Using the Structure of a Conceptual Network in Computing Semantic Relatedness

Iryna Gurevych

EML Research gGmbH, Schloss-Wolfsbrunnenweg 33, 69118, Heidelberg, Germany
http://www.eml-research.de/∼gurevych

Abstract. We present a new method for computing semantic relatedness of concepts. The method relies solely on the structure of a conceptual network and eliminates the need for performing additional corpus analysis. The network structure is employed to generate artificial conceptual glosses. They replace textual definitions *proper* written by humans and are processed by a dictionary based metric of semantic relatedness [1]. We implemented the metric on the basis of GermaNet, the German counterpart of WordNet, and evaluated the results on a German dataset of 57 word pairs rated by human subjects for their semantic relatedness. Our approach can be easily applied to compute semantic relatedness based on alternative conceptual networks, e.g. in the domain of life sciences.

1 Introduction

Semantic relatedness of words represents important information for many applications dealing with processing of natural language. A more narrowly defined phenomenon of semantic similarity has been extensively studied in psychology, cognitive science, artificial intelligence, and computational linguistics. In the context of linguistics, it is typically defined via the lexical relation of synonymy. While synonymy is indeed an example of extreme similarity (suggesting that two words are interchangeable in a certain context), many natural language processing applications require knowledge about semantic relatedness rather than just similarity [2]. Departing from that, we define semantic relatedness as any kind of lexical or functional association that may exist between two words. For example, the words "car" and "journey" apparently display a close semantic relationship, while they are not synonymous.

Many natural language processing applications, e.g. word sense disambiguation or information retrieval, do not need to determine the exact type of a semantic relation, but rather to judge if two words are closely semantically related or not. For example, for an application in the domain of career consultancy it might be important to conclude that the words "baker" and "bagel" are closely related, while the exact type of a semantic relation does not need to be assigned.

Metrics of semantic relatedness are increasingly embedded into natural language processing applications for English due to the availability of free software, e.g. [3] and pre-computed information content values from English corpora. The evaluation of all approaches to compute semantic relatedness has so far been done for the task of semantic similarity. The underlying data was based on the English language [4,5]. We propose the following classification of the metrics of semantic relatedness:

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 767–778, 2005.

- *intrinsic or extrinsic*. Intrinsic metrics employ no external evidence, i.e. no knowledge sources except for the conceptual network itself [1,6,7]. Extrinsic metrics require additional knowledge, e.g. information content values of concepts computed from corpora [8,9,10].
- *the type of knowledge source employed, e.g. a dictionary or a conceptual network.* Metrics can either employ a machine readable dictionary, i.e. textual definitions of words therein as an underlying knowledge base [1,11], or operate on the structure of a conceptual network, whereby textual definitions themselves are not available [9,7].

Researchers working on the processing of languages such as English, for which many resources exist, have a large choice of options for choosing a metric or a knowledge source. This is, however, not the case for many other languages. Extrinsic metrics relying on a conceptual network and additional corpus data cannot always be applied. It is difficult and time-consuming to find and process a corpus, which is substantially large to compute information content of concepts for a new language. The same is true for domain-specific corpora, e.g. in the domain of life sciences. Before information content of domain concepts, e.g. protein names can be computed, a considerable effort has to be invested in compiling and processing a substantially large corpus in the respective domain. This makes it difficult to apply corpus-based metrics. At the same time, many domains already have a kind of domain model in the form of thesauri or at least taxonomies, which appear to be instances of conceptual networks.

For dictionary based metrics, the difficulty is that textual definitions of word senses in dictionaries are often inconsistent. [12] note that dictionary definitions often do not contain enough words to be effectively deployed in dictionary based metrics. Furthermore, an important dimension is the portability of a metric to new domains, i.e. whether the metric can be applied e.g. to medical or biological domains. They typically have well developed taxonomies, but are lacking language based descriptions (definitions). Therefore, the application of both extrinsic and dictionary based metrics is problematic.

