# Word Sense and Subjectivity

**Janyce Wiebe**
Department of Computer Science
University of Pittsburgh
wiebe@cs.pitt.edu

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

## Abstract

Subjectivity and meaning are both important properties of language. This paper explores their interaction, and brings empirical evidence in support of the hypotheses that (1) subjectivity is a property that can be associated with word senses, and (2) word sense disambiguation can directly benefit from subjectivity annotations.

## 1 Introduction

There is growing interest in the automatic extraction of opinions, emotions, and sentiments in text (*subjectivity*), to provide tools and support for various NLP applications. Similarly, there is continuous interest in the task of *word sense disambiguation*, with sense-annotated resources being developed for many languages, and a growing number of research groups participating in large-scale evaluations such as SENSEVAL.

Though both of these areas are concerned with the *semantics* of a text, over time there has been little interaction, if any, between them. In this paper, we address this gap, and explore possible interactions between subjectivity and word sense.

There are several benefits that would motivate such a joint exploration. First, at the resource level, the augmentation of lexical resources such as WordNet (Miller, 1995) with subjectivity labels could support better subjectivity analysis tools, and principled methods for refining word senses and clustering similar meanings. Second, at the tool level, an explicit link between subjectivity and word sense could help improve methods for each, by integrating features learned from one into the other in a pipeline approach, or through joint simultaneous learning.

In this paper we address two questions about word sense and subjectivity. First, can subjectivity labels be assigned to word senses? To address this question, we perform two studies. The first (Section 3) investigates agreement between annotators who manually assign the labels *subjective*, *objective*, or *both* to WordNet senses. The second study (Section 4) evaluates a method for automatic assignment of subjectivity labels to word senses. We devise an algorithm relying on distributionally similar words to calculate a subjectivity score, and show how it can be used to automatically assess the subjectivity of a word sense.

Second, can automatic subjectivity analysis be used to improve word sense disambiguation? To address this question, the output of a subjectivity sentence classifier is input to a word-sense disambiguation system, which is in turn evaluated on the nouns from the SENSEVAL-3 English lexical sample task (Section 5). The results of this experiment show that a subjectivity feature can significantly improve the accuracy of a word sense disambiguation system for those words that have both subjective and objective senses.

A third obvious question is, can word sense disambiguation help automatic subjectivity analysis? However, due to space limitations, we do not address this question here, but rather leave it for future work.

## 2 Background

Subjective expressions are words and phrases being used to express opinions, emotions, evaluations, speculations, etc. (Wiebe et al., 2005). A general covering term for such states is *private state*, "a state that is not open to objective obser-

vation or verification" (Quirk et al., 1985).[1] There are three main types of subjective expressions:[2]

(1) references to private states:

> His **alarm** grew.
> He **absorbed** the information quickly.
> He was **boiling with anger**.

(2) references to speech (or writing) events expressing private states:

> UCC/Disciples leaders **roundly condemned** the Iranian President's **verbal assault** on Israel.
>
> The editors of the left-leaning paper **attacked** the new House Speaker.

(3) expressive subjective elements:

> He would be **quite a catch**.
> **What's the catch?**
> That doctor is a **quack.**

Work on automatic subjectivity analysis falls into three main areas. The first is identifying words and phrases that are associated with subjectivity, for example, that *think* is associated with private states and that *beautiful* is associated with positive sentiments (e.g., (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Kamps and Marx, 2002; Turney, 2002; Esuli and Sebastiani, 2005)). Such judgments are made for words. In contrast, our end task (in Section 4) is to assign subjectivity labels to word *senses*.

The second is subjectivity classification of sentences, clauses, phrases, or word instances in the context of a particular text or conversation, either subjective/objective classifications or positive/negative sentiment classifications (e.g.,(Riloff and Wiebe, 2003; Yu and Hatzivassiloglou, 2003; Dave et al., 2003; Hu and Liu, 2004)).

