

Exploring Resources for Lexical Chaining: A Comparison of Automated Semantic Relatedness Measures and Human Judgments

Irene Cramer, Tonio Wandmacher, and Ulli Waltinger

Abstract In the past decade various semantic relatedness, similarity, and distance measures have been proposed which play a crucial role in many NLP-applications. Researchers compete for better algorithms (and resources to base the algorithms on), and often only few percentage points seem to suffice in order to prove a new measure (or resource) more accurate than an older one. However, it is still unclear which of them performs best under what conditions. In this work we therefore present a study comparing various relatedness measures. We evaluate them on the basis of a human judgment experiment and also examine several practical issues, such as run time and coverage. We show that the performance of all measures – as compared to human estimates – is still mediocre and argue that the definition of a shared task might bring us considerably closer to results of high quality.

1 Motivation

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 in-

Irene Cramer
Faculty of Cultural Studies, TU Dortmund University, e-mail: irene.cramer@udo.edu

Tonio Wandmacher
Institute of Cognitive Science, University of Osnabrück, e-mail: tonio.wandmacher@uni-osnabrueck.de

Ulli Waltinger
Text Technology, Bielefeld University, e-mail: ulli.marc.waltinger@uni-bielefeld.de

fluence the semantic processing of each other (e.g. 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 [12] for example groups lexical items according to their SR in a conceptual graph. Similar ideas can be found in Anderson's ACT theory [2].

The question that we want to discuss here 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 terms. 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 over 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 SR measures. We evaluate sixteen different algorithms involving four different resources based on a human judgment experiment, and we analyze the algorithms from a theoretical and practical point of view.

We perform this evaluation in the context of *lexical chaining*, a task that aims to determine sequences of terms which are semantically interrelated in a given text. Such term sequences (or chains) represent an important intermediate structure for purposes involving higher-order semantic analysis.

The following sections are organized as follows: The principle concepts of lexical chaining are introduced in Section 2. Section 3 outlines related work by means of two different human judgment experiments. The various SR measures used in our experiment are described in Section 4. The results are presented in Section 5.

2 Lexical Chaining

Based on the notion of lexical *cohesion*, as described by Halliday and Hasan in 1976 [20], computational linguists, e.g. Morrish and Hirst [32], developed in the 1990s a method to compute partial text representations: *lexical chains*. To illustrate the idea of lexical chains, an annotated text passage is given in Figure 1. These chains (e.g. *sit down - rest - tired - fall asleep*) consist of semantically related terms, and they describe the cohesive structure of a given text. They can be constructed automatically by linking lexical items with respect to the SR holding between them. Many approaches of lexical chaining employ a lexical-semantic resource such as Princeton *WordNet* (cf. [16], which has been used in the majority of cases, e.g. [21],

[18], [45]), *Rogert's Thesaurus* (e.g. [32]), or the open encyclopedia *Wikipedia* and its offshoot *Wiktionary*¹ (e.g. [51]).

However, since the construction of lexical chains does not necessarily depend on explicit relation types, distributional SR measures (such as PMI or LSA, cf. section 4.2) represent an alternative resource for the calculation of lexical chains.

Jan sat down to rest at the foot of a huge beech-tree. Now he was so tired that he soon fell asleep; and a leaf fell on him, and then another, and then another, and before long he was covered all over with leaves, yellow, golden and brown.

Chain 1: sat down, rest, tired, fell asleep

Chain 2: beech-tree, leaf, leaves

Unsystematic relations not yet considered in resource for lexical chaining: foot / huge – beech-tree; yellow / golden / brown – leaves

Fig. 1 Chaining example adapted from Halliday and Hasan's work [20]

A variety of NLP-applications, namely text summarization (e.g. [4], [44]), malapropism recognition (e.g. [21]), automatic hyperlink generation (e.g. [18]), question answering (e.g. [35]), and topic detection or tracking (e.g. [8]), benefit from lexical chaining as a valuable resource and preprocessing step.

In order to formally evaluate the performance of a lexical chaining system, a standardized test set would be required. However, the output of a chainer is normally assessed with respect to an application; although in several works more general evaluation criteria have been proposed (e.g. [6], [14]), no consensus among researchers could yet be achieved, and, consequently, the different sets of criteria have not yet been systematically applied.

