Improving the Impact of Subjectivity Word Sense Disambiguation on Contextual Opinion Analysis

Cem Akkaya, Janyce Wiebe, Alexander Conrad
University of Pittsburgh
Pittsburgh PA, 15260, USA
{cem,wiebe,conrada}@cs.pitt.edu

Rada Mihalcea
University of North Texas
Denton TX, 76207, USA
rada@cs.unt.edu

Abstract

Subjectivity word sense disambiguation (SWSD) is automatically determining which word instances in a corpus are being used with subjective senses, and which are being used with objective senses. SWSD has been shown to improve the performance of contextual opinion analysis, but only on a small scale and using manually developed integration rules. In this paper, we scale up the integration of SWSD into contextual opinion analysis and still obtain improvements in performance, by successfully gathering data annotated by non-expert annotators. Further, by improving the method for integrating SWSD into contextual opinion analysis, even greater benefits from SWSD are achieved than in previous work. We thus more firmly demonstrate the potential of SWSD to improve contextual opinion analysis.

1 Introduction

Often, methods for opinion, sentiment, and subjectivity analysis rely on lexicons of subjective (opinion-carrying) words (e.g., (Turney, 2002; Whitelaw et al., 2005; Riloff and Wiebe, 2003; Yu and Hatzivassiloglou, 2003; Kim and Hovy, 2004; Bloom et al., 2007; Andreevskaia and Bergler, 2008; Agarwal et al., 2009)). 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**.

However, even manually developed subjectivity lexicons have significant degrees of subjectivity sense ambiguity (Su and Markert, 2008; Gyamfi et al., 2009). That is, many clues in these lexicons have both subjective and objective senses. This ambiguity leads to errors in opinion and sentiment analysis, because objective instances represent false hits of subjectivity clues. For example, the following sentence contains the keywords from (1) used with objective senses:

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

Recently, in (Akkaya et al., 2009), we introduced the task of *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 developed a supervised system for SWSD, and exploited the SWSD output to improve the performance of multiple contextual opinion analysis tasks.

Although the reported results are promising, there are three obvious shortcomings. First, we were able to apply SWSD to contextual opinion analysis only on a very small scale, due to a shortage of annotated data. While the experiments show that SWSD improves contextual opinion analysis, this was only on the small amount of opinion-annotated data that was in the coverage of our system. Two questions arise: is it feasible to obtain greater amounts of the needed data, and do SWSD performance improvements on contextual opinion analysis hold on a

Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 87–96,
Portland, Oregon, USA, 23–24 June 2011.

larger scale. Second, the annotations in (Akkaya et al., 2009) are piggy-backed on SENSEVAL sense-tagged data, which are fine-grained word sense annotations created by trained annotators. A concern is that SWSD performance improvements on contextual opinion analysis can only be achieved using such fine-grained expert annotations, the availability of which is limited. Third, (Akkaya et al., 2009) uses manual rules to apply SWSD to contextual opinion analysis. Although these rules have the advantage that they transparently show the effects of SWSD, they are somewhat ad hoc. Likely, they are not optimal and are holding back the potential of SWSD to improve contextual opinion analysis.

To address these shortcomings, in this paper, we investigate (1) the feasibility of obtaining a substantial amount of annotated data, (2) whether performance improvements on contextual opinion analysis can be realized on a larger scale, and (3) whether those improvements can be realized with subjectivity sense tagged data that is not built on expert full-inventory sense annotations. In addition, we explore better methods for applying SWSD to contextual opinion analysis.

2 Subjectivity Word Sense Disambiguation

2.1 Annotation Tasks

We adopt the definitions of *subjective* (*S*) and *objective* (*O*) from (Wiebe et al., 2005; Wiebe and Mihalcea, 2006; Wilson, 2007). 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. Objective expressions instead are words and phrases that lack subjectivity.

