Towards a text analysis system for political debates

Dieu-Thu Le
IMS, Institute for NLP
University of Stuttgart
Germany

Ngoc Thang Vu
IMS, Institute for NLP
University of Stuttgart
Germany

Andre Blessing
IMS, Institute for NLP
University of Stuttgart
Germany

{dieu-thu.le,thang.vu,andre.blessing}@ims.uni-stuttgart.de

Abstract

Social scientists and journalists nowadays have to deal with an increasingly large amount of data. It usually requires expensive searching and annotation effort to find insight in a sea of information. Our goal is to build a discourse analysis system which can be applied to large text collections. This system can help social scientists and journalists to analyze data and validate their research theories by providing them with tailored machine learning methods to alleviate the annotation effort and exploratory facilities and visualization tools. We report initial experimental results in a case study related to discourse analysis in political debates.

1 Introduction

The overall goal of our project is to develop an interactive research environment for text collections that (a) puts state-of-the-art text analysis models from Computational Linguistics in the hands of social scientists or data journalists, allowing them to quickly tailor search facilities and filters to their research goal, i.e., finding and categorizing textual passages in the collection that instantiate a relevant position towards an issue under exploration. The environment furthermore (b) relates the categorized positions, or claims, to the uttering actors, capturing dates of utterance, the relation to relevant mentioned entities, and (c) provides exploratory facilities and visualization tools for performing time-series analysis and network analysis on aggregated text-analytical results, including differential analysis against trends observed in previous legislation processes. By keeping all backward links from aggregated results to the individual underlying text sources, the environment readily supports (d) a critical assessment of the analysis and (e) a transparent presentation of the data basis of a news story.

A major side-effect of the project is to engage in an exchange among two different explorative points of view towards large heterogeneous data collections: social scientists and journalists on the one hand have certain intuitions and strategies how to proceed when they first approach a collection which they suspect to contain some newsworthy evidence. They cannot know however which substeps in their strategy can be supported or taken over by sufficiently reliable automatic means. Computational linguists on the other hand have a wide range of analytical tools at their disposal, they know how to adapt them to specifics of some application context, and they are able to combine tools to solve more complex structural questions about a text. However, ideas for completely novel types of complex analytical questions about a text collection have to come from outside of Computational Linguistics - so professional investigators of novel questions are highly interesting partners for developing explorative strategies.

In the next sections, we will report the first experimental results, which were carried out on an already annotated dataset to illustrate how the system could be used to assist social scientists and journalists to analyze data.

2 Approach

Argumentation mining is an arising research topic (Peldszus and Stede, 2013; Moens, 2013) which models argumentation in textual content. Most theories propose that each argumentation consists of two parts: i) the premise and ii) the conclusion/claim. For discourse network analysis only claims and the actor behind is relevant. Further-

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 134–139,
Berlin, Germany, August 11, 2016.

more, our first analysis of existing labeled data showed that there are large divergences in the way claims are annotated in the different communities. Thus, we have chosen a task-driven approach, instead of a theory-driven approach, which is defined by actual questions of the journalists and social scientists on large text collections. Which means, that we follow a supervised approach since we use a seed of already annotated text segments. Nevertheless the annotation[1] process is also well-defined by complex codebooks (Koopmans, 2002).

3 Case Study: The debate of nuclear power phase-out

In March 2011, Japanese earthquake and tsunami caused a nuclear accident in Fukushima, which prompted a critical re-thinking of nuclear power. Germany witnessed a radical political change towards an accelerated phasing out of nuclear reactors as an immediate reaction to the disaster. The sudden changes in decisions could not be explained by traditional political science theories. A few months before the accident, an agreement related to prolonging of nuclear energy use had been made, but was quickly withdrawn after the energy debate and set the final exit date to the year 2022.

A political science group in Bremen (Haunss et al., 2013) has proposed using discourse network analysis to find a plausible explanation. They examined articles in two Germany newspapers published during this time. They argued that actor centrality, consistency and cohesion of discourse coalition could be used to explain the fast development in political changes.

4 Problem statement

The problems of identifying factors for text analysis of the political science group could be stated in machine learning tasks as follows (Figure 1):

Claim vs. Non-claim classification In our case study, claims are defined to be sentences related to political opinions and decisions of actors, while non-claims are general statements without content about political decision. In the first step, claims are extracted from articles. We train a claim classification that learns from some pre-annotated claims and help the annotators to automatically find other relevant claims.

[1]Social scientists use often the term *coding* instead of *annotation*.

