From Social Networks To Distributional Properties: A Comparative Study On Computing Semantic Relatedness

Ulli Waltinger (ulli_marc.waltinger@uni-bielefeld.de)
Text Technology, Bielefeld University

Irene Cramer (irene.cramer@udo.edu)
Faculty of Cultural Studies, TU Dortmund University

Tonio Wandmacher (tonio.wandmacher@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück

Abstract

In recent years a variety of approaches in computing semantic relatedness have been proposed. However, the algorithms and resources employed differ strongly, as well as the results obtained under different experimental conditions. This article investigates the quality of various semantic relatedness measures in a comparative study. We conducted an extensive experiment using a broad variety of measures operating on social networks, lexical-semantic nets and co-occurrence in text corpora. For two sample data sets we obtained human relatedness judgements which were compared to the estimates of the automated measures. We also analyzed the algorithms implemented and resources employed from a theoretical point of view, and we examined several practical issues, such as run time and coverage. While the performance of all measures is still mediocre, we could observe that in terms of of coverage and correlation distributional measures operating on controlled corpora perform best.

Keywords: Semantic Relatedness; Semantic Similarity; Human Judgement; Social Networks; WordNet; LSA;

Introduction

The computation of semantic relatedness (SR) has become an important task in many NLP applications such as spelling error detection, automatic summarization, word sense disambiguation, and information extraction. In recent years a large variety of approaches in computing SR has been proposed. However, algorithms and results differ depending on resources and experimental setup. It is obvious that SR plays a crucial role in the lexical retrieval of humans. In various priming experiments it could be shown that semantically related terms influence the semantic processing of one another. For example, if "*bread*" is primed by "*butter*" it is recognized more quickly. Moreover, many theories of memory are based on the notion of SR. The spreading activation theory of (Collins & Loftus, 1975) for example groups lexical items according to their SR in a conceptual graph. Similar ideas can be found in the ACT theory of Anderson (Anderson, 1983). The question that arises for us is, how this kind of relatedness can be determined by automatic means. In the literature the notion of SR is often confounded with semantic similarity; there is however a clear distinction between these notions. Two terms are semantically similar if they behave similarly in a given context and if they share some aspects of meaning (e.g. in the case of synonyms or hypernyms). On the other hand two terms can be semantically strongly related without behaving similarly. For example they can show a strong associative relationship (e.g. *ball - goal*), and they can be related across different linguistic categories (e.g. *milk - white, dog - bark*). With respect to the automatic computation of SR, however, many research questions remain unanswered. As stated above, many algorithms were presented in the past decade, but thorough evaluations and comparisons of their ability to capture SR in a human-like manner are still rare. In this work we therefore present a study comparing various semantic relatedness measures. We evaluate sixteen different algorithms involving four different resources based on a human judgement experiment, and we analyze the algorithms from a theoretical and practical point of view. The paper is organized as follows: the subsequent section describes two works representing the methodological basis for our study. The various semantic relatedness measures employed in our experiment are described in Section *Semantic Relatedness Measures*. The experimental setup as well as the results obtained are presented in Section *Evaluation*.

Related Work

The task of estimating SR between two given lexical items can be performed by humans in an effortless and intuitive manner. However, this notion is very difficult to formalize from a psycholinguistic or computational point of view. In terms of an evaluation of SR algorithms, most commonly human judgement experiments are conducted. The performance of an SR measure is determined by directly comparing the automatic computed results with those gained from the human judgements via correlation. As a most prominent example Budanitsky and Hirst (Budanitsky & Hirst, 2006) presented a comparison of five semantic relatedness measures for the English language. They recommended a three-level evaluation including theoretical examination, comparison with human judgements and evaluation with respect to a given NLP-application. The measures were evaluated on two different data sets: The first data set was compiled by Rubenstein and Goodenough (Rubenstein & Goodenough, 1965); it contained 65 word-pairs. The second set, containing 30 word pairs, was compiled by Miller and Charles (Miller & Charles, 1991). For each of the five measures, Budanitsky and Hirst reported correlation coefficients between 0.78 and 0.83. Boyd-Graber et al. (Boyd-Graber, Fellbaum, Osherson, & Schapire, 2006) presented a list of 120,000 concept pairs, which were rated by 20 subjects with respect to their *evocation* - how much one concept brings to mind the other. Volunteers were given manual instruction before the experiment and were trained on a sample of 1000 randomly selected pairs. However, it has to be pointed out that this approach focuses on constructing new relations within the lexical resource WordNet (Fellbaum, 1998) rather than assessing semantic relatedness. Still, four different semantic relatedness measures were compared. Results (correlation coefficients) ranged only between 0.008 and 0.131 Nevertheless, the approach of Boyd-Graber et al. (2006) is based to an important extent on human involvement. We argue that this is a crucial condition for all approaches that aim to model aspects of human lexical processing (such as computing SR), therefore we also make use of this kind of evaluation. Moreover, many works presenting new SR algorithms prove their accuracy with respect to one or two similar approaches. A large and standardized evaluation campaign is however still missing. For this reason we consider in our study a large variety of SR measures, and evaluate them with respect to a human judgement experiment (on German data), first presented by Cramer&Finthammer (Cramer, 2008).