The contextual opinion analysis experiments described in Section 3 include both *S*/*O* and polarity (positive,negative, neutral) classifications. The opinion-annotated data used in those experiments is from the MPQA Corpus (Wiebe et al., 2005; Wilson, 2007),[1] which consists of news articles annotated for subjective expressions, including polarity.

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

2.1.1 Subjectivity Sense Labeling

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.

Both (Wiebe and Mihalcea, 2006) and (Su and Markert, 2008) performed agreement studies of the scheme and report that good agreement can be achieved between human annotators labeling the subjectivity of senses (κ values of 0.74 and 0.79, respectively).

(Akkaya et al., 2009) followed the same annotation scheme to annotate the senses of the words used in the experiments. For this paper, we again use the same scheme and annotate WordNet senses of 90 new words (the process of selecting the words is described in Section 2.4).

2.1.2 Subjectivity Sense Tagging

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 there was no such tagged data at the time, (Akkaya et al., 2009) created a data set by combining two types of sense annotations: (1) labels of senses within a dictionary as *S* or *O* (i.e., the subjectivity sense labels of the previous section), and (2) sense tags of word instances in a corpus (i.e., SENSEVAL sense-tagged data).[2] The subjectivity sense labels were used to collapse the sense labels in the sense-tagged data into the two new senses, *S* and *O*. The target words (Akkaya et al., 2009) chose are the words tagged in SENSEVAL that are also members

[2] Please see the paper for details on the SENSEVAL data used in the experiments.

Sense_Set1 (Subjective)

{ **attack**, round, assail, lash_out, snipe, assault } – attack in speech or writing; "The editors attacked the House Speaker"

{ assail, assault, set_on, **attack** } – attack someone emotionally; "Nightmares assailed him regularly"

Sense_Set2 (Objective)

{ **attack** } – begin to injure; "The cancer cells are attacking his liver"; "Rust is attacking the metal"

{ **attack**, aggress } – take the initiative and go on the offensive; "The visiting team started to attack"

Figure 1: Sense sets for target word "attack" (abridged).

of the subjectivity lexicon of (Wilson et al., 2005; Wilson, 2007).[3] There are 39 such words. (Akkaya et al., 2009) chose words from a subjectivity lexicon because such words are known to have subjective usages.

For this paper, subjectivity sense-tagged data was obtained from the MTurk workers using the annotation scheme of (Akkaya et al., 2010). A goal is to keep the annotation task as simple as possible. Thus, the workers are not directly asked if the instance of a target word has a subjective or an objective sense, because the concept of subjectivity would be difficult to explain in this setting. Instead the workers are shown two sets of senses – one subjective set and one objective set – for a specific target word and a text passage in which the target word appears. Their job is to select the set that best reflects the meaning of the target word in the text passage. The set they choose gives us the subjectivity label of the instance.

A sample annotation task is shown below. An MTurk worker has access to two sense sets of the target word "attack" as seen in Figure 1. The S and O labels appear here only for the purpose of this paper; the workers do not see them. The worker is presented with the following text passage holding the target word "attack":

> Ivkovic had been a target of intra-party feuding that has shaken the party. He was **attacked** by Milosevic for attempting to carve out a new party from the Socialists.

In this passage, the use of "attack" is most similar to the first entry in sense set one; thus, the correct answer for this problem is Sense_Set-1.

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

(Akkaya et al., 2010) carried out a pilot study where a subjectivity sense-tagged dataset was created for eight SENSEVAL words through MTurk. (Akkaya et al., 2010) evaluated the non-expert label quality against gold-standard expert labels which were obtained from (Akkaya et al., 2009) relying on SENSEVAL. The non-expert annotations are reliable, achieving κ scores around 0.74 with the expert annotations.

For some words, there may not be a clean split between the subjective and objective senses. For these, we opted for another strategy for obtaining MTurk annotations. Rather than presenting the workers with WordNet senses, we show them a set of objective usages, a set of subjective usages, and a text passage in which the target word appears. The workers' job is to judge which set of usages the target instance is most similar to.