3 Related Work

Lexical cohesion constitutes a concept, which can be observed practically in every text or discourse; while it is easy to approach intuitively, it is rather difficult to detect and analyze in a formal manner. Although many researchers of different scientific communities have proposed various approaches intended to formalize and compute structures of lexical cohesion, there are nevertheless only few prominent strands of research which differ especially with respect to the names attributed to the phenomenon, the features and types of relations subsumed, and last but not least the

¹ <http://www.wikipedia.org>

methods to deal with it. We argue that all these works might best be split into two groups: firstly, research intended to understand, describe, and finally formalize the underlying concepts (cf. [33]), and secondly, studies mainly focused on technical aspects, namely efficient algorithms (cf. [44]) and promising resources (cf. [51]). In the following we describe in depth one prominent work of each of the two strands of research, in order to illustrate the central issues under discussion.

3.1 *Budanitsky and Hirst*

Budanitsky and Hirst’s work [6] aims at an extensive comparison of the performance of various SR measures, i.e. different algorithms. For this purpose, Budanitsky & Hirst indicate 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. In their opinion the second and third method are the most appropriate ones; they therefore focus on them in the empirical work presented. As a basis for the second evaluation method, i.e. the comparison between SR measures and human judgments, they use two lists of word pairs: the first has been compiled by Rubenstein and Goodenough [39] and contains 65 word pairs², while the second, containing 30 word pairs, has been created by Miller and Charles [30]. In order to evaluate the performance of five different measures, Budanitsky and Hirst [6] compute respective relatedness values for the word pairs, and they compare them with the human judgments. In this way they determine the correlation coefficients summarized in Table 1.

Table 1 Correlation Coefficients by Budantisky and Hirst

r	Leacock-	Hirst-	Jiang-		
	Chodorow	St-Onge	Resnik	Conrad	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	.78	0.82	0.82

In examining the results of this comparison, Budanitsky & Hirst identify several limitations of this evaluation method: i.e. they stress that the amount of data available (65 word pairs) is inadequate for real NLP-applications, however, the development of a large-scale data set would be time-consuming and expensive (cf.

² Rubenstein & Goodenough [39] 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 [30] 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 [39].

Section 5.1). Moreover, they argue that the experiments by Rubenstein and Goodenough [39] as well as Miller and Charles [30] focus on relations between words rather than word-senses (concepts), which would be more appropriate for most NLP-applications. Nevertheless, they consider it difficult to trigger a specific concept without biasing the subjects.

3.2 *Boyd-Graber et al.*

In contrast to the above-mentioned experiments by Budanitsky and Hirst [6], the research reported by Boyd-Graber et al. [5] strives for the development of a new, conceptually different layer of word net relation types and is motivated by three widely acknowledged, yet unsolved challenges:

- The lack of cross-POS links connecting the noun, verb, and adjective sub-graphs respectively.
- The low density of relations in the sub-graphs, i.e. potentially missing relation types.
- The absence of weights assigned to the relations, i.e. representing the degrees of semantic distance.

Unlike Rubenstein and Goodenough [39] or Miller and Charles [30], Boyd-Graber et al. do not restrict themselves to systematic relation types but introduce the concept of *evocation*³. While systematic relations are well defined, evocation seemingly represents a diffuse accumulation of various aspects, which intuitively account for lexical cohesion but can only in parts be precisely characterized.

Table 2 Correlation Coefficients by Boyd-Graber et al. [5]

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

In their experiment, Boyd-Graber et al. asked 20 subjects to rate evocation in 120,000 pairs of words (these pairs form a random selection of all possible word pairs stemming from 1000 core synsets in WordNet). The subjects were given a detailed manual explaining the task, and they were trained on a sample of 1000 (two sets of 500) randomly selected pairs. Although the research objective of their work is to construct a new layer of relations for WordNet rather than to evaluate SR measures, Boyd-Graber et al. compare the results of their human judgment experiment

³ Boyd-Graber et al. define this term in a rather loose sense as “*how much one concept evokes or brings to mind the other*”.

with the relatedness values of four different semantic measures. The correlation coefficients of this comparison are summarized in Table 2.

Boyd-Graber et al. arrive at the conclusion that – given the obvious lack of correlation (cf. Table 2) – 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).

While the first work mentioned above discusses in detail various SR algorithms, provides a survey of evaluation methods, and finally presents their applicability from a technical point of view, the second additionally sheds light on the linguistic and psycholinguistic aspects of the set-up of a human assessment experiment and the comparison between relatedness values and human judgments.

4 Semantic Relatedness Measures

4.1 Net-based Measures

The following eight SR measures draw on a lexical-semantic net like Princeton *WordNet* or its German counterpart *GermaNet* [26]. Although all of these measures are based on the same resource, they use different features (some additionally rely on a word frequency list⁴) and therefore also cover different aspects of SR.

