How Well Do Semantic Relatedness Measures Perform? A Meta-Study

Irene Cramer

Dortmund University of Technology (Germany)

email: irene.cramer@udo.edu

Abstract

Various semantic relatedness, similarity, and distance measures have been proposed in the past decade and many NLP-applications strongly rely on these semantic measures. Researchers compete for better algorithms and normally only few percentage points seem to suffice in order to prove a new measure outperforms an older one. In this paper we present a meta-study comparing various semantic measures and their correlation with human judgments. We show that the results are rather inconsistent and ask for detailed analyses as well as clarification. We argue that the definition of a shared task might bring us considerably closer to understanding the concept of semantic relatedness.

1 Introduction

Various applications in Natural Language Processing, such as Question Answering (Novischi and Moldovan, 2006), Topic Detection (Carthy, 2004), and Text Summarization (Barzilay and Elhadad, 1997), rely on semantic relatedness (similarity or distance)[1] measures either based on word nets and/or corpus statistics as a resource. In the HyTex project, funded by the German Research Foundation, we develop strategies for the text-to-hypertext conversion using text-grammatical features. One strand of research in this project consists of topic-based linking methods using lexical chaining as a resource (Cramer and Finthammer, 2008). Lexical chaining is a well-known method to calculate partial text representations; it relies on semantic relatedness values as basic input. We therefore implemented[2] eight semantic relatedness measures — (Hirst and St-Onge, 1998; Jiang and Conrath, 1997; Leacock and Chodorow, 1998; Lin, 1998; Resnik, 1995; Wu and Palmer, 1994) — based on GermaNet[3] Lemnitzer and Kunze (2002) and three based on Google co-occurrence counts (Cilibrasi and Vitanyi, 2007). In order to evaluate the performance of these measures we conducted two human judgment experiments and computed the correlation between the human judgment and the values of the eleven semantic measures. We also compared our results with those reported in the literature and found that the correlations between human judgments and semantic measures are extremely scattered. In this paper we compare the correlation of our own human judgment experiments and the results of three similar studies. In our opinion this comparison points to the necessity of a thorough analysis of the methods used in these experiments. We argue that this analysis should aim at answering the following questions:

- How does the setting of the human judgment experiment influence the results?
- How does the selection of the word-pairs influence the results?
- Which aspects of semantic relatedness are included in human judgments? Thus, what do these experiments actually measure?
- Are the semantic relatedness measures proposed in the literature able to capture all of these aspects?

In this paper we intend to open the above mentioned analysis and therefore assembled a set of aspects which we consider to be important in order to answer these questions. Consequently, the remainder of this paper is structured as follows: In Section 2 we

[1] The notions of semantic relatedness, similarity, and distance measure are controversially discussed in the literature, e.g. Budanitsky and Hirst (2006). However, semantic similarity and relatedness seem to be the predominant terms in this context. Budanitsky and Hirst (2006) define them as follows: word-pairs are considered to be semantically similar if a synonymy or hypernymy relation holds. In contrast, word-pairs are considered to be semantically related if a systematic relation, such as synonymy, antonymy, hypernymy, holonymy, or an unsystematic relation holds. Thus relatedness is the more general (broader) concept since it includes intuitive associations as well as linguistically formalized relations between words (or concepts). The focus of this paper is on relatedness.

[2] Since GermaNet — e.g. in terms of internal structure — slightly differs from Princeton WordNet we could not simply use the measure implementations of the latter and therefore had to reimplement and adapt them for GermaNet.

[3] GermaNet is the German counterpart of WordNet (Fellbaum, 1998).

present our own human judgment experiments. In Section 3 we describe three similar studies, two conducted with English data and one with German. In Section 4 we compare the results of the four studies and discuss (with respect to the experimental setting and goals) potential differences and possible causes for the observed inconsistency of the results. Finally, we summarize our work and outline some ideas for future research.

