Computing Semantic Relatedness in German with Revised Information Content Metrics

Iryna Gurevych and Hendrik Niederlich
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/~gurevych

Abstract

The paper presents an application of information content based metrics to compute semantic relatedness of word senses in German. The main contributions are: an annotation study based on a revised definition of semantic relatedness beyond synonymy, an extension of Resnik's (1995) procedure for computing information content of concepts for strongly inflected languages, an application of information content based metrics to compute semantic relatedness of German word senses defined in GermaNet (Kunze, 2004) and a new interpretation and normalization function for Jiang & Conrath's (1997) distance metric. Semantic relatedness metrics consistently outperform two baselines: a Lesk based algorithm, and one using Google word co-occurrence statistics.

1 Introduction

Systems computing semantic relatedness should allow to approximate human intuitions about lexical semantic relations existing between words. For example, given the words *Glass*, *Mug* and *Jewel*, we note that while *Glass* and *Mug* display a fairly close semantic relatedness, the relation between *Glass* and *Jewel* is less close according to human judgements. Numerous metrics were proposed to compute semantic similarity. However, as noted by Hirst & Budanitsky (2005) similarity is a special case of a more general notion, semantic relatedness, which encompasses additional lexical semantic relations, such as meronymy, antonymy, functional association and more. Also, relatedness is more often required by NLP applications than just similarity, e.g. in information retrieval. Therefore, we focus on semantic relatedness and apply information content metrics to this more general task.

No extensive studies or large-scale evaluations of semantic relatedness algorithms for languages other than English have been conducted so far, to our knowledge. As a consequence, we know little about the applicability of semantic relatedness metrics to other languages. We study the performance of semantic relatedness metrics across several parameters. We touch upon methodological issues in computing semantic relatedness *proper* (as opposed to semantic similarity) as an NLP task – what is about the human performance, lower and upper bounds for evaluation? The next parameter is WordNet versus other x-Nets – can we expect the same performance with resources constructed following WordNet principles, but divergent in some design decisions and in the coverage? Departing from that, we check the applicability of the methods developed for English to other languages – is the performance of the methods language-specifi c or will it be similar for other natural languages?

The remainder of the paper will be structured as follows: Section 2 presents the design of a German dataset with human semantic relatedness judgments, which is followed by a description of the information content based semantic relatedness metrics in Section 3. Then, our experiments on computing information content of GermaNet concepts and semantic relatedness of word senses are presented in Section 4. The results are compared with two baselines in Section 5 and are followed by conclusions and an outline of future work.

2 Experiments with Subjects

Human judgments of semantic relatedness provide a *gold standard* for evaluating the results of automatic methods. The inter-annotator agreement defi nes an upper bound for the evaluation of automatic methods (Resnik, 1995). The main issues while designing a dataset in our study are the following: the choice of lexical units comp-

rising the word pairs to be evaluated, the number of word pairs and human subjects, defining semantic relatedness and the rating scale for human judgments.

The choice of lexical units is challenging as there is no well-defined criterion for that. If we would choose e.g. 65 items randomly, we would run into the risk of getting an unbalanced dataset with skewed value distributions, potentially leading to unreliable evaluations at a later stage. Another option is to have a very large dataset chosen randomly. In this case it becomes problematic to have a large number of human judges rating them manually for semantic relatedness. Therefore, we decided to keep the word pairs from the psycholinguistic experiment by Rubenstein & Goodenough (1965) and translate them into German. Advantages of this are: our results can be related (although not directly compared) to the results for English based on that dataset, the number of word pairs in the dataset by Rubenstein and Goodenough, i.e. 65 is reasonably large to generalize. A disadvantage is that we also include only nouns in the evaluation of semantic relatedness.

We asked 24 subjects (native speakers of German) to rate 65 word pairs on a scale from 0 to 4 for semantic relatedness. Semantic relatedness was defined in a broader sense than just similarity. To determine the upper bound of performance for automatic semantic relatedness algorithms, we computed a summarized correlation coefficient for a set of 24 judges. This is based on the interclass reliability analysis in statistics. To get the average, we computed the bivariate correlations for all judges pairwise and then pooled them using Fisher's z transformation, yielding $z = 1.1266$. This number is transformed back to a correlation coefficient, yielding $r = .8098$, which is statistically significant. We observe that the correlation coefficient in our study, i.e. the upper bound for evaluating the system's performance on the relatedness task, is lower than what had been reported by Resnik (1995) for the similarity task, $r = .8848$. This is caused by relaxed rating criteria based on a broader definition of semantic relatedness. Though this leads to more diverse human judgments, they are reliable to serve as an evaluation dataset for computational methods.

