Integrating Semantic Knowledge into Text Similarity and Information Retrieval

Christof Müller, Iryna Gurevych Max Mühlhäuser
Ubiquitous Knowledge Processing Lab Telecooperation
Darmstadt University of Technology
Hochschulstr. 10, 64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de
{mueller,gurevych,max}@tk.informatik.tu-darmstadt.de

Abstract

This paper studies the influence of lexical semantic knowledge upon two related tasks: ad-hoc information retrieval and text similarity. For this purpose, we compare the performance of two algorithms: (i) using semantic relatedness, and (ii) using a conventional extended Boolean model [12]. For the evaluation, we use two different test collections in the German language: (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 task. We found that integrating lexical semantic knowledge improves 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 preprocessing of natural language essays employed as topics.

1 Introduction

An often occurring problem in information retrieval (IR) is the gap between the vocabulary used in formulating the topics,[1] and the vocabulary used in writing the documents of the collection to be queried. An example for this problem is the domain of electronic career guidance.[2] 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[3] 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 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.

In [11], an overview is presented of how lexical semantic knowledge can be integrated into IR. The authors describe an algorithm utilizing a measure of semantic relatedness (SR) in IR operating on the German wordnet GermaNet [6]. The algorithm is evaluated on the GIRT corpus, a standard

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

[2] A detailed description of electronic career guidance including the employment of SR measures based on Wikipedia can be found in [4].

[3] `http://www.interesse-beruf.de`

benchmark provided by the CLEF conference.[4] Employing topics and relevance judgments from CLEF'2004 and CLEF'2005, significant increases in IR performance could only be found for the semantic model on CLEF'2005 data.

While evaluations on standard benchmarks enable a generalizable comparison of results across different IR systems, various studies reported that the performance of IR critically depends on the type of queries submitted to such a system [9, 1]. This implies that the results obtained on such a benchmark cannot be generalized to cover a great variety of IR application scenarios [2], but should always be related to the properties of the corpus underlying the evaluation. For this reason, we extend the previous work in this paper by studying the performance of IR models across two different tasks: (i) IR on the GIRT and BERUFEnet[5] 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 [3].[6] We also report on runs of the EB model with query expansion using (i) synonyms, and (ii) hyponyms, extracted from GermaNet.

Several works investigated the integration of lexical semantic knowledge in IR. In [16] Voorhees is using 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. showed in [8] 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. 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].

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

2 Data

2.1 GIRT benchmark

GIRT is employed in the German domain-specific task at CLEF.

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.

	#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

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

	#doc	*#token*	*#unique token*	*#token/doc (mean)*
CLEF2005 Topics				
Title	25	44	43	1.76
Description	25	173	97	6.64
Narration	25	484	263	19.36
CLEF2004 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

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

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

Gold Standard A portion of GIRT documents is annotated with relevance judgments for each topic by using the *pooling method*[17].

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, as described in Section 1.

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 (*of-*

[4] http://www.clef-campaign.org

[5] http://berufenet.arbeitsamt.de/

[6] 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.

fice, *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. The topics contain, on average, 130 words. Table 2 shows descriptive statistics about the topics.

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.

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 for the topic. Therefore, we apply an automatic method, which uses the knowledge base employed by the GFLO, described in Section 1. To obtain the gold standard, we first annotate each essay with relevant keywords from the tagset of 41 and retrieve a ranked list of professions, which were assigned one or more keywords by domain experts.

Example annotation translated to English
programming, writing, laboratory, workshop, electronics, technical installations

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.

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

Table 3. Example of the knowledge-based ranking.

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, 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 [13]. We finally split compounds into their constituents, and add both, constituents and compounds, to the index.[7]

3.2 Extended Boolean Model

Lucene[8] 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:

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

[7] http://www.drni.de/niels/cl/BananaSplit/
[8] http://lucene.apache.org

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, and $lengthNorm(d)$ is a normalization value of document d, given the number of terms within the document.

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 semantic relatedness 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 [7]. 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 [14]. 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*[9] 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:

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

[9] `http://www.taz.de`

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 using Equation 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 i-th document token, $t_{q,j}$ the j-th 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*:

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

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

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

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 final computation by the SR model.

4.1 IR

The evaluation metrics used for the IR task are *mean average precision* [10] (MAP), and *the number of relevant returned documents*.

[10] 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.

