A Hybrid Approach to Domain-Specific Entity Linking

[Extended Abstract]

Alex Olieman Jaap Kamps Maarten Marx Arjan Nusselder
University of Amsterdam, Amsterdam, The Netherlands
{olieman|kamps|maartenmarx}@uva.nl arjan@nusselder.eu

1. INTRODUCTION

In the Entity Linking (EL) task, textual mentions are linked to corresponding Knowledge Base (KB) entries. The majority of state-of-the-art EL systems utilize one or more open-domain KBs, such as Wikipedia, DBpedia, Freebase, or YAGO, as basis for learning their entity recognition and disambiguation models [3]. The results of annotating a domain-specific corpus disappoint, however, when using a domain-agnostic EL system. We propose to use specialized linkers for salient entity types within the corpus' domain, which can work in concert with a generally trained model. Our approach is applied to conversational text, in particular parliamentary proceedings. The techniques that we have investigated are designed to be applicable to written records of any kind of conversation.

2. DOMAIN-SPECIFIC ENTITY LINKING

The specialist linkers are developed to target specific entity types that are mentioned frequently in the target corpus. These linkers capitalize on a small amount of background knowledge, and achieve entity recognition and disambiguation by means of pattern detection, string matching, and structured queries against the corpus. The simplest way that we have considered to annotate entities of a specific type is based on exact string matching. In our corpus we target Dutch political parties (*n=155*), because they are highly relevant as well as unambiguously named.

Our second linker applies to ambiguous entity types, and targets mentioned persons. It utilizes information about which people were present during a conversation, and about the period(s) during which a person was active, for disambiguation. Government and parliament members (*n=3,664*) are targeted in our corpus, and some knowledge of debating etiquette assists in detecting where they are mentioned. We also detect where government members are mentioned by their role, by means of a temporal index which maps roles to persons.

3. BENCHMARK

We have selected a sample of Dutch parliamentary proceedings from the period 1999–2012, which was subsequently annotated by DBpedia Spotlight (DBpS) [1], UvA Semanticizer (F+S) [2], and the specialist linkers. In order to assess the quality of these annotations against a consistent gold standard, we employed two human annotators for an independent and a consensus-building annotation round. To combine the output of multiple systems, we employ a preference ordering: the most specialized (i.e. estimated high-precision) system is asked to link a phrase first, and only if it doesn't the second system in the order is asked, and so on. By adding a generalist EL system at the end of the chain, the phrases that mention non-domain-specific entities also have their chance at being linked.

4. RESULTS

The results show that the specialist linkers were able to generate a larger number of accurate annotations for the corpus than either of the baseline systems, whilst limited to two specific entity types. F+S is the more precise of the baselines, but DBpS produces a greater number of potentially useful links. Our approach of combining a relatively simple custom-made EL system with an off-the-shelf EL system has also proven to be successful. This combination strategy produced a significantly better result than any of the systems could by themselves.

5. CONCLUSIONS

The current state-of-the-art entity linking systems aim to be open-domain solutions for corpora that are as heterogeneous as the Web. An unfortunate effect of this aim is that such generalist EL systems often disappoint when they are used on domain-specific corpora. We have outlined the prerequisites for, and development of, a lightweight linking system that targets salient entity types in a specific corpus. The specialist system, two baseline generalist systems, and hybrid combinations thereof have been evaluated against a gold standard, which is available as an open-data benchmark for the EL community at `http://datahub.io/dataset/el-bm-nl-9912`.

Our results show that the specialist system offers competitive performance to the two baseline systems, even though it is limited to two highly specific entity types. Moreover, by combining the specialist linkers with one or both generalist EL systems, recall can be significantly increased at a modest precision cost.

6. REFERENCES

[1] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proc. of I-Semantics 2013*, pages 3–6, Austria, Graz, 2013.

[2] D. Odijk, E. Meij, and M. de Rijke. Feeding the Second Screen: Semantic Linking based on Subtitles. In *OAIR 2013*, 2013.

[3] W. Shen, J. Wang, and J. Han. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 4347(2):443–460, 2014.