Integrating Knowledge for Subjectivity Sense Labeling

Yaw Gyamfi and **Janyce Wiebe**
University of Pittsburgh
{anti,wiebe}@cs.pitt.edu

Rada Mihalcea
University of North Texas
rada@cs.unt.edu

Cem Akkaya
University of Pittsburgh
cem@cs.pitt.edu

Abstract

This paper introduces an integrative approach to automatic word sense subjectivity annotation. We use features that exploit the hierarchical structure and domain information in lexical resources such as WordNet, as well as other types of features that measure the similarity of glosses and the overlap among sets of semantically related words. Integrated in a machine learning framework, the entire set of features is found to give better results than any individual type of feature.

1 Introduction

Automatic extraction of opinions, emotions, and sentiments in text (*subjectivity analysis*) to support applications such as product review mining, summarization, question answering, and information extraction is an active area of research in NLP.

Many approaches to opinion, sentiment, and subjectivity analysis rely on lexicons of words that may be used to express subjectivity. However, words may have both subjective and objective senses, which is a source of ambiguity in subjectivity and sentiment analysis. We show that even words judged in previous work to be reliable clues of subjectivity have significant degrees of subjectivity sense ambiguity.

To address this ambiguity, we present a method for automatically assigning subjectivity labels to word senses in a taxonomy, which uses new features and integrates more diverse types of knowledge than in previous work. We focus on nouns, which are challenging and have received less attention in automatic subjectivity and sentiment analysis.

A common approach to building lexicons for subjectivity analysis is to begin with a small set of seeds which are prototypically subjective (or positive/negative, in sentiment analysis), and then follow semantic links in WordNet-like resources. By far, the emphasis has been on horizontal relations, such as *synonymy* and *antonymy*. Exploiting vertical links opens the door to taking into account the information content of ancestor concepts of senses with known and unknown subjectivity. We develop novel features that measure the similarity of a target word sense with a seed set of senses known to be subjective, where the similarity between two concepts is determined by the extent to which they share information, measured by the information content associated with their least common subsumer (LCS). Further, particularizing the LCS features to domain greatly reduces calculation while still maintaining effective features.

We find that our new features do lead to significant improvements over methods proposed in previous work, and that the combination of all features gives significantly better performance than any single type of feature alone.

We also ask, given that there are many approaches to finding subjective words, if it would make sense for word- and sense-level approaches to work in tandem, or should we best view them as competing approaches? We give evidence suggesting that first identifying subjective words and then disambiguating their senses would be an effective approach to subjectivity sense labeling.

There are several motivations for assigning subjectivity labels to senses. First, (Wiebe and Mihalcea, 2006) provide evidence that word sense labels, together with contextual subjectivity analysis, can be exploited to improve performance in word sense disambiguation. Similarly, given subjectivity sense labels, word-sense disambiguation may potentially help contextual subjectivity analysis. In addition, as lexical resources such as WordNet are developed further, subjectivity labels would provide principled criteria for refining word senses, as well as for clustering similar meanings to create more course-grained sense inventories.

For many opinion mining applications, polarity (positive, negative) is also important. The overall framework we envision is a layered approach: classifying instances as objective or subjective, and further classifying the subjective instances by polarity. Decomposing the problem into subproblems has been found to be effective for opinion mining. This paper addresses the first of these subproblems.

2 Background

We adopt the definitions of *subjective* and *objective* from Wiebe and Mihalcea (2006) (hereafter *WM*). Subjective expressions are words and phrases being used to express opinions, emotions, speculations, etc. WM give the following examples:

His **alarm** grew.
He **absorbed** the information quickly.
UCC/Disciples leaders **roundly condemned** the Iranian President's **verbal assault** on Israel.
What's the catch?

Polarity (also called *semantic orientation*) is also important to NLP applications in sentiment analysis and opinion extraction. In review mining, for example, we want to know whether an opinion about a product is positive or negative. Even so, we believe there are strong motivations for a separate subjective/objective (S/O) classification as well.

First, expressions may be subjective but not have any particular polarity. An example given by (Wilson et al., 2005) is *Jerome says the hospital* **feels** *no different than a hospital in the states.* An NLP application system may want to find a wide range of private states attributed to a person, such as their motivations, thoughts, and speculations, in addition to their positive and negative sentiments.

