What to be? - Electronic Career Guidance Based on Semantic Relatedness

Iryna Gurevych, Christof Müller and Torsten Zesch
Ubiquitous Knowledge Processing Group
Telecooperation, Darmstadt University of Technology
Hochschulstr. 10, 64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de
{gurevych,mueller,zesch}@tk.informatik.tu-darmstadt.de

Abstract

We present a study aimed at investigating the use of semantic information in a novel NLP application, Electronic Career Guidance (ECG), in German. ECG is formulated as an information retrieval (IR) task, whereby textual descriptions of professions (*documents*) are ranked for their relevance to natural language descriptions of a person's professional interests (*the topic*). We compare the performance of two semantic IR models: (IR-1) utilizing semantic relatedness (SR) measures based on either wordnet or Wikipedia and a set of heuristics, and (IR-2) measuring the similarity between the topic and documents based on Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007). We evaluate the performance of SR measures intrinsically on the tasks of (T-1) computing SR, and (T-2) solving Reader's Digest Word Power (RDWP) questions.

1 Electronic Career Guidance

Career guidance is important both for the person involved and for the state. Not well informed decisions may cause people to drop the training program they are enrolled in, yielding loss of time and financial investments. However, there is a mismatch between what people know about existing professions and the variety of professions, which exist in reality. Some studies report that school leavers typically choose the professions known to them, such as *policeman*, *nurse*, etc. Many other professions, which can possibly match the interests of the person very well, are not chosen, as their titles are unknown and people seeking career advice do not know about their existence, e.g. *electronics installer*, or *chemical laboratory worker*. However, people are very good at describing their professional interests in natural language. That is why they are even asked to write a short essay prior to an appointment with a career guidance expert.

Electronic career guidance is, thus, 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[1] is based on vocational trainings, manually annotated using a tagset of 41 keywords. The user must select appropriate keywords according to her interests. In reply, the system consults a knowledge base with professions manually annotated with the keywords by career guidance 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.

At the same time, GFLO maintains an extensive database with textual descriptions of professions,

[1] http://www.interesse-beruf.de/

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 1032–1039,
Prague, Czech Republic, June 2007.

called BERUFEnet.[2] Therefore, we cast the problem of ECG as an IR task, trying to remove the disadvantages of conventional ECG outlined above by letting the user describe her interests in a short natural language essay, called *a professional profile*.

```
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.
```

Textual descriptions of professions are ranked given such an essay by using NLP and IR techniques. As essays and descriptions of professions display a mismatch between the vocabularies of topics and documents and there is lack of contextual information, due to the documents being fairly short as compared to standard IR scenarios, lexical semantic information should be especially beneficial to an IR system. For example, the profile can contain words about some objects or activities related to the profession, but not directly mentioned in the description, e.g. *oven*, *cakes* in the profile and *pastries*, *baker*, or *confectioner* in the document. Therefore, we propose to utilize semantic relatedness as a ranking function instead of conventional IR techniques, as will be substantiated below.

2 System Architecture

Integrating lexical semantic knowledge in ECG requires the existence of knowledge bases encoding domain and lexical knowledge. In this paper, we investigate the utility of two knowledge bases: (i) a German wordnet, GermaNet (Kunze, 2004), and (ii) the German portion of Wikipedia.[3] A large body of research exists on using wordnets in NLP applications and in particular in IR (Moldovan and Mihalcea, 2000). The knowledge in wordnets has been typically utilized by expanding queries with related terms (Vorhees, 1994; Smeaton et al., 1994), concept indexing (Gonzalo et al., 1998), or similarity measures as ranking functions (Smeaton et al., 1994; Müller and Gurevych, 2006). Recently, Wikipedia has been discovered as a promising lexical semantic resource and successfully used in such different NLP tasks as question answering (Ahn et al., 2004), named entity disambiguation (Bunescu and Pasca, 2006), and information retrieval (Katz et al., 2005). Further research (Zesch et al., 2007b) indicates that German wordnet and Wikipedia show different performance depending on the task at hand.

