Exploring the Potential of Semantic Relatedness in Information Retrieval

Christof Müller and **Iryna Gurevych**
Ubiquitous Knowledge Processing Group
Telecooperation, Darmstadt University of Technology
Hochschulstr. 10, 64289 Darmstadt, Germany
http://www.cre-elearning.tu-darmstadt.de/elearning/sir/

Abstract

Employing lexical-semantic knowledge in information retrieval (IR) is recognised as a promising way to go beyond bag-of-words approaches to IR. However, it has not yet become a standard component of IR systems due to many difficulties which arise when knowledge-based methods are applied in IR. In this paper, we explore the use of semantic relatedness in IR computed on the basis of GermaNet, a German wordnet [Kunze, 2004]. In particular, we present several experiments on the German IR benchmarks GIRT'2005 (training set) and GIRT'2004 (test set) aimed at investigating the potential of semantic relatedness in IR as opposed to bag-of-words models, as implemented e.g. in Lucene [Gospodnetic and Hatcher, 2005]. These experiments shed some light upon how to combine the strengths of both models in our future work. Our evaluation results show some improvement in IR performance over the bag-of-words model, i.e. a significant increase in mean average precision of about 5 percent points for the training set, but only 1 percent increase for our test set.

1 Introduction

It is often assumed that the use of linguistic, in particular lexical-semantic information, should improve the performance of bag-of-words IR systems, which are based on string matching. The problems with the bag-of-words IR systems arise due to polysemy and synonymy in the natural language. Polysemy of words creates ambiguity and can lead to poor precision due to the words, which are not sense disambiguated. If the synonymy is not taken into account, the recall of the system would be poor, as it does not find relevant documents containing terms, which are synonymous to the search term.

Multiple attempts have been made to address these issues by employing Natural Language Processing (NLP) methods in IR with so far limited success. In many cases, the use of semantic knowledge captured in computer-readable resources like WordNet [Fellbaum, 1998] or FrameNet [Baker *et al.*, 1998] has been explored for the task of disambiguating or selecting related words and thereby improving the performance of IR systems. There exist multiple ways to incorporate lexical-semantic knowledge into IR systems:

- *query expansion* where the query is extended by semantically related terms;
- *indexing concepts* instead of words;
- *document ranking functions* based on lexical-semantic knowledge.

In the paper, we describe a set of experiments aimed to integrate lexical-semantic knowledge into an IR system by using semantic relatedness as the model of relevance between query and documents. Section 2 describes the application domain and corpora used in our experiments. In Section 3, we introduce our baseline system and the newly developed semantic relatedness retrieval model including necessary preprocessing steps of documents and queries. Evaluation results follow in Section 4. After that, we put our work into context by extensively reviewing the state-of-the-art on integrating semantic knowledge in information retrieval in Section 5. Finally, we draw some conclusions in Section 6 and develop some ideas for further research.

2 Corpus Data

In this paper, we conduct experiments with the GIRT corpus, which is a domain-specific corpus devoted to the domain of social science and a standard information retrieval benchmark for German [Kluck, 2004]. It is used in the German domain-specific task at CLEF, which allows to make cross-system comparisons for this task. The corpus consists of abstracts of scientific papers in social science, together with the author and title information and several keywords. The experiments described in Section 4 use the topics and relevance assessments of CLEF'2005. A topic is a natural language statement of information need which is used to create a query for an IR system. For CLEF'2005 there are 25 topics for the GIRT corpus in the German language. Each topic consists of three different parts: a title (keywords), a description (a sentence), and a narration (exact type specification of documents to retrieve). A portion of GIRT documents is annotated with relevance judgements for each topic by using the *pooling technique*. Table 1 shows some statistics about the corpus and topics.

3 System Architecture

In this study, we compare two kinds of IR models on the GIRT corpus: an IR model as implemented by Lucene[1] as the baseline and a model integrating semantic relatedness.