The third exploits automatic subjectivity analysis in applications such as review classification (e.g., (Turney, 2002; Pang and Lee, 2004)), mining texts for product reviews (e.g., (Yi et al., 2003; Hu and Liu, 2004; Popescu and Etzioni, 2005)), summarization (e.g., (Kim and Hovy, 2004)), information extraction (e.g., (Riloff et al., 2005)), and question answering (e.g., (Yu and Hatzivassiloglou, 2003; Stoyanov et al., 2005)).

Most manual subjectivity annotation research has focused on annotating words, out of context (e.g., (Heise, 2001)), or sentences and phrases in the context of a text or conversation (e.g., (Wiebe et al., 2005)). The new annotations in this paper are instead targeting the annotation of word *senses*.

## 3 Human Judgment of Word Sense Subjectivity

To explore our hypothesis that subjectivity may be associated with word senses, we developed a manual annotation scheme for assigning subjectivity labels to WordNet senses,[3] and performed an inter-annotator agreement study to assess its reliability. Senses are classified as *S(ubjective)*, *O(bjective)*, or *B(oth)*. 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.

We saw a number of subjective expressions in Section 2. A subset is repeated here, along with relevant WordNet senses. In the display of each sense, the first part shows the synset, gloss, and any examples. The second part (marked with =>) shows the immediate hypernym.

---

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))

---

He was **boiling** with anger.
seethe, boil – (be in an agitated emotional state; "The customer was seething with anger")
=> be – (have the quality of being; (copula, used with an adjective or a predicate noun); "John is rich"; "This is not a good answer")

---

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")

---

That doctor is a **quack.**
quack – (an untrained person who pretends to be a physician and who dispenses medical advice)
=> doctor, doc, physician, MD, Dr., medico

---

Before specifying what we mean by an objective sense, we give examples.

[1] Note that *sentiment*, the focus of much recent work in the area, is a type of subjectivity, specifically involving positive or negative opinion, emotion, or evaluation.

[2] These distinctions are not strictly needed for this paper, but may help the reader appreciate the examples given below.

[3] All our examples and data used in the experiments are from WordNet 2.0.

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")

The water **boiled**.
boil – (come to the boiling point and change from a liquid to vapor; "Water boils at 100 degrees Celsius")
=> change state, turn – (undergo a transformation or a change of position or action; "We turned from Socialism to Capitalism"; "The people turned against the President when he stole the election")

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)

The duck's **quack** was loud and brief.
quack – (the harsh sound of a duck)
=> sound – (the sudden occurrence of an audible event; "the sound awakened them")

While we expect phrases or sentences containing subjective senses to be subjective, we do not necessarily expect phrases or sentences containing objective senses to be objective. Consider the following examples:

> Will someone shut that damn **alarm** off?
> Can't you even **boil** water?

While these sentences contain objective senses of *alarm* and *boil*, the sentences are subjective nonetheless. But they are not subjective due to *alarm* and *boil*, but rather to punctuation, sentence forms, and other words in the sentence. Thus, 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.

Finally, classifying a sense as *B* means it covers both subjective and objective usages, e.g.:

absorb, suck, imbibe, soak up, sop up, suck up, draw, take in, take up – (take in, also metaphorically; "The sponge absorbs water well"; "She drew strength from the minister's words")

Manual subjectivity judgments were added to a total of 354 senses (64 words). One annotator, Judge 1 (a co-author), tagged all of them. A second annotator (Judge 2, who is not a co-author) tagged a subset for an agreement study, presented next.

