Computing semantic similarity from bilingual dictionaries

Steffen Eger [1], Ineta Sejane [2]

[1] Institut für Deutsche Sprache – Mannheim – Germany

[2] Ruprecht-Karls-Universität Heidelberg – Heidelberg – Germany

Abstract

In this paper, we address the task of calculating mono- and bilingual semantic similarity. We introduce a method that, in order to arrive at a measure of semantic relatedness, exploits the information implicitly contained in bilingual dictionaries. Through experiments we show that our method performs well, with a performance comparable to approaches based on hierarchical knowledge bases and corpus statistics. The advantage of our approach is that it solely relies on easily available bilingual dictionaries and that it is capable of computing mono- and bilingual semantic relatedness at the same time.

Keywords: semantic similarity, bilingual semantic similarity, multilingual semantic similarity, cross-lingual semantic similarity, semantic relatedness, word relatedness, bilingual dictionaries

1. Introduction

In computational linguistics, the topic of computing semantic similarity (or semantic relatedness) has been widely studied in a number of different layouts during the last twenty years. While the issue of concern was mostly monolingual semantic similarity (e.g. How similar is English 'car' to English 'automobile'? How can this similarity be computed?), the data used to address this question was usually derived from monolingual dictionaries, encyclopedias, taxonomies, and wordnets such as WordNet or GermaNet. Moreover, some sort of probabilistic component (e.g. corpus statistics to quantify the probability of, say, concepts in a taxonomy) was typically added to the models developed. Very recently, Hassan and Mihalcea (2009) set out to tackle the problem of *multilingual* similarity (or *cross-lingual* similarity, as they call it), which is concerned with evaluating the relatedness of words from different languages. [1]

In the following, we present a model for computing semantic similarity that, contrary to most other models discussed in the literature so far, is not based on any sort of hierarchical or encyclopedic knowledge but relies solely on the information implicitly contained in bi- or multilingual dictionaries. The model rests on the simple idea that instead of merely looking up the translations of a word in a dictionary, one can also look up the translations of these translations, and so forth. In this way, one will find all words *u* (in any of the involved languages) 'similar' to the original word *w* in that there holds a (possibly extended) 'translation relation' between *u*

[1] Although the EuroWordNet project, which was carried out at the end of the 1990s, was concerned with multilingual wordnets and although some research was conducted there in the area of cross-lingual text retrieval (e.g. Gonzalo et al., 1999), no noteworthy publications have emerged from this project in the field of (cross-lingual) semantic similarity.

and *w*. This model allows both the computation of monolingual and bi- or multilingual semantic similarity at the same time. [2]

The paper is structured as follows. We discuss our method in the following section. In Section 3, we give an overview over the data we use, a German-Latin bilingual dictionary. In Section 4, we present an evaluation of our method that discusses the type of semantic relatedness (synonymy, hypernymy, etc.) that is captured using our approach, and that provides a comparison with human judgments of semantic similarity. Moreover, we illustrate the usefulness of our approach by pointing out several interesting applications in Section 5: automatically detecting candidates for missing entries in a bilingual dictionary; providing the user with more 'distant' translations of a word; word-sense differentiation; and language comparison. Before concluding in Section 7, we discuss previous and related work on semantic similarity in Section 6. This includes a contrasting of our approach with that presented in Hassan and Mihalcea (2009).

2. The Method

Assume that there are two natural languages A and B and a bilingual dictionary mapping between them. For the following discussion assume that X is the language A or B and $\overline{X}$ is the 'complement' of X, i.e. $\overline{X} = B$ if $X = A$ and $\overline{X} = A$ if $X = B$. If the language X word α is a translation of the language $\overline{X}$ word w, we will denote this by $wT\alpha$ ("w translates into α") and we will assume that the translation relation T is symmetric, i.e. xTy if and only if yTx.

