International Journal of Semantic Computing
Vol. 2, No. 2 (2008) 253–272

World Scientific
www.worldscientific.com

CLOSING THE VOCABULARY GAP FOR COMPUTING TEXT SIMILARITY AND INFORMATION RETRIEVAL*

CHRISTOF MÜLLER

Ubiquitous Knowledge Processing Lab, Computer Science Department
Technische Universität Darmstadt
Hochschulstr. 10, 64289 Darmstadt, Germany
mueller@tk.informatik.tu-darmstadt.de
http://www.ukp.tu-darmstadt.de

IRYNA GUREVYCH

Ubiquitous Knowledge Processing Lab, Computer Science Department
Technische Universität Darmstadt
gurevych@tk.informatik.tu-darmstadt.de
http://www.ukp.tu-darmstadt.de

MAX MÜHLHÄUSER

Telecooperation, Computer Science Department
Technische Universität Darmstadt
max@tk.informatik.tu-darmstadt.de
http://www.tk.informatik.tu-darmstadt.de

Received 9 November 2007
Revised 16 January 2008
Accepted 19 February 2008

This paper studies the integration of lexical semantic knowledge in two related semantic computing tasks: ad-hoc information retrieval and computing text similarity. For this purpose, we compare the performance of two algorithms: (i) using semantic relatedness, and (ii) using a conventional extended Boolean model [13] with additional query expansion. For the evaluation, we use two different test collections in the German language especially suitable to study the *vocabulary gap* problem: (i) GIRT [5] for the information retrieval task, and (ii) a collection of descriptions of professions built to evaluate a system for electronic career guidance in the information retrieval and text similarity tasks. We found that integrating lexical semantic knowledge increases the performance for both tasks. On the GIRT corpus, the performance is improved only for short queries. The performance on the collection of professional descriptions is improved, but crucially depends on the accurate preprocessing of the natural language essays employed as topics.

Keywords: Information Retrieval; Text Similarity; Semantic Relatedness.

*Extended version of "Integrating Semantic Knowledge into Text Similarity and Information Retrieval", published in *Proceedings of the First IEEE International Conference on Semantic Computing* (ICSC), 2007.

1. Introduction

Semantic computing deals with utilizing semantic technologies to connect the intentions of users with computational content. Thereby, the intentions of users and the computational content are often formulated by means of natural language. Establishing the connection between the users' information needs, called topics[a] and the relevant information in the documents of the collection to be queried is the task known as information retrieval (IR).

This task has been around for a long time, but has made relatively little use of semantic information so far. Several works investigated the integration of lexical semantic knowledge in IR. Voorhees [19] uses WordNet for expanding queries from TREC collections. Even by using manually selected terms, the performance could only be improved on short queries. Mandala *et al.* [9] showed that by combining a WordNet based thesaurus with a co-occurrence and a predicate-argument-based thesaurus and by using expansion term weighting, the retrieval performance on several data collections can be improved. Smeaton [17] reports about several experiments on using WordNet in IR. A large-scale experiment in which WordNet is used for computing a measure of query-document similarity yields a low retrieval performance due to malicious word sense disambiguation and unanalyzed proper nouns. A follow-up experiment uses a collection of image captions and yields a significant improvement over the baseline. The application of word-based semantic similarity for measuring text similarity on a paraphrase data set has been shown to yield a significant performance improvement in [10].

Managing vast amounts of information has now become crucial as we are faced with a rapidly expanding universe of personal information. Significant advances in the key technologies, such as computing text similarity and IR, are needed to manage, organize and find copious textual information. Additionally, web 2.0 leads to the proliferation of user generated content. This makes a lot of previously personal knowledge widely available to a large community of users. Forums, blogs, wikis, or books etc.[b] are widely spread in electronic form and build a wide body of knowledge to be accessed by IR techniques.

User generated content displays some special features as opposed to conventional web content. On the one hand, it is rather poorly linked. Therefore, it cannot be searched effectively by conventional IR algorithms, such as PageRank [12], which define the relevance of information by analysing the hyperlink structure. On the other hand, user generated content lacks editorial control. Therefore, the variability of the vocabulary employed by the users to describe the same things is extremely high. This challenge is commonly known as the *vocabulary gap* in the IR literature.

[a] A topic is a natural language statement of the user's information need, which is used to create a query for an IR system.

[b] http://www.nabble.com, http://www.blogger.com, http://www.wikipedia.org, http://www.gutenberg.org

In order to provide efficient access to information in poorly linked environments, while the variability of the vocabulary employed by users is high, information retrieval algorithms should integrate semantic information. For this purpose, we need: (i) large scale knowledge bases to deliver the knowledge necessary to fill the gap, and (ii) sophisticated algorithms to integrate the knowledge into the process of information search. In this paper, we present a set of experiments aimed at semantic IR and evaluate the proposed models utilizing the knowledge from the German wordnet, GermaNet [6], on two German IR benchmarks described in Sec. 2. These benchmarks are especially suitable to study the *vocabulary gap* problem as they display a great variability in the vocabularies employed by user topics and the underlying document collection. We study the performance of IR models across two different tasks: (i) IR on the GIRT and BERUFEnet[c] based corpora, and (ii) text similarity on the BERUFEnet based corpus. The semantic IR model is compared with the conventional extended Boolean (EB) model as implemented by Lucene [2].[d] We also report on runs of the EB model with query expansion using (i) synonyms and (ii) hyponyms, extracted from GermaNet.

The remainder of this paper is structured as follows: In Sec. 2, we will describe the two test collections, the respective topics and gold standards. This is followed by a description of the employed algorithms in Sec. 3. The experiments and the analysis of results are described in Sec. 4. Finally, we draw our conclusions in Sec. 5.

2. Data

2.1. *GIRT Benchmark*

GIRT is employed in the domain-specific track at the Cross Language Evaluation Forum (CLEF).[e]

Document Collection. The corpus consists of 151,319 documents containing abstracts of scientific papers in social science, together with the author and title information and several keywords. Table 1 shows descriptive statistics about the corpus.

Table 1. Descriptive statistics of test collections (after preprocessing).

