Sources of Evidence for Automatic Indexing of Political Texts

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

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

Abstract. Political texts on the Web, documenting laws and policies and the process leading to them, are of key importance to government, industry, and every individual citizen. Yet access to such texts is difficult due to the ever increasing volume and complexity of the content, prompting the need for indexing or annotating them with a common controlled vocabulary or ontology. In this paper, we investigate the effectiveness of different sources of evidence—such as the labeled training data, textual glosses of descriptor terms, and the thesaurus structure—for automatically indexing political texts. Our main findings are the following. First, using a learning to rank (LTR) approach integrating all features, we observe significantly better performance than previous systems. Second, the analysis of feature weights reveals the relative importance of various sources of evidence, also giving insight in the underlying classification problem. Third, a lean-and-mean system using only four features (text, title, descriptor glosses, descriptor term popularity) is able to perform at 97% of the large LTR model.

Keywords: Automatical Indexing, Political Texts, Learning to Rank

1 Introduction

Political texts are pervasive on the Web, with a multitude of laws and policies in national and supranational jurisdictions, and the law making process as captured in debate notes of national and local governments. Access to this data is crucial for government transparency and accountability to the population, yet notoriously hard due to the intricate relations between these documents. Indexing documents with a controlled vocabulary is a proven approach to facilitate access to these special data sources [9]. There are serious challenges in the increased production of the political text, making human indexing very costly and error-prone.[3] 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. For example, descriptor terms can be expanded by their textual descriptions or glosses, which is useful for calculating the similarity of a descriptor term with the content of documents [7]. Also the structure of thesauri, if existing, could be another

[3] Iivonen [2] focuses on search (with the same information mediators that do subject cataloguing), and lists 32.1% pairwise agreement on the chosen terms, but 87.6% agreement when taking into account terms that are close in terms of the thesauri relations.

useful source for finding the semantic relations between descriptor terms and taking these relations into account [6, 7]. One of the main sources of evidence is to use a set of annotated documents, with the descriptor terms assigned. These documents are considered as train data in supervised methods [4, 5].

The main research problem of this paper is: How effective are different sources of evidence—such as the labeled training data, textual glosses of descriptor terms, and the thesaurus structure—for automatically indexing political texts? Our approach is based on learning to rank (LTR) as a means to take advantage of all sources of evidence, similar to [11], considering 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 [8] and compare our results with JEX [9] which is one of the state of the art systems developed for annotating political text documents. JEX treats the problem of indexing document as a profile-based category ranking task and uses textual features of documents as well as description of categories to index documents.

Our first research question is: How effective is a learning to rank approach integrating a variety of sources of information as features? We use LTR also as an analytic tool, leading to our second research question: What is the relative importance of each of these sources of information for indexing political text? Finally, based on the analysis of feature importance, we study our third research question: Can we select a small number of features that approximate the effectiveness of the large LTR system?

2 Sources of Evidence

In this section, we briefly introduce the sources of evidence used: 1) labeled documents, 2) textual glosses of descriptor terms,and 3) the thesaurus structure. We construct formal models of documents and descriptor terms, and use them to extract features.

Models are based on both title and body text of documents, which are available in all political document collections. The constructed model of documents is as follows:

$$Model_D = \langle M(title_D), M(text_D) \rangle, \tag{1}$$

where $Model_D$ is the model generated for the document D. This model is composed of different submodels: $M(title_D)$ based on only the title and $M(text_D)$ based on all text in the document (including titles). To construct these models, title and text of the document are considered as bag of words with stopword removal and stemming.

Similarly, the model of a descriptor terms is defined as:

$$Model_{DT} = \langle M(title_{DT}), M(text_{DT}), M(gloss_{DT}), M(anc\_gloss_{DT}) \rangle, \tag{2}$$

where $M(title_{DT})$ and $M(text_{DT})$ are the union of the title models and text models of all documents annotated by descriptor term DT. $M(gloss_{DT})$ is the descriptor model of DT and defined as the bag of words representation of glossary text of DT. $M(anc\_gloss_{DT})$ is considers all descriptor terms that are ancestors of the descriptor term DT in the thesaurus hierarchy, and takes the union of their descriptor models.

These models lead to eight possible combinations of a document and descriptor term submodel (2 times 4, respectively). For each combination, we employ three IR

measures: a) language modeling similarity based on KL-divergence using Dirichlet smoothing, b) the same run using Jelinek-Mercer smoothing, and c) Okapi-BM25.

In addition, we define a number of features for reflecting the characteristics of descriptor terms independent of documents. First, the statistics of the descriptor terms in train data is considered as the prior knowledge for determining what is the likelihood of selecting a descriptor term for annotating documents. That is, we define the number of times that a descriptor term has been selected for annotating documents in training data as its *popularity*. Second, in automatic indexing of documents, the degree of ambiguity of a descriptor term implicitly affects its chance for being assigned to the documents. We have modeled *ambiguity* with two different features, the number of parents of a descriptor term in thesaurus hierarchy graph and the number of its children. Another factor for determining the chance of a descriptor term for being an annotation of a given document is its *generality*. We quantify the generality of a descriptor term as its level in the thesaurus hierarchy. We consider the level of a descriptor term as the length of its shortest path to the root of thesaurus hierarchy.

