Using Data Mining and the CLARIN Infrastructure to Extend Corpus-based Linguistic Research

Christian Pölitz[1], Thomas Bartz[2]
TU Dortmund University,
[1]Artificial Intelligence Group
[2]Department of German Language and Literature
44227 Dortmund, Germany
E-mail: christian.poelitz@tu-dortmund.de, thomas.bartz@tu-dortmund.de

Keywords: corpus-based linguistic and lexicographic studies, data mining, disambiguation

1. Introduction

Large digital corpora of written language such as those that are held by the CLARIN-D centres in Berlin, Mannheim or Tübingen offer excellent possibilities for linguistic research on authentic language data. The size of the corpora mostly allows remarkable insights into the distribution of interesting phenomena of language usage with respect to time and/or domain-specific aspects. Not least thanks to the efforts been done in CLARIN, analysing and query tools are getting more and more sophisticated, enabling researchers to search for word forms or constructions and filter the results with regard to part of speech types or morphosyntactic aspects. But despite these advances been made, the large number of hits that can be retrieved from the corpora often may also lead to challenges in concrete linguistic research settings. This is particularly the case, if the queried word forms or constructions are (semantically) ambiguous. Researchers in the linguistics usually are not examining word forms but terms representing relations of word forms and their meanings. That is why the word form-based filtering carried out by the query tools is not sufficient in many cases and leads to an unpredictable number of false positives. Depending on the amount of data, intense manual effort has to be done for cleaning and disambiguation tasks as, for example, described by Storrer (2011). Many research questions even cannot be addressed for this reason.

The project "KobRA" ("Korpus-basierte linguistische Recherche und Analyse mithilfe von Data-Mining", Eng.: Corpus-based Linguistic Research and Analysis using Data Mining), a project funded by the German BMBF (Federal Ministry of Education and Research), is therefore investigating benefits and issues of using machine learning technology in order to perform after-retrieval cleaning and disambiguation tasks automatically. To this end, German linguists, computational linguists and computer scientists closely cooperate on concrete corpus-based case studies in the fields of lexicography, diachronic linguistics and variational linguistics. The case studies reflect the research actually carried out in these fields. Three major German corpus providers (Berlin-Brandenburg Academy of Sciences and Humanities, BBAW; Institute for the German Language, IDS; Department of Linguistics at Tübingen University, SfS), that all are CLARIN-D centres as well, take part in the project, provide the corpus data and will integrate the project results into the existing infrastructure.

In our presentation, we will introduce the aims and the approach of the project KobRA as well as findings obtained so far. In particular, we will focus on case studies in the field of corpus-based lexicography, where we use topic modeling on top of corpus query result lists to automatically disambiguate queried words with more than one meaning. In the following, we go into detail on experiments with a choice of German words that are interesting from a linguist's point of view. As topic models operate independently from language, we suppose the method to be suitable for other languages than German as well. We therefore also ran experiments with English language data derived from the Leipzig Corpora Collection (also part of the CLARIN-D infrastructure).

In the following, we describe the corpora and queries used as well as our topic modeling approach, related work and our obtained results.

2. Words of interest and queried corpora

The basis of our experiments are query result lists for the German words "Leiter" and "zeitnah" derived from the DWDS core corpus of the 20th century (for "zeitnah" also from the DWDS newspaper corpus *Die ZEIT*; for details, see below). In order to investigate the fit of the proposed method for English language data, we also queried the English corpora of the Leipzig Corpora Collection (news, wikipedia and web data; for details see below) for the word "cloud".

The chosen words are supposed to provide interesting insights into processes of language change: "Der Leiter" (chief, director) and "die Leiter" (ladder) are homonyms with possible further senses *Energieleiter* (conducting medium) and *Tonleiter* (scale, in music), whereby "der Leiter" competes against borrowings like "Boss" or "Chef". In order to investigate the popular hypothesis that borrowings like these might replace indigenous German words over the decades of the 20th century, we need to disambiguate the word "Leiter". "Zeitnah", a polyseme meaning *zeitgenössisch* (contemporary), *zeitkritisch* (critical of the times) as well as *unverzüglich* (prompt), seems to have acquired the latter meaning as a new sense not before the second half of the last century. For this

word as well, semantic disambiguation on top of the used language data could enable linguists to trace its development. The same is true for the English word "cloud". Alongside the meaning *mass of condensed water, smoke, dust or other elements*, a new sense seems to become common: "cloud" with the meaning *remote server network* like the "Amazon AWS cloud" or "cloud computing".

