Building a fine-grained subjectivity lexicon from a web corpus

Isa Maks and Piek Vossen

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

Abstract

In this paper we propose a method to build fine-grained subjectivity lexicons including nouns, verbs and adjectives. The method, which is applied for Dutch, is based on the comparison of word frequencies of three corpora: Wikipedia, News and News comments. Comparison of the corpora is carried out with two measures: log-likelihood ratio and a percentage difference calculation. The first step of the method involves subjectivity identification, i.e. determining if a word is subjective or not. The second step aims at the identification of more fine-grained subjectivity which is the distinction between actor subjectivity and speaker / writer subjectivity. The results suggest that this approach can be usefully applied producing subjectivity lexicons of high quality.

Keywords: subjectivity, sentiment, lexicon building

1.   Introduction

In this paper we present a method to build a fine-grained subjectivity lexicon for Dutch on the basis of three corpora, while making use of the dissimilarities between these corpora.

Early work in automatic building of subjectivity lexicons focused mainly on polarity identification which is knowing whether a lexical item is negative, positive or neutral (Kamps et al., 2004; Hatzivassiloglou and McKeown, 1997; Esuli and Sebastiani, 2006). In more recent work it has been argued that the classification of subjectivity vs. objectivity needs to be done independently from the polarity identification (Gyamfi et al., 2009; Su and Markert, 2009). Items may be subjective without having polarity. For example, words like *know*, *feeling*, and *interested* are subjective since they express thoughts, speculations and emotions, but they do not refer to positive or negative sentiment (Gyamfi et al., 2009). Moreover, research has shown that sentiment analysis benefits from subdividing the annotation task in distinguishing subjective and objective instances prior to polarity classification (Yu and Hatzivassiloglou, 2003; Wilson et al., 2005). Annotation schemas have been developed which show that human annotators can annotate words and word senses in a reliable manner, both with regard to polarity and with regard to subjectivity annotation (Wiebe and Mihalcea, 2006; Su and Markert, 2008). In recent annotation studies (Maks and Vossen, 2010; 2011), we argued that a third task which we call 'attitude holder identification', needs to be defined. Attitude holder identification regards the classification of subjective items in a subclass of subjectivity: the distinction between so-called speaker/writer and actor subjectivity where the speaker/writer is the producer of the utterance and the actor is some other person referred to or quoted in the text.

This distinction is illustrated by the following examples:

Ex.(1)   Bush is angry over Obama's leeking of private conversation ..... [attitude/judgment of Bush on Obama]

Ex.(2) Bush is bad for the economy …. [attitude/judgement of Speaker/Writer on Bush]

The subjectivity cues *angry* and *bad* imply different kinds of information relevant to subjectivity analysis: (a) they are – as opposed to other words in the sentence like *economy* and *Obama* – subjective as they refer to somebody's private state; (b) they both have polarity; and (c) they give information about the person whose private state is expressed, the attitude holder. In example (1), representing actor subjectivity, the attitude holder is the person the adjective is attributed to, in this case represented by the logical subject of the sentence. In example (2), representing speaker/writer subjectivity, the attitude of the implicit speaker/writer is expressed.

These different kinds of subjectivity can be found with verbs, nouns and adjectives as illustrated below:

Ex. (3) He is my hero (Positive attitude of Speaker/Writer towards 'he')

Ex. (4) They are nagging all day long (Negative attitude of Speaker/Writer towards 'they')

Ex. (5) His hatred for religion …. (Negative Attitude of 'his' (he) towards 'religion')

Ex. (6) T.F. <u>criticizes</u> her daughter … (Negative Attitude of 'T.F.' towards 'her daughter')

Ex. (7) He is boasting about his work … (Positive Attitude of 'he' towards 'his work'; Negative attitude of Speaker/writer towards 'he')

Sentences (3) and (4) are illustrations of speaker subjectivity. Both express the speaker's attitude towards the participants, i.e., 'he' and 'they', respectively. Sentences (5) and (6) are illustrations of actor subjectivity as they both express attitudes from persons referred to in the text. Sentence (7) shows that speaker's and actor's attitude can also be combined in one word: 'boast', which expresses both the attitude of the actor ('he') and the attitude of the speaker. These different perspectives, and the possible inferences implied by them, are part of the semantics of the word itself. As the subjectivity type, in combination with the argument structure, can be used to identify the attitude holder, we suggest that this is an important word feature to be stored in a subjectivity lexicon.

