Automatic Classification by Topic Domain for Meta Data Generation, Web Corpus Evaluation, and Corpus Comparison

Roland Schäfer[1] and Felix Bildhauer[2]

[1]Ling. Web Characterization (DFG), FU Berlin
[2]Institut für Deutsche Sprache, Mannheim

10th Web as Corpus Workshop, ACL 2016, Berlin
August 12, 2016

Background

- Reliable metadata: not available for large crawled web corpora
- Topic domain (and genre/register): essential for many corpus linguists
- Also important for corpus evaluation and corpus comparison

- Automatic classification by genre/register: in unrestricted domains, disappointing results, even in recent experiments.
- Biber and Egbert (2016): acc.=0.42, prec.=0.27, rec.=0.3

Automatic classification by content

- Promising results years ago already (Sebastiani, 2002).
- Data-driven induction of topics: a very objective way of organizing a collection of documents by content.
- Topic classification through internal criteria: also advocated in the EAGLES (1996) guidelines

But:

- Topic modeling: no category labels
- from a linguist's viewpoint: categories should be 'intuitively' interpretable

Experiment

Idea

1. Infer a topic distribution over a corpus using topic modeling algorithms (unsupervised)
2. Do not interprete the inferred topical structure directly.
3. Instead, learn a small set of topic domains from the documents' assignment to the topics (supervised)

Experiment

Idea

1. Infer a topic distribution over a corpus using topic modeling algorithms (unsupervised)
2. Do not interprete the inferred topical structure directly.
3. Instead, learn a small set of topic domains from the documents' assignment to the topics (supervised)

Goals

- Development of a suitable annotation scheme for topic domain, grounded in lexical distributions
- Corpus comparison: web corpus vs. newspaper corpus (very little is known about the composition of crawled web corpora)

COWCat

Text classification schema (Schäfer and Bildhauer, 2012)

- No complex categories such as *genre*, *register* etc.
- Instead simple categories: *Aim*, *Mode*, *Topic Domain*
- Builds on previous work by Sharoff (2006)

COWCat

Text classification schema (Schäfer and Bildhauer, 2012)

- No complex categories such as *genre*, *register* etc.
- Instead simple categories: *Aim*, *Mode*, *Topic Domain*
- Builds on previous work by Sharoff (2006)

Topic Domain

- Design goal: moderate number (about 10–20) categories
- Basis for our classifcation experiments: 13 categories
- Developed in a cyclic fashion
  (repeated annotation processes, annotator feedback)

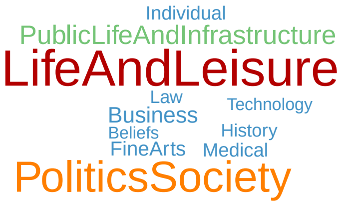
Step 1: Creating a gold standard data set

- 870 documents from DECOW14, crawled web corpus (Schäfer and Bildhauer, 2012; Schäfer, 2015)
- 886 documents from DeReKo, mostly newspaper texts (Kupietz et al., 2010)
- Manually annotated with COWCat for topic domain (COWCat)

- Annotators: Sarah Dietzfelbinger, Lea Helmers, Theresia Lehner, Kim Maser, Samuel Reichert, Luise Rißmann (FU Berlin); Monica Fürbacher (IDS Mannheim)

Distribution of topic domains

Comparison of DeReKo and DECOW14



Step 2: Topic modeling

- Starting point: term-document matrix
- Documents: weighted assignment to topics
- Topics: defined by a set of weighted words

Step 2: Topic modeling

- Starting point: term-document matrix
- Documents: weighted assignment to topics
- Topics: defined by a set of weighted words

Our experiment:

- LSI (Landauer and Dumais, 1994)
  LDA (Blei et al., 2003)
  as implemented in `Gensim` (Řehůřek and Sojka, 2010)
- LDA topic distributions unstable (small gold standard corpora)
- Incrementally add other documents from the source copora