3.1 Document and Query Preprocessing

During the preprocessing of documents and queries we apply several NLP methods which are commonly used in many state-of-the-art IR systems. These include tokenisation, stopword removing, stemming, lemmatisation and compound splitting.

[1] http://lucene.apache.org

	#docs	*#tokens*	*#distinct tokens*	*#tokens/doc (mean)*
GIRT4	151,319	14,312,116	525745	94.58
Topics CLEF2005:	25	–	–	–
Title	–	45	44	1.8
Narration	–	181	104	7.24
Description	–	517	281	20.68

Table 1: Corpus and query statistics (number of tokens counted after removing stopwords and lemmatising).

We perform both, stemming and lemmatisation, but use only lemmas for query and index building. In the future, we will compare the retrieval performance with lemmatised and stemmed indexes. There have already been a number of studies about the usefulness of morphological normalisation in IR. Some of the most recent ones are [Hollink *et al.*, 2004] and [Airio, 2006]. They confirm the positive impact which morphological normalisation has, especially for German. However, they find almost no difference in performance between stemming and lemmatisation. For our system, we use the Snowball Stemmer[2] and the lemmatiser of the TreeTagger [Schmid, 1994].

The third morphological normalisation we perform is the decomposition of compounds. The algorithm we use is based on [Langer, 1998] and uses GermaNet as the lexicon. Decompounding shows significant gain in performance for German in [Hollink *et al.*, 2004]. However, [Airio, 2006] can find almost no difference in performance.

3.2 Lucene-based IR

Lucene is an open source text search library based on an extended boolean (EB) model [Salton *et al.*, 1983]. The preprocessed topics are converted to a Boolean query, whereby separate terms are combined with the operator OR.

3.3 Semantic Relatedness

Semantic relatedness is defined as *any* kind of lexical-semantic or functional association that exists between two words [Gurevych, 2005b]. In order to compute semantic relatedness, lexical-semantic knowledge is required. This knowledge can be derived from a range of resources like computer-readable dictionaries, thesauri, or corpora. The experiments presented in this paper employ the German wordnet GermaNet as the knowledge base. Currently, GermaNet includes about 40000 synsets with more than 60000 word senses modelling nouns, verbs and adjectives. In previous work, the application of different semantic relatedness metrics to GermaNet has been explored [Gurevych and Niederlich, 2005]. The results suggested that the information content based metric introduced by [Lin, 1998] showed better performance than a dictionary-based metric by [Gurevych, 2005b]. Therefore, we integrated the metric by [Lin, 1998] in our information retrieval system. Sometimes, it is called a *universal* semantic similarity metric, as it is supposed to be application-, domain-, and resource independent. However, we should be aware of the fact that semantic similarity takes only synonymy and hyperonymy relations between two concepts into account. Our future work should extend this metric to other types of semantic relations. For computing the information content of concepts, the German newspaper corpus *taz*[3] was used. This corpus covers a wide variety of topics and has about 172 million tokens.

3.4 IR based on Semantic Relatedness

Computing semantic relatedness as described in Section 3.3 allows to quantify the relatedness between two semantic concepts. In order to apply the metric to the task of IR, the relevance of documents to a given query should be computed based on semantic relatedness for the concept pairs. Therefore, we first map all document and query terms except stopwords to concepts in the GermaNet structure receiving two sets of concepts K_d and K_q respectively. As a simple first approach we compute the similarities between a query and a document as the sum of the semantic relatedness values for each pair of query and document terms:

$$sim(d,q) = \sum_{i=1}^{n_d} \sum_{j=1}^{n_q} s(t_{d,i}, t_{q,j}) \quad (1)$$

4 Experimental Work

We conducted several experiments. After each experiment we performed a qualitative analysis of the results in order to understand the strengths and weaknesses of the proposed methods and derive improvements for our method.

Experiment 1 In the first experiment, we used the configuration explained above. Figure 1 depicts *mean average precision*[4] (MAP) for the different runs depending on the query length and the index type (lemmas or lemmas with compounds and decomposed parts) used.

