Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval

Christof Müller and Iryna Gurevych

Ubiquitous Knowledge Processing Lab, Computer Science Department
Technische Universität Darmstadt,
Hochschulstr. 10, D-64289 Darmstadt, Germany
{mueller, gurevych}@tk.informatik.tu-darmstadt.de

Abstract

The main objective of our experiments in the domain-specific track at CLEF 2008 is utilizing semantic knowledge from collaborative knowledge bases such as Wikipedia and Wiktionary to improve the effectiveness of information retrieval. While Wikipedia has already been used in IR, the application of Wiktionary in this task is new. We evaluate two retrieval models, i.e. SR-Text and SR-Word, based on semantic relatedness by comparing their performance to a statistical model as implemented by Lucene. When Lucene is combined with the semantic models the mean average precision increases by 14% for German, 9% for English, and 16% for Russian. In the bilingual task, we translate the English topics into the document language, i.e. German, by using machine translation. For SR-Text, we alternatively perform the translation process by using cross-language links in Wikipedia, whereby the terms are directly mapped to concept vectors in the target language. The evaluation shows that the latter approach especially improves the retrieval performance in cases where the machine translation system incorrectly translates query terms. When Lucene is combined with SR-Text, the mean average precision increases by 34%.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing – Thesauruses; H.3.3 Information Search and Retrieval – Retrieval models; H.3.4 Systems and Software – Performance evaluation (efficiency and effectiveness); H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Semantic Relatedness, Collaborative Knowledge Bases, Cross-Language Information Retrieval

1 Introduction

Statistical models are most frequently used in domain-specific information retrieval (**IR**). One of the disadvantages of these models is their lack of flexibility concerning synonymy, i.e. expressing a concept with different terms. There exist several approaches of tackling the problem of synonymy divided into local and global methods.

Local methods like relevance and pseudo-relevance feedback try to refine the representation of the user's information need by using manual respectively automatic feedback about already returned documents. However, these methods require that the relevant documents show a significant term overlap, and that the term overlap between relevant and irrelevant documents is small. Also they are not able to close the gap between the vocabulary used in queries and in documents, i.e. query terms which do not occur in the document collection can not be expanded with related terms.

Global methods expand the query with related terms using either automatically built thesauri based on the document collection or external linguistic knowledge bases (**LKBs**) like WordNet [2]. Using thesauri which are based on the document collection also suffers from the inability to close the vocabulary gap, if query terms do not occur in the document collection. The use of LKBs for query expansion has shown inconclusive results so far. Voorhees [27] could improve retrieval performance only in some cases even for manually selected expansion terms, while Mandala *et al.* [15] improved the performance on several test collections by combining a LKB with different types of thesauri built from the underlying text collections. The general problem of query expansion is that in fact it is able to improve recall in certain situations, but at the same time precision degrades as also irrelevant terms are added to the query.

Another semantic approach to tackle the problem of synonymy is to use retrieval models which are based on semantic relatedness (**SR**) between query and document terms computed by using LKBs. Although first results of employing SR in IR were inconclusive [23], there have also been several promising results [14, 3, 18]. The main problem with using LKBs for semantically enhanced IR is the low coverage of domain-specific vocabulary and proper nouns.

A new form of resources, so called collaborative knowledge bases (**CKBs**) have the potential to overcome these limitations. Enabled by Web 2.0 technologies, CKBs are constructed by volunteers on the web and have reached a size which makes them promising for improving IR performance. The most widely used and probably largest CKB is Wikipedia[1]. It contains encyclopedic knowledge in a broad range of domains and has been recently employed as a knowledge base (**KB**) in IR with very positive results [12, 8, 22, 19].

For our experiments in the domain-specific track at CLEF 2008, we employ Wikipedia and for the first time Wiktionary as knowledge bases for SR-based IR models. We compare their performance to a statistical model and also combine all three models by combining their respective relevance scores for each document. We perform the experiments for the languages English, German, and Russian.

For bilingual IR experiments using English topics on a German document collection we use (i) machine translation methods for statistical and semantic models, and (ii) cross-language links in Wikipedia for one semantic model.

