# Subjectivity Word Sense Disambiguation

**Cem Akkaya** and **Janyce Wiebe**
University of Pittsburgh
{cem,wiebe}@cs.pitt.edu

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

## Abstract

This paper investigates a new task, *subjectivity word sense disambiguation (SWSD)*, which is to automatically determine which word instances in a corpus are being used with subjective senses, and which are being used with objective senses. We provide empirical evidence that SWSD is more feasible than full word sense disambiguation, and that it can be exploited to improve the performance of contextual subjectivity and sentiment analysis systems.

## 1 Introduction

The 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. Examples of such words are the following (in bold):

(1) He is a **disease** to every team he has gone to.
Converting to SMF is a **headache**.
The concert left me **cold**.
That guy is such a **pain**.

Knowing the meaning (and thus subjectivity) of these words would help a system recognize the negative sentiments in these sentences.

Most subjectivity lexicons are compiled as lists of keywords, rather than word meanings (senses). However, many keywords have both subjective and objective senses. False hits – subjectivity clues used with objective senses – are a significant source of error in subjectivity and sentiment analysis. For example, even though the following sentence contains all of the negative keywords above, it is nevertheless objective, as they are all false hits:

(2) Early symptoms of the **disease** include severe **headaches**, red eyes, fevers and **cold** chills, body **pain**, and vomiting.

To tackle this source of error, we define a new task, *subjectivity word sense disambiguation (SWSD)*, which is to automatically determine which word instances in a corpus are being used with subjective senses, and which are being used with objective senses. We hypothesize that SWSD is more feasible than full word sense disambiguation, because it is more coarse grained – often, the exact sense need not be pinpointed. We also hypothesize that SWSD can be exploited to improve the performance of contextual subjectivity analysis systems via sense-aware classification.

The paper consists of two parts. In the first part, we build and evaluate a targeted supervised SWSD system that aims to disambiguate members of a subjectivity lexicon. It labels clue instances as having a subjective sense or an objective sense in context. The system relies on common machine learning features for word sense disambiguation (WSD). The performance is substantially above both baseline and the performance of full WSD on the same data, suggesting that the task is feasible, and that subjectivity provides a natural coarse-grained grouping of senses.

The second part demonstrates the promise of SWSD for contextual subjectivity analysis. First, we show that subjectivity sense ambiguity is highly prevalent in the MPQA opinion-annotated corpus (Wiebe et al., 2005; Wilson, 2008), thus establishing the potential benefit of performing SWSD. Then, we exploit SWSD to improve performance on several subjectivity analysis tasks, from subjective/objective sentence-level classification to positive/negative/neutral expression-level classification. To our knowledge, this is the

first attempt to explicitly use sense-level subjectivity tags in contextual subjectivity and sentiment analysis.

## 2 Background

We adopt the definitions of *subjective* and *objective* from (Wiebe et al., 2005; Wiebe and Mihalcea, 2006; Wilson, 2008). Subjective expressions are words and phrases being used to express mental and emotional states, such as speculations, evaluations, sentiments, and beliefs. A general covering term for such states is *private state* (Quirk et al., 1985), an internal state that cannot be directly observed or verified by others. (Wiebe and Mihalcea, 2006) give the following examples:

(3) 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 review mining, for example, we want to know whether an opinion about a product is positive or negative. Nonetheless, as argued by (Wiebe and Mihalcea, 2006; Su and Markert, 2008), there are also motivations for a separate subjective/objective (*S/O*) classification.

First, expressions may be subjective but not have any particular polarity. An example given by (Wilson et al., 2005a) 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, benefits for sentiment analysis can be realized by decomposing the problem into *S*/*O* (or neutral versus polar) and polarity classification (Yu and Hatzivassiloglou, 2003; Pang and Lee, 2004; Wilson et al., 2005a; Kim and Hovy, 2006). We will see further evidence of this in Section 4.2.3 in this paper.

The contextual subjectivity analysis experiments in Section 4 include both *S*/*O* and polarity classifications. The data used in those experiments is from the MPQA Corpus (Wiebe et al., 2005; Wilson, 2008),[1] which consists of texts from the world press annotated for subjective expressions. In the MPQA Corpus, subjective expressions of varying lengths are marked, from single words to long phrases. In addition, other properties are annotated, including polarity.