	#doc	#token	#unique token	#token/doc (mean)
GIRT	151,319	13,961,046	540,721	92.26
BERUFEnet	529	222,912	34,346	421.38

[c]http://berufenet.arbeitsamt.de/
[d]We also ran experiments with Okapi BM25 model as implemented in the Terrier framework, but the results were worse than those by EB model. Therefore, we limit our discussion to the latter.
[e]http://www.clef-campaign.org

Example document

```
<DOC>
<DOCNO>GIRT-DE19909106</DOCNO>
<DOCID>GIRT-DE19909106</DOCID>
<TITLE-DE>Politiker einer ethnischen Gruppe im Kongreß:
Deutsch-amerikanische Fallstudien zur Interaktion von Ethnizität,
Nationalität und demokratischer Regierung, 1865-1930</TITLE-DE>
<AUTHOR>Adams, Willi Paul</AUTHOR>
<PUBLICATION-YEAR>1990</PUBLICATION-YEAR>
<LANGUAGE-CODE>DE</LANGUAGE-CODE>
<CONTROLLED-TERM-DE>ethnische Gruppe</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Nordamerika</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>USA</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>politischer Einfluß</CONTROLLED-TERM-DE>
<METHOD-TERM-DE>beschreibend</METHOD-TERM-DE>
<METHOD-TERM-DE>historisch</METHOD-TERM-DE>
<CLASSIFICATION-TEXT-DE>Sozialgeschichte</CLASSIFICATION-TEXT-DE>
<ABSTRACT-DE>Klärung des Einflusses einer ethnischen Gruppe und ihrer
gewählten Vertreter auf den demokratischen Proze und auf das
Selbstverständnis der amerikanischen Gesellschaft als multiethnischer
Nationalstaat.</ABSTRACT-DE>
</DOC>
```

Topics. The experiments described in Sec. 4 use the topics and relevance assessments of CLEF 2004 and 2005. Each topic consists of three different fields: a title (keywords), a description (a sentence), and a narration (exact specification of relevant information). Table 2 shows descriptive statistics about the topics.

Example topic

```
<top>
<num>137 </num>
<DE-title> Ehre und Gesellschaft </DE-title>
<DE-desc> Finde Dokumente, die Ehre als soziale Handlungskategorie des
gesellschaftlichen Wertesystems diskutieren. </DE-desc>
<DE-narr> Relevante Dokumente berichten darüber, welche Rolle der Ehre als
Wertkategorie gesellschaftlicher Systeme des 20. und 21. Jahrhunderts
zukommt. Relevante sind Dokumente, die Ehre als Motiv fr ein bestimmtes
Handeln oder Denken oder für bestimmte Einstellungen oder Haltungen
identifizieren. </DE-narr>
</top>
```

Table 2. Descriptive statistics of topics (after preprocessing).

	#doc	#token	#unique token	#token/doc (mean)
CLEF 2005 Topics				
Title	25	44	43	1.76
Description	25	173	97	6.64
Narration	25	484	263	19.36
CLEF 2004 Topics				
Title	25	47	46	1.88
Description	25	181	105	7.24
Narration	25	483	287	19.32
Professional Profiles	30	1,140	715	38.00

Gold Standard. A portion of GIRT documents is annotated with relevance judgments for each topic by using the *pooling method* [18].

2.2. *BERUFEnet data*

The second benchmark employed in our experiments was built based on a real-life task based scenario in the domain of electronic career guidance.[f] Electronic career guidance is a supplement to career guidance by human experts, helping young people to decide which profession to choose. The goal is to automatically compute a ranked list of professions according to the user's interests. A current system employed by the German Federal Labour Office (GFLO) in their automatic career guidance front-end[g] is based on vocational trainings, manually annotated with a tagset of 41 keywords. The user selects appropriate keywords according to her interests. In reply, the system consults a knowledge base with professions manually annotated with the keywords by domain experts. Thereafter, it outputs a list of the best matching professions to the user. This approach has two significant disadvantages. Firstly, the knowledge base has to be maintained and steadily updated, as the number of professions and keywords associated with them is continuously changing. Secondly, the user has to describe her interests in a very restricted way. By applying IR methods to the task of electronic career guidance, we try to remove the disadvantages by letting the user describe her interests in natural language, i.e. by writing a short essay. An important observation about essays and descriptions of professions is a mismatch between the vocabularies of topics and documents and the lack of contextual information, as the documents are fairly short. Typically, people seeking career advice use different words for describing their professional preferences as those employed in the professionally prepared descriptions of professions. Therefore, lexical semantic knowledge and *soft matching*, i.e. matching not

[f] A detailed description of electronic career guidance including the employment of SR measures based on Wikipedia can be found in [3].
[g] http://www.interesse-beruf.de

only exact terms, must be especially beneficial to such a system, where semantically close words should be related. For example, a person may be writing about *cakes*, while the description of the profession contains the words *pastries* and *confectioner*. Also, the topics are longer than those typically employed in IR tasks. Considering the expected output and length of topics, we define the task of electronic career guidance not as classical ad-hoc IR, but as computing text similarity.

Document collection. The document collection is extracted from BERUFEnet, a database created by the GFLO. It contains textual descriptions of about 1,800 vocational trainings, e.g. *Elderly care nurse*, and 4,000 descriptions of professions, e.g. *Biomedical Engineering*. We restrict the collection to a subset of BERUFEnet documents, consisting of 529 descriptions of vocational trainings, due to the process necessary to obtain a gold standard, as described below. The documents contain not only details of professions, but also a lot of information concerning the training, and administrative issues. In present experiments, we only use those portions of the descriptions, which characterize the profession itself, e.g. typical objects (*computer*, *plant*), activities (*programming*, *drawing*), or working places (*office*, *fabric*). Table 1 shows descriptive statistics about the corpus.

Topics. We collected real natural language topics by asking 30 human subjects to write an essay about their professional interests. Table 2 shows descriptive statistics about the topics. Below is an example topic translated to English.

Example essay translated to English

I would like to work with animals, to treat and look after them, but I cannot stand the sight of blood and take too much pity on them. On the other hand, I like to work on the computer, can program in C, Python and VB and so I could consider software development as an appropriate profession. I cannot imagine working in a kindergarden, as a social worker or as a teacher, as I am not very good at asserting myself.

[German original]

Ich würde gerne mit Tieren arbeiten, sie behandeln, für sie sorgen, aber ich kann kein Blut sehen und ich habe zu viel Mitleid mit kranken Tieren. Andererseits arbeite ich besonders gerne am Computer, kann programmieren in C, Python und VB und könnte mir daher auch in der Software-Entwicklung einen passenden Beruf vorstellen. Ich kann mir nur schlecht vorstellen in einem Kindergarten, als Sozialberater oder als Lehrer zu arbeiten, da ich mich nicht besonders gut durchsetzen kann.

