Opinion Mining Using SentiWordNet

Julia Kreutzer & Neele Witte
Semantic Analysis
HT 2013/14
Uppsala University

Contents

1. Introduction

One approach to sentiment analysis is to collect affective words in a sentiment lexicon.
However these lexicons are limited to their domain and do not take into account the relation between words.
This is what motivates the authors of SentiWordNet to extend the large and frequently used WordNet resource by sentiment scores. In this manner NLP applications can access both semantic and sentiment information relying on one resource.

In this essay we first report about WordNet, then introduce SentiWordNet and finally provide an example application using the new sentiment resource for opinion mining. We draw conclusions about SentiWordNet's qualities, weaknesses and applicabilities.

2. WordNet

a. Creation and purpose of WordNet

WordNet's development started in 1985 at the Cognitive Science Laboratory of Princeton University. WordNet and its database can be downloaded and used freely. Development still continues and the current version is WordNet 3.0. WordNet can also be assessed through an online interface.
The idea behind WordNet is to create a "dictionary of meaning" integrating the functions of dictionaries and thesauruses.
Lexical information is not organized in word forms, but in word meanings which is consistent with the human representations of meaning and their processing in the brain.
Besides creating a innovatively organized lexical semantic resource, the researchers aim furthermore to support and promote automatic text analysis for applications in the field of artificial intelligence.

b. Structure and contents

WordNet contains English nouns, verbs, adjectives and adverbs.
They form so called "synsets", i.e. sets of distinctive cognitive synonyms, which glosses, i.e. descriptions of the synsets with sample expressions or sentences, are attached to.
What constitutes the "net"-like structure of WordNet are the links between the synsets. Synsets that have a certain lexical or conceptual relation are linked.

- Nouns can be connected through hyperonymy/hyponymy and meronymy/holonymy relations which can also be inherited.They form a hierarchy which all goes back up to one root. There is also a differentiation between types (common nouns) and instances (persons, entities).
- Verbs are organized via troponym, hypernym and entailment relations.
- Adjectives are linked to their antonyms, and relational adjectives point to their related nouns.

- Adverbs make up the smallest group of synsets. They are mostly derived from adjectives and are linked to them via a pertainym relation.

Additionally there are very few cross-POS relations. Morphosemantic links connect words that share the same stem, as for many adverbs and adjectives. Some noun-verb pairs are furthermore annotated for semantic roles.

In the current version there are 82115 distinct noun synsets, 13767 for verbs, 18156 for adjectives and 3621 for adverbs, which sums up to 117659 synsets composed by 155287 unique words all in all.[1]
The 20-volume Oxford English Dictionary records 171476 words in current use.[2] Estimating the number of words in the English language by this number, WordNet already covers the major part.

c. Application and usage
Besides the English WordNet there are a large number of projects that aim to realize similar systems for over 45 different languages, also multilingually.[3]
In NLP applications WordNet is a popular knowledge source for word sense disambiguation and word similarity calculations. Most appealing is the fact that it offers differentiated hierarchies of semantically organized word groups without suffering losses in terms of coverage and reliability.

Still, a complete semantic analysis of a phrase, text or corpus may require detecting and processing sentiments which form the emotional and directional parts of its meaning.
It is possible to disambiguate senses of words, measure their relatedness to others, define and describe their meaning with WordNet, but in order to cope with sentiment directed tasks as well, WordNet would need to be extended by additional information.
This is the problem that the authors of SentiWordNet tackle - aiming to expand the application of WordNet in NLP tasks to another dimension.

3. SentiWordNet

a. The purpose of SentiWordNet
The aim of SentiWordNet is to provide an extension for WordNet, such that all synsets can be associated with a value concerning the negative, positive or objective connotation. SentiWordNet 3.0 is the improved version of SentiWordNet 1.0 and publicly freely available for research purpose with a webinterface.[4]
This extension labels each synset with a value for each category between 0.0 and 1.0. The sum of the three values is always 1.0, so each synset can have a nonzero value for each sentiment,

---
[1] WordNet statistics: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html (22.11.2013)
[2] Entries in the Oxford English Dictionary: http://www.oxforddictionaries.com/words/how-many-words-are-there-in-the-english-language (22.11.2013)
[3] The Global WordNet Association: http://globalwordnet.org (22.11.2013)
[4] http://sentiwordnet.isti.cnr.it

because some synsets can be positive, negative or objective depending on the context in which they are used. The web interface allows the user to search for any synset belonging to WordNet with its associated SentiWordNet scores. Additionally the user is able to see a visualisation of those scores. Each category is linked to a color, which is red for negativity, blue for objectivity and green for positivity. The visualisation of the synset good#5 can be seen in figure 1.
The advantage of using synsets instead of terms is to offer different sentiment scores for each sense of one word, because the connotations can differ in one word depending on the sense.