The DWDS core corpus of the 20th century, constructed at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), contains approximately 100 million running words, balanced chronologically (over the decades of the 20th century) and by text genre (journalism, literary texts, scientific literature, other nonfiction). The newspaper corpus *Die ZEIT* covers all the issues of the German weekly newspaper *Die ZEIT* from 1946 to 2009, approximately 460 million running words (Klein & Geyken, 2010; Geyken, 2007). The Leipzig Corpora Collection consists of corpora in different languages that contain randomly selected sentences from newspaper texts and a web sample (Quasthoff, Richter & Biemann, 2006). We used the English corpus with language data from newspapers and the English Wikipedia covering the time span from 2005 to 2010. The corpus queries provide text snippets with occurrences of the investigated words (with inflected forms) in a context of three sentences (in a context of one sentence as regards the Leipzig Corpora Collection). In addition, the publication dates and, for the DWDS core corpus, information about the text genre classes the snippets are attributed to, are given for each snippet.

For evaluation purposes, 30 percent of the retrieved result snippets for the queried words were disambiguated manually by two independent annotators, that reached an agreement (kappa: Cohen, 1960) of 0.97 for "Leiter", 0.91 for "zeitnah" and 0.92 for "cloud".

3. Topic Models

We expect that certain distributions of words in each query result snippet and over the query result list correspond to certain meanings of a queried word. For snippets with the word "Leiter", for instance, we anticipate cooccurrences of words like "steigen" (climb), "auf" (upon), "hohe" (high) etc. being more likely for the meaning *ladder* than for other senses. Based on the query result lists, we train topic models to disambiguate the queried words. We use the probabilistic topic models called Latent Dirichlet Allocation (LDA) as introduced by Blei et al. (2003). LDA models the probability distributions of the words and the snippets from the result lists. The probability distributions are scattered over a number of so-called latent topics that correspond to different meanings of the queried word. We will interpret these topics as meanings like those mentioned above: *der Leiter*/*Person in leitender Position* (chief, director), *die Leiter*/*Sprossenstiege* (ladder) etc. Further, we expect to even find meanings that we have not considered beforehand.

The probability distributions of the topics for a given word or snippet are multinomial distributions φ respectively θ. These distributions are drawn from a Dirichlet distribution Dir(β) respectively Dir(α) for the meta parameter α and β. The Dirichlet distribution is a distribution over distributions.

The estimation of the distributions is done via a Gibbs sampler as proposed by Griffiths and Steyvers (2004). The Gibbs sampler models the process of assigning a word or snippet to a certain topic based on the topic distributions of all other topics. This is a Markov chain process and converges to the true topic distributions for given words and snippets.

An important aspect, that we investigate, is the possibility to integrate further information into the generation of the topic models. Steyvers et al. (2004), for instance, integrate additional information like authorships of documents in the topic models. We use their approach to integrate information about the text genre classes the query result snippets are attributed to. This enables an additional investigation of how topics, words and snippets distribute over these classes. Moreover, the integration of the publication dates provided with the snippets are interesting for us. Blei & Lafferty (2006), for example, introduced a dynamic topic model that facilitates analysing the development of the found topics over time.

In our experiments, we investigate how the approach described above can be used on query result lists from the DWDS corpora or from an English corpus of the Leipzig Corpora Collection to disambiguate the retrieved result snippets having regard to semantic change over time.

4. Experiments and Evaluation

We evaluate the topic models based on the manually disambiguated result snippets (see above). Therefore, we generate the topic models to extract the meanings of the queried words' occurrences and compare the results to the labels attributed by the annotators. We estimate the Normalized Mutual Information (NMI) as score for the goodness of the method. NMI measures how many snippets that are manually annotated as having different meanings are attributed to the same topic extracted by LDA. It is defined as the fraction of the entropies of the annotations' and the disambiguation results' distributions plus the entropy of annotations' and disambiguation results' joint distribution (Manning et al., 2008). Further, we use one of the standard measures to estimate the goodness of a word sense disambiguation result, the F_1 score. The F_1 score is the weighted average of the disambiguation results' precision and recall in relation to the given annotations. This and further evaluation methods are described by Navigli & Crisafulli (2010).

5. Results

In the following, we show the results achieved using the approach described above. The tables list the evaluation scores for the investigated words. For the English word "cloud" (see table 3), we were barely not able to reach the accuracy level that we achieved for the German data. This could be due to the smaller word context of only one

sentence in the data from the Leipzig Corpora Collection (see section 2) and has to be further investigated. Nonetheless, also for the English word "cloud", the evaluation scores show the use of the proposed method in general.

From the figures one can see benefits of the integration of the query snippet's publication dates into the generation of the topic models: Researchers investigating semantic change are enabled to easily track the use of disambiguated word forms over time. Going into more detail, it becomes apparent, for example, that the English borrowing "Boss" does not seem to replace the German word "Leiter", but to develop in parallel (see figure 1). The use of "zeitnah" with the meaning *unverzüglich* (prompt) is obviously highly increasing by the end of the 20th century (see figure 2), while "cloud", as well, seems to develop a new sense by the end of the 21st century's first decade (see figure 3).

