A Study on the Semantic Relatedness of Query and Document Terms in Information Retrieval

Christof Müller and **Iryna Gurevych**
Ubiquitous Knowledge Processing (UKP) Lab
Computer Science Department
Technische Universität Darmstadt, Hochschulstraße 10
D-64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de/

Abstract

The use of lexical semantic knowledge in information retrieval has been a field of active study for a long time. Collaborative knowledge bases like Wikipedia and Wiktionary, which have been applied in computational methods only recently, offer new possibilities to enhance information retrieval. In order to find the most beneficial way to employ these resources, we analyze the lexical semantic relations that hold among query and document terms and compare how these relations are represented by a measure for semantic relatedness. We explore the potential of different indicators of document relevance that are based on semantic relatedness and compare the characteristics and performance of the knowledge bases Wikipedia, Wiktionary and WordNet.

1 Introduction

Today we face a rapidly growing number of electronic documents in all areas of life. This demands for more effective and efficient ways of searching these documents for information. Especially user-generated content on the web is a growing source of huge amounts of data that poses special difficulties to IR. The precise wording is often difficult to predict and current information retrieval (IR) systems are mainly based on the assumption that the meaning of a document can be inferred from the occurrence or absence of terms in it. In order to yield a good retrieval performance, i.e., retrieving all relevant documents without retrieving non-relevant documents, the query has to be formulated by the user in an appropriate way. Blair and Maron (1985) showed that with larger growing document collections, it gets impossible for the user to anticipate the terms that occur in all relevant documents, but not in non-relevant ones.

The use of semantic knowledge for improving IR by compensating non-optimal queries has been a field of study for a long time. First experiments on query expansion by Voorhees (1994) using lexical-semantic relations extracted from a linguistic knowledge base (LKB), namely WordNet (Fellbaum, 1998), showed sginificant improvements in performance only for manually selected expansion terms. The combination of WordNet with thesauri built from the underlying document collections by Mandala et al. (1998) improved the performance on several test collections. Mandala et al. (1998) identified missing relations, especially cross part of speech relations and insufficient lexical coverage as reasons for the low performance improvement when using only WordNet.

In recent work, collaborative knowledge bases (CKB) like Wikipedia have been used in IR for judging the document relevance by computing the semantic relatedness (SR) of queries and documents (Gurevych et al., 2007; Egozi et al., 2008; Müller and Gurevych, 2008) and have shown promising results. These resources have a high coverage of general and domain-specific terms. They are employed in several SR measures such as *Explicit Semantic Analysis* (ESA) (Gabrilovich and Markovitch, 2007) that allow the cross part of speech computation of SR and are not restricted to standard lexical semantic relations.

The goal of this paper is to shed light on the role of lexical semantics in IR and the way it can improve the performance of retrieval systems. There exist different kinds of resources for lexical semantic knowledge and different ways to embed this knowledge into IR. Wikipedia and Wiktionary, which have been applied in computational methods only recently, offer new possibilities to enhance IR. They have already shown an excellent performance in computing the SR of word pairs (Strube and Ponzetto, 2006; Gabrilovich and

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1338–1347,
Singapore, 6-7 August 2009.

Markovitch, 2007; Zesch et al., 2008). However, it is not yet clearly understood, what the most beneficial method is to employ SR using these resources in IR. We therefore perform a comparative study on an IR benchmark. We particularly analyze the contribution of SR of query and document terms to this task. To motivate those experiments we first prove that there exists a vocabulary gap in the test collection between queries and documents and show that the gap can be reduced by using lexical semantic knowledge. As the vocabulary coverage of knowledge bases is a crucial factor for being effective in IR, we compare the coverage of Wikipedia, Wiktionary and WordNet. We then analyze the lexical semantic relations that hold among query and document terms and how they are represented by the values of a SR measure. Finally, we explore the potential of different SR-based indicators of document relevance.

The remainder of this paper is structured as follows: In Section 2 we give a short overview of the LKBs and CKBs and the measure of SR we employ in this paper. The test collection we use in our experiments is described in Section 3. In Section 4 we analyze the vocabulary of the test collection and determine the coverage of the knowledge bases. This is followed by the examination of lexical semantic relations and the analysis of the SR of query terms in relevant and non-relevant documents in Section 5.

