Enhancing the possibilities of corpus-based investigations: Word sense disambiguation on query results of large text corpora

Christian Poelitz
Technical University Dortmund
Artificial Intelligence Group
44227 Dortmund, Germany
poelitz@tu-dortmund.de

Thomas Bartz
Technical University Dortmund
Institute of German Language and Literature
44227 Dortmund, Germany
bartz@tu-dortmund.de

Abstract

Common large digital text corpora do not distinguish between different meanings of word forms, intense manual effort has to be done for disambiguation tasks when querying for homonyms or polysemes. To improve this situation, we ran experiments with automatic word sense disambiguation methods operating directly on the output of the corpus query. In this paper, we present experiments with topic models to cluster search result snippets in order to separate occurrences of homonymous or polysemous queried words by their meanings.

1 Introduction

Large digital text corpora contain text documents from different sources, genres and periods of time as well as often structural and linguistic markups. Nowadays, they provide novel and enhanced possibilities of exploring research questions at the basis of authentic language usage not only in the field of linguistics, but for humanities and social sciences in general. But even though tools for query and analysis are getting more and more flexible and sophisticated (not least thanks to the efforts been done in infrastructure projects like CLARIN), automatically obtained data have to be reviewed manually in most cases because of false positives. Depending on the amount of data, intense manual effort has to be done for cleaning, classification or disambiguation tasks. Hence, many research questions cannot be addressed because of time constraints (Storrer, 2011). A project funded by the German BMBF (Bundesministerium für Bildung und Forschung, ”Federal Ministry of Education and Research”), therefore, is investigating benefits and issues of using machine learning technology in order to perform these tasks automatically. In this paper, we focus on the disambiguation task, which is an issue known for a long time in the field of corpus-based lexicography (Engelberg and Lemnitzer, 2009), but has not been satisfactorily solved, yet, and is still highly relevant also to social scientists or historians. In the humanities, researchers usually are not examining word forms, but terms representing relations of word forms and their meanings. While the common large corpora do not distinguish between different meanings of word forms, the disambiguation task has to be carried out manually most of the times. To improve this situation, we ran experiments with word sense disambiguation methods operating directly on the output of the corpus queries, i.e. search result lists containing small snippets with the occurrences of the search keyword, each in a context of about only three sentences. In particular, we used topic modelling to automatically detect clusters of keyword occurrences with similar contexts, that we consider corresponding to a certain meaning of the keyword. In the following, we report our findings from experiments with the German terms *Leiter* and *zeitnah*, both 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*. *Zeitnah*, a polyseme meaning *zeitgenssisch* ”contemporary”, *zeitkritisch* ”critical of the times” as well as *unverzglich* ”prompt”, seems to have acquired the latter meaning as a new sense not until the second half of the last century. The basis of our experiments are search result lists derived from the DWDS Kernkorpus core corpus of the 20th century (for *Leiter*) and, in addition, from the ZEIT corpus (for *zeitnah*). The DWDS Kernkorpus, constructed at the Berlin-Brandenburg Academy of Sciences (BBAW), contains approximately 100

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 42–46,
Gothenburg, Sweden, April 26 2014. ©2014 Association for Computational Linguistics

million running words, balanced chronologically (over the decades of the 20th century) and by text genre (over the genres journalism, literary texts, scientific literature and other nonfiction; (Geyken, 2007)). The ZEIT corpus covers all the issues of the German weekly newspaper *Die Zeit* from 1946 to 2009, approximately 460 million running words (http://www.dwds.de/ressourcen/korpora).

2 Related Work

Word sense disambiguation is a well studied problem in Machine Learning and Natural Language Processing. For a given word, later mentioned as word of interest, we expect that there exist several meanings. The differences in the meanings are reflected by different words occurring and frequencies together with the word to be disambiguated. A very early statistical approach was proposed by (Brown et al., 1991). The authors proposed to estimate the probability distribution of senses for given words from annotated examples. A general survey about the topic can be found in (Navigli, 2009). Latent Dirichlet Allocation (LDA) introduced by (Blei et al., 2003) can be used to estimate topic distributions for a given document corpus. Each topic represent a sense in which the documents, respectively the words, appear. (Griffiths and Steyvers, 2004) proposed efficient training for LDA using Monte Carlo sampling. They used Gibbs sampling to estimate the topic distribution. The authors in (Brody and Lapata, 2009) extend the generative model by LDA by many parallel feature representations. Hence, beside the pure words, additional features like part of speech tags can be used. Further, the authors perform analysis with different context sizes. Investigations of word sense disambiguation on small snippets have been done before on search engine results. The snippets retrieved after a query has been sent to a search engine are used for disambiguation. In (Navigli and Crisafulli, 2010) for instance, the authors search for word senses of web search results using retrieved snippets.

