Discovering Links between Political Debates and Media

Damir Juric[1,3], Laura Hollink[2], and Geert-Jan Houben[1]

[1] Delft University of Technology
[2] VU University Amsterdam
[3] FER University of Zagreb

Abstract. Politics and media are heavily intertwined and both play a role in the discussion on policy proposals and current affairs. However, a dataset that allows a joint analysis of the two does not yet exist. In this paper we take the first step by discovering links between parliamentary debates in a political dataset and newspaper articles in a media dataset. Our approach consists of 3 steps. We first discover topics discussed in the debates. Second, we query a newspaper archive for relevant articles using a combination of debate elements: dates, actors, topics, and named entities of the debates. Finally, we discover links, represent them in RDF, and make them available for download. An evaluation of various versions of this approach shows that the topic detection adds to the quality of the discovered links, as well as the use of the semantic structure of the debate, such as headers and a division into smaller events.

Keywords: RDF, parliamentary debates, NER, topic modeling, linking.

1 Introduction

In this paper we present our work on the linking of political debates to media articles. We present the design choices for a model in which we capture parliamentary debates, including how they are covered by various media, and we describe the method for automatic linking between the speeches inside the debates and various media articles.

To make comparisons between different types of media outlets, links between datasets would need to be produced. Such links could, for example, support researchers that want to know how political debates are represented in the media and how the representation of topics and people change over time. We aim to facilitate this kind of analysis by providing links between datasets of political debate events and media data.

The presented method for link discovery aims to connect debate content on a speech level with relevant articles that contain not just the mentions of speakers but also mentions of speakers in a context of topics that politicians tackled in their speech in parliament. This goal made task much harder, because context of the consequential speeches is often very similar (politicians are speaking about same general topic of the day). We had to extract enough semantics from each particular speech so that we could retrieve more articles reporting on the particularities of politician's speech and

F. Daniel, P. Dolog, and Q. Li (Eds.): ICWE 2013, LNCS 7977, pp. 367–375, 2013.

not just mentioning his name in the context of general topics of the debate (which is also often the case). We used semantic and information retrieval techniques to generate automatic queries that contain the context of the parliamentary speeches and to search newspaper dataset for the connections between speeches and newspaper articles that are covering them. Since the debate transcripts that we use as a source posses structural elements such as debate and conversation descriptions, we wanted to explore if using these elements will help us in our goal of creating the debate-media dataset.

This paper is organized as follows. First, we describe the PoliMedia project in which this work is carried out. In Section 2 we present our method for discovering links between speeches and the media articles. In Section 3 we evaluate the method on a randomly generated dataset, and finally in Section 4 we conclude our work.

1.1 Related Work

We draw on previous work from various domains: other projects using parliamentary debate data, event modeling, relatedness discovery, topic modeling, and entity linking.

[1] presents an approach that extends existing metadata enrichment processes with a method to discover historical events. In [2], the authors put events as the central elements in the representation of data from domains such as history, cultural heritage, multimedia and geography. The Simple Event Model (SEM) is created to model events in these various domains, without making assumptions about the domain-specific vocabularies used. In [3] the authors describe a real life problem using SEM. The problem of link discovery is tackled in [4]: it presents a validation approach of detected alignment links between dialog transcript and discussed documents, in the context of a multimodal document alignment framework of multimedia events (meetings and lectures). In [5] the authors present a function that discovers relatedness between news articles across four aspects: relevance, novelty, connection clarity, and transition smoothness. Although, our work does not perform the same task (we do not have a knowledge base, and we are interested in topics and not just in the named entities), this field of work is related to ours. In [6] authors describe the system that disambiguate entity mentions in text and link them to a knowledge base. They approach readily ports to knowledge bases other than Wikipedia. The Text Analytics Conference on Knowledge Base Population (TAC-KBP) included the task of entity linking [7]. Some of the examples include the use of information retrieval techniques for retrieving the correct knowledge base entry, such as query expansion [8], and generative clustering models for entities in text based on knowledge base entries [9]. In [10] authors presented algorithm for linking between two different archives, the news archive as a source and multimedia archive as target. Although the problem is similar to ours, in this task target archive is heavily annotated by human annotators. Domain experts extracted entities and topics manually, which is different from our case. In [11] retrieval techniques are used to link between four different encyclopedias. They report 40% precision at 100% recall. In our case, we do not have the similar or same type of articles (like encyclopedia articles) but noisy spoken text and concise newspaper articles.

