Subjectivity Recognition on Word Senses via Semi-supervised Mincuts

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

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

Abstract

We supplement WordNet entries with information on the subjectivity of its word senses. Supervised classifiers that operate on word sense definitions in the same way that text classifiers operate on web or newspaper texts need large amounts of training data. The resulting data sparseness problem is aggravated by the fact that dictionary definitions are very short. We propose a semi-supervised minimum cut framework that makes use of both WordNet definitions and its relation structure. The experimental results show that it outperforms supervised minimum cut as well as standard supervised, non-graph classification, reducing the error rate by 40%. In addition, the semi-supervised approach achieves the same results as the supervised framework with less than 20% of the training data.

1 Introduction

There is considerable academic and commercial interest in processing subjective content in text, where subjective content refers to any expression of a private state such as an opinion or belief (Wiebe et al., 2005). Important strands of work include the identification of subjective content and the determination of its *polarity*, i.e. whether a favourable or unfavourable opinion is expressed.

Automatic identification of subjective content often relies on word indicators, such as unigrams (Pang et al., 2002) or predetermined sentiment lexica (Wilson et al., 2005). Thus, the word *positive* in the sentence *"This deal is a positive development for our company."* gives a strong indication that the sentence contains a favourable opinion. However, such word-based indicators can be misleading for two reasons. First, contextual indicators such as irony and negation can reverse subjectivity or polarity indications (Polanyi and Zaenen, 2004). Second, different word senses of a single word can actually be of different subjectivity or polarity. A typical *subjectivity-ambiguous* word, i.e. a word that has at least one subjective and at least one objective sense, is *positive*, as shown by the two example senses given below.[1]

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

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

We concentrate on this latter problem by automatically creating lists of subjective senses, instead of subjective words, via adding subjectivity labels for senses to electronic lexica, using the example of WordNet. This is important as the problem of subjectivity-ambiguity is frequent: We (Su and Markert, 2008) find that over 30% of words in our dataset are subjectivity-ambiguous. Information on subjectivity of senses can also improve other tasks such as word sense disambiguation (Wiebe and Mihalcea, 2006). Moreover, Andreevskaia and Bergler (2006) show that the performance of automatic annotation of subjectivity at the *word* level can be hurt by the presence of subjectivity-ambiguous words in the training sets they use.

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

Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pages 1–9,
Boulder, Colorado, June 2009.

We propose a semi-supervised approach based on minimum cut in a lexical relation graph to assign subjectivity (subjective/objective) labels to word senses.[2] Our algorithm outperforms supervised minimum cuts and standard supervised, non-graph classification algorithms (like SVM), reducing the error rate by up to 40%. In addition, the semi-supervised approach achieves the same results as the supervised framework with less than 20% of the training data. Our approach also outperforms prior approaches to the subjectivity recognition of word senses and performs well across two different data sets.

The remainder of this paper is organized as follows. Section 2 discusses previous work. Section 3 describes our proposed semi-supervised minimum cut framework in detail. Section 4 presents the experimental results and evaluation, followed by conclusions and future work in Section 5.

2 Related Work

There has been a large and diverse body of research in opinion mining, with most research at the text (Pang et al., 2002; Pang and Lee, 2004; Popescu and Etzioni, 2005; Ounis et al., 2006), sentence (Kim and Hovy, 2005; Kudo and Matsumoto, 2004; Riloff et al., 2003; Yu and Hatzivassiloglou, 2003) or word (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Kim and Hovy, 2004; Takamura et al., 2005; Andreevskaia and Bergler, 2006; Kaji and Kitsuregawa, 2007) level. An up-to-date overview is given in Pang and Lee (2008).

Graph-based algorithms for classification into subjective/objective or positive/negative language units have been mostly used at the sentence and document level (Pang and Lee, 2004; Agarwal and Bhattacharyya, 2005; Thomas et al., 2006), instead of aiming at dictionary annotation as we do. We also cannot use prior graph construction methods for the document level (such as physical proximity of sentences, used in Pang and Lee (2004)) at the word sense level. At the word level Takamura et al. (2005) use a semi-supervised spin model for word polarity determination, where the graph is constructed using a variety of information such as gloss co-occurrences and WordNet links. Apart from using a different graph-based model from ours, they assume that subjectivity recognition has already been achieved prior to polarity recognition and test against word lists containing subjective words only. However, Kim and Hovy (2004) and Andreevskaia and Bergler (2006) show that subjectivity recognition might be the harder problem with lower human agreement and automatic performance. In addition, we deal with classification at the *word sense* level, treating also subjectivity-ambiguous words, which goes beyond the work in Takamura et al. (2005).