2.2 SWSD System

We follow the same approach as in (Akkaya et al., 2009) to build our SWSD system. We train a different supervised SWSD classifier for each target word separately. This means the overall SWSD system consists of as many SWSD classifiers as there are target words. We utilize the same machine learning features as in (Akkaya et al., 2009), which are commonly used in *Word Sense Disambiguation (WSD)*.

2.3 Expert SWSD vs. Non-expert SWSD

Before creating a large subjectivity sense-tagged corpus via MTurk, we want to make sure that non-expert annotations are good enough to train reliable SWSD classifiers. Thus, we decided to compare the performance of a SWSD system trained on non-expert annotations and on expert annotations. For this purpose, we need a subjectivity sense-tagged corpus where word instances are tagged both by expert and non-expert annotations. Fortunately, we have such a corpus. As discussed in Section 3, (Akkaya et al., 2009) created a subjecvitivity sense-tagged corpus piggybacked on SENSEVAL. This gives us a gold-standard corpus tagged by experts. There is also a small subjectivity sense-tagged corpus consisting of eight target words obtained from non-expert annotators in (Akkaya et al., 2010). This corpus is a subset of the gold-standard corpus from (Akkaya et al., 2009) and it consists of 60 tagged

	Acc	p-value
$SWSD_{GOLD}$	79.2	-
$SWSD_{MJL}$	78.4	0.542
$SWSD_{MJC}$	78.8	0.754

Table 1: Comparison of SWSD systems

instances for each target word.

Actually, (Akkaya et al., 2010) gathered three labels for each instance. This gives us two options to train the non-expert SWSD system: (1) training the system on the majority vote labels *($SWSD_{MJL}$)* (2) training three systems on the three separate label sets and taking the majority vote prediction *($SWSD_{MJC}$)*. Additionally, we train an expert SWSD system *($SWSD_{GOLD}$)* – a system trained on gold standard expert annotations. All these systems are trained on 60 instances of the eight target words for which we have both non-expert and expert annotations and are evaluated on the remaining instances of the gold-standard corpus. This makes a total of 923 test instances for the eight target words with a majority class baseline of 61.8.

Table 1 reports micro-average accuracy of each system and the two-tailed p-value between the expert SWSD system and the two non-expert SWSD systems. The p-value is calculated with McNemar's test. It shows that there is no statistically significant difference between classifiers trained on expert gold-standard annotations and non-expert annotations. We adopt $SWSD_{MJL}$ in all our following experiments, because it is more efficient.

2.4 Corpus Creation

For our experiments, we have multiple goals, which effect our decisions on how to create the subjectivity sense-tagged corpus via MTurk. First, we want to be able to disambiguate more target words than (Akkaya et al., 2009). This way, SWSD will be able to disambiguate a larger portion of the MPQA Corpus allowing us to evaluate the effect of SWSD on contextual opinion analysis on a larger scale. This will also allow us to investigate additional integration methods of SWSD into contextual opinion analysis rather than simple ad hoc manual rules utilized in (Akkaya et al., 2009). Second, we want to show that we can rely on non-expert annotations instead of expert annotations, which will make an annotation effort on a larger-scale both practical and feasible, timewise and costwise. Optimally, we could have annotated via MTurk the same subjectivity sense-tagged corpus from (Akkaya et al., 2009) in order to compare the effect of a non-expert SWSD system on contextual opinion analysis directly with the results reported for an expert SWSD system in (Akkaya et al., 2009). But, this would have diverted our resources to reproduce the same corpus and contradict our goal to extend the subjectivity sense-tagged corpus to new target words. Moreover, we have already shown in Section 2.3 that non-expert annotations can be utilized to train reliable SWSD classifiers. It is reasonable to believe that similar performance on the SWSD task will reflect to similar improvements on contextual opinion analysis. Thus, we decided to prioritize creating a subjectivity sense-tagged corpus for a totally new set of words. We aim to show that the favourable results reported in (Akkaya et al., 2009) will still hold on new target words relying on non-expert annotations.