The remainder of this paper is structured as follows: In Section 2, we give a short overview about the employed knowledge bases. In Section 3, we explain the SR-based and the statistical IR model. The test collections used in our experiments are described in Section 4. This is followed by Section 5 where the results of the experiments are presented and discussed. Finally, we draw some conclusions in Section 6.

2 Collaborative Knowledge Bases

The development of Web 2.0 technology in recent years has led to a vastly increasing amount of user generated content on the world wide web, which is created and controlled by decentralized communities of volunteers with diverse personal backgrounds and fields of expertise. Most CKBs are freely available and they do not suffer from the various disadvantages of LKBs such as:

- their coverage and size are limited;
- they are mainly restricted to common vocabulary;

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

- continuous maintenance is often not feasible;
- the content is often quickly out-dated;
- only major languages are typically supported.

However, the downside of CKBs is that they mostly contain semi- or unstructured text which first needs to be transformed into structured knowledge in order to be used as a KB. A potential problem with CKBs is the quality of their content. Most CKBs lack editorial quality control. However, it has been found, e.g. even without any explicit process of quality control, the factual quality in Wikipedia is high [7]. On the other hand, the quality of LKBs like WordNet also has been criticized in the past [11].

2.1 Wikipedia

Started in 2001 as a *multilingual, Web-based, free content encyclopedia project*, Wikipedia is probably the largest collection of freely available knowledge. It contains about 10 million articles in more than 250 languages. The English language version of Wikipedia is by far the largest with almost 2.4 million articles, followed by the German language version with 754,000 articles. The knowledge stored in Wikipedia which can be exploited for computational methods consists not only of the articles' text itself. For example, in [26, 30] the hierarchy of categories that Wikipedia articles are tagged with is used for computing semantic relatedness of word pairs. Milne *et al.* [16] employ the link structure of Wikipedia articles for extracting a domain-specific thesaurus and use it for query expansion in IR. Schönhofen *et al.* [22] improve cross-lingual IR by using redirecting links of Wikipedia articles to identify synonyms. They also employ links between articles on the same topic in different languages to find term and phrase translations.

Especially for IR, it is important that Wikipedia contains a lot of named entities which are usually missing in LKBs.

2.2 Wiktionary

Wiktionary is a multilingual dictionary and a sister project of Wikipedia. Unlike Wikipedia, it focuses on lexical instead of encyclopedic knowledge, which makes Wikipedia and Wiktionary complementary knowledge sources. Wiktionary contains many types of information also found in LKBs, like definitions, synonyms, and hyponyms, and also additional types of information, e.g. abbreviations, compounds or contractions, which are usually not found in LKBs. Another difference to LKBs is that each language-specific edition of Wiktionary contains not only entries for words in that particular language, but also for words in other languages. Wiktionary has about 3.7 million word entries in 171 language editions in total. The English and French language versions are the largest with roughly 800,000 entries. Compared to this, the German edition is rather small consisting of less than 80,000 entries.

Wiktionary has been employed for tasks such as sentiment analysis [1] or ontology learning [28], but we are not aware of any work that employed it in IR before.

3 Information Retrieval Models

3.1 Preprocessing

Besides applying standard preprocessing steps like tokenization and stopword removal, we use lemmatization employing the TreeTagger [21] for all tasks. For the German test data, we also split compounds into their constituents [13], and we use both, constituents and compounds in the retrieval process.

3.2 Statistical Model

The statistical IR model we use in our experiments is the model as implemented by Lucene[2], an open source text search library.

The model first extracts the relevant documents from the collection by matching the query against the index. In the second step, the actual ranking of the relevant documents is computed by using a vector space model according to the following equation:

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

where n_q is the number of query terms, $tf(t_q,d)$ is the term frequency factor for term t_q in document d, $idf(t_q)$ is the inverse document frequency of the term, and $norm(d)$ is a normalization value of document d, given the number of terms within the document.

