Inferring Topic Domains from Topics in Newspaper and Web Data

Anonymous ACL submission

Abstract

1 Introduction

In this paper, we describe preliminary encouraging 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/register, is among the meta data categories most sought for by (corpus-)linguists. From a corpus engineering perspective, it is therefore desirable to add as much linguistically relevant meta data as possible. While lack of reliable meta data in general is often mentioned as a major drawback of large, crawled web corpora, high-level annotations are not usually available for most traditional corpora either. Given the size of many modern corpora, these classifying tasks can only be accomplished in an automated fashion. However, poor results have recently been reported for automatic classification of genre/register (**REF Globbe**), and it is unclear if such categories can be operationalized to a point where a reliable annotation is even possible for humans. By contrast, automatic text categorization according to content (i. e., topic) has yielded much more promising results (Sebastiani, 2002), **perhaps something more recent?**). In this connection, while 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), the category labels that can be inferred from such topics do not necessarily meet the needs of linguists. In this paper, we explore to what extend a number of traditional, pre-established topic domains can be learned from the topics induced by Latent Semantic Indexing (Landauer and Dumais, 1994; Landauer and Dumais, 1997) and Latent Dirichlet Allocation (Blei et al., 2003). Since we are classifying and comparing large German corpora with respect to their topic domain distributions, the paper also touches on the question of corpus comparison, which is another important issue in corpus linguistics, both for corpus builders and users (Kilgarriff, 2001; Biemann et al., 2013; Schäfer and Bildhauer, 2013).

2 Gold standard Data

Gold standard data was prepared by manually annotating documents from two different corpora of German:

- 870 randomly selected documents from DECOW12, a crawled web corpus (Schäfer and Bildhauer, 2012; Schäfer, 2015)
- 886 documents randomly selected from *DeReKo*, consisting predominantly of newspaper texts (Kupietz et al., 2010).

The choice of corpora was motivated by fact that we expected some overlap w. r. t. to topics covered in them, but also some important differences.

All documents were classified according to a custom annotation scheme, *COWCat 2013* **(REF?)**, which builds on work by (Sharoff, 2006) and provides guidelines for classifying text with respect to a number of dimensions. The design goal for the topic dimension was to a have moderate number (about 10–20) of topic *domains* that can be thought of as subsuming more fine-grained topic distinctions. The scheme was developed in a cyclic fashion, taking into account annotator feedback after repeated annotation processes. The version used in this experiment distinguishes 13 topic domains (while the current version has undergone some revision, disambiguating and elaborating on some of these domains): *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 (Section 2) using topic modelling algorithms as a first step. In the second step, we used the resulting document–topic matrix to infer topic domains for the documents from their assignment to the topics. To achieve this, supervised classifiers were used. Through permutation of 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. Because 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 both *Latent Semantic Indexing* (LSI) (Landauer and Dumais, 1994; Landauer et al., 2007) and *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003) as implemented in the Gensim toolkit (Řehůřek and Sojka, 2010). The LDA topic distribution in first experiments was highly unstable, and results were generally unusable. This 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 have to 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. We systematically increased the number of mixed-in document in increments of roughly half as many documents as contained in the gold standard corpora.

Linguistic pre-processing was simplified because both corpora have already tokenized, lemmatized, POS-tagged, etc. by their creators. 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. Vocabularies were filtered to contain only terms with a term-document frequency above 2. Terms which occurred in more than 50% of the documents were also removed. Preliminary experiments showed that the exact cutoffs were not crucial, however.

4 Results

Figure 1 shows the classification accuracy using 20 to 90 LSI topics. Each line corresponds to one sub-experiment (slightly different pre-processing 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 allows the method to pick up idiosyncratic features of single documents or very small clusters, leading to extreme overtraining.

The four panels show results based of 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), for example, 1,200 documents were added to the 870 gold standard documents. While the overtraining effect is alleviated by mixing in more documents, the maximum achieved accuracy does not significantly improve. We continued the experiment (further results not shown here), mixing in up to 8,000 additional documents 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 parallel plot for the DeReKo data is shown in Figure 3, and maximally best results are also given in Table 1. The picture is essentially the same as for the COW data set. The added accu-

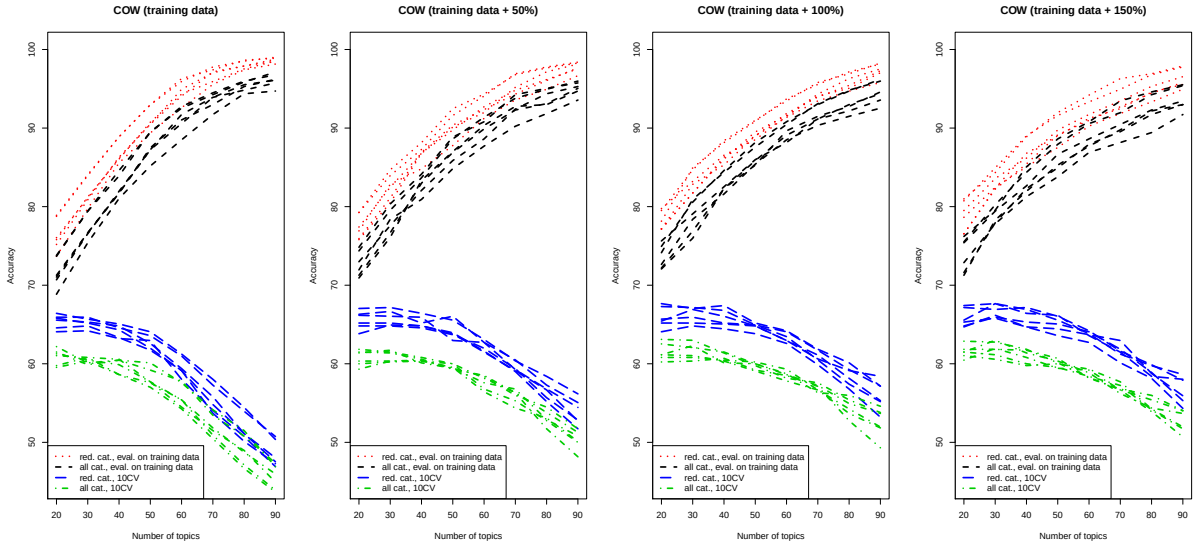

Figure 1: Accuracy with different numbers of topics for COW-only dataset

Corpus	Mixed-in	Attribute	Topics	Accuracy	Precision	Recall	F-Measure
COW	∼3,200	token	20	68.765%	0.688	0.688	0.674
DeReKo	∼3,600	lemma + POS	40	72.162%	0.716	0.722	0.686
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; Precision, Recall, and F-Measure are weighted averages across all categories

racy (3.397% 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 $\kappa_{\text{COW}} = 0.575$ and $\kappa_{\text{DeReKo}} = 0.569$ show that achieving a higher accuracy for the COW data is actually harder in comparison with the DeReKo data.

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 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 becomes clear what causes these below average results on the side of the SVM classifier when looking at the confusion matrices, cf. Table 2. The dominant modal category is *Life and Leisure* in the annotated COW gold standard (panel a). However, the distribution of topic domains is not too skewed, and the confusion is distributed roughly uniformly across categories. The DeReKo data set (panel b) consists mainly of two clusters of documents in the domains *Politics and Society* (276 of 837) and *Life and Leisure* (350 of 837). 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 definitely show, however, that topic domains are represented quite differently in newspaper and web corpora.

5 Conclusions and Outlook

The results presented here are preliminary, but highly encouraging (over 90% accuracy on training data and over 70% accuracy in cross-validation on some data sets), and they 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 experi-

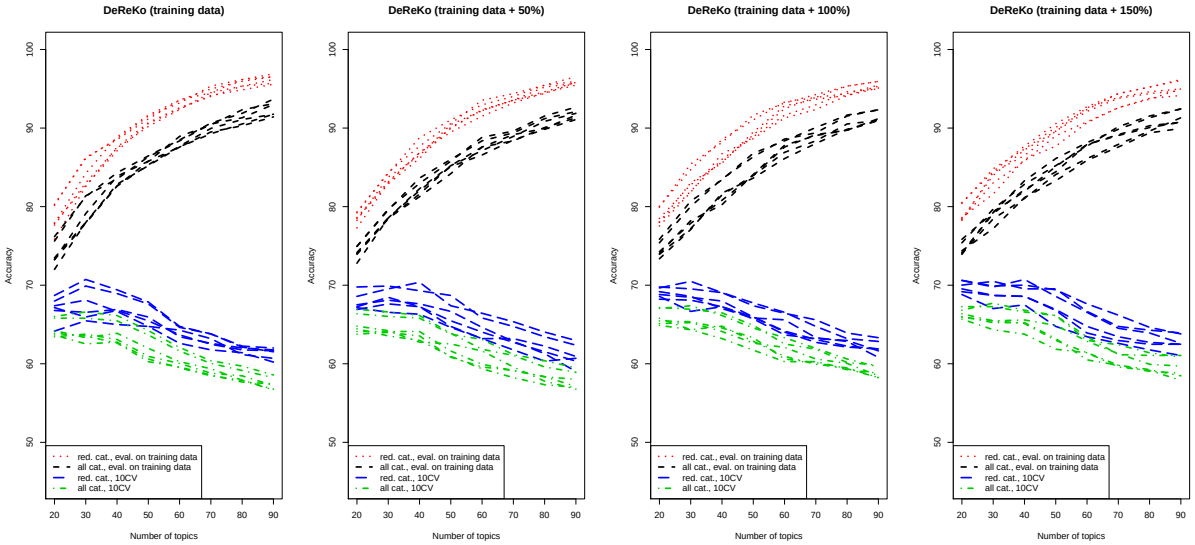

Figure 2: Accuracy with different numbers of topics for DeReKo-only dataset

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	Arts	Public	Law	Beliefs	Hist
Annotated	PolSoc	**222**	5	41	0	8	0	0	0
	Busi	19	**25**	8	0	1	0	0	0
	Life	27	1	**321**	0	1	0	0	0
	Arts	2	0	29	**5**	0	0	0	0
	Public	41	0	27	0	**31**	0	0	0
	Law	10	0	0	0	0	**0**	0	0
	Beliefs	0	0	3	0	0	0	**0**	0
	Hist	2	0	7	0	1	0	0	**0**

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

ments 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 data sets 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

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,

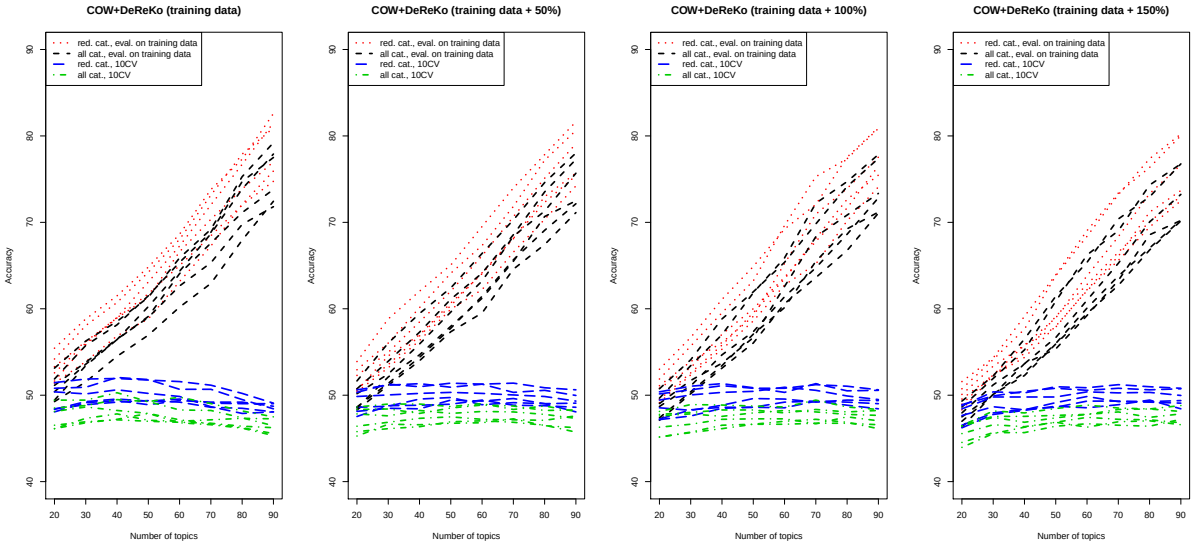

Figure 3: Accuracy with different numbers of topics for COW+DeReKo dataset

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.

Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, editors. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah.

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 and Felix Bildhauer. 2013. *Web Corpus Construction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool, San Francisco etc. Im Druck.

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.