2 Our Human Judgement Experiments

In order to evaluate the quality of a semantic measure, a set of pre-classified (i.e. judged with respect to their semantic relatedness by subjects) word-pairs is necessary. In previous work for English data, most researchers used the word-pair list by Rubenstein and Goodenough (1965) as well as the list by Miller and Charles (1991) as an evaluation resource. For German there are — to our knowledge — two research groups, which compiled lists of word-pairs with respective human judgment:

- Gurevych et al. constructed three lists (a translation of Rubenstein and Goodenough's list (Gurevych, 2005), a manually generated set of word-pairs, and a semi-automatically generated one (Zesch and Gurevych, 2006)).
- While investigating lexical chaining for German corpora, we additionally compiled a total of six lists, each of which consists of 100 word-pairs with respective human judgments.

The goal of our experiments was to cover a wide range of relatedness types, i.e. systematic and unsystematic relations, and relatedness levels, i.e. various degrees of relation strength. However, we only included nouns in the construction of our sets of word-pairs, since we consider cross-part-of-speech (cross-POS) relations to be an additional challenge[4], which we intend to address in a continuative experiment. Furthermore, in order to identify a potential bias of the lists and the impact of this bias on the results, we applied two different methods for the compilation of word-pairs.

For our first human judgment experiment (Cramer and Finthammer, 2008) we collected nouns (analytically)[5] of diverse semantic classes, e.g. abstract nouns, such as `das Wissen` (Engl. knowledge), and concrete nouns, such as `das Bügeleisen` (Engl. flat-iron). By this means, we constructed a list of approximately 300 word-pairs. We picked approximately 75 and randomized them. For the remaining 25 word-pairs, we selected five words and constructed word-pairs such as `Sonne-Wind` (Engl. sun-wind), `Sonne-Wärme` (Engl. sun-warmth), `Sonne-Wetter` (Engl. sun-weather) etc. We arranged these 25 pairs into sequences in order to focus our subjects' attention on small semantic relatedness distinctions.

[4] Since in most word nets cross-POS relations are very sparse, researchers currently investigate relation types able to connect the noun, verb, and adjective sub-graphs (e.g. Marrafa and Mendes (2006) or Lemnitzer et al. (2008)). However, these new relations are not yet integrated on a large scale and therefore should not (or even cannot) be used in semantic relatedness measures. Furthermore, calculating semantic relatedness between words with different POS might introduce additional challenges potentially as yet unidentified, which calls for a careful exploration.

[5] In this paper and in most comparable studies, the term *analytical* means that the word-pairs are handpicked. Obviously, the disadvantage of this approach is its sensibility to idiosyncrasies, which might extremely bias the outcome of the experiments.

For the five remaining lists (WP2-WP6), we applied a different method: firstly, we again analytically collected word-pairs which are part of collocations, i.e. the two nouns `Rat` and `Tat` (`mit Rat und Tat helfen`, Engl. to help with words and deeds) or `Qual` and `Wahl` (`die Qual der Wahl haben`, Engl. to be spoilt for choice). Secondly, we assembled word-pairs which feature association relations, i.e. `Afrika` (Engl. Africa) and `Tiger` (Engl. tiger) or `Weihnachten` (Engl. Christmas) and `Zimt` (Engl. cinnamon). Thirdly, we automatically constructed a list of random word-pairs using the Wacky corpus (Baroni and Bernardini, 2006) as a resource and manually excluded ad-hoc-constructions. Finally, out of these three resources we compiled five sets of 100 randomized word-pairs with no more than 20% of the collocation and association word-pairs.

We asked subjects to rate the word-pairs on a 5-level scale (0 = *not related* to 4 = *strongly related*). The subjects were instructed to base the rating on their intuition about any kind of conceivable relation between the two words. WP1 was rated by 35 subjects and WP2 to WP6 were each rated by 15 subjects. We then calculated the average judgment per word-pair and ranked the word-pairs accordingly.

The correlation between the human judgments and the eleven semantic measures is shown in Table 1. The difference between the correlation coefficients of WP1 and WP2-WP6 suggests that the method of construction might have an impact on the results of the experiments. The manual compilation of word-pairs seems to lead to better correlation coefficients and might therefore cause an overestimation of the performance of the semantic measures. Furthermore, with respect to the list construction methods, the two resources and respective measures, namely GermaNet (TreePath–Lin) and Google (GoogleQ–GooglePMI), seem to respond differently: whereas the correlation coefficients of the eight GermaNet based measures drop to a greater or lesser extend (Table 1: *r* for WP1 and *r* for WP2-WP6), the correlation coefficients of the three Google based measures approximately level off.