3.3 Semantic Models

An obvious solution to the problem of synonymy and also polysemy in IR is to perform the retrieval process also on the semantic level rather than only on the level of surface forms. One such method is Latent Semantic Indexing [5] where the term vectors of query and documents are mapped into a (lower dimensional) conceptual space. In [20], a similar method is used to build a similarity thesaurus for query expansion. However, as the conceptual space is derived from the relations of terms and documents in the collection, these methods cannot solve a mismatch of the vocabulary in queries and the document collection. Instead of using the document collection for deriving the conceptual space, an external document collection with a large number of documents from a wide range of domains can be used. Koberstein *et al.* [12] use Wikipedia as corpus to calculate word similarities by applying different measures based on the co-occurrence of the terms in the same Wikipedia article. The retrieved word clusters are then applied to compute sentence-based document similarity.

Gabrilovich and Markovitch [6] propose a similar approach where they refer to Wikipedia articles as concepts. Thereby, each term contained in Wikipedia is represented in the concept space as a vector of tf.idf values [25], derived from the term's occurrence in the respective Wikipedia articles. The similarity of two documents is then computed using a centroid-based classifier [9]. The concept vector of each term in the document is weighted with the term's tf.idf value. From these weighted concept vectors an average vector is calculated which represents the respective document in the concept space. The similarity score of two documents is then computed using the cosine metric. This method was successfully employed for IR in the domain of electronic career guidance [8] and in a multilingual IR model [19].

In our experiments, we use the method proposed by Gabrilovich and Markovitch as a SR-based IR model and refer to it as **SR-Text**. Additionally, we employ a retrieval model proposed in [17] to which we refer as **SR-Word**. We extended the model by also taking into account the idf value of document terms and the tf value of query and document terms. The formula of this model is as follows:

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

where n_d is the number of unique terms in the document, n_q the number of unique terms in the query, $t_{d,i}$ the i-th unique document term, $t_{q,j}$ the j-th unique query term, $s(t_{d,i},t_{q,j})$ the SR score for the respective document and query term (using the cosine of the respective terms' concept vectors as score), n_{nsm} the number of unique query terms not exactly contained in the document, and n_{nr} the number of unique query terms which do not contribute a SR score above the threshold. For SR-Text and SR-Word, we compute tf and idf as follows:

$$tf(t) = 1 + \log f(t)$$

[2] `http://lucene.apache.org`

where $f(t)$ is the frequency of term t in the corresponding document or query, and

$$idf(t) = \frac{n_{docs}}{df(t)}$$

where n_{docs} is the number of documents in the collection and $df(t)$ is the number of documents in the collection containing term t.[3]

Besides Wikipedia we use Wiktionary as KB for the IR models. Thereby, we refer to each word entry in these KBs as a concept, and use the entry's information as the textual representation of a concept analogous to the text of Wikipedia articles (for details see [32]). In order to improve retrieval effectiveness, we combine the concept space of Wikipedia and Wiktionary, so that the concept vector of one term consists of concepts from both KBs. When using Wikipedia as KB we remove concepts where the respective Wikipedia articles have less than 100 words or fewer than 5 in- or outlinks. For both, Wikipedia and Wiktionary, we remove concepts from a term's concept vector if the tf.idf value is below the predefined threshold of *0.01*. The pruning methods are applied to achieve noise reduction and better performance. For accessing the CKBs we use freely available Java-based APIs described in [31].

3.4 Combination of Models

As the statistical and semantic models use different types of information represented in queries, documents, and possibly external knowledge we hypothesize that the combination of the models will increase the retrieval effectiveness. We therefore combine their relevance scores computed in separate retrieval runs into one relevance score for each document per query. For computing the combined relevance score, we use the *CombSUM* method which was introduced by Fox and Shaw [4] where the combined relevance score is set to the sum of the individual relevance scores. This method has been shown to outperform other methods for English, German, and other European languages [10]. Before combining the scores, they are normalized using the formula:

$$r_{norm} = \frac{r_{orig} - r_{min}}{r_{max} - r_{min}}$$

where r_{orig} is the original relevance score, r_{min} is the minimal and r_{max} is the maximal occurring score for the query. This normalization method was one of the top performing approaches in [29].