Semantic Relatedness Measures

In general, we split all implemented algorithms on the basis of their resources into three different groups. Net-based measures make use of a lexical-semantic net like the already mentioned WordNet, which has been developed for many different languages (e.g. *GermaNet* (Lemnitzer & Kunze, 2002)). Most of the implemented algorithms use a hyponym-tree induced from the given lexical-semantic net. Since such a resource models only systematic semantic relations such as hyponymy or meronymy, unsystematic connections (i.e. associations) can not be directly computed. Distributional measures consider semantics on the basis of similar distributional properties of words in large text corpora. Such approaches deduce SR on the basis of co-occurences of features from text or web data. As a third group we regard social networks such as the online encyclopedia *Wikipedia*. Wikipedia driven approaches are able not only to comprise statistics from the entire text collection, but are also able to induce the category taxonomy using classical graph algorithms. The following three sections outline the sixteen different algorithms that we have implemented for our evaluation.

Net-based Measures

With the development of lexical-semantic nets (such as the Princeton WordNet) in the mid-1990's various measures for computing SR have been proposed. The eight most prominent algorithms were implemented using GermaNet (Lemnitzer & Kunze, 2002) as a resource.

- **Leacock-Chodorow** (Leacock & Chodorow, 1998): This measure computes the length of the shortest path between two synonym sets and scales it by the depth of the complete hyponym-tree.

  $$\mathrm{rel}_{\mathrm{LC}}(s_1,s_2) = -\log \frac{2\cdot \mathrm{sp}(s_1,s_2)}{2\cdot D_{Tree}} \tag{1}$$

  s_1 and s_2: the two synonym sets examined; $\mathrm{sp}(s_1,s_2)$: length of shortest path between s_1 and s_2 in the hyponym-tree; D_{Tree}: depth of the hyponym-tree

- **Wu-Palmer** (Wu & Palmer, 1994): The Wu-Palmer measure utilizes the least common subsumer in order to compute the similarity between two synonym sets in a hyponym-tree.

  $$\mathrm{rel}_{\mathrm{WP}}(s_1,s_2) = \frac{2\cdot \mathrm{depth}(\mathrm{lcs}(s_1,s_2))}{\mathrm{depth}(s_1)+\mathrm{depth}(s_2)} \tag{2}$$

  $\mathrm{depth}(s)$: length of the shortest path form root to vertex s; $\mathrm{lcs}(s)$: least common subsumer of s

- **Resnik** (Resnik, 1995): Given a hyponym-tree and a frequency list, the Resnik measure utilizes the information content in order to compute the similarity between two synonym sets.

  $$\mathrm{p}(s) := \frac{\sum_{w\in W(s)} \mathrm{freq}(w)}{TotalFreq} \tag{3}$$

  $$\mathrm{IC}(s) := -\log \mathrm{p}(s) \tag{4}$$

  $$\mathrm{rel}_{\mathrm{Res}}(s_1,s_2) = \mathrm{IC}(\mathrm{lcs}(s_1,s_2)) \tag{5}$$

  $\mathrm{freq}(w)$: frequency of a word within a corpus; $W(s)$: set of the synonym set s and all its direct/indirect hyponym synonym sets; *TotalFreq*: sum of the frequencies of all words in GermaNet; $\mathrm{IC}(s)$: information content of the synonym set s

- **Jiang-Conrath** (Jiang & Conrath, 1997): Given a hyponym-tree and a frequency list, the Jiang-Conrath measure computes the distance of two synonym sets.

  $$\mathrm{dist}_{\mathrm{JC}}(s_1,s_2) = \mathrm{IC}(s_1)+\mathrm{IC}(s_2) - 2\cdot \mathrm{IC}(\mathrm{lcs}(s_1,s_2)) \tag{6}$$