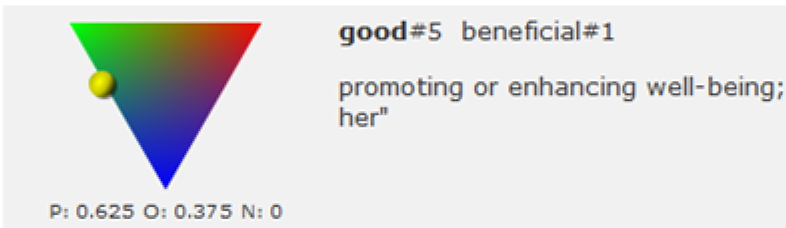


Figure 1: visualisation of a scored synset in SentiWordNet 3.0

b. The methods of building SentiWordNet

Figure 2 shows the main structure of SentiWordNet. The system was built in two main steps. In the **semi-supervised learning step** 8 classifiers were established to decide for each synset belonging to WordNet if it is negative, positive or objective.
This provides on the one hand a higher generalization factor and a low risk of overfitting, on the other hand the different classification results help to give the synsets a tendency of being more positive or negative, rather than just one opportunity: namely that a synset can positive, negative and objective to a certain extend.
In the second step, the **random-walk step** the scores for the positive and negative scores are due to the “defiens-defiendum” relationship.
Through averaging of all the classification results a value between 0.0 and 1.0 can be obtained for each category for each synset. If all classifiers will decide on the same category, this sentiment will have the maximum value, which is 1.0.

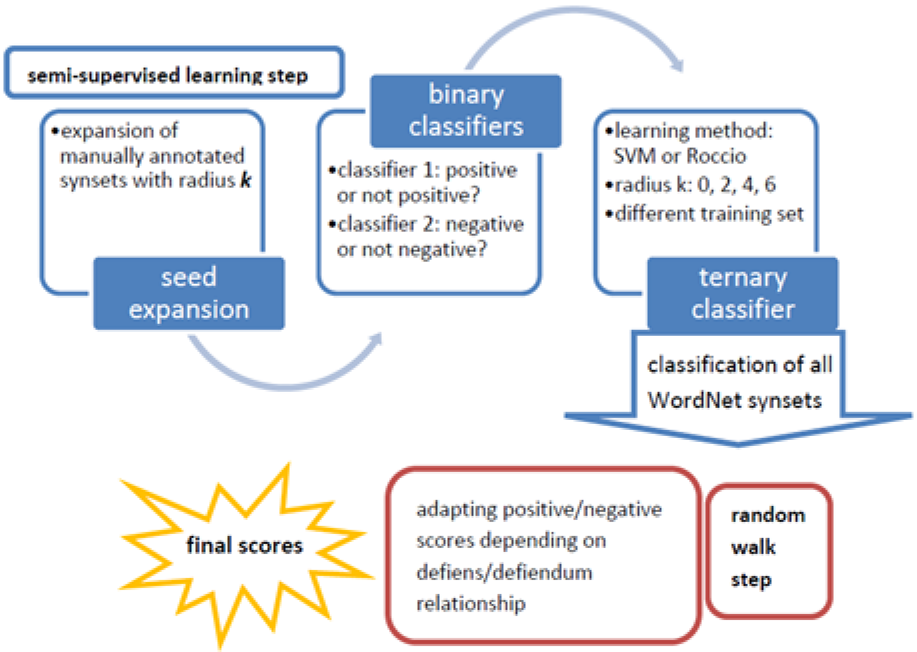


Figure 2: the methods to build SentiWordnet 3.0

c. The semi-supervised learning step

In the semi-supervised step the training set for each of the outcoming 8 classifiers is created.

In the beginning small set of synsets is labelled manually and then expanded by automatic annotation. This can be quite efficient and helps to avoid labelling errors. On these training sets the classifiers are trained with two learning methods.

Figure 2 shows how the seed expansion step works in detail. Seven positive and seven negative instances are chosen from WordNet by human annotators. These are very clearly either negative or positive. Then these two sets of negative and positive synsets are expanded with the help of the WordNet relations. All synsets with the relation “also see” and “similar to” to a positive synset in one class are added to this class, all synsets with the relations “antonym”, i.e. the opposite, are added to the negative class. (same counts for negative examples)

So in the example above the synset “unhappy” has a “also see relation” to the synsets “sad” and “sorrowful” which are added to the negative set. The antonym “happy” is added to the positive set and so on. This is done for a maximum of k iterations, each classifier uses a different numbers of iteration.