Departing from this, we first compare two semantic relatedness (SR) measures based on the information either in the German wordnet (Lin, 1998) called **LIN**, or in Wikipedia (Gabrilovich and Markovitch, 2007) called Explicit Semantic Analysis, or **ESA**. We evaluate their performance intrinsically on the tasks of (T-1) computing semantic relatedness, and (T-2) solving Reader's Digest Word Power (RDWP) questions and make conclusions about the ability of the measures to model certain aspects of semantic relatedness and their coverage. Furthermore, we follow the approach by Müller and Gurevych (2006), who proposed to utilize the LIN measure and a set of heuristics as an IR model (IR-1).

Additionally, we utilize the ESA measure in a semantic information retrieval model, as this measure is significantly better at vocabulary coverage and at modelling cross part-of-speech relations (Gabrilovich and Markovitch, 2007). We compare the performance of ESA and LIN measures in a task-based IR evaluation and analyze their strengths and limitations. Finally, we apply ESA to directly compute text similarities between topics and documents (IR-2) and compare the performance of two semantic IR models and a baseline Extended Boolean (EB) model (Salton et al., 1983) with query expansion.[4]

To summarize, the *contributions of this paper* are three-fold: (i) we present a novel system, utilizing NLP and IR techniques to perform Electronic Career Guidance, (ii) we study the properties and intrinsically evaluate two SR measures based on GermaNet and Wikipedia for the tasks of computing semantic relatedness and solving Reader's Digest Word Power Game questions, and (iii) we investigate the performance of two semantic IR models in a task based evaluation.

[2] http://infobub.arbeitsagentur.de/berufe/

[3] http://de.wikipedia.org/

[4] We also ran experiments with Okapi BM25 model as implemented in the Terrier framework, but the results were worse than those with the EB model. Therefore, we limit our discussion to the latter.

3 Computing Semantic Relatedness

3.1 SR Measures

GermaNet based measures GermaNet is a German wordnet, which adopted the major properties and database technology from Princeton's WordNet (Fellbaum, 1998). However, GermaNet displays some structural differences and content oriented modifications. Its designers relied mainly on linguistic evidence, such as corpus frequency, rather than psycholinguistic motivations. Also, GermaNet employs 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 modelling nouns, verbs and adjectives.

We use the semantic relatedness measure by Lin (1998) (referred to as LIN), as it consistently is among the best performing wordnet based measures (Gurevych and Niederlich, 2005; Budanitsky and Hirst, 2006). Lin defined semantic similarity using a formula derived from information theory. This measure is sometimes called a universal semantic similarity measure as it is supposed to be application, domain, and resource independent. *Lin* is computed as:

$$sim_{c_1,c_2} = \frac{2 \times \log p(LCS(c_1, c_2))}{\log p(c_1) + \log p(c_2)}$$

where c_1 and c_2 are concepts (word senses) corresponding to w_1 and w_2, $\log p(c)$ is the information content, and $LCS(c_1, c_2)$ is the lowest common subsumer of the two concepts. The probability p is computed as the relative frequency of words (representing that concept) in the taz[5] corpus.

Wikipedia based measures Wikipedia is a free online encyclopedia that is constructed in a collaborative effort of voluntary contributors and still grows exponentially. During this process, Wikipedia has probably become the largest collection of freely available knowledge. Wikipedia shares many of its properties with other well known lexical semantic resources (like dictionaries, thesauri, semantic wordnets or conventional encyclopedias) (Zesch et al., 2007a). As Wikipedia also models relatedness between concepts, it is better suited for computing semantic relatedness than GermaNet (Zesch et al., 2007b).

In very recent work, Gabrilovich and Markovitch (2007) introduce a SR measure called *Explicit Semantic Analysis* (ESA). The ESA measure represents the meaning of a term as a high-dimensional concept vector. The concept vector is derived from Wikipedia articles, as each article focuses on a certain topic, and can thus be viewed as expressing a concept. The dimension of the concept vector is the number of Wikipedia articles. Each element of the vector is associated with a certain Wikipedia article (or concept). If the term can be found in this article, the term's tfidf score (Salton and McGill, 1983) in this article is assigned to the vector element. Otherwise, 0 is assigned. As a result, a term's concept vector represents the importance of the term for each concept. Semantic relatedness of two terms can then be easily computed as the cosine of their corresponding concept vectors. If we want to measure the semantic relatedness of texts instead of terms, we can also use ESA concept vectors. A text is represented as the average concept vector of its terms' concept vectors. Then, the relatedness of two texts is computed as the cosine of their average concept vectors.