For SWSD, we need the notions of subjective and objective *senses* of words in a dictionary. We adopt the definitions from (Wiebe and Mihalcea, 2006), who describe the annotation scheme as follows. Classifying a sense as *S* means that, when the sense is used in a text or conversation, one expects it to express subjectivity, and also that the phrase or sentence containing it expresses subjectivity. As noted in (Wiebe and Mihalcea, 2006), sentences containing objective senses may not be objective. Thus, objective senses are defined as follows: Classifying a sense as *O* means that, when the sense is used in a text or conversation, one does not expect it to express subjectivity and, if the phrase or sentence containing it is subjective, the subjectivity is due to something else. Finally, classifying a sense as *B* means it covers both subjective and objective usages.

The following subjective examples are given in (Wiebe and Mihalcea, 2006):

---

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")

---

They give the following objective examples:

---

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)

---

Wiebe and Mihalcea performed an agreement study and report that good agreement ($\kappa$=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 ($\kappa$=0.79).

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

## 3 Subjectivity Word Sense Disambiguation

### 3.1 Task Definition and Method

We now turn to SWSD, and our method for performing it.

Note that SWSD is midway between pure dictionary classification and pure contextual interpretation. For SWSD, the context of the word is considered in order to *perform* the task, but the subjectivity is determined solely by the dictionary. In contrast, full contextual interpretation can deviate from a sense's subjectivity label in the dictionary. As noted above, words used with objective senses may appear in subjective expressions. For example, an SWSD system would label the following examples of alarm as *S*, *O* and *O*, respectively. On the other hand, a sentence-level subjectivity classifier would label the sentences as *S*, *S*, and *O*, respectively.

(4) His **alarm** grew.
Will someone shut that darn **alarm** off?
The **alarm** went off.

We use a supervised approach to SWSD. We train a different classifier for each lexicon entry for which we have training data. Thus, our approach is like targeted WSD (in contrast to all-words WSD), with two labels: *S* and *O*.

We borrow machine learning features which have been successfully used in WSD. Specifically, given an ambiguous target word, we use the following features from (Mihalcea, 2002):

**CW** : the target word itself
**CP** : POS of the target word
**CF** : surrounding context of 3 words and their POS
**HNP** : the head of the noun phrase to which the target word belongs
**NB** : the first noun before the target word
**VB** : the first verb before the target word
**NA** : the first noun after the target word
**VA** : the first verb after the target word
**SK** : at most 10 context words occurring at least 5 times; determined for each sense

### 3.2 Lexicon and Data

Our target words are members of a subjectivity lexicon, because, since they are in such a lexicon, we know they have subjective usages. Specifically, we use the lexicon of (Wilson et al., 2005b; Wilson, 2008).[2] The entries have been divided into those that are strongly subjective (*strongsubj*) and those that are weakly subjective (*weaksubj*), reflecting their reliability as subjectivity clues. The sources of the entries in the lexicon are identified in (Wilson, 2008). In the second part of this paper, we evaluate systems against the MPQA corpus. Wilson also uses this corpus for her evaluations. To enable this, entries were added to the lexicon independently from the MPQA corpus (that is, none of the entries were derived using the MPQA corpus).

The training and test data for SWSD consists of word instances in a corpus labeled as *S* or *O*, indicating whether they are used with a subjective or objective sense. Because we do not have data labeled with the *S*/*O* coarse-grained senses and we did not want to undertake the annotation effort at this stage, we created an annotated corpus by combining two types of sense annotations: (1) labels of senses within a dictionary as *S* or *O* (i.e., subjectivity sense labels), and (2) sense tags of word instances in a corpus (i.e., sense-tagged data). The subjectivity sense labels are used to collapse the sense labels in the sense-tagged data into the two new senses, *S* and *O*.

Our sense-tagged data are the lexical sample corpora (training and test data) from SENSEVAL1 (Kilgarriff and Palmer, 2000), SENSEVAL2 (Preiss and Yarowsky, 2001), and SENSEVAL3 (Mihalcea and Edmonds, 2004). We selected all of the SENSEVAL words that are also in the subjectivity lexicon, and labeled their dictionary senses as *S*, *O*, or *B* according to the annotation scheme described above in Section 2. We did this subjectivity sense labeling according to the sense inventory of the underlying corpus (Hector for SENSEVAL1; WordNet1.7 for SENSEVAL2; and WordNet1.7.1 for SENSEVAL3).