The semantic relatedness (SR) model performs worse than the extended boolean (EB) model in all configurations. SR and EB system work best for short queries, longer queries seem to add noise and the better performance of a combination of *Title* and *Description* over *Description* suggests that some relevant search terms are missing in *Description* or their weighting is changed in the combination of *Title* and *Description*. The combination of lemmas, compounds and compound elements yields the best performance for the EB model, but for the SR model we can observe a decrease when using compound splitting. However follow-up experiments showed the superiority of decomposed compounds also for the SR model. We therefore give only the results for the runs using compound splitting and short queries (*Title*) for the follow-up experiments. Figure 2 shows MAP, the number of retrieved and relevant documents, and the precision after 10 documents have been retrieved (P10) for each experiment and the best EB run.

In order to identify weak points of the semantic relatedness method and to improve it, we examined the results of single topics and searched for possible errors in the relevance judgement of the system. The following shows an example for one topic.

Topic No. 131 (Title): *Zweisprachige Erziehung* (bilingual education) For this topic, the SR model performs better than the EB model. It ranks many relevant documents higher and can even retrieve some documents not found by Lucene. The documents which are not found by

[2] http://snowball.tartarus.org

[3] www.taz.de

[4] Mean average precision is the mean of the average precision for each query.

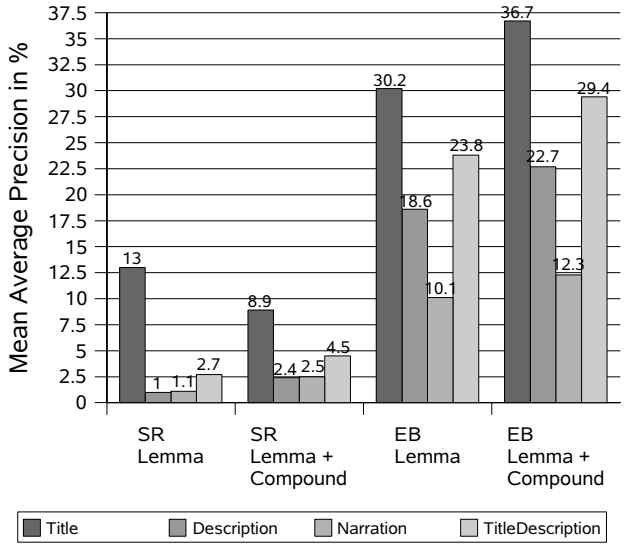

Figure 1: Mean Average Precision for experiment 1 (SR=semantic relatedness model, EB=extended boolean model).

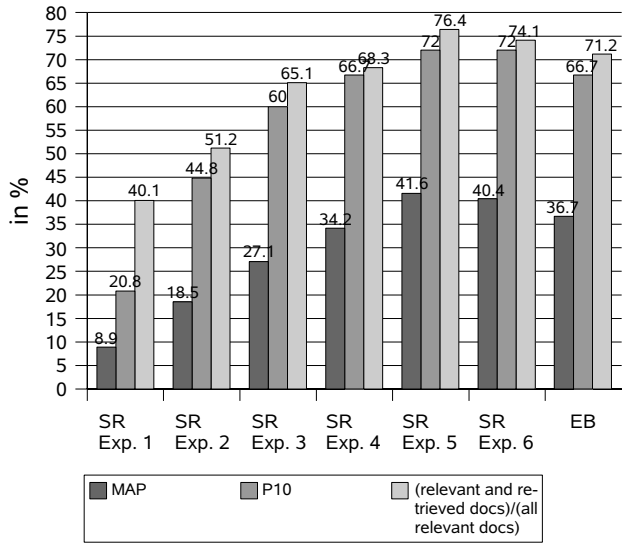

Figure 2: Results for different experiments using lemmas, compound splitting, and short queries (*Title*).