Now, intuitively, if we want to judge the *bilingual similarity* between some language X word α and some language $\overline{X}$ word w, we can just count the minimal number of times we have to apply the translation relation 'operator' T in order to arrive at α starting from w, i.e. we determine the length of the shortest sequence (if indeed there exists such a sequence) $w_1T\alpha_1,\alpha_1Tw_2, \ldots, w_nT\alpha_n$ such that $w_1 = w$ and $\alpha_n = \alpha$ – this yields a *distance measure* $d \equiv d(w,\alpha)$ between w and α, from which we can easily derive a *similarity measure* $\mathrm{sim}(w,\alpha)$ by means of an inverse transformation, say $\mathrm{sim}(w,\alpha) = \exp(-d)$ or $\mathrm{sim}(w,\alpha) = 1/d$. We shall also call this process of repeatedly looking-up the translations of words the process of *repeated look-up*. We exemplify this concept in Fig. 1.

Note that the rationale for this definition of bilingual semantic similarity is that if w is a translation of α, then certainly w and α should be semantically very similar since translating can be considered as an act of providing a (partially) synonymous expression. Moreover, a 'long distance' – in terms of intermediate translations – between w and α should indicate decreasing semantic similarity because during each translation some degree of relatedness is lost.

Formally, we can define the just described distance measure $d(w,\alpha)$ between w and α as the length of the shortest path from w to α in the undirected graph $G = (V,E)$ where V consists of all words (of both languages) in the bilingual dictionary that maps from A to B, and E contains all edges $\{u,v\}$ such that uTv, $u, v \in V$. We shall also call this graph the *graph induced by the bilingual dictionary*. Note that in this way, by likewise computing shortest paths in the named graph, we can also determine the *monolingual similarity* between two words w and w' of the same language.

[2] Note, however, that if bilingual (or multilingual) resources are used for determining monolingual semantic relatedness, then the computed similarities will, to some degree, depend on both (or all) languages involved and their interplay, see also Figure 2. While for some applications this may be considered an undesirable side-effect, it can be very useful for others, see Section 5.

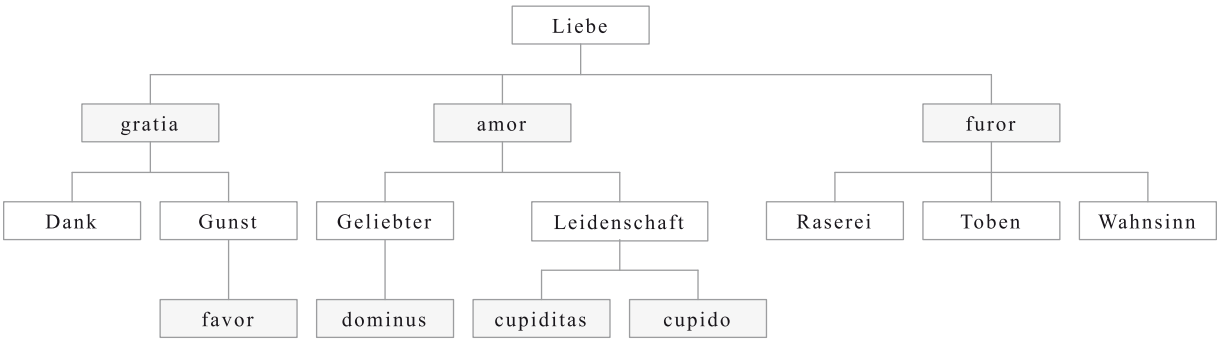

Figure 1: Assume that for the languages German and Latin, the hypothetical translational structure as depicted in the picture holds (indeed, this example is taken from our database), i.e. German Liebe (love) translates into Latin gratia (gratefulness), amor (sexuality) and furor (madness), which in turn have the indicated translations. Then we arrive at e.g. dominus (master) given Liebe by repeatedly looking up translations of previously translated words, starting with Liebe.

As a concrete example, in Fig. 1, the distance between w = *Liebe* (love) and α = *dominus* (master) is 3 and the distance between w = *Liebe* (love) and w' = *Leidenschaft* (passion) is 2 (or 1, if, on the shortest path from w to w', one only counts the number of words in the *same* language as the starting word, as we will usually do in the monolingual case; see also Fig. 2).