Second, distinguishing *S* and *O* instances has often proven more difficult than subsequent polarity classification. Researchers have found this at various levels of analysis, including the manual annotation of phrases (Takamura et al., 2006), sentiment classification of phrases (Wilson et al., 2005), sentiment tagging of words (Andreevskaia and Bergler, 2006b), and sentiment tagging of word senses (Esuli and Sebastiani, 2006a). Thus, effective methods for *S*/*O* classification promise to improve performance for sentiment classification. In fact, researchers in sentiment analysis have realized benefits by decomposing the problem into *S*/*O* and polarity classification (Yu and Hatzivassiloglou, 2003; Pang and Lee, 2004; Wilson et al., 2005; Kim and Hovy, 2006). One reason is that different features may be relevant for the two subproblems. For example, negation features are more important for polarity classification than for subjectivity classification.

Note that some of our features require vertical links that are present in WordNet for nouns and verbs but not for other parts of speech. Thus we address nouns (leaving verbs to future work). There are other motivations for focusing on nouns. Relatively little work in subjectivity and sentiment analysis has focused on subjective nouns. Also, a study (Bruce and Wiebe, 1999) showed that, of the major parts of speech, nouns are the most ambiguous with respect to the subjectivity of their instances.

Turning to word senses, we adopt the definitions from WM. First, subjective: "Classifying a sense as *S* means that, when the sense is used in a text or conversation, we expect it to express subjectivity; we also expect the phrase or sentence containing it to be subjective [WM, pp. 2-3]."

In WM, it is noted that sentences containing objective senses may not be objective, as in the sentence *Will someone shut that darn* **alarm** *off?* Thus, objective senses are defined as follows: "Classifying a sense as *O* means that, when the sense is used in a text or conversation, we do not expect it to express subjectivity and, if the phrase or sentence containing it is subjective, the subjectivity is due to something else [WM, p 3]."

The following subjective examples are given in

WM:

His **alarm** grew.
alarm, dismay, consternation – (fear resulting from the awareness of danger)
=> fear, fearfulness, fright – (an emotion experienced in anticipation of some specific pain or danger (usually accompanied by a desire to flee or fight))

What's the **catch**?
catch – (a hidden drawback; "it sounds good but what's the catch?")
=> drawback – (the quality of being a hindrance; "he pointed out all the drawbacks to my plan")

The following objective examples are given in WM:

The **alarm** went off.
alarm, warning device, alarm system – (a device that signals the occurrence of some undesirable event)
=> device – (an instrumentality invented for a particular purpose; "the device is small enough to wear on your wrist"; "a device intended to conserve water")

He sold his **catch** at the market.
catch, haul – (the quantity that was caught; "the catch was only 10 fish")
=> indefinite quantity – (an estimated quantity)

WM performed an agreement study and report that good agreement (κ=0.74) can be achieved between human annotators labeling the subjectivity of senses. For a similar task, (Su and Markert, 2008) also report good agreement.

3 Related Work

Many methods have been developed for automatically identifying subjective (*opinion*, *sentiment*, *attitude*, *affect-bearing*, etc.) words, e.g., (Turney, 2002; Riloff and Wiebe, 2003; Kim and Hovy, 2004; Taboada et al., 2006; Takamura et al., 2006).

Five groups have worked on subjectivity sense labeling. WM and Su and Markert (2008) (hereafter *SM*) assign *S*/*O* labels to senses, while Esuli and Sebastiani (hereafter *ES*) (2006a; 2007), Andreevskaia and Bergler (hereafter *AB*) (2006b; 2006a), and (Valitutti et al., 2004) assign polarity labels.

WM, SM, and ES have evaluated their systems against manually annotated word-sense data. WM's annotations are described above; SM's are similar. In the scheme ES use (Cerini et al., 2007), senses are assigned three scores, for positivity, negativity, and neutrality. There is no unambiguous mapping between the labels of WM/SM and ES, first because WM/SM use distinct classes and ES use numerical ratings, and second because WM/SM distinguish between objective senses on the one hand and neutral subjective senses on the other, while those are both neutral in the scheme used by ES.

