UKParl: A Data Set for Topic Detection with Semantically Annotated Text

Federico Nanni[1], Mahmoud Osman[1], Yi-Ru Cheng[1], Simone Paolo Ponzetto[1], Laura Dietz[2]
[1]Data and Web Science Group - University of Mannheim, Germany
[2]Computer Science Department - University of New Hampshire, USA

Abstract

We present a dataset created from the Hansard House of Commons archived debates of the UK parliament (2013-2016). The resource includes fine-grained topic annotations at the document level and is enriched with additional semantic information such as the one provided by entity links. We assess the quality and usefulness of this corpus with two benchmarks on topic classification and ranking.

Keywords: topic detection, text as data, text classification, entity linking, ranking

1. Introduction

In recent years, the prompt availability of digital collection of political texts (Koehn, 2005; Vinciarelli et al., 2009; Bachmann, 2011; Cullen et al., 2014; Merz et al., 2016; van Aggelen et al., 2017) has fostered much work in the field of computational social science (CSS), an interdisciplinary field where political science scholars adopt – among other methodologies – Natural Language Processing (NLP) approaches for studying the act and content of political communication (Grimmer and Stewart, 2013).

A task that has attracted large interest in the Computational Social Science community (CSS) is the automatic detection of topics in unstructured text, since this can, in turn, support higher-level tasks such as fine-grained political campaign analyses (Nanni et al., 2016), measuring the agreement between political leaders (Menini et al., 2017) and quantify political attention (Quinn et al., 2010), to name a few.

However, while there is such large availability of digital collections of transcript of campaign speeches and parliamentary debates, social media posts on political events or datasets of party manifestos, most of these collections lack fine-grained annotations of the topics they cover. This limits both the types of analysis that researchers can conduct employing such corpora and the development of benchmarks and evaluation campaigns for testing topic detection algorithms in the political science domain.

Contributions. Consequently, in order to address these issues, we provide the research community with: *a)* a political corpus that we have constructed from the UK parliament Hansard House of Commons archived debates (2013-2016), including fine-grained topic annotations at the document level and entity links; *b)* two different topic prediction benchmarks, in order to foster further research on textual topic detection in the political domain.

2. Related Corpora

One of the first machine-readable resources of transcript of political speeches available to the research community is the well-known *EuroParl* corpus (Koehn, 2005), a collection of parallel texts in 11 languages (later extended to 21 languages (Islam and Mehler, 2012)) created from the proceedings of the European Parliament (EP)[1]. The same collection has been recently made available as linked open data (van Aggelen et al., 2017): *LinkedEP*[2] offers translation of the reports of the plenary meetings of the EP, together with additional metadata information such as the political affiliation of the parliament members, for instance, which is organized in over 25 million triples. Similar resources can be found on the government websites of the United Kingdom[3] and of Italy[4]; regarding the case of the United States, Thomas et al. (2006) presented a corpus of speeches from the US Congress. However, despite thee availability and usefulness for NLP research of such collections (cf. *EuroParl* historically being a core resource for the development of statistical machine translation systems), none of these resources offer fine-grained annotations of the topics addressed in the speeches.

Apart from transcripts of parliamentary debates, another relevant collection of political text is the Manifesto Corpus (Merz et al., 2016)[5], a resource presenting digitized and topically annotated electoral programs that is based on the coding of the Manifesto Project (7 broad categories and more than a hundred fine-grained type of annotations). While researchers have pointed out inconsistencies in the annotations (Mikhaylov et al., 2012), this resource is considered to be one of the biggest human-coded, multilingual, cross-national, open-access corpora in the field of political science. The corpus provides more than 1,800 machine-readable documents, containing more than 600,000 annotated statements as well as metadata like political party affiliations and election year. However, for evaluating a topic detection system the topical annotations of the Manifesto Project remain too coarse-grained: as a matter of fact, instead of describing directly the topic addressed in text (e.g., "refugee crisis"), they map the content to a pre-defined fine-grained category like, for instance "freedom and human rights".