3 Information Content Based Metrics

Typically, ICMs employ the structure of the wordnet (in our case, GermaNet) hierarchy together with additional corpus-based evidence, which is called *information content*. Information content values of GermaNet concepts are required to compute the semantic relatedness score of a concept pair. Resnik (1995) introduced the notion of information content and the first metric based on it (in the following abbreviated as *res*). Semantic similarity between two words w_1 and w_2 is defined as the information content value of their lowest common subsumer (LCS) as given in Equation 1:[1]

$$sim_{c_1,c_2} = \max_{c \in S(c_1,c_2)} [-\log p(c)] \qquad (1)$$

where $S(c_1, c_2)$ is the set of concepts which subsume both c_1 and c_2 and $-\log p(c)$ is the information content. The probability p is computed as the relative frequency of words (representing that concept) in a corpus (the discussion of this follows in Sections 4.1 and 4.3, Equation 5). In evaluating Resnik's metric, we use the GermaNet hierarchy to determine the lowest super class for a pair of concepts. If multiple inheritance occurs, we select the LCS with the highest information content as this is the one maximizing their semantic relatedness.

Jiang & Conrath (1997) proposed to combine edge- and node-based techniques in counting the edges and enhancing it by the node-based calculation of the information content as introduced by Resnik (1995). The method is abbreviated as *jcn*. The distance between two concepts c_1 and c_2 is formalized as given in Equation 2:

$$dist_{c_1,c_2} = IC_{(c_1)} + IC_{(c_2)} - 2 \times IC(LCS(c_1, c_2)) \qquad (2)$$

where IC is the information content value of the concept, and $LCS(c_1, c_2)$ is the lowest common subsumer of the two concepts.

The third method is that of Lin (1998) (referred to as *lin*). He defined semantic similarity using a formula derived from information theory. This metric is sometimes called a universal semantic similarity metric as it is supposed to be

[1] For all methods, c_1 and c_2 are concepts (word senses) corresponding to w_1 and w_2.

application-, domain-, and resource independent. According to it, similarity is given in Equation 3:

$$sim_{c_1,c_2} = \frac{2 \times \log p(LCS(c_1, c_2))}{\log p(c_1) + \log p(c_2)} \quad (3)$$

4 Experiments with ICMs for German

4.1 Frequency Estimations

Frequencies of concepts (synsets) in GermaNet were estimated using a German newspaper corpus *taz* (see www.taz.de). This corpus covers a wide variety of topics and has about 172 million tokens. The number of tokens representing nouns is 52,490,873. They correspond to 2,326,397 distinct word forms and 1,912,577 unique stems.

We produced a part of speech (POS) tagged version of the corpus using the TreeTagger (Schmid, 1997). Resnik (1995) counted each word (noun) in the corpus as an occurrence of each synset subsuming it in a conceptual hierarchy. However, applying this procedure to a German corpus poses a problem, as the words are highly inflected. This means that any inflected form of a word which does not coincide with a base form represented in GermaNet will be missed. We employed a German version of the Porter algorithm to stem both the word forms in word frequency lists and the GermaNet word senses.[2] As a result of stemming, word frequency lists are transformed into stem frequency lists. The reduction of different word types to a single stem may influence the data undesirably. However, as the effect is marginal, it is unlikely to bias the results considerably and can so far be neglected. To solve the issue of stem disambiguation, a special preprocessing component would be required.