Most of the relatedness measures mentioned in this section are continuous, with the exception of *Hirst-StOnge*, *Tree-Path*, and *Graph-Path* which are discrete.

All of the measures range in a closed interval between 0 (not related) and a maximum value (mostly 1), or they can be normalized: the distance value calculated by the three distance measures (*Jiang-Conrath*, *Tree-Path*, and *Graph-Path*) is mapped into a closed range relatedness value by subtracting it from the theoretical maximum distance.

The first four measures use a hyponym-tree induced from a given lexical-semantic net, i.e. all other edges except the hyponym links are disregarded. The resulting unconnected trees are subsequently reconnected by an artificial root in order to construct the required hyponym-tree.

- **Leacock-Chodorow** [25]: Given a hyponym-tree, this measure computes the length of the shortest path between two synonym sets and scales it by the depth of the complete tree.

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

⁴ We used a word frequency list computed by Dr. Sabine Schulte im Walde on the basis of the *Huge German Corpus* (see <http://www.schulteimwalde.de/resource.html>). We thank Dr. Schulte im Walde for kindly permitting us to use this resource in the framework of our project.

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

- **Wu-Palmer** [48]: Given a hyponym-tree, the *Wu-Palmer* measure utilizes the least common subsumer in order to compute the similarity between two synonym sets. The least common subsumer is the deepest vertex which is a direct or indirect hypernym of both synonym sets.

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

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

- **Resnik** [38]: Given a hyponym-tree and the frequency list mentioned above, the *Resnik* measure utilizes the information content in order to compute the similarity between two synonym sets. As typically defined in information theory, the information content is the negative logarithm of the probability. Here the probability is calculated on the basis of subsumed frequencies. A subsumed frequency of a synonym set is the sum of frequencies of the set of *all* words which are in this synonym set, *or* a direct or indirect hyponym synonym set.

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

$$\text{IC}(s) := -\log p(s) \quad (4)$$

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

$\text{freq}(w)$: 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 the respective lexical-semantic net; $\text{IC}(s)$: information content of the synonym set s

- **Jiang-Conrath** [22]: Given a hyponym-tree and the frequency list mentioned above, the *Jiang-Conrath* measure computes the distance (as opposed to similarity) of two synonym sets. The information content of each synonym set is included separately in this distances value, while the information content of the least common subsumer of the two synonym sets is subtracted.

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

- **Lin** [28]: Given a hyponym-tree and the frequency list mentioned above, the *Lin* measure computes the SR of two synonym sets. As the formula clearly shows, the same expressions are used as in *Jiang-Conrath*. However, the structure is different, as the expressions are divided and not subtracted.

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

- **Hirst-StOnge** [21]: In contrast to the four above-mentioned methods, the *Hirst-StOnge* measure computes the semantic relatedness on the basis of the whole graph structure. It classifies the relations considered into the following four classes: *extra strongly related* (the two words are identical), *strongly related* (the two words are e.g. synonym or antonym), *medium strongly related* (there is a relevant path between the two), and *not related* (there is no relevant path between the two). The relatedness values in the case of extra strong and strong relations are fixed values, whereas the medium strong relation is calculated based on the path length and the number of changes in direction.
- **Tree-Path** (Baseline 1): Given a hyponym-tree, the simple *Tree-Path* measure computes the length of a shortest path between two synonym sets. Due to its simplicity, the Tree-Path measure serves as a baseline for more sophisticated similarity measures.

$$\text{dist}_{\text{Tree}}(s_1, s_2) = \text{sp}(s_1, s_2) \quad (8)$$

- **Graph-Path** (Baseline 2): Given the whole graph structure of a lexical-semantic net, the *Graph-Path* measure calculates the length of a shortest path between two synonym sets in the whole graph, i.e. the path can make use of all relations available. Analogous to Tree-Path, the Graph-Path measure gives us a very rough baseline for other relatedness measures.

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

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

In the task of determining SR, humans do not seem to distinguish between systematic relations (e.g. synonymy or hyponymy) and unsystematic one: either a pair of words is (more or less) related or it is not [cp. [30] and [34]. However, a net like *GermaNet* only models systematic relations such as hyponymy or meronymy. Unsystematic (i.e. associative) connections are not directly taken into account in any of the measures mentioned above. We therefore expect all of them to produce many false negatives, i.e. low relation values for word pairs which are judged by humans to be (strongly) related.