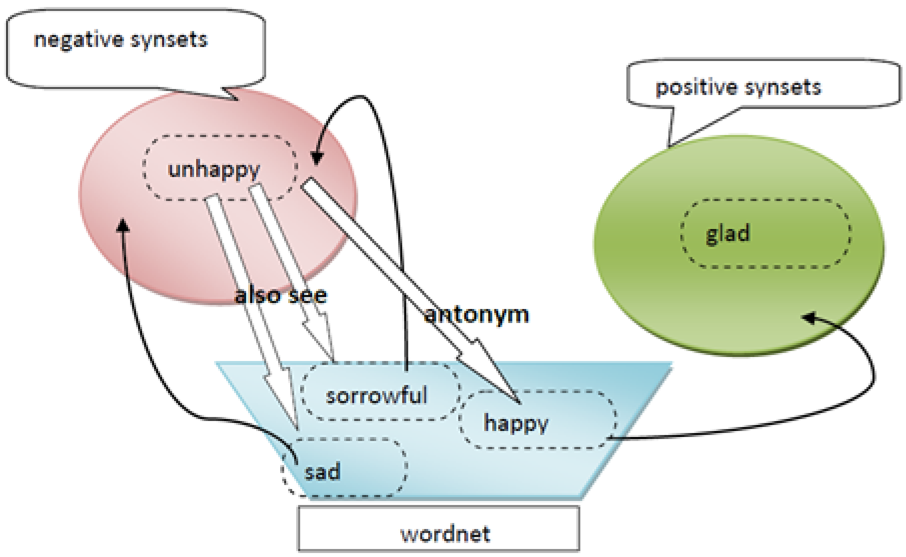


Figure 3: seed expansion for training set

For the objective set of synsets those synsets are used which could not be classified as either positive or negative through the WordNet relations. For an additional hint of objectivity those were also checked in the General Inquirer Lexicon. This dataset provides sentiment scores for a list of words and is often used in sentiment detection.[5] Thus it can be double checked if any objective synset has no positive or negative sentiment for at all.

In the next step two binary classifiers are learned on the previously constructed training set. One of them can classify the synsets as either being positive or not positive, the other can classify them as either being negative or not negative. If a synset is classified as either both positive and negative or not positive and not negative it is concerned to be objective.
The resulting classifier is a ternary classifier which is able to categorize a synset as being negative, objective or positive. It is applied to the whole WordNet to classify each synset. Each classifier uses either the Roccio or SVM learning method and for each training set a different radius k between 0 and 6 was used.

d. The Random Walk step

The random walk algorithm is a graph based model that gives more weight to those nodes that gain more incoming links. The process is running iteratively on WordNet as a graph. In this step the relation between synsets and glosses of other synsets is observed. Because the glosses in WordNet consist of a sequence of ambigue synsets, those are disambiguated with the help of the Princeton WordNet Gloss-Corpus before the process takes place.
The basic assumption for this step is the following: if two synsets share a common context it is very likely that they probably have the same sentiment. This is realized by “walking” from synset

---

[5] Potts, Christopher (2011): Sentiment Symposium Tutorial: Lexicons, Stanford Linguistics. http://sentiment.christopherpotts.net/lexicons.html#inquirer (28.11.2013)

to synset and comparing the relationship to it's neighbours. If a synset (*defiens*) occurs in the gloss of another synset (*defiendum*) they tend to have the same sentiment polarity. So by this comparison, links of positivity and negativity are set up between the synsets. The more positive links are pointing to a certain synset, the higher its positive value will be in the end. So you can see positivity and negativity links between all synsets. Those that have a high number of positive or negative incoming links will have a greater value of positivity or negativity. As the values were too small after the random walk process, the final scores were normalized.

e. Example Scores

Table 3 shows some example scores of the most positive and the most negative synsets.

synsets	positive	negative	objective
good#1	0.75	0	0.25
divine#1	0.875	0	0.125
solid#1	0.875	0	0.125
superb#2	0.875	0	0.125
abject#2	0	1	0
pitiful#2	0	1	0
bad#1	0	0.625	0.325
unfortunate#1	0	0.125	0.875

Table 1[6]: scores for the most positive/negative synsets in SentiWordNet 3.0

It is mentionable that all scores will have one of the following values: 0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1. This shows the effect of weighting all 8 classifiers equally and averaging their classifications.

f. Evaluation

The newer version of SentiWordNet was evaluated on the same test set as the older versions. The Micro-WN (OP) test set consists of 1105 WordNet synsets and had to be changed for the purpose of evaluating SentiWordNet 3.0 via a mapping methods because it was based on an old version of WordNet. The gold standard was obtained with the help of 5 human annotators. Synsets 1-110 were annotated by all of them together to gain a common semantic understanding, the other synsets were annotated independently. With a distance measure the scores of SentiWordNet were compared to the gold standard. Evaluation revealed that an improvement of 19.48 % for the ranking by positivity and 21,69% for the ranking by negativity was gained.