WM use an unsupervised corpus-based approach, in which subjectivity labels are assigned to word senses based on a set of distributionally similar words in a corpus annotated with subjective expressions. SM explore methods that use existing resources that do not require manually annotated data; they also implement a supervised system for comparison, which we will call *SMsup*. The other three groups start with positive and negative seed sets and expand them by adding synonyms and antonyms, and traversing horizontal links in WordNet. AB, ES, and SMsup additionally use information contained in glosses; AB also use hyponyms; SMsup also uses relation and POS features. AB perform multiple runs of their system to assign fuzzy categories to senses. ES use a semi-supervised, multiple-classifier learning approach. In a later paper, (Esuli and Sebastiani, 2007), ES again use information in glosses, applying a random walk ranking algorithm to a graph in which synsets are linked if a member of the first synset appears in the gloss of the second.

Like ES and SMsup, we use machine learning, but with more diverse sources of knowledge. Further, several of our features are novel for the task. The LCS features (Section 6.1) detect subjectivity by measuring the similarity of a candidate word sense with a seed set. WM also use a similarity measure, but as a way to filter the output of a measure of distributional similarity (selecting words for a given word sense), not as we do to cumulatively calculate the subjectivity of a word sense. Another novel aspect of our similarity features is that they are particularized to domain, which greatly reduces calculation. The domain subjectivity LCS features (Section 6.2) are also novel for our task. So is augmenting seed sets with monosemous words, for greater coverage without requiring human intervention or sacrificing quality. Note that none of our features as we specifically define them has been used in previous work; combining them together, our approach outperforms previous approaches.

4 Lexicon and Annotations

We use the subjectivity lexicon of (Wiebe and Riloff, 2005)[1] both to create a subjective seed set and to create the experimental data sets. The lexicon is a list of words and phrases that have subjective uses, though only word entries are used in this paper (i.e., we do not address phrases at this point). Some entries are from manually developed resources, including the General Inquirer, while others were derived from corpora using automatic methods.

Through manual review and empirical testing on data, (Wiebe and Riloff, 2005) divided the clues into strong (*strongsubj*) and weak (*weaksubj*) subjectivity clues. *Strongsubj* clues have subjective meanings with high probability, and *weaksubj* clues have subjective meanings with lower probability.

To support our experiments, we annotated the senses[2] of polysemous nouns selected from the lexicon, using WM's annotation scheme described in Section 2. Due to time constraints, only some of the data was labeled through consensus labeling by two annotators; the rest was labeled by one annotator.

Overall, 2875 senses for 882 words were annotated. Even though all are senses of words from the subjectivity lexicon, only 1383 (48%) of the senses are subjective.

The words labeled *strongsubj* are in fact less ambiguous than those labeled *weaksubj* in our analysis, thus supporting the reliability classifications in the lexicon. 55% (1038/1924) of the senses of *strongsubj* words are subjective, while only 36% (345/951) of the senses of *weaksubj* words are subjective.

For the analysis in Section 7.3, we form subsets of the data annotated here to test performance of our method on different data compositions.

5 Seed Sets

Both subjective and objective seed sets are used to define the features described below. For seeds, a large number is desirable for greater coverage, although high quality is also important. We begin to build our subjective seed set by adding the monosemous *strongsubj* nouns of the subjectivity lexicon (there are 397 of these). Since they are monosemous, they pose no problem of sense ambiguity. We then expand the set with their hyponyms, as they were found useful in previous work by AB (2006b; 2006a). This yields a subjective seed set of 645 senses. After removing the word senses that belong to the same synset, so that only one word sense per synset is left, we ended up with 603 senses.

To create the objective seed set, two annotators manually annotated 800 random senses from WordNet, and selected for the objective seed set the ones they both agreed are clearly objective. This creates an objective seed set of 727. Again we removed multiple senses from the same synset leaving us with 722. The other 73 senses they annotated are added to the mixed data set described below. As this sampling shows, WordNet nouns are highly skewed toward objective senses, so finding an objective seed set is not difficult.

6 Features

6.1 Sense Subjectivity LCS Feature

This feature measures the similarity of a target sense with members of the subjective seed set. Here, similarity between two senses is determined by the extent to which they share information, measured by using the information content associated with their least common subsumer. For an intuition behind this feature, consider this example. In WordNet, the hypernym of the "strong criticism" sense of *attack* is *criticism*. Several other negative subjective senses are descendants of *criticism*, including the relevant senses of *fire*, *thrust*, and *rebuke*. Going up one more level, the hypernym of *criticism* is the "expression of disapproval" meaning of *disapproval*, which has several additional negative subjective descendants, such as the "expression of opposition and disapproval" sense of *discouragement*. Our hypothesis is that the cases where subjectivity is preserved in the hypernym structure, or where hypernyms do lead from subjective senses to others, *are* the ones that have the highest least common subsumer score with the seed set of known subjective senses.