2 Linking Speeches from the Debates to Media Articles

Our PoliMedia method consists of three steps. First, we enrich the existing debate metadata with topics. Second, for each speech, we search the archive for candidate articles based on when they were published (7 days after the debate) and occurrence of the name of the speaker of the speech. Finally, we rank these candidate articles based on similarity to the query (automatically created from speech text) by comparing vectors of topics and named entities. We create links between a speech and an article if the similarity score is above a threshold t.

2.1 Topic Modeling

For each speech inside a debate segment (called *PartOfDebate* in our method) we extract ten words that represent one topic discussed inside the speech. Also all speeches contained inside one debate segment are concatenated into one text and the set of ten words that represent one topic of the debate segment as a whole is then extracted from that text. Debate is queried for all *PartOfDebate* identifiers, which represents a number of different topics that are being discussed in this particular debate. The *PartOfDebate* identifiers contain properties that lead to the actual speeches and their descriptions (*DebateContext*). At this point we create two vectors that are populated with named entity values (those values are objects in statements '*DebateContext mentions* $NEs_{context}$' and '*Speech mentions* NEs_{speech}'). The objects of the *hasSpokenText* and *hasText* properties are taken and sent for preprocessing. In the preprocessing step text we remove all the words that have high frequency but bring a small amount of information. In [12] it is stated that probabilistic topic models are a popular tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus, and that it can be used to check models, summarize the corpus, and guide exploration of its contents. Topic models lead to semantically meaningful decompositions of text because they tend to place high probability on words that represent concepts, and documents are represented as expressions of those concepts. We used a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, etc., called Mallet [13]. Mallet uses a fast and highly scalable implementation of Gibbs sampling. [14].

2.2 Search for Candidate Articles and Ranking

In the second step, we preselect data by fetching all available media from secondary datasets by using the Search and Retrieve protocol (SRU). Preselected data contains only those articles in which the name of the speaker from the debate can be found in a time span of seven days after the speech has been spoken in the parliament.

In the third step, the transcript of the parliamentary debate (in XML format) is used as a primary source for the task of finding links between the speeches that the debate consists of and the media articles. Each debate contains some basic metadata that depicts a debate as a whole (it should be noticed that one debate can actually contain more than one part with different topics (spoken on a same day) that we call *part of the debate* in this article), of which the date when the speeches were spoken in the Dutch parliament is the most important one. Each part of the debate (collection of speeches that are all about single theme) has its description that consists of an unstructured text. Each article is treated as a document D for which we calculate the similarity with the previously automatically created query Q_{exp} (that should represent the context of the debate). We pose our task of finding relevant newspaper articles for a speech in a debate as an information retrieval problem using the vector space model, where we consider the speech as the query Q and the newspaper article as the document D. Fetched candidate articles are tokenized, stripped of stop words, and indexed. Each article D is represented as a term vector $\vec{d}$ of length n, where n is the length of the total number of terms in our corpus of candidate articles. The elements of $\vec{d}$ are term frequency–inverse document frequency (TF-IDF) scores. We create vectors for each speech in the debate, made up of topic sets as discussed in Section 2.1 and named entities (NEs) that are associated to the speeches with the *polivoc:mentions* property. Similarly, we create vectors for topic sets and NEs derived from the *debateContext* of the speech. Element *debateContext* is a short description of the subjects that will be addressed in the forthcoming debate segment, that is read by the chairman (*voorzitter*) of the debate. Given the *debateContext*, we detect topic sets from the text of all speeches that fall under the same *debateContext*. We select NEs mentioned in the introductory text of the *debateContext*. In this way, the semantic structure of the debate enables us to treat the speeches not as isolated pieces of text but as part of a broader conversation.

In total, this process results in 4 vectors for each speech: $\mathrm{NEs}_{\mathrm{speech}}$, $NEs_{context}$, $\mathrm{Topics}_{\mathrm{speech}}$, $Topics_{context}$.

For each speech, we measure similarity between the article vectors of the candidate articles and the four debate speech vectors that represent the speech. We use two state-of-the-art measures: the cosine similarity measure and the BM25 similarity measure. Standard cosine similarity is used because it is proven to work best for this type of comparisons [4]. Both measures produce similar matches but with different rankings. Since ranking is not our primary concern (our goal is to populate the dataset with relevant documents, regardless of their ranking) we choose the cosine similarity measure because it suits better to our needs. Because cosine similarity values range between 0 and 1 it was easier to find a threshold for what we take as a relevant document, then with BM25. Articles with a score above a threshold (0.01) are linked to the speech with the *polivoc:coveredIn* property. We exclude articles where only one (or few) vector contributes to the high similarity score by means of an overlap measure. The overlap coefficient is related to the Jaccard index that computes the overlap between two sets. The method pipeline is presented in Fig. 1. The debate transcript serves as an entry point.