We chose our target words from the subjectivity lexicon of (Wilson et al., 2005), because we know they have subjective usages. The contextual opinion systems we want to improve rely on this lexicon. We call the words in the lexicon *subjectivity clues*. At this stage, we want to concentrate on the frequent and ambiguous subjectivity clues. We chose frequent ones, because they will have larger coverage in the MPQA Corpus. We chose ambiguous ones, because these clues are the ones that are most important for SWSD. Choosing most frequent and ambiguous subjectivity clues guarantees that we utilize our limited resources in the most efficient way. We judge a clue to be ambiguous if it appears more than 25% and less than 75% of the times in a subjective expression. We get these statistics by simply counting occurrences in the MPQA Corpus inside and outside of subjective expressions.

There are 680 subjectivity clues that appear in the MPQA Corpus and are ambiguous. Out of those, we selected the 90 most frequent that have to some extent distinct objective and subjective senses in WordNet, as judged by the co-authors. The co-authors annotated the WordNet senses of those 90 target words. For each target word, we selected approximately 120 instances randomly from the *GIGAWORD Corpus*. In a first phase, we collected three sets of MTurk an-

notations for the selected instances. In this phase, MTurk workers base their judgements on two sense sets they observe. This way, we get training data to build SWSD classifiers for these 90 target words.

The quality of these classifiers is important, because we will exploit them for contextual opinion analysis. Thus, we evaluate them by 10-fold cross-validation. We split the target words into three groups. If the majority class baseline of a word is higher than 90%, it is considered as *skewed* (skewed words have a performance at least as good as the majority class baseline). If a target word improves over its majority class baseline by 25% in accuracy, it is considered as *good*. Otherwise, it is considered as *mediocre*. This way, we end up with 24 skewed, 35 good, and 31 mediocre words. There are many possible reasons for the less reliable performance for the mediocre group. We hypothesize that a major problem is the similarity between the objective and subjective sense sets of a word, thus leading to poor annotation quality. To check this, we calculate the agreement between three annotation sets and report averages. The agreement in the mediocre group is 78.68%, with a κ value of 0.57, whereas the average agreement in the good group is 87.51%, with a κ value of 0.75. These findings support our hypothesis. Thus, the co-authors created usage inventories for the words in the mediocre group as described in Section 2.1.1. We initiated a second phase of MTurk annotations. We collect for the mediocre group another three sets of MTurk annotations for 120 instances, this time utilizing usage inventories. The 10-fold cross-validation experiments show that nine of the 31 words in the mediocre group shift to the good group. Only for these nine words, we accept the annotations collected via usage inventories. For all other words, we use the annotations collected via sense inventories. From now on, we will refer to this non-expert subjectivity sense-tagged corpus consisting of the tagged data for all 90 target words as the *MTurkSWSD Corpus* (agreement on the entire MTurkSWSD corpus is 85.54%, κ:0.71).

3 SWSD Integration

Now that we have the MTurkSWSD Corpus, we are ready to evaluate the effect of SWSD on contextual opinion analysis. In this section, we apply our SWSD system trained on MTurkSWSD to both expression-level classifiers from (Akkaya et al., 2009): (1) the subjective/objective (S/O) classifier and (2) the contextual polarity classifier. Both classifiers are introduced in Section 3.1

Our SWSD system can disambiguate 90 target words, which have 3737 instances in the MPQA Corpus. We refer to this subset of the MPQA Corpus as *MTurkMPQA*. This subset makes up the coverage of our SWSD system. Note that MTurkMPQA is 5.2 times larger than the covered MPQA subset in (Akkaya et al., 2009) referred as *senMPQA*. We try different strategies to integrate SWSD into the contextual classifiers. In Section 3.2, we follow the same rule-based strategy as in (Akkaya et al., 2009) for completeness. In Section 3.3, we introduce two new learning strategies for SWSD integration outperforming existing rule-based strategy. We evaluate the improvement gained by SWSD on MTurkMPQA.