As ESA uses all textual information in Wikipedia, the measure shows excellent coverage. Therefore, we select it as the second measure for integration into our IR system.

3.2 Datasets

Semantic relatedness datasets for German employed in our study are presented in Table 1. Gurevych (2005) conducted experiments with two datasets: i) a German translation of the English dataset by Rubenstein and Goodenough (1965) (**Gur65**), and ii) a larger dataset containing 350 word pairs (**Gur350**). Zesch and Gurevych (2006) created a third dataset from domain-specific corpora using a semi-automatic process (**ZG222**). Gur65 is rather small and contains only noun-noun pairs connected by either synonymy or hypernymy. Gur350 contains nouns, verbs and adjectives that are connected by classical and non-classical relations (Morris and Hirst, 2004). However, word pairs for this dataset are biased towards strong classical relations, as they were manually selected from a corpus.

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

DATASET	YEAR	LANGUAGE	# PAIRS	POS	SCORES	# SUBJECTS	CORRELATION r INTER	CORRELATION r INTRA
Gur65	2005	German	65	N	discrete {0,1,2,3,4}	24	.810	-
Gur350	2006	German	350	N, V, A	discrete {0,1,2,3,4}	8	.690	-
ZG222	2006	German	222	N, V, A	discrete {0,1,2,3,4}	21	.490	.647

Table 1: Comparison of datasets used for evaluating semantic relatedness in German.

ZG222 does not have this bias.

Following the work by Jarmasz and Szpakowicz (2003) and Turney (2006), we created a second dataset containing multiple choice questions. We collected 1072 multiple-choice word analogy questions from the German **Reader's Digest Word Power Game** (RDWP) from January 2001 to December 2005 (Wallace and Wallace, 2005). We discarded 44 questions that had more than one correct answer, and 20 questions that used a phrase instead of a single term as query. The resulting 1008 questions form our evaluation dataset. An example question is given below:

Muffin (**muffin**)
a) Kleingebäck (small cake)
b) Spenglerwerkzeug (plumbing tool)
c) Miesepeter (killjoy)
d) Wildschaf (moufflon)

The task is to find the correct choice - 'a)' in this case.

This dataset is significantly larger than any of the previous datasets employed in this type of evaluation. Also, it is not restricted to synonym questions, as in the work by Jarmasz and Szpakowicz (2003), but also includes hypernymy/hyponymy, and few non-classical relations.

3.3 Analysis of Results

Table 2 gives the results of evaluation on the task of correlating the results of an SR measure with human judgments using Pearson correlation. The GermaNet based LIN measure outperforms ESA on the Gur65 dataset. On the other datasets, ESA is better than LIN. This is clearly due to the fact, that Gur65 contains only noun-noun word pairs connected by classical semantic relations, while the other datasets also contain cross part-of-speech pairs connected by non-classical relations. The Wikipedia based ESA measure can better capture such relations. Additionally, Table 3 shows that ESA also covers almost all word pairs in each dataset, while GermaNet is much lower for Gur350 and ZG222. ESA performs even better on the Reader's Digest task (see Table 4). It shows high coverage and near human performance regarding the relative number of correctly solved questions.[6] Given the high performance and coverage of the Wikipedia based ESA measure, we expect it to yield better IR results than LIN.

		GUR65	GUR350	ZG222
# covered word pairs		53	116	55
Upper bound		0.80	0.64	0.44
GermaNet	Lin	0.73	0.50	0.08
Wikipedia	ESA	0.56	0.52	0.32

Table 2: Pearson correlation r of human judgments with SR measures on word pairs covered by GermaNet and Wikipedia.

DATASET	# PAIRS	COVERED PAIRS LIN	COVERED PAIRS ESA
Gur65	65	60	65
Gur350	350	208	333
ZG222	222	88	205

Table 3: Number of covered word pairs based on Lin or ESA measure on different datasets.

4 Information Retrieval

4.1 IR Models

Preprocessing For creating the search index for IR models, we apply first tokenization and then remove stop words. We use a general German stop

[6] Values for human performance are for one subject. Thus, they only indicate the approximate difficulty of the task. We plan to use this dataset with a much larger group of subjects.