Word Sense Level: There are three prior approaches addressing *word sense* subjectivity or polarity classification. Esuli and Sebastiani (2006) determine the polarity (positive/negative/objective) of word senses in WordNet. However, there is no evaluation as to the accuracy of their approach. They then extend their work (Esuli and Sebastiani, 2007) by applying the Page Rank algorithm to rank the WordNet senses in terms of how strongly a sense possesses a given semantic property (e.g., positive or negative). Apart from us tackling subjectivity instead of polarity, their Page Rank graph is also constructed focusing on WordNet glosses (linking glosses containing the same words), whereas we concentrate on the use of WordNet relations.

Both Wiebe and Mihalcea (2006) and our prior work (Su and Markert, 2008) present an annotation scheme for word sense subjectivity and algorithms for automatic classification. Wiebe and Mihalcea (2006) use an algorithm relying on distributional similarity and an independent, large manually annotated opinion corpus (MPQA) (Wiebe et al., 2005). One of the disadvantages of their algorithm is that it is restricted to senses that have distributionally similar words in the MPQA corpus, excluding 23% of their test data from automatic classification. Su and Markert (2008) present supervised classifiers, which rely mostly on WordNet glosses and do not effectively exploit WordNet's relation structure.

3 Semi-Supervised Mincuts

3.1 Minimum Cuts: The Main Idea

Binary classification with minimum cuts (Mincuts) in graphs is based on the idea that similar items

[2] It can be argued that subjectivity labels are maybe rather more graded than the clear-cut binary distinction we assign. However, in Su and Markert (2008a) as well as Wiebe and Mihalcea (2006) we find that human can assign the binary distinction to word senses with a high level of reliability.

should be grouped in the same cut. All items in the training/test data are seen as vertices in a graph with undirected weighted edges between them specifying how strong the similarity/association between two vertices is. We use minimum s-t cuts: the graph contains two particular vertices s (source, corresponds to subjective) and t (sink, corresponds to objective) and each vertex u is connected to s and t via a weighted edge that can express how likely u is to be classified as s or t in isolation.

Binary classification of the vertices is equivalent to splitting the graph into two disconnected subsets of all vertices, S and T with $s \in S$ and $t \in T$. This corresponds to removing a set of edges from the graph. As similar items should be in the same part of the split, the best split is one which removes edges with low weights. In other words, a minimum cut problem is to find a partition of the graph which minimizes the following formula, where $w(u, v)$ expresses the weight of an edge between two vertices.

$$W(S,T) = \sum_{u \in S, v \in T} w(u,v)$$

Globally optimal minimum cuts can be found in polynomial time and near-linear running time in practice, using the maximum flow algorithm (Pang and Lee, 2004; Cormen et al., 2002).

3.2 Why might Semi-supervised Minimum Cuts Work?

We propose semi-supervised mincuts for subjectivity recognition on senses for several reasons.

First, our problem satisfies two major conditions necessary for using minimum cuts. It is a binary classification problem (subjective vs. objective senses) as is needed to divide the graph into two components. Our dataset also lends itself naturally to s-t Mincuts as we have two different views on the data. Thus, the edges of a vertex (=sense) to the source/sink can be seen as the probability of a sense being subjective or objective without taking similarity to other senses into account, for example via considering only the sense gloss. In contrast, the edges between two senses can incorporate the WordNet relation hierarchy, which is a good source of similarity for our problem as many WordNet relations are *subjectivity-preserving*, i.e. if two senses are connected via such a relation they are likely to be both subjective or both objective.[3] An example here is the antonym relation, where two antonyms such as *good—morally admirable* and *evil, wicked—morally bad or wrong* are both subjective.

Second, Mincuts can be easily expanded into a semi-supervised framework (Blum and Chawla, 2001). This is essential as the existing labeled datasets for our problem are small. In addition, glosses are short, leading to sparse high dimensional vectors in standard feature representations. Also, WordNet connections between different parts of the WordNet hierarchy can also be sparse, leading to relatively isolated senses in a graph in a supervised framework. Semi-supervised Mincuts allow us to import unlabeled data that can serve as bridges to isolated components. More importantly, as the unlabeled data can be chosen to be related to the labeled and test data, they might help pull test data to the right cuts (categories).