Gold Standard. Creating a gold standard to evaluate the electronic career guidance system requires domain expertise, as the descriptions of professions have to be ranked according to their relevance to the topic. Therefore, we apply an automatic method as shown in Fig. 1, which uses the knowledge base employed by the GFLO, described in Sec. 1. To obtain the gold standard, we first annotate each essay with

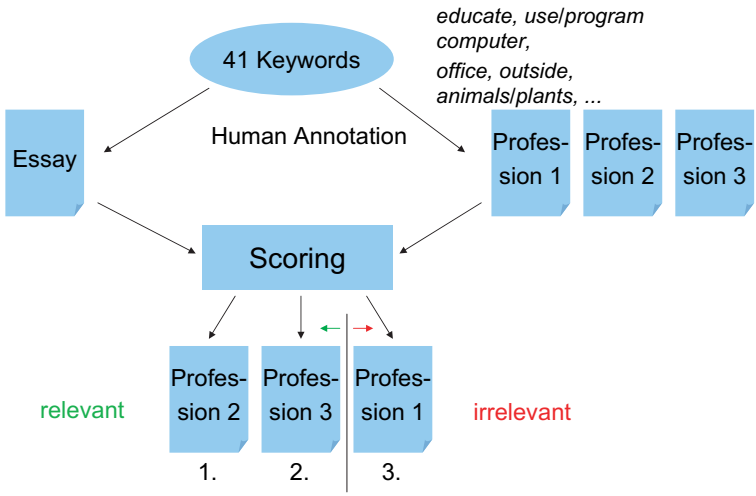

Fig. 1. Automatic creation of relevance judgements.

relevant keywords from the tagset of 41 and retrieve a ranked list of professions, which were assigned one or more keywords by domain experts. Examples of such keywords are shown below.

Example annotation translated to English
programming, writing, laboratory, workshop, electronics, technical installations
[German original]
programmieren, schreiben, Labor, Werkstatt, Elektronik, technische Anlagen

A ranked list retrieved for the above annotation is shown in Table 3. To obtain relevance judgments for the IR task, we map the ranked list to a set of relevant and irrelevant professions by setting a threshold of 3 keyword matches between profile and job description annotations, above which job descriptions will be judged relevant to a given profile. This threshold was suggested by domain experts. Using the threshold yields on average 93 relevant documents per topic.

The quality of the automatically created gold standard depends on the quality of the applied knowledge base. As the knowledge base was created by domain experts and is at the core of the electronic career guidance system of the GFLO, we assume that the quality is adequate to ensure a reliable evaluation.

3. Models

3.1. *Preprocessing*

For creating the search index for IR models, we apply first tokenization and then remove stopwords. For the GIRT data, we use a general German stopword list,

Table 3. Example of the knowledge-based ranking.

Rank	Profession	Score
1	Elektrotechnische/r Assistent/in	4
2	Energieelektroniker/in, Anlagentechnik	4
3	Energieelektroniker/in, Betriebstechnik	4
4	Industrieelektroniker/in, Produktionstechnik	4
5	Prozessleitelektroniker/in	4
6	Beamt(er/in) — Wetterdienst (mittl. Dienst)	3
7	Chemikant/in	3
8	Elektroanlagenmonteur/in	3
9	Fachkraft für Lagerwirtschaft	3
10	Film- und Videolaborant/in	3
11	Fotolaborant/in	3
12	Informationselektroniker/in	3
13	Ingenieurassistent/in, Maschinenbautechnik	3
14	IT-System-Elektroniker/in	3
15	Kommunikationselektroniker/in, Informationstechnik	3
16	Mechatroniker/in	3
17	Mikrotechnologe/-technologin	3
18	Pharmakant/in	3
19	Schilder- und Lichtreklamehersteller/in	3
20	Technische/r Assistent/in für Konstruktions- und Fertigungstechnik	3

while for the BERUFEnet data, the list is extended with highly frequent domain specific terms. Before adding the remaining words to the index, they are lemmatized employing the TreeTagger [14]. There have already been a number of studies about the usefulness of morphological normalisation in IR. Some of the most recent ones are [4] and [1]. They confirm the positive impact which morphological normalisation has, especially for German. However, they find almost no difference in performance between stemming and lemmatisation. We finally split compounds into their constituents [7], and add both, constituents and compounds, to the index.

3.2. *Extended boolean model*

Lucene[h] is an open source text search library based on an EB model. After matching the preprocessed queries against the index, the document collection is divided into a set of relevant and irrelevant documents. The set of relevant documents is, then, ranked according to the formula given in the following equation:

$$r_{EB}(d,q) = \sum_{i=1}^{n_q} tf(t_q,d) \cdot idf(t_q) \cdot lengthNorm(d)$$

where n_q is the number of terms in the query, $tf(t_q,d)$ is the term frequency factor for term t_q in document d, $idf(t_q)$ is the inverse document frequency of the term,

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

and $lengthNorm(d)$ is a normalization value of document d, given the number of terms within the document. We added a simple query expansion algorithm using (i) synonyms and (ii) hyponyms, extracted from GermaNet.

3.3. *Semantic relatedness model*

SR is defined as *any* kind of lexical semantic or functional association that exists between two words. There exist several different methods, which calculate a numerical score that gives a measure for the SR between a word pair. The required lexical semantic knowledge can be derived from a range of resources like computer-readable dictionaries, thesauri, or corpora.

For integrating semantic knowledge into IR and text similarity, we follow the approach proposed in [11]. The algorithm is based on Lin's information-content based SR metric described in [8]. Thereby, we use the German wordnet GermaNet as a knowledge base. The structure of GermaNet is very similar to that of WordNet, but shows differences in some of the design principles. Discrepancies between GermaNet and WordNet are e.g. that GermaNet employs additionally artificial, i.e. non-lexicalized concepts, and adjectives are structured hierarchically as opposed to WordNet. Currently, GermaNet includes about 40000 synsets with more than 60000 word senses modeling nouns, verbs and adjectives.

Lin's metric incorporates not only the knowledge of the wordnet, but also some corpus-based evidence. In particular, it integrates the notion of information content as defined in [15]. Information content of concepts in a semantic network is defined as the negative logarithm of the likelihood of concept c:

$$ic(c) = -\log p(c).$$

We compute the likelihood of concept c from a corpus, in which we count the number of occurrences n_c of the concept. Given the number N of all tokens in the corpus, the likelihood is computed as:

$$p(c) = \frac{n_c}{N}.$$

Therefore, a more sparsely occurring concept has a higher information content than a more often occurring one. For computing the information content of concepts, the German newspaper corpus *taz*[i] was used. This corpus covers a wide variety of topics and has about 172 million tokens. Defining LCS_{c_1,c_2} as the lowest common subsumer of the two concepts c_1 and c_2 which is the first common ancestor in the GermaNet taxonomy, Lin's metric can be defined as:

$$s(c_1, c_2) = \frac{2 \cdot \log p(LCS_{c_1,c_2})}{\log p(c_1) + \log p(c_2)}. \quad (1)$$

We compute the similarities between a query and a document as a function of the sum of semantic relatedness values for each pair of query and document terms

[i] http://www.taz.de

using Eq. (1). Scores above a predefined threshold are summed up and weighted by different factors, which boost or lower the scores for documents, depending on how many query terms are contained exactly or contribute a high enough SR score. Several heuristics described in [11] were introduced to improve the performance of this scoring approach. In order to integrate the strengths of traditional IR models, the inverse document frequency *idf* is considered, which measures the general importance of a term for predicting the content of a document. The final formula of the model is as follows:

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

where n_d is the number of tokens in the document, n_q the number of tokens in the query, $t_{d,i}$ the ith document token, $t_{q,j}$ the jth query token, $s(t_{d,i}, t_{q,j})$ the SR score for the respective document and query term, n_{nsm} the number of query terms not exactly contained in the document, n_{nr} the number of query tokens which do not contribute a SR score above the threshold. We use two different types of *idf*:

$$idf(t) = \frac{1}{f_t} \tag{2}$$

where f_t is the number of documents in the collection containing term t, and *idf* calculated by Lucene

$$idf = log(\frac{n_{docs}}{f_t + 1}) + 1 \tag{3}$$

taking into account the number of documents in the collection n_{docs}.

We extend the work reported in [11] by considering the influence, which variable document length inside the document collection can have on the retrieval performance. We experimented with different document length and query length normalization schemes for SR values and the heuristics.

4. Analysis of Results

We report the results with the two best performing thresholds (.85 and .98) for the scores employed in the final computation by the SR model.

4.1. *Information retrieval*

The evaluation metrics used for the IR task are *mean average precision*[j] (MAP), and *the number of relevant returned documents*.

4.1.1. *GIRT*

We used two types of topics: titles and descriptions. In Table 4, we summarize the results. Recall-Precision curves are depicted in Fig. 2.

[j]After each relevant document is retrieved, the precision is calculated. These values are averaged for each query. The average over all queries is the mean average precision.

Table 4. IR performance on the GIRT collection. The best performance on a given benchmark is shown in bold.

Corpus	EB			SR		
	MAP	#Rel.Ret.	Type	MAP	#Rel.Ret.	Thresh.
CLEF 2004	**0.34**	**1100**	**EB**	0.33	1076	0.85
Title	0.34	1077	EB+SYN	**0.37**	**1156**	**0.98**
	0.34	1089	EB+HYPO			
CLEF 2004	**0.22**	**976**	**EB**	0.16	864	0.85
Description	0.19	866	EB+SYN	**0.22**	**980**	**0.98**
	0.09	631	EB+HYPO			
CLEF 2005	**0.39**	**1996**	**EB**	0.37	1988	0.85
Title	0.38	1963	EB+SYN	**0.43**	**2130**	**0.98**
	0.37	1928	EB+HYPO			
CLEF 2005	**0.23**	**1614**	**EB**	0.17	1413	0.85
Description	0.19	1421	EB+SYN	**0.20**	**1631**	**0.98**
	0.13	1137	EB+HYPO			

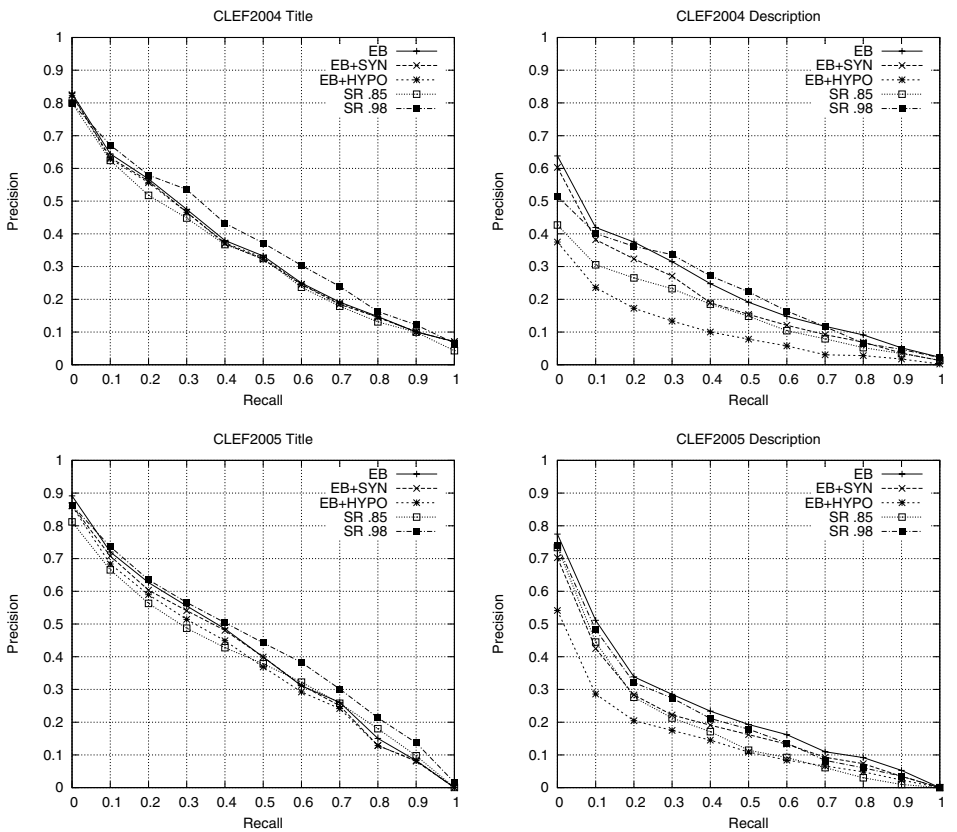

Fig. 2. Recall-Precision curves for the GIRT corpus.

EB Model using Query Expansion. The use of query expansion in the EB model yields no performance increase. For short queries the performance is at best the same as for the pure EB model. For longer queries the performance decreases. The results are similar to the ones found in [19]. Query expansion using synonyms yields better results than by using hyponyms.

EB Model vs. SR Model. The SR model outperforms the EB model on most topic types. Only for the CLEF 2005 topics using the description part, the performance of the EB model is better. We observe that the SR model performs better on the topics represented by titles than descriptions. This suggests that semantic information is especially useful for short queries, lacking contextual information as compared to longer queries.

We also analysed precision and recall on the query-level. Figs. 3 and 4 show average precision and the number of relevant retrieved documents for each topic using the title field.

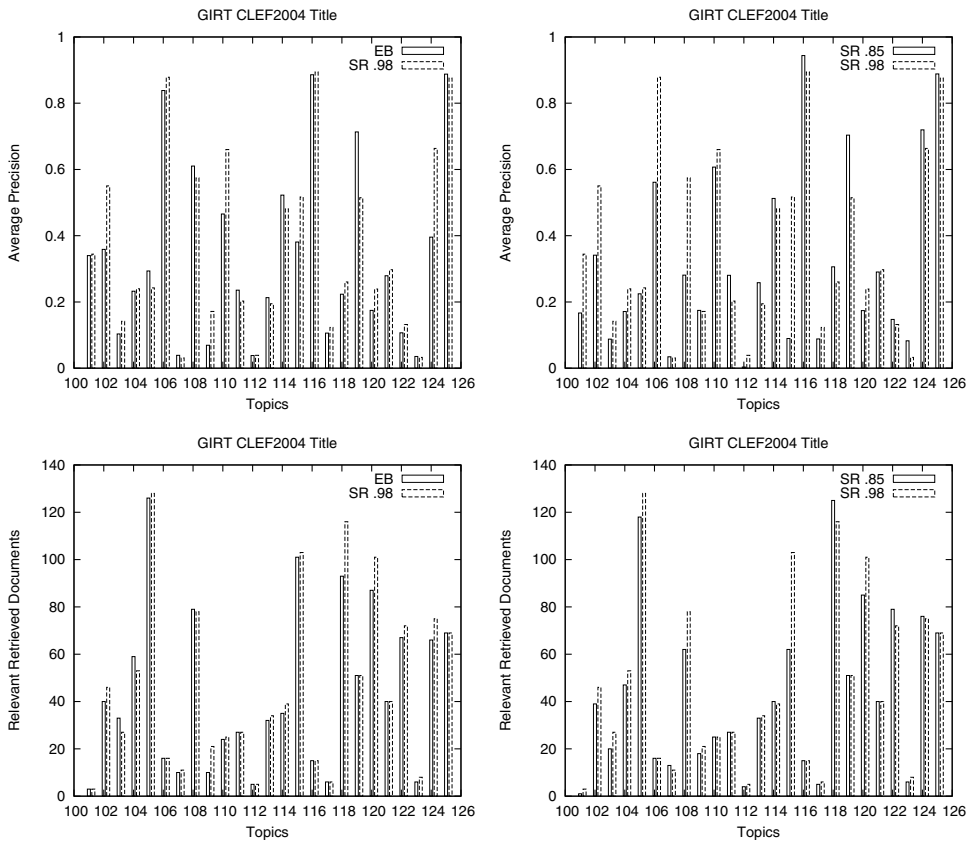

Fig. 3. Precision and recall for CLEF 2004 topics on the GIRT corpus.

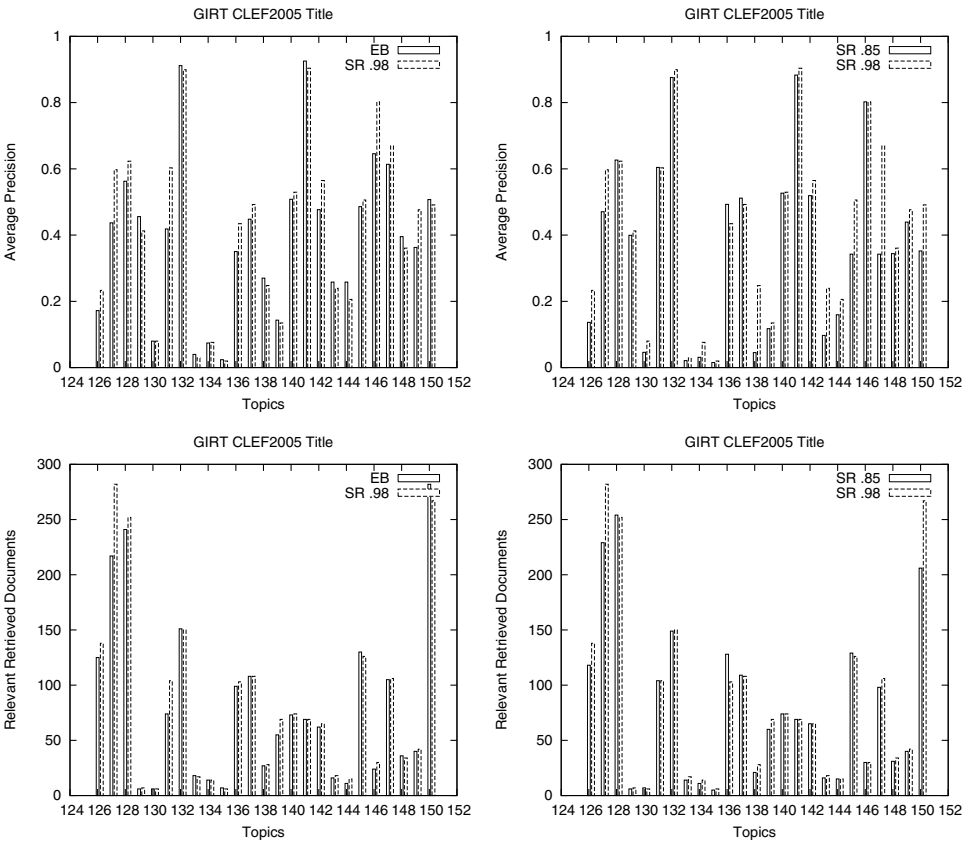

Fig. 4. Precision and recall for CLEF 2005 topics on the GIRT corpus.

