Building a Cross-lingual Relatedness Thesaurus using a Graph Similarity Measure

Lukas Michelbacher, Florian Laws, Beate Dorow, Ulrich Heid and Hinrich Schütze

Institute for Natural Language Processing
Universität Stuttgart
michells,lawsfn,dorowbe,heid@ims.uni-stuttgart.de, hs999@ifnlp.org

Abstract

The Internet is an ever growing source of information stored in documents of different languages. Hence, cross-lingual resources are needed for more and more NLP applications. This paper presents (i) a graph-based method for creating one such resource and (ii) a resource created using the method, a cross-lingual relatedness thesaurus. Given a word in one language, the thesaurus suggests words in a second language that are semantically related. The method requires two monolingual corpora and a basic dictionary. Our general approach is to build two monolingual word graphs, with nodes representing words and edges representing linguistic relations between words. A bilingual dictionary containing basic vocabulary provides seed translations relating nodes from both graphs. We then use an inter-graph node-similarity algorithm to discover related words. Evaluation with three human judges revealed that 49% of the English and 57% of the German words discovered by our method are semantically related to the target words. We publish two resources in conjunction with this paper. First, noun coordinations extracted from the German and English Wikipedias. Second, the cross-lingual relatedness thesaurus which can be used in experiments involving interactive cross-lingual query expansion.

1. Introduction

The Internet is an ever growing source of information stored in documents of different languages. Hence, cross-lingual resources are needed for more and more NLP applications. This paper presents (i) a graph-based method for creating one such resource and (ii) a resource created using the method, a *cross-lingual relatedness thesaurus*. Given a word in one language, the thesaurus suggests words in a second language that are semantically related. A cross-lingual relatedness thesaurus is valuable for a number of applications, e.g., for query expansion in cross-language information retrieval (Grefenstette, 1998).

For the German word *Löwe* (*lion*), for example, the method described below identifies the following ten words as most related: *cheetah*, *panther*, *rhino(ceros)*, *tiger*, *jaguar*, *leopard*, *hyena*, and *cub* as well as the actual translation, all of which are wild animals.

The method requires two monolingual corpora and a basic dictionary. Our general approach is to build two monolingual word graphs, with nodes representing words and edges representing linguistic relations between words. A bilingual dictionary containing basic vocabulary provides seed translations relating nodes from both graphs. We then use an inter-graph node-similarity algorithm to discover related words.

We build the graphs based on noun coordinations because coordinations are well suited to model the semantic relatedness of nouns. We believe, however, that our method is applicable to other parts-of-speech as well as using the appropriate linguistic relations. We make two resources, the noun coordinations and the cross-lingual relatedness thesaurus, available to the public (Section 6.).

2. Related Work

Hassan and Mihalcea (2009) presented a method that calculates semantic relatedness of words across languages. In contrast to our approach, in their method words are represented using explicit semantic analysis (Gabrilovich and Markovitch, 2007). The approach uses Wikipedia's inter-language links to map concept vectors across languages.

The IR system presented by Hsu et al. (2008) implements cross-lingual query expansion via a two-stage process. First, queries are translated using online translation services and Wikipedia inter-language links. The subsequent query expansion step incorporates the anchor text of normal Wikipedia links. In our system, cross-lingual query expansion is integrated into one step. A two-stage process that separates the translation and the expansion step may compound error of individual steps.

Defrancq (2008) conducted a contrastive study on semantic relatedness between verbs in different languages which also uses monolingual corpora. A set of English, French, Dutch and Spanish verbs are compared using KL-divergence (Lee, 1999). The underlying distributions are based on the verbs' cooccurrences with interrogative elements (e.g. *He said* how *it happened*). The method identifies 69 out of 99 pre-defined verb equivalences. It differs from our experiment in a number of ways: the test words are manually selected (less than 10 verbs per language), it is restricted to newswire, the extraction of cooccurrences is semi-automatic, the total of cooccurrences is small (just over 10,000). Furthermore, it focuses on a very specific phenomenon, namely verbs that appear with interrogative elements. Our approach avoids the aforementioned restrictions and has a broader focus.

