# Inferring Topic Domains from Topics in Newspaper and Web Data

Anonymous ACL submission

## Abstract

## 1 Introduction

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

Accuracy and $\kappa$ for Cow, Dereko, Coreko and plus variants

4 plots

## 5 Conclusions and Outlook

The results presented here are preliminary, but highly encouraging (over 90% accuracy on training data and 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.

## 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).