Among the words, we found that 11 are not ambiguous - either they have only *S* or only *O* senses (in the corresponding sense inventory), or the senses of their instances in the SENSEVAL data are all *S* or all *O*. So as not to inflate our results, we removed those 11 from the data, leaving 39 words. In addition, we excluded the senses labeled *B* (a total of 10 senses). This leaves a total of 372 senses: 9 words (64 senses) from SENSEVAL1, 18 words (201 senses) from SENSEVAL2, and 12 words (107 senses) from SENSEVAL3.

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

| | Base | Acc | SP | SR | SF | OP | OR | OF | IB | EB(%) |
|---|---|---|---|---|---|---|---|---|---|---|
| All | 79.9 | 88.3 | 89.3 | 89.1 | 89.2 | 87.1 | 87.4 | 87.2 | 8.4 | 41.8 |
| S1 | 57.9 | 80.7 | 81.1 | 78.3 | 79.7 | 80.2 | 82.9 | 81.5 | 22.8 | 54.2 |
| S2 | 81.1 | 87.3 | 86.5 | 85.2 | 85.8 | 87.9 | 89.0 | 88.4 | 6.2 | 32.8 |
| S3 | 95.0 | 96.4 | 96.5 | 99.0 | 97.7 | 96.3 | 87.8 | 91.8 | 1.4 | 28.0 |

Table 1: Overall SWSD results (micro averages). *Base* is majority-class baseline; *Acc* is accuracy; *SP*, *SR*, and *SF* are subjective precision, recall and F-measure; similarly for *OP*, *OR*, and *OF*. *IB* is absolute improvement in Acc over Base; *EB* is percent error reduction in Acc.

### 3.3 SWSD Experiments

In this section, we evaluate our SWSD system, and compare its performance to an WSD system on the same data.

Note that, although generally in the SENSEVAL datasets, training and test data are provided separately, a few target words from SENSEVAL1 do not have both training and testing data. Thus, we opted to combine the training and test data into one dataset, and then perform 10-fold cross validation experiments.

For our classifier, we use the SVM classifier from the Weka package (Witten and Frank., 2005) with its default settings.

We were interested in how well the system would perform on more and less ambiguous words. Thus, we split the words into three subsets according to their majority-class baselines, and report separate results: *S1* (9 words), *S2* (18 words), and *S3* (12 words) have majority-class baselines in the intervals [50%,70%) , [70%,90%), and [90%,100%), respectively.

Table 1 contains the results, giving the overall results (micro averages), as well as results for the subsets *S1*, *S2*, and *S3*.

The improvement for SWSD over baseline is especially high for the less skewed set, *S1*. This is very encouraging because these words are the more ambiguous words, and thus are the ones that most need SWSD (assuming the SENSEVAL priors are similar to the priors in the corpus). The average error reduction over baseline for *S1* words is 54.2%. Even for the more skewed sets *S2* and *S3*, reductions are 32.8% and 28.0%, respectively, with an overall reduction of 41.8%.

To compare SWSD with WSD, we re-ran the 10-fold cross validation experiments, but this time using the original sense labels, rather than *S* and *O*. The (micro-averaged) accuracy is 67.9%, much lower than the overall accuracy for SWSD (88.3%).

The positive results provide evidence that SWSD is a feasible variant of WSD, and that the *S*/*O* sense groupings are natural ones, since the system is able to learn to distinguish between them with high accuracy. There is also potential for improvement by using a richer feature set, including subjectivity features.

## 4 Opinion Analysis with Subjectivity Word Sense Disambiguation

In this section, we explore the promise of SWSD for contextual subjectivity analysis. First, we provide evidence that a subjectivity lexicon can have substantial coverage of the subjective expressions in a corpus, yet still be responsible for significant subjectivity sense ambiguity in that corpus. Then, we exploit SWSD in several contextual opinion analysis systems, comparing the performance of sense-aware and non-sense-aware versions. They are all variations of components of the OpinionFinder opinion recognition system.[3]

### 4.1 Coverage and Ambiguity of Lexicon Entries in the MPQA Corpus

In this section, we consider the distribution of lexicon entries in the MPQA corpus.

The lexicon covers a substantial subset of the subjective expressions in the corpus: 67.1% of the subjective expressions contain one or more lexicon entries.