3.3 Formulation of Semi-supervised Mincuts

The formulation of our semi-supervised Mincut for sense subjectivity classification involves the following steps, which we later describe in more detail.

1. We define two vertices s (source) and t (sink), which correspond to the "subjective" and "objective" category, respectively. Following the definition in Blum and Chawla (2001), we call the vertices s and t *classification vertices*, and all other vertices (labeled, test, and unlabeled data) *example vertices*. Each example vertex corresponds to one WordNet sense and is connected to both s and t via a weighted edge. The latter guarantees that the graph is connected.
2. For the test and unlabeled examples, we see the edges to the classification vertices as the probability of them being subjective/objective disregarding other example vertices. We use a supervised classifier to set these edge weights. For the labeled training examples, they are connected by edges with a high constant weight to the classification vertices that they belong to.
3. WordNet relations are used to construct the edges *between two example vertices*. Such

[3] See Kamps et al. (2004) for an early indication of such properties for some WordNet relations.

edges can exist between any pair of example vertices, for example between two unlabeled examples.

4. After graph construction we then employ a maximum-flow algorithm to find the minimum s-t cuts of the graph. The cut in which the source vertex s lies is classified as "subjective", and the cut in which the sink vertex t lies is "objective".

We now describe the above steps in more detail.

Selection of unlabeled data: Random selection of unlabeled data might hurt the performance of Mincuts, as they might not be related to any sense in our training/test data (denoted by A). Thus a basic principle is that the selected unlabeled senses should be related to the training/test data by WordNet relations. We therefore simply scan each sense in A, and collect all senses related to it via one of the WordNet relations in Table 1. All such senses that are not in A are collected in the unlabeled data set.

Weighting of edges to the classification vertices: The edge weight to s and t represents how likely it is that an example vertex is initially put in the cut in which s (subjective) or t (objective) lies. For unlabeled and test vertices, we use a supervised classifier (SVM[4]) with the labeled data as training data to assign the edge weights. The SVM is also used as a baseline and its features are described in Section 4.3. As we do not wish the Mincut to reverse labels of the labeled training data, we assign a high constant weight of 5 to the edge between a labeled vertex and its corresponding classification vertex, and a low weight of 0.01 to the edge to the other classification vertex.

Assigning weights to WordNet relations: We connect two vertices that are linked by one of the ten WordNet relations in Table 1 via an edge. Not all WordNet relations we use are subjectivity-preserving to the same degree: for example, hyponyms (such as *simpleton*) of objective senses (such as *person*) do not have to be objective. However, we aim for high graph connectivity and we can assign different weights to different relations to reflect the degree to which they are subjectivity-preserving. Therefore, we experiment with two methods of weight assignment. Method 1 (NoSL) assigns the same constant weight of 1.0 to all WordNet relations.

Method 2 (SL) reflects different degrees of preserving subjectivity. To do this, we adapt an unsupervised method of generating a large noisy set of subjective and objective senses from our previous work (Su and Markert, 2008). This method uses a list of subjective words (SL)[5] to classify each WordNet sense with at least two subjective words in its gloss as subjective and all other senses as objective. We then count how often two senses related via a given relation have the same or a different subjectivity label. The weight is computed by #same/(#same+#different). Results are listed in Table 1.

Table 1: Relation weights (Method 2)

Method	#Same	#Different	Weight
Antonym	2,808	309	0.90
Similar-to	6,887	1,614	0.81
Derived-from	4,630	947	0.83
Direct-Hypernym	71,915	8,600	0.89
Direct-Hyponym	71,915	8,600	0.89
Attribute	350	109	0.76
Also-see	1,037	337	0.75
Extended-Antonym	6,917	1,651	0.81
Domain	4,387	892	0.83
Domain-member	4,387	892	0.83

Example graph: An example graph is shown in Figure 1. The three example vertices correspond to the senses **religious**—*extremely scrupulous and conscientious,* **scrupulous**—*having scruples; arising from a sense of right and wrong; principled*; and **flicker**, *spark, glint—a momentary flash of light* respectively. The vertex "scrupulous" is unlabeled data derived from the vertex "religious"(a test item) by the relation "similar-to".

4 Experiments and Evaluation

4.1 Datasets

We conduct the experiments on two different gold standard datasets. One is the Micro-WNOp corpus,

[4]We employ LIBSVM, available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`. Linear kernel and probability estimates are used in this work.

