Different Approaches to Automatic Polarity Annotation at Synset Level

Isa Maks, Piek Vossen

Vu University, Faculty of Arts
De Boelelaan 1105, 1081 HV Amsterdam
E-mail: e.maks@let.vu.nl, p.vossen@let.vu.nl

Abstract

In this paper we explore two approaches for the automatic annotation of polarity (positive, negative and neutral) of adjective synsets in Dutch. Both approaches focus on the creation of a Dutch polarity lexicon at word sense level using wordnet as a lexical resource. The first method is based upon the simple transfer of an English sentiment lexicon (Sentiwordnet 1.0) into Dutch. The second approach regards the use of a wordnet based propagation algorithm with different settings with respect to the quality and length of the seed lists. Results are validated against manually compiled gold standards and compared with results of similar approaches generating polarity lexicons for English.

1. Introduction

The automatic extraction of opinions, emotions, and sentiments in text to support applications such as product, hotel and film review mining, analysis of opinionated text like news, forum posts, and blogs is an active area of research in natural language processing. Many approaches to opinion and sentiment analysis rely on lexicons or lists of words that may be used to express sentiment. Knowing the polarity (positive, negative or neutral) of these words helps a system recognize the positive and negative sentiments in these sentences.
Many subjectivity lexicons are compiled as lists of keywords, rather than word meanings. However, words may have positive, negative and neutral meanings (cf. ex. (1a) and ex. (1b)) which may cause major errors if incorrectly tagged in the applications they are used in.

Ex. (1) wreed (cool, cruel)
(a) een wrede despoot *a cruel tyrant*
(b) ze rijden daar in vet wrede auto's rond *They drive around in really cool cars*

The example shows that the Dutch word *wreed* has two different meanings (properly translated into *cruel* and *cool* , respectively), with opposite (negative and postive) polarity.
Most studies, nowadays, recognize the importance of sentiment scores at meaning level (Esuli and Sebastiani (2006), Andreesvkaia and Bergler (2006), Wiebe and Mihalcea (2006), Su and Markert (2008). Although these approaches are widely used in English, little is known about how they perform at synset level as opposed to word level. More recently, a number of approaches have been tested to build subjectivity lexicons at synset level (Gyamfi et al. (2009); Su et al. (2009)). They focus, however, on subjectivity classification, a task that slighly differs from ours, as it aims at the classification of word senses as subjective or objective.
For Dutch, the only existing polarity lexicon - to our knowledge - is built by Jijkoun and Hofmann (2009). Their approach is, like ours, wordnet based, but produced a list of words (instead of synsets).
In this paper, we focus on the creation of subjectivity lexicons at word sense level using wordnet as a lexical resource where word senses are organized in synsets.
We explore two methods for polarity annotation of Dutch adjective wordnet entries, leaving the nouns and verbs for future work. The first method relies on the transfer of polarity values from an English sentiment lexicon, Sentiwordnet 1.0 (Esuli and Sebastiani (2006)) to the Dutch Wordnet.
The second approach consists of the implementation of a propagation algorithm that starts with a seed list of synsets of known sentiment and sends polarity through the wordnet making use of its lexical relations. Experiments with different seed lists are performed : the General Inquirer word list (Stone et al., 1966) translated into Dutch, and two different manually compiled synset lists following a method that might be used when no manually compiled seed lists exist.

The remainder of this paper is organised as follows. In the next session we briefly discuss the lexical resources and gold standards referred to in this paper. Sections 3, 4 and 5 present the two different approaches to polarity annotations and their results. In Section 6 the results are compared with other studies.

2. Descriptions of Lexical Resources and Gold Standards

2.1 Lexical resources

- **Dutch**

We make use of two lexical resources for Dutch: the Dutch Wordnet and the Dutch Reference Lexicon which both are part of the Cornetto database (Vossen et al. 2008). The two combined resources have different semantic organisations: the Dutch Wordnet has, like the Princeton Wordnet, a synset organisation and the Dutch Reference Lexicon is organised in form-meaning composites or lexical units. The description of the lexical units includes definitions, usage constraints, selectional restrictions, syntactic behaviours, illustrative contexts, etc. Within the Cornetto Database, each synonym in a synset is linked to the corresponding lexical unit of the Dutch Reference Lexicon. Synsets are linked by translation equivalents links to the Princeton Wordnet (versions 2.0 and 3.0); these translation links have been derived automatically and are then manually corrected.
The Cornetto database is semi-automatically compiled and manually corected afterwards. As the manual correction is still in progress , the status of the synsets with regard to the number of lexical relations like hyponyms, near-synonyms, hypernyms and antonyms (LR) and/or translation equivalent links (Equi) may differ. Table 1 presents the statistics of the adjective part of Cornetto. Part ADJ1 consists of 3,616 synsets which have both lexical and translation equivalent relations; Part ADJ2 consists of 2,109 synsets which have translation equivalent relations only; part ADJ3 consists of 733 synsets which have lexical relations only ; and part ADJ4 consists of synsets lacking both lexical and translation equivalent relations.