3. The Data

The Internet project Adiuvaris [3] is a Web 2.0 project focusing on the relationships between the languages German and classical Latin. In particular, it provides morphological information on Latin words and translational relationships between Latin and German entries. In the current study, we concentrate on this second aspect of information; in other words, our data consists of a simple bilingual database linking German words with their Latin translations. However, with regard to determining semantic similarity, we only consider nouns and adjectives because, in semantic respects, they allow for an easier comparison. It must be said, too, that our bilingual database is fairly small (only about 1000 nouns and adjectives each), and the quality of the translational links provided is – compared with first class bilingual dictionaries – rather poor because the Adiuvaris project is a joint effort of generally linguistically untrained users. Still, we decided for this dictionary because it is freely available and because our method might prove particularly valuable in the situation just described (see also Section 5). Tab. 1 lists some important details on our data set.

As already indicated in the previous section, when computing the semantic similarity between words in our database, we are interested in the distance associated with the shortest path between the respective words in the graph induced by the bilingual dictionary. This graph can be used for determining bilingual semantic similarity or monolingual semantic similarity.

4. Evaluation

4.1. Semantic Similarity

Because there are many more speakers fluent in German than there are speakers fluent in both German and Latin, we decided to test the quality of our method in terms of monolingual semantic

[3] http://www.adiuvaris.de.

similarity. Moreover, since the vocabulary of our Latin-German data set is considerably different from the vocabulary of standard reference data sets used for evaluating semantic similarity (e.g. the Miller-Charles data set (Miller and Charles, 1991) or the WordSimilarity-353 data set (Finkelstein et al., 2001)), we determined to make up our own evaluation set. For this, we selected 10 German nouns and 10 German adjectives – 5 of each category were chosen randomly, the other 5 were chosen in accordance with what we thought to be good candidates for a wide semantic field associated with them (e.g. *weiß* (white), *schwarz* (black), *Liebe* (love), *Tod* (death), etc.). For each such German word, we determined the sets of German words at distances 1, 2, 3, etc. in the monolingual graph induced by our dictionary, and from each of these sets we randomly picked up to 3 words. Additionally, we selected for each German word 3 more German words which, according to our method, had a similarity value of zero, i.e. there was no path between them in the monolingual graph. Altogether, we thus arrived at approximately 1000 word pairs which we had 13 native speakers of German evaluate. We thereby used a definition of semantic similarity similar to that employed by Gurevych and Niederlich (2005), and allowed the human subjects to consider any type of semantic relation including association when judging semantic relatedness. The scale for judging this similarity was from 0 (not related) to 5 (high degree of relatedness/identity).

	Nouns	*Adjectives*
Number (German)	1233	617
Average monolingual degree of node (German)	2.8	4.0
Number of connected components (German)	338	92
Number of connected components (German) > 10	6	2
Longest distance in graph (German)	21	15

***Table 1**: The table summarizes important information on our database and the noun and adjective graphs induced by it; since in the evaluation section of this paper, we focus on monolingual semantic similarity, we give here numbers related to German only. The* monolingual degree of a node *denotes the number of edges leaving a node in the monolingual graphs induced by our dictionary; this variable is related to the number of translations given for a word in the dictionary. In our case, the low average values (*2.8 *and* 4.0*) indicate that our dictionary may have a lot of missing translations. The number of* connected components *in the named graphs denotes the number of (maximal) word clusters that are unrelated to other clusters, in the sense that there is no path between the words in different clusters. We are certain that if the dictionary were more complete, i.e. if there were more translations provided for a given word, there would ultimately be a path from each word to every other word. The table also shows that there are only few large clusters (with more than* 10 *members). The* longest distance in the graph *denotes the length of the longest shortest path in the noun and adjective graphs, respectively. For nouns, we have* 21 *as the distance between* Feldfrucht *(crop) and* Fälschung *(fake); for adjectives, we have* 15 *as the distance between* bunt *(colorful) and* gewohnt *(familiar).*

Finally, we compared the human judgments of semantic similarity with the semantic similarity computed by our method. Here we used, as described above, as a distance measure between two words w and w' the distance of the shortest path between these words in the monolingual graph induced by the bilingual dictionary. The similarity between w and w' was then calculated as the inverse path length,

$$\mathrm{sim}(w,w') = \begin{cases} \dfrac{1}{d(w,w')} & \text{if } d(w,w') < \infty \\ 0 & \text{else} \end{cases}$$

where $d(w,w')$ is the distance between w and w' (this value is defined as infinite if there is no path between w and w').

