Automatically creating datasets for measures of semantic relatedness

Torsten Zesch and **Iryna Gurevych**
Department of Telecooperation
Darmstadt University of Technology
D-64289 Darmstadt, Germany
{zesch,gurevych} (at) tk.informatik.tu-darmstadt.de

Abstract

Semantic relatedness is a special form of linguistic distance between words. Evaluating semantic relatedness measures is usually performed by comparison with human judgments. Previous test datasets had been created analytically and were limited in size. We propose a corpus-based system for automatically creating test datasets.[1] Experiments with human subjects show that the resulting datasets cover all degrees of relatedness. As a result of the corpus-based approach, test datasets cover all types of lexical-semantic relations and contain domain-specific words naturally occurring in texts.

1 Introduction

Linguistic distance plays an important role in many applications like information retrieval, word sense disambiguation, text summarization or spelling correction. It is defined on different kinds of textual units, e.g. documents, parts of a document (e.g. words and their surrounding context), words or concepts (Lebart and Rajman, 2000).[2] Linguistic distance between words is inverse to their semantic similarity or relatedness.

Semantic similarity is typically defined via the lexical relations of synonymy ($automobile$ – car) and hypernymy ($vehicle$ – car), while semantic relatedness (SR) is defined to cover any kind of lexical or functional association that may exist between two words (Gurevych, 2005).[3] Dissimilar words can be semantically related, e.g. via functional relationships ($night$ – $dark$) or when they are antonyms ($high$ – low). Many NLP applications require knowledge about semantic relatedness rather than just similarity (Budanitsky and Hirst, 2006).

A number of competing approaches for computing semantic relatedness of words have been developed (see Section 2). A commonly accepted method for evaluating these approaches is to compare their results with a gold standard based on human judgments on word pairs. For that purpose, relatedness scores for each word pair have to be determined experimentally. Creating test datasets for such experiments has so far been a labor-intensive manual process.

We propose a corpus-based system to automatically create test datasets for semantic relatedness experiments. Previous datasets were created analytically, preventing their use to gain insights into the nature of SR and also not necessarily reflecting the reality found in a corpus. They were also limited in size. We provide a larger annotated test set that is used to better analyze the connections and differences between the approaches for computing semantic relatedness.

The remainder of this paper is organized as follows: we first focus on the notion of semantic relatedness and how it can be evaluated. Section 3 reviews related work. Section 4 describes our system for automatically extracting word pairs from a corpus. Furthermore, the experimental setup leading to human judgments of semantic relatedness

[1] In the near future, we are planning to make the software available to interested researchers.

[2] In this paper, *word* denotes the graphemic form of a token and *concept* refers to a particular sense of a word.

[3] Nevertheless the two terms are often (mis)used interchangeably. We will use semantic relatedness in the remainder of this paper, as it is the more general term that subsumes semantic similarity.

Proceedings of the Workshop on Linguistic Distances, pages 16–24,
Sydney, July 2006. ©2006 Association for Computational Linguistics

is presented. Section 5 discusses the results, and finally we draw some conclusions in Section 6.

2 Evaluating SR measures

Various approaches for computing semantic relatedness of words or concepts have been proposed, e.g. dictionary-based (Lesk, 1986), ontology-based (Wu and Palmer, 1994; Leacock and Chodorow, 1998), information-based (Resnik, 1995; Jiang and Conrath, 1997) or distributional (Weeds and Weir, 2005). The knowledge sources used for computing relatedness can be as different as dictionaries, ontologies or large corpora.

According to Budanitsky and Hirst (2006), there are three prevalent approaches for evaluating SR measures: mathematical analysis, application-specific evaluation and comparison with human judgments.

Mathematical analysis can assess a measure with respect to some formal properties, e.g. whether a measure is a metric (Lin, 1998).[4] However, mathematical analysis cannot tell us whether a measure closely resembles human judgments or whether it performs best when used in a certain application.