Simultaneously, in this and in many other cases elaborated conceptual networks or taxonomies are available. One of the most prominent examples of that is the Open Directory Project (http://www.dmoz.com). Therefore, we propose a method for computing semantic relatedness which overcomes the constraints related to previous metrics and is completely intrinsic. It exploits solely the structure of a conceptual network, while the need for both external statistical data and textual definitions of concepts is completely eliminated. Thus, our method is language independent and can be applied to different knowledge bases represented as conceptual networks.

We conducted an experiment with human subjects to determine the upper bound of performance for automatic metrics of semantic relatedness. While this task is more difficult than computing semantic similarity, human judgments display a high interclass correlation. We evaluated our approach against this dataset, see Section 4, and compared it with baseline systems. The proposed metric achieves the same performance as a popular extrinsic (information content based) metric by [8], and is significantly better than the results of a conventional dictionary based measure [1] employing a machine-readable dictionary compiled by human writers. To exhaustively evaluate the metric, we introduced an additional baseline based on co-occurrences of words in the Web. A summary of the results is given in Section 5.

2 Definitions in Dictionaries and Conceptual Networks

Definitions of words and their distinct senses are essential to our approach. What is the definition of a good definition? This question has been disputed in philosophy since the days of Platon and Euklid, recently also in the disciplines such as cognitive science and linguistics. Different types of definitions were proposed, whose names are expressed in various terminologies, e.g. lexical, theoretical, circular and the definition by genus and difference to name just a few. Lexical definitions are what we typically find in a dictionary. They often suffer from inaccuracies as they are confined to established meanings and can prove to be confusing for example in legal matters. According to [13], a good definition should include several indisposable components: a *functional part* describing what the concept is intended for, the characteristics of the definiendum contrasting *the general* with *particular*, and *context* (time, place, cultural and mental). For example, "window" is a planar discontinuity in a solid artificial (context) surface (genus), which allows to look through it, or for the penetration of light or air (when not covered or open) (differentia). Without the differentia – with the genus alone – the definition can well fit the door; without the context, the definition can well fit a hole in a rock.

When human writers create definitions, they take care of the structural elements and requirements described above. On the other hand, when creating conceptual networks, dictionaries and language based examples are often employed as knowledge sources to determine lexical and semantic relations between words. Therefore, information about functions, general terms, and context is integrated into a conceptual network. The main idea explored in the present paper is, then, the possibility to extract knowledge about concepts from the conceptual network based on known properties of definitions and how they are encoded in the network. We call extracted pieces of knowledge *pseudo glosses*. Pseudo glosses can be used in the situations when textual definitions *proper* are not available. The information encoded in the network as lexical semantic relations is transformed into artificially generated glosses. Those can be employed in NLP applications. An additional advantage of pseudo glosses as opposed to real glosses is the possibility to include or exclude certain types of information from a gloss. This way, glosses can be easily tailored to a specific task at hand. In our application, this amounts to experimentally determining the types of information crucial for computing semantic relatedness.

The knowledge base employed in our experiments is GermaNet [14], the German counterpart of WordNet [15]. Direct re-implementation of semantic relatedness metrics developed for WordNet on the basis of GermaNet is not a trivial task. While sharing many design principles with WordNet, GermaNet displays a number of divergent features [16]. Some of them, such as the lack of conceptual glosses, make it impossible to apply dictionary based metrics in a straightforward manner. Therefore, pseudo glosses are generated directly from the conceptual network.

