Semantifying the UK Hansard (1918-2018)

Federico Nanni
Data and Web Science Group
University of Mannheim
Germany
federico@informatik.uni-mannheim.de

Stefano Menini*
Digital Humanities Group
Bruno Kessler Foundation
Italy
menini@fbk.eu

Sara Tonelli
Digital Humanities Group
Bruno Kessler Foundation
Italy
satonelli@fbk.eu

Simone Paolo Ponzetto
Data and Web Science Group
University of Mannheim
Germany
simone@informatik.uni-mannheim.de

ABSTRACT

The transcripts of UK parliamentary debates, offered by the Hansard Online collection are a major resource for historians and political scientists. To foster their use, we provide *a)* semantic annotations of over one hundred years of debated motions in the form of disambiguated and entity-linked speakers, *b)* topic annotations, and *c)* topical-clusters of the most frequently addressed issues.

KEYWORDS

UK Hansard, topic extraction, entity linking, parliamentary corpora

ACM Reference Format:
Federico Nanni, Stefano Menini, Sara Tonelli, and Simone Paolo Ponzetto. 2019. Semantifying the UK Hansard (1918-2018). In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'19).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Parliamentary corpora are a very relevant language resource for political scientists, sociologists and historians as well as for computational linguists. One of the first machine-readable resources of transcripts of political speeches is, in fact, the well-known *EuroParl* corpus [4], a collection of parallel texts in 11 languages (later extended to 21 languages [3]) generated from the proceedings of the European Parliament (EP).[1] During the last decade this corpus has become one of the most used resources in Natural Language Processing (NLP) for statistical machine translation, word sense disambiguation, information extraction, anaphora resolution.

The same collection has recently been made available as linked open data [9]: *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, which is organized in over 25 million triples. For debates from the United States, Thomas et al. [8] presented a corpus of speeches from the US Congress. The CLARIN ERIC infrastructure currently provides access to 18 parliamentary corpora, most of them available for direct download,[3] and works toward building a research community by organizing workshops on the topic.[4]

Hansard. Arguably, the largest diachronic corpus of parliamentary proceedings available online is known as *Hansard*, and comprises debates from the UK Parliament (Houses of Lords and Commons) since 1803. This collection has been extended and enriched by different projects: as part of the SAMUELS project, in 2015 the University of Glasgow presented a vast collection under the name "Hansard Corpus"[5] comprising all the parliamentary speeches of the Members of Parliament (MP) from 1803 to 2005, together with a semantic search-tool supporting synonym search, speaker-based search, etc. The corpus is not directly available for download, but it is offered by the Hansard Archives[6] of the UK Parliament. The same archive also provides access to a different collection covering debates in the period 1988 to 2016, organized by date, speaker name, Session and Bill. The platform Hansard Online[7] is another resource offering access to the debates directly from the UK Parliament; initially, it provided speeches dating back to 2010, but recently it has been merged with the Hansard Archives and now it offers query search over more than two hundred years of textual data. Another project that has worked on extending access to the Hansard is DiLiPaD, which provides an interface for semantic search (filtering based on topics, speakers, parties) over the entire proceedings, up until 2014. However, this enriched corpus is not available for download. Finally, there are also private actors that provide advanced access to the Hansard collection. Among them, the website *TheyWorkForYou*, run by the UK charity mySociety, offers House of Commons debates since 1918, with disambiguated names for the speakers.[8]

*The first two authors equally contributed to the paper.
[1] http://www.statmt.org/europarl/
[2] http://purl.org/linkedpolitics

JCDL'19, June 2019, Urbana-Champaign, Illinois USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[3] https://www.clarin.eu/resource-families/parliamentary-corpora
[4] https://www.clarin.eu/ParlaCLARIN
[5] https://www.hansard-corpus.org/
[6] http://www.hansard-archive.parliament.uk/
[7] https://hansard.parliament.uk/
[8] Debates are available for download here: https://www.theyworkforyou.com/pwdata/scrapedxml/debates/

Table 1: Examples of the extracted topics.