4.2 GermaNet Coverage

GermaNet is a German wordnet which adopted the major properties and database technology from Princeton's WordNet. However, GermaNet displays some structural differencies and content oriented modifications. Its designers relied mainly on linguistic evidence rather than psycholinguistic motivations. Example of discrepancies between GermaNet and WordNet are e.g. that GermaNet employs artificial, i.e. non-lexicalized concepts, and adjectives are structured hierarchically as opposed to WordNet. Currently, GermaNet includes about 40000 synsets with more than 60000 word senses modelling nouns, verbs and adjectives. E.g., the entry for "Brot" (Engl. *bread*) looks as follows: Sense 1 Brot => Backware => ?festes Nahrungsmittel => Nahrung, Nahrungsmittel, Essen, Speisen => Objekt.[3]

Based on the stemming process described above, we analyzed the GermaNet coverage of the newspaper corpus. The results of this analysis are summarized in Table 1, where "+" before a number corresponds to the attribute "found" and "-" means not found. We present the percentage of tokens and stems which could or could not be mapped to a GermaNet word sense, and the proportion of synsets that could be assigned an information content value. Striking about the numbers is the low coverage of stems. The coverage is significantly higher for tokens, especially for nouns and adjectives. We can explain this as follows: The main source of the stems not found in GermaNet are compounds, which is an extremely productive form of word composition in German, e.g. *Beschäftigungsverhältnis* (Engl. *employment relation*), *Informationsmarketing* (Engl. *marketing of information*), *Betriebssystemkenntnis* (Engl. *knowledge about operating systems*). The most compounds are low frequent words in German, whereas non-compound words are frequent and are covered rather well by GermaNet. Nevertheless, considering that the vast part of synsets in GermaNet finally get an IC value assigned, the calculation of information content in general seems not to be affected heavily by this effect.

	Stems	*Tokens*	*Synsets*
Nouns	+1.4% / -98.6%	+65% / -35%	+83% / -17%
Verbs	+8% / -92%	+34% / -66%	+91% / -9%
Adject.	+2.2% / -97.8%	+73% / -27%	+98% / -2%
Total	+2% / -98%	+60% / -40%	+88% / -12%

Table 1: GermaNet coverage of the corpus.

4.3 Computing Information Content Values

Resnik (1995) defined the procedure to compute information content from word frequencies.

[2] http://snowball.tartarus.org/german/stemmer.html.

[3] => stands for *is-a-kind-of*, "?" stands for a non-lexicalized concept.

However, as we showed, the formula has to be rewritten for stems in strongly inflected languages, such as German. The resulting scheme is given in Equation 4, where $count(n)$ is a function which returns the sum of occurrences for a particular stem n, and $stems(c)$ is the set of stems subsumed by a concept c. The mapping to GermaNet is implemented on the basis of a stem associated with a certain POS. On the other hand, stemmed string representations of word senses in GermaNet are associated with distinct part of speech, too.

$$freq(c) = \sum_{n \in stems(c)} count(n) \quad (4)$$

$$p(c) = freq(c)/N \quad (5)$$

The concept probabilities are then computed according to Resnik's original proposal as the relative frequency, s. Equation 5, where N is the total number of stems (equal to the total number of original words as stems are in fact "pointing" to them), which could be found in GermaNet.

4.4 Application of ICMs

We applied Resnik's and Lin's metrics in a straightforward manner based on the information content values of GermaNet concepts computed as outlined above. The results are based on 57 out of 65 word pairs in the evaluation dataset as the rest were missing in GermaNet. They are summarized in terms of interclass correlation coefficient, yielding $r_{res} = .7152$ and $r_{lin} = .7337$. As noted in Hirst & Budanitsky (2005), jcn returns semantic distance, rather than a similarity value. Its implementation poses several additional questions: (1) What is the distance of two word senses belonging to the same synset? (2) What is the distance of two word senses with no lowest common subsumer? (3) How to compute the distance between a hyperhym and its hyponym? (4) How to evaluate the distance scores as opposed to relatedness scores output by alternative metrics?

(1) The distance from the word sense to itself is represented by zero. Therefore, we believe that the distance of two word senses belonging to the same synset should be non-zero. This number can be defined as the smallest (minimum) distance value given a specific concept hierarchy. In order to compute it, we adopt the proposal by Sid Patwardhan outlined in http://groups.yahoo.com/group/wn-similarity/message/8. We assign the minimum distance value (given the GermaNet hierarchy and a specific corpus as the basis of information content) to synonymous word senses (this value should approach 0). In doing that, we look for the concept with the lowest IC value in the corpus. Typically, this is the concept *Object* in GermaNet. The frequency count of this concept minus 1 models a very similar concept, which is not identical to the one under consideration. From this, we can compute the IC value of an artificial concept and obtain the lowest value of distance possible for GermaNet given our particular corpus. E.g. the distance between "Edelstein" (Engl. *gem*) and "Juwel" (Engl. *Jewel*) is set to 3.4787E-08 according to this.