We calculate similarity using the information-content based measure proposed in (Resnik, 1995), as implemented in the WordNet::Similarity package (using the default option in which LCS values are computed over the SemCor corpus).[3] Given a

[1] Available at http://www.cs.pitt.edu/mpqa

[2] In WordNet 2.0

[3] http://search.cpan.org/dist/WordNet-Similarity/

taxonomy such as WordNet, the information content associated with a concept is determined as the likelihood of encountering that concept, defined as $-log(p(C))$, where $p(C)$ is the probability of seeing concept C in a corpus. The similarity between two concepts is then defined in terms of information content as: $LCS_s(C_1, C_2) = max[-log(p(C))]$, where C is the concept that subsumes both C_1 and C_2 and has the highest information content (i.e., it is the *least common subsumer (LCS)*).

For this feature, a score is assigned to a target sense based on its semantic similarity to the members of a seed set; in particular, the maximum such similarity is used.

For a target sense t and a seed set S, we could have used the following score:

$$Score(t, S) = \max_{s \in S} LCS_s(t, s)$$

However, several researchers have noted that subjectivity may be domain specific. A version of WordNet exists, WordNet Domains (Gliozzo et al., 2005), which associates each synset with one of the domains in the Dewey Decimal library classification. After sorting our subjective seed set into different domains, we observed that over 80% of the subjective seed senses are concentrated in six domains (the rest are distributed among 35 domains).

Thus, we decided to particularize the semantic similarity feature to domain, such that only the subset of the seed set in the same domain as the target sense is used to compute the feature. This involves much less calculation, as LCS values are calculated only with respect to a subset of the seed set. We hypothesized that this would still be an effective feature, while being more efficient to calculate. This will be important when this method is applied to large resources such as the entire WordNet.

Thus, for seed set S and target sense t which is in domain D, the feature is defined as the following score:

$$SenseLCSscore(t, D, S) = \max_{d \in D \cap S} LCS_s(t, d)$$

The seed set is a parameter, so we could have defined a feature reflecting similarity to the objective seed set as well. Since WordNet is already highly skewed toward objective noun senses, any naive classifier need only guess the majority class for high accuracy for the objective senses. We included only a subjective feature to put more emphasis on the subjective senses. In the future, features could be defined with respect to objectivity, as well as polarity and other properties of subjectivity.

6.2 Domain Subjectivity LCS Score

We also include a feature reflecting the subjectivity of the domain of the target sense. Domains are assigned scores as follows. For domain D and seed set S:

$$DomainLCSscore(D, S) = ave_{d \in D \cap S} MemLCSscore(d, D, S)$$

where:

$$MemLCSscore(d, D, S) = \max_{d_i \in D \cap S, d_i \neq d} LCS_s(d, d_i)$$

The value of this feature for a sense is the score assigned to that sense's domain.

6.3 Common Related Senses

This feature is based on the intersection between the set of senses related (via WordNet relations) to the target sense and the set of senses related to members of a seed set. First, for the target sense and each member of the seed set, a set of related senses is formed consisting of its synonyms, antonyms and direct hypernyms as defined by WordNet. For a sense s, $R(s)$ is s together with its related senses.

Then, given a target sense t and a seed set S we compute an average percentage overlap as follows:

$$RelOverlap(t, S) = \frac{\sum_{s_i \in S} \frac{|R(t) \cap R(s_i)|}{\max(|R(t)|, |R(s_i)|)}}{|S|}$$

The value of a feature is its score. Two features are included in the experiments below, one for each of the subjective and objective seed sets.

6.4 Gloss-based features

These features are Lesk-style features (Lesk, 1986) that exploit overlaps between glosses of target and seed senses. We include two types in our work.

6.4.1 Average Percentage Gloss Overlap Features

For a sense s, $gloss(s)$ is the set of stems in the gloss of s (excluding stop words). Then, given a tar-

get sense t and a seed set S, we compute an average percentage overlap as follows:

$$GlOverlap(t,S) = \frac{\sum_{s_i \in S} \frac{\left|gloss(t) \cap \cup_{r \in R(s_i)} gloss(r)\right|}{\max(|gloss(t)|, |\cup_{r \in R(s_i)} gloss(r)|)}}{|S|}$$

As above, $R(s)$ is considered for each seed sense s, but now only the target sense t is considered, not $R(t)$. We did this because we hypothesized that the gloss can provide sufficient context for a given target sense, so that the addition of related words is not necessary.