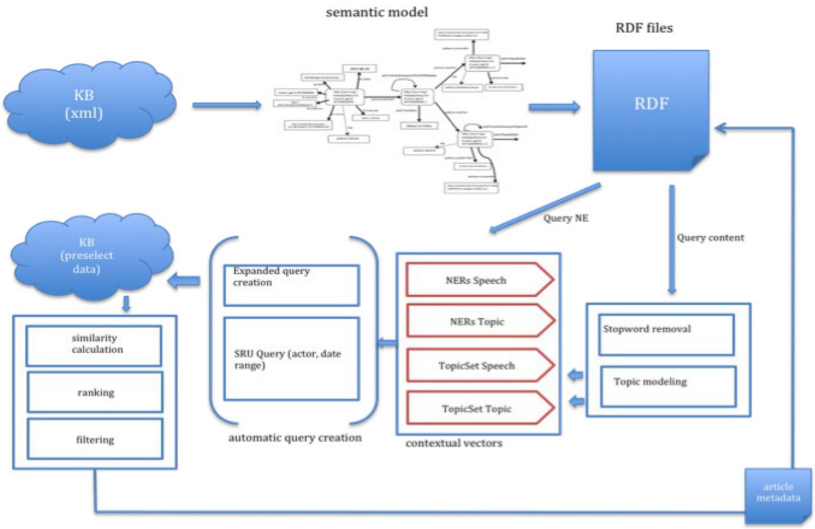


Fig. 1. Method pipeline

3 Experiments

To gain insight into the quality and added value of the various steps of the linking method described in the previous section, we have performed experiments with three versions of the method. Specifically, we have varied which information is used to rank the candidate articles (named entities (NEs), topics) and whether the *partOf* relations between speeches and larger parts of debates are used to also include information associated to these larger parts (debate segments).

Experiment 1: NEs in speech In the most simple form of our method, we rank articles only based on the NEs found in the speech.
Experiment 2: NEs + topics in speech Here, we include not only NEs but also topics detected for the speech.
Experiment 3: NEs + topics in speech and context Finally, we include not just NEs and topics extracted from the speech itself but combine those with NEs extracted from the debate context and topics extracted from all speeches in this context.

The method to query the media archive and select candidate articles is kept constant: we query the archive for articles that mention the name of the speaker and were published within 7 days after the date of the debate. The value 7 is based on our own estimation of the time media takes to write about political debates. Since it is kept constant, a potential suboptimal value will not affect the results. In the future, we intend to investigate what is the optimal value to produce the highest quality of links.

For the experiments, we have randomly selected 20 debates from our dataset of 10,924 debates. The subjects of those debates ranged from fraud in the social system to the European elections. In all three experiments we have linked speeches from within these 20 debates, thus limiting the effect of variations in debate topics on the quality of the resulting links. Second, in each of the three experiments, we have randomly selected 50 speeches from the 20 debates, and linked these to newspaper articles. One speech can be linked to multiple articles, but for evaluation purposes we have randomly selected one linked article for each speech. As a result, we have 150 speech-article pairs, namely 3 sets of 50 each.

We used two independent evaluators to read the speeches and articles linked to them and manually assess their relatedness. Rating was done on a 3-point scale. A score of *0* score is given if the name of the politician is mentioned in the newspaper article in a context that is unrelated to the subjects of the speech the politician gave in parliament; A score of *1* is given if the article mentions the politician in the context of the debate as a whole, but not specifically in the context of the particular speech; A score of *2* is given to all the articles that mention name of the politician in the context of his particular speech (X is given when evaluators can't decide which score to give).

The left part of Table 1 (dataset **Eval$_{\text{NESpeech}}$**) shows the average number of relevant, partially relevant and unrelated links found in Experiment 1. The complete dataset created with these parameters (using just NEs from the speech) is considerably bigger than other two datasets, as a result of using the least specific query (the whole dataset contains 5887 linked articles). Also, for the same reason this dataset contains a large number of unrelated articles. In [15], the authors stated that NEs play an important role in news documents. They wanted to exploit that characteristic by considering them as the only distinguishing features of the documents. In our experiments we found out that using just NEs is not enough to distinguish between newspaper articles. For that reason we included an additional element, the topics from the speech. Results of the evaluation of our method with that additional parameter can be seen in the middle part of Table 1 (dataset **Eval$_{\text{NESpeechTSpeech}}$**). It is visible that the second dataset represents an improvement over the first dataset in terms of quality (this dataset contained 4449 linked articles in total).