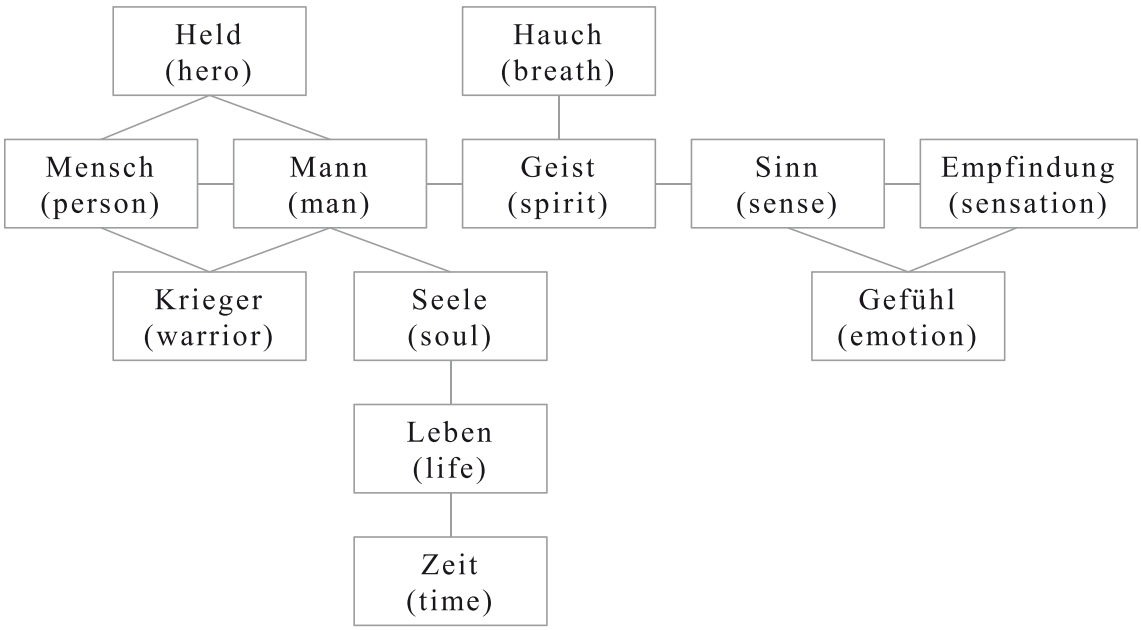

Figure 2: Given a graph network induced by a bilingual dictionary, one can leave out all nodes labeled with one of the languages (in this case, Latin language nodes were left out) in order to obtain a network that depicts monolingual word relatedness. In our case, we see e.g. that the shortest distance between Mensch and Mann is 1 (indicating very close similarity), whereas the distance between Mensch and Zeit is 4, which indicates a more distant relatedness. Note also how in this network (partial) synonymy between adjacent words translates into considerable semantic difference between words further apart. Moreover, note how the illustrated mono-lingual structuring of words is affected by the bilingual perspective: for example, the closeness between Mann and Held seems to be inherited from the translations in the Latin language (and does possibly not reflect the conceptualizations of modern German).

Tab. 2 lists the results of our evaluation. Overall, our method succeeds in capturing the semantic similarity between words. While the correlation between our approach and the human evaluators is comparatively low (compared e.g. with the correlation values for the German language in Gurevych and Niederlich (2005)), human-human correlation is in our situation likewise quite low, and our method almost comes up to that upper bound. Moreover, when one scales up respectively – human-human correlation is by a factor of about 1.5 higher in e.g. Gurevych and Niederlich (2005) – our method nearly matches the best-performing systems for determining semantic similarity. When one considers that these systems usually rely on cost-intensive taxonomies and non-negligible employment of corpus statistics, our dictionary-based approach looks like an attractive alternative; additionally, we are confident that as dictionary size and quality increases, our methodology's performance will increase, too. Also, it should not be forgotten that our approach allows both the computation of mono- and bilingual semantic relatedness, an advantage most other systems cannot offer.

