Using Wiktionary for Computing Semantic Relatedness

Torsten Zesch and Christof Müller and Iryna Gurevych
Ubiquitous Knowledge Processing Lab
Computer Science Department
Technische Universität Darmstadt, Hochschulstraße 10
D-64289 Darmstadt, Germany
{zesch,mueller,gurevych} (at) tk.informatik.tu-darmstadt.de

Abstract

We introduce Wiktionary as an emerging lexical semantic resource that can be used as a substitute for expert-made resources in AI applications. We evaluate Wiktionary on the pervasive task of computing semantic relatedness for English and German by means of correlation with human rankings and solving word choice problems. For the first time, we apply a concept vector based measure to a set of different concept representations like Wiktionary pseudo glosses, the first paragraph of Wikipedia articles, English WordNet glosses, and GermaNet pseudo glosses. We show that: (i) Wiktionary is the best lexical semantic resource in the ranking task and performs comparably to other resources in the word choice task, and (ii) the concept vector based approach yields the best results on all datasets in both evaluations.

Introduction

Many natural language processing (NLP) tasks require external sources of lexical semantic knowledge such as WordNet (Fellbaum 1998). Traditionally, these resources have been built manually by experts in a time consuming and expensive manner. Recently, emerging Web 2.0 technologies have enabled user communities to collaboratively create new kinds of resources. One such emerging resource is Wiktionary, a freely available, web-based multilingual dictionary. It has been previously applied in NLP research for sentiment classification (Chesley et al. 2006) and diachronic phonology (Bouchard et al. 2007), but has not yet been considered as a substitute for expert-made resources.

In this paper, we systematically study the applicability of Wiktionary as a lexical semantic resource for AI applications by employing it to compute *semantic relatedness* (**SR** henceforth) which is a pervasive task with applications in word sense disambiguation (Patwardhan and Pedersen 2006), semantic information retrieval (Gurevych, Müller, and Zesch 2007), or information extraction (Stevenson and Greenwood 2005).

We use two SR measures that are applicable to all lexical semantic resources in this study: Path length based measures (Rada et al. 1989) and concept vector based measures (Qiu and Frei 1993). So far, only Wikipedia has been applied as a lexical semantic resource in a concept vector based approach (Gabrilovich and Markovitch 2007). We generalize this approach by using concept representations like Wiktionary pseudo glosses, the first paragraph of Wikipedia articles, English WordNet glosses, and GermaNet pseudo glosses. Additionally, we study the effect of using shorter but more precise textual representations by considering only the first paragraph of a Wikipedia article instead of the full article text.

We compare the performance of Wiktionary with expert-made wordnets, like Princeton WordNet and GermaNet (Kunze 2004), and with Wikipedia as another collaboratively constructed resource. In order to study the effects of the coverage of lexical semantic resources, we conduct our experiments on English and German datasets, as Wiktionary offers substantially higher coverage for English than for German.

Wiktionary

Wiktionary[1] is a multilingual, web-based, freely available *dictionary*, *thesaurus* and *phrase book*. Although expert-made dictionaries or wordnets have been used in NLP for a long time (Wilks et al. 1990; Leacock and Chodorow 1998), the collaboratively constructed Wiktionary differs considerably from them. In this paper, we focus on the differences that are most relevant with respect to Wiktionary's applicability in AI. We refer to (Zesch, Müller, and Gurevych 2008) for a more detailed comparison.

Relation types Wiktionary shows many commonalities with expert-made resources, since they all contain concepts, which are connected by lexical semantic relations, and described by a gloss giving a short definition. Some common relation types can be found in most resources, e.g. synonymy, hypernymy, or antonymy, whereas others are specific to a knowledge resource, e.g. etymology, translations, and quotations in Wiktionary, or evocation in WordNet (Boyd-Graber et al. 2006).

