of the MTL models on the rhetorical analysis tasks (these can be compared to the respective single-task model performances from Tables 6.7 and 6.8.

When coupled in MTL settings with ACI using the joint loss formulation of Kendall

[12] In the multi-task settings, the early stopping criterion was based on the auxiliary task score.

et al. (2018), the results significantly[13] improve for all rhetorical analysis tasks and models (except for SAC with the *Hierarchical* model), in comparison with the respective single-task models. However, the performance for the argumentation component identification does not improve in MTL. In other words, the extraction of fine-grained argumentative components seems to inform higher-level rhetorical analysis tasks, but not vice-versa. This indeed supports the hypothesis that argumentation guides scientific writing and influences rhetorical structure of publications. Furthermore, our results support the findings of Schulz et al. (2018) who show that, opposed to initial results of Martínez Alonso and Plank (2017), MTL can yield performance gains for higher-level semantic tasks.

6.1.7 Conclusion

Acknowledging the argumentative nature of scientific text and the issue of complementarity of knowledge across argumentative analysis tasks (**C3**), in this Section, we investigated the role of argumentation in the rhetorical analysis of scientific publications. We first extended an existing corpus annotated with four different layers of rhetorical information with annotations of argumentative components and relations, creating the largest argumentation-labeled corpus of scientific text in English. We first presented an annotation scheme for argumentation analysis in scientific publications. We annotated the *Dr. Inventor Corpus* (Fisas et al., 2015, 2016) with an argumentation layer. The resulting corpus, which is, to the best of our knowledge, the first argument-annotated corpus of scientific publications in English, enables (1) computational analysis of argumentation in scientific writing and (2) integrated analysis of argumentation and other rhetorical aspects of scientific text. We further provided corpus statistics and graph-based analysis of the argumentative structure of the annotated publications and analyzed the dependencies between different rhetorical aspects, which can inform computational models aiming to jointly address multiple aspects of scientific discourse. Employing the corpus, we explored intuitive neural architectures with recurrent encoders for argument extraction and rhetorical analysis tasks and showed significant improvements over traditional machine learning models. We then coupled argument extraction with different rhetorical analysis tasks in MTL models with dynamic loss weighting and demonstrated that the argumentative signal has a positive impact on high-level rhetorical analysis tasks.[14] Admittedly, the corpus we used in this work is limited to the domain of computer graphics. Nonetheless, we believe that our findings relating to the argumentative nature of scientific text and links between argumentation and other rhetorical aspects generalize to other domains too. This is also supported by the comparable IAA between expert and non-expert annotators.

In the next Section, we leave the specific case of scientific argumentation. Instead, we study the complementarity of knowledge in contextualized language representations for computational AQ assessment in multiple domains of online argumentation.

[13] Significant at $\alpha < 0.05$, tested using the non-parametric stratified shuffling test (Yeh, 2000).

[14] The recurrent encoder employed in this study could naturally be replaced with a pretrained contextualized language representations, e.g., BERT (Devlin et al., 2019).

6.2 Complementarity of Knowledge across Argument Quality Dimensions

[*]Envisioned CA applications include systems, which automatically assess the quality of argumentative texts in order to support users in improving their argumentative writing. Though preceding work in computational argument quality (AQ) mostly focuses on assessing overall AQ or specific conceptualizations of AQ, researchers agree that writers would benefit from feedback targeting individual dimensions of argumentation theory as described in Subsection 2.1.2. However, a large-scale theory-based corpus and corresponding computational models are missing. In this Section, we address this research gap by conducting an extensive analysis covering three diverse domains of online argumentative writing and presenting GAQCorpus: the first large-scale English multi-domain (community questions and answers forums, debate forums, review forums) corpus annotated with theory-based AQ scores. We then propose the first computational approaches to theory-based assessment, which can serve as strong baselines for future work. We demonstrate the feasibility of large-scale AQ annotation, show that exploiting the complementarity of knowledge between dimensions (**C3**) yields performance improvements, and explore the synergies between theory-based prediction and practical AQ assessment.

6.2.1 Introduction

Providing relevant and sufficient justifications for a claim and using clear and appropriate language to express reasoning are important features of everyday argumentative writing. These features relate to the notion of *argument quality (AQ)*, which has been studied in many domains, such as student essays (Wachsmuth et al., 2016), news editorials (El Baff et al., 2018), and online debate forums (Lukin et al., 2017).

Preceding work in NLP and CL has mostly focused on practical AQ assessment,[15] considering either the *overall quality* of arguments (Toledo et al., 2019; Gretz et al., 2020, inter alia) or a single specific conceptualization of AQ, e.g., *argument strength* (Persing and Ng, 2015), *convincingness* (Habernal and Gurevych, 2016), and *relevance* (Wachsmuth et al., 2017d). However, Gretz et al. (2020) note the need to predict quality in terms of fine-grained aspects. Fine-grained prediction enables a deeper understanding of argumentation and offers specific feedback to authors aiming to improve their argumentative writing skills. For instance, authors might want to know whether their premises are *sufficient* with regard to their claim(s) or whether their language is *appropriate*. As explained in Subsection 2.1.2, Wachsmuth et al. (2017b) surveyed and synthesized theory-based

[*]Adapted from: (1) **Anne Lauscher**, Lily Ng, Courtney Napoles, and Joel Tetreault. Rhetoric, Logic, and Dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4563–4574, Barcelona, Spain (Online), December 2020, International Committee on Computational Linguistics. (2) Lily Ng, **Anne Lauscher**, Joel Tetreault, and Courtney Napoles. Creating a domain- diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Work- shop on Argument Mining (ArgMining)*, pages 117–126, Online, December 2020, Association for Computational Linguistics.

[15]We adopt the terminology of Wachsmuth et al. (2017a) who refer to task-driven approaches, which often also focus on the *relative* assessment of AQ, as "practical".

dimensions of AQ into a taxonomy consisting of several fine-grained aspects under three main dimensions: Cogency (Logic), Effectiveness (Rhetoric), and Reasonableness (Dialectic). The taxonomy enables for theory-based AQ assessment, which provides a more targeted and informative perspective for researchers and end users. However, this holistic approach comes with the downside of higher complexity, especially when it comes to annotating textual corpora, which are required for training and developing common computational approaches (e.g., Gretz et al., 2020). In a small study, Wachsmuth et al. (2017a) demonstrate that theory-based AQ annotations can be done both by trained experts and by crowd annotators, though the authors acknowledge the high complexity and subjectivity of the problem. Accordingly, the authors call for the simplification of theory-based AQ annotation in order to reliably create larger-scale corpora. Given the overall feasibility of annotation and the recognized need for fine-grained dimensions in AQ assessment, it is surprising that no further efforts in NLP and CL have been made. There is no attempt on simplifying the task, no large scale annotated corpus and, consequently, no computational model. Furthermore, although intuitively there are interrelations between the different AQ dimensions, complementarity of knowledge between those (**C3**, see Section 3.3) has not been studied yet. In this Section, we aim to fill this research gap by conducting an in-depth analysis of theory-based AQ assessment covering overall AQ and the three dimensions (logic, rhetoric, and dialectic) of the Wachsmuth et al. taxonomy, and three diverse domains of online argumentative writing (Community Questions and Answers forums, debate forums, and review forums).

Drawing on existing AQ theories, we address five research questions (**RQs**) to inform and fuel future AQ annotation studies and computational AQ research:

(RQ1) *Can we develop a large-scale theory-based AQ corpus?* Building on Wachsmuth et al. (2017a), we modify the complex task of annotating theory-based AQ dimensions to be suitable for both experts and the crowd while preserving the theoretical basis of the taxonomy. We collect and annotate argumentative texts from web debate forums, as well as community questions and answers (CQA) forums, and review forum texts, which are still understudied in computational AQ. The latter domains can consist of rather non-canonical arguments in that they exhibit a lack of explicitness of certain argumentative components; are topic-wise more subjective; or consist of longer, more convoluted text. This makes assessing the quality of such arguments even more challenging, but downstream can result in a more robust model of computational AQ.

Given all these challenges, we work closely with trained linguists to adapt the annotation task, iterating over how best to approach these novel domains and simplify the annotation guidelines for crowdsourcing, allowing us to collect a large number of judgments efficiently. Our efforts result in GAQCorpus, the first large-scale multi-domain English corpus annotated with theory-based AQ scores. In total, GAQCorpus consists of 5,295 arguments from three domains of online argumentative writing.

(RQ2) *Are we able to develop computational models that can do theory-based AQ assessment in varying domains?* Based on GAQCorpus, we are the first to propose computational approaches to theory-based AQ assessment and show that it is possible to develop models for this task. Our models can serve as strong baselines for future research and enable the field to investigate follow-up research questions.

(RQ3) *Can the interrelations between the different AQ dimensions be exploited in a computational setup?* Inspired by the hierarchical structure of the taxonomy, We explore whether the relationships between dimensions can be computationally exploited. In addition to simple single-task learning approaches, we study the complementarity of knowledge in theory-based AQ assessment. To this end, we jointly predict AQ dimensions in two MTL (see Section 6.2.5) variants (*flat* vs. *hierarchical*) and find that combining the training signals of all four aspects benefits theory-based AQ assessment.

(RQ4) *Does the corpus support training a single unified model for multi-domain evaluation?* Relating back to our discussions about domain-specificity (**C2**) in Section 3.2 and Chapter 5, training on in-domain data is typically preferred over multi-domain data assuming that domain-specificity of language representations results in performance improvements. However, as we have seen before, there exists a trade-off: a higher degree of domain-specificity may imply a smaller amount of data and, accordingly, does not always result in better performance. Larger amounts of data are especially useful for complex model architectures currently prominent in NLP (e.g., BERT (Devlin et al., 2019), GPT2 (Radford et al., 2019)). We study these two mutually opposing effects on GAQCorpus and show that our corpus supports training a single unified model across all three domains, with improved performances in individual domains.

(RQ5) *Can we empirically substantiate the idea that theory-based and practical AQ assessment can learn from each other?* Wachsmuth et al. (2017a) suggest that both the practical and the theory-based views can learn from each other, but so far, this has been only tested manually. Employing our models, we go one step further and conduct a bi-directional experiment employing a practical AQ corpus. We demonstrate two concrete ways how theory-based and practical AQ research can profit from their combination.

6.2.2 Related Work

Earlier work in AQ assessment can be divided into practical and theory-based approaches.

Practical approaches. Recently, the field of computational AQ research has been mostly driven by practical approaches that each target an individual domain. Accordingly, past approaches tackle either overall quality (Toledo et al., 2019) or specific subqualities of argumentation, such as convincingness (Habernal and Gurevych, 2016) and relevance (Wachsmuth et al., 2017d). The popularity of practical approaches can partly be attributed to the relative simplicity of crowd-sourcing annotations.

Much prior work has focused on aspects of student essays, including essay clarity (Persing and Ng, 2013), organization (Persing et al., 2010), prompt adherence (Persing and Ng, 2014), and argument strength (Persing and Ng, 2015). Later, Wachsmuth et al. (2016) present an approach driven by detecting argumentative units, thereby demonstrating the usefulness of argument mining techniques to the problem. Similarly, Stab and Gurevych (2016) predict the absence of opposing arguments and in subsequent work (2017b) predict insufficient premise support in arguments. Another well-studied domain is web debates. Wachsmuth et al. (2017d) adapt PageRank to identify argument relevance. Habernal and Gurevych (2016) conduct pairwise comparison of the convincingness of debate

arguments. Additionally, Persing and Ng (2017) predict why an argument receives a low persuasive power score. By explaining flaws in argumentation, they highlight the importance of explainability and specific author feedback. Other approaches take into account properties of the source, i.e., the author (Durmus and Cardie, 2019) or the audience (El Baff et al., 2018; Durmus and Cardie, 2018). In contrast, we assume that a system may not have much knowledge about the authors or audience and thus our models operate solely on the text. Most recently, Toledo et al. (2019) and Gretz et al. (2020) crowd-sourced overall argument quality by presenting pairwise arguments to annotators, who then had to select the argument "they would recommend a friend to use that argument as is in a speech supporting/contesting the topic." This is an extreme simplification of the task, which does not seem to lead to better IAA: the authors report an average IAA of $\kappa = 0.12$ and attribute the low score to the high subjectivity of the task (Gretz et al., 2020). These corpora, on which they train computational models, cover a variety of topics, but only within single domains. The authors emphasize that research on theory-based approaches could further advance the field of computational AQ.

Theory-based approaches. Rooted in classic argumentation theory, the works can according to Wachsmuth et al. (2017b), be categorized based on whether they related to the *logical* (Johnson and Blair, 2006; Hamblin, 1970), *rhetorical* (Aristotle, ca. 350 B.C.E./ translated 2006), or *dialectical* (Perelman et al., 1969; Eemeren and Grootendorst, 2003) properties of an argument. Wachsmuth et al. (2017b) were the first to survey and highlight the importance of the theory-based approach to computational AQ and synthesized the argumentation-theoretic literature into a taxonomy, which we introduced in Subsection 2.1.2. Wachsmuth et al. (2017a) conducted a study in which crowd workers annotated 304 arguments for all 15 quality dimensions following Wachsmuth et al. (2017b), and demonstrated that the theory-based and practical AQ assessment match to a large extent and that the two views can learn from each other, for instance, when it comes to more practical annotation processes for theory-based AQ annotations.

However, until now, no further research on computational theory-based AQ assessment in NLP has been conducted, no larger-scale annotated corpus has been presented, and thus no computational model that would allow further investigation into the concrete synergies between the two perspectives exists.

6.2.3 Annotation Study

Wachsmuth et al. (2017a) suggest that large-scale annotation of theory-based AQ dimensions is possible. We test this finding and take it one step further by asking whether we can develop a large-scale theory-based AQ corpus (**RQ1**). This section presents GAQCorpus, the result of the first study annotating theory-based dimensions, including $5,285$ arguments from three diverse domains of real-world argumentative writing.

Simplifying the task

In designing our annotation task, we start from the annotation guidelines of Wachsmuth et al. (2017a), henceforth TvsP, which reflect the full taxonomy in Figure 6.3, but which

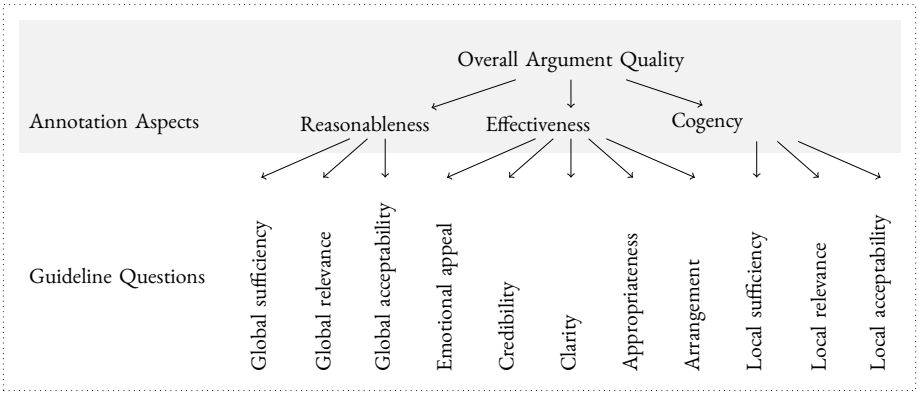


Figure 6.3: Scored dimensions and guideline questions based on the taxonomy of theory-based argument quality (Wachsmuth et al., 2017b). Annotators were guided by questions relating to all aspects for assessing the higher-level dimensions.

the authors posited was too complex for crowd-sourcing. Before collecting any crowd-sourced annotations, we conducted 14 pilot experiments with a group of four "expert" annotators, simplifying the TvsP task design through their feedback and observations, as they provided both a deep understanding of the argumentation theory and practical experience annotating the arguments. Each expert annotator was a fluent or native English speaker with an advanced degree in linguistics. Experts underwent training, which included studying guidelines and participating in calibration tasks to analyze debate arguments from three sources: Dagstuhl-ArgQuality-Corpus-V2,[16] originally from UKPConvArgRank (Habernal and Gurevych, 2016); the Internet Argument Corpus V2[17] (IAC; Abbott et al., 2016); and ChangeMyView,[18] a Reddit forum. Through the pilot studies and subsequent debriefs with the experts, we made the following modifications to the annotation task of Wachsmuth et al. (2017a):

(1) Reduce taxonomy complexity. While TvsP defined the task to score all 11 AQ subaspects (Local Acceptability, Local Relevance, etc.), 3 dimensions (Cogency, Effectiveness, Reasonableness), and overall AQ, we reduced the number of qualities scored by only focusing on the 3 higher-level dimensions plus overall AQ. As a result, annotators assessed an argumentative text in terms of 4 scores instead of 15 scores, and instead of 3 different AQ levels, the simplified taxonomy is reduced to 2.

(2) Instruction modifications. We reworded the TvsP dimension descriptions and added several examples to make the guidelines more understandable. As the annotators were not rating the 11 AQ subaspects, we experimented with different methods to in-

[16] `http://argumentation.bplaced.net/arguana/data`
[17] `https://nlds.soe.ucsc.edu/iac2`
[18] `https://www.reddit.com/r/changemyview/`

Dimension	Subdimension	Question
Cogency	Local Acceptability	Are the justifications for the argument acceptable/believable?
	Local Relevance	Are the justifications relevant to the author's point?
	Local Sufficiency	Do the justifications provide enough support to draw a conclusion?
Effectiveness	Credibility	Is the author qualified to be making the argument?
	Emotional Appeal	Does the argument evoke emotions that make the audience more likely to agree with the author?
	Clarity	Does the author's language make it easy for you to understand what they are arguing for or against?
	Appropriateness	Is the author's argument and delivery appropriate for an online forum?
	Arrangement	Did the author present their argument in an order that makes sense?
Reasonableness	Global Acceptability	Would the target audience accept the argument and the way it is stated?
	Global Relevance	Does the argument contribute to the resolution of the given issue?
	Global Sufficiency	Does the argument address and adequately rebut counterarguments?

Table 6.10: AQ subdimensions represented as questions in the annotation task of debates.

corporate the subaspects into the guidelines. Instead of explaining the subdimensions in the guidelines and trusting crowd annotators to bear them in mind, we represented each subdimension as a yes/no question in the annotation task itself (Table 6.10). Our pilot experiments showed that presenting the questions without asking for a response eased the perceived complexity of the task while not affecting agreement.

(3) Five-point scale. While TvsP collected judgments with a three-point rating scale (low, medium, high), we employ a five-point scale (very low, low, medium, high, very high, plus *cannot judge*) to allow for more nuanced judgments, as the expert annotators found the distance between the items on a three-point scale too large. Scales with 5–9 items have been shown to be optimal, balancing the informational needs of the researcher and the capacity of the raters (Cox, 1980). We experimented with both three- and five-point scales and found that the larger scale did not negatively affect inter-annotator agreement.

Our finalized task design is as follows: first, annotators decide whether a text is argumentative. Next, if *yes*, each of the three high-level dimensions is scored on a five-point scale and subaspect questions are presented to guide the annotator's judgment. Finally, overall AQ is scored, also on a five-point scale.

	Cogency	Effectiveness	Reasonableness	Overall
Ours	**0.46**	**0.48**	**0.48**	**0.55**
TvsP	0.27	0.38	0.13	0.43

Table 6.11: Agreement between the Dagstuhl "gold" annotations and our crowd-sourced annotations (Ours) compared to the agreement of TvsP.

Validating the Task Design

Before collecting annotations from the crowd, we validated our modifications subjectively and objectively. First, we ran a series of pilot tasks with our expert annotators. They initially annotated using the TvsP guidelines, and next worked with the simplified taxonomy. In follow-up discussions, the experts confirmed that the new task design reduced the cognitive load necessary to rate arguments, and that the guidelines were more understandable. This makes the task more approachable, which is vital when presenting it to (untrained) crowd-workers for larger-scale annotation.

We validated the simplifications quantitatively by reproducing the study of TvsP, which compared their crowd and "expert" annotations. To this end, we randomly sampled 200 arguments from Dagstuhl-ArgQuality-Corpus-V2, which come with author-annotated "gold" ratings. We collected ratings from a crowd (10 ratings per item), following our simplified design.[19] All crowd contributors were native or fluent English speakers engaged through Appen (formerly Figure Eight). Crowd contributors did not participate in calibration meetings, and all feedback was relayed to contributors through a liaison.

We average the crowd ratings to obtain a single score for each argument and computed the IAA with the "gold" annotations using Krippendorff's α (Krippendorff, 2007). The results are shown in Table 6.11. Even though the annotation scores are not strong, the IAA between our crowd annotators and the gold annotations generally surpasses the agreement scores reported by TvsP. This is a highly nuanced and subjective task, which is reflected in the agreement levels. Based on these results and annotator observations, we conclude that our task guidelines and design allow for better (or at least comparable) quality crowd-sourcing of theory-based AQ annotations.

Data

We investigate different domains to obtain a deeper understanding of real-world AQ and the feasibility of the annotation scheme in different settings. We include three domains in our study: CQA forum posts (*CQA*), debate forum posts (*Debates*), and business review forum posts (*Reviews*). While Debates are generally well-explored in computational AQ assessment, we are unaware of any work involving CQA and Reviews. For each of these domains, we first identified items likely to be argumentative and then adjusted the guidelines in consultation with expert annotators, as described below.

[19] The only difference is that we used a 3-point scale to more fairly compare to the gold.

Debate forums. Out of the three domains we investigate, *Debates* is the most straightforward to annotate. Given a topic or motion, users can define their stance (*pro/contra*) and write an argument which supports it. We included data from two online debate forums. ConvinceMe (CM) is a subset of the IAC, where users share their *Stance* on a topic and discuss their point of view, with replies aiming to change the view of the original poster. Change My View (CMV) is a Reddit forum in which participants post their opinion on a topic and ask others to post replies to change their mind. We sampled original posts from CMV, skipping any moderator posts, and the first reply to an original post from CM, in order to limit the context that annotators must consider when evaluating arguments. CMV posts always include the author's perspective in the title, while CM posts may or may not include a stance in the title. In the guidelines, we instruct annotators to judge a post by how successfully it justifies the author's claim.

CQA. In community questions and answers forums, users post questions or ask for advice, which other users can address. We experimented with arguments from Yahoo! Answers.[20] When posting a question, users can provide background information (*context*) and can later indicate which response is the *best answer* to their question. The forum's looser structure provides for a wide variety of content, which is appealing as a potential source of non-standard arguments, but challenging as many of the posts do not contain any arguments. Through manual analysis, we identified three categories that frequently contained controversial topics, hypothesizing they would have a higher incidence of debates: *Social Science* > *Sociology*, *Society & Culture* > *Other*, and *Politics & Government* > *Law & Ethics*. We empirically selected the category with the highest proportion of arguments in a study on Amazon Mechanical Turk. Qualified annotators[21] decided if question and best-answer pairs were argumentative. We collected 10 judgments for 100 pairs from each category and aggregated judgments with a simple majority. *Law & Ethics* had the most argumentative posts (70%, compared to *Sociology* with 40% and *Society & Culture* with 34%), so we sampled posts from this category to annotate.

In the guidelines for this domain, we asked annotators to judge the argumentative strength of an answer with respect to how well it addressed the given question. The guidelines and subdimension questions were altered to encourage this. One obstacle in pilot studies with expert annotators was posts offering advice, as many users solicited legal support in the Law & Ethics forum. We decided to consider advice-giving posts as argumentative as long as the author supported the advice with justifications, which mirrors our general approach to the Argumentative dimension.

Reviews. The third domain consists of restaurant reviews from the Yelp-Challenge-Dataset.[22] On Yelp, users write reviews of businesses and rate the quality of their experience from 1 (low) to 5 (high) stars. Unlike the Debate and CQA forums, the format of Yelp does not support dialogue between users (i.e., users cannot directly reply to other

[20] `https://answers.yahoo.com/`

[21] HIT approval rate >= 97; HITs approved > 500; Location = US

[22] `https://www.yelp.com/dataset`

	Crowd	Experts			Overlap	
# Annotators	10	1	2	3	11–13	**Total size**
CQA	1,334	626	–	625	500	**2,085**
Debates	1,438	600	–	600	538	**2,100**
Reviews	600	200	400	–	100	**1,100**

Table 6.12: Number of arguments annotated by experts and the crowd and the number of overlapping instances (annotated by both experts and the crowd) by domain.

users or posts), and so it is possible to present each post in isolation as a self-contained argument. As most posts do not explicitly state a claim, we pose the star rating as a claim the user is making about the business, and the review as the argument supporting it.

Yelp reviews can be highly subjective as each review is based on a single user's experience. For instance, a user may rate a restaurant as 5-stars and write only *The food was delicious*. To address this subjectivity, we asked annotators to judge the argumentative quality of each review with respect to how well it supported the rating provided. Another challenge was defining what constituted a counterargument, as these have a very different character than counterarguments in debates (e.g., *Everyone says that the pizza crust is too thin here but that's authentic!*). In consultation with our experts, we defined counterarguments by the following characteristics: (1) addressing and rebutting the viewpoints of other reviews, (2) addressing and rebutting points that discredit the author's rating, and (3) bringing up favorable points in an unfavorable review and vice versa.

Experts completed a series of pilots before each domain was presented to the crowd, using the simplified task design. Expert agreement on novel domains (CQA and Reviews) is shown in Table 6.13. Feedback on the task and guidelines was gathered during calibration meetings, and guidelines were iteratively altered to be more clear and specific.

Data Analysis

Applying the annotation task design and data selection described above, we created GAQ-Corpus, containing 5,285 arguments across three domains of online writing, annotated for theory-based dimensions. All arguments were limited to have a length between 70 and 200 characters. Ratings were provided by the two groups of annotators mentioned above, Expert and the Crowd. Each group judged 3,000 arguments, with about 1,000 arguments annotated by both groups for comparison. The size of the corpus is described in Table 6.12. Annotators worked with the domains in the following order: Debate forums, CQA forums, and Review forums. Before switching to a new domain, annotators completed a small study for calibration. All data and guidelines are available from `https://github.com/grammarly/gaqcorpus`.

Domain	Cogency	Effective.	Reasonable.	Overall
CQA	0.16	0.31	0.36	0.29
Debates	0.22	0.33	0.20	0.33
Reviews	0.41	0.19	0.21	0.34

Table 6.13: Agreement (Krippendorff's α) between experts on pilot studies for CQA, Debates, and Reviews (146, 150, and 50 arguments, respectively).

	Cogency	Effective.	Reasonable.	Overall
CQA	0.42	0.52	0.52	0.53
Debates	0.14	0.11	0.21	0.19
Reviews	0.32	0.32	0.31	0.33

Table 6.14: IAA between the mean expert and crowd scores for Cogency, Effectiveness, Reasonableness, and Overall AQ.

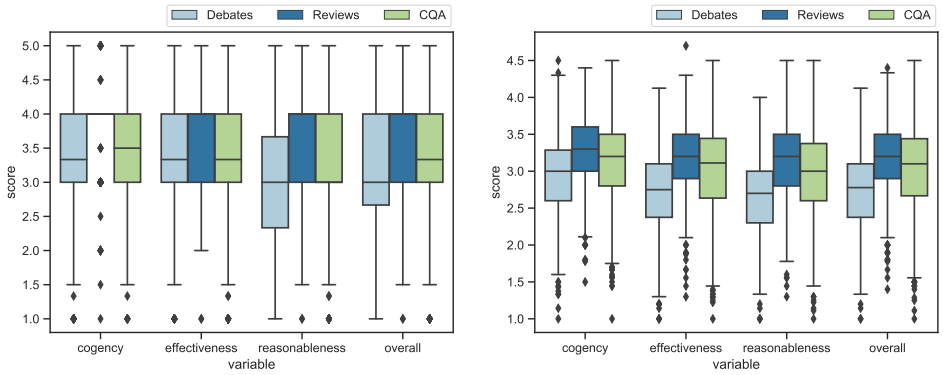
***Title**: Should 'blogging' be a capital crime? Iran is considering it...*
***Stance**: A government has the right to censor speech (...)*

***Text**: My government doesn't give me freedom of speech, so I have to argue for this side. Freedom of speech is bad because ... um ... then Our Leader's beliefs could be challenged. No one wants that. I mean, if everyone would just say and believe what Our Leader says to, we wouldn't need those firing squads altogether! Everyone wins.*

	Cogency	Effectiveness	Reasonableness	Overall
Annotator 1	4	1	1	2
Annotator 2	4	5	3	4
Annotator 3	2	2	2	2

Figure 6.4: Example argument exhibiting disagreement in the Effectiveness dimension.

Inter-Annotator Agreement. We assessed the quality of the crowd annotations by calculating the IAA between the experts and crowd workers on the overlapping portions of GAQCorpus using the mean scores (Table 6.14). For debate forums, the agreement is weak with $\alpha \leq 0.21$, while for the CQA forums, the agreement is higher: 0.42–0.53. These results suggest that the difficulty of the task is highly dependent on the domain. While our Debates data and the Dagstuhl-ArgQuality-Corpus-V2 data both consist of web debate arguments, the difference in IAA is high, which might be attributed to different complexities of the web debates data. While TvsP only look at single arguments in isolation, often consisting of a single sentence only, we look at web posts, which mostly consist of multiple sentences. One area of disagreement centered on arguments, which were sarcastic, ironic, or included rhetorical questions. Consider the argument given in Figure 6.4, over which the expert annotators expressed disagreement. This argument appears to support the stance that a government has the right to censor speech, but several linguistic cues indicate that the argument might be ironic: (a) Punctuation: ellipsis indicates thinking/searching for justifications; similarly, (b) the filler *um*; (c) capitalization: the noun phrase *Our Leader* is capitalized, indicating hyperbolic apotheosis; and finally, (d) the phrase *(...) so I have to argue for this side.* acts like an apologia, which is put in front of the actual argument. Annotators 1 and 2 based their judgments on an interpretation



(a) Expert annotations.

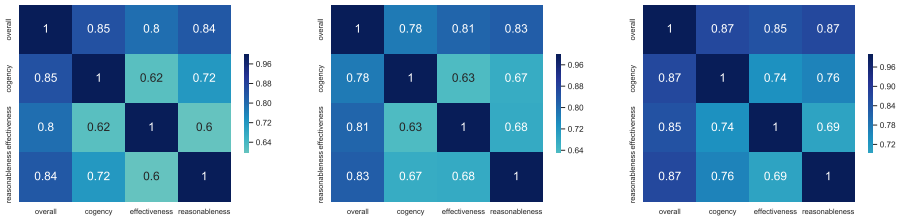
(b) Crowd annotations.

Figure 6.5: Score distributions by domain for expert and crowd annotators.

of this text that related to the estimated degree of irony in the post. While Annotator 1 did not perceive irony and judged the argument as *very weak* in *Effectiveness*, Annotator 2 considered it to be highly effective as in their view, the irony positively underlined the perceived stance. Annotator 3 gave medium scores across the board. Such disagreements were regularly discussed and usually revealed that multiple opinions may exist according to how the texts were interpreted, highlighting the high subjectivity of the task.

Another area of disagreement was how to judge arguments on topics that were deemed "less worthy" of being discussed, and which usually were rather humorous in nature or had trivial consequences, such as *Batman vs. Superman*, in which users argued for the the superiority of either superhero. In our pilot studies, some experts provided lower ratings of arguments on these topic that they considered less worthy. In contrast, others thought that writing a strong, serious argument on a less worthy topic was especially difficult, and thus provided higher ratings for such arguments.

Analysis of the Scores. The distributions of mean scores across domains and annotator groups in GAQCorpus are depicted in Figures 6.5a and 6.5b. In general, the interquartile range of the expert scores was higher than the crowd, suggesting that experts were more specific when scoring items, which is also reflected in the medians: while the crowd exhibits a tendency to score variables equally, expert annotations exhibit more differentiation. To understand the interrelations between Overall AQ and the dimensions, we compute Pearson correlations between the mean scores (Figure 6.6). Generally, the trends are similar across all three domains. For instance, for Debates (Figures 6.6d and 6.6a), the crowd annotations exhibit stronger correlations between the different dimension scores than the experts, with $0.83 \leq r \leq 0.96$. Interestingly, the variance among the Pearson scores is lower, indicating that the crowd tends to distribute ratings for a single instance more consistently while the experts seem to put more weight on differentiating the dimensions. Expert ratings of Overall AQ have substantially stronger correlation with the dimensions than any of the dimension scores with each other, further



(a) Experts on Debates. (b) Experts on Reviews. (c) Experts on CQA.

