From Words to Senses: A Case Study of Subjectivity Recognition

Fangzhong Su
School of Computing
University of Leeds, UK
fzsu@comp.leeds.ac.uk

Katja Markert
School of Computing
University of Leeds, UK
markert@comp.leeds.ac.uk

Abstract

We determine the subjectivity of word senses. To avoid costly annotation, we evaluate how useful existing resources established in opinion mining are for this task. We show that results achieved with existing resources that are *not* tailored towards word sense subjectivity classification can rival results achieved with supervision on a manually annotated training set. However, results with different resources vary substantially and are dependent on the different definitions of subjectivity used in the establishment of the resources.

1 Introduction

In recent years, subjectivity analysis and opinion mining have attracted considerable attention in the NLP community. Unlike traditional information extraction and document classification tasks which usually focus on extracting facts or categorizing documents into topics (e.g., "sports", "politics", "medicine"), subjectivity analysis focuses on determining whether a language unit (such as a word, sentence or document) expresses a *private state, opinion or attitude* and, if so, what *polarity* is expressed, i.e. a positive or negative attitude.

Inspired by Esuli and Sebastiani (2006) and Wiebe and Mihalcea (2006), we explore the automatic detection of the subjectivity of *word senses*, in contrast to the more frequently explored task of determining the subjectivity of *words* (see Section 2). This is motivated by many words being *subjectivity-ambiguous*, i.e. having both subjective and objective senses, such as the word *positive* with its two example senses given below.[1]

(1) positive, electropositive—having a positive electric charge;"protons are positive" (*objective*)

(2) plus, positive—involving advantage or good; "a plus (or positive) factor" (*subjective*)

Subjectivity labels for senses add an additional layer of annotation to electronic lexica and allow to group many fine-grained senses into higher-level classes based on subjectivity/objectivity. This can increase the lexica's usability. As an example, Wiebe and Mihalcea (2006) prove that subjectivity information for WordNet senses can improve word sense disambiguation tasks for subjectivity-ambiguous words (such as *positive*). In addition, Andreevskaia and Bergler (2006) show that the performance of automatic annotation of subjectivity at the *word* level can be hurt by the presence of subjectivity-ambiguous words in the training sets they use. Moreover, the prevalence of different word senses in different domains also means that a subjective or an objective sense of a word might be dominant in different domains; thus, in a science text *positive* is likely not to have a subjective reading. The annotation of words as subjective and objective or positive and negative independent of sense or domain does not capture such distinctions.

In this paper, we validate whether word sense subjectivity labeling can be achieved with existing resources for subjectivity analysis at the word and sentence level *without creating a dedicated, manually annotated training set of WordNet senses labeled for subjectivity*.[2] We show that such an ap-

[1] All examples in this paper are from WordNet 2.0.

[2] We use a subset of WordNet senses that are manually annotated for subjectivity as test set (see Section 3).

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 825–832
Manchester, August 2008

proach — even using a simple rule-based unsupervised algorithm — can compete with a standard supervised approach and also compares well to prior research on word sense subjectivity labeling. However, success depends to a large degree on the definition of subjectivity used in the establishment of the prior resources.

The remainder of this paper is organized as follows. Section 2 discusses previous work. Section 3 introduces our human annotation scheme for word sense subjectivity and also shows that subjectivity-ambiguous words are frequent. Section 4 describes our proposed classification algorithms in detail. Section 5 presents the experimental results and evaluation, followed by conclusions and future work in Section 6.

2 Related Work

There has been extensive research in opinion mining at the document level, for example on product and movie reviews (Pang et al., 2002; Pang and Lee, 2004; Dave et al., 2003; Popescu and Etzioni, 2005). Several other approaches focus on the subjectivity classification of sentences (Kim and Hovy, 2005; Kudo and Matsumoto, 2004; Riloff and Wiebe, 2003). They often build on the presence of subjective words in the sentence to be classified.

Closer to our work is the large body of work on the automatic, context-independent classification of words according to their polarity, i.e as positive or negative (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Kim and Hovy, 2004; Takamura et al., 2005). They use either co-occurrence patterns in corpora or dictionary-based methods. Many papers assume that subjectivity recognition, i.e. separating subjective from objective words, has already been achieved prior to polarity recognition and test against word lists containing subjective words only (Hatzivassiloglou and McKeown, 1997; Takamura et al., 2005). However, Kim and Hovy (2004) and Andreevskaia and Bergler (2006) also address the classification into subjective/objective words and show this to be a potentially harder task than polarity classification with lower human agreement and automatic performance.