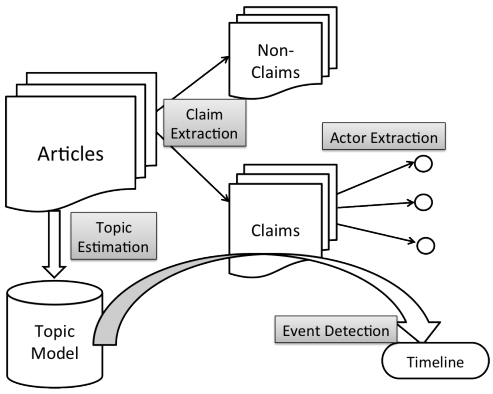

Figure 1: A computational linguistic pipeline for text analysis with main steps: Claim, actor extraction and event detection

Actor extraction One major part of the discourse analysis is to identify actors associated to each claim. We argue that using Named Entity recognition, the system can propose possible candidates for each claim and help annotators to select correct actors faster. The names of actors are usually mentioned within a claim itself or within the article where the claim is stated. By proposing a ranked list of named entities of type Person and Organization, the annotators can browse through the list of suggestions and select the correct one.

Topic estimation, trend and event detection In this pipeline, we use topic models (Blei et al., 2003) as a way to browse and summarize articles by dates and find out which topics/events are important. Firstly, a topic model is estimated from all articles. After that, we use this model to infer topics for claims grouped by dates. The topic distribution over time can be used to detect important events and to have an overview of what topics were discussed during which time.

5 Models and experiments

5.1 Term extraction

Figure 2 shows top terms that appear in claims and non-claims using term frequency (TF) and term extraction (TE). In term frequency, we counted how many times a term appears in all claims or non-claims. In term extraction, we compare how important a term is in the dataset in compared to the term appearing in a reference corpus, which is a collection of online German news articles.

The first glance at the top extracted terms from claims and non-claims suggests that terms in both categories are very similar. A traditional bag-of-

Claims		Non-claims	
TF	TE	TF	TE
Ausstieg	Salzstocks	Deutschland	Kernkraftwerke
Deutschland	designierte	Merkel	japanischen
Energien	Suchraum	Grünen	Kraftwerke
Kernenergie	Standortsuche	Japan	Teyssen
Merkel	unumkehrbaren	CDU	Reaktoren
Kernkraftwerke	geologische	Prozent	Atomkraftwerke
Atomausstieg	potenziellen	Bundesregierung	Moratoriums
Energiewende	gesetzliche	deutschen	Kernkraftwerken
deutschen	einzuspeisen	Jahr	Sicherheitsstandards
CDU	Endlagers	Ausstieg	Atomkraftwerken
Bundesregierung	Beeckens	FDP	warnte
Netz	Entsorgungskommission	Regierung	Fukushima
Meiler	inhaltlichen	Euro	nuklearen
müssen	erweitere	Atomausstieg	Huhne
Atomkraftwerke	Kaltreserve	Fukushima	bayerischen

Figure 2: Term extraction from claims and unlabeled data

word approach may not be sufficient to distinguish them to suggest appropriate claims for the annotators. Following, we present our claim classification method using deep learning to automatically detect important features for finding claims.

5.2 Claim classification

5.2.1 Settings

Claim classification can be considered as a sentence classification task. Hence, we applied convolutional neural networks (CNNs) - a state-of-the-art method (Kalchbrenner et al., 2014; Kim, 2014) for this task. CNNs perform a discrete convolution on an input matrix with a set of different filters. The input matrix represents a sentence, i.e. each column of the matrix stores the word embedding of the corresponding word. Word embedding can be randomly initialised or pre-trained with unsupervised training method. In both cases, we fine-tuned the embeddings during the network training. By applying a filter with a width of e.g. three columns, three neighbouring words (trigram) are convolved. Afterwards, the convolution results are pooled. In this work, our model used filters of width 3-5 with 100 filters each. Following (Collobert et al., 2011), we perform max-pooling which extracts the maximum value for each filter and, thus, the most informative n-gram for the following steps. Finally, the resulting values are concatenated and used for claim classification. To train the network we used stochastic gradient descent with a mini-batch size of 50 and AdaDelta (Zeiler, 2012) to adapt learning rate after each epoch. We pre-trained word embeddings with word2vec[2] using 99M German sentences collected from the news and Wikipedia. Motivated by the fact that claims are independent from person or organization, we replaced all named entities with NE tags to improve the generalization of the network.

[2]https://code.google.com/archive/p/word2vec/

5.2.2 Results

In total, we have 1,837 sentences which are manually annotated as claims and 12,033 non-claim sentences. It is, however, not clear whether non-claim sentences are manually cross checked (if all non-claim sentences contain no claim at all). Furthermore to balance the claims:non-claim ratio, we randomly picked only 1,837 non-claim sentences. Table 1 summarized the average F1-scores on a 10-fold cross-validation with different experimental setups. Our results revealed that using pre-trained word embeddings and replacing all named entities with their corresponding tags are useful to improve the final performance.