		#ANSWERED	#CORRECT	RATIO
Human		1008	874	0.87
GermaNet	Lin	298	153	0.51
Wikipedia	ESA	789	572	0.72

Table 4: Evaluation results on multiple-choice word analogy questions.

word list extended with highly frequent domain specific terms. Before adding the remaining words to the index, they are lemmatized. We finally split compounds into their constituents, and add both, constituents and compounds, to the index.

EB model Lucene[7] is an open source text search library based on an EB model. After matching the preprocessed queries against the index, the document collection is divided into a set of relevant and irrelevant documents. The set of relevant documents is, then, ranked according to the formula given in the following equation:

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

where n_q is the number of terms in the query, $tf(t_q,d)$ is the term frequency factor for term t_q in document d, $idf(t_q)$ is the inverse document frequency of the term, and $lengthNorm(d)$ is a normalization value of document d, given the number of terms within the document. We added a simple query expansion algorithm using (i) synonyms, and (ii) hyponyms, extracted from GermaNet.

IR based on SR For the (IR-1) model, we utilize two SR measures and a set of heuristics: (i) the Lin measure based on GermaNet (LIN), and (ii) the ESA measure based on Wikipedia (ESA-Word). This algorithm was applied to the German IR benchmark with positive results by Müller and Gurevych (2006). The algorithm computes a SR score for each query and document term pair. 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. 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.

For the (IR-2) model, we apply the ESA method for directly comparing the query with documents, as described in Section 3.1.

4.2 Data

The corpus employed in our experiments was built based on a real-life IR scenario in the domain of ECG, as described in Section 1. The **document collection** is extracted from BERUFEnet,[8] a database created by the GFLO. It contains textual descriptions of about 1,800 vocational trainings, and 4,000 descriptions of professions. We restrict the collection to a subset of BERUFEnet documents, consisting of 529 descriptions of vocational trainings, due to the process necessary to obtain relevance judgments, as described below. The documents contain not only details of professions, but also a lot of information concerning the training and administrative issues. We only use those portions of the descriptions, which characterize the profession itself.

We collected real natural language **topics** by asking 30 human subjects to write an essay about their professional interests. The topics contain 130 words, on average. Making **relevance judgments** for ECG requires domain expertise. Therefore, we applied an automatic method, which uses the knowledge base employed by the GFLO, described in Section 1. To obtain relevance judgments, 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. To map the ranked list to a set of relevant and irrelevant professions, we use a threshold of 3, as suggested by career guidance experts. This setting 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

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

[8] `http://berufenet.arbeitsamt.de/`

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.

4.3 Analysis of Results

In Table 5, we summarize the results of the experiments applying different IR models on the BERUFEnet data. We build queries from natural language essays by (QT-1) extracting nouns, verbs, and adjectives, (QT-2) using only nouns, and (QT-3) manually assigning suitable keywords from the tagset with 41 keywords to each topic. We report the results with two different thresholds (.85 and .98) for the Lin model, and with three different thresholds (.11, .13 and .24) for the ESA-Word models. The evaluation metrics used are *mean average precision* (MAP), *precision after ten documents* (P10), *the number of relevant returned documents* (#RRD). We compute the absolute value of *Spearman's rank correlation coefficient* (SRCC) by comparing the relevance ranking of our system with the relevance ranking of the knowledge base employed by the GFLO.

Using query expansion for the EB model decreases the retrieval performance for most configurations. The SR based models outperform the EB model in all configurations and evaluation metrics, except for P10 on the keyword based queries. The Lin model is always outperformed by at least one of the ESA models, except for (QT-3). (IR-2) performs best on longer queries using nouns, verbs, adjectives or just nouns.

Comparing the number of relevant retrieved documents, we observe that the IR models based on SR are able to return more relevant documents than the EB model. This supports the claim that semantic knowledge is especially helpful for the vocabulary mismatch problem, which cannot be addressed by conventional IR models. E.g., only SR-based models can find the job *information technician* for a profile which contains the sentence *My interests and skills are in the field of languages and IT*. The job could only be judged as relevant, as the semantic relation between *IT* in the profile and *information technology* in the professional description could be found.

