Detecting Topic Domains from Topic Distributions

Anonymous ACL submission

1 Introduction

In this paper, we describe preliminary results from an ongoing experiment wherein we classify large unstructured text corpora by *topic domain*. *Topic domain*—along with other high-level classifications such as genre or register—is among the types of meta data most essential to many corpus linguists. Therefore, the lack of reliable meta data in general is often mentioned as a major drawback of large, crawled web corpora. It must be noted, however, that such high-level annotations are not usually available for large traditional corpora (such as newspaper corpora), either. Given the size of many modern corpora (traditional or web corpora), automatic approaches to meta data generation are a general desideratum. When it comes to the automatic identification of register, even very recent approaches (Biber and Egbert, 2016) cannot deliver satisfying accuracy, and it is unclear if categories such as register and genre can be operationalized such that a reliable annotation is even possible for humans. By contrast, automatic text categorization according to content has yielded much more promising results (Sebastiani, 2002). Data-driven induction of topics has proven quite successful and is in many respects a more objective way of organizing a collection of documents by content (EAGLES, 1996). Still, the category labels that can be inferred from such topics are not necessarily useful for linguistic corpus users. In this paper, we explore the possibility of inferring a small, more traditional set of *topic domains* (or *subject areas*) from the topics induced in an unsupervised manner by Latent Semantic Indexing (Landauer and Dumais, 1994; Landauer and Dumais, 1997) and Latent Dirichlet Allocation (Blei et al., 2003). Since we classify and compare two large German corpora with respect to their distribution of topic domains, our paper also contributes to the area of corpus comparison, another important issue in corpus linguistics (Kilgarriff, 2001; Biemann et al., 2013).

2 Gold standard Data

The gold standard corpora were prepared by manually annotating documents from two large German corpora. The first data set consists of 870 randomly selected documents from DECOW14A, a crawled web corpus (Schäfer and Bildhauer, 2012; Schäfer, 2015). The second data set contains 886 documents randomly selected from *DeReKo*, a corpus composed predominantly of newspaper texts (Kupietz et al., 2010). Our choice of corpora was motivated by fact that we expected some overlap w. r. t. to topics covered in them, but also some major differences. The documents in these gold standard corpora were classified according to a custom annotation scheme for topic domain which builds on previous work (Sharoff, 2006). The design goal was to a have moderate number (about 10–20) of topic domains that can be thought of as subsuming more fine-grained topic distinctions. We developed the annotation scheme in a cyclic fashion, taking into account annotator feedback after repeated annotation processes. Currently, we use a version that distinguishes 13 topic domains, namely *Science*, *Technology*, *Medical*, *Public Life and Infrastructure*, *Politics and Society*, *History*, *Business*, *Law*, *Fine Arts*, *Philosophy*, *Beliefs*, *Life and Leisure*, *Individuals*.

3 Experiment Setup

Our general approach was to infer a topic distribution over a corpus using topic modelling algorithms as a first step. In the second step, we used the resulting document–topic matrix to learn topic domain distinctions for the documents from their assignment to the topics. To achieve this, super-

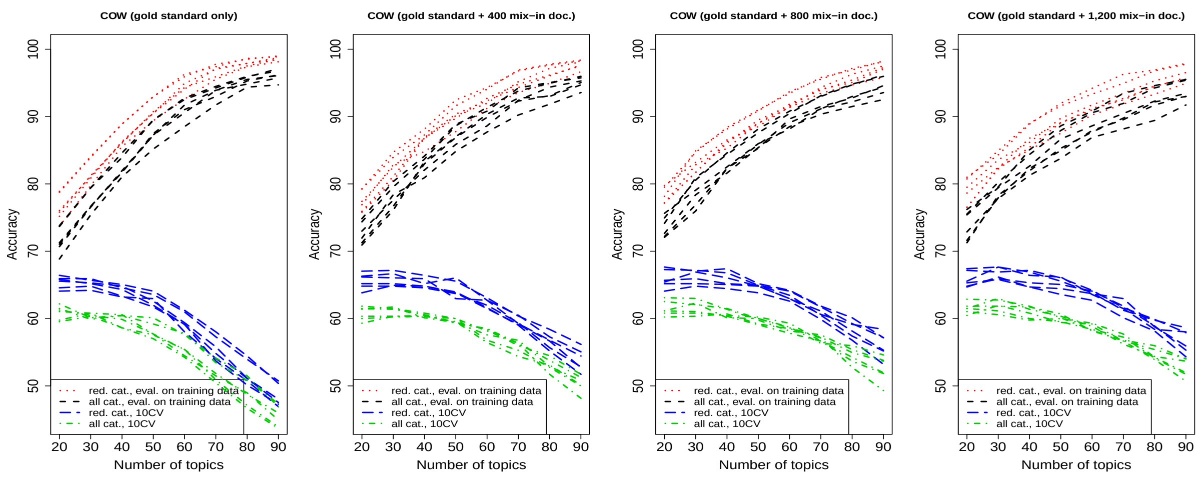

Figure 1: Accuracy with different numbers of topics for COW-only dataset

vised classifiers were used. Through permutation of virtually all available classifiers (with the appropriate capabilities) in the Weka toolkit (Hall and Witten, 2011), LM Trees (Landwehr et al., 2005) and SVMs with a Pearson VII kernel (Üstün et al., 2006) were found to be most accurate. Due to minimally higher accuracy, SVMs were used in all subsequent experiments. Some topic domains occurred only rarely in the gold standard, and we did not expect the classifier to be able to generalize well from just a few instances. Therefore, we evaluated the results on the *full* data set and a *reduced* data set with rare categories removed.

For the underlying topic inference, we used LSI and LDA as implemented in the Gensim toolkit (Řehůřek and Sojka, 2010). In previous experiments, the LDA topic distribution was unstable, and results were generally unusable, which was probably due to the comparatively small corpora used. We consequently only report LSI results here and will return to LDA in further experiments. However, for any topic modelling algorithm, our corpora can be considered small. Therefore, we inferred topics not just based on the annotated gold standard data sets, but also on larger datasets which consisted of the gold standard mixed with additional documents from the source corpora. For the training of the SVM classifiers, the documents that had been mixed in were removed again. We systematically increased the number of mixed-in document in increments of roughly half as many documents as contained in the gold standard corpora.

We pre-processed both corpora in exactly the same way (tokenization, lemmatization, POS-tagging, named entity recognition). Using the lemma and the simplified POS tags (such as *kindergarten_nn*) as terms in combination with some filters (use only lower-cased purely alphabetic common and proper noun lemmas between 4 and 30 characters long) mostly gave the best results.

4 Results

Figure 1 shows the classification accuracy using 20 to 90 LSI topics. Each line corresponds to one sub-experiment (with slightly different preprocessing options), and the lines form well distinguishable bands. The highest accuracy is achieved with the reduced set of topic domains (minor categories removed) when the evaluation is performed on the training data. The full set of topic domains leads to a drop in accuracy of about 5%. The two lower bands show the classification accuracy in a 10-fold cross-validation (10CV), again with the reduced set of topic domains performing roughly 5% better. While a higher number of topics improves results on the training data, the accuracy in the cross-validation drops. Too large numbers of topics obviously allow the method to pick up idiosyncratic features of single documents or very small clusters of documents, leading to extreme overtraining.

The four panels show results based on different topic models. Panel (a) uses a topic model inferred only from the 870 gold standard documents. Results in panel (b) through (d) are based on topic models inferred on larger data sets as described in Section 3. In the experiment reported in panel (d),

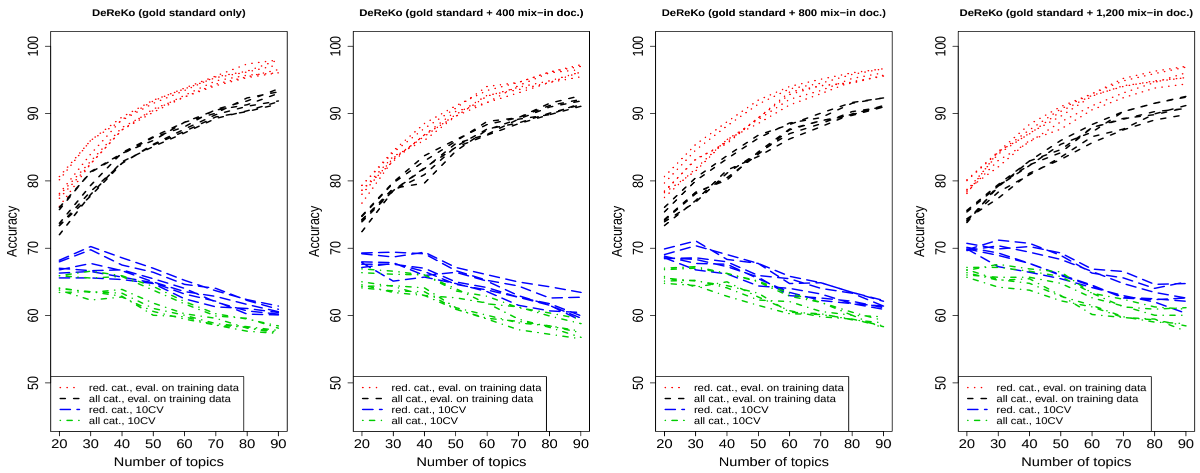

Figure 2: Accuracy with different numbers of topics for DeReKo-only dataset

Corpus	Mixed-in	Attribute	Topics	Accuracy	Precision[1]	Recall[1]	F-Measure[1]
COW	3,200	token	20	68.765%	0.688	0.688	0.674
DeReKo	3,600	lemma + POS	40	72.999%	0.725	0.730	0.696
COW + DeReKo	0	lemma + POS	30	51.872%	0.431	0.519	0.417

Table 1: Evaluation at best achievable accuracy with the reduced set of topic domains in 10-fold cross-validation ([1]weighted average across all categories)

for example, 1,200 documents were added to the 870 gold standard documents. While the results of the 10CV are slightly improved by mixing in more documents, the maximum achieved accuracy does not change significantly. We mixed in up to 8,000 additional documents (not all results shown here) with no significant change compared to panel (d) in Figure 1. We consider the maximum 10CV accuracy with the reduced set of topic domains most informative w. r. t. the potential quality of the classifier, and we report it in Table 1.

A very similar plot for the DeReKo data is shown in Figure 3. The best results are also given in Table 1. The added accuracy (4.23% according to Table 1) is a side effect of the more skewed distribution of topic domains in the DeReKo gold standard data. The κ statistic for the COW and DeReKo results from Table 1 of $\kappa_{\mathrm{COW}} = 0.575$ and $\kappa_{\mathrm{DeReKo}} = 0.582$ indicate that achieving a higher accuracy for the COW data is actually harder than for the DeReKo data (see also analysis of confusion matrices below).

When the COW and DeReKo data are pooled, however, quality drops below any acceptable level, cf. Figure 3 and Table 1. Mixing in more documents (panels b–d) improves the evaluation results on the training data, but the 10CV results remains steady at around 50%. This is remarkable because larger training data sets should lead to increased, not degraded accuracy. While a deeper analysis of the LSI topic distributions remains to be undertaken, it is evident what causes these below average results on the side of the SVM classifier when looking at the confusion matrices, cf. Table 2. In the COW gold standard (panel a), the dominant modal category is *Life and Leisure*. The distribution of topic domains is not too skewed, and the confusion is distributed roughly uniformly across categories. The DeReKo gold standard (panel b) consists mainly of two clusters of documents in the domains *Politics and Society* and *Life and Leisure*. In the joint data set (panel c), this leads to a situation in which the classifier tips over and assigns most documents to *Life and Leisure* and the rest mostly to *Politics and Society*. This indicates that for such skewed distributions of topic domains, larger gold standard data sets are required. It is not indicative of a general failure of the method or a general incompatibility of newspaper and web data in the context of our method. The confusion matrices in Table 2 clearly indicate, however, that topic domains are represented quite differently in newspaper and web corpora.

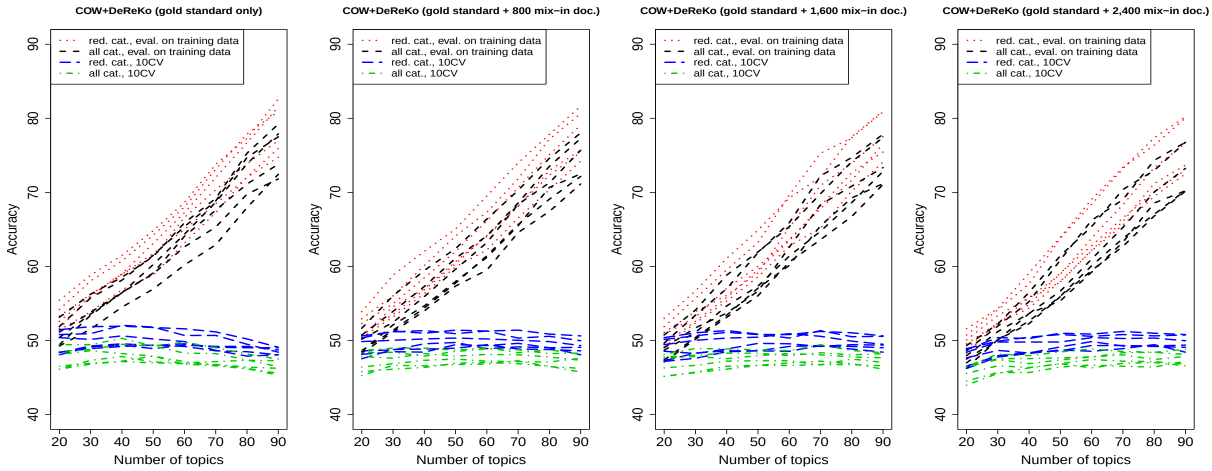

Figure 3: Accuracy with different numbers of topics for COW+DeReKo dataset

COW		Classified							
		PolSoc	Busi	Life	Arts	Public	Law	Beliefs	Hist
Annotated	PolSoc	**26**	12	10	1	1	0	1	0
	Busi	5	**105**	40	7	1	2	1	1
	Life	3	14	**286**	6	4	1	1	1
	Arts	3	2	36	**78**	1	0	2	6
	Public	0	3	11	0	**9**	1	0	0
	Law	3	9	8	0	1	**8**	0	0
	Beliefs	4	3	11	6	1	0	**30**	1
	Hist	9	0	9	7	1	1	2	**15**

DeReKo		Classified					
		PolSoc	Busi	Life	Indiv	Arts	Public
Annotated	PolSoc	223	6	39	0	0	8
	Busi	20	24	9	0	0	0
	Life	24	1	324	0	0	1
	Indiv	5	0	17	0	0	1
	Arts	2	0	28	0	6	0
	Public	35	0	30	0	0	34

Joint		Classified								
		PolSoc	Busi	Medical	Life	Arts	Public	Law	Beliefs	Hist
Annotated	PolSoc	**199**	7	0	109	0	12	0	0	0
	Busi	18	**23**	0	172	0	2	0	0	0
	Medical	6	0	**0**	29	0	1	0	0	0
	Life	25	4	0	**632**	0	5	0	0	0
	Arts	2	2	0	160	**0**	0	0	0	0
	Public	46	2	0	56	0	**19**	0	0	0
	Law	8	0	0	31	0	0	**0**	0	0
	Beliefs	0	0	0	59	0	0	0	**0**	0
	Hist	4	0	0	50	0	0	0	0	**0**

Table 2: Confusion matrices for the best achievable results on the COW (a), DeReKo (b), and joint (c) data sets as reported in Table 1; different sets of categories are the result of excluding low-frequency topic domains (below 20 for COW and DeReKo, below 30 for joint data)

5 Conclusions and Outlook

The results presented here are preliminary but highly encouraging, and they clearly indicate the route to be taken in further experiments. First of all, there appears to be a connection between induced topic distributions and more general topic domains. The decreased performance in cross-validation experiments indicates that larger gold standard data sets are required. Such data sets are currently being annotated under our supervision. Secondly, there appears to be a significant difference in the topic distribution and the topic/domain mapping in newspaper and web corpora. This might be one of the reasons behind the collapse of the classifier when newspaper and web data are pooled. In future experiments, it remains to be discovered whether larger gold standard corpora can alleviate such problems. This will eventually enable us to decide whether separate models or joint models for the two kinds of corpora are more appropriate. Thirdly, the highly skewed topic distributions in both newspaper and web corpora indicate that splitting up some topic domains might lead to a better fit. In fact, annotators have independently asked whether splitting up *Politics and Society* and *Life and Leisure*—the critical categories which make the classifier collapse (cf. Section 4)—could not be split up into at least two categories.

References

Douglas Biber and Jesse Egbert. 2016. Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2:3–36.

Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, and Torsten Zesch. 2013. Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics*, 28(2):23–60.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

EAGLES. 1996. Preliminary recommendations on text typology. Technical report EAG-TCWG-TTYP/P, EAGLES.

Mark Hall and Ian H. Witten. 2011. *Data mining: practical machine learning tools and techniques*. Kaufmann, Burlington, 3rd edition.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German reference corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, pages 1848–1854, Valletta, Malta. European Language Resources Association (ELRA).

Thomas K. Landauer and Susan T. Dumais. 1994. Latent semantic analysis and the measurement of knowledge. In R. M. Kaplan and J. C. Burstein, editors, *Princeton, NJ*. Educational Testing Service, Princeton, NJ.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic model trees. *Machine Learning*, 95(1–2):161–205.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul. ELRA.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, and Andreas Witt, editors, *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster. UCREL.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *Wacky! Working papers on the Web as Corpus*. GEDIT, Bologna.

Bülent Üstün, Willem J. Melssen, and Lutgarde M.C. Buydens. 2006. Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems*, 81:29–40.