Experiments on Hybrid Corpus-Based Sentiment Lexicon Acquisition

Goran Glavaš, Jan Šnajder and Bojana Dalbelo Bašić
Faculty of Electrical Engineering and Computing
University of Zagreb
Zagreb, Croatia
{goran.glavas, jan.snajder, bojana.dalbelo}@fer.hr

Abstract

Numerous sentiment analysis applications make usage of a sentiment lexicon. In this paper we present experiments on hybrid sentiment lexicon acquisition. The approach is corpus-based and thus suitable for languages lacking general dictionary-based resources. The approach is a hybrid two-step process that combines semi-supervised graph-based algorithms and supervised models. We evaluate the performance on three tasks that capture different aspects of a sentiment lexicon: polarity ranking task, polarity regression task, and sentiment classification task. Extensive evaluation shows that the results are comparable to those of a well-known sentiment lexicon SentiWordNet on the polarity ranking task. On the sentiment classification task, the results are also comparable to SentiWordNet when restricted to *monosentimous* (all senses carry the same sentiment) words. This is satisfactory, given the absence of explicit semantic relations between words in the corpus.

1 Introduction

Knowing someone's attitude towards events, entities, and phenomena can be very important in various areas of human activity. Sentiment analysis is an area of computational linguistics that aims to recognize the subjectivity and attitude expressed in natural language texts. Applications of sentiment analysis are numerous, including sentiment-based document classification (Riloff et al., 2006), opinion-oriented information extraction (Hu and Liu, 2004), and question answering (Somasundaran et al., 2007).

Sentiment analysis combines subjectivity analysis and polarity analysis. Subjectivity analysis answers whether the text unit is subjective or neutral, while polarity analysis determines whether a subjective text unit is positive or negative. The majority of research approaches (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Wilson et al., 2009) see subjectivity and polarity as categorical terms (i.e., classification problems). Intuitively, not all words express the sentiment with the same intensity. Accordingly, there has been some research effort in assessing subjectivity and polarity as graded values (Baccianella et al., 2010; Andreevskaia and Bergler, 2006). Most of the work on sentence or document level sentiment makes usage of sentiment annotated lexicon providing subjectivity and polarity information for individual words (Wilson et al., 2009; Taboada et al., 2011).

In this paper we present a hybrid approach for automated acquisition of sentiment lexicon. The method is language independent and corpus-based and therefore suitable for languages lacking general lexical resources such as WordNet (Fellbaum, 2010). The two-step hybrid process combines semi-supervised graph-based algorithms and supervised learning models.

We consider three different tasks, each capturing different aspect of a sentiment lexicon:

1. Polarity ranking task – determine the relative rankings of words, i.e., order lexicon items descendingly by positivity and negativity;
2. Polarity regression task – assign each word absolute scores (between 0 and 1) for positivity and negativity;
3. Sentiment classification task – classify each

word into one of the three sentiment classes (*positive*, *negative*, or *neutral*).

Accordingly, we evaluate our method using three different measures – one to evaluate the quality of the ordering by positivity and negativity, other to evaluate the absolute sentiment scores assigned to each corpus word, and another to evaluate the classification performance.

The rest of the paper is structured as follows. In Section 2 we present the related work on sentiment lexicon acquisition. Section 3 discusses the semi-supervised step of the hybrid approach. In Section 4 we explain the supervised step in more detail. In Section 5 the experimental setup, the evaluation procedure, and the results of the approach are discussed. Section 6 concludes the paper and outlines future work.

2 Related Work

Several approaches have been proposed for determining the prior polarity of words. Most of the approaches can be classified as either dictionary-based (Kamps et al., 2004; Esuli and Sebastiani, 2007; Baccianella et al., 2010) or corpus-based (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003). Regardless of the resource used, most of the approaches focus on bootstrapping, starting from a small seed set of manually labeled words (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Esuli and Sebastiani, 2007). In this paper we also follow this idea of the semi-supervised bootstrapping as the first step of the sentiment lexicon acquisition.

Dictionary-based approaches grow the seed sets according to the explicit paradigmatic semantic relations (synonymy, antonymy, hyponymy, etc.) between words in the dictionary. Kamps et al. (2004) build a graph of adjectives based on synonymy relations gathered from WordNet. They determine the polarity of the adjective based on its shortest path distances from positive and negative seed adjectives *good* and *bad*. Esuli and Sebastiani (2007) first build a graph based on a gloss relation (i.e., *definiens* – *definiendum* relation) from WordNet. Afterwards they perform a variation of the PageRank algorithm (Page et al., 1999) in two runs. In the first run positive PageRank value is assigned to the vertices of the synsets from the positive seed set and zero value to all other vertices. In the second run the same is done for the synsets from the negative seed set. Word's polarity is then decided based on the difference between its PageRank values of the two runs. We also believe that graph is the appropriate structure for the propagation of sentiment properties of words. Unfortunately, for many languages a precompiled lexical resource like WordNet does not exist. In such a case, semantic relations between words may be extracted from corpus.

In their pioneering work, Hatzivassiloglou and McKeown (1997) attempt to determine the polarity of adjectives based on their co-occurrences in conjunctions. They start with a small manually labeled seed set and build on the observation that adjectives of the same polarity are often conjoined with the conjunction *and*, while adjectives of the opposite polarity are conjoined with the conjunction *but*. Turney and Littman (2003) use pointwise mutual information (PMI) (Church and Hanks, 1990) and latent semantic analysis (LSA) (Dumais, 2004) to determine the similarity of the word of unknown polarity with the words in both positive and negative seed sets. The aforementioned work presumes that there is a correlation between lexical semantics and sentiment. We base our work on the same assumption, but instead of directly comparing the words with the seed sets, we use distributional semantics to build a word similarity graph. In contrast to the approaches above, this allows us to potentially account for similarities between all pairs of words from corpus. To the best of our knowledge, such an approach that combines corpus-based lexical semantics with graph-based propagation has not yet been applied to the task of building sentiment lexicon. However, similar approaches have been proven rather efficient on other tasks such as document level sentiment classification (Goldberg and Zhu, 2006) and word sense disambiguation (Agirre et al., 2006).

3 Semi-supervised Graph-based Methods

The structure of a graph in general provides a good framework for propagation of object properties, which, in our case, are the sentiment values of the words. In a word similarity graph, weights of edges represent the degree of semantic similarity between words.

In the work presented in this paper we build graphs from corpus, using different notions of

word similarity. Each vertex in the graph represents a word from corpus. Weights of the edges are calculated in several different ways, using measures of word co-occurrence (co-occurrence frequency and pointwise mutual information) and distributional semantic models (latent semantic analysis and random indexing). We manually compiled positive and negative seed sets, each consisting of 15 words:

> positiveSeeds = {*good, best, excellent, happy, well, new, great, nice, smart, beautiful, smile, win, hope, love, friend*}
>
> negativeSeeds = {*bad, worst, violence, die, poor, terrible, death, war, enemy, accident, murder, lose, wrong, attack, loss*}

In addition to these, we compiled the third seed set consisting of neutral words to serve as sentiment sinks for the employed label propagation algorithm:

> neutralSeeds = {*time, place, company, work, city, house, man, world, woman, country, building, number, system, object, room*}

Once we have built the graph, we label the vertices belonging to the words from the polar seed set with the sentiment score of 1. All other vertices are initially unlabeled (i.e., assigned a sentiment score of 0). We then use the structure of the graph and one of the two random-walk algorithms to propagate the labels from the labeled seed set vertices to the unlabeled ones. The random walk algorithm is executed twice: once with the words from the positive seed set being initially labeled and once with the words from the negative seed set being initially labeled. Once the random walk algorithm converges, all unlabeled vertices will be assigned a sentiment label. However, the final sentiment values obtained after the convergence of the random-walk algorithm are directly dependent on the size of the graph (which, in turn, depends on the size of the corpus), the size of the seed set, and the choice of the seed set words. Thus, they should be interpreted as relative rather than absolute sentiment scores. Nevertheless, the scores obtained from the graph can be used to rank the words by their positivity and negativity.

3.1 Similarity Based on Corpus Co-occurrence

If the two words co-occur in the corpus within a window of a given size, an edge in the graph between their corresponding vertices is added. The weight of the edge should represent the measure of the degree to which the two words co-occur.

There are many word collocation measures that may be used to calculate the weights of edges (Evert, 2008). In this work, we use raw co-occurrence frequency and pointwise mutual information (PMI) (Church and Hanks, 1990). In the former case the edge between two words is assigned a weight indicating a total number of co-occurrences of the corresponding words in the corpus within the window of a given size. In the latter case, we use PMI to account for the individual frequencies of each of the two words along with their co-occurrence frequency. The most frequent corpus words tend to frequently co-occur with most other words in the corpus, including words from both positive and negative seed sets. PMI compensates for this shortcoming of the raw co-occurrence frequency measure.

3.2 Similarity Based on Latent Semantic Analysis

Latent semantic analysis is a well-known technique for identifying semantically related concepts and dimensionality reduction in large vector spaces (Dumais, 2004). The first step is to create a sparse word-document matrix. Matrix elements are frequencies of words occurring in documents, usually transformed using some weighting scheme (e.g., *tf-idf*). The word-document matrix is then decomposed using singular value decomposition (SVD), a well-known linear algebra procedure. Finally, the dimensionality reduction is performed by approximating the original matrix using only the top k largest singular values.

We build two different word-document matrices using different weighting schemes. The elements of the first matrix were calculated using the *tf-idf* weighting scheme, while for the second matrix the *log-entropy* weighting scheme was used. In the *log-entropy* scheme, each matrix element, $m_{w,d}$, is calculated using logarithmic value of word-document frequency and the global word entropy (entropy of word frequency across the documents), as follows:

$$m_{w,d} = \log\left(tf_{w,d} + 1\right) \cdot g_e(w)$$

with

$$g_e(w) = 1 + \frac{1}{\log n} \sum_{d' \in D} \frac{tf_{w,d'}}{gf_w} \log \frac{tf_{w,d'}}{gf_w}$$

where $tf_{w,d}$ represents occurrence frequency of word w in document d, parameter gf_w represents global frequency of word w in corpus D, and n is the number of documents in corpus D. Next, we decompose each of the two matrices using SVD in order to obtain a vector for each word in the vector space of reduced dimensionality k ($k \ll n$). LSA vectors tend to express semantic properties of words. Moreover, the similarity between the LSA vectors may be used as a measure of semantic similarity between the corresponding words. We compute this similarity using the cosine between the LSA vectors and use the obtained values as weights of graph edges. Because running random-walk algorithms on a complete graph would be computationally intractable, we decided to reduce the number of edges by thresholding the similarity values.

3.3 Similarity Based on Random Indexing

Random Indexing (RI) is another word space approach, which presents an efficient and scalable alternative to more commonly used word space methods such as LSA. Random indexing is a dimensionality reduction technique in which a random matrix is used to project the original word-context matrix into the vector space of lower dimensionality. Each context is represented by its *index vector*, a sparse vector with a small number of randomly distributed $+1$ and -1 values, the remaining values being 0 (Sahlgren, 2006). For each corpus word its *context vector* is constructed by summing index vectors of all context elements occurring within contexts of all of its occurrences in the corpus. The semantic similarity of the two words is then expressed as the similarity between its context vectors.

We use two different definitions for the context and context relation. In the first case (referred to as *RI with document context*), each corpus document is considered as a separate context and the word is considered to be in a context relation if it occurs in the document. The context vector of each word is then simply the sum of random index vectors of the documents in which the word occurs. In the second case (referred to as *RI with window context*), each corpus word is considered as a context itself, and the two words are considered to be in a context relation if they co-occur in the corpus within the window of a given size. The context vector of each corpus word is then computed as the sum of random index vectors of all words with which it co-occurs in the corpus inside the window of a given size. Like in the LSA approach, we use the cosine of the angle between the context vectors as a measure of semantic similarity between the word pairs. To reduce the number of edges, we again perform the thresholding of the similarity values.

3.4 Random-Walk Algorithms

Once the graph building phase is done, we start propagating the sentiment scores from the vertices of the seed set words to the unlabeled vertices. To this end, one can use several semi-supervised learning algorithms. The most commonly used algorithm for dictionary-based sentiment lexicon acquisition is PageRank. Along with the PageRank we employ another random-walk algorithm called harmonic function learning.

PageRank

PageRank (Page et al., 1999) was initially designed for ranking web pages by their relevance. The intuition behind PageRank is that a vertex v should have a high score if it has many high-scoring neighbours and these neighbours do not have many other neighbours except the vertex v. Let $\mathbf{W}$ be the weighted row-normalized adjacency matrix of graph G. The algorithm iteratively computes the vector of vertex scores $\mathbf{a}$ in the following way:

$$\mathbf{a^{(k)}} = \alpha \mathbf{a^{(k-1)}} \mathbf{W} + (\mathbf{1} - \alpha)\mathbf{e}$$

where α is the PageRank damping factor. Vector $\mathbf{e}$ models the normalized internal source of score for all vertices and its elements sum up to 1. We assign the value of e_i to be $\frac{1}{|SeedSet|}$ for the vertices whose corresponding words belong to the seed set and $e_i = 0$ for all other vertices.

Harmonic Function

The second graph-based semi-supervised learning algorithm we use is the harmonic func-

tion label propagation (also known as absorbing random walk) (Zhu and Goldberg, 2009). Harmonic function tries to propagate labels between sources and sinks of sentiment. We perform two runs of the algorithm: one for positive sentiment, in which we use the words from the positive seed set as sentiment sources, and one for the negative sentiment, in which we use the words from the negative seed set as sentiment sources. In both cases, we use the precompiled seed set of neutral words as sentiment sinks. Note that we could not have used positive seed set words as sources and negative seed set words as sinks (or vice versa) because we aim to predict the positive and negative sentiment scores separately.

The value of the harmonic function for a labeled vertex remains the same as initially labeled, whereas for an unlabeled vertex the value is computed as the weighted average of its neighbours' values (Zhu and Goldberg, 2009):

$$f(v_k) = \frac{\sum_{j \in |V|} w_{kj} \cdot f(v_j)}{\sum_{j \in |V|} w_{kj}}$$

where V is the set of vertices of graph G and w_{kj} is the weight of the edge between the vertices v_k and v_j. If there is no graph edge between vertices v_k and v_j, the value of the weight w_{kj} is 0. This equation also represents the update rule for the iterative computation of the harmonic function. However, it can be shown that there is a closed-form solution of the harmonic function. Let W be the unnormalized weighted adjacency matrix of the graph G, and let D be the diagonal matrix with the element $D_{ii} = \sum_{j \in |V|} w_{ij}$ being the weighted degree of the vertex v_i. Then the unnormalized graph Laplacian is defined with $L = D - W$. Assuming that the labeled seed set vertices are ordered before the unlabeled ones, the graph Laplacian can be partitioned in the following way:

$$L = \begin{pmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{pmatrix}$$

The closed form solution for the harmonic function of the unlabeled vertices is then given by:

$$\mathbf{f_u} = -\mathbf{L_{uu}^{-1}} \mathbf{L_{ul}} \mathbf{y_l}$$

where $\mathbf{y_l}$ if the vector of labels of the seed set vertices (Zhu and Goldberg, 2009).

4 Supervised Step Hybridization

The sentiment scores obtained by the semi-supervised graph-based approaches described above are relative because they depend on the graph size as well as on the size and content of the seed sets. As such, these values can be used to rank the words by positivity or negativity, but not as absolute positivity and negativity scores. Thus, in the second step of our hybrid approach, we use supervised learning to obtain the absolute sentiment scores (polarity regression task) and the sentiment labels (sentiment classification task).

Each score obtained on each graph represents a single feature for supervised learning. There are altogether 24 different semi-supervised features used as input for the supervised learners. These features are both positive and negative labels generated from six different semi-supervised graphs (co-occurence frequency, co-occurrence PMI, LSA log-entropy, LSA tf-idf, random indexing with document context, and random indexing with window context) using two different random-walk algorithms (harmonic function and PageRank). We used the occurrence frequency of words in corpus as an additional feature.

For polarity regression, learning must be performed twice: once for the negative and once for the positive sentiment score. We performed the regression using SVM with radial-basis kernel. The same set of features used for regression was used for sentiment classification, but the goal was to predict the class of the word (positive, negative, or neutral) instead of separate positivity or negativity scores. SVM with radial-basis kernel was used to perform classification learning as well.

5 Evaluation and Results

All the experiments were performed on the excerpt of the New York Times corpus (years 2002–2007), containing 434,494 articles. The corpus was preprocessed (tokenized, lemmatized, and POS tagged) and only the content lemmas (nouns, verbs, adjectives, and adverbs) occurring at least 80 times in the corpus were considered. Lemmas occurring less than 80 were mainly named entities or their derivatives. The final sentiment lexicon consists of 41,359 lemmas annotated with positivity and negativity scores and sentiment class.[1]

[1] Sentiment lexicon is freely available at `http://takelab.fer.hr/sentilex`

5.1 Sentiment Annotations

To evaluate our methods on the three tasks, we compare the results against the Micro-WN(Op) dataset (Cerini et al., 2007). Micro-WN(Op) contains sentiment annotations for 1105 WordNet 2.0 synsets. Each synset s is manually annotated with the degree of positivity $Pos(s)$ and negativity $Neg(s)$, where $0 \leq Pos(s) \leq 1$, $0 \leq Neg(s) \leq 1$, and $Pos(s) + Neg(s) \leq 1$. Objectivity score is defined as $Obj(s) = 1 - (Pos(s) + Neg(s))$.

This gives us a list of 2800 word-sense pairs with their sentiment annotations. For reasons that we explain below, we retain from this list only those words for which all senses from WordNet have been sentiment-annotated, which leaves us with a list of 1645 word-sense pairs. From this list we then filter out all words that occur less than 80 times in our corpus, leaving us with a list of 1125 word-sense pairs (365 distinct words, of which 152 are monosemous). We refer to this set of 1125 sentiment-annotated word-sense pairs as Micro-WN(Op)-0.

Because our corpus-based methods are unable to discriminate among various senses of a polysemous word, we wish to be able to eliminate the negative effect of polysemy in our evaluation. The motivation for this is twofold: first, it gives us a way of measuring how much polysemy influences our results. Secondly, it provides us with the answer how well our method could perform in an ideal case where all the words from corpus have been pre-disambiguated. Because each of the words in Micro-WN(Op)-0 has all its senses sentiment-annotated, we can determine for each of these words how sentiment depends on its sense. Expectedly, there are words whose sentiment differs radically across its senses or parts-of-speech (e.g., *catch*, *nest*, *shark*, or *hot*), but also words whose sentiment is constant or similar across all its senses. To eliminate the effect of polysemy on sentiment prediction, we further filter the Micro-WN(Op)-0 list by retaining only the words whose sentiment is constant or nearly constant across all their senses. We refer to such words as *monosentimous*. We consider a word to be monosentimous iff (1) pairwise differences between all sentiment scores across senses are less than 0.25 (separately for both positive and negative sentiment) and (2) the sign of the difference between positive and negative sentiment score is constant across all senses. Note that every monosemous word is by definition monosentimous. Out of 365 words in Micro-WN(Op)-0, 225 of them are monosentimous. To obtain the sentiment scores of monosentimous words, we simply average the scores across their senses. We refer to the so-obtained set of 225 sentiment-annotated words as Micro-WN(Op)-1.

5.2 Semi-supervised Step Evaluation

The semi-supervised step was designed to propagate sentiment properties of the labeled words, ordering the words according to their positivity or negativity. Therefore, we decided to use the evaluation metric that measures the quality of the ranking in ordered lists, Kendall τ distance. The performance of the semi-supervised graph-based methods was evaluated both on the Micro-WN(Op)-1 and Micro-WN(Op)-0 sets.

In order to be able to compare our results to SentiWordNet (Baccianella et al., 2010), the *de facto* standard sentiment lexicon for English, we use the p-normalized Kendall τ distance between the rankings generated by our semi-supervised graph-based methods and the gold standard rankings. The p-normalized Kendall τ distance (Fagin et al., 2004) is a version of the standard Kendall τ distance that accounts for ties in the ordering:

$$\tau = \frac{n_d + p \cdot n_t}{Z}$$

where n_d is the number of pairs in disagreement (i.e., pairs of words ordered one way in the gold standard and the opposite way in the ranking under evaluation), n_t is the number of pairs which are ordered in the gold standard and tied in the ranking under evaluation, p is the penalization factor to be assigned to each of the n_t pairs (usually set to $p = \frac{1}{2}$), and Z is the number of pairs of words that are ordered in the gold standard. Table 1 presents the results for each of the methods used to build the sentiment graph and for both random-walk algorithms. The results were obtained by evaluating the relative rankings of words against the Micro-WN(Op)-1 as gold standard. For comparison, the p-normalized Kendall τ scores for SentiWordNet 1.0 and SentiWordNet 3.0 are extracted from (Baccianella et al., 2010).

Rankings for the negative scores are consistently better across all methods and algorithms. We believe that the negative rankings are better

Table 1: The results on the polarity ranking task

	Harmonic function		PageRank	
	Positive	Negative	Positive	Negative
Co-occurrence freq.	0.395	0.298	0.540	0.544
LSA log-entropy	0.425	0.308	0.434	0.370
LSA tf-idf	0.396	0.320	0.417	0.424
Co-occurrence PMI	**0.321**	**0.256**	0.550	0.576
Random indexing document context	0.402	0.433	0.534	0.557
Random indexing window context	0.455	0.398	0.491	0.436
	Positive		Negative	
SentiWordNet 1.0	0.349		0.296	
SentiWordNet 3.0	0.281		0.231	

for two reasons. Firstly, the corpus contains many more articles describing negative events such as wars and accidents than the articles describing positive events such as celebrations and victories. In short, the distribution of articles is significantly skewed towards "negative" events. Secondly, the lemma *new*, which was included in the positive seed set, occurs in the corpus very frequently as a part of named entity collocations such as "New York" and "New Jersey" in which it does not reflect its dominant sense. The harmonic function label propagation generally outperforms the PageRank algorithm. The best performance on the Micro-WN(Op)-0 set was 0.380 for the positive ranking and 0.270 for the negative ranking, showing that the performance deteriorates when polysemy is present. However, the drop in performance, especially for the negative ranking, is not substantial. Our best method (graph built based on PMI of corpus words used in combination with harmonic function label propagation) outperforms SentiWordNet 1.0 and performs slightly worse than SentiWordNet 3.0 for both positive and negative rankings.

5.3 Evaluation of the Supervised Step

Supervised step deals with the polarity regression task and the sentiment classification task. Polarity regression maps the "virtual" sentiment scores obtained on graphs to the absolute sentiment scores (on a scale from 0 to 1). The regression was performed twice: once for the positive scores and once for the negative scores. We evaluate the performance of the polarity regression against the Micro-WN(Op)-0 gold standard in terms of root mean square error (RMSE). We used the average of the labeled polarity scores (positive and negative) of all monosentimous words in Micro-WN(Op)-1 as a baseline for this task.

Sentiment classification uses the scores obtained on graphs as features in order to assign each word with one of the three sentiment labels (*positive*, *negative*, and *neutral*). The classification performance is evaluated in terms of micro-F1 measure. The labels for the classification are assigned according to the positivity and negativity scores (the label *neutral* is assigned if $Obj(s) = 1 - Pos(s) - Neg(s)$ is larger than both $Pos(s)$ and $Neg(s)$). The majority class predictor was used as a baseline for the classification task.

Due to the small size of the labeled sets (e.g., 225 for Micro-WN(Op)-1) we performed the 10 $\times$ 10 CV evaluation (10 cross-validation trials, each on randomly permuted data) (Bouckaert, 2003) both for regression and classification. For comparison, we evaluated the SentiWordNet in the same way – we averaged the SentiWordNet scores for all the senses of monosentimous words from the Micro-WN(Op)-1.

Although the semi-supervised step itself was not designed to deal with polarity regression task and sentiment classification task, we decided to evaluate the results gained from graphs on these tasks as well. This gives us an insight to how much the supervised step adds in terms of performance. The positivity and negativity scores obtained from graphs were directly evaluated on the regression task measuring the RMSE against the gold standard. Classification labels were deter-

mined by comparing the positive rank of the word against the negative rank of the word. The word was classified as *neutral* if the absolute difference between its positive and negative rank was below the given treshold t. Empirically determined optimal value of the treshold was $t = 1000$.

Table 2 we present the results of the hybrid method on both the regression (for both positive and negative scores) and classification tasks compared with the performance of the SentiWordNet and the baselines. Additionally, we present the results obtained using only the semi-supervised step. On both the regression and classification task our method outperforms the baseline. The performance is comparable to SentiWordNet on the sentiment classification task. However, the performance of our corpus-based approach is significantly lower than SentiWordNet on the polarity regression task – a more detailed analysis is required to determine the cause of this. The hybrid approach performs significantly better than the semi-supervised method alone, confirming the importance of the supervised step.

Models trained on the Micro-WN(Op)-1 were applied on the set of words from the Micro-WN(Op)-0 not present in the Micro-WN(Op)-1 (i.e., the difference between the two sets) in order to test the performance on non-monosentimous words. The obtained results on this set are, surprisingly, slightly better (positivity regression – 0.337; negativity regression – 0.313; and classification – 57.55%). This is most likely due to the fact that, although not all senses have the same sentiment, most of them have similar sentiment, which is often also the sentiment of the dominant sense in the corpus.

6 Conclusion

We have described a hybrid approach to sentiment lexicon acquisition from corpus. On one hand, the approach combines corpus-based lexical semantics with graph-based label propagation, while on the other hand it combines semi-supervised and supervised learning. We have evaluated the performance on three sentiment prediction tasks: polarity ranking task, polarity regression task, and sentiment classification task. Our experiments suggest that the results on the polarity ranking task are comparable to SentiWordNet. On the sentiment classification task, the results are also comparable to SentiWordNet when restricted to monosentimous words. On the polarity regression task, our results are worse than SentiWordNet, although still above the baseline.

Unlike with the WordNet-based approaches, in which sentiment is predicted based on sentiment-preserving semantic relations between synsets, the corpus-based approach operates at the level of words and thus suffers from two major limitations. Firstly, the semantic relations extracted from corpus are inherently unstructured, vague, and – besides paradigmatic relations – also include syntagmatic and very loose topical relations. Thus, sentiment labels propagate in a less controlled manner and get influenced more easily by the context. For example, words "understandable" and "justifiable" get labeled as predominately negative, because they usually occur in negative contexts. Secondly, in the approach we described, polysemy is not accounted for, which introduces sentiment prediction errors for words that are not monosentimous. It remains to be seen whether this could be remedied by employing WSD prior to sentiment lexicon acquisition.

For future work we intend to investigate how syntax-based information can be used to introduce more semantic structure into the graph. We will experiment with other hybridization approaches that combine semantic links from WordNet with corpus-derived semantic relations.

Acknowledgments

We thank the anonymous reviewers for their useful comments. This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia under the Grant 036-1300646-1986.

References

E. Agirre, D. Martínez, O.L. de Lacalle, and A. Soroa. 2006. Two graph-based algorithms for state-of-the-art wsd. In *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 585–593. Association for Computational Linguistics.

A. Andreevskaia and S. Bergler. 2006. Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *Proc. of EACL*, volume 6, pages 209–216.

S. Baccianella, A. Esuli, and F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In

Table 2: The performance on the polarity regression task and sentiment classification task

	Regression (RMSE)		Classification (micro-F1)
	Positivity	Negativity	
Hybrid approach	0.363 ± 0.005	0.387 ± 0.003	0.548 ± 0.126
Baseline	0.383	0.413	0.427
Semi-supervised	0.443	0.466	0.484
SentiWordNet	0.284	0.294	0.582

Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

R.R. Bouckaert. 2003. Choosing between two learning algorithms based on calibrated tests. In *Machine learning-International workshop then conference-*, volume 20, pages 51–58.

S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini. 2007. Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*, pages 200–210.

K.W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

S.T. Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230.

A. Esuli and F. Sebastiani. 2007. Pageranking wordnet synsets: An application to opinion mining. In *Annual meeting-association for computational linguistics*, volume 45, pages 424–431.

S. Evert. 2008. Corpora and collocations. *Corpus Linguistics. An International Handbook*, pages 1212–1248.

R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. 2004. Comparing and aggregating rankings with ties. In *Proc. of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 47–58. ACM.

C. Fellbaum. 2010. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243.

A.B. Goldberg and X. Zhu. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proc. of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52. Association for Computational Linguistics.

V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proc. of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics.

M. Hu and B. Liu. 2004. Mining opinion features in customer reviews. In *Proc. of the National Conference on Artificial Intelligence*, pages 755–760.

J. Kamps, MJ Marx, R.J. Mokken, and M. De Rijke. 2004. Using WordNet to measure semantic orientations of adjectives.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. The PageRank citation ranking: Bringing order to the web.

E. Riloff, S. Patwardhan, and J. Wiebe. 2006. Feature subsumption for opinion analysis. In *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 440–448. Association for Computational Linguistics.

M. Sahlgren. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University, Stockholm, Sweden.

S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov. 2007. Qa with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *Proc. of the International Conference on Weblogs and Social Media (ICWSM)*. Citeseer.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, (Early Access):1–41.

P. Turney and M.L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. In *ACM Transactions on Information Systems (TOIS)*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

X. Zhu and A.B. Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.