Lucene contain several different terms as a substitute for the query term *Zweisprachige*, which yield a high semantic relatedness score, e.g.:

- *Mehrsprachigkeit* (multilingualism) 0.98
- *sprachlich* (linguistic) 0.83
- *Vielsprachigkeit* (multilingualism) 0.86
- *bilingual* (bilingual) 1.0

Table 2 shows examples of relevant documents for this topic. The EB system retrieves the first two documents as they contain the query term *Erziehung* (education) several times, though they are both ranked low. The third document is not found by the EB system as it contains neither exactly *Zweisprachige* (bilingual) nor *Erziehung*. In this case, only the SR system is able to retrieve the relevant document by using lexical-semantic knowledge. One drawback of our system we observed was that many documents which relate only to one query term, e.g. *Erziehung*, but not to both query terms are ranked very high due to a high frequency of the occurring query term. This causes many relevant documents to be ranked much lower or not to be retrieved at all. To address this issue, we introduced a heuristic in a follow-up experiment.

Experiment 2 We extended the semantic relatedness model in the following way: for the documents which do not contain *all* of the query terms, i.e. not all of the query terms contribute a semantic relatedness score of > 0.8, the similarity score is multiplied by the factor $1/(1 + Number\_of\_not\_related\_query\_terms)$. The following shows the modified Equation 1:

$$sim(d, q) = \frac{\sum_{i=1}^{n_d} \sum_{j=1}^{n_q} s(t_{d,i}, t_{q,j})}{1 + n_{nr}} \qquad (2)$$

where n_{nr} is the number of the query terms which are not semantically related to any of the document terms. This heuristic pushes documents which contain the maximum number of query terms as opposed to those which contain a smaller number of query terms occurring many times.

For the SR system this heuristic yields a precision increase of 9.6 percent points and the number of relevant documents retrieved is increased by 297. This effect can also be seen in Table 2, where the first two documents are ranked much higher after applying this heuristic.

Experiment 3 Our analysis also suggested that the threshold of 0.8 for semantic relatedness might be set too low. We found several pairs of document and query terms with high scores which add noise to the retrieval system, e.g.:

- *Abfallwirtschaft* (waste industry) – *Nutzung* (use) 0.83
- *Werbung* (advertisement) – *Vorschlag* (suggestion) 0.81
- *Politik* (politics) – *Vorgehensweise* (approach) 0.89

We therefore experimented with different values for this threshold and found the optimal value to be 0.98, so that only highly related terms are taken into account for computing the relevance of documents. Increasing the threshold to 0.98 yields a performance improvement of 8.6 percent points and retrieves 372 relevant documents more than in the last experiment. As 0.98 is a very high threshold, we need to perform further analysis on threshold settings.

Experiment 4 In order to motivate this experiment, we will first discuss another example in our data.

Topic No. 147 (Title): *Fußball und Gesellschaft* (soccer and society) Many documents are ranked high which contain the exact query term *Gesellschaft* and the highly *Fußball*-related term Sport (0.84). However, these documents were neither annotated as relevant nor as irrelevant, and judging by title and abstract of some of the documents it can not be concluded if soccer is addressed in

Document ID and Title	*Relevant Terms*		*Relevance Judgement*	*Rank*						
	EB	*SR*		*EB*	*SR*					
					E1	*E2*	*E3*	*E4*	*E5*	*E6*
Ideologie und Realität : interkulturelle Erziehung auf Irrwegen GIRT-DE19980101311	*Erziehung* (education)	*Erziehung*, *Mehrsprachigkeit* (multilingualism), *bilingual*	relevant	117	**50**	**9**	5	5	20	16
Kurzinformation über Modellprojekte : schulische Betreuung der Kinder von Einwanderern und Interkulturelle Erziehung GIRT-DE19910106855	*Erziehung*	*Erziehung*, *Mehrsprachigkeit*, *zweisprachig* (bilingual)	relevant	450	**81**	**13**	8	8	21	23
Sozialpsychologische Grundlagen des schulischen Zweitspracherwerbs bei MigrantenschlerInnen... GIRT-DE19970112872	–	*zweisprachig*, *Mehrsprachigkeit*, *sprachlich* (linguistic)	relevant	—	**21**	**4**	9	9	3	7