Table 1: F1 score for claim classification

Systems	F1-score
using random initialized word embs	67.5%
+ replace NEs	68.5%
using pretrained word embs	70.3%
+ replace NEs	70.6%

5.3 Named Entities

We applied Named Entity recognition using Conditional Random Field explained in (Finkel et al., 2005) and the German model prepared by (Faruqui and Padó, 2010) to recognize entities in all claims. We used Person and Organization named entities to prepare a list of suggested actors for each claim.

We carried out two experiments: in the first one, only sentences where claims are annotated were used to extract named entities from; and in the second one, we further expanded to all sentences in articles that contain claims. The results are shown in Table 2, where 71.2% of actors could be found within the suggested named entity list extracted from articles where claims are annotated.

Table 2: Percentage of actors detected using NER in claims

using only sentences containing claims	51%
using articles containing claims	71.2%

5.4 Topic browsing - trend detection

Firstly, we estimated a topic model with 20 topics from all articles. Then we grouped claims by dates and inferred topics for these claims. We provide a visualization tool for social scientists to perform time-series analysis. Figures 3, 4, 5 show the topic distribution of claims over time. Figure

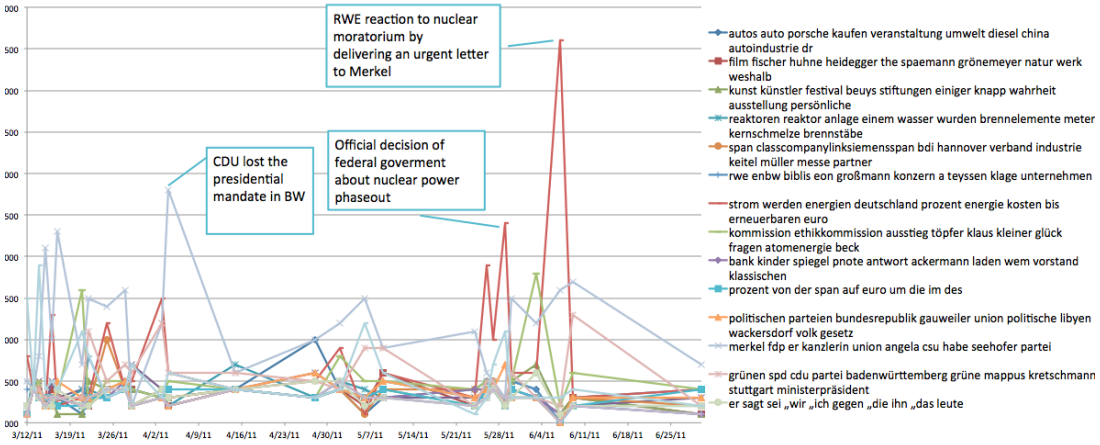

Figure 5: Topic timeline of claims related to CDU and Angela Merkel

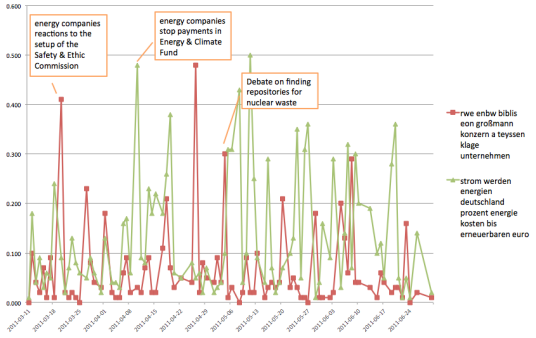

Figure 3: Discussion related to energy changing and energy companies

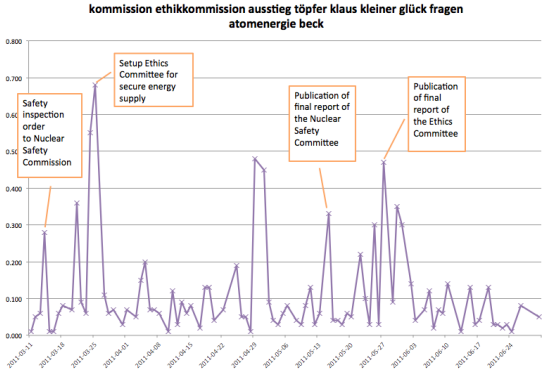

Figure 4: Timeline of discussion related to security and ethic commissions

3 shows that discussions related to the topic of energy changing heated up after the nuclear catastrophe in Japan, which involves statements of energy companies, their reactions and debates on problems such as payments in the energy and climate funds, finding repositories for nuclear waste. Important events related to the setup of security and ethic commissions to examine the safety of nuclear reactors can be spotted from Figure 4.