[6] All values taken from: SentiWordNet, 2010: http://sentiwordnet.isti.cnr.it/ (04.12.2013)

4. Application: Opinion Mining

SentiWord extends WordNet's usability by another dimension, it can now used for sentiment analysis as well.
Applications can still use the synsets, glosses and their hierarchical structure but now are able to assess information about the sentiment the synsets contain. Thus applications that already base their semantic analysis on WordNet can enlarge their analysis.
Also applications that only aim for sentiment detection and analysis can profit from this resource.

a. Motivation

Opinion mining (or sentiment analysis) generally aims to extract and analyse a speaker's or writer's opinion.
Especially through the internet it has become very easy to express one's own opinion or read other's opinions today. There are forums where discussions about all kind of subjects take place, blogs where people communicate their thoughts about events, products, people or attitudes, and platforms like twitter, where the users share their feelings and experiences in only a few words.
According to surveys (reported in Pang and Lee, 2008), more than 80% percent of American internet users have researched online for a product at least once, 20% even daily. Over two third of those state that the online reviews or opinions have a significant influence on their purchase.
This shows to which large degree other peoples' opinion stated on the internet influences our own opinion about a product or similar. And our opinion furthermore controls our will to spend money on certain products or services.

Companies therefore have a great interest in their customers' feedback and if many of those customers express their opinion online in text format, companies prefer to find and analyse them automatically. With the help of automatic opinion mining they can adapt their future plans to their customers' opinions and needs and consequently increase their profit.

b. Subtasks of opinion mining

The authors of SentiWordNet name three categories of tasks in the field of opinion mining (Esuli and Sebastiani, 2006):

1. subjectivity-objectivity-polarity: determine whether a text is subjective or objective
2. positivity-negativity-polarity: determine whether a text is positive or negative
3. strength of the positivity-negativity-polarity: determine how positive or negative a text is

In the following example we demonstrate how automatic opinion mining based on SentiWordNet information could be realized.

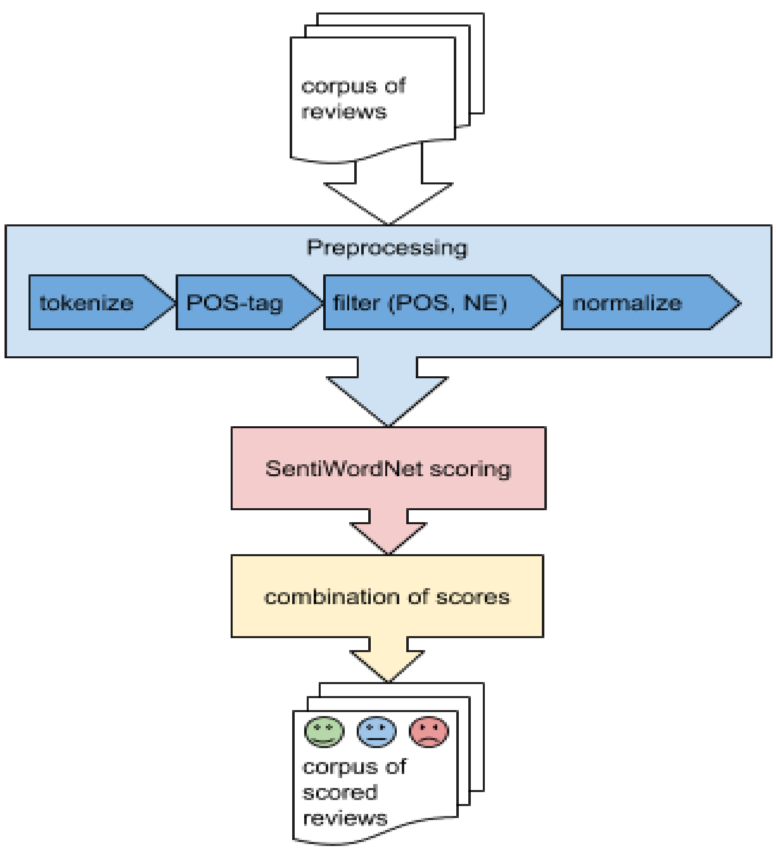
c. Scoring review words

Depending on the specific domain of application reviews or reports have to be collected.
These reports will be categorized or labelled with scores according to their positivity/negativity with the help of SentiWordNet.
The further proceeding is illustrated in figure 3.

SentiWordNet is not able handle multi word queries, so we suggest preprocessing them in the following way:

1. tokenization
2. POS-tagging
3. reduce text to nouns, adjectives, verbs, adverbs (optionally filtering out named entities)
4. normalization: stemming and/or lemmatization

After preprocessing the text is reduced to its contents words in a normalized form.
Now they are ready to be fed into the SentiWordNet system in order to collect sentiment scores for the single word.
For each of the words, SentiWordNet retrieves the synsets that contain that word.



If SentiWordNet does not find any suiting synset, the sentiment scores for this word simply are all zero.
If more than one synset (or several synsets with differing sentiment scores) are returned the system needs to perform word sense disambiguation.
No external resources have to be consulted as the WordNet information part can be used.
There are several ways to perform word sense disambiguation with WordNet, for instance with one of the Lesk algorithms which disambiguate calculating overlaps of the context words (or it's glosses) and the synsets' glosses.