In earlier studies we showed that attitude holder annotation can be done in a reliable manner manually (Maks and Vossen, 2010; 2011). In this paper we explore whether this task can also be performed reliably in an automatic way. We present a method to perform the two subtasks of subjectivity identification (subjective vs. objective) and attitude holder identification. First, a distinction is made between subjective and objective words. In addition, we make a distinction between actor and speaker/writer subjectivity, leaving polarity identification for future work.

In the following section, we explain in more detail the general idea behind the method, the corpus comparison measures and the design of the corpus. In section 3, we briefly discuss the gold standards used for the evaluation of the obtained lexicons. In section 4 we present the results of the lexicon-building process and in section 5 we conclude with a discussion of the results.

2. Subjective word extraction method

The main assumption of the proposed method is that words that express different types of subjectivity are distributed differently depending on the text type. The method is based on comparison between these texts using and testing two different calculations: the log-likelihood technique developed by Garson et al. (2000) and the percentage difference calculations (DIFF, from now on) developed by Gabrielatos and Marchi (2011). Both measures have been used for keyword extraction and allow identification of words that are significantly more frequent in one corpus than in the other. The higher the value, the more representative the word is for that particular corpus

and so, in our view, the more indicative for that specific type of subjectivity.

2.1 Corpus composition

For this kind of method, it is important to design an appropriate corpus. In our case, we need a corpus which first helps to distinguish subjective from objective words, and then to distinguish actor from speaker/writer subjectivity. We propose that both types of subjectivity can be found in texts related to news. In our view, the ideal corpus would consist of one part containing news articles and another part containing user comments on these news articles. With respect to subjective words, the first part primarily includes words carrying actor subjectivity expressing emotions and attitude of the persons in the text (e.g. politicians, experts, victims, etc.) reported on by journalists. The second part primarily includes words carrying speaker/writer subjectivity that refer to speaker/writer attitudes towards the issues and persons reported on in these news articles. In this design, the two parts of the corpus have in common the objective words which are used to refer to the news issues and their differences ca be found in the subjective words that represent two types, i.e., speaker/writer and actor subjectivity. The common par would then be filtered out by our corpus comparison method leaving the subjective words which can be distinguished by futher constrasting of the two corpus parts. However, it proved to be difficult to compose a corpus that is sufficiently large for our purposes and that includes news articles and their user comments. Therefore, we decided on a slightly different design. To distinguish between speaker/writer and actor subjectivity, we collected news articles and user comments in the same period but not directly related to each other. In addition we added a large amount of Wikipedia articles as an extra help to filter out objective words.

The corpus consists of three components: a random selection of Dutch Wikipedia articles (DWIKI); a collection of news articles from four Dutch newspapers (DNEWS) and a collection of reader comments on a part of the news articles (DCOMM). Texts included in the news and reader comment corpus have been published in 2010 and 2011 and were collected directly from the respective website's archive. Table 1 gives an overview of the three components of the corpus.

.

	nr of Articles	nr of Tokens	nr of Types
Dutch Wikipedia	36K	14,8M	240K
Dutch News	25K	4,9M	135K
Dutch News Comments	338K	13,9M	225K
Total	400K	33, 6M	600K

Table 1: Corpus Overview (K= thousand; M= million)

Concerning the language use in the different corpus components we have the following assumptions:

A. Subjectivity is expressed differently in the three components with respect to both the amount and the type of subjective expressions;

B. Wikipedia articles contain mainly facts and can hence be considered as maximally objective;

C. Both news articles and news comments contain subjective text;

D. News articles contain a large amount of actor subjectivity expressions;

E. News comments contain a large amount of speaker/writer subjectivity expressions.

In view of a comparison of normalised frequencies in the three corpora, the assumption is that if a word occurs more frequently in Wikipedia than in the other two corpora, it is most likely not to be a subjective word; if a words occurs more frequently in Comments than in the other two corpora, it is most likely to be a word expressing speaker/writer subjectivity; and if a word occurs more frequently in News than in the other two corpora, it is most likely to be a word expressing actor subjectivity.