In our analysis of the BERUFEnet results obtained on (QT-1), we noticed that many errors were due to the topics expressed in free natural language essays. Some subjects deviated from the given task to describe their professional interests and described facts that are rather irrelevant to the task of ECG, e.g. *It is important to speak different languages in the growing European Union*. If all content words are extracted to build a query, a lot of noise is introduced.

Therefore, we conducted further experiments with (QT-2) and (QT-3): building the query using only nouns, and using manually assigned keywords based on the tagset of 41 keywords. For example, the following query is built for the professional profile given in Section 1.

```
Keywords assigned:
care for/nurse/educate/teach; use/program computer;
office; outside:  outside facilities/natural
environment; animals/plants
```

IR results obtained on (QT-2) and (QT-3) show that the performance is better for nouns, and significantly better for the queries built of keywords. This suggests that in order to achieve high IR performance for the task of Electronic Career Guidance, it is necessary to preprocess the topics by performing information extraction to remove the noise from free text essays. As a result of the preprocessing, natural language essays should be mapped to a set of keywords relevant for describing a person's interests. Our results suggest that the word-based semantic relatedness IR model (IR-1) performs significantly better in this setting.

5 Conclusions

We presented a system for Electronic Career Guidance utilizing NLP and IR techniques. Given a natural language professional profile, relevant professions are computed based on the information about semantic relatedness. We intrinsically evaluated and analyzed the properties of two semantic relatedness measures utilizing the lexical semantic information in a German wordnet and Wikipedia on the tasks of estimating semantic relatedness scores and answering multiple-choice questions. Furthermore, we applied these measures to an IR task, whereby they were used either in combination with a set of heuristics or the Wikipedia based measure was used to directly compute semantic relatedness of topics and

Model	(QT-1) Nouns, Verbs, Adj.				(QT-2) Nouns				(QT-3) Keywords			
	MAP	P10	#RRD	SRCC	MAP	P10	#RRD	SRCC	MAP	P10	#RRD	SRCC
EB	.39	.58	2581	.306	.38	.58	2297	.335	.54	.76	2755	.497
EB+SYN	.37	.56	2589	.288	.38	.57	2310	.331	.54	.73	2768	.530
EB+HYPO	.34	.47	2702	.275	.38	.56	2328	.327	.47	.65	2782	.399
Lin .85	.41	.56	**2787**	.338	.40	.59	2770	.320	.59	.73	**2787**	**.578**
Lin .98	.41	.61	2753	.326	.42	.59	2677	.341	.58	.74	2783	.563
ESA-Word .11	.39	.56	**2787**	.309	.44	.63	**2787**	.355	.60	**.77**	**2787**	.535
ESA-Word .13	.38	.59	**2787**	.282	.43	.62	**2787**	.338	**.62**	.76	**2787**	**.550**
ESA-Word .24	.40	.60	**2787**	.259	.43	.60	2699	.306	.54	.73	2772	.482
ESA-Text	**.47**	**.62**	**2787**	**.368**	**.55**	**.71**	**2787**	**.462**	.56	.74	**2787**	.489

Table 5: Information Retrieval performance on the BERUFEnet dataset.

documents. We experimented with three different query types, which were built from the topics by: (QT-1) extracting nouns, verbs, adjectives, (QT-2) extracting only nouns, or (QT-3) manually assigning several keywords to each topic from a tagset of 41 keywords.

In an intrinsic evaluation of LIN and ESA measures on the task of computing semantic relatedness, we found that ESA captures the information about semantic relatedness and non-classical semantic relations considerably better than LIN, which operates on an *is-a* hierarchy and, thus, better captures the information about semantic similarity. On the task of solving RDWP questions, the ESA measure significantly outperformed the LIN measure in terms of correctness. On both tasks, the coverage of ESA is much better. Despite this, the performance of LIN and ESA as part of an IR model is only slightly different. ESA performs better for all lengths of queries, but the differences are not as significant as in the intrinsic evaluation. This indicates that the information provided by both measures, based on different knowledge bases, might be complementary for the IR task.