Table 1: Our Correlation Coefficients: Correlation between Average Human Judgment and Semantic Measure Values

r	**Tree Path**	**Graph Path**	**Wu-Palm.**	**Leac.-Chod.**	**Hirst-St-O.**	**Resnik**	**Jiang-Conr.**	**Lin**	**Google Norm.**	**Google Quot.**	**Google PMI**
WP1	0.41	0.42	0.36	0.48	0.47	0.44	0.45	0.48	0.27	0.37	0.37
WP2	0.09	0.31	0.33	0.16	0.26	0.37	0.18	0.36	0.24	0.29	0.27
WP3	0.03	0.22	0.24	0.11	0.28	0.19	0.15	0.26	0.46	0.45	0.40
WP4	0.07	0.39	0.11	0.11	0.31	0.11	0.25	0.16	0.34	0.38	0.34
WP5	0.27	0.39	0.26	0.32	0.38	0.31	0.41	0.34	0.19	0.32	0.28
WP6	0.09	0.27	0.15	0.17	0.39	0.24	0.29	0.25	0.26	0.38	0.43
mean	**0.16**	**0.33**	**0.24**	**0.23**	**0.35**	**0.28**	**0.29**	**0.31**	**0.29**	**0.36**	**0.35**

In any case, since the correlation coefficients are rather low, there is much room for improvement. However, as all measures scatter in the same range — independently of the precise algorithm or resource used, as it seems — we argue that the reason for this critical performance might be one of the following two aspects (most probably a combination of both):

- Word nets (and/or corpora) do not cover the (all) types of semantic information required.
- Human judgment experiments are (without clear and standardized specification of the experimental setup) an inappropriate way to evaluate semantic measures.

Both aspects are discussed in Section 4. However, we first should have a look at three similar studies, two for English and one for German.

3 Three Similar Studies

As mentioned above various researchers rely on human judgment experiments as an evaluation resource for semantic relatedness measures. In this section, three such studies are summarized in order to identify differences with respect to the methods adopted and results obtained.[6]

3.1 Budanitsky and Hirst

Budanitsky and Hirst (2006) specify the purpose of their paper *Evaluating WordNet-based Measures of Lexical Semantic Relatedness* as a comparison of the performance of various relatedness measures. Accordingly, they sketch a number of measures and identify three evaluation methods: firstly, the theoretical examination (of e.g. the mathematical properties of the respective measure); secondly, the comparison with human judgments; thirdly, the evaluation of a measure with respect to a given NLP-application. They regard the second and third method as being the most appropriate ones and therefore focus on them in their empirical work presented in the paper. As a basis for the second evaluation method, i.e. the comparison between semantic measure and human judgments, they use two word-pair lists: the first compiled by Rubenstein and Goodenough (1965) and containing 65 word-pairs[7], the second compiled by Miller and Charles (1991) and containing 30 word-pairs. In order to evaluate the performance of five different measures (and potentially in order to find a ranking), Budanitsky and Hirst (2006) compute the semantic relatedness values for the word-pairs and compare them with the human judgments. They thus find the correlation coefficients summarized in Table 2.

Budanitsky and Hirst (2006) regard this evaluation method, i.e. comparing measure values and human judgments, as the ideal approach. However, in examining the results of this comparison, they identify several limitations; i.e. they point out that the amount of data available (65 word-pairs) might be inadequate for real NLP-applications. They additionally emphasize that the development of a large-scale data set would be time-consuming and expensive. Moreover, they argue that the experiments by Rubenstein and Goodenough (1965) as well as Miller and Charles (1991) focus on relations between words rather than relations between word-senses (concepts), which would be

[6] There are many more relevant studies; however, they all point to the same issue, namely, the incompatibility of the results.