	Synset	LR	Equi
ADJ1	3,616	+	+
ADJ2	2,109	-	+
ADJ3	733	+	-
ADJ4	1,440	-	-
Totals	7,898	4,349	5,725

Table 1. Number of Adjective Synsets and Lexical Units in Cornetto (situation 2010)

Because of the different stages of elaboration of the synsets, the two approaches discussed in this paper are relevant for the Dutch wordnet as they may complement each other. Synsets that have translation equivalent links to the English wordnet are covered by the transfer approach and synsets that have lexical relations are covered by the propagation method.

2.2 Gold Standards

- **Dutch**

For the evaluation of the results for Dutch we use the gold standard developed by Maks and Vossen (2010b). The gold standard includes annotations for subjectivity (subjective vs. objective), attitude holder (SpeakerWriter or AgentExperiencer) and polarity (positive/negative/ neutral). Only the latter category will be used in this study. We use the synset level variant of the gold standard which includes 512 synsets (gs-ss-512).
Reported inter-annotator agreement for polarity, is 86.3% with a Cohen kappa (κ) of 0.80. The polarity annotations are distributed as follows: 37% negative, 35% positive, 28 % neutral.
In section 6 we refer to a word level gold standard for Dutch (w-1916) compiled by Jijkoun and Hoffman (2009), which consists of 1916 words annotated by two annotators with positive (50%), negative (29%) and neutral (21%) polarity. Interannotator agreement is 76% (κ=0.62); we use a version where disagreements are abjudicated by a third annotator.

- **English**

For English, the Micro-WNOp corpus (Cerini et al., 2007) is used as a gold standard to evaluate Sentiwordnet. The Micro-WNOp corpus is a – publicly available - list of about 1000 WordNet synsets (285 adjective synsets) annotated with polarity values. The raters manually assigned a duplet of numerical scores to each synset which represent the strength of positivity and negativity, respectively. Thus, a synset could have a non-zero rating on both negativity and positivity. The gold standard does not provide a abjudicated judgment for each synset but the lists with judgments by all different annotators can be downloaded. The gold standard consists of 285 adjective synsets divided into three groups: a common part of 29 adjective synsets with one abjudicated annotation judgment; group 1 consisting of 147 synsets with 2 annotation judgments for each synset and group2 consisting of 138 synsets with 3 annotation judgments for each synset.
For our purposes, we converted the numerical scores to categorical ones (positive , negative and neutral) by assigning 'positive' to synsets where the positive score is larger than the negative score and 'negative' where the negative score is larger than the positive one. The rest of the synsets (i.e. where the positive and negative scores are equal, including zero) is considered 'neutral'. We then derived one judgment for each adjective synset when there was agreement between at least two annotators. The remaining 12 synsets on which all (2 or 3) annotators disagreed were eliminated from the gold standard. Thus,

the final 'simplified and categorical' gold standard which will be called WNO-273 in the remainder of this paper, consists of 273 synsets (78 negative; 70 neutral, 125 positive).

3. Method I: Sentiwordnet translated

The transfer of Sentiwordnet 1.0 to Dutch consists of the copying of the sentiment values from the English synsets to the Dutch synsets through the translation equivalents which exist between the English and the Dutch wordnet. We evaluate the English and the Dutch version and compare the results.

3.1 Method I

Sentiwordnet1.0 (Esuli and Sebastiani (2006)) is a resource with automatically determined polarity of word senses in WordNet produced via bootstrapping from a small manually compiled seed set. Each synset has two scores assigned, representing the positive, and negative polarity. The method used to develop Sentiwordnet is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification.
Table 2 shows the statistics of the adjective part of the English Sentiwordnet (SWN) in relation to the adjective part of the Dutch Wordnet (DWN). As can be seen from the first row, the English Sentiwordnet (18,563 synsets) is considerable larger than the Dutch wordnet (7,898 synsets). The wordnets are connected to each other by 17,754 translation equivalent links. Dutch translated synsets have an average of 3.1 translation links per synset.