(2) Contrary to that, the distance of two word senses with no LCS should be assigned the maximum distance value $dist_{max}$. We compute $dist_{max}$ according to Equation 2, where we assume that the two concepts compared are the ones with the highest IC in our data (17.3441) and their lowest common superclass is the one with the smallest IC value (.17) in our data. This maximizes the possible distance value und results in $dist_{max} = 34.348$ for our data.

(3) Another special case of semantic distance occurs, if a subclass, e.g. "Forst" (Engl. *forest*) is compared with a superclass, e.g. "Wald" (Engl. *wood*). Then, the distance becomes negative. We take the absolute value for distance, as the direction is not essential in this case.

(4) Hirst & Budanitsky (2005) suggest evaluating by applying a correlation measure to distance values, whereas a negative correlation value is obtained (as distance is the opposite of similarity). Following this suggestion, we obtained the correlation $r = -.5292$. We investigated this matter and noticed that the bad correlation coefficient is due to 16 (!) out of 57 word pairs in our data, which do not have a lowest common superclass. According to the *jcn* algorithm, we assign $dist_{max} = 34.348$ to them. The next

highest value of semantic distance is 21.97, and the distribution of scores is skewed. Pedersen et al. (2004) convert semantic distance into similarity by taking the reciprocal of semantic distance $sim_{c1,c2} = 1/dist_{c1,c2}$. This changes the correlation to a positive number. However, the transformation is not linear and thus affects the distribution of scores, yielding a worse correlation coeffi cient on our data of $r = .2915$.

In order to minimize the role of extremely small and extremely large numbers in the distance values, which skew the distribution of scores, we need a special normalization function. This function should map the scores to the range from 0 to 1, with small numbers for little and large numbers for high semantic relatedness. We defi ne the normalization function given in Equation 6 using the hyperbolic tangent function:

$$sim_{c_1,c_2} = 1 - (tanh(dist_{c1,c2} \times c)) \quad (6)$$

where $dist_{c1,c2}$ is a distance score for a pair of concepts and c is a special constant. This constant can be determined by solving Equation 6, which results in Equation 7. In this transformation, we assume that $dist_{avg}$ is the distance value of a word pair with an average human semantic relatedness score in a given dataset (5.2 for our data):[4] This value should be mapped to $sim_{c_1,c_2} = .5$ in our application. Given this:

$$c = atanh(.5)/dist_{avg} \approx .55/dist_{avg} \quad (7)$$

The correlation coeffi cient for jcn is $r = -.5292$ if no normalization is performed and $r = .7379$ if the scores are normalized as described above.

5 Evaluation

We designed and implemented two baselines to determine semantic relatedness. The fi rst baseline compares the performance of information content metrics to a dictionary based approach, the Lesk algorithm (Lesk, 1986) operating on the glosses from traditional dictionaries written by human authors. The correlation between the number of stem overlaps in textual defi nitions of word senses and human judgments of semantic relatedness yielded $r = .5307$. The second baseline for the evaluation was represented by word co-occurrence counts obtained from querying the Web through Google. Semantic relatedness was computed according to Equation 8, where $hits_{w1}$ and $hits_{w2}$ are the frequencies of words $w1$ and $w2$. The correlation of Google based results with human judgments of semantic relatedness was .5723. The result seems quite impressive (the Lesk based baseline yielded $r = .5307$), if we consider that the method does not require any sophisticated knowledge sources and is conceptually simple. It should be noted that we tried several other established measures of lexical association, e.g., PMI and log-likelihood on Google counts, but the results were always worse than those achieved by Equation 8.

$$sim_{w1,w2} = hits_j/hits_{w1} + hits_j/hits_{w2} \quad (8)$$

In Figure 1, we summarize the results of our experimental work. All information content based metrics perform well and approach the human performance. This is generally consistent with the fi ndings of other researchers for English. However, there are a couple of reasons why the absolute results are lower than those in previous studies with WordNet, where ICMs achieved a correlation of about .8 with human judgments.