[7] Rubenstein and Goodenough (1965) investigated the relationship between 'similarity of context' and 'similarity of meaning'. They asked 51 subjects to rate on a scale of 0 to 4 the similarity of meaning for the 65 word-pairs. Miller and Charles (1991) selected 30 out of the 65 original word-pairs (according to their relatedness strength) and asked 38 subjects to rate this list. They used the same experimental setup as Rubenstein and Goodenough (1965).

Table 2: Correlation Coefficients by Budantisky and Hirst

r	**Leac.-Chod.**	**Hirst-StO.**	**Resnik**	**Jiang-Conr.**	Lin
M&C	0.816	0.744	0.774	0.850	0.82
R&G	0.838	0.786	0.779	0.781	0.819
mean	**0.83**	**0.77**	**0.78**	**0.82**	**0.82**

— especially when taking potential NLP-applications into account — more appropriate. They note that it might however be difficult to trigger a specific concept without biasing the subjects.

3.2 Boyd-Graber, Fellbaum, Osherson, and Schapire

In contrast to the above mentioned experiments by Budanitsky and Hirst (2006), the research reported in *Adding Dense, Weighted Connections to WordNet* aims at the development of a new, conceptually different layer of relations to be included into a word net. Boyd-Graber et al. (2006) are motivated in their work by three widely acknowledged shortcomings of word nets:

- The lack of cross-POS links connecting the sub-graphs containing nouns, verbs, or adjectives, respectively.
- The low density of relations in the sub-graphs, i.e. potentially missing types of relations such as 'actor' or 'instrument'.
- The absence of weights assigned to the relations, i.e. representing the degrees of semantic distance of different subordinates of the same superordinate.

In order to address these shortcomings, Boyd-Graber et al. ask subjects to assign values of 'evocation' representing the relations between 1,000 synsets. They ask 20 subjects to rate evocation in 120,000 pairs of synsets (these pairs form a random selection of all possible pairs of the above mentioned 1,000 core synsets considered in the experiment). The subjects are given a manual explaining a couple of details about the task and are trained on a sample of 1,000 (two sets of 500) randomly selected pairs. Although the research objective of the work presented in this paper is to construct a new relations layer for Princeton WordNet rather than to evaluate semantic relatedness measures, Boyd-Graber et al. compare the results of their human judgment experiment with the relatedness values of four different semantic measures. The correlation coefficients of this comparison are summarized in Table 3.

Boyd-Graber et al. arrive at the conclusion that — given the obvious lack of correlation (see Table 3) — evocation constitutes an empirically supported semantic relation type which is still not captured by the semantic measures (at least not by those considered in this experiment).

3.3 Gurevych et al.

Similar to the study by Budanitsky and Hirst (2006), Gurevych (2005) gives insight into a human judgment experiment conducted in order to compare the performance

Table 3: Correlation Coefficients by Boyd-Graber et al.

r	**Lesk**	**Path**	**LC**	**LSA**
all	0.008			
verbs		0.046		
nouns		0.013	0.013	
closest				0.131

Table 4: Correlation Coefficients by Gurevych (with Lesk_1 = Lesk (DWDS); Lesk_2 = Lesk (radial); Lesk_3 = Lesk (hypernym); Resn. = Resnik)

r	**Google**	**Lesk_1**	**Lesk_2**	**Lesk_3**	**Resn.**
R&G German	0.57	0.53	0.55	0.60	0.72

of her own semantic relatedness measure[8] with established ones. For this purpose (Gurevych, 2005) translates the word-pair list by Rubenstein and Goodenough (1965) and asks 24 native speakers of German to rate the word-pairs with respect to their semantic relatedness on a 5-level scale; she thus replicates the study by Rubenstein and Goodenough (1965) for German. Gurevych (2005) finally compares the human judgments with several semantic measures. The correlation coefficients of this comparison are summarized in Table 4.