4.2 Distributional Measures

A recent branch of lexical semantics aims to exploit statistics of word usage to derive meaning. Based on the assumption that words with similar distributional properties have similar meanings, such approaches infer semantic relatedness from the co-occurrence of words in text corpora. Distributional similarity can be defined in (at least) two ways: One group of measures establishes relatedness on direct co-occurrence in text (1^{st} order relatedness); many of these measures can be related to standard statistical tests. The other group aims to compare the similarity of contexts in which two terms occur (2^{nd} order relatedness); such measures usually operate on

the vector space. The following gives an overview of the different 1st and 2nd order measures is given:

- **Pointwise Mutual Information:** A typical representative of a 1st order measure is *pointwise mutual information* (PMI) [10]. Here, the co-occurrence probability of two terms is set in relation to the probability of the singular terms. In [37] and [46] it could be shown that 1st order measures are able to determine semantically related terms, even though the relations tend to be of syntagmatic nature.

$$rel_{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i) \times P(w_j)} \quad (10)$$

where $P(w_i)$, $P(w_i, w_j)$ is the probability estimate of a key word w_i or a word pair w_i, w_j . Since the reliability of the collected parameters usually grows with the size of the training corpus (cf. for example [7]), it was also proposed to use the web as a corpus. In such settings *hit counts* (number of pages found for a given query) are determined from a search engine for a given query (word or set of words) then replace the probabilities.

- **Normalized Search Distance (NSD)** [11]: This measure is inherently based upon the idea of using hit counts from a search engine. As the web-based PMI measure, NSD is calculated from the singular ($hc(w_i)$) and the joined ($hc(w_i, w_j)$) hit counts as well as the total number of pages M .

$$rel_{NSD}(w_i, w_j) = \frac{\max[\log hc(w_i) \log hc(w_j)] - \log hc(w_i, w_j)}{\log M - \min[\log hc(w_i), \log hc(w_j)]} \quad (11)$$

- **Google Quotient:** Another measure has been proposed by [13], the *Google quotient*. It is defined as follows:

$$rel_{GQ}(w_i, w_j) = \frac{2 \cdot hc(w_i, w_j)}{hc(w_i) + hc(w_j)} \quad (12)$$

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 .

- **Latent Semantic Analysis (LSA)** [15]: Among the 2nd order approaches *Latent Semantic Analysis* has obtained particular attention, due to its success in a large variety of tasks involving semantic processing. When it was first presented by Deerwester et al. [15], it aimed mainly at improving the vector space model in information retrieval (cf. [40]), but in the meantime it has become a helpful tool in NLP as well as in cognitive science (cf. [24]). As the vector space model, LSA is based on a term×context matrix A , displaying the occurrences of each word in each context. When only term relationships are considered, a slightly different setting, as described by Schütze [42] and Cederberg & Widdows [9] is more appropriate; here the original matrix is not based on occurrences of terms in documents, but on other cooccurring terms (term×term-matrix). We thus count the frequency with which a given term occurs with others in a predefined context window ($\pm 10 - 100$ words).

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. The high-dimensional input matrix is hereby reduced to a subspace of k dimensions ($k \approx 100$ to 300). After applying SVD, each word is represented as a k -dimensional vector, and for every word pair w_i, w_j of our vocabulary we can calculate a relatedness value $rel_{LSA}(w_i, w_j)$, based on the *cosine* measure. The cosine of the angle between any two vectors \mathbf{w}_i and \mathbf{w}_j of dimensionality m with components w_{ik} and w_{jk} , $k \leq m$ is defined as follows:

$$\cos(\mathbf{w}_i, \mathbf{w}_j) = \frac{\sum_{k=1}^m w_{ik}w_{jk}}{\sqrt{\sum_{k=1}^m w_{ik}^2 \sum_{k=1}^m w_{jk}^2}} \quad (13)$$

Since the denominator normalizes the vector length, frequency influences are leveled out. In addition, the result becomes standardized ($[-1; 1]$), which facilitates further comparisons.

- **Semantic Vectors (Sem.Vec.)** [1]: The open source *Semantic-Vectors* package⁵ creates a word space model from a term-document matrix using a random projection algorithm. It is supposed to perform similarly to techniques like LSA but it does not rely on complex procedures such as SVD, making it a more scalable technique. Word similarity is performed by producing a query vector and calculating its distance to the term vectors (using the cosine).