We experimented with different parameters that control which concepts are included in a pseudo gloss:

- *size* determines the length of a path for the hypernyms to be included in a pseudo gloss. The values of *size* range over the interval $[1, depth_{max}]$, where $depth_{max}$ is the maximum path length in a conceptual network. The depth is equivalent to the height in this context.
- *limit* determines the length of a path from the root node of a hierarchy (i.e. the most abstract concept) towards a given concept. The concepts of the path are excluded from the pseudo gloss. The values of *limit* range over the interval $[0, depth_{max}]$. Given $limit = 0$ no concepts will be excluded from the pseudo gloss, and given $limit = depth_{max}$ the resulting pseudo gloss contains solely the given word sense itself. If $size$ and $limit$ are conflicting (e.g. the concept A should be included according to $size$, and excluded according to $limit$), the latter takes precedence over the former.
- *one_sense_per_synset* (OSPS) parameter, either true or false. A synset is often represented by multiple synonymous word senses. If the parameter is set to true, only one word sense from a synset will be included into a pseudo gloss (this is also the case in paper dictionaries). Otherwise, all word senses of a synset are included.
- *lexical semantic relations* control the type of relations in a conceptual network which are involved in generating pseudo glosses, i.e. hypernymy, hyponymy, synonymy, meronymy, association, etc.

Table 1 presents examples of pseudo glosses generated according to two different system configurations: a radial gloss (all lexical semantic relations of a given concept are taken into account, except hyponyms, $OSPS = true$, $size = 3$), and a hypernym gloss (only hypernymy relation is considered, $OSPS = true$, $size = 3$, $limit = 2$).

Table 1. Examples of pseudo glosses for “Bruder – Bursche”

Radial glosses	*Hypernym glosses*
Bursche	
1. junger Mensch, Erwachsener, Bursche, Bub, Junge, Knabe, Bube, Kind, Jüngling	1. Bursche, Junge, Kind
Bruder	
1. Bruder, Geschwister, Mitmensch, Familie, Verwandter	1. Bruder
2. LaienpredigerIn, Fachkraft, unausgebildeter Mensch, Geistlicher, Prediger, ausgebildeter Mensch, Bruder, Berufstätiger, Laie, Laienprediger	2. unausgebildeter Mensch, Geistlicher, Prediger, Laie, Laienprediger
3. christlicher Sakralbau, Kloster, Geistlicher, Mönch, Bruder, Mönchskloster, Ordensangehöriger, Berufstätiger, Glaubensgemeinschaft, Orden, Laie	3. Geistlicher, Ordensangehöriger, Mönch

3 Dictionary Based Metrics

Dictionary based metrics of semantic relatedness were introduced by [1] and received a lot of attention in the context of work on word sense disambiguation. The main idea of this work is to permutate all textual definitions of the senses of two words and to assign them a score based on the number of word overlaps in glosses. Thus, the context which matches best the combination of the two words is assumed to be the disambiguated sense. This can also be viewed as a metric of how the two words are semantically related.

A dictionary gloss is typically represented by textual definitions of word senses corresponding to a given word. E.g. in the Digital Dictionary of the German Language[1] we find the following definitions of "Bruder" (Engl. *brother*), s. Example (1) and "Bursche" (Engl. *fellow* or *lad*), s. Example (2).

(1) *Bruder, der; -s, Brüder /Verkl.: Brüderchen, Brüderlein/*
(1_1) jede männliche Person einer Geschwisterreihe in ihrer Beziehung zu jedem anderen Kind derselben Geschwisterreihe
(1_2) a) enger Freund, Gesinnungsgenosse: b) scherzh. unter Brüdern offen, ehrlich gesprochen: c) Rel. kath. Mönch: d) scherzh. /bezeichnet allgemein einen Mann/ e) salopp abwertend Bursche, Kerl

(2) *Bursche, der; -n, -n /Verkl.: Bürschchen, Bürschlein/*
(2_1) männliche Person a) Knabe, Junge b) junger Mann c) vertraul. alter B. (Freund)! d) Studentenspr. Student einer Verbindung e) veraltend Diener, der einem anderen für persönliche Dienstleistungen zur Verfügung steht
(2_2) umg. kräftiges Tier

