Learning to Combine Sources of Evidence for Indexing Political Texts

Extended Abstract

Mostafa Dehghani[1] Hosein Azarbonyad[2] Maarten Marx[2] Jaap Kamps[1]

[1]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[2]Informatics Institute, University of Amsterdam, The Netherlands
{dehghani,h.azarbonyad,kamps,maartenmarx}@uva.nl

ABSTRACT

Political texts are pervasive on the Web and access to this data is crucial for government transparency and accountability to the population. However, access to such texts is notoriously hard due to the ever increasing volume, complexity of the content and intricate relations between these documents. Indexing documents with a controlled vocabulary is a proven approach to facilitate access to these special data sources. However, increasing production of the political text makes human indexing very costly and error-prone. Thus, technology-assisted indexing is needed which scale and can automatically index any volume of texts.

There are different sources of evidence for the selection of appropriate indexing terms for political documents, including variant document and descriptor term representations, the structure of thesauri, if existing, and the set of annotated documents with the descriptor terms assigned as training data.

The main goal of this research is to investigate the effectiveness of different sources of evidence—such as the labeled training data, textual glosses of descriptor terms, and the thesaurus structure—for indexing political texts and combine these sources to have a better performance. We break down our main goal into three concrete research questions:

RQ1 *How effective is a learning to rank approach integrating a variety of sources of information as features?*

RQ2 *What is the relative importance of each of these sources of information for indexing political text?*

RQ3 *Can we select a small number of features that approximate the effectiveness of the large LTR system?*

We make use of learning to rank (LTR) as a means to not only take advantage of all sources of evidence effectively, but also analyse the importance of each source. To do so, we consider each document to be annotated as a query, and using all text associated with a descriptor term as documents. We evaluate the performance of the proposed LTR approach on the English version of JRC-Acquis [3] and compare our results with JEX [4] which is one of the state of the art systems developed for annotating political text.

1. EXPERIMENTS AND ANALYSIS

For our experiments, we have used English documents of last five years (from 2002 to 2006) of JRC-Acquis dataset [3]. The documents of this corpus have been manually labeled with EuroVoc concepts [2].

To address RQ1, using a LTR approach for integrating all features, we observe significantly better performance than previous systems. Table 1 shows the evaluation results of the proposed method compared to the baseline systems. Furthermore, we define features based on similarity of titles, texts, and descriptions between documents and descriptor terms as well as structural features, i.e. popularity, ambiguity, and generality, and then use LTR as a analytic tool to address our second research question. The analysis of feature weights reveals the relative importance of various sources of evidence, also gives insight in the underlying classification problem (Figure 1). Finally, based on the analysis of feature importance, we study RQ3. We suggest a lean-and-mean system using only four features (text, title, descriptor glosses, descriptor term popularity) which is able to perform at 97% of the large LTR model. The result of the lean-and-mean system is also presented in Table 1.

Table 1: Performance of JEX, best single feature, LTR, and lean-and-mean system. We report "incremental" improvement and significance (▲ indicates t-test, one-tailed, p-value < 0.05)

Method	P@5 (%Diff.)	Recall@5 (%Diff.)
JEX	0.4353	0.4863
BM25-TITLES	0.4798 (10%)▲	0.5064 (4%)▲
LTR-ALL	0.5206 (9%)▲	0.5467 (8%)▲
LTR-TTGP	0.5058 (-3%)	0.5301 (-3%)

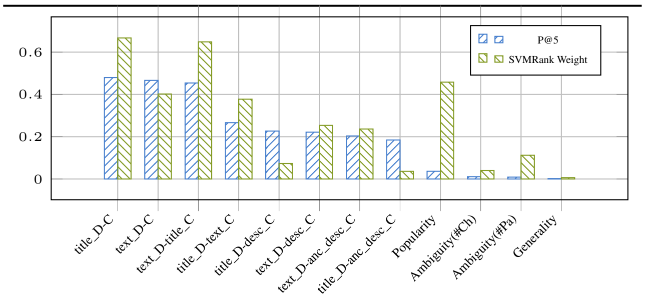

Figure 1: Feature importance: 1) P@5 of individual features, 2) weights in SVM-Rank model

Acknowledgments The full version of this paper is available as [1]. This research is funded by the Netherlands Organization for Scientific Research (WebART project, NWO CATCH # 640.005.001; ExPoSe project, NWO CI # 314.99.108; DiLiPaD project, NWO Digging into Data # 600.006.014).

References

[1] M. Dehghani, H. Azarbonyad, M. Marx, and J. Kamps. Sources of evidence for automatic indexing of political texts. In *Proceedings ECIR*, pages 568–573, 2015. URL http://dx.doi.org/10.1007/978-3-319-16354-3_63.

[2] EuroVoc. Multilingual thesaurus of the european union. http://eurovoc.europa.eu/.

[3] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufis. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *LREC*, pages 2142–2147, 2006.

[4] R. Steinberger, M. Ebrahim, and M. Turchi. JRC EuroVoc indexer JEX-A freely available multi-label categorisation tool. In *LREC*, pages 798–805, 2012.