3.1 Contextual Classifiers

The original contextual polarity classifier is introduced in (Wilson et al., 2005). We use the same implementation as in (Akkaya et al., 2009). This classifier labels clue instances in text as contextually negative/positive/neutral. The gold standard is defined on the MPQA Corpus as follows. If a clue instance appears in a positive expression, it is contextually positive (Ps). If it appears in a negative expression, it is contextually negative (Ng). If it is in an objective expression or in a neutral subjective expression, it is contextually neutral (N). The contextual polarity classifier consists of two separate steps. The first step is an expression-level neutral/polar (N/P) classifier. The second step classifies only polar instances further into positive and negative classes. This way, the overall system performs a three-way classification (Ng/Ps/N).

The subjective/objective classifier is introduced in (Akkaya et al., 2009). It relies on the same machine learning features as the N/P classifier (i.e. the first step of the contextual polarity classifier). The only difference is that the classes are S/O instead of N/P. The gold standard is defined on the MPQA Corpus in the following way. If a clue instance appears in a subjective expression, it is contextually S. If it appears in an objective expression, it is contextually O. Both contextual classifiers are supervised.

	Baseline		Acc	OF	SF
MTurkMPQA	52.4% (O)	$O_{S/O}$	67.1	68.9	65.0
		R1R2	**71.1**	72.7	69.2
senMPQA	63.1% (O)	$O_{S/O}$	75.4	65.4	80.9
		R1R2	81.3	75.9	84.8

Table 2: S/O classifier with and without SWSD.

	Baseline		Acc	NF	PF
MTurkMPQA	70.6% (P)	$O_{N/P}$	72.3	82.0	39.8
		R4	**74.5**	84.0	37.8
senMPQA	73.9% (P)	$O_{N/P}$	79.0	86.7	50.3
		R4	81.6	88.6	52.3

Table 3: N/P classifier with and without SWSD

3.2 Rule-Based SWSD Integration

(Akkaya et al., 2009) integrates SWSD into a contextual classifier by simple rules. The rules flip the output of the contextual classifier if some conditions hold. They make use of following information: (1) SWSD output, (2) the contextual classifier's confidence and (3) the presence of another subjectivity clue – any clue from the subjectivity lexicon – in the same expression.

For the contextual S/O classifier, (Akkaya et al., 2009) defines two rules: one flipping the S/O classifier's output from O to S *(R1)* and one flipping from S to O *(R2)*. R1 is defined as follows : if the contextual classifier decides a target word instance is contextually O and SWSD decides that it is used in a S sense, then SWSD overrules the contextual S/O classifier's output and flips it from O to S, because an instance in a S sense will make the surrounding expression subjective. R2 is a little bit more complex. It is defined as follows: If the contextual classifier labels a clue instance as S but (1) SWSD decides that it is used in an O sense, (2) the contextual classifier's confidence is low, and (3) there is no other subjectivity clue in the same expression, then R2 flips the contextual classifier's output from S to O. The rationale behind R2 is that even if the target word instance has an O sense, there might be another reason (e.g. the presence of another subjectivity clue in the same expression) for the expression enclosing it to be subjective.

We use the exact same rules and adopt the same confidence threshold. Table 2 holds the comparison of the original contextual classifier and the classifier with SWSD support on senMPQA as reported in (Akkaya et al., 2009) and on MTurkMPQA. $O_{S/O}$ is the original S/O classifier; R1R2 is the system with SWSD support utilizing both rules. We report only R1R2, since (Akkaya et al., 2009) gets highest improvement utilizing both rules.