There are only two prior approaches addressing *word sense* subjectivity or polarity classification. Esuli and Sebastiani (2006) determine the polarity of word senses in WordNet, distinguishing among positive, negative and objective. They expand a small, manually determined seed set of strongly positive/negative WordNet senses by following WordNet relations and use the resulting larger training set for supervised classification. The resulting labeled WordNet gives three scores for each sense, representing the positive, negative and objective score respectively. However, there is no evaluation as to the accuracy of their approach. They then extend their work (Esuli and Sebastiani, 2007) by applying the Page Rank algorithm to ranking the WordNet senses in terms of how strongly a sense possesses a given semantic property (e.g., positive or negative).

Wiebe and Mihalcea (2006) label word senses in WordNet as subjective or objective. They use a method relying on distributional similarity as well as an independent, large manually annotated opinion corpus (MPQA) (Wiebe et al., 2005) for determining subjectivity. One of the disadvantages of their algorithm is that it is restricted to senses that have distributionally similar words in the MPQA corpus, excluding 23.2% of their test data from automatic classification.

3 Human Annotation of Word Sense Subjectivity and Polarity

In contrast to other researchers (Hatzivassiloglou and McKeown, 1997; Takamura et al., 2005), we do not see polarity as a category that is dependent on prior subjectivity assignment and therefore applicable to subjective senses only. We follow Wiebe and Mihalcea (2006) in that we see subjective expressions as private states "that are not open to objective observation or verification". This includes direct references to emotions, beliefs and judgements (such as *anger, criticise*) as well as expressions that let a private state be inferred, for example by referring to a doctor as a *quack*. In contrast, polarity refers to positive or negative associations of a word or sense. Whereas there *is* a dependency in that most subjective senses have a relatively clear polarity, polarity can be attached to objective words/senses as well. For example, *tuberculosis* is not subjective — it does not describe a private state, is objectively verifiable and would not cause a sentence containing it to carry an opinion, but it does carry negative associations for the vast majority of people.

We therefore annotate subjectivity of word senses similarly to Wiebe and Mihalcea (2006),

distinguishing between *subjective (S), objective (O)* or *both (B)*. *Both* is used if a literal and metaphoric sense of a word are collapsed into one WordNet synset or if a WordNet synset contains both opinionated and objective expressions (such as *bastard* and *illegitimate child* in Ex. 3 below).

We expand their annotation scheme by also annotating polarity, using the labels positive (*P*), negative (*N*) and varied (*V*). The latter is used when a sense's polarity varies strongly with the context, such as Example 8 below, where we would expect *uncompromising* to be a judgement but this judgement will be positive or negative depending on what a person is uncompromising about. To avoid prevalence of personalised associations, annotators were told to only annotate polarity for subjective senses, as well as objective senses that carry a strong association likely to be shared by most people at least in Western culture (such as the negative polarity for words referring to diseases and crime). Other objective senses would receive the label *O:NoPol*.

Therefore, we have 7 sub categories in total: *O:NoPol*, *O:P*, *O:N*, *S:P*, *S:N*, *S:V*, and *B*. The notation before and after the colon represents the subjectivity and polarity label respectively. We list some annotated examples below.

(3) bastard, by-blow, love child, illegitimate child, illegitimate, whoreson— the illegitimate offspring of unmarried parents (*B*)

(4) atrophy—undergo atrophy; "Muscles that are not used will atrophy" (*O:N*)

(5) guard, safety, safety device—a device designed to prevent injury (*O:P*)

(6) nasty, awful—offensive or even (of persons) malicious;"in a nasty mood";"a nasty accident"; "a nasty shock" (*S:N*)

(7) happy—enjoying or showing or marked by joy or pleasure or good fortune; "a happy smile";"spent many happy days on the beach"; "a happy marriage" (*S:P*)

(8) uncompromising, inflexible—not making concessions; "took an uncompromising stance in the peace talks" (*S:V*)

As far as we are aware, this is the first annotation scheme for both subjectivity and polarity of word senses. We believe both are relevant for opinion extraction: subjectivity for finding and analysing directly expressed opinions, and polarity for either classifying these further or extracting objective words that, however, serve to "colour" a text or present bias rather than explicitly stated opinions. Su and Markert (2008) describe the annotation scheme and agreement study in full.