Exploiting LTR enables us to learn an effective way to combine features and generate a final ranking list using all features. Finally the top-k (typically 5) descriptor terms in the ranking list are selected as the labels of a document.

3 Experiments

In this section, we detail the experimental settings (data, parameters and pre-processing), followed by the experimental results and analysis.

3.1 Experimental settings

We use JRC-Acquis dataset [8], a widely used collection for automatic indexing of political texts. The documents of this corpus have been manually labeled with EuroVoc concepts [1]. EuroVoc contains 6,796 hierarchically structured concepts, used to annotate political documents and news within the EU and in national governments. Since the structure of documents has changed over the years, we only use the documents of the last five years: from 2002 to 2006. We use the English version of JRC-Acquis, which contains 16,824 documents, each labeled with 5.4 concepts on average.

In order to evaluate the proposed methods, we divide the collection respecting its chronological order. The first part which contains the 70% oldest of documents is used to construct the models of descriptor terms (as documents in LTR). The remaining 30% of the collection is used to construct the test and train data (train and test query in LTR). To avoid missing information, in the second part we have removed descriptor terms that do not exist in the first part as annotation. This leads to 1,639 different descriptor terms in our dataset. We do 5-fold cross validation on the second part. To have a comparable evaluation, for 5-fold cross validation on JEX, we added the first 70% part of the collection to the training data used in each fold, to train its model. We have trained the ranking model using different LTR algorithms. Among them, AdaRank [10] has a slightly better performance and we report the results of this method.

We compare our results with JEX [9]. The pre-processing done in this paper is same as in JEX. We employ Porter stemmer and consider the 100 top frequent words in the collection as stopwords. We use different parameters for similarity functions according

Table 1. Performance of JEX, best single feature, and LTR methods. We report incremental improvement and significance ($^{\blacktriangle}$ 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%)$^{\blacktriangle}$	0.5064 (4%)$^{\blacktriangle}$
LTR-ALL	0.5206 (20%)$^{\blacktriangle}$	0.5467 (12%)$^{\blacktriangle}$

to the type of queries and documents. Based on pilot experiments, for short queries (considering titles of documents as queries) we use these parameters: $\mu = 1,000$ for LM-Dirichlet, $\lambda = 0.2$ for LM-JM, and $b = 0.65$ and $k_1 = 1.2$ for Okapi BM25. For long queries (the text of documents) we use these parameters: $\mu = 2,000$ for LM-Dirichlet, $\lambda = 0.6$ for LM-JM, and $b = 0.75$ and $k_1 = 1.2$ for Okapi BM25.

3.2 Experimental results

We now discuss our results, following the three research questions.

Effectiveness of LTR Starting with our first research question: How effective is a learning to rank approach integrating a variety of sources of information as features? Table 1 shows the evaluation results of the proposed method compared to the baseline system and JEX in terms of P@5, Recall@5. We use P@5 as the main measure to evaluate different methods, since the average number of descriptor terms per document in our dataset is about 5. Therefore, P@5 approximately could be considered as R-Precision as well. BM25-TITLES ranks the descriptor terms based on the similarities of them with title submodels of documents. This is the best performing single feature, and significantly better than JEX. The proposed LTR-ALL method significantly outperforms both BM25-TITLES and JEX. This demonstrates that the additional sources of evidence are effective for the indexing task.

Importance of Different Information Sources Next, we continue with our second research question: What is the relative importance of each of these sources of information for indexing political text? We use the trained model of SVM-Rank [3] as well as the $P@5$ of employing each individual feature. SVM-Rank tries to learn weights of features and combines them linearly based on their weights. For feature analysis, we assume the weight of each feature is a reflection of its importance. Figure 1 illustrates the importance of a selected set of exploited features. We pick only one of the similarity methods (BM25) from each feature type since the other two get very similar scores.

Similarity of titles of documents and descriptor terms is the most efficient feature. The performance is statistically better than the performance of the feature defined using text models of both descriptor terms and documents. Similarity of text of the given document and titles of the descriptor terms is also efficient. Therefore titles can be considered as a succinct predictor of classes. Titles of political documents tend to be directly descriptive of the content, making the title the most informative part of the document. In addition, to human annotators will pay considerable attention to the titles.