The latter question is tackled by application-specific evaluation, where a measure is tested within the framework of a certain application, e.g. word sense disambiguation (Patwardhan et al., 2003) or malapropism detection (Budanitsky and Hirst, 2006). Lebart and Rajman (2000) argue for application-specific evaluation of similarity measures, because measures are always used for some task. But they also note that evaluating a measure as part of a usually complex application only indirectly assesses its quality. A certain measure may work well in one application, but not in another. Application-based evaluation can only state the fact, but give little explanation about the reasons.

The remaining approach - comparison with human judgments - is best suited for application independent evaluation of relatedness measures. Human annotators are asked to judge the relatedness of presented word pairs. Results from these experiments are used as a gold standard for evaluation. A further advantage of comparison with human judgments is the possibility to gain deeper insights into the nature of semantic relatedness.

However, creating datasets for evaluation has so far been limited in a number of respects. Only a small number of word pairs was manually selected, with semantic similarity instead of relatedness in mind. Word pairs consisted only of noun-noun combinations and only general terms were included. Polysemous and homonymous words were not disambiguated to concepts, i.e. humans annotated semantic relatedness of words rather than concepts.

3 Related work

In the seminal work by Rubenstein and Goodenough (1965), similarity judgments were obtained from 51 test subjects on 65 noun pairs written on paper cards. Test subjects were instructed to order the cards according to the “similarity of meaning” and then assign a continuous similarity value (0.0 - 4.0) to each card. Miller and Charles (1991) replicated the experiment with 38 test subjects judging on a subset of 30 pairs taken from the original 65 pairs. This experiment was again replicated by Resnik (1995) with 10 subjects. Table 1 summarizes previous experiments.

A comprehensive evaluation of SR measures requires a higher number of word pairs. However, the original experimental setup is not scalable as ordering several hundred paper cards is a cumbersome task. Furthermore, semantic relatedness is an intuitive concept and being forced to assign fine-grained continuous values is felt to overstrain the test subjects. Gurevych (2005) replicated the experiment of Rubenstein and Goodenough with the original 65 word pairs translated into German. She used an adapted experimental setup where test subjects had to assign discrete values {0,1,2,3,4} and word pairs were presented in isolation. This setup is also scalable to a higher number of word pairs (350) as was shown in Gurevych (2006). Finkelstein et al. (2002) annotated a larger set of word pairs (353), too. They used a 0-10 range of relatedness scores, but did not give further details about their experimental setup. In psycholinguistics, relatedness of words can also be determined through association tests (Schulte im Walde and Melinger, 2005). Results of such experiments are hard to quantify and cannot easily serve as the basis for evaluating SR measures.

Rubenstein and Goodenough selected word pairs analytically to cover the whole spectrum of

[4] That means, whether it fulfills some mathematical criteria: $d(x, y) \geq 0$; $d(x, y) = 0 \Leftrightarrow x = y$; $d(x, y) = d(y, x)$; $d(x, z) \leq d(x, y) + d(y, z)$.

							Correlation	
Paper	Language	Pairs	POS	Rel-type	Scores	# Subjects	Inter	Intra
R/G (1965)	English	65	N	sim	continuous 0–4	51	-	.850
M/C (1991)	English	30	N	sim	continuous 0–4	38	-	-
Res (1995)	English	30	N	sim	continuous 0–4	10	.903	-
Fin (2002)	English	353	N, V, A	relat	continuous 0–10	16	-	-
Gur (2005)	German	65	N	sim	discrete {0,1,2,3,4}	24	.810	-
Gur (2006)	German	350	N, V, A	relat	discrete {0,1,2,3,4}	8	.690	-
Z/G (2006)	German	328	N, V, A	relat	discrete {0,1,2,3,4}	21	.478	.647