In Table 2, we present the results of the Lesk algorithm applied to this word pair. The overlaps are counted on the basis of stems because the German language is highly inflected. The sense combination $1_2 - 2_1$ turns out to be the best fitting one resulting in three overlaps of stems *friend, man, lad.* [17] adopts the algorithm by Lesk and apply it to the task of computing semantic relatedness of WordNet concepts. Their metric is based on the number of shared words in the glosses of concepts available through WordNet. They extend the metric to include the glosses of other concepts, to which they are related according to the WordNet hierarchy. Those are encoded in WordNet as semantic relations, but can be found in any dictionary via synonyms, antonyms, and see-also references. The relatedness score rel_{w_1,w_2} is formally defined in Equation 3:

Table 2. The Lesk algorithm applied to "Bruder–Bursche"

Sense combin.	*Stem overlaps*	*Score*
$1_1 - 2_1$	männlich, Person	2
$1_1 - 2_2$	–	0
$1_2 - 2_1$	Freund, Mann, Bursch	3
$1_2 - 2_2$	Bursch	1

$$rel_{c_1,c_2} = \sum score(R_1(c_1), R_2(c_2)) \qquad (3)$$

where c_1 and c_2 are the compared word senses, R is a set of lexical semantic relations, $score()$ is a function which receives the definitions of word senses and their related

[1] http://www.iai.uni-sb.de/iaide/de/dwds.htm

concepts and returns a numeric score of word overlaps in them. [17] reports a correlation of .67 to the Miller and Charles human study, and one of .60 to the Rubenstein and Goodenough's experiment, which is below the performance of other semantic relatedness metrics reported in [2]. E.g. the metric by Resnik yielded a correlation of .774 and .779 on the datasets respectively.

4 Experimental Work

In this section, we detail the process of generating artificial glosses from GermaNet. We discuss the set of parameters, their consequences for the creation of glosses and the application of artificially generated glosses to the task of computing semantic relatedness. The evaluation, then, measures the performance of semantic relatedness algorithms based on pseudo glosses with respect to human judgments of semantic relatedness. We apply the Lesk algorithm to pseudo glosses and compute a semantic relatedness score for each sense combination of a word pair. The scores are related to average human ratings by means of interclass correlation analysis.

As no datasets for evaluating semantic relatedness are available, we translated 65 word pairs from the dataset by [4] to German. Their word pairs were selected to cover a range of semantic distances. We asked 24 subjects (native speakers of German) to rate the word pairs on the scale from 0 to 4 for their semantic relatedness. Semantic relatedness was defined in a broader sense than just similarity. To determine the upper bound of performance for automatic semantic relatedness metrics, we computed a summarized correlation coefficient for a set of 24 judges. This means that we computed correlations for all judges pairwise, transformed them to a Z-score, computed the average and transformed back to a correlation coefficient yielding $r = .8098$, which is statistically significant at $p = .01$. The evaluation results for relatedness metrics are reported on the basis of 57 from 65 word pairs in the test dataset compared with average human judgments. The remaining words were not covered in GermaNet.

4.1 Experiment 1

We evaluated different methods to generate pseudo glosses according to the parameters: *lexical semantic relations* and *size*. The range of values for *size* was set from 2 to 6. The four system configurations for generating pseudo glosses were the following:

1. a *radial* gloss based on all types of lexical semantic relations in GermaNet;
2. a *hypernymy* gloss based exclusively on the hypernymy relation;
3. a *hyponymy* gloss utilizing the hyponymy relation only;
4. a gloss consisting of *coordinate sisters* of a given concept, i.e. the immediate hyponyms of the concepts' hypernyms.

Table 3 presents the results for the four system configurations (Column "Config."), whereby the best result for each configuration is printed in bold. *Radial* and *hypernymy* glosses yield better results as *hyponymy* and *coordinate sisters* glosses. This happens because pseudo glosses generated by these system configurations resemble the structural components of conventional glosses more accurately. In Examples 1 and 2, we hardly find hyponyms in the definitions. At the same time, e.g. *Lausbube* would be included in the gloss as a hyponym and *Boy* as a coordinate of *Bursche*. Summarizing this experiment, we conclude that the information on hyponymy and coordinate word senses for computing semantic relatedness is not very relevant. The remaining system configurations will be further analyzed in the following.