Among the query-independent features, generality and ambiguity do not help a lot while popularity stands out. Investigating the hierarchy graph of the concepts, we see that there is little variation in generality: the average number of levels in the hierarchy

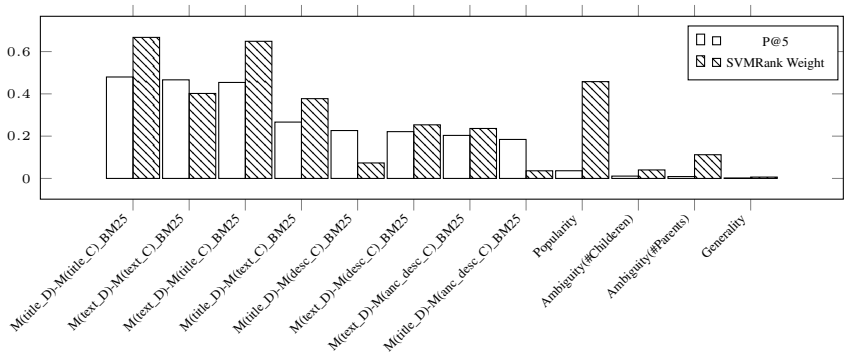

Fig. 1. Feature importance: 1) $P@5$ of individual features, 2) weights in SVM-Rank model

Table 2. Performance of LTR on all feature, and on four selected features

Method	P@5 (%Diff.)	Recall@5 (%Diff.)
LTR-ALL	0.5206 (-)	0.5467 (-)
LTR-TTGP	0.5058 (-3%)	0.5301 (-3%)

is 3.85 and its standard deviation is 1.29. There is considerable difference in ambiguity: the average number of children is 4.94 (standard deviation is 4.96) and the average number of parents is 1.08 (standard deviation is 0.25). Ambiguity may have low importance because it is not discriminative on this data. Although popularity of classes cannot achieve a high performance by itself, it gets a high weight in SVM-Rank model. It means that considering the fact that a descriptor term is frequently assigned in general, increases the quality along with other features. This feature is important due to skewness of assigned descriptor term frequency in JRC-Acquis [1].

Lean and Mean Approach Based on the feature analysis, we now continue with our third research question: Can we select a small number of features that approximate the effectiveness of the large LTR system? The designed LTR-ALL uses a large set of features that is very complex, hence we try to carve out a lean-and-mean system which has a better efficiency/effectiveness trade-off.

Our lean-and-mean system is an LTR trained system on four selected features: the BM25 similarities of text submodel of documents with all text, titles only, and textual glosses of descriptor terms, and popularity of descriptor terms. Table 2 indicates the performance of this LTR-TTGP approach using only four features. The LTR-TTGP approach is significantly better than JEX and BM25-TITLES before. Although the performance of LTR-ALL is significantly better than the LTR-TTGP method, the performance of LTR-TTGP is 97% of the large LTR-ALL system. Therefore, making the selective LTR approach a computationally attractive alternative to the full LTR-ALL approach.

4 Conclusion and Future Work

Our broad motivation is to build connections between political data from different national and international jurisdictions (such as EU versus national laws and parliamentary debates, or between different national parliaments). Such connections are essential

for researchers, both at the level of whole documents and individual document parts. This paper addresses an important initial step, trying to replicate the human indexing of EU laws and policies based on the EuroVoc vocabulary functioning as pivot language.

Our main findings are the following. First, using a learning to rank (LTR) approach integrating all features, we observe significantly better performance than previous systems. Second, the analysis of feature weights reveals the relative importance of various sources of evidence, also giving insight in the underlying classification problem. Third, a lean-and-mean system using only four features (text, title, descriptor glosses, descriptor term popularity) is able to perform at 97% of the large LTR model.

Are the proposed systems "good enough" for the motivating task at hand. Clearly we are far from exactly replicating the choices of the human indexer. However, considering the inter-indexer agreement and the (soft) upperbound of the full LTR approach, the room for improvement seems limited. However, as Iivonen [2] observes, indexers that disagree pick terms that are near to each other in the concept hierarchy. Anecdotal inspection of our automatic indexing reveals the same: wrong descriptors tend to be conceptually close to the gold standard indexing term. Hence, this give support to the utility of the current systems for discovering conceptual cross-connections in political texts, as well as suggests ways to improve the current approaches by clustering and propagating descriptors to similar terms.

Acknowledgements This research was supported by the Netherlands Organization for Scientific Research (ExPoSe project, NWO CI # 314.99.108; DiLiPaD project, NWO Digging into Data # 600.006.014) and by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement ENVRI, number 283465.

5 References

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

[2] M. Iivonen. Consistency in the selection of search concepts and search terms. *IPM*, 31: 173–190, 1995.

[3] T. Joachims. Training linear svms in linear time. In *SIGKDD*, pages 217–226, 2006.

[4] J. Nam, J. Kim, I. Gurevych, and J. Furnkranz. Large-scale multi-label text classification - revisiting neural networks. In *ECML PKDD*, pages 437–452, 2014.

[5] B. Pouliquen, R. Steinberger, and C. Ignat. Automatic annotation of multilingual text collections with a conceptual thesaurus. In *EUROLAN*, pages 9–28, 2003.

[6] Z. Ren, M.-H. Peetz, S. Liang, van Willemijn Dolen, and M. de Rijke. Hierarchical multi-label classification of social text streams. In *SIGIR*, pages 213–222, 2014.

[7] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *J. Mach. Learn. Res.*, 7:1601–1626, 2006.

[8] 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.

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

[10] J. Xu and H. Li. Adarank: A boosting algorithm for information retrieval. In *SIGIR*, pages 391–398, 2007.

[11] Y. Yang and S. Gopal. Multilabel classification with meta-level features in a learning-to-rank framework. *Machine Learning*, 88(1-2):47–68, 2012.