- **Lin** (Lin, 1998): Given a hyponym-tree and a frequency list, the Lin measure computes the semantic relatedness of two synonym sets.

  $$\mathrm{rel}_{\mathrm{Lin}}(s_1,s_2) = \frac{2\cdot \mathrm{IC}(\mathrm{lcs}(s_1,s_2))}{\mathrm{IC}(s_1)+\mathrm{IC}(s_2)} \tag{7}$$

- **Hirst-StOnge** (Hirst & St-Onge, 1998): This measure computes the semantic relatedness on the basis of the entire GermaNet graph structure. It classifies the relations considered into 4 classes: *extra strongly related*, *strongly related*, *medium strongly related*, and *not related*.

- **Tree-Path**: The tree-path measure computes the length of a shortest path between two synonym sets in a hyponym-tree.

  $$\mathrm{dist}_{\mathrm{Tree}}(s_1,s_2) = \mathrm{sp}(s_1,s_2) \tag{8}$$

- **Graph-Path**: The graph-path measure calculates the length of a shortest path between two synonym sets in the whole graph.

$$\text{dist}_{\text{Graph}}(s_1, s_2) = \text{sp}_{Graph}(s_1, s_2) \tag{9}$$

  $\text{sp}_{Graph}(s_1, s_2)$: Length of a shortest path between s_1 and s_2 in the GermaNet graph

Using GermaNet as a lexical resource in computing semantic relatedness, lexical-semantic relations such as hyponymy are considered only. However, it seems that for determining SR humans do not distinguish between systematic and unsystematic relations. Since GermaNet does not incorporate unsystematic relations, we expect all of the above measures to produce many false negatives; i.e. word pairs with low relation values in GermaNet, but strongly related ranked by humans.

Distributional Measures

Based on the assumption that words with similar distributional properties have similar meaning, distributional approaches infer semantic relatedness considering co-occurrences of words in text corpora. Distributional similarity can be determined in two major ways: One group of measures establishes relatedness on direct co-occurrence in text (1^{st} order relatedness). The other group aims to compare the similarity of contexts in which two terms occur (2^{nd} order relatedness). In 1^{st} order approaches, the co-occurrence probability of two terms is set in relation to the probability of the singular terms. In recent times a number of co-occurrence measures were proposed that use hit counts from large search engines (Google/Yahoo). We have implemented four different SR measures using hit counts:

- **Pointwise Mutual Information (PMI)**: The point wise mutual information (PMI) measure on hit counts for example can be defined as follows:

$$rel_{GPMI}(w_i, w_j) = \log M + \log \frac{hc(w_i, w_j)}{hc(w_i) \times hc(w_j)} \tag{10}$$

  where $hc(w_i)$, $hc(w_i, w_j)$ are the hit counts of a key word w_i or a word pair w_i, w_j and M is the total number of pages indexed by the search engine.

- **Google Quotient** (Cramer, 2008): It is defined as follows:

$$rel_{GQ}(w_i, w_j) = \frac{2 \cdot hc(w_i, w_j)}{hc(w_i) + hc(w_j)} \tag{11}$$

  Again, $hc(w_i)$, $hc(w_i, w_j)$ are the hit counts of a key word w_i or a word pair w_i, w_j.

- **Normalised Google Distance (NSD)** (Cilibrasi & Vitanyi, 2007):

$$rel_{NGD}(w_i, w_j) = \frac{max[\log hc(w_i) \log hc(w_j)] - \log hc(w_i, w_j)}{\log M - \min[\log hc(w_i), \log hc(w_j)]} \tag{12}$$

- **Normalised Wiki Distance (NSD Wiki)**: We adapted the approach of (Cilibrasi & Vitanyi, 2007), but restricted the corpus index to a social network by means of Wikpedia. That is, the normalised distance is derived on basis of the Apache Lucene index of Wikpedia articles only

Among the 2^{nd} order approaches one model has obtained particular attention, due to its success in a large variety of tasks involving semantic processing: *Latent Semantic Analysis* (LSA).

- **Latent Semantic Analysis (LSA)** (Deerwester, 1990): LSA is based on a term×context matrix A, displaying the occurrences of each word in each context. The decisive step in the LSA process is then a *singular value decomposition* (SVD) of the matrix, which enhances the contrast between reliable and unreliable relations. To measure the distance of the word vectors, the cosine measure is most often used, since it normalizes for length.