Adjectives	**SWN**	**DWN**
number of synsets	18,563	7,898
translated synsets	8,217	5,725
equivalent links	17,754	17,754

Table 2. Statistics DWN and SWN

The transfer of the polarity values from the English to the Dutch wordnet consists of the following steps: (1) Copy the set of sentiment scores (positive and negative) from a SWN synset into the equivalent Dutch synsets (2) Calculate one set of scores for each DWN synset by counting up the positive scores and negative scores, respectively. As can be seen from table 2, many Dutch synsets have more than one translation equivalent which results in multiple sets of scores per synset. (3) Translate the two accumulated scores into one categorical value by attributing positive value if the positive_score is larger than the negative_score, and negative value if the positive_score is smaller than the negative_score. A synset is considered neutral if both scores are equal (being zero or larger than zero). (4) Assign neutral polarity to all synsets that are not covered by the transfer method , i.e. all synsets that do not have translation links with the English Wordnet.
To be able to compare the quality of the source and the target lexicons, Sentiwordnet1.0 was evaluated against the 'simplified' WNO-273 (cf. section 2.2). Different versions of Sentiwordnet have already been evaluated against Micro-WNOp by other studies (Baccianella et al. (2010)), but these evaluations use scalar values. For the present study, we converted the numerical scores of Sentiwordnet into categorical ones by applying the same rules as described above for the conversion of Micro-WNOp's numerical values.

3.2 Method I: Results and Discussion

The results of the transfer are presented in the following table. The first column (name) gives the name of the lexicon, e.g. SWN for the English Sentiwordnet and DSWN for the derived Dutch Sentiwordnet. The second column (gs) gives the gold standard against which the results are evaluated. The third column gives precision (P), recall(R) and weighted average (F) for all polarity (pol) categories together and for each one separately. By default, all other synsets are considered neutral and evaluated as such.

name	gs	pol	P	R	F
SWN (eng)	WNO-273	**All**	**0.62**	**0.62**	**0.62**
		POS	0.72	0.70	0.71
		NEG	0.58	0.63	0.60
		NTR	0.48	0.47	0.47
DSWN (dut)	ss-512	**All**	**0.58**	**0.58**	**0.58**
		POS	0.58	0.64	0.61
		NEG	0.61	0.61	0.61
		NTR	0.54	0.47	0.50
SWN-retro	WNO-273	**All**	**0.67**	**0.67**	**0.67**
		POS	0.74	0.85	0.79
		NEG	0.64	0.72	0.67
		NTR	0.54	0.31	0.40

Table 3. Evaluation Results English and Dutch Sentiwordnet

When comparing the scores of the source Sentiwordnet1.0 and the target Dutch resource, we see that overall performance drops with 4% precision (from 62% to 58%). Interestingly, however, precision scores of individual categories may also rise (cf. negative polarity which rises from 58% to 61%).
A closer look at the data shows that different factors affect the outcome. Conceivably, a substantial number of the errors may be due to incorrect annotations in the source lexicon. One single incorrect annotation in the source lexicon can affect large quantities of synsets in the target lexicon if they have many translation equivalent links. For example, more than ten Dutch synsets have a translation equivalent link with [comfy#a#1 comfortable#a#1] which is incorrectly tagged as 'negative'.
However, the transfer method has also positive

side-effects: if a word sense has many translation equivalents, incorrect annotations may be solved by correct ones. For example, *behulpzaam* (helpful) has 3 related English synsets which are correctly tagged 'positive' and one synset that is incorrectly tagged as 'neutral' (nice#a#7). The 'neutral' *nice* will in this case be overruled by the correct polarity values of the other synsets. The following experiment shows how powerful this multiple translation effect can be. We transferred the derived Dutch sentiwordnet back into English and replaced the scores of the translated English synsets (i.e. 8,217 synsets, cf. Table 2) with the newly obtained scores. Table 3 (SWN-retro) shows that both overall performance (from 0.62 to 0.67) and precision rates for each polarity category (from 0.72 to 0.74, from 0.58 to 0.64 and from 0.48 to 0.54 for positive, negative and neutral polarity, respectively) increase.