In Table 2 we see that R1R2 achieves 4% percentage points improvement in accuracy over $O_{S/O}$ on MTurkMPQA. The improvement is statistically significant at the $p < .01$ level with McNemar's test. It is accompanied with improvements both in subjective F-measure (SF) and objective F-measure (OF). It is not possible to directly compare improvements on senMPQA and MTurkMPQA since they are different subsets of the MPQA Corpus. SWSD support brings 24% error reduction on senMPQA over the original S/O classifier. In comparison, on MTurkMPQA, the error reduction is 12%. We see that the improvements on the large MTurkMPQA set still hold, but not as strong as in (Akkaya et al., 2009).

(Akkaya et al., 2009) uses a similar rule to make the contextual polarity classifier sense-aware. Specifically, the rule is applied to the output of the first step (N/P classifier). The rule, R4, flips P to N and is analogous to R2. If the contextual classifier labels a clue instance as P but (1) SWSD decides that it is used in an O sense, (2) the contextual classifier's confidence is low, and (3) there is no other clue instance in the same expression, then R4 flips the contextual classifier's output from P to N.

Table 3 holds the comparison of the original N/P classifier with and without SWSD support on senMPQA as reported in (Akkaya et al., 2009) and on MTurkMPQA. $O_{N/P}$ is the original N/P classifier; R4 is the system with SWSD support utilizing rule R4. Since our main focus is not rule-based integration, we did not run the second step of the polarity classifier. We report the second step result below for the learning-based SWSD integration in section 3.4.

In Table 3, we see that R4 achieves 2.2 percentage points improvement in accuracy over $O_{N/P}$ on MTurkMPQA. The improvement is statistically significant at the $p < .01$ level with McNemar's test. It is accompanied with improvement only in objective F-measure (OF). SWSD support brings 12.4% error reduction on senMPQA (Akkaya et al., 2009).

On MTurkMPQA, the error reduction is 8%. We see that the rule-based SWSD integration still improves both contextual classifiers on MTurkMPQA, but the gain is not as large as on senMPQA. This might be due to the brittleness of the rule-based integration.

3.3 Learning SWSD Integration

Now that we can disambiguate a larger portion of the MPQA Corpus than in (Akkaya et al., 2009), we can investigate machine learning methods for SWSD integration to deal with the brittleness of the rule-based integration. In this section, we introduce two learning methods to apply SWSD to the contextual classifiers. For the learning methods, we rely on exactly the same information as the rule-based integration: (1) SWSD output, (2) the contextual classifier's output, (3) the contextual classifier's confidence, and (4) the presence of another clue instance in the same expression. The rationale is the same as for the rule-based integration, namely to relate sense subjectivity and contextual subjectivity.

3.3.1 Method1

In the first method, we extend the machine learning features of the underlying contextual classifiers by adding (1) and (4) from above. We evaluate the extended contextual classifiers on MTurkMPQA via 10-fold cross-validation. Tables 4 and 5 hold the comparison of Method1 ($EXT_{S/O}$, $EXT_{N/P}$) to the original contextual classifiers ($O_{S/O}$, $O_{N/P}$) and to the rule-based SWSD integration (R1R2, R4). We see substantial improvement for Method1. It achieves 39% error reduction over $O_{S/O}$ and 25% error reduction over $O_{N/P}$. For both classifiers, the improvement in accuracy over the rule-based integration is statistically significant at the $p < .01$ level with McNemar's test.