### 3.1 Agreement Study

For the agreement study, Judges 1 and 2 independently annotated 32 words (138 senses). 16 words have both *S* and *O* senses and 16 do not (according to Judge 1). Among the 16 that do not have both S and O senses, 8 have only *S* senses and 8 have only *O* senses. All of the subsets are balanced between nouns and verbs. Table 1 shows the contingency table for the two annotators' judgments on this data. In addition to *S*, *O*, and *B*, the annotation scheme also permits *U(ncertain)* tags.

| | S | O | B | U | Total |
|---|---|---|---|---|---|
| S | 39 | O | O | 4 | 43 |
| O | 3 | 73 | 2 | 4 | 82 |
| B | 1 | O | 3 | 1 | 5 |
| U | 3 | 2 | O | 3 | 8 |
| Total | 46 | 75 | 5 | 12 | 138 |

Table 1: Agreement on balanced set (Agreement: 85.5%, $\kappa$: 0.74)

Overall agreement is 85.5%, with a Kappa ($\kappa$) value of 0.74. For 12.3% of the senses, at least one annotator's tag is *U*. If we consider these cases to be borderline and exclude them from the study, percent agreement increases to 95% and $\kappa$ rises to 0.90. Thus, annotator agreement is especially high when both are certain.

Considering only the 16-word subset with both *S* and *O* senses (according to Judge 1), $\kappa$ is .75, and for the 16-word subset for which Judge 1 gave only *S* or only *O* senses, $\kappa$ is .73. Thus, the two subsets are of comparable difficulty.

The two annotators also independently annotated the 20 ambiguous nouns (117 senses) of the SENSEVAL-3 English lexical sample task used in Section 5. For this tagging task, *U* tags were not allowed, to create a definitive gold standard for the experiments. Even so, the $\kappa$ value for them is 0.71, which is not substantially lower. The distributions of Judge 1's tags for all 20 words can be found in Table 3 below.

We conclude this section with examples of disagreements that illustrate sources of uncertainty. First, uncertainty arises when subjective senses are missing from the dictionary. The labels for the senses of noun *assault* are (*O:O,O:O,O:O,O:UO*).[4] For verb *assault* there is a subjective sense:

attack, round, assail, lash out, snipe, assault (attack in speech or writing) "The editors of the left-leaning paper attacked the new House Speaker"

However, there is no corresponding sense for

[4] I.e., the first three were labeled *O* by both annotators. For the fourth sense, the second annotator was not sure but was leaning toward *O*.

noun *assault*. A missing sense may lead an annotator to try to see subjectivity in an objective sense.

Second, uncertainty can arise in weighing hypernym against sense. It is fine for a synset to imply just *S* or *O*, while the hypernym implies both (the synset specializes the more general concept). However, consider the following, which was tagged (*O:UB*).

attack – (a sudden occurrence of an uncontrollable condition; "an attack of diarrhea")
=> affliction – (a cause of great suffering and distress)

While the sense is only about the condition, the hypernym highlights subjective reactions to the condition. One annotator judged only the sense (giving tag *O*), while the second considered the hypernym as well (giving tag *UB*).

## 4 Automatic Assessment of Word Sense Subjectivity

Encouraged by the results of the agreement study, we devised a method targeting the automatic annotation of word senses for subjectivity.

The main idea behind our method is that we can derive information about a word sense based on information drawn from words that are distributionally similar to the given word sense. This idea relates to the unsupervised word sense ranking algorithm described in (McCarthy et al., 2004). Note, however, that (McCarthy et al., 2004) used the information about distributionally similar words to approximate corpus frequencies for word senses, whereas we target the estimation of a *property* of a given word sense (the "subjectivity").

Starting with a given ambiguous word $w$, we first find the distributionally similar words using the method of (Lin, 1998) applied to the automatically parsed texts of the British National Corpus. Let $DSW$ = $dsw_1$, $dsw_2$, ..., $dsw_n$ be the list of top-ranked distributionally similar words, sorted in decreasing order of their similarity.

Next, for each sense $ws_i$ of the word $w$, we determine the similarity with each of the words in the list $DSW$, using a WordNet-based measure of semantic similarity ($wnss$). Although a large number of such word-to-word similarity measures exist, we chose to use the (Jiang and Conrath, 1997) measure, since it was found both to be efficient and to provide the best results in previous experiments involving word sense ranking (McCarthy et al., 2004)[5]. For distributionally similar words

[5]Note that unlike the above measure of *distributional* similarity which measures similarity between words, rather than word senses, here we needed a similarity measure that also takes into account word senses as defined in a sense inventory such as WordNet.