Finally, also the quality of the translation equivalent links show impact on the results. As the automatically generated translation equivalent links between the Dutch and the English Wordnet are not yet all manually corrected, the Dutch Wordnet consists of synsets with high quality – manually corrected – links and synsets automatically derived links. We divided the gold standard in synsets with manually correct links (202 items) and synsets with automatically derived links (303 items), and measured performance on the Dutch gold standard.We obtain 0.60 for the manually corrected items and 0.56 for automatically derived items which leads to the conclusion the quality of the derived Dutch sentiwordnet will increase when all translation links are manually corrected.

3.3 Method I: Conclusion

It seems that the transfer of coarse-grained sentiment like positive and negative polarity between wordnets of different languages can be done in a reliable manner, since the decrease in performance – after transfer - is rather low with 4%. Important factors that bear effect on the outcome are the quality of the source lexicon and the quality of the translation links.

Moreover, as demonstrated by translating the lexicons back and forth, the transfer process not only worsens but also improves the polarity scores.

4. Method II: Seed propagation

The seed propagation approach relies on the assumption that the concepts that are represented by synsets that are closely related by semantic links, have similar meaning and thus similar sentiment. Many versions of this approach have been implemented for English (Andreevskaia and Bergler (2006), Esuli and Sebastiani (2006)).

Also for Dutch a similar approach has been used by Jijkoun and Hofmann (2009). They generated, however, a word level polarity lexicon whereas our approach is aimed at generating a synset level lexicon.

4.1 Method II

We start with a set of seed synsets of known polarity (positive, negative and neutral) which is propagated through the wordnet making use of the lexical relations between synsets. The synset seed list is augmented during each iteration by adding near-synonym, antonym, hyponym and hypernym synsets. After each iteration the augmented list is used as seed list for the next step until convergence is achieved and no new synsets are added to the result list. The synsets that not have been added to the result list are considered 'neutral'. We did several experiments varying the type of lexical relations, the number of iterations, and the size and the composition of the seed list.

4.2 Seedlist Composition and Size

Andreevskaia and Bergler (2006) showed that the composition of the seed list has a considerable impact on the performance of the system. They did 58 runs of their sentiment tagging system on unique non-intersecting seed lists and found that the accuracy ranged from 47.6% to 87.5%. They attribute these variation to the fact that the used seed lists consisted of words, and not synsets or word senses, and that several words have both neutral and sentiment laden meanings whereas only one of them is included.

We think, however, that this is not the only reason for the variation, but that also the size and composition of the seed list are of considerable importance. To test this, we did experiments with three different seed lists: a high quality one, a low quality one and a large one of mixed quality.

- a 'high quality' seed synset list (sds-HQ)

Our hypothesis is that a carefully selected list of seed synsets taking into account the number of lexical relations (synonyms, near_synonyms, antonyms, hyponyms) with other members, may produce better results than a randomly chosen seed list. A large number of semantic ties with other members in the field proves that the involved synsets represent sentiment bearing concepts that are central and prototypical (Andreevskaia and Bergler (2006)). Thus, core members are identifiable in a wordnet by the number of lexical relations (LRs) links they have. This assumption is confirmed by the fact that typical evaluative sentiment bearing words have many synonym links as they tend to group together in large synsets, as shown by Maks and Vossen (2010a). A 'high quality' seed synset list is composed, as follows: (1) select 250 adjective synsets with more than 8 LRs (2) annotate this list manually with positive, negative and neutral polarity and (3) exclude synsets that have synonyms with mixed – positive, negative and/or neutral polarity – members as they produce noise because of their ambiguity.

- a 'low quality' seed synset list (sds-LQ)

A seed list of equal size but 'low' quality is composed. This list includes 250 synsets which have less than 3 LRs.

- a large seed synset list (sds-GI)

To complete the experiment we produced a large seed synset list of mixed quality. We use the General Inquirer Lexicon (Stone et al., 1966) as the starting point for this seed list. The list consists of 2,558 unique adjective words with neutral (1,203), negative (800) or neutral (771) polarity. We then use the online Google translation service to translate this list of words into Dutch. The seed words are related to the appropriate synsets. This procedure results in 1,411 labeled Dutch synsets, (428 neutral, 422 positive and 561 negative) 'of mixed quality'. The list includes both low quality seeds with less than 3 LRs (315 synsets) and high quality seeds with more 8 LRs (322 synsets).