4.2 Experiment 2

The aim of the experiment was to examine the performance of *radial* glosses under different experimental conditions. We observed that many high scores of semantic relatedness were caused by a large number of hyponyms that some of the words, e.g. *Edelstein* (Engl. *gem*) and *Juwel* (Engl. *jewel*) have. Due to this effect, e.g. at $size = 3$, the number of overlaps for *radial* glosses increases to 199, whereas it is only 7 for *hypernym* glosses. As there exist no well-defined guidelines or criteria as to what number of hyponyms a synset should have, it is rather arbitrary and heavily skews the distribution of semantic relatedness scores.

Another parameter $one\_sense\_per\_synset$ (OSPS) controls whether all of the word senses belonging to a synset are included in a pseudo gloss or only one. The motivation to investigate this effect is similar to the one described previously, i.e. the number of synonyms (word senses) for a given synset is arbitrary. This means that counting overlaps according to the Lesk algorithm will "favour" those word pairs, whose synsets have a large number of word senses.

Table 4 shows the results of the $hyponyms = true/false$ parameter variation. We also varied the $size$ while $OSPS$ was set to *true* and *false*. The data makes evident that ignoring hyponyms in generating pseudo glosses consistently improves the performance. It eliminates the bias in the counts of overlaps due to hyponyms, which do not skew the overall distribution of scores any more. Table 5 further explores the effects of the $size$ parameter and $OSPS = true/false$ variation. For *radial* glosses, the results of $OSPS = true$ are better for $size$ from 3 to 6.

Table 3. Evaluation results for different types of pseudo glosses

Config.	*Size=2*	*Size=3*	*Size=4*	*Size=5*	*Size=6*
1	.3885	**.4570**	.4377	.4027	.4021
2	.5936	.6350	**.6682**	.6072	.6279
3	.3167	.3296	.3244	.3322	**.3538**
4	**.3140**	.2560	.2474	.2062	.1305

Table 4. Hyponyms true/false for *radial* glosses

	Hypo_true	*Hypo_false*
OSPS_true, size 1	.3939	.4513
OSPS_true, size 2	.4235	.5494
OSPS_false, size 1	.3885	.4945
OSPS_false, size 2	.4570	**.5567**

4.3 Experiment 3

We explored the application of *hypernym* glosses to computing semantic relatedness. Several issues were checked, such as the use of the *one_sense_per_synset* true versus false in glosses, the use and interaction of *size* and *limit* as well as the optimal settings for those. The results of this experiment are presented in Figure 1, which shows the correlation values for different combinations of *size* and *limit* settings. The value $OSPS = true$ leads to consistently better results for the parameter combinations of *size* and *limit*. Furthermore, the performance of the system rapidly drops for the *limit* range [4,6]. However, the use of limit is generally justified as it improves the performance for $size = 2$ and $size = 3$. At these points the performance is most optimal, s. Table 6 for exact results.

If neither *size* nor *limit* are in use (which corresponds to the case when all hypernyms of a concept become part of the gloss), the correlation drops to .568. These parameters, therefore, turn out to be necessary to control the choice of concepts in a gloss. As a consequence, the most abstract (close to the root) concepts of the hierarchy will not appear in the gloss.