2 Knowledge Sources and Semantic Relatedness Measure

2.1 Linguistic Knowledge Bases

LKBs are mainly created by trained linguists following clearly defined guidelines. Therefore, their content is typically of high quality. This labor and cost intensive approach, however, yields a number of disadvantages for LKBs:

- their coverage and size are limited;
- they lack domain-specific vocabulary;
- continuous maintenance is often not feasible;
- the content can quickly be out-dated;
- only major languages are typically supported.

The most common types of LKBs are (i) dictionaries, which alphabetically list words and their senses of a certain language along with their definitions and possibly some additional information and (ii) thesauri, which group words with similar meaning together and define further semantic relations between the words, e.g., antonymy. The most widely used LKB is WordNet, which is a combination of dictionary and thesaurus. Since the hypernym and hyponym relations between noun groups form an *is-a* hierarchy, WordNet can also be seen as an ontology. The current version 3.0 of WordNet, which we use in our experiments, contains over 155,000 English words organized into almost 118,000 so called synsets, i.e., groups of synonymous words. WordNet covers mainly general vocabulary terms and its strongest part is the noun hierarchy.

2.2 Collaborative Knowledge Bases

Enabled by the development of Web 2.0 technology and created by communities of volunteers, CKBs have emerged as a new source of lexical semantic knowledge in recent years. In contrast to LKBs, they are created by persons with diverse personal backgrounds and fields of expertise. CKBs have the advantage of being freely available unlike many LKBs. However, the content of CKBs is mainly semi- or unstructured text which initially requires the extraction of explicit knowledge that can then be used in computational methods.

One of the CKBs we use in this paper is Wikipedia, a freely available encyclopedia. It currently contains more than 12 million articles in 265 languages. Besides articles, Wikipedia also offers other forms of knowledge that can be used in computational methods. This includes the hierarchy of article categories (Strube and Ponzetto, 2006; Zesch et al., 2007) and links between articles in the same language (Milne and Witten, 2008) and across languages (Schönhofen et al., 2007; Potthast et al., 2008; Sorg and Cimiano, 2008; Müller and Gurevych, 2008). Due to its encyclopedic character, Wikipedia contains many named entities and domain-specific terms which are not found in WordNet. In our experiments we used the Wikipedia dump of February 6th, 2007.

The second CKB we use is Wiktionary which is a multilingual dictionary and an affiliated project of Wikipedia. It resembles WordNet by containing synonym and hyponym information. It also contains information usually not found in LKBs like abbreviations, compounds, contractions, and the etymology of words. The 171 language-specific

editions of Wiktionary contain more than 5 million entries. Note that each language-specific edition contains not only entries for words of that particular language, but also for words of foreign languages. Wiktionary has been used in IR (Müller and Gurevych, 2008; Bernhard and Gurevych, 2009) and other tasks like sentiment analysis (Chesley et al., 2006) or ontology learning (Weber and Buitelaar, 2006). In our experiments we used the Wiktionary dump of Oct 16, 2007.

2.3 Semantic Relatedness Measure

A wide range of methods for measuring the SR of term pairs are discussed in the literature. In our experiments, we employ ESA as it can be used with all three knowledge bases in our experiments and has shown an excellent performance in related work. ESA was introduced by Gabrilovich and Markovitch (2007) employing Wikipedia as a knowledge base. Zesch et al. (2008) explored its performance using Wiktionary and WordNet as knowledge bases.

The idea of ESA is to express a term's meaning by computing its relation to Wikipedia articles. Each article title in Wikipedia is referred to as a concept and the article's text as the textual representation of this concept. A term is represented as a high dimensional concept vector where each value corresponds to the term's frequency in the respective Wikipedia article. The SR of two terms is then measured by computing the cosine between the respective concept vectors. When applying ESA to Wiktionary and WordNet, each word and synset entry, respectively, is referred to as a distinct concept, and the entry's information[1] is used as the textual representation of the concept.

In our experiments, we apply pruning methods as proposed by Gabrilovich and Markovitch (2007) with the goal of reducing noise and computational costs. Wikipedia concepts are not taken into account where the respective Wikipedia articles have less than 100 words or fewer than 5 in- or outlinks. For all three knowledge bases, concepts are removed from a term's concept vector if their normalized values are below a predefined threshold (empirically set to *0.01*).

[1]For WordNet, the glosses and example sentences of the synsets are used. Wiktionary does not contain glosses for all entries due to instance incompleteness. Therefore, a concatenation of selected information from each entry is used. See Zesch et al. (2008) for details.