The important advantage of 2^{nd} order approaches is that they are better able to capture paradigmatic relations such as synonymy or hyponymy, since paradigmatically similar words tend to occur in similar contexts. However, they also have a disadvantage with respect to direct co-occurrence measures, because the matrix computations are computationally demanding, so that they cannot be performed online. This means that usually far smaller training corpora must be used.

4.3 Wikipedia-based Measures

With regard to *Wikipedia*-based SR computation, some approaches have been proposed which mainly focus either on the hyperlink structure [31], the vector space model (VSM), or on category concepts for graph related measures [36, 50]. We have implemented three different algorithms using *Wikipedia* as a resource in computing semantic relatedness:

- **Explicit Semantic Analysis (ESA)** [17]: 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 defined as *Wikipedia* articles.

⁵ <http://code.google.com/p/semanticvectors/>

Each concept is represented as an attribute vector of terms occurring in the corresponding article (weighted by a *tf.idf* scheme [41]). Semantic relatedness of a pair of terms is computed by comparing their respective concept vectors using the cosine metric (cf. equation 13). We have adopted the approach of Gabrilovich and Markovitch [17] to the German Wikipedia data (lemmatized). We have also removed small and overly specific concepts (articles having fewer than 100 words and fewer than 5 hyperlinks), leaving 126,475 articles on which the inverted index was built.

- **Wikipedia Graph-Path:** This measure operates on the Wikipedia hyperlink graph $G_w = (V, E)$, where Wikipedia articles denote a set of vertices V , and hyperlinks between articles, and the 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 .

$$distW_{G_w}(v1, v2) = sp_{G_w}(v1, v2) \quad (14)$$

- **Category Concept Analysis (CCA):** For this measure an inverted concept-term matrix is constructed on the full Wikipedia corpus (lemmatized). In contrast to [17], concepts are defined as Wikipedia categories, i.e. we assigned each article to its categories in Wikipedia. For term weighting the *tf.idf* scheme was used. Small articles have been removed using a threshold value for a minimum length of the termvector (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).

5 Evaluation

5.1 Method

In order to evaluate the quality of a SR measure, a set of pre-classified word pairs is needed. As mentioned above, in previous work on English data, most researchers used the word-pair list by Rubenstein and Goodenough [39] as well as the list by Miller and Charles [30] as an evaluation resource. For German there are - to our knowledge - two research groups, who have compiled lists of word-pairs with respective human judgment: Gurevych et al. constructed three lists (a translation of Rubenstein and Goodenough's list [19], a manually generated set of word pairs, and a semi-automatically generated one [49]).

Cramer and Finthammer (cf. [14], [13]) compiled two lists of word pairs for which they obtained human judgments.⁶ We make use of these two lists by Cramer and Finthammer, since they cover a wide range of relatedness types, i.e. systematic and unsystematic relations, and relatedness levels, i.e. various degrees of relation strength. However, they only include nouns, since cross-part-of-speech (cross-POS) relations can be considered to be an additional challenge⁷. In order to better understand the impact (and the potentially included bias) of the construction method of a list, two different methods were applied for the compilation of the word pairs.

For the first list nouns were collected manually from diverse semantic classes, e.g. abstract nouns, such as *Wissen* (Engl. knowledge), and concrete nouns, such as *Bügeleisen* (Engl. *flat-iron*; cf. [14] and [13] for further information). This list of 100 word pairs represents our test set A.

A different method was applied for the second list (set B): firstly, word pairs which are part of collocations were again manually collected, i.e. the two nouns *Rat* and *Tat* (*mit Rat und Tat helfen*; "to help with words and deeds") or *Qual* and *Wahl* (*die Qual der Wahl haben*; "to be spoilt for choice"). Secondly, word pairs featuring association relations were assembled, i.e. *Afrika* ('Africa') and *Tiger* ('tiger') or *Weihnachten* ('christmas') and *Zimt* ('cinnamon'). Thirdly, a list of random word pairs was automatically constructed using the *Wacky* corpus [3] as a resource; *ad hoc* constructions were manually excluded. Finally, out of these three resources a set of 500 pairs of words was compiled with no more than 20% of the collocation and association word pairs.

Subjects were asked 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. Thus, in contrast to the experimental set-up by Boyd-Graber et al., subjects had no manual, and they were not trained beforehand. Set A was rated by 35 subjects and set B was rated by 75 subjects. For each word pair a human judgment score was calculated by averaging the singular judgments of all subjects. Secondly, the Pearson correlation coefficients were calculated comparing the human scores with each of the measures on the test sets A and B.

⁶ Cramer and Finthammer actually worked on 6 separate lists; we merged them into two according to their method of compilation.