Gurevych (2005) comments on (among others) the following four issues: firstly, she emphasizes the difference between semantic similarity and relatedness; she argues that most word-pair lists were constructed in order to measure semantic similarity rather than relatedness and that these lists might therefore be inappropriate for the task at hand. Secondly, Gurevych (2005) observes that, in contrast to the concept of semantic similarity, semantic relatedness is not well defined. Thirdly, as the experiments are based on words rather than concepts, the results attained thus far might exhibit additional noise. Finally, she notes that the amount of data is too limited in size and that analytically created word-pair lists are inherently biased. Accordingly, Zesch and Gurevych (2006) propose a corpus based method for automatically constructing test data and list a number of advantages of this approach: i.e. lexical-semantic cohesion in texts accounts for various relation types, domain-specific and technical terms can easily be included, and, in contrast to manually constructed, corpus based lists are probably more objective.

4 Meta-Level Evaluation

Table 5 shows the minimum, maximum, and mean correlations reported in the three studies as well as our own results. The table illustrates the broad statistical spread: the mean correlation coefficients range between 0.8 and 0.04 for English and between 0.61 and 0.29 for German. Admittedly, the experimental setup and the goals of the

[8] Her measure is able to manage limitations of some of the previously published measures.

Table 5: Comparison of the Correlation Coefficients of the Different Experiments (with B&G: Budanitsky and Hirst / B-G et al.: Boyd-Graber et al. / G et al.: Gurevych et al. / C&F, C: our results)

	B&H	**B-G et al.**	**G et al.**	**C&F, C**
max	0.83	0.131	0.72	0.36
min	0.77	0.008	0.53	0.16
mean	**0.80**	**0.04**	**0.61**	**0.29**
stdv	0.03	0.05	0.08	0.06

four studies differ in several aspects[9]. However, the principle idea — i.e. using human judgments as a baseline or evaluation resource — is the same.

We argue that — given the statistical spread shown in Table 5 — as long as the reasons for this discrepancy have not been determined and the methods have not been harmonized as far as possible, the results of these experiments should not be used as a basis for e.g. the evaluation or comparison of semantic measures. As mentioned in Section 1 we suspect that (no fewer than) the following aspects influence the results of the human judgment experiments and thus the correlation between humans and semantic measures:

- **Research objective:** The goals of the studies differ with respect to several aspects. Firstly, some studies, e.g. Budanitsky and Hirst (2006), aim at comparing the performance of different semantic (relatedness) measures, whereas Boyd-Graber et al. (2006) intend to construct a new relations layer (potentially able to substitute or complement established relatedness measures). Secondly, in some cases, e.g. Cramer and Finthammer (2008), relations between words are considered, whereas e.g. Boyd-Graber et al. (2006) examine relations between concepts. Thirdly, it seems to be unclear which types of relations are actually searched for (relatedness, similarity, evocation, distance) and in what aspects these correspond or differ. Interestingly, in computational linguistics and psycholinguistics there is an additional strand of research investigating the so-called 'association relation', e.g. Schulte im Walde and Melinger (2005) and Roth and Schulte im Walde (2008), which is not yet considered or integrated in the research on semantic relatedness measures. We argue that such an integration might be fruitful for both research strands.

- **Setting of the human judgment experiment:** In all studies summarized above, the subjects are students (mostly of linguistics, computer sciences, and computational linguistics). In most cases, they are given a short manual explaining the task, which certainly differs in many aspects, e.g. due to the above mentioned fact that the relation type searched for is a still unsettled issue. Furthermore, no training phase is included in most of the studies except the one by Boyd-Graber

[9] It seems unfeasible to determine all possible differences of the studies because, among other things, the papers do not specify the experimental setup in detail. We therefore assume that the definition of a shared task might bring us considerably closer to understanding the questions raised in this paper.

et al. (2006), who are therefore able to identify potential training effects. Again only Boyd-Graber et al. (2006) account for the handling of idiosyncrasies.