The title field of topic no. 131 contains the two keywords *zweisprachige Erziehung* (bilingual education). For this topic, the SR model performs much better than the EB model both in terms of precision and recall. In the documents which are not found by the EB model the query term *zweisprachige* (bilingual) is often substituted by different terms with a similar meaning. These terms have a high semantic relatedness score when compared to the original query term, e.g.:

- *bilingual* (bilingual) 1.0
- *mehrsprachig* (multilingual) 0.983
- *Mehrsprachigkeit* (multilingualism) 0.983

Thus, documents missing the original query term are either not found at all by the EB model or get ranked with a lower score than by the SR model.

Also morphologically related terms like *Wahl* (vote) and *Wähler* (voter) are matched by the SR model which can improve average precision and recall as in the case of topic no. 127 for which the title field consists of the term *Wählerverhalten*

(electoral behavior). Several relevant documents contain the term *Wahl*, but not the decomposed query term *Wähler* (voter), and can therefore be found by the SR model.

However, the use of SR is not beneficial in all cases. For topic no. 144 with the title field *Radio und Internet* (radio and internet), the EB model shows a higher average precision compared to the SR model. The SR model matches *Radio* with the related terms *Funk* (radio) and *Rundfunk* (broadcasting). This results in highly ranking the documents containing a high number of terms related to the query term *Radio*, but not necessarily containing a high number of occurrences of the second query term *Internet*. Many of these documents are irrelevant and do not deal with listening radio over the internet as is the intention of the topic stated in the description and narration field.

Threshold Settings. The threshold .98 performs systematically better for all kinds of topics. This indicates that the information about strong SR is especially valuable to IR. The threshold .85 seems to introduce too much noise in the process, when word pairs are not strongly related, e.g.:

- *Politik* (politics) — *Vorgehensweise* (approach) 0.89
- *Lernen* (study) — *Behälter* (container) 0.88
- *Sportwettkampf* (sports competition) — *Konflikt* (conflict) 0.88
- *Unternehmen* (company) — *Start* (start) 0.91

As can be seen from Fig. 4, for topic no. 138 using the title field, which only contains the keyword *Unternehmensinsolvenz* (insolvent companies), the higher threshold setting of 0.98 yields a much higher average precision than with the threshold of 0.85. This is due to the high semantic relatedness of 0.94 between the query term *Insolvenz* (insolvency) and the term *Armut* (poverty). As a consequence numerous documents are retrieved that deal with poverty, but not insolvent companies. Excluding this relatedness by setting a threshold of 0.98 not only increases the average precision, but also increases the number of relevant documents which are retrieved, as generally only the first one thousand retrieved documents are taken into account in the evaluation.

Other Settings. Our results on the GIRT data are generally better than those reported in [11]. We believe this is due to a different stop word list and the normalization schemes, which we used in the present paper.

The influence of the application of different document length and query length normalization schemes for SR values and the heuristics and the selection of the *idf* type depends on the data set. For the GIRT data, the use of Eq. (2) for *idf* computation yields better results and the application of length normalization decreases performance.

Table 5. IR performance on the BERUFEnet collection. The best performance on a given benchmark is shown in bold.

Corpus	EB			SR		
	MAP	#Rel.Ret.	Type	MAP	#Rel.Ret.	Thresh.
BERUFEnet	**0.39**	**2581**	**EB**	**0.41**	**2787**	**0.85**
N,V,Adj	0.37	2589	EB+SYN	0.41	2753	0.98
	0.34	2702	EB+HYPO			
BERUFEnet	0.38	2297	EB	0.40	2770	0.85
N	0.38	2310	EB+SYN	**0.42**	**2677**	**0.98**
	0.38	**2328**	**EB+HYPO**			
BERUFEnet	0.54	2755	EB	**0.59**	**2787**	**0.85**
Keywords	**0.54**	**2768**	**EB+SYN**	0.58	2783	0.98
	0.47	2782	EB+HYPO			

BERUFEnet. We built queries from natural language essays by (i) extracting nouns, verbs, and adjectives, (ii) using only nouns, and (iii) suitable keywords from the tagset of 41 assigned to each topic. The last type was introduced in order to simulate a well performing information extraction system, which extracts professional features from the topics. This enables us to estimate the possible performance increase a better preprocessing could yield. The results are shown in Table 5 and Fig. 5.

EB Model using Query Expansion. The results for using query expansion in the EB model are consistent with the results in Sec. 4.1.1. The use of query expansion in the EB model yields no performance increase.

EB vs. SR. Comparing the number of relevant retrieved documents, we observe that the IR model based on SR is able to return more relevant documents, especially remarkable on the BERUFEnet data. This supports our hypothesis that semantic knowledge is especially helpful for the *vocabulary mismatch problem*, which cannot be addressed by conventional IR models.

In our analysis of the BERUFEnet results, we noticed that many errorneous results were due to the topics, which are free natural language essays. Some subjects deviated from the given task to describe their professional interests and described the facts that are rather irrelevant to the task of electronic career guidance, e.g. *It is important to speak different language in the growing European Union.* If all content words are extracted to build a query, a lot of noise is introduced.

Therefore, we experimented with two further system configurations: building the query using only nouns, and using manually assigned keywords based on the tagset of 41 keywords. Results obtained in these system configurations show that the performance is better for nouns, and significantly better for the queries built of keywords. This suggests that in order to achieve a high performance in the given application scenario, it is necessary to preprocess the topics by performing