We include two features, one for each of the subjective and objective seed sets.

6.4.2 Vector Gloss Overlap Features

For this feature we also consider overlaps of stems in glosses (excluding stop words). The overlaps considered are between the gloss of the target sense t and the glosses of $R(s)$ for all s in a seed set (for convenience, we will refer to these as $seedRelationSets$).

A vector of stems is created, one for each stem (excluding stop words) that appears in a gloss of a member of $seedRelationSets$. If a stem in the gloss of the target sense appears in this vector, then the vector entry for that stem is the total count of that stem in the glosses of the target sense and all members of $seedRelationSets$.

A feature is created for each vector entry whose value is the count at that position. Thus, these features consider counts of individual stems, rather than average proportions of overlaps, as for the previous type of gloss feature.

Two vectors of features are used, one where the seed set is the subjective seed set, and one where it is the objective seed set.

6.5 Summary

In summary, we use the following features (here, SS is the subjective seed set and OS is the objective one).

1. $SenseLCSscore(t, D, SS)$
2. $DomainLCSscore(D, SS)$
3. $RelOverlap(t, SS)$
4. $RelOverlap(t, OS)$
5. $GlOverlap(t, SS)$
6. $GlOverlap(t, OS)$

Features	Acc	P	R	F
All	77.3	72.8	74.3	73.5
Standalone Ablation Results				
All	77.3	72.8	74.3	73.5
LCS	68.2	69.3	44.2	54.0
Gloss vector	74.3	71.2	68.5	69.8
Overlaps	69.4	75.8	40.6	52.9
Leave-One-Out Ablation Results				
All	77.3	72.8	74.3	73.5
LCS	75.2	70.9	70.6	70.7
Gloss vector	75.0	74.4	61.8	67.5
Overlaps	74.8	71.9	73.8	72.8

Table 1: Results for the mixed corpus (2354 senses, 57.82% O))

7. *Vector of gloss words (SS)*
8. *Vector of gloss words (OS)*

7 Experiments

We perform 10-fold cross validation experiments on several data sets, using SVM_light (Joachims, 1999)[4] under its default settings.

Based on our random sampling of WordNet, it appears that WordNet nouns are highly skewed toward objective senses. (Esuli and Sebastiani, 2007) argue that random sampling from WordNet would yield a corpus mostly consisting of objective (neutral) senses, which would be "pretty useless as a benchmark for testing derived lexical resources for opinion mining [p. 428]." So, they use a mixture of subjective and objective senses in their data set.

To create a mixed corpus for our task, we annotated a second random sample from WordNet (which is as skewed as the previously mentioned one). We added together all of the senses of words in the lexicon which we annotated, the leftover senses from the selection of objective seed senses, and this new sample. We removed duplicates, multiple senses from the same synset, and any senses belonging to the same synset in either of the seed sets. This resulted in a corpus of 2354 senses, 993 (42.18%) of which are subjective and 1361 (57.82%) of which are objective.

The results with all of our features on this mixed corpus are given in Row 1 of Table 1. In Table 1, the

[4] http://svmlight.joachims.org/

first column identifies the features, which in this case is all of them. The next three columns show overall accuracy, and precision and recall for finding subjective senses. The baseline accuracy for the mixed data set (guessing the more frequent class, which is objective) is 57.82%. As the table shows, the accuracy is substantially above baseline.[5]

7.1 Analysis and Discussion

In this section, we seek to gain insights by performing ablation studies, evaluating our method on different data compositions, and comparing our results to previous results.

7.2 Ablation Studies

Since there are several features, we divided them into sets for the ablation studies. The vector-of-gloss-words features are the most similar to ones used in previous work. Thus, we opted to treat them as one ablation group (*Gloss vector*). The *Overlaps* group includes the $RelOverlap(t, SS)$, $RelOverlap(t, OS)$, $GlOverlap(t, SS)$, and $GlOverlap(t, OS)$ features. Finally, the *LCS* group includes the $SenseLCSscore$ and the $DomainLCSscore$ features.