Figure 3: flowchart for opinion mining process applied on reviews using SentiWordNet

d. Combining scores

Once the scores have been retrieved, they have to be combined to classify the text as a whole. This can be done in several ways and we inspect these alternatives:

a) sum up all scores
b) average all scores
c) as in a), only for adjectives
d) as in b), only for adjectives
e) average of all non-zero scores
f) majority vote

The three different scoring numbers for each word can also be assembled into votes, i.e. if a word's negative score is higher than the positive and the objective one, it casts the vote "-1", in case it is the positive score would it be "1", if it is objective it is "0".
These votes can furthermore be combined and interpreted in the different ways described above.
According to the overall scores, the text can finally be classified in either neutral, negative or positive. If wanting to compare texts, it might be sensible to not label the texts but keep their numerical overall scores for each dimension (objective, negative, positive).

e. Example analysis

The effects of choosing one of the above scores will be demonstrated on the following example. We take a product review from amazon.com which was rated with five stars[7]. This indicates that the above described opinion mining procedure should return a very positive score.

"This cute little set is not only sturdy and realistic, it was also a wonderful introduction to preparing food for our 3 year old daughter. Ever since her Grandpa bought this for her, she's made everything from a cheese sandwhich to a triple decker salami club! She had so much fun playing with this toy, that she started to become interested in how I prepared meals. She now is very eager to spread peanut butter and jam, layer turkey and cheese and help mix cake batter. [...] A very cute gift to give a girl or boy."

We assume that tokenization is simply performed by splitting the text at whitespaces. POS-Tags are marked with different colours: nouns, adjectives, verbs, and adverbs.
Step 4) of the preprocessing pipeline turns words like "was" into "is", "preparing" into "prepare", "brought" into "buy" and so on.
Word sense disambiguation for this example is manually performed, e.g. "spread" is assigned to the synset '"cover by spreading something over; "spread the bread with cheese"'.

[7] five star amazon user review on Learning Resources Sandwich Set: http://www.amazon.com/review/R124QNSECSLRPO/ref=cm_cr_dp_title?ie=UTF8&ASIN=B00004WKVF&nodeID=165793011&store=toys-and-games (23.11.2013)

The scores are collected in a spreadsheet (see attached) which returns the results as shown in table 2.

	Positivity	Objectivity	Negativity	Votes
a) SUM	7.125	48.75	2.125	4
b) AVG	0.123	0.841	0.037	0.069
c) SUM (adj)	4.125	5.625	1.25	4
d) AVG (adj)	0.375	0.511	0.114	0.364
e) AVG (nonzero)	0.324	0.855	0.097	0.8
f) majority vote		x		

Table 2: example scores for a positive review; highlighted are the highest scores per row

If we take the average and stay on the positivity/objectivity/negativity scale is the text to 84% objective, 12% positive, around 4% negative (b). This does not agree with the five star rating and our intuition that reviews always are rather subjective text.
As highlighted in the table, the scores for objectivity are always the highest. But at what level turns a text subjective? Which percentage of its words has to be subjective? Here, 84% objectivity are assigned to an obviously subjective text. It remains to be assessed where to draw the line.

If we want a binary decision excluding the objectivity scores the review is classified as positive (b). But it is still hard to say how positive the text is. According to the figures, it is around three times more positive than negative, which is not the usual understanding of measuring the strength of positivity-negativity-polarity.

Since adjectives are the group of words that carry the most notions of sentiment we try to obtain some better results only considering adjectives (c,d). Nevertheless is the the text still over 50% objective and the relation of positivity and negativity stays the same.

Even if we only inspect the non-zero scores (e), the objectivity score is highest although the difference between positivity and negativity scores is increased.

Considering the majority vote (f) the text is likewise classified as objective because most words vote for objectivity.

If we take a look at the votes ("-1","0","1") column, the scores are all positives, which means that the positive votes outnumber the negative ones. This is the first hint that clearly leads to a positive classification. But if we want to compare several texts this score is not reliable, as a longer text may naturally have a higher number of positive words. Therefore we need to inspect the averages. The average of the non-zero votes (0.8) indicates best that the text is positive.

We apply the same procedure now to a negative (2 star) example review[8] of the same product to test the score combining methods onto the opposite sentiment:

"This toy does not look like the picture at all. The bread, tomato, and lettuce were all hard. The only parts that are a bit realistic are the cheese and meats. This toy came with only one tomato and one lettuce. That's not enough objects to make multiple full sandwiches. I would recommend not buying this toy if you want your money's worth. "

	Positivity	Objectivity	Negativity	Votes
a) SUM	1.625	33.125	1.25	-1
b) AVG	0.045	0.920	0.035	-0.028
c) SUM (adj)	0.5	4.875	0.625	-1
d) AVG (adj)	0.083	0.813	0.104	-0.167
e) AVG (nonzero)	0.181	0.945	0.111	-0.111
f) majority vote		x		

Table 3: example scores for a negative review; highlighted are the highest scores per row

Here again, the text is prevailingly classified as objective, as the scores for objectivity are the highest (92% objective, 5% positive, 3% negative).
When comparing only positivity and negativity scores for all words (b) and the nonzero scores (e), it is observable that the negativity average scores are lower than those for the positivity, which does not agree with the two star rating implying negative sentiment.
We obtain better results by only taking adjectives (d) such that the negativity score is higher than the positivity score.
Using the "votes" (last column in the table), all figures are below zero. This leads to the conclusion that the use of this voting system (positive against negative) is most useful for the positivity-negativity-polarity detection.