Table 1: Comparison of previous experiments. R/G=Rubenstein and Goodenough, M/C=Miller and Charles, Res=Resnik, Fin=Finkelstein, Gur=Gurevych, Z/G=Zesch and Gurevych

similarity from "not similar" to "synonymous". This elaborate process is not feasible for a larger dataset or if domain-specific test sets should be compiled quickly. Therefore, we automatically create word pairs using a corpus-based approach. We assume that due to lexical-semantic cohesion, texts contain a sufficient number of words related by means of different lexical and semantic relations. Resulting from our corpus-based approach, test sets will also contain domain-specific terms. Previous studies only included general terms as opposed to domain-specific vocabularies and therefore failed to produce datasets that can be used to evaluate the ability of a measure to cope with domain-specific or technical terms. This is an important property if semantic relatedness is used in information retrieval where users tend to use specific search terms (*Porsche*) rather than general ones (*car*).

Furthermore, manually selected word pairs are often biased towards highly related pairs (Gurevych, 2006), because human annotators tend to select only highly related pairs connected by relations they are aware of. Automatic corpus-based selection of word pairs is more objective, leading to a balanced dataset with pairs connected by all kinds of lexical-semantic relations. Morris and Hirst (2004) pointed out that many relations between words in a text are non-classical (i.e. other than typical taxonomic relations like synonymy or hypernymy) and therefore not covered by semantic similarity.

Previous studies only considered semantic relatedness (or similarity) of *words* rather than concepts. However, polysemous or homonymous words should be annotated on the level of *concepts*. If we assume that *bank* has two meanings ("financial institution" vs. "river bank")[5] and it is paired with *money*, the result is two sense qualified pairs ($(bank_{financial} - money)$) and ($bank_{river}$ – $money$). It is obvious that the judgments on the two concept pairs should differ considerably. Concept annotated datasets can be used to test the ability of a measure to differentiate between senses when determining the relatedness of polysemous words. To our knowledge, this study is the first to include concept pairs and to automatically generate the test dataset.

In our experiment, we annotated a high number of pairs similar in size to the test sets by Finkelstein (2002) and Gurevych (2006). We used the revised experimental setup (Gurevych, 2005), based on discrete relatedness scores and presentation of word pairs in isolation, that is scalable to the higher number of pairs. We annotated semantic relatedness instead of similarity and included also non noun-noun pairs. Additionally, our corpus-based approach includes domain-specific technical terms and enables evaluation of the robustness of a measure.

4 Experiment

4.1 System architecture

Figure 1 gives an overview of our automatic corpus-based system for creating test datasets for evaluating SR measures.

In the first step, a source corpus is preprocessed using tokenization, POS-tagging and lemmatization resulting in a list of POS-tagged lemmas. Randomly generating word pairs from this list would result in too many unrelated pairs, yielding an unbalanced dataset. Thus, we assign weights to each word (e.g. using tf.idf-weighting). The most important document-specific words get the highest weights and due to lexical cohesion of the documents many related words can be found among the top rated. Therefore, we randomly generate a user-defined number of word pairs from the r words with the highest weights for each document.

[5] WordNet lists 10 meanings.

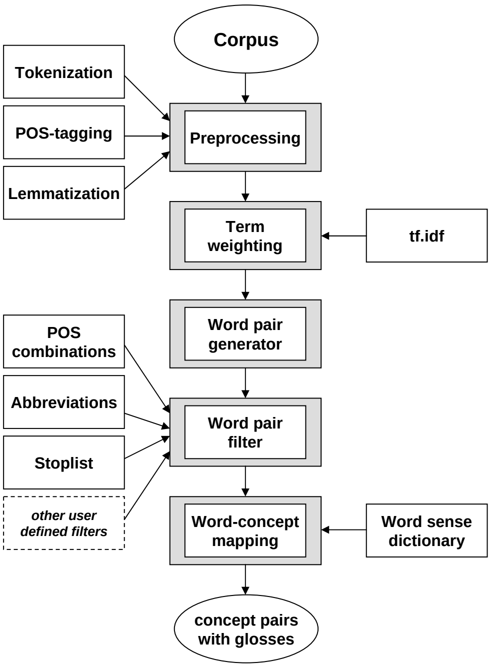


Figure 1: System architecture for extraction of concept pairs.