**Algorithm 1** Word Sense Subjectivity Score

**Input:** Word sense $w_i$
**Input:** Distributionally similar words $DSW = \{dsw_j | j = 1..n\}$
**Output:** Subjectivity score subj($w_i$)

```
1:  subj(w_i) = 0
2:  total_sim = 0
3:  for j = 1 to n do
4:     Insts_j = all instances of dsw_j in the MPQA corpus
5:     for k in Insts_j do
6:        if k is in a subj. expr. in MPQA corpus then
7:           subj(w_i) += sim(w_i,dsw_j)
8:        else if k is not in a subj. expr. in MPQA corpus
          then
9:           subj(w_i) -= sim(w_i,dsw_j)
10:       end if
11:       total_sim += sim(w_i,dsw_j)
12:    end for
13: end for
14: subj(w_i) = subj(w_i) / total_sim
```

that are themselves ambiguous, we use the sense that maximizes the similarity score. The similarity scores associated with each word $dsw_j$ are normalized so that they add up to one across all possible senses of $w$, which results in a score described by the following formula:

$$sim(ws_i, dsw_j) = \frac{wnss(ws_i, dsw_j)}{\sum_{i' \in senses(w)} wnss(ws_{i'}, dsw_j)}$$

where

$$wnss(ws_i, dsw_j) = \max_{k \in senses(dsw_j)} wnss(ws_i, dsw_j^k)$$

A selection process can also be applied so that a distributionally similar word belongs only to one sense. In this case, for a given sense $w_i$ we use only those distributionally similar words with whom $w_i$ has the highest similarity score across all the senses of $w$. We refer to this case as *similarity-selected*, as opposed to *similarity-all*, which refers to the use of all distributionally similar words for all senses.

Once we have a list of similar words associated with each sense $ws_i$ and the corresponding similarity scores $sim(ws_i, dsw_j)$, we use an annotated corpus to assign subjectivity scores to the senses. The corpus we use is the *MPQA* Opinion Corpus, which consists of over 10,000 sentences from the world press annotated for subjective expressions (all three types of subjective expressions described in Section 2).[6]

[6]The MPQA corpus is described in (Wiebe et al., 2005) and available at www.cs.pitt.edu/mpqa/databaserelease/.

Algorithm 1 is our method for calculating sense subjectivity scores. The subjectivity score is a value in the interval [-1,+1] with +1 corresponding to highly subjective and -1 corresponding to highly objective. It is a sum of $sim$ scores, where $sim(w_i,dsw_j)$ is added for each instance of $dsw_j$ that is in a subjective expression, and subtracted for each instance that is not in a subjective expression.

Note that the annotations in the MPQA corpus are for subjective expressions in context. Thus, the data is somewhat noisy for our task, because, as discussed in Section 3, objective senses may appear in subjective expressions. Nonetheless, we hypothesized that subjective senses tend to appear more often in subjective expressions than objective senses do, and use the appearance of words in subjective expressions as evidence of sense subjectivity.

(Wiebe, 2000) also makes use of an annotated corpus, but in a different approach: given a word *w* and a set of distributionally similar words *DSW*, that method assigns a subjectivity score to *w* equal to the conditional probability that any member of *DSW* is in a subjective expression. Moreover, the end task of that work was to annotate words, while our end task is the more difficult problem of annotating word *senses* for subjectivity.

### 4.1 Evaluation

The evaluation of the algorithm is performed against the gold standard of 64 words (354 word senses) using Judge 1's annotations, as described in Section 3.