3.3.2 Method2

This method defines a third classifier that accepts as input the contextual classifier's output and the SWSD output and predicts what the contextual classifier's output should have been. We can think of this third classifier as the learning counterpart of the manual rules from Section 3.2, since it actually learns when to flip the contextual classifier's output considering SWSD evidence. Specifically, this merger classifier relies on four machine learning features (1), (2), (3), (4) from above (the exact same information used in rule-based integration). Because it is a supervised classifier, we need training data where we have clue instances with corresponding contextual classifier and SWSD predictions. Fortunately, we can use senMPQA for this purpose. We train our merger classifier on senMPQA (we get contextual classifier predictions via 10-fold cross-validation on the MPQA Corpus) and apply it to MTurkMPQA. We use SVM classifier from the Weka package (Witten and Frank., 2005) with its default settings. Tables 4 and 5 hold the comparison of Method2 ($MERGER_{S/O}$, $MERGER_{N/P}$) to the original contextual classifiers (Oo/s, $O_{N/P}$) and the rule-based SWSD integration (R1R2, R4). It achieves 29% error reduction over $O_{S/O}$ and 29% error reduction over $O_{N/P}$. The improvement on the rule-based integration is statistically significant at the $p < .01$ level with McNemar's test. Method2 performs better (statistically significant at the $p < .05$ level) than Method1 for the N/P classifier but worse (statistically significant at the $p < .01$ level) for the S/O classifier.

	Acc	OF	SF
$O_{S/O}$	67.1	68.9	65.0
R1R2	71.1	72.7	69.2
$EXT_{S/O}$	**80.0**	81.4	78.3
$MERGER_{S/O}$	78.2	80.3	75.5

Table 4: S/O classifier with learned SWSD integration

	Acc	NF	PF
$O_{N/P}$	72.3	82.0	39.8
R4	74.5	84.0	37.8
$EXT_{N/P}$	79.1	85.7	61.1
$MERGER_{N/P}$	**80.4**	86.7	62.8

Table 5: N/P classifier with learned SWSD integration

3.4 Improving Contextual Polarity Classification

We have seen that Method2 is the best method to improve the N/P classifier, which is the first step of the contextual polarity classifier. To assess the overall improvement in polarity classification, we run the second step of the contextual polarity classifier after correcting the first step with Method2. Table 6 summarizes the improvement propagated to

		Acc	NF	NgF	PsF
MTurkMPQA	$O_{Ps/Ng/N}$	72.1	83.0	34.2	15.0
	$MERGER_{N/P}$	**77.8**	87.4	53.0	27.7
senMPQA	$O_{Ps/Ng/N}$	77.6	87.2	39.5	40.0
	R4	80.6	89.1	43.2	44.0

Table 6: Polarity classifier with and without SWSD.

Ps/Ng/N classification. For comparison, we also include results from (Akkaya et al., 2009) on senMPQA. Method2 results in 20% error reduction in accuracy over $O_{Ps/Ng/N}$ (R4 achieves 13.4% error reduction on senMPQA). The improvement on the rule-based integration is statistically significant at the $p < .01$ level with McNemar's test. More importantly, the F-measure for all the labels improves. This indicates that non-expert MTurk annotations can replace expert annotations for our end-goal – improving contextual opinion analysis – while reducing time and cost requirements by a large margin. Moreover, we see that the improvements in (Akkaya et al., 2009) scale up to new subjectivity clues.

4 Related Work

One related line of research is to automatically assign subjectivity and/or polarity labels to word senses in a dictionary (Valitutti et al., 2004; Andreevskaia and Bergler, 2006; Wiebe and Mihalcea, 2006; Esuli and Sebastiani, 2007; Su and Markert, 2009). In contrast, the task in our paper is to automatically assign labels to word instances in a corpus.

Recently, some researchers have exploited full word sense disambiguation in methods for opinion-related tasks. For example, (Martín-Wanton et al., 2010) exploit WSD for recognizing quotation polarities, and (Rentoumi et al., 2009; Martín-Wanton et al., 2010) exploit WSD for recognizing headline polarities. None of this previous work investigates performing a coarse-grained variation of WSD such as SWSD to improve their application results, as we do in this work.

A notable exception is (Su and Markert, 2010), who exploit SWSD to improve the performance on a contextual NLP task, as we do. While the task in our paper is subjectivity and sentiment analysis, their task is English-Chinese lexical substitution. As (Akkaya et al., 2009) did, they annotated word senses, and exploited SENSEVAL data as training data for SWSD. They did not directly annotate words in context with S/O labels, as we do in our work. Further, they did not separately evaluate a SWSD system component.

