Generating a Non-English Subjectivity Lexicon: Relations That Matter

Valentin Jijkoun and Katja Hofmann
ISLA, University of Amsterdam
Amsterdam, The Netherlands
{jijkoun,k.hofmann}@uva.nl

Abstract

We describe a method for creating a non-English subjectivity lexicon based on an English lexicon, an online translation service and a general purpose thesaurus: Wordnet. We use a PageRank-like algorithm to bootstrap from the translation of the English lexicon and rank the words in the thesaurus by polarity using the network of lexical relations in Wordnet. We apply our method to the Dutch language. The best results are achieved when using synonymy and antonymy relations only, and ranking positive and negative words simultaneously. Our method achieves an accuracy of 0.82 at the top 3,000 negative words, and 0.62 at the top 3,000 positive words.

1 Introduction

One of the key tasks in subjectivity analysis is the automatic detection of subjective (as opposed to objective, factual) statements in written documents (Mihalcea and Liu, 2006). This task is essential for applications such as online marketing research, where companies want to know what customers say about the companies, their products, specific products' features, and whether comments made are positive or negative. Another application is in political research, where public opinion could be assessed by analyzing user-generated online data (blogs, discussion forums, etc.).

Most current methods for subjectivity identification rely on *subjectivity lexicons*, which list words that are usually associated with positive or negative sentiments or opinions (i.e., words with *polarity*). Such a lexicon can be used, e.g., to classify individual sentences or phrases as subjective or not, and as bearing positive or negative sentiments (Pang et al., 2002; Kim and Hovy, 2004; Wilson et al., 2005a). For English, manually created subjectivity lexicons have been available for a while, but for many other languages such resources are still missing.

We describe a language-independent method for automatically bootstrapping a subjectivity lexicon, and apply and evaluate it for the Dutch language. The method starts with an English lexicon of positive and negative words, automatically translated into the target language (Dutch in our case). A PageRank-like algorithm is applied to the Dutch wordnet in order to filter and expand the set of words obtained through translation. The Dutch lexicon is then created from the resulting ranking of the wordnet nodes. Our method has several benefits:

- It is applicable to any language for which a wordnet and an automatic translation service or a machine-readable dictionary (from English) are available. For example, the EuroWordnet project (Vossen, 1998), e.g., provides wordnets for 7 languages, and free online translation services such as the one we have used in this paper are available for many other languages as well.

- The method ranks all (or almost all) entries of a wordnet by polarity (positive or negative), which makes it possible to experiment with different settings of the precision/coverage threshold in applications that use the lexicon.

We apply our method to the most recent version of Cornetto (Vossen et al., 2007), an extension of the Dutch WordNet, and we experiment with various parameters of the algorithm, in order to arrive at a good setting for porting the method to other languages. Specifically, we evaluate the quality of the resulting Dutch subjectivity lexicon using different subsets of wordnet relations and information in the glosses (definitions). We also examine

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 398–405,
Athens, Greece, 30 March – 3 April 2009.

the effect of the number of iterations on the performance of our method. We find that best performance is achieved when using only synonymy and antonymy relations and, moreover, the algorithm converges after about 10 iterations.

The remainder of the paper is organized as follows. We summarize related work in section 2, present our method in section 3 and describe the manual assessment of the lexicon in section 4. We discuss experimental results in section 5 and conclude in section 6.

2 Related work

Creating subjectivity lexicons for languages other than English has only recently attracted attention of the research community. (Mihalcea et al., 2007) describes experiments with subjectivity classification for Romanian. The authors start with an English subjectivity lexicon with 6,856 entries, OpinionFinder (Wiebe and Riloff, 2005), and automatically translate it into Romanian using two bilingual dictionaries, obtaining a Romanian lexicon with 4,983 entries. A manual evaluation of a sample of 123 entries of this lexicon showed that 50% of the entries do indicate subjectivity.

In (Banea et al., 2008) a different approach based on boostrapping was explored for Romanian. The method starts with a small seed set of 60 words, which is iteratively (1) expanded by adding synonyms from an online Romanian dictionary, and (2) filtered by removing words which are not similar (at a preset threshold) to the original seed, according to an LSA-based similarity measure computed on a half-million word corpus of Romanian. The lexicon obtained after 5 iterations of the method was used for sentence-level sentiment classification, indicating an 18% improvement over the lexicon of (Mihalcea et al., 2007).

Both these approaches produce unordered sets of positive and negative words. Our method, on the other hand, assigns polarity scores to words and produces a ranking of words by polarity, which provides a more flexible experimental framework for applications that will use the lexicon.

Esuli and Sebastiani (Esuli and Sebastiani, 2007) apply an algorithm based on PageRank to rank synsets in English WordNet according to positive and negativite sentiments. The authors view WordNet as a graph where nodes are synsets and synsets are linked with the synsets of terms used in their glosses (definitions). The algorithm is initialized with positivity/negativity scores provided in SentiWordNet (Esuli and Sebastiani, 2006), an English sentiment lexicon. The weights are then distributed through the graph using an the algorithm similar to PageRank. Authors conclude that larger initial seed sets result in a better ranking produced by the method. The algorithm is always run twice, once for positivity scores, and once for negativity scores; this is different in our approach, which ranks words from negative to positive in one run. See section 5.4 for a more detailed comparison between the existing approaches outlined above and our approach.

3 Approach

Our approach extends the techniques used in (Esuli and Sebastiani, 2007; Banea et al., 2008) for mining English and Romanian subjectivity lexicons.

3.1 Boostrapping algorithm

We hypothesize that concepts (synsets) that are closely related in a wordnet have similar meaning and thus similar polarity. To determine relatedness between concepts, we view a wordnet as a graph of lexical relations between words and synsets:

- nodes correspond to lexical units (words) and synsets; and
- directed arcs correspond to relations between synsets (hyponymy, meronymy, etc.) and between synsets and words they contain; in one of our experiments, following (Esuli and Sebastiani, 2007), we also include relations between synsets and all words that occur in their glosses (definitions).

Nodes and arcs of such a graph are assigned weights, which are then propagated through the graph by iteratively applying a PageRank-like algorithm.

Initially, weights are assigned to nodes and arcs in the graph using translations from an English polarity lexicon as follows:

- words that are translations of the positive words from the English lexicon are assigned a weight of 1, words that are translations of the negative words are initialized to -1; in general, weight of a word indicates its polarity;

- All arcs are assigned a weight of 1, except for antonymy relations which are assigned a weight of -1; the intuition behind the arc weights is simple: arcs with weight 1 would usually connect synsets of the same (or similar) polarity, while arcs with weight -1 would connect synsets with opposite polarities.

We use the following notation. Our algorithm is iterative and $k = 0, 1, \ldots$ denotes an iteration. Let a_i^k be the weight of the node i at the k-th iteration. Let w_{jm} be the weight of the arc that connects node j with node m; we assume the weight is 0 if the arc does not exist. Finally, α is a damping factor of the PageRank algorithm, set to 0.8. This factor balances the impact of the initial weight of a node with the impact of weight received through connections to other nodes.

The algorithm proceeds by updating the weights of nodes iteratively as follows:

$$a_i^{k+1} = \alpha \cdot \sum_j \frac{a_j^k \cdot w_{ji}}{\sum_m |w_{jm}|} + (1 - \alpha) \cdot a_i^0$$

Furthermore, at each iterarion, all weights a_i^{k+1} are normalized by $\max_j |a_j^{k+1}|$.

The equation above is a straightforward extension of the PageRank method for the case when arcs of the graph are weighted. Nodes propagate their polarity mass to neighbours through outgoing arcs. The mass transferred depends on the weight of the arcs. Note that for arcs with negative weight (in our case, antonymy relation), the polarity of transferred mass is inverted: i.e., synsets with negative polarity will enforce positive polarity in their antonyms.

We iterate the algorithm and read off the resulting weight of the word nodes. We assume words with the lowest resulting weight to have negative polarity, and word nodes with the highest weight positive polarity. The output of the algorithm is a list of words ordered by polarity score.

3.2 Resources used

We use an English subjectivity lexicon of OpinionFinder (Wilson et al., 2005b) as the starting point of our method. The lexicon contains 2,718 English words with positive polarity and 4,910 words with negative polarity. We use a free online translation service[1] to translate positive and negative polarity words into Dutch, resulting in 974 and 1,523 Dutch words, respectively. We assumed that a word was translated into Dutch successfully if the translation occurred in the Dutch wordnet (therefore, the result of the translation is smaller than the original English lexicon).

The Dutch wordnet we used in our experiments is the most recent version of Cornetto (Vossen et al., 2007). This wordnet contains 103,734 lexical units (words), 70,192 synsets, and 157,679 relations between synsets.

[1] `http://translate.google.com`

4 Manual assessments

To assess the quality of our method we re-used assessments made for earlier work on comparing two resources in terms of their usefulness for automatically generating subjectivity lexicons (Jijkoun and Hofmann, 2008). In this setting, the goal was to compare two versions of the Dutch Wordnet: the first from 2001 and the other from 2008. We applied the method described in section 3 to both resources and generated two subjectivity rankings. From each ranking, we selected the 2000 words ranked as most negative and the 1500 words ranked as most positive, respectively. More negative than positive words were chosen to reflect the original distribution of positive vs. negative words. In addition, we selected words for assessment from the remaining parts of the ranked lists, randomly sampling chunks of 3000 words at intervals of 10000 words with a sampling rate of 10%. The selection was made in this way because we were mostly interested in negative and positive words, i.e., the words near either end of the rankings.

4.1 Assessment procedure

Human annotators were presented with a list of words in random order, for each word its part-of-speech tag was indicated. Annotators were asked to identify positive and negative words in this list, i.e., words that indicate positive (negative) emotions, evaluations, or positions.

Annotators were asked to classify each word on the list into one of five classes:

++ the word is positive in most contexts (*strongly positive*)

\+ the word is positive in some contexts (*weakly positive*)

0 the word is hardly ever positive or negative (*neutral*)

– the a word is negative in some contexts (*weakly negative*)

–– the word is negative in most contexts (*strongly negative*)

Cases where assessors were unable to assign a word to one of the classes, were separately marked as such.

For the purpose of this study we were only interested in identifying subjective words without considering subjectivity strength. Furthermore, a pilot study showed assessments of the strength of subjectivity to be a much harder task (54% inter-annotator agreement) than distinguishing between positive, neutral and negative words only (72% agreement). We therefore collapsed the classes of strongly and weakly subjective words for evaluation. These results for three classes are reported and used in the remainder of this paper.

4.2 Annotators

The data were annotated by two undergraduate university students, both native speakers of Dutch. Annotators were recruited through a university mailing list. Assessment took a total of 32 working hours (annotating at approximately 450-500 words per hour) which were distributed over a total of 8 annotation sessions.

4.3 Inter-annotator Agreement

In total, 9,089 unique words were assessed, of which 6,680 words were assessed by both annotators. For 205 words, one or both assessors could not assign an appropriate class; these words were excluded from the subsequent study, leaving us with 6,475 words with double assessments.

Table 1 shows the number of assessed words and inter-annotator agreement overall and per part-of-speech. Overall agreement is 69% (Cohen's κ=0.52). The highest agreement is for adjectives, at 76% (κ=0.62) . This is the same level of agreement as reported in (Kim and Hovy, 2004) for English. Agreement is lowest for verbs (55%, κ=0.29) and adverbs (56%, κ=0.18), which is slightly less than the 62% agreement on verbs reported by Kim and Hovy. Overall we judge agreement to be reasonable.

Table 2 shows the confusion matrix between the two assessors. We see that one assessor judged more words as subjective overall, and that more words are judged as negative than positive (this can be explained by our sampling method described above).

POS	Count	% agreement	κ
noun	3670	70%	0.51
adjective	1697	76%	0.62
adverb	25	56%	0.18
verb	1083	55%	0.29
overall	6475	69%	0.52

Table 1: Inter-annotator agreement per part-of-speech.

	–	0	+	Total
–	1803	137	39	1979
0	1011	1857	649	3517
+	81	108	790	979
Total	2895	2102	1478	6475

Table 2: Contingency table for all words assessed by two annotators.

5 Experiments and results

We evaluated several versions of the method of section 3 in order to find the best setting.

Our **baseline** is a ranking of all words in the wordnet with the weight -1 assigned to the translations of English negative polarity words, 1 assigned to the translations of positive words, and 0 assigned to the remaining words. This corresponds to simply translating the English subjectivity lexicon.

In the run **all.100** we applied our method to all words, synsets and relations from the Dutch Wordnet to create a graph with 153,386 nodes (70,192 synsets, 83,194 words) and 362,868 directed arcs (103,734 word-to-synset, 103,734 synset-to-word, 155,400 synset-to-synset relations). We used 100 iterations of the PageRank algorihm for this run (and all runs below, unless indicated otherwise).

In the run **syn.100** we only used synset-to-word, word-to-synset relations and 2,850 near-synonymy relations between synsets. We added 1,459 near-antonym relations to the graph to produce the run **syn+ant.100**. In the run **syn+hyp.100** we added 66,993 hyponymy and 66,993 hyperonymy relations to those used in run syn.100.

We also experimented with the information provided in the definitions (glosses) of synset. The glosses were available for 68,122 of the 70,192

synsets. Following (Esuli and Sebastiani, 2007), we assumed that there is a semantic relationship between a synset and each word used in its gloss. Thus, the run **gloss.100** uses a graph with 70,192 synsets, 83,194 words and 350,855 directed arcs from synsets to lemmas of all words in their glosses. To create these arcs, glosses were lemmatized and lemmas not found in the wordnet were ignored.

To see if the information in the glosses can complement the wordnet relations, we also generated a hybrid run **syn+ant+gloss.100** that used arcs derived from word-to-synset, synset-to-word, synonymy, antonymy relations and glosses.

Finally, we experimented with the number of iterations of PageRank in two setting: using all wordnet relations and using only synonyms and antonyms.

5.1 Evaluation measures

We used several measures to evaluate the quality of the word rankings produced by our method.

We consider the evaluation of a ranking parallel to the evaluation for a binary classification problem, where words are classified as positive (resp. negative) if the assigned score exceeds a certain threshold value. We can select a specific threshold and classify all words exceeding this score as positive. There will be a certain amount of correctly classified words (true positives), and some incorrectly classified words (false positives). As we move the threshold to include a larger portion of the ranking, both the number of true positives and the number of false positives increase.

We can visualize the quality of rankings by plotting their *ROC curves*, which show the relation between true positive rate (portion of the data correctly labeled as positive instances) and false positive rate (portion of the data incorrectly labeled as positive instances) at all possible threshold settings.

To compare rankings, we compute the *area under the ROC curve (AUC)*, a measure frequently used to evaluate the performance of ranking classifiers. The AUC value corresponds to the probability that a randomly drawn positive instance will be ranked higher than a randomly drawn negative instance. Thus, an AUC of 0.5 corresponds to random performance, a value of 1.0 corresponds to perfect performance. When evaluating word rankings, we compute AUC^- and AUC^+ as evaluation measures for the tasks of identifying words with negative (resp., positive) polarity.

Run	τ_k	D_k	AUC^-	AUC^+
baseline	0.395	0.303	0.701	0.733
syn.10	0.641	0.180	0.829	0.837
gloss.100	0.637	0.181	0.829	0.835
all.100	0.565	0.218	0.792	0.787
syn.100	0.645	0.177	0.831	0.839
syn+ant.100	**0.650**	**0.175**	**0.833**	**0.841**
syn+ant+gloss.100	0.643	0.178	0.831	0.838
syn+hyp.100	0.594	0.203	0.807	0.810

Table 3: Evaluation results

Other measures commonly used to evaluate rankings are Kendall's rank correlation, or Kendall's tau coefficient, and Kendall's distance (Fagin et al., 2004; Esuli and Sebastiani, 2007). When comparing rankings, Kendall's measures look at the number of pairs of ranked items that agree or disagree with the ordering in the gold standard. The measures can deal with partially ordered sets (i.e., rankings with ties): only pairs that are ordered in the gold standard are used. Let $T = \{(a_i, b_i)\}_i$ denote the set of pairs ordered in the gold standard, i.e., $a_i \prec_g b_i$. Let $C = \{(a, b) \in T \mid a \prec_r b\}$ be the set of concordant pairs, i.e., pairs ordered the same way in the gold standard and in the ranking. Let $D = \{(a, b) \in T \mid b \prec_r a\}$ be the set of discordant pairs and $U = T \setminus (C \cup D)$ the set of pairs ordered in the gold standard, but tied in the ranking. Kendall's rank correlation coefficient τ_k and Kendall's distance D_k are defined as follows:

$$\tau_k = \frac{|C| - |D|}{|T|} \quad D_k = \frac{|D| + p \cdot |U|}{|T|}$$

where p is a penalization factor for ties, which we set to 0.5, following (Esuli and Sebastiani, 2007).

The value of τ_k ranges from -1 (perfect disagreement) to 1 (perfect agreement), with 0 indicating an almost random ranking. The value of D_k ranges from 0 (perfect agreement) to 1 (perfect disagreement).

When applying Kendall's measures we assume that the gold standard defines a partial order: for two words a and b, $a \prec_g b$ holds when $a \in N_g, b \in U_g \cup P_g$ or when $a \in U_g, b \in P_g$; here N_g, U_g, P_g are sets of words judged as negative, neutral and positive, respectively, by human assessors.

5.2 Types of wordnet relations

The results in Table 3 indicate that the method performs best when only synonymy and antonymy

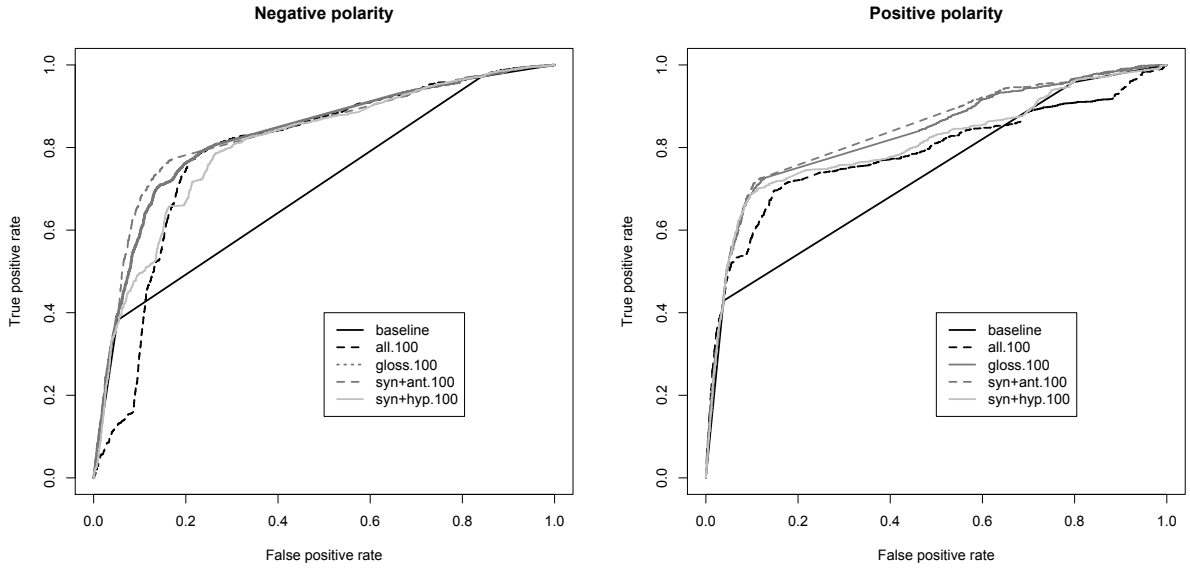

Figure 1: ROC curves showing the impact of using different sets of relations for negative and positive polarity. Graphs were generated using ROCR (Sing et al., 2005).

relations are considered for ranking. Adding hyponyms and hyperonyms, or adding relations between synsets and words in their glosses substantially decrease the performance, according to all four evaluation measures. With all relations, the performance degrades even further. Our hypothesis is that with many relations the polarity mass of the seed words is distributed too broadly. This is supported by the drop in the performance early in the ranking at the "negative" side of runs with all relations and with hyponyms (Figure 1, left). Another possible explanation can be that words with many incoming arcs (but without strong connections to the seed words) get substantial weights, thereby decreasing the quality of the ranking.

Antonymy relations also prove useful, as using them in addition to synonyms results in a small improvement. This justifies our modification of the PageRank algorithm, when we allow negative node and arc weights.

In the best setting (syn+ant.100), our method achieves an accuracy of 0.82 at top 3,000 negative words, and 0.62 at top 3,000 positive words (estimated from manual assessments of a sample, see section 4). Moreover, Figure 1 indicates that the accuracy of the seed set (i.e., the baseline translations of the English lexicon) is maintained at the positive and negative ends of the ranking for most variants of the method.

5.3 The number of iterations

In Figure 2 we plot how the AUC^{-} measure changes when the number of PageRank iterations increases (for positive polarity; the plots are almost identical for negative polarity). Although the absolute maximum of AUC is achieved at 110 iteration (60 iterations for positive polarity), the AUC clearly converges after 20 iterations. We conclude that after 20 iterations all useful information has been propagated through the graph. Moreover, our version of PageRank reaches a stable weight distribution and, at the same time, produces the best ranking.

5.4 Comparison to previous work

Although the values in the evaluation results are, obviously, language-dependent, we tried to replicate the methods used in the literature for Romanian and English (section 2), to the degree possible.

Our **baseline** replicates the method of (Mihalcea et al., 2007): i.e., a simple translation of the English lexicon into the target language. The run **syn.10** is similar to the iterative method used in (Banea et al., 2008), except that we do not perform a corpus-based filtering. We run PageRank for 10 iterations, so that polarity is propagated from the seed words to all their 5-step-synonymy neighbours. Table 3 indicates that increasing the number of iterations in the method of (Banea et

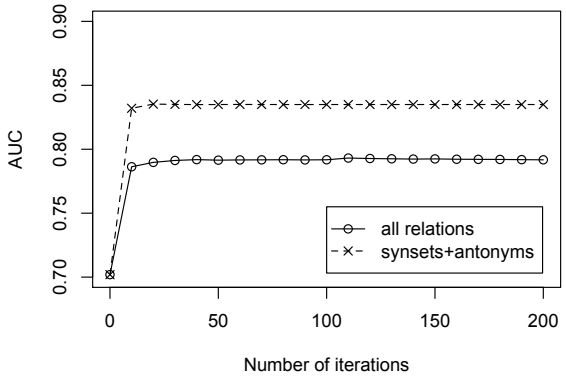

Figure 2: The number of iterations and the ranking quality (AUC), for positive polarity. Rankings for negative polarity behave similarly.

al., 2008) might help to generate a better subjectivity lexicon.

The run **gloss.100** is similar to the PageRank-based method of (Esuli and Sebastiani, 2007). The main difference is that Esuli and Sebastiani used the extended English WordNet, where words in all glosses are manually assigned to their correct synsets: the PageRank method then uses relations between synsets and *synsets of words in their glosses*. Since such a resource is not available for our target language (Dutch), we used relations between synsets and *words in their glosses*, instead. With this simplification, the PageRank method using glosses produces worse results than the method using synonyms. Further experiments with the extended English WordNet are necessary to investigate whether this decrease can be attributed to the lack of disambiguation for glosses.

An important difference between our method and (Esuli and Sebastiani, 2007) is that the latter produces two independent rankings: one for positive and one for negative words. To evaluate the effect of this choice, we generated runs **gloss.100.N** and **gloss.100.P** that used only negative (resp., only positive) seed words. We compare these runs with the run **gloss.100** (that starts with both positive and negative seeds) in Table 4. To allow a fair comparison of the generated rankings, the evaluation measures in this case are calculated separately for two binary classification problems: words with negative polarity versus all words, and words with positive polarity versus all.

The results in Table 4 clearly indicate that information about words of one polarity class helps to identify words of the other polarity: negative words are unlikely to be also positive, and vice versa. This supports our design choice: ranking words from negative to positive in one run of the method.

Run	τ_k^-	D_k^-	AUC^-
gloss.100	**0.669**	**0.166**	**0.829**
gloss.100.N	0.562	0.219	0.782
	τ_k^+	D_k^+	AUC^+
gloss.100	**0.665**	**0.167**	**0.835**
gloss.100.P	0.580	0.210	0.795

Table 4: Comparison of separate and simultaneous rankings of negative and positive words.

6 Conclusion

We have presented a PageRank-like algorithm that bootstraps a subjectivity lexicon from a list of initial seed examples (automatic translations of words in an English subjectivity lexicon). The algorithm views a wordnet as a graph where words and concepts are connected by relations such as synonymy, hyponymy, meronymy etc. We initialize the algorithm by assigning high weights to positive seed examples and low weights to negative seed examples. These weights are then propagated through the wordnet graph via the relations. After a number of iterations words are ranked according to their weight. We assume that words with lower weights are likely negative and words with high weights are likely positive.

We evaluated several variants of the method for the Dutch language, using the most recent version of Cornetto, an extension of Dutch WordNet. The evaluation was based on the manual assessment of 9,089 words (with inter-annotator agreement 69%, κ=0.52). Best results were achieved when the method used only synonymy and antonymy relations, and was ranking positive and negative words simultaneously. In this setting, the method achieves an accuracy of 0.82 at the top 3,000 negative words, and 0.62 at the top 3,000 positive words.

Our method is language-independent and can easily be applied to other languages for which wordnets exist. We plan to make the implementation of the method publicly available.

An additional important outcome of our experiments is the first (to our knowledge) manually annotated sentiment lexicon for the Dutch language.

The lexicon contains 2,836 negative polarity and 1,628 positive polarity words. The lexicon will be made publicly available as well. Our future work will focus on using the lexicon for sentence- and phrase-level sentiment extraction for Dutch.

Acknowledgments

This work was supported by projects DuOMAn and Cornetto, carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (`http://www.stevin-tst.org`), and by the Netherlands Organization for Scientific Research (NWO) under project number 612.061.814.

References

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *LREC*.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC 2006*, pages 417–422.

Andrea Esuli and Fabrizio Sebastiani. 2007. Pageranking wordnet synsets: An application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 424—431.

Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. 2004. Comparing and aggregating rankings with ties. In *PODS '04: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 47–58, New York, NY, USA. ACM.

Valentin Jijkoun and Katja Hofmann. 2008. Task-based Evaluation Report: Building a Dutch Subjectivity Lexicon. Technical report. Technical report, University of Amsterdam. `http://ilps.science.uva.nl/biblio/cornetto-subjectivity-lexicon`.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

R. Mihalcea and H. Liu. 2006. A corpus-based approach to finding happiness. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs*.

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.

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

T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941.

P. Vossen, K. Hofman, M. De Rijke, E. Tjong Kim Sang, and K. Deschacht. 2007. The cornetto database: Architecture and user-scenarios. In *Proceedings of 7th Dutch-Belgian Information Retrieval Workshop DIR2007*.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics*, volume 3406 of *Lecture Notes in Computer Science*, pages 475–486. Springer-Verlag.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005a. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 347–354.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005b. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLTEMNLP 2005*.