Table 5. Evaluation results with OSPS true/false for *radial* glosses

	Size 1	Size 2	Size 3	Size 4	Size 5	Size 6
OSPS_false	.4945	**.5567**	.5507	.5299	.5247	.4871
OSPS_true	.4513	.5494	**.5525**	.5444	.5309	.5075

Table 6. Scope/limit in *hypernym* pseudo glosses

Scope;limit	(2;0)	(2;1)	(2;2)	(2;3)
r	.6192	.6456	.6768	.6581
Scope;limit	(3;0)	(3;1)	(3;2)	(3;3)
r	.6682	.6735	**.6914**	.6842

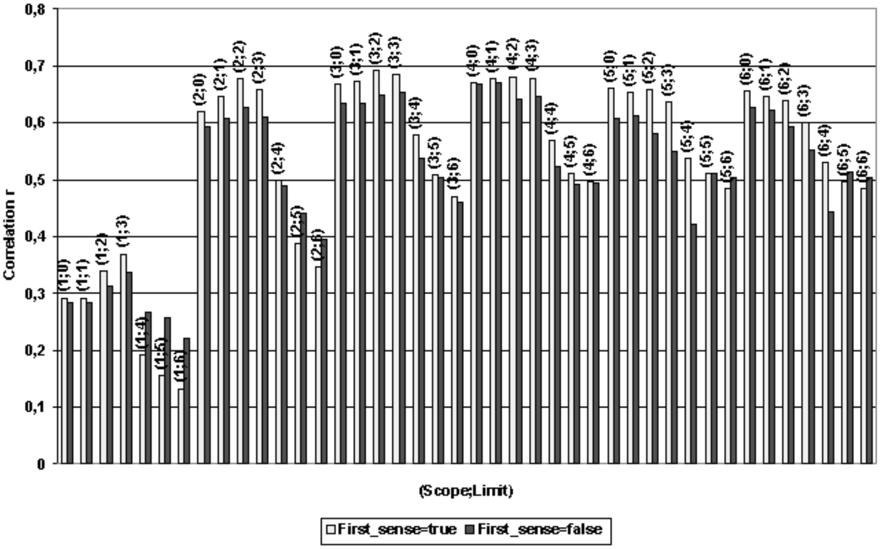

Fig. 1. Correlation r for $one\_per\_synset = true/false$. $r_{(scope,limit)}$ is plotted for $scope = [1;6]$, $limit = [0;6]$.

5 Evaluation Summary

To our knowledge, no datasets are available for validating the results of semantic relatedness metrics.[2] The results obtained for semantic similarity with WordNet are not directly comparable due to differences in the underlying knowledge bases, and – most importantly – in the task definition (changed from similarity to relatedness). The dataset designed during the present study is based on the German language. We opted for GermaNet as the underlying semantic network. Testing the results of metrics for another language, e.g. English, would involve an experiment with native speakers to collect judgments of semantic relatedness and employing an appropriate semantic resource for the chosen language and the domain of interest. However, our evaluation results can be extrapolated to other languages and similar semantic resources, as semantic relatedness metrics themselves are not tied to a particular language or resource. Experimental verification of this fact remains beyond the scope of our paper.

To better understand the performance of the semantic relatedness metrics, we designed and implemented several baselines. The first baseline compares the performance of our system to the original version of the Lesk algorithm operating on the glosses from traditional dictionaries written by human authors. As GermaNet itself does not contain a sufficient number of textual definitions of word senses, they were retrieved from the Digital Dictionary of the German Language.[3] We excluded all additional information from definitions, such as citations and examples of usage. The remaining "pure" definitions were stemmed. The resulting correlation with human judgments yielded $r = .5307$.

The second baseline for the evaluation was represented by word co-occurrence counts obtained by means of querying the Web through Google. This baseline is based on the assumption that the Web can be used as a corpus. [18] provide estimates of the Web size in words, as indexed by Altavista, for the German language: 7,035,850,000. This exceeds the size of freely available German corpora by a large margin. We constructed Google queries, where the query string was represented by a particular word pair. Semantic relatedness of words was, then, computed according to Equation 4, 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, which is quite impressive if we consider that the method does not employ any sophisticated knowledge sources. 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 worse than those achieved by Equation 4.

$$sim_{w1,w2} = hits_{joint}/hits_{w1} + hits_{joint}/hits_{w2} \tag{4}$$

The third baseline is a conventional approach by [8] to compute semantic relatedness via the information content of the lowest common subsumer of the two concepts. Information content of concepts was calculated on the basis of a German newspaper corpus with 172 million tokens (www.taz.de).