(d) Crowd on Debates. (e) Crowd on Reviews. (f) Crowd on CQA.

Figure 6.6: Mean score correlations between the different dimensions for expert and crowd annotators across the three domains (Pearson's r).

indicating that experts are more discerning in their scores than the crowd. Across both annotator groups and all domains, the correlation between Overall AQ and Reasonableness is highest, which is consistent with earlier observations (Wachsmuth et al., 2017b).

Qualitative Analysis. We next examine low-scoring arguments from all domains to understand how AQ is perceived differently, focusing on the *Reasonableness* dimension (Table 6.15). The Debate argument raises a counterargument but does not rebut it and additionally neglects to address an obvious counterargument (i.e., the many ethical implications of such a policy). On the other hand, the CQA and Review arguments do not raise or address any counterarguments and are not judged Reasonable for other reasons: the CQA argument jokes about the original poster's question and accuses the poster of malignant behavior, while the Review argument delves into a personal experience that does not contribute to the discussion about the quality of the business.

Standard Split

We provide and use a standard split for each domain, which is composed as follows: The training and development sets consist of the instances which were *either* annotated by our linguistic experts or the crowd workers. In contrast, the test sets encompass only instances scored by *both* experts and the crowd. For each instance and group, we obtain a single score by averaging the annotators' votes. In addition to the group-specific annotations (*expert* and *crowd*), we also compute a *mix* score which is the average of the two group-specific scores. This way, we train on a mix of expert and crowd annotations, where the dominant

Debates		
Cogency	2.0	**Title:** Should you need to pass an IQ test to have kids?– **Stance:** Dumb parents lead to more dumb kids. **Text:** I have a strong opinion that before having children, the prospective parents should have to pass a series of background and IQ tests. Kids being brought into this world need a good foundation to start a successful life with. You may have that limited case where the parents are morons and the kids strive to be different then their failure parents, but in most cases it is an endless line of parasites on our world. We need more smart people.
Effectiveness	1.7	
Reasonableness	1.0	
Overall	1.3	
CQA		
Cogency	2.7	**Question:** Bounced CHECK? **Context:** Does the company holding the bounced check have to send you a certified letter before issuing a warrant for your arrest. I feel almost certain that they do but i am not sure. **Answer:** I always make sure my checks are not printed on rubber. they are just too expensive and not worth it. We all make a mistake from time to time, and usually it is no big deal except for the extreme annoyance and all the bounced check fees. But if you are worried about an arrest warrant then I am sure you are doing this deliberately and trying to defraud the company. You have probably sent them a couple of bad checks already in an attempt to string them along so your guilt is probably pretty well established. You can hope that you do not have to share a jail cell with a gross deviate of some sort.
Effectiveness	2.0	
Reasonableness	1.7	
Overall	2.0	
Reviews		
Cogency	1.0	**Title:** Business review: 2.0 Stars. **Business name:** Cook Out. **City:** Charlotte. **Categories:** Restaurants, Desserts, Food, Fast Food, American (Traditional), Hot Dogs, Burgers **Review:** Burgers are good but I like those other 5 guys burgers instead oh and I guess if your not from around here don't even think about going thru the drive thru it's like the biggest most unreadable confusing hurried crazy thing ever if I ever go again hell with drive thru until I've lived here for at least 5 maybe 10 years and can be a veteran drive thru person I'm walking in it's like if I mix up all the letters in this review and give you 1 minute to read it and figure it out then you gotta move on.
Effectiveness	1.0	
Reasonableness	1.0	
Overall	1.0	

Table 6.15: Low-scoring arguments from all domains.

portion comes from the crowd, and test on overlapping instances, enabling us to compare model performance to both expert and crowd ratings on a static set of instances. The numbers of instances in each portion are given in Table 6.16.

6.2.4 Models

Having developed GAQCorpus to enable computational AQ assessment (**RQ1**), we address the remaining research questions by experimenting with AQ models. To determine whether we can develop a computational theory-based AQ model (**RQ2**), we employ a naive length baseline, three different support vector regression (SVR) models, and a BERT-based (Devlin et al., 2019) model. We next investigate whether the interrelations between AQ dimensions can be exploited in a computational setup (**RQ3**), employing two MTL BERT-based models. For the BERT-based models, we transform each argument into a "BERT-compatible" format, i.e., into a sequence of WordPiece (Johnson et al., 2017) tokens and prepend the sequence with BERT's start token (`[CLS]`). The

Domain	Total	Train	Dev	Test
CQA	2,085	1,109	476	500
Debates	2,100	1,093	469	538
Reviews	1,100	700	300	100
All	5,285	2,902	1,245	1,138

Table 6.16: Number of instances in the train, development, and test sets of GAQCorpus.

pooled hidden representation of the latter corresponds to the aggregated document representation. The specific details of each model are described below.

Argument Length (ARG LENGTH). To estimate the task difficulty and to measure a potential length bias in our data set, our naive baseline is the correlation between the argument's character length and quality scores.

SVR with Lexical Features (SVR_{TFIDF}). We run a simple SVR with TF–IDF representations and test to what extend quality is reflected by purely lexical features.

SVR with Semantic Features (SVR_{EMBD}). We represent each argument as the average of the FASTTEXT (Bojanowski et al., 2017) embedding[23] representations of each word.

Feature-rich SVR (WACHSMUTH_{CFS}). We reimplement the approach of Wachsmuth et al. (2016), who employ standard features (token n-grams, part-of-speech tags, etc.) and higher-level features (sentiment flows, argumentative units, etc.). We run correlation-based feature selection on the training set and include only the most predictive features.

Single-task Learning Setting (BERT ST). For each AQ dimension t, e.g., Effectiveness, we train an individual regressor. Our AQ predictor is a simple linear regression layer in which we feed the pooled document representation. The loss L_t is then simply the mean squared error over the k instances in the training batch.

Flat Multi-Task Learning Setting (BERT MT_{FLAT}). We explore whether a joint training setup would improve the individual score predictions. For each quality dimension, we employ an individual prediction layer as described above and compute an individual task loss. We then define the total training loss as the sum of the task losses.

Hierarchical Multi-Task Learning Setting (BERT MT_{HIER}). We propose a hierarchical MTL setting to exploit the hierarchical relationship between the scores suggested by the taxonomy. Similar to above, we first learn jointly the lower-level tasks (Cogency,

[23] `https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M-subword.vec.zip`

	Model	CQA	D	R		Model	CQA	D	R
	ARG LENGTH	0.406	0.420	0.365		ARG LENGTH	0.390	0.399	0.372
	SVR_{TFIDF}	0.389	0.265	0.450		SVR_{TFIDF}	0.411	0.120	0.340
	SVR_{EMBD}	0.278	0.388	0.265		SVR_{EMBD}	0.293	0.403	0.187
Overall	$WACHSMUTH_{CFS}$	0.492	0.432	0.533	**Effective.**	$WACHSMUTH_{CFS}$	0.523	0.450	0.432
	BERT ST	**0.652**	**0.511**	**0.605**		BERT ST	**0.612**	**0.542**	**0.555**
	BERT MT_{FLAT}	**0.667**	**0.537**	0.588		BERT MT_{FLAT}	**0.671**	**0.570**	0.514
	BERT MT_{HIER}	0.661	0.494	0.593		BERT MT_{HIER}	0.670	0.532	0.486
	ARG LENGTH	0.420	0.437	0.340		ARG LENGTH	0.396	0.377	0.405
	SVR_{TFIDF}	0.444	0.257	0.384		SVR_{TFIDF}	0.457	0.247	0.452
	SVR_{EMBD}	0.261	0.333	0.103		SVR_{EMBD}	0.379	0.258	0.234
Cogency	$WACHSMUTH_{CFS}$	0.503	0.429	0.464	**Reasonable.**	$WACHSMUTH_{CFS}$	0.476	0.399	0.432
	BERT ST	**0.587**	**0.503**	**0.554**		BERT ST	**0.665**	**0.418**	**0.609**
	BERT MT_{FLAT}	0.633	**0.541**	**0.561**		BERT MT_{FLAT}	0.644	**0.473**	0.610
	BERT MT_{HIER}	**0.638**	0.474	0.541		BERT MT_{HIER}	0.626	0.408	**0.611**

Table 6.17: Pearson correlations of our model predictions with the annotation scores on the mix test annotations when training on in-domain data for Community Q&A (CQA), Debates (D), and Reviews (R). Numbers in bold indicate best performances.

Effectiveness, Reasonableness) resulting in three scores $\hat{y}_{\text{Cog}}$, $\hat{y}_{\text{Eff}}$, and $\hat{y}_{\text{Rea}}$. Next, we employ these scores for informing the overall AQ predictor by concatenating these with the hidden document representation $\mathbf{h}_D$: $\mathbf{h}_{informed} = \mathbf{h}_{\widehat{D}}[\hat{y}_{\text{Cog}}, \hat{y}_{\text{Eff}}, \hat{y}_{\text{Rea}}]$. The resulting vector $\mathbf{h}_{informed}$ serves as input to the overall AQ predictor as defined in before.

6.2.5 Experiments

We employ the proposed architectures above to answer research questions RQ2–RQ5.

RQ2: Computational theory-based AQ assessment

To test whether our corpus supports the development of theory-based AQ assessment models, this experiment employs all single-task models presented in Subsection 6.2.4 (ARG LENGTH, SVR_{TFIDF}, SVR_{EMBD} $WACHSMUTH_{CFS}$, and BERT ST). We train and predict on the domain-specific data sets and report the results on the *mix* test set per AQ dimension for each domain.[24] Details on the grid search we conduct for hyperparameter optimization can be found in Part B of the supplementary material.

Results. The respective Pearson correlation scores for AQ dimensions on the three domain-specific test sets are shown in Table 6.17. Generally, we reach medium to high Pearson correlation scores of up to nearly 0.67. However, like the IAA, performance varies across domains: on Debates, the best model, BERT ST, achieves a correlation coefficient with the annotation scores for reasonableness of 0.42 and on the CQA forums,

[24] Trends for the other evaluation sets (crowd and expert) are similar. Full results can be found in Part B of the supplementary material.

it achieves a performance of 0.67. The BERT-based regressor outperforms the other methods, showing that we can successfully utilize a large-scale corpus with theory-based AQ dimensions to train models for automatic AQ assessment (**RQ2**). Note that ARG LENGTH is relatively high across all domains and properties and often outperforms SVR_{TFIDF} and SVR_{EMBD}, indicating a slight length bias in the corpus. However, BERT ST outperforms this baseline in all cases by a large margin, demonstrating this model's ability to capture useful information beyond pure length.

RQ3: Effect of AQ dimension interrelations

Next, we seek to determine whether it is possible to exploit the interrelations between the three dimensions and the overall AQ as suggested by the taxonomy by conducting experiments on GAQCorpus. We compare the MTL architectures, BERT MT_{FLAT} and BERT MT_{HIER}, against the results of the BERT ST model, the best performing single-task model. Again, we train and predict on the domain-specific data splits.

Results. Table 6.17 shows the respective Pearson correlation scores for the four AQ dimensions on each domain. The MTL models outperform the single-task model in 9 out of 12 cases,which suggests that the interrelations between the AQ dimensions and overall AQ can be exploited to improve model performance (**RQ3**). More specifically, the best method is BERT MT_{FLAT}, which outperforms the other methods in 7 cases. BERT ST and BERT MT_{HIER} are best in 3 and 2 cases, respectively.

RQ4: Unified multi-domain model

Relating back to our experiments on domain-specificity from before (see Chapter 5), we examine whether our corpus supports training a unified multi-domain model (**C2**). To this end, we train the BERT-based models on the joint training set covering all domains and test performance on each individual domain, thereby including out-of-domain data in training. Similarly, we optimize the hyperparameters on the joint development set. We compare with the best in-domain score from Table 6.17.

Results. The respective results for the four argument quality dimensions can be seen in Table 6.18. In 11 out of 12 cases, training on all domains increases the performance compared to the best in-domain model. While the resulting models are less domain-specific, the increased amount of data leads to better convergence and leads to gains up to 5 percentage points. This is in-line with our findings on the trade-off between larger and more heterogeneous vs. smaller and more homogeneous corpora from Chapter 5.

RQ5: Synergies between practical and theory-driven AQ

To empirically test the hypothesis that synergies exist between practical and theory-based computational AQ assessment, we conduct a bi-directional experiment with the recently released IBM-Rank-30k corpus (Gretz et al., 2020).

	Model	CQA	Debates	Reviews
Overall	Best in-domain	0.667	0.537	0.605
	BERT ST	0.676	0.545	0.596
	BERT MT$_{FLAT}$	**0.681**	**0.562**	**0.633**
	BERT MT$_{HIER}$	0.665	**0.562**	0.622
Cogency	Best in-domain	0.638	0.541	0.561
	BERT ST	0.608	0.515	0.563
	BERT MT$_{FLAT}$	**0.653**	0.542	0.570
	BERT MT$_{HIER}$	0.638	**0.552**	**0.599**
Effective.	Best in-domain	0.671	0.570	0.555
	BERT ST	**0.686**	**0.598**	0.601
	BERT MT$_{FLAT}$	0.670	0.578	**0.603**
	BERT MT$_{HIER}$	0.653	0.592	0.576
Reasonable.	Best in-domain	**0.665**	0.473	0.611
	BERT ST	0.635	**0.487**	0.603
	BERT MT$_{FLAT}$	0.657	0.486	0.631
	BERT MT$_{HIER}$	0.633	0.483	**0.643**

Table 6.18: Pearson correlations of the model predictions with the annotation scores when training on the joint training sets of all domains. We compare with the best result of the in-domain setting.

Domain	Dimension	r	ρ
BERT IBM	–	0.492	0.456
Gretz et al. (2020)	–	0.52	0.48
All	Overall	**0.313**	**0.303**
	Cogency	0.311	0.300
	Effectiveness	**0.313**	**0.303**
	Reasonableness	0.304	0.298
CQA	Overall	0.258	0.224
	Cogency	**0.269**	**0.228**
	Effectiveness	0.262	0.225
	Reasonableness	0.262	0.226
Debates	Overall	**0.336**	**0.326**
	Cogency	0.331	0.321
	Effectiveness	**0.336**	**0.326**
	Reasonableness	0.333	0.319
Reviews	Overall	0.150	0.145
	Cogency	0.139	0.138
	Effectiveness	**0.152**	**0.151**
	Reasonableness	0.149	0.148

Table 6.19: Performance of BERT MT$_{FLAT}$ trained on GAQCorpus, predicting on IBM-Rank-30k evaluated against the weighted average score.

Experimental setup. IBM-Rank-30k consists of 30,497 crowd-sourced arguments relating to 71 topics, where each argument is restricted to 35–210 characters. The corpus has binary judgments indicating whether raters would recommend the argument to a friend. Based on these ratings, a score for each argument was computed, either using MACE or weighted average of all ratings. Compared to GAQCorpus, IBM-Rank-30k is much larger but the arguments are much shorter and more artificial than real world texts. Manual inspection revealed that the nature of the texts substantially differs from each those in GAQCorpus, i.e., arguments mainly cover reasons for higher-level claims. For example, in IBM-Rank-30k for the topic *"We should end racial profiling"*, a highly rated argument is *"racial profiling unfairly targets minorities and the poor"*. We conduct three experiments in two directions: (E1) train on GAQCorpus, then predict on IBM-Rank-30k, (E2) train on IBM-Rank-30k, then predict on GAQCorpus, and finally, (E3) train on IBM-Rank-30k, next, train on GAQCorpus, and then, predict on GAQCorpus.

For **experiment (E1)**, we take the (already trained) BERT MT$_{FLAT}$ models trained on each domain of GAQCorpus and predict on the test portion of IBM-Rank-30k. This enables us to determine which one of our domains and dimensions are closest to the data and annotations in IBM-Rank-30k. We compare against the best score reported in the Gretz et al. (2020) as well as against our own reimplementation using BERT$_{BASE}$, dubbed BERT IBM.[25] We optimize the BERT IBM baseline by grid searching for the learning rate $\lambda \in \{2e-5, 3e-5\}$ and the number of training epochs $\in \{3, 4\}$ on the

[25]Note that Gretz et al. (2020) do not indicate whether they employ BERT$_{BASE}$ or BERT$_{LARGE}$.

		CQA	Debates	Reviews
Overall	BERT IBM	0.392	0.317	0.154
	BERT IBM MT_{FLAT}	0.666	0.543	0.568
	BERT MT_{FLAT}	**0.681**	**0.562**	**0.633**
Cogency	BERT IBM	0.368	0.274	0.149
	BERT IBM MT_{FLAT}	0.639	0.518	0.541
	BERT MT_{FLAT}	**0.653**	**0.542**	**0.570**
Effectiveness	BERT IBM	0.426	0.378	0.195
	BERT IBM MT_{FLAT}	**0.678**	**0.594**	0.545
	BERT MT_{FLAT}	0.670	0.578	**0.603**
Reasonableness	BERT IBM	0.348	0.246	0.151
	BERT IBM MT_{FLAT}	0.637	0.465	0.581
	BERT MT_{FLAT}	**0.657**	**0.486**	**0.631**

Table 6.20: Pearson correlations on GAQCorpus when predicting with BERT IBM (trained on IBM-Rank-30k) and BERT IBM MT_{FLAT} trained on IBM-Rank-30k in STILT setup fine-tuned on GAQCorpus in comparison to BERT MT_{FLAT}.

IBM-Rank-30k development set. For the already trained models from Sections 6.2.5 and 6.2.5, no further optimization is necessary. In **experiment (E2)**, we reverse the direction of (E1): We train a BERT-based regressor as defined before on the MACE-P aggregated annotations of IBM-Rank-30k.[26] We predict on GAQCorpus and correlate the scores with our annotations. Finally for **experiment (E3)**, in order to flatten out expected losses from the zero-shot domain transfer, inspired by Phang et al. (2018) we use IBM-Rank-30k in the STILT setup, which we discussed in Section 2.2.3. Concretely, we take the trained BERT IBM encoder and continue training the model as BERT IBM MT_{FLAT} in the all-domain setup. We compare both models from (2) and (3) with the BERT MT_{FLAT}.

Results. The results of experiment (E1) are depicted in Table 6.19. As expected, the zero-shot domain transfer results in a large drop compared to training on the associated train set of IBM-Rank-30k. Quite surprisingly, the model trained on the debate forums reaches the highest correlation scores – even higher than the model trained on *all-domains*. Further, in most cases, the effectiveness predictions correlate best with the annotations provided by Gretz et al. (2020). This is in-line with the authors' observations, who manually had to annotate the data for the theory-based scores.

Table 6.20 displays the results of (E2)–(E3). Experiment (E2), draws a similar picture: the zero-shot domain transfer using BERT IBM results in a huge loss in performance compared to BERT MT_{FLAT}. Finally, the results in (E3) indicate potential for using resources drawn from practical approaches in a theory-based AQ assessment scenario: when reusing the encoder in the STILT setup, BERT IBM MT_{FLAT}, the losses originating from the zero-shot domain transfer can be flattened out – in some cases even outperforming

[26] This corresponds to our BERT IBM baseline from before.

BERT MT_{FLAT}. This is especially the case when correlating the predictions with our annotations for the Effectiveness dimension. To sum up, our experiment (E1)–(E3) yield the following findings: (1) large-scale predictions, obtained from a theory-based AQ model on a large (practical) AQ data set, correlate mostly with the Effectiveness dimension. (2) The transferred knowledge obtained in the STILT-setup on IBM-Rank-30k in BERT IBM MT_{FLAT} improves the performance on GAQCorpus for Effectiveness the most. These two facts match Gretz et al. (2020)'s hypothesis that their annotations mostly captured Effectiveness. To summarize, with these experiments, we empirically substantiate the idea (without any manual effort) that, on the one hand, a theory-based approach can inform practical AQ research and increase interpretability of practically-driven research outcomes. On the other hand, the practical approach can increase the efficacy of theory-based AQ models when targeting a matching domain and dimension.

6.2.6 Conclusion

Specific assessment of the rhetorical, logical, and dialectical perspectives on argumentative texts can inform researchers, e.g., about phenomena captured within their annotation study, and help people improve their writing skills by providing targeted feedback. However, the field of computational AQ assessment has been almost exclusively driven by practical approaches. Aiming to fill this research gap, in this Section, we advanced theory-based computational AQ research with the following contributions: we performed a large-scale annotation study on English argumentative texts covering debate forums, CQA forums, and business review forums. We thereby presented GAQCorpus, the largest and first multi-domain corpus annotated with theory-based AQ scores (**RQ1**). Next, we proposed the first computational theory-based AQ models (**RQ2**) and demonstrated that jointly predicting AQ scores can improve the performance of the models (**RQ3**) thereby exploiting the complementarity of knowledge across the AQ assessment dimensions (**C3**). Furthermore, we showed that in most cases, models benefit from including out-of-domain training data (**RQ4**, **C2**). Finally, we investigated concrete synergies between the practical and the theory-based approach to AQ assessment in a bi-directional experimental setup (**RQ5**). The theory-based models can help to increase the interpretability of practical approaches, and practical approaches can be employed to increase the performance of the theory-based models, another example of the complementarity of knowledge in language representations for computational AQ assessment (**C3**).

In this Chapter, we have presented two case studies that focus on understanding the complementarity of knowledge (**C3**) across two argumentative understanding problems, (1) analysis of scitorics, and (2) AQ assessment. In both cases, we have demonstrated performance improvements when coupling different CA tasks using inductive transfer learning techniques. In the next Chapter, we focus on multilinguality (**C4**).

Chapter 7

Multilinguality

[*]Given that argumentation is supposed to exist in most, if not all, human civilizations, a challenge for language representations in computational argumentation is multilinguality (**C4**, see Section 3.4). This issue can be addressed by employing cross-lingual transfer, which is in its most extreme case when no data for the target task in the target language is employed, termed zero-shot transfer. Here, massively multilingual transformers pretrained via language modeling (e.g., mBERT, XLM-R) have become a default paradigm in NLP, offering unmatched transfer performance. Current evaluations, however, verify their efficacy in transfers (a) to languages with sufficiently large pretraining corpora and (b) between close languages. In this work, we analyze the limitations of downstream language transfer with MMTs, showing that, much like cross-lingual word embeddings, they are substantially less effective in resource-lean scenarios and for distant languages. Our experiments, encompassing two higher-level semantic tasks with NLI as an instance of argumentative reasoning (see Section 2.1.4), plus question answering (QA), empirically correlate transfer performance with linguistic proximity between source and target languages, but also with the size of target language corpora used in MMT pretraining. Finally, we demonstrate that inexpensive few-shot transfer (i.e., additional fine-tuning on a few target-language instances) is effective across the board, warranting more research efforts reaching beyond the limiting zero-shot conditions.

7.1 Introduction

Labeled data sets of sufficient size support supervised learning in CA and NLP. The notorious tediousness, subjectivity, and cost of linguistic annotation (Dandapat et al., 2009; Sabou et al., 2012; Fort, 2016), coupled with plethora of structurally different NLP tasks, lead to existence of such data sets only for a handful of resource-rich languages (Bender, 2011; Ponti et al., 2019a; Joshi et al., 2020). This data scarcity renders the need for effective

[*]This Chapter is adapted from: **Anne Lauscher**, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From Zero to Hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online, November 2020, Association for Computational Linguistics. The published version also includes results of experiments on lower-level tasks carried out by Vinit Ravishankar.

cross-lingual transfer strategies (see Section 2.2.3): how can we exploit abundant labeled data from resource-rich languages to make predictions in resource-lean languages? In the most extreme scenario, termed *zero-shot cross-lingual transfer*, not a single labeled instance exists for a target language. Recent work has placed much emphasis on this scenario exactly; in theory, it offers the widest portability across the world's 7,000+ languages (Pires et al., 2019; Artetxe et al., 2020b; Lin et al., 2019; Cao et al., 2020; Hu et al., 2020).

The current mainstay of cross-lingual transfer in NLP are approaches based on continuous cross-lingual representation spaces such as cross-lingual word embedding spaces (Ruder et al., 2019) and, most recently, massively multilingual transformer (MMT) networks, pretrained on multilingual corpora with language modeling (LM) objectives (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020a). The latter have *de facto* become the default language transfer paradigm, with multiple studies reporting their unparalleled transfer performance (Pires et al., 2019; Wu and Dredze, 2019; Rönnqvist et al., 2019; K et al., 2020; Conneau et al., 2020b).

Key Questions and Contributions. In this Chapter, we dissect the current state-of-the-art MMT-based approach to (zero-shot) cross-lingual transfer and analyze a variety of conditions and factors that critically impact or limit effective cross-lingual transfer. Our aim is to provide answers to the following crucial questions.

(RQ1) *What is the effect of language (dis)similarity and language-specific corpora size in pretraining on the zero-shot transfer performance?*

Current cross-lingual transfer via MMTs is still primarily focused on either (1) languages that are typologically or etymologically close to English (e.g., German, Scandinavian languages, French, Spanish), or (2) languages with large monolingual corpora, well-represented in the multilingual pretraining corpora (e.g., Arabic, Hindi, Chinese). Conneau et al. (2020b) suggest that LM-pretrained transformers, much like static word embeddings models, produce topologically similar representation spaces that can easily be aligned between languages, offering this as explanation of language transfer efficacy of MMTs. However, transfer with static CLWEs has been shown ineffective between dissimilar languages (Søgaard et al., 2018; Vulić et al., 2019) or languages with small corpora (Vulić et al., 2020). We thus scrutinize MMTs in diverse zero-shot transfer settings and find, in line with prior work on CLWEs, that MMTs' transfer performance critically depends on (1) linguistic (dis)similarity between the source and target language and (2) size of the pretraining corpus of the target language.

(RQ2) *Can we (even) predict transfer performance?*

Running a simple regression on available transfer results, we show that we can (roughly) predict the transfer performance from the combination of language proximity and size of target-language pretraining corpora for our two high-level semantic tasks.

(RQ3) *Should we focus more on few-shot transfer scenarios and quick annotation cycles?*

Complementing the efforts on improving zero-shot transfer (Cao et al., 2020), we point to few-shot transfer as a very effective mechanism for improving target-language performance. Similar to the seminal "pre-neural" work of Garrette and Baldridge (2013), our results suggest that only several hours (or even minutes) of annotation work can "buy"

substantial performance gains for low-resource target languages. For both tasks in our study, we obtain substantial improvements with minimal annotation effort. Crucially, the few-shot gains are most pronounced exactly where zero-shot transfer fails: for distant target languages with small monolingual corpora.

7.2 Related Work

For completeness and as a reminder on the language representations and the cross-lingual transfer fundamentals discussed in Sections 2.2.2 and 2.2.3, we provide a brief overview of **1)** cross-lingual transfer approaches, with a focus on **2)** MMT models, and then **3)** position our work w.r.t. other studies that examine different properties of MMTs.

7.2.1 Cross-Lingual Transfer Paradigms

Language transfer entails representing texts from both the source and target language in a shared cross-lingual space. Transfer paradigms based on discrete language representations include *machine translation (MT)* of target language text to the source language (or vice-versa) (Mayhew et al., 2017; Eger et al., 2018), and grounding texts from both languages in *multilingual knowledge bases* KBs (Navigli and Ponzetto, 2012; Lehmann et al., 2015). While reliable MT hinges on availability of large parallel corpora, transfer via multilingual KBs (Camacho-Collados et al., 2016; Mrkšić et al., 2017) is impaired by the limited KB coverage and inaccurate entity linking (Moro et al., 2014; Raiman and Raiman, 2018).

Therefore, recent years have seen a surge of language transfer methods based on continuous representation spaces. The previous state-of-the-art, CLWEs (Mikolov et al., 2013b; Ammar et al., 2016; Artetxe et al., 2017; Smith et al., 2017; Glavaš et al., 2019; Vulić et al., 2019; Glavaš and Vulić, 2020) and sentence embeddings (Artetxe and Schwenk, 2019), have most recently been replaced by MMTs pretrained with LM objectives (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020a).

7.2.2 Massively Multilingual Transformers

Multilingual BERT (mBERT). As we have already discussed in Section 2.2.2, BERT's (Devlin et al., 2019) core is a multi-layer transformer network (Vaswani et al., 2017), parameters of which are pretrained using language modeling objectives, MLM and NSP. Liu et al. (2019) introduce RoBERTa, a more robust instance of BERT trained on larger corpora using only the MLM objective. mBERT is an instance of BERT trained on concatenation of the 104 largest Wikipedias. The effects of underfitting for languages with small Wikipedias and overfitting to languages with large Wikipedias are respectively attenuated with exponentially smoothed up-sampling and down-sampling of the data.

XLM-RoBERTa (XLM-R). Conneau et al. (2020a) present XLM-R, an instance of RoBERTa, which is robustly trained on a large multilingual CommonCrawl-100 (CC-100) corpus (Wenzek et al., 2020) covering 100 languages. mBERT's pretraining corpus and CC-100 share 88 languages, with the corresponding portions of CC-100 being much larger than the Wikipedias employed to train mBERT.

The "Curse of Multilinguality". For XLM-R, Conneau et al. (2020a) observe that for a fixed model capacity, downstream cross-lingual transfer improves with more pretraining languages up to a point after which adding more pretraining languages hurts the downstream transfer. This effect, termed the "curse of multilinguality", can be mitigated by increasing the capacity of the model (Artetxe et al., 2020b) or additional training for particular language pairs (Pfeiffer et al., 2020). This points to MMTs' capacity (i.e., computational budgets), as a critical factor for effective zero-shot transfer.

In contrast, we identify few-shot target-language cross-lingual transfer as a much more cost-effective strategy for improving downstream target language performance (Section 7.4). We show for a number of target languages and two downstream tasks that one can obtain consistent performance gains with very small annotation cost, without having to pretrain from scratch an MMT of larger capacity.

7.2.3 Cross-Lingual Transfer with MMTs

A body of recent work probed the knowledge encoded in MMTs, primarily MBERT. Libovický et al. (2020) analyze language-specific versus language-universal knowledge encoded in MBERT. Pires et al. (2019) demonstrate MBERT to be effective for part-of-speech POS tagging and named entity recognition (NER) zero-shot transfer between related languages. Wu and Dredze (2019) extend this analysis to more tasks and languages, and show that MBERT-based transfer is on a par with the best task-specific zero-shot transfer approaches. Similarly, K et al. (2020) prove MBERT to be effective for NER and NLI transfer to Hindi, Spanish, and Russian.[1] Importantly, they show that transfer effectiveness does not depend on the vocabulary overlap between the languages.

In most recent work, concurrent to this, Hu et al. (2020) introduce XTREME, a benchmark for evaluating multilingual encoders encompassing 9 tasks and 40 languages.[2] While the primary focus is a large-scale zero-shot transfer evaluation, they also experiment with target-language fine-tuning (1,000 instances for POS and NER). While Hu et al. (2020) focus on the evaluation aspects and protocols, in this work, we provide a more detailed analysis of the factors that hinder effective zero-shot transfer across several tasks.[3] We also put more emphasis on few-shot transfer and approach it differently: by sequentially fine-tuning MMTs, first on (larger) source language training data and then on few target-language instances. Artetxe et al. (2020b) and Conneau et al. (2020b) analyze different monolingual BERTs to explain transfer efficacy of MBERT. They find topological similarities between monolingual spaces, suggesting these are responsible for effective language transfer with MMTs. In essence, their work recasts the well-known assumption of approximate isomorphism of monolingual representation spaces (Søgaard et al., 2018). For CLWEs, this assumption does not hold for distant languages (Søgaard et al., 2018; Vulić et al., 2019), and in face of monolingual corpora of small size (Vulić et al., 2020). We demonstrate that the same is the case for language transfer with MMTs.

[1] Note that all three are high-resource Indo-European languages with large Wikipedias.

[2] None of the individual tasks in XTREME covers all 40 languages, but much smaller language subsets.

[3] We leave an even more general analysis that combines transfer both across tasks (Pruksachatkun et al., 2020; Glavaš and Vulić, 2020) *and* across languages for future work.