Languages Wiktionary is a *multilingual* resource available for many languages where corresponding entries are linked between different Wiktionary language editions. Each language edition comprises a multilingual dictionary with a substantial amount of entries in different languages. For example, the English Wiktionary currently contains ap-

[1] http://www.wiktionary.org

prox 10,000 entries about German words (e.g. the German term "Haus" is explained in English as meaning "house"). Expert-made resources are usually designed for a specific language, though the EuroWordNet framework (Vossen 1998) allows to interconnect wordnets by linking them to an Inter-Lingual-Index, based on Princeton WordNet.

Size The size of a particular language edition of Wiktionary largely depends on how active the particular community is. The largest language edition is French (730,193 entries) closely followed by English (682,982 entries).[2] Other major languages like German (71,399 entries) or Spanish (31,652 entries) are not among the ten largest language editions. As each language edition is a multilingual dictionary, not all entries are about words in the target language. Out of the 682,982 entries in the English Wiktionary only about 175,000 refer to English words. However, this still exceeds the size of WordNet 3.0 which contains about 150,000 words. In contrast, the German Wiktionary edition only contains about 20,000 German words compared to about 70,000 lexical units in GermaNet 5.0.

Instance structure Wiktionary allows to easily create, edit, and link HTML pages on the web using a simple markup language. For most language editions, the user community has introduced a layout standard acting as a data schema to enforce a uniform structure of the entries. As schemas evolve over time, older entries are possibly not updated. Moreover, as no contributor is forced to follow the schema, the structure of entries is fairly inconsistent. Additionally, schemas are specific to each language edition. Layout decisions for expert-made wordnets are made in the beginning and changed only with caution afterwards. The compliance of entries with the layout decisions is enforced.

Instance incompleteness Even if a Wiktionary entry follows the schema posed by a layout standard, the entry might be a stub, where most relation types are empty. Wiktionary also does not include any mechanism to enforce symmetrically defined relations (e.g. synonymy) to hold in both directions. Instance incompleteness is not a major concern for expert-made wordnets as new entries are usually entered along with all relevant relation types.

Quality In contrast to incompleteness and inconsistency described above, *quality* refers to the correctness of the encoded information itself. To our knowledge, there are no studies on the quality of the information in Wiktionary. However, the collaborative construction approach has been argued to yield remarkable factual quality in Wikipedia (Giles 2005), and the quality of expert-made resources like WordNet has also been target of criticism (Kaplan and Schubert 2001).

Experimental Setup

Semantic Relatedness Measures

A multitude of SR measures has been introduced in the literature (refer to Budanitsky and Hirst (2006) for an overview). As we aim to evaluate the SR measures on a wide range of lexical semantic resources with quite different properties, the selection of measures is restricted to those which are applicable to WordNet, GermaNet, Wikipedia, and Wiktionary. In particular, this precludes most SR measures that rely on a lowest common subsumer (Jiang and Conrath 1997) or need to determine the depth of a resource (Leacock and Chodorow 1998). Thus, we chose (i) a path based approach (Rada et al. 1989) as it can be utilized with any resource containing concepts connected by lexical semantic relations, and (ii) a concept vector based approach (Qiu and Frei 1993). The latter approach has already been applied to Wikipedia (Gabrilovich and Markovitch 2007) showing excellent results. We generalize this approach to work on each resource which offers a textual representation of a concept as shown below.

The path length (**PL** henceforth) based measure determines the length of a path between nodes representing concepts c_i in a lexical semantic resource. The resource is treated as a graph, where the nodes represent concepts and edges are established due to lexical semantic relations between the concepts. The measure is formalized as $rel_{\text{PL}}(c_1, c_2) = l_{max} - l(c_1, c_2)$, where l_{max} is the length of the longest non-cyclic path in the graph, and $l(c_1, c_2)$ returns the number of edges on the path from concept c_1 to c_2.

In a concept vector (**CV** henceforth) based measure, the meaning of a word w is represented as a high dimensional concept vector $\vec{v}(w) = (v_1, \ldots, v_n)$, where n is the number of documents.[3] The value of v_i depends on the occurrence of the word w in the document d_i. If the word w can be found in the document, the word's tf.idf score (Spärck Jones 1972) in the document d_i is assigned to the CV element v_i. Otherwise, v_i is 0. As a result, the vector $\vec{v}(w)$ represents the word w in a concept space. The SR of two words can then be computed as the cosine of their concept vectors.

