Discovering Links in Political Conversations

Talk of Europe, Creative Camp #3 on"Linking Government Data"

Mostafa Dehghani, Alex Olieman, Hosein Azarbonyad
{dehghani,olieman,h.azarbonyad}@uva.nl

Abstract

Entity recognition and disambiguation is a strong tool for enriching parliamentary debates and linking them to external sources. However, the general-purpose entity linkers are often not able to accurately detect entities which are particularly relevant for the parliamentary domain. Therefore, in the Talk of Europe event, Creative Camp#3, we are going to design and develop a specialized entity linker for political debates. Furthermore, we will address challenges of combining the designed specialized entity linker with generic entity linkers in our proposed method.

1 Introduction

Parliamentary proceedings capture our history from centuries ago up to today, reflecting any event of significance in the world: times of war and of peace, times of economic crisis and of prosperity, etc. Creating enriched annotated proceedings by linking contained concepts and named entities to external resources allows government, industry, journalists and even every individual citizen to investigate information in a straightforward way. According to the fact that political proceedings are considered as domain-specific texts ("political") with specific characteristics ("conversation"), and since entity linking systems typically perform very differently for different data sets and domains [7], there is something to be gained by specializing entity recognition and disambiguation for parliamentary proceedings, using the particular features, structures and already existing information in this data.

While general-purpose entity linking systems are usually able to link to a large number of entities from various domains, they are not necessarily able to link the most salient entities in a given domain, and may confuse salient entities with similarly named counterparts from popular culture. So, we propose an entity linker in which a specialized linker is combined with general entity linking systems to achieve a promising precision as well as an acceptable recall. Furthermore, we employ indirect linking, to build links to any particular external resources.

2 Linking Approach

In the proposed approach, using the specialized linker, we design a system that capitalizes on a small amount of background knowledge, and achieves entity recognition and disambiguation by means of pattern detection, string matching, and structured queries against the corpus. In this linker, we use special characteristics, features and the structure of the genre of political conversation to improve the precision. Due to the fact that in political texts most of the entities are addressed in a special way, (e.g. dictated by etiquette and/or jargon), the quality of the entity recognition can be improved considering these rules. Moreover, taking the *temporal aspect* and *being situated* of conversational text into consideration can facilitate disambiguation.

Combining the specialized linker with a general system like DBpedia Spotlight[4], we are able to work with an effective system in terms of precision and recall of discovered links. However, the problem would be "how to link detected entities to the particular external resources, like a specific news archives?". To address this problem, as the envisage solution, we propose *indirect linking* approach, in which using a mapping of documents that are linked using the linker to the documents in the particular resource, detected entities are linked to the resource. This mapping is learned based on the similarity of the language models of documents along with other features indicating conceptual connections [1].

3 Data that we plan to link

The goal of developing this linker is to semantically enrich the Dutch parliamentary proceedings available as open data [6, 3, 2]. This is the annotated proceedings of the Dutch parliamentary debates from 1814 until yesterday, available as open data in both XML and RDF [6].

As the external sources, we will connect the detected names of "persons" in the parliamentary proceedings who played a role in the parliament or government, to their bibliography in the biographical archives including: Biographical Portal of the Netherlands[1], BiographyNet [2], and Biographical Archive of PDC[3].

Furthermore, we are going to create links on European Parliamentary Debates [4] by focusing on the Dutch language and specific Dutch entities.

4 Envisioned Usage of the Linked Data

Linking parliamentary proceedings as a corpus of digitized heritage texts to the external resources and turning them into a connected network of information is a key step toward having tools for exploratory search and semantic inference.

[1] http://www.biografischportaal.nl/

[2] http://www.biographynet.nl/

[3] http://www.parlementairdocumentatiecentrum.nl/9360000/1/j9vvhd2jcta8fqs/vhdjk0fbh1ap

[4] http://linkedpolitics.ops.few.vu.nl/home

The obtained networked structure could be exploited to develop powerful tools that allow searcher to explore the rich content, by interactively constructing complex queries or search strategies, and interactively exploring the results of each stage [3]. This will have many desirable consequences:

- For human scholars, this makes it easy to understand and recognize events, entities and concepts in the data surfing through the links.
- It lays the grounds for building tools that focus not only "what" is said, but also by "who" and "to whom", and "why"?
- It can be used for geo-referencing and visualization [5].

5 Intended Outcome of Participation

Creative camp event is a highly related event to the ExPoSe, which we are working on it. Attending this event would be the best way of sharing the research and ideas we have had so far. After the event, we will make all the annotated data available through the Political Mashup project website [6] as well as the implementation as the open source tool at one of the public repositories.

6 Team members who would like to attend the Creative Camp

- Mostafa Dehghani, PhD Student at University of Amsterdam. Researcher on the Exploratory Political Search Project [3].

  Email: `dehghani@uva.nl`
- Alex Olieman, Research Programmer on the Exploratory Political Search Project [3].

  Email: `olieman@uva.nl`
- Hosein Azarbonyad, PhD Student at University of Amsterdam. Researcher on the Exploratory Political Search Project [3].

  Email: `h.azarbonyad@uva.nl`

References

[1] Mostafa Dehghani, Hosein Azarbonyad, Maarten Marx, and Jaap Kamps. Sources of evidence for automatic indexing of political texts. In *Proceedings of 37th European Conference on Information Retrieval*, ECIR'15, pages 568–573. 2015.

[2] Digging into linked parliamentary data. `http://dilipad.history.ac.uk/`, 2015. Digging into Data Challenge.

[3] Exploratory political search project. http://humanities.uva.nl/~kamps/expose/, 2015. Netherlands Organization for Scientific Research.

[4] Alex Olieman, Hosein Azarbonyad, Mostafa Dehghani, Jaap Kamps, and Maarten Marx. Entity linking by focusing dbpedia candidate entities. In *Proceedings of the First International Workshop on Entity Recognition & Disambiguation*, ERD '14, pages 13–24, 2014.

[5] Alex Olieman, Jaap Kamps, and Rosa Merino Claros. Loclinkvis: A geographic information retrieval-based system for large-scale exploratory search. In *Demo papers: SEMANTiCS*, 2014.

[6] Political mashup project. http://politicalmashup.nl/ and http://schema.politicalmashup.nl/, 2015. Netherlands Organization for Scientific Research.

[7] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460, 2014.