Our method consists of a preprocessing step where the original files of different source formats are converted into structured text format, converting them into lower case, cleaning them by removing hyperlinks, html tags, and the like. During this step we keep track of information such as author, title, subject, date of publication, whenever these are available. Linguistic preprocessing consists of a shallow and limited process of lemmatization where the corpus tokens are linked to their basic lemma form using a 430,000 word form lexicon for Dutch (e-Lex1.1).

2.2 Corpus comparison

To build the lexicon we used a two-step procedure:
• Step 1: Identifying subjective words in the corpus by comparing Wikipedia with News and Comments. Words that are over-represented, i.e., relatively more frequent, in News and Comments are considered subjective and will be used in the next step. Words that are over-represented in Wikipedia are considered objective and not taken into further consideration.
• Step 2: Comparing the remaining (subjective) words of News and Comments. Words that are over-represented in News are considered having

speaker subjectivity; words that are over-represented in Comments are considered having actor subjectivity.

The over-use of words in the different corpora when compared to each other is calculated by two metrics used for corpus comparison, i.e. the log-likelihood test (Garson et al., 2000) and the DIFF calculation (Gabrielatos and Marchi, 2011). The information needed for the log-likelihood (LL) test is the frequency of the word in the one corpus, the frequency of the word in the other corpus and the total amount of words in both corpora. If the LL of the result is greater than 3.84, 6.63 or 10.83, the probability of the result happening by chance is less than 5 % ($p < 0,05$), 1% ($p<0.01$) or 0.1% ($p<0,001$), respectively. The higher the log-likelihood, the more significant the difference between the two frequency scores and the more we expect that the difference between the two corpora actually means something and helps us to identify the words we are interested in.

The percentage difference is calculated using the following formula: $(((fr1 – fr2) * 100) / fr2)$ where fr1 is the word's normalized frequency in the target corpus and fr2 is the word's normalized frequency in the reference corpus. The higher the DIFF score, the higher the frequency difference between the corpora.

Moreover, these experiments only include words with a frequency > 2 within the compared corpora to exclude misspellings and the like.

3. Gold standards

For the evaluation of the results for Dutch we use the gold standard for nouns, verbs and adjectives developed by Maks and Vossen (2010, 2011). The gold standard includes word-sense level annotations for subjectivity (subjective (S) vs. objective (O)), and attitude holdership (Speaker/Writer (SW) vs. Actor (AC) vs. Objective (O)). Inter-annotator agreement for the subjectivity vs. objectivity version is 89% (89%, 90% and 86% for adjectives, nouns and verbs, respectively) with a Cohen kappa of 0.79. Inter-annotator agreement for the attitude holder gold standard is 85% (87%, 87% and 83% for adjectives, nouns and verbs, respectively) with a Cohen kappa of 0.77.

For the purpose of this study, we derived word level standards from these word sense level gold standards, with one annotation for each word. The subjective-objective gold standard (gs-so) consists of 406 objective and 610 subjective words. The attitude holder standard sub-categorises the subjective words into speaker/writer subjectivity (SW) and actor subjectivity (AC) and also includes the 406 objective words. If a word has senses with different annotations, the annotations of the (alleged) most

frequent sense is chosen. The total number of words included in these gold standards is 1016 (cf. table 2) distributed among the different parts-of-speech nouns, verbs and adjectives.

	O	S	SW	AC	total
GS-SO	406	610			1016
GS-SWACO	406		426	184	1016

Table 2: Composition Gold Standards

Each word received one annotation in one gold standard (cf. table 3). This is a simplification with respect to the examples given earlier (cf. section 1, ex. 7) where the verb *boast* expresses both SW and AC subjectivity. In the present gold standards – with one value for each word - we preferred the SW value over AC with the assumption that SW expresses a stronger sentiment than AC.

	gs-so	gs-swaco
held (hero)	S	SW
weigeren (refuse)	S	AC
slecht (bad)	S	SW
boos (angry)	S	AC
huis (house)	O	O
opscheppen (boast)	S	SW

Table 3: Sample of Gold Standard entries

4. Experiments and Results

In this section we present the results of the consecutive steps of the method and hold them against the gold standards.