7.3 Zero-Shot Transfer: Analyses

We first address RQ1: we conduct zero-shot language transfer experiments for our two tasks and analyze the factors behind the varying performance drops across languages.

7.3.1 Experimental Setup

Tasks and Languages. We experiment with two high-level NLU tasks: NLI, a task in argumentative reasoning, and QA, which similarly requires deep semantic knowledge.

Cross-Lingual Natural Language Inference (XNLI). We evaluate on the XNLI corpus (Conneau et al., 2018), which was created by translating the development and test portions of the English MNLI data (Williams et al., 2018) by professional translators. XNLI covers 14 languages (French (FR), Spanish (ES), German (DE), Greek (EL), Bulgarian (BG), Russian (RU), Turkish (TR), Arabic (AR), Vietnamese (VI), Thai (TH), Chinese (ZH), Hindi (HI), Swahili (SW), and Urdu (UR)).

Cross-lingual Question Answering Dataset (XQuAD). We rely on the XQuAD data set (Artetxe et al., 2020b), created by translating the 240 dev paragraphs (from 48 documents) and corresponding 1,190 QA pairs of SQuAD V1.1 (Rajpurkar et al., 2016) to 11 languages (ES, DE, EL, RU, TR, AR, VI, TH, ZH, and HI). In order to allow for a comparison between zero-shot and few-shot transfer (see Section 7.4), we reserve 10 documents as the development set for our experiments and evaluate on the remaining 38 articles.[4]

Fine-tuning. We perform standard downstream fine-tuning of LM-pretrained MBERT and XLM-R.[5] We add the following task-specific architectures on top of the two MMTs: for XNLI, we apply a simple softmax classifier on the vector of the sequence start token (`[CLS]` for MBERT; `<s>` for XLM-R); in the case of XQuAD, we pool the MMT's representations of all input subwords and forward these to a span classification head – a linear layer computing the start and the end of the answer.

Training and Evaluation Details. We experiment with MBERT_{BASE} in the *cased* version and XLM-R_{BASE}, both with $L = 12$ transformer layers, hidden size of $H = 768$, and $A = 12$ self-attention heads. For XNLI, we limit the inputs to $T = 128$ subword tokens and train in batches of 32 instances. For XQuAD, we limit paragraphs to $T = 384$ tokens and questions to $Q = 64$ tokens. We slide over paragraphs with a window of 128 tokens and train in batches of size 12. For both of our tasks, we search in the following hyperparameter grid: learning rate $\lambda \in \{5 \cdot 10^{-5}, 3 \cdot 10^{-5}\}$; training epochs $n \in \{2, 3\}$. We optimize all models with Adam (Kingma and Ba, 2015).

[4] As a general note, while the effects of "translationese" might have some impact on the absolute numbers (Artetxe et al., 2020a), they are not prominent enough to have any impact on the relative trends in the reported results. For both XNLI and XQuAD, the translations were done completely manually and not via post-editing of MT (which would pose a higher "translationese" risk). Moreover, having an independently created test set in each language would impede comparability across languages.

[5] We tokenize the input for each model with the corresponding pretrained fixed-vocabulary tokenizer: WordPiece tokenizer (Johnson et al., 2017) with the vocabulary of 110K tokens for MBERT, and the SentencePiece byte-pair encoding tokenizer (Sennrich et al., 2016) with the vocabulary of 250K tokens for XLM-R.

Task	Model	EN	ZH Δ	TR Δ	RU Δ	AR Δ	HI Δ	VI Δ	TH Δ	ES Δ	EL Δ	DE Δ	FR Δ	BG Δ	SW Δ	UR Δ
XNLI	B	82.8	-13.6	-20.6	-13.5	-17.3	-21.3	-11.9	-28.1	-8.1	-14.1	-10.5	-7.8	-13.3	**-33.0**	-23.4
	X	84.3	-11.0	-11.3	-9.0	-13.0	-14.2	-9.7	-12.3	-5.8	-8.9	-7.8	-6.1	-6.6	**-20.2**	-17.3
XQuAD	B	71.1	-22.9	-34.2	-19.2	-24.7	-28.6	-22.1	**-43.2**	-16.6	-28.2	-14.8	-	-	-	-
	X	72.5	**-26.2**	-18.7	-15.4	-24.1	-22.8	-19.7	-14.8	-14.5	-15.7	-16.2	-	-	-	-

Table 7.1: Zero-shot cross-lingual transfer performance on XNLI, and XQuAD with MBERT (B) and XLM-R (X). We show the monolingual EN performance and report drops in performance relative to EN for all target languages. Numbers in bold indicate the largest zero-shot performance drops for each task.

7.3.2 Results and Preliminary Discussion

A summary of the zero-shot cross-lingual transfer results, per target language, is provided in Table 7.1. For XNLI we report accuracy, and for XQuAD, we report the Exact Match (EM) score. As expected, we observe drops in performance for all tasks and all target languages w.r.t. reference EN performance. However, the drops vary greatly across languages. For example, XNLI transfer with XLM-R yields a moderate 6.1 percentage points drop for FR, but a large 20 percentage points drop for SW, and, similarly, for XQuAD with MBERT we note a moderate drop of 14.8 percentage points for DE, but a huge 43.2 percentage points drop for TH. At first glance, it appears – as suggested in prior work – that the transfer drops primarily correlate with language proximity: they are more pronounced for languages that are more distant from EN (e.g., ZH, AR, TH, SW). But we also see that language proximity alone does not explain many of the XNLI and XQuAD results. For instance, RU XNLI (for both MBERT and XLM-R) is comparable to that of ZH, and lower than that for HI and UR: this is despite the fact that, as Indo-European languages, RU, HI, and UR are linguistically closer to EN than ZH. Similarly, we observe comparable performance on XQuAD for TH, RU, and ES.

7.3.3 Analysis

For both tasks, we now analyze the correlations between transfer performance and **a)** several measures of linguistic proximity (i.e., similarity) between languages and **b)** the size of MMT pretraining corpora of each target language.

Language Vectors and Corpora Sizes. For estimates of linguistic similarity, we rely on language vectors from LANG2VEC, which encode various linguistic features from the URIEL database (Littell et al., 2017). We consider the following LANG2VEC vectors: `syntax` (SYN) vectors encode syntactic properties, e.g., if a subject appears before or after a verb; `phonology` (PHON) vectors encode phonological properties such as the consonant-vowel ratio; `inventory` (INV) vectors denote presence or absence of natural classes of sounds (e.g., voiced uvulars); FAM vectors encode memberships in `language families`; and GEO vectors express orthodromic distances for languages w.r.t. fixed points on the Earth's surface. Language proximity is computed as cosine similarity between the languages' corresponding LANG2VEC vectors: each vector type (e.g., SYN)

		SYN		PHON		INV		FAM		GEO		SIZE	
Task	**Model**	Pears	Spear	Pears	Spear	Pears	Spear	Pears	Spear	Pears	Spear	Pears	Spear
XNLI	XLM-R	**0.88**	**0.90**	0.29	0.27	0.31	-0.11	0.63	0.54	0.54	0.74	0.70	0.76
	MBERT	**0.87**	0.86	0.21	0.08	0.29	0.04	0.61	0.47	0.55	0.67	0.77	**0.91**
XQuAD	XLM-R	0.69	0.53	**0.85**	**0.81**	0.62	-0.01	**0.81**	0.54	0.43	0.50	**0.81**	0.55
	MBERT	0.84	0.89	0.56	0.48	0.55	0.22	0.79	0.64	0.51	0.55	**0.89**	**0.96**

Table 7.2: Correlations between zero-shot transfer performance with MBERT and XLM-R for XNLI and XQuAD with linguistic proximity features (SYN, PHON, INV, FAM and GEO) and pretraining size of target-language corpora (SIZE). Results reported in terms of Pearson (Pears) and Spearman (Spear) correlation coefficients.

Task	**Model**	**Selected features**	Pears	Spear	MAE
XNLI	XLM-R	SYN (.51); SIZE (.49)	0.84	0.85	2.01
	MBERT	SYN (.35); SIZE (.34); FAM (.31)	0.89	0.90	2.78
XQuAD	XLM-R	PHON (.99)	0.95	0.83	2.89
	MBERT	SIZE (.99)	0.89	0.93	4.76

Table 7.3: Results of the meta-regression analysis, i.e., predicting zero-shot transfer performance for MBERT and XLM-R. For each task-model pair we list only features with weights ≥ 0.01. Pears=Pearson; Spear=Spearman; MAE=Mean Absolute Error.

produces one similarity score (i.e., feature). We couple LANG2VEC features with the z-normalized size of the target language corpus used in MMT pretraining (SIZE).[6]

Correlation Analysis. We first correlate individual features with the zero-shot transfer scores for each task and show the results in Table 7.2. SYN correlates well with all transfer results except with XLM-R results on XQuAD. Somewhat surprisingly, the phonological language similarity (PHON) correlates best with transfer performance with XLM-R for XQuAD. For both tasks and both MMTs, we observe very high correlations between the transfer performance and the size of pretraining corpora of the target language (SIZE). We believe that this reflects the fact that high-level NLU tasks, such as argumentative reasoning, rely on rich representations of semantic phenomena of a language for which it takes a large amount of distributional data to acquire.

Meta-Regression. Across the tasks, we observe high correlations between zero-shot transfer results and several features (e.g.,SYN, PHON and SIZE). We next test if we can predict the transfer performance for a new language by (linearly) combining individual features. For each task, we fit a linear SVR using transfer results for target languages as labels. With only between 11 and 14 target languages (i.e., instances for fitting the regressor) per task, we resort to leave-one-out cross-validation to obtain correlations for feature combinations. We perform greedy forward feature selection: in each iteration, we add the

[6]For XLM-R, we take reported sizes of language-specific CC-100 portions (Conneau et al., 2020a); for MBERT, we work with sizes of language-specific Wikipedias.

		k	$k=10$		$k=50$		$k=100$		$k=500$		$k=1000$	
Task	Model	$k=0$	score	Δ	score	Δ	score	Δ	score	Δ	score	Δ
XNLI	mBERT	65.92	65.89	-0.03	65.08	-0.84	64.92	-1.00	67.41	1.49	68.16	2.24
	XLM-R	73.32	73.73	0.41	73.76	0.45	75.03	1.71	75.34	2.02	75.84	2.52
			$k=2$		$k=4$		$k=6$		$k=8$		$k=10$	
XQuAD	mBERT	45.62	48.12	2.50	48.66	3.04	49.34	3.72	49.91	4.29	50.19	4.57
	XLM-R	53.68	53.73	0.05	53.84	0.17	54.76	1.08	55.56	1.88	55.78	2.10

Table 7.4: Results of the few-shot experiments with varying numbers of target-language examples k. For each k, we report the performance averaged across all languages and the difference (Δ) with respect to the zero-shot setting.

feature which boosts correlation (obtained via leave-one-out cross-validation) the most; we stop when none of the remaining features further improves the Pearson correlation.

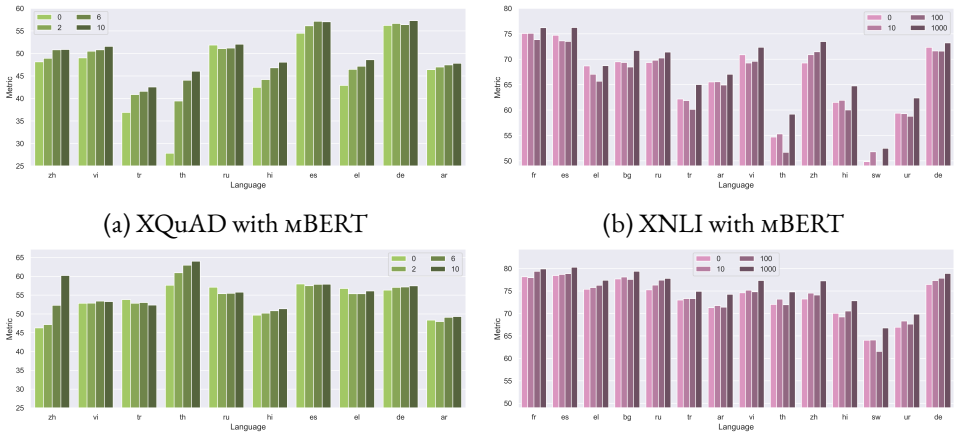
We summarize the results of this meta-regression analysis in Table 7.3. For each task-model pair, we list features selected with the greedy feature selection and show (normalized) weights assigned to each feature. Combinations of features manage to yield higher correlations with zero-shot transfer results than any of the features on their own. These results empirically confirm our previous intuition that linguistic proximity between the source and target language only partially explains zero-short transfer performance. On XNLI, transfer performance is best explained with the combination of structural similarity between languages (SYN) and the size of the target-language pretraining corpora (SIZE); on XQuAD with mBERT, SIZE alone best explains zero-short transfer scores. Note that the features are mutually quite correlated as well (e.g., languages closer to EN also tend to have larger pretraining corpora): thus, if the regressor selects only one feature, this does not mean that other features do not correlate with transfer performance (as shown by Table 7.2). The coefficients in Table 7.3 again indicate the importance of SIZE for the language understanding tasks and highlight our core finding: pretraining corpora sizes are strong features for predicting zero-shot performance in higher-level semantic.

7.4 Few-Shot Target-Language Fine-Tuning

Motivated by the low zero-shot transfer performance for many languages obtained on both tasks in Section 7.3, we now investigate Q3 from Section 7.1: we aim to mitigate transfer losses with inexpensive few-shot cross-lingual transfer.

Experimental Setup. We rely on the same models, tasks, and evaluation protocols as described in Subsection 7.3.1. However, instead of fine-tuning the MMTs on task-specific data in EN only, we continue the fine-tuning process by feeding k additional training examples randomly chosen from reserved target language data portions, disjoint with the test sets.[7] For both tasks, we run the experiments five times and report the average scores.

[7]Note that for XQuAD, we performed the split on the article level to avoid topical overlap. Consequently, for XQuAD k refers to the number of articles.



(a) XQuAD with MBERT

(b) XNLI with MBERT

(c) XQuAD with XLM-R

(d) XNLI with XLM-R

Figure 7.1: Few-shot transfer results for each language with varying k for a) XQuAD with MBERT, b) XNLI with MBERT, c) XQuAD with XLM-R, d) XNLI with XLM-R. For XNLI k denotes the number of sampled sentences, for XQuAD, the number of articles.

7.4.1 Results and Discussion

The results on the two tasks, conditioned on the number of few-shot examples k and averaged across all target languages, are presented in Table 7.4. We note consistent performance improvements in few-shot learning setups for both tasks. The maximum gains for XNLI and XQuAD after seeing $k = 1,000$ target-language instances and 10 articles, respectively, are between 2.52 (XLM-R) and 4.57 points (MBERT).

Figure 7.1 illustrates few-shot performance for individual languages for XNLI, and XQuAD for different values of k.[8] Across languages, we see a clear trend – more distant target languages benefit much more from the few-shot data. Observe, e.g., DE for XQuAD with MBERT. It is closely related to EN, exhibits high zero-shot transfer performance, and benefits only marginally from few in-language instances. We hypothesize that for such closely related languages, with enough pretraining data, MMT is able to extrapolate the missing language-specific knowledge from few in-language examples; its priors for languages close to EN are already quite sensible and *a priori* offer less room for improvements. In stark contrast, TH for XQuAD with MBERT, for example, exhibits poor zero-shot performance and understandably so, given their linguistic distance to EN. Given in-language data, however, it sees rapid leaps in performance, displaying gains of almost 5 percentage points, and we observe already substantial improvement from only 2 in-language documents. This can be seen as MMTs' ability to rapidly learn to utilize the multilingual space to adjust its task-specific knowledge for the target language.

In sum, we see the largest gains from few-shot transfer exactly for languages for which the zero-shot transfer setup yields the largest performance drops: languages distant from EN and represented with small corpora in MMT pretraining.

[8] Exact numbers are provided in Part C of the supplementary material.

Task	Number of Instances	Cost estimate	Δ MBERT	Δ XLM-R
NLI	1,000 sentence pairs	$10	+2.24	+2.54
QA	10 docs	$30	+4.5	+2.1

Table 7.5: Conversion rates between target language annotation costs and corresponding average performance gains from MMT-based few-shot language transfer.

Direct Target Language Few-Shot Fine-Tuning. We additionally ran a set of control experiments in which we bypass the task-specific fine-tuning on the English data and directly fine-tune the MMTs on the few target language instances. Expectedly, fine-tuning the MMTs with only a handful of target language examples (i.e., *without* prior fine-tuning in English) yields subpar performance with respect to the corresponding model variant that had been previously fine-tuned on English data. For instance, direct few-shot target language fine-tuning of MBERT yields the average XNLI performance of 33.95 for $k = 100$ and 40.19 for $k = 1,000$, respectively (compared to 64.92 and 68.16, respectively, when prior fine-tuning on English data is performed). These findings suggest that fine-tuning with abundant (English) in-task data plus fine-tuning with scarce in-language in-task data yields a truly synergistic effect: the small number of examples in the target language is not sufficient to adapt the MMT directly, but they can provide a substantial edge over fine-tuning only on the English data (i.e., zero-shot transfer).

7.4.2 Cost of Language Transfer Gains

As shown in Subsection 7.4.1, moving to few-shot target-language transfer can improve the performance and reduce the gaps observed with zero-shot transfer, especially for low-resource languages. While additional fine-tuning on few target-language examples is computationally cheap, data annotation may be expensive, especially for minor languages. What are the annotation costs, and how do they translate into performance gains? Table 7.5 provides ballpark estimates for both evaluation tasks; the estimates are based on annotation costs from the literature (Marelli et al., 2014; Rajpurkar et al., 2016).

Natural Language Inference. Marelli et al. (2014) reportedly paid $\$2,030$ for 200k judgements, which would amount to $\$0.01015$ per NLI instance and, in turn, to $\$10.15$ for $1,000$ annotations. In our few-shot experiments this would yield an average improvement of 2.24 and 2.52 accuracy points for MBERT and XLM-R, respectively. It is also possible to translate the English data directly via professional translation services as done with the XNLI data set and XQuAD: platforms for hiring professionals, e.g., Upwork, show that it is possible to find qualified translators even for lower-resource languages: e.g., the translation cost estimate for Zulu is $\$12.5$-$\$16$/h, or $\$19$/h for the Basque language.

Question Answering. Rajpurkar et al. (2016) report a payment cost of $9 per hour and a time effort of 4 minutes per paragraph. With an average of 5 paragraphs per article, our few-shot scenario (10 articles) roughly requires 50 paragraphs-level annotations, i.e., 200 minutes of annotation effort and would in total cost around $30 (for respective performance improvements of 4.6 and 2.1 points for MBERT and XLM-R).

A provocative high-level question that calls for further discussion in future work can be framed as: are GPU hours effectively more costly[9] than data annotations are in the long run? While MMTs are extremely useful as general-purpose models of language, their potential for some (target) languages can be quickly unlocked by pairing them with a small number of annotated target-language examples. Effectively, this suggests leveraging the best of both worlds, i.e., coupling knowledge encoded in large MMTs with a small annotation effort to foster inclusive and sustainable language representations for CA.

7.5 Conclusion

A vital challenge on the intersection of CA and language representations is multilinguality (**C4**, see Section 3.4). Here, research on zero-shot language transfer is motivated by inherent data scarcity: the fact that most languages have no annotated data for most CA and NLP tasks. Massively multilingual transformer models have recently been praised for their zero-shot transfer capabilities that mitigate the data scarcity issue. In this Chapter, we have demonstrated that, similar to earlier language transfer paradigms, MMTs perform poorly in zero-shot transfer to distant target languages and to languages with smaller monolingual corpora available for exploitation in MMT pretraining. We have presented a detailed empirical analysis of factors affecting zero-shot transfer performance of MMTs across two tasks and multiple diverse languages. Our results have revealed that the pretraining corpora size of the target language is crucial for explaining transfer results for higher-level language understanding tasks, i.e., natural language inference and question answering. Finally, we have shown that the MMT potential on distant and low-resource target languages can be quickly unlocked if they are provided a handful of annotated instances in the target language. This finding provides a strong incentive for intensifying future research efforts that focus on cheap or naturally occurring supervision (Vulić et al., 2019; Artetxe et al., 2020c; Marchisio et al., 2020), quick and simple annotation procedure, and the more effective few-shot transfer learning setups.

Next, we move to our last challenge, which deals with ethical considerations with regard to language representations (**C5**). Here, we focus on the issue of stereotypical bias.

[9] Financially, but also ecologically (Strubell et al., 2019).

Chapter 8

Ethical Considerations

As discussed in Section 3.5, previous research has noted several ethical issues in the context of language representations (**C5**). In light of these challenges, we have already addressed two problems that arise in relation to computational argumentation: (1) to foster *inclusion* of speakers of languages other than English in CA technologies, we have acknowledged the inherently multilingual nature of argumentation and analyzed the size of the performance gaps arising in the current state-of-the-art zero-shot cross-lingual transfer paradigm. We then proposed a resource-lean approach for attenuating those losses. This approach, few-shot target-language fine-tuning, accounts for the (2) *ecological impact* of language technologies. Big transformer-based language representation models require a large amount of training resources, which results in a large carbon footprint of these representations. By proposing resource-lean methods, we can (partially) account for this. For the same reason, we have proposed an approach for the injection of external knowledge, which does not require pretraining from scratch due to relying on the efficiency of adapter layers. In this Chapter, we aim to mitigate potential harm arising from CA technologies due to *unfair stereotypical bias* in language representations. Unfair stereotypical bias may arise due to co-occurrence biases in the pretraining data coupled with the distributional nature of language representations (see Section 2.2.4). This has been pointed out as an essential challenge for CA (Spliethöver and Wachsmuth, 2020). To account for this, we (1) first present XWEAT, a resource based on which we conduct a multi-dimensional analysis of biases in language representations. We then (2) present a general framework that synthesizes previous work on bias evaluation and mitigation in static word embeddings. Within this framework, we propose a new bias measure (Bias Analogy Test (BAT)) and three bias mitigation methods (General Bias Direction Debiasing (GBDD), Bias Alignment Method (BAM), and Explicit Neural Debiasing (DebiasNet)).

8.1 Multidimensional Bias Analysis in Word Embeddings

*As discussed, word embeddings have recently been shown to reflect many of the pronounced societal biases (e.g., gender bias or racial bias), which poses a challenge for CA. Existing studies are, however, limited in scope and do not investigate the consistency of biases across relevant dimensions like embedding models, types of texts, and different languages. In this Section, we present a systematic study of biases encoded in distributional word vector spaces: we analyze how consistent the bias effects are across languages, corpora, and embedding models. Furthermore, we analyze the cross-lingual biases encoded in bilingual embedding spaces, indicative of the effects of bias transfer encompassed in cross-lingual transfer of NLP models. Our study yields some unexpected findings, e.g., that biases can be emphasized or downplayed by different embedding models or that user-generated content may be less biased than encyclopedic text. We hope our work catalyzes bias research in NLP and informs the development of bias reduction techniques.

8.1.1 Introduction

Recent work demonstrated that word embeddings induced from large text collections encode many human biases (e.g., Bolukbasi et al., 2016; Caliskan et al., 2017). As we briefly outlined in Section 2.2.4, this finding is not particularly surprising given that (1) we are likely to project our biases in the text that we produce and (2) these biases in text are bound to be encoded in word vectors due to the distributional nature (Harris, 1954) of the word embedding models (Mikolov et al., 2013b; Pennington et al., 2014; Bojanowski et al., 2017). For illustration, consider the famous analogy-based gender bias example from Bolukbasi et al. (2016): *"man is to computer programmer as woman is to homemaker"*. This bias will be reflected in the text (i.e., the word *man* will co-occur more often with words like *programmer* or *engineer*, whereas *woman* will more often appear next to *homemaker* or *nurse*), and will, in turn, be captured by word embeddings built from such biased texts. While biases encoded in word embeddings can be a useful data source for diachronic analyses of societal biases (e.g., Garg et al., 2018), they may cause ethical problems for many downstream applications and NLP models. For CA, Spliethöver and Wachsmuth (2020) showed popular argumentative corpora to contain such stereotypical biases, and Dev et al. (2020) demonstrated that argumentative downstream tasks, as in natural language inference, biases in language representations may result in stereotypical inferences.

In order to measure the extent to which various societal biases are captured in static language representations, Caliskan et al. (2017) proposed the *Word Embedding Association Test (WEAT)*. WEAT measures semantic similarity, computed through word embeddings, between two sets of *target* words (e.g., insects vs. flowers) and two sets of *attribute* words (e.g., pleasant vs. unpleasant words). While they test a number of biases, the analysis is limited in scope to English as the only language, GloVe (Pennington

*This Section is adapted from: **Anne Lauscher** and Goran Glavaš. Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 85–91, Minneapolis, Minnesota, June 2019, Association for Computational Linguistics.

et al., 2014) as the embedding model, and Common Crawl as the type of text. Following the same methodology, McCurdy and Serbetci (2017) extend the analysis to three more languages (German, Dutch, Spanish) but test only for gender bias.

In this Section, we present the most comprehensive study of biases captured by distributional word vectors to date. We create Cross-lingual WEAT (XWEAT), a collection of multilingual and cross-lingual versions of the WEAT data set, by translating WEAT to six other languages and offer a comparative analysis of biases over seven diverse languages. We thereby, as in the previous Chapter, account for the challenge of multilinguality in CA (**C4**). Furthermore, we measure the consistency of WEAT biases across different embedding models and types of corpora. What is more, given the recent surge of models for inducing cross-lingual embedding spaces (Mikolov et al., 2013b; Hermann and Blunsom, 2014; Smith et al., 2017; Lample et al., 2018; Artetxe et al., 2018; Hoshen and Wolf, 2018, *inter alia*) and their ubiquitous application in cross-lingual transfer of NLP models for downstream tasks, we investigate cross-lingual biases encoded in cross-lingual embedding spaces and compare them to bias effects present of corresponding monolingual embeddings. Our analysis yields some interesting findings: the amount of the biases depends on the embedding model, and, quite surprisingly, stereotypical bias seems to be less pronounced in embeddings trained on social media texts. Furthermore, we find that the effects (i.e., amount) of bias in cross-lingual embedding spaces can roughly be predicted from the bias effects of the corresponding monolingual embedding spaces.

8.1.2 Data for Measuring Biases

We first introduce the WEAT data set (Caliskan et al., 2017) and then describe XWEAT, our multilingual and cross-lingual extension of WEAT designed for comparative bias analyses across languages and in cross-lingual embedding spaces.

WEAT

The Word Embedding Association Test (WEAT) (Caliskan et al., 2017) is an adaptation of the Implicit Association Test (IAT) (Nosek et al., 2002). Whereas IAT measures biases based on response times of human subjects to provided stimuli, WEAT quantifies the biases using semantic similarities between word embeddings of the same stimuli. For each bias test, WEAT specifies four stimuli sets: two sets of *target* words and two sets of *attribute* words. The sets of target words represent stimuli *between* which we want to measure the bias (e.g., for gender biases, one target set could contain male names and the other female names). The *attribute* words, on the other hand, represent stimuli *towards* which the bias should be measured (e.g., one list could contain pleasant stimuli like *health* and *love* and the other negative *war* and *death*). The WEAT data set defines ten bias tests, each containing two target and two attribute sets.[1] Table 8.1 enumerates the WEAT tests and provides examples of the respective target and attribute words.

[1]Some of the target and attribute sets are shared across multiple tests.

Test	Target Set #1	Target Set #2	Attribute Set #1	Attribute Set #2
T1	Flowers (e.g., *aster*, *tulip*)	Insects (e.g., *ant*, *flea*)	Pleasant (e.g., *health*)	Unpleasant (e.g., *abuse*)
T2	Instruments (e.g., *cello*, *guitar*)	Weapons (e.g., *gun*, *sword*)	Pleasant	Unpleasant
T3	Euro-American names	Afro-American names	Pleasant (e.g., *caress*)	Unpleasant (e.g., *abuse*)
T4	Euro-American names	Afro-American names	Pleasant	Unpleasant
T5	Euro-American names	Afro-American names	Pleasant (e.g., *joy*)	Unpleasant (e.g., *agony*)
T6	Male names (e.g., *John*)	Female names (e.g., *Lisa*)	Career (e.g. *management*)	Family (e.g., *children*)
T7	Math (e.g., *algebra*, *geometry*)	Arts (e.g., *poetry*, *dance*)	Male (e.g., *brother*, *son*)	Female (e.g., *woman*)
T8	Science (e.g., *experiment*)	Arts	Male	Female
T9	Physical condition (e.g., *virus*)	Mental condition (e.g., *sad*)	Long-term (e.g., *always*)	Short-term (e.g., *occasional*)
T10	Older names (e.g., *Gertrude*)	Younger names	Pleasant	Unpleasant

Table 8.1: WEAT bias tests.

Multilingual and Cross-Lingual WEAT

We port the WEAT test term sets to the multilingual and cross-lingual settings by translating the test vocabularies consisting of attribute and target terms from English to six other languages: German (DE), Spanish (ES), Italian (IT), Russian (RU), Croatian (HR), and Turkish (TR). To this end, we first automatically translate the vocabularies and then let native speakers of the respective languages (also fluent in English) fix the incorrect automatic translations (or introduce better fitting ones). Our aim was to translate the WEAT vocabularies to languages from diverse language families[2] for which we also had access to native speakers. Whenever the translation of an English term indicated the gender in a target language (e.g., *Freund* vs. *Freundin* in DE), we asked the respective translator to provide both male and female forms, and we included both forms in the final test vocabularies. This helps to avoid artificially amplifying the gender bias stemming from the grammatically masculine or feminine word forms.

The monolingual tests are created by simply using the corresponding translations of target and attribute sets in those languages. For every two languages, L1 and L2 (e.g., DE and IT), we additionally create two cross-lingual bias tests: we pair (1) target translations in L1 with L2 translations of attributes (e.g., for T2 we combine DE target sets {*Klavier*, *Cello*, *Gitarre*, ...} and {*Gewehr*, *Schwert*, *Schleuder*, ...} with IT attribute sets {*salute*, *amore*, *pace*, ...} and {*abuso*, *omicidio*, *tragedia*, ...}), and vice versa, (2) target translations in L2 with attribute translations in L1 (e.g., T2 IT target sets with DE attribute sets). We did not translate or modify proper names from WEAT sets 3–6. In our multilingual and cross-lingual experiments we, however, discard the (translations of) WEAT tests for which we cannot find more than 20% of words from some target or attribute set in the embedding vocabulary of the respective language. This strategy eliminates tests 3–5 and 10 which include proper American names, majority of which can not be found in distributional vocabularies of other languages. The exception to this is test 6, containing frequent English first names (e.g., *Paul*, *Lisa*), which we do find in distributional vocabularies of other languages as well. In summary, for languages other than EN and for cross-lingual settings, we execute six bias tests (T1, T2, T6–T9).

[2] EN and DE from the Germanic branch of Indo-European languages, IT and ES from the Romance branch, RU and HR from the Slavic branch, and finally TR as a non-Indo-European language.

8.1.3 Methodology

We adopt the general bias-testing framework from Caliskan et al. (2017), but we span our study over multiple dimensions: (1) corpora – we analyze the consistency of biases across distributional vectors induced from different types of text; (2) embedding models – we compare biases across distributional vectors induced by different embedding models (on the same corpora); and (3) languages – we measure biases for word embeddings of different languages, trained from comparable corpora. Furthermore, unlike Caliskan et al. (2017), we test whether biases depend on the selection of the similarity metric. Finally, given the ubiquitous adoption of cross-lingual embeddings (Ruder et al., 2019; Glavaš et al., 2019), we investigate biases in a variety of bilingual embedding spaces.