Immigration	fair_immigration, illegal_immigration, immigration_act, primary_immigration immigration_control, immigration_officer, immigration_policy, immigration_rules, immigration_system
Social Security	disability_living_allowance, invalid_care_allowance, invalidity_benefit child_benefit, children, married_women
Finance	income_tax, purchase_tax, selective_employment, selective_employment_tax british_economy, second_finance, year_'s_finance
Northern Ireland	fight_against_terrorism, irish_terrorism, prevention_of_terrorism rural_development, northern_ireland_agriculture, department_of_agriculture

2 SEMANTIC ENRICHMENT

As described above, while there exist many online interfaces for searching the UK Hansard Collection, only a few of them permit download of the corpus for further research. These are, for instance, the Hansard Archives, which offer the Hansard Corpus (1803-2005) and the website *TheyWorkForYou*, providing access to debates from 1918 until today. In this work, we expand this second collection, adding new layers of semantic annotation to the debated motions.

Entity-Linked Speakers. The corpus that *TheyWorkForYou* makes available for download is divided into daily XML files, which offer the debates collected directly from the Hansard Online. Each speech has been enriched with meta information about the speaker, identified with a unique ID; for recent MPs, additional data about the party affiliation and the constituency of the speaker are offered on the *TheyWorkForYou* website. Starting from these disambiguated entries and knowing each member of the UK Parliament in every legislation thanks to a highly curated Wikipedia Category on the topic,[9] we have integrated the two resources obtaining, for every speaker in the corpus since 1918, several additional pieces of information such as their party affiliation (and change of party over time), their constituency, and a link to their Wikipedia page.

Annotated Topics. We then follow previous efforts in organizing the Hansard debates into coherent topics [7]. To do so, inspired by the work conducted by Abercrombie and Batista-Navarro [2], we also consider motions as the central unit of analysis, due to the fact that they are proposals that a parliamentarian puts to the other Members and they are essential for understanding the opinions expressed by the MPs during all subsequent speeches.

With a series of hand-crafted rules (as in [1]), we start by detecting the beginning of a motion in the daily Hansard and consider the following debate under the same topic (statistics in Table 2). We then employ the content of the motion alone to identify each debated topic; we do so, due to the fact that motions clearly state what will be subsequently discussed, while the following debate could be shifted in different directions by the speakers. We extract and rank the most relevant key-concepts for each motion using the unsupervised tool Keyphrase Digger [6]; we enrich each motion with the ten most relevant key-concepts as new metadata for supporting the retrieval of the results.[10]

[9]https://en.wikipedia.org/wiki/Category:Lists_of_MPs_elected_in_United_Kingdom_general_elections

[10]The new dataset together with a Jupyter Notebook to browse it are available here: https://federiconanni.com/semantifying-hansard/

Clustered Topics. Next, we employ the obtained results as inputs for Key-Concept Clustering,[11] an algorithm for unsupervised topic detection that we have introduced in a previous work [5]. This approach relies on a recursive procedure that merges key-concepts into meaningful clusters, based on the semantic similarity of their word vectors. This allows us to identify topics related for instance to "immigration", "finance", and "social security" (see Table 1).

Table 2: Statistics of the dataset.

Motions	72k
Speeches	4.9 m
Tokens / Motion	488
Tokens / Speech	170

3 CONCLUSION

We presented three new layers of semantic annotations that allow advanced access to the UK Hansard digital collection. To foster their use in research, we provide a Jupyter Notebook that allows the user to explore a corpus of discussed motions and retrieve debates based on information regarding the speaker, the party affiliation or the addressed topic.

REFERENCES

[1] Gavin Abercrombie and Riza Batista-Navarro. 2018. 'Aye' or 'No'? Speech-level Sentiment Analysis of Hansard UK Parliamentary Debate Transcripts. In *LREC*.
[2] Gavin Abercrombie and Riza Theresa Batista-Navarro. 2018. Identifying Opinion-Topics and Polarity of Parliamentary Debate Motions. In *WASSA*.
[3] Zahurul Islam and Alexander Mehler. 2012. Customization of the Europarl Corpus for Translation Studies.. In *LREC*.
[4] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*.
[5] Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-based agreement and disagreement in us electoral manifestos. In *EMNLP*.
[6] Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the dirt: Extracting keyphrases from texts with kd. *CLiC-IT* (2015).
[7] Federico Nanni, Mahmoud Osman, Yi-Ru Cheng, Simone Paolo Ponzetto, and Laura Dietz. 2018. UKParl: A Data Set for Topic Detection with Semantically Annotated Text. ParlaCLARIN at LREC.
[8] Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *EMNLP*.
[9] Astrid Van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2017. The debates of the european parliament as linked open data. *Semantic Web* (2017).

[11]https://dh.fbk.eu/technologies/key-concept-clustering