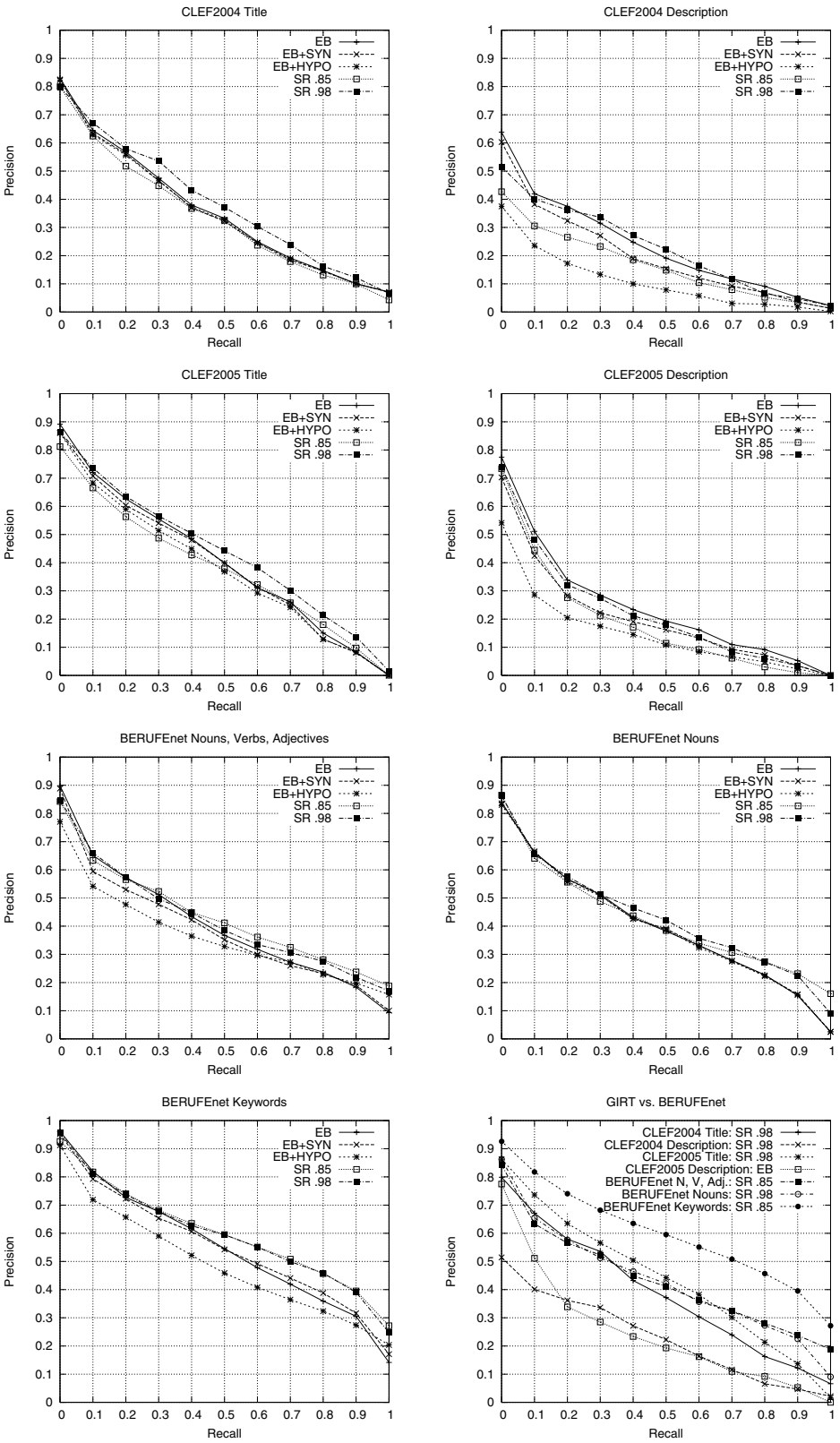

Figure 1. Recall-Precision curves for the IR task.

Corpus	EB		EB+QE			SR		
	MAP	*#Rel.Ret.*	*MAP*	*#Rel.Ret.*	*Type*	*MAP*	*#Rel.Ret.*	*Thresh.*
CLEF2004 Title	0.34	1100	0.34	1077	SYN	0.33	1076	0.85
			0.34	1089	HYPO	0.37	1156	0.98
CLEF2004 Description	0.22	976	0.19	866	SYN	0.16	864	0.85
			0.09	631	HYPO	0.22	980	0.98
CLEF2005 Title	0.39	1996	0.38	1963	SYN	0.37	1988	0.85
			0.37	1928	HYPO	0.43	2130	0.98
CLEF2005 Description	0.23	1614	0.19	1421	SYN	0.17	1413	0.85
			0.13	1137	HYPO	0.20	1631	0.98

Table 4. IR performance on the GIRT collection.