- **Construction of experimental data:** In Boyd-Graber et al. (2006) the concept-pairs were randomly selected, whereas the word-pairs used by Budanitsky and Hirst (2006) were constructed analytically. In the studies by Gurevych (2005), Zesch and Gurevych (2006), and Cramer and Finthammer (2008), some were analytically constructed and some randomly (semi-automatically) selected. In addition, the data sets vary with respect to their size: Budanitsky and Hirst (2006), Gurevych (2005), and Cramer and Finthammer (2008) only use small sets of word-pairs (concept-pairs), i.e. a few hundred pairs, whereas Boyd-Graber et al. (2006) investigate a huge amount of data; their experiment therefore certainly constitutes the most representative one. All studies also indicate the (mean/median) inter-subject correlation[10] which varies from 0.48 (concept-pair based) in Zesch and Gurevych (2006) and 0.72 (concept-pair based) in Boyd-Graber et al. (2006) to 0.85 (word-pair based) in Budanitsky and Hirst (2006).

We think that this comparison of the various experiments points to two aspects which probably cause the large statistical spread shown in Table 5: the selection of the word-pairs (concept-pairs) and the type of relation (relatedness, similarity, evocation, distance). We assume that it should be possible to condense the comparison into one (more or less simple) rule: the narrower the relation concept (similarity $<$ relatedness $<$ evocation) and the narrower the data considered (lexical semantic selection rule $<$ any kind of selection rule $<$ random selection) the better the correlation between human judgment and semantic measure[11]. In any case, it seems essential to determine which relation types the subjects (knowingly or unknowingly) bear in mind when they judge word-pairs with respect to semantic relatedness. In order to achieve this goal and be able to integrate all relevant relations into the resources used for calculating semantic relatedness, the human judgments collected in the above-mentioned studies should be dissected into components (i.e. components for which systematic/unsystematic lexical semantic relations account etc.); such a decomposition certainly also helps render more precisely the definition of semantic relatedness.

Furthermore, it is — in our opinion — an unsettled issue whether the three types of semantic relation at hand, thus the relations

1. represented in a word net or corpus (both computed via semantic measure),

2. existing between any given word-pair in a text (which is mostly relevant for NLP-applications),

3. and the one assigned by subjects in a human judgment experiment

[10] The inter-subject correlation depends on various parameters, e.g. the complexity of the task, the subjects (and their background, age, etc.) as well as the experimental setup (task definition, training phase, etc.).

[11] ... and obviously the easier the task!

correspond at all. In principle, word nets, corpus statistics, and human judgments should be related to (theoretically even represent) the (at least partially) shared knowledge of humans about the underlying 'lexical semantic system', whereas relations between words in a concrete text represent an instantiation of a system. From this point of view, at least the human judgments should correspond to the semantics encoded in a word net (or corpus statistics). Instead of using human judgments as an evaluation resource (for e.g. word net based semantic measures), they might as well be directly integrated into the word net as a (preferably dense) layer of (potentially cognitively relevant, weighted but unlabeled) semantic relations, which is best adopted in Boyd-Graber et al. (2006), as summarized in Section 3. This approach has several advantages: firstly, the calculation of a semantic relatedness value is — given such a layer — trivial, since it merely consists in a look-up procedure. Secondly, NLP-applications using word nets as a resource would certainly benefit from the thus enhanced density of relations, i.e. cross-POS relations. Thirdly, an elaborate and standardized experimental setup for human judgment experiments could be used for the construction of such a layer in different languages (and domains) and would also guarantee the modeling quality. Finally, such a new word net layer would hopefully resolve the above mentioned open issue of the diverging correlation coefficients.

Alternatively, since it is completely unclear if the evocation relation can really act as a substitute for classical semantic relatedness measures in NLP-applications, current word nets should be enhanced by systematically augmenting existing relation types and integrating new ones. On that condition and given that a common evaluation framework exists, it should be possible to determine which semantic relatedness measure performs best under what conditions.

Last but not least, in order to determine the relation between an underlying semantic system (represented by a semantic measure or as mentioned above the evocation layer) and the instantiation of this system in a concrete text, a study similar to the one reported in Zesch and Gurevych (2006) should be conducted. Such a study probably also shows if the evocation relation is able to substitute (or at least complement) semantic relatedness measures typically used in NLP-applications.

