Creating dictionaries for argument identification by reference data

The creation of dictionaries is an important task to conceptualize and operationalize research questions in content analysis (Neuendorf, 2002). One can define concepts for coding operationalized variables in the form of mutual exclusive categories or decide if the content of documents is relevant for coding within the research task by the formalization of meaning trough a dictionary (Krippendorff, 2004). Dictionaries are often defined on the basis of a “theory of meaning that reflects a research question or the vocabulary of an academic discipline” (Krippendorff, 2004). Thus, we can think of dictionaries as operationalized representations of historical, sociological, cultural or political theories that are investigated within humanities research.

In contrast to manual dictionary creation from a small set of selected sample documents we present our approach to automatically extract dictionaries from a reference corpus of arbitrary size. For our goal of identifying arguments for a political science research task we create two dictionaries. One semantic dictionary on the utilization of topic models (Blei/Ng/Jordan, 2003; Teh/Jordan, 2010) to identify thematically relevant documents; and one rather syntactical dictionary based on term similarities of linguistic markers to identify a high density of argument structures. We present the idea, results and an example application of the extracted dictionaries for relevancy judging of retrieval results in large digital document collections.

Semantic dictionaries via topic models

Domain experts easily can compile a small reference corpus of paradigmatic documents containing contents of their interest. On this reference corpus we apply a topic model based on the Pitman-Yor Process (Teh, 2006). It employs Poisson instead of Dirichlet distributions which better approximate distributions of natural language data. One of the key properties of topic modeling is the inference of not directly observable variables considered as latent topics. A distribution over these latent topics (classes of co-occurring terms) is allocated to each of the documents within a digital text collection. Another hidden variable describes each of those topics in form of a probability distribution over the vocabulary of the text collection. On the basis of the assumption that all of the topics, extracted in a certain abstraction level controlled by the model parameters, represent the meaning and content of a digital text collection in a compressed form we created our dictionary extraction process. For this process we utilize the set of all resulting topics ***z*** to calculate scores for each word in the vocabulary within the collection. Since we have the property that only a few terms in a topic have high probability we use only a limited number of the most probable words in each topic. The score for each word is calculated by

$$score(w_n)=\log(F(w_n))\sum_{k=1}^{K} p(w_n|z_k) \quad ,$$

where $p(w_n|z_k)$ is the probability of the nth word in the vocabulary within the kth topic of the model and $F(w_n)$ is the absolute word frequency of the term w_n within the text collection. The idea behind this formula is that terms of high probability within a topic have significance for the meaning of the text collection. Furthermore we take the frequency of the word into account because a high frequent use of a term and a high probability within a topic induce a prototypical usage within the texts. Using topic models further allows for filtering of unwanted semantical structures when creating dictionaries from the collections. In our application we identified a foreign language 'topic' and a topic thematically not related to our research question in the reference texts and could easily exclude them from our *k* topics before applying the score calculation.

Syntactic dictionaries via term similarities

Additionally to our dictionary containing semantic information related to theoretical aspects of the political science research task we created a second dictionary of linguistic markers which can be employed to identify argumentative structures. We took a list of 46 German linguistic markers from another research project on causality and textual coherence (Breindl/Walter, 2009) as a starting point. This list was incrementally extended up to 144 terms by automatically computed synonyms of the markers retrieved from the database of the *“Projekt Deutscher Wortschatz”* (Quasthoff/Eckart, 2009), a representative corpus of German language.

Application

We applied these dictionaries for retrieval of documents in a large collection of newspaper articles to identify argumentative texts with a certain ideological framing. The retrieved texts then are subject of a close reading process of political scientists which utilize the dictionaries for qualitative coding schemes.

References

- Alsumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic Significance Ranking of LDA Generative Models. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I (S. 67–82). Berlin, Heidelberg: Springer-Verlag.
- Breindl, Eva / Walter, Maik (2009): Der Ausdruck von Kausalität im Deutschen. Amades - Arbeitspapiere zur deutschen Sprache, Mannheim.
- Blei, D.M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3, 993–1022.
- Krippendorff, K. (2004). Content analysis: an introduction to its methodology (2nd ed.). Thousand Oaks Calif.: Sage.
- Neuendorf, K. A. (2002). The content analysis guidebook. Thousand Oaks, Calif: Sage Publications.
- Niekler, A., & Jähnichen, P. (2012). Matching Results of Latent Dirichlet Allocation for Text. In Proceedings of ICCM 2012, 11th International Conference on Cognitive Modeling (S. 317–322). Universitätsverlag der TU Berlin.
- Quasthoff, Uwe / Eckart, Thomas (2009): Corpus Building Process of the Project "Deutscher Wortschatz". In: Lingustic Processing Pipelines Workshop at GSCL 2009.
- Teh, Y. W., & Jordan, M. I. (2010). Hierarchical Bayesian Nonparametric Models with Applications. In N. Hjort, C. Holmes, P. Müller, & S. Walker (Hrsg.), Bayesian Nonparametrics: Principles and Practice. Cambridge University Press.
- Teh, Yee Whye. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In Proceedings of the 21st International Conference on Computational Linguistics, 985–992.