As to the low value of human-human correlation, we think that this might be mainly due to the fact that while most other studies rely on human-defined evaluation sets where the semantic relations to be considered can be chosen in a way that allows 'easy' assessment, our approach of randomly picking word pairs may have complicated the task for the human subjects.

Average correlation	*Human$_1$*	*Human$_2$*	*RLU*	*Contextual*	*JC*
Human$_1$	0.54		0.46 (0.85)	0.42 (0.77)	
Human$_2$		0.81			0.74 (0.91)

Table 2: *Average Pearson correlation coefficient between human-human judgments and between human-computer judgments. Note that the human-human correlation represents the upper bound of performance in the evaluation of the computer-based approaches. 'Human$_1$' and 'Human$_2$' refer to the human subjects assessing the evaluation sets in our evaluation and the evaluation of Gurevych and Niederlich (2005), respectively. 'RLU' is the method of repeated look-up described in this paper. 'Contextual' is a similarity measure based on contextual similarity of words described in Keibel and Belica (2007). 'JC' is the best-performing method for computing semantic similarity for German in Gurevych and Niederlich (2005). For the computer systems, we give in brackets the value of their performance relative to their respective upper bound, i.e., the human-human coefficient.*

4.2. Semantic Relations

In order to find out about the type of semantic relations that are captured with our method, we had a linguistically trained expert analyze the above described evaluation set. The task was to classify the relation between each word pair with one of the categories 'S' (synonymy or quasi-synonymy), 'H' (hypernymy or meronymy), 'RT' (arbitrarily related term), 'A' (antonymy or quasi-antonymy), and 'NA' (no relation).

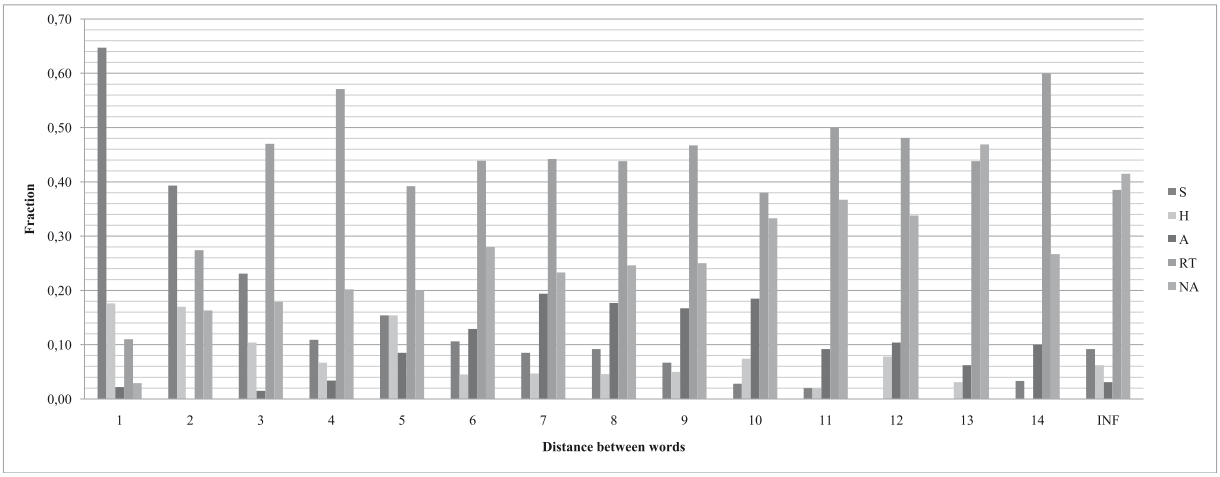

Figure 3: The types of semantic relations that words have at different distances (lengths of shortest paths between the words). The columns at the far right depict the semantic relations between words when there was no path connecting them, i.e., when our method postulated a similarity value of zero