Bias-Testing Framework. We first describe the WEAT framework (Caliskan et al., 2017). Let X and Y be two term sets of *targets*, and A and B two term sets of *attributes* (see Subsection 8.1.2). The tested statistic is the difference between X and Y in average similarity of their terms with the terms from A and B:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)\,, \tag{8.1}$$

with the association difference for a term t computed as:

$$s(t, A, B) = \frac{1}{|A|} \sum_{a \in A} f(\mathbf{t}, \mathbf{a}) - \frac{1}{|B|} \sum_{b \in B} f(\mathbf{t}, \mathbf{b})\,, \tag{8.2}$$

where $\mathbf{t}$ is the distributional vector of term t and f is a similarity or distance metric, fixed to cosine similarity in the original work (Caliskan et al., 2017). The significance of the test statistic is validated by comparing the score $s(X, Y, A, B)$ with the scores $s(X_i, Y_i, A, B)$ obtained for different equally sized partitions $\{X_i, Y_i\}_i$ of the set $X \cup Y$. The p-value of this permutation test is then measured as the probability of $s(X_i, Y_i, A, B) > s(X, Y, A, B)$ computed over all possible permutations $\{X_i, Y_i\}_i$.[3] Finally, the effect size, i.e., the "amount of bias", is computed as the normalized measure of separation between the association distributions:

$$\frac{\mu\left(\{s(x, A, B)\}_{x \in X}\right) - \mu\left(\{s(y, A, B)\}_{y \in Y}\right)}{\sigma\left(\{s(w, A, B)\}_{w \in X \cup Y}\right)}\,, \tag{8.3}$$

where μ denotes the mean and σ the standard deviation.

Dimensions of Bias Analysis. We analyze the bias effects across multiple dimensions. First, we analyze the effect that different embedding models have: we compare biases in distributional spaces induced from the English Wikipedia, using the CBOW (Mikolov et al., 2013c), GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017), and Dict2Vec algorithms (Tissier et al., 2017). Secondly, we investigate the effect of

[3] If f is a distance metric, we measure the probability of $s(X_i, Y_i, A, B) < s(X, Y, A, B)$.

Metric	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Cosine	1.7	1.6	-0.1*	-0.2*	-0.2*	1.8	1.3	1.3	1.7	-0.6*
Euclidean	1.7	1.6	-0.1*	-0.2*	-0.1*	1.8	1.3	1.3	1.7	-0.7*

Table 8.2: WEAT bias effects in EN Wikipedia FASTTEXT embeddings for cosine similarity and Euclidean distance. Asterisks indicate bias effects that are insignificant at $\alpha < 0.05$.

employing different corpora for inducing the language representations: we compare biases between embeddings trained on the Common Crawl, Wikipedia, and a corpus of tweets. Finally, and (arguably) most interestingly, we test the consistency of biases across seven languages (see Subsection 8.1.2). To this end, we test for biases in seven monolingual FASTTEXT spaces trained on Wikipedia dumps of the respective languages.

Biases in Cross-lingual Embeddings. Cross-lingual word embeddings (CLWEs) are widely used in multilingual NLP and CA and for cross-lingual transfer of NLP and CA models. Despite the ubiquitous usage of CLWEs, the biases they potentially encode have not been analyzed so far. We analyze projection-based CLWEs (Glavaš et al., 2019), induced through post hoc linear projections between monolingual embedding spaces (Mikolov et al., 2013b; Artetxe et al., 2016; Smith et al., 2017). The projection is commonly learned through supervision with a few thousand word translation pairs. Most recently, however, a number of models have been proposed that learn the projection without any bilingual signal (Artetxe et al., 2018; Lample et al., 2018; Hoshen and Wolf, 2018; Alvarez-Melis and Jaakkola, 2018, *inter alia*). Let $\mathbf{X}$ and $\mathbf{Y}$ be, respectively, the distributional spaces of the source (S) and target (T) language and let $D = \{w_S^{(i)}, w_T^{(i)}\}_i$ be the word translation dictionary. Let $(\mathbf{X}_S, \mathbf{X}_T)$ be the aligned subsets of monolingual embeddings, corresponding to word-aligned pairs from D. We then compute the orthogonal matrix $\mathbf{W}$ that minimizes the Euclidean distance between $\mathbf{X_S W}$ and $\mathbf{X_T}$ (Smith et al., 2017): $\mathbf{W} = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \text{SVD}(\mathbf{X}_T\mathbf{X}_S{}^\top)$. We create comparable bilingual dictionaries D by translating the 5K most frequent EN words to the other six languages and induce a bilingual space for all 21 language pairs.

8.1.4 Findings

Here, we report and discuss the results of our multi-dimensional analysis. Table 8.2 shows the effect sizes for WEAT T1–T10 based on Euclidean or cosine similarity between word vector representations trained on the EN Wikipedia using FASTTEXT. We observe the highest bias effects for T6 (Male/Female – Career/Family), T9 (Physical/Mental diseases – Long-term/Short-term), and T1 (Insects/Flowers – Positive/Negative). Importantly, the results show that biases do not depend on the similarity metric. We observe nearly identical effects for cosine similarity and Euclidean distance for all WEAT tests. In the following experiments, we thus analyze biases only for cosine similarity.

WEAT	**CBOW**	**GloVe**	**FastText**	**Dict2Vec**
T1	1.20	1.41	1.67	1.35
T2	1.38	1.45	1.55	1.66
T3	-0.28*	1.16	-0.09*	–
T4	-0.35*	1.36	-0.17*	–
T5	-0.36*	1.40	-0.18*	–
T6	1.78	1.75	1.83	–
T7	1.28	1.16	1.30	1.48
T8	0.39*	1.28	1.30	1.30
T9	1.55	1.35	1.72	1.69
T10	0.09*	1.17	-0.61*	–

Table 8.3: WEAT bias effects for language representation spaces induced (on EN Wikipedia) with different embedding models: CBOW, GloVe, fastText, and Dict2Vec methods. Asterisks indicate bias effects that are insignificant at $\alpha < 0.05$.

Word Embedding Models. Table 8.3 compares biases in embedding spaces induced with different models: CBOW, GloVe, fastText, and Dict2Vec. While the first three embedding methods are trained on Wikipedia only, Dict2Vec employs definitions from dictionaries (e.g., Oxford dictionary) as additional resources for identifying strongly related terms.[4] We only report WEAT test results T1, T2, and T7–T9 for Dict2Vec, as the Dict2Vec's vocabulary does not cover most of the proper names from the remaining tests. Somewhat surprisingly, the bias effects seem to vary greatly across embedding models. While GloVe embeddings are biased according to all tests,[5] fastText and especially CBOW exhibit significant biases only for a subset of the tests. We hypothesize that the bias effect sizes reflected in the distributional space depend on the preprocessing steps of the embedding model. E.g., fastText relies on embedding subword information to avoid issues with representations of out-of-vocabulary and underrepresented terms: additional reliance on morpho-syntactic signal may make fastText more resilient to biases stemming from the distributional signal (i.e., word co-occurrences). The fact that the embedding space induced with Dict2Vec exhibits larger bias effects may seem counterintuitive at first since the dictionaries used for vector training should be more objective and therefore less biased than encyclopedic text. We believe, however, that the additional dictionary-based training objective only propagates the distributional biases across definitionally related words. Generally, we find these results to be important as they indicate that embedding models may accentuate or diminish biases expressed in text.

Corpora. In Table 8.4 we compare the biases of embeddings trained with the same model (GloVe) but on different corpora: Common Crawl (i.e., noisy web content), Wikipedia (i.e., encyclopedic text) and a corpus of tweets (i.e., user-generated content). Expectedly, the biases are slightly more pronounced for embeddings trained on Common Crawl than for those obtained on Wikipedia. Countering our intuition, the corpus of

[4] Terms A and B are strongly related if B appears in the definition of A and vice versa (Tissier et al., 2017).

[5] This is consistent with the original results obtained by Caliskan et al. (2017).

Corpus	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
WIKI	1.4	1.5	1.2	1.4	1.4	1.8	1.2	1.3	1.3	1.2
CC	1.5	1.6	1.5	1.6	1.4	1.9	1.1	1.3	1.4	1.3
TWEETS	1.2	1.0	1.1	1.2	1.2	1.2	-0.2*	0.6*	0.7*	0.8*

Table 8.4: WEAT bias effect sizes for GLOVE embedding spaces trained on different corpora: Wikipedia (WIKI), Common Crawl (CC), and corpus of tweets (TWEETS). Asterisks indicate bias effects that are insignificant at $\alpha < 0.05$.

XW	**EN**	**DE**	**ES**	**IT**	**HR**	**RU**	**TR**
T1	1.67	1.36	1.47	1.28	1.45	1.28	1.21
T2	1.55	1.25	1.47	1.36	1.10	1.46	0.83
T6	1.83	1.59	1.67	1.72	1.83	1.87	1.85
T7	1.30	0.46*	1.47	1.00	0.72*	0.59*	-0.88
T8	1.30	0.05*	1.16	0.10*	0.13*	0.37*	1.72
T9	1.72	0.82*	1.71	1.57	-0.40*	1.73	1.09*
Avg_{all}	1.56	0.92	1.49	1.17	0.81	1.22	0.88
Avg_{sig}	1.68	1.4	1.54	1.45	1.46	1.54	1.30

Table 8.5: XWEAT effect sizes across seven languages (FASTTEXT embedding spaces trained on Wikipedias). Avg_{all}: average effect size over all tests; Avg_{sig}: average effect size over the subset of tests yielding significant bias effect sizes for all languages. Asterisks indicate bias effects that are insignificant at $\alpha < 0.05$.

XW	**EN**	**DE**	**ES**	**IT**	**HR**	**RU**	**TR**
EN	–	1.09*	1.58	1.49	0.72*	1.17*	1.20*
DE	1.53	–	1.50	1.45	0.55*	1.35	1.07*
ES	1.52	0.79*	–	1.38*	0.60*	1.37*	1.09*
IT	1.33*	0.69*	1.27	–	0.53*	0.82*	0.80*
HR	1.47	1.30*	1.29	1.18*	–	1.14*	1.11*
RU	1.47	0.72*	1.35	1.35	0.77*	–	0.80*
TR	1.41	0.90*	1.37*	1.45	0.29*	0.64*	–

Table 8.6: XWEAT bias effects (aggregated over all six tests) for cross-lingual word embedding spaces. Rows: *targets* language; columns: *attributes* language. Asterisks indicate the inclusion of bias effects sizes in the aggregation that were insignificant at $\alpha < 0.05$.

tweets seems to be consistently less biased (across all tests) than Wikipedia. In fact, the biases covered by tests T7–T10 are not even significantly present in the vectors trained on tweets. This finding is indeed surprising and warrants further investigation.

Multilingual Comparison. Table 8.5 compares the bias effects across the seven different languages. Whereas many of the biases are significant in all languages, DE, HR,

and TR consistently display smaller effect sizes. Intuitively, the amount of bias should be proportional to the size of the corpus.[6] Wikipedias in TR and HR are the two smallest ones – thus, they are expected to contain the least biased statements. DE Wikipedia, on the other hand, is the second largest and low bias effects here suggest that German texts are indeed less biased than texts in other languages. Additionally, for (X)WEAT T2, which defines a universally accepted bias (Instruments vs. Weapons), TR and HR exhibit the smallest effect sizes, while the highest bias is observed for EN and IT. We measure the highest gender bias, according to (X)WEAT T6, for TR and RU, and the lowest for DE.

Biases in Cross-Lingual Embeddings. We report bias effects for all 21 bilingual embedding spaces in Table 8.6. For brevity, here we report the bias effects averaged over all six XWEAT tests (we provide results detailing bias effects for each of the tests separately in Section D.1 of the supplementary material). Generally, the bias effects of bilingual spaces are in between the bias effects of the two corresponding monolingual spaces (cf. Table 8.5): this means that we can roughly predict the amount of bias in a cross-lingual embedding space from the same bias effects of corresponding monolingual spaces. For example, effects in cross-lingual spaces increase over monolingual effects for low-bias languages (HR and TR), and decrease for high-bias languages (EN and ES).

8.1.5 Conclusion

In this Section, we have presented the largest study on unfair stereotypical biases encoded in static language representations to date. To this end, we have extended previous analyses based on the WEAT test (Caliskan et al., 2017; McCurdy and Serbetci, 2017) in multiple dimensions: across seven languages, four embedding models, and three different types of text. We find that different language representation models may produce embeddings with very different biases, which stresses the importance of embedding model selection when fair language representations are to be created. Surprisingly, we find that user-generated texts, e.g., tweets, may be less biased than redacted content. Furthermore, we have investigated the bias effects in cross-lingual embedding spaces and have shown that they may be predicted from the biases of corresponding monolingual embeddings. We make the XWEAT data set and the testing code publicly available.[7]

WEAT, which we extended in this Section to XWEAT, is able to measure *explicit* biases (Gonen and Goldberg, 2019). In the next Section, we explain the difference between *explicit* and *implicit* biases and present a framework that provides a broader perspective on biases by including bias measures for both bias types plus testing for the semantic quality of embedding spaces. We further introduce three new debiasing methods.

[6] The larger the corpus, the larger is the overall number of contexts in which some bias may be expressed.
[7] At `https://github.com/umanlp/XWEAT`.

8.2 Implicit and Explicit Debiasing of Word Embeddings

[*]In response to the issue of unfair bias in language representations, which we also dealt with in the previous Section, a number of methods for attenuating stereotypical biases have been proposed. However, existing models and studies (1) operate on under-specified and mutually differing bias definitions, (2) are tailored for a particular bias (e.g., gender bias), and (3) have been evaluated inconsistently and non-rigorously. In this Section, we introduce a general framework for debiasing word embeddings to further address the challenge of bias in language representations for CA (**C5**). We operationalize the definition of a bias by discerning two types of bias specification: explicit and implicit. We then propose three debiasing models that operate on explicit or implicit bias specifications and that can be composed towards more robust debiasing. Next, we devise a full-fledged evaluation framework in which we couple existing bias metrics with newly proposed ones. Experimental findings across three embedding methods suggest that the proposed debiasing models are robust and widely applicable: they often completely remove the bias both implicitly and explicitly without degradation of semantic information encoded in any of the input distributional spaces. Moreover, we successfully transfer debiasing models, by means of cross-lingual embedding spaces, and remove or attenuate biases in distributional word vector spaces of languages that lack readily available bias specifications by which we implicitly also address the challenge of multilinguality in CA (**C4**). Finally, in addition to the intrinsic evaluation provided by our evaluation framework, we extrinsically test the effects of debiasing in an argumentative downstream application: with the task of NLI, we show that a model employing one of our debiased spaces produces the smallest amount of stereotypically biased inferences. However, the results also indicate that debiasing effects may be overwritten by large amounts of training data.

8.2.1 Introduction

Distributional word vector spaces have been recently shown to encode prominent human biases related to, e.g., gender or race (Bolukbasi et al., 2016; Caliskan et al., 2017; Manzini et al., 2019). Such biases are observed across languages and embedding methods (see Section 8.1), both in static and contextualized language representations (Zhao et al., 2019). While this issue requires remedy, the finding itself is hardly surprising: we project our biases, in terms of biased word co-occurrences, into the texts we produce. Consequently, this is propagated to the embedding models, both static (Mikolov et al., 2013c; Pennington et al., 2014; Bojanowski et al., 2017) and contextualized (Peters et al., 2018) alike, by virtue of the distributional hypothesis (Harris, 1954).[8] While biases may be useful for diachronic or sociological analyses (Garg et al., 2018), they (1) raise ethical issues, since biases are amplified by machine learning models using embeddings as input (Zhao et al., 2017), and

[*]This Section is adapted from: **Anne Lauscher**, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 8131–8138, New York, New York, U.S., January 2020, AAAI Press.

[8]Borrowing the famous example (Bolukbasi et al., 2016), *man* will be found more often in the same context with *programmer*, and *woman* with *homemaker* in any sufficiently large corpus.

(2) impede tasks like coreference resolution (Zhao et al., 2018a; Rudinger et al., 2018) and abusive language detection (Park et al., 2018). As we have outlined in previous Sections (see Sections 2.2.4, 3.5, 8.1), bias in language representations is a specifically crucial issue for the area of CA (Spliethöver and Wachsmuth, 2020).

A number of methods for attenuating and eliminating human-like biases in static word vector spaces have been proposed recently (Bolukbasi et al., 2016; Zhao et al., 2018a,b; Dev and Phillips, 2019). While they address the same types of bias – primarily the gender bias – they start from different bias "specifications" and either lack proper empirical evaluation (Bolukbasi et al., 2016) or employ different evaluation procedures, both hindering a direct comparison of the "debiasing abilities" of the methods (Zhao et al., 2019; Dev and Phillips, 2019; Manzini et al., 2019). What is more, the most prominent debiasing models (Bolukbasi et al., 2016; Zhao et al., 2018b) have been criticized recently for merely masking the bias instead of removing it (Gonen and Goldberg, 2019). To resolve these inconsistencies in the current debiasing research and evaluation, in this Section, we propose a general debiasing framework DebIE (**DEB**iasing embeddings **I**mplicitly and **E**xplicitly), which operationalizes bias specifications, groups the debiasing models according to the bias specification type they operate on, and evaluates the abilities of the models to remove unfair stereotypical biases both explicitly and implicitly (Gonen and Goldberg, 2019).

We first define two types of bias specifications – *implicit* and *explicit* – and propose a method of augmenting bias specifications with the help of embeddings specialized for semantic similarity (Mrkšić et al., 2017; Ponti et al., 2018). We then introduce the main contributions of this Section: first, we present three novel debiasing models. (1) We adjust the linear projection method of Dev and Phillips (2019), an extension of the debiasing model of Bolukbasi et al. (2016), to operate on the augmented bias specifications. (2) We then propose an alternative model that projects the embedding space to itself using the term sets from implicit bias specifications as the projection signal. (3) Next, we propose an effective neural debiasing model, which is, to the best of our knowledge, the first debiasing model that operates on an explicit bias specification. All three models perform *post hoc* debiasing: they can be applied to any pretrained word vector space.[9] As another contribution, we combine existing bias metrics with newly proposed ones and assemble an evaluation suite that tests word vectors for explicit biases, implicit biases, and (preservation of) semantic quality. Furthermore, by coupling the proposed debiasing models with the cross-lingual embedding spaces (Ruder et al., 2019; Glavaš et al., 2019), we facilitate cross-lingual debiasing transfer: we successfully debias embedding spaces in target languages without bias specifications in those languages. Finally, to complement the intrinsic analysis provided by our evaluation framework, we seek to understand the effect of debiasing in a downstream evaluation focusing on NLI. To this end, we follow Dev et al. (2020) and create a synthetic data set that tests the models' for gender-biased inferences. The least amount of biased inferences is produced by a model employing one of the debiased spaces, but in many cases, the debiasing effects seem to be overwritten.[10]

[9] In contrast, debiasing models like GN-GloVe (Zhao et al., 2018b) integrate debiasing constraints into objectives of embedding models like GloVe (Pennington et al., 2014). The downside of these approaches is that they cannot be directly ported to other embedding models.

[10] The code is available at `https://github.com/umanlp/DEBIE`.

Initial T_1	*science*, *technology*, *physics*, *chemistry*, *Einstein*, *NASA*, *experiment*, *astronomy*
Initial T_2	*poetry*, *art*, *Shakespeare*, *dance*, *literature*, *novel*, *symphony*, *drama*
Initial A_1	*brother*, *father*, *uncle*, *grandfather*, *son*, *he*, *his*, *him*
Initial A_2	*sister*, *mother*, *aunt*, *grandmother*, *daughter*, *she*, *hers*, *her*
Augmentation T_1	*automation*, *radiochemistry*, *test*, *biophysics*, *learning*, *electrodynamics*, *biochemistry*, *astrophysics*, *astrometry*
Augmentation T_1	*orchestra*, *artistry*, *dramaturgy*, *poesy*, *philharmonic*, *craft*, *untried*, *hop*, *poem*, *dancing*, *dissertation*, *treatise*
Augmentation A_1	*beget*, *buddy*, *forefather*, *man*, *nephew*, *own*, *himself*, *theirs*, *boy*, *crony*, *cousin*, *grandpa*, *granddad*
Augmentation A_2	*niece*, *girl*, *parent*, *grandma*, *granny*, *woman*, *theirs*, *sire*, *auntie*, *sibling*, *herself*, *jealously*, *stepmother*, *wife*

Table 8.7: Initial and augmented gender bias specifications. Test T8 from WEAT.

8.2.2 General Debiasing Framework

In what follows, we first formalize two types of bias specifications – implicit and explicit. We then introduce new debiasing models: two operate on the implicit bias specification and the third on the explicit bias specification. Finally, we show how to debias word embeddings in a variety of target languages via cross-lingual embeddings.

Bias Specifications

An *implicit bias specification* $B_I = (T_1, T_2)$ consists of two sets of *target* terms between which a bias is expected to exist in the embedding space. For example, two sets of science and art terms, T_1 = {*physics*, *chemistry*, *experiment*} and $T_2 = \{$*poetry*, *dance*, *drama*$\}$ constitute an implicit specification of the gender bias. Strictly speaking, B_I does not specify a bias directly – it merely specifies two categories of concepts for which we *implicitly* assume that there exists some set of reference terms A (e.g., male terms *man*, *father* and/or female terms like *woman*, *girl*) with respect to which T_1 and T_2 exhibit differences. Most existing debiasing models (Bolukbasi et al., 2016; Zhao et al., 2018b; Dev and Phillips, 2019; Manzini et al., 2019) operate on $B_I = (T_1, T_2)$, i.e., not requiring terms A.

An *explicit bias specification* B_E defines, in addition to T_1 and T_2, one or more *attribute* sets. We consider an explicit bias specification with a single attribute set, $B_E = (T_1, T_2, A)$ (as employed by our DebiasNet model),[11] and also with two (opposing) attribute sets, $B_E = (T_1, T_2, A_1, A_2)$, as used in WEAT tests (Caliskan et al., 2017).

Augmentation of Bias Specifications. The initial bias specification (B_I or B_E) commonly contains only a handful of words in each target and attribute set. These are commonly the most representative words of a category (e.g., *man*, *boy*, *father* to represent the category *male*). However, in order to provide a finer-grained bias specification, we

[11] The attribute set A can be any set of attributes towards which the bias is to be removed. In our experiments, we joined the WEAT test specification attribute sets A_1 and A_2.

propose to augment each term set with synonyms and semantically similar words of the initial terms. We therefore extract nearest neighbours of initial terms from an embedding space specialized to accentuate true semantic similarity and attenuate other types of semantic association (Faruqui et al., 2015; Vulić et al., 2018; Glavaš and Vulić, 2018, *inter alia*). For the augmentation process, we rely on the recent state-of-the-art similarity specialization method of Ponti et al. (2018): for more details, see the original work.

Given a bias specification B_I or B_E and a similarity-specialized word vector space $\mathbf{X_{sim}}$, we augment each of the term sets in the specification by retrieving the top k most (cosine-)similar terms from $\mathbf{X_{sim}}$ for each of the initial terms.[12] Extending bias specification sets using a similarity-specialized word vector space – as opposed to a regular distributional space – reduces the noisy augmentation stemming from the semantic relatedness instead of true semantic similarity, as discussed in Section 4.1.[13] Table 8.7 illustrates the initial bias specification and the corresponding augmentation (showing $k = 2$ nearest neighbors, without the initial terms) for one explicitly defined gender bias.

Debiasing Models

We present three novel debiasing models, two of which operate on an implicit bias specification $B_I = (T_1, T_2)$ and one on the explicit bias specification $B_E = (T_1, T_2, A)$.

General Bias Direction Debiasing (GBDD) focuses on B_I as a generalization of the linear projection model proposed by Dev and Phillips (2019), itself, in turn, an extension of the hard-debiasing model of Bolukbasi et al. (2016).

The model of Dev and Phillips (2019) requires a stricter bias specification than our B_I: it requires T_1 and T_2 to be ordered lists of equal length L, so that the so-called equivalence pairs $\{(t_1^{(l)}, t_2^{(l)})\}_{l=1}^{L}$ can be created. For instance, T_1 ={*man*, *father*, *boy*} and T_2 ={*woman*, *mother*, *girl*} give rise to the following equivalence pairs: (*man*, *woman*), (*father*, *mother*), and (*boy*, *girl*). For each equivalence pair $(t_1^{(l)}, t_2^{(l)})$ they compute the *bias direction vector* $\mathbf{b_l}$ by subtracting the vector of term t_2^l from the vector of term $t_1^{(l)}$. We find this bias specification overly restrictive: it requires an additional effort to create true equivalence pairs from T_1 and T_2 *and* it produces only L partial bias direction vectors. In contrast, we propose to create one bias direction vector $\mathbf{b_{ij}}$ for each pair $(t_1^{(i)}, t_2^{(j)})$, $t_1^{(i)} \in T_1, t_2^{(j)} \in T_2$. If T_1 and T_2 truly specify categories that are opposite in some regard (e.g., gender-wise), then any pair $(t_1^{(i)}, t_2^{(j)})$ should induce a meaningful partial bias direction vector. This way we also obtain a much larger number of partial bias direction vectors (e.g., L^2 if T_1 and T_2 are of the same length L): this should result in a more reliable *general bias direction vector*, computed as follows. We stack all of the obtained bias direction vectors $\mathbf{b_{ij}}$ corresponding to pairs $(t_1^{(i)}, t_2^{(j)})$, $t_1^{(i)} \in T_1$,

[12] We discard nearest neighbors that are initially present in other sets of the same bias specification: for instance, if we retrieve an augmentation candidate *woman* for an initial T_1 term *man*, *woman* will not be added to T_1 if it already exists in one of the target term sets T_1, T_2, or in any attribute set A.

[13] We also considered using clean lexical knowledge from WordNet (Miller, 1995) directly, but this resulted in much lower recall as well as less accurate augmentation candidates.

$t_2^{(j)} \in T_2$ to form a bias direction matrix $\mathbf{B}$. We then obtain the *global* bias direction vector $\mathbf{b}$ as the top singular vector of $\mathbf{B}$, i.e., as the first row of matrix $\mathbf{V}$, where $\mathbf{U\Sigma V}^\top$ is the singular value decomposition of $\mathbf{B}$. Let $\mathbf{x}$ be the ℓ_2-normalized d-dimensional vector from a biased input vector space. Its debiased version is then computed as:

$$\text{GBDD}(\mathbf{x}) = \mathbf{x} - (\mathbf{x} \cdot \mathbf{b})\mathbf{b}\,. \tag{8.4}$$

In other words, the closer the vector $\mathbf{x}$ is to the global bias direction $\mathbf{b}$, the more it is bias-corrected (i.e., the larger portion of $\mathbf{b}$ is subtracted from $\mathbf{x}$). Vectors orthogonal to the bias direction $\mathbf{b}$ remain unchanged (zero dot-product with the bias vector $\mathbf{b}$).

Bias Alignment Method (BAM). An alternative to computing a bias direction vector $\mathbf{b}$ is to use target-term pairs $(t_1^{(i)}, t_2^{(j)})$, $t_1^{(i)} \in T_1, t_2^{(j)} \in T_2$ to learn a projection of the biased embedding space $\mathbf{X} \in \mathbb{R}^d$ to itself that (approximately) aligns T_1 and T_2. The idea behind this model stems from the research on projection-based CLWE spaces (see also Section 8.1), where an orthogonal mapping between monolingual embedding spaces is learned from a set of word translations (Smith et al., 2017; Glavaš et al., 2019).[14]

Here, we use bias term pairs $(t_1^{(i)}, t_2^{(j)})$ to learn the debiasing projection of $\mathbf{X}$ with respect to itself. Let $\mathbf{X}_{T_1}$ and $\mathbf{X}_{T_2}$ be the matrices obtained by stacking the (biased) vectors of left and right words of pairs $(t_1^{(i)}, t_2^{(j)})$, respectively. We then learn the orthogonal mapping matrix $\mathbf{W_X} = \mathbf{UV}^\top$, where $\mathbf{U\Sigma V}^\top$ is the singular value decomposition of $\mathbf{X}_{T_2}\mathbf{X}_{T_1}^\top$. Since $\mathbf{W_X}$ is orthogonal, the projection $\mathbf{X}' = \mathbf{XW_X}$ is isomorphic to the original space $\mathbf{X}$, and thus equally biased. However, the transformation (specified by $\mathbf{W_X}$) defines the angle and direction of debiasing. We obtain the debiased space by averaging the original space $\mathbf{X}$ and the projected space $\mathbf{X}'$:

$$\text{BAM}(\mathbf{X}) = \frac{1}{2}\left(\mathbf{X} + \mathbf{XW}_X\right). \tag{8.5}$$

Explicit Neural Debiasing (DebiasNet). The final model, dubbed DebiasNet (in Tables referred to with its function abbreviation DBN), is the first neural model that operates on an explicit bias specification B_E. It is inspired by previous work on semantic specialization of static language represention (e.g., Vulić et al., 2018; Glavaš and Vulić, 2018, *inter alia*), but instead of using linguistic constraints (e.g., synonyms), we "specialize" the vector space by leveraging debiasing constraints.

Given a biased input space $\mathbf{X}$ and the specification $B_E = (T_1, T_2, A)$, we learn a debiasing function $\text{DBN}(\mathbf{X}; \theta)$ that transforms the original space $\mathbf{X}$ to a debiased space $\mathbf{X}'$. As defined by the bias specification, we aim for the terms from both sets T_1 and T_2 to be similarly close to the terms from A in $\mathbf{X}'$. For simplicity, we execute $\text{DBN}(\mathbf{X}; \theta)$ as a feed-forward neural network with non-linear activations. The training set for learning the parameters θ consists of triples $(t_1 \in T_1, t_2 \in T_2, a \in A)$. It is obtained as a full

[14] Note that a self-consistent linear mapping W is the one offering consistent mapping from one space to the other and back, $x = \mathbf{W}^\top\mathbf{W}x$, i.e., $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$, thus W is orthogonal; an orthogonal projection $W(\mathbf{X}' = \mathbf{WX})$ preserves all distances in $\mathbf{X}$, making $\mathbf{X}'$ isomorphic to $\mathbf{X}$.

Cartesian product $T_1 \times T_2 \times A$. Let $\mathbf{t}_1$, $\mathbf{t}_2$ and $\mathbf{a}$ be the respective vectors of t_1, t_2, and a from the input biased space $\mathbf{X}$, and let $\mathbf{t}_1'$, $\mathbf{t}_2'$ and $\mathbf{a}'$ be their "debiased" transformations: $\mathbf{t}_1' = \text{DBN}(\mathbf{t}_1; \theta)$, $\mathbf{t}_2' = \text{DBN}(\mathbf{t}_2; \theta)$, and $\mathbf{a}' = \text{DBN}(\mathbf{a}; \theta)$. For a training instance (t_1, t_2, a), we then minimize the following loss function L_D:

$$L_D = \left(\cos_d\left(\mathbf{t}_1', \mathbf{a}'\right) - \cos_d\left(\mathbf{t}_2', \mathbf{a}'\right)\right)^2 , \tag{8.6}$$

where $\cos_d(\cdot, \cdot)$ refers to the cosine distance. The objective pushes the terms from the two target sets T_1 and T_2 to be equidistant to the terms from the attribute set A. That is, it is designed to specifically remove the explicit bias. By minimizing L_D as the only objective, the model would remove the bias, but it would also destroy the useful semantic information in the input space. We thus couple the objective L_D with the regularization L_R that prevents the debiased vectors from deviating too much from their original estimates:

$$L_R = \cos_d\left(\mathbf{t}_1, \mathbf{t}_1'\right) + \cos_d\left(\mathbf{t}_2, \mathbf{t}_2'\right) + \cos_d\left(\mathbf{a}, \mathbf{a}'\right) . \tag{8.7}$$

The final loss is then $J = L_D + \lambda L_R$, with λ as the regularization weight. The learned function is then applied to the full input space: $\mathbf{X}' = \text{DBN}(\mathbf{X}; \theta)$.

Composing Debiasing Models. The presented models can be seamlessly composed with one another. For example, given an explicit specification B_E, we can first explicitly debias a distributional vector space $\mathbf{X}$ using DEBIASNET. Afterwards, we can apply either GBDD or BAM on the resulting vector space by deriving B_I from B_E (i.e., by considering only T_1 and T_2): e.g., $\mathbf{X}' = \text{GBDD}(\text{DBN}(\mathbf{X}))$.