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

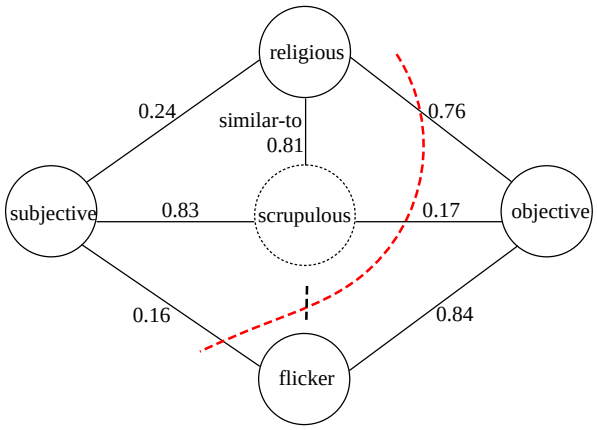

Figure 1: Graph of Word Senses

which is representative of the part-of-speech distribution in WordNet [6]. It includes 298 words with 703 objective and 358 subjective WordNet senses. The second one is the dataset created by Wiebe and Mihalcea (2006).[7] It only contains noun and verb senses, and includes 60 words with 236 objective and 92 subjective WordNet senses. As the Micro-WNOp set is larger and also contains adjective and adverb senses, we describe our results in more detail on that corpus in the Section 4.3 and 4.4. In Section 4.5, we shortly discuss results on Wiebe&Mihalcea's dataset.

4.2 Baseline and Evaluation

We compare to a baseline that assigns the most frequent category *objective* to all senses, which achieves an accuracy of 66.3% and 72.0% on Micro-WNOp and Wiebe&Mihalcea's dataset respectively. We use the McNemar test at the significance level of 5% for significance statements. All evaluations are carried out by 10-fold cross-validation.

4.3 Standard Supervised Learning

We use an SVM classifier to compare our proposed semi-supervised Mincut approach to a reasonable baseline.[8] Three different feature types are used.

Lexical Features (L): a bag-of-words representation of the sense glosses with stop word filtering.

Relation Features (R): First, we use two features for each of the ten WordNet relations in Table 1, describing how many relations of that type the sense has to senses in the subjective or objective part of the training set, respectively. This provides a non-graph summary of subjectivity-preserving links. Second, we manually collected a small set (denoted by $SubjSet$) of seven subjective verb and noun senses which are close to the root in WordNet's hypernym tree. A typical example element of $SubjSet$ is *psychological feature —a feature of the mental life of a living organism*, which indicates subjectivity for its hyponyms such as *hope — the general feeling that some desire will be fulfilled.* A binary feature describes whether a noun/verb sense is a hyponym of an element of $SubjSet$.

Monosemous Feature (M): for each sense, we scan if a monosemous word is part of its synset. If so, we further check if the monosemous word is collected in the subjective word list (SL). The intuition is that if a monosemous word is subjective, obviously its (single) sense is subjective. For example, the sense *uncompromising, inflexible—not making concessions* is subjective, as "uncompromising" is a monosemous word and also in SL.

We experiment with different combinations of features and the results are listed in Table 2, prefixed by "SVM". All combinations perform significantly better than the more frequent category baseline and similarly to the supervised Naive Bayes classifier (see S&M in Table 2) we used in Su and Markert (2008). However, improvements by adding more features remain small.

In addition, we compare to a supervised classifier (see Lesk in Table 2) that just assigns each sense the subjectivity label of its most similar sense in the training data, using Lesk's similarity measure from Pedersen's WordNet similarity package[9]. We use Lesk as it is one of the few measures applicable across all parts-of-speech.

[6] Available at `http://www.comp.leeds.ac.uk/markert/data`. This dataset was first used with a different annotation scheme in Esuli and Sebastiani (2007) and we also used it in Su and Markert (2008).

[7] Available at `http://www.cs.pitt.edu/~wiebe/pubs/papers/goldstandard.total.acl06`.

[8] This SVM is also used to provide the edge weights to the *classification vertices* in the Mincut approach.

[9] Available at `http://www.d.umn.edu/~tpederse/similarity.html`.