3.1 Agreement Study

We used the Micro-WNOp corpus containing 1105 WordNet synsets to test our annotation scheme.[3] The Micro-WNOp corpus is representative of the part-of-speech distribution in WordNet.

Two annotators (both near-native English speakers) independently annotated 606 synsets of the Micro-WNOp corpus for subjectivity and polarity. One annotator is the second author of this paper whereas the other is not a linguist. The overall agreement using all 7 categories is 84.6%, with a kappa of 0.77, showing high reliability for a difficult pragmatic task. This at first seems at odds with the notion of sentiment as a fuzzy category as expressed in (Andreevskaia and Bergler, 2006) but we believe is due to three factors:

- The annotation of senses instead of words splits most subjectivity-ambiguous words into several senses, removing one source of annotation difficulty.
- The annotation of senses in a dictionary provided the annotators with sense descriptions in form of Wordnet glosses as well as related senses, providing more information than a pure word annotation task.
- The split of subjectivity and polarity annotation made the task clearer and the annotation of only very strong connotations for objective word senses "de-individualized" the task.

As in this paper we are only interested in subjectivity recognition, we collapse *S:V*, *S:P*, and *S:N* into a single label *S* and *O:NoPol*, *O:P*, and *O:N* into a single label *O*. Label *B* remains unchanged. For this three-way annotation overall percentage agreement is 90.1%, with a kappa of 0.79.

3.2 Gold Standard

After cases with disagreement were negotiated between the two annotators, a gold standard annotation was agreed upon. Our test set consists of this agreed set as well as the remainder of the Micro-WNOp corpus annotated by one of the annotators alone after agreement was established. This set is available for research purposes at `http://www.comp.leeds.ac.uk/markert/data`.

[3]The corpus has originally been annotated by the providers (Esuli and Sebastiani, 2007) with scores for positive, negative and objective/no polarity, thus a mixture of subjectivity and polarity annotation. We re-annotated the corpus with our annotation scheme.

How many words are subjectivity-ambiguous? As the number of senses increases with word frequency, we expect rare words to be less likely to be subjectivity-ambiguous than frequent words. The Micro-WNOp corpus contains relatively frequent words so we will get an overestimation of subjective-ambiguous word types from this corpus, though not necessarily of word tokens. It includes 298 different words with all their synsets in WordNet 2.0. Of all words, 97 (32.5%) are subjectivity-ambiguous, a substantial number.

4 Algorithms

In this section, we present experiments using five different resources as training sets or clue sets for this task. The first is the Micro-WNOp corpus with our own dedicated word sense subjectivity annotation which is used in a standard supervised approach as training and test set via 10-fold cross-validation. This technique presupposes a manual annotation effort tailored directly to our task to provide training data. As it is costly to create such training sets, we investigate whether existing resources such as two different subjective sentence lists (Section 4.2) and two different subjective word lists (Section 4.3) can be adapted to provide training data or clue sets although they do not provide *any* information about word senses. All resources are used to create training data for supervised approaches; the subjective word lists are also used in a simple rule-based unsupervised approach.

All algorithms were tested on the Micro-WNOp corpus by comparing to the human gold standard annotation. However, we excluded all senses with the label *both* from Micro-WNOp for testing the automatic algorithms, resulting in a final 1061 senses, with 703 objective and 358 subjective senses. We also compare all algorithms to a baseline of always assigning the most frequent category (objective) to each sense, which results in an overall accuracy of 66.3%.

4.1 Standard Supervised Approach: 10-fold Cross-validation (CV) on Micro-WNOp

We use 10-fold cross validation for training and testing on the annotated synsets in the Micro-WNOp corpus. We applied a Naive Bayes classifier, [4] using the following three types of features:

Lexical Features: These are unigrams in the glosses. We use a bag-of-words approach and filter out stop words.