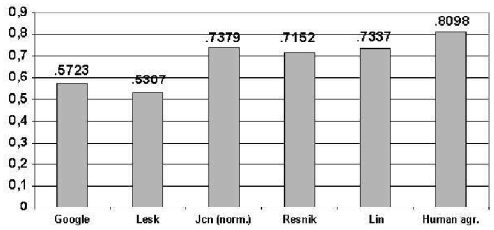

Figure 1: Evaluation results.

Information content based metrics perform poorly for the word pairs, which do not have a common subsumer, such as "car" and "journey", "food" and "rooster". GermaNet displays quite a large proportion of such cases where no LCS is found (28%). Though we work on translated English word pairs, the number of cases without LCS is higher than for WordNet, as GermaNet is not modeled in the same detail. This causes a certain decrease in performance.

In fact, coverage is a general problem with existing wordnets. Semantic relations beyond

[4]Not necessarily corresponding to an arithmetic average.

hypernymy are covered insufficiently and sometimes even inconsistently. As they are rather part of world knowledge, further developments of lexical-semantic nets in the direction of ontology should be pushed forward. An error source for semantic relatedness metrics is also the idiosyncracy of some modeling decisions in GermaNet. For example, "Coast" is modeled as a subclass of *Geographical area* and *Location*, whereas "Shore" is a subclass of *Border* and *Attribute*. The same happens to "Hill" and "Mountain". Being parts of different hierarchies, these words do not have an LCS and receive a zero score of semantic relatedness.

One of the major inherent drawbacks of lCMs is that they make semantic relatedness dependent on the subsumption hierarchy. This way, they minimize the role of lexical-semantic relations beyond hypernymy, which are essential to relatedness in general. Therefore, not only we need that such relations are modeled in sufficient detail (which is still not the case), but also that the metrics are extended to include those relations into the models of semantic relatedness.

6 Conclusions

We explored the applicability of information content based metrics to compute semantic relatedness of German words. We revised the calculation of information content for German concepts based on frequency counts for stems rather than words. The implementation of the *jcn* metric was revised and a new normalization function introduced. Our reported results compare favorably with a Lesk and a Google based baselines and are consistent with the findings for English. A somewhat lower performance of the metrics can be explained by a broader definition of the task as relatedness and discrepancies in the underlying knowledge bases. While some extensions may still become necessary, the hypothesis of the applicability of the methods to strongly inflected languages has been generally confirmed.

We found out that stemming is not optimal to handle complex compositional morphological structure of German. To achieve accurate mappings from word frequency counts to word senses, a component for morphological analysis should be employed. A sort of contextual analysis (word sense disambiguation) should be done to associate different "senses" of a word with their individual counts in frequency lists. Our current work is aimed at a considerably larger dataset with 350 word pairs of different parts of speech. The performance of semantic relatedness metrics based on the new dataset has to be studied for a number of parameters, such as parts of speech and different types of relatedness. We have to investigate to what extent the measures are applicable across parts of speech, e.g. for a verb and a noun.

Acknowledgments

This work has been funded by the Klaus Tschira Foundation. We thank Michael Strube for his valuable comments concerning this work.

References

Hirst, Graeme & Alexander Budanitsky (2005). Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111.

Jiang, Jay J. & David W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING)*. Tapei, Taiwan.

Kunze, Claudia (2004). Lexikalisch-semantische Wortnetze. In K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde & H. Langer (Eds.), *Computerlinguistik und Sprachtechnologie. Eine Einführung*, pp. 423–431. Heidelberg, Germany: Spektrum Akademischer Verlag.

Lesk, Michael (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, Toronto, Ontario, Canada, June, pp. 24–26.

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

Pedersen, Ted, Siddharth Patwardhan & Jason Michelizzi (2004). WordNet::Similarity – Measuring the relatedness of concepts. In *Demonstrations of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, Mass., 2–7 May 2004, pp. 267–270.

Resnik, Phil (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montréal, Canada, 20–25 August 1995, Vol. 1, pp. 448–453.

Rubenstein, Herbert & John Goodenough (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Schmid, Helmut (1997). Probabilistic part-of-speech tagging using decision trees. In Daniel Jones & Harold Somers (Eds.), *New Methods in Language Processing*, Studies in Computational Linguistics, pp. 154–164. London, UK: UCL Press.