Table 2: Results of SVM and Mincuts with different settings of feature

Method	Subjective			Objective			Accuracy
	Precision	Recall	F-score	Precision	Recall	F-score	
Baseline	N/A	0	N/A	66.3%	**100%**	79.7%	66.3%
S&M	66.2%	64.5%	65.3%	82.2%	83.2%	82.7%	76.9%
Lesk	65.6%	50.3%	56.9%	77.5%	86.6%	81.8%	74.4%
SVM-L	69.6%	37.7%	48.9%	74.3%	91.6%	82.0%	73.4%
L-SL	82.0%	43.3%	56.7%	76.7%	95.2%	85.0%	77.7%
L-NoSL	80.8%	43.6%	56.6%	76.7%	94.7%	84.8%	77.5%
SVM-LM	68.9%	42.2%	52.3%	75.4%	90.3%	82.2%	74.1%
LM-SL	83.2%	44.4%	57.9%	77.1%	95.4%	85.3%	78.2%
LM-NoSL	83.6%	44.1%	57.8%	77.1%	95.6%	85.3%	78.2%
SVM-LR	68.4%	45.3%	54.5%	76.2%	89.3%	82.3%	74.5%
LR-SL	82.7%	65.4%	73.0%	84.1%	93.0%	88.3%	83.7%
LR-NoSL	82.4%	65.4%	72.9%	84.0%	92.9%	88.2%	83.6%
SVM-LRM	69.8%	47.2%	56.3%	76.9%	89.6%	82.8%	75.3%
LRM-SL	**85.5%**	65.6%	**74.2%**	**84.4%**	94.3%	**89.1%**	**84.6%**
LRM-NoSL	84.6%	**65.9%**	74.1%	**84.4%**	93.9%	88.9%	84.4%

[1] L, R and M correspond to the lexical, relation and monosemous features respectively.

[2] SVM-L corresponds to using lexical features only for the SVM classifier. Likewise, SVM-LRM corresponds to using a combination for lexical, relation, and monosemous features for the SVM classifier.

[3] L-SL corresponds to the Mincut that uses only lexical features for the SVM classifier, and subjective list (SL) to infer the weight of WordNet relations. Likewise, LM-NoSL corresponds to the Mincut algorithm that uses lexical and monosemous features for the SVM, and predefined constants for WordNet relations (without subjective list).

4.4 Semi-supervised Graph Mincuts

Using our formulation in Section 3.3, we import 3,220 senses linked by the ten WordNet relations to any senses in Micro-WNOp as unlabeled data. We construct edge weights to classification vertices using the SVM discussed above and use WordNet relations for links between example vertices, weighted by either constants (NoSL) or via the method illustrated in Table 1 (SL). The results are also summarized in Table 2. Semi-supervised Mincuts always significantly outperform the corresponding SVM classifiers, regardless of whether the subjectivity list is used for setting edge weights. We can also see that we achieve good results without using any other knowledge sources (setting LR-NoSL).

The example in Figure 1 explains why semi-supervised Mincuts outperforms the supervised approach. The vertex "religious" is initially assigned the subjective/objective probabilities 0.24/0.76 by the SVM classifier, leading to a wrong classification. However, in our graph-based Mincut framework, the vertex "religious" might link to other vertices (for example, it links to the vertex "scrupulous" in the unlabeled data by the relation "similar-to"). The mincut algorithm will put vertices "religious" and "scrupulous" in the same cut (subjective category) as this results in the least cost 0.93 (ignoring the cost of assigning the unrelated sense of "flicker"). In other words, the edges between the vertices are likely to correct some initially *wrong* classification and pull the vertices into the *right* cuts.

In the following we will analyze the best minimum cut algorithm LRM-SL in more detail. We measure its accuracy for each part-of-speech in the Micro-WNOp dataset. The number of noun, adjective, adverb and verb senses in Micro-WNOp is 484, 265, 31 and 281, respectively. The result is listed in Table 3. The significantly better performance of semi-supervised mincuts holds across all parts-of-speech but the small set of adverbs, where there is no significant difference between the baseline, SVM and the Mincut algorithm.

Table 3: Accuracy for Different Part-Of-Speech

Method	Noun	Adjective	Adverb	Verb
Baseline	76.9%	61.1%	77.4%	72.6%
SVM	81.4%	63.4%	**83.9%**	75.1%
Mincut	**88.6%**	**78.9%**	77.4%	**84.0%**

We will now investigate how LRM-SL performs with different sizes of labeled and unlabeled data. All learning curves are generated via averaging 10 learning curves from 10-fold cross-validation.