In the next step, user defined filters are applied to the initial list of word pairs. For example, a filter can remove all pairs containing only uppercase letters (mostly acronyms). Another filter can enforce a certain fraction of POS combinations to be present in the result set.

As we want to obtain judgment scores for semantic relatedness of concepts instead of words, we have to include all word sense combinations of a pair in the list. An external dictionary of word senses is necessary for this step. It is also used to add a gloss for each word sense that enables test subjects to distinguish between senses.

If differences in meaning between senses are very fine-grained, distinguishing between them is hard even for humans (Mihalcea and Moldovan, 2001).[6] Pairs containing such words are not suitable for evaluation. To limit their impact on the experiment, a threshold for the maximal number of senses can be defined. Words with a number of senses above the threshold are removed from the list.

The result of the extraction process is a list of sense disambiguated, POS-tagged pairs of concepts.

4.2 Experimental setup

4.2.1 Extraction of concept pairs

We extracted word pairs from three different domain-specific corpora (see Table 2). This is motivated by the aim to enable research in information retrieval incorporating SR measures. In particular, the “Semantic Information Retrieval” project (SIR Project, 2006) systematically investigates the use of lexical-semantic relations between words or concepts for improving the performance of information retrieval systems.

The *BERUFEnet* (BN) corpus[7] consists of descriptions of 5,800 professions in Germany and therefore contains many terms specific to professional training. Evaluating semantic relatedness on a test set based on this corpus may reveal the ability of a measure to adapt to a very special domain. The *GIRT* (German Indexing and Retrieval Testdatabase) corpus (Kluck, 2004) is a collection of abstracts of social science papers. It is a standard corpus for evaluating German information retrieval systems. The third corpus is compiled from 106 arbitrarily selected *scientific PowerPoint presentations* (SPP). They cover a wide range of topics from bio genetics to computer science and contain many technical terms. Due to the special structure of presentations, this corpus will be particularly demanding with respect to the required preprocessing components of an information retrieval system.

The three preprocessing steps (tokenization, POS-tagging, lemmatization) are performed using TreeTagger (Schmid, 1995). The resulting list of POS-tagged lemmas is weighted using the SMART ‘ltc’[8] tf.idf-weighting scheme (Salton, 1989).

We implemented a set of filters for word pairs. One group of filters removed unwanted word pairs. Word pairs are filtered if they contain at least one word that a) has less than three letters b) contains only uppercase letters (mostly acronyms) or c) can be found in a stoplist. Another filter enforced a specified fraction of combinations of nouns (N), verbs (V) and adjectives (A) to be present in the result set. We used the following parameters: $NN = 0.5$, $NV = 0.15$, $NA = 0.15$, $VV = 0.1$, $VA = 0.05$, $AA = 0.05$. That means 50% of the resulting word pairs for each corpus

[6] E.g. the German verb “halten” that can be translated as hold, maintain, present, sustain, etc. has 26 senses in GermaNet.

[7] `http://berufenet.arbeitsagentur.de`

[8] l=logarithmic term frequency, t=logarithmic inverse document frequency, c=cosine normalization.

CORPUS	# DOCS	# TOKENS	DOMAIN
BN	9,022	7,728,501	descriptions of professions
GIRT	151,319	19,645,417	abstracts of social science papers
SPP	106	144,074	scientific .ppt presentations

Table 2: Corpus statistics.

were noun-noun pairs, 15% noun-verb pairs and so on.

Word pairs containing polysemous words are expanded to concept pairs using GermaNet (Kunze, 2004), the German equivalent to WordNet, as a sense inventory for each word. It is the most complete resource of this type for German.