For a given word, our method suggests semantically related words, mainly (co-)hyponyms and hypernyms and exact translations. However, no information about the nature of the lexical relation between a source word and its related items is given. A recently presented system (Baroni et al., 2009) goes in this direction, trying to induce concepts and properties as well as conceptual hierarchies from POS-tagged text. Their method is not based on graph-theory and

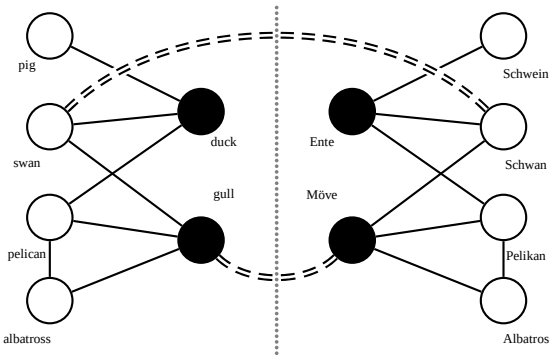

Figure 1: Similarity through seed translations

it is only applied monolingually.

3. Method

3.1. Model

Our method for finding related words uses two main building blocks: graphs representing words and the relationships between them and a measure of similarity between words based on these relationship graphs. We use separate graphs for each language, with words represented as nodes and relationships as edges, but we compute word similarity across the two monolingual graphs with an inter-graph similarity algorithm.

This algorithm is based on SimRank (Jeh and Widom, 2002). SimRank recursively computes node similarities based on the similarity scores of neighboring nodes within a graph. Dorow et al. (2009) proposed an extension that computes node-similarities across two graphs and allows for weighted graph edges.

SimRank is a recursive algorithm that is based on the idea that two nodes in a graph are similar if the neighbors are similar. We extend this notion to inter-graph similarity. We think of two words as being related if they have neighboring words that are also related, or belong to a set of initial node-to-node correspondences between the two graphs. Correspondences are translations ("seed translations") provided by a dictionary. These node pairs are assigned the similarity value 1 (maximum similarity). The similarity then "spreads" to neighboring bilingual node pairs, and by repeated application of the algorithm reaches all nodes.

Figure 1 illustrates this idea. Double lines indicate seed translations. The nodes *duck* and *Ente* occur in coordinations with the same nouns in the two languages; one of these (*swan – Schwan*) is a seed translation. This coordination relationship contributes to the similarity of *duck – Ente*. Also, *pelican* and *Pelikan* are similar (because of *gull – Möwe*) and this similarity will also contribute to *duck – Ente* in a later iteration.

3.2. SimRank algorithm

SimRank (Jeh and Widom, 2002) computes similarity scores S_{ij} of a node pair ij as the average pairwise similarity of neighboring nodes:

$$S_{ij} = \frac{c}{|N(i)|\ |N(j)|} \sum_{k \in N(i), l \in N(j)} S_{kl}.$$

where $N(i)$ and $N(j)$ are the sets of i's and j's neighbors. The constant c ($0 < c < 1$) attenuates the contribution of nodes further away. Following Jeh and Widom (2002), we use $c = 0.8$.

Since the original formulation of SimRank only allows for monolingual similarity calculations, we use the formulation proposed by Dorow et al. (2009) for two graphs. In this case, the nodes i and j simply come from two separate graphs A and B. As the basis for the recursion, the initial node-to-node correspondences are used such that $S_{ij} = 1$ if i and j are a pair in the set of correspondences (seed translation pairs). Furthermore, the formulation also allows for weighted graphs by simply multiplying the similarity with the entries in the weighted adjacency matrices A and B.

$$S_{ij} = \frac{c}{|N_A(i)|\ |N_B(j)|} \sum_{k \in N_A(i), l \in N_B(j)} A_{ik}\ B_{jl}\ S_{kl}$$

where S_{ij} is the similarity of the nodes i and j of graphs A and B, respectively. A and B are the weighted adjacency matrices of the graphs A and B, and $N_A(i)$ and $N_B(j)$ are the sets of neighbors.

See (Dorow et al., 2009) for details, and also for an equivalent formulation of the iteration using matrix multiplications, which we also used for the experiments.