Our approach differs from these previous ones since we concentrate on snippets from a text corpus for linguistic and lexicographic research purposes (see Section 1). Unlike results from search engines, that refer to documents whose topics are strongly related to the search keyword, result lists from text corpora contain snippets with occurrences of the keyword in each document of the corpus, irrespective of the document topic. That is why keywords can occur in less typical, semantically less definite contexts. In the literary documents, they are not infrequently used as metaphors.

3 Snippet Representation

In order to properly apply Machine Learning methods for word sense disambiguation we need to encode the snippets in an appropriate way. Therefore, we represent each snippet as bag-of-words. This means we build a large vector that contains at the component i the number of times word i - from the overall vocabulary of the document corpus - appears in the snippet. These vectors are very sparse and can be efficiently saved as hash tables.

Since we want to investigate different context information for the disambiguation, we generate for each snippet many different bag-of-words representations. First, we use only those words that appear in close proximity to the word we want to disambiguate. This means, we place a window on the text, that contains a certain number of words that appear before and after the word of interest. Next, we filter out words that are not immediate constituents (or immediate constituents of the 1st, 2nd, nth superordinate node) of the word of interest. In this case the proximity is not crucial but the syntactical relatedness to the word of interest.

These word vectors are used for the word sense disambiguation.

4 Disambiguation

For the word sense disambiguation we use Latent Dirichlet Allocation (LDA) as introduced by (Blei et al., 2003). LDA estimates the probability distributions of words and documents, respectively snippet, over a number of different topics. The topics will be used to disambiguate the word of interest. These distributions are drawn from Dirichlet distributions that depend on given meta parameters α and β.

The probability of a topic, given a snippet is modelled as Multinomial distribution that depends on a Dirichlet distributed distribution of the snippets over the topics. Formally we have: $\phi \sim Dirichlet(\beta)$ the probability distribution of a snippet and $p(z_i|\phi(j)) \sim Multi(\phi(j))$ the probability of topic z_i for a given snippet j.

To estimate the distributions we use a Gibbs

Leiter	w10	w40	w80	all	syntax
NMI	0.2086	0.2579	0.2414	0.2573	0.1944
zeitnah	w10	w40	w80	all	syntax
NMI	0.1012	0.1926	0.1656	0.2230	0.0456

Table 1: NMI of the extracted senses with respect to the given annotations of the text snippets.

Leiter	w10	w40	w80	all	syntax
F1	0.7271	0.7487	0.7405	0.7416	0.6904
zeitnah	w10	w40	w80	all	syntax
F1	0.7773	0.6919	0.7630	0.7488	0.4584

Table 2: F1 score of the extracted senses with respect to the given annotations of the text snippets.

sampler as proposed by (Griffiths and Steyvers, 2004). The Gibbs sampler models the probability distributions of a given topic z_i, depending on all other topics and the words in the snippet as Markov chain. This Markov chain converges to the posterior distribution of the topics given the words in a certain snippet. This posterior can be used to estimate the most likely topic for a given snippet.

Further, we use the author topic model as introduced by (Steyvers et al., 2004). This model integrates additional indications about the author for each snippet into the topic modelling process. This method can also be used to model the text categories instead of authors. We simply treat the categories as the authors. Now, the probability distribution of the topics additionally depends on the random variable c over the categories. This can be leveraged to estimate the probability of category c for a given topic z_i, hence $p(c|z_i)$.

Using the author topic model, we estimate the topic distribution over words and categories. Based on these distributions the stochastic process of generating topics is simulated. Depending on the number of times a topic is drawn for a given snippet and category, we extract the most likely words and categories for the topics. The topics represent the different senses of the word of interest.

5 Experiments

We performed experiments on two data sets that consist of short snippets retrieved by corpus queries for the words *Leiter* and *zeitnah* in the DWDS Kernkorpus `www.dwds.de` and the ZEIT corpus (see Section 1). Each snippet consists of the three sentences, whereby the second sentence contains the search keyword (the word to disambiguate) in each case. The snippets belong to the different text categories covered by the mentioned corpora: journalism, literary texts, scientific literature and other nonfiction (see Section 1). For each snippet, we have information to which category it belongs to. This information is used only for validation, not for the topic extraction. For each data set, 30 percent of snippets were disambiguated manually by two independent annotators, whereby doubtful cases were clarified by a third person. The annotations are not used for disambiguation, but for the validation of the method.

For each snippet we generate bag-of-words vectors using contexts of 10, 40, 80 or all words around the word of interest. Hence, for context size 10 we use the ten words before the token, the token itself and the ten following tokens, as representation of the snippet. For further experiments we used the Stanford Constituent Parser (Klein and Manning, 2003) to get only the words that syntactically depend on the words of interest. For the extraction of the topics and distribution over the text categories we used the Gibbs sampler for LDA and the author topic model from the Matlab library Topictoolbox (Griffiths and Steyvers, 2004).