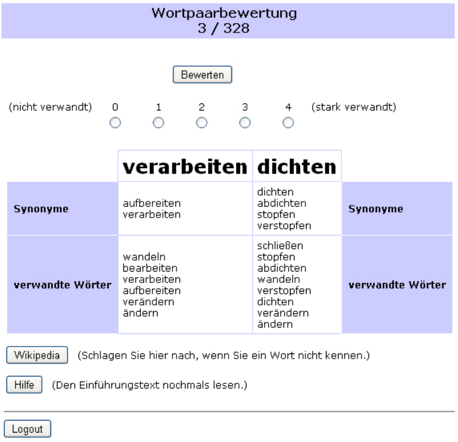
GermaNet contains only a few conceptual glosses. As they are required to enable test subjects to distinguish between senses, we use artificial glosses composed from synonyms and hypernyms as a surrogate, e.g. for *brother*: "brother, male sibling" vs. "brother, comrade, friend" (Gurevych, 2005). We removed words which had more than three senses.

Marginal manual post-processing was necessary, since the lemmatization process introduced some errors. Foreign words were translated into German, unless they are common technical terminology. We initially selected 100 word pairs from each corpus. 11 word pairs were removed because they comprised non-words. Expanding the word list to a concept list increased the size of the list. Thus, the final dataset contained 328 automatically created concept pairs.

4.2.2 Graphical User Interface

We developed a web-based interface to obtain human judgments of semantic relatedness for each automatically generated concept pair. Test subjects were invited via email to participate in the experiment. Thus, they were not supervised during the experiment.

Gurevych (2006) observed that some annotators were not familiar with the exact definition of semantic relatedness. Their results differed particularly in cases of antonymy or distributionally related pairs. We created a manual with a detailed introduction to SR stressing the crucial points. The manual was presented to the subjects before the experiment and could be re-accessed at any time.



Wortpaarbewertung
3 / 328

Bewerten

(nicht verwandt) 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ (stark verwandt)

	verarbeiten	dichten	
Synonyme	aufbereiten verarbeiten	dichten abdichten stopfen verstopfen	Synonyme
verwandte Wörter	wandeln bearbeiten verarbeiten aufbereiten verändern ändern	schließen stopfen abdichten wandeln verstopfen dichten verändern ändern	verwandte Wörter

Wikipedia (Schlagen Sie hier nach, wenn Sie ein Wort nicht kennen.)

Hilfe (Den Einführungstext nochmals lesen.)

Logout

Figure 2: Screenshot of the GUI. Polysemous words are defined by means of synonyms and related words.

During the experiment, one concept pair at a time was presented to the test subjects in random ordering. Subjects had to assign a discrete relatedness value {0,1,2,3,4} to each pair. Figure 2 shows the system's GUI.

In case of a polysemous word, synonyms or related words were presented to enable test subjects to understand the sense of a presented concept. Because this additional information can lead to undesirable priming effects, test subjects were instructed to deliberately decide only about the relatedness of a concept pair and use the gloss solely to understand the sense of the presented concept.

Since our corpus-based approach includes domain-specific vocabulary, we could not assume that the subjects were familiar with all words. Thus, they were instructed to look up unknown words in the German Wikipedia.[9]

Several test subjects were asked to repeat the experiment with a minimum break of one day. Results from the repetition can be used to measure intra-subject correlation. They can also be used to obtain some hints on varying difficulty of judgment for special concept pairs or parts-of-speech.

5 Results and discussion

21 test subjects (13 males, 8 females) participated in the experiment, two of them repeated it. The average age of the subjects was 26 years. Most subjects had an IT background. The experiment took 39 minutes on average, leaving about 7 seconds for rating each concept pair.

The summarized inter-subject correlation between 21 subjects was r=.478 (cf. Table 3), which

[9] http://www.wikipedia.de

	Concepts		Words	
	Inter	Intra	Inter	Intra
all	.478	.647	.490	.675
BN	.469	.695	.501	.718
GIRT	.451	.598	.463	.625
SPP	.535	.649	.523	.679
AA	.556	.890	.597	.887
NA	.547	.773	.511	.758
NV	.510	.658	.540	.647
NN	.463	.620	.476	.661
VA	.317	.318	.391	.212
VV	.278	.494	.301	.476

Table 3: Summarized correlation coefficients for all pairs, grouped by corpus and grouped by POS combinations.

is statistically significant at $p < .05$. This correlation coefficient is an upper bound of performance for automatic SR measures applied on the same dataset.