For each sense of each word in the set of 64 ambiguous words, we use Algorithm 1 to determine a subjectivity score. A subjectivity label is then assigned depending on the value of this score with respect to a pre-selected threshold. While a threshold of 0 seems like a sensible choice, we perform the evaluation for different thresholds ranging across the [-1,+1] interval, and correspondingly determine the precision of the algorithm at different points of recall[7]. Note that the word senses for which none of the distributionally similar words are found in the *MPQA* corpus are not included in this evaluation (excluding 82 senses), since in this case a subjectivity score cannot be calculated. The evaluation is therefore performed on a total of 272 word senses.

As a baseline, we use an "informed" random assignment of subjectivity labels, which randomly assigns *S* labels to word senses in the data set, such that the maximum number of *S* assignments equals the number of correct S labels in the gold standard data set. This baseline guarantees a maximum recall of 1 (which under true random conditions might not be achievable). Correspondingly, given the controlled distribution of *S* labels across the data set in the baseline setting, the precision is equal for all eleven recall points, and is determined as the total number of correct subjective assignments divided by the size of the data set[8].

| Algorithm | Number of DSW | Break-even point |
|---|---|---|
| similarity-all | 100 | 0.41 |
| similarity-selected | 100 | **0.50** |
| similarity-all | 160 | 0.43 |
| similarity-selected | 160 | **0.50** |
| baseline | - | 0.27 |

Table 2: Break-even point for different algorithm and parameter settings

There are two aspects of the sense subjectivity scoring algorithm that can influence the label assignment, and correspondingly their evaluation.

First, as indicated above, after calculating the semantic similarity of the distributionally similar words with each sense, we can either use *all* the distributionally similar words for the calculation of the subjectivity score of each sense (*similarity-all*), or we can use only those that lead to the highest similarity (*similarity-selected*). Interestingly, this aspect can drastically affect the algorithm accuracy. The setting where a distributionally similar word can belong only to one sense significantly improves the algorithm performance. Figure 1 plots the interpolated precision for eleven points of recall, for *similarity-all*, *similarity-selected*, and *baseline*. As shown in this figure, the precision-recall curves for our algorithm are clearly above the "informed" baseline, indicating the ability of our algorithm to automatically identify subjective word senses.

Second, the number of distributionally similar words considered in the first stage of the algorithm can vary, and might therefore influence the

[7]Specifically, in the list of word senses ranked by their subjectivity score, we assign a subjectivity label to the top $N$ word senses. The precision is then determined as the number of correct subjectivity label assignments out of all $N$ assignments, while the recall is measured as the correct subjective senses out of all the subjective senses in the gold standard data set. By varying the value of $N$ from 1 to the total number of senses in the corpus, we can derive precision and recall curves.

[8]In other words, this fraction represents the probability of making the correct subjective label assignment by chance.

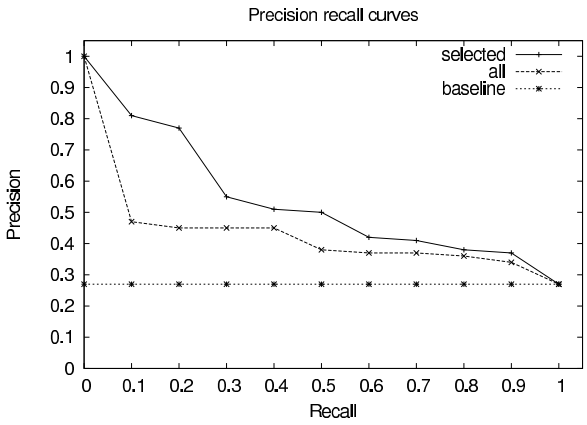


Figure 1: Precision and recall for automatic subjectivity annotations of word senses (DSW=160).

output of the algorithm. We experiment with two different values, namely 100 and 160 top-ranked distributionally similar words. Table 2 shows the break-even points for the four different settings that were evaluated,[9] with results that are almost double compared to the informed baseline. As it turns out, for weaker versions of the algorithm (i.e., *similarity-all*), the size of the set of distributionally similar words can significantly impact the performance of the algorithm. However, for the already improved *similarity-selected* algorithm version, this parameter does not seem to have influence, as similar results are obtained regardless of the number of distributionally similar words. This is in agreement with the finding of (McCarthy et al., 2004) that, in their word sense ranking method, a larger set of neighbors did not influence the algorithm accuracy.