⁷ 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. [29] or [27]). However, these new relations are not yet integrated on a large scale and therefore should not (or even cannot) be used in SR measures. Furthermore, calculating SR between words with different POS might introduce additional challenges potentially as yet unidentified, which calls for a careful exploration.

5.2 Results

Net-based measures

The net-based measures were calculated on *GermaNet* v. 5.0 using *GermaNet Pathfinder* v. 0.83⁸. Table 3 lists the correlations (Pearson) for test sets A and B, as well as the coverage (percentage of word pairs for which a measure could be calculated) and the average processing time per word pair⁹.

Table 3 Correlations (*Pearson* coeff. to human estimates), coverage, and processing time per pair of the GermaNet-based measures tested

Test set	WordNet-based measures							
	Leacock & Chodorow	Wu & Palmer	Jiang & Resnik	Jiang & Conrath	Hirst & Lin	Tree St-Onge	Graph 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

Distributional measures

The three web-based (first 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*¹⁰ v. 0.8.6 on a newspaper corpus (*Süddeutsche Zeitung*) of 145 million words, which had been lemmatized¹¹. The co-occurrence matrix (window size: ± 75 words) comprised $80,000 \times 3,000$ terms and was reduced by SVD to 300 dimensions. For the vector comparisons the cosine measure was applied. Table 4 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* package¹² utilizes the *Apache Lucene* library. *ESA* and *Graph Path* are implemented in C++ using *Trolltech Qt*.

⁸ http://www.hytext.info/030_ergebnisse/030_tools/index_eng.html

⁹ The computation was performed on an AMD Athlon XP 2400+, 2.0 GHz and 1GB of RAM.

¹⁰ <http://infomap-nlp.sourceforge.net/>

¹¹ We used our own lemmatizer, to be described in [?]

¹² <http://code.google.com/p/semanticvectors/>

Table 4 Correlations (*Pearson* coeff. to human estimates), coverage, and processing time per pair of the distributional measures tested

Test set	PMI <i>Google</i>	<i>Google</i> Quotient	NSD <i>Google</i>	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

For both *CCA* and *ESA* 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 certain threshold (>300). Building the *NSD* measures, we have directly connected to the special page *search* of Wikipedia (<http://de.wikipedia.org/wiki/Spezial:Suche>).

Furthermore, we calculated an LSA word space on Wikipedia, again on an $80,000 \times 3,000$ -matrix using a window of ± 75 terms; however, due to computational limitations we had to use only a subcorpus, by taking the first 800 words of each article (148 mill. tokens in total). Table 5 lists the results for all Wikipedia-based measures.

Table 5 Correlations (*Pearson* coeff. to human estimates), coverage and processing time per pair of the Wikipedia-based measures tested

Test set	NSD (Wiki)	CCA	Sem. Vec. (Wiki)	ESA	Wiki Graph Path	LSA (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

Comparing the correlation results shown in Tables 3, 4, and 5, it can be observed that the net-based measures show rather low coefficients ($r = 0.11 - 0.48$); interestingly they score quite similarly 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 is a 2nd order approach, being able to establish more paradigmatic relations, or the hit counts, obtained from *Google* are insufficiently precise indicators of co-occurrence. Among the Wikipedia measures the *WikiSearch Distance* scores significantly better than the others (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 fraction of random word pairs (80%), a probable expla-

nation is that such measures tend to overestimate relatedness, i.e. they cannot well discriminate between related 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 a small, hand-crafted and structured resource such as a word net is clearly inferior to a large and semi-structured (Wikipedia) or even completely unstructured resource such as 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 strongly 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*), for the times can reach up to several seconds per calculation.

In general, we observed that the distributional measures, especially LSA, perform 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, does not establish SR 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. particularly 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 [23]).

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.

Considering all the results above, it can be stated that the calculation of semantic relatedness is far from being solved in a satisfying manner. 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.