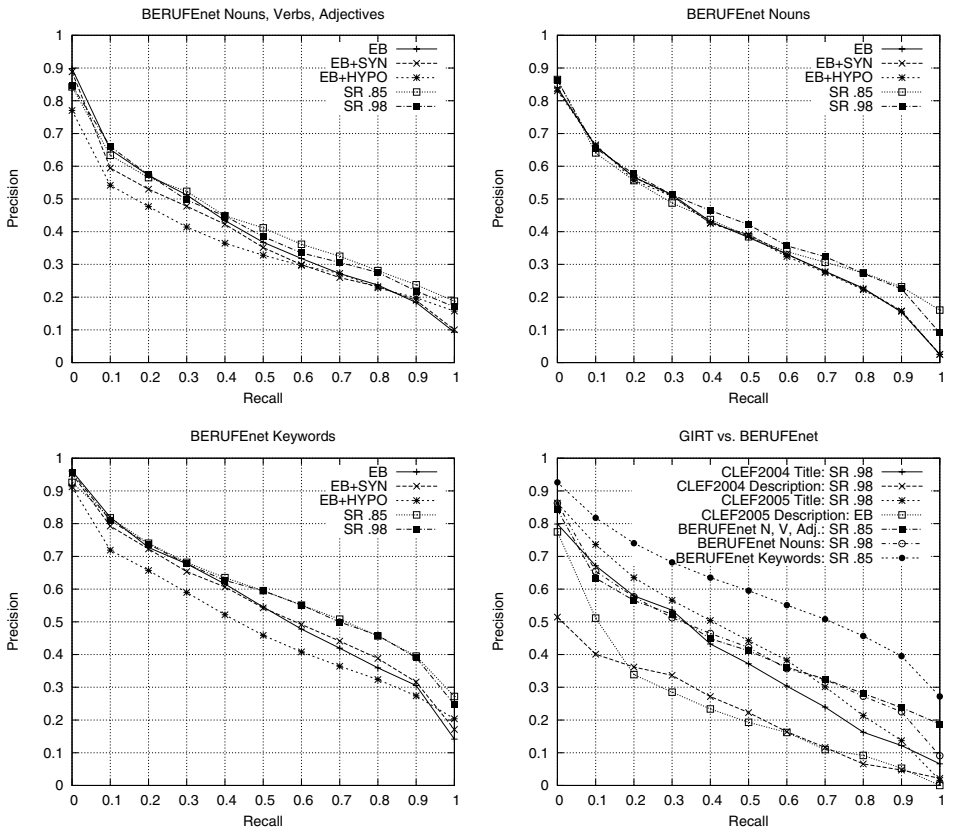

Fig. 5. Recall-Precision curves for the BERUFEnet corpus and summary for both corpora.

information extraction. In this process, natural language essays should be mapped to a set of features relevant for describing a person's interests. Our results suggest that SR model performs significantly better in this setting.

As can be seen from Fig. 6, the SR model is able to improve precision for most of the topics on all query types.

Threshold Setting. The value of the threshold seems to have less influence on the retrieval performance for this data set. This might be also due to the employment of a domain specific stopword list. If it is not applied, the results are significantly worse.

Other Settings. The influence of document length normalization and *idf* is different on this benchmark compared to GIRT: Eq. (3) for *idf* computation yields a better performance and applying the document length normalization increases the performance. Inconsistent impact on performance might be caused by differences in the document length, query length, and the type of documents in the benchmarks.

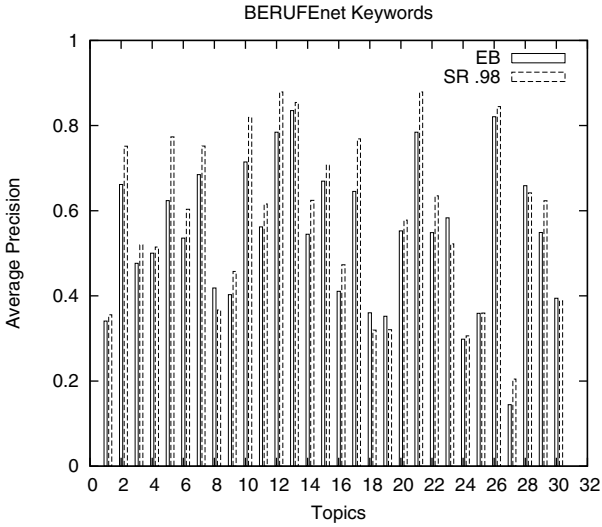

Fig. 6. Precision for using keywords as query on the BERUFEnet corpus.

4.1.2. *Overall results*

The lower right diagram in Fig. 5 depicts the Recall-Precision curves of the best system configurations for all benchmarks. It shows that the employment of SR is especially beneficial for short queries.

4.2. *Text similarity*

In this task, we measured the similarity between the descriptions of professions in the BERUFEnet corpus with the natural language essays by (i) extracting nouns, verbs, and adjectives, (ii) using only nouns, and (iii) suitable keywords from the tagset of 41 assigned for each topic, as done in the IR task. The gold standard consists of not merely relevance judgments dividing the set of documents into relevant and irrelevant documents, as in IR, but is a list of possible professions ranked according to their relevance score to a given profile (see Sec. 2.2). To evaluate the performance of the text similarity algorithm we, therefore, use a rank correlation measure, i.e. Spearman's rank correlation coefficient [16]. For each query, we calculated the correlation coefficient. By using Fisher's z transformation, we compute the average over all queries, yielding one coefficient expressing the correlation between the rankings of the gold standard and the text similarity system. Table 6 and Fig. 7 show the results of the text similarity task.

EB using Query Expansion. The query expansion can only improve the performance of the EB model for the keyword-based approach using synonyms of the query terms for expansion, but cannot reach the performance of the SR model.

Table 6. Text Similarity performance on the BERUFEnet dataset.

Corpus	EB	EB+QE		SR	
	|Rank Corr.|	|Rank Corr.|	Type	|Rank Corr.|	Thresh.
BERUFEnet N,V,Adj	0.306	0.288	SYN	0.338	0.85
		0.275	HYPO	0.326	0.98
BERUFEnet N	0.335	0.331	SYN	0.320	0.85
		0.327	HYPO	0.341	0.98
BERUFEnet Keywords	0.497	0.530	SYN	0.580	0.85
		0.399	HYPO	0.563	0.98

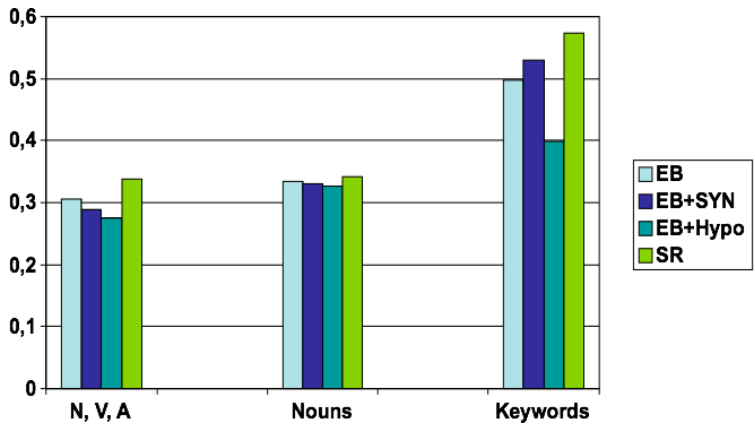

Fig. 7. Text Similarity performance on the BERUFEnet dataset.