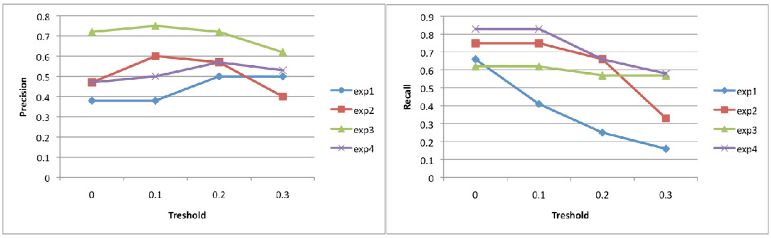
Table 1. Evaluation of produced links based on NEs from speech and on NEs and topics from speech speech and debate description element

Eval$_{\text{NESpeech}}$	Eval1	Eval2	**Eval$_{\text{NESpeechTSpeech}}$**	Eval1	Eval2	**Eval$_{\text{All}}$**	Eval 1	Eval 2
0	20/40%	18/36%	0	10/20%	13/26%	0	3/ 6%	9/18%
1	13/26%	16/32%	1	16/32%	20/40%	1	**17/34%**	**19/38%**
2	11/22%	8/16%	2	15/30%	11/22%	2	**24/48%**	**20/40%**
X	6/12%	8/16%	X	9/18%	6/12%	X	6/12%	2/ 4%

Finally, we produced the third dataset by harvesting the debate structure. We used NEs and topics from debate descriptions to create a query that is more specific than both previous queries. In **Eval$_{\text{All}}$** we can see that the resulting dataset has the best quality (for our purpose that means the biggest number of relevant links, with scores 1 or 2). This dataset contained 3804 linked articles in total. Evaluator agreement (Cohen's Kappa) was 0.5207, which represents a moderate agreement.

Table 2. Evaluation of produced links based on speech and debate description element

Recall - To calculate recall we had to conduct a different kind of evaluation. Since for each speech we have a different query, only way to calculate meaningful recall was to analyze the speech, create the query manually and then search the library portal in the same time span as our algorithm. Evaluator task was to analyze five arbitrarily chosen speeches and to manually create a query that should retrieve all articles containing those terms in the given context. Then evaluator had to analyze articles retrieved from using the manual queries (115 newspaper articles) and to decide how many of them are relevant to the particular speech. With settings as in experiment nr.3 recall was 62% with precision 75% (using the same threshold as for previous evaluation). Lowering the threshold didn't change our recall but precision fall to 72% (Fig. 2ab, exp 3). We discovered that the only way to make the recall higher is to remove vector with NEs from debate description from the system. This vector is used as a control of topic drift, so without that vector we got highest recall but with low precision of 50% (Fig. 2ab, exp4). Since we aim to have more quality than quantity in our final dataset we decided to use the vector with NEs from debate descriptions. With settings as from experiments nr.1 and nr.2 we got again lower precision but higher recall in some cases (Fig. 2ab).



(a) (b)

Fig. 2. Precision and relative recall

After this evaluation we can conclude that the query representation of the speech as a combination of named entities (from speech itself and debate descriptions) and topics (from speech itself and the whole conversation) in combination with used similarity measures works best for our goal of discovering media articles that covers the topics from the parliament. The structural elements contained in the transcripts (like possibility to distinguish segments of the debates that represents one conversation between many actors) played important part in formalizing a speech into a complex query. Also, great deal of help was the metadata available from the transcripts, which allowed us to preselect newspaper archive using predefined time span. Evaluation showed what we expected, that treating particular speech as a part of the bigger context (conversation) and creating a query that is a mixture of elements from both structures will retrieve higher number or relevant articles, because the newspaper articles from this domain are usually written in a form of report, where general subject discussed in the debate are mentioned intertwined with topics from particular speeches. Extracting topics from the speech was crucial for producing better recall and using just named entities from the speech produces very low precision. But extracting additional topics from collection of speeches contained under one description and additional extracting named entities from that description, gave us enough semantics to retrieve articles connected to the speech with reasonable precision and recall.

4 Conclusion and Next Steps

In this paper we have studied the creation of links between a dataset of political debates and a media archive. We have presented a linking method that takes advantage of metadata associated to the debates, NEs mentioned in the debate, topics detected in the debates, and the semantic *partOf* structure of the debates. We succeeded to create a pipeline for linking two very different types of text: debate transcripts containing spoken language full of digressions and different entities, and short and concise newspapers. We analyzed each parliamentary speech as a part of a bigger debate, i.e. taking into account the structure of the debate. The links we produced are of a different nature than those produced by e.g. ontology alignment tools.