3.5 Methods for Bilingual Retrieval

For the cross-lingual IR runs we use machine translation (Systran Translator[4]) for translating the query into the language of the documents. For the SR-Text model, we additionally explore a different method using the cross-language links between different language editions of Wikipedia. A cross-language link points from an article in one language to the same article in a different language, e.g. an English article might point to its German counterpart. Using these links in a similar way as proposed in [19], we are able to map a concept vector whose concepts are represented by articles in the English Wikipedia into a concept vector whose concepts are represented by articles in the German Wikipedia. Thus, by transforming the concept vector of an English query using cross-language links, the similarity between the English query and the German documents is computed by the SR-Text model without actually translating the query.[5]

4 Overview of Test Collections

The test collections consist of structured bibliographic data for social sciences. The following corpora are used: (i) for German, the GIRT-4 database consisting of 151,319 documents, (ii)

[3] As it is possible for a query term to not appear in the document collection at all, we set the document frequency to 1 instead of 0 for calculating idf in these cases.

[4] `http://babelfish.yahoo.com/`

[5] As we do not actually translate the query terms, we are not able to compute the idf of the terms in the document collection. Instead we use the idf in Wikipedia.

for English, a translation of the GIRT-4 corpus and the database of Sociological Abstracts from Cambridge Scientific Abstracts (CSA) containing 20,000 documents, and (iii) for Russian, the INION corpus ISISS with 145,802 documents. Each document contains the title, the author, the abstract, and the source information of a publication along with the subject metadata from controlled vocabularies. For building the document index, we use the complete information except for the author, the year of publication, and the identification number.

The queries are created from 25 topics available in the languages English, German, and Russian. Each topic consists of three fields. The *title* field (**T**) contains a few keywords describing the user's information need. The *description* field (**D**) contains one sentence characterizing the information need in more detail. The *narrative* field (**N**) contains several sentences which specify in even more detail what a document should or should not contain to be judged relevant for this query.

5 Evaluation

We experiment with several query types by using different combinations of the topic fields. In our test runs using topics from the past CLEF workshops, we found that the retrieval effectiveness improved when query terms are weighted depending on the field in which they occur. We therefore use the following weights for query terms in all experiments: 1 for title, 0.8 for description, and 0.6 for narrative.

In the following sections, we present several tables which contain the *mean average precision* (**MAP**) values of official and inofficial runs for the respective task. For each run the tables also contain the ID, the type (official/inofficial), the topic fields which were used for generating the query, the IR model, the MAP value for a single model, and the MAP value for the combination of models. For the semantic models, we mention the employed KBs. For SR-Word, we also give the value of the predefined threshold for SR values. In the bilingual task, we also point out the translation method which was used. The highest MAP value for each task is in bold.

5.1 Monolingual Retrieval

5.1.1 English

We submitted two official runs which are combinations of all three models, one run using only the title and description fields, the other one using all three topic fields. Table 1 shows the official runs along with several inofficial ones.

The statistical model outperforms the semantic models for both query types. However, the best performance in terms of MAP is reached when all three models are combined. For the query type *TDN*, MAP improves from 0.2987 to 0.3242 when combining the models. Overall, the inclusion of the narrative field of the topics improves MAP. However, for the SR-Word model we found in our training runs that MAP decreases when the narrative field is taken into account. Therefore, we only use the title and description fields for this model. The SR-Word model outperforms the SR-Text model. We hypothesize that this happens because the SR-Word model also accounts for direct string matching.

5.1.2 German

In the monolingual task for German, we submitted two official runs using the combinations TD and TDN of the topic fields. Table 2 shows the results of the official and inofficial runs. The results are similar to the monolingual runs for English. Generally, we yield higher MAP values, and for the TD query type the SR-Word model outperforms the statistical model. Again the combination of all three models performs best, and for the query type TDN the combination of all models improves the MAP value from 0.3536 to 0.3950. Also for this task, the SR-Word model outperforms the SR-Text model.

ID	*Official*	*Query*	*Model*	*Single MAP*	*Combined MAP*
964		TD	Lucene	0.2983	0.2781
		TD	SR-Text: WP+WKT	0.2040	
983		TD	Lucene	0.2983	0.3017
		TD	SR-Word: WP+WKT(0.25)	0.2536	
984		TD	SR-Text: WP+WKT	0.2040	0.2789
		TD	SR-Word: WP+WKT(0.25)	0.2536	
969	x	TD	Lucene	0.2983	0.3053
		TD	SR-Text: WP+WKT	0.2040	
		TD	SR-Word: WP+WKT(0.25)	0.2536	
975		TDN	Lucene	0.2987	0.2948
		TDN	SR-Text: WP+WKT	0.2289	
985		TDN	Lucene	0.2987	0.3158
		TD	SR-Word: WP+WKT(0.25)	0.2536	
986		TDN	SR-Text: WP+WKT	0.2289	0.3010
		TD	SR-Word: WP+WKT(0.25)	0.2536	
976	x	TDN	Lucene	0.2987	**0.3242**
		TDN	SR-Text: WP+WKT	0.2289	
		TD	SR-Word: WP+WKT(0.25)	0.2536	

Table 1: Results of monolingual runs for English.