Table 2: Topic No. 131 (Title): *Zweisprachige Erziehung* (bilingual education)

these documents or not. Despite that these documents seem to be more relevant than some documents highly ranked by Lucene containing only the query term *Gesellschaft*, they have a disturbing influence on the retrieval performance. Relevant documents which contain both query terms *Gesellschaft* und *Fußball*, but with a smaller frequency, receive a lower score and rank. In order to boost these documents to a higher rank, we 'punish' documents which do not contain the exact string representation of all the query terms. Similar to experiment 2 we therefore multiply the similarity score by the factor $1/(1 + Number\_of\_not\_string\_matched\_query\_terms)$.

$$sim(d,q) = \frac{\sum_{i=1}^{n_d}\sum_{j=1}^{n_q} s(t_{d,i}, t_{q,j})}{(1+n_{nsm})\cdot(1+n_{nr})} \quad (3)$$

This gives us a 7.1 percent points improvement in MAP and lets us retrieve 85 more relevant documents.

Experiment 5 Lucene is using the inverse document frequency *idf* which measures the general importance of a term for predicting the content of a document. We tried to integrate idf into our system and experimented with different measures and approaches and found that the following modification of Equation 3 brought the best improvement:

$$sim(d,q) = \frac{\sum_{i=1}^{n_d}\sum_{j=1}^{n_q} idf(t_{q,j}) \cdot s(t_{d,i}, t_{q,j})}{(1+n_{nsm})\cdot(1+n_{nr})} \quad (4)$$

with $idf(t) = 1/f_t$ where f_t is the number of documents in the collection containing term t. We yield an improvement in MAP of 7.4 percent points and retrieve 219 more relevant documents. With this result we outperform the EB system by 4.9 percent points and retrieve 141 relevant documents more than the best EB run.

Experiment 6 We assumed that by combining the SR approach with the EB model we would be able to improve the retrieval performance. Several methods for combining the similarity scores of different IR systems have been evaluated in the past. We adopted a very simple method which just calculates the sum of the scores of both systems:

$$sim(d,q) = sim_{EB}(d,q) + sim_{SR}(d,q) \quad (5)$$

In the evaluation of [Lee, 1997] this method performed not significantly worse than the best approach. As the semantic relatedness scores are not normalised, we normalise the scores by the minimum and maximum score for each query before applying Equation 5:

$$sim_{SR,norm}(d,q) = \frac{sim_{SR}(d,q) - sim_{SR,min}(q)}{sim_{SR,max}(q) - sim_{SR,min}(q)} \quad (6)$$

For combining the SR system of experiment 5 with the best run of the EB system we found no performance increase, but a slight decrease of MAP compared to experiment 5.

Test Set We used the topics of CLEF'2004 as test set and repeated experiment 5, experiment 6 and the best EB run. Table 3 shows the results and Table 4 shows a comparison of some runs of the SR and EB system on average precision (AP) and P10, using a paired T-Test. Despite the success on our training set we yield an insignificant performance increase of only 1.1 percent MAP for experiment 5 compared with the best EB run. In our future work, we plan to study the impact of semantic relatedness in IR, on multiple datasets to see, under which experimental conditions semantic relatedness is most appropriate.

Run	*MAP*	*P10*	*#Rel.+retr. docs*
SR Experiment 5	34.4	56.0	1074
SR Experiment 6	33.1	57.2	1044
EB	33.3	56.0	1088

Table 3: Results for the test set CLEF'2004.

Paired T-Test(p)		*AP*	*P10*
CLEF'2005	(SR Exp.5,EB)	**0.046**	0.103
	(SR Exp.6,EB)	**0.0040**	0.062
CLEF'2004	(SR Exp.5,EB)	0.755	1.0
	(SR Exp.6,EB)	0.937	0.798

Table 4: Paired T-Test (two-tailed distribution) between Exp.5/Exp.6 and baseline; statistically significant results are highlighted.