Documents	
Number of documents	319115
Number of unique terms	400194
Ave. document length	256.23
Queries	
Number of queries	50
Number of unique terms	117
Ave. query length	2.44

Table 1: Statistics of the test data (after preprocessing).

3 Data

For our study we use parts of the data from the HARD track at the TREC 2003 conference[2]. The document collection consists of newswire text data in English from the year 1999, drawn from the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service.[3] As we did not have access to the other document collections in the track, we restrict our experiments to the newswire text data.

From the 50 available topics of that track, we use only the title field, which consists of a few keywords describing the information need of a user. Table 1 shows some descriptive statistics of the documents and topics. The topics cover general themes like *animal protection*, *Y2K crisis* or *Academy Awards ceremony*. For the preprocessing of topics and documents we use tokenization, stopword removal and lemmatization employing the TreeTagger (Schmid, 1994). In our study, we rely on the relevance assessments performed at TREC to distinguish between relevant and non-relevant documents for each topic.

4 Vocabulary Mismatch

To confirm the intuition that there exists a vocabulary mismatch between queries and relevant documents, we computed the overlap of the terms in queries and relevant documents. The results are shown in the column *String-based* in Table 2. Averaged over all 50 topics, 35.5% of the relevant documents do contain all terms of the query, and 86.5% contain at least one of the query terms. However, this means that 13.5% of the relevant documents do not contain any query term and

[2]`http://trec.nist.gov/`

[3]AQUAINT Corpus, Linguistic Data Consortium (LDC) catalog number LDC2002T31

Measure	*String-based*	*SR-Wikipedia*		*SR-Wiktionary*		*SR-WordNet*	
Threshold		*0.0*	*0.05*	*0.0*	*0.05*	*0.0*	*0.05*
Ave. number of documents where all query terms are matched (in %)	35.5	91.2	72.2	82.6	65.2	74.8	50.8
Ave. number of documents where at least one query term is matched (in %)	86.5	100.0	99.1	97.7	97.2	94.9	92.9
Ave. number of query terms matched per document (in%)	55.8	95.6	84.0	87.0	76.8	79.2	65.7

Table 2: Statistics about the matching of the terms of queries and relevant documents.

cannot be retrieved by simple string-matching retrieval methods. In average, each relevant document matches 55.8% of the query terms. With an average query length of 2.44 (see Table 1), this means that in general, only one of two query terms occurs in the relevant documents which significantly lowers the probability of these documents to have a high ranking in the retrieval result.

In a second experiment, we proved the effectiveness of the SR measure and knowledge bases in reducing the vocabulary gap by counting the number of query terms that match the terms in the relevant documents as string or are semantically related to them. The results are shown in Table 2 for the different knowledge bases in the columns *SR-Wikipedia*, *SR-Wiktionary* and *SR-WordNet*. In order to analyse the performance of the SR measure when excluding very low SR values that might be caused by noise, we additionally applied a threshold of 0.05, i.e. only values above this threshold were taken into account. The SR values range between 0 and 1. However, the majority of SR values lie between 0 and 0.1.

Without threshold, using Wikipedia as knowledge base, in 91.2% of the relevant documents all query terms were matched. For Wiktionary with 82.6% and WordNet with 74.8% the number is lower, but still more than twice as high as for the string-based matching. Wikipedia matches in all relevant documents at least one query term. The average number of query terms matched per document is also increased for all three knowledge bases. Applying a threshold of 0.05, the values decrease, but are still above the ones for string-based matching.

The sufficient coverage of query and document terms is crucial for the effectiveness of knowledge bases in IR. It was found that LKBs do not necessarily provide a sufficient coverage (Mandala et al., 1998). Table 3 shows the amount of terms in queries and documents that are contained in Wikipedia, Wiktionary and WordNet. Wikipedia contains almost all query terms and also shows the best coverage for the document terms, followed by Wiktionary and WordNet. The values for all three knowledge bases are all higher than 80% except for the percentage of queries or documents where all terms are covered and the number of covered unique terms. The low percentage of covered unique document terms for even Wikipedia is mostly due to named entities, misspellings, identification codes and compounds.