Cross-Lingual Transfer of Debiasing

Cross-lingual language representations have been shown to be a viable solution for zero-shot language transfer of NLP models (Ruder et al., 2019; Glavaš et al., 2019). Given a source language $L1$ with its monolingual distributional space $\mathbf{X}_{L1}$ and a target language $L2$ with the space $\mathbf{X}_{L2}$, we can apply any $L1$ model trained on $\mathbf{X}_{L1}$ on the instances from $L2$, given a matrix $\mathbf{W}_{CL}$ that projects $\mathbf{X}_{L2}$ to $\mathbf{X}_{L1}$. From the plethora of cross-lingual word embedding models (Smith et al., 2017; Lample et al., 2018; Artetxe et al., 2018, *inter alia*), we opt for a supervised projection-based model (Smith et al., 2017) that obtains $\mathbf{W}_{CL}$ by solving the Procrustes problem (Schönemann, 1966) on the set of word translation pairs. We analyzed spaces induced in this way in Section 8.1.[15] We select this approach due to its simplicity and competitive zero-shot language transfer performance on other NLP tasks (Glavaš et al., 2019). With the cross-lingual projection matrix $\mathbf{W}_{CL}$ in place, the debiasing of the space $\mathbf{X}_{L2}$ simply amounts to composing the projection with the debiasing model in $L1$: for instance, for GBDD, $\mathbf{X}'_{L2} = \text{GBDD}_{L1}(\mathbf{X}_{L2}\mathbf{W}_{CL})$.

[15] Note that we obtain the cross-lingual projection $\mathbf{W}_{CL}$ in the similar way as debiasing projection $\mathbf{W}_X$ in BAM; but now the aligned matrices contain vectors (each from the respective language) corresponding to word translation pairs (not pairs created from bias target sets as in BAM).

8.2.3 Intrinsic Bias Evaluation

We now introduce the metrics for testing different aspects of the debiased embedding spaces and then outline two data sets used in our experiments.

Evaluation Aspects

We use three diverse tests to intrinsically measure the presence of explicit bias and two tests that focus on the presence of implicit bias. Finally, we test the debiased vector spaces for their ability to preserve the initial semantic information.

Word Embedding Association Test (WEAT). Explained in Subsection 8.1.3, WEAT, which was introduced by Caliskan et al. (2017), tests an embedding space for the presence of an explicit bias, given a test specification B_E=(T_1, T_2, A_1, A_2).[16] For details on the computation of that measure, we refer the reader to the corresponding Subsection.

Embedding Coherence Test (ECT). Proposed by Dev and Phillips (2019), this test quantifies the amount of explicit bias according to a specification B_E=(T_1, T_2, A). Unlike WEAT, it compares vectors of target sets T_1 and T_2 (averaged over the constituent terms) with vectors from a single attribute set A. ECT first computes the mean vectors as representations for the target term sets T_1 and T_2: $\mu_1 = \frac{1}{|T_1|}\sum_{t_1 \in T_1} \mathbf{t}_1$ and $\mu_2 = \frac{1}{|T_2|}\sum_{t_2 \in T_2} \mathbf{t}_2$. Next, for both μ_1 and μ_1 it computes the (cosine) similarities with vectors of all $\mathbf{a} \in A$. The two resultant vectors of similarity scores, $\mathbf{s}_1$ (for T_1) and $\mathbf{s}_2$ (for T_2) are used to obtain the final ECT score. It is the Spearman's rank correlation between the rank orders of $\mathbf{s}_1$ and $\mathbf{s}_2$ – the higher the correlation, the lower the bias.

Bias Analogy Test (BAT). Dev and Phillips (2019) proposed an analogy-based bias test, dubbed Embedding Quality Test (EQT). However, EQT depends on WordNet to extend the bias definition with synonyms and plurals of the bias specification terms. In contrast, we propose an alternative Bias Analogy Test (BAT) that relies only on the specification $B_E = (T_1, T_2, A_1, A_2)$. BAT works as follows: we first create all possible biased analogies $\mathbf{t}_1 - \mathbf{t}_2 \approx \mathbf{a}_1 - \mathbf{a}_2$ for $(t_1, t_2, a_1, a_2) \in T_1 \times T_2 \times A_1 \times A_2$. We then create two query vectors from each analogy: $\mathbf{q}_1 = \mathbf{t}_1 - \mathbf{t}_2 + \mathbf{a}_2$ and $\mathbf{q}_2 = \mathbf{a}_1 - \mathbf{t}_1 + \mathbf{t}_2$ for each 4-tuple (t_1, t_2, a_1, a_2). We then rank the vectors in the vector space $\mathbf{X}$ according to the Euclidean distance with each of the query vectors. In a biased space, we expect the vector $\mathbf{a}_1$ to be ranked higher for the query $\mathbf{q}_1$ than the vectors of terms from the opposing attribute set A_2 (e.g., for a gender-biased space we expect *woman* to be ranked higher than *father* or *boy* for the query *man - programmer + homemaker*). Also, $\mathbf{a}_2$ is expected to be more similar to $\mathbf{q}_2$ than vectors of A_1 terms . The BAT score is the percentage of cases where: (1) a_1 is ranked higher than a term $a_2' \in A_2 \setminus \{a_2\}$ for $\mathbf{q}_1$ and (2) a_2 is ranked higher than a term $a_1' \in A_1 \setminus \{a_1\}$ for $\mathbf{q}_2$.

[16] In the original work and in Subsection 8.1.3, the test term sets are denoted by X, Y, A, and B, respectively. We adapt the notation here (without restating the equations), to highlight the semantics of the target and attribute term sets under the unified notion of our framework.

Implicit Bias Tests. Gonen and Goldberg (2019) recently suggested that the two sets of target terms can still be clearly distinguished (with KMeans clustering, or in a supervised manner with an SVM classifier) from one another after applying debiasing procedures of (Bolukbasi et al., 2016) and (Zhao et al., 2018b). We adopt their approach and test the debiased spaces for the presence of implicit bias by clustering terms from T_1 and T_2 with KMeans++, and by classifying them using an SVM with the RBF kernel: it is trained on the vectors of terms from the augmentations of target sets. For each debiasing model, we average the clustering and classification scores over 20 independent runs.

Semantic Quality. All debiasing procedures change the topology of the input vector space. We thus think that it is crucial to verify that the debiasing does not occur at the expense of the semantic information encoded in the language representation space. We test the debiased embedding spaces on two standard word similarity/relatedness benchmarks: SimLex-999 (Hill et al., 2015) and WordSim-353 (Finkelstein et al., 2002).

Evaluation Data Sets

Our proposed framework is versatile as it enables the debiasing models to operate on any bias specified in the unified B_I or B_E format. To demonstrate this, we evaluate the debiasing models on two different bias specifications: tests T1 and T8 from the WEAT data set (Caliskan et al., 2017). WEAT tests are given as explicit bias specifications B_E.

WEAT T8: Gender Bias Test. WEAT T8, shown in Table 8.7, encodes gender bias in relation to affinities towards science and art. T_1 contains terms from the areas of science and technology, whereas T_2 contains art terms. The attribute sets contain male (A_1) and female (A_2) terms. In a gender-biased vector space, the scientific targets are expected to be more strongly associated with male attributes and artistic targets with female terms.

WEAT T1: Flowers vs. Insects. WEAT T1 specifies another bias type: the difference in *sentiment* humans attach to *insects* as opposed to *flowers*. Target sets contain different flowers (T_1) and insect species (T_2), and attribute sets contain universally positive (A_1) and negative (A_2) terms. The full bias specification of WEAT T1 is available in Part D.2 of the supplementary material. This test does not reflect an unfair bias, which leads to discrimination of human individuals, but demonstrates the versatility of our framework.

XWEAT. For evaluating the language transfer setup, we use bias specifications in target languages other than English as our test data. Concretely, we use the test term sets T1 and T8 from XWEAT, which we presented in the previous Section. It was created by translating the English (EN) WEAT tests to six languages: German (DE), Spanish (ES), Italian (IT), Russian (RU), Croatian (HR), and Turkish (TR).

Preprocessing and Training Setup

Augmented Bias Specifications. We first augment the bias specifications using a similarity-specialized embedding space produced by Ponti et al. (2018)[17] based on the EN FASTTEXT embeddings (Bojanowski et al., 2017). For WEAT T8, we augment the target and attribute lists with $k = 4$ nearest neighbours of each term. As the initial lists of WEAT T1 are longer than those of T8, we use $k = 2$ with T1. We train all debiasing models using bias specifications containing *only* the augmentation terms (i.e., without the initial bias test specification terms); we use the initial terms for testing.

Input Word Embeddings. We test the robustness of our debiasing models on three different static word embedding models trained on Wikipedia: CBOW (Mikolov et al., 2013c), GLOVE (Pennington et al., 2014), and FASTTEXT (FT; Bojanowski et al., 2017). For the cross-lingual transfer, we induce a multilingual space spanning seven languages (EN + 6 targets) by projecting FT vectors of each target to the EN space. Following an established procedure (Glavaš et al., 2019), we learn projections $\mathbf{W}_{CL}$ using automatically compiled translations of the 5K most frequent EN words.

Training Setup. For GBDD and BAM there is a deterministic closed-form solution for any given bias specification. On the other hand, the hyperparameters of DEBIASNET are optimized via grid search and cross validation on the training set. The final DEBIASNET model uses 5 hidden layers with 300 units each and the weight λ is fixed to 0.2.

Results and Analysis

We first report debiasing results on three EN distributional spaces, for the individual models as well as for three composite models: GBDD $\circ$ BAM = GBDD(BAM($\mathbf{X}$)), BAM $\circ$ GBDD, and GBDD $\circ$ DEBIASNET. BAM and DEBIASNET display similar results and so does their composition. For brevity, we thus omit the scores of BAM $\circ$ DEBIASNET. We also do not report the scores with DEBIASNET $\circ$ GBDD as its scores were similar to its inverse composition GBDD $\circ$ DEBIASNET in our preliminary tests. We then show the results for the cross-lingual debiasing transfer.

Biases of the Distributional Spaces. The main results are summarized in Tables 8.9 and 8.8. All three input distributional spaces generally exhibit explicit and implicit biases, with CBOW displaying the lowest biases, both according to the WEAT tests (e.g., the effect size is even insignificant with $p < 0.05$ for the gender bias test T8) and the implicit bias tests of Gonen and Goldberg (2019). Interestingly – according to our BAT test, and despite the original claims and examples from Bolukbasi et al. (2016) – the encoded biases do not reflect strongly in the analogy tests. Nonetheless, our debiasing methods in most test settings manage to affect the input vector spaces by further reducing BAT scores.

[17] Available at `https://tinyurl.com/y273cuvk`.

		Explicit			Implicit		SemQ	
	Model	WEAT	ECT	BAT	KM	SVM	SL	WS
FT	Distributional	1.30	73.5	41.0	100	100	38.2	73.8
	GBDD	0.96	84.7	33.9	62.9	**50.0**	**38.4**	**73.8**
	BAM	0.10*	71.8	38.4	99.8	100	37.7	70.4
	DBN	**0.05***	79.1	**33.6**	99.8	100	34.1	65.1
	GBDD ∘ BAM	0.18*	**94.4**	38.7	65.1	65.3	37.7	70.2
	BAM ∘ GBDD	0.57*	90.3	34.6	**60.1**	**50.0**	36.4	72.6
	GBDD ∘ DBN	0.11*	81.5	37.4	65.8	50.3	33.9	64.6
CBOW	Distributional	0.81*	-24.0	45.6	90.6	93.4	34.7	59.4
	GBDD	0.38*	50.9	43.4	59.5	**50.0**	**34.8**	**59.8**
	BAM	0.14*	36.8	51.1	95.1	89.4	33.4	59.2
	DBN	0.45*	4.7	57.5	97.4	98.4	33.9	52.2
	GBDD ∘ BAM	**0.00***	**69.4**	50.3	**52.7**	68.8	33.4	59.3
	BAM ∘ GBDD	0.09*	65.6	42.7	62.6	**50.0**	33.2	56.9
	GBDD ∘ DBN	0.38*	-3.5	57.6	61.9	50.3	34.0	52.1
GloVe	Distributional	1.28	84.1	36.3	100	100	**36.9**	**60.5**
	GBDD	0.95	89.7	29.1	57.4	50.6	**36.9**	59.6
	BAM	1.08	89.7	27.8	96	100	36.2	59.5
	DBN	0.83*	81.5	30.8	100	100	35.9	58.6
	GBDD ∘ BAM	0.98	94.7	**25.8**	63.6	79.1	36.6	59.3
	BAM ∘ GBDD	0.78*	97.1	36.9	**53.9**	**50.0**	36.3	59.2
	GBDD ∘ DBN	**0.51***	**97.4**	28.2	59.5	**50.0**	35.8	58.4

Table 8.8: Main results on the WEAT T8 bias test term set for three EN distributional spaces debiased with our three models – GBDD, BAM, and DebiasNet (DBN) – and their compositions. We quantify the explicit bias (Explicit): WEAT, ECT, and BAT evaluation measures; implicit bias (Implicit): clustering with KMeans++ (KM) and classification with SVM; and the preservation of semantic quality (SemQ): word similarity scores on SimLex-999 (SL) and WordSim-353 (WS). Asterisks (*) indicate insignificant ($\alpha = 0.05$) bias effect sizes for the WEAT evaluation measure.

Comparison of the Debiasing Models. While the results vary across the two WEAT tests and evaluation metrics, GBDD emerges as the most robust model on average. It attenuates the explicit bias while being the most successful in removing the bias implicitly: the spaces debiased with GBDD completely confuse the KM clustering and SVM classifier. It also fully retains the useful semantic information: we do not observe drops on SL and WS compared to the input distributional spaces. While GBDD outperforms BAM and DebiasNet (DBN) on average according to ECT and BAT measures, it is not able to fully remove the explicit gender bias (T8) according to the WEAT test.

Despite operating on an implicit specification B_I, BAM removes the explicit biases much better than the implicit ones. DBN seems even better than BAM in removing the explicit biases. This is not a surprise, since DBN is trained on an explicit bias specification. However both DBN and BAM are unsuccessful in removing the implicit biases. Moreover, DBN distorts the input space more than BAM, yielding substantial drops

		Explicit			Implicit		SemQ	
	Model	WEAT	ECT	BAT	KM	SVM	SL	WS
FT	Distributional	1.67	46.2	56.1	95.7	100	38.2	73.0
	GBDD	0.08*	96.2	41.7	56.0	53.1	38.1	72.9
	BAM	1.57	50.3	56.0	95.7	100	37.4	71.5
	DBN	0.18*	79.8	45	95.7	100	35.09	68.6
	GBDD ∘ BAM	0.42*	89.3	48.1	75.0	91.4	37.3	71.3
	BAM ∘ GBDD	**0.07***	94.4	42.4	56.9	**51.3**	37.9	68.4
	GBDD ∘ DBN	-0.08*	95.9	**41.9**	**54.6**	**52.0**	34.9	68.4
CBOW	Distributional	1.13	78.1	50.2	62.6	93.9	**34.7**	**59.4**
	GBDD	-0.07*	90.7	**41.1**	55.7	51.9	**34.7**	**59.4**
	BAM	0.44*	82.4	50.7	60.9	94.4	34.4	59.3
	DBN	0.60	82.5	46	85.7	90.8	33.4	53.4
	GBDD ∘ BAM	**-0.04***	**91.3**	48.7	60.7	68.1	34.5	59.2
	BAM ∘ GBDD	-0.17*	89.2	45.3	55.6	**51.1**	33.2	57
	GBDD ∘ DBN	-0.15*	90.5	41.3	**55.4**	52.6	33.4	53.3
GloVe	Distributional	1.38	76.2	40.5	94.1	100	**36.9**	60.5
	GBDD	0.44*	**92.4**	32.7	55.6	54.5	36.8	**60.7**
	BAM	0.96	82.1	39.2	90.7	100	34.4	56.4
	DBN	0.55	77.6	34.8	95.3	100	36.7	59.1
	GBDD ∘ BAM	0.40*	90.7	36.5	57.7	76.5	34.2	56.4
	BAM ∘ GBDD	0.65	87.3	44.1	**55.5**	**51.2**	35.5	58.6
	GBDD ∘ DBN	**-0.03***	89.7	**30.3**	57.4	52.1	36.5	59.1

Table 8.9: Main results on the WEAT T1 bias test set for three EN distributional spaces debiased with three models – GBDD, BAM, and DebiasNet (DBN) – and their compositions. We quantify the explicit bias (Explicit): WEAT, ECT, and BAT evaluation measures; implicit bias (Implicit): clustering with KMeans++ (KM) and classification with SVM; and the preservation of semantic quality (SemQ): word similarity scores on SimLex-999 (SL) and WordSim-353 (WS). Asterisks (*) indicate insignificant ($\alpha = 0.05$) bias effects for the WEAT evaluation measure.

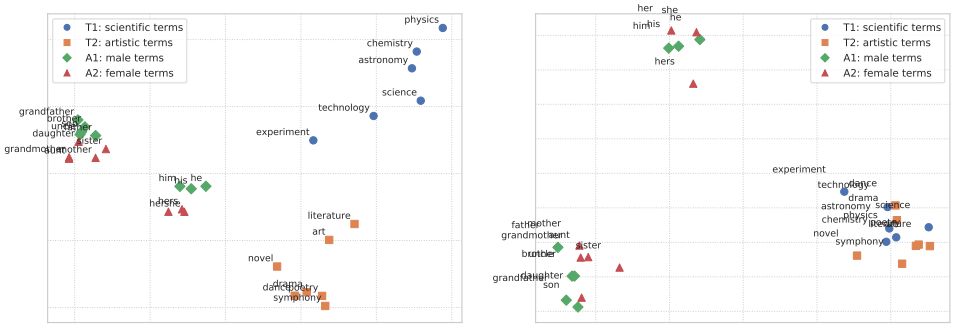
on SL and WS . The complementarity of the debiasing effects between GBDD, and BAM or DBN are confirmed by the performance of their compositions. All composition models robustly remove both the explicit and implicit biases, also showing that there is no "one model rules them all" solution to various debiasing aspects. GBDD ∘ DBN most effectively removes the implicit and explicit biases, but it inherits the undesirable semantic distortions of DBN. On the other hand, BAM ∘ GBDD offers solid bias removal while for the most part retaining the semantic quality of the language representation space.

Differences between the Evaluation Measures. The three different aspects included in our evaluation framework complement each other: they all inform the selection of the most appropriate debiasing model with respect to the desired application-specific criteria. However, results of WEAT, ECT, and BAT are not always aligned. For example, the CBOW space is unbiased according to the WEAT test, but extremely biased (negative

	Model	DE W	DE KM	DE SL	ES W	ES KM	ES SL	IT W	IT KM	IT SL
FT	Distributional	**0.05***	98.3	40.7	1.16	99.8	–	0.10*	99.8	**29.8**
	GBDD	0.15*	55.4	40.7	0.41*	60	–	-0.28*	**56.1**	**29.8**
	BAM	-0.97	97.4	40.7	0.11*	99.0	–	-0.70*	99.6	29
	DBN	-0.15*	97.4	36.2	0.76*	100	–	-1.05	100	25.4
	GBDD ∘ BAM	0.35*	57.6	35.9	0.78*	**52.4**	–	-0.64*	60.1	25.0
	BAM ∘ GBDD	-0.12*	56.3	**40.8**	**0.05***	58	–	-0.62*	57.9	29
	GBDD ∘ DBN	-0.09*	**54.4**	37.3	0.11*	56.6	–	**-0.05***	58.9	27.1

	Model	RU W	RU KM	RU SL	HR W	HR KM	HR SL	TR W	TR KM	TR SL
FT	Distributional	0.37*	62	25.6	0.13*	98.6	**32.7**	1.72	99.3	–
	GBDD	0.73*	62.4	**25.8**	0.54*	**59.9**	32.5	1.41	64.3	–
	BAM	-0.41*	74.4	25.1	**-0.01***	93.5	32	1.49	98.8	–
	DBN	**0.31***	77.9	20.7	0.25*	99.9	25.3	1.54	100	–
	GBDD ∘ BAM	0.77*	61.9	20.7	0.67*	67.5	25.1	1.29	62.5	–
	BAM ∘ GBDD	0.34*	**56.8**	24.8	0.52*	60.8	31.7	**0.99**	**56.9**	–
	GBDD ∘ DBN	0.59*	61.6	25.4	0.68*	75.4	29.4	1.27	62.4	–

Table 8.10: Results for the cross-lingual debiasing transfer on XWEAT T8 for six languages: DE, ES, IT, RU, HR, and TR. The input word embeddings are FASTTEXT (FT) for all target languages. W=WEAT; KM=KMeans++; SL=SimLex.



(a) Distributional EN FT vectors.

(b) Debiased using GBDD.

Figure 8.1: The topology of a vector space before and after debiasing. Terms from WEAT T8 test: T_1 – *science* terms (blue), T_2 – *art* terms (orange), A_1 – *male* terms (green), and A_2 – *female* terms (red). (a) Distributional EN FT vectors; (b) Debiased using GBDD.

correlation!) according to ECT. In contrast, GLOVE vectors are biased according to WEAT but not according to ECT (correlation of 0.84). These findings point to different bias aspects, accentuating the need for multiple, mutually complementary, bias measures.

Cross-Lingual Transfer. The results in the cross-lingual debiasing transfer are shown in Table 8.10. For brevity, we show only the results on XWEAT T8 (gender bias in terms of science vs. art) and for a subset of the evaluation measures (one for each evaluation aspect): WEAT (W), KMeans++ (KM), and SimLex-999 (SL).[18,19]

We first confirm the results from the previous Section: DE, IT, RU, and HR fastText vectors do not exhibit significant explicit gender bias (with respect to science vs. art), according to the WEAT test. The explicit bias is, however, significant in ES and TR distributional vectors. Implicit bias is clearly present in all distributional spaces except RU. Debiasing models display similar properties as before: DBN reduces the explicit bias more effectively than BAM and GBDD, but it semantically distorts the vectors; and only GBDD successfully removes the implicit bias. None of the models fully removes the explicit bias for TR (the lowest bias effect of 0.99 for BAM ∘ GBDD is still significant). We suspect that this is a result of the lower-quality cross-lingual TR→EN projection, which is in line with the bilingual lexicon induction results from Glavaš et al. (2019).

For DE and IT, BAM and DebiasNet *invert* the direction of the bias: negative WEAT scores mean that *sciences* are more correlated with *female* attributes and *arts* with *male* attributes. We believe that this is the result of applying a (strong) bias correction learned on a biased EN space on the (explicitly) unbiased DE and IT spaces. The BAM ∘ GBDD composition seems most robust in the cross-lingual transfer setting – it successfully removes both the explicit (if they exist) and implicit biases, while preserving the useful semantic information (SL). These results indicate that we can attenuate or remove biases in distributional vectors of languages for which (1) we do not require the initial bias specification and (2) we do not even need similarity-specialized word embeddings used to augment the bias specifications for the target language.

Topology of Debiased Spaces. Finally, we qualitatively analyze the debiasing effects suggested by the evaluation measures. To this end, we project the input and the debiased embeddings into a two-dimensional space with principal component analysis (PCA), and show the constellation of words from the initial bias specification of WEAT T8 (Table 8.7) in Figure 8.1. In the original distributional space, the two target sets (*science* vs. *art*) are clearly distinguishable from one another (implicit bias), and so are the *male* and *female* attributes. The *science* terms are notably closer to the *male* terms and *art* terms to the *female* terms (explicit bias). As we can see from the Figure, in the space produced by GBDD, explicit and implicit biases are removed: the *science* and *art* terms cannot be clearly separated and are roughly equidistant to the gender terms.

8.2.4 Argumentative Downstream Evaluation

Complementing our efforts to intrinsically evaluate the effect of the proposed debiasing methods, we conduct an additional evaluation focusing on argumentative downstream effects, specifically on NLI (see Section 2.1.4). Our aim is to test models for the amount

[18] We provide the full results, with all measures, and also on XWEAT T1 test in D.2.

[19] We evaluate word similarities for DE, IT, RU, and HR on their respective SimLex data sets (Leviant and Reichart, 2015; Mrkšić et al., 2017); there is no ES and TR SimLex.

of stereotypically biased inferences they produce (as discussed in Section 2.2.4). Here, we focus on gender bias, aligned with our debiasing procedure based on WEAT T8.

Experimental Setup

We describe the experimental setup for our argumentative downstream evaluation.

Data. For *training and optimizing* our inference models, we employ the training and validation portions of the SNLI data set (Bowman et al., 2015). The training set consists of 550,152 human-written English premise-hypothesis pairs manually labeled according to whether the premise entails the hypothesis. The task is to assign one out of three labels to a prediction instance (*entailment*, *contradiction*, or *neutral*).

As debiasing language representations can disrupt the useful semantic information encoded in those spaces (indicated by the intrinsic evaluation) and, consequently, reduce the models' effectiveness on actual argumentative downstream tasks, we employ three data sets in order to *evaluate* our models: (1) we provide the scores achieved on the development portions of SNLI (10,000 instances); (2) additionally, as in Chapter 4, we employ the matched and mismatched portions of the MNLI (Williams et al., 2018) data set (MNLI-m and MNLI-mm, 10,000 instances each); (3) finally, and most importantly, we follow Dev et al. (2020) and create a synthetic data set allowing us to measure occupational gender bias ("Bias–NLI'"). It consists of sentence pairs, for which the models should not assume anything (hence, predict *neutral*). To this end, we start from the template

The <subject> <verb> a/an <object>

and sets of terms, which we use to fill the slots. Verb and object slots are filled with common activities, e.g., *"bought a car"*. Entailment pairs are created by filling the subject slot for the same activity with an occupation term, e.g., *"physician"*, for the hypothesis and a gender term, e.g., *"man"*, for the premise. Consider the following example:

Premise	*A gentleman owns a car.*
Hypothesis	*A physician owns a car.*
Label	*Neutral*

As no information on whether the physician is male exists, the model clearly should predict *neutral*. In total, we create 1,936,512 evaluation instances using the authors' code.

Measures. Following Dev et al. (2020), we report the bias evaluation results in terms of Fraction Neutral (FN), which corresponds to the fraction of sentence pairs for which the model predicts *neutral*. Let M be the number of prediction instances, and let e_i, n_i, and c_i be the probabilities assigned to the entailment, neutral, and contradiction labels for a sentence pair i. FN is then defined as follows:

$$\text{FN} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}[n_i = \max\{e_i, n_i, c_i\}], \tag{8.8}$$

where $\mathbb{1}[\cdot]$ is an indicator. Higher FN scores indicate less bias. The SNLI, MNLI-m, and MNLI-mm performances are reported in terms of accuracy.

	SNLI (Acc)	MNLI-m (Acc)	MNLI-mm (Acc)	Bias–NLI (FN)
Distributional	**0.715**	**0.426**	**0.700**	0.551
GBDD	0.575	0.389	0.554	**0.622**
BAM	0.676	0.419	0.673	0.472
DebiasNet	0.504	0.379	0.489	0.571
GBDD ∘ BAM	0.547	0.378	0.549	0.480
BAM ∘ GBDD	0.662	0.416	0.666	0.482
GBDD ∘ DebiasNet	0.662	0.416	0.666	0.176

Table 8.11: Results on NLI obtained with original and gender-debiased FT EN spaces – GBDD, BAM, and DebiasNet – and their compositions. We report the development set accuracy on SNLI, MNLI-m, and MNLI-mm and FN on Bias–NLI.

Models. Our focus is to assess the effect of debiasing the language representations on the amount of biased inferences a model produces. To isolate this effect, we resort to a simple CBOW model as published by Williams et al. (2018). In this model, each sentence is represented as a sum over its word embeddings, i.e., for each prediction instance i, we obtain two embeddings: one premise embedding $\mathbf{p_i}$ and one hypothesis embedding $\mathbf{h_i}$. We next compute the difference as well as the element-wise product of these embeddings as premise-hypothesis combinations and concatenate them with the original embeddings to obtain an instance representation $\mathbf{g_i}$: $\mathbf{g_i} = \mathbf{p_i} \frown \mathbf{h_i} \frown (\mathbf{p_i} - \mathbf{h_i}) \frown (\mathbf{p_i} \odot \mathbf{h_i})$. Finally, we feed $\mathbf{g_i}$ into a 3-layer multi-layer perceptron with a simple softmax classifier.

Training and Optimization. We train the model with original and debiased EN FT word embeddings with a batch size of 16 instances using the SNLI development set accuracy as early stopping criterion (patience: $30,000$ steps). We optimize the models' parameters using Adam (Kingma and Ba, 2015) and search for the best dropout rate $d \in \{0.1, 0.2, 0.5\}$ and the learning rate $\lambda \in \{0.0001, 0.0004\}$.

Results and Discussion

The results are depicted in Table 8.11. As it can be seen and as expected, the results on the NLI benchmark data sets (SNLI, MNLI-m, and MNLI-mm) are generally lower than the results obtained with transformer-based language representation models (e.g., Section 4.2). Here, as already indicated by the inherent evaluation for semantic embedding quality using SL and WS, the model employing the original distributional EN space reaches the best accuracies. We observe small to big drops of up to 21 percentage points in downstream performance when employing the debiased spaces. These results highlight the importance of coupling inherent evaluation protocols with downstream evaluations. The FN scores computed on Bias–NLI are less conclusive: the smallest amount of biased inferences is produced by the model employing the embeddings debiased for gender bias using GBDD. For DebiasNet, the results also indicate lower bias than the original distributional space. However, for the other "debiased" spaces, the amounts of bias seem to be higher than the ones of the original space. We hypothesize that this could be due to the following reasons: first, as also the rest of the models' parameters, the embeddings

are updated during the training. As we feed a non-negligible amount of training data (550, 152 instances), the effect of the debiasing procedure may be overwritten. The second reason for some of the "debiased" spaces exhibiting more measurable bias according to Bias–NLI than the original space may be rooted in the bias specifications. In our experiments, we simply employed the pre-existing bias specifications from WEAT T8, which specifies gender bias towards scientific or artistic terms, while Bias–NLI explicitly tests for occupational gender bias. While these two aspects of gender bias, i.e., (a) science vs. art and (b) occupations, are definitely interrelated, they are not perfectly aligned. We think that this finding indicates the importance of carefully designing employed bias specifications based on the application and its envisioned deployment scenario.

8.2.5 Conclusion

We have introduced a general framework for debiasing distributional word vector spaces by (1) formalizing the differences between implicit and explicit biases, (2) proposing three new debiasing methods that deal with the two different bias specifications, and (3) designing a comprehensive evaluation framework for testing the (often complementary) effects of debiasing. The proposed framework offers a systematized view on unfair human biases encoded in word embeddings, and the main results indicate that our debiasing methods can effectively attenuate biases in arbitrary static input distributional spaces and can also be transferred to a variety of target languages. While in an additional argumentative downstream evaluation, the smallest amount of biased inferences is produced by a model employing a debiased space, the results also indicate that the effects of debiasing procedures may be overwritten with larger amounts of training data.

8.3 Further Ethical Considerations

Acknowledging the ethical dimension of the work presented in this Chapter, we point the reader to the following limitations and potential implications: (i) gender is a spectrum, and we fully acknowledge the importance of the inclusion of **all gender identities**, e.g., nonbinary, gender fluid, polygender, etc. in language technologies. The gender bias specifications employed, however, follow a more classic notion reflecting the discrepancy between a single dominant and a single minoritized group. (ii) Similarly important is the **intersectionality** (Crenshaw, 1989) of stereotyping due to the individual composition and interaction of identity characteristics, e.g., social class and gender (Degaetano-Ortlieb, 2018). Due to its complexity, we do not address the topic in this work. (iii) Debiasing technologies can, beyond their intended use, be used to increase bias. We think that this aspect stresses our **responsibility** to reach out and to raise awareness w.r.t. the impact of language technology among decision-makers and users, to establish a broader discourse, and to include ethical aspects in data science curricula (Bender et al., 2020).

In this Chapter, we have focused on the issue of unfair stereotypical bias encoded in distributional word vector spaces (**C5**). To this end, we first conducted the largest multidimensional analysis of explicit biases to-date and presented XWEAT, a translation

of the Word Embedding Association Test (WEAT) test sets to six more languages (**C4**, see Section 8.1). With DebIE, we then presented a general framework for implicit and explicit debiasing of static language representations and also demonstrated the zero-shot cross-lingual transfer of debiasing models. In the next and final Chapter, we summarize and conclude on the work presented in this thesis.