ID	*Official*	*Query*	*Model*	*Single MAP*	*Combined MAP*
962		TD	Lucene	0.3318	0.3602
		TD	SR-Text: WP+WKT	0.3223	
987		TD	Lucene	0.3318	0.3613
		TD	SR-Word: WP+WKT (0.11)	0.3447	
988		TD	SR-Text: WP+WKT	0.3223	0.3698
		TD	SR-Word: WP+WKT (0.11)	0.3447	
970	x	TD	Lucene	0.3318	0.3770
		TD	SR-Text: WP+WKT	0.3223	
		TD	SR-Word: WP+WKT (0.11)	0.3447	
973		TDN	Lucene	0.3536	0.3778
		TDN	SR-Text: WP+WKT	0.3329	
989		TDN	Lucene	0.3536	0.3835
		TD	SR-Word: WP+WKT (0.11)	0.3447	
1012		TDN	SR-Text: WP+WKT	0.3329	0.3781
		TD	SR-Word: WP+WKT (0.11)	0.3447	
974	x	TDN	Lucene	0.3536	**0.3950**
		TDN	SR-Text: WP+WKT	0.3329	
		TD	SR-Word: WP+WKT (0.11)	0.3447	

Table 2: Results of the monolingual runs for German.

ID	*Official*	*Query*	*Model*	*Single MAP*	*Combined MAP*
965		TD	Lucene	0.1254	0.1224
		TD	SR-Text: WP	0.1016	
991		TD	Lucene	0.1254	0.1386
		T	SR-Word: WP (0.23)	0.1211	
992		TD	SR-Text: WP	0.1016	0.1208
		T	SR-Word: WP (0.23)	0.1211	
970	x	TD	Lucene	0.1254	0.1356
		TD	SR-Text: WP	0.1016	
		T	SR-Word: WP (0.23)	0.1211	
977		TDN	Lucene	0.1286	0.1268
		TDN	SR-Text: WP	0.0749	
993		TDN	Lucene	0.1286	**0.1491**
		T	SR-Word: WP (0.23)	0.1211	
994		TDN	SR-Text: WP	0.0749	0.1179
		T	SR-Word: WP (0.23)	0.1211	
978	x	TDN	Lucene	0.1286	0.1400
		TDN	SR-Text: WP	0.0749	
		T	SR-Word: WP (0.23)	0.1211	

Table 3: Results of the monolingual runs for Russian.

5.1.3 Russian

Table 3 shows the results of the official and inofficial runs for Russian. In the two official runs using the topic field combinations TD and TDN respectively, MAP is generally much lower than for English and German. In this setting, Lucene and SR-Word outperform the SR-Text model, too. SR-Word performs almost as well as Lucene. In contrast to the runs in the German and English tasks, the use of the topic field *narrative* decreases MAP for SR-Text. Also we found in our training runs, that it is beneficial to only use the title field as the query for the SR-Word model. As for the English and German tasks, the combination of the three models increases MAP. The combination of Lucene and SR-Word shows the best performance, increasing MAP by about 16% compared to using Lucene alone.

5.2 Bilingual Retrieval

For the bilingual retrieval, we submitted four runs where we used English topics with the German document collection. As described in Section 3.5, the English topics were translated into German using machine translation (**MT**). In the case of SR-Text, we also mapped the concept vector of English terms directly to its German counterpart without first translating the term itself using cross-language links in Wikipedia (**CLL**). As Wiktionary also has cross-language links and furthermore many of the word entries contain translations of the term into other languages, it is in principle possible to apply the CLL method to both Wikipedia and Wiktionary. As we only implemented this method for Wikipedia, the reported runs using CLL employ only Wikipedia as KB.

