An Argument-Annotated Corpus of Scientific Publications

Anne Lauscher,¹,² Goran Glavaš,¹ and Simone Paolo Ponzetto¹
¹Data and Web Science Research Group
University of Mannheim, Germany
²Web-based Information Systems and Services
Stuttgart Media University, Germany
{anne, goran, simone}@informatik.uni-mannheim.de
lauscher@hdm-stuttgart.de

Abstract

Argumentation is an essential feature of scientific language. We present an annotation study resulting in a corpus of scientific publications annotated with argumentative components and relations. The argumentative annotations have been added to the existing Dr. Inventor Corpus, 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.

1 Introduction

With the rapidly growing amount of scientific literature (Bornmann and Mutz, 2015), computational methods for analyzing scientific writing are becoming paramount. To support learning-based models for automated analysis of scientific publications, potentially leading to better understanding of the different rhetorical aspects of scientific language (which we dub scitorics), researchers publish manually-annotated corpora. To date, existing manually-annotated scientific corpora already reflect several of these aspects, such as sentential discourse roles (Fisas et al., 2015), argumentative zones (Teufel et al., 1999, 2009; Liakata et al., 2010), subjective aspects (Fisas et al., 2016), and citation polarity and purpose (Jochim and Schütte, 2012; Jha et al., 2017; Fisas et al., 2016).

As tools of persuasion (Gilbert, 1976, 1977), scientific publications are abundant with argumentation. Yet, somewhat surprisingly, there is no publicly available corpus of scientific publications (in English), annotated with fine-grained argumentative structures. In order to support comprehensive analyses of rhetorics in scientific text (i.e., scitorics), argumentative structure of scientific publications should not be studied in isolation, but rather in relation to other rhetorical aspects, such as the discourse structure. This is why in this work we contribute a new argumentation annotation layer to an existing Dr. Inventor Corpus (Fisas et al., 2016), already annotated for several rhetorical aspects.

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 components and relations. Our 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.¹ Finally, we offer an extensive statistical and information-theoretic analysis of the corpus.

2 Related Work

Researchers have offered a plethora of argument annotation schemes and corpora for various domains, including Wikipedia discussions (Biran and Rambow, 2011), on-line debates (e.g., Abbott et al., 2016; Habernal and Gurevych, 2016), e-markets (e.g., Islam, 2007), persuasive essays (Stab and Gurevych, 2017), news editorials (Al Khatib et al., 2016), and law (Wyner et al., 2010). The corpus of Reed et al. (2008) covers multiple domains, including news and political debates.

The work on argumentative annotations in scientific writing is, however, much scarcer. 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 publica-

¹http://data.dws.informatik.uni-mannheim.de/sci-arg/compiled_corpus.zip
tions 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. 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 on what are the claims (and what is not a claim) and how they are mutually connected. Green et al. (2014); Green (2014, 2015, 2016) proposed methods for identifying and annotating argumentative structures in scientific publications, but released no publicly available annotated corpus. In the effort most similar to ours, Kirschner et al. (2015) annotated arguments in a corpus of educational research publications. Besides being quite small, this corpus is also written in German.

3 Annotation Scheme

A number of theoretical frameworks of argumentation have been proposed (Walton et al., 2008; Anscombe and Ducrot, 1983, inter alia). Among the most widely used is the model of Toulmin (2003), from which we start in this work as well, because of its relative simplicity and adoption in artificial intelligence and argument mining (Bench-Capon, 1998; Verheij, 2005; Kirschner et al., 2015). The Toulmin model, originally developed for the legal domain, recognizes six types of argumentative components: claim, data, warrant, backing, qualifier, and rebuttal.

We conducted a preliminary annotation study using the Toulmin model with two expert annotators on a small corpus subset. Annotators did not identify any warrant, backing, qualifier, nor rebuttal components. The annotators also pointed to the interlinked argumentative structure of publications in which claim were often used as ground for (supporting or conflicting) another claim. Not foreseen by the Toulmin model, we realized that the relations between argumentative components can be of different nature. Finally, the annotators recognized two distinct claim types: those presented as common knowledge (or state of the art) in the research area and those relating to authors’ own research.

Following the above observations from the preliminary annotation, we simplify the annotation scheme by removing the non-observed component types. Our final annotation scheme has the following types of argumentative component:

(1) Own Claim is an argumentative statement that closely relates to the authors’ own work, e.g.:

“Furthermore, we show that by simply changing
the initialization and target velocity, the
same optimization procedure leads to running
controllers.”

(2) Background Claim is an argumentative statement relating to the background of authors’ work, e.g., about related work or common practices in the respective research field, e.g.:

“Despite the efforts, accurate modeling of human
motion remains a challenging tasks.”

(3) Data component represents a fact that serves as evidence for or against a claim. Note that references or (factual) examples can also serve as data, e.g.:

“[..], due to memory and graphics hardware
constraints nearly all video game character
animation is still done using traditional SSD.”

We follow Bench-Capon (1998) and allow for links between the arguments. We introduce three different relations types, similar to Dung (1995).

(1) A Supports relation holds between components $a$ and $b$ if the assumed veracity of $b$ increases with the veracity of $a$;

(2) A Contradicts relation holds between components $a$ and $b$ if the assumed veracity of $b$ decreases with the veracity of $a$;

(3) The Semantically Same relation is annotated between two mentions of effectively the same claim or data component. This relation can be seen as argument coreference, analogous to entity (Lee et al., 2011, 2017) and event coreference (Glavaš and Šnajder, 2013; Lu and Ng, 2018).