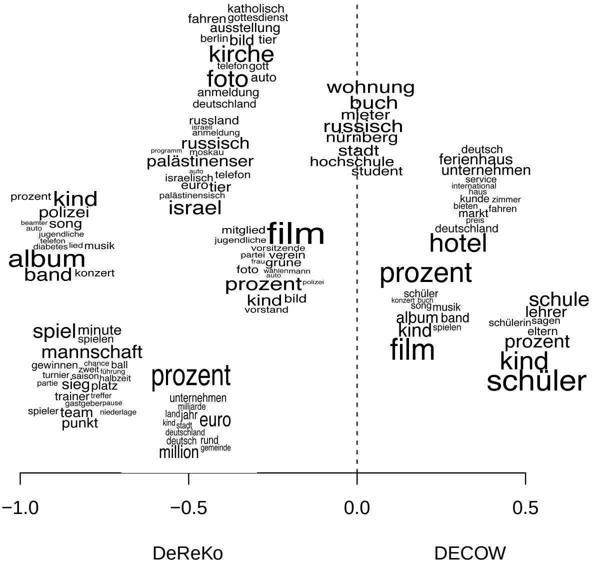
Step 2: Topic modeling

- Starting point: term-document matrix
- Documents: weighted assignment to topics
- Topics: defined by a set of weighted words

Our experiment:

- LSI (Landauer and Dumais, 1994)
  LDA (Blei et al., 2003)
  as implemented in `Gensim` (Řehůřek and Sojka, 2010)
- LDA topic distributions unstable (small gold standard corpora)
- Incrementally add other documents from the source copora
- Input terms: lemma + simplified POS tag (*kindergarten_nn*)
- Filtering: best results with lower-cased, purely alphabetic noun lemmas, 4–30 chars long

Corpus comparison: distribution of (selected) LSI-topics



Step 3: Learning topic domains from LSI-topics

- Supervised classifiers
- Permutation of virtually all available classifiers in `Weka` (Hall and Witten, 2011)
- Highest accuracy: SVMs with a Pearson VII kernel (Üstün et al., 2006)

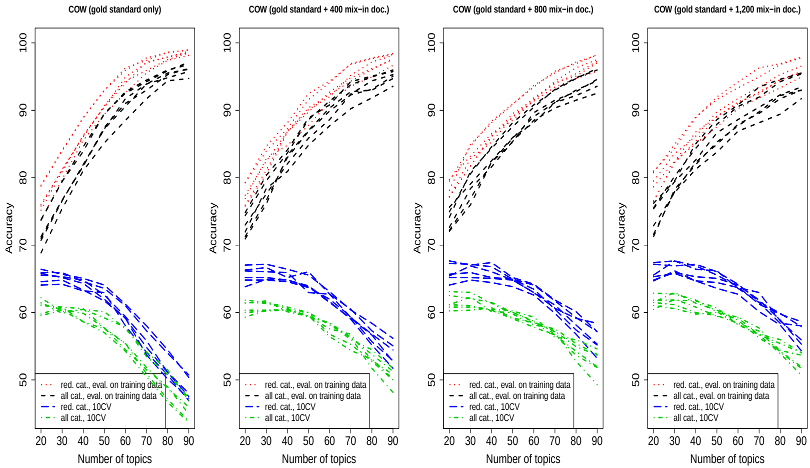
Step 3: Learning topic domains from LSI-topics

- Supervised classifiers
- Permutation of virtually all available classifiers in `Weka` (Hall and Witten, 2011)
- Highest accuracy: SVMs with a Pearson VII kernel (Üstün et al., 2006)

Set of experiments with:

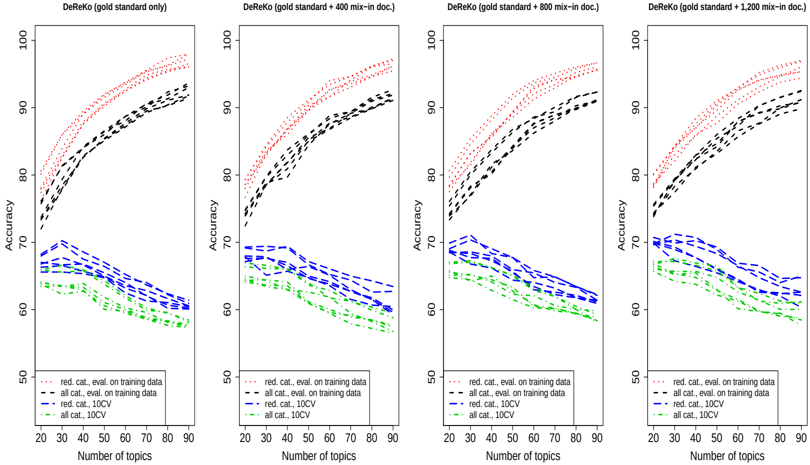
- varying number of LSI-topics
- topics induced from
  - gold standard data plus varying amounts of additional documents
  - several pre-processing variants
- evaluation on the *full* data set and on a *reduced* data set (with rare categories removed)

Results: Web (accuracy)



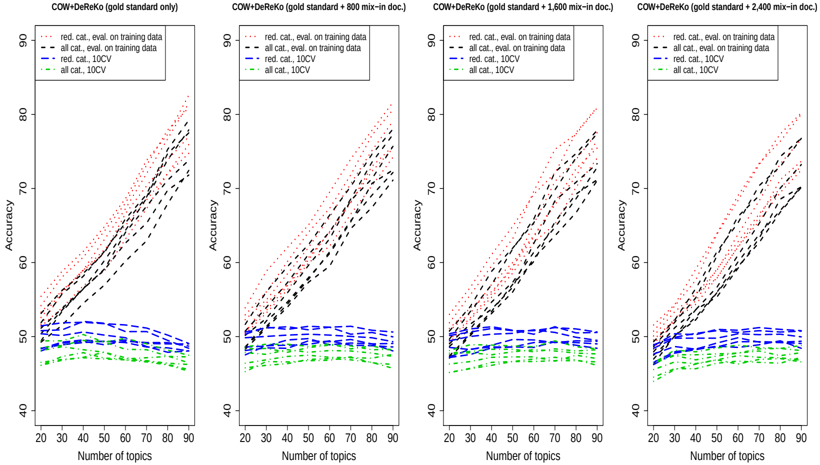
Mixed-in	Attribute	Topics	Accuracy	Precision	Recall	F-Measure
3,200	token	20	68.765%	0.688	0.688	0.674

Results: News (accuracy)



Mixed-in	Attribute	Topics	Accuracy	Precision	Recall	F-Measure
3,600	lemma + POS	40	72.999%	0.725	0.730	0.696

Results: Web + News (accuracy)



Mixed-in	Attribute	Topics	Accuracy	Precision	Recall	F-Measure
0	lemma + POS	30	51.872%	0.431	0.519	0.417

Results: all

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

- Web + News: larger training set does not increase accuracy
- Web + News: mixing in more documents for topic modeling does not increase accuracy
- News: higher accuracy (4.23%) probably a side effect of the more skewed distribution of topic domains in gold standard data

Confusion matrices

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**

References I

Biber, Douglas and Egbert, Jesse. 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.

Blei, David M., Ng, Andrew Y. and Jordan, Michael I. 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.

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

Kupietz, Marc, Belica, Cyril, Keibel, Holger and Witt, Andreas. 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 (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, pages 1848–1854, Valletta, Malta: European Language Resources Association (ELRA).

References II

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

Řehůřek, Radim and Sojka, Petr. 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.

Schäfer, Roland. 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 (eds.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, UCREL, Lancaster.

Schäfer, Roland and Bildhauer, Felix. 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 (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul: ELRA.

References III

Sebastiani, Fabrizio. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1), 1–47.

Sharoff, Serge. 2006. Creating General-Purpose Corpora Using Automated Search Engine Queries. In Marco Baroni and Silvia Bernardini (eds.), *Wacky! Working papers on the Web as Corpus*, Bologna: GEDIT.

Üstün, Bülent, Melssen, Willem J. and Buydens, Lutgarde M.C. 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.