Gabrilovich and Markovitch (2007) have applied the CV based approach using Wikipedia articles to create the concept space. However, CV based measures can be applied to any lexical semantic resource that offers a textual representation of a concept. We adapt the approach to WordNet by using glosses and example sentences as short documents representing a concept. As GermaNet does not contain glosses for most entries, we create pseudo glosses by concatenating concepts that are in close relation (synonymy, hypernymy, meronymy, etc.) to the original concept as proposed by Gurevych (2005). This is based on the observation that most content words in glosses are in close lexical semantic relation to the described concept. For example, the pseudo gloss for the concept *tree (plant)* would be *"woody plant, ligneous plant, tree stump, crown, tree branch, trunk, ..."* showing a high overlap with its WordNet gloss *"a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown"*. Wiktionary also does not contain glosses for each entry due to instance incompleteness. Therefore, we construct pseudo glosses similarly to the approach used for GermaNet by concatenating all information that is available for a Wiktionary entry. To our

[2] As of February 29, 2008.

[3] *Concept vector* based approaches (Qiu and Frei 1993) represent a word in a document vector space, while *context vector* based approaches (Patwardhan and Pedersen 2006) represents a word in a word vector space relying on word co-occurrence counts.

knowledge, we are the first to employ a CV based measure to compute SR using WordNet, GermaNet, and Wiktionary as lexical semantic resources.

Configuration of SR Measures

WordNet We use WordNet 3.0 and the measures as available in the WordNet::Similarity package (Patwardhan and Pedersen 2006). For constructing the concept vectors, we treat each WordNet synset as a concept, and its gloss (together with the example sentences) as the concept's textual representation. We access WordNet glosses using the JWNL WordNet API.[4]

GermaNet We have adapted the PL and CV measures using the freely available GermaNet API[5] applied to GermaNet 5.0. As GermaNet does not contain glosses, we construct pseudo glosses by concatenating the lemmas of all concepts that are reachable within a radius of three from the original concept.[6]

Wikipedia We use the freely available JWPL Wikipedia API (Zesch, Müller, and Gurevych 2008) to access the English and German Wikipedia dumps from February 6th, 2007. When adapting the CV measure, we differentiate between considering the full Wikipedia article or only the first paragraph as the concept's textual representation. The first paragraph usually contains a definition of the concept described in the article. As some words in the latter parts of an article are likely to describe less important or even contradictory topics, we expect a CV based measure that uses only the first paragraph to yield a better precision by trading in some recall. We abbreviate this measure as CV_{first}.

When using the full Wikipedia articles, we prune the concept space for performance reasons by only considering articles as concepts if they contain at least 100 words and have more than 5 inlinks and 5 outlinks.[7] In the experiments with the PL measure operating on the English Wikipedia, we limit the search for a shortest path to 5 edges for performance reasons.

Wiktionary For accessing Wiktionary, we have created a Java-based API (Zesch, Müller, and Gurevych 2008) called JWKTL that is based on Wiktionary dumps which are freely available for each language edition.[8] We have used the English dump from Oct 16, 2007 and the German dump from Oct 9, 2007. Textual representations for the CV based measure are created by concatenating the contents of all relation types offered by JWKTL for each Wiktionary entry.[9]

Evaluation

The prevalent approaches for evaluating SR measures are: (i) application-specific evaluation (Budanitsky and Hirst 2006), (ii) correlation with human rankings (Gabrilovich and Markovitch 2007; Pedersen et al. 2007), and (iii) solving word choice problems (Mohammad et al. 2007). An application-specific evaluation tests a measure within the framework of a usually complex application, which entails influence of parameters besides the SR measure being tested. Thus, we will only use the remaining approaches, correlation with human rankings and solving word choice problems. We conduct experiments on English and German datasets, in order to study the impact of Wiktionary's coverage on the performance.