[8]two star amazon user review on Learning Resources Sandwich Set: http://www.amazon.com/review/R12KIYPIGST7LZ/ref=cm_cr_pr_cmt?ie=UTF8&ASIN=B00004WKVF&linkCode=&nodeID=&tag=#wasThisHelpful

To put it in a nutshell, the main challenge of SentiWordNet's usage for opinion mining is to find an optimal way of combining the word scores to an overall score for whole documents. Our experiments have revealed that the SentiWordNet scores can solve the task of detecting positivity-negativity-polarity, best by summing up positive or negative votes, but struggle at the two remaining tasks of identifying the strength of this polarity and of detecting subjectivity-objectivity-polarity.

f. Related research

Ohana, Bruno and Brendan Tierney (2009) followed a similar approach and aimed for using SentiWordNet in order to rate movie reviews in a larger scale experiment. Besides trying several scoring methods and facing similar problems as we did, they apply machine learning techniques to improve the results. In fact they exclude SentiWordNet from the final classification step but use it for feature extraction to receive features for training a support vector machine classifier.
On the polarity movie data set[9] they obtain an overall accuracy of 65.85% by using the positivity and negativity scores of SentiWordNet and manage to increase it up to an accuracy of 69.35% by additionally applying machine learning.
Compared to the accuracy of 67.80% for methods using word lexicons the improvement is so little that it is debatable whether SentiWordNet really is advantageous over standard sentiment word lexicons.

5. Problems

In this section we discuss the observed problems from section 4. Analyzing the difficulties with the use of SentiWordNet requires a careful inspection of all the methods and assumptions that were included in the process of building the final system. We stress those points that affect the performance of the system most.

a. WordNet

First of all SentiWordNet is built on the WordNet resource, so we cannot expect SentiWordNet to skip problems that occur with WordNet.
It is a commonly observed problem[10] that WordNet senses appear to be too fine-grained such that some distinctions between the senses of a word cannot be captured by humans anymore. Consequently this makes the scoring of synsets even more difficult. And if some sense distinctions cannot be verified by human intuition, the scores will not be verified either.
As reported in section 1b adjectives and adverbs represent the smallest POS-groups in WordNet and furthermore they are least linked synsets. However these are the group of words that contain the most of a text's sentiment. This leads back to the question whether the WordNet resource is the right choice to build sentiment scores upon.

b. SentiWordNet

[9] movie review data set: http://www.cs.cornell.edu/people/pabo/movie-review-data/ (4.12.2013)
[10] as our own experiences in the Semantic Analysis course have proven

There are several points in the construction of SentiWordnet that could be improved in order to gain a better performance. These can also help to find out why the use of SentiWordNet in the example application above failed to classify the very distinct reviews.

The authors of SentiWordNet decided to choose a combination of classifiers based on different learning methods and training to obtain the sentiment scores. This is good as it provides a wide range of learning approaches so the final averaged scores can be more precise with a greater generalisation factor. But in Esuli and Sebastiani (2006) the authors state that both learning methods tend to have the same behaviour. Both have a high precision and a low recall when based on a small radius (e.g. k = 0) which is obvious because then they're both trained on a very small training set and tend to overfitting. With a higher radius (e.g k=6) all classifiers have a lower precision and a high recall[11]. So in fact the classifiers don't differ very much no matter which learning method was used. That makes the use of different learning methods somewhat reasonless because there is not a wider range of results concerned.
It would've been interesting to have classifiers based on other learning methods in order to gain more precise sentiment scores. On top of that one could have taken those classifiers that are more reliable, i.e. the ones with a higher precision, more into account.

The main assumption of the authors about how related synsets and glosses affect each others sentiment is a good idea so far, but unfortunately it is not always the case that related synsets or even terms in one synset have the same sentiment. There are related synsets that contain objective and positive or negative connotated terms.
The example above showed that even one synset can consist of terms with different connotations. For the synset “Grandpa” (grandfather, gramps, granddad, granddad, granddaddy, grandpa) the score is objective. Whereas this is obvious for the word “grandfather” the other words, like “grandpa”, “granddaddy” and especially “gramps” have a more positive connotation. Thus it is important to keep in mind that there are no pure synonyms and synonymous words furthermore tend to have different connotations although they refer to the same object.