3.3. Data

We use the English and German Wikipedias as corpora, processed with JWPL (Zesch et al., 2008). Both corpora were lemmatized and part-of-speech-tagged (Schmid, 1994). In the graph, nouns (nodes) are connected if they appeared together in coordinations (e.g., *men and women* or *communism, collectivism, or participatory economics*). For coordinations with n elements, we select all $\binom{n}{2}$ combinations of word pairs. We discard information about the order in which the elements appear by sorting all pairs alphabetically. We do not distinguish between the pairs (w_1, w_2) and (w_2, w_1). For this experiment, we focus on the cross-lingual aspect of semantic relatedness, ignoring potential benefits of an asymmetrical view (Michelbacher et al., 2007). Graph theory does, however, allow for asymmetric association via directed edges. We leave this asymmetrical aspect of relatedness for future research.

We used log-likelihood association scores as edge weights, and removed edges with scores below 3.84, the critical value at significance level 0.05 of the log-likelihood test (Snedecor and Cochran, 1989). As seed translations, we used 4220 pairs from the dict.cc English-German online dictionary[1].

3.4. Evaluation

We evaluated our method on a test set proposed by Rapp (1999). We selected the 53 nouns contained in the test set. Manual evaluation was carried out by three students (two German native speakers, one English-German bilingual) each annotating the complete test set. The annotators were given a print-out of the test words for each language together with the top ten suggested words in the other language.

[1] `http://www.dict.cc/` (May 5th 2008)

The annotators were asked to mark cohyponyms (C), hypernyms (R), hyponyms (H) and exact translations (E) among the top ten list. We also offered the category "other" (O) for words that were semantically related in a way that does not fit into one of the aforementioned categories. The annotators were allowed to use an English-German dictionary.
In the annotation instructions we defined cohyponyms as two words that share a hypernym, for example *cheese* and *yoghurt* which have a common hypernym, namely *dairy product*. The annotators were made aware that technically, two concepts can always be considered cohyponyms since they have a common trivial hypernym (e.g. *thing*) but it was made clear that this was not the desired interpretation for the experiment.

4. Results and Discussion

Before we turn to the systematic evaluation with human subjects, consider a real example of related words suggested by our method. Table 1 shows the example pair *anger* and its German translation *Zorn*. All suggested words describe emotions with either a clear or conceivable negative connotation. This observation is reflected in the annotators' assessment. All German suggestions were consistently labeled (C) by all subjects (with one (E) for *Erregung* by one subject). A look at the English suggestions for *Zorn* reveals a similar picture. All annotators agreed that the suggestions are category (K). In addition, they classified three suggestions as exact translations. Unanimously, *fury* was labeled (E); *wrath* and *rage* received two votes, respectively. One annotator assessed *hate* as a true translation of *Zorn*.
In the systematic evaluation, the method yields promising results. As shown in Table 2, 57% of the top ten ranked words for DE $\rightarrow$ EN (i.e., the test word was German and the suggested words were English), and 49% words for EN $\rightarrow$ DE are semantically related. Most of the related words are cohyponyms, followed by "other" semantic relations. Examples of category (O) include part-of relations such as *moon - galaxy*, but also more abstract concepts such as *man - manhood*. Sometimes, the annotators also chose category (O) for less specific cohyponyms. For example for *butter*, one annotator chose milk products such as *yoghurt* as proper cohyponyms, and (O) for other foods such as *honey*.
We verified that the annotators chose non-trivial cohyponyms by using WordNet (Fellbaum, 1998). We looked up the length of the path from the root node of the WordNet taxonomy to the lowest common subsuming hypernym (LCS). For example, in Figure 2, *apple* and *strawberry* have the lowest common subsumer *fruit.n.01*, which is eight nodes away from the root node *entity.n.01*. The average path length is over 5. This confirms that most cohyponyms are non-trivial. In the few cases where there were short path lengths, we manually checked the cohyponym pairs in question and found that the assumed common hypernym was not part of the paths to WordNet's root node. E.g. for *fruit* and *seafood*, the common hypernym chosen by the annotators, *food*, is not an LCS in WordNet, even when checking all senses of the words.
We calculated inter-annotator agreement using Cohen's κ (Artstein and Poesio, 2008). The average of the pairwise inter-annotator agreement is 0.57 for EN $\rightarrow$ DE and 0.49 for DE $\rightarrow$ EN. There was a noticeable discrepancy between the two directions in the experiment. On the one hand, the annotators annotated more semantic relations when the suggested words were in English but on the other, the average agreement among them was better when the suggested words were German. We believe that the annotators were able to make more consistent judgements with the suggested words in their native language. Further experimentation is needed to determine the cause of the lower performance with the suggested words in English.
When leaving out the (O) category, performance decreases to 43% for DE $\rightarrow$ EN and 39% for EN $\rightarrow$ DE, but inter-annotator agreement rises to 0.54 for DE $\rightarrow$ EN and 0.62 for EN $\rightarrow$ DE. This seems plausible since the definition of the (O) category is broader, allowing matches more easily but also allowing more disagreeing interpretation. The fact that agreement for (E) is 0.85 (EN $\rightarrow$ DE) and 0.81 (DE $\rightarrow$ EN) supports this claim since exact translations leave less room for interpretation.
Unrelated words among the suggestions are often caused by polysemy. For example, Table 3 shows the top ten suggested words for *chair*, which are predominantly financial terms, as opposed to pieces of furniture, as one might expect. This is likely caused by *chair* being strongly associated with *bench*, whose German translation *Bank* also has the meaning of *bank*, the financial institution. We describe a possible remedy for this problem in the next section.

5. Possible Improvements

We intend to reduce the kind of ambiguity mentioned above by taking the *contexts* of words into account instead of individual words. In the graph model, edges between words provide context information that can be used for sense disambiguation (e.g., *"chair and table"*).
This information can be incorporated into the graph similarity framework by constructing the incidence graph $\mathcal{I}(\mathcal{A})$ out of the original graph $\mathcal{A}$ (cf. Figure 3).
Given a graph $\mathcal{A}$, we can construct a new graph $\mathcal{I}(\mathcal{A})$ by putting a new node on each link of $\mathcal{A}$. The resulting graph is bipartite. Its two vertex sets are the nodes and the links in $\mathcal{A}$ respectively, and its edges connect the links in $\mathcal{A}$ with their two endpoints. The incidence graph has as many nodes as

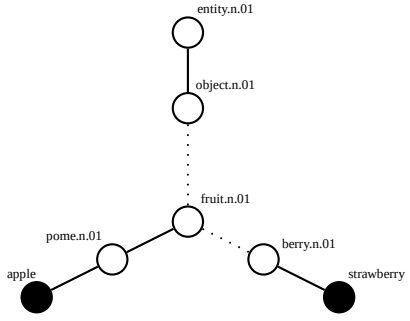

Figure 2: Cohyponym check using WordNet

rank	related word
1	Schrecken (*horror*)
2	Scham (*shame*)
3	Erregung (*enragement*)
4	Unzufriedenheit (*discontent*)
5	Sorge (*sorrow*)
6	Mißtrauen (*distrust*)
7	Mitleid (*compassion*)
8	Eitelkeit (*vanity*)
9	Unsicherheit (*insecurity*)
10	Verzweiflung (*despair*)

rank	related word
1	fury
2	wrath
3	avarice
4	dread
5	jealousy
6	hate
7	rage
8	envy
9	indignation
10	insecurity

Table 1: Ten most related words to the translation pair *anger* (left) and *Zorn* (right)

	cohyponyms (C)	hyponyms (H)	hypernyms (R)	exact (E)	other (O)	**total**
DE → EN	28%	5%	2%	7%	15%	**57%**
EN → DE	22%	5%	3%	8%	11%	**49%**

Table 2: Percentage of semantically related items

rank	related word
1	Sparkasse (*savings bank*)
2	Versicherung (*insurance*)
3	Börse (*stock market*)
4	Einlage (*investment*)
5	Zentralbank (*central bank*)
6	Kreditinstitut (*credit institution*)
7	Eingabe (*input*)
8	Corporation (*corporation*)
9	Konzern (*corporate group*)
10	Tisch (*table*)

Table 3: Ten most related German words to *chair*

there are nodes and edges in the original graph, and twice as many edges as the original graph.

The bilingual SimRank algorithm can then be run on the incidence graphs. For effective sense disambiguation, however, we need to provide sense-discriminating seed translations. For this, we will abandon word equivalences in favor of link equivalences which are established through translations of pairs of words that appear in coordinations in both languages. For example the equivalence $(\textit{rock}, \textit{gravel}) - (\textit{Fels}, \textit{Kies})$ would be part of the new seed set.

With this light-weight word sense disambiguation our approach can also be adapted to the task of bilingual lexicon extraction.

6. Resources

In the course of the experiments described in this paper we prepared two data sets that we believe to be useful to the research community. We thus made these data sets available for free download at `http://www.ifnlp.org/wiki/extern/WordGraph`

Noun coordinations We extracted lists of noun coordinations for the experiments described above. There are approximately 5 million English coordinations and 2 million German coordinations.

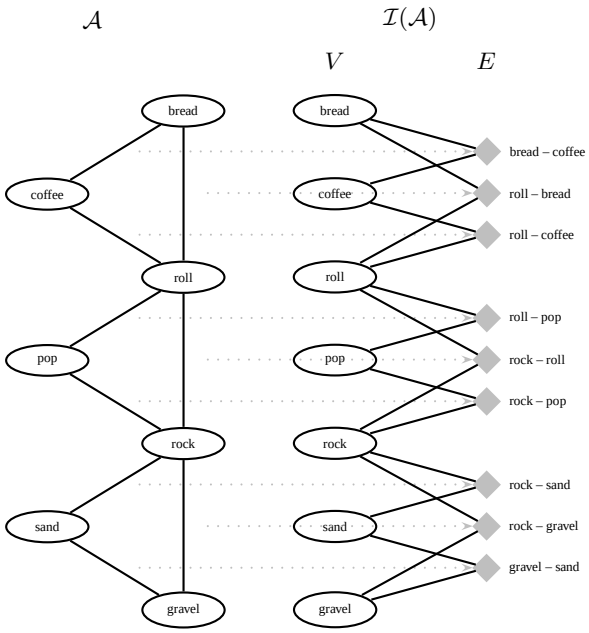

Figure 3: Example of a graph $\mathcal{A}$ and its incidence graph $\mathcal{I}(\mathcal{A})$

Thesaurus data set The thesaurus of related words produced by our method is available for experiments. The data set contains top-ten lists for 9000 English words (EN → GE) and 6000 German words (GE → EN).

7. Conclusion

We have presented a method for creating a cross-lingual relatedness thesaurus. With our approach, cross-lingual query expansion can be carried out in one step. Evaluation with three human judges revealed that 49% of the

English and 57% of the German words discovered by our method are semantically related to the target words. We publish two resources in conjunction with this paper. First, the noun coordinations extracted from the German and English Wikipedias. Second, the cross-lingual relatedness thesaurus which can be used in experiments involving interactive cross-lingual query expansion.

8. Acknowledgement

This research was funded by the *German Research Foundation* (DFG) within the project *A graph-theoretic approach to lexicon acquisition*.

9. References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), December.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2009. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*.

Bart Defrancq. 2008. Establishing cross-linguistic semantic relatedness through monolingual corpora. *International Journal of Corpus Linguistics*, 13(4):465–490.

Beate Dorow, Florian Laws, Lukas Michelbacher, Christian Scheible, and Jason Utt. 2009. A graph-theoretic algorithm for automatic extension of translation lexicons. In *EACL 2009 Workshop on Geometrical Models of Natural Language Semantics*.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI 2007*.

Gregory Grefenstette, editor. 1998. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, Boston, MA.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *EMNLP 2009*. Association for Computational Linguistics.

Chih-Chuan Hsu, Yu-Te Li, You-Wei Chen, and Shih-Hung Wu. 2008. Query expansion via link analysis of wikipedia for clir. In *7th NTCIR Workshop*, Tokyo, Japan.

Glen Jeh and Jennifer Widom. 2002. Simrank: A measure of structural-context similarity. In *KDD '02*, pages 538–543.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, College Park, Maryland, USA, June. Association for Computational Linguistics.

Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. 2007. Asymmetric association measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2007*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *COLING 1999*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

George Waddel Snedecor and William G. Cochran. 1989. *Statistical methods*. Iowa State University Press.

Torsten Zesch, C. Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.