Resnik (1995) reported a correlation of r=.9026.[10] The results are not directly comparable, because he only used noun-noun pairs, words instead of concepts, a much smaller dataset, and measured semantic similarity instead of semantic relatedness. Finkelstein et al. (2002) did not report inter-subject correlation for their larger dataset. Gurevych (2006) reported a correlation of r=.69. Test subjects were trained students of computational linguistics, and word pairs were selected analytically.

Evaluating the influence of using concept pairs instead of word pairs is complicated because word level judgments are not directly available. Therefore, we computed a lower and an upper bound for correlation coefficients. For the lower bound, we always selected the concept pair with highest standard deviation from each set of corresponding concept pairs. The upper bound is computed by selecting the concept pair with the lowest standard deviation. The differences between correlation coefficient for concepts and words are not significant. Table 3 shows only the lower bounds.

Correlation coefficients for experiments measuring semantic relatedness are expected to be lower than results for semantic similarity, since the former also includes additional relations (like cooccurrence of words) and is thus a more complicated task. Judgments for such relations strongly depend on experience and cultural background of the test subjects. While most people may agree that $(car - vehicle)$ are highly related, a strong connection between $(parts - speech)$ may only be established by a certain group. Due to the corpus-based approach, many domain-specific concept pairs are introduced into the test set. Therefore, inter-subject correlation is lower than the results obtained by Gurevych (2006).

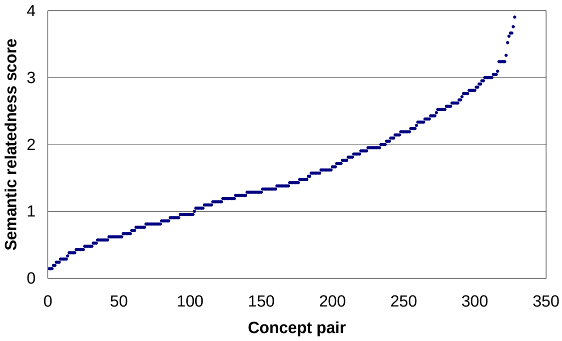


Figure 3: Distribution of averaged human judgments.

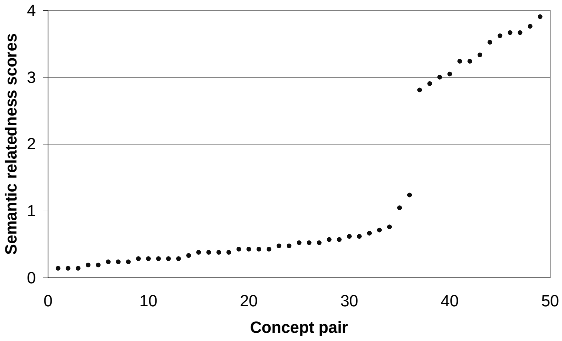


Figure 4: Distribution of averaged human judgments with standard deviation < 0.8.

In our experiment, intra-subject correlation was r=.670 for the first and r=.623 for the second individual who repeated the experiment, yielding a summarized intra-subject correlation of r=.647. Rubenstein and Goodenough (1965) reported an intra-subject correlation of r=.85 for 15 subjects judging the similarity of a subset (36) of the original 65 word pairs. The values may again not be compared directly. Furthermore, we cannot generalize from these results, because the number of participants which repeated our experiment was too low.

The distribution of averaged human judgments on the whole test set (see Figure 3) is almost balanced with a slight underrepresentation of highly related concepts. To create more highly related concept pairs, more sophisticated weighting schemes or selection on the basis of lexical chain-

[10] Note that Resnik used the averaged correlation coefficient. We computed the summarized correlation coefficient using a Fisher Z-value transformation.