Generally, the MAP values in our bilingual runs are much lower compared to the monolingual German runs as both methods, MT and CLL, add noise to the retrieval process. For the query type TD, SR-Word is the best performing single model. For the query type TDN, Lucene performs slightly better than SR-Word. On first sight, the MT method seems to yield better results for SR-Text than the CLL method. When combined with the Lucene model, SR-Text using CLL outperforms SR-Text using MT by about 0.04. In these cases MAP decreases when SR-Word is added to the combination. The best performing run with a MAP of 0.2342 is using the query type TDN and a combination of Lucene and SR-Text with CLL. Compared to using Lucene alone, MAP increases by almost 34%. Analyzing the results on the query level, we found that the CLL method is especially beneficial in cases of substantial translation errors of query terms. In topic

ID	*Official*	*Query*	*Translation*	*Model*	*Single MAP*	*Comb. MAP*
966		TD	MT	Lucene	0.1638	0.2167
		TD	CLL	SR-Text: WP	0.1339	
1013		TD	MT	Lucene	0.1638	0.1775
		TD	MT	SR-Word: WP+WKT (0.11)	0.1723	
1014		TD	CLL	SR-Text: WP	0.1339	0.2038
		TD	MT	SR-Word: WP+WKT (0.11)	0.1723	
971	x	TD	MT	Lucene	0.1638	0.2060
		TD	CLL	SR-Text: WP	0.1339	
		TD	MT	SR-Word: WP+WKT (0.11)	0.1723	
967		TD	MT	Lucene	0.1638	0.1776
		TD	MT	SR-Text: WP+WKT	0.1499	
1015		TD	MT	SR-Text: WP+WKT	0.1499	0.1850
		TD	MT	SR-Word: WP+WKT (0.11)	0.1723	
972	x	TD	MT	Lucene	0.1638	0.1875
		TD	MT	SR-Text: WP+WKT	0.1499	
		TD	MT	SR-Word: WP+WKT (0.11)	0.1723	
979		TDN	MT	Lucene	0.1746	**0.2342**
		TD	CLL	SR-Text: WP	0.1399	
1016		TDN	MT	Lucene	0.1746	0.1958
		TD	MT	SR-Word: WP+WKT (0.11)	0.1723	
980	x	TDN	MT	Lucene	0.1746	0.2231
		TD	CLL	SR-Text: WP	0.1399	
		TD	MT	SR-Word: WP+WKT (0.11)	0.1723	
981		TDN	MT	Lucene	0.1746	0.1935
		TD	MT	SR-Text: WP+WKT	0.1499	
982	x	TDN	MT	Lucene	0.1746	0.2001
		TD	MT	SR-Text: WP+WKT	0.1499	
		TD	MT	SR-Word: WP+WKT (0.11)	0.1723	

Table 4: Results of the bilingual runs using the English topics and the German documents.

no. 209 where the English title field contains the terms *Doping and sports* the correct German translation of *Doping* would be the same term *Doping*. Instead, it is incorrectly translated by the machine translation system to *Lackieren* which has the meaning of *painting* or *to lacquer*. As the Lucene model relies on the translation with the MT system, the combination with SR-Text using the CLL method especially improves the retrieval in these cases. The generally lower performance of SR-Text when using CLL instead of MT might be caused by the differences in employed KBs, as SR-Text using MT employs not only Wikipedia, but additionally Wiktionary. Another reason could be the missing cross-language links between articles in the German and English Wikipedia. Not even half of the articles in the German Wikipedia link to the respective articles in the English Wikipedia. As a possible solution, automatic methods for enriching the cross-language link structure in Wikipedia as proposed in [24] could be applied.

6 Conclusion

In our experiments, we have explored the integration of semantic knowledge from collaborative knowledge bases into IR. For the first time, we have employed Wiktionary in combination with Wikipedia for this task. We have evaluated two IR models, i.e. SR-Text and SR-Word, based on semantic relatedness by comparing their performance to a statistical model as implemented by Lucene. In both semantic models, the articles in Wikipedia and the word entries in Wiktionary are employed as textual representations of concepts. The SR-Text model computes the similarity of a query and document using a centroid-based classifier. The SR-Word model combines individual similarities of each query and document term pair that are above a predefined threshold and then

applies a set of heuristics.