seeds	ss-512	ss-complement	
sds-HQ	0.69	0.65	414 synsets
sds-LQ	0.55	0.55	498 synsets
sds-GI	0.75	0.67	236 synsets

Table 4 propagation with different seed lists

Table 4 shows the results obtained after propagation of the seed lists through the wordnet. The results have been evaluated against the complete gold standard (column ss-512) and against reduced versions of the gold standard from which the intersection between gold standard and seed list is removed resulting in 3 different test sets of 498, 414 and 236 items respectively (cf. Column ss-complement). By doing both evaluations we know if scores are due to larger overlaps of manually annotated seed list items and gold standard items or if they may be ascribed to the quality of the seed list.
The scores confirm our hypothesis that the number of LRs is indicative for the performance: the sds-HQ scores better than sds-LQ on both the full test set and the reduced version (0.69 vs. 0.65 and 0.65 vs. 0.55, respectively). However, the large seed list (sds-GI) performs even better and outperforms the high quality list on both versions of the test set. The fact that sds-GI scores better than sds-HQ even on the reduced version (0.67 vs. 0.65), suggests that the number of seeds might be even more important than the quality.
For further experiments with the propagation algorithm (cf. following sections) we use the sds-GI as it is the best scoring seed list.

4.3 Polarity Values

The performance of the propagation algorithm differs with regard to the different polarity categories (cf. Table 5.

Seeds	Gs	pol	P	R	F
sdsGI	ss-512	**All**	**0.75**	**0.75**	**0.75**
		POS	0.78	0.76	0.77
		NEG	0.76	0.82	0.79
		NTR	0.72	0.68	0.70

Table 5: Performance of different polarity categories

The scores of the neutral items, especially recall, are lower than those of the sentiment laden items. This is probably due to the fact that, although the number of seeds is almost equal for the different polarity categories, neutral items have less quality (cf. previous section) than the other categories as they have fewer lexical relations. Table 6 shows that 428 neutral seeds have an average of 2.5 synonyms (column SYN) and 2.5 other lexical relations (column SAHH: near-synonyms, antonyms, hypernyms and hyponyms) per synset whereas the negative and positive seeds have an average of 3.4 to 3.6 for both.

Pol	nr of seed synsets	SAHH	SYN
POS	422	3.4	3.4
NEG	560	3.6	3.5
NTR	428	2.5	2.5

Table 6 Average of LRs per synset

4.4 Number of Iterations

We experimented with the number of iterations (i) given in the first column of Table 7. Best balance between precision and recall is achieved with 5 iterations. With 10 iterations convergence is achieved. It is this last setting that is used throughout this paper.

i	Gs	Pol	P	R	F
0	-ss-512	**All**	**0.64**	**0.64**	**0.64**
		POS	0.87	0.51	0.64
		NEG	0.81	0.58	0.67
		NTR	0.49	0.85	0.62
1		**All**	**0.73**	**0.73**	**0.73**
		POS	0.81	0.70	0.75
		NEG	0.78	0.74	0.76
		NTR	0.63	0.74	0.68
5		**All**	**0.76**	**0.76**	**0.76**
		POS	0.78	0.74	0.76
		NEG	0.77	0.81	0.79
		NTR	0.71	0.70	0.71
10		**All**	**0.75**	**0.75**	**0.75**
		POS	0.78	0.76	0.77
		NEG	0.76	0.82	0.79
		NTR	0.72	0.68	0.70

Table 7: Various numbers of iterations (I)

With each iteration, recall increases while precision decreases as far as negative and positive polarity items are concerned. Varying the number of iterations can thus be used to produce small lists of lexical units with high precision rates with regard to positive and negative sentiment.

4.5 Lexical Relations (LRs)

In order to propagate the seeds through the wordnet near_synonym (comparable to similar_to in Princeton Wordnet), antonym, hyponym, and hypernym relations are used. The adjective part of the Dutch Wordnet includes 3,119 hypernym/hyponym relations, 1,070 antonym relations and 703 near_synonym relations. The hierarchy is rather flat with many top nodes and only few synsets that have both hypernym and hyponym relations.

	Lexical relation	F
1	Ant(onym)	0.66
2	Hyper(nym)	0.66
3	Syn (near synonym)	0.67
4	Hypo(nym)	0.71
5	Syn+Ant+Hyper	0.69
6	Hyper+Hypo	0.73
7	Ant+Hypo+Hyper	0.74
8	Syn+Ant+Hypo	0.75
9	Syn+Hypo+Hyper	0.75
10	Syn+Ant+Hypo+Hyper	0.75

Table 8 Various types of Lexical Relations

Table 8 (row 4) shows that the best scoring relation is the hyponym relation with 0.71 whereas the other relations (cf. row 1-3) hardly outperform each other. Combinations of links score equally good (0.75) as long as the near-synonym (Syn) and hyponym (Hypo) relations are included (cf. row 8-10). When all relations are used, the impact of the hypernym relations is nihil (cf. row 8 and 10). The same holds for the antonym relation (cf. row 9 and 10): when all other relations are used the antonym relations do not affect the outcome.

Our conclusion is that there are no LRs which decrease performance. The combination of LRs scores best but only until a certain limit is reached.

These results will differ between wordnets. For example, as in the Princeton Wordnet there are no hyponym/hypernym relations between adjectives, the existing lexical relations will score differently.

4.6 Method II: Conclusions

We conclude that the performance of the propagation approach is determined by the number of iterations, the type and number of lexical relations and the type of seed list. The most important factor in determining the outcome of the propagation algorithm is the size of the seed list, i.e. the larger the better. Another important factor is the quality of the seed list; we proposed a set of rules which can be used to compile a well reasoned seed list.

5 Comparison of Method I and Method II

seed set	lexicon items	F
Transfer	Synsets	0.58
propagation-sdsGI	Synsets	0.75
propagation-sdsHQ	Synsets	0.65
propagation-sdsLQ	Synsets	0.55
Combi II + I	Synsets	0.74

Table 9 Results of transfer (I) and propagation (II) method

The results (copied in Table 9 from earlier sections for reader's convenience) show that the propagation method performs better than the transfer method. The results of the propagation method (0.75) outperform the transfer method (0.58) with 17%. Only with regard to the short and low quality seed list (0.55), the transfer method performs better than the propagation method.

We already mentioned that the two methods might complement each other as they cover different parts of the Dutch Wordnet (cf. Section 2.1). Therefore, the results of the two methods are combined, by taking the scores of the – best scoring – propagation method and replacing that part (2,109 synsets) that lacks lexical relations with the scores of the transfer method. The results show that the overall score degrades with 1%. This means that for those synsets which lack lexical relations, the default value 'neutral' performs better than the transfer method.

6 Comparison with other polarity lexicons

- vs. a word level lexicon for Dutch

To be able to compare the synset level results with other word level polarity lexicons, we generate a word level version of our lexicon. The results are evaluated against the 1,916 Dutch positive, negative and neutral words of the gold standard w-1916 (cf. section 2.2) and have a performance of 74%. This means that the extra step to bring the polarity values from synsets to words causes a small decline (1%) only.

seed set	language	lexicon items	gs	F
sds-GI	Dut	Words	w-1916	0.74
UvaLex	Dut	Words	w-1916	0.72

Table 10: Results at word level

The scores are compared with the scores of Jijkoun and Hofmann (2009)) who built a polarity lexicon (UvaLex) for Dutch at word level. Their approach is also wordnet based and makes use of lexical relations like synonyms and antonyms and of word-to-word links. Their results are comparable (table 10, row UvaLex) with ours (0.72 vs.

0.74 F-measure). Interestingly, an approach like the one of Jijkoun and Hofmann (2009) which is aimed at polarity annotation at word level, and therefore uses word-to-word relations to propagate the sentiment through the wordnet, does not perform better on word level than our system which is primarily meant for synset level annotation.

- vs. an English word level polarity lexicon

Secondly, the word level results are compared with an English polarity lexicon. Andreevskaia and Bergler (2006) whose annotations are at synset level and then aggregated to the word level evaluated their results against General Inquirer (Stone et al., 1966), and report an overall precision of 66.5%, for all 22,000 adjectives in the English Wordnet. For a smaller selection of 1,828 words with positive or negative polarity only, they report 83% precision. This is comparable with our scores; if we make smaller selections by applying fewer iterations and focus on positive and negative polarity only, we measure 82% precision for 2,530 adjective words. So, overall scores for the complete English wordnet are considerably lower than for the complete Dutch wordnet but with regard to smaller selections, Dutch and English perform equally good.

- vs. an English synset level polarity lexicon

The English Sentiwordnet1.0 (2006) is the only freely available polarity lexicon which covers all synsets of the Princeton Wordnet. A more recent version, Sentiwordnet3.0 which has higher scores than the previous versions, but is not publicly available (Baccianella et al. (2010)).

We measured on Sentiwordnet1.0 an overall performance of 62% (cf. section 3.2 above) which is considerably lower than our scores (0.75 and 0.69 for both seed lists, respectively). However, also in the case of Sentiwordnet1.0, smaller selections produce better results. For example, on a selection of 1648 high scoring positive and negative synsets, 84% precision is achieved.

A weakness of this study is that the results are not tested against one single gold standard. However, since we want to compare lexicons of different languages and different lexicon items (words vs. synsets), this is clearly impossible. We think that observed differences between English and Dutch are due to the considerable difference in size of the English wordnet and the Dutch wordnet (18.563 and 7.898 adjective synsets respectively) . The assumption is supported by the fact that small selections of high scoring items perform equally good across the two wordnets.

7 Conclusions

In this paper we described two approaches to generate synset level polarity lexicons for Dutch. The first approach builds a Dutch language polarity lexicon by translating the English Sentiwordnet into Dutch using translation equivalent links between the Dutch and the English Wordnet. The second approach generates a Dutch polarity lexicon at synset level propagating a seed list of known seeds through the wordnet using lexical relations.

It seems that the transfer of coarse-grained sentiment like positive and negative polarity between wordnets of different languages can be done in a reliable manner, since the decrease in performance – after transfer - is rather low with 4%. Important factors that bear effect on the outcome are the quality of the source lexicon and the quality of the translation links.

However, in the case of the Dutch Wordnet, we found that the propagation method considerably outperforms the transfer method. The best scoring seed list is a large seed list of 1,411 seed synsets, but a smaller 'a high quality' seed synset list, i.e. a list of synsets with many lexical relations, produces rather high scores as well.

Another objective of our study was to find out how methods designed for generating a synset level polarity lexicon perform at word level. Our conclusion is that the differences between the word level and synset level results are so small that they may be considered negligible.

8 Acknowledgements

This research has been carried out within the project From Text To Political Positions (http: //www2.let.vu.nl/oz/cltl/t2pp/). It is funded by the VUA Interfaculty Research Institute CAMeRA

9 References

Andreevskaia, Alina and Sabine Bergler (2006). Sentiment Tagging of Adjectives at the Meaning Level. In *LNAI 4013:. Advances in Artificial Intelligence. 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI-2006*, Springer-Verlag, Heidelberg and Berlin, Germany.

Baccianella, S. , Andea Esuli, F. Sebastiani (2010) Sentiwordnet 3.0 : an enhanced lexical resource for sentiment analysis and opinion mining . Lrec,Malta

Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. (2007). *Language resources and linguistic theory: Typology, second language acquisition, English linguistics (Forthcoming)*, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT.

Esuli, Andrea and Fabrizio Sebastiani. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of LREC-2006*, Genova, Italy.

Fellbaum, Chrisrtiane (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Gyamfi, Y., J. Wiebe, R.Mihalcea, C. Akkaya (2009). Integrating knowledge for subjectivity sense labeling. In *Proceedings of HLT-NAACL2009,* Boulder, Colorado.

Jijkoun, V. and K. Hofmann (2009) Generating a Non-English Subjectivity Lexicon: Relations That Matter. In *Proceedings of EACL-2009*, Athens, Greece.

Maks, I.and P. Vossen (2010a) Modeling Attitude, Polarity and Subjectivity in Wordnet. In *Proceedings of Fifth Global Wordnet Conference*, Mumbai, India.

Maks, I. and P. Vossen (2010b) . Annotation Scheme and Gold Standard for Dutch Subjective Adjectives. In *Proceedings of LREC-2010.* Valletta, Malta.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Su, F.and K. Markert (2008). Eliciting Subjectivity and Polarity Judgements on Word Senses. In *Proceedings of Coling-2008*, Manchester, UK.

Su, F; Markert, K. (2009). Subjectivity Recognition on Word Senses via Semi-supervised Mincuts. In: *Proceedings of NAACL-2009:* Boulder, Colorado.

Wiebe, Janyce and Rada Micalcea.(2006) . Word Sense and Subjectivity. In *Proceedings of ACL'06,* Sydney, Australia.

Vossen, P., I.Maks, R. Segers and H. van der Vliet (2008). Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database. In *Proceedings of LREC-2008*, Marrakech, Morocco