Correlation with Human Rankings

This task evaluates the ability of a SR measure to rank a set of word pairs according to human judgments of SR. Evaluation datasets for this task are created by asking human annotators to judge the SR of presented word pairs. The resulting dataset is then correlated with the ranking produced on the basis of a particular measure using Spearman's rank correlation coefficient ρ (Siegel and Castellan 1988), where 0 means no correlation and 1 stands for perfect correlation.

We now give an overview of the evaluation datasets employed in our study. Rubenstein and Goodenough (1965) created a dataset with 65 English noun pairs (**RG–65** for short). A subset of this dataset has been used for experiments by Miller and Charles (1991) (**MC–30**). Finkelstein et al. (2002) created a larger dataset for English containing 353 word pairs. However, this dataset consists of two subsets, which have been annotated by different annotators, and have a different inter-annotator agreement. Therefore, we treat them as independent datasets, **Fin1–153** and **Fin2–200** henceforth. Yang and Powers (2006) created a dataset (**YP–130**) containing 130 verb pairs that will be particularly informative with respect to the ability of a SR measure to estimate verb relatedness. Gurevych (2005) conducted experiments with a German translation (**Gur–65**) of the English RG–65 dataset, and a larger dataset (**Gur–350**) containing 350 word pairs containing nouns, verbs and adjectives that are connected by classical and non-classical relations (Morris and Hirst 2004).

If a term from a word pair in these datasets can be found in a lexical semantic resource, it is said to be *covered* by the resource. Hence, we define the coverage of a resource as the percentage of word pairs in a dataset where both terms are covered.

Results Table 1 displays the results obtained for this task. It shows that Wiktionary outperforms all other lexical semantic resources except for the English verb dataset YP130, where WordNet yields slightly better results. However, the differences are not statistically significant.[10]

When analyzing coverage, we find that all *English* lexical semantic resources including Wiktionary cover the datasets almost perfectly (coverage ranging from 98% to 100%). Hence, we only report detailed results on the German datasets in Table 2. We find that the coverage of German

[4] http://jwordnet.sourceforge.net/

[5] http://projects.villa-bosch.de/nlpsoft/gn_api/

[6] Optimized configuration as reported by Gurevych (2005).

[7] Same configuration as in (Gabrilovich and Markovitch 2007).

[8] http://dumps.wikimedia.org/

[9] Used relation types: Antonyms, categories, characteristic word combinations, coordinate terms, derived terms, examples, glosses, holonyms, hypernyms, hyponyms, meronyms, "see also" terms, synonyms, and troponyms.

[10] Fisher-Z transformation; two-tailed t-test with $\alpha = .05$

Dataset		English					German	
		MC–30	**RG–65**	**Fin1–153**	**Fin2–200**	**YP–130**	**Gur–65**	**Gur–350**
Word pairs used		30	65	144	190	80	52	131
WN / GN	PL	.77	.82	.44	.38	**.71**	.77	.40
	CV	.78	.82	.61	.56	**.71**	.80	.59
Wikipedia	PL	.30*	.36	.43	.30	.01*	.50	.39
	CV_{first}	.68	.76	**.70**	.50	.29	.40	.62
	CV	.67	.69	.62	.31	.28	.65	.65
Wiktionary	PL	.54	.68	.52	.04*	.37	-	-
	CV	**.84**	**.84**	**.70**	**.60**	.65	**.83**	**.67**

Table 1: Spearman's rank correlation coefficients on English and German datasets. Best values for each dataset are in bold. Non-significant values are marked with a '*' (two tailed t-test, $\alpha = .05$). We only used the subset of word pairs covered by all resources to ensure a fair comparison.

Dataset		**Gur–65**	**Gur–350**
GN	PL	0.88	0.70
	CV	0.88	0.70
Wikipedia	PL	0.94	0.52
	CV_{first}	**1.00**	0.81
	CV	**1.00**	**0.96**
Wiktionary	PL	0.42	0.33
	CV	**1.00**	0.73

Table 2: Coverage of resources on German datasets.