	SR-Wikipedia	*SR-Wiktionary*	*SR-WordNet*
Queries			
Percentage of queries where all terms are covered	98.0	78.0	62.0
Percentage of covered terms	99.2	89.3	80.3
Percentage of covered unique terms	99.1	88.9	80.3
Ave. percentage of covered terms per query	99.6	89.2	80.1
Ave. percentage of covered unique terms per query	99.6	89.2	80.1
Documents			
Percentage of documents where all terms are covered	7.9	0.3	0.2
Percentage of covered terms	96.5	88.5	84.3
Percentage of covered unique terms	34.5	12.9	10.0
Ave. Percentage of terms covered per document	97.4	91.8	88.8
Ave. percentage of covered unique terms per document	96.3	88.0	83.6

Table 3: Statistics about the coverage of the knowledge bases.

Judging from the number of covered query and document terms alone, one would expect Wikipedia to yield a better performance when applied in IR than Wiktionary and especially WordNet. The higher coverage of Wikipedia is due to its nature of being an encyclopedia featuring arbitrarily long articles whereas entries in WordNet, and also Wiktionary, have a rather short length following specific guidelines. The high coverage alone is however not the only important factor for the effectiveness of a resource. It was shown by Zesch et al. (2008) that Wiktionary outperforms Wikipedia

in the task of ranking word pairs by their semantic relatedness when taking into account only word pairs that are covered by both resources.

5 Comparison of Semantic Relatedness in Relevant and Non-Relevant Documents

We have shown in Section 4 that a mismatch between the vocabulary of queries and relevant documents exists and that the SR measure and knowledge bases can be used to address this gap. In order to further study the SR of query and document terms with the goal to find SR-based indicators for document relevance, we created sets of relevant and non-relevant documents and compared their characteristic values concerning SR.

5.1 Document Selection

For analysing the impact of SR in the retrieval process, we compare relevant and non-relevant documents that were assigned similar relevance scores by a standard IR system. For the document selection we followed a method employed by Vechtomova et al. (2005). We created two sets of documents for each topic: one for relevant and one for non-relevant documents. We first retrieved up to 1000 documents for the topic using the BM25 IR model[4] (Spärck Jones et al., 2000) as implemented by Terrier[5]. The relevant retrieved documents constituted the first set. For the second set we selected for each relevant retrieved document a non-relevant document which had the closest score to the relevant document. After selecting an equal number of relevant and non-relevant documents, we computed the mean average and the standard deviation for the scores of each set. If there was a substantial difference between the values of more than 20%, the sets were rearranged by exchanging non-relevant documents or excluding pairs of relevant and non-relevant documents. If this was not possible, we excluded the corresponding topic from the experiments.

Table 4 shows the statistics for the final sets. From the original 50 topics, 13 were excluded for the above stated reasons or because no relevant documents were retrieved. The average length of about 345 terms for relevant documents is almost 40% larger than the length of non-relevant documents. Also the average number of query term instances is 6.93 in cotrast to 4.64 for non-relevant documents. The large difference of average document length and query term instances suggests a larger difference of the average relevance scores than 20%. However, in the BM25 model the relevance score is decreased with increasing document length and additional occurrences of a query term have little impact after three or four occurrences.

	Rel.	Nonrel.	Diff. (%)
Number of queries	37	37	0
Number of documents	1771	1771	0
Mean BM25 document score	6.388	6.239	2.39
Stdev BM25 document score	1.442	1.288	12.00
Ave. query length	2.32	2.32	0
Ave. document length	345.22	248.89	38.70
Ave. query term instances in documents	6.93	4.64	49.35

Table 4: Data characteristics that are independent of the chosen knowledge base and threshold.

5.2 Types of Lexical Semantic Relations

The most common classical lexical semantic relations between words are synonymy, hyponymy and a couple of others. In order to analyze the importance of these relations in the retrieval process, we automatically annotated the relations that hold between query and document terms using WordNet. Table 5 shows the percentage of lexical semantic relations between query and document terms (normalized by the number of query and document terms). The table also shows the coverage of the relations by the SR measure, i.e. the percentage of annotated relations for which the SR measure computed a value above 0 or the threshold 0.05, respectively. The percentage of relation types in general is higher for relevant documents. Cohyponymy and synonymy are by far the most frequently occurring relation types with up to almost 6%. Hypernyms and hyponyms have both a percentage of less than 1%. Holonymy and meronymy do almost not occur.

When applying no threshold, the SR measure covers up to 21% of the synonyms and cohyponyms and up to 12% of the hyper- and hyponyms in relevant documents. Using Wiktionary as knowledge base, the SR measure shows a better coverage than with Wikipedia. This is consistent with the findings in Zesch et al. (2008).

[4] We used the default values for the constants of the model ($k_1 = 1.2, b = 0.75$).

