Freie Universität Berlin

DFG

INSTITUT FÜR DEUTSCHE SPRACHE

Induktive Topikmodellierung und extrinsische Topikdomänen

Felix Bildhauer[1] und Roland Schäfer[2]

[1]Abt. Grammatik IDS Mannheim, [2]Ling. Webcharakterisierung (DFG) FU Berlin

IDS Jahrestagung 2016, Mannheim

Hintergrund

- Textklassifikation als Metadaten für sehr große Korpora
- Kovariation grammatischer, lexikalischer und externer Merkmale
- Korpusevaluation und Korpusvergleich
  [Kilgarriff, 2001, Biemann et al., 2013, Schäfer and Bildhauer, 2013]

- linguistischer Wunsch nach Interpretierbarkeit

- schlechte Qualität bei automatischer Auszeichnung
  komplexer Genre- und Register-Kategorien [Biber and Egbert, 2016]

COWCat

- Textklassifikationsschema mit möglichst einfachen Kategorien
- keine komplexen Kategorien wie Genres, Register usw.
- *Autorschaft*, *Äußerungsabsicht*, *Äußerungsmodus*, *Topikdomäne*

- Basis für automatische Klassifikationsversuche

- Sharoff [2006], Schäfer and Bildhauer [2012]
  `http://corporafromtheweb.org/cowcat-categorization-scheme/`

Korpora und Goldstandards

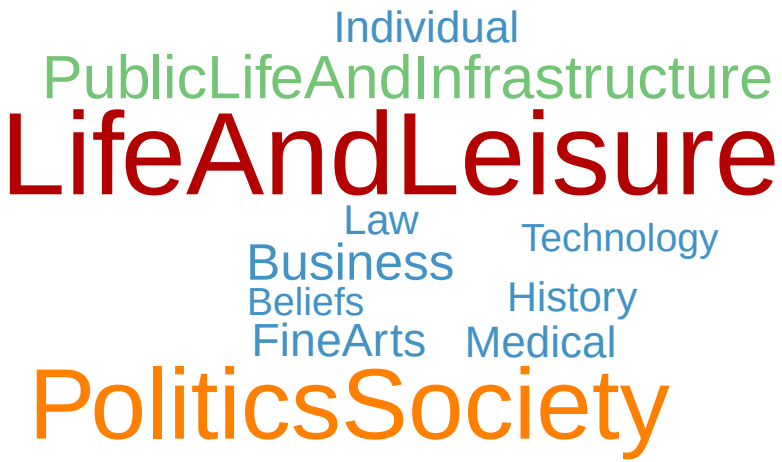
- DECOW14A: 17 Mrd. Wörter Webdaten
  [Schäfer and Bildhauer, 2012, 2013, Schäfer, 2015]
- DeReKo: 28 Mrd. Wörter überwiegend Zeitungstexte
  [Kupietz et al., 2010]

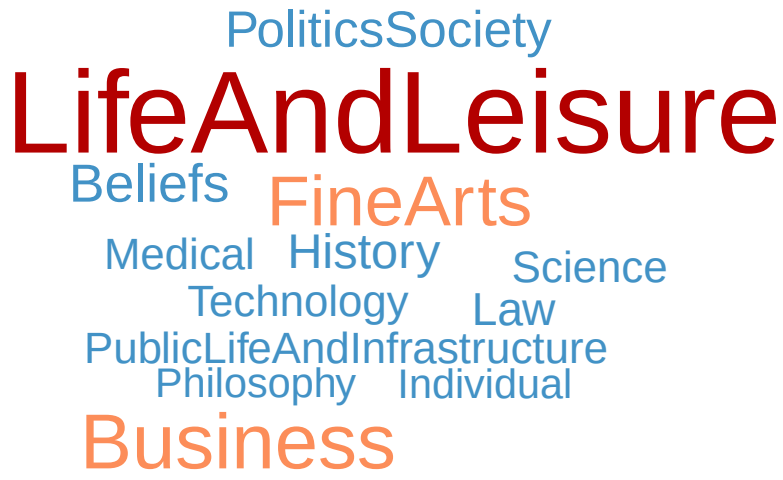
- aus beiden Korpora: manuell annotierte Goldstandards
  für Topikdomänen mit ca. 850 Dokumenten

- Dank an unsere AnnotatorInnen Sarah Dietzfelbinger, Lea Helmers, Theresia Lehner, Kim Maser, Samuel Reichert, Luise Rißmann (FU Berlin); Monica Fürbacher (IDS Mannheim)

Verteilung der Topikdomänen

Wie ähnlich sind sich die beiden Korpora?