Wiktionary exceeds that of GermaNet even though Wiktionary has much less German word entries than GermaNet (cf. section *Wiktionary*). This is due to the CV based measure using additional information drawn from glosses, derived terms, characteristic word combinations, etc. that can be found in Wiktionary, but not in GermaNet. These results show that Wiktionary can substitute expert-made lexical semantic resources with respect to coverage and the performance on the task of computing SR.

When analyzing the correlation with respect to the measure type, we find that the CV based measures outperform the PL based measures consistently over all lexical semantic resources and most datasets. The performance gains for the datasets (Fin1–153, Fin2–200 and Gur–350) which contain also word pairs connected by non-classical lexical semantic relations are generally higher.[11] Thus, the CV based measure appears to be better suited to estimate non-classical relationships between concepts. Moreover, the performance gains of the CV based measures over the PL based measures are higher when operating on collaboratively constructed resources. This was to be expected as the amount of additional information that the CV based measures can use is significantly higher in Wiktionary (additional relation types) and Wikipedia (long article texts), than in WordNet or GermaNet.

When comparing our results with previously obtained values, we find that Patwardhan and Pedersen (2006) report slightly higher values on the MC–30 and RG–65 datasets (.91 and .90), but the difference to our best results is not statistically significant. Gabrilovich and Markovitch (2007) report a correlation of σ=.75 on a dataset consisting of Fin1–153 and Fin2–200. However, we obtained only σ=.62 and σ=.31 on the two subsets using a more recent Wikipedia version and a reimplementation of their method. Yang and Powers (2006) report a Pearson correlation coefficient of r=.84 for their YP–130 dataset that cannot be directly compared to Spearman's rank correlation coefficient reported in this paper.

[11] Statistically significant at the $\alpha = .05$ level.

Solving Word Choice Problems

A different approach to evaluate the performance of SR measures relies on word choice problems consisting of a target word and four candidate words or phrases (Jarmasz and Szpakowicz 2003). The objective is to pick the one that is most closely related to the target. An example problem is given below, the correct choice is 'a)' in this case.

beret
a) round cap b) cap with horizontal peak
c) wedge cap d) helmet

We run experiments on a dataset for English and German each. The English dataset contains 300 word choice problems collected by Jarmasz and Szpakowicz (2003). The German dataset contains 1,008 word choice problems collected by Mohammad et al. (2007). We lemmatize the target word and all candidates. This is especially beneficial for German words that can be highly inflected.

Following the approach by Jarmasz and Szpakowicz (2003), we compute the relatedness between the target and each of the candidates, and select the candidate with the maximum SR value. If two or more candidates are equally related to the target, then the candidates are said to be tied. If one of the tied candidates is the correct answer, then the problem is counted as correctly solved, but the corresponding score s_i is reduced to $\frac{1}{\text{\# of tied candidates}}$ (in effect approximating the score obtained by randomly guessing one of the tied candidates). Thus, a correctly solved problem without ties is assigned a score of 1.

If a phrase or multiword expression is used as a candidate and cannot be found in the lexical semantic resource, we remove stopwords (prepositions, articles, etc.) and split the candidate phrase into component words. For example, the target 'beret' in the above example has 'cap with horizontal peak' as one of its answer candidates. The candidate phrase

Language	Resource	Measure	Attempted	Score	# Ties	P	R	F_1
English	WordNet	PL	196	121.9	25	.62	.65	.64
		CV	152	131.3	3	**.86**	.51	.64
	Wikipedia	PL	226	88.33	96	.39	.75	.51
		CV_{first}	152	131.33	3	**.86**	.51	.64
		CV	288	165.83	2	.58	**.96**	**.72**
	Wiktionary	PL	201	103.7	55	.52	.67	.58
		CV	174	147.3	3	.85	.58	.69
German	GermaNet	PL	386	214.4	35	.56	.38	.45
		CV	304	193.3	3	.64	.30	.41
	Wikipedia	PL	711	326.8	174	.46	.71	.56
		CV_{first}	268	230.0	0	.86	.27	.41
		CV	807	574.3	4	.71	**.80**	**.75**
	Wiktionary	PL	194	84.8	30	.44	.19	.27
		CV	307	273.8	2	**.89**	.30	.45