## 5 Automatic Subjectivity Annotations for Word Sense Disambiguation

The final question we address is concerned with the potential impact of subjectivity on the quality of a word sense classifier. To answer this question, we augment an existing data-driven word sense disambiguation system with a feature reflecting the subjectivity of the examples where the ambiguous word occurs, and evaluate the performance of the new subjectivity-aware classifier as compared to the traditional context-based sense classifier.

We use a word sense disambiguation system that integrates both local and topical features. Specifically, we use the current word and its part-of-speech, a local context of three words to the left and right of the ambiguous word, the parts-of-speech of the surrounding words, and a global context implemented through sense-specific keywords determined as a list of at most five words occurring at least three times in the contexts defining a certain word sense. This feature set is similar to the one used by (Ng and Lee, 1996), as well as by a number of SENSEVAL systems. The parameters for sense-specific keyword selection were determined through cross-fold validation on the training set. The features are integrated in a Naive Bayes classifier, which was selected mainly for its performance in previous work showing that it can lead to a state-of-the-art disambiguation system given the features we consider (Lee and Ng, 2002).

The experiments are performed on the set of ambiguous nouns from the SENSEVAL-3 English lexical sample evaluation (Mihalcea et al., 2004). We use the rule-based subjective sentence classifier of (Riloff and Wiebe, 2003) to assign an *S*, *O*, or *B* label to all the training and test examples pertaining to these ambiguous words. This subjectivity annotation tool targets sentences, rather than words or paragraphs, and therefore the tool is fed with sentences. We also include a surrounding context of two additional sentences, because the classifier considers some contextual information.

Our hypothesis motivating the use of a sentence-level subjectivity classifier is that instances of subjective *senses* are more likely to be in subjective *sentences*, and thus that sentence subjectivity is an informative feature for the disambiguation of words having both subjective and objective senses.

For each ambiguous word, we perform two separate runs: one using the basic disambiguation system described earlier, and another using the subjectivity-aware system that includes the additional subjectivity feature. Table 3 shows the results obtained for these 20 nouns, including word sense disambiguation accuracy for the two different systems, the most frequent sense baseline, and the subjectivity/objectivity split among the word senses (according to Judge 1). The words in the top half of the table are the ones that have both *S* and *O* senses, and those in the bottom are the ones that do not. If we were to use Judge 2's tags instead of Judge 1's, only one word would change: *source* would move from the top to the bottom of the table.

[9]The break-even point (Lewis, 1992) is a standard measure used in conjunction with precision-recall evaluations. It represents the value where precision and recall become equal.

