D-12b, D-13b
Task-based Evaluation Report: Building a Dutch Subjectivity Lexicon

Technical Report
Version 1
2 September 2008

Project STE05039
http://www.let.vu.nl/onderzoek/projectsites/cornetto/

Funded by the Stevin framework (http://taalunieversum.org/stevin/)

Authors:
Valentin Jijkoun, Katja Hofmann

Abstract

We describe a method for creating a Dutch subjectivity lexicon based on an English subjectivity lexicon, an online translation service and a Dutch general purpose thesaurus: Wordnet. We use a PageRank-like algorithm to bootstrap from the Dutch translation of the English lexicon and rank the words in the Dutch thesaurus by polarity. Two versions of the Dutch Wordnet are used in the experiments: the 2001 version and the 2008 version developed within the Cornetto project. We present the evaluation results based on human assessment of the top 2000 negative words and the top 1500 positive words in the resulting lexicons. We find that using Cornetto results in a 7% improvement in accuracy. Between 70% to 86% of this improvement can be attributed to the larger size of Cornetto, the remaining improvement is attributed to the larger set of relations between words.

Contents

1 Introduction

Subjectivity identification aims to develop technologies that can automatically detect subjective statements in written documents. Applications of this technology include, for example, marketing research, where companies want to know what customers say about the company online, and whether positive or negative comments about their products are made, and 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 so-called *subjectivity lexicons*, which contain words with positive and negative polarity. For English, such lexicons have been available for a while (those were created manually), but a similar resource is missing for Dutch. This lack of a suitable resource seriously limits the applicability of subjectivity identification methods to Dutch.

In this report we outline a method for bootstrapping a subjectivity lexicon for Dutch. Based on an English subjectivity lexicon automatically translated to Dutch, and either Dutch WordNet (DWN) or Cornetto, our method ranks Dutch words according to their polarity. We evaluate the resulting lists and determine whether DWN or Cornetto produces cleaner lexicon. A cleaner automatically generated lexicon is expected to either reduce manual effort (in case it is used as a basis for manually created lexicons), or achieve better results when directly applied to subjectivity identification.

Specifically, we answer the following research questions:

- What is the quality of the lexicon created by the method?
- Do the increased size and added lexical relations of Cornetto result in a better subjectivity lexicon?

The remainder of this report is organized as follows: we discuss our approach in section 2, present the evaluation results in section 3, and conclude in section 4.

2 Approach

Our approach extends the techniques used in [2, 1] for mining English and Romanian subjectivity lexicons.

2.1 Boostrapping algorithm

We hypothesize that concepts (i.e. synsets) that are closely related in a wordnet have similar meaning and thus similar polarity. To determine ”relatedness” between concepts, we consider the wordnet as a graph of lexical relations between words and synsets. We initialize weights in this network using translations from an English polarity lexicon and then apply a PageRank-like algorithm to propagate these initial weights throughout the network.

In our algorithm, we view Wordnet as a network consisting of:

- Nodes: literals and synsets
- Directed arcs: relations between synsets (hyponymy, meronymy etc.) and between synsets and words they contain.

Nodes and arcs in the graph are assigned weights:

- words that are translations of the positive words from the English lexicon are initially assigned weight 1, words that are translations of the negative words get -1; in general, weight of a word indicates its polarity;
- All arcs are assigned weight 1, except antonymy relations that are assigned weight -1; the intuition behind the arc weights is simple: arcs with weight 1 would usually relate synsets of the same (or similar) polarity, while arcs with weight -1 would relate synsets with opposite polarities.

We will use the following notations:

- Our algorithm is iterative, and $k = 0, 1, \ldots$ denotes an iteration;
- a_i^k is weight of node i of the graph at the k-th iteration;
- w_{jm} is the weight of the arc that goes from node j to node m; we assume the weight is 0 if there is no arc;
- α is a damping factor (as used in the original PageRank algorithm) set to 0.8.

Our 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.

We ran the agorithm for N=100 iterations and considered word nodes 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 ordered list of words, with (hopefully) negative polarity words at the top and positive polarity words at the bottom.

2.2 Data

2.2.1 English lexicon and Dutch translation

We took the English lexicon used in [4] as the starting point of our method. From this lexicon we extracted 2719 English words with positive polarity and 4913 words with negative polarity.

We used a free online translation service Google Translate[1] to translate positive and negative polarity words into Dutch, resulting in 974 and 1523 Dutch words, respectively. (We assumed that a word was translated successfully if the translation occurred in the lexicon of the Dutch Wordnet or in the Cornetto lexicon).

2.2.2 Dutch Wordnet and Cornetto

The Dutch Wordnet database we used in the experiments contained 70329 lexical units, 44115 synsets and 111741 relations between synsets.

From the Cornetto database we extracted 103734 lexical units, 70192 synsets, and 157679 relations between synsets.

3 Experiments and Results

3.1 Inter-annotator Agreement

We first conducted a small-scale pilot study to test our annotation guidelines (see Appendix A) and assess inter-annotator agreement. For this purpose we randomly selected 50 words each from the 1000 words scored most positive and most negative by our method using DWN and 100 iterations. The sample was evenly distributed with respect to the part of speech: verbs, nouns and adjectives.

The resulting list of 100 words was randomized and presented to two native Dutch speakers who independently assigned each word to one of five classes:

- strongly positive ($++$)
- weakly positive ($+$)
- neutral (0)
- weakly negative ($-$)
- strongly negative ($--$)

Cases where assessors were unable to assign a word to one of the classes, were separately marked with "?". This happened in three cases, which were excluded from subsequent analysis. Results are presented for the remaining 97 words.

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

Overall agreement between assessors is 54%, which we consider rather low. From Table 1 we see that there are no cases where one annotator assigned a positive and the other assigned a negative label. Thus, the decision of whether a word is positive or negative appears to be relatively easy for human annotators. Distinguishing neutral from weakly subjective words appears to be a more difficult task, 23% of the disagreement between annotators was in this category. This is, however, relative to a very large number of words in the class neutral. Almost the same amount of disagreement (21%) occurred between strongly and weakly subjective words.

	−−	−	0	+	++	Totals
−−	10	9	1			20
−	3	5	7			15
0		2	29	6		37
+			7	5	3	15
++			2	5	3	10
Totals	13	16	49	16	6	97

Table 1: Contingency table for two assessors assigning five classes. Empty cells mean zero.

We see that the distinction between strong/weak polarity is difficult to make for assessors. However, when the strong/weak distinction is dropped, the two assessors agree in 72% of the cases (see Table 2). Our assessment of agreement conforms to the "strict agreement" in [3]: there, the human agreement is 76.19% for adjectives and 62.35% for verbs; thus the agreement we achieve is comparable to the results for English.

	−	0	+	Totals
−	27	8		35
0	2	29	6	37
+		9	16	25
Totals	29	46	22	97

Table 2: Contingency table for two assessors when collapsing the two positive and negative classes respectively.

In the subsequent data collection for evaluation we decided to maintain the distinction between strongly and weakly positive and negative words respectively, in order to maintain the proportion of subjective and neutral words. With only three categories to choose from we would expect that more words would be assigned to the neutral class, while we aimed at a high level of recall of subjective words.

3.2 DWN vs Cornetto

To compare the quality of the subjectivity lexicons generated using two DWN and Cornetto, we ran our approach (section 2) using each resource separately for 100 iterations. The initial seed data set for both variants was the same, the automatically translated English subjectivity lexicon.

To evaluate the resulting rankings we took the 2,000 words judged most negative, and the 1,500 words judged the most positive by either variant, as well as the initial seed data. After duplicate removal and randomization this list of words was labeled by an independent annotator, who assigned each word one of the five class labels (see section 3.1).

We assess the quality of an automatic method by determining accuracy at the top-N words. In other words, we directly evaluate the quality of the lexicon by counting how many negative (resp., positive) words are in the top (resp., bottom) N words in the ranking produced by the algorithm. We also compare the accuracy of our algorithm to the accuracy of the initial seed list, which, except for noise introduced through the translation process, can be considered as a human-labeled standard.

Method	Accuracy (number of words)	Accuracy (percent)
Seed set	814	**53%**
DWN	918	46%
Cornetto	**1069**	**53%**

Table 3: Accuracy for the top 2000 negative words identified using DWN and Cornetto. For comparison, accuracy of the initial seed set is included (based on 1523 words overlapping with DWN/Cornetto). Best results are highlighted in bold.

Method	Accuracy (number of words)	Accuracy (percent)
Seed set	438	**45%**
DWN	423	28%
Cornetto	**522**	35%

Table 4: Accuracy for the top 1500 positive words identified using DWN and Cornetto. For comparison, accuracy of the initial seed set is included (based on 974 words overlapping with DWN/Cornetto). Best results are highlighted in bold.

Results for the subjectivity lexicons generated using DWN and Cornetto, as well as the accuracy of the initial seed data set are shown in tables 3 and 4. For both positive and negative words, using Cornetto instead of DWN results in a 7% increase in accuracy. The accuracy of the seed example ranges between 45% and 53%. For negative words, the run using Cornetto achieves the same accuracy, while it is lower for positive words.

For both positive and negative classes, the run based on Cornetto produces the most correctly labeled instances. It produces 151 more negative words than the run

using DWN, and 99 more positive words, which constitutes an 18%-22% relative improvement.

Interestingly, the top 2000 negative and top 1500 positive words produced by the runs based on DWN and Cornetto show relatively little overlap. Specifically, at the negative end, Cornetto identifies 565 negative words not contained in the top 2000 negative words produced using DWN, of which 395 words are not even present in the DWN. The lexicon produced using DWN contained 414 negative words that were not present in the top 2000 produced using Cornetto. At the positive end, 236 words produced using Cornetto are not contained in the top 1500 produced using DWN, of which 203 not in DWN at all. The top 1500 of the DWN run contained 137 words not produced by the Cornetto run.

The amount of overlap we find indicates that between 70%-86% of the improvement in accuracy results from Cornetto's larger lexicon. The remaining improvement (14%-30%) is judged to result from better ranking due to Cornetto's larger set of relations.

4 Conclusion

We have presented an algorithm that bootstraps a subjectivity lexicon from a list of initial seed examples. The algorithm considers a wordnet as a graph structure where similar concepts are connected by relations such as synonymy, hyponymy, 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 in the graph. After a specified number of iterations words are ranked according to their weight. Words at the top of this ranked list are assumed to be positive and words at the bottom of the list are assumed to be negative.

The algorithm was implemented and run using two different wordnets available for Dutch: DWN, and Cornetto. Cornetto is an extension of DWN, containing more words and more relations between these words.

We found that using Cornetto instead of DWN resulted in a 7% improvement of classification accuracy in the top-1500 positive words and in the top-2000 negative words. Between 70% to 86% of this improvement can be attributed to the larger size of Cornetto, the remaining improvement is attributed to the larger set of relations between words.

While the main focus of this report is on the comparison between use of the two resources DWN and Cornetto, a second important outcome of our work is a clean subjectivity lexicon for Dutch. In the future we will apply this resource to sentiment identification using existing approached and for the development of new approaches.

Appendix A: Subjectivity classification guidelines

You are asked to annotate text files, in which each line contains a word with its part-of-speech tag (v=verb, n=noun, a=adjective, b=adverb). You need to identify "positive", "negative" and "neutral" words. A word is "positive" in a specific context if it indicates positive emotions (e.g., "happy" in "I'm happy"), evaluations (e.g., "great" in "Great idea!") or positions (e.g., "supports" in "She supports the bill"). A word is "negative" if it indicates negative emotions ("I'm sad"), evaluations ("Bad idea!") or positions ("She opposed the bill"). If none of these apply, then a word is neutral.

When making the judgement try to guess the sentiment of the author or persons mentioned in the text, not your own associations with the word.

The task is to assign a label to a word based on how likely the word occurs in positive or negative contexts. Think of a few typical sentences with this word. Would the word be mostly positive, mostly negative, or mostly neutral?

You need to classify words by adding one of the following five labels at the end of each line:

$++$ indicates that the word is positive in most contexts

$+$ indicates that the word is positive in some contexts

0 indicates that the word is hardly ever positive or negative

$-$ indicates that the word is negative in some contexts

$--$ indicates that the word is negative in most contexts

Please classify each word. In case you are uncertain, pick the label that you think comes closest. If you do not know a word, place a "?".

References

[1] BANEA, C., MIHALCEA, R., AND WIEBE, J. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *LREC* (2008).

[2] ESULI, A., AND SEBASTIANI, F. Pageranking wordnet synsets: An application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (2007), pp. 424–431.

[3] KIM, S.-M., AND HOVY, E. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)* (2004).

[4] WILSON, T., WIEBE, J., AND HOFFMANN, P. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP 2005* (2005).