It is important to emphasize that we do not bind the spans of our argumentative components to sentence boundaries, but rather allow for argumentative components of arbitrary span lengths, ranging from a single token to multiple sentences.

4 Annotation study

Dataset. Believing that argumentation needs to be studied in combination with other rhetorical aspects of scientific writing, we enriched the existing Dr. Inventor corpus (Fisas et al., 2015, 2016), consisting of 40 publications from computer graphics,
with argumentative information. The Dr. Inventor corpus has four layers of rhetorical annotations: (1) discourse roles, (2) citation purposes with associated citation contexts, (3) judgments of subjective aspects, and (4) annotations of sentence relevance for a summary. Table 1 summarizes the different annotation layers and their label distributions.

**Annotation Process.** We hired one expert and three non-expert annotators for our annotation study. We trained the annotators in a calibration phase, consisting of five iterations, in each of which all annotators annotated one publication. After each iteration we computed the inter-annotator agreement (IAA), discussed the disagreements, and, if needed, adjourned the annotation guidelines. We measured the IAA in terms of the $F_1$-measure because (1) it is easily interpretable and straight-forward to compute and (2) it can account for spans of varying length, allowing for computing relaxed agreements in terms of partial overlaps.

The evolution of IAA over the five calibration iterations is depicted in Figure 1, 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. 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 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.

### 5 Corpus Analysis

We first study the argumentation layer we annotated in isolation. Afterwards, we focus on the interrelations with other rhetorical annotation layers.

**Analysis of Argumentation Annotations.** Table 2 lists the number of components and relations in total and on average per publication. The number of own claims roughly doubles the amount instances grows (Hripcsak and Rothschild, 2005).

<table>
<thead>
<tr>
<th>Annotation Layer</th>
<th>Labels</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discourse Role</td>
<td>Background</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Challenge</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Approach</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Outcome</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Future Work</td>
<td>2</td>
</tr>
<tr>
<td>Citation Purpose</td>
<td>Criticism</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Use</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Substantiation</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Basis</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>53</td>
</tr>
<tr>
<td>Subjective Aspect</td>
<td>Advantage</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Disadvantage</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Adv.-Disadv.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Novelty</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Common Practice</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Limitation</td>
<td>2</td>
</tr>
<tr>
<td>Summarization Relevance</td>
<td>Totally irrelevant</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Should not appear</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>May appear</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Relevant</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Very relevant</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 1: Annotation layers of the Dr. Inventor Corpus with label distributions.

---

3 A researcher in computational linguistics, not in computer graphics.

4 Humanities and social sciences scholars.

5 http://data.dws.informatik.uni-mannheim.de/sci-arg/annotation_guidelines.pdf

6 Note that the chance-corrected measures, e.g., Cohen’s Kappa, approach $F_1$-measure when the number of negative
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, as authors mainly argue by providing supporting evidence for their own claims.

Table 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 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.

Table 4: Graph-based analysis of the argumentative structures identified in the extended Dr. Inventor Corpus (per publication).

Table 2: Total and per-publication distributions of labels of argumentative components and relations in the extended Dr. Inventor Corpus.

Table 3: Statistics on length of argumentative components (in number of characters) in the extended Dr. Inventor Corpus.
Table 5: Claims with maximum PageRank score in a publication.

<table>
<thead>
<tr>
<th>Type</th>
<th>Pub.</th>
<th>Claim with maximal PageRank score</th>
</tr>
</thead>
<tbody>
<tr>
<td>background claim</td>
<td>A13</td>
<td>'physical validity is often sacrificed for performance'</td>
</tr>
<tr>
<td></td>
<td>A21</td>
<td>'a tremendous variety of materials exhibit this type of behavior'</td>
</tr>
<tr>
<td>own claim</td>
<td>A39</td>
<td>'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'</td>
</tr>
</tbody>
</table>

Table 6: Normalized mutual information between different annotation layers.

<table>
<thead>
<tr>
<th></th>
<th>AC</th>
<th>DR</th>
<th>SA</th>
<th>SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>0.22</td>
<td>0.11</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>DR</td>
<td>0.08</td>
<td>0.10</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td>SA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.18</td>
<td>0.10</td>
<td>0.04</td>
<td>0.01</td>
</tr>
</tbody>
</table>

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. In Table 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 argumentative components (AC) and discourse roles (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 discourse role. 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 typically reference other publications and in order to motivate their work:

"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. At the same time, the underlined text can be seen as the citation context with the reference as target.

6 Conclusion

We 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. Finally, we analyzed the dependencies between different rhetorical aspects, which can inform computational models aiming to jointly address multiple aspects of scientific discourse. In the future, we plan to extend the corpus with publications from other domains and develop computational models for the integrated analysis of scientific writing.

Acknowledgments

This research was partly funded by the German Research Foundation (DFG), grant number EC 477/5-1 (LOC-DB). We thank our annotators for their very dedicated annotation effort.
References


Nancy Green. 2016. Implementing argumentation schemes as logic programs. In The 16th Workshop on Computational Models of Natural Argument, volume 30, New York, USA. CEUR-WS.

Nancy Green, E Cabrio, S Villata, and A Wyner. 2014. 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, pages 21–25, Forl-Cesina, Italy. CEUR-WS.