- **Semantic Vectors (Sem.Vec.)** (Widdows & Ferraro, 2008): The open source *Semantic-Vectors* package creates a word space model from a term-document matrix using positional indexing. Word similarity is performed by producing a query vector and calculate its distance to the term vectors (using the cosine).

The important advantage of 2^{nd} order approaches, is that they are usually better able to capture paradigmatic relations such as synonymy or hyponymy, since paradigmatically similar words tend to occur in similar contexts.

Wikipedia-based Measures

In terms of *Wikipedia* based semantic interpretation some approaches have been proposed which mainly focus either on the hyperlink structure (Milne, 2007), the vector space model (VSM) or on category concepts for graph-related measures (Ponzetto & Strube, 2006; Zesch, Gurevych, & Mühlhäuser, 2007). We have implemented three different algorithms using Wikipedia as a resource in computing semantic relatedness:

- **Explicit Semantic Analysis (ESA)** (Gabrilovich & Markovitch, 2007): This method represents term similarity by an inverted term-document index in a high-dimensional space of concepts derived from Wikipedia. In this case, concepts are represented by Wikipedia articles. Each concept corresponds to an attribute vector of terms occurring in the respective article (weighted by a TFIDF scheme (Salton & McGill, 1983)). Semantic relatedness of a pair of terms is computed by comparing the concept vector A with B using the cosine metric. Since Gabrilovich & Markovitch reported their results for experiments on English, we adopted their approach to the lemmatized German Wikipedia data set. We also removed small and overly specific concepts (articles having fewer than 100 words and fewer than 5 hyperlinks), leaving 126,475 articles for building the inverted index.

- **Wikipedia Graph-Path**: Given the entire Wikipedia hyperlink graph $G_w = (V, E)$, where Wikipedia articles denote a set of vertices V and hyperlinks between articles, and categories denote a set of edges $E \subseteq V^2$. The Wikipedia Graph-Path distance calculates the length of the shortest path (sp) between two articles in G_w.

$$distW_{Gw}(v1, v2) = \mathrm{sp}_{Gw}(v1, v2) \tag{13}$$

- **Category Concept Analysis (CCA)**:

  Given a lemmatized Wikipedia dump an inverted concept-term matrix is constructed. Different to Gabrilovich & Markovitch (2007), concepts are defined as Wikipedia categories, i.e. we assigned each article to its categories in Wikipedia. For term weighting the TFIDF scheme was used. Small concepts have been removed using a threshold value for a minimum length of the term vector (more than 400 lemmata). The relatedness computation was performed using the cosine metric, the dice coefficient and the jaccard similarity coefficient, utilizing a maximum length of 20,224 as the category concepts vectors (A and B).

Evaluation

Method

In order to assess the above mentioned algorithms, we evaluated their performance on the basis of a human judgement experiment. For the German language there are – to our knowledge – three human judgement data sets available: (1) a translation of the Rubenstein list (Gurevych, 2005), (2) a semi-automatically generated list compiled by Zesch & Gurevych (2006) and (3) two lists assembled by Cramer and Finthammer (2008) , comprising a total of 600 word pairs. Since the data sets of Cramer and Finthammer (2008) cover not only a wide range of relation types (random connections, associations, synonyms etc.), but also various degrees of relation strengths, we decided to use their lists to evaluate the algorithms implemented.[1]

The first test set (A) contains 100 word pairs (nouns of diverse semantic classes comprising abstract and concrete concepts). The test set B contains 500 randomized word pairs with not more than 20 % of collocations and associations. Set A was rated by 35 subjects and set B was rated by 75 subjects. Volunteers rated the word-pairs on a 5-level scale and received no further instruction (apart from estimating the semantic relatedness of the given terms).

Net-based measures The net-based measures were calculated on *GermaNet* v. 5.0 using *GermaNet Pathfinder v. 0.83*. Table 1 lists the correlations (Pearson) for test sets A and B, as well as the coverage and the average processing time per word pair[2].

Distributional measures The three web-based (1^{st} order) measures obtained their hit counts via the *Google* API; all counts were calculated beforehand and stored in a repository. The LSA word space was calculated using the *Infomap toolkit*[3] v. 0.8.6 on a newspaper corpus (Süddeutsche Zeitung) of 145 million words, which had been lemmatised by the lemmatizer presented in (Waltinger & Mehler, 2009). The co-occurrence matrix (window size: ±75 words) comprised 80.000×3.000 terms and was reduced by SVD to 300 dimensions. For the vector comparisons the cosine measure was applied. Table 2 shows the results (correlation, coverage and processing time) for all distributional measures tested.