Table 3: Results on word choice problems. Best precision, recall, and F_1 values for each language and resource are in bold.

is split into its component content words 'cap', 'horizontal', and 'peak'. We compute the SR between the target and each phrasal component and select the maximum value as the relatedness between the target and the candidate. If the target or all candidates cannot be found in the lexical semantic resource, a SR measure does not attempt to solve the problem. The overall score S of a SR measure is the sum of the scores yielded on the single problems $S = \sum_{wp \in A} s(wp)$, where A is the set of word choice problems that were attempted by the measure, and wp is a certain word choice problem.[12]

We define precision as $P = \frac{S}{|A|}$, recall as $R = \frac{|A|}{n}$, and F-measure as $F_1 = \frac{2PR}{P+R}$, where S is the overall score as defined above, $|A|$ is the number of word choice problems that were attempted by the SR measure, and n is the total number of word choice problems.[13]

Results When looking at the precision values in Table 3, we find that all resources except GermaNet perform comparably. Thus, the task of solving word choice problems in general is a matter of recall. Consequently, Wikipedia clearly outperforms all other lexical semantic resources with respect to overall performance (F-measure) on the German dataset due to its much higher recall. The differences are smaller for the English dataset, as the English Wiktionary and WordNet are more developed than their German counterparts.

When analyzing the performance with respect to the measure type, we find that CV based measures outperform PL based measures when using Wiktionary or Wikipedia. This is due to the higher recall of CV based measures using the large amount of additional information that Wiktionary or Wikipedia offer. PL based measures also produce a lot more ties as there is only a limited number of discrete PL values.

[12]Jarmasz and Szpakowicz (2003) use the overall score S. However, with this approach, a measure that attempts more problems may get a higher score just from random guessing.

[13]Note that the definition of recall is different from the definition by Mohammad et al. (2007). They computed $R = \frac{S}{n}$ making the recall dependent on the precision.

Important parameters of the CV measure are the length and the quality of the textual representations used to create the vector space. Using the full Wikipedia article yields the best recall (Eng. R=.96, Ger. R=.80) with reasonable precision (Eng. P=.58, Ger. P=.71). Using only the first paragraph yields the highest precision (Eng. P=.86, Ger. P=.86), while the recall is quite low in both cases (Eng. R=.51, Ger. R=.27). This is consistent with our previously described intuition, and allows us to configure the CV based measure according to whether high precision (CV_{first}) or high recall with reasonable precision (CV) is needed for an application.

Conclusions

In this paper, we have introduced Wiktionary as an emerging lexical semantic resource, and applied it to the task of ranking word pairs for English and German according to their semantic relatedness. We found that Wiktionary outperforms expert-made wordnets and Wikipedia in this task except for the special case of verb pairs where Wiktionary performs comparably to WordNet. In the second experiment of solving word choice problems, Wiktionary shows a precision equal to that of the other lexical semantic resources. The F_1 score obtained on the English dataset is comparable to that obtained by using other lexical semantic resources. On the German dataset, Wikipedia outperforms Wiktionary due to its higher recall. These results show that Wiktionary is a very promising and valuable lexical semantic resource that can be used as a substitute for expert-made lexical semantic resources, which are not easily available for many languages, and are expensive to create and maintain.

We generalized a concept vector based SR measure to work on each lexical semantic resource which provides a textual representation of a concept. For the first time, we applied a concept vector based measure to the following representations: Wiktionary pseudo glosses, the first paragraph of Wikipedia articles, the English WordNet glosses, and the GermaNet based pseudo glosses. We found that the CV measure consistently outperforms the standard PL measure due to its ability to capture also implicitly expressed lexical semantic relations. We also found that the CV based measure can be adapted to yield a better precision by trading

in some recall, when using shorter but more precise textual representations. Using this effect, we can configure the CV measure according to whether precision or recall is more important in an AI application.