Many researchers work on reducing the granularity of sense inventories for WSD (e.g., (Palmer et al., 2004; Navigli, 2006; Snow et al., 2007; Hovy et al., 2006)). Their criteria for grouping senses are syntactic and semantic similarities, while the groupings in work on SWSD are driven by the goals to improve contextual subjectivity and sentiment analysis.

5 Conclusions and Future Work

In this paper, we utilized a large pool of non-expert annotators (MTurk) to collect subjectivity sense-tagged data for SWSD. We showed that non-expert annotations are as good as expert annotations for training SWSD classifiers. Moreover, we demonstrated that SWSD classifiers trained on non-expert annotations can be exploited to improve contextual opinion analysis.

The additional subjectivity sense-tagged data enabled us to evaluate the benefits of SWSD on contextual opinion analysis on a corpus of opinion-annotated data that is five times larger. Using the same rule-based integration strategies as in (Akkaya et al., 2009), we found that contextual opinion analysis is improved by SWSD on the larger datasets. We also experimented with new learning strategies for integrating SWSD into contextual opinion analysis. With the learning strategies, we achieved greater benefits from SWSD than the rule-based integration strategies on all of the contextual opinion analysis tasks.

Overall, we more firmly demonstrated the potential of SWSD to improve contextual opinion analysis. We will continue to gather subjectivity sense-tagged data, using sense inventories for words that are well represented in WordNet for our purposes, and with usage inventories for those that are not.

6 Acknowledgments

This material is based in part upon work supported by National Science Foundation awards #0917170 and #0916046.

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 24–32. Association for Computational Linguistics.

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 190–199, Singapore, August. Association for Computational Linguistics.

Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 195–203, Los Angeles, June. Association for Computational Linguistics.

Alina Andreevskaia and Sabine Bergler. 2006. 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 (EACL-2006)*.

Alina Andreevskaia and Sabine Bergler. 2008. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *Proceedings of ACL-08: HLT*, pages 290–298, Columbus, Ohio, June. Association for Computational Linguistics.

Kenneth Bloom, Navendu Garg, and Shlomo Argamon. 2007. Extracting appraisal expressions. In *HLT-NAACL 2007*, pages 308–315, Rochester, NY.

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. Association for Computational Linguistics.

Yaw Gyamfi, Janyce Wiebe, Rada Mihalcea, and Cem Akkaya. 2009. Integrating knowledge for subjectivity sense labeling. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2009)*, pages 10–18, Boulder, Colorado, June. Association for Computational Linguistics.

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.

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

Tamara Martín-Wanton, Aurora Pons-Porrata, Andrés Montoyo-Guijarro, and Alexandra Balahur. 2010. Opinion polarity detection - using word sense disambiguation to determine the polarity of opinions. In *ICAART 2010 - Proceedings of the International Conference on Agents and Artificial Intelligence, Volume 1*, pages 483–486.

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.

Randolph Quirk, Sidney Greenbaum, Geoffry Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, New York.

Vassiliki Rentoumi, George Giannakopoulos, Vangelis Karkaletsis, and George A. Vouros. 2009. Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proceedings of the International Conference RANLP-2009*, pages 370–375, Borovets, Bulgaria, September. Association for Computational Linguistics.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 105–112, Sapporo, Japan.

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.

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 (COLING-2008)*, Manchester.

Fangzhong Su and Katja Markert. 2009. Subjectivity recognition on word senses via semi-supervised mincuts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Fangzhong Su and Katja Markert. 2010. Word sense subjectivity for cross-lingual lexical substitution. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 357–360, Los Angeles, California, June. Association for Computational Linguistics.

Peter 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-02)*, pages 417–424, Philadelphia, Pennsylvania.

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

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal taxonomies for sentiment analysis. In *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*, Bremen, DE.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia, July. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210.

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 (HLT/EMNLP-2005)*, pages 347–354, Vancouver, Canada.

Theresa Wilson. 2007. *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.

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