On the other hand, fully 42.9% of the instances of the lexicon entries in the MPQA corpus are not in subjective expressions. An instance that is not in a subjective expression is, by definition, being used with an objective sense. Thus, these instances are false hits of subjectivity clues. As mentioned above, the entries in the lexicon have been pre-classified as either more (*strongsubj*) or less (*weaksubj*) reliable. We see this difference reflected in their degree of ambiguity – 53% of the

[3] Available at http://www.cs.pitt.edu/opin

*weaksubj* instances are false hits, while only 22% of the *strongsubj* instances are.

The high coverage of the lexicon demonstrates its potential usefulness for opinion analysis systems, while its degree of ambiguity, in the form of false hits in a subjectivity annotated corpus, shows the potential benefit to opinion analysis of performing SWSD.

As mentioned above, our experiments involve only lexicon entries that are covered by the SENSEVAL data, as we did not perform manual sense tagging for this work. We have hope to expand the system's coverage in the future, as more word-sense tagged data is produced (e.g., ONTONOTES (Hovy et al., 2006)). We also have evidence that a moderate amount of manual annotation would be worth the effort. For example, let us order the lexicon entries from highest to lowest by frequency in the MPQA corpus. The top 20 are responsible for 25% of all false hits in the corpus; the top 40 are responsible for 34%; and the top 80 are responsible for 44%. If the SWSD system could be trained for these words, the potential impact on reducing false hits could be substantial, especially considering the good performance of the SWSD system on the more ambiguous words. Note that we do not want to simply discard these clues. The top 20 cover 9.4% of all subjective expressions; the top 40 cover 15.4%; and the top 80 cover 29.5%. Note that SWSD only needs the data annotated with the coarse-grained binary labels, which should be less time consuming to produce than full word sense tags.

## 4.2 Contextual Classification

We found in Section 3.3 that SWSD is a feasible task and then in Section 4.1 that there is a great deal of subjectivity sense ambiguity in a standard subjectivity-annotated corpus (MPQA). We now turn to exploiting the results of SWSD to automatically recognize subjectivity and sentiment in the MPQA corpus.

A motivation for using the MPQA data is that many types of classifiers have been evaluated on it, and we can directly test the effect of SWSD on these classifiers.

Note that, for the SWSD experiments, the number of words does not limit the amount of data, as SENSEVAL provides data for each word. However, the only parts of the MPQA corpus for which SWSD could affect performance is the subset containing instances of the words in the SWSD system's coverage. Thus, for the classifiers in this section, the data used is the *SenMPQA* dataset, which consists of the sentences in the MPQA Corpus that contain at least one instance of the 39 keywords. There are 689 such sentences (containing, in total, 723 instances of the 39 keywords).

Even though this dataset is smaller than the one used above, it gives us enough data to draw conclusions according to McNemar's test for statistical significance.

### 4.2.1 Rule-based Classifier

We first apply SWSD to the rule-based classifier from (Riloff and Wiebe, 2003). The classifier, which is a sentence-level *S/O* classifier, has low subjective and objective recall but high subjective and objective precision. It is useful for creating training data for subsequent processing by applying it to large amounts of unannotated data.

The classifier is a good candidate for directly measuring the effects of SWSD on contextual subjectivity analysis, because it classifies sentences only by looking for the presence of subjectivity keywords. Performance will improve if false hits can be ignored.

The classifier labels a sentence as *S* if it contains two or more *strongsubj* clues. On the other hand, it considers three conditions to classify a sentence as *O*: there are no *strongsubj* clues in the current sentence, there are together at most one *strongsubj* clue in the previous and next sentence, and there are together at most 2 *weaksubj* clues in the current, previous, and next sentence. A sentence that is not labeled *S* or *O* is labeled *unknown*.

The rule-based classifier is made sense aware by making it blind to the target word instances labeled *O* by the SWSD system, as these represent false hits of subjectivity keywords. We compare this sense-aware method (*SE*), with the original classifier ($O_{RB}$), in order to see if SWSD would improve performance. We also built another modified rule-based classifier *RE* to demonstrate the effect of randomly ignoring subjectivity keywords. *RE* ignores a keyword instance randomly with a probability of 0.429, the expected value of false hits in the MPQA corpus. The results are listed in Table 2.

