Learning Subjective Language

Janyce Wiebe*, Theresa Wilson*,

Rebecca Bruce†, Matthew Bell*,

Melanie Martin‡

University of Pittsburgh*, University of North Carolina at Asheville†,

New Mexico State University‡

Subjectivity in natural language refers to aspects of language used to express opinions, evaluations, and speculations. There are numerous NLP applications for which subjectivity analysis is relevant, including information extraction and text categorization. The goal of this work is learning subjective language from corpora. Clues of subjectivity are generated and tested, including low-frequency words, collocations, and adjectives and verbs identified using distributional similarity. The features are also examined working together in concert. The features, generated from different datasets using different procedures, exhibit consistency in performance in that they all do better and worse on the same datasets. In addition, this paper shows that the density of subjectivity clues in the surrounding context strongly affects how likely it is that a word is subjective, and gives the results of an annotation study assessing the subjectivity of sentences with high-density features. Finally, the clues are used to perform opinion-piece recognition (a type of text categorization and genre detection), to demonstrate the utility of the knowledge acquired in this paper.

1 Introduction

Subjectivity in natural language refers to aspects of language used to express opinions, evaluations, and speculations (Banfield, 1982; Wiebe, 1994). Many natural language processing applications could benefit from being able to distinguish subjective language from language used to objectively present factual information. Current extraction and retrieval technology focuses almost exclusively on the subject matter of documents. However, additional aspects of a document influence its relevance, including evidential status and attitude (Kessler, Nunberg, and Schütze, 1997). Information extraction systems should be able to distinguish between factