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 (such as web corpora) by *topic domain*. While topics modelling . . . linguists are often interested in high-level classifications such as genre, register, or topic domain.

Why topic domain?

Automatic meta data: desirable not JUST for web data.

Short comments on register scene and poor results in recent Globbe paper.

Mention corpus comparison as important field: Kilgarriff, WCC, "Biemann et al."

2 Gold standard Data

Corpora (Kupietz et al., 2010), SchBi 2012, Sch 2015

Annotation scheme: Sharoff; mention that it was developed in repeated annotation processes based on annotator feedback; mention that design goal was roughly 10 to 20 topic domains

3 Experiment Setup

Pre-processing

Algorithms (LSI/LDA)

Gold and plus versions

Filters and lexicon thresholds

Numbers of topics

SVM vs. Trees; SVM kernel selection

Elimination of very small topic domains

4 Results

Figure 1 shows the classification accuracy for 20 to 90 LSI topics. Each line corresponds to one sub-experiment, and the lines form well distinguishable bands. The highest accuracy is achieved with a reduced set of topic domains when the evaluation is performed on the training data (upper six dotted lines). The full set of topic domains leads to a drop in accuracy of about 5% (six upper dashed lines). The two lower bands show the classification accuracy in a 10-fold cross-validation (10CV), again with the reduced set of topic domains faring 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 (larger in increments of roughly 50% of the original number of documents) in order to stabilize the topic distribution. 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, maximally best results are also given in Table 1. The picture is essentially the same as for the COW data set. The added accuracy (3.397% according to Table 1) is a side effect of

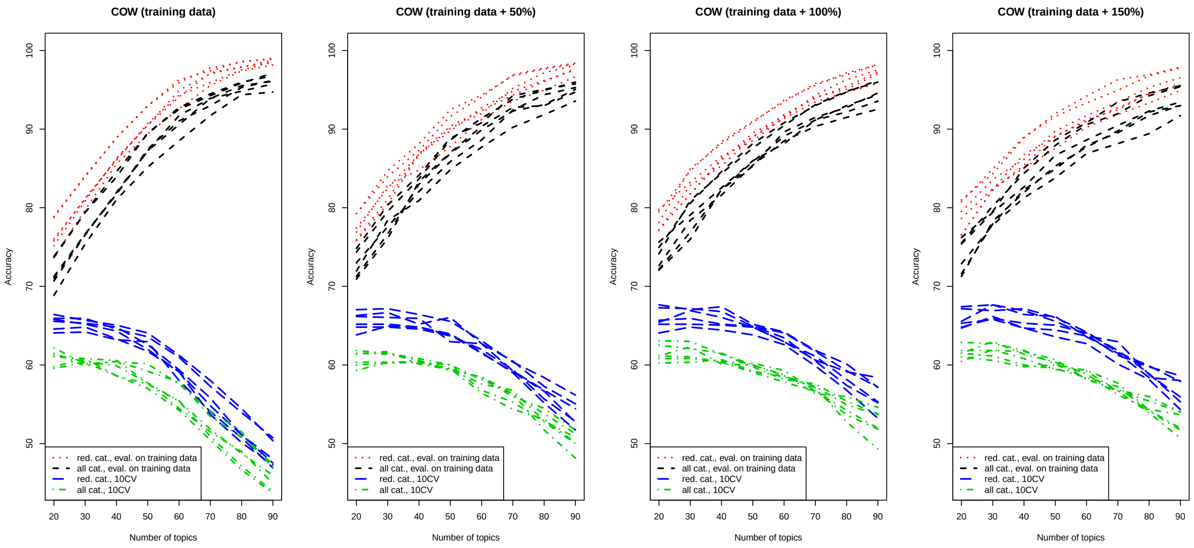

Figure 1: Accuracy with different numbers of topics for COW-only dataset

Corpus	Mixed-in	Filters	Attribute	Topics	Accuracy	Precision	Recall	F-Measure
COW	~3,200	set 2	token	20	68.765%	0.688	0.688	0.674
DeReKo	~3,600	set 2	lemma + POS	40	72.162%	0.716	0.722	0.686
COW + DeReKo	0	set 2	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

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—which give us a more realistic idea of the real-world performance of the classifier—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. While the COW data set show a less than perfect but

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 solid connection between induced topic distributions and externally defined topic domains. The relative poor performance in cross-validation experiments indicates that larger gold standard data sets are required. Such data sets are currently being annotated. Secondly, there appears to be a significant difference in the topic distribution and the topic/domain mapping in newspaper and web corpora. This is indicated by the drop in classification accuracy when newspaper and web data are pooled. As such, In future experiments, it remains to be discovered whether larger corpora can alleviate this divergence, finally enabling us to decide whether separate models are required or joint models can be trained. Thirdly, the highly skewed topic distributions in both newspaper and web data sets as well as comments from annotators indicate that splitting up some topic domains might lead to a better fit.

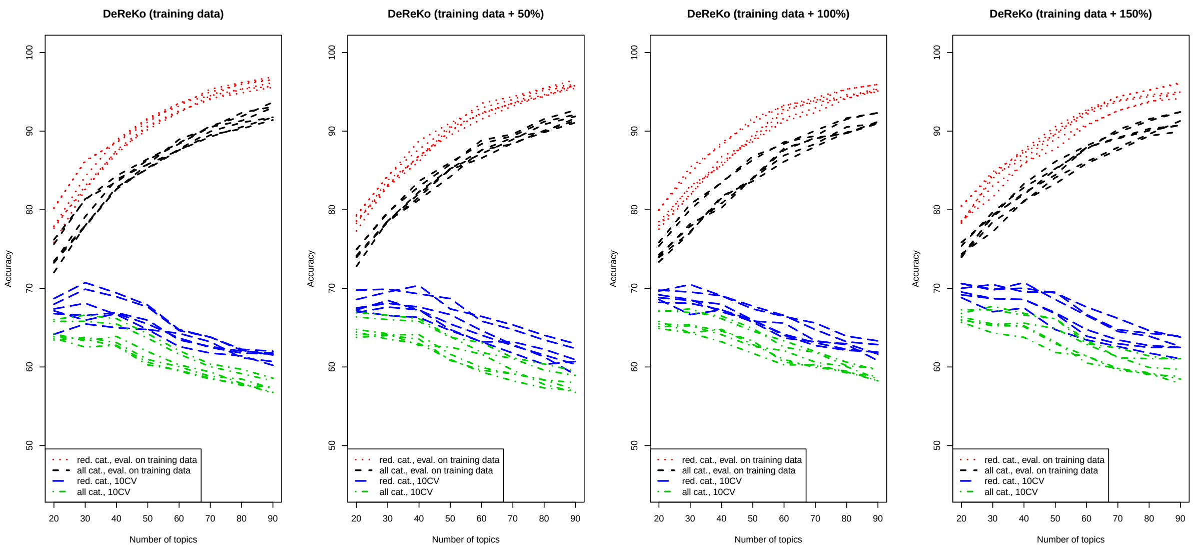

Figure 2: Accuracy with different numbers of topics for DeReKo-only dataset

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 results on the COW (a), DeReKo (b), and joint (c) data sets as reported in Table 1

References

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).

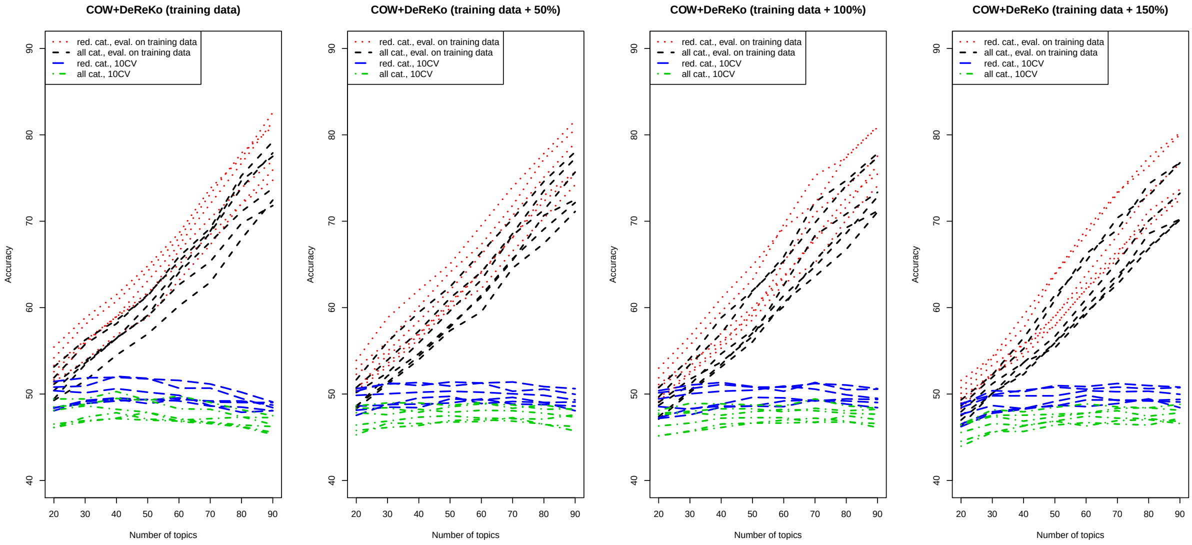

Figure 3: Accuracy with different numbers of topics for COW+DeReKo dataset