[5] http://ir.dcs.gla.ac.uk/terrier/

Relation Type	Percentage	SR-Wikipedia 0.0	SR-Wikipedia 0.05	SR-Wiktionary 0.0	SR-Wiktionary 0.05	SR-WordNet 0.0	SR-WordNet 0.05
Relevant Documents							
synonymy	3.61	17.81	13.13	18.33	13.78	15.28	12.18
hypernymy	0.86	8.57	2.30	12.18	3.02	11.69	2.26
hyponymy	0.88	5.72	1.28	6.33	1.67	6.54	1.02
cohyponymy	5.64	19.49	10.49	21.04	10.05	16.85	8.14
holonymy	0.02	0.61	0.17	0.74	0.17	0.53	0.00
meronymy	0.07	1.94	0.78	2.23	0.74	1.88	0.76
non-classical	—	58.80	6.62	23.22	3.13	12.77	2.56
Non-Relevant Documents							
synonymy	3.41	15.84	12.41	16.46	12.90	14.19	11.44
hypernymy	0.56	6.10	1.95	9.43	2.10	8.93	1.57
hyponymy	0.74	4.77	1.00	6.35	1.40	5.90	0.78
cohyponymy	5.42	17.42	9.91	19.23	9.71	15.38	7.66
holonymy	0.02	0.39	0.09	0.49	0.09	0.32	0.00
meronymy	0.10	1.88	0.55	1.84	0.65	1.57	0.57
non-classical	—	57.33	5.54	21.92	2.59	11.77	2.15

Table 5: Percentage of lexical semantic relations between query and document terms and their coverage by SR scores above threshold 0.00 and 0.05 in percent.

The reason for this is the method for constructing the textual representation of the concepts in the SR measure, where synonyms and other related words are concatenated. Also SR-WordNet outperforms Wikipedia for hypernymy and hyponymy. In contrast to Wiktionary, no direct information about related words is used to construct the textual representation of concepts. However, the very short and specific representations are built from glosses and examples which often contain hypernym-hyponym pairs. As WordNet is used for both, the automatic annotation of lexical semantic relations and the computation of SR values, its lower term coverage in general has not much impact on this experiment, as only the relations between terms contained in WordNet are annotated.

More than half of the SR values using Wikipedia are computed for term pairs which were not annotated with a classical relation. This is depicted in Table 5 as non-classical relation. These non-classical relations can be for example functional relations (*pencil* and *paper*) (Budanitsky and Hirst, 2006). However, as WordNet covers only a small part of the terms in the test collection, some of the SR values refered to as non-classical relations might actually be classical relations. For Wiktionary and WordNet, the number of non-classical relations is much lower, due to their smaller size and the way the textual representations of concepts are constructed. In general, the average number of SR scores for classical and non-classical relations are almost consistently higher for relevant documents which suggests that the comparison of SR scores could be beneficial in the IR process.

Relation Type	SR-Wikipedia 0.0	SR-Wikipedia 0.05	SR-Wiktionary 0.0	SR-Wiktionary 0.05	SR-WordNet 0.0	SR-WordNet 0.05
Relevant Documents						
synonymy	0.362	0.371	0.372	0.374	0.366	0.368
hypernymy	0.021	0.021	0.021	0.019	0.016	0.019
hyponymy	0.017	0.021	0.008	0.012	0.007	0.015
cohyponymy	0.270	0.334	0.315	0.353	0.312	0.363
holonymy	0.001	0.000	0.001	0.000	0.000	0.000
meronymy	0.004	0.003	0.003	0.002	0.003	0.002
non-classical	0.045	0.356	0.098	0.491	0.205	0.599
Non-Relevant Documents						
synonymy	0.344	0.348	0.349	0.350	0.343	0.344
hypernymy	0.027	0.030	0.025	0.029	0.022	0.028
hyponymy	0.019	0.029	0.012	0.023	0.009	0.025
cohyponymy	0.250	0.295	0.277	0.312	0.295	0.334
holonymy	0.001	0.000	0.000	0.000	0.000	0.000
meronymy	0.003	0.002	0.002	0.002	0.003	0.002
non-classical	0.041	0.374	0.103	0.538	0.222	0.643

Table 6: Average values of SR scores corresponding to lexical semantic relations between query and document terms above threshold 0.00 and 0.05 in percent.