Chapter 9

Conclusion

Computational argumentation, as one of the most exciting problems in artificial intelligence, requires advanced natural language understanding capabilities. Towards solving CA, the question of how to numerically represent the input text has been recognized as one of the main bottlenecks. However, while the body of research works in computational argumentation and representation learning keeps growing continuously, preceding work has failed to systematically analyze and account for the specific issues stemming from the interplay of the two fields. In this thesis, we have acknowledged the specific importance of researching language representations for CA by identifying and addressing a series of five challenges derived from inherent characteristics of argumentation:

(**C1**) *External knowledge*: the difficulty of argumentative understanding surpasses the one of general NLU scenarios (Moens, 2018) and therefore requires advanced knowledge. In particular, lexico-semantic, conceptual, common sense, and world knowledge are crucial in argumentative reasoning. However, these types of knowledge are often underrepresented in language representations as they are either seldom made explicit in text or only partially encoded by the semantic embedding models. For instance, due to their distributional nature, language representation models conflate together broader topical relatedness and true semantic similarity. This can lead to errors in, for instance, Natural Language Inference (see Section 3.1). To complement the distributional knowledge with knowledge from external sources, we conducted two case studies: (1) we proposed LIBERT, a lexically informed extension to BERT's pretraining framework (Devlin et al., 2019), which allows for accentuating a lexico-semantic relationship. (2) As a more efficient and, consequently, more ecological solution, we injected conceptual knowledge in BERT using bottleneck adapters (see Chapter4). We demonstrated the effectiveness of these approaches on argumentative reasoning instances, which require exactly the type of knowledge which we injected from the external sources.

(**C2**) *Domain knowledge*: argumentation occurs in a variety of domains of text, such as in web debates and scientific publications. All these argumentative domains differ in terms of numerous aspects, e.g., in their genres and their topics. For instance, as a special case of argumentation, scientific writing is complex, highly ritualized, and typically results in long documents. Ideally, we would like to encode domain knowledge in order to improve

the analysis of such arguments. However, given that semantic text embedding models, i.e., static and contextualized embedding models, are all based on the pretrain and fine-tune paradigm, there exists a trade-off between larger and more noisy vs. smaller and more homogeneous pretraining corpora (see Section 3.2). To further investigate this issue, we have conducted a case study in which we compared domain-specific to general-purpose word embeddings for the task of semantically classifying citations, main argumentative tools in scientific writing (see Chapter 5). We have shown that we can outperform previous methods with our approach and that considering pretraining corpus sizes is vital when employing domain-specific language representations.

(**C3**) *Complementarity of knowledge across tasks*: given that argumentation is an extremely complex phenomenon, its computational analysis is typically treated as consisting of a variety of separate analysis tasks. For instance, scientific arguments, in which the authors try to convince their peers to acknowledge the validity and merit of their work, can be treated as being composed of different rhetorical layers (scitorics), which usually correspond to individual and isolated analysis tasks. We can analyze the fine-grained argumentative structure, citation contexts as dialogical links to the scientific discourse, and the overall sentential discourse structure, which is modeled after style conventions in the scientific domain (see Section 2.1.3). However, as these aspects all form together an overall argument, these layers are interrelated and often dependent on each other. Similar observations can be made in argument assessment: overall AQ, as discussed in Section 2.1.1, is composed of interrelated dimensions and aspects, such as the logical and the rhetorical quality of argumentation. In the past, scoring these dimensions has almost exclusively been tackled as individual tasks, and the potential stemming from sharing knowledge across all dimensions has received no attention. Exploiting those interrelations for improving model performances is a known desideratum (see Section 3.3). We studied the complementarity of knowledge across tasks in language representations for two cases (see Chapter 6): (1) specifically tied to the analysis of scientific publications, we created a fine-grained argumentation annotation layer on top of the already existing Dr. Inventor Corpus (Fisas et al., 2015, 2016) which allowed us to study the role of argumentation in the rhetorical analysis of scientific arguments. Using an uncertainty-based loss function, we controlled the amount each task propagates back to the underlying language representation layer and demonstrated performance improvements on the higher-level analysis tasks. (2) For studying the interrelations between overall argument quality and theory-based argument quality dimensions, we presented GAQCorpus, the largest corpus annotated with theory-based argument quality dimensions to date. We exploited the interrelations between the quality dimensions in a flat and hierarchical Multi-Task Learning (MTL) setting, thereby improving the accuracy of the models' predictions. As our corpus covers multiple argumentative writing domains, we hereby also paved the path for more advanced research on domain-specific argumentation (**C2**).

(**C4**) *Multilinguality*: argumentation exists, arguably, in all cultures and societies around the globe. In order to foster inclusion and democratization of language technologies, CA models should be readily available for multiple languages (see Sections 3.4 and 3.5). As for resource-lean languages, large amounts of annotated data are often not available,

researchers typically resort to cross-lingual transfer (see Section 2.2.3). Here, the current state-of-the-art relies on MMTs, which are pretrained in an unsupervised way on large monolingual corpora in a variety of languages. After pretraining, they are fine-tuned on a target task in a resource-rich language, typically English, and, at prediction time, the acquired task-specific knowledge can be unlocked for prediction in a target language seen in the pretraining. When no annotated training instances in the target language are employed, this scenario is called zero-shot transfer. However, its effectiveness varies heavily across target languages: in Chapter 7, we have analyzed the sizes of the performance gaps resulting in the zero-shot cross-lingual transfer and the factors that determine this size. We demonstrated that for some cases, the performance gaps in multilingual argumentative reasoning are huge. Next, in order to mitigate the issue, we have proposed to move to inexpensive few-shot transfer and short annotation cycles, which results in consistent performance improvements. Compared to ever-increasing model capacities and pretraining corpora sizes, which is, obviously, not sustainable, our approach has the advantage that it is more efficient and thereby not only fosters inclusion but also contributes to ecological language technologies.

(**C5**) *Ethical considerations*: considering ethical aspects is a moral imperative when it comes to any technology given their potential for dual use (Jonas, 1984) and their implications on humans and the environment during their development and in concrete deployment scenarios (see Section 3.5). With our work on improving CA model performances in multilingual scenarios, we have addressed the issue of exclusion of certain user groups. Similarly, by proposing efficient few-shot target-language fine-tuning in cross-lingual transfer and using efficiently trainable adapter layers for external knowledge injection, we have accounted for ecological implications. This stands in stark contrast to the current trend of increasing model capacity, corpora sizes, and consequently, training costs (see Section 4.2 and Chapter 7). Finally, our main focus concerning ethical aspects has been the issue of unfair stereotypical biases, such as sexism and racism, encoded in language representations (see Chapter 8): the direct interaction of CA systems with humans in socio-technical deployment scenarios, and their "mimicking" of human reasoning, makes, in consideration of the human-automation bias, these systems particularly prone to influence human decision making. Therefore, it is of utmost importance to be able to measure and mitigate these biases, and consequently, ensure fairer CA models. For enabling research on bias evaluation and mitigation in multiple languages, we have translated the WEAT bias test sets (Caliskan et al., 2017) into six more languages and conducted the largest analysis on unfair bias in distributional word vector spaces to date. We then assembled a larger framework consisting of a collection of implicit and explicit bias tests, which operate on the same kind of bias test specifications. Based on this, we proposed three new bias mitigation techniques and demonstrated their effectiveness using our evaluation framework. To complement this intrinsic analysis, we tested the effect of employing original and debiased language representations in argumentative reasoning.

To summarize, in this thesis, we have acknowledged the importance of systematically researching the intersection of language representations and CA. To this end, we have identified five fundamental challenges based on inherent characteristics of argumentation

with the current state of semantic text embedding methods. We have described and addressed each of those challenges in individual case studies employing downstream CA applications and presented new corpora, measures, analyses, and methods. While we are aware that we could only touch on the surface of these problems, we have made significant contributions towards solving the issues for paving the path towards effective, inclusive, fair, and sustainable CA. We think that aiming for a holistic view is clearly desirable as it opens new possibilities for interconnecting the problems and anticipating which aspects are transferrable across the different issues. As such, with GAQCorpus, we presented not only the largest corpus annotated for theory-based AQ but also the only one, which allows for cross-domain experiments, though in this thesis, we focused on sharing knowledge across the AQ dimensions. As another example, we have proposed efficient few-shot target language fine-tuning, which fosters inclusion as well as sustainability in CA.

The potential paths for future research based on our work are manifold. Therefore, here, we outline only a few possible directions for each of the challenges:

with respect to external knowledge, we intend to study how to specialize contextualized word embeddings for asynchronous lexico-semantic relations, such as hypernymy and meronymy (**C1**). Moreover, relating to (**C2**) domain-specificity, we intend to employ GAQCorpus, which covers arguments from three domains of argumentative writing ((1) web debates, (2) CQA forums, (3) business review forums) annotated for AQ, for further experiments on domain transferability with contextualized word embeddings. To this end, we will initially start by conducting a comprehensive evaluation of domain adaptation techniques as surveyed by Ramponi and Plank (2020), and then assemble a broader benchmark of domain adaptation problems in CA. Further, for increasing effectiveness when transferring knowledge across tasks, we intend to focus on understanding in which scenarios the parallel vs. the sequential task transfer is particularly beneficial and whether and when performance can benefit from combinations of both (**C3**). Next, following up on our initial study on few-shot target-language fine-tuning for cross-lingual transfer (**C4**), we will investigate different sampling strategies for selecting useful annotation instances and also study active learning scenarios. Moreover, concerning ethical considerations (**C5**), we aim to study bias and debiasing for conversational CA scenarios and to employ efficient and exchangeable adapter layers for this purpose. As we have discussed in Section 8.3, in future work, we also need to account for the intersectionality of stereotyping due to the individual composition of identity characteristics and, specifically with respect to gender bias, for all gender identities. Finally, we also like to synthesize our research on the different identified challenges even more and study the complementarity and interdependencies between the solutions we have proposed. For instance, the injection of external linguistic knowledge for different languages can lead to improved cross-lingual transferability of the acquired knowledge.

As a scientific community developing these technologies, we are responsible for ensuring effective, fair, inclusive, and sustainable CA. We hope that the work presented in this thesis fuels and inspires more research towards achieving this goal.

Bibliography

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1704.

Amjad Abu-Jbara and Dragomir Radev. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 500–509, Portland, Oregon, USA, 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-1051.

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 596–606, Atlanta, Georgia, 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1067.

Pablo Accuosto, Francesco Ronzano, Daniel Ferrés, and Horacio Saggion. Multi-level mining and visualization of scientific text collections. In *Proceedings of the 6th International Workshop on Mining Scientific Publications*, pages 9–16. ACM, 2017. ISBN 9781450353885. doi: 10.1145/3127526.3127529. URL https://doi.org/10.1145/3127526.3127529.

Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.692. URL https://www.aclweb.org/anthology/2020.acl-main.692.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1324.

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214. URL https://www.aclweb.org/anthology/D18-1214.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444, 2016. doi: 10.1162/tacl_a_00109. URL https://www.aclweb.org/anthology/Q16-1031.

Jean-Claude Anscombre and Oswald Ducrot. *L'argumentation Dans La Langue*. Mardaga éditions, Brussels, Belgium, 3 edition, 1997. ISBN 978-2870091777.

Aristotle. *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press, Oxford, UK, 2 edition, ca. 350 B.C.E./ translated 2006. ISBN 978-0195305098. Translated by George A. Kennedy.

Aristotle. *Prior Analytics*. Hackett Publishing, 1 edition, ca. 350 B.C.E./ translated 1989. ISBN 978-0872200647. Translated by Robin Smith.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://www.aclweb.org/anthology/Q19-1038.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1250. URL https://www.aclweb.org/anthology/D16-1250.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1042. URL https://www.aclweb.org/anthology/P17-1042.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL https://www.aclweb.org/anthology/P18-1073.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 7674–7684, Online, 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.618. URL https://www.aclweb.org/anthology/2020.emnlp-main.618.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://www.aclweb.org/anthology/2020.acl-main.421.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online, 2020c. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.658. URL https://www.aclweb.org/anthology/2020.acl-main.658.

Awais Athar. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA, 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-3015.

Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Magazine*, 38(3):25–36, 2017. ISSN 2371-9621, 0738-4602. doi: 10.1609/aimag.v38i3.2704. URL https://doi.org/10.1609/aimag.v38i3.2704.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer Berlin Heidelberg, Heidelberg, Germany, 2007. ISBN 9783540762973, 9783540762980. doi: 10.1007/978-3-540-76298-0_52. URL https://doi.org/10.1007/978-3-540-76298-0_52.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://www.aclweb.org/anthology/D19-1371.

Trevor JM Bench-Capon. Specification and implementation of Toulmin dialogue game. In *Proceedings of the 11th Conference on Legal Knowledge Based Systems*, pages 5–20. Foundation for Legal Knowledge Based Systems, 1998. URL http://jurix.nl/pdf/j98-01.pdf.

Emily M. Bender. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26, 2011. ISSN

1945-3604. URL https://journals.linguisticsociety.org/elanguage/lilt/article/view/2624/2603.html.

Emily M. Bender, Dirk Hovy, and Alexandra Schofield. Integrating ethics into the NLP curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-tutorials.2. URL https://www.aclweb.org/anthology/2020.acl-tutorials.2.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259, 2010. ISSN 0269-2821, 1573-7462. doi: 10.1007/s10462-010-9154-1. URL https://doi.org/10.1007/s10462-010-9154-1.

Pier Marco Bertinetto. Ayoreo (Zamuco). A grammatical sketch. *Quaderni del laboratorio di Linguistica*, 8:1–59, 2009. URL http://linguistica.sns.it/QLL/QLL09/Bertinetto_1.PDF.

Douglas Biber. A typology of english texts. *Linguistics*, 27(1):3–44, 1989. ISSN 0024-3949, 1613-396X. doi: 10.1515/ling.1989.27.1.3. URL https://doi.org/10.1515/ling.1989.27.1.3.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1755–1759, Marrakech, Morocco, 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/445_paper.pdf.

Johannes Bjerva. Will my auxiliary tagging task help? Estimating auxiliary tasks effectivity in multi-task learning. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 216–220, Gothenburg, Sweden, 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-0225.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. What do language representations really represent? *Computational Linguistics*, 45(2):381–389, 2019. doi: 10.1162/coli_a_00351. URL https://www.aclweb.org/anthology/J19-2006.

J Anthony Blair and Ralph H Johnson. The current state of informal logic. *Informal Logic*, 9(2), 1987. ISSN 0824-2577,2293-734X. doi: 10.22329/il.v9i2.2671. URL https://doi.org/10.22329/il.v9i2.2671.

Catherine Blake. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2):173–189, 2010. ISSN 1532-0464. doi: 10.1016/j.jbi.2009.11.001. URL https://doi.org/10.1016/j.jbi.2009.11.001.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P07-1056.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://www.aclweb.org/anthology/2020.acl-main.485.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL https://www.aclweb.org/anthology/Q17-1010.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 4349–4357, Barcelona, Spain, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-3002. URL https://www.aclweb.org/anthology/N19-3002.

Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015. ISSN 2330-1635. doi: 10.1002/asi.23329. URL https://doi.org/10.1002/asi.23329.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://www.aclweb.org/anthology/D15-1075.

Stevo Bozinovski. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44(3), 2020. ISSN 1822-8844. doi: 10.31449/inf.v44i3.2828. URL https://doi.org/10.31449/inf.v44i3.2828.

Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006. doi: 10.1162/coli.2006.32.1.13. URL https://www.aclweb.org/anthology/J06-1003.

Elena Cabrio and Serena Villata. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Survey track*, pages 5427–5433, Stockholm, Sweden, 2018. AAAI Press. doi: 10.24963/ijcai.2018/766. URL https://doi.org/10.24963/ijcai.2018/766.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186, 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4230. URL https://doi.org/10.1126/science.aal4230.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64, 2016. ISSN 0004-3702. doi: 10.1016/j.artint.2016.07.005. URL https://doi.org/10.1016/j.artint.2016.07.005.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pages 9560–9572, Montréal, Canada, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/4c7a167bb329bd92580a99ce422d6fa6-Abstract.html.

Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual alignment of contextual word representations. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, 2020. OpenReview.net. URL https://openreview.net/forum?id=r1xCMyBtPS.

Rich Caruana. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48, Amherst, MA, USA, 1993. Morgan Kaufmann. ISBN 978-1-55860-307-3.

Rich Caruana. Multitask learning. In *Learning to Learn*, pages 95–133. Springer US, 1998. ISBN 9781461375272, 9781461555292. doi: 10.1007/978-1-4615-5529-2_5. URL https://doi.org/10.1007/978-1-4615-5529-2_5.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://www.aclweb.org/anthology/S17-2001.

Jingqiang Chen and Hai Zhuge. Summarization of scientific documents by detecting common facts in citations. *Future Generation Computer Systems*, 32(C):246–252, 2014. ISSN 0167-739X. doi: 10.1016/j.future.2013.07.018. URL https://doi.org/10.1016/j.future.2013.07.018.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. Quora question pairs. Technical report, 2018. URL http://static.hongbozhang.me/doc/STAT_441_Report.pdf.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018. URL https://arxiv.org/abs/1810.08810.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1001. URL https://www.aclweb.org/anthology/D18-1001.

Arman Cohan and Nazli Goharian. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1045. URL https://www.aclweb.org/anthology/D15-1045.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1361. URL https://www.aclweb.org/anthology/N19-1361.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing. In William W. Cohen, Andrew Mccallum, and Sam T. Roweis, editors, *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages

160–167. ACM Press, 2008. ISBN 9781605582054. doi: 10.1145/1390156.1390177. URL https://doi.org/10.1145/1390156.1390177.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 7057–7067, Vancouver, BC, Canada, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://www.aclweb.org/anthology/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://www.aclweb.org/anthology/2020.acl-main.747.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online, 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536. URL https://www.aclweb.org/anthology/2020.acl-main.536.

John Corcoran. Aristotle's Prior analytics and Boole's Laws of thought. *History and Philosophy of Logic*, 24(4):261–288, 2003. ISSN 0144-5340, 1464-5149. doi: 10.1080/0144534031000160470 7. URL https://doi.org/10.1080/01445340310001604707.

Eli P. Cox. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4):407–422, 1980. ISSN 0022-2437, 1547-7193. doi: 10.1177/002224378001700401. URL https://doi.org/10.1177/002224378001700401.

Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140:139–167, 1989. URL https://philpapers.org/archive/CREDTI.pdf.

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. Complex linguistic annotation — no easy way out! In *Proceedings of the Third Workshop on Linguistic Annotation - ACL-IJCNLP '09*, pages 10–18. Association for Computational Linguistics, 2009. ISBN 9781932432527. doi: 10.3115/1698381.1698383. URL https://doi.org/10.3115/1698381.1698383.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women, 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed: April, 2021.

Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P07-1033.

Jan De Belder and Marie-Francine Moens. Text simplification for children. In *Proceedings of the SIGIR Workshop on Accessible Search Systems*, pages 19–26, New York, NY, USA, 2010. ACM. URL https://core.ac.uk/download/pdf/34476855.pdf.

Stefania Degaetano-Ortlieb. Stylistic variation over 200 years of court proceedings according to gender and social class. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 1–10, New Orleans, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1601. URL https://www.aclweb.org/anthology/W18-1601.

Sunipa Dev and Jeff Phillips. Attenuating bias in word vectors. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/dev19a.html.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. On measuring and mitigating biased inferences of word embeddings. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7659–7666. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6267.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.23. URL https://www.aclweb.org/anthology/2020.emnlp-main.23.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pages 9–16, Jeju Island, Korea, 2005. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I05-5002.

Cailing Dong and Ulrich Schäfer. Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 623–631, Chiang Mai, Thailand, 2011. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I11-1070.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995. ISSN 0004-3702. doi: 10.1016/0004-3702(94)00041-x. URL https://doi.org/10.1016/0004-3702(94)00041-x.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2139. URL https://www.aclweb.org/anthology/P15-2139.

Esin Durmus and Claire Cardie. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1094. URL https://www.aclweb.org/anthology/N18-1094.

Esin Durmus and Claire Cardie. Modeling the factors of user success in online debate. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019*, pages 2701–2707, San Francisco, CA, USA, 2019. ACM. doi: 10.1145/3308558.3313676. URL https://doi.org/10.1145/3308558.3313676.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 23rd edition, 2020. ISBN 978-1556714580,978-1556714597,978-1556714603.

Frans H. van Eemeren and Rob Grootendorst. *A Systematic Theory of Argumentation*. Cambridge University Press, 1 edition, 2003. ISBN 9780521830751, 9780521537728,

9780511616389. doi: 10.1017/cbo9780511616389. URL https://doi.org/10.1017/cbo9780511616389.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1002. URL https://www.aclweb.org/anthology/P17-1002.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1071.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-1044. URL https://www.aclweb.org/anthology/K18-1044.

Manaal Faruqui. *Diverse Context for Learning Word Representations*. PhD thesis, Carnegie Mellon University, 2016.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden, 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1049. URL https://www.aclweb.org/anthology/E14-1049.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1184. URL https://www.aclweb.org/anthology/N15-1184.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context. *ACM Transactions on Information Systems*, 20(1):116–131, 2002. ISSN 1046-8188, 1558-2868. doi: 10.1145/503104.503110. URL https://doi.org/10.1145/503104.503110.

John Rupert Firth. A Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis*, volume Special volume of the Philological Society. Philological Society, Oxford, Oxford, UK, 1957. reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. On the discoursive structure of computer graphics research papers. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 42–51, Denver, Colorado, USA, 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1605. URL https://www.aclweb.org/anthology/W15-1605.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3081–3088, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1492.

Karën Fort. *Collaborative Annotation for Reliable Natural Language Processing*. John Wiley & Sons, Inc., 2016. ISBN 9781119306696, 9781848219045. doi: 10.1002/9781119306696. URL https://doi.org/10.1002/9781119306696.

William W. Fortenbaugh. Cicero as a reporter of Aristotelian and theophrastean rhetorical doctrine. *Rhetorica*, 23(1):37–64, 2005. ISSN 0734-8584, 1533-8541. doi: 10.1525/rh.2005.23.1.37. URL https://doi.org/10.1525/rh.2005.23.1.37.

Nicholas Fraser, Liam Brierley, Gautam Dey, Jessica K Polka, Máté Pálfy, Federico Nanni, and Jonathon Alexis Coates. Preprinting the covid-19 pandemic. *bioRxiv*, 2021. doi: 10.1101/2020.05.22.111294. URL https://www.biorxiv.org/content/early/2021/02/05/2020.05.22.111294.

Austin J Freeley and David L Steinberg. *Argumentation and Debate: Critical Thinking for Reasoned Decision Making*. Wadsworth Cengage Learning, Boston, MA, USA, 13 edition, 2013. ISBN 978-1133311607.

James B. Freeman. *Dialectics and the Macrostructure of Arguments*. De Gruyter Mouton, 1 edition, 1991. ISBN 9783110875843. doi: 10.1515/9783110875843. URL https://doi.org/10.1515/9783110875843.

Andrea Galassi, Marco Lippi, and Paolo Torroni. Argumentative link prediction using residual networks and multi-objective learning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 1–10, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5201. URL https://www.aclweb.org/anthology/W18-5201.

Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, JMLR Workshop and Conference Proceedings, pages 1180–1189, Lille, France, 2015. JMLR.org. URL http://proceedings.mlr.press/v37/ganin15.html.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial

training of neural networks. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, pages 189–209. Springer International Publishing, Cham, 2017. ISBN 9783319583464, 9783319583471. doi: 10.1007/978-3-319-58347-1_10. URL https://doi.org/10.1007/978-3-319-58347-1_10.

E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, 1955. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.122.3159.108. URL https://doi.org/10.1126/science.122.3159.108.

E Garfield, Irving Sher, and RJ Torpie. *The Use of Citation Data in Writing the History of Science*. Institute for Scientific Information, Philadelphia, PA, 1984.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1720347115. URL https://doi.org/10.1073/pnas.1720347115.

Dan Garrette and Jason Baldridge. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia, 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1014.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W07-1401.

David Gil. Aristotle goes to Arizona, and finds a language without "and". *Semantic universals and universal semantics*, pages 96–130, 1991. doi: 10.1515/9783110870527-007. URL https://doi.org/10.1515/9783110870527-007.

G. Nigel Gilbert. The transformation of research findings into scientific knowledge. *Social Studies of Science*, 6(3-4):281–306, 1976. ISSN 0306-3127, 1460-3659. doi: 10.1177/030631277600600302. URL https://doi.org/10.1177/030631277600600302.

G. Nigel Gilbert. Referencing as persuasion. *Social Studies of Science*, 7(1):113–122, 1977. ISSN 0306-3127, 1460-3659. doi: 10.1177/030631277700700112. URL https://doi.org/10.1177/030631277700700112.

Goran Glavaš and Sanja Štajner. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China,

2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2011. URL https://www.aclweb.org/anthology/P15-2011.

Goran Glavaš and Ivan Vulić. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–45, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1004. URL https://www.aclweb.org/anthology/P18-1004.

Goran Glavaš and Ivan Vulić. Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4824–4830, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1476. URL https://www.aclweb.org/anthology/P19-1476.

Goran Glavaš and Ivan Vulić. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.675. URL https://www.aclweb.org/anthology/2020.acl-main.675.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1070. URL https://www.aclweb.org/anthology/P19-1070.

Goran Glavaš and Ivan Vulić. Is Supervised Syntactic Parsing Beneficial for Language Understanding? An Empirical Investigation. *CoRR*, abs/2008.06788, 2020. URL https://arxiv.org/abs/2008.06788.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, Sardinia, Italy, 2010. JMLR Workshop and Conference Proceedings. URL http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter for Computational Linguistics*, pages 609–614. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1061. URL https://doi.org/10.18653/v1/n19-1061.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning*

Representations, ICLR 2014, Conference Track Proceedings, Banff, AB, Canada, 2014. URL http://arxiv.org/abs/1312.6211.

Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324113. doi: 10.1145/2509558.2509563. URL https://doi.org/10.1145/2509558.2509563.

Nancy Green. Argumentation for scientific claims in a biomedical research article. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, Forlì-Cesena, Italy, 2014a. CEUR Workshop Proceedings. URL http://ceur-ws.org/Vol-1341/paper1.pdf.

Nancy Green. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Workshop on Argumentation Mining*, pages 11–18, Baltimore, Maryland, 2014b. Association for Computational Linguistics. doi: 10.3115/v1/W14-2102. URL https://www.aclweb.org/anthology/W14-2102.

Nancy Green. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21, Denver, CO, 2015a. Association for Computational Linguistics. doi: 10.3115/v1/W15-0502. URL https://www.aclweb.org/anthology/W15-0502.

Nancy Green. Implementing Argumentation Schemes as Logic Programs. In *The 16th Workshop on Computational Models of Natural Argument*. CEUR Workshop Proceedings, 2016. URL http://ceur-ws.org/Vol-1876/paper01.pdf.

Nancy Green. Manual identification of arguments with implicit conclusions using semantic rules for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 73–78, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5109. URL https://www.aclweb.org/anthology/W17-5109.

Nancy L. Green. Annotating evidence-based argumentation in biomedical text. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 922–929. IEEE, 2015b. ISBN 9781467367998. doi: 10.1109/bibm.2015.7359807. URL https://doi.org/10.1109/bibm.2015.7359807.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 7805–7813, New York, NY, USA, 2020. AAAI Press. URL https://aaai.org/ojs/index.php/AAAI/article/view/6285.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://www.aclweb.org/anthology/2020.acl-main.740.

Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(11):307–361, 2012. URL http://jmlr.org/papers/v13/gutmann12a.html.

Taru Haapala. "That in the opinion of this House": The parliamentary culture of debate in the nineteenth-century Cambridge and Oxford Union Societies. *Jyväskylä studies in education, psychology and social research*, (456), 2012. ISSN 0075-4625.

Ivan Habernal and Iryna Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1150. URL https://www.aclweb.org/anthology/P16-1150.

Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, 2017. doi: 10.1162/COLI_a_00276. URL https://www.aclweb.org/anthology/J17-1004.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation Mining on the Web from Information Seeking Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, Forlì-Cesena, Italy, 2014. URL http://ceur-ws.org/Vol-1341/paper4.pdf.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1175. URL https://www.aclweb.org/anthology/N18-1175.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1530. URL https://www.aclweb.org/anthology/D19-1530.

Charles L Hamblin. *Fallacies*. Methuen young books, London, UK, 1 edition, 1970. ISBN 978-0-416-14570-0.

Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1433. URL https://www.aclweb.org/anthology/D19-1433.

Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. ISSN 0043-7956, 2373-5112. doi: 10.1080/00437956.1954.11659520. URL https://doi.org/10.1080/00437956.1954.11659520.

Martie G. Haselton, Daniel Nettle, and Damian R. Murray. *The Evolution of Cognitive Bias*, chapter 41, pages 1–20. American Cancer Society, 2015. ISBN 9781119125563. doi: https://doi.org/10.1002/9781119125563.evpsych241. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119125563.evpsych241.

Thomas Hellström, Virginia Dignum, and Suna Bensch. Bias in Machine Learning – What is it Good for? 2020. URL https://arxiv.org/abs/2004.00686.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arxiv*, 2016. URL https://arxiv.org/pdf/1606.08415.pdf.

Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 58–68, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1006. URL https://www.aclweb.org/anthology/P14-1006.

Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015. doi: 10.1162/COLI_a_00237. URL https://www.aclweb.org/anthology/J15-4004.

Perry Hinton. Implicit stereotypes and the predictive brain: Cognition and culture in "biased" person perception. *Palgrave Communications*, 3(1):1–9, 2017. ISSN 2055-1045. doi: 10.1057/palcomms.2017.86. URL https://doi.org/10.1057/palcomms.2017.86.

J.E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0507655102. URL https://doi.org/10.1073/pnas.0507655102.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Colby Horn, Cathryn Manduca, and David Kauchak. Learning a lexical simplifier using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2075. URL https://www.aclweb.org/anthology/P14-2075.

Yedid Hoshen and Lior Wolf. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1043. URL https://www.aclweb.org/anthology/D18-1043.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799, Long Beach, CA, USA, 2019. PMLR. URL http://proceedings.mlr.press/v97/houlsby19a.html.

Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2079. URL https://www.aclweb.org/anthology/P15-2079.

Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL https://www.aclweb.org/anthology/P16-2096.

G. Hripcsak. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298, 2005. ISSN 1067-5027, 1527-974X. doi: 10.1197/jamia.m1733. URL https://doi.org/10.1197/jamia.m1733.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/hu20b.html.

Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. Overview of the CL-SciSumm 2016 shared task. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 93–102, 2016. URL https://www.aclweb.org/anthology/W16-1511.

Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R. Radev. NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, 23(1):93–130, 2016. ISSN 1351-3249, 1469-8110. doi: 10.1017/s1351324915000443. URL https://doi.org/10.1017/s1351324915000443.

Hwiyeol Jo and Stanley Jungkyu Choi. Extrofitting: Enriching word representation and its vector space with semantic lexicons. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 24–29, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-3003. URL https://www.aclweb.org/anthology/W18-3003.

Charles Jochim and Hinrich Schütze. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of COLING 2012*, pages 1343–1358, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL https://www.aclweb.org/anthology/C12-1082.

Charles Jochim and Hinrich Schütze. Improving citation polarity classification with product reviews. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–48, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2008. URL https://www.aclweb.org/anthology/P14-2008.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. doi: 10.1162/tacl_a_00065. URL https://www.aclweb.org/anthology/Q17-1024.

Ralph Henry Johnson and J Anthony Blair. *Logical Self-Defense*. International Debate Education Association, New York, NY, USA, 1 edition, 2006. ISBN 978-1932716184.

Hans Jonas. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. University of Chicago Press, 1 edition, 1984. ISBN 0-226-40597-4. Original in German: Prinzip Verantwortung.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.

acl-main.560. URL https://www.aclweb.org/anthology/2020.acl-main.560.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, 2020. OpenReview.net. URL https://openreview.net/forum?id=HJeT3yrtDr.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1062. URL https://www.aclweb.org/anthology/P14-1062.

Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 72–83, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4310. URL https://www.aclweb.org/anthology/W19-4310.