As glosses are usually quite short, using a bag-of-word feature representation will result in high-dimensional and sparse feature vectors, which often deteriorate classification performance. In order to address this problem to some degree, we also explored other features which are available as training and test instances are WordNet synsets.

Part-of-Speech (POS) Features: each sense gets its POS as a feature (adjective, noun, verb or adverb).

Relation Features: WordNet relations are good indicators for determining subjectivity as many of them are *subjectivity-preserving*. For example, if sense A is subjective, then its antonym sense B is likely to be subjective. We employ 8 relations here—antonym, similar-to, derived-from, attribute, also-see, direct-hyponym, direct-hypernym, and extended-antonym. Each relation R leads to 2 features that describe for a sense A how many links of that type it has to synsets in the subjective or the objective training set respectively.

Finally, we represent the feature weights through a TF*IDF measure.

Considering the size of WordNet (115,424 synsets in WordNet 2.0), the labeled Micro-WNOp corpus is small. Therefore, the question arises whether it is possible to adapt other data sources that provide subjectivity information to our task.

4.2 Sentence Collections: Movie and MPQA

It is reasonable to cast word sense subjectivity classification as a sentence classification task, with the glosses that WordNet provides for each sense as the sentences to be classified. Then we can in theory feed any collection of annotated subjective and objective sentences as training data into our classifier while the annotated Micro-WNOp corpus is used as test data. We experimented with two different available data sets to test this assumption.

Movie-domain Subjectivity Data Set (Movie): Pang and Lee (2004) used a collection of labeled subjective and objective sentences in their work on review classification.[5] The data set contains 5000 subjective sentences, extracted from movie reviews collected from the Rotten Tomatoes web

[4] We also experimented with KNN, Maximum Entropy, Rocchio and SVM algorithms and overall Naive Bayes performed best.

[5] Available at `http://www.cs.cornell.edu/People/pabo/movie-review-data/`

site.[6] The 5000 objective sentences were collected from movie plot summaries from the Internet Movie Database (IMDB). The assumption is that all the snippets from the Rotten Tomatoes pages are subjective (as they come from a review site), while all the sentences from IMDB are objective (as they focus on movie plot descriptions).

The **MPQA Corpus** contains news articles manually annotated at the phrase level for opinions, their polarity and their strength. The corpus (Version 1.2) contains 11,112 sentences. We convert it into a corpus of subjective and objective sentences following exactly the approach in (Riloff et al., 2003; Riloff and Wiebe, 2003) and obtain 6127 subjective and 4985 objective sentences respectively. Basically any sentence that contains at least one strong subjective annotation at the phrase level is seen as a subjective sentence.

We again use a Naive Bayes algorithm with lexical unigram features. Note that part-of-speech and relation features are not applicable here as the training set consists of corpus sentences, not WordNet synsets.

4.3 Word Lists: General Inquirer and Subjectivity List

Several word lists annotated for subjectivity or polarity such as the General Inquirer (GI) [7] or the subjectivity clues list (SL) collated by Janyce Wiebe and her colleagues[8] are available.

The **General Inquirer (GI)** was developed by Philip Stone and colleagues in the 1960s. It concentrates on word polarity. Here we make the simple assumption that both positive and negative words in the GI list are subjective clues whereas all other words are objective.

The **Subjectivity Lexicon (SL)** centers on subjectivity so that it is ideally suited for our task. It provides fine-grained information for each clue, such as part-of-speech, subjectivity strength (strong/weak), and prior polarity (positive, negative, or neutral). For example, *object(verb)* is a subjective clue whereas *object(noun)* is objective. Regarding strength, the adjective *evil* is marked as strong subjective whereas the adjective *exposed* is marked as a weak subjective clue.

Both lexica *do not* include any information about word senses and therefore cannot be used directly for subjectivity assignment at the sense level. For example, at least one sense of any subjectivity-ambiguous word will be labeled incorrectly if we just adopt a word-based label. In addition, these lists are far from complete: compared to the over 100,000 synsets in WordNet, GI contains 11,788 words marked for polarity (1915 positive, 2291 negative and 7582 no-polarity words) and the SL list contains about 8,000 subjective words.