When applying a threshold of 0.05, the most visible effect is that the percentage of non-classical relations is decreasing much stronger than the percentage of classical relations. The comparison of the average SR values for each relation type in Table 6 confirms that this is due to the fact that the SR measure assigns on average higher values to the classical relations than to the non-classical relations. After applying a threshold of 0.05 the average SR values corresponding to non-classical relations increase and are equal to or higher than the values for classical relations. The values for classical relations are in general higher for relevant documents, whereas the values for non-classical relations are lower.

5.3 SR-based Indicators for Document Relevance

For each topic and document in one of the sets we computed the SR between the query and document terms. We then computed the arithmetic mean of the following characteristic values for each set: the sum of SR scores, the number of SR scores, the number of terms which are semantically related to a query term and the average SR score. In order to eliminate the difference in document length and average number of query term instances between the relevant and non-relevant sets, we normalized all values, except for the average SR score, by the document length and excluded the SR scores of query term instances.

Figure 1 shows the average difference of these values between relevant and non-relevant document sets for SR-thresholds from 0 to 0.6 (step-size=0.01). As the majority of the SR scores have a low value, there is not much change for thresholds above 0.5.

Except for the average SR score, the differences have a peak at thresholds between 0.01 and 0.09 and decrease afterwards to a constant value. The SR scores computed using Wikipedia show the highest differences. Wiktionary and WordNet perform almost equally, but show lower differences than Wikipedia, especially for the sum of scores. All three knowledge bases show higher differences for the number of scores and number of related terms than for the sum of scores. The differences at the peaks are statistically significant[6], except for the differences of the sum of scores for Wiktionary and WordNet.

For the average SR score, the differences are mostly negative at low thresholds and increase to a low positive value for higher thresholds. A higher number of very low SR values is computed for the relevant documents, which causes the average SR score to be lower than for the non-relevant documents at low thresholds.

Additionally, Figure 2 shows the percentage of topics where the mean value of the relevant document set is higher than the one of the non-relevant document set. Wikipedia shows the highest percentage with about 86% for the number of scores and the number of related terms. Wiktionary and WordNet have a low percentage for the sum of scores, but reach up to 75% for the number of scores and the number of related terms.

[6]We used the Wilcoxon test at a significance level of 0.05.

The analysis of the SR of query and document terms shows that there are significant differences for relevant and non-relevant documents that can be measured by computing SR scores with any of the three knowledge bases. Especially when using Wiktionary and WordNet, the number of SR scores and the number of related terms might be better indicators for the document relevance than the sum of SR scores.

6 Conclusions

The vocabulary mismatch of queries and documents is a common problem in IR, which becomes even more serious the larger the document collection grows. CKBs like Wikipedia and Wiktionary, which have been applied in computational methods only recently, offer new possibilities to tackle this problem. In order to find the most beneficial way to employ these resources, we studied the semantic relatedness of query and document terms of an IR benchmark and compared the characteristics and performance of the CKBs Wikipedia and Wiktionary to the LKB WordNet.

We first proved that there exists a vocabulary gap in the test collection between queries and documents and that it can be reduced by employing a concept vector based measure for SR with any of the three knowledge bases. Using WordNet to automatically annotate the lexical semantic relations of query and document terms, we found that cohyponymy and synonymy are the most frequent classical relation types. Although the percentage of annotated relations for which also the SR measure computed values above a predefined threshold was at best 21%, the average number of SR scores for classical and non-classical relations were almost consistently higher for relevant documents.

Comparing the number and the value of SR scores of query and document terms, a significant difference between relevant and non-relevant documents was observed by using any of the three knowledge bases. Although Wikipedia had the best coverage of collection terms and showed the best perfomance in our experiments, Wiktionary and Wikipedia also seem to have a sufficient size for being beneficial in IR. In comparison to our previous work where the sum of SR scores was used as an indicator for document relevance (Müller and Gurevych, 2008), the results suggest that the number of SR scores and the number of related terms might show a better performance, es-