The rule-based classifier looks for the presence of the keywords to find subjective sentences and for the absence of the keywords to find objective sentences. It is obvious that a variant working on

| | Acc | OP | OR | OF | SP | SR | SF |
|---|---|---|---|---|---|---|---|
| $O_{RB}$ | 27.0 | 50.0 | 4.1 | 7.6 | 92.7 | 36.0 | 51.8 |
| SE | 28.3 | 62.1 | 9.3 | 16.1 | 92.7 | 35.8 | 51.6 |
| RE | 27.6 | 48.4 | 7.7 | 13.3 | 92.6 | 35.4 | 51.2 |

Table 2: Effect of SWSD on the rule-based classifiers.

fewer keyword instances than $O_{RB}$ will always have the same or higher objective recall and the same or lower subjective recall than $O_{RB}$. That is the case for both *SE* and *RE*. The real benefit we see is in objective precision, which is substantially higher for *SE* than $O_{RB}$. For our experiments, *OP* gives a better idea of the impact of SWSD, because most of the keyword instances SWSD disambiguates are *weaksubj* clues, and *weaksubj* keywords figure more prominently in objective classification. On the other hand, *RE* has both lower *OP* and *SP* than $O_{RB}$. Note that accuracy for all three systems is low, because all *unknown* predictions are counted as incorrect.

These findings suggest that SWSD performs well on disambiguating keyword instances in the MPQA corpus,[4] and demonstrates a positive impact of SWSD on sentence-level subjectivity classification.

#### 4.2.2 Subjective/Objective Classifier

We now move to more fine-grained expression-level subjectivity classification. Since sentences often contain multiple subjective expressions, expression-level classification is more informative than sentence-level classification.

The classifier in this section is an implementation of the *neutral/polar* supervised classifier of (Wilson et al., 2005a) (using the same features), except that the classes are *S*/*O* rather than *neutral/polar*. These classifiers label instances of lexicon entries. The gold standard is defined on the MPQA Corpus as follows: If an instance is in a subjective expression, it is contextually *S*. If the instance is in an objective expression, it is contextually *O*. We evaluate the system on the 723 clue instances in the SenMPQA dataset.

We incorporate SWSD information into the contextual subjectivity classifier in a straightforward fashion: outputs are modified according to simple, intuitive rules.

[4]which we cannot evaluate directly, as the MPQA corpus is not sense tagged.

Our strategy is defined by the relation between sense subjectivity and contextual subjectivity and involves two rules, *R1* and *R2*.

We know that a keyword instance used with a *S* sense must be in a subjective expression. *R1* is to simply trust SWSD: If the contextual classifier labels an instance as *O*, but SWSD determines that it has an *S* sense, then *R1* flips the contextual classifier's label to *S*.

Things are not as simple in the case of *O* senses, since they may appear in both subjective and objective expressions. We will state *R2*, and then explain it: If the contextual classifier labels an instance as *S*, but (1) SWSD determines that it has an *O* sense, (2) the contextual classifier's confidence is low, and (3) there is no other subjective keyword in the same expression, then *R2* flips the contextual classifier's label to *O*. First, consider confidence: though a keyword with an *O* sense may appear in either subjective or objective expressions, it is more likely to appear in an objective expression. We assume that this is reflected to some extent in the contextual classifier's confidence. Second, if a keyword with an *O* sense appears in a subjective expression, then the subjectivity is not due to that keyword but rather due to something else. Thus, the presence of another lexicon entry "explains away" the presence of the *O* sense in the subjective expression, and we do not want SWSD to overrule the contextual classifier. Only when the contextual classifier isn't certain and only when there isn't another keyword does *R2* flip the label to *O*.

Our definition of low confidence is in terms of the label weights assigned by BoosTexter (Schapire and Singer, 2000), which is the underlying machine learning algorithm of the classifier. We use the difference between the largest label weight and the second largest label weight as a measure of confidence, as suggested in the BoosTexter documentation. The threshold we use is 0.0008.[5]

We apply the contextual classifier and the SWSD system to the data, and compare the performance of the original system ($O_{S/O}$) and three sense-aware variants: one using only *R1*, one us-

[5]As will be noted below, we experimented with three thresholds for the classifier in Section 4.2.3, with no significant difference in accuracy. Here, we simply adopt 0.0008, without further experimentation. In addition, we did not experiment with other conditions than those incorporated in the two rules in this section and the two rules in Section 4.2.3 below.