The work closest to ours is that of Bachmann (2011), where the authors conduct a corpus-driven semantic analysis of discourses about same-sex relationships in the UK Parliament. To this end, they create a corpus from the UK Hansard

[1] http://www.statmt.org/europarl/

[2] http://purl.org/linkedpolitics

[3] http://lda.data.parliament.uk/

[4] http://dati.camera.it/

[5] https://manifesto-project.wzb.eu/

Table 1: Corpus Statistics.

Session	# Speech	# Topic	# Token	# Entity
2013-14	23,935	2,343	175,604	72,791
2014-15	19,439	1,987	166,777	72,248
2015-16	26,605	1,923	169,119	74,678
Total	69,979	5,634	354,403	125,886

Archives[6] consisting of 16 electronic debates transcripts from both houses of the parliament: 9 debates from the House of Lords, and 7 from the House of Commons. In our work, we consider the same archive, but we collected all materials available between 2013 and 2016, which sum up to around 70,000 speeches and more than 5,600 topics, as presented in Table 1.

3. Corpus Overview

In order to create the corpus, we collected all transcripts of speeches made on the House of Commons floor between 2013 and 2016. Speeches have been manually associated with a single topic (e.g., 'Isis', 'Zika Virus', 'Greece Financial Crisis', etc.) by the curators of the corpus. In order to enable analyses leveraging background knowledge, we additionally aligned each topic, whenever possible, with the related Wikipedia page, for instance 'Isis' with /wiki/Islamic_State_of_Iraq_and_the_Levant. Given the large number of speeches in the corpus and the fact that the associated topics are often clearly defined (e.g., 'EU Sanctions (Russia)', 'Northern Ireland Political Situation'), this was done automatically by employing the topic as a query and matching it with the first retrieved page, using the Wikipedia search-tool. However, we are aware that for potentially ambiguous topics (e.g., 'Voting System', 'Foreign Students') or topics without a related Wikipedia page (e.g., 'Wi-fi in Hospitals') this approach could generate inconsistencies. We aim to address this issue in the future with the support of human annotators.

The dataset[7] follows the structure of the original collection and it is organized in three sessions: 2013-14, 2014-15 and 2015-16. Each session is divided into a set of topics, where for each topic-speech pair we provide i) the original text of the speech; and ii) the list of entities that were identified in text (we use TagMe (Ferragina and Scaiella, 2010) with standard settings). The number of unique speeches, topics, tokens and entities in the corpus are presented in Table 1. The alignment between the topic and the related Wikipedia page is provided in an accompanying file.

4. Topic Classification Benchmark

Numerous supervised models have been proposed in the past for the classification of political text (Purpura and Hillard, 2006; Stewart and Zhukov, 2009; Verberne et al., 2014; Karan et al., 2016; Zirn et al., 2016; Glavaš et al., 2017a, *inter alia*). Inspired by these works, we test different feature vector representations of text and classification algorithms to provide a benchmark for this task on our corpus.

[6] http://www.parliament.uk/business/publications/

[7] http://federiconanni.com/ukparl

4.1. Feature vector representations

We compare four different ways of processing the text and transforming it into feature vectors.

TF-IDF (words). Standard TF-IDF (logarithmic, L2-normalised variant) vectors of documents (tokenized and lemmatized).

TF-IDF (entities). We used TagMe! to identify, disambiguate and link entities in text. We then compute entity-based TF-IDF vectors by considering each document as a bag of entities.

Word embeddings. As in previous work (Glavaš et al., 2017b), the document embedding representation of each speech is computed as the element-wise average of the embeddings of the words in the text. Let W be the set of unique words in a document D. The embedding of D is then computed as:

$$\frac{1}{N}\sum_{w \in W} \text{freq}(w) \cdot \vec{v}_w$$

where $\text{freq}(w)$ is the frequency of word w, $\vec{v}_w$ is its embedding vector, and N is the total number of unique words in D. For this, we use the state-of-the-art pre-computed GloVe word embeddings (300d)[8].