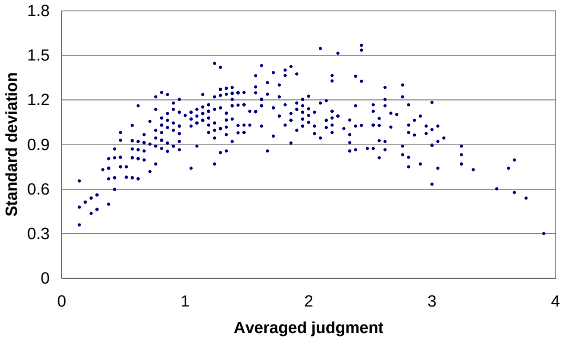


Figure 5: Averaged judgments and standard deviation for all concept pairs. Low deviations are only observed for low or high judgments.

ing could be used. However, even with the present setup, automatic extraction of concept pairs performs remarkably well and can be used to quickly create balanced test datasets.

Budanitsky and Hirst (2006) pointed out that distribution plots of judgments for the word pairs used by Rubenstein and Goodenough display an empty horizontal band that could be used to separate related and unrelated pairs. This empty band is not observed here. However, Figure 4 shows the distribution of averaged judgments with the highest agreement between annotators (standard deviation < 0.8). The plot clearly shows an empty horizontal band with no judgments. The connection between averaged judgments and standard deviation is plotted in Figure 5.

When analyzing the concept pairs with lowest deviation there is a clear tendency for particularly highly related pairs, e.g. hypernymy: *Universität – Bildungseinrichtung* (*university – educational institution*); functional relation: *Tätigkeit – ausführen* (*task – perform*); or pairs that are obviously not connected, e.g. *logisch – Juni* (*logical – June*). Table 4 lists some example concept pairs along with averaged judgments and standard deviation.

Concept pairs with high deviations between judgments often contain polysemous words. For example, $Quelle$ ($source$) was disambiguated to $Wasserquelle$ ($spring$) and paired with $Text$ ($text$). The data shows a clear distinction between one group that rated the pair low (0) and another group that rated the pair high (3 or 4). The latter group obviously missed the point that *textual source* was not an option here. High deviations were also common among special technical terms like ($Mips$ – $Core$), proper names ($Georg$ – $August$ – two common first names in German) or functionally related pairs ($agieren$ – $mobil$). Human experience and cultural background clearly influence the judgment of such pairs.

The effect observed here and the effect noted by Budanitsky and Hirst is probably caused by the same underlying principle. Human agreement on semantic relatedness is only reliable if two words or concepts are highly related or almost unrelated. Intuitively, this means that classifying word pairs as related or unrelated is much easier than numerically rating semantic relatedness. For an information retrieval task, such a classification might be sufficient.

Differences in correlation coefficients for the three corpora are not significant indicating that the phenomenon is not domain-specific. Differences in correlation coefficients for different parts-of-speech are significant (see Table 3). Verb-verb and verb-adjective pairs have the lowest correlation. A high fraction of these pairs is in the problematic medium relatedness area. Adjective-adjective pairs have the highest correlation. Most of these pairs are either highly related or not related at all.

6 Conclusion

We proposed a system for automatically creating datasets for evaluating semantic relatedness measures. We have shown that our corpus-based approach enables fast development of large domain-specific datasets that cover all types of lexical and semantic relations. We conducted an experiment to obtain human judgments of semantic relatedness on concept pairs. Results show that averaged human judgments cover all degrees of relatedness with a slight underrepresentation of highly related concept pairs. More highly related concept pairs could be generated by using more sophisticated weighting schemes or selecting concept pairs on the basis of lexical chaining.