Still, it is a reasonable assumption that any gloss that contains several subjective words indicates a subjective sense overall. This intuition is strengthened by the characteristics of glosses. They normally are short and concise without a complex syntactic structure, thus the occurrence of subjective words in such a short string is likely to indicate a subjective sense overall. This contrasts, for example, with sentences in newspapers where one clause might express an opinion, whereas other parts of the sentence are objective.

Therefore, for the **rule-based unsupervised algorithm** we lemmatized and POS-tagged the glosses in the Micro-WNOp test set. Then we compute a subjectivity score *S* for each synset by summing up the weight values of all subjectivity clues in its gloss. If *S* is equal or higher than an agreed threshold *T*, then the synset is classified as subjective, otherwise as objective. For the GI lexicon, all subjectivity clues have the same weight 1, whereas for the SL list we assign a weight value 2 to strongly subjective clues and 1 to weakly subjective clues. We experimented with several thresholds T and report here the results for the best thresholds, which were 2 for SL and 4 for the GI word list. The corresponding methods are called Rule-SL and Rule-GI.

This approach does not allow us to easily integrate relational WordNet features. It might also suffer from the incompleteness of the lexica and the fact that it has to make decisions for borderline cases (at the value of the threshold set). We therefore explored instead to **generate larger, more reliable training data consisting of WordNet synsets** from the word lists. To achieve this, we assign a subjectivity score *S* as above to *all* WordNet synsets (excluding synsets in the test set). If *S* is higher or equal to a threshold T_1 it is added to the subjective training set, if it is lower or equal to T_2 it is added to the objective training set. This allows us to choose quite clear thresholds so that borderline cases with a score between T_1 and T_2 are not in the training set. It also allows to use part-

[6] http://www.rottentomatoes.com/
[7] http://www.wjh.harvard.edu/~inquirer/
[8] http://www.cs.pitt.edu/mpqa/

of-speech and relational features as the training set then consists of WordNet synsets. In this way, we can automatically generate (potentially noisy) training data of WordNet senses marked for subjectivity without annotating any WordNet senses manually for subjectivity.

We experimented with several different threshold sets but we found that they actually have a minimal impact on the final results. We report here the best results for a threshold T_1 of 4 and T_2 of 2 for the SL lexicon and of 3 and 1 respectively for the GI word list.

5 Experiments and Evaluation

We measure the classification performance with overall accuracy as well as precision, recall and balanced F-score for both categories (objective and subjective). All results are summarised in Table 1. Results are compared to the baseline of majority classification using a McNemar test at the significance level of 5%.

5.1 Experimental Results

Table 1 shows that SL* performs best among all the methodologies. All CV, Rule-SL and SL methods significantly beat the baseline. In addition, if we compare the results of methods with and without additional parts-of-speech and WordNet relation features, we see a small but consistent improvement when we use additional features. It is also worthwhile to expand the rule-based unsupervised method into a method for generating training data and use additional features as SL* significantly outperforms Rule-SL.

5.2 Discussion

Word Lists. Surprisingly, using SL greatly outperforms GI, regardless of whether we use the supervised or unsupervised method or whether we use lexical features only or the other features as well.[9] There are several reasons for this. First, the GI lexicon is annotated for polarity, not subjectivity. More specifically, words that we see as objective but with a strong positive or negative association (such as words for crimes) and words that we see as subjective are annotated with the same polarity label in the GI lexicon. Therefore, the GI definition of subjectivity does not match ours. Also, the GI lexicon does not operate with a clearly expressed polarity definition, leading to conflicting annotations and casting doubt on its widespread use in the opinion mining community as a gold standard (Turney and Littman, 2003; Takamura et al., 2005; Andreevskaia and Bergler, 2006). For example, *amelioration* is seen as non-polar in GI but *improvement* is annotated with positive polarity. Second, in contrast to SL, GI does not consider different parts-of-speech of a word and subjectivity strength (strong/weak subjectivity). Third, GI contains many fewer subjective clues than SL.

Sentence Data. When using the Movie dataset and MPQA corpus as training data, the results are not satisfactory. We first checked the purity of these two datasets to see whether they are too noisy. For this purpose, we used a naive Bayes algorithm with unigram features and conducted a 10-fold cross validation experiment on recognizing subjective/objective sentences within the Movie dataset and MPQA independently. Interestingly, the accuracy for the Movie dataset and MPQA corpus achieved 91% and 76% respectively. Considering that they are balanced datasets with a most frequent category baseline of about 50%, this accuracy is high, especially for the Movie dataset.