Even if the number of entries is relatively low for the German Wiktionary, we showed that by using the concept vector based SR measure the coverage of Wiktionary exceeds the coverage of GermaNet. As Wikipedia has a high coverage of proper nouns, and Wiktionary covers more common vocabulary, we expect the two lexical semantic resources to be complementary. Thus, in the future, we will investigate whether the combination of Wiktionary and Wikipedia can further improve the performance of SR measures.

Finally, we created a Java based Wiktionary API that we have made freely available[14] for research purposes enabling researchers to use Wiktionary as a knowledge source in AI.

Acknowledgments

Parts of this work were supported by the German Research Foundation under grant GU 798/1-2. Many thanks to Lizhen Qu, who implemented important parts of JWKTL. We thank Alistair Kennedy, Giuseppe Pirro, Nuno Seco, Stan Szpakowicz, and Dongqiang Yang for providing evaluation data used in this paper.

References

Bouchard, A.; Liang, P.; Griffiths, T.; and Klein, D. 2007. A probabilistic approach to diachronic phonology. In *Proceedings of EMNLP-CoNLL*, 887–896.

Boyd-Graber, J.; Fellbaum, C.; Osherson, D.; and Shapire, R. 2006. Adding dense, weighted, connections to WordNet. In *Proceedings of the Third Global WordNet Meeting*.

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

Chesley, P.; Vincent, B.; Xu, L.; and Srihari, R. 2006. Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of AAAI-CAAW-06*.

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

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; and Wolfman, G. 2002. Placing Search in Context: The Concept Revisited. *ACM TOIS* 20(1):116–131.

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

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

Gurevych, I.; Müller, C.; and Zesch, T. 2007. What to be? - Electronic Career Guidance Based on Semantic Relatedness. In *Proceedings of ACL*, 1032–1039. Association for Computational Linguistics.

Gurevych, I. 2005. Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In *Proceedings of IJCNLP*, 767–778.

Jarmasz, M., and Szpakowicz, S. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of Recent Advances in Natural Language Processing*, 111–120.

[14] http://www.ukp.tu-darmstadt.de/software/

Jiang, J. J., and Conrath, D. W. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics*.

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

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

Leacock, C., and Chodorow, M. 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press. chapter Combining Local Context and WordNet Similarity for Word Sense Identification, 265–283.

Miller, G. A., and Charles, W. G. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes* 6(1):1–28.

Mohammad, S.; Gurevych, I.; Hirst, G.; and Zesch, T. 2007. Cross-lingual Distributional Profiles of Concepts for Measuring Semantic Distance. In *Proceedings of EMNLP-CoNLL*, 571–580.

Morris, J., and Hirst, G. 2004. Non-Classical Lexical Semantic Relations. In *Workshop on Computational Lexical Semantics, HLT-NAACL*, 46–51.

Patwardhan, S., and Pedersen, T. 2006. Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, 1–8.

Pedersen, T.; Pakhomov, S. V. S.; Patwardhan, S.; and Chute, C. G. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics* 40(3):288–299.

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

Rada, R.; Mili, H.; Bicknell, E.; and Blettner, M. 1989. Development and Application of a Metric on Semantic Nets. *IEEE Trans. on Systems, Man, and Cybernetics,* 19(1):17–30.

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

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

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

Stevenson, M., and Greenwood, M. A. 2005. A semantic approach to IE pattern induction. In *Proceedings of ACL*, 379–386. Association for Computational Linguistics.

Vossen, P. 1998. Introduction to EuroWordNet. *Computers and the Humanities. Special Issue on EuroWordNet.* 32(2–3):73–89.

Wilks, Y.; Fass, D.; Guo, C.-M.; McDonald, J.; Plate, T.; and Slator, B. 1990. Providing machine tractable dictionary tools. *Journal of Machine Translation* 5(2):99–151.

Yang, D., and Powers, D. M. W. 2006. Verb Similarity on the Taxonomy of WordNet. In *Proceedings of GWC-06*, 121–128.

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