Based on the annotation mentioned above we can estimate the Normalized Mutual Information (NMI) as score for the goodness of the method. NMI measures how many snippets that are annotated as being from different topics are placed into the same topic based on the extracted topics from LDA. It is defined as the fraction of the sum of the entropies of the distributions of the annotations and the disambiguation results, and the entropy of the joint distribution of annotations and results (Manning et al., 2008) (p. 357f). Further, we use one of the standard measures to estimate the goodness of a word sense disambiguation result, the F1 score. The F1 score is the weighted average of the precision and recall of the disambiguation results for the given annotations. This and further evaluation methods are described in (Navigli and Vannella, 2013).

In the Tables 1 and 2 we show the NMI and F1 score for the extracted topics, respectively senses, by LDA. We tested different context sizes from 10 to 80 words around the word of interest. Compared to the results when we use the whole snippets, we see that a context size of 40 results in the

Sense 1	Sense 2	Sense 3	Sense 4
music	standing	GDR [1]	government
Berlin	saw	SED [2]	got
Prof	up	party	Berlin
Comp	above	political	ZK [3]

Table 3: Translation of the most frequent words for each of the extracted senses for the word *Leiter*.

Sense 1	Sense 2	Sense 3	Sense 4
question	society	German	publisher
DM [4]	just	time	book
years	examples	film	literature
music	questions	Berlin	year

Table 4: Translation of the most frequent words for each of the extracted senses for the word *zeitnah*.

best performance. Less context decrease the performance and the filtering by constituencies give the worst results. The experiments show that a windowing approach is well suited to represent documents for a word sense disambiguation task. The size of the window seems to be crucial and must be chosen a priori. Optimal window size could be found by cross validation techniques using annotated snippets.

Next, we investigated the distribution of the topics over the text categories. We used the author topic model as described above to estimate how the categories distribute over the sense. Tables 3 and 4 show the most likely words to appear in the corresponding senses translated into English for four extracted topics. In the Tables 5 and 6 the distribution of the senses over the given categories are presented. Based on the posterior distribution of the categories, we simulated the process of assigning topics to categories for each word in the snippets. In the tables we present the number of times we assign sense i to category c.

For the word *Leiter* in Table 5, we see that in each category always one certain sense for the word is prominent. For instance sense 2, here *Leiter* appears in the context of a ladder. In this context, the word is more likely to appear in a fictional text than in the other categories. For *zeitnah* in Table 6 the results are not very clear. First, the word is most likely to appear in news papers rather than in literature or science articles. This is due to the fact that we have much more snippets from news papers. Only in sense 3, the word is also likely to appear in other categories. This context seems to be German films. In contrast, we see sense 2 that is about social questions appears only in news papers.

Leiter	Sense 1	Sense 2	Sense 3	Sense 4
Literature	597	23818	7464	6718
Non-fiction	3031	5295	63708	8733
Science	41564	3269	1216	1046
Journalism	5527	8845	23104	78645

Table 5: The distribution of the senses among the text categories during the simulation for the word *Leiter*.

zeitnah	Sense 1	Sense 2	Sense 3	Sense 4
Literature	23	0	12	6
Non-fiction	1	0	574	10
Science	211	0	478	1
Journalism	2150	2438	1691	2924

Table 6: The distribution of the senses among the text categories during the simulation for the word *zeitnah*.

6 Conclusion and Future Work

We used topic models to cluster search result snippets received by queries in two large digital text corpora in order to separate occurrences of homonymous or polysemous queried words by their meanings. We showed that LDA performs well in extracting the senses in which the words appear. Finally, we found that the author topic model can be used to estimate how the extracted senses distribute over document categories.

For the future, we want to further investigate the distribution of the topics over different categories and time periods, as first experiments showed potential benefit of the author topic model. An important point for future work is, moreover, the integration of syntactic features not only for filtering important words but also for enhancement of our simple bag-of-words representation. Especially, the integration of constituency and dependency information will be further investigated.

Aknowledgements

The authors are supported by the Bundesministerium für Bildung und Forschung (BMBF) in project KobRA[5].

[5] http://www.kobra.tu-dortmund.de/

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 103–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, pages 264–270, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stefan Engelberg and Lothar Lemnitzer. 2009. *Lexikographie und Woerterbuchbenutzung*. Stauffenburg, Tuebingen.

Alexander Geyken. 2007. The DWDS corpus. A reference corpus for the German language of the twentieth century. In Christiane Fellbaum, editor, *Idioms and collocations. corpus-based linguistic and lexicographic studies*, pages 23–40. Continuum, London.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Roberto Navigli and Giuseppe Crisafulli. 2010. Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 116–126, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roberto Navigli and Daniele Vannella. 2013. Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69, February.

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 306–315, New York, NY, USA. ACM.

Angelika Storrer. 2011. Korpusgesttzte sprachanalyse in lexikographie und phraseologie. In Karlfried Knapp et al., editor, *Angewandte Linguistik. Ein Lehrbuch*, pages 216–239. Francke Verlag, Tuebingen.