There are two types of ablation studies. In the first, one group of features at a time is included. Those results are in the middle section of Table 1. Thus, for example, the row labeled *LCS* in this section is for an experiment using only the *LCS* features. In comparison to performance when all features are used, F-measure for the *Overlaps* and *LCS* ablations is significantly different at the $p < .01$ level, and, for the *Gloss Vector* ablation, it is significantly different at the $p = .052$ level (one-tailed t-test). Thus, all of the features together have better performance than any single type of feature alone.

In the second type of ablation study, we use all the features minus one group of features at a time. The results are in the bottom section of Table 1. Thus, for example, the row labeled *LCS* in this section is for an experiment using all but the *LCS* features. F-measures for *LCS* and *Gloss vector* are significantly different at the $p = .056$ and $p = .014$ levels, respectively. However, F-measure for the *Overlaps* ablation is not significantly different ($p = .39$).

[5]Note that, because the majority class is *O*, baseline recall (and thus F-measure) is 0.

Data (#senses)	Acc	P	R	F
mixed (2354 57.8% O)	77.3	72.8	74.3	73.5
strong+weak (1132)	77.7	76.8	78.9	77.8
weaksubj (566)	71.3	70.3	71.1	70.7
strongsubj (566)	78.6	78.8	78.6	78.7

Table 2: Results for different data sets (all are 50% S, unless otherwise notes)

These results provide evidence that *LCS* and *Gloss vector* are better together than either of them alone.

7.3 Results on Different Data Sets

Several methods have been developed for identifying subjective words. Perhaps an effective strategy would be to begin with a word-level subjectivity lexicon, and then perform subjectivity sense labeling to sort the subjective from objective senses of those words. We also wondered about the relative effectiveness of our method on *strongsubj* versus *weaksubj* clues.

To answer these questions, we apply the full model (again in 10-fold cross validation experiments) to data sets composed of senses of polysemous words in the subjectivity lexicon. To support comparison, all of the data sets in this section have a 50%-50% objective/subjective distribution.[6] The results are presented in Table 2.

For comparison, the first row repeats the results for the mixed corpus from Table 1. The second row shows results for a corpus of senses of a mixture of *strongsubj* and *weaksubj* words. The corpus was created by selecting a mixture of *strongsubj* and *weaksubj* words, extracting their senses and the *S*/*O* labels applied to them in Section 4, and then randomly removing senses of the more frequent class until the distribution is uniform. We see that the results on this corpus are better than on the mixed data set, even though the baseline accuracy is lower and the corpus is smaller. This supports the idea that an effective strategy would be to first identify opinion-bearing words, and then apply our method to those words to sort out their subjective and objective senses.

The third row shows results for a *weaksubj* subset

[6]As with the mixed data set, we removed from these data sets multiple senses from the same synset and any senses in the same synset in either of the seed sets.

Method	P	R	F
Our method	56.8	66.0	61.1
WM, 60% recall	44.0	66.0	52.8
SentiWordNet mapping	60.0	17.3	26.8

Table 3: Results for WM Corpus (212 senses, 76% O)

Method	A	P	R	F
Our Method	81.3%	60.3%	63.3%	61.8%
SM CV*	82.4%	70.8%	41.1%	52.0%
SM SL*	78.3%	53.0%	57.4%	54.9%

Table 4: Results for SM Corpus (484 senses, 76.9% O)

of the *strong+weak* corpus and the fourth shows results for a *strongsubj* subset that is of the same size. As expected, the results for the *weaksubj* senses are lower while those for the *strongsubj* senses are higher, as *weaksubj* clues are more ambiguous.

7.4 Comparisons with Previous Work

WM and SM address the same task as we do. To compare our results to theirs, we apply our full model (in 10-fold cross validation experiments) to their data sets.[7]

Table 3 has the WM data set results. WM rank their senses and present their results in the form of precision recall curves. The second row of Table 3 shows their results at the recall level achieved by our method (66%). Their precision at that level is substantially below ours.

Turning to ES, to create *S*/*O* annotations, we applied the following heuristic mapping (which is also used by SM for the purpose of comparison): any sense for which the sum of positive and negative scores is greater than or equal to 0.5 is S, otherwise it is O. We then evaluate the mapped tags against the gold standard of WM. The results are in Row 3 of Table 3. Note that this mapping is not fair to SentiWordNet, as the tasks are quite different, and we do not believe any conclusions can be drawn. We include the results to eliminate the possibility that their method is as good ours on our task, despite the differences between the tasks.