Wikipedia-based measures The calculation of the Wikipedia measures is based upon the German version of Wikipedia (october 2007). The *Semantic Vector* package[4] utilizes the *Apache Lucene* library. Explicit Semantic, Category Concept Analysis and Wikpedia Graph Path are implemented in C++ using *Trolltech Qt*. For both *Category Concept Analysis* and *Explicit Semantic Analysis* we had to reduce the matrices on the lemma-dimension for computational reasons, i.e. when building the matrix we excluded those lemmata whose corpus frequency did not exceed a threshold of 300. Building the *Normalized Search Distance* measures, we have directly connected to the special page *search* of Wikipedia (http://de.wikipedia.org/wiki/Spezial:Suche). Furthermore, we also calculated an LSA word space on Wikipedia; however, due to computational limitations we had to utilize a subcorpus only, by taking the first 800 words of each article (148 mill. words in total). Table 3 lists the results for all Wikipedia-based measures.

Results

Comparing the correlation results shown in Tables 1, 2 and 3 it can be observed that the net-based measures have the lowest scores (r= 0.11 - 0.48); interestingly they score quite similar within one test set, despite their rather different calculation. For the distributional measures a clear difference can be seen between the three web-based techniques (0.27 - 0.37) and the LSA results (scoring up to 0.64); this may either be due to the fact that LSA (being a 2^{nd} order approach) is able to establish more paradigmatic relations such as synonymy or hyponymy, or the hit counts, obtained from *Google* are not sufficiently precise indicators of co-occurrence. Among the Wikipedia measures the *WikiSearch Distance* scores significantly better than the other measures (up to 0.69). A second observation of the results concerns the differences between the correlations of the test sets A and B. Especially the net-based measures, but also most of the Wikipedia-based show significantly worse correlations for set B. Recalling that set B contains a large part of random word pairs (80%), a probable explanation is that such measures tend to overestimate relatedness, i.e. they cannot well discriminate between related

[1] See (Cramer & Finthammer, 2008) for detailed information about the experiment and the data sets.

[2] The computation was performed on an AMD Athlon XP 2400+, 2,0 GHz and 1GB of RAM.

[3] http://infomap-nlp.sourceforge.net/

[4] http://code.google.com/p/semanticvectors/

WordNet-based measures

Test set	Leacock & Chodorow	Wu & Palmer	Resnik	Jiang & Conrath	Lin	Hirst & St-Onge	Tree path	Graph path
r Set *A*	0.48	0.36	0.44	0.46	0.48	0.47	0.41	0.42
r Set *B*	0.17	0.21	0.24	0.25	0.27	0.32	0.11	0.31
Coverage	86.9%	86.9%	86.9%	86.9%	86.9%	86.9%	86.9%	86.9%
t/pair (ms)	<10	<10	<10	<10	<10	1110	<10	3649

Table 1: Correlations (Pearson), coverage and processing time per pair of the net-based measures tested

Test set	PMI *Google*	*Google* Quotient	NSD *Google*	LSA (newspaper)
r Set *A*	0.37	0.27	0.37	0.64
r Set *B*	0.34	0.31	0.36	0.63
Coverage	100%	100%	100%	87.0%
t/pair (ms)	<10	<10	<10	<10

Table 2: Correlations (Pearson), coverage and processing time per pair of the distributional measures tested

and unrelated word pairs. The differences between the approaches tested clearly show how important the influence of the resource is. One conclusion that may be drawn from our results is that for determining SR a small, hand-crafted and structured resource such as a lexical-semantic net is clearly inferior to a large and semi-structured (Wikipedia) or even completely unstructured resource (plain text). With respect to coverage, the web-based measures (including the *WikiSearch Distance* clearly outperform all other approaches. This is not astonishing, given the fact that they operate on the largest vocabulary available. The off-line approaches on the other hand are not as sparse as one might have imagined, the lowest scores are still over 75%, and the net-based as well as the LSA approach achieve a coverage of approximately 87%. The processing time (per word pair) however differs quite strongly. It is also to be taken with a grain of salt, since it depends on the implementation chosen. Most of the approaches show almost negligible processing times (<10 ms), however if complex tree or graph traversals are involved (e.g. *GermaNet* or *Wiki graph path*), the processing times can reach up to several seconds. Comparing all these different measures and resources, we observed that the distributional measures, especially those based on a 2^{nd} order approach (such as LSA), perform significantly better than the net-based measures and those using explicit categorial information (ESA, CCA). We therefore conclude that the use of explicit structural information, in the form of semantic links, categories or of hyperlink graphs, establishes semantic relatedness not as well as distributional information. Secondly we could clearly see that the choice of the resource plays an important role. Interestingly, those measures using the web as a corpus were inferior to those operating on smaller but better controlled training corpora (cf. the important difference between the web-based and the wikipedia-based NSD). With respect to corpus choice we can conclude that quality is more important than quantity, an observation which is in line with Kilgarriff (2007). Considering all the results above it can be stated that the calculation of semantic relatedness is far from being solved. Each of the resources that we used certainly captures an important part of lexical meaning; however, it seems that this is not yet sufficient for describing the complex nature of SR between any two terms. Secondly, a factor that we disregarded in our study is the influence of context. It is quite obvious that SR is not a static and independent size. On the contrary, it is dynamically interrelated with the current lexical, syntactic and semantic context, and a proper theory of (or algorithm computing) SR will have to take it into account.