One big problem of SentiWordNet 3.0 is its problem of evaluation. The authors state that it is impossible to test the accuracy of the system as it would require a annotation of the whole WordNet. So it is difficult to give evidence or reasons of how good SentiWordNet can work and thus might be useful in further applications.
Besides the evaluation problem in general, for the evaluation of SentiWordNet 3.0, the authors used a test set that is based on an older version of WordNet which they adapted to the newer version with a mapping method. They state that this might have led to annotation errors and that the comparison between the older system (SentiWordNet 1.0) and the new one is only partly reliable.[12] One could have developed a new test set based on the latest version of WordNet to

[11] Esuli, Andrea, and Fabrizio Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." *Proceedings of LREC*. Vol. 6. 2006.

[12] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." *LREC*. Vol. 10. 2010.

avoid annotation errors to provide more reliable results.
Another problem was the manual annotation of the test set.The gold standard was annotated by five humans who started to annotate the first 110 words together and then annotated the rest of the test set independently. The question is how reliable the human annotation is, as they had to annotate each synset with a score for each sentiment. It is quite difficult to decide to what percentage a synset is positive, negative or objective and it therefore might not be enough to take so few annotators into account. They already tried to solve that problem by allowing the user to evaluate the calculated score of a term with either “the scores are ok” or the opportunity to propose better scores. The comparison with the older version of SentiWordnet showed that the new version had a better result though.

c. Informativeness of the scores

A more general question is whether every opinion can be classified by a negativity, a positivity and a objectivity score. The authors chose this way because they wanted to find a solution for the three tasks of opinion mining listed in section 4b. By having three figures, summing up to 1, they assumed it would help to first detect a subjectivity examining the objectivity score, then to compare the positivity and negativity scores to classify the subjectivity and finally to even draw conclusions about this polarity’s strength.
This works in theory, but as our experiment has shown, the objectivity scores always prevail even in texts with strong sentiment. Without defining a threshold different from the intuitive 0.33 or finding a way of smoothing the objectivity scores, the subjectivity-objectivity-polarity detection fails.

On the one hand side this three dimensional score allows to maintain some space for ambiguities, but on the other hand side it remains doubtful whether a rather balanced score, for instance a score of P:0.33, O:0.33, N:0.33, can express any polarity.

Besides the synsets with balanced scores there are also synsets where the score can simply not be verified by our intuition. When running the experiments above, we were surprised to find that the synset “male_child#1 boy#1” (P:0.25, O:0.75, N:0) has a more positive score than the synset “little_girl#1 girl#2 female_child#1” (P:0.125, O:0.874, N:0)”. This implies that the score calculation methods clearly need some improvement as described above.

6. Conclusion

All in all the idea of extending WordNet, one of the biggest lexical databases available in the web, with Sentiment Scores can be very helpful. But the main task for which this extension can be used is opinion mining. The problem is that the scores SentiWordNet calculates are not easily applicable for such a task as explained in section 5. The objective score is mainly too high to classify a text as either positive or negative, even when the text is obviously belonging to one category.

We suggest for future work to further try to find a way to benefit from the original qualities of WordNet and on the same time integrate some sentiment information. The basic idea to use the relations between words to draw conclusions about the relation of their sentiment seems convincing, but the scoring system did not work well for practical application.
It would be helpful to expand a database like WordNet by more adjectives or affective words or to focus on another database which just contains those affective words. It would also be helpful to integrate some colloquial expressions, abbreviations and words used in “online” language to cope with the challenges of opinion mining. Nowadays opinions are mainly expressed online, and WordNet does not cover these relatively new expressions which makes SentiWordNet unable to detect a sentiment in it.
Once the database has sufficient coverage the significance of the scores still needs to be improved. This could for example be realized by applying different learning methods or adapting the scores especially concerning the subjective-objective polarity.

References

a. Web sources

- Entries in the Oxford English Dictionary
http://www.oxforddictionaries.com/words/how-many-words-are-there-in-the-english-language (22.11.2013)

- WordNet Statistics
http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html (22.11.2013)

- The Global WordNet Association
http://globalwordnet.org (22.11.2013)

- SentiWordNet Online Application
http://sentiwordnet.isti.cnr.it/ (04.12.2013)

- Potts, Christopher (2011): Sentiment Symposium Tutorial: Lexicons, Stanford Linguistics
http://sentiment.christopherpotts.net/lexicons.html#inquirer (28.11.2013)

- Bo Pang’s and Lillian Lee’s movie review database
http://www.cs.cornell.edu/people/pabo/movie-review-data/ (04.12.2013)

➢ amazon.com user reviews on Learning Resources Sandwich Set: http://www.amazon.com/review/R124QNSECSLRPO/ref=cm_cr_dp_title?ie=UTF8&ASIN=B00004WKVF&nodeID=165793011&store=toys-and-games (23.11.2013) http://www.amazon.com/review/R12KIYPIGST7LZ/ref=cm_cr_pr_cmt?ie=UTF8&ASIN=B00004WKVF&linkCode=&nodeID=&tag=#wasThisHelpful (04.12.2013)

b. Papers

➢ Miller, George A., et al. "Introduction to wordnet: An on-line lexical database*." *International journal of lexicography* 3.4 (1990): 235-244.