In the monolingual task, we found that SR-Word outperformed SR-Text in most experiments. SR-Word outperformed Lucene only in one experiment. However, when Lucene was combined with the semantic models by using the CombSUM method, the MAP increased by 14% for German, 9% for English, and 16% for Russian.

In the bilingual task, we translated the English topics into the document language, i.e. German, by using machine translation. For SR-Text, we additionally explored a different method using the cross-language links between different language editions of Wikipedia. This approach especially improved the retrieval performance in cases where the machine translation system incorrectly translated terms. When Lucene was combined with SR-Text, the MAP increased by 34%. In our future work, we will additionally use the cross-language links in Wiktionary to further improve the IR effectiveness. We also plan to integrate the cross-language links into the SR-Word model.

7 Acknowledgement

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 798/1-2 and GU 798/1-3.

References

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

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

[3] G. Fliedner. A Generalised Similarity Measure for Question Answering. In *Proceedings of NLDB 2005*, volume 3513 of *LNCS*, pages 380–383, Alicante, 2005.

[4] E. Fox and J. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.

[5] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–480, New York, NY, USA, 1988. ACM.

[6] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of IJCAI*, pages 1606–1611, 2007.

[7] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[8] 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*, pages 1032–1039, Prague, Czech Republic, June 2007.

[9] E.-H. Han and G. Karypis. Centroid-Based Document Classification: Analysis and Experimental Results. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 424–431, London, UK, 2000. Springer-Verlag.

[10] J. Kamps and M. de Rijke. The effectiveness of combining information retrieval strategies for European languages. In *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*, pages 1073–1077, New York, NY, USA, 2004. ACM.

[11] A. N. Kaplan and L. K. Schubert. Measuring and improving the quality of world knowledge extracted from WordNet. Tech. Rep. 751 14627-0226, Dept. of Computer Science, Univ. of Rochester, 2001.

[12] J. Koberstein and Y.-K. Ng. Using Word Clusters to Detect Similar Web Documents. In J. Lang, F. Lin, and J. Wang, editors, *KSEM*, volume 4092 of *LNCS*, pages 215–228. Springer, 2006.

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

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

[15] R. Mandala, T. Tokunaga, and H. Tanaka. The Use of WordNet in Information Retrieval. In S. 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, 1998.

[16] D. N. Milne, I. H. Witten, and D. M. Nichols. A Knowledge-Based Search Engine Powered by Wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Information and knowledge management*, pages 445–454, New York, NY, USA, 2007. ACM.

[17] C. Müller and I. Gurevych. Exploring the Potential of Semantic Relatedness in Information Retrieval. In M. Schaaf and K.-D. Althoff, editors, *LWA 2006 Lernen - Wissensentdeckung - Adaptivität, 9.-11.10.2006 in Hildesheim*, Hildesheimer Informatikberichte, pages 126–131, Hildesheim, Germany, 2006. GI-Fachgruppe Information Retrieval, Universität Hildesheim.

[18] C. Müller, I. Gurevych, and M. Mühlhäuser. Integrating Semantic Knowledge into Text Similarity and Information Retrieval. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC)*, pages 257–264, Irvine, CA, USA, 2007.

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

[20] Y. Qiu and H. Frei. Concept Based Query Expansion. In *Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval*, 1993.

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

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

[23] A. Smeaton. Using NLP or NLP Resources for Information Retrieval Tasks. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 99–111. Kluwer Academic Publishers, 1999.

[24] P. Sorg and P. Cimiano. Enriching the crosslingual link structure of Wikipedia - A classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, 2008.

[25] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

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

[27] 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.

[28] N. Weber and P. Buitelaar. 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 2006.

[29] S. Wu, F. Crestani, and Y. Bi. Evaluating Score Normalisation Methods in Data Fusion. In *Proceedings of AIRS 2006, the 3rd Asia Information Retrieval Symposium*, 2006.

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

[31] T. Zesch, C. Müller, and I. Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, 2008.

[32] T. Zesch, C. Müller, and I. Gurevych. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of AAAI*, pages (861–867), 2008.