GIRT We used two types of topics: titles and descriptions. In Table 4, we summarize the results. Recall-Precision curves are depicted in Figure 1.

The SR model outperforms the EB model on most topic types. Only for the CLEF2005 topics using the description part, the performance of the EB model is better.

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 [16]. Query expansion using synonyms yields better results than by using hyponyms.

We observe that 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.

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.

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 Equation 2 for *idf* computation yields better results and the application of length normalization decreases performance.

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 Figure 1.

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.

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 errornous 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 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.

The influence of document length normalization and *idf* is different on this benchmark compared to the GIRT: Equation 3 for *idf* computation yields a better performance and applying the document length normalization increases the

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

Table 5. IR performance on the BERUFEnet collection.

performance. Inconsistent impacts on performance might be caused by differences in document length, query length, and the type of documents in the benchmarks.

The lower right diagram in Figure 1 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 not merely of relevance judgments dividing the set of documents into relevant and irrelevant documents, as in IR, but is a list of possible professions ranked by their relevance score to a given profile (see Section 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 [15]. 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 text similarity system. Table 6 shows the results of the text similarity task.

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.

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 to the performance of the SR model.

Though our results cannot directly be compared to the ones of Mihalcea et al. in [10], the interpretation of the results is similar: the use of semantic relatedness improves the conventional lexical matching.

5 Conclusions

In this paper, 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 [16] 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 relationship types, especially cross part of speech relationships;
- insufficient lexical coverage of thesauri.

Mandala et al. addressed these points and could improve IR performance as described in Section 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 and implicitly performs word sense disambiguation. In order to further increase the performance of

Corpus	*EB*	*EB+QE*		*SR*	
	$\|RankCorr.\|$	$\|RankCorr.\|$	*Type*	$\|RankCorr.\|$	*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

Table 6. Text Similarity performance on the BERUFEnet dataset.

our model, we also need to address other types of semantic relations and increase the coverage of the applied knowledge base. First attempts in this direction can be found in [4], where the authors proposed an algorithm for computing SR using Wikipedia[11] as a background knowledge source and using this in IR.

Acknowledgements

This work was supported by the German Research Foundation under 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] N. J. Belkin, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and C. Cool. Query length in interactive information retrieval. In *Proceedings of SIGIR '03*. ACM Press, 2003.

[2] S. Bhavnani, K. Drabenstott, and D. Radev. Towards a unified framework of IR tasks and strategies. *ASIST*, November 2001.

[3] O. Gospodnetic and E. Hatcher. *Lucene in Action*. Manning Publications Co., 2005.

[4] 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 (ACL'2007)*, page (to appear), Prague, Czech Republic, June 2007.

[5] M. Kluck. The girt data in the evaluation of clir systems from 1997 until 2003. In *Comparative Evaluation of Multilingual Information Access Systems.*, volume 3237 of *Lecture Notes in Computer Science*. Springer, 2004.

[6] C. Kunze. *Computerlinguistik und Sprachtechnologie. Eine Einführung*, chapter Lexikalisch-semantische Wortnetze. Spektrum, 2004.

[7] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, 1998.

[8] R. Mandala, T. Tokunaga, and H. Tanaka. The use of WordNet in information retrieval. In S. Harabagiu, editor, *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, pages 31–37. Association for Computational Linguistics, Somerset, New Jersey, 1998.

[9] T. Mandl and C. Womser-Hacker. Linguistic and statistical analysis of the CLEF topics, 2002.

[10] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the American Association for Artificial Intelligence (AAAI 2006)*, Boston, July 2006.

[11] C. Müller and I. Gurevych. Exploring the Potential of Semantic Relatedness in Information Retrieval. In *Proceedings of LWA 2006 Lernen - Wissensentdeckung - Adaptivität: Information Retrieval*, pages 126–131, Hildesheim, Germany, 2006. GI-Fachgruppe Information Retrieval.

[12] G. Salton, E. Fox, and H. Wu. Extended Boolean Information Retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.

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

[14] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 & 623–656, July & October 1948.

[15] S. Siegel and N. J. Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988.

[16] E. 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*, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[17] E. M. Voorhees and D. K. Harman. Overview of the 6th text retrieval conference (TREC-6). In *Proceedings of the Sixth Text REtrieval Conference*, pages 1–24, Gaithsburg, MD, USA, 1997. NIST Special Publication.

[11] `http://www.wikipedia.org`