5 Related Work

There have been several attempts in the past to integrate lexical-semantic knowledge in IR systems. Table 5 gives an overview.

[Leveling, 2005] has used *Multilayered Extended Semantic Network* (MultiNet) representations of queries and documents in the CLEF domain-specific track for several years with mixed results.

[Smeaton, 1999] reports about several experiments on using WordNet in IR. A large-scale experiment yields a low retrieval performance due to malicious word sense disambiguation and unanalyzed proper nouns, but a small-scale follow-up experiment shows a significant improvement.

[Gurevych, 2005a] uses the German BERUFENet corpus, a collection of descriptions of 5800 professions in Germany [Bundesagentur für Arbeit, 2006], and investigates

Paper	*Queries*	*Documents*	*Method*	*Result*	*Explanation*
[Leveling, 2005]	CLEF2003/ CLEF2004 CLEF2005	GIRT3/ GIRT4/ GIRT4	query expansion/ indexing/ query construction	small improvement/ low performance/ inconclusive	knowledge not sufficient/ spelling and grammatical errors and sentence-based matching
[Smeaton, 1999]	Trec-3	portion of Trec-3, category B, Wall Street Journal	semantic similarity using WordNet	low performance	WSD errors, unanalysed proper nouns
	self-built user queries	captions for 4000 images	semantic similarity	encouraging performance	small scale, manual WSD
[Gurevych, 2005a]	essays about job preferences	BERUFENet elect. job counceling	query expansion/ semantic relatedness	performance increase/ no improvement	only for hyponomy/ no advanced pre-processing GermaNet coverage insufficient
[Arampatzis *et al.*, 2000]	—	—	semantic similarity	—	theoretical
[Fliedner, 2005]	—	Süddeutsche Zeitung, 1700 sentences	semantic similarity	encouraging	no extensive evaluation
[Sanderson, 1994]	subject code of documents	Reuters text categorisation collection	WSD influence on IR	insensitive to ambiguity, very sensitive to WSD errors	—
[Gonzalo *et al.*, 1998]	summaries of documents	derived from SEMCOR	WSD influence on IR	sensitive to ambiguity sensitive to WSD errors	—
[Gonzalo *et al.*, 1998]	summaries of documents	derived from SEMCOR	indexing WordNet synsets	high performance improvement	manual WSD
[Lytinen *et al.*, 2000]	153 test questions	600 frequently asked question files	semantic similarity	good performance	no exclusive evaluation of semantic similarity

Table 5: Summary of related work.

the use of query expansion and semantic relatedness using GermaNet as the underlying knowledge base. Query expansion yields a slightly increased performance. Incorrect analysis resulting from using stemming when mapping words to GermaNet entries and a missing word sense disambiguation (WSD) component are the main reasons for that. The retrieval model using semantic relatedness shows no significant performance gain over the baseline model. However, the system does not use any advanced preprocessing components, such as compound splitting and detection of negative preference statements referring to professions. It is stated that the coverage of the special terminology in GermaNet is still insufficient to be used as a knowledge resource in specialised domains.

A general linguistically motivated retrieval system is proposed by [Arampatzis *et al.*, 2000]. Among others, the model includes semantic expansion of queries and incorporates a semantic similarity measure into the retrieval function which performs a *fuzzy matching* of query and document terms. Unfortunately, no empirical evaluation of the model is reported.

[Fliedner, 2005] develops a question answering system, which incorporates linguistic knowledge from different resources, such as GermaNet and a German FrameNet currently under development in the SALSA project [Burchardt *et al.*, 2006]. The integration of the lexical-semantic knowledge is based on a *Generalised Similarity Measure*. However, no extensive evaluation of the question answering system is reported.