Finally, we grouped claims based on actors and do topic inference for these claims over time. Figure 5 shows an example of a topic timeline for the CDU party and Angela Merkel. Some events related to the election results and nuclear company reactions to the government can be spotted from the timeline (e.g., election in Baden-Württemberg (BW) - the first time CDU lost the presidential mandate, final decision of the federal state regarding nuclear phaseout, an energy company suing the government).

6 Related work

Textual content analysis in social science is still a handcrafted discipline which requires manual annotations (Baumgartner et al., 2008; Bruycker and Beyers, 2015; Koopmans and Statham, 1999). The main drawback besides the expensive manual work is that for each research questions the whole process has to be repeated. In contrast to other content analysis systems (Bamman and Smith, 2015; Qiu et al., 2015; Levy et al., 2014; Slonim et al., 2014) our approach can be seen as a bottom-up task-driven approach instead of a top-down approach based on the theory of argumentation (Moens, 2013).

7 Conclusions

In this paper, we have presented our first experimental results on building a tool to facilitate research in political and social science using discourse analysis. In particular, we focus on three tasks involving claim extraction, actor identifica-

tion and timeline visualization for detecting important events and topics. In our case study, all data has been manually annotated. Our initial results show that this manual annotation process can be accelerated with the assistance of tailored state-of-the-art machine learning systems: for claim extraction, a fine-tuned word embedding system can achieve up to 70% F1-score when taking into account automatically tagged persons and organizations; for actor extraction, 71% of actors can be found using named entity recognition. Finally, we show how topic timelines could be used to spot important events related to the debate.

Acknowledgments

This research was supported by CRETA - Center for Reflected Text Analytics funded by the German Federal Ministry of Education and Research (BMBF) and by the project DebateExplorer funded by the VolkswagenStiftung.

References

David Bamman and Noah A. Smith. 2015. Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 76–85, Lisbon, Portugal, September. Association for Computational Linguistics.

Frank R Baumgartner, Suzanna L De Boef, and Amber E Boydstun. 2008. *The decline of the death penalty and the discovery of innocence*. Cambridge University Press.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Iskander De Bruycker and Jan Beyers. 2015. Balanced or biased? interest groups and legislative lobbying in the european news media. *Political Communication*, 32(3):453–474.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych, and Seid Muhie Yimam. 2014. Webanno: a flexible, web-based annotation tool for clarin. *Proceedings of the CLARIN Annual Conference (CAC) 2014*, October.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sebastian Haunss, Matthias Dietz, and Frank Nullmeier. 2013. Der Ausstieg aus der Atomenergie. Diskursnetzwerkanalyse als Beitrag zur Erklärung einer radikalen Politikwende. *Zeitschrift für Diskursforschung*, 1(3):288–316.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Ruud Koopmans and Paul Statham. 1999. Political claims analysis: integrating protest event and political discourse approaches. *Mobilization: An International Quarterly*, 4(2):203–221.

Ruud Koopmans. 2002. Codebook for the analysis of political mobilisation and communication in european public spheres. http://europub.wzb.eu/Data/Codebooks

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *COLING*, pages 1489–1500.

Marie-Francine Moens. 2013. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation*, page 2. ACM.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Minghui Qiu, Yanchuan Sim, Noah A. Smith, and Jing Jiang. 2015. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In Suresh Venkatasubramanian and Jieping Ye, editors, *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, pages 855–863. SIAM.

Srikrishna Raamadhurai, Oskar Kohonen, and Teemu Ruokolainen. 2014. Creating custom taggers by integrating web page annotation and machine learning. In *Proceedings of the Conference System Demonstrations , COLING*, pages 15–19.

Noam Slonim, Ehud Aharoni, Carlos Alzate Perez, Roy Bar-Haim, Yonatan Bilu, Lena Dankin, Iris Eiron, Daniel Hershcovich, Shay Hummel, Mitesh M. Khapra, Tamar Lavee, Ran Levy, Paul Matchen, Anatoly Polnarov, Vikas C. Raykar, Ruty Rinott, Amrita Saha, Naama Zwerdling, David Konopnicki, and Dan Gutfreund. 2014. Claims on demand - an initial demonstration of a system for automatic detection and polarity identification of context dependent claims in massive corpora. In Lamia Tounsi and Rafal Rak, editors, *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, August 23-29, 2014, Dublin, Ireland*, pages 6–9. ACL.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.