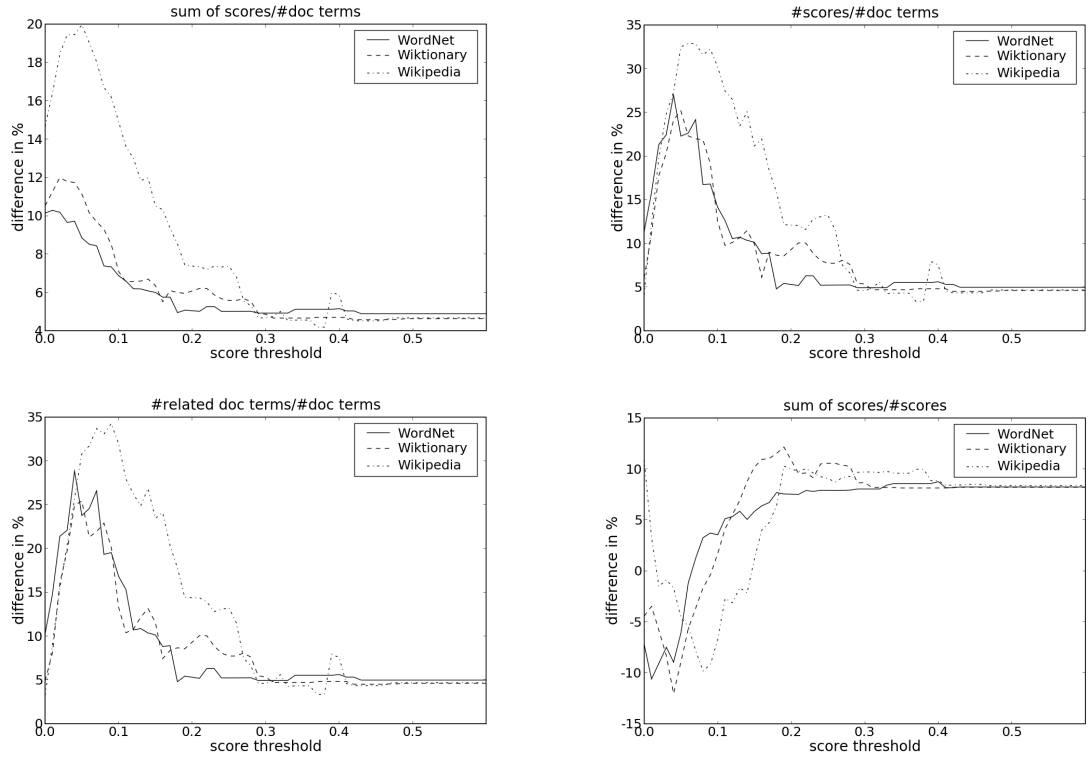

Figure 1: Differences between mean values of relevant and non-relevant document sets.