Conclusions and Future Work

We presented a study comparing sixteen different SR measures on various lexical resources. The measures made use of information from lexical-semantic nets, co-occurrence distribution and the structure as well as the content of a large social network (Wikipedia). We conducted an extensive evaluation on the basis of a human judgement experiment. Morever, the implemented algorithms and employed resources were analyzed with respect to practical issues such as run time and resource coverage. In terms of coverage and correlation we could observe that distributional measures perform best, however, results show that even the best performing algorithms leave a lot of room for improvement. For the future, we want to propose the definition of a shared task which might bring us considerably closer to results of high performance but also to better understand the complex characteristics of SR.

Acknowledgement

Financial support of the German Research Foundation (DFG) through the Excellence Cluster 277 *Cognitive Interaction Technology*, the Research Group 437 *Text Technological Information Modeling* and the LIS-Project *P2P-Agents for Thematic Structuring and Search Optimisation in Digital Libraries* at Bielefeld University is gratefully acknowledged.

Test	NSD		Sem. Vec.		Wiki Graph	LSA
set	(Wiki)	CCA	(Wiki)	ESA	Path	(Wiki)
r Set *A*	0.69	0.57	0.51	0.52	0.49	0.65
r Set *B*	0.61	0.36	0.28	0.44	0.37	0.57
Coverage	100%	79.8%	99.1%	75.9%	92.0%	83.8%
t/pair (ms)	850	<10	1299	240	2301	<10

Table 3: Correlations (Pearson), coverage and processing time per pair of the Wikipedia-based measures tested

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Leaning and Verbal Behaviour*, *22*, 261–295.

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

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

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

Collins, A., & Loftus, E. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407-428.

Cramer, I. (2008). How well do semantic relatedness measures perform? a meta-study. In *Proceedings of the symposium on semantics in systems for text processing.*

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

Deerwester, S. (1990). Indexing by latent semantic analysis. *J. Ameri. Soci. Inf. Sci*, *41*(6), 391407.

Fellbaum, C. (Ed.). (1998). *Wordnet. an electronic lexical database*. The MIT Press.

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 6–12.

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., & St-Onge, D. (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., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of rocling x* (pp. 19–33).

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, *33*(1), 147–151.

Leacock, C., & Chodorow, M. (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., & Kunze, C. (2002). Germanet - representation, visualization, application. In *Proceedings of the 4th language resources and evaluation conference* (pp. 1485–1491).

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on machine learning* (pp. 296–304).

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

Milne, D. (2007). Computing semantic relatedness using wikipedia link structure. In *Proc. of nzcsrsc07.*

Ponzetto, S., & Strube, M. (2006, July). Wikirelate! computing semantic relatedness using wikipedia.

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

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

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw Hill.

Waltinger, U., & Mehler, A. (2009). *Web as preprocessed corpus: Building large annotated corpora from heterogeneous web document data.* In preparation.

Widdows, D., & Ferraro, K. (2008, May). Semantic vectors: a scalable open source package and online technology management application. In E. L. R. A. (ELRA) (Ed.), *Proceedings of the sixth international language resources and evaluation (lrec'08).* Marrakech, Morocco.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the association for computational linguistics* (pp. 133–138).

Zesch, T., & Gurevych, I. (2006). Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the workshop on linguistic distances at coling/acl 2006* (pp. 16–24).

Zesch, T., Gurevych, I., & Mühlhäuser, M. (2007). Comparing wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets. In *In proc. of naacl-hlt.*