| | Acc | OP | OR | OF | SP | SR | SF |
|---|---|---|---|---|---|---|---|
| $O_{S/O}$ | 75.4 | 68.0 | 62.9 | 65.4 | 79.2 | 82.7 | 80.9 |
| R1 | 77.7 | 75.5 | 58.8 | 66.1 | 78.6 | 88.8 | 83.4 |
| R2 | 79.0 | 67.3 | 83.9 | 74.7 | 89.0 | 76.1 | 82.0 |
| R1R2 | 81.3 | 72.5 | 79.8 | 75.9 | 87.4 | 82.2 | 84.8 |

Table 3: Effect of SWSD on the subjective/objective classifier

ing only *R2*, and one using both (*R1R2*). The results are in Table 3. The *R1* variant shows an improvement of 2.3 points in accuracy (a 9.4% error reduction). The *R2* variant shows an improvement of 3.6 points in accuracy (a 14.6% error reduction). Applying both rules (*R1R2*) gives an improvement of 5.9 percentage points in accuracy (a 24% error reduction).

In our case, a paired t-test is not appropriate to measure statistical significance, as we are not doing multiple runs. Thus, we apply McNemar's test, which is a non-parametric method for algorithms that can be executed only once, meaning training once and testing once (Dietterich, 1998). For *R1*, the improvement in accuracy is statistically significant at the $p < .05$ level. For *R2* and *R1R2*, the improvement in accuracy is statistically significant at the $p < .01$ level. Moreover, in all cases, we see improvement in both objective and subjective F-measure.

#### 4.2.3 Contextual Polarity Classifier

We now apply SWSD to contextual polarity classification (positive/negative/neutral), in the hope that avoiding false hits of subjectivity keywords will also lead to performance improvement in contextual sentiment analysis.

We use an implementation of the classifier of (Wilson et al., 2005a). This classifier labels instances of lexicon entries. The gold standard is defined on the MPQA Corpus as follows: If an instance is in a positive subjective expression, it is contextually positive (*Ps*); if in a negative subjective expression, it is contextually negative (*Ng*); and if it is in an objective expression or a neutral subjective expression, then it is contextually *N(eutral)*. As above, we evaluate the system on the keyword instances in the SenMPQA dataset.

Wilson et al. use a two step approach. The first step classifies keyword instances as being in a polar (positive or negative) or a neutral context. The first step is performed by the neutral/polar classifier mentioned above in Section 4.2.2. The second step decides the contextual polarity (positive or negative) of the instances classified as polar in the first step, and is performed by a separate classifier.

To make a sense-aware version of the system, we use rules to change some of the answers of the neutral/polar classifier.

Unfortunately, we cannot simply trust SWSD when it labels a keyword as an *S* sense, because an *S* sense might be in a *N(eutral)* expression (since there are neutral subjective expressions). But, an *S* sense is more likely to appear in a *P(olar)* expression. Thus, we consider confidence (rule *R3*): If the contextual classifier labels an instance as *N*, but SWSD determines it has an *S* sense and the contextual classifier's confidence is low,[6] then *R3* flips the contextual classifier's label to *P*.

Rule *R4* is analogous to *R2* in the previous section: If the contextual classifier labels an instance as *P*, but (1) SWSD determines that it has an *O* sense, (2) the contextual classifier's confidence is low, and (3) there is no other subjective keyword in the same expression, then *R2* flips the contextual classifier's label to *N*.

We compare the performance of the original neutral/polar classifier ($O_{N/P}$) and sense-aware variants using *R3* and *R4*. The results are in Table 4. This time, the table does not include a combined method, because only *R4* improves performance. This is consistent with the finding in (Wilson et al., 2005a) that most errors are caused by subjectivity keywords with non-neutral prior polarity appearing in phrases with neutral contextual polarity. *R4* targets these cases. It is promising to see that SWSD provides enough information to fix some of them. There is a 2.6 point improvement in accuracy (a 12.4% error reduction). The improvement in accuracy is statistically significant at the $p < .01$ level with McNemar's test. The improvement in accuracy is accompanied by improvements in both neutral and polar F-measure.