4.1 Objective vs. Subjective

First, we created a baseline that assigns the most frequent category 'subjective' to all words of the corpus. The baseline achieves an accuracy of 59% on the gold standard gs-so (cf. table 4, base-so). Second, we checked our assumptions (cf. section 3.1) which state that objective words are relatively frequent in Wikipedia (assumption B) and subjective words are relatively frequent in News and Comments (assumption C). We built a word list including all words more frequent in the News and Comments corpus than in Wikipedia and label them as subjective. This 'zero word list' makes use of the frequency differences between the corpus parts, but no LL or DIFF threshold is applied to make smaller selections. Accuracy of this list on gold standard

gs-so is calculated for all three part-of-speeches (ANV) together and for adjectives (A), nouns (N) and verbs (V) separately. The accuracy rate is 72% (cf. table 4: zero) which outperforms the baseline with 13%. Hence, we conclude that the assumptions are correct.

Next, we applied the log-likelihood tests and DIFF calculations in order to build smaller lexicons with higher accuracy. Table 4 presents the results for the different log-likelihood thresholds (LL) evaluated against the gs-so. The columns 'gscov' and 'lsize' report the overlap of the gold standard and the lexicon and the size of the lexicon, respectively. Accuracy is calculated for all parts-of-speeches together and separately. As can be seen from table 4, accuracy increases with 14% from 59% (the baseline) to 73% when applying the LL threshold of 3.84. Although there is a performance increase with all parts-of-speech, high scores on subjectivity are only found with adjectives, while nouns and verbs do not surpass 66% and 63%, respectively. Moreover, the application of the log-likelihood ratio does not raise the scores with respect to the zero line: the lexicons get smaller but accuracy on subjectivity increases only with 1%.

As can be seen from table (4), the application of the DIFF calculation is quite successful. With the heightening of the threshold from 25 to 100 (cf. table (4), rows dif25, dif50 and dif100), the size of the lexicon decreases and accuracy increases from 73 to 80 %. High performance is achieved with respect to all parts-of speech, with 90, 73 and 76 % for adjectives, nouns and verbs, respectively.

	ANV	A	N	V	gscov	lsize
base-so	59	66	55	54		
zero	72	84	66	63		
LL 3.84	73	88	66	63	364	74K
LL 6.63	72	89	65	63	331	47K
LL 10.83	71	88	64	63	295	32K
DIFF 25	73	86	67	65	443	54K
DIFF 50	76	88	70	68	406	51K
DIFF 100	80	90	73	76	322	46K

DIFF=percentage difference
LL=log-likelihood ratio
gscov= gold standard coverage
lsize= lexicon size
A=adjectives N=nouns   V=verbs
Table 4: Results tested against gs-so

We conclude that the log-likelihood method works fine with respect to adjectives, but subjectivity identification on verbs and nouns seems more difficult. However, subjectivity identification using DIFF calculations performs quite well with respect to all parts-of-speech.

4.2 Actor (AC) vs. speaker/writer (SW) subjectivity

The next step aims at a further classification of the subjective words into the subcategories AC and SW. We performed two sets of evaluations. The first set starts from the assumption that the previous step was completely successful. The second set of evaluations starts with the true results of the subjectivity identification and presents a more realistic scenario as it takes into account the errors of the previous step.

- Evaluation I

In order to test for the potential of the method in case of a 100% successful identification of subjective words, we deleted the objective words from the gold standard gs-swaco and tested our results on the restricted gold standard which is called gs-swaco-subjective. We created a baseline that assigns the most frequent category SW to all words of the news and comments corpus. On gs-swaco-subjective the baseline achieves an accuracy of 53% (cf. table 5: base-sw-I).

	ANV	A	N	V
base-SW-I	53	65	53	42
zero	74	71	81	66

Table 5: Results tested against gs-swaco-subjective

Assumptions D and E (cf. section 3.1) are tested by compiling a word list with all words relatively more frequent in News labelled as AC and all words relatively more frequent in Comments labelled as SW. Accuracy of this list on the gold standard swaco-subjective (cf. table 5: zero) is 74% which outperforms the baseline with 21%. We conclude that assumptions D and E are correct.