Inter-subject correlation in this experiment is lower than the results from previous studies due to several reasons. We measured semantic relatedness instead of semantic similarity. The former is a more complicated task for annotators because its definition includes all kinds of lexical-semantic relations not just synonymy. In addition, concept pairs were automatically selected eliminating the bias towards strong classical relations with high agreement that is introduced into the dataset by a manual selection process. Furthermore, our dataset contains many domain-specific

Pair				
German	English	Corpus	Avg	St-dev
Universität – Bildungseinrichtung	university – educational institution	GIRT	3.90	0.30
Tätigkeit – ausführen	task – to perform	BN	3.67	0.58
strafen – Paragraph	to punish – paragraph	GIRT	3.00	1.18
Quelle – Text	spring – text	GIRT	2.43	1.57
Mips – Core	mips – core	SPP	2.10	1.55
elektronisch – neu	electronic – new	GIRT	1.71	1.15
verarbeiten – dichten	to manipulate – to caulk	BN	1.29	1.42
Leopold – Institut	Leopold – institute	SPP	0.81	1.25
Outfit – Strom	outfit – electricity	GIRT	0.24	0.44
logisch – Juni	logical – June	SPP	0.14	0.48

Table 4: Example concept pairs with averaged judgments and standard deviation. Only one sense is listed for polysemous words. Conceptual glosses are omitted due to space limitations.

concept pairs which have been rated very differently by test subjects depending on their experience. Future experiments should ensure that domain-specific pairs are judged by domain experts to reduce disagreement between annotators caused by varying degrees of familiarity with the domain.

An analysis of the data shows that test subjects more often agreed on highly related or unrelated concept pairs, while they often disagreed on pairs with a medium relatedness value. This result raises the question whether human judgments of semantic relatedness with medium scores are reliable and should be used for evaluating semantic relatedness measures. We plan to investigate the impact of this outcome on the evaluation of semantic relatedness measures. Additionally, for some applications like information retrieval it may be sufficient to detect highly related pairs rather than accurately rating word pairs with medium values.

There is also a significant difference between the correlation coefficient for different POS combinations. Further investigations are needed to elucidate whether these differences are caused by the new procedure for corpus-based selection of word pairs proposed in this paper or are due to inherent properties of semantic relations existing between word classes.

Acknowledgments

We would like to thank Sabine Schulte im Walde for her remarks on experimental setups. We are grateful to the *Bundesagentur für Arbeit* for providing the BERUFEnet corpus. This work was carried out as part of the "Semantic Information Retrieval" (SIR) project funded by the German Research Foundation.

References

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

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, and Gadi Wolfman. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

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.

Iryna Gurevych. 2006. Computing Semantic Relatedness Across Parts of Speech. Technical report, Darmstadt University of Technology, Germany, Department of Computer Science, Telecooperation.

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

Michael Kluck. 2004. The GIRT Data in the Evaluation of CLIR Systems - from 1997 Until 2003. *Lecture Notes in Computer Science*, 3237:376–390, January.

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

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

Ludovic Lebart and Martin Rajman. 2000. Computing Similarity. In Robert Dale, editor, *Handbook of NLP*. Dekker: Basel.

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, Toronto, Ontario, Canada.

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. In *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin.

Rada Mihalcea and Dan Moldovan. 2001. Automatic Generation of a Coarse Grained WordNet. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA, June.

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

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.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada.

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

Gerard Salton. 1989. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing, Boston, MA, USA.

Helmut Schmid. 1995. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Sabine Schulte im Walde and Alissa Melinger. 2005. Identifying Semantic Relations and Functional Properties of Human Verb Associations. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in NLP*, pages 612–619, Vancouver, Canada.

SIR Project. 2006. Project 'Semantic Information Retrieval'. URL `http://www.cre-elearning.tu-darmstadt.de/elearning/sir/`.

Julie Weeds and David Weir. 2005. Co-occurrence Retrieval: A Flexible Framework For Lexical Distributional Similarity. *Computational Linguistics*, 31(4):439–475, December.

Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. In *32nd Annual Meeting of the ACL*, pages 133–138, New Mexico State University, Las Cruces, New Mexico.