Entity embeddings. As in the case of word embeddings, we computed the vector as the element-wise average average of the embeddings of the unique entities in the text. We use state-of-the-art pre-computed RDF entities embeddings (Ristoski et al., 2016).

4.2. Classifiers

We compare the performance of four different classifiers all implemented in the Python library Scikit-Learn[9].

NB. A standard multinomial Naive-Bayes classifier.

Nearest Centroid. This memory-based classifier first creates a centroid for each topic, and then assigns each example to the topic whose centroid is closest, based on the euclidean distance between the feature vectors.

***k*-NN.** A standard k-Nearest Neighbors classifier that labels each example with the majority class of the k[10] most similar labeled documents, based on the euclidean distance between the feature vectors.

SVM. A Support Vector Machine using a linear kernel, with standard parameters (C=1.0).

4.3. Dataset

We evaluate the performance of each pair of document representation and classifier on two different sets of speeches.

2015-16. We first select for testing the largest subset of our collection, namely all speeches addressed in the session 2015-16. Among the most relevant topics there are

[8] https://nlp.stanford.edu/projects/glove/

[9] http://scikit-learn.org/

[10] During testing we obtain consistently good performance using 10 neighbors.

Table 2: Results on topic prediction (2015-2016 subset)

Doc. Representation	Classifier	Topic Prediction			
		Macro			Micro
		P	R	F_1	F_1
TF-IDF (words)	NB	0.18	0.14	0.15	0.17
	NearestCentroid	**0.52**	**0.49**	**0.50**	**0.46**
	k-NN	0.41	0.42	0.41	0.42
	SVM	0.49	0.39	0.43	0.44
TF-IDF (entities)	NB	0.10	0.09	0.09	0.10
	NearestCentroid	**0.30**	**0.30**	**0.30**	**0.28**
	k-NN	0.22	0.23	0.22	0.24
	SVM	0.27	0.25	0.25	**0.28**
Word embeddings	NB	0.31	0.28	0.29	0.24
	NearestCentroid	0.33	**0.33**	**0.33**	0.33
	k-NN	0.26	0.27	0.26	0.29
	SVM	**0.36**	0.31	**0.33**	**0.38**
Entity embeddings	NB	0.16	0.18	0.16	0.15
	NearestCentroid	0.23	0.23	0.23	0.21
	k-NN	0.17	0.18	0.17	0.20
	SVM	**0.27**	**0.25**	**0.25**	**0.28**

Table 3: Results on topic prediction (complete corpus).

Doc. Representation	Classifier	Topic Prediction			
		Macro			Micro
		P	R	F_1	F_1
TF-IDF (words)	NB	0.12	0.10	0.10	0.13
	NearestCentroid	**0.36**	0.33	**0.34**	0.36
	k-NN	0.22	0.22	0.22	0.27
	SVM	0.34	**0.35**	**0.34**	**0.38**
TF-IDF (entities)	NB	0.06	0.06	0.06	0.07
	NearestCentroid	**0.17**	**0.16**	**0.16**	**0.17**
	k-NN	0.10	0.11	0.10	0.13
	SVM	0.15	0.16	0.15	0.16
Word embeddings	NB	0.16	0.16	0.16	0.17
	NearestCentroid	**0.22**	**0.22**	**0.22**	0.22
	k-NN	0.15	0.16	0.15	0.19
	SVM	0.19	0.17	0.16	**0.25**
Entity embeddings	NB	0.09	0.10	0.08	0.09
	NearestCentroid	0.13	0.13	0.13	0.13
	k-NN	0.10	0.11	0.09	0.13
	SVM	**0.14**	**0.14**	**0.14**	**0.19**

the Scotland Bill, Brexit, the war in Syria, immigration in UK and the economic crisis in Greece. We excluded the general topics 'Topical Questions', 'Business of the House' and 'Engagements' and topics with less than 10 speeches; the final dataset is composed by 490 topics and more than 15,000 speeches.