[Sanderson, 1994] takes a closer look at the relationship between word sense disambiguation and information retrieval. He introduces ambiguity into documents by using pseudo-words. The results show that: i.) word sense ambiguity is only a problem for very short queries; ii.) word sense disambiguation with an accuracy of less than 90% has a negative effect on the retrieval performance.[5]

[Gonzalo *et al.*, 1999] on the other hand show with their experiments that word sense disambiguation can be beneficial to IR, even with an accuracy of less than 90%. Additionally, indexing with WordNet synsets is examined by [Gonzalo *et al.*, 1998]. Information retrieval results improve on a manually disambiguated corpus, but also with a disambiguation accuracy of less than 90% an improvement is still observed.

[Lytinen *et al.*, 2000] show that word sense disambiguation of even around 60% accuracy can be helpful in IR. They use a WordNet-based semantic similarity metric for relevance ranking in a question answering system. The similarity metric is combined with a metric based on the *Vector Space Model*. Unfortunately, the impact of the similarity metric on the retrieval performance is not evaluated separately.

Summarising related work, we can see that there is no clear proof for the usefulness of lexical-semantic knowledge in information retrieval. One of the reasons for this is an insufficient coverage of terms by the knowledge bases. They often contain either general vocabulary and thus cannot be effectively applied in specific domains (the case of GermaNet), or model narrow domains and cannot be applied on a broad scale (hand-crafted ontologies). However, if the domain-specific vocabulary is modelled in a knowledge resource and the information retrieval is limited to this particular domain, successful results can be found. A way to overcome the insufficient coverage is by combining several knowledge resources in one system. Unrobust analysis and processing methods also have a negative influence on the performance of IR systems. Finally, word sense disambiguation seems to play an important role when incorporating the lexical-semantic knowledge in IR. Even if word sense disambiguation is not perfect, it seems to be possible to employ information retrieval methods which require word sense disambiguation and achieve positive results.

6 Conclusions and Future Work

In this paper, we explored the potential of lexical-semantic knowledge in IR by using semantic relatedness. Our experiments on the GIRT corpus show that semantic relatedness has the potential to outperform a traditional bag-of-words approach as implemented by Lucene.

Comparing the best run of our system (using *Title*) to IR systems which took part in the domain-specific German monolingual track of CLEF2005 (using *Title* and *Description*), our system would be ranked in the middle-field on the third rank.

The experiments presented in this paper were conducted

[5] The state-of-the art word sense disambiguation systems typically display between 65% and 70% accuracy rates, which is far below 90%.

on the GIRT corpus. In our future work, we would like to study how the models perform for different information retrieval scenarios. The Semantic Information Retrieval (SIR) Project investigates the application of NLP and IR techniques in the domain of electronic career guidance. We collected natural language essays about career preferences of school leavers. Based on these natural language essays, queries for the information retrieval system are generated which use the BERUFENet corpus. Pilot information retrieval experiments with the system based on semantic relatedness have been described by [Gurevych, 2005a].

Another interesting domain for the application of our information retrieval system is eLearning. Educational presentation slides usually contain phrases and keywords rather than complete sentences and feature a complex structure and layout, e.g. figures, tables or diagrams. These experiments can provide useful insights about the applicability of semantically enhanced information retrieval across different domains and different types of information retrieval scenarios.

Acknowledgements

We are grateful to the *Bundesagentur für Arbeit* for providing the BERUFENet corpus. We thank Torsten Zesch for his helpful comments and contributions, and we also thank Niels Ott for providing the compound splitter. This work is carried out as part of the "Semantic Information Retrieval" (SIR) project funded by the German Research Foundation.

References

[Airio, 2006] Eija Airio. Word normalization and decompounding in mono- and bilingual IR. *Information Retrieval*, 9(3):249 – 271, June 2006.

[Arampatzis *et al.*, 2000] Avi Arampatzis, Th.P. van der Weide, P. van Bommel, and C.H.A. Koster. Linguistically motivated information retrieval. *Encyclopedia of Library and Information Science*, 69, 2000.