When ESA is applied to directly compute semantic relatedness between topics and documents, it outperforms IR-1 and the baseline EB model by a large margin for QT-1 and QT-2 queries. For QT-3, i.e., the shortest type of query, it performs worse than IR-1 utilizing ESA and a set of heuristics. Also, the performance of the baseline EB model is very strong in this experimental setting. This result indicates that IR-2 utilizing conventional information retrieval techniques and semantic information from Wikipedia is better suited for longer queries providing enough context. For shorter queries, *soft* matching techniques utilizing semantic relatedness tend to be beneficial.

It should be born in mind, that the construction of QT-3 queries involved a manual step of assigning the keywords to a given essay. In this experimental setting, all models show the best performance. This indicates that professional profiles contain a lot of noise, so that more sophisticated NLP analysis of topics is required. This will be improved in our future work, whereby the system will incorporate an information extraction component for automatically mapping the professional profile to a set of keywords. We will also integrate a component for analyzing the sentiment structure of the profiles. We believe that the findings from our work on applying IR techniques to the task of Electronic Career Guidance generalize to similar application domains, where topics and documents display similar properties (with respect to their length, free-text structure and mismatch of vocabularies) and domain and lexical knowledge is required to achieve high levels of performance.

Acknowledgments

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. We would like to thank the anonymous reviewers for valuable feedback on this paper. We would also like to thank Piklu Gupta for helpful comments.

References

David Ahn, Valentin Jijkoun, Gilad Mishne, Karin Müller, Maarten de Rijke, and Stefan Schlobach. 2004. Using Wikipedia at the TREC QA Track. In *Proceedings of TREC 2004*.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Semantic Distance. *Computational Linguistics*, 32(1).

Razvan Bunescu and Marius Pasca. 2006. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of ACL*, pages 9–16, Trento, Italy.

Christiane Fellbaum. 1998. *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, January.

Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran. 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the Coling-ACL '98 Workshop Usage of WordNet in Natural Language Processing Systems,* Montreal, Canada, August.

Iryna Gurevych and Hendrik Niederlich. 2005. Computing semantic relatedness in german with revised information content metrics. In *Proceedings of "OntoLex 2005 - Ontologies and Lexical Resources" IJCNLP'05 Workshop*, pages 28–33, October 11 – 13.

Iryna Gurevych. 2005. Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pages 767–778, Jeju Island, Republic of Korea.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *RANLP*, pages 111–120.

Boris Katz, Gregory Marton, Gary Borchardt, Alexis Brownell, Sue Felshin, Daniel Loreto, Jesse Louis-Rosenberg, Ben Lu, Federico Mora, Stephan Stiller, Ozlem Uzuner, and Angela Wilcox. 2005. External knowledge sources for question answering. In *Proceedings of the 14th Annual Text REtrieval Conference (TREC'2005)*, November.

Claudia Kunze, 2004. *Lexikalisch-semantische Wortnetze*, chapter Computerlinguistik und Sprachtechnologie, pages 423–431. Spektrum Akademischer Verlag.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Dan Moldovan and Rada Mihalcea. 2000. Using WordNet and lexical operators to improve Internet searches. *IEEE Internet Computing*, 4(1):34–43.

Jane Morris and Graeme Hirst. 2004. Non-Classical Lexical Semantic Relations. In *Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the ACL*, Boston.

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

Herbert Rubenstein and John B. Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Gerard Salton, Edward Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Communication of the ACM*, 26(11):1022–1036.

Alan F. Smeaton, Fergus Kelledy, and Ruari O'Donell. 1994. TREC-4 Experiments at Dublin City University: Thresholding posting lists, query expansion with WordNet and POS tagging of Spanish. In *Proceedings of TREC-4*, pages 373–390.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Ellen Vorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69.

DeWitt Wallace and Lila Acheson Wallace. 2005. *Reader's Digest, das Beste für Deutschland*. Jan 2001–Dec 2005. Verlag Das Beste, Stuttgart.

Torsten Zesch and Iryna Gurevych. 2006. Automatically Creating Datasets for Measures of Semantic Relatedness. In *Proceedings of the Workshop on Linguistic Distances*, pages 16–24, Sydney, Australia, July. Association for Computational Linguistics.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007a. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In *Biannual Conference of the Society for Computational Linguistics and Language Technology*, pages 213–221, Tuebingen, Germany.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007b. Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In *Proceedings of NAACL-HLT*.