| Word | Senses | Sense subjectivity | Data train | Data test | Baseline | Classifier basic | Classifier +subj. |
|---|---|---|---|---|---|---|---|
| Words with subjective senses | | | | | | | |
| argument | 5 | 3-S 2-O | 221 | 111 | 49.4% | 51.4% | 54.1% |
| atmosphere | 6 | 2-S 4-O | 161 | 81 | 65.4% | 65.4% | 66.7% |
| difference | 5 | 2-S 3-O | 226 | 114 | 40.4% | 54.4% | 57.0% |
| difficulty | 4 | 2-S 2-O | 46 | 23 | 17.4% | 47.8% | 52.2% |
| image | 7 | 2-S 5-O | 146 | 74 | 36.5% | 41.2% | 43.2% |
| interest | 7 | 1-S 5-O 1-B | 185 | 93 | 41.9% | 67.7% | 68.8% |
| judgment | 7 | 5-S 2-O | 62 | 32 | 28.1% | 40.6% | 43.8% |
| plan | 3 | 1-S 2-O | 166 | 84 | 81.0% | 81.0% | 81.0% |
| sort | 4 | 1-S 2-O 1-B | 190 | 96 | 65.6% | 66.7% | 67.7% |
| source | 9 | 1-S 8-O | 64 | 32 | 40.6% | 40.6% | 40.6% |
| Average | | | | | 46.6% | 55.6% | 57.5% |
| Words with no subjective senses | | | | | | | |
| arm | 6 | 6-O | 266 | 133 | 82.0% | 85.0% | 84.2% |
| audience | 4 | 4-O | 200 | 100 | 67.0% | 74.0% | 74.0% |
| bank | 10 | 10-O | 262 | 132 | 62.6% | 62.6% | 62.6% |
| degree | 7 | 5-O 2-B | 256 | 128 | 60.9% | 71.1% | 71.1% |
| disc | 4 | 4-O | 200 | 100 | 38.0% | 65.6% | 66.4% |
| organization | 7 | 7-O | 112 | 56 | 64.3% | 64.3% | 64.3% |
| paper | 7 | 7-O | 232 | 117 | 25.6% | 49.6% | 48.0% |
| party | 5 | 5-O | 230 | 116 | 62.1% | 62.9% | 62.9% |
| performance | 5 | 5-O | 172 | 87 | 26.4% | 34.5% | 34.5% |
| shelter | 5 | 5-O | 196 | 98 | 44.9% | 65.3% | 65.3% |
| Average | | | | | 53.3% | 63.5% | 63.3% |
| Average for all words | | | | | 50.0% | 59.5% | 60.4% |

Table 3: Word Sense Disambiguation with and without subjectivity information, for the set of ambiguous nouns in SENSEVAL-3

For the words that have both *S* and *O* senses, the addition of the subjectivity feature alone can bring a significant error rate reduction of 4.3% ($p < 0.05$ paired t-test). Interestingly, no improvements are observed for the words with no subjective senses; on the contrary, the addition of the subjectivity feature results in a small degradation. Overall for the entire set of ambiguous words, the error reduction is measured at 2.2% (significant at $p < 0.1$ paired t-test).

In almost all cases, the words with both *S* and *O* senses show improvement, while the others show small degradation or no change. This suggests that if a subjectivity label is available for the words in a lexical resource (e.g. using Algorithm 1 from Section 4), such information can be used to decide on using a subjectivity-aware system, thereby improving disambiguation accuracy.

One of the exceptions is *disc*, which had a small benefit, despite not having any subjective senses. As it happens, the first sense of *disc* is *phonograph record*.

phonograph record, phonograph recording, record, disk, disc, platter – (sound recording consisting of a disc with continuous grooves; formerly used to reproduce music by rotating while a phonograph needle tracked in the grooves)

The improvement can be explained by observing that many of the training and test sentences containing this sense are labeled subjective by the classifier, and indeed this sense frequently occurs in subjective sentences such as "This is anyway a stunning **disc**."

Another exception is the noun *plan*, which did not benefit from the subjectivity feature, although it does have a subjective sense. This can perhaps be explained by the data set for this word, which seems to be particularly difficult, as the basic classifier itself could not improve over the most frequent sense baseline.

The other word that did not benefit from the subjectivity feature is the noun *source*, for which its only subjective sense did not appear in the sense-annotated data, leading therefore to an "objective only" set of examples.

## 6 Conclusion and Future Work

The questions posed in the introduction concerning the possible interaction between subjectivity and word sense found answers throughout the paper. As it turns out, a correlation can indeed be established between these two semantic properties of language.

Addressing the first question of whether subjectivity is a property that can be assigned to word senses, we showed that good agreement ($\kappa$=0.74) can be achieved between human annotators labeling the subjectivity of senses. When uncertain cases are removed, the $\kappa$ value is even higher (0.90). Moreover, the automatic subjectivity scoring mechanism that we devised was able to successfully assign subjectivity labels to senses, significantly outperforming an "informed" baseline associated with the task. While much work remains to be done, this first attempt has proved the feasibility of correctly assigning subjectivity labels to the fine-grained level of word senses.

