Sentiment Translation through Lexicon Induction

Christian Scheible
Institute for Natural Language Processing
University of Stuttgart
scheibcn@ims.uni-stuttgart.de

Abstract

The translation of sentiment information is a task from which sentiment analysis systems can benefit. We present a novel, graph-based approach using SimRank, a well-established vertex similarity algorithm to transfer sentiment information between a source language and a target language graph. We evaluate this method in comparison with SO-PMI.

1 Introduction

Sentiment analysis is an important topic in computational linguistics that is of theoretical interest but also implies many real-world applications. Usually, two aspects are of importance in sentiment analysis. The first is the detection of subjectivity, i.e. whether a text or an expression is meant to express sentiment at all; the second is the determination of sentiment orientation, i.e. what sentiment is to be expressed in a structure that is considered subjective.

Work on sentiment analysis most often covers resources or analysis methods in a single language, usually English. However, the transfer of sentiment analysis between languages can be advantageous by making use of resources for a source language to improve the analysis of the target language.

This paper presents an approach to the transfer of sentiment information between languages. It is built around an algorithm that has been successfully applied for the acquisition of bilingual lexicons. One of the main benefits of the method is its ability of handling sparse data well.

Our experiments are carried out using English as a source language and German as a target language.

2 Related Work

The translation of sentiment information has been the topic of multiple publications.

Mihalcea et al. (2007) propose two methods for translating sentiment lexicons. The first method simply uses bilingual dictionaries to translate an English sentiment lexicon. A sentence-based classifier built with this list achieved high precision but low recall on a small Romanian test set. The second method is based on parallel corpora. The source language in the corpus is annotated with sentiment information, and the information is then projected to the target language. Problems arise due to mistranslations, e.g., because irony is not recognized.

Banea et al. (2008) use machine translation for multilingual sentiment analysis. Given a corpus annotated with sentiment information in one language, machine translation is used to produce an annotated corpus in the target language, by preserving the annotations. The original annotations can be produced either manually or automatically.

Wan (2009) constructs a multilingual classifier using co-training. In co-training, one classifier produces additional training data for a second classifier. In this case, an English classifier assists in training a Chinese classifier.

The induction of a sentiment lexicon is the subject of early work by (Hatzivassiloglou and McKeown, 1997). They construct graphs from coordination data from large corpora based on the intuition that adjectives with the same sentiment orientation are likely to be coordinated. For example, *fresh and delicious* is more likely than *rotten and delicious*. They then apply a graph clustering algorithm to find groups of adjectives with the same orientation. Finally, they assign the same label to all adjectives that belong to the same cluster. The authors note that some words cannot be assigned a unique label since their sentiment depends on con-

Proceedings of the ACL 2010 Student Research Workshop, pages 25–30,
Uppsala, Sweden, 13 July 2010.

text.

Turney (2002) suggests a corpus-based extraction method based on his pointwise mutual information (PMI) synonymy measure He assumes that the sentiment orientation of a phrase can be determined by comparing its pointwise mutual information with a positive (*excellent*) and a negative phrase (*poor*). An introduction to SO-PMI is given in Section 5.1

3 Bilingual Lexicon Induction

Typical approaches to the induction of bilingual lexicons involve gathering new information from a small set of known identities between the languages which is called a *seed lexicon* and incorporating intralingual sources of information (e.g. cooccurrence counts). Two examples of such methods are a graph-based approach by Dorow et al. (2009) and a vector-space based approach by Rapp (1999). In this paper, we will employ the graph-based method.

SimRank was first introduced by Jeh and Widom (2002). It is an iterative algorithm that measures the similarity between all vertices in a graph. In SimRank, two nodes are similar if their neighbors are similar. This defines a recursive process that ends when the two nodes compared are identical. As proposed by Dorow et al. (2009), we will apply it to a graph $\mathcal{G}$ in which vertices represent words and edges represent relations between words. SimRank will then yield similarity values between vertices that indicate the degree of relatedness between them with regard to the property encoded through the edges. For two nodes i and j in $\mathcal{G}$, similarity according to SimRank is defined as

$$\text{sim}(i, j) = \frac{c}{|N(i)||N(j)|} \sum_{k \in N(i), l \in N(j)} \text{sim}(k, l),$$

where $N(x)$ is the neighborhood of x and c is a weight factor that determines the influence of neighbors that are farther away. The initial condition for the recursion is $\text{sim}(i, i) = 1$.

Dorow et al. (2009) further propose the application of the SimRank algorithm for the calculation of similarities between a source graph $\mathcal{S}$ and a target graph $\mathcal{T}$. Initially, some relations between the two graphs need to be known. When operating on word graphs, these can be taken from a bilingual lexicon. This provides us with a framework for the induction of a bilingual lexicon which can be constructed based on the obtained similarity values between the vertices of the two graphs.

One problem of SimRank observed in experiments by Laws et al. (2010) was that while words with high similarity were semantically related, they often were not exact translations of each other but instead often fell into the categories of hyponymy, hypernomy, holonymy, or meronymy. However, this makes the similarity values applicable for the translation of sentiment since it is a property that does not depend on exact synonymy.

4 Sentiment Transfer

Although unsupervised methods for the design of sentiment analysis systems exist, any approach can benefit from using resources that have been established in other languages. The main problem that we aim to deal with in this paper is the transfer of such information between languages. The SimRank lexicon induction method is suitable for this purpose since it can produce useful similarity values even with a small seed lexicon.

First, we build a graph for each language. The vertices of these graphs will represent adjectives while the edges are coordination relations between these adjectives. An example for such a graph is given in Figure 1.

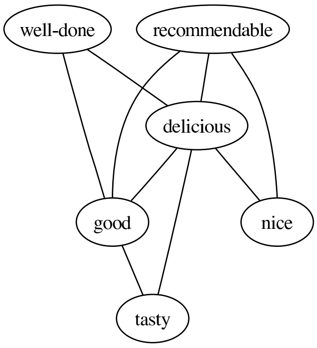

Figure 1: Sample graph showing English coordination relations.

The use of coordination information has been shown to be beneficial for example in early work by Hatzivassiloglou and McKeown (1997).

Seed links between those graphs will be taken from a universal dictionary. Figure 2 shows an example graph. Here, intralingual coordination relations are represented as black lines, seed relations as solid grey lines, and relations that are induced through SimRank as dashed grey lines.

After computing similarities in this graph, we

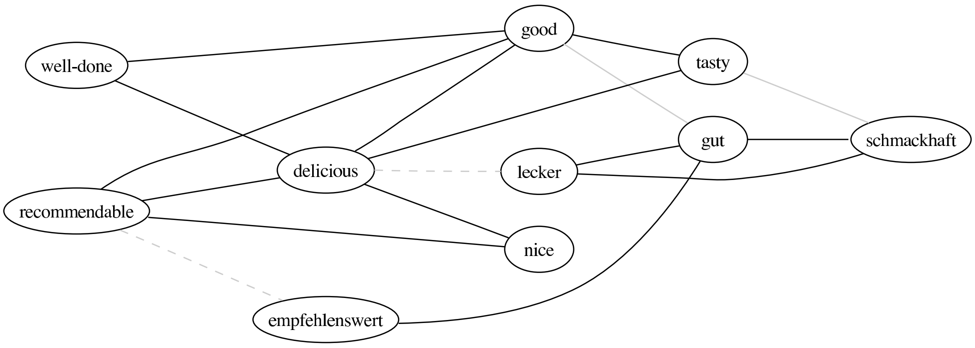

Figure 2: Sample graph showing English and German coordination relations. Solid black lines represent coordinations, solid grey lines represent seed relations, and dashed grey lines show induced relations.

need to obtain sentiment values. We will define the sentiment score (sent) as

$$\text{sent}(n_t) = \sum_{n_s \in \mathcal{S}} \text{sim}_{\text{norm}}(n_s, n_t)\ \text{sent}(n_s),$$

where n_t is a node in the target graph $\mathcal{T}$, and $\mathcal{S}$ the source graph. This way, the sentiment score of each node is an average over all nodes in $\mathcal{S}$ weighted by their normalized similarity, sim_{norm}.

We define the normalized similarity as

$$\text{sim}_{\text{norm}}(n_s, n_t) = \frac{\text{sim}(n_s, n_t)}{\sum_{n_s \in \mathcal{S}} \text{sim}(n_s, n_t)}.$$

Normalization guarantees that all sentiment scores lie within a specified range. Scores are not a direct indicator for orientation since the similarities still include a lot of noise. Therefore, we interpret the scores by assigning each word to a category by finding score thresholds between the categories.

5 Experiments

5.1 Baseline Method (SO-PMI)

We will compare our method to the well-established SO-PMI algorithm by Turney (2002) to show an improvement over an unsupervised method. The algorithm works with cooccurrence counts on large corpora. To determine the semantic orientation of a word w, the hits near positive ($Pwords$) and negative ($Nwords$) seed words is used. The SO-PMI equation is given as

$$\text{SO-PMI}(word) = \\ log_2\Big(\frac{\prod_{pword \in Pwords} hits(word\,\text{NEAR}\,pword)}{\prod_{nword \in Nwords} hits(word\,\text{NEAR}\,nword)} \\ \times \frac{\prod_{nword \in Nwords} hits(nword)}{\prod_{pword \in Pwords} hits(pword)}\Big)$$

5.2 Data Acquisition

We used the English and German Wikipedia branches as our corpora. We extracted coordinations from the corpus using a simple CQP pattern search (Christ et al., 1999). For our experiments, we looked only at coordinations with *and*. For the English corpus, we used the pattern `[pos = "JJ"] ([pos = ","] [pos = "JJ"])*([pos = ","]? "and" [pos = "JJ"])+`, and for the German corpus, the pattern `[pos = "ADJ.*"] ([pos = ","] [pos = "ADJ.*"])* ("und" [pos = "ADJ"])+` was used. This yielded $477{,}291$ pairs of coordinated English adjectives and $44{,}245$ German pairs. We used the dict.cc dictionary[1] as a seed dictionary. It contained a total of $30{,}551$ adjectives.

After building a graph out of this data as described in Section 4, we apply the SimRank algorithm using 7 iterations.

Data for the SO-PMI method had to be collected from queries to search engines since the information available in the Wikipedia corpus was too sparse. Since Google does not provide a stable `NEAR` operator, we used coordinations instead. For each of the test words `w` and the SO-PMI seed words `s` we made two queries `+"w und s"` and `+"s und w"` to Google. The quotes and `+` were added to ensure that no spelling correction or synonym replacements took place. Since the original experiments were designed for an English corpus, a set of German seed words had to be constructed. We chose *gut*, *nett*, *richtig*, *schön*, *ordentlich*, *angenehm*, *aufrichtig*, *gewissenhaft*, and *hervorragend* as positive seeds, and *schlecht*, *teuer*, *falsch*, *böse*, *feindlich*, *verhasst*, *widerlich*, *fehlerhaft*, and

[1] `http://www.dict.cc/`

word	value
strongpos	1.0
weakpos	0.5
neutral	0.0
weakneg	-0.5
strongneg	-1.0

Table 1: Assigned values for positivity labels

mangelhaft as negative seeds.

We constructed a test set by randomly selecting 200 German adjectives that occurred in a coordination in Wikipedia. We then eliminated adjectives that we deemed uncommon or too difficult to understand or that were mislabeled as adjectives. This resulted in a 150 word test set. To determine the sentiment of these adjectives, we asked 9 human judges, all native German speakers, to annotate them given the classes *neutral*, *slightly negative*, *very negative*, *slightly positive*, and *very positive*, reflecting the categories from the training data. In the annotation process, another 7 adjectives had to be discarded because one or more annotators marked them as unknown.

Since human judges tend to interpret scales differently, we examine their agreement using Kendall's coefficient of concordance (W) including correction for ties (Legendre, 2005) which takes ranks into account. The agreement was calculated as $W = 0.674$ with a significant confidence ($p < .001$), which is usually interpreted as substantial agreement. Manual examination of the data showed that most disagreement between the annotators occurred with adjectives that are tied to political implications, for example *nuklear* (*nuclear*).

5.3 Sentiment Lexicon Induction

For our experiments, we used the polarity lexicon of Wilson et al. (2005). It includes annotations of positivity in the form of the categories *neutral*, weakly positive (*weakpos*), strongly positive (*strongpos*), weakly negative (*weakneg*), and strongly positive (*strongneg*). In order to conduct arithmetic operations on these annotations, mapped them to values from the interval $[-1, 1]$ by using the assignments given in Table 1.

5.4 Results

To compare the two methods to the human raters, we first reproduce the evaluation by Turney (2002) and examine the correlation coefficients. Both methods will be compared to an average over the human rater values. These values are calculated on values asserted based on Table 1. The correlation coefficients between the automatic systems and the human ratings, SO-PMI yields $r = 0.551$, and SimRank yields $r = 0.587$ which are not significantly different. This shows that SO and SR have about the same performance on this broad measure.

Since many adjectives do not express sentiment at all, the correct categorization of neutral adjectives is as important as the scalar rating. Thus, we divide the adjectives into three categories – positive, neutral, and negative. Due to disagreements between the human judges there exists no clear threshold between these categories. In order to try different thresholds, we assume that sentiment is symmetrically distributed with mean 0 on the human scores. For $x \in \{\frac{i}{20} | 0 \leq i \leq 19\}$, we then assign word w with human rating $score(w)$ to negative if $score(w) \leq -x$, to neutral if $-x < score(w) < x$ and to positive otherwise. This gives us a three-category gold standard for each x that is then the basis for computing evaluation measures. Each category contains a certain percentile of the list of adjectives. By mapping these percentiles to the rank-ordered scores for SO-PMI and SimRank, we can create three-category partitions for them. For example if for $x = 0.35$ 21% of the adjectives are negative, then the 21% of adjectives with the lowest SO-PMI scores are deemed to have been rated negative by SO-PMI.

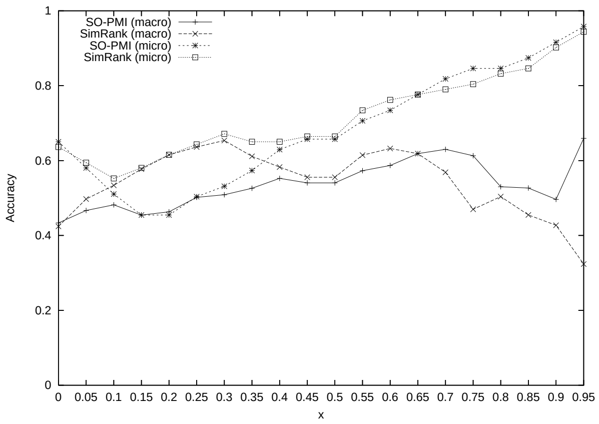

Figure 3: Macro- and micro-averaged Accuracy

First, we will look at the macro- and micro-averaged accuracies for both methods (cf. Figure 3). Overall, SimRank performs better for x

between 0.05 and 0.4 which is a plausible interval for the neutral threshold on the human ratings. The results diverge for very low and high values of x, however these values can be considered unrealistic since they implicate neutral areas that are too small or too large. When comparing the accuracies for each of the classes (cf. Figure 4), we observe that in the aforementioned interval, SimRank has higher accuracy values than SO-PMI for all of them.

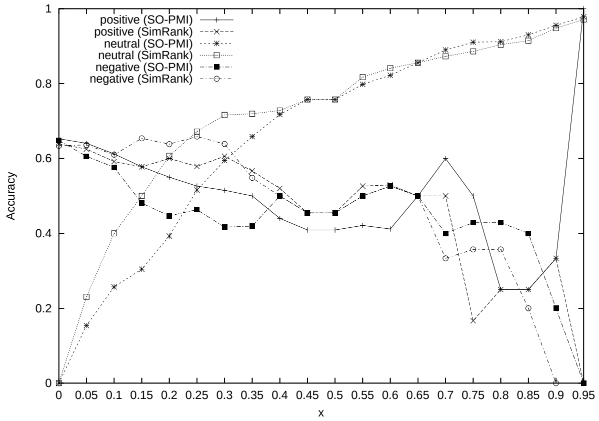

Figure 4: Accuracy for individual classes

Table 2 lists some interesting example words including their human ratings and SO-PMI and SimRank scores which illustrate advantages and possible shortcomings of the two methods. The medians of SO-PMI and SimRank scores are -15.58 and -0.05, respectively. The mean values are -9.57 for SO-PMI and 0.08 for SimRank, the standard deviations are 13.75 and 0.22. SimRank values range between -0.67 and 0.41, SO-PMI ranges between -46.21 and 46.59. We will assume that the medians mark the center of the set of neutral adjectives.

Ausdrucksvoll receives a positive score from SO-PMI which matches the human rating, however not from SimRank, which assigns a score close to 0 and would likely be considered neutral. This error can be explained by examining the similarity distribution for *ausdrucksvoll* which reveals that there are no nodes that are similar to this node, which was most likely caused by its low degree. *Auferstanden* (resurrected) is perceived as a positive adjective by the human judges, however it is misclassified by SimRank as negative due to its occurrence with words like *gestorben* (*deceased*) and *gekreuzigt* (*crucified*) which have negative associations. This suggests that coordinations are sometimes misleading and should not be used as the only data source. *Grafisch* (*graphics-related*) is an example for a neutral word misclassified by SO-PMI due to its occurrence in positive contexts on the web. Since SimRank is not restricted to relations between an adjective and a seed word, all adjective-adjective coordinations are used for the estimation of a sentiment score. *Kriminell* is also misclassified by SO-PMI for the same reason.

word (translation)	**SR**	**SO**	**judges**
ausdrucksvoll (expressive)	0.069	22.93	0.39
grafisch (graphic)	-0.050	-4.75	0.00
kriminell (criminal)	-0.389	-15.98	-0.94
auferstanden (resurrected)	-0.338	-10.97	0.34

Table 2: Example adjectives including translation, and their scores

6 Conclusion and Outlook

We presented a novel approach to the translation of sentiment information that outperforms SO-PMI, an established method. In particular, we could show that SimRank outperforms SO-PMI for values of the threshold x in an interval that most likely leads to the correct separation of positive, neutral, and negative adjectives. We intend to compare our system to other available work in the future. In addition to our findings, we created an initial gold standard set of sentiment-annotated German adjectives that will be publicly available.

The two methods are very different in nature; while SO-PMI is suitable for languages in which very large corpora exist, this might not be the case for knowledge-sparse languages. For some German words (e.g. *schwerstkrank* (*seriously ill*)), SO-PMI lacked sufficient results on the web whereas SimRank correctly assigned negative sentiment. SimRank can leverage knowledge from neighbor words to circumvent this problem. In turn, this information can turn out to be misleading (cf. *auferstanden*). An advantage of our method is that it uses existing resources from another language and can thus be applied without much knowledge about the target language. Our future work will include a further examination of the merits of its application for knowledge-sparse languages.

The underlying graph structure provides a foundation for many conceivable extensions. In this paper, we presented a fairly simple experiment restricted to adjectives only. However, the method is suitable to include arbitrary parts of speech as well as phrases, as used by Turney (2002). Another conceivable application would be the direct combination of the SimRank-based model with a statistical model.

Currently, our input sentiment list exists only of prior sentiment values, however work by Wilson et al. (2009) has advanced the notion of contextual polarity lists. The automatic translation of this information could be beneficial for sentiment analysis in other languages.

Another important problem in sentiment analysis is the treatment of ambiguity. The sentiment expressed by a word or phrase is context-dependent and is for example related to word sense (Akkaya et al., 2009). Based on regularities in graph structure and similarity, ambiguity resolution might become possible.

References

C. Akkaya, J. Wiebe, and R. Mihalcea. 2009. Subjectivity Word Sense Disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 190–199.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135, Honolulu, Hawaii, October. Association for Computational Linguistics.

O. Christ, B.M. Schulze, A. Hofmann, and E. Koenig. 1999. The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. *University of Stuttgart, March*, 8:1999.

Beate Dorow, Florian Laws, Lukas Michelbacher, Christian Scheible, and Jason Utt. 2009. A graph-theoretic algorithm for automatic extension of translation lexicons. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 91–95, Athens, Greece, March. Association for Computational Linguistics.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, July. Association for Computational Linguistics.

Glen Jeh and Jennifer Widom. 2002. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, New York, NY, USA. ACM.

F. Laws, L. Michelbacher, B. Dorow, U. Heid, and H. Schütze. 2010. Building a Cross-lingual Relatedness Thesaurus Using a Graph Similarity Measure. *Submitted on Nov 7, 2009, to the International Conference on Language Resources and Evaluation (LREC).*

P. Legendre. 2005. Species associations: the Kendall coefficient of concordance revisited. *Journal of Agricultural Biological and Environment Statistics*, 10(2):226–245.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic, June. Association for Computational Linguistics.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, College Park, Maryland, USA, June. Association for Computational Linguistics.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, Suntec, Singapore, August. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing Contextual Polarity: an Exploration of Features for Phrase-level Sentiment Analysis. *Computational Linguistics*, 35(3):399–433.