Comparison to Information Retrieval. The performance of the text similarity ranking shows similar trends as the IR performance on the same data collection. The SR model outperforms the EB model for all query types. The preprocessing of topics has also a great influence on the performance in this task. Especially for the third query type where assigned keywords are used, the SR model shows a significant improvement of the rank correlation.

Though our results cannot directly be compared to the ones of Mihalcea *et al.* [10], the interpretation of the results is similar: the use of SR improves the conventional bag-of-words models.

5. Conclusions

In this paper, we investigated the integration of semantic information into IR algorithms. Semantic information is especially valuable for IR on user generated content where the variability of the vocabulary employed by users is high. We compared the performance of an EB model and a model based on SR for two tasks: ad-hoc IR and text similarity. For the IR task we used the standard IR benchmark GIRT and a test

collection that is employed in a system for electronic career guidance determining relevant professions, given a natural language essay about a person's interests. The collection was extracted from the BERUFEnet corpus. The latter collection was also employed in the text similarity task. We found that both IR models display similar performance across the different corpora and tasks. However, the SR model is almost consistently stronger, especially for shorter queries. A fairly high threshold of SR scores .98 showed the best results, which indicates that the information about strong SR is especially valuable to IR.

In the experiments with the BERUFEnet data and electronic career guidance, we found that preprocessing the topics is essential in this application scenario. Simple query building techniques used in IR introduce too much noise. Therefore, better analysis and more accurate information extraction are required in the preprocessing.

Mandala *et al.* analyzed the methods of query expansion applied in [19] and other works. Some reasons identified as a cause for missing performance improvement in these works are:

- insufficient or missing weighting methods for expansion terms;
- missing word sense disambiguation;
- missing relations, especially cross part of speech relations;
- insufficient lexical coverage of thesauri.

Mandala *et al.* addressed these points and could improve IR performance as described in Sec. 1. The use of a SR measure in our work can be seen as an implicit way of query expansion. The SR measure is used for weighting expansion terms. In order to further increase the performance of our model, we also need to address non-classical types of semantic relations and increase the coverage of the applied knowledge base. First attempts in this direction can be found in [3], where we proposed an algorithm for computing SR using Wikipedia as a background knowledge source and using this in IR. The results show a significant performance improvement for the Wikipedia-based IR model. Thus, the rapidly growing amount of user generated content on the World Wide Web poses not only a challenge to IR, but can also help to improve the effectiveness of IR by providing valuable semantic information.

Acknowledgements

This work was supported by the German Research Foundation under the grant "Semantic Information Retrieval from Texts in the Example Domain Electronic Career Guidance", GU 798/1-2. We are grateful to the *Bundesagentur für Arbeit* for providing the BERUFEnet corpus.

References

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

[2] O. Gospodnetic and E. Hatcher, *Lucene in Action*, Manning Publications Co., 2005.
[3] I. Gurevych, C. Müller and T. Zesch, What to be? — electronic career guidance based on semantic relatedness, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 1032–1039.
[4] V. Hollink, J. Kamps, C. Monz and M. de Rijke, Monolingual document retrieval for european languages, *Information Retrieval* **7**(12) (2004) 33–52.
[5] M. Kluck, The girt data in the evaluation of CLIR systems from 1997 until 2003, in *Comparative Evaluation of Multilingual Information Access Systems*, LNCS 3237, Springer, 2004.
[6] C. Kunze, *Computerlinguistik und Sprachtechnologie*: *Eine Einführung*, in Lexikalisch-semantische Wortnetze, eds. K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klablunde and H. Langer, Akademischer Verlag, Heidelberg, Berlin, 2004.
[7] S. Langer, Zur Morphologie und Semantik von Nominalkomposita, in *Proceedings of Konferenz zur Verarbeitung natürlicher Sprache*, Bonn, Germany, October 1998, pp. 83–97.
[8] D. Lin, An information-theoretic definition of similarity, in *Proceedings of the International Conference on Machine Learning*, Madison, USA 1998, pp. 296–304.
[9] R. Mandala, T. Tokunaga and H. Tanaka, The use of WordNet in information retrieval, in *Proceeding of the COLING-ACL Workshop on Usage of Word Net in Natural Language Processing*, S. Harabagiu (ed.), Somerset, New Jersey, USA, 1998, pp. 31–37.
[10] R. Mihalcea, C. Corley and C. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity, in *Proceedings of the 21st National Conference on Artificial Intelligence* (AAAI–2006), Boston, July 2006.
[11] C. Müller and I. Gurevych, Exploring the Potential of Semantic Relatedness in Information Retrieval, in *LWA 2006 Lernen — Wissensentdeckung — Adaptivität, 9.-11.10.2006 in Hildesheim*, M. Schaaf and K.-D. Althoff (eds.), Hildesheimer Informatikberichte, pp. 126–131, Hildesheim, Germany, 2006. GI-Fachgruppe Information Retrieval, Universität Hildesheim.
[12] L. Page, S. Brin, R. Motwani and T. Winograd, The pagerank citation ranking: Bringing order to the web, Technical report, Stanford Digital Library Technologies Project, 1998.
[13] G. Salton, E. Fox and H. Wu, Extended Boolean Information Retrieval, *Communications of the ACM* **26**(11) (1983) 1022–1036.
[14] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in *Proceedings of the Conference on New Methods in Language Processing*, 1994.
[15] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal* **27** (1948) 379–423; 623–656.
[16] S. Siegel and N. J. Jr. Castellan, *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, 1988.
[17] A. F. Smeaton. *Natural Language Information Retrieval*, chapter Using NLP or NLP Resources for Information Retrieval Tasks, ed. T. Strzalkowski, Kluwer Academic Publishers, 1999, pp. 99–111.
[18] E. M. Voorhees and D. K. Harman, Overview of the 6th text retrieval conference (TREC-6). In *Proceedings of the Sixth Text REtrieval Conference*, Gaithsburg, MD, USA, 1997, pp. 1–24.
[19] Ellen M. Voorhees, Query expansion using lexical-semantic relations, in *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1994, pp. 61–69.