Martin Kay. Machine translation. *American Journal of Computational Linguistics*, 8(2): 74–78, 1982. URL https://www.aclweb.org/anthology/J82-2005.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7482–7491. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00781. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Kendall_Multi-Task_Learning_Using_CVPR_2018_paper.html.

George A. Kennedy. *A New History of Classical Rhetoric*. Princeton University Press, 2009. ISBN 9781400821471. doi: 10.1515/9781400821471. URL https://doi.org/10.1515/9781400821471.

Douwe Kiela, Felix Hill, and Stephen Clark. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1242. URL https://www.aclweb.org/anthology/D15-1242.

In Cheol Kim and George R. Thoma. Automated classification of author's sentiments in citation using machine learning techniques: A preliminary study. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2015. ISBN 9781479969265. doi: 10.1109/cibcb.2015.7300319. URL https://doi.org/10.1109/cibcb.2015.7300319.

Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. Intent detection using semantically enriched word embeddings. In *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016. ISBN 9781509049035. doi: 10.1109/slt.2016.7846297. URL https://doi.org/10.1109/slt.2016.7846297.

Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. Extracting domain-specific words - a statistical approach. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 94–98, Sydney, Australia, December 2009. URL https://www.aclweb.org/anthology/U09-1013.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://www.aclweb.org/anthology/D14-1181.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, San Diego, CA, USA, 2015. URL http://arxiv.org/abs/1412.6980.

Barbara Ann Kipfer. *Roget's 21st Century Thesaurus (3rd Edition)*. Bantam Doubleday Dell Publishing Group, 3 edition, 2005. ISBN 9780440242697.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073/pnas.1611835114.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO, 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0501. URL https://www.aclweb.org/anthology/W15-0501.

David Klahr and Kevin Dunbar. Dual space search during scientific reasoning. *Cognitive Science*, 12(1):1–48, 1988. ISSN 0364-0213. doi: 10.1207/s15516709cog1201_1. URL https://doi.org/10.1207/s15516709cog1201_1.

Petr Knoth and Drahomira Herrmannova. Towards semantometrics: A new semantic similarity based measure for assessing a research publication's contribution. *D-Lib Magazine*, 20(11/12), 2014. ISSN 1082-9873. doi: 10.1045/november2014-knoth. URL https://doi.org/10.1045/november2014-knoth.

Petr Knoth and Zdenek Zdrahal. CORE: Three access levels to underpin open access. *D-Lib Magazine*, 18(11/12), 2012. ISSN 1082-9873. doi: 10.1045/november2012-knoth. URL https://doi.org/10.1045/november2012-knoth.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. Technical report, University of Pennsylvania, Annenberg School for Communication, 2007. URL https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers.

Deanna Kuhn, John Black, Alla Keselman, and Danielle Kaplan. The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, 18(4):495–523, 2000. ISSN 0737-0008, 1532-690X. doi: 10.1207/s1532690xci1804_3. URL https://doi.org/10.1207/s1532690xci1804_3.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289, Williamstown, MA, USA, 2001. Morgan Kaufmann.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1549. URL https://www.aclweb.org/anthology/D18-1549.

Bruno Latour and Steve Woolgar. *Laboratory Life*. Princeton University Press, 1987. ISBN 9781400820412. doi: 10.1515/9781400820412. URL https://doi.org/10.1515/9781400820412.

John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, 2019. doi: 10.1162/coli_a_00364. URL https://www.aclweb.org/anthology/J19-4006.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. In Michael A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–258. MIT Press, Cambridge, MA, USA, 1998. ISBN 0-262-51102-9.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015. ISSN 1570-0844. doi: 10.3233/sw-140134. URL https://doi.org/10.3233/sw-140134.

Ira Leviant and Roi Reichart. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*, 2015. URL https://arxiv.org/pdf/1508.00106.pdf.

Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia,

2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2005. URL https://www.aclweb.org/anthology/P18-2005.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2054–2061, Valletta, Malta, 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/644_Paper.pdf.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Methods of Biochemical Analysis*, 28(7):991–1000, 2012. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/bts071. URL https://doi.org/10.1093/bioinformatics/bts071.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.150. URL https://www.aclweb.org/anthology/2020.findings-emnlp.150.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1301. URL https://www.aclweb.org/anthology/P19-1301.

Marco Lippi and Paolo Torroni. Argument mining: A machine learning perspective. In *Theory and Applications of Formal Argumentation*, pages 163–176. Springer International Publishing, 2015. ISBN 9783319284590, 9783319284606. doi: 10.1007/978-3-319-28460-6_10. URL https://doi.org/10.1007/978-3-319-28460-6_10.

Marco Lippi and Paolo Torroni. MARGOT: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303, 2016a. ISSN 0957-4174. doi: 10.1016/j.eswa.2016.08.050. URL https://doi.org/10.1016/j.eswa.2016.08.050.

Marco Lippi and Paolo Torroni. Argumentation mining. *ACM Transactions on Internet Technology*, 16(2):1–25, 2016b. ISSN 1533-5399, 1557-6051. doi: 10.1145/2850417. URL https://doi.org/10.1145/2850417.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter*

of the Association for Computational Linguistics: Volume 2, Short Papers, pages 8–14, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2002.

H Liu and P Singh. ConceptNet – a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004. ISSN 1358-3948. doi: 10.1023/b:bttj.0000047600.45421.6d. URL https://doi.org/10.1023/b:bttj.0000047600.45421.6d.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1501–1511, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1145. URL https://www.aclweb.org/anthology/P15-1145.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 2901–2908, New York, NY, USA, 2020. AAAI Press. URL https://aaai.org/ojs/index.php/AAAI/article/view/5681.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*, 2019. URL https://arxiv.org/pdf/1907.11692.pdf.

Richard Lockwood. *The reader's figure: epideictic rhetoric in Plato, Aristotle, Bossuet, Racine and Pascal*, volume 351. Librairie Droz, 1 edition, 1996. ISBN 9782600001403.

Nicole Loraux. *The children of Athena: Athenian ideas about citizenship and the division between the sexes*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1994. ISBN 9780691037622.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer International Publishing, 2020. ISBN 9783030620769, 9783030620776. doi: 10.1007/978-3-030-62077-6_14. URL https://doi.org/10.1007/978-3-030-62077-6_14.

H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957. doi: 10.1147/rd.14.0309. URL https://doi.org/10.1147/rd.14.0309.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain, 2017. Association

for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1070.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1062. URL https://www.aclweb.org/anthology/N19-1062.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.68.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2001. URL https://www.aclweb.org/anthology/S14-2001.

Héctor Martínez Alonso and Barbara Plank. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 44–53, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1005.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://www.aclweb.org/anthology/N19-1063.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1269. URL https://www.aclweb.org/anthology/D17-1269.

Katherine McCurdy and Oguz Serbetci. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. In *Proceedings of the first WiNLP Workshop*, Vancouver, Canada, 2017. URL http://www.winlp.org/wp-content/uploads/2017/final_papers_2017/46_Paper.pdf.

Kathy McKeown, Hal Daume, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R. Fleischmann, Luis Gravano, Rahul Jha, Ben King, Kevin McInerney, Taesun Moon, Arvind Neelakantan, Diarmuid O'Seaghdha, Dragomir Radev, Clay Templeton, and Simone Teufel. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, 67(11):2684–2696, 2016. ISSN 2330-1635. doi: 10.1002/asi.23612. URL https://doi.org/10.1002/asi.23612.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019. URL https://arxiv.org/pdf/1908.09635.pdf.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR: Proceeding of the International Conference on Learning Representations Workshop Track*, page 1301–3781, Arizona, USA, 2013a. URL https://arxiv.org/pdf/1301.3781.pdf.

Tomáš Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR, abs/1309.4168*, 2013b. URL https://arxiv.org/pdf/1309.4168.pdf.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 3111–3119, Lake Tahoe, Nevada, USA, 2013c. URL https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. ISSN 0001-0782, 1557-7317. doi: 10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939, Las Vegas, NV, USA, 2016. IEEE. doi: 10.1109/CVPR.2016.320. URL https://ieeexplore.ieee.org/document/7780689.

Tom H. Mitchell. *The Need for Biases in Learning Generalizations*, chapter Special topics, pages 184–191. Cambridge University Press, 1980. ISBN 9780511811692. doi: 10.1017/cbo9780511811692.015. URL https://doi.org/10.1017/cbo9780511811692.015. Book published in 1990.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. ISBN 978-0-07-042807-2.

Marie-Francine Moens. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument & Computation*, 9(1):1–14, 2018. ISSN 1946-2174, 1946-2166. doi: 10.3233/aac-170025. URL https://doi.org/10.3233/aac-170025.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law - ICAIL '07*, ICAIL '07, pages 225–230. ACM Press, 2007. ISBN 9781595936806. doi: 10.1145/1276318.1276362. URL https://doi.org/10.1145/1276318.1276362.

Gaku Morio and Katsuhide Fujita. On the role of syntactic graph convolutions for identifying and classifying argument components. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 9997–9998, Honolulu, Hawaii, USA, 2019. AAAI Press. doi: 10.1609/aaai.v33i01.33019997. URL https://doi.org/10.1609/aaai.v33i01.33019997.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244, 2014. doi: 10.1162/tacl_a_00179. URL https://www.aclweb.org/anthology/Q14-1019.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1018. URL https://www.aclweb.org/anthology/N16-1018.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324, 2017. doi: 10.1162/tacl_a_00063. URL https://www.aclweb.org/anthology/Q17-1022.

Preslav I. Nakov, Ariel S Schwartz, and Marti Hearst. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on search and discovery in bioinformatics*, volume 4, pages 81–88. ACM Press, 2004. URL https://biotext.berkeley.edu/papers/citances-nlpbio04.pdf.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P10-1023.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. ISSN 0004-3702. doi: 10.1016/j.artint.2012.07.001. URL https://doi.org/10.1016/j.artint.2012.07.001.

Douglas L. Nelson, Valerie S. Reed, and John R. Walling. Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory*, 2(5):523–528, 1976. ISSN 0096-1515. doi: 10.1037/0278-7393.2.5.523. URL https://doi.org/10.1037/0278-7393.2.5.523.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2074. URL https://www.aclweb.org/anthology/P16-2074.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1022. URL https://www.aclweb.org/anthology/D17-1022.

Vlad Niculae, Joonsuk Park, and Claire Cardie. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1091. URL https://www.aclweb.org/anthology/P17-1091.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-5001.

Brian A. Nosek, Mahzarin R. Banaji, and Anthony G. Greenwald. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101–115, 2002. ISSN 1930-7802, 1089-2699. doi: 10.1037/1089-2699.6.1.101. URL https://doi.org/10.1037/1089-2699.6.1.101.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1356, 2020. ISSN 1942-4787,

1942-4795. doi: 10.1002/widm.1356. URL https://doi.org/10.1002/widm.1356.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019. ISSN 2624-909X. doi: 10.3389/fdata.2019.00013. URL https://doi.org/10.3389/fdata.2019.00013.

Dominique Osborne, Shashi Narayan, and Shay B. Cohen. Encoding prior knowledge with eigenword embeddings. *Transactions of the Association for Computational Linguistics*, 4:417–430, 2016. doi: 10.1162/tacl_a_00108. URL https://www.aclweb.org/anthology/Q16-1030.

Jonathan Osborne. Arguing to learn in science: The role of collaborative, critical discourse. *Science*, 328(5977):463–466, 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1183944. URL https://doi.org/10.1126/science.1183944.

Gustavo Paetzold and Lucia Specia. Benchmarking lexical simplification systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3074–3080, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1491.

Gustavo H. Paetzold and Lucia Specia. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593, 2017. ISSN 1076-9757. doi: 10.1613/jair.5526. URL https://doi.org/10.1613/jair.5526.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999. URL http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law - ICAIL '09*, pages 98–107. ACM, ACM Press, 2009. ISBN 9781605585970. doi: 10.1145/1568234.1568246. URL https://doi.org/10.1145/1568234.1568246.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. ISSN 1041-4347. doi: 10.1109/tkde.2009.191. URL https://doi.org/10.1109/tkde.2009.191.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1302. URL https://www.aclweb.org/anthology/D18-1302.

Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural*

Intelligence (IJCINI), 7(1):1–31, 2013. doi: 10.4018/jcini.2013010101. URL https://doi.org/10.4018/jcini.2013010101.

Andreas Peldszus and Manfred Stede. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1110. URL https://www.aclweb.org/anthology/D15-1110.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

Chaïm Perelman, Lucie Olbrechts-Tyteca, John Wilkinson, and Purcell Weaver. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press, online edition, 1969. doi: 10.2307/j.ctvpj74xx. URL https://doi.org/10.2307/j.ctvpj74xx.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710. ACM, 2014. ISBN 9781450329569. doi: 10.1145/2623330.2623732. URL https://doi.org/10.1145/2623330.2623732.

Isaac Persing and Vincent Ng. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-1026.

Isaac Persing and Vincent Ng. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1144. URL https://www.aclweb.org/anthology/P14-1144.

Isaac Persing and Vincent Ng. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1053. URL https://www.aclweb.org/anthology/P15-1053.

Isaac Persing and Vincent Ng. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1164. URL https://www.aclweb.org/anthology/N16-1164.

Isaac Persing and Vincent Ng. Why can't you convince me? Modeling weaknesses in unpersuasive arguments. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4082–4088. ijcai.org, 2017. doi: 10.24963/ijcai.2017/570. URL https://doi.org/10.24963/ijcai.2017/570.

Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA, 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D10-1023.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1005. URL https://www.aclweb.org/anthology/D19-1005.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy, 2019b. Association for Computational Linguistics. doi: 10.18653/v1/W19-4302. URL https://www.aclweb.org/anthology/W19-4302.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://www.aclweb.org/anthology/D19-1250.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://www.aclweb.org/anthology/2020.emnlp-main.617.

Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *CoRR*, abs/1811.01088, 2018. URL https://arxiv.org/pdf/1811.01088.pdf.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://www.aclweb.org/anthology/P19-1493.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1026. URL https://www.aclweb.org/anthology/D18-1026.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601, 2019a. doi: 10.1162/coli_a_00357. URL https://www.aclweb.org/anthology/J19-3005.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2206–2217, Hong Kong, China, 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1226. URL https://www.aclweb.org/anthology/D19-1226.

Karl Raimund Popper. *Logik Der Forschung*. JCB Mohr Tübingen, Tübingen, Germany, 1935, edition 2002. ISBN 978-3161478376.

Peter Potash, Alexey Romanov, and Anna Rumshisky. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1364–1373, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1143. URL https://www.aclweb.org/anthology/D17-1143.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.467. URL https://www.aclweb.org/anthology/2020.acl-main.467.

Jipeng Qiang, Yun Li, Zhu Yi, Yunhao Yuan, and Xindong Wu. Lexical simplification with pretrained encoders. *Thirty-Fourth AAAI Conference on Artificial Intelligence*, page 8649–8656, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6389/6245.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. *OpenAI Technical Report*, 2018. URL https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019. URL http://www.persagen.com/files/misc/radford2019language.pdf.

Jonathan Raiman and Olivier Raiman. Deeptype: Multilingual entity linking by neural type system evolution. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 5406–5413, New Orleans, Louisiana, USA, 2018. AAAI Press. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17148.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://www.aclweb.org/anthology/D16-1264.

Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.603. URL https://www.aclweb.org/anthology/2020.coling-main.603.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8119–8127. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00847. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Rebuffi_Efficient_Parametrization_of_CVPR_2018_paper.html.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1299. URL https://www.aclweb.org/anthology/D18-1299.

Petar Ristoski and Heiko Paulheim. RDF2Vec: RDF graph embeddings for data mining. In *International Semantic Web Conference*, pages 498–514. Springer, 2016. URL https://www.springerprofessional.de/rdf2vec-rdf-graph-embeddings-for-data-mining/10816474.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00349. URL https://doi.org/10.1162/tacl_a_00349.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland, 2019. Linköping University Electronic Press. URL https://www.aclweb.org/anthology/W19-6204.

Francesco Ronzano and Horacio Saggion. Dr. Inventor Framework: Extracting Structured Information from Scientific Publications. In *Discovery Science*, Lecture Notes in Computer Science, pages 209–220. Springer, Cham, 2015. ISBN 978-3-319-24281-1 978-3-319-24282-8. doi: 10.1007/978-3-319-24282-8_18.

Francesco Ronzano and Horacio Saggion. Knowledge Extraction and Modeling from Scientific Publications. In *Semantics, Analytics, Visualization. Enhancing Scholarly Data*, Lecture Notes in Computer Science, pages 11–25. Springer, Cham, 2016. ISBN 978-3-319-53636-1 978-3-319-53637-8. doi: 10.1007/978-3-319-53637-8_2.

Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway, 2019.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640. URL https://doi.org/10.1613/jair.1.11640.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL https://www.aclweb.org/anthology/N18-2002.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. Crowdsourcing research opportunities. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '12*, pages 1–8. ACM Press, 2012. ISBN 9781450312424. doi: 10.1145/2362456.2362479. URL https://doi.org/10.1145/2362456.2362479.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1156. URL https://www.aclweb.org/anthology/P18-1156.

Marina Santini. State-of-the-art on automatic genre identification. *Information Technology Research Institute Technical Report Series*, 04, 2004. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.7680&rep=rep1&type=pdf.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2006. URL https://www.aclweb.org/anthology/N18-2006.

Roy Schwartz, Roi Reichart, and Ari Rappoport. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 258–267, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-1026. URL https://www.aclweb.org/anthology/K15-1026.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. ISSN 0033-3123, 1860-0980. doi: 10.1007/bf02289451. URL https://doi.org/10.1007/bf02289451.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://www.aclweb.org/anthology/P19-1282.

Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto, editors, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 959–962. ACM, 2015. doi: 10.1145/2766462.2767830. URL https://doi.org/10.1145/2766462.2767830.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2106.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In Robert Meersman and Zahir Tari, editors, *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36124-4.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=r1Aab85gg.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Anders Søgaard and Yoav Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2038. URL https://www.aclweb.org/anthology/P16-2038.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL https://www.aclweb.org/anthology/P18-1072.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. ISSN 0022-0418. doi: 10.1108/eb026526. URL https://doi.org/10.1108/eb026526.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.

Maximilian Spliethöver and Henning Wachsmuth. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.argmining-1.9.

Christian Stab and Iryna Gurevych. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2813. URL https://www.aclweb.org/anthology/W16-2813.

Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017a. doi: 10.1162/COLI_a_00295. URL https://www.aclweb.org/anthology/J17-3005.

Christian Stab and Iryna Gurevych. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain, 2017b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1092.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 21–25, Forlì-Cesena, Italy, 2014. CEUR Workshop Proceedings. URL http://ceur-ws.org/Vol-1341/paper5.pdf.

Manfred Stede and Jodi Schneider. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191, 2018. ISSN 1947-4040, 1947-4059. doi: 10.2200/s00883ed1v01y201811hlt040. URL https://doi.org/10.2200/s00883ed1v01y201811hlt040.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(null): 583–617, March 2003. ISSN 1532-4435. doi: 10.1162/153244303321897735. URL https://doi.org/10.1162/153244303321897735.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://www.aclweb.org/anthology/P19-1355.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*

- *WWW '07*, pages 697–706. ACM, ACM Press, 2007. ISBN 9781595936547. doi: 10.1145/1242572.1242667. URL https://doi.org/10.1145/1242572.1242667.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2058–2065, Phoenix, Arizona, USA, 2016. AAAI Press. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12443.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 2.0: A continual pre-training framework for language understanding. *AAAI*, 34(05):8968–8975, 2020. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i05.6428. URL https://doi.org/10.1609/aaai.v34i05.6428.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019. URL https://arxiv.org/pdf/1901.10002.pdf.

John M Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge, England, 13 edition, 1990/ edition 2008. ISBN 978-0-521-32869-2.

Wilson L. Taylor. "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433, 1953. ISSN 0022-5533. doi: 10.1177/107769905303000401. URL https://doi.org/10.1177/107769905303000401.

Simone Teufel. Scientific Argumentation Detection as Limited-domain Intention Recognition. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, Forlì-Cesena, Italy, 2014. CEUR Workshop Proceedings. URL http://ceur-ws.org/Vol-1341/paper14.pdf.

Simone Teufel and Marc Moens. Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting. In *Advances in Automatic Text Summarization*, pages 155–171. MIT Press, 1999a. URL https://www.cl.cam.ac.uk/~sht25/papers/aits.pdf.

Simone Teufel and Marc Moens. Discourse-level argumentation in scientific articles: Human and automatic annotation. In *Towards Standards and Tools for Discourse Tagging, Workshop*, Maryland, MA, USA, 1999b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W99-0311.

Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002. ISSN 0891-2017, 1530-9312. doi: 10.1162/089120102762671936. URL https://doi.org/10.1162/089120102762671936.

Simone Teufel, Jean Carletta, and Marc Moens. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway, 1999. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E99-1015.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia, 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W06-1613.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards discipline-independent argumentative zoning. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 3 - EMNLP '09*, pages 1493–1502. Association for Computational Linguistics, 2009. ISBN 9781932432633. doi: 10.3115/1699648.1699696. URL https://doi.org/10.3115/1699648.1699696.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. Technical Report 2, 2012. URL http://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf.

Christopher W. Tindale. *Fallacies and Argument Appraisal*. Critical Reasoning and Argumentation. Cambridge University Press, Cambridge, UK, 2007. ISBN 9780511806544. doi: 10.1017/cbo9780511806544. URL https://doi.org/10.1017/cbo9780511806544.

Julien Tissier, Christophe Gravier, and Amaury Habrard. Dict2vec : Learning word embeddings using lexical dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1024. URL https://www.aclweb.org/anthology/D17-1024.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1564. URL https://www.aclweb.org/anthology/D19-1564.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.29. URL https://www.aclweb.org/anthology/2020.findings-emnlp.29.

Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, Cambridge, UK, 2 edition, 1958, 2003 edition. ISBN 9780521827485, 9780521534833, 9780511840005. doi: 10.1017/cbo9780511840005. URL https://doi.org/10.1017/cbo9780511840005.

Amos Tversky and Daniel Kahneman. Belief in the law of small numbers. *Psychological Bulletin Journal*, 76(2):105–110, 1971. ISSN 1939-1455, 0033-2909. doi: 10.1037/h0031322. URL https://doi.org/10.1037/h0031322.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.185.4157.1124. URL https://doi.org/10.1126/science.185.4157.1124.

Anahuac Valero Haro, Omid Noroozi, Harm Biemans, and Martin Mulder. Argumentation competence: Students' argumentation knowledge, behavior and attitude and their relationships with domain-specific knowledge acquisition. *Journal of Constructivist Psychology*, 0(0):1–23, 2020. ISSN 1072-0537, 1521-0650. doi: 10.1080/10720537.2020.1734995. URL https://doi.org/10.1080/10720537.2020.1734995.

Marlies van der Wees, Arianna Bisazza, Wouter Weerkamp, and Christof Monz. What's in a domain? Analyzing genre and topic differences in statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 560–566, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2092. URL https://www.aclweb.org/anthology/P15-2092.

Frans H. van Eemeren and Rob Grootendorst. In *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*, Cambridge, UK, 2010. Cambridge University Press. ISBN 978-0-511-61638-9. URL https://doi.org/10.1017/CBO9780511616389.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008, 2017. URL https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Bart Verheij. The toulmin argument model in artificial intelligence. In Guillermo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 219–238. Springer US, Boston, MA, 2009. ISBN 9780387981963, 9780387981970. doi: 10.1007/978-0-387-98197-0_11. URL https://doi.org/10.1007/978-0-387-98197-0_11.

Jean-Pierre Vernant. Espace et organisation politique en Grèce ancienne. In *Annales. Histoire, Sciences Sociales*, volume 20, pages 576–595, Cambridge, UK, 1965. Cambridge University Press. URL https://www.persee.fr/doc/ahess_0395-2649_1965_num_20_3_421305.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commununications of the ACM*, 57(10):78–85, 2014. ISSN 0001-0782, 1557-7317. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

Ivan Vulić. Injecting lexical contrast into word vectors by guiding vector space specialisation. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 137–143, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-3018. URL https://www.aclweb.org/anthology/W18-3018.

Ivan Vulić and Nikola Mrkšić. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1103. URL https://www.aclweb.org/anthology/N18-1103.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 516–527, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1048. URL https://www.aclweb.org/anthology/N18-1048.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1449. URL https://www.aclweb.org/anthology/D19-1449.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.257. URL https://www.aclweb.org/anthology/2020.emnlp-main.257.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. Sentiment flow – a general model of web review argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 601–611, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1072. URL https://www.aclweb.org/anthology/D15-1072.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1158.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada, 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-2039. URL https://www.aclweb.org/anthology/P17-2039.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain, 2017b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1017.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark, 2017c. Association for Computational Linguistics. doi: 10.18653/v1/W17-5106. URL https://www.aclweb.org/anthology/W17-5106.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain, 2017d. Association for Computational Linguistics. doi: 10.18653/V1/E17-1105. URL https://www.aclweb.org/anthology/E17-1105.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 454–463, Oxford, UK, 2015. AAAI Press. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/download/10585/10528.

Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 1. edition, 2008. doi: 10.1017/CBO9780511802034. URL https://doi.org/10.1017/CBO9780511802034.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language

understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020. URL https://arxiv.org/pdf/2002.01808.pdf.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290. URL https://www.aclweb.org/anthology/Q19-1040.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. Measuring and Reducing Gendered Correlations in Pre-trained Models. *arXiv preprint arXiv:2010.06032*, 2020. URL https://arxiv.org/pdf/2010.06032.pdf.

Mark Weinstein. Towards an account of argumentation in science. *Argumentation*, 4(3):269–298, 1990. ISSN 0920-427X, 1572-8374. doi: 10.1007/bf00173968. URL https://doi.org/10.1007/bf00173968.

Joseph Weizenbaum. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman & Co., USA, 1. edition, 1976. ISBN 0-7167-0464-1.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.494.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358, 2015. doi: 10.1162/tacl_a_00143. URL https://www.aclweb.org/anthology/Q15-1025.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://www.aclweb.org/anthology/N18-1101.

Michael Wojatzki and Torsten Zesch. Stance-based argument Mining–Modeling implicit argumentation using stance. *Proceedings of the KONVENS, Bochum, Germany*,

pages 313–322, 2016. doi: 10.17185/duepublico/46445. URL https://doi.org/10.17185/duepublico/46445.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771, 2019.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077. URL https://www.aclweb.org/anthology/D19-1077.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. URL https://arxiv.org/pdf/1609.08144.pdf.

Han Xu, Eric Martin, and Ashesh Mahidadia. Using heterogeneous features for scientific citation classification. In *Proceedings of the 13th Conference of the Pacific Association for Computational Linguistics*, 2013. URL https://www.researchgate.net/publication/255673224_Using_Heterogeneous_Features_for_Scientific_Citation_Classification.

Yongxin Yang and Timothy M Hospedales. Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038*, 2016. URL https://arxiv.org/pdf/1606.04038.pdf.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 5754–5764, Vancouver, BC, Canada, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1174. URL https://www.aclweb.org/anthology/N16-1174.

Alexander Yeh. More accurate tests for the statistical significance of result differences. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, 2000. URL https://www.aclweb.org/anthology/C00-2137.

Mo Yu and Mark Dredze. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 545–550, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2089. URL https://www.aclweb.org/anthology/P14-2089.

Jingwei Zhang, Jeremy Salwen, Michael Glass, and Alfio Gliozzo. Word semantic representations using Bayesian probabilistic tensor factorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1522–1531, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1161. URL https://www.aclweb.org/anthology/D14-1161.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1139. URL https://www.aclweb.org/anthology/P19-1139.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://www.aclweb.org/anthology/D17-1323.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://www.aclweb.org/anthology/N18-2003.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL https://www.aclweb.org/anthology/D18-1521.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634,

Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1064. URL https://www.aclweb.org/anthology/N19-1064.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.151. URL https://www.aclweb.org/anthology/2020.acl-main.151.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.11. URL https://doi.org/10.1109/ICCV.2015.11.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://www.aclweb.org/anthology/P19-1161.

Appendix A

Published Resources

In the Table A.1, we provide an overview of the resources published with this thesis.

Chapter	Resource Name	Type	Location	Challenges
4.1	LIBERT	Code	https://github.com/anlausch/LIBERT	C1
4.2	CN-Adapt	Code	https://github.com/Wluper/Retrograph	C1
5	Scientific Embeddings	Model	https://github.com/anlausch/scientific-domain-embeddings	C2
6.1	Argument Annotations	Corpus	http://data.dws.informatik.uni-mannheim.de/sci-arg/compiled_corpus.zip	C3, C2
6.1	Argument Analysis	Code	https://github.com/anlausch/multitask_sciarg	C3, C2
6.1	MT for Scitorics	Code	https://github.com/anlausch/sciarg_resource_analysis	C3, C2
6.2	GAQCorpus	Corpus	https://github.com/grammarly/gaqcorpus	C3, C2
7	Zero2Hero	Code	https://github.com/anlausch/CLZeroShotTransferLimitations	C4
8.1	XWEAT	Corpus	https://github.com/anlausch/XWEAT	C4, C5
8.2	DebIE	Code	https://github.com/anlausch/DEBIE	C5

Table A.1: Overview of all resources published in the context of this thesis.

Appendix B

Experimental Details for Section 6.2

B.1 Hyperparameter Optimization

Model Type	Hyperparameter	Values
SVR	Regularization c	$0.001, 0.01, 0.1, 1.0, 10$
	Epsilon-tube specifier ϵ	$0.001, 0.01, 0.1, 1.0$
BERT	Learning rate λ	$2 \cdot 10^{-5}, 3 \cdot 10^{-5}$
	Number of epochs	3,4

Table B.1: Search values per model type and hyperparameter employed in the experiments.

For each experiment, we conducted a grid search on the corresponding development portion of the employed training set. The search spaces are depicted in Table B.1.

B.2 Full Experimental Results for RQ2–RQ4

We list the full experimental results on GAQCorpus with respect to RQ2–RQ4.

		CQA forums			Debate Forums			Review Forums		
	Model	**Crowd**	**Expert**	**Mix**	**Crowd**	**Expert**	**Mix**	**Crowd**	**Expert**	**Mix**
Overall	ARG LENGTH	0.498	0.236	0.406	0.542	0.232	0.420	0.486	0.190	0.365
	SVR_{TFIDF}	0.381	0.323	0.389	0.299	0.179	0.265	0.446	0.340	0.450
	SVR_{EMBD}	0.323	0.180	0.278	0.467	0.239	0.388	0.223	0.227	0.265
	$WACHSMUTH_{CFS}$	0.550	0.340	0.492	0.524	0.264	0.432	**0.619**	0.342	0.533
	BERT ST	**0.681**	**0.498**	**0.652**	**0.575**	**0.346**	**0.511**	0.611	**0.450**	**0.605**
Cogency	ARG LENGTH	0.502	0.227	0.420	**0.574**	0.225	0.437	0.491	0.125	0.340
	SVR_{TFIDF}	0.449	0.330	0.444	0.295	0.164	0.257	0.409	0.264	0.384
	SVR_{EMBD}	0.301	0.154	0.261	0.404	0.196	0.333	0.264	-0.059	0.103
	$WACHSMUTH_{CFS}$	0.565	0.311	0.503	0.548	0.232	0.429	0.611	0.223	0.464
	BERT ST	**0.623**	**0.405**	**0.587**	0.556	**0.337**	**0.503**	**0.618**	**0.359**	**0.554**
Effectiveness	ARG LENGTH	0.475	0.237	0.390	0.502	0.225	0.399	0.425	0.251	0.372
	SVR_{TFIDF}	0.432	0.313	0.411	0.141	0.074	0.120	0.354	0.253	0.340
	SVR_{EMBD}	0.328	0.204	0.293	0.456	0.264	0.403	0.186	0.144	0.187
	$WACHSMUTH_{CFS}$	0.555	0.393	0.523	0.528	0.281	0.450	0.567	0.246	0.432
	BERT ST	**0.596**	**0.509**	**0.612**	**0.548**	**0.405**	**0.542**	**0.639**	**0.370**	**0.555**
Reasonableness	ARG LENGTH	0.480	0.245	0.396	0.535	0.170	0.377	0.496	0.241	0.405
	SVR_{TFIDF}	0.466	0.364	0.457	0.292	0.153	0.247	0.435	0.345	0.452
	SVR_{EMBD}	0.411	0.278	0.379	0.393	0.096	0.258	0.205	0.191	0.234
	$WACHSMUTH_{CFS}$	0.543	0.326	0.476	**.549**	0.192	0.399	0.524	0.261	0.432
	BERT ST	**0.696**	**0.512**	**0.665**	0.544	**0.222**	**0.418**	**0.556**	**0.484**	**0.609**

Table B.2: Pearson correlations of our model predictions with the annotation scores for the four AQ dimensions on the three different test annotations (Crowd, Expert, Mix) when training on in-domain data. Numbers in bold indicate best performances.