5.3 *Meta-level Evaluation*

Given the statistical spread shown in Table 3 to Table 5 as well as the obvious discrepancies of the various experimental results exposed by Cramer [13], we argue

that the calculation of SR should be considered a continuous problem. 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:** Seemingly, most studies intend to model the same, i.e. a phenomenon observable in natural language which accounts for lexical cohesion and which is called – depending on the specific research community – semantic similarity/relatedness, association (e.g. [43]), evocation, or semantic distance. However, practically none of these concepts is well defined; there is no consensus among researchers as to which types of relation are to be included and whether these are to be established between words, any kind of lexical unit, concepts etc.
- **Setting of the human judgment experiment:** The studies summarized above differ with respect to the subjects asked to participate, their background and training, as well as the manuals used to explain the task. The different experimental set-ups therefore represent an uncontrolled parameter and might seriously influence the results.
- **Construction of experimental data:** As mentioned above, different methods may be employed in order to construct the experimental data, i.e. randomly selected word-/concept-pairs vs. analytically constructed ones. In addition, data sets might considerably vary with respect to their size, i.e. only a few hand-picked vs. several thousand pairs.
- **Evaluation method:** Finally, the results might be influenced by the specific statistical methods, i.e. different correlation coefficient algorithms, drawn on to determine the correlation between humans and semantic measures, the inter-subject, and the intra-subject correlation¹³.

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, ontology, corpus etc. (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

correspond at all. In principle, word nets, ontologies, corpus statistics, and human judgments should be related to (theoretically even represent) the potentially underlying at least partially shared (lexical) semantic system encoding the collective knowledge of humans, while the relations between words in a concrete text represent more than just an instantiation of this semantic system. That means, the comprehension individually evolving while reading a text might considerably alter the relation strength perceived by the reader between a pair of words. Thus, a (in the sense of the semantic system) moderately related word pair, might be strongly related given a specific context. And consequently, there are at least two aspects which need to

¹³ The inter-subject as well as the intra-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.).

be considered in order to model and successfully compute lexical cohesion: firstly, the shared knowledge of the relations in principle, secondly, the concrete relations in a given context. Thus, it is – in our opinion – vital to distinguish between the first step, that is the calculation of SR, for which, as mentioned above, controlled but unstructured resources and distributional (2^{nd} order) methods seem to perform best, and the second, that is the calculation of lexical cohesion, for which additional text-grammatical features also need to be taken into account (q.v. [34]).

6 Conclusions and Future Work

We presented a study comparing eighteen different semantic relatedness measures on various lexical resources, which we classified into WordNet-based, distributional and Wikipedia-based SR measures. The algorithms implemented and resources employed were analyzed with respect to practical issues, e.g. run time and coverage. Furthermore, we conducted an extensive evaluation on the basis of a human judgment experiment using Pearson's coefficient to measure correlation. We found that the distributional measures perform best – in terms of coverage and correlation. However, our experiments also show that none of the algorithms, proposed in the literature and implemented for this study, performs outstanding. Taking our experimental results into account, we conclude that the less structure (e.g. semantically typed relations or links) a resource exhibits and the more controlled (e.g. in terms of quality of the documents, language and so forth) it is, the better seem to be the correlation coefficients. In the future, we therefore argue that the research should continue on three different levels: firstly, the concept of semantic relatedness should be stated more precisely, which also means that research results of psycholinguistics, linguistics, cognitive science, and computational linguistics should be integrated. Secondly, on the basis of the thus elaborated concept of semantic relatedness, one or more resources can be determined which best fit the requirements at hand; if more than one resource needs to be considered, a method to combine them needs to be devised in addition. Thirdly, given a substantiated concept of SR and the most suitable resources, an (or a family of) algorithm(s) need to be developed, adapted correspondingly, and evaluated again e.g. on the basis of a human judgment experiment. Hence, we propose the definition of a shared task which might bring the research community interested in SR considerably closer to results of high performance. In addition to this, we plan to experiment with a combination of the above-mentioned relatedness measures; more precisely, we intend to feed the various elements of the measures in Sections 4.1 to 4.3 into a machine learning toolkit such as Weka (cf. [47]). A pilot study already demonstrated that it is thus possible to enhance the performance by at least 10% compared with the currently best performing measure mentioned in Section 5.1.

Acknowledgements The authors would like to thank Christiane Fellbaum, Lothar Lemnitzer, Alexander Mehler, Sabine Schulte im Walde, Angelika Storrer, Armin Wegner, and Torsten Zesch for their helpful comments. This research was partially funded by the DFG Research Group 437.