All. The second benchmark is instead composed of speeches from all four sessions. Here we also removed the general topics mentioned above and those with less than 10 speeches. The final data collection consists of a total of 1,341 topics and more than 41,000 speeches.

4.4. Results

The results of our benchmark (precision, recall and F1-Score) are presented in Table 2 and 3. As it can be seen, in both cases the use of lexical features (TF-IDF) outperforms semantic approaches based on word embeddings or the use of entity links. This is mainly due to the size of the documents analyzed, which makes it difficult to represent them with a single embedding vector maintaining their meaning. Among the different classifiers that we tested, the best performance have been achieved in both datasets by the Nearest Centroid and the Support Vector Machine.

Table 4: Topical Ranking task on dataset.

	MAP	P@1
Baseline (Random)	0.13	0.04
Entity frequency	**0.37**	**0.24**
Entity TF-IDF	**0.37**	**0.24**
Centroid (embeddings)	0.20	0.10
Position (doc. order)	0.23	0.09
Position + frequency	0.24	0.14
Position + TF-IDF	0.22	0.11
Position + centroid	0.22	0.12

5. Topic Ranking Benchmark

There are many different ways of predicting in an unsupervised way the topic addressed in a political text (Grimmer and Stewart, 2013). In our setting, the topic of each document is represented by its aligned (Wikipedia) entity, such as, for instance, /wiki/European_Migrant_Crisis. This task has been already approached by the NLP community, for example in Hulpus et al. (2013) and in Lauscher et al. (2016) by combining entity linking and topic models. As already noticed in previous work (Hulpus et al., 2013), it is often the case that the topic of the document is not directly mentioned in the text. In our case we noticed that only 22% of the documents (15,581 documents) in our collection mention the entity that is assigned as its topic label. When considering this subset, the task of predicting the topic is similar to that of the entity salience (Dunietz and Gillick, 2014).

5.1. Ranking Approaches

Inspired by previous works, we present the results of our evaluation regarding topic-label ranking comparing different baseline approaches over the Topic Ranking benchmark.

Entity frequency. We rank entities in the document by their frequency of mentions. This follows the intuition that the topic of a document is probably often mentioned in a text.

Entity TF-IDF. Following previous work (Lauscher et al., 2016), we additionally weight the raw frequency of entities by their inverse document frequency (i.e., standard TF-IDF).

Centroid (embeddings). We compute for each document its centroid on the basis of its entity embeddings (Ristoski et al., 2016). Entities are ranked by their distance to the centroid.

Position-based ranking. Inspired by Dunietz and Gillick (2014), we consider entities mentioned at the beginning of the document (in our case the first 10 entities), and rank them by their order of appearance (**Position**). We additionally experiment with alternative ranking functions, namely on the basis of raw frequency of occurrence (**Position + frequency**), a standard TF-IDF weighting scheme (**Position + TF-IDF**), or distance to their centroid computed on the basis of the entity embeddings (**Position + Centroid**).

5.2. Results

We present the results of our benchmark on topic ranking in Table 4, where we quantify performance using standard ranking-sensitive metrics like Mean Average Precision (MAP) and Precision@1. As it can be noticed, for both

metrics the best baseline approaches rely on ranking entities based on raw or weighted (TF-IDF) frequency. Instead, the use of a centroid as well as the adoption of the heuristic presented in Dunietz and Gillick (2014) do not lead to good results, showing the complexity of the task. Based on these initial findings, we will explore in future works how to identify the topic of a document when this is not explicitly mentioned in the content. A possible approach could, for instance, employ relatedness measures to retrieve additional entities from the knowledge base, as already done in similar tasks by Hulpus et al. (2013) and Weiland et al. (2016).