Table 4 has the results for the noun subset of SM's data set, which is the data set used by ES, reannotated by SM. CV* is their supervised system and SL* is their best non-supervised one. Our method has higher F-measure than the others.[8] Note that the focus of SM's work is not supervised machine learning.

8 Conclusions

In this paper, we introduced an integrative approach to automatic subjectivity word sense labeling which combines features exploiting the hierarchical structure and domain information of WordNet, as well as similarity of glosses and overlap among sets of semantically related words. There are several contributions. First, we learn several things. We found (in Section 4) that even reliable lists of subjective (opinion-bearing) words have many objective senses. We asked if word- and sense-level approaches could be used effectively in tandem, and found (in Section 7.3) that an effective strategy is to first identify opinion-bearing words, and then apply our method to sort out their subjective and objective senses. We also found (in Section 7.2) that the entire set of features gives better results than any individual type of feature alone.

Second, several of the features are novel for our task, including those exploiting the hierarchical structure of a lexical resource, domain information, and relations to seed sets expanded with monosemous senses.

Finally, the combination of our particular features is effective. For example, on senses of words from a subjectivity lexicon, accuracies range from 20 to 29 percentage points above baseline. Further, our combination of features outperforms previous approaches.

Acknowledgments

This work was supported in part by National Science Foundation awards #0840632 and #0840608. The authors are grateful to Fangzhong Su and Katja Markert for making their data set available, and to the three paper reviewers for their helpful suggestions.

[7] The WM data set is available at http://www.cs.pitt.edu/www.cs.pitt.edu/~wiebe. ES applied their method in (2006b) to WordNet, and made the results available as *SentiWordNet* at http://sentiwordnet.isti.cnr.it/.

[8] We performed the same type of evaluation as in SM's paper. That is, we assign a subjectivity label to one word sense for each synset, which is the same as applying a subjectivity label to a synset as a whole as done by SM.

References

Alina Andreevskaia and Sabine Bergler. 2006a. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *Proceedings of the 11rd Conference of the European Chapter of the Association for Computational Linguistics*.

Alina Andreevskaia and Sabine Bergler. 2006b. Sentiment tag extraction from wordnet glosses. In *Proceedings of 5th International Conference on Language Resources and Evaluation*.

Rebecca Bruce and Janyce Wiebe. 1999. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5(2):187–205.

S. Cerini, V. Campagnoni, A. Demontis, M. Formentelli, and C. Gandini. 2007. Micro-wnop: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*. Milano.

Andrea Esuli and Fabrizio Sebastiani. 2006a. Determining term subjectivity and term orientation for opinion mining. In *11th Meeting of the European Chapter of the Association for Computational Linguistics*.

Andrea Esuli and Fabrizio Sebastiani. 2006b. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, Genova, IT.

Andrea Esuli and Fabrizio Sebastiani. 2007. PageRanking wordnet synsets: An application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 424–431, Prague, Czech Republic, June.

A. Gliozzo, C. Strapparava, E. d'Avanzo, and B. Magnini. 2005. Automatic acquisition of domain specific lexicons. Tech. report, IRST, Italy.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Scholkopf, C. Burgess, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, Cambridge, MA. MIT-Press.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the Twentieth International Conference on Computational Linguistics*, pages 1267–1373, Geneva, Switzerland.

Soo-Min Kim and Eduard Hovy. 2006. Identifying and analyzing judgment opinions. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 200–207, New York.

M.E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference 1986*, Toronto, June.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* , pages 271–278, Barcelona, ES. Association for Computational Linguistics.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. International Joint Conference on Artificial Intelligence.*

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Conference on Empirical Methods in Natural Language Processing*, pages 105–112.

Fangzhong Su and Katja Markert. 2008. From word to sense: a case study of subjectivity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester.

M. Taboada, C. Anthony, and K. Voll. 2006. Methods for creating semantic orientation databases. In *Proceedings of 5th International Conference on Language Resources and Evaluation* .

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2006. Latent variable models for semantic orientations of phrases. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics* , Trento, Italy.

P. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia.

Alessandro Valitutti, Carlo Strapparava, and Oliviero Stock. 2004. Developing affective lexical resources. *PsychNology Journal*, 2(1):61–83.

J. Wiebe and R. Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics* , pages 486–497, Mexico City, Mexico.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing* , pages 347–354, Vancouver, Canada.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Conference on Empirical Methods in Natural Language Processing* , pages 129–136, Sapporo, Japan.