In three experiments we have shown the added value of topics and debate structure. The results showed that using the NEs we can discover related media articles, but since the automatically generated queries are not specific enough their usage produces a dataset that contains a large amount of articles that are not related to the context from speech nor debate as a whole. We concluded that using topic modeling together with NEs we can create query representation of speeches that contains enough amount of context needed to retrieve and discover related articles with satisfactory precision.

These results provide leads for further research into automatic discovery of links between politics and media. At present, our method results are relatively coarsely typed links; we are able to discover that a speech and an article are linked, but we remain unclear about the nature and strength of the link. While it would be easy for us to *represent* a finer distinction of link types in RDF, the interpretation and usefulness of various types of links requires further study and will necessarily be an interdisciplinary effort. In future work, we aim to look into the direction of the links – whether politics influences media, or media influences politics – and the strength of the links, including how the strength varies as more time passes between the date of the political event and the publication of the media article.

While the presented method is designed to use the specific structure of the data and metadata at hand, the general idea of combining *who* (actors), *what* (named entities and topics) and *when* (dates) to find documents related to an event is applicable also outside the domain of Dutch parliamentary data. Future work includes generalization of the linking method to other (political) topics and other media collections such as televised news and radio bulletins (for which the first experiments look promising).

A virtual research environment will be built that allows the exploration of the debate topics and media coverage thereof via search and browsing. Next to the use of standard information retrieval libraries (Lucene), navigation options will be implemented that will allow users to browse through the linked datasets of debates and different types of media (newspapers, radio and video content).

References

1. van Erp, Marieke, et al.: Automatic Heritage Metadata Enrichment with Historic Events. In: Trant, J., Bearman, D. (eds.) Museums and the Web 2011: Proceedings. Archives & Museum Informatics, Toronto (2011)
2. van Hage, W., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the Simple Event Model (SEM). J. Web Semantics (2011)
3. van Hage, W.R., Malaisé, V., de Vries, G., Schreiber, G., van Someren, M.: Combining Ship Trajectories and Semantics with the Simple Event Model (SEM). In: Proceedings of the 1st ACM International Workshop on Events in Multimedia, pp. 73–80 (2009)
4. Mekhaldi, D., Lalanne, D.: Multimodal Document Alignment: Feature-based Validation to Strengthen Thematic Links. JMPT 1(1), 30–46 (2010)
5. Lv, Y., Moon, T., Kolari, P., Zheng, Z., Wang, X., Chang, Y.: Learning to model relatedness for news recommendation. In: WWW (2011)
6. Rao, D., McNamee, P., Dreze, M.: Entity Linking: Finding Extracted Entities in a Knowledge Base. Springer Lecture Notes in Computer Science: Multisource, Multilingual Information Extraction and Summarization (2011)
7. Gottipati, S., Jiang, J.: SMU-SIS at TAC 2010 - KBP Track Entity Linking. In: Proceedings of Text Analysis Conference (TAC 2010) Workshop (2010)
8. Gottipati, S., Jiang, J.: Linking entities to a knowledge base with query expansion. Empirical Methods in Natural Language Processing, EMNLP (2011)
9. Han, X., Sun, L.: A generative entity-mention model for linking entities with knowledge base. Association for Computational Linguistics (2011)
10. Bron, M., Huurnink, B., de Rijke, M.: Linking Archives Using Document Enrichment and Term Selection. In: Gradmann, S., Borri, F., Meghini, C., Schuldt, H. (eds.) TPDL 2011. LNCS, vol. 6966, pp. 360–371. Springer, Heidelberg (2011)
11. Kern, R., Granitzer, M.: German encyclopedia alignment based on information retrieval techniques. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 315–326. Springer, Heidelberg (2010)
12. Chang, J., Boyd-Graber, J.L., Gerrish, S., Wang, C., Blei, D.M.: Reading Tea Leaves: How Humans Interpret Topic Models. In: Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems (2009)
13. McCallum, Andrew Kachites: MALLET: A Machine Learning for Language Toolkit (2002), http://mallet.cs.umass.edu
14. Darling, W.M.: A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling (2011)
15. Montalvo, S., Martínez, R., Casillas, A., Fresno, V.: Bilingual news clustering using named entities and fuzzy similarity. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 107–114. Springer, Heidelberg (2007)