- Evaluation II

The second set of evaluations is more realistic as it builds on the real outcome of the subjective vs. objective classification. In this scenario, we cannot achieve accuracy scores higher than 73 % and 84 % which are the best performances of the previous step with respect to LL and DIFF respectively. The baseline which assigns the most frequent category SW to all words has an accuracy of 46% on gs-swaco (cf. table. 6, base-sw-II).

Table (6) presents the results obtained with the different log-likelihood thresholds evaluated against gs-swaco. In addition to accuracy (acc), we report precision rates for SW (p-sw) and AC (p-aw). Highest accuracy (59%) is achieved with a log-likelihood threshold of 3.86 and outperforms the baseline considerably with 13%. However, precision

on AC is not high with 45% and not all part-of-speeches have equal performance as in particular the verbs score rather poorly. Again, the application of the log-likelihood threshold seems to have a low impact on the results, as accuracy hardly changes when the threshold is raised.

		ANV	A	N	V	gscov	lsize
base-sw-II	acc	46	59	33	31		
LL 3.84/3.84	acc	59	78	60	40	238	15K
	p-sw	62	81	65	41		
	p-ac	45	43	50	39		
LL 6.63/6.63	acc	58	79	59	38	213	11K
	p-sw	61	82	63	37		
	p-ac	49	50	52	44		

acc=accuracy p-sw=precision on SW p-ac=precision on AC
Table 6: LL results tested against gs-swaco

With respect to DIFF calculations, we experiment with different settings. As can be seen from table (7), accuracy increases when the DIFF score is raises.

The higher the performance of the subjectivity lexicon (i.e., the result of the previous step), the higher the scores of the AC/SW lexicon on the gold standard. Precision on SW has the same trend, ranging from 66 to 77%. However, precision on AC does not reflect this trend and remains very low with scores between 30 and 36 %.

	acc	p-sw	p-ac	gscov	lsize
DIFF25/25	**58**	66	33	379	49K
DIFF 25/100	64	71	35	289	43K
DIFF 25/200	66	71	28	207	39K
DIFF 100/25	63	70	30	288	42K
DIFF 100/25	67	72	36	267	40K
DIFF 100/100	69	71	31	235	37K
DIFF 200/25	68	75	26	229	38K
DIFF 200/50	71	77	32	216	36K

d100/25= DIFF>100% on S vs. O combined with
DIFF>25% on AC vs. SW
Table 7: Diff results tested against gs-swaco

Table (8) presents detailed results of the combinations of the lowest (DIFF25/25) and the highest scores (d200/50). In all settings, verbs perform considerably poorer as compared to adjectives and nouns, and AC scores low on all parts-of-speech.

		ANV	A	N	V	gscov	lsize
DIFF25/25	acc	58	67	63	46	379	49K
	p-sw	66	79	72	50		
	p-ac	33	29	41	28		
DIFF200/50	acc	71	75	72	66	216	36K
	p-sw	77	84	80	68		
	p-ac	32	31	30	40		

Table 8: Detailed DIFF results

5. Discussion

We tested our hypotheses B, C, D and E and found that they were correct, suggesting that the general idea behind the method and the design of the corpus is appropriate for the task of coarse-grained (subjective vs. objective) and fine-grained (actor vs. accspeaker/writer) subjectivity identification.

We saw that the application of a log-likelihood treshold to improve accuracy does not perform well, neither with respect to the identification of coarse-grained subjectivity nor with respect to the identification of fine-grained subjectivity. However, DIFF calculations perform quite well on the first task, giving high scores on subjectivity across all parts-of-speech. The identification of fine-grained subjectivity is more problematic. With the DIFF tresholds, high accuracy is achieved but this is largely due to high precision on SW subjectivity while AC subjectivity scores rather low.

Whether the results are useful depends on what they are intended for. We think that coarse-grained subjectivity identification performs quite well and produces useful results without further processing. If we consider the results of the fine-grained subjectivity identification as a preliminary identification with a certain degree of errors, they may be useful for further automatic or manual processing.