However, again the subjectivity definition in the Movie corpus does not seem to match ours. Recall that we see a word sense or a sentence as subjective if it expresses a private state (i.e., emotion, opinion, sentiment, etc.), and objective otherwise. Inspecting the movie data set, we found that indeed the sentences included in its subjective set would mostly be seen as subjective in our sense as well as they contain opinions *about* the movie such as *it desperately wants to be a wacky , screwball comedy , but the most screwy thing here is how so many talented people were convinced to waste their time*. It is also true that the sentences (plot descriptions) in its "objective" data set relatively rarely contain opinions *about the movie*. However, they still contain other opinionated content like opinions and emotions of the characters in the movie such as the obsession of a character with John Lennon in *the beatles fan is a drama about Albert, a psychotic prisoner who is a devoted fan of John Lennon and the beatles*. Since the data set's definition of subjective sentences is closer to ours than the one for objective sentences, we conducted a one-class learning approach (Li and Liu, 2003) using Movie's subjective sentences as

[9] This pattern is repeated for all threshold combinations, which are not reported here.

Table 1: Results

Method	Subjective			Objective			Accuracy
	Precision	Recall	F-score	Precision	Recall	F-score	
Baseline	N/A	0	N/A	66.3%	**100%**	79.7%	66.3%
CV	65.2%	52.8%	58.3%	78.1%	85.6%	81.7%	74.6%†
CV*	**69.5%**	55.3%	61.6%	79.4%	87.6%	**83.3%**	76.7%†
Movie	43.8%	60.1%	50.6%	74.9%	60.7%	67.1%	60.5%
MPQA	44.5%	**78.5%**	56.8%	82.1%	50.1%	62.2%	59.7%
GI	50.4%	39.4%	44.2%	72.2%	80.2%	76.0%	66.4%
GI*	54.5%	33.5%	41.5%	71.7%	85.8%	78.1%	68.1%
SL	64.3%	62.8%	63.6%	81.3%	82.2%	81.8%	75.7%†
SL*	66.2%	64.5%	**65.3%**	82.2%	83.2%	82.7%	**76.9%**†
Rule-GI	38.5%	5.6%	9.8%	66.5%	95.4%	78.4%	65.1%
Rule-SL	59.7%	70.4%	64.6%	**83.4%**	75.8%	79.4%	74.0%†

[1] CV, GI and SL correspond to methods using lexical features only.
[2] CV*, GI* and SL* correspond to methods using a feature combination of lexical, part-of-speech, and WordNet relations.
[3] † indicates results significantly better than the baseline.

the only training data. The algorithm [10] combines Expectation Maximization and Naive Bayes algorithms, and we used randomly extracted 50,000 unlabeled synsets in WordNet as the necessary unlabeled data. This approach achieves an accuracy of 69.4% on Micro-WNOp, which is significantly better than the baseline.

The subjectivity definition in the MPQA corpus is quite close to ours. However, our mapping from its phrase annotation to sentence annotation might be too coarse-grained as many sentences in the corpus span several clauses containing both opinions and factual description. We assume that this is possibly also the reason why its purity is lower than in the Movie dataset. We therefore experimented again with a one-class learning approach using just the *subjective phrases* in MPQA as training data. The accuracy does improve to 67.6% but is still not significantly higher than the baseline.

5.3 Comparison to Prior Approaches

Esuli and Sebastiani (2006) make their labeled WordNet *SentiWordNet 1.0* publically available.[11] Recall that they actually use polarity classification: however, as there is a dependency between polarity and subjectivity classification for subjective senses, we map their polarity scores to our subjectivity labels as follows. If the sum of positive and negative scores of a sense in SentiWordNet is more than or equal to 0.5, then it is subjective and otherwise objective.[12] Using this mapping, it achieves an accuracy of 75.3% on the Micro-WNOp corpus, compared to our gold standard. Therefore our methods CV* and SL* perform slightly better than theirs, although the improvement is not significant.