5 Conclusions and Future Work

We have presented our own human judgment experiments for German and compared them with three similar studies. This comparison illustrates that the results of these studies are incompatible. We therefore argue that the experimental setup should be clarified and, if possible, harmonized. We also think that the notion of association should be considered carefully, since it is an established concept for measuring related phenomena in several psycholinguistic and computational linguistic communities.

We now plan to continue our work on three levels. Firstly, we intend to conduct a study similar to the one reported by Boyd-Graber et al. (2006) with a small amount of German data. We hope that this will provide us with insight into some of the open issues mentioned in Section 1 and Section 4. Secondly, we plan to investigate if the evocation relation is able to substitute the semantic relatedness measures typically used in lexical chaining and similar NLP-applications. Thirdly, we intend to run experiments using the database of noun associations in German constructed by Melinger and Weber (2006) as a resource for the evaluation of semantic relatedness measures.

Acknowledgements The author would like to thank Michael Beißwenger, Christiane Fellbaum, Sabine Schulte im Walde, Angelika Storrer, Tonio Wandmacher, Torsten Zesch, and the anonymous reviewers for their helpful comments. This research was funded by the DFG Research Group 437 (HyTex).

References

Baroni, M. and S. Bernardini (Eds.) (2006). *Wacky! Working papers on the Web as Corpus*. GEDIT, Bologna.

Barzilay, R. and M. Elhadad (1997). Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop*, pp. 10–17.

Boyd-Graber, J., C. Fellbaum, D. Osherson, and R. Schapire (2006). Adding dense, weighted, connections to wordnet. In *Proceedings of the 3rd Global WordNet Meeting*, pp. 29–35.

Budanitsky, A. and G. Hirst (2006). Evaluating wordnet-based measures of semantic relatedness. *Computational Linguistics 32 (1)*, 13–47.

Carthy, J. (2004). Lexical chains versus keywords for topic tracking. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pp. 507–510. Springer.

Cilibrasi, R. and P. M. B. Vitanyi (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering 19*(3), 370–383.

Cramer, I. and M. Finthammer (2008). An evaluation procedure for word net based lexical chaining: Methods and issues. In *Proceedings of the 4th Global WordNet Meeting*, pp. 120–147.

Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the IJCNLP 2005*, pp. 767–778.

Hirst, G. and D. St-Onge (1998). Lexical chains as representation of context for the detection and correction malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, pp. 305–332. The MIT Press.

Jiang, J. J. and D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of ROCLING X*, pp. 19–33.

Leacock, C. and M. Chodorow (1998). Combining local context and wordnet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, pp. 265–284. The MIT Press.

Lemnitzer, L. and C. Kunze (2002). Germanet - representation, visualization, application. In *Proceedings of the 4th Language Resources and Evaluation Conference*, pp. 1485–1491.

Lemnitzer, L., H. Wunsch, and P. Gupta (2008). Enriching germanet with verb-noun relations - a case study of lexical acquisition. In *Proceedings of the 6th International Language Resources and Evaluation*.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304.

Marrafa, P. and S. Mendes (2006). Modeling adjectives in computational relational lexica. In *Proceedings of the COLING/ACL 2006 poster session*, pp. 555–562.

Melinger, A. and A. Weber (2006). A database of noun associations in german. Online available database: http://www.coli.uni-saarland.de/projects/nag.

Miller, G. A. and W. G. Charles (1991). Contextual correlates of semantic similiarity. *Language and Cognitive Processes 6*(1), 1–28.

Novischi, A. and D. Moldovan (2006). Question answering with lexical chains propagating verb arguments. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 897–904.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the IJCAI 1995*, pp. 448–453.

Roth, M. and S. Schulte im Walde (2008). Corpus co-occurrence, dictionary and wikipedia entries as resources for semantic relatedness information. In *Proceedings of the 6th Conference on Language Resources and Evaluation*.

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

Schulte im Walde, S. and A. Melinger (2005). Identifying semantic relations and functional properties of human verb associations. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 612–619.

Wu, Z. and M. Palmer (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138.

Zesch, T. and I. Gurevych (2006). Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the Workshop on Linguistic Distances at COLING/ACL 2006*, pp. 16–24.