A closer look at the data shows that errors are due to different causes. They may be due to the fact that subjectivity labels are associated with words instead of word senses. For example, according to our gold standard *zeepbel* (*soap bubble*) is an objective word. However, in the corpus *zeepbel* is mostly used in the figurative and subjective sense (similar to English *bubble* or *house of cards*). This contrast between gold standard and actual use may be due to the fact that in the gold standard the annotation of the alleged most frequent sense is preferred, but it may also be caused by novel and creative uses of the word.

With respect to the confusion of actor and speaker/writer subjectivity, corpus design may play a role. Many of the AC words refer to emotions. For example, in *he is angry or he hates* the attitude is attributed to the actor (*he*) in the sentence. However, these attitudes can also be attributed to the personal pronoun *I* , as in *I am angry* and *I hate* and in that case the combination expresses the attitude of the speaker being the *I* in the sentence. Thus, words typically expressing actor subjectivity, may frequently occur in news comment when explicitly attributed to the actor (*I*) and then be incorrectly classified by the system as SW.

The resulting lexicons are rather large, ranging from 11.000 to 56.000 words. This is partly due to the fact that we used only shallow NLP techniques in the tagging and lemmatizing procedures. As a consequence, most of the lexicon items are word forms instead of lemmas. Moreover, we did not apply techniques to filter out html-tags, typos, English words, and so on. We think that, if lemmatizing and text cleaning were performed properly, the size of the lexicons may be reduced by 50 %.

6. Comparison to other work

There are some studies for English which identify objective and subjective words and word senses in a lexicon, in particular WordNet, which report high accuracy rates. For example, Su and Markert (2009) make use of both Wordnet definitions and Wordnet relations and achieve an accuracy of 84.6% on all parts-of-speech. However, these studies differ from this study as they are not aimed at finding novel words.

There are many studies which perform a slightly different task to ours, by creating polarity (positive and negative) lexicons from a corpus. For example, Kaji and Kitsuregawa (2007) report 80% precision with respect to positive and negative polarity on adjective and adjective phrases extracted from a corpus. Their performance seems to be comparable with ours as we achieved 80 % accuracy with respect to all parts-of-speech (cf. table 4).
To our knowledge, no comparable studies exist as far as fine-grained subjectivity classification is concerned.

7. Conclusion

In this paper we presented a method to build a fine-grained subjectivity lexicon for Dutch on the basis of three corpora. We used two different comparison measures and found that DIFF performs better than log-likelihood. The results on subjectivity identification are quite promising with accuracy rates of up to 80%. We think that the obtained subjectivity lexicon can be used in applications without further automatic or manual processing.

Identification of more fine-grained type of subjectivity, i.e., actor vs. speaker/writer subjectivity proves to be more problematic. In particular, the identification of actor subjectivity scores rather low. Future work regards the development of additional techniques to improve the identification of actor subjectivity. Moreover, we will test the obtained lexicons also within sentiment analysis and opinion mining applications.

8. Acknowledgements

9. References

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

Gabrielatos, C. and A. Marchi (2011). Keyness : Matching metrics to definitions. Paper presented at *Theoretical- methodological challenges in corpus approaches to discourse studies: and some ways of addressing them,* Portsmouth. Powerpoint presentation available online: http://eprints.lancs.ac.uk/_51449/

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

Kaji, N. and M. Kitsuregawa (2007). Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings EMNLP-CoNLL.*

Kim, S.M. and E.H. Hovy. (2006) Identifying and Analyzing Judgment Opinions. In *Proceedings of the Human Language Technology*, North American Association of Computational Linguistics conference (HLT-NAACL 2006). New York, NY.

Hatzivassiloglou, V., McKeown, K.B. (1997) *Predicting the semantic orientation of adjectives*. In *Proceedings of ACL-97*, Madrid, Spain.

Kamps, J., R. J. Mokken, M. Marx, and M. de Rijke (2004). Using WordNet to measure semantic orientation of adjectives. In *Proceedings LREC-2004, Paris*.

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

Maks, I. and P. Vossen (2011) A verb lexicon model for deep sentiment analysis and opinion mining applications. In *Proceedings of 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, Portland, USA.

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 NAACL-2009:* Boulder, Colorado.

Yu, Hong and Vasileios Hatzivassiloglou (2003). Towards Answering Opinion Questions: Separating Facts from   Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings EMNLP'03*.

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

Wilson, T., Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Joint Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing* (HLT-EMNLP-2005).