Performance with different sizes of labeled data: we randomly generate subsets of labeled data A_1, A_2... A_n, and guarantee that $A_1 \subset A_2 ... \subset A_n$. Results for the best SVM (LRM) and the best minimum cut (LRM-SL) are listed in Table 4, and the corresponding learning curve is shown in Figure 2. As can be seen, the semi-supervised Mincuts is consistently better than SVM. Moreover, the semi-supervised Mincut with only 200 labeled data items performs even better than SVM with 954 training items (78.9% *vs* 75.3%), showing that our semi-supervised framework allows for a training data reduction of more than 80%.

Table 4: Accuracy with different sizes of labeled data

# labeled data	SVM	Mincuts
100	69.1%	72.2%
200	72.6%	78.9%
400	74.4%	82.7%
600	75.5%	83.7%
800	76.0%	84.1%
900	75.6%	84.8%
954 (all)	75.3%	84.6%

Performance with different sizes of unlabeled data: We propose two different settings.

Option1: Use a subset of the ten relations to generate the unlabeled data (and edges between example vertices). For example, we first use {antonym, similar-to} only to obtain a unlabeled dataset U_1, then use a larger subset of the relations like {antonym, similar-to, direct-hyponym, direct-hypernym} to generate another unlabeled dataset U_2, and so forth. Obviously, U_i is a subset of U_{i+1}.

Option2: Use all the ten relations to generate the unlabeled data U. We then randomly select subsets of U, such as subset U_1, U_2 and U_3, and guarantee that $U_1 \subset U_2 \subset U_3 \subset \ldots U$.

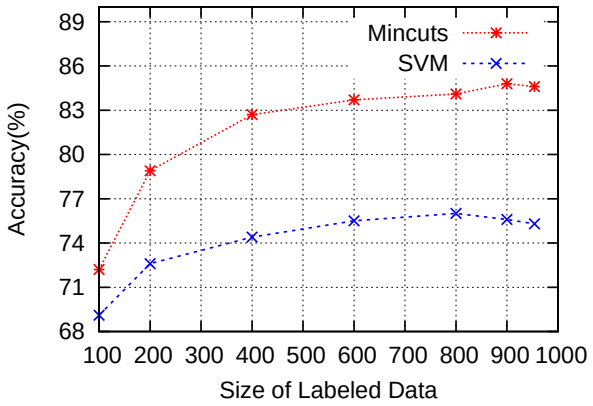

Figure 2: Learning curve with different sizes of labeled data

The results are listed in Table 5 and Table 6 respectively. The corresponding learning curves are shown in Figure 3. We see that performance improves with the increase of unlabeled data. In addition, the curves seem to converge when the size of unlabeled data is larger than 3,000. From the results in Tabel 5 one can also see that hyponymy is the relation accounting for the largest increase.

Table 6: Accuracy with different sizes of unlabeled data (random selection)

# unlabeled data	Accuracy
0	75.9%
200	76.5%
500	78.6%
1000	80.2%
2000	82.8%
3000	84.0%
3220	84.6%

Furthermore, these results also show that a supervised mincut without unlabeled data performs only on a par with other supervised classifiers (75.9%). The reason is that if we exclude the unlabeled data, there are only 67 WordNet relations/edges between senses in the small Micro-WNOp dataset. In contrast, the use of unlabeled data adds more edges (4,586) to the graph, which strongly affects the graph cut partition (see also Figure 1).

4.5 Comparison to Prior Approaches

In our previous work (Su and Markert, 2008), we report 76.9% as the best accuracy on the same Micro-

Table 5: Accuracy with different sizes of unlabeled data from WordNet relation

Relation	# unlabeled data	Accuracy
{∅}	0	75.3%
{similar-to}	418	79.1%
{similar-to, antonym}	514	79.5%
{similar-to, antonym, direct-hypernym, direct-hyponym}	2,721	84.4%
{similar-to, antonym, direct-hypernym, direct-hyponym, also-see, extended-antonym}	3,004	84.4%
{similar-to, antonym, direct-hypernym, direct-hyponym, also-see, extended-antonym, derived-from, attribute, domain, domain-member}	3,220	84.6%