Fig. 3 shows that when words have a short distance (distance 1 or 2) in the monolingual graph induced by the bilingual dictionary, then these words are usually synonymously related, in the judgments of a human expert. Moreover, when the distance between words increases, then the probability of synonymy decreases, while antonymy becomes more prevalent. [4] The hierarchical relations of hypernymy and meronymy have a similar behavior as synonymy. Fig. 3 also shows that even words with a very long distance in the named graph (greater than 10) are frequently judged to be related at least in some way by a human expert, although the

[4] An example of how partial synonymy translates into antonymy is the following transition, taken from our data: *weiß* (white) ↔ *aufrichtig* (candid) ↔ *gewissenhaft* (conscientious) ↔ *ernst* (earnest) ↔ *traurig* (sad) ↔ *schwarz* (black).

probability of relatedness (in the eye of the human expert) decreases with distance. When our system postulates a similarity of zero between two words, then in almost half of the cases the human judge is not able to discern any kind of similarity as well.

5. Possible Applications

Automatically determining mono- and bilingual semantic similarity from a dictionary can be useful in a variety of different situations. First of all, when dictionaries are created not by linguistically trained experts but, say, 'ordinary' Internet users (as, for example, in the case of Web 2.0), bilingual semantic similarity can help the *developers* find 'omitted' or 'forgotten' translations by offering a set of words in the 'target' language that have a high similarity value to the given word in the 'source' language. Using our method, no external resources are needed for this goal but the dictionary itself. Secondly, our method may be helpful by providing not only translations for a given word but showing the *users* all kinds of related words to the word in question. These relations may include synonymy, antonymy, hypernymy, meronymy, etc., and our method might thus assist the translator or dictionary user in finding a translation that is possibly more adequate than the one he had in mind. Again, this happens by automatically exploiting the structure that is implicitly contained within the dictionary itself.

While we have now stressed applications related to the *bilingual* structure of a dictionary, it must be said that (our approach to determining) semantic similarity can also be used to differentiate meaning potentials of a word in *one* of the languages. For example, Keibel and Belica (2007) use contextual information of words to calculate their respective similarity and then select for a given word those words that are most similar to it in order to contrast these "near-synonyms" by means of a self-organizing map (SOM); in this way, some sort of word-sense differentiation is performed. With our approach, the same is feasible by declaring those words with a low distance to the word under consideration – recall from Fig. 3 that these words are usually synonymously related to the given word – as this word's near-synonyms whose meaning aspects can then be discriminated by a SOM. Note that in this situation, when our approach is taken, the near-synonyms do not only depend on the quality of the dictionary (a decisive factor in our approach) but also on the respective other language in the bilingual dictionary. For example, the near-synonyms of a given German word might be at least partially different when a German-Latin dictionary is used from the near-synonyms when a German-Chinese dictionary is used. Thus our method might shed light both on the lexical meanings of German words and the different aspects of conceptualizations in different languages.

To be a little bit more precise on this last issue, if there exists a multilingual database of dictionaries with a distinguished reference language serving as an 'intermediary' (say, bilingual dictionaries German-Latin, German-Chinese, German-French, etc.), then by computing monolingual word similarities for this reference language for varying partner languages in the database, it should be possible to contrast and compare these partner languages in terms of single items (e.g. How similar is *Mann* (man) to *Krieger* (warrior) using a German-Latin and a German-Chinese dictionary, respectively?) or in terms of the whole vocabulary. Moreover, the approach then taken would be bottom-up and exclusively data-driven, as opposed to the top-down and introspective approach taken by other systems (e.g. the diverse wordnets).

A few more applications of mono- and bilingual semantic similarity will be discussed in the next section.

6. Related Work

As already indicated in Section 1, past measures of word relatedness usually exhibited the following features: 1) they investigated monolingual semantic similarity, 2) they were based on hierarchical knowledge bases or encyclopedias and statistics (e.g. Lesk, 1986; Niwa and Nitta, 1994; Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998; Roth and Schulte im Walde, 2008). The similarity measures were applied to a variety of NLP applications, including word-sense disambiguation (e.g. McCarthy et al., 2004), coreference resolution (e.g. Strube and Ponzetto, 2006), and many others.

As mentioned, most of these studies were concerned with *monolingual* semantic relatedness. Only recently have Hassan and Mihalcea (2009), applying corpus statistics to encyclopedic knowledge derived from Wikipedia, investigated the possibilities of *multilingual* relatedness, which can also be implemented with our approach. They claimed that their work is most closely related to cross-lingual information retrieval, where the task is to retrieve documents in one language for queries posed in another. Moreover, they claimed that their approach is superior to traditional dictionary-based approaches in this field in that it is not only able to consider direct translations of words, but the "entire spectrum of relatedness". To use their own example, if one is seeking documents in Italian that match the English query word *factory*, a method based on translations in bilingual dictionaries would, they say, typically return documents related to *fabbrica* (It.) but oversee the similarity between *lavoratore* (It.; 'worker' in English) and the original *factory*. We showed that a dictionary-based approach is, quite contrarily, capable of detecting word relatedness that goes beyond a direct translation and that the knowledge of such similarity lies within the dictionary itself.

Further areas related to the computation of semantic similarity (and in particular to our dictionary-based approach) are the areas of word alignment for machine translation (e.g. Och and Ney, 2000) and induction of translation lexicons (e.g. Schafer and Yarowsky, 2002).

7. Conclusions

In this paper, we addressed the problem of determining mono- and bilingual semantic similarity. We introduced a method that, contrary to most other previous work, does neither rely on encyclopedias, taxonomies, wordnets, nor any other kind of hierarchical knowledge base. In our approach, we exploit the information implicitly contained in bilingual dictionaries by counting the 'translational distances' between words in these dictionaries as a measure of semantic similarity. Experiments performed on the language pair Latin-German showed that our method is effective at capturing semantic similarity and can compete both with human judgments of semantic relatedness and with other computer systems. The advantages of our approach are that dictionaries are in general much more easily available than taxonomies and are cheaper to produce – which makes our methodology attractive e.g. for languages for which there exist no or only few sophisticated electronically available language resources such as hierarchical knowledge bases – and that it is suitable for computing both mono- and bilingual semantic similarity in one step. Concerning applications, our method can be used for word-sense differentiation and for contrasting languages by computing the monolingual word similarities of a reference language for varying partner languages in a multilingual database. Additionally, it is useful in automatically suggesting missing translations in a dictionary and providing the user with a multitude of related words instead of only the bare translation of the input word.

In future work, it would be interesting to incorporate further language pairs in our system and thus extend it to handle multilingual semantic similarity. Moreover, corpus statistics could aid in improving system performance by suggesting links not present in the dictionaries.

References

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G. and Ruppin E. (2001). Placing search in context: the concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, *Hong Kong*, pp. 406-414.

Gonzalo J., Verdejo F. and Chugur I. (1999). Using EuroWordNet in a concept-based approach to cross-language text retrieval. *Applied Artificial Intelligence*, 13(7): 647-678.

Gurevych I. and Niederlich H. (2005). Computing semantic relatedness of GermaNet concepts. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pp. 5-8.

Hassan S. and Mihalcea R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1192-1201.

Jiang J. and Conrath D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING X)*.

Keibel H. and Belica C. (2007). CCDB: A corpus-linguistic research and development workbench. In *Proceedings of the 4th Corpus Linguistics conference*.

Lesk M. (1986). Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pp. 24-26.

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

McCarthy D., Koeling R., Weeds J. and Carroll J. (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 280-287.

Miller G. and Charles W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1): 1-28.

Niwa Y. and Nitta Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th conference on Computational linguistics*, Vol. 1, pp. 304-309.

Och F. and Ney H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*.

Resnik P. (1995). Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

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

Schafer C. and Yarowsky D. (2002). Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL 2003)*.

Strube M. and Ponzetto S. P. (2006). Wikirelate! computing semantic relatedness using Wikipedia. In *AAAI'06*, pp. 1419-1424.