Appendix

References

- 1.
2. John R. Anderson. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behaviour*, 22:261–295, 1983.
3. M. Baroni and S. Bernardini, editors. *Wacky! Working papers on the web as corpus*. GEDIT, Bologna, Italy, 2006.
4. Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop*, pages 10–17, 1997.
5. J. Boyd-Graber, C. Fellbaum, D. Osherson, and R. Schapire. Adding dense, weighted, connections to wordnet. In *Proceedings of the 3rd Global WordNet Meeting*, pages 29–35, 2006.
6. Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of semantic relatedness. *Computational Linguistics*, 32 (1):13–47, 2006.
7. J. A. Bullinaria and journal = Behavior Research Methods year = 2007 volume = 39 number = 1 pages = 510–526 owner = twandmac Levy, J. P. title = Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study.
8. Joe Carthy. Lexical chains versus keywords for topic tracking. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 507–510. Springer, 2004.
9. S. Cederberg and D. Widdows. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy. In *Proc. of CoNLL’03*, 2003.
10. K. Church and P. Hanks. Word association norms, mutual information and lexicography. In *Proceedings of the 27th ACL*, number 27, pages 76–83, 1989.
11. Rudi Cilibrasi and Paul M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.
12. A.M. Collins and E.F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, 82:407–428, 1975.
13. Irene Cramer. How Well Do Semantic Relatedness Measures Perform? A Meta-Study. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 59–70. College Publications, 2008.
14. Irene Cramer and Marc Finthammer. An evaluation procedure for word net based lexical chaining: Methods and issues. In *Proceedings of the 4th Global WordNet Meeting*, pages 120–147, 2008.
15. S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
16. Christiane Fellbaum, editor. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.
17. E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.
18. Stephen J. Green. Building hypertext links by computing semantic similarity. *IEEE Transactions on Knowledge and Data Engineering*, 11(5), 1999.
19. Iryna Gurevych. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the IJCNLP 2005*, pages 767–778, 2005.

20. M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.
21. Graeme Hirst and David St-Onge. Lexical chains as representation of context for the detection and correction malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. The MIT Press, 1998.
22. Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of ROCLING X*, pages 19–33, 1997.
23. A. Kilgarriff. Googleology is bad science. *Computational Linguistics*, 33(1):147–151, 2007.
24. T. Landauer and S. Dumais. A solution to Plato’s problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104(1):211–240, 1997.
25. Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–284. The MIT Press, 1998.
26. Lothar Lemnitzer and Claudia Kunze. Germanet - representation, visualization, application. In *Proceedings of the 4th Language Resources and Evaluation Conference*, pages 1485–1491, 2002.
27. Lothar Lemnitzer, Holger Wunsch, and Piklu Gupta. Enriching germanet with verb-noun relations - a case study of lexical acquisition. In *Proceedings of the 6th International Language Resources and Evaluation*, 2008.
28. Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.
29. Palmira Marrafa and Sara Mendes. Modeling adjectives in computational relational lexica. In *Proceedings of the COLING/ACL 2006 poster session*, pages 555–562, 2006.
30. George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.
31. David Milne. Computing semantic relatedness using wikipedia link structure. In *Proc. of NZCSRSC07*, 2007.
32. Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1), 1991.
33. Jane Morris and Graeme Hirst. Non-classical lexical semantic relations. In *Proc. of HLT-NAACL Workshop on Computational Lexical Semantics*, 2004.
34. Jane Morris and Graeme Hirst. The subjectivity of lexical cohesion in text. In J. C. Chanahan, C. Qu, and J. Wiebe, editors, *Computing attitude and affect in text*. Springer, 2005.
35. Adrian Novischi and Dan Moldovan. 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*, pages 897–904, 2006.
36. Simone Ponzetto and Michael Strube. Wikirelate! computing semantic relatedness using wikipedia. July 2006.
37. R. Rapp. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of COLING’02*, Taipei, Taiwan, 2002.
38. Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the IJCAI 1995*, pages 448–453, 1995.
39. Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.
40. G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.
41. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, 1983.
42. H. Schtze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):pp. 97–124, 1998.
43. Sabine Schulte im Walde and Alissa Melinger. 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*, pages 612–619, 2005.

44. Gregory H. Silber and Kathleen F. McCoy. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4), 2002.
45. Elke Teich and Peter Fankhauser. Wordnet for lexical cohesion analysis. In *Proc. of the 2nd Global WordNet Conference (GWC2004)*, 2004.
46. T. Wandmacher. How semantic is Latent Semantic Analysis? In *Proceedings of TALN/RECITAL'05*, Dourdan, France, 2005.
47. Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2nd edition, 2005.
48. Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, 1994.
49. Torsten Zesch and Iryna Gurevych. Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the Workshop on Linguistic Distances at COLING/ACL 2006*, pages 16–24, 2006.
50. Torsten Zesch, Iryna Gurevych, and Max Mhlhuser. Comparing wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets. In *In Proc. of NAACL-HLT*, 2007.
51. Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC), electronic proceedings*, Mai 2008.