➢ Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38.11 (1995): 39-41.

➢ Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." *LREC*. Vol. 10. 2010.

➢ Esuli, Andrea, and Fabrizio Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." *Proceedings of LREC*. Vol. 6. 2006.

➢ Indurkhya, Nitin, and Fred J. Damerau, eds. *Handbook of natural language processing*. Vol. 2. CRC Press, 2010.

➢ Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and trends in information retrieval* 2.1-2 (2008): 1-135.

➢ Ohana, Bruno, and Brendan Tierney. "Sentiment classification of reviews using SentiWordNet." *9th. IT & T Conference*. 2009.

➢ Ohana, Bruno. "Opinion mining with the SentiWordNet lexical resource." (2009)

Appendix: score calculations for the example reviews

positive review						**negative**				

						review				
5 stars						2 stars				
word in text	**P**	**O**	**N**	**vote**		**word in text**	**P**	**O**	**N**	**vote**
cute	0.5	0.5	0	1		toy	0	1	0	0
little	0	0.625	0.375	0		does	0	1	0	0
set	0.125	0.875	0	0		look	0	1	0	0
is	0.25	0.625	0.125	0		picture	0	1	0	0
only	0	1	0	0		bread	0	1	0	0
sturdy	0.375	0.625	0	0		tomato	0	1	0	0
realistic	0.375	0	0.625	-1		lettuce	0	1	0	0
was	0.25	0.625	0.125	0		all	0.375	0.625	0	0
also	0	1	0	0		hard	0	1	0	0
wonderful	0.75	0.25	0	1		only	0	1	0	0
introduction	0	1	0	0		parts	0	1	0	0
preparing	0	1	0	0		are	0.25	0.625	0.125	0
food	0	1	0	0		bit	0	0.875	0.125	0
year	0	1	0	0		realistic	0.375	0	0.625	-1
old	0.375	0.625	0	0		are	0.25	0.625	0.125	0
daughter	0	1	0	0		cheese	0	1	0	0
ever	0	1	0	0		meats	0	1	0	0
grandpa	0	1	0	0		toy	0	1	0	0
bought	0	1	0	0		came	0	1	0	0
has made	0	1	0	0		only	0	1	0	0
cheese	0	1	0	0		tomato	0	1	0	0
sandwich	0	1	0	0		lettuce	0	1	0	0
triple	0	1	0	0		is	0.25	0.625	0.125	0
decker	0.25	0.75	0	0		enough	0.125	0.875	0	0
salami	0.375	0.625	0	0		objects	0	1	0	0
club	0	1	0	0		make	0	1	0	0
had	0	1	0	0		multiple	0	1	0	0
so	0	1	0	0		full	0	1	0	0
much	0	1	0	0		sandwiches	0	1	0	0
fun	0.375	0.625	0	0		recommend	0	0.875	0.125	0
playing	0	1	0	0		buying	0	1	0	0
toy	0	1	0	0		toy	0	1	0	0
started	0	1	0	0		want	0	1	0	0
become	0	1	0	0		money	0	1	0	0
interested	0.625	0.375	0	1		worth	0	1	0	0
prepared	0	1	0	0		only	0	1	0	0

meals	0	1	0	0						
now	0	1	0	0						
is	0.25	0.625	0.125	0						
very	0.25	0.5	0.25	0						
eager	0.625	0.125	0.25	1						
spread	0	1	0	0						
peanut	0	1	0	0						
butter	0	1	0	0						
jam	0	1	0	0						
layer	0	1	0	0						
turkey	0	1	0	0						
cheese	0	1	0	0						
help	0	1	0	0						
mix	0	1	0	0						
cake	0	1	0	0						
batter	0	1	0	0						
very	0.25	0.5	0.25	0						
cute	0.5	0.5	0	1						
gift	0.25	0.75	0	0						
give	0	1	0	0						
girl	0.125	0.875	0	0						
boy	0.25	0.75	0	0						
SUM	7.13	48.75	2.13	4.00		**SUM**	1.63	33.13	1.25	-1.00
AVG	0.12	0.84	0.04	0.07		**AVG**	0.05	0.92	0.03	-0.03
average of those who are not 0	0.32	0.86	0.10	0.80		average of those who are not 0	0.18	0.94	0.11	-0.11
0 countings	36.00	1.00	36.00	53.00		0 countings	31.00	1.00	31.00	36.00
0 percentage	0.62	0.02	0.62	0.91		0 percentage	0.84	0.03	0.84	0.97
only adjectives SUM	4.13	5.63	1.25	4.00		only adjectives SUM	0.50	4.88	0.63	-1.00
only adjectives AVG	0.38	0.51	0.11	0.36		only adjectives AVG	0.08	0.81	0.10	-0.17