Sentiment Translation through Multi-Edge Graphs

Christian Scheible, Florian Laws, Lukas Michelbacher, and **Hinrich Schütze**
Institute for Natural Language Processing
University of Stuttgart
{scheibcn, lawsfn, michells}@ims.uni-stuttgart.de

Abstract

Sentiment analysis systems can benefit from the *translation* of sentiment information. We present a novel, graph-based approach using SimRank, a well-established graph-theoretic algorithm, to transfer sentiment information from a source language to a target language. We evaluate this method in comparison with semantic orientation using pointwise mutual information (SO-PMI), an established unsupervised method for learning the sentiment of phrases.

1 Introduction

Sentiment analysis is an important topic in computational linguistics that is of theoretical interest but is also useful in many practical 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 two languages that does not rely on resources with limited availability like parallel corpora. It is built around SimRank, a graph similarity algorithm that has successfully been applied to the acquisition of bilingual lexicons (Laws et al., 2010) and semantic similarity (Michelbacher et al., 2010). It uses linguistic relations extracted from two monolingual corpora to determine the similarity of words in different languages. One of the main benefits of our method is its ability to handle sparse data about the relations between the languages well (i.e., a small seed lexicon). Further, we experiment with combining multiple types of linguistic relations for graph-based translation. Our experiments are carried out using English as a source language and German as a target language. We evaluate our method using a hand-annotated set of German adjectives which we intend to publish.

In the following section, related work is discussed. Section 3.1 gives an introduction to SimRank and its application to lexicon induction, while section 3.2 reviews SO-PMI (Turney, 2002), an unsupervised baseline method for the generation of sentiment lexicons. In section 4, we define our sentiment transfer method which we apply in experiments in section 5.

2 Related Work

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.

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

Coling 2010: Poster Volume, pages 1104–1112,
Beijing, August 2010

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.

Corpus work and bilingual dictionaries are promising resources for translating sentiment. In contrast to previous approaches, the work presented in this paper uses corpora that are not annotated with sentiment.

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 this method is given in Section 3.2.

3 Background

3.1 Lexicon Induction via SimRank

We use the extension of the SimRank (Jeh and Widom, 2002) node similarity algorithm proposed by Dorow et al. (2009). Given two graphs $\mathcal{A}$ and $\mathcal{B}$, the similarity between two nodes a in $\mathcal{A}$ and b in $\mathcal{B}$ is computed in each iteration as:

$$S(a,b) = \frac{c}{|N_A(a)||N_B(b)|} \sum_{k \in N_A(a), l \in N_B(b)} S(k,l).$$

$N_X(x)$ is the neighborhood of node x in graph X. To compute similarities between two graphs, some initial links between these graphs have to be given, called seed links. These form the recursion basis which sets $S(a,b) = 1$ if there is a seed link between a and b. At the beginning of each iteration, all known equivalences between nodes are reset to 1.

Multi-Edge Extraction (MEE). MEE is an extension of SimRank that, in each iteration, computes the average node-node similarity of several different SimRank matrices. In our case, we use two different SimRank matrices, one for coordinations and one for adjective modification. See (Dorow et al., 2009) for details. We also used the node degree normalization function $h(n) = \sqrt{n} \times \sqrt{\max_k(|N(k)|)}$ (where n is the node degree, and $N(k)$ the degree of node k) to decrease the harmful effect of high-degree nodes on final similarity values. See (Laws et al., 2010) for details.

3.2 SO-PMI

Semantic orientation using pointwise mutual information (SO-PMI) (Turney, 2002) is an algorithm for the unsupervised learning of semantic orientation of words or phrases. A word has positive (resp. negative) orientation if it is associated with positive (resp. negative) terms more frequently than with negative (resp. positive) terms. Association of terms is measured using their pointwise mutual information (PMI) which is defined for two words w_1 and w_2 as follows:

$$\text{PMI}(w_1, w_2) = log\left(\frac{p(w_1, w_2)}{p(w_1)p(w_2)}\right)$$

Using PMI, Turney defines SO-PMI for a word w as

$$\text{SO-PMI}(w) =$$

$$log\frac{\prod_{p \in P} hits(word \text{ NEAR } p) \times \prod_{n \in N} hits(n)}{\prod_{n \in N} hits(word \text{ NEAR } n) \times \prod_{p \in P} hits(p)}$$

$hits$ is a function that returns the number of hits in a search engine given the query. P is a set of known positive words, N a set of known negative words, and `NEAR` an operator of a search engine that returns documents in which the operands occur within a close range of each other.