[2] English datasets were designed with semantic similarity in mind as described in Section 4.

[3] In fact, any other machine-readable dictionary for German could have been employed instead.

In Figure 2, we summarize the results of our experimental work. These results are based on the most optimal system configurations as described in Section 4: $OSPS = true$, $hyponyms = false$, $size = 3$ for radial glosses, $OSPS = true$, $size = 4$, $limit = 2$ for hypernym glosses. The results show that radial pseudo glosses perform approximately on the same level ($r = .5525$) as the stemmed glosses created by humans ($r = .5307$). This suggests that radial pseudo glosses mimic the behavior of "real" glosses rather well. Hypernym pseudo glosses outperform their radial counterparts and both the Lesk and the Google based baselines by a large margin, yielding $r = .6914$. Their performance is comparable to that of a conventional method by [8] based on external corpus evidence ($r = .7152$).

As the method operates exclusively on pseudo glosses generated on the basis of the hypernymy relation, this type of information from definitions turns out to be the most important one for computing semantic relatedness. In other words, *definitio per genus proximum* is superior to *definitio per differentia specifica* in this task. It should be noted that the situation can change for different types of tasks where the knowledge from a conceptual network is employed, e.g. in word sense disambiguation. As opposed to using textual definitions from traditional dictionaries, generating them automatically from a conceptual network has a great advantage: we can easily control the usage of specific types of information with the help of a set of parameters. In naturally occurring texts, this is problematic, as the required information should be first extracted from free text, a task not trivial to achieve.

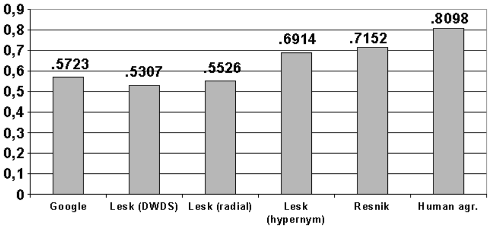

Fig. 2. A summary of evaluation results

If compared to results reported for the task of computing semantic similarity on the basis of WordNet, we note that our numbers are lower due to a number of reasons:

- The upper bound for the performance of computational algorithms is lower for our task ($r = .8089$) as the one given by [8] ($r = .8848$). Semantic relatedness is not as well defined as semantic similarity. The results have to be interpreted according to this lower upper bound.
- The performance of the metrics is dependent on the underlying knowledge base, i.e. WordNet versus GermaNet. Apparently, GermaNet has a lower coverage than WordNet. E.g. no lowest common subsumers are found for some word pairs whereas those exist in WordNet, and some links are missing. Of course, the quality of a conceptual knowledge base and the quality of resulting glosses are strongly correlated.

6 Conclusions

We proposed a method to generate artificial definitions of concepts from a conceptual network. The method was applied to the task of computing semantic relatedness of words and tested on the basis of word senses defined in GermaNet. This approach bridges the gap between gloss based algorithms and the cases, when textual definitions of concepts are not available. This is the case for languages, which do not have well developed machine readable dictionaries, and in many applications which do have domain-specific taxonomies, but no additional descriptions of concepts. The main idea is to compensate for the lack of definitions in a conceptual hierarchy by generating a textual definition of the concept automatically from a knowledge base. NLP applications can then employ the resulting glosses.

We have restricted ourselves to nouns in this work, since this part of speech is very important in NLP and thus represents a good starting point. However, the metrics are applicable to other parts of speech represented in a conceptual network. The results of a semantic relatedness metric operating on automatically generated glosses correlate very well with human judgments of semantic relatedness. The metric performs significantly better than the Lesk algorithm itself, employing a traditional dictionary, and the baseline based on word co-occurrences in Web pages (Google hits). It performs on the same scale as the information content based metric, while no additional processing of corpus data is necessary.