The task definition in Wiebe and Mihalcea (2006) is much more similar to ours but they use different annotated test data, which is not publically available, so an exact comparison is not possible. Both data sets, however, seem to include relatively frequent words. One disadvantage of their method is that it is not applicable to all WordNet senses as it is dependent on distributionally similar words being available in the MPQA. Thus, 23% of their test data is excluded from evaluation, whereas our methods can be used on any WordNet sense. They measure precision and recall for subjective senses in a precision/recall curve: Precision is about 48/9% at a recall of 60% for subjective senses whereas our best SL* method has a precision of 66% at about the same recall. Although this suggests better performance of our method, it is not possible to draw final conclusions from this comparison due to the data set differences.

[10] Available at http://www.cs.uic.edu/∼liub/LPU/.

[11] Available at `http://sentiwordnet.isti.cnr.it/`

[12] We experimented with slightly different mappings but this mapping gave SentiWordNet the best possible result. There is a relatively large number of cases with a 0.5/0.5 split in SentiWordNet, making it hard to decide between subjective and objective senses.

6 Conclusion and Future Work

We proposed different ways of extracting training data and clue sets for word sense subjectivity labeling from existing opinion mining resources. The effectiveness of the resulting algorithms depends greatly on the generated training data, more specifically on the different definitions of subjectivity used in resource creation. However, we were able to show that at least one of these methods (based on the SL word list) resulted in a classifier that performed on a par with a supervised classifier that used dedicated training data developed for this task (CV). Thus, it is possible to avoid any manual annotation for the subjectivity classification of word senses.

Our future work will explore new methodologies in feature representation by importing more background information (e.g., syntactic information). Furthermore, our current method of integrating the rich relation information in WordNet (using them as standard features) does not use joint classification of several senses. Instead, we think it will be more promising to use the relations to construct graphs for semi-supervised graph-based learning of word sense subjectivity. In addition, we will also explore whether the derived sense labels improve applications such as sentence classification and clustering WordNet senses.

References

Andreevskaia, Alina and Sabine Bergler. 2006. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. *Proceedings of EACL'06*.

Dave, Kushal, Steve Lawrence, and David Pennock. 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *Proceedings of WWW'03*.

Esuli, Andrea and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of LREC'06*.

Esuli, Andrea and Fabrizio Sebastiani. 2007. PageRanking WordNet Synsets: An application to Opinion Mining. *Proceedings of ACL'07*.

Hatzivassiloglou, Vasileios and Kathleen McKeown. 1997. Predicting the Semantic Orientation of Adjectives. *Proceedings of ACL'97*.

Kim, Soo-Min and Eduard Hovy. 2004. Determining the Sentiment of Opinions. *Proceedings of COLING'04*.

Kim, Soo-Min and Eduard Hovy. 2005. Automatic Detection of Opinion Bearing Words and Sentences. *Proceedings of ICJNLP'05*.

Kudo, Taku and Yuji Matsumoto. 2004. A Boosting Algorithm for Classification of Semi-structured Text. *Proceedings of EMNLP'04*.

Li, Xiaoli and Bing Liu. 2003. Learning to classify text using positive and unlabeled data. *Proceedings of IJCAI'03*.

Pang, Bo and Lillian Lee. 2004. A Sentiment Education: Sentiment Analysis Using Subjectivity summarization Based on Minimum Cuts. *Proceedings of ACL'04*.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of EMNLP'02*.

Popescu, Ana-Maria and Oren Etzioni. 2003. Extracting Product Fatures and Opinions from Reviews *Proceedings of EMNLP'05*.

Riloff, Ellen, Janyce Wiebe, and Theresa Wilson. 2003. Learning Subjective Nouns using Extraction Pattern Bootstrapping. *Proceedings of CoNLL'03*

Riloff, Ellen and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. *Proceedings of EMNLP'03*.

Su, Fangzhong and Katja Markert. 2008. Eliciting Subjectivity and Polarity Judgements on Word Senses. *Proceedings of Coling'08 workshop of Human Judgements in Computational Linguistics*.

Takamura, Hiroya, Takashi Inui, and Manabu Okumura. 2005. Extracting Semantic Orientations of Words using Spin Model. *Proceedings of ACL'05*.

Turney, Peter and Michael Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transaction on Information Systems*.

Wiebe, Janyce and Rada Micalcea. 2006. Word Sense and Subjectivity. *Proceedings of ACL'06*.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*.