[Baker *et al.*, 1998] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of ACL*, pages 86–90, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[Bundesagentur für Arbeit, 2006] Bundesagentur für Arbeit. BERUFENet. http://infobub.arbeitsagentur.de/berufe/index.jsp, July 2006.

[Burchardt *et al.*, 2006] A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Pado, and M. Pinkal. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of LREC 2006*, Genoa, Italy, 2006.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[Fliedner, 2005] Gerhard Fliedner. A generalised similarity measure for question answering. In *Proceedings of NLDB 2005*, volume 3513 of *LNCS*, pages 380–383, Alicante, 2005.

[Gonzalo *et al.*, 1998] Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP*, pages 38–44, Montreal, Canada, 1998.

[Gonzalo *et al.*, 1999] J. Gonzalo, A. Pefias, and F. Verdejo. Lexical ambiguity and information retrieval revisited. In *Proceedings of EMNLP/VLC*, 1999.

[Gospodnetic and Hatcher, 2005] Otis Gospodnetic and Erik Hatcher. *Lucene in Action*. Manning Publications Co., 2005.

[Gurevych and Niederlich, 2005] Iryna Gurevych and Hendrik Niederlich. Computing semantic relatedness in german with revised information content metrics. In *Proceedings of "OntoLex 2005 - Ontologies and Lexical Resources" IJCNLP'05 Workshop*, Jeju Island, Republic of Korea, 2005.

[Gurevych, 2005a] Iryna Gurevych. Anwendungen des semantischen Wissens über Konzepte im Information Retrieval. In *Proceedings of Knowledge eXtended: Die Kooperation von Wissenschaftlern, Bibliothekaren und IT-Spezialisten*, Jülich, Germany, 2. - 4. November 2005.

[Gurevych, 2005b] Iryna Gurevych. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of IJCNLP'05*, Jeju Island, Republic of Korea, 2005.

[Hollink *et al.*, 2004] Vera Hollink, Jaap Kamps, Christof Monz, and Maarten de Rijke. Monolingual document retrieval for european languages. *Information Retrieval*, 7(1 - 2):33 – 52, Jan 2004.

[Kluck, 2004] Michael Kluck. The girt data in the evaluation of clir systems from 1997 until 2003. In C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003 Trondheim, Norway, Revised Selected Papers*, volume 3237 of *Lecture Notes in Computer Science*, pages 379–393. Springer, Trondheim, Norway, 2004.

[Kunze, 2004] Claudia Kunze. *Computerlinguistik und Sprachtechnologie. Eine Einführung*, chapter Lexikalisch-semantische Wortnetze, pages 423–431. Spektrum Akademsicher Verlag, second edition, 2004.

[Langer, 1998] Stefan Langer. Zur Morphologie und Semantik von Nominalkomposita. In *Proceedings of KONVENS*, page 8397, 1998.

[Lee, 1997] John Ho Lee. Analyses of multiple evidence combination. In *Proceedings of ACM-SIGIR*, pages 267–276, 1997.

[Leveling, 2005] Johannes Leveling. University of hagen at clef 2005: Towards a better baseline for nlp methods in domain-specific information retrieval. In *Results of CLEF 2005, Working Notes for the CLEF 2005 Workshop*, Wien, Österreich, 2005.

[Lin, 1998] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.

[Lytinen *et al.*, 2000] S. Lytinen, N. Tomuro, and T. Repede. The use of wordnet sense tagging in faqfinder. In *Proceedings of the AAAI-2000 workshop on AI and Web Search*, Austin, TX, July 2000.

[Salton *et al.*, 1983] G. Salton, E. Fox, and H. Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.

[Sanderson, 1994] Mark Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of ACM-SIGIR*, pages 49–57, Dublin, IE, 1994.

[Schmid, 1994] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, 1994.

[Smeaton, 1999] A.F. Smeaton. *Natural Language Information Retrieval*, chapter Using NLP or NLP Resources for Information Retrieval Tasks, pages 99–111. Kluwer Academic Publishers, 1999.