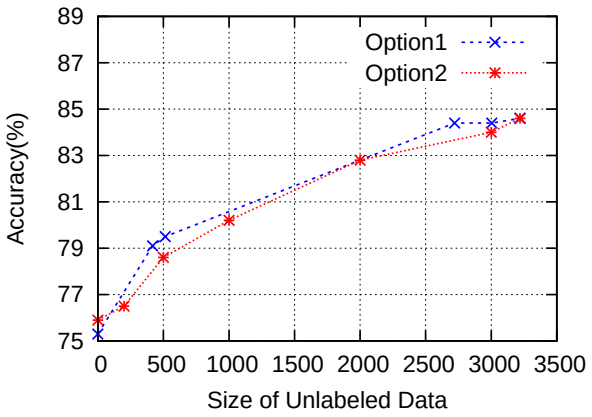

Figure 3: Learning curve with different sizes of unlabeled data

WNOp dataset used in the previous sections, using a supervised Naive Bayes (S&M in Tabel 2). Our best result from Mincuts is significantly better at 84.6% (see LRM-SL in Table 2).

For comparison to Wiebe and Mihalcea (2006), we use their dataset for testing, henceforth called `Wiebe` (see Section 4.1 for a description). Wiebe and Mihalcea (2006) report their results in precision and recall curves for subjective senses, such as a precision of about 55% at a recall of 50% for subjective senses. Their F-score for subjective senses seems to remain relatively static at 0.52 throughout their precision/recall curve.

We run our best Mincut LRM-SL algorithm with two different settings on `Wiebe`. Using Micro-WNOp as training set and `Wiebe` as test set, we achieve an accuracy of 83.2%, which is similar to the results on the Micro-WNOp dataset. At the recall of 50% we achieve a precision of 83.6% (in comparison to their precision of 55% at the same recall). Our F-score is 0.63 (vs. 0.52).

To check whether the high performance is just due to our larger training set, we also conduct 10-fold cross-validation on `Wiebe`. The accuracy achieved is 81.1% and the F-score 0.56 (vs. 0.52), suggesting that our algorithm performs better. Our algorithm can be used on all WordNet senses whereas theirs is restricted to senses that have distributionally similar words in the MPQA corpus (see Section 2). However, they use an unsupervised algorithm i.e. they do not need labeled word senses, although they do need a large, manually annotated opinion corpus.

5 Conclusion and Future Work

We propose a semi-supervised minimum cut algorithm for subjectivity recognition on word senses. The experimental results show that our proposed approach is significantly better than a standard supervised classification framework as well as a supervised Mincut. Overall, we achieve a 40% reduction in error rates (from an error rate of about 25% to an error rate of 15%). To achieve the results of standard supervised approaches with our model, we need less than 20% of their training data. In addition, we compare our algorithm to previous state-of-the-art approaches, showing that our model performs better on the same datasets.

Future work will explore other graph construction methods, such as the use of morphological relations as well as thesaurus and distributional similarity measures. We will also explore other semi-supervised algorithms.

References

Alekh Agarwal and Pushpak Bhattacharyya. 2005. Sentiment Analysis: A new Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified. *Proceedings of ICON'05*.

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

Avrim Blum and Shuchi Chawla. 2001. Learning from Labeled and Unlabeled Data using Graph Mincuts. *Proceedings of ICML'01*.

Thomas Cormen, Charles Leiserson, Ronald Rivest and Clifford Stein. 2002. Introduction to Algorithms. *Second Edition, the MIT Press*.

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

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

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

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

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. *Proceedings of EMNLP'07*.

Japp Kamps, Maarten Marx, Robert Mokken, and Maarten de Rijke. 2004. Using WordNet to Measure Semantic Orientation of Adjectives. *Proceedings of LREC'04*.

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

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

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

Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne and Ian Soboroff. 2006. Overview of the TREC-2006 Blog Track. *Proceedings of TREC'06*.

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

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval 2(1-2)*.

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

Livia Polanyi and Annie Zaenen. 2004. Contextual Valence Shifters. *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting Product Features and Opinions from Reviews *Proceedings of EMNLP'05*.

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

Fangzhong Su and Katja Markert. 2008. From Words to Senses: A Case Study in Subjectivity Recognition. *Proceedings of COLING'08*.

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

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

Matt Thomas, Bo Pang and Lilian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *Proceedings of EMNLP'06*.

Peter Turney. 2002. Thumbs up or Thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of ACL'02*.

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

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proceedings of EMNLP'03*.

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

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

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of HLT/EMNLP'05*.