The second question was also positively answered: the quality of a word sense disambiguation system can be improved with the addition of subjectivity information. Section 5 provided evidence that automatic subjectivity classification may improve word sense disambiguation performance, but mainly for words with both subjective and objective senses. As we saw, performance may even degrade for words that do not. Tying the pieces of this paper together, once the senses in a dictionary have been assigned subjectivity labels, a word sense disambiguation system could consult them to decide whether it should consider or ignore the subjectivity feature.

There are several other ways our results could impact future work. Subjectivity labels would be a useful source of information when manually augmenting the lexical knowledge in a dictionary,

e.g., when choosing hypernyms for senses or deciding which senses to eliminate when defining a coarse-grained sense inventory (if there is a subjective sense, at least one should be retained).

Adding subjectivity labels to WordNet could also support automatic subjectivity analysis. First, the input corpus could be sense tagged and the subjectivity labels of the assigned senses could be exploited by a subjectivity recognition tool. Second, a number of methods for subjectivity or sentiment analysis start with a set of seed words and then search through WordNet to find other subjective words (Kamps and Marx, 2002; Yu and Hatzivassiloglou, 2003; Hu and Liu, 2004; Kim and Hovy, 2004; Esuli and Sebastiani, 2005). However, such searches may veer off course down objective paths. The subjectivity labels assigned to senses could be consulted to keep the search traveling along subjective paths.

Finally, there could be different strategies for exploiting subjectivity annotations and word sense. While the current setting considered a pipeline approach, where the output of a subjectivity annotation system was fed to the input of a method for semantic disambiguation, future work could also consider the role of word senses as a possible way of improving subjectivity analysis, or simultaneous annotations of subjectivity and word meanings, as done in the past for other language processing problems.

**Acknowledgments** We would like to thank Theresa Wilson for annotating senses, and the anonymous reviewers for their helpful comments. This work was partially supported by ARDA AQUAINT and by the NSF (award IIS-0208798).

## References

K. Dave, S. Lawrence, and D. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proc. WWW-2003*, Budapest, Hungary. Available at http://www2003.org.

A. Esuli and F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss analysis. In *Proc. CIKM-2005*.

V. Hatzivassiloglou and K. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proc. ACL-97*, pages 174–181.

D. Heise. 2001. Project magellan: Collecting cross-cultural affective meanings via the internet. *Electronic Journal of Sociology*, 5(3).

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD*.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical tax onomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan.

J. Kamps and M. Marx. 2002. Words with attitude. In *Proc. 1st International WordNet Conference*.

S.M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proc. Coling 2004*.

Y.K. Lee and H.T. Ng. 2002. An empirical evaluation of knowledge sources and learning algo rithms for word sense disambiguation. In *Proc. EMNLP 2002*.

D. Lewis. 1992. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of ACM SIGIR*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*, Montreal, Canada.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant senses in untagged text. In *Proc. ACL 2004*.

R. Mihalcea, T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proc. ACL/SIGLEX Senseval-3*.

G. Miller. 1995. Wordnet: A lexical database. *Communication of the ACM*, 38(11):39–41.

H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word se nse: An examplar-based approach. In *Proc. ACL 1996*.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summariza tion based on minimum cuts. In *Proc. ACL 2004*.

A. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proc. of HLT/EMNLP 2005*.

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 *Proc. EMNLP 2003*.

E. Riloff, J. Wiebe, and W. Phillips. 2005. Exploiting subjectivity classification to improve information ex traction. In *Proc. AAAI 2005*.

V. Stoyanov, C. Cardie, and J. Wiebe. 2005. Multi-perspective question answering using the opqa corpus. In *Proc. HLT/EMNLP 2005*.

P. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proc. ACL 2002*.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).

J. Wiebe. 2000. Learning subjective adjectives from corpora. In *Proc. AAAI 2000*.

J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natu ral language processing techniques. In *Proc. ICDM 2003*.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proc. EMNLP 2003*.