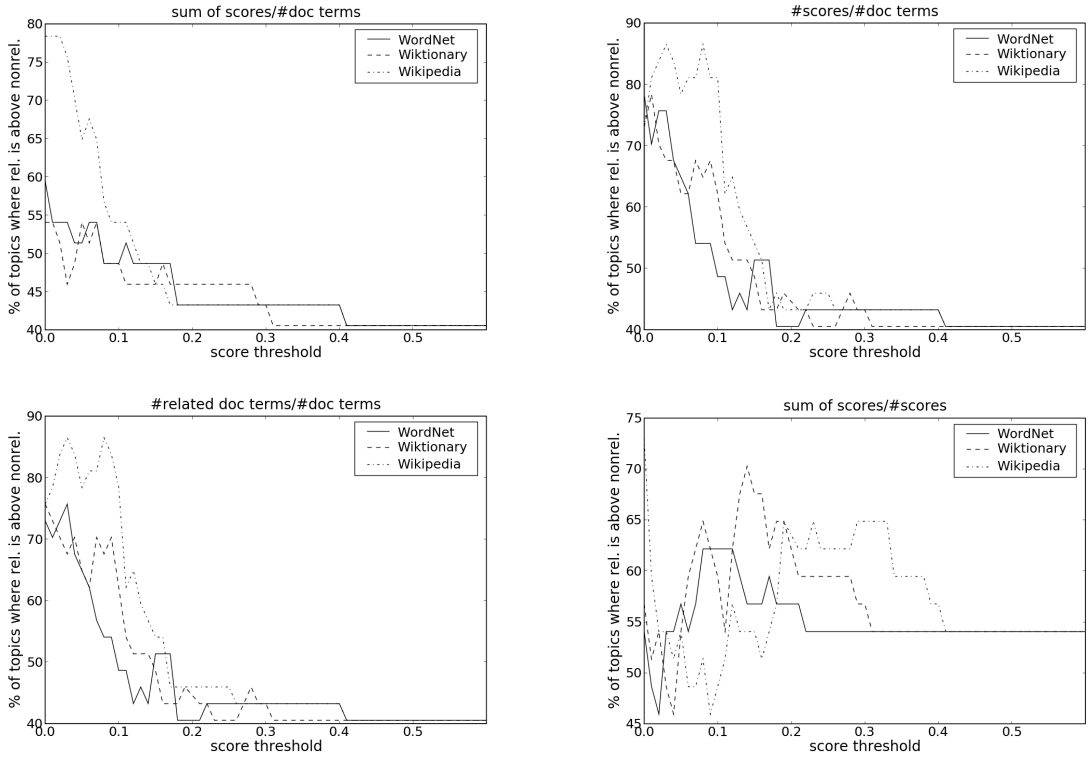

Figure 2: Percentage of topics where the mean value of the relevant document sets is higher than the one of the non-relevant document sets.

pecially for Wiktionary and WordNet.

In our future work, we plan to extend our analysis to other test collections and to query expansion methods in order to generalize our conclusions. As the problem of language ambiguity has a high impact on the use of SR measures, we will also consider word sense disambiguation in our future experiments.

Acknowledgments

This work was supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806 and by the German Research Foundation under grant No. GU 798/1-3. We would like to thank Aljoscha Burchardt for his helpful comments and the anonymous reviewers for valuable feedback on this paper.

References

D. Bernhard and I. Gurevych. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Singapore, Aug.

D. C. Blair and M. E. Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM*, 28(3):289–299.

A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based Measures of Semantic Distance. *Computational Linguistics*, 32(1):13–47.

P. Chesley, B. Vincent, L. Xu, and R. Srihari. 2006. Using Verbs and Adjectives to Automatically Classify Blog Sentiment. In *Proceedings of AAAI-CAAW-06*.

O. Egozi, E. Gabrilovich, and S. Markovitch. 2008. Concept-Based Feature Generation and Selection for Information Retrieval. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, Chicago, IL.

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

E. Gabrilovich and S. Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence*, pages 1606–1611, Hyderabad, India.

I. Gurevych, C. Müller, and T. Zesch. 2007. What to be? - Electronic Career Guidance Based on Semantic Relatedness. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 1032–1039, Prague, Czech Republic, June.

R. Mandala, T. Tokunaga, and H. Tanaka. 1998. The Use of WordNet in Information Retrieval. In Sanda Harabagiu, editor, *Proceedings of the COLING-ACL workshop on Usage of WordNet in Natural Language Processing*, pages 31–37. Association for Computational Linguistics, Somerset, New Jersey.

D. Milne and I. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Wikipedia and AI workshop at the AAAI-08 Conference (WikiAI08)*, Chicago, USA.

C. Müller and I. Gurevych. 2008. Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In F. Borri, A. Nardi, and C. Peters, editors, *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, Sep.

M. Potthast, B. Stein, and M. Anderka. 2008. A Wikipedia-Based Multilingual Retrieval Model. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors, *30th European Conference on IR Research, ECIR 2008, Glasgow*, volume 4956 of *LNCS*, pages 522–530. Springer.

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

P. Schönhofen, I. Biro, A. A. Benczur, and K. Csalogany. 2007. Performing Cross Language Retrieval with Wikipedia. In *Working Notes for the CLEF 2007 Workshop*.

P. Sorg and P. Cimiano. 2008. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In F. Borri, A. Nardi, and C. Peters, editors, *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, Sep.

K. Spärck Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36(6):779–808 (Part 1); 809–840 (Part 2).

M. Strube and S. P. Ponzetto. 2006. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *Proceedings of AAAI*, pages 1419–1424.

O. Vechtomova, M. Karamuftuoglu, and S. E. Robertson. 2005. A Study of Document Relevance and Lexical Cohesion between Query Terms. In *Proceedings of the Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-World Applications (ELECTRA 2005), the 28th Annual International ACM SIGIR Conference*, pages 18–25, Salvador, Brazil, August.

E. M. Voorhees. 1994. 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. Springer-Verlag New York, Inc.

N. Weber and P. Buitelaar. 2006. Web-based Ontology Learning with ISOLDE. In *Proc. of the Workshop on Web Content Mining with Human Language at the International Semantic Web Conference*, Athens GA, USA, 11.

T. Zesch, I. Gurevych, and M. Mühlhäuser. 2007. Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In *Proceedings of HLT-NAACL*, pages 205–208.

T. Zesch, C. Müller, and I. Gurevych. 2008. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008*, pages (861–867), Chicago, Illinois, USA.