We wanted to see if the improvements in the

[6] As in the previous section, low confidence is defined in terms of the difference between the largest label weight and the second largest label weight assigned by BoosTexter. We tried three thresholds, 0.0007, 0.0008, and 0.0009, resulting in only a slight difference in accuracy: 0.0007 and 0.0009 both give 81.5 accuracy compared to 81.6 accuracy for 0.0008. We report results using 0.0008, though the accuracy using the other thresholds is statistically significantly better than the accuracy of the original classifier at the same level.

| | Acc | NP | NR | NF | NgP | NgR | NgF | PsP | PsR | PsF |
|---|---|---|---|---|---|---|---|---|---|---|
| $O_{Ps/Ng/N}$ | 77.6 | 80.9 | 94.6 | 87.2 | 60.4 | 29.4 | 39.5 | 52.2 | 32.4 | 40.0 |
| R4 | 80.6 | 81.2 | 98.7 | 89.1 | 82.1 | 29.4 | 43.2 | 68.6 | 32.4 | 44.0 |

Table 5: Effect of SWSD on the contextual polarity classifier

| | Acc | NP | NR | NF | PP | PR | PF |
|---|---|---|---|---|---|---|---|
| $O_{N/P}$ | 79.0 | 81.5 | 92.5 | 86.7 | 65.8 | 40.7 | 50.3 |
| R3 | 70.0 | 83.7 | 73.8 | 78.4 | 44.4 | 59.3 | 50.8 |
| R4 | 81.6 | 81.7 | 96.8 | 88.6 | 81.1 | 38.6 | 52.3 |

Table 4: Effect of SWSD on the neutral/polar classifier

first step of Wilson et al's system can be propagated to their second step, yielding an overall improvement in positive /negative/neutral (*Ps*/*Ng*/*N*) classification.

The sense-aware variant of the overall two-part system is the same as the original except that we apply *R4* to the output of the first step (flipping some of the neutral/polar classifier's *P* labels to *N*). Thus, since the second step in Wilson et al.'s classifier processes only those instances labeled *P* in the first step, in the sense-aware system, fewer instances are passed from the first to the second step.

Table 5 reports results for the original system ($O_{Ps/Ng/N}$) and the sense-aware variant (*R4*). These results are for the entire SenMPQA dataset, not just those labeled *P* in the first step.

The accuracy improves 3 percentage points (a 13.4% error reduction). The improvement in accuracy is statistically significant at the $p < .01$ level with McNemar's test. We see the real benefit when we look at the precision of the positive and negative classes. Negative precision goes from 60.4 to 82.1 and positive precision goes from 52.2 to 68.6, with no loss in recall. This is evidence that the SWSD system is doing a good job of removing some false hits of subjectivity clues that harm the original version of the system.

## 5 Comparisons to Previous Work

Several researchers exploit lexical resources for contextual subjectivity and sentiment analysis. These systems typically look for the presence of subjective or sentiment-bearing words in the text. They may rely only on this information (e.g., (Turney, 2002; Whitelaw et al., 2005; Riloff and Wiebe, 2003)), or they may combine it with additional information as well (e.g., (Yu and Hatzivassiloglou, 2003; Kim and Hovy, 2004; Bloom et al., 2007; Wilson et al., 2005a)). We apply SWSD to some of those systems to show the effect of SWSD on contextual subjectivity and sentiment analysis.

Another set of related work is on subjectivity and polarity labeling of word senses (e.g. (Esuli and Sebastiani, 2006; Andreevskaia and Bergler, 2006; Wiebe and Mihalcea, 2006; Su and Markert, 2008)). They label senses of words in a dictionary. In comparison, we label senses of word instances in a corpus.

Moreover, our work extends findings in (Wiebe and Mihalcea, 2006) and (Su and Markert, 2008). (Wiebe and Mihalcea, 2006) demonstrates that subjectivity is a property that can be associated with word senses. We show that it is a natural grouping of word senses and that it provides a principled way for clustering senses. They also demonstrate that subjectivity helps with WSD. We show that a coarse-grained WSD variant (SWSD) helps with subjectivity and sentiment analysis. Both (Wiebe and Mihalcea, 2006) and (Su and Markert, 2008) show that even reliable subjectivity clues have objective senses. We demonstrate that this ambiguity is also prevalent in a corpus.