"Leiter"	chief, director	ladder
F_1 score	0.97	0.85
NMI	0.300	

Table 1: Evaluation results for "Leiter".

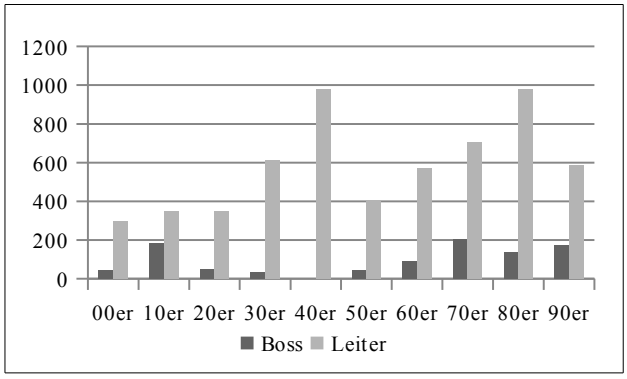

Figure 1: Occurrences of the words "Leiter" with the meaning *Chef, Vorgesetzter* (chief, director) and "Boss" in the decades of the 20th century.

"zeitnah"	contemporary	prompt
F_1 score	0.88	0.88
NMI	0.418	

Table 2: Evaluation results for "zeitnah".

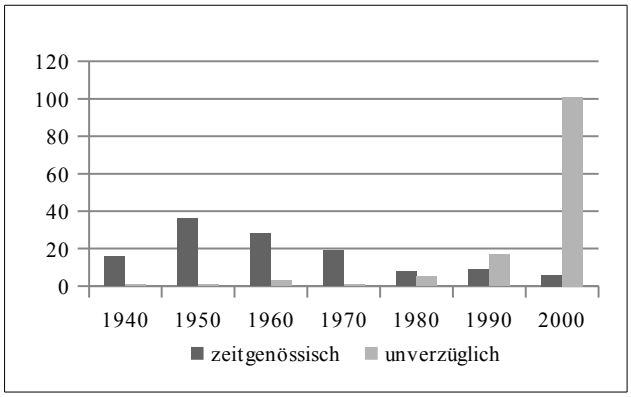

Figure 2: Occurrences of the word "zeitnah" with the senses *zeitgenössisch* (contemporary), *zeitkritisch* (critical of the times) and *unverzüglich* (prompt) in the time span 1940–2000.

"cloud"	mass of condensed water etc.	network
F_1 score	0.85	0.63
NMI	0.366	

Table 3: Evaluation results for "cloud".

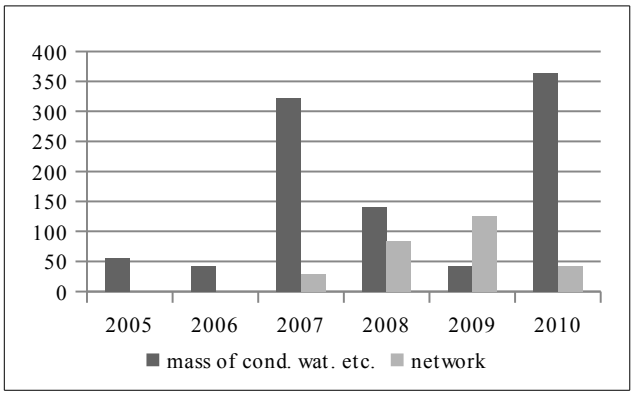

Figure 3: Occurrences of the word "cloud" with the senses *mass of condensed water, smoke, dust or other elements* and *remote server network* in the time span 2005–2010.

6. Acknowledgements

This work is funded by the German BMBF (Federal Ministry of Education and Research) under the grant 01UG1245A.

7. References

Blei, D.M., Lafferty, J.D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. New York: ACM, pp. 113--120.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), pp. 993--1022.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), pp. 37--46.

Geyken, A. (2007). The DWDS corpus. A reference corpus for the German language of the twentieth

century. In C. Fellbaum (Ed.), *Idioms and collocations. corpus-based linguistic and lexicographic studies.* London: Continuum, pp. 23--40.

Griffiths, T.L., Steyvers, M. (2004). Finding scientific topics, *Proceedings of the National Academy of Sciences* 101 (suppl 1), pp. 5228--5235.

Klein, W., Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). *Lexikographica*, 29(1), pp. 79--93.

Manning, C.D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press.

Navigli, R., Crisafulli, G. (2010). Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, pp. 116--126.

Quasthoff, U., Richter, M., Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation*. Genoa, pp. 1799--1802.

Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: ACM, pp. 306--315.

Storrer, A. (2011). Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In K. Knapp et al. (Eds.), *Angewandte Linguistik. Ein Lehrbuch.* Tübingen: Franke, pp. 216--239.