6. Conclusion

In this paper we presented a dataset of political speeches addressed at the UK House of Commons (2013-2016), with fine-grained topic annotations at the document level and enriched with entity links. The corpus is accompanied by two benchmarks on topic classification and ranking.
We envision the use of this dataset and benchmarks for supporting future interactions between the NLP and CSS communities in developing and testing together new algorithms for addressing the topic detection task.

Acknowledgments

This work was funded in part by a scholarship of the Eliteprogramm for Postdocs of the Baden-Württemberg Stiftung (project "Knowledge Consolidation and Organization for Query-specific Wikipedia Construction") and was also supported by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (project C4), funded by the German Research Foundation (DFG).

7. Bibliographical References

Bachmann, I. (2011). Civil partnership - "Gay marriage in all but name". *Corpora*, 6(1).

Cullen, A., Hines, A., and Harte, N. (2014). Building a Database of Political Speech: Does Culture Matter in Charisma Annotations? *Proc. of Audio/Visual Emotion Challenge*.

Dunietz, J. and Gillick, D. (2014). A new entity salience task with millions of training examples. In *Proc. of EACL*, pages 205–209.

Ferragina, P. and Scaiella, U. (2010). TAGME: One-the-fly Annotation of Short Text Fragmetns (by Wikiepdia Entities). *Proc. of CIKM*.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017a). Cross-lingual classification of topics in political texts. In *Proc. of NLP+CSS*.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017b). Unsupervised cross-lingual scaling of political texts. In *Proc. of EACL*.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D. (2013). Unsupervised graph-based topic labelling using dbpedia. In *Proc. of WSDM*, pages 465–474.

Islam, Z. and Mehler, A. (2012). Customization of the europarl corpus for translation studies. In *Proc. of LREC*.

Karan, M., Širinić, D., Šnajder, J., and Glavaš, G. (2016). Analysis of policy agendas: Lessons learned from automatic topic classification of Croatian political texts. In *Proc. of LaTeCH*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit*, pages 79–86.

Lauscher, A., Nanni, F., Ruiz Fabo, P., and Ponzetto, S. P. (2016). Entities as topic labels: combining entity linking and labeled lda to improve topic interpretability and evaluability. *IJCol-Italian journal of computational linguistics*, 2(2):67–88.

Menini, S., Nanni, F., Ponzetto, S. P., and Tonelli, S. (2017). Topic-based agreement and disagreement in us electoral manifestos. In *Proc. of EMNLP*.

Merz, N., Regel, S., and Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, 3(2):1–8.

Mikhaylov, S., Laver, M., and Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91.

Nanni, F., Zirn, C., Glavaš, G., Eichorst, J., and Ponzetto, S. P. (2016). TopFish: topic-based analysis of political position in US electoral campaigns. In *Proc. of PolText*.

Purpura, S. and Hillard, D. (2006). Automated classification of congressional legislation. In *Proc. of dg.o*, pages 219–225.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Ristoski, P., Rosati, J., Di, T., Leone, R. D., and Paulheim, H. (2016). RDF2Vec : RDF Graph Embeddings and Their Applications. *IOS Press*.

Stewart, B. M. and Zhukov, Y. M. (2009). Use of force and civil–military relations in russia: an automated content analysis. *Small Wars & Insurgencies*, 20(2):319–343.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proc. of EMNLP*.

van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., and Beunders, H. (2017). The debates of the european parliament as linked open data. *Semantic Web*, 8(2).

Verberne, S., Dhondt, E., van den Bosch, A., and Marx, M. (2014). Automatic thematic classification of election manifestos. *Information Processing & Management*, 50(4):554–567.

Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009). Canal9: A database of political debates for analysis of social interactions. In *Proc. of ACII*, pages 1–4.

Weiland, L., Hulpus, I., Ponzetto, S. P., and Dietz, L. (2016). Understanding the message of images with knowledge base traversals. In *Proc. of ICTIR*, pages 199–208. ACM.

Zirn, C., Glavaš, G., Nanni, F., Eichorts, J., and Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. In *Proc. of PolText*.