Several researchers (e.g., (Palmer et al., 2004; Navigli, 2006; Snow et al., 2007; Hovy et al., 2006)) work on reducing the granularity of sense inventories for WSD. They aim for a more coarse-grained sense inventory to overcome performance shortcomings related to fine-grained sense distinctions. Our work is similar in the sense that we reduce all senses of a word to two senses (*S*/*O*). The difference is the criterion driving the grouping. Related work concentrates on syntactic and semantic similarity between senses to group them. In contrast, our grouping is driven by subjectivity with a specific application area in mind, namely subjectivity and sentiment analysis.

## 6 Conclusions and Future Work

We introduced the task of subjectivity word sense disambiguation (SWSD), and evaluated a supervised method inspired by research in WSD. The

system achieves high accuracy, especially on highly ambiguous words, and substantially outperforms WSD on the same data. The positive results provide evidence that SWSD is a feasible variant of WSD, and that the *S*/*O* sense groupings are natural ones.

We also explored the promise of SWSD for contextual subjectivity analysis. We showed that a subjectivity lexicon can have substantial coverage of the subjective expressions in the corpus, yet still be responsible for significant sense ambiguity. This demonstrates the potential benefit to opinion analysis of performing SWSD. We then exploit SWSD in several contextual opinion analysis systems, including positive/negative/neutral sentiment classification. Improvements in performance were realized for all of the systems.

We plan several future directions which promise to further increase the impact of SWSD on subjectivity and sentiment analysis. We will manually annotate a moderate number of strategically chosen words, namely frequent ones which are highly ambiguous. In addition, we will add features to the SWSD system reflecting the subjectivity of the surrounding context. Finally, there are more sophisticated strategies to explore for improving subjectivity and sentiment analysis via SWSD than the simple, intuitive rules we began with in this paper.

## Acknowledgments

This material is based in part upon work supported by National Science Foundation awards #0840632 and #0840608. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

A. Andreevskaia and S. Bergler. 2006. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *(EACL-2006)*.

K. Bloom, N. Garg, and S. Argamon. 2007. Extracting appraisal expressions. In *HLT-NAACL 2007*, pages 308–315, Rochester, NY.

T. G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.

A. Esuli and F. Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *(LREC-06)*, Genova, IT.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York City.

A. Kilgarriff and M. Palmer, editors. 2000. *Computer and the Humanities. Special issue: SENSEVAL. Evaluating Word Sense Disambiguation programs*, volume 34, April.

S.-M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *(COLING 2004)*, pages 1267–1373, Geneva, Switzerland.

S.-M. Kim and E. Hovy. 2006. Identifying and analyzing judgment opinions. In *(HLT/NAACL-06)*, pages 200–207, New York, New York.

R. Mihalcea and P. Edmonds, editors. 2004. *Proceedings of SENSEVAL-3, Association for Computational Linguistics Workshop*, Barcelona, Spain.

R. Mihalcea. 2002. Instance based learning with automatic feature selection applied to Word Sense Disambiguation. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, August.

R. Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia.

M. Palmer, O. Babko-Malaya, and H. T. Dang. 2004. Different sense granularities for different applications. In *HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding*, Boston, Massachusetts.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *(ACL-04)*, pages 271–278, Barcelona, ES. Association for Computational Linguistics.

J. Preiss and D. Yarowsky, editors. 2001. *Proceedings of SENSEVAL-2, Association for Computational Linguistics Workshop*, Toulouse, France.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, New York.

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *(EMNLP-2003)*, pages 105–112, Sapporo, Japan.

R. E. Schapire and Y. Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

R. Snow, S. Prakash, D. Jurafsky, and A. Ng. 2007. Learning to merge word senses. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic.

F. Su and K. Markert. 2008. From word to sense: a case study of subjectivity recognition. In *(COLING-2008)*, Manchester.

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 (ACL 2002)*, pages 417–424, Philadelphia.

C. Whitelaw, N. Garg, and S. Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*, Bremen, DE.

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

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, 39(2/3):164–210.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005a. Recognizing contextual polarity in phrase-level sentiment analysis. In *(HLT/EMNLP-2005)*, pages 347–354, Vancouver, Canada.

T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. 2005b. OpinionFinder: A system for subjectivity analysis. In *Proc. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005) Companion Volume (software demonstration)*.

T. Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of private states*. Ph.D. thesis, Intelligent Systems Program, University of Pittsburgh.

I. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, June.

H. Yu and V. 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 (EMNLP-03)*, pages 129–136, Sapporo, Japan.