		CQA forums			Debate Forums			Review Forums		
	Model	**Crowd**	**Expert**	**Mix**	**Crowd**	**Expert**	**Mix**	**Crowd**	**Expert**	**Mix**
Overall	BERT ST	**0.681**	0.498	0.652	0.575	0.346	0.511	**0.611**	0.450	**0.605**
	BERT MT_{FLAT}	0.671	**0.535**	**0.667**	**0.607**	**0.362**	**0.537**	0.534	**0.478**	0.588
	BERT MT_{HIER}	0.668	0.528	0.661	0.480	0.393	0.494	0.563	0.465	0.593
Cogency	BERT ST	0.623	0.405	0.587	0.556	0.337	0.503	**0.618**	0.359	0.554
	BERT MT_{FLAT}	**0.651**	0.457	0.633	**0.622**	0.343	**0.541**	0.533	**0.440**	**0.561**
	BERT MT_{HIER}	0.650	**0.468**	**0.638**	0.476	**0.353**	0.474	0.559	0.388	0.541
Effectiveness	BERT ST	0.596	0.509	0.612	0.548	0.405	0.542	**0.639**	0.370	**0.555**
	BERT MT_{FLAT}	**0.663**	0.549	**0.671**	**0.599**	0.408	**0.570**	0.522	**0.389**	0.514
	BERT MT_{HIER}	0.656	**0.552**	0.670	0.477	**0.443**	0.532	0.466	0.388	0.486
Reasonableness	BERT ST	**0.696**	**0.512**	**0.665**	0.544	0.222	0.418	**0.556**	0.484	0.609
	BERT MT_{FLAT}	0.672	0.499	0.644	**0.587**	0.273	**0.473**	0.550	**0.489**	0.610
	BERT MT_{HIER}	0.660	0.478	0.626	0.445	**0.280**	0.408	0.555	0.488	**0.611**

Table B.3: Pearson correlations of our model predictions with the annotation scores. We compare single-task versus multi-task learning setups training on in-domain data only.

		CQA forums			Debate Forums			Review Forums		
	Model	**Crowd**	**Expert**	**Mix**	**Crowd**	**Expert**	**Mix**	**Crowd**	**Expert**	**Mix**
Overall	Best in-domain	0.681	**0.535**	0.667	**0.607**	0.362	0.537	0.619	**0.478**	0.605
	BERT ST	0.693	0.530	0.676	0.571	0.401	0.545	0.650	0.409	0.596
	BERT MT_{FLAT}	**0.697**	**0.535**	**0.681**	0.574	**0.425**	**0.562**	**0.678**	0.443	**0.633**
	BERT MT_{HIER}	0.680	0.522	0.665	0.576	0.424	**0.562**	0.618	0.469	0.622
Cogency	Best in-domain	0.651	0.468	0.638	**0.622**	0.353	0.541	**0.618**	0.440	0.561
	BERT ST	0.639	0.426	0.608	0.540	0.367	0.515	0.601	0.386	0.563
	BERT MT_{FLAT}	**0.673**	**0.472**	**0.653**	0.560	0.392	0.542	0.610	0.391	0.570
	BERT MT_{HIER}	0.662	0.455	0.638	0.573	**0.397**	**0.552**	0.577	**0.465**	**0.599**
Effectiveness	Best in-domain	0.656	0.552	0.671	**0.599**	0.443	0.570	0.639	0.389	0.555
	BERT ST	0.664	**0.574**	**0.686**	0.544	**0.492**	**0.598**	**0.711**	0.387	0.601
	BERT MT_{FLAT}	**0.676**	0.536	0.670	0.569	0.444	0.578	0.683	**0.409**	**0.603**
	BERT MT_{HIER}	0.657	0.523	0.653	0.573	0.462	0.592	0.644	0.396	0.576
Reasonableness	Best in-domain	**0.696**	**0.512**	**0.665**	**0.587**	0.280	0.473	0.556	**0.489**	0.611
	BERT ST	0.658	0.495	0.635	0.550	0.320	**0.487**	0.616	0.437	0.603
	BERT MT_{FLAT}	0.691	0.503	0.657	0.538	**0.328**	0.486	**0.667**	0.443	0.631
	BERT MT_{HIER}	0.665	0.485	0.633	0.554	0.312	0.483	0.642	0.476	**0.643**

Table B.4: Pearson correlations with the annotation scores when training on the joint training sets of all domains. We compare with the best result of the in-domain setting.

Appendix C

Experimental Details for Chapter 7

C.1 Reproducibility

We first provide details on where to obtain datasets and code used in this work.

Codebase	MMT	**Vocab**	**Params**	**URL**
HF Trans.	–	–	–	`https://github.com/huggingface/transformers`
	MBERT	119K	125M	`https://huggingface.co/bert-base-multilingual-cased`
	XLM-R	250K	125M	`https://huggingface.co/xlm-roberta-base`

Table C.1: Links to codebases and pretrained models used in this work. We built our models directly on top of the HuggingFace (HF) Transformers library.

Task	**Dataset**	**URL**
Natural Language Inference	XNLI	`https://github.com/facebookresearch/XNLI`
Question Answering	XQuAD	`https://github.com/deepmind/xquad`

Table C.2: Links to the datasets used in our work.

Code and Dependencies. Our code directly builds on top of the HuggingFace Transformers framework (Wolf et al., 2019). We provide links to all code dependencies and to the pretrained models we used in Table C.1.

Datasets. Table C.2 provides links to all datasets that we used in our study.

C.2 Full Per-Language Few-Shot Results

We show full per-language few-shot transfer results for MBERT and XLM-R in Tables C.3 and C.4, respectively.

XNLI	fr	es	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	de
0	75.05	74.71	68.68	69.50	69.34	62.18	65.53	70.88	54.69	69.26	61.50	49.84	59.38	72.34
10	75.09	73.62	67.04	69.35	69.80	61.86	65.56	69.26	55.30	70.89	61.92	51.79	59.28	71.63
50	74.60	73.91	66.44	68.37	69.05	60.99	64.63	70.29	51.17	71.32	60.08	49.95	58.83	71.43
100	73.85	73.50	65.67	68.47	70.24	60.13	64.93	69.59	51.68	71.46	60.01	48.96	58.78	71.60
500	75.36	74.97	68.04	71.03	70.59	63.21	66.71	72.38	58.12	72.81	64.06	52.26	61.15	73.09
1000	76.20	76.24	68.73	71.73	71.41	65.01	67.04	72.35	59.19	73.47	64.75	52.47	62.38	73.21

XQuAD	zh	vi	tr	th	ru	hi	es	el	de	ar
0	48.14	49.02	36.90	27.84	51.86	42.47	54.48	42.90	56.22	46.40
2	48.93	50.50	40.87	39.43	51.07	44.19	56.14	46.46	56.66	46.99
4	49.72	51.38	40.22	41.24	51.33	45.90	56.62	47.25	56.38	46.57
6	50.81	50.81	41.59	44.04	51.20	46.81	57.14	47.16	56.40	47.45
8	51.53	51.29	41.99	45.28	51.29	47.10	57.45	47.95	57.07	48.21
10	50.87	51.57	42.55	46.05	52.05	48.06	57.03	48.60	57.29	47.82

Table C.3: Detailed per-language few-shot language results for XNLI and XQuAD with MBERT for different number of target-language data instances k.

XNLI	fr	es	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	de
0	78.16	78.44	75.39	77.68	75.25	72.99	71.28	74.59	72.00	73.21	70.02	64.03	66.93	76.45
10	77.96	78.67	75.77	78.11	76.32	73.31	71.75	75.17	73.18	74.53	69.23	64.09	68.32	77.32
50	78.69	79.81	76.13	77.57	76.16	73.96	71.20	75.01	71.74	74.47	69.84	61.98	68.06	77.60
100	79.37	78.87	76.28	77.58	77.42	73.31	71.40	74.83	71.94	74.10	70.54	61.55	67.63	77.84
200	79.29	79.84	77.01	78.94	77.54	74.81	73.22	76.52	73.91	76.37	71.54	64.00	68.98	78.42
500	79.65	79.95	77.34	79.09	77.78	74.08	73.6	77.22	74.32	77.03	71.75	65.37	68.85	78.71
1000	79.91	80.29	77.39	79.39	77.80	74.92	74.26	77.34	74.80	77.26	72.83	66.77	69.84	78.91

XQuAD	zh	vi	tr	th	ru	hi	es	el	de	ar
0	46.29	52.84	53.82	57.64	57.10	49.67	57.97	56.77	56.33	48.36
2	47.16	52.86	52.84	60.96	55.39	50.20	57.51	55.37	57.05	47.97
4	48.06	53.43	51.88	61.57	54.21	50.28	57.62	55.68	56.72	49.00
6	52.29	53.41	53.03	62.97	55.48	50.85	57.88	55.37	57.16	49.10
8	57.88	53.49	52.47	63.73	55.87	50.96	58.25	55.83	57.05	50.09
10	60.22	53.28	52.36	64.02	55.79	51.38	57.90	56.11	57.47	49.30

Table C.4: Detailed per-language few-shot language results for XNLI and XQuAD with XLM-R for different number of target-language data instances k.

Appendix D

Experimental Details for Chapter 8

D.1 Experimental Details for Section 8.1

For completeness, we report detailed results on bias effects for each of the six XWEAT tests and bilingual word embedding spaces for all 21 language pairs. Tables D.1 to D.6 show bias effects for XWEAT tests T1, T2, and T6–T9.

XW1	EN	DE	ES	IT	HR	RU	TR
EN	–	1.28	1.63	1.62	1.59	1.49	1.32
DE	1.55	–	1.28	1.45	1.41	1.03	1.29
ES	1.45	1.25	–	1.28	1.21	1.31	1.09
IT	1.18	1.10	1.28	–	1.29	0.61	1.09
HR	1.57	1.62	1.59	1.62	–	1.62	1.63
RU	1.41	1.12	1.20	1.38	1.46	–	1.29
TR	1.23	1.21	1.06	1.26	1.24	1.04	–

Table D.1: XWEAT T1 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

XW2	EN	DE	ES	IT	HR	RU	TR
EN	–	1.35	1.51	1.48	1.60	1.56	1.15
DE	1.37	–	1.25	1.19	1.31	1.47	1.16
ES	1.55	1.50	–	1.53	1.50	1.57	1.22
IT	1.54	1.37	1.28	–	1.47	1.39	1.27
HR	1.19	1.25	0.72	1.09	–	1.26	0.81
RU	1.46	1.26	1.23	1.08	1.13	–	0.71
TR	1.29	1.44	1.21	1.4	1.25	1.57	–

Table D.2: XWEAT T2 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

XW6	EN	DE	ES	IT	HR	RU	TR
EN	–	1.77	1.81	1.88	1.83	1.78	1.89
DE	1.82	–	1.77	1.85	1.84	1.74	1.86
ES	1.71	0.95	–	1.81	1.80	1.61	1.50
IT	1.76	1.58	1.70	–	1.72	1.77	1.76
HR	1.68	1.65	1.66	1.43	–	1.74	1.73
RU	1.86	1.74	1.74	1.82	1.86	–	1.80
TR	1.90	1.66	1.77	1.82	1.77	1.55	–

Table D.3: XWEAT T6 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

XW7	EN	DE	ES	IT	HR	RU	TR
EN	–	0.34*	1.36	1.33	0.26*	0.46*	0.49*
DE	1.51	–	1.60	1.42	0.23*	1.33	-0.62*
ES	1.63	0.24*	–	1.26	0.60*	1.29	1.55
IT	1.12	0.65*	1.01	–	0.51*	-0.20*	-1.08
HR	1.46	0.94	0.95	1.27	–	0.62*	0.00*
RU	1.19	-0.51*	1.30	1.09	0.81*	–	-0.79*
TR	1.22	0.07*	0.81*	1.30	-0.23*	-0.48*	–

Table D.4: XWEAT T7 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

XW8	EN	DE	ES	IT	HR	RU	TR
EN	–	0.68*	1.49	1.01	-0.38*	-0.06*	0.71*
DE	1.17	–	1.43	1.10	-0.09*	1.06	1.16
ES	1.13	-0.69*	–	0.61*	-0.19*	0.67*	-0.18*
IT	0.75*	-0.76*	0.87	–	-0.18*	-0.52*	0.04*
HR	1.36	0.42*	0.92	0.76*	–	-0.16*	0.90
RU	1.09	-0.84*	0.96	0.99	0.19*	–	1.00
TR	0.93	0.06*	1.49	1.21	-0.47*	-0.43*	–

Table D.5: XWEAT T8 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

XW9	EN	DE	ES	IT	HR	RU	TR
EN	–	1.12	1.66	1.61	-0.59*	1.76	1.65
DE	1.74	–	1.68	1.66	-1.39	1.46	1.57
ES	1.64	1.48	–	1.79	-1.34	1.75	1.37
IT	1.62	0.19*	1.47	–	-1.63	1.87	1.74
HR	1.54	1.89	1.87	0.96*	–	1.73	1.59
RU	1.82	1.54	1.64	1.72	-0.84*	–	0.80*
TR	1.88	0.98*	1.88	1.70	-1.80	0.58*	–

Table D.6: XWEAT T9 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

D.2 Experimental Details for Section 8.2

D.2.1 Full Experimental Results

We provide the complete experimental results of the cross-lingual debiasing transfer.

		DE							ES						
		Explicit			Implicit		SemQ		Explicit			Implicit		SemQ	
	Model	WEAT	ECT	BAT	KM	SVM	SL	WS	WEAT	ECT	BAT	KM	SVM	SL	WS
W1	Distributional	1.36	41.7	59.9	98.9	75.7	40.7	68.0	1.47	61.8	48.1	100	57.5	–	–
	GBDD	0.42*	77.7	48.2	90.5	51	40.7	**68.1**	0.56	**89.4**	**34.4**	96.8	50.3	–	–
	BAM	1.39	50.6	54	95	94.3	39	64.5	1.12	62.9	42.2	97.7	95.3	–	–
	DN	0.42*	48.1	48.3	98.9	53	39.9	61.9	0.96	55.8	41.6	97.7	34.4	–	–
	GBDD ∘ BAM	0.61	81.1	**44.3**	93.2	88.4	39.1	64.7	0.56	76.4	38.2	98.4	77	–	–
	BAM ∘ GBDD	0.75	74.3	52.4	90.8	50	**40.8**	64.9	**0.48***	85.3	42.8	**94.1**	49.5	–	–
	GBDD ∘ DN	**0.30***	**82.8**	45.7	**86.6**	**42.9**	39.6	61.9	0.69	75.1	38	96.2	38.3	–	–
W8	Distributional	**0.05***	34.1	37.2	98.3	50	40.7	68	1.16	67.8	36.4	99.8	50	–	–
	GBDD	0.15*	**85.3**	30.5	55.4	50	40.7	67.7	0.41*	70.9	31.1	60	50	–	–
	BAM	-0.97	41.5	33.6	97.4	100	40.7	65.8	0.11*	70.9	34.4	99	100	–	–
	DN	-0.1*	67.1	37.4	97.4	50	36.2	62	0.76*	74	48.1	100	50	–	–
	GBDD ∘ BAM	-0.12*	83.2	35.2	56.3	50	**40.8**	65.6	**0.05***	83.7	33.1	58	50	–	–
	BAM ∘ GBDD	**-0.09***	84.4	**28.5**	**54.4**	50	37.3	66.7	0.11*	85.9	**28.1**	56.6	50	–	–
	GBDD ∘ DN	0.35*	73.4	35.7	57.6	50	35.9	61.1	0.78*	**88.5**	46.4	**52.4**	50	–	–

Table D.7: Complete cross-lingual debiasing transfer results for transfer to German (DE) Spanish (ES). Results obtained on the XWEAT T1 and T8 tests of respective languages.

		IT							RU						
		Explicit			Implicit		SemQ		Explicit			Implicit		SemQ	
	Model	WEAT	ECT	BAT	KM	SVM	SL	WS	WEAT	ECT	BAT	KM	SVM	SL	WS
W1	Distributional	1.28	57.7	57.2	97	54.8	29.8	64.2	1.28	57.6	43.5	96.7	54.3	25.6	59.2
	GBDD	0.02*	81.8	**44**	77.3	51.1	**29.8**	**64**	0.67	79.8	**35.3**	93.5	49.9	**25.4**	**59**
	BAM	1.35	54	55.5	95.9	95.6	27.3	62.2	1.20	66	44.4	94.4	94.3	24.2	55.5
	DN	0.53	62.8	51.9	99.8	55.5	25.7	58.5	**0.44***	57.7	42.7	96.5	56.3	24.3	52.6
	GBDD ∘ BAM	0.44*	70.9	51.4	87.7	86.2	27.3	62.2	0.6	**80.7**	40.1	93.5	89	24.2	55.4
	BAM ∘ GBDD	0.29*	76.5	48.6	**73.4**	50.2	28.2	62.4	0.65	80.2	37.7	92.8	49.6	25	56.3
	GBDD ∘ DN	0.2*	**83.5**	48	88.1	57.6	25.8	58.3	**0.36***	75	40.7	**91.1**	52.4	24.1	52.5
W8	Distributional	**0.10***	**92.5**	25.9	99.8	50	29.8	64.2	0.37*	49.9	32.1	62	50	25.6	59.2
	GBDD	-0.28*	86.4	25.9	**56.1**	50	**29.8**	**63.4**	0.73*	49.5	32	62.4	50	**25.8**	**58.3**
	BAM	-0.70*	57.4	23	99.6	100	29	61	-0.41*	44.6	**25.9**	74.4	100	25.1	56.8
	DN	-1.05	40.7	**14.1**	100	50	25.4	57.7	**0.31***	46.8	35.5	77.9	50	20.7	56.9
	GBDD ∘ BAM	-0.62*	67	23.1	57.9	50	29	60	0.34*	72.7	30.8	**56.8**	50	24.8	55.8
	BAM ∘ GBDD	**-0.05***	82.3	28.9	58.9	50	27.1	60.2	0.59*	**83.7**	31	61.6	50	25.4	57.5
	GBDD ∘ DN	-0.64*	51.2	18.7	60.1	50	25	56.7	0.77*	69.7	38.3	61.9	50	20.7	55.1

Table D.8: Complete cross-lingual debiasing transfer results for transfer to Italian (IT) and Russian (RU). Results obtained on the XWEAT T1 and T8 tests of respective languages.

D.2.2 Bias Specifications

We provide the full term sets of the bias specifications and their augmentations for different k employed in our study in Tables D.10 and D.11.

		HR							TR						
		Explicit			Implicit		SemQ		Explicit			Implicit		SemQ	
	Model	WEAT	ECT	BAT	KM	SVM	SL	WS	WEAT	ECT	BAT	KM	SVM	SL	WS
W1	Distributional	1.45	56.3	63.4	57	51.7	32.7	–	1.21	69.6	47.9	86.3	50.6	–	–
	GBDD	0.85	81.2	60.5	63.2	49.8	**32.8**	–	0.64	83.9	40.9	79.7	51.4	–	–
	BAM	1.35	50.8	63.8	**59.5**	90.5	31.2	–	0.89	64.8	39.1	84.3	90.6	–	–
	DN	0.86	74.8	67.2	87.4	35.8	28.4	–	0.78	73.3	36.9	88.1	58.3	–	–
	GBDD ∘ BAM	0.82	63.6	**57.1**	55.1	77.5	31.3	–	**0.19***	80	34.5	**72**	73.2	–	–
	BAM ∘ GBDD	0.71	86.8	63	68.7	50	30.9	–	0.76	82.3	53	75	51.1	–	–
	GBDD ∘ DN	0.56*	**85.9**	65.5	61.4	**44**	28.5	–	0.63	81.5	**33**	74.7	54.9	–	–
W8	Distributional	0.13*	53.2	39.4	98.6	50	32.7	–	1.72	39.6	64.5	99.3	50	–	–
	GBDD	0.54*	59.7	40.2	**59.9**	50	**32.5**	–	1.41	71.9	66.5	64.3	50	–	–
	BAM	**-0.01***	30.3	41.1	93.5	100	32	–	1.49	62.1	59.5	98.8	100	–	–
	DN	0.25*	81.7	52.8	99.9	50	25.3	–	1.54	44.6	65.5	100	50	–	–
	GBDD ∘ BAM	0.52*	73.8	47	60.8	50	31.7	–	**0.99**	85.3	**56**	**56.9**	50	–	–
	BAM ∘ GBDD	0.68*	60.9	44.5	75.4	50	29.4	–	1.27	59.3	76	62.4	50	–	–
	GBDD ∘ DN	0.67*	**88.5**	**56.6**	67.5	50	25.1	–	1.29	**86.7**	65	62.5	50	–	–

Table D.9: Complete cross-lingual debiasing transfer results for Croatian (HR) and Turkish (TR). Results obtained on the XWEAT T1 and T8 tests of respective languages.

k=0	T_1	*aster clover hyacinth marigold poppy azalea crocus iris orchid rose blue-bell daffodil lilac pansy tulip buttercup daisy lily peony violet carnation gladiola magnolia petunia zinnia*
	T_2	*ant caterpillar flea locust spider bedbug centipede fly maggot tarantula bee cockroach gnat mosquito termite beetle cricket hornet moth wasp blackfly dragonfly horsefly roach weevil*
	A_1	*caress freedom health love peace cheer friend heaven loyal pleasure diamond gentle honest lucky rainbow diploma gift honor miracle sunrise family happy laughter paradise vacation*
	A_2	*abuse crash filth murder sickness accident death grief poison stink assault disaster hatred pollute tragedy divorce jail poverty ugly cancer kill rotten vomit agony prison*
k=2	T_1	*glovers gladiolus nance crowfoot meadowsweet dianthus pinkish dolly poppies cyclamen tulips sapphire azaleas wisteria camellia asters trefoil sissy olive penstemon candlewood prunella primula mauve opium buddleja taupe magenta veronica hyacinths magnolias watercress minaj cowslip lilies tulipa orchis daffodils scarlet jasmine faggot marigolds orchids*
	T_2	*caterpillars gnats termites avenger ants bumblebee arachnid sticking cricketing flit tarantulas pyralidae harrier millipede centipedes mosquitos vermin worm cockroaches locusts wasps insect snook larva scoot gracillariidae weevils grasshopper undershot fathead whitefly louse batsman dragonflies*
	A_1	*donation liberty tranquility fortunate mild laugh diamonds holiday truthful endowment untried fitness colleague credentials lineage gurgling honour faithful cheerfulness auspicious affection prism genuine esteem moonlight newfound vacations gem eden peacefulness gladden wellness partner glad cuddle cherish joy liege diplomas phenomenon fondle autonomy prodigy tickled enjoyment clement utopia tribe*
	A_2	*misuse collision stench destitution demise anguish annihilate estrangement illness incarcerate sorrow mistreat infection destroy separation slaughter antipathy penitentiary smash regurgitate malady misery decease dirt calamity impoverishment spew stinking toxin enmity imprison tainted massacre gaol sinister horrible defile contaminate reek prostate catastrophe crud casualty mishap leukemia invasion misadventure onslaught*
k=3	T_1	*faggot cornflower meadowsweet cowslip camellia cress weeknd orchidaceae watercress trefoil pinkish magnoliaceae orchids lilies dianthus hyacinths primula willowherb daffodils mauve penstemon azaleas fleabane magenta wisteria jessie licorice lilacs polly peonies magnolias candlewood amaranthus jasmine opium bluish poppies sapphire orchis sissy buddleja tangerine olive clovers marigolds lavender dandelions tulipa taupe tulips poof crowfoot gladiolus prunella dandelion veronica dolly asters cyclamen scarlet minaj nance*
	T_2	*projected avenger grasshopper vermin scamper worm cockroaches fathead harrier batsman weevils snook whitefly bug noctuidae scorpion mayfly tarantulas louse roaches cricketing bumblebee gnats curculionidae arachnid mosquitoes wasps dragonflies scoot termites larva millipede corsair flit gracillariidae locusts wicket hive insect caterpillars mosquitos parasitoid undershot sticking centipedes ants pyralidae fleas*
	A_1	*fortunate colleague auspicious peacefulness untried jewel propitious cherish joy truthful stunner hug dearest partner comrade honour gladden glad bliss delight encourage mild eden laugh moonlight genuine tickled joyful diamonds gem gratuity sabbatical enjoyment lineage endowment liberty certificate newfound liege wellness gurgling credentials clement utopia autonomy faithful tribe chuckle vacations prism holiday serenity sincere phenomenon diplomas homage rainbows donation cuddle welfare tranquility affection allegiant independency tranquil prodigy esteem fondle cheerfulness ancestry fitness untested*
	A_2	*severance reek imprison onslaught surly destroy massacre invasion complaint spew dirt casualty heartbreak slaying stinking catastrophe penitentiary demise slaughter privation toxin illness impoverishment annihilate calamity contaminate separation collision outrage grime stench disgorge mishap collide hate regurgitate crud misuse malady contagion sinister infection smash attack leukemia tumour tainted anguish defile stinky ailment gaol decease extinguish enmity sorrow misadventure expiration pollutes antipathy estrangement misery incarcerate horrible prostate destitution mistreat*
k=4	T_1	*scarlet bluebell cornflower delphinium fleabane amaranthus dianthus chromatic poof peonies orchidaceae orchis azaleas mauve tangerine nance tulipa camellia taupe willowherb hyacinths minaj periwinkle helianthemum poppies lilies cress magnolias macklemore dolly sissy sapphire orchids buddleja licorice jasmine faggot tulips lavender opium dandelion weeknd wisteria cowslip prunella thyme alfalfa lilacs daffodils magnoliaceae pinkish watercress crowfoot veronica primula carrie bluish cryptanthus trefoil asters jessie polly olive clovers meadowsweet fuchsia penstemon candlewood marigolds dandelions cyclamen snowberry purplish sassafras gladiolus epiphyte magenta*
	T_2	*caterpillars wasps corsair whitefly insect bumblebee bowler noctuidae yellowjacket mayfly curculionidae cockroaches dragonflies avenger mulligan pilotless roundworm undershot protruding grasshopper crambidae damselfly louse projected cricketing vermin parasitoid tarantulas wicket sticking scorpion gnats hellcat mosquitoes sawfly hive arachnid larva locusts centipedes snook batsman weevils dart flit bug fleas gracillariidae harrier burrowing scamper roaches hickory mosquitos scoot tractor fathead worm bumblebees millipede pyralidae termites leafhopper ants*
	A_1	*independency rhombus daybreak endowment enliven vacationing cheerful tribe partner privilege truthful rainbows gem gratification gratuity affection phenomenon delight untried daydream mirth fondle tranquility prism gladden enjoyment esteem stunner certificate genuine holiday glad sabbatical encourage autonomy cherish baccalaureate favorable credentials donation tranquil fitness wellness mild reverence hug benefaction gracious diplomas ancestry nirvana staunch chuckle vacations cuddle marvel propitious liege gurgling serenity peacefulness honour kiss allegiant utopia welfare sincere clement jewel eden fortunate faithful joyful prodigy moonlight homage diamonds tickled laugh dearest sidekick colleague untested bliss cheerfulness lineage liberty parentage idolize calmness authentic comrade joy auspicious newfound wellbeing*
	A_2	*stinky protest mistreat sorrow disease maltreatment taint remand horrible casualty contaminate smash misery misuse annihilate imprison crud raid grime pollutes contagion barf infection hate decease slaughter destroy calamity sinister breakup expiration enmity carnage hideous demise regurgitate stench tainted outrage stockade dying separation invasion shatter antipathy happening extinguish privation spew tumour ailment complaint attack destitution exterminate rancid massacre impoverishment slaying heartache misfortune incarcerate disgorge surly malady catastrophe onslaught collide misadventure defile gaol prostate dirt penitentiary anguish dearth animosity muck heartbreak reek severance contamination collision estrangement illness leukemia tumor mishap toxin stinking*

Table D.10: Bias specification of WEAT T1: sentiment attached to flowers (T_1) vs. insects (T_2). Original terms from Caliskan et al. (2017) and augmented list for different k.

k=0	T_1	*science technology physics chemistry Einstein NASA experiment astronomy*
	T_2	*poetry art Shakespeare dance literature novel symphony drama*
	A_1	*brother father uncle grandfather son he his him*
	A_2	*sister mother aunt grandmother daughter she hers her*
k=2	T_1	*automation radiochemistry test biophysics learning electrodynamics biochemistry astrophysics erudition astrometry technologies experimentation*
	T_2	*orchestra artistry dramaturgy poesy philharmonic craft untried hop poem dancing dissertation treatise new dramatics*
	A_1	*beget buddy forefather man nephew own himself theirs boy helium crony cousin grandpa granddad herself*
	A_2	*niece girl parent grandma granny woman theirs sire auntie sibling herself jealously stepmother wife*
k=3	T_1	*technologies biochemistry astrophysics engineering electrodynamics radiochemistry astronomer erudition education automation biophysics chromodynamics research learning experimentation test astrometry biology*
	T_2	*groundbreaking craftsmanship dissertation new literatures dramatization philharmonic sinfonietta artistry untried poems dramaturgy dancing dramatics poem poesy craft hop treatise orchestra waltz*
	A_1	*granddad granddaddy man helium grandpa own himself forefather themself kinsman theirs sire beget boy buddy herself comrade who crony nephew grandson cousin*
	A_2	*sire beget stepmother aunty parent woman grandma herself own stepsister female girl jealously sibling auntie theirs granny niece wife*
k=4	T_1	*physicists test electrochemistry automation engineering biophysics education learning chromodynamics technologies radiochemistry examination biology technological astronomer astrophysics experimentation biochemistry research lore electrodynamics astrobiology astrometry erudition*
	T_2	*dramaturgy monograph untried dances poesy dissertation craftsmanship orchestra treatise skill waltz poem literatures dramatization poems theatre dancing newfound hop artistry new verse craft philharmonic concerto groundbreaking dramatics sinfonietta*
	A_1	*grandad theirs grandson buddy themself stepbrother forefather ironically crony granddaddy grandpa sidekick boy heir granddad cousin who male man sire parent beget kinsman nephew herself own comrade himself helium*
	A_2	*auntie fiance theirs female stepmother grandma woman procreate stepsister widow aunty grandmothers mimi granny sibling wife sire parent beget niece herself own girl jealously siblings*
k=5	T_1	*experimentation lore research chromodynamics astrobiology technological technologies physicists education investigation engineering examination radiochemistry biology astrophysics astrology chemistries learning biochemistry electrochemistry biophysics astronomer test scholarship electrodynamics biotechnology erudition automation astrometry*
	T_2	*new untried literatures rhyme sinfonietta monograph philharmonic hop expertise craft dancing theater dances newfound artistry dramatics untested writing orchestra dramatization poesy craftsmanship dramaturgy jitterbug theatre treatise concerto poem orchestral verse poems waltz dissertation groundbreaking skill*
	A_1	*granddad crony its granddaddy male helium herself forefather heir granduncle own sidekick grandson comrade grandfathers sire nephew man stepbrother grandad theirs cousin who hesitates themself parent grandpa kinsman ironically himself boy buddy spawn beget*
	A_2	*female wife kinswoman girl herself stepsisters stepsister grandmothers own granny stepmother affections woman sire spouse lady theirs fiance aunty procreate progenitor parent jealously sisters siblings niece widow mimi auntie matriarch sibling grandma beget*

Table D.11: Bias specification of WEAT T8: female vs. male attributes attached to science (T_1) vs. art (T_2). Original terms and augmented list for different k.