We expect to enhance the work presented here in a number of respects. First of all, we are working on a considerably larger dataset including 350 word pairs with corresponding semantic relatedness judgments. The word pairs involve not only nouns, but verbs and adjectives as well. The reliability of human judgments as well as the performance of semantic relatedness metrics based on the new dataset remain to be studied. Also, we have to find our what kind of modifications may be necessary to make the metrics applicable across different parts-of-speech.

Acknowledgments

This work has been funded by the Klaus Tschira Foundation. I would like to thank Hendrik Niederlich who contributed implementations as a part of an internship and Michael Strube for his valuable comments concerning draft versions of this paper.

References

1. Lesk, Michael: 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, 1986, pages 24–26.
2. Hirst, Graeme and Budanitsky, Alexander: Correcting real-word spelling errors by restoring lexical cohesion. In Natural Language Engineering, 11(1):87–111, 2005.
3. Pedersen, Ted and Patwardhan, Siddharth and Michelizzi, Jason: WordNet::Similarity – Measuring the relatedness of concepts. In Intelligent Systems Demonstrations of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), *San Jose, CA, 25–29 July 2004.*

4. Rubenstein, Herbert and Goodenough, John: Contextual Correlates of Synonymy. In Communications of the ACM, 8(10), 1965, pages 627–633.
5. Miller, George A. and Charles, Walter G.: Contextual correlates of semantic similarity. In Language and Cognitive Processes, 6(1), 1991, pages 1–28.
6. Leacock, Claudia and Chodorow, Martin: Combining local context and WordNet similarity for word sense identification. In Fellbaum, Christiane (Ed.) WordNet: An Electronic Lexical Database, Cambridge: MIT Press, 1998, pages 265–283.
7. Seco, Nuno and Veale, Tony and Hayes, Jer: An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In Proceedings of the 16th European Conference on Artificial Intelligence, *Valencia, Spain, 22–27 August 2004*, pages 1089–1090.
8. Resnik, Phil: 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*, Volume 1, pages 448–453.
9. Jiang, Jay J. and Conrath, David W.: Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING), *Tapei, Taiwan*, 1997.
10. Lin, Dekang: An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, *San Francisco, Cal.*, pages 296–304, 1998.
11. Patwardhan, Siddharth and Banerjee, Satanjeev and Pedersen, Ted: Using measures of semantic relatedness for word sense disambiguation. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, *Mexico City, Mexico*, pages 241–257, 2003.
12. Ekedahl, Jonas and Golub, Koraljka: Word Sense Disambiguation using WordNet and the Lesk algorithm, http://www.cs.lth.se/EDA171/Reports/2004/jonas_koraljka.pdf, 2004.
13. Vaknin, Sam: The definition of definitions, http://samvak.tripod.com/define.html, 2005.
14. Kunze, Claudia: Lexikalisch-semantische Wortnetze. In Carstensen, K.-U. and Ebert, C. and Endriss, C. and Jekat, S. and Klabunde, R. and Langer, H. (eds.) Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg, Germany: Spektrum Akademischer Verlag, 2004, pages 423–431.
15. Fellbaum, Christiane (Ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass., 1998.
16. Kunze, Claudia and Lemnitzer, Lothar: GermaNet - representation, visualization, application. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), *Las Palmas, Canary Islands, Spain, 29 - 31 May*, 2002, pages 1485-1491.
17. Banerjee, Satanjeev and Pedersen, Ted: Extended gloss overlap as a measure of semantic relatedness. In Proceedings of the 13th International Joint Conference on Artificial Intelligence, *Chambery, France, 28 August – 3 September*, 1993.
18. Kilgarriff, Adam and Grefenstette, Gregory: Introduction to the special issue on the Web as a corpus. In Computational Linguistics, 29(3), 2003, pages 333–348.