Topikmodellierung

- Ziel: unüberwachte Induktion von Topiks
- Ausgangspunkt der Induktion: Term-Dokument-Matrix
- Dokumente gewichtet den Topiks zugeordnet
- Topiks definiert durch gewichtete Wörter
- Gegensatz zu a priori gegebenen Taxonomien:
  nicht strittig und nicht diskret

Topikmodellierung

- Ziel: unüberwachte Induktion von Topiks
- Ausgangspunkt der Induktion: Term-Dokument-Matrix
- Dokumente gewichtet den Topiks zugeordnet
- Topiks definiert durch gewichtete Wörter
- Gegensatz zu a priori gegebenen Taxonomien:
  nicht strittig und nicht diskret

- z. B. Latent Semantic Indexing [Landauer and Dumais, 1994],
  Latent Dirichlet Allocation [Blei et al., 2003]

- für unser Experiment: LSI (Gensim; Řehůřek and Sojka, 2010)

Topiks auf Topikdomänen abbilden

Idee

- Fundierung und Optimierung oft umstrittener gegebener Klassifikationsschemata (Topikdomänen) anhand lexikalischer Verteilungen in den Texten
- gewichtete Zuordnung der Dokumente zu Topiks als Grundlage für überwachtes Lernen von Topikdomänen

Experimente

- identische Vorverarbeitung (COW-Toolchain)
- Ausfilterung schwach repräsentierter Kategorien
- Stapeltests mit verschiedenen Klassifikatoren

Ergebnisse und Probleme

Evaluation

Corpus	Accuracy	Precision	Recall	F-Measure
COW	68.765%	0.688	0.688	0.674
DeReKo	72.999%	0.725	0.730	0.696
COW + DeReKo				

Ergebnisse und Probleme

Evaluation

Corpus	Accuracy	Precision	Recall	F-Measure
COW	68.765%	0.688	0.688	0.674
DeReKo	72.999%	0.725	0.730	0.696
COW + DeReKo				

Ergebnisse und Probleme

Evaluation

Corpus	Accuracy	Precision	Recall	F-Measure
COW	68.765%	0.688	0.688	0.674
DeReKo	72.999%	0.725	0.730	0.696
COW + DeReKo				

Ergebnisse und Probleme

Evaluation

Corpus	Accuracy	Precision	Recall	F-Measure
COW	68.765%	0.688	0.688	0.674
DeReKo	72.999%	0.725	0.730	0.696
COW + DeReKo	?	?	?	?

Ergebnisse und Probleme

Evaluation

Corpus	Accuracy	Precision	Recall	F-Measure
COW	68.765%	0.688	0.688	0.674
DeReKo	72.999%	0.725	0.730	0.696
COW + DeReKo	?	?	?	?

Ergebnisse und Probleme

Evaluation

Corpus	Accuracy	Precision	Recall	F-Measure
COW	68.765%	0.688	0.688	0.674
DeReKo	72.999%	0.725	0.730	0.696
COW + DeReKo	?	?	?	?

Ergebnisse und Probleme

Evaluation

Corpus	Accuracy	Precision	Recall	F-Measure
COW	68.765%	0.688	0.688	0.674
DeReKo	72.999%	0.725	0.730	0.696
COW + DeReKo	?	?	?	?

References I

D. Biber and J. Egbert. 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, 2016.

C. Biemann, F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, J. Simon, L. Swiezinski, and T. Zesch. Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics*, 28(2):23–60, 2013.

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

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

M. Kupietz, C. Belica, H. Keibel, and A. Witt. The German reference corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, pages 1848–1854, Valletta, Malta, 2010. European Language Resources Association (ELRA).

References II

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

R. Řehůřek and P. Sojka. 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, 2010. ELRA.

R. Schäfer. Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, and A. Witt, editors, *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster, 2015. UCREL.

R. Schäfer and F. Bildhauer. Building large corpora from the web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, 2012. ELRA.

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

References III

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