4 Sentiment Translation

Unsupervised methods like SO-PMI are suitable to acquire basic sentiment information in a language. However, since hand-annotated resources for sentiment analysis exist in other languages, it seems plausible to use automatic translation of sentiment information to leverage these resources. In order to translate sentiment, we will use multiple sources of information that we represent in a MEE graph as given in Section 3.1.

In our first experiments (Scheible, 2010), coordinated adjectives were used as the sole training source. Two adjectives are coordinated if they are linked with a conjunction like *and* or *but*. The intuition behind using coordinations – based on work by Hatzivassiloglou and McKeown (1997) and Widdows and Dorow (2002) – was that words which are coordinated share properties. In particular, coordinated adjectives usually express similar sentiments even though there are exceptions (e.g., "The movie was both good and bad").

In this paper, we focus on using multiple edge types for sentiment translation. In particular, the graph we will use contains two types of relations, *coordinations* and *adjective-noun modification*. In the sentence "The movie was enjoyable and fun", *enjoyable* and *fun* are coordinated. In *This is an enjoyable movie*, the adjective *enjoyable* modifies the noun *movie*.

We selected these two relation types for two reasons. First, the two types provide clues for sentiment analysis. Coordination information is an established source for sentiment similarity (e.g. Hatzivassiloglou and McKeown (1997)) while adjective-noun relations provide a different type of information on sentiment. For example, nouns with positive associations (*vacation*) tend to occur with positive adjectives and nouns with negative associations (*pain*) tend to occur with negative adjectives. Second, we have successfully used these two types for a similar acquisition task, the acquisition of word-to-word translation pairs (Laws et al., 2010).

In the resulting graph, adjectives and nouns are represented as nodes, each containing a word and its part of speech, and relations are represented as links which are distinguished by their edge types. Two graphs, one in the source language and one in the target language, are needed to translate words between those languages. Figure 1 shows an example for such a setup. Black links in this graph are coordinations, grey links are seed relations.

In order to calculate sentiment for all nodes in the target language, we apply the SimRank algorithm to the graphs which gives us similarities between all nodes in the source graph and all nodes in the target graph. Using the similarity $S(n_s, n_t)$ between a node n_s in the source language graph $\mathcal{S}$ and a node n_t in the target language graph $\mathcal{T}$, the sentiment score ($\text{sent}(n_t)$) is the similarity-weighted average of all sentiment scores in the target language:

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

We assume that sentiment scores in the source language are expressed on a numeric scale. The normalized similarity sim_{norm} is defined as

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

The normalization assures that all resulting sentiment values are within $[-1, 1]$, with -1 being the most negative sentiment and 1 the most positive.

5 Experiments

5.1 Data Acquisition

For our experiments, we needed coordination data to build weighted graphs and a bilingual lexicon to define seed relations between those graphs. Coordinations were extracted from the English and German versions of Wikipedia[1] by applying pattern-based search using the Corpus Query Processor (CQP) (Christ et al., 1999). We annotated both corpora with parts of speech using the Tree Tagger (Schmid, 1994). A total of 477,291 English coordinations and 112,738 German coordinations were collected. A sample of this data is given in Figure 2. We restrict these experiments to the use of *and*/*und* since other coordinations

[1] `http://www.wikipedia.org/` (01/19/2009)

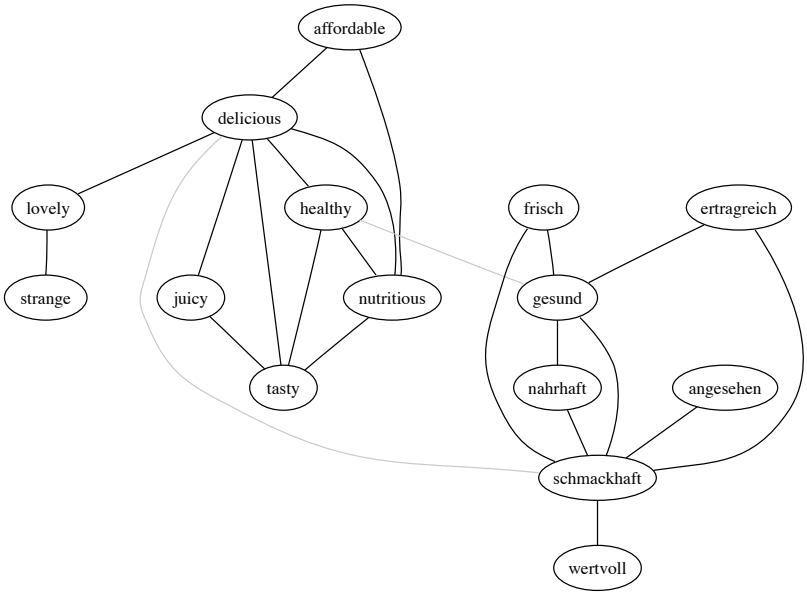

Figure 1: A German and an English graph with coordinated adjectives including seed links

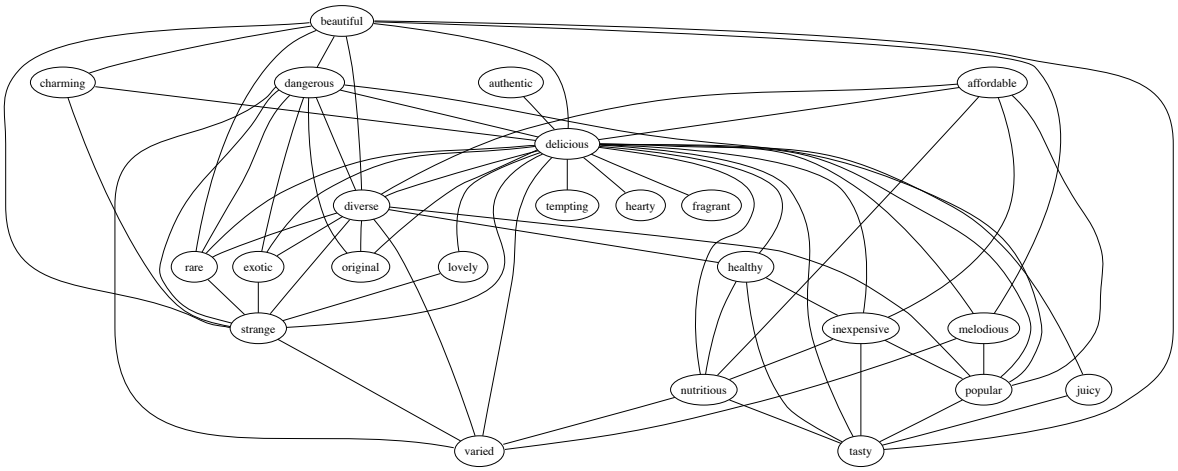

Figure 2: English sample coordinations (adjectives)

behave differently and might even express dissimilarity (e.g. *Was the weather good or bad?*).

The seed lexicon was constructed from the dict.cc dictionary[2]. While the complete dictionary contains 30,551 adjective pairs, we reduced the number of pairs used in the experiments to 1,576.

To produce a smaller seed lexicon which still makes sense from a semantic point of view, we used the General Service List (GSL) (West, 1953) which contains about 2000 words the author considered central to the English language. More specifically, a revised list was used[3].

SO-PMI needs a larger amount of training data. Since Wikipedia does not satisfy this need, we collected additional coordination data from the web using search result counts from Google. In Turney's original paper, he uses the `NEAR` operator, which returns documents that contain two search terms that are within a certain distance of each other, to collect collocations. Unfortunately, Google does not support this operator, so instead, we searched for coordinations using the queries

`+ "w and s"` and
`+ "w und s"`

for English and German, respectively. We added the quotes and the + operator to make sure that both spelling correction and synonym replacements were disabled.

The original experiments were made for English, so we had to construct our own set of seed words. For German, we chose *gut (good)*, *nett (nice)*, *richtig (right)*, *schön (beautiful)*, *ordentlich (neat)*, *angenehm (pleasant)*, *aufrichtig (honest)*, *gewissenhaft (faithful)*, and *hervorragend (excellent)* as positive seed words, and *schlecht (bad)*, *teuer (expensive)*, *falsch (wrong)*, *böse (evil)*, *feindlich (hostile)*, *verhasst (invidious)*, *widerlich (disgusting)*, *fehlerhaft (faulty)*, and *mangelhaft (flawed)* as negative ones.

5.2 Sentiment Lexicon

For our experiments, we used two different polarity lexicons. The lexicon of Wilson et al. (2005) contains sentiment annotations for 8,221 words which are tagged as *positive*, *neutral*, or *negative*. A few words are tagged as *weakneg*, implying weak negativity. These categorial annotations are mapped to the range [-1,1] using the assignment scheme given in Table 1.

annotation	value
positive	1.0
weakpos	0.5
neutral	0.0
weakneg	−0.5
negative	−1.0

Table 1: Assigned values for Wilson et al. set

5.3 Human Ratings

In order to manually annotate a test set, we chose 200 German adjectives that occurred in the Wikipedia corpus and that were part of a coordination. From these words, we removed those which we deemed uncommon, too complicated, or which were mislabeled as adjectives by the tagger. The test set contained 150 adjectives of which seven were excluded after annotators discarded them.

We asked 9 native speakers of German to annotate the adjectives. Possible annotations were *very positive*, *slightly positive*, *neutral*, *slightly negative*, or *very negative*. These categories are the same as the ones used in the training data.

In order to capture the general sentiment, i.e., sentiment that is not related to a specific context, the judges were asked to stay objective and not let their personal opinions influence the annotation. However, some words with strong political implications were annotated by some judges as non-neutral which led to disagreement beyond the usual level. *Nuklear* (*nuclear*) is an example for such a word. We measured the agreement of the judges with Kendall's coefficient of concordance (W) with tie correction (Legendre, 2005), yielding $W = 0.674$ with a high level of significance ($p < .001$); thus, inter-annotator agreement was high (Landis and Koch, 1977).

5.4 Experimental Setup

Given the relations extracted from Wikipedia, we built a German and an English graph by setting

[2] `http://www.dict.cc`
[3] `http://jbauman.com/aboutgsl.html`

Method	r
MEE	0.63
MEE-GSL	0.47
SR	0.63
SR-GSL	0.48
SO-PMI	0.58

Table 2: Correlation with human ratings

the weight of each link to the log-likelihood ratio of the two words it connects according to the corpus frequencies. There are two properties of the graph transfer algorithm that we intend to investigate. First, we are interested in the merits of applying multi edge extraction (MEE) for sentiment transfer. Second, we are interested in how the transfer quality changes when the seed lexicon is reduced in size. This way, a sparse data situation is simulated where large dictionaries are unavailable. Having these two properties in mind, four possible setups are evaluated: (i) using the full seed lexicon with all $30{,}551$ entries, but using only coordination data (*SR*), (ii) reducing the seed lexicon to $1{,}576$ entries from the General Service List (*SR-GSL*), (iii) applying MEE by adding adjective modification data (*MEE*), and (iv) using MEE with a reduced seed lexicon (*MEE-GSL*). SimRank was run for 6 iterations in all experiments. All experiments use the weight function h as described above. We show that this function improves similarities and thus lexicon induction in Laws et al. (2010).

Correlation. First, we will examine the correlation between the automatic methods (SO-PMI and the aforementioned SimRank variations) and the gold standard as done by Turney in his evaluation. For this purpose, the human ratings are mapped to float values following Table 1 and the average rating over all judges for each word is used. The correlation coefficients r are given in Table 2. Judging from these results, the ordering of SR and MEE matches the human ratings better than SO-PMI, however it decreases when using any of the GSL variations instead which can be attributed to using less data.

Classification. The correct identification of the classes *positive*, *neutral*, and *negative* is more important than the correct assignment of values on a scale since the rank ordering is debatable – this becomes apparent when measuring the agreement of human annotators. Since the assignments made by the human judges are not unanimous in most cases, the averages are distributed across the interval [-1,1]; this means that the borders between the three distinct categories are not clear. Since there is no standard evaluation for this particular problem, we need to devise a way to make the range of the neutral category dynamic. In order to find possible borders, we first assume that sentiment is distributed symmetrically around 0. We then define a threshold x which assumes the values $x \in \{\frac{i}{20} | 0 \leq i \leq 20\}$, covering the interval [0,0.5]. Since 0.5 is *slightly positive*, we do not believe that values above it are plausible. Then, each word w is positive if its human rating $\text{score}_h(w) \geq x$, negative if $\text{score}_h(w) \leq -x$, and neutral if $-x < \text{score}_h(w) < x$. The result of this process is a gold standard for the three categories for each of the values for x. The percentiles of the sizes of those categories are mapped to the values produced by the automatic methods. For example, if $x = 0.35$ means that the top 21% of all adjectives are in the positive class, the top 21% of all adjectives as assigned by SO-PMI and the SimRank varieties are positive as well.

The size of the neutral category increases the larger x becomes. Thus, high values for x are unlikely to produce a correct partitioning of the data. Since *slightly positive* was defined as 0.5, we expect the highest plausible value for x to be below that. The size of the neutral category for each value of x is given in Table 3. (Recall that the total size of the set is 143.)

We can then compute the assignment accuracy on the positive, neutral, and negative classes, as well macro- and micro-averages over these classes.

5.5 Results and Discussion

Figures 3 and 4 show the macro- and micro-averaged accuracies over the positive, negative, and neutral class for each automatic method, respectively. Overall, the SimRank variations perform better for x in the interval $[0, 0.3]$. In particular, MEE has a slightly higher accuracy than SR,

x	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
# neutral	0	13	35	46	56	64	74	82	92	99	99

Table 3: Size of neutral category given x

word (translation)	**humans**	**SO**	**MEE**	**MEE-GSL**	**SR**	**SR-GSL**
chemisch (chemical)	0.00	-20.20	0.185	0.185	0.186	0.184
auferstanden (resurrected)	0.39	-10.96	-0.075	-0.577	-0.057	-0.493
intelligent (intelligent)	0.94	46.59	0.915	0.939	0.834	0.876
versiert (skilled)	0.67 -5.26 0.953 0.447 0.902 0.404					
mean	-0.04	-9.58	0.003	0.146	0.010	0.142
median	0.00	-15.60	0.110	0.157	0.114	0.157

Table 4: Example adjectives including translation, and their scores

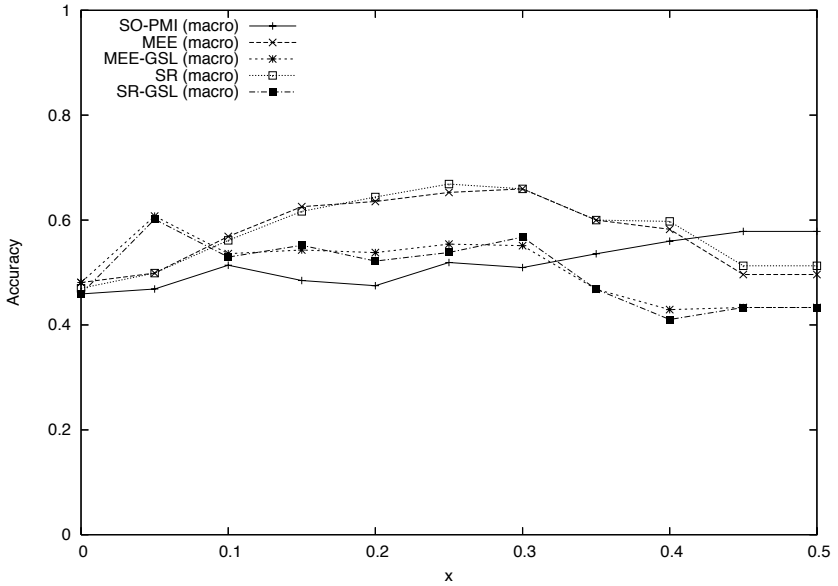

Figure 3: Macro-averaged Accuracy

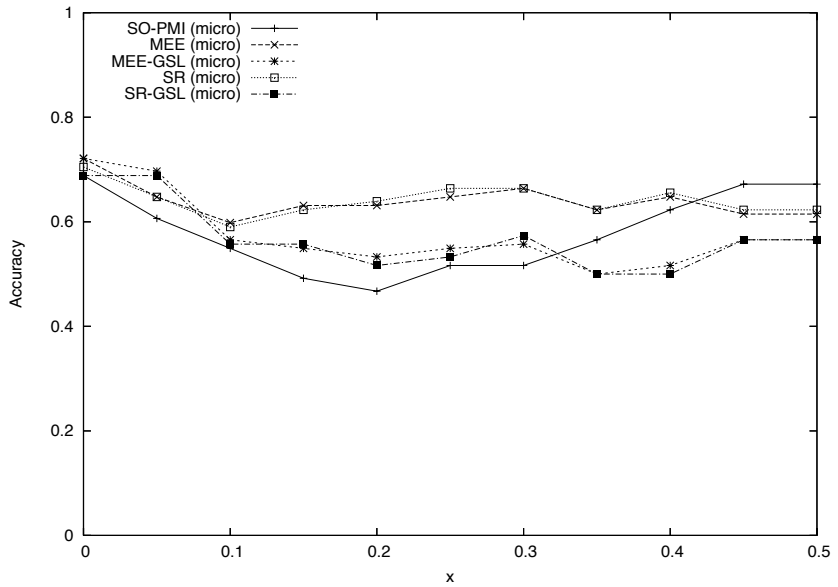

Figure 4: Micro-averaged Accuracy

however, not significantly.

Table 4 shows selected example words with their scores. These values can be understood better together with the means and medians of the respective methods which are given in the table as well. These values give us an idea of where we might expect the neutral point of a particular distribution of polarities.

Chemisch (chemical) is misclassified by SO-PMI since it occurs in negative contexts on the web. SimRank in turn was able to recognize that most words similar to *chemisch* are neutral, the most similar one being its literal translation, *chemical*. *Auferstanden (resurrected)* is an example for misclassification by SimRank which happens because the word is usually coordinated with words that have negative sentiment, e.g. *gestorben (deceased)* and *gekreuzigt (crucified)*. This problem could not be fixed by including adjective-noun modification data since the coordinations produced high log-likelihood values which lead to *dead* being the most similar word to *auferstanden*. *Intelligent* receives a score close to neutral with the original (coordination-only) training method, which could be corrected by applying MEE simply because the ordering of similar words changes through the new weighting method. Nouns modified by *intelligent* include *Leben (life)* and *Wesen (being)* whose translations are modified by positive adjectives. Many words, such as *versiert (skilled)* are classified more accurately due to the new weighting method when compared to our previous experiments (Scheible, 2010) where it received a SimRank polarity of only 0.224.

The inclusion of adjective modifications does not improve the classification results as often as we had hoped. For some cases (cf. *intelligent* mentioned above), the scores do improve, but the overall impact is limited.

6 Conclusion and Outlook

We were able to show that sentiment translation with SimRank is able to classify adjectives more accurately than SO-PMI, an unsupervised baseline method. We demonstrated that SO-PMI is outperformed by SimRank when choosing a reasonable region of neutral adjectives. In addition, we showed that the improvements of SimRank lead to better accuracy in sentiment translation in some cases. In future work, we will apply a sentiment lexicon generated with SimRank in a sentiment classification task for reviews.

The algorithms we compared are different in their purpose of application. While SO-PMI is applicable when large corpora are available for a language, it fails when used in a sparse-data situation, as noted by Turney (2002). We showed that despite reducing the seed lexicon for SimRank to a small fraction of its original size, it still performs better than SO-PMI.

Currently, our experiments are limited by the choice of using adjectives for our test set. While the examination of adjectives is highly important for sentiment analysis (as shown by Pang et al. (2002) who were able to achieve high accuracy even when using only adjectives), the application of our algorithms to a broader set of linguistic units is an important goal for future work.

Acknowledgments. We are grateful to Deutsche Forschungsgemeinschaft for funding this research as part of the WordGraph project.

References

Banea, Carmen, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Empirical Methods in Natural Language Processing*, pages 127–135.

Christ, O., 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.

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

Hatzivassiloglou, Vasileios 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.

Jeh, Glen and Jennifer Widom. 2002. Simrank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD Interna-*

tional Conference on Knowledge Discovery and Data Mining, pages 538–543.

Landis, J.R. and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Laws, Florian, Lukas Michelbacher, Beate Dorow, Christian Scheible, Ulrich Heid, and Hinrich Schütze. 2010. A linguistically grounded graph model for bilingual lexicon extraction. In *Proceedings of the 23nd International Conference on Computational Linguistics*.

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

Michelbacher, Lukas, Florian Laws, Beate Dorow, Ulrich Heid, and Hinrich Schütze. 2010. Building a cross-lingual relatedness thesaurus using a graph similarity measure. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*.

Mihalcea, Rada, 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.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Scheible, Christian. 2010. Sentiment translation through lexicon induction. In *Proceedings of the ACL 2010 Student Research Workshop*, Uppsala, Sweden. Association for Computational Linguistics.

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

Turney, Peter. 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.

Wan, Xiaojun. 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.

West, Michael. 1953. A general service list of english words.

Widdows, Dominic and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *COLING*.

Wilson, Theresa, 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, October.