A semi-supervised approach to classifying political agenda issues

Tim Kreutz and **Walter Daelemans**
CLiPS - Computational Linguistics Group
Department of Linguistics
University of Antwerp
{tim.kreutz,walter.daelemans}@uantwerpen.be

Abstract

This paper presents a semi-supervised approach to classifying political texts with the Comparative Agendas Project coding scheme. Starting with limited domain knowledge in the form of ten seed words that are central to the meaning of a topic, new candidate textual indicators are found using a graph propagation algorithm over a semantic network of words and phrases. We show that there is a balance between precision and recall when it comes to the number of candidates to add to a lexicon for each topic, and optimize this balance on the basis of a development dataset. The automatically generated lexica substantially outperform the handmade CAP-lexicon in four tested genres: political party manifestos, news articles, parliamentary documents and social media texts. Besides having better discriminatory qualities, these lexica require less resources to generate and are more genre-independent than their handmade counterparts.

1 Introduction

Political experts can analyze newspapers or television channel and summarize the attention given to certain political issues in the media. The Comparative Agendas Project (CAP) (CAP)[1] provides coding schemes in many languages to aid such analyses and make them comparative. But even with clear guidelines, manual coding of political texts becomes prohibitively time consuming.

Recently, more research has focused on automatic content analysis to help expert annotation, especially in the social and political sciences. Two main methods have been proposed: dictionary-based approaches and supervised learning approaches. In dictionary-based approaches an expert-made lexicon is constructed of high precision indicators that are linked to political topics. These indicators are words or other language units. Despite the insight and expertise contained in these topic lexicons, they usually suffer from low coverage over the instances; a dictionary-based system cannot make decisions about documents that contain none of the dictionary words.

In a supervised learning approach, annotators assign labels to a large collection of documents and a machine learning algorithm learns weights between features in the documents and the labels. For example, Purpura and Hillard (2006) classified US congressional legislation using support vector machines and score 88.7% accuracy on major topics and 81.0% on subtopics. This is close to human agreement on the task. There are drawbacks to supervised learning too. A system trained on congressional legislation will perform differently on newspaper articles or social media messages. To achieve consistent results across genres, the classifier would have to be retrained on additional annotated in-genre documents.

We introduce a hybrid solution in this paper; a semi-supervised approach to classifying political texts according the CAP coding scheme for political agendas. Our contribution does not require expert annotation yet improves an existing lexicon-based approach with regards to recall. We obtain these results for two languages (Dutch and English) and for four different genres of political texts namely party manifestos, news articles, parliamentary reports and tweets.

2 Related work

Most previous work on automatic content analysis of political texts with regards to political issues used the coding scheme developed by the Policy Agendas Project (PAP) (John, 2006) and the successive Comparative Agendas Project. The codebook developed by CAP discerns 20 major political top-

[1] https://www.comparativeagendas.net/

ics.

2.1 Dictionary-based

Sevenans et al. (2014) created a Dutch and an English dictionary by taking topic indicators from the respective CAP coding schemes and adding synonyms and related terms by hand. The topic indicators that identify a certain topic can be words or partial words (suffixes, infixes or affixes). In the English lexicon, for example, the topic macroeconomics contains "econom" which will match "economy", "economist", "noneconomic" and many others.

The classification performance of the lexicons differed greatly between the topics and the languages. For the English lexicon, a considerable number of the parliamentary questions did not contain any of the dictionary words (22%) and could not be classified. Interestingly, the parliamentary questions in Dutch did not receive a class in only 5% of the cases. The authors go into detail on the quality of the lexicon for specific topics, but on average, performance was low compared to human annotations: 0.43 recall, 0.52 precision and 0.61 recall, 0.60 precision for English and Dutch, respectively.

Praet et al. (2018) apply the Dutch CAP-lexicon to tweets by Flemish politicians. More than half (54%) of the tweets did not match with any dictionary word, leading to very low classification accuracy.

2.2 Supervised learning

Supervised classification with the CAP coding scheme has been applied to US congressional legislation (Purpura and Hillard, 2006), Norwegian news texts (Hagen, 2012), Kroatian news headlines (Karan et al., 2016), and tweets by US state legislators and governmental bodies (Li, 2016; Qi et al., 2017). There is also a body of work on the supervised classification of party manifestos, which draws its labels from the separate but similar coding-scheme in the Comparative Manifesto Project (CMP) (Zirn et al., 2016; Glavaš et al., 2017).

These standalone applications of machine learning architectures work well in general. Congressional documents are assigned the right major topics in almost 90% of cases while performance drops with shorter texts such as media headlines (0.77% accuracy) and tweets (around 65% accuracy).

As far as we know there has not been an extensive study on cross-domain portability of the supervised classification systems. Rihiu Li (2016) trains a CNN on tweets from state legislators from Iowa and Nebraska. They note that there is a drop in performance when training on data from one state and testing on the other, but this drop could also be due to a smaller training set compared to training on both. Grimmer and Steward (2013) note that supervised machine learning systems are inherently domain- and problem-specific but see this as an advantage over multi-purpose dictionary-based systems.

2.3 Semi-supervised learning

Semi-supervised approaches bootstrap minimal domain knowledge to learn about a problem. As such, the invested expert knowledge and effort are far less than in supervised learning.

Semi-supervised approaches have seen frequent use in sentiment analysis. Rao and Ravichandran (2009) use synonym and hypernym relationships from WordNet to deduce sentiment information. Starting with positive and negative seed terms obtained from the General Inquirer[2] lexicon, polarities are propagated over the word graph using a label propagation algorithm (Zhu and Ghahramani, 2002). Even with as few as ten seeds terms, word polarity scores could accurately be predicted using semi-supervised learning: "*label propagation is especially suited when annotation data is extremely sparse*" (Rao and Ravichandran, 2009).

Kreutz and Daelemans (2018) induce polarity scores without using handmade knowledge graphs. Instead, they take seed words from an existing sentiment lexicon and propagate sentiments to candidate words that appear in similar contexts. The additions made to the existing lexicons improved sentiment analysis for two different domains.

3 Data

3.1 CAP lexicon

We take the CAP-lexicons, which were developed for the Flemish and United States contexts by (Sevenans et al., 2014), as a starting point for semi-supervised learning. The number of indicators linked to a topic ranges from 35 to 102 and 28 to 143 in Dutch and English respectively.

[2] http://www.wjh.harvard.edu/~inquirer/

3.2 Text genres

The datasets for testing were obtained from the Comparative Agendas Project [3] and were created for the Flemish and U.S. CAP-subprojects [4]. We selected the datasets in Table 1 to reflect the diversity in genres while having comparable types of documents across Flemish and U.S. contexts. All documents have major topic codes from the CAP codebook.

4 Methods

4.1 Seed selection

To demonstrate that our approach requires only limited domain knowledge the initial dictionary is restricted to only ten indicators per topic. These seed terms needs to both precisely and frequently denote a topic. We calculate degree centrality between topic indicators in our data sets and select seeds based on this score since it is found to be an effective measure for quantifying such keyword like qualities (Boudin, 2013).

4.2 Extending seed terms

We use a network of words to extend seed terms with other candidates. Edges between words are based on distributional semantics, in which words that occur in similar contexts are more strongly connected. We use the well-established Word2Vec (Mikolov et al., 2013) algorithm to calculate cosine similarities between candidates (words or phrases) to use as edge weights. Our Word2Vec models were trained on the Google News dataset for English[5] and an unpublished corpus of Dutch newspaper and online news data with 300-dimensional vectors and negative sampling.

Suitable candidates are found by doing random walks over the network of words starting from the selected seeds. This method is adapted from SentProp (Hamilton et al., 2016), a package originally intended for inducing sentiment lexicons. We adopt its default settings of connecting words to their ten nearest neighbors. A word or phrase which is found often by a random walk from a seed get a higher score for that topic, while a penalty is applied if the word or phrase is found from a seed term that is linked to another topic. A ranking of candidates by scores then determines in which order they be added to the lexicon.

4.3 Determining a cut-off

We split the annotated data in a stratified development and test set (50% each). The development set is used to determine a cut-off for candidate words. As seed terms and candidates become more dissimilar, adding more candidates can harm the discriminative performance of the dictionary. Figure 1 shows precision, recall and F-score for the Civil Rights topic on development data at different numbers of added candidates. Although recall improves when more indicators are being added for this topic (more documents are classified as belonging to Civil Rights), precision suffers. The cut-off is determined as the highest harmonic mean of precision and recall (the F-score optimum).

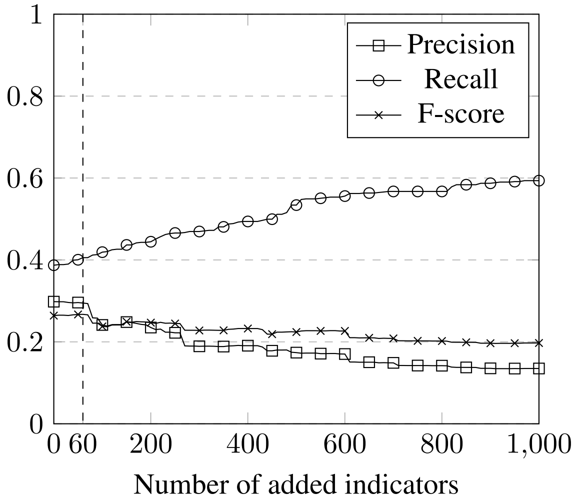

Figure 1: As the number of words added to *Civil Rights* increases, recall improves and precision decreases. The optimal number of added words (60) lies at the F-score optimum.

4.4 Classification algorithm

Classifying a text is done by simply checking if any of its words appear in the lexicon for one of the topics. We assign the document a label based on the topic that occurred most often. Although a more refined algorithm could be used to take into account class distribution or to assign different weights to words, using a simple classification algorithm ensures that each entry precisely denotes a topic and does not greatly offset the decision boundaries.

[3] https://www.comparativeagendas.net/datasets_codebooks

[4] Full citations for the Belgian datasets, the U.S Democratic and Republican party platform datasets, the U.S. bills dataset and NYT dataset in acknowledgements 1, 2, 3 and 4 respectively. (excluded here for anonymous review)

[5] Available at https://code.google.com/archive/p/word2vec/

	Belgium		U.S.	
Domain	Type	# Documents	Type	# Documents
Manifestos	Party manifesto excerpts	5,147	Party manifesto excerpts	7,296
News media	De Standaard abstracts	17,981	New York Times abstracts	17,216
Parliament	Bills	4,868	Congressional bills	52,366
Social media	Tweets by politicians	6,027	Tweets by state legislators	16,988

Table 1: Data spanning four genres and two contexts is used to tune and evaluate the semi-supervised approach.

	U.S.				Belgium			
Method	Precision	Recall	F1	Entries	Precision	Recall	F1	Entries
Original lexicon	0.308	0.207	0.210	3,610	0.446	0.408	0.378	8,353
Seed selection	0.311	0.262	0.246	200	0.428	0.367	0.346	200
Induced lexicon	**0.321**	**0.286**	**0.278**	9,976	**0.448**	**0.422**	**0.387**	18,650

Table 2: The macro-averaged results of the lexica created with SentProp compared to the original hand-made lexicon and the lexica containing only seed terms.

5 Results

Adding the optimal number of candidate indicators to the seeds results in the final induced lexica for testing. The lexica contain 9,976 and 18,650 words for the U.S. and Belgian context respectively. Table 2 lists their results compared to using the original lexicon of hand-picked words and the seed lexicon on the test data.

The induced lexica outperform the original lexica both in the U.S. and Belgian context and not only with regards to recall. Surprisingly, the added candidates also more precisely denote a topic compared to a handmade lexicon. We regard this latter result mainly as a demonstration of how difficult it is, even for experts, to construct a dictionary that can distinguish between topics in a real-world setting by hand.

In absolute terms, results are still rather poor. This is to be expected considering that a lexicon distinguishes 20 different major political topics, and that some genres, social media texts in particular, contain very little information. Another problem is the genre independence that a classifier needs to work on party manifestos, news, bills as well as social media texts. We believe a supervised classifier trained on other political texts will face the same difficulty in deciding on the right label, although future work will have to compare these approaches in detail.

6 Conclusion

We introduced an easy to use semi-supervised approach for inducing dictionaries suitable for classifying diverse political texts. The induced dictionaries outperform a handmade lexicon from the Comparative Agendas Project across contexts (U.S. and Belgium) and genres (political party manifestos, news articles, bills and social media texts).

Creating lexica in an automatic way is less time consuming while remaining as interpretable and easily adaptable as existing dictionary-based approaches; the words or phrases can be inspected and changed by experts when necessary. Future work should compare the semi-supervised method with supervised models, both in terms of overall performance and in diverse cross-genre settings.

6.1 Data and code availability

All datasets used in this paper, except for the tweets which cannot be freely shared due to GDPR[6] restrictions, are available from the CAP website.

To enable replicability and direct comparison in future work, we publish our method in a public code repository. Alongside the code we present the induced lexica for both U.S. and Belgian contexts here: `https://github.com/clips/lextension`.

References

Florian Boudin. 2013. A comparison of centrality measures for graph-based keyphrase extraction. In *Proceedings of the sixth international joint conference on natural language processing*, pages 834–838.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-lingual classification of top-

[6] `https://gdpr.eu`

ics in political texts. Association for Computational Linguistics (ACL).

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

Thomas M. Hagen. 2012. Automatic topic classification of a large newspaper corpus. In Gisle Andersen, editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, chapter 6, pages 111–130. John Benjamins Publishing, Oxford.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 595. NIH Public Access.

Peter John. 2006. The policy agendas project: a review. *Journal of European Public Policy*, 13(7):975–986.

Mladen Karan, Jan Šnajder, Daniela Širinić, and Goran Glavaš. 2016. Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 12–21.

Tim Kreutz and Walter Daelemans. 2018. Enhancing general sentiment lexicons for domain-specific use. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1056–1064.

Rihui Li. 2016. Classification of tweets into policy agenda topics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Stiene Praet, Walter Daelemans, Tim Kreutz, Peter Van Aelst, Stefaan Walgrave, and David Martens. 2018. Issue communication by political parties on twitter. In *Data Science, Journalism & Media 2018, August 20, 2018, London, UK*, pages 1–8.

Stephen Purpura and Dustin Hillard. 2006. Automated classification of congressional legislation. In *Proceedings of the 2006 international conference on Digital government research*, pages 219–225.

Lei Qi, Rihui Li, Johnny Wong, Wallapak Tavanapong, and David AM Peterson. 2017. Social media in state politics: Mining policy agendas topics. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 274–277.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics.

Julie Sevenans, Quinn Albaugh, Tal Shahaf, Stuart Soroka, and Stefaan Walgrave. 2014. The automated coding of policy agendas: A dictionary based approach (v. 2.0.). In *CAP Conference*, pages 12–14.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. *CMU-CALD-02-107*.

Căcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos.