"Vreselijk mooi!" (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives.

Tom De Smedt, Walter Daelemans
CLiPS Computational Linguistics Group
Universiteit Antwerpen
Antwerpen, Belgium

tom@organisms.be, walter.daelemans@ua.ac.be

Abstract

We present a new open source subjectivity lexicon for Dutch adjectives. The lexicon is a dictionary of 1,100 adjectives that occur frequently in online product reviews, manually annotated with polarity strength, subjectivity and intensity, for each word sense. We discuss two machine learning methods (using distributional extraction and synset relations) to automatically expand the lexicon to 5,500 words. We evaluate the lexicon by comparing it to the user-given star rating of online product reviews. We show promising results in both in-domain and cross-domain evaluation. The lexicon is publicly available as part of the PATTERN software package (http://www.clips.ua.ac.be/pages/pattern).

Keywords: sentiment analysis, subjectivity lexicon, Dutch

1. Introduction

Textual information can be broadly categorized into two types: objective facts and subjective opinions (Liu, 2010). Opinions carry people's sentiments, appraisals and feelings toward the world. Before the World Wide Web, opinions were acquired by asking friends and families, or by opinion polls and surveys. Since, the online availability of opinionated text has grown substantially. Such a resource is interesting for marketing or sociological research. Sentiment Analysis (or Opinion Mining) is a field that in its more mature work focuses on two main approaches. The first approach is based on subjectivity lexicons (Taboada et al., 2011), dictionaries of words associated with a positive or negative sentiment score ("polarity"). Such lexicons can be used to classify sentences or phrases as subjective or objective, positive or negative. The second approach is by using machine learning text classification (Pang et al., 2002).

In this paper we describe a new open source subjectivity lexicon for Dutch adjectives, integrated with the main lexical semantic resource for Dutch: CORNETTO, an extension of the Dutch WordNet (Vossen et al., 2007). In Esuli & Sebastiani, (2006) it is noted that adverbs and adjectives are classified more frequently as subjective (40% and 36%) than verbs (11%). In our approach, we focus on adjectives, with possible expansion to other words in future research. We first extracted adjectives from online Dutch book reviews and manually annotated them for polarity, subjectivity and intensity strength (section 2.1). The results are described in section 2.2. We then experimented with two machine learning methods for expanding the initial lexicon: one semi-supervised (section 3.1) and one supervised (section 3.2). Each of the book reviews has an accompanying, user-given "star rating" (1–5), which we used to evaluate the lexicon (see section 4.1).

2. Manual Annotation

2.1. Assessment Procedure

As adjectives with high subjectivity will occur more frequently in text that expresses a sentiment or opinion (for example a customer product review or a fan movie review) we collected 14,000 online Dutch book reviews (bol.com), in which approximately 4,200 Dutch adjective forms occurred. The texts were mined with PATTERN (De Smedt & Daelemans, 2011) and part-of-speech tagged with FROG (Van den Bosch et al., 2007). We did not apply lemmatization at this stage and therefore some adjectives occur both in citation form and in inflected form, e.g., *goede* vs. *goed* (good). A small number of words is incorrectly tagged as adjective by FROG.

As shown in Figure 1, the adjective frequency in the book reviews approximates a Zipf distribution, with *goed* (good, 6380 occurrences) being the most frequent, followed by *echt* (real, 4682) and *heel* (very, 3632). The top 10% constitutes roughly 90% of all occurrences. We took the top 1,100 most frequent adjectives, i.e. all adjectives that occurred more than four times.

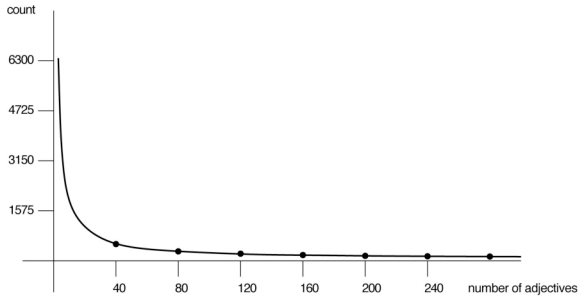

Figure 1: Frequency distribution of adjectives in Dutch book reviews.

Seven human annotators were presented with the list in random order and asked to classify each adjective in terms of positive-negative polarity and subjectivity. In Esuli & Sebastiani (2006), adjectives are not only classified in terms of polarity (i.e., positive or negative sentiment) but also in terms of subjectivity (i.e., objective vs. subjective). They also classified adjectives per sense, as different senses of the same term may have different opinion-related properties, as in crazy–*insane* (negative) vs. *crazy*–enamored (positive).

We use a similar approach where annotators were asked to assess word senses using a triangle representation (Figure 2), where the horizontal axis represents grades of positive-negative polarity and the vertical axis represents the objectivity-subjectivity strength. This approach entails that more positive/negative adjectives are also more subjective. But not all subjective adjectives are necessarily positive or negative: e.g., *blijkbaar* (apparently). We used CORNETTO to retrieve the different senses of each adjective (97% of our list occurs in CORNETTO). To our knowledge, this is the first Dutch subjectivity lexicon that applies sense discrimination.

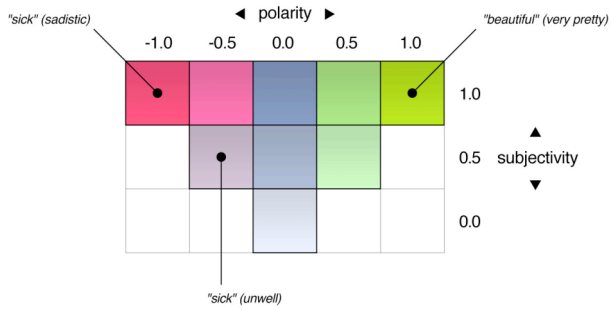

Figure 2: Triangle representation with polarity and subjectivity axes.

Dutch adjectives can be used as adverbs, where in English the ending -ly is usually required. For example: *ongelooflijk goed*, which is understood as *incredibly good* and not as *unbelievable* + *good*. The sentiment expressed is very positive, whereas *ongelooflijk–unbelievable* in itself could also be negative. To represent this, annotators where asked to provide an additional "intensity" value, which can be used as a multiplier for the successive adjective's polarity.

To summarize, we classified adjectives per word sense using grades on a Polarity/Subjectivity/Intensity-scale. The intensity strength may be useful to map Dutch adjectives to English adverbs.

2.2. Annotators

The data was annotated by the two researchers, four graduate linguistics students, and one graduate art student specialized in typography. All of them are native speakers of Dutch.

2.3. Annotator Agreement

In total, 1,100 adjectives (1,584 unique word senses) were assessed by all annotators, meaning 7 votes for polarity, subjectivity and intensity for each word sense. We removed a number of inflected adjectives, spelling errors and adverbs, bringing the final **`gold1000`** lexicon to 1,044 adjectives (1,526 word senses) with the average scores of the 7 annotators. We manually corrected the spelling of 12 words (e.g., *sexueel* => *seksueel*, *poetisch* => *poëtisch*). The lexicon contains 740 positive assessments (48%), 544 negative (36%) and 242 neutral (16%). Figure 3 shows a breakdown of the distribution. Table 1 shows the inter-annotator agreement calculated using Fleiss' kappa, which measures reliability in ratings given by different voters. We attain the highest agreement for positive-neutral-negative polarity without considering polarity strength (κ=0.63). Assessment of subjectivity strength is shown to be a much harder task (κ=0.34), perhaps because the task is more vague than classifying positive vs. negative.

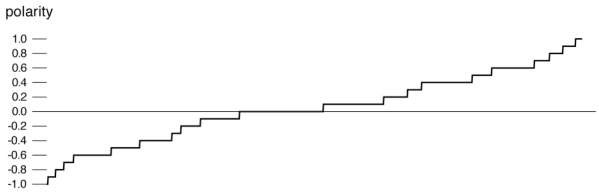

Figure 3: Distribution of positive-negative polarity strength in `gold1000`.

Rating	***K***
polarity (`-1` or `0` or `+1`)	+0.63
polarity (`-1.0` => `+1.0`)	+0.47
polarity + subjectivity	+0.30
subjectivity	+0.34
intensity	+0.32

Table 1: Agreement for 7 annotators.

2.4. Comparison to the DUOMAN Lexicon

In Jijkoun & Hofmann (2009) a subjectivity lexicon is used (DUOMAN), containing 5,276 words with `++`, `+`, `0`, `-`, `--` polarity assessments by two annotators. 47% of the adjectives in `gold1000` also occur in DUOMAN. We compared the positive-neutral-negative polarity (without strength) of the adjectives in `gold1000`, using the average of word senses, to those that also occur in DUOMAN. Agreement is κ=82%.

27 adjectives[1] are positive in `gold1000` but negative in DUOMAN, or vice-versa. One explanation is the different semantics between Dutch in the Netherlands and Dutch in Flanders for some adjectives (e.g., *maf*). Agreement increases to κ=93% when the aberrant 27 adjectives are omitted from the measurement.

[1] For completeness, the 27 adjectives are: *apart* (separate), *breed* (broad), *diep* (deep), *droog* (dry), *eindeloos* (endless), *extreem* (extreme), *fanatiek* (fanatical), *gek* (crazy), *gewend* (accustomed), *gewoon* (ordinary), *grenzeloos* (boundless), *groen* (green), *ironisch* (ironic), *laaiend* (blazing), *maf* (slack), *naast* (next), *onbeschrijfelijk* (indescribable), *onconventioneel* (unconventional), *ondeugend* (naughty), *ongelooflijk* (incredible), *ontzettend* (tremendous), *onvoorspelbaar* (unpredictable), *raadselachtig* (enigmatic), *sarcastisch* (sarcastic), *sober* (sober), *strak* (tight), *waanzinnig* (insane).

3. Automatic Expansion

3.1. Using Distributional Extraction

There is a well-known approach in computational linguistics in which semantic relatedness between words is extracted from distributional information (see e.g. Van de Cruys 2010 for an example for Dutch). The method uses a vector space model with adjectives as vectors (i.e., matrix rows) and nouns as vector features (i.e., matrix columns). The value for each vector feature represents the frequency an adjective precedes a noun. It is then possible to take the cosine of the angle between two vectors as a measure of similarity (cosine similarity). In other words, adjectives followed by the same nouns are more semantically related than adjectives followed by different adjectives. The method then uses dimensionality reduction and clustering by cosine distance to create groups of semantically related words.

We applied this approach to automatically annotate new adjectives, based on their semantic relatedness to `gold1000` adjectives. From the TWNC (Ordelman et al., 2002), we analyzed 3,000,000 words and selected the top 2,500 most frequent nouns. For each adjective in TWNC that is also in the CORNETTO database, we counted the number of times it directly precedes one or more of these top nouns, resulting in 5,784 adjective vectors with 2,500 vector features. For each `gold1000` adjective we then used cosine similarity to retrieve the top 20 most similar nearest neighbors. For *fantastisch* (fantastic) the top five nearest neighbors are: *geweldig* (great, 70%), *prachtig* (beautiful, 51%), *uitstekend* (excellent, 50%), *prima* (fine, 50%), *mooi* (nice, 49%) and *goed* (good, 47%).

The best nearest neighbors for each `gold1000` adjective were then handpicked by 2 annotators in order to reduce antonymy (e.g., *fantastic* vs. *horrible*) and noise (e.g., *fantastic* vs. *electoral*), and for word-sense disambiguation (e.g., *fantastic* is sentimentally related to *great* in the sense of *wonderful*, but not in the sense of *big*). These nearest neighbors inherit polarity, subjectivity and intensity from their `gold1000` parent.

In the **`auto3000`** lexicon (consisting of the `gold1000` adjectives + 2,077 selected nearest neighbors) we added a "reliability" strength, where reliability=1.0 identifies a manually annotated adjective, or an automatically annotated adjective selected by both annotators, which appears in the same CORNETTO synset as its `gold1000` parent. Adjectives with reliability=0.9 were selected by only one annotator but also appear in the same CORNETTO synset as their parent. Adjectives with reliability=0.8 matched their parent by CORNETTO description. For example: for *uitzonderlijk* the CORNETTO description is *exceptioneel*, and for *exceptioneel* it is *uitzonderlijk* – even though they may not appear in the same synset.

Adjectives with reliability=0.7 inherit their polarity, subjectivity and intensity from the average of all their selected parents. For example, *schalks* (roguish) was selected as nearest neighbor for both *ondeugend* (naughty, polarity +0.14) and *ironisch* (ironic, polarity +0.21) so its polarity strength is (0.14 + 0.21) / 2 = +0.16.

3.2. Using CORNETTO Relations

The `auto3000` lexicon contains 3,121 adjectives (3,713 word senses). In the **`auto5500`** lexicon we iteratively extend the lexicon by traversing relations between CORNETTO synsets. In CORNETTO or in WORDNET, word senses with a similar meaning are grouped in synsets (i.e., concepts). Different synsets are related to each other by synonymy (same-as), antonymy (opposite-of), hyponymy (type-of), etc. For each `auto3000` adjective word sense, we take its CORNETTO synset and inherit its polarity, subjectivity and intensity to related word senses, decreasing reliability by a factor *p* as follows: p=0.7 for word senses in the same synset, p=0.6 for synonyms and antonyms (where polarity is reversed) in other synsets, and p=0.5 for near-synonyms in other synsets.

In three iterations we can spread out to 2,286 new adjectives (2,962 word senses) before reliability is lower than 0.1. The `auto5500` lexicon then contains 5,407 adjectives (6,675 word senses). Figure 4 shows a sample entry in the `auto5500` lexicon, in XML-format.

```
<sentiment language="nl">
  <word form="razend"
        sense="geweldig"
        cornetto_id="r_a-14522"
        wordnet_id="a-01387319"
        polarity="0.8"
        subjectivity="1.0"
        intensity="1.9"
        reliability="1.0" />
</sentiment>
```

Figure 4: `auto5500` sample in XML-format.

4. Evaluation

4.1. Dutch Product Reviews

For evaluation we tested with a set of 2,000 Dutch book reviews, which were not used to extract the initial `gold1000` lexicon, and which were evenly distributed over negative opinion (star rating 1 and 2) and positive opinion (4 and 5). For each review, we then scored polarity for those adjectives that occur in the lexicon and compared the average strength to the original star rating. We took polarity >= +0.1 as a positive observation and polarity < +0.1 as a negative observation. This is a form of binary classification in which there will be a number of correctly classified words (true positives and negatives) and incorrectly classified words (false positives and negatives) by which we can calculate precision (`TP/TP+FP`) and recall (`TP/TP+FN`). The polarity threshold can be lowered or raised, but +0.1 yields the best results. Overall we attain a precision of 0.72 and a recall of 0.78 (15% `FP` and 11% `FN`).

In a second run we also used the intensity strength. Instead of scoring *echt teleurgesteld* (truly disappointed) as *echt* (true) + *teleurgesteld* (disappointed) = 0.2 + -0.4 = -0.2, we now used *echt* (intensity 1.6) x *teleurgesteld* = 1.6 x -0.4 = -0.64. This increases recall to 0.82.

To provide a cross-domain measurement we repeated the test with a new set of 2,000 Dutch music CD reviews evenly distributed by star rating. Here we attain a precision of 0.70 and a recall of 0.77.

The results for the `gold1000` and `auto3000` lexicons are near identical. The reason that precision and recall do not increase by adding more adjectives is that 90% of top frequent adjectives is already covered in `gold1000`, adding more words has a minimal coverage effect. For `auto5500`, F1-score is less (-2%). The reason for this is that the adjectives *een* (united), *in* (hip) and *op* (exhausted) are part of the expanded lexicon. These words can also function as a common determiner (*een* = a/an) and common prepositions (in = in, op = on) in Dutch. Without these three, the scores for `auto5500` come close to the results for `gold1000` and `auto3000`. This suggests that the automatic expansion can benefit from a manual correction, and that prediction in general can benefit from part-of-speech tagging.

positive >= 0.1	**Books.2000**				
	# adj.	A	P	R	F1
`gold1000`	794	0.75	0.72	0.82	0.77
`auto3000`	1,085	0.75	0.72	0.82	0.77
`auto5500`	1,286	0.74	0.72	0.79	0.75

positive >= 0.1	**Music.2000**				
	# adj.	A	P	R	F1
`gold1000`	480	0.72	0.69	0.78	0.73
`auto3000`	613	0.72	0.70	0.77	0.73
`auto5500`	678	0.71	0.70	0.71	0.71

Table 2: Number of unique adjectives rated, accuracy, precision, recall and F1-scores for opinion prediction.

We also experimented with a simple algorithm for negation. We looked for the words *niet* (not), *nooit* (never) and *geen* (none). Reversing the polarity of the successive adjective can be used to raise precision by 2-3%.

4.2. English Product Reviews

Many word senses in the CORNETTO database have inter-language relations to WORDNET. For example, the Dutch adjective *briljant* has an `is-near-synonym` relation to the English synset containing the adjectives *bright*, *brilliant* and *vivid*. We took advantage of this to map the polarity and subjectivity scores in the Dutch lexicon to an English lexicon. We then tested our English lexicon against Pang & Lee's `polarity dataset v2.0` containing 1,000 positive and 1,000 negative IMDb movie reviews (imdb.com).

82% of word forms in our `gold1000` lexicon have an `is-near-synonym` relation. The best results are attained by mapping them to the first (main) synonym in the related English synset (e.g., *briljant* = *bright*). However, if we look at the 1,000 top frequent adjectives in 3,500 random English IMDb movie reviews, only 32% overlaps with the Dutch most frequent adjectives. Initial test results were therefore poor: a precision of 0.66 and a recall of 0.54.

We then manually annotated 560 frequent English adjectives (1,643 word senses). This was done by a single annotator, but the effect is apparent: precision increases to 0.72 and recall to 0.71.

positive >= 0.1	**Movies.2000 (Pang&Lee)**				
	# adj.	A	P	R	F1
`english1250`	1,121	0.72	0.72	0.71	0.72

Table 3. Number of unique adjectives rated, accuracy, precision, recall and F1-scores for English.

4.3. Analysis of the Dutch Test Data

Overall in 4.1, positive predictions (57%) were made more frequently than negative predictions (43%). We offer three potential reasons, by examining the test data, as to why it may be harder to identify negative opinions:

- **Comparison**: some negative opinions defend their viewpoint by referring to other instances, for example: "*dat boek was grappig, origineel, pakkend, maar dit boek vond ik op al die punten tegenvallen*" (that book was funny, inspiring, moving, but this book fails on all those points). All the adjectives rate as positive but the review is negative.
- **Feature-based opinions**: in "*de eerste tien pagina's zijn sterk, maar dan zakt het als een pudding in elkaar*" (the first ten pages are quite good, but it collapses in ruins from there) the positive opinion accounts for a specific feature of the book (first ten pages), while the general negative opinion is carried by a figure of speech (*to collapse in ruins*).
- **Sarcasm**: for example, "*zou niet weten wat ik met mijn leven moest als ik dit geweldige boek gemist had*" (wouldn't know what to do with my life if I had missed this awesome book).

Other indicators of opinion we encountered include interjections such as *bwah* (meh) and *tjongejonge* (boy oh boy), and subjective verbs (*to struggle*) and nouns (*turn-off*, *abomination*).

5. Conclusions and Future Work

We annotated a compact lexicon of Dutch adjectives frequently found in opinionated text. The lexicon has been released with a PDDL public domain license as part of the PATTERN[2] software package. Overall, we judge the results to be useful, and a good basis for more robust opinion prediction systems. The distributional approach for automatic expansion described in 3.1 worked well, as illustrated by the figures in Table 2. Automatic expansion by exploiting the relations in lexical databases such as CORNETTO and WORDNET slightly lowers the results. Manual corrections can be beneficial to this approach. In future work it may be interesting to adopt a similar approach for frequent Dutch nouns and verbs.

It is possible to translate the work to other languages (e.g., English) using the inter-language relations in CORNETTO. The usefulness of this approach depends on how many frequent adjectives are covered in the target language.

[2] http://www.clips.ua.ac.be/pages/pattern

6. Acknowledgements

We would like to thank the annotators of the Dutch lexicon: Tim De Cort, Nic De Houwer, Lore Douws, Aaricia Sobrino Fernández, Kimberly Ivens.

7. References

Esuli, A., Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of LREC 2006*, pp. 417--422.

Jijkoun, V., Hofmann, K. (2009). Generating a Non-English Subjectivity Lexicon: Relations That Matter. *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 398--405.

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, Second Edition.

Ordelman, R., de Jong, F., van Hessen, A., Hondorp, H. (2007). TwNC: a Multifaceted Dutch News Corpus. *ELRA Newsletter*, 12(3-4).

Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 79--86.

Pang, B., Lee, L., (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the Association for Computational Linguistics (ACL 2004)*.

Taboada, M., Brooks, J., Tofiloski, M., Voll, K., Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), pp. 267--307.

Van de Cruys, T. (2010). Mining for Meaning: The Extraction of Lexico-semantic Knowledge from Text. *Groningen Dissertations in Linguistics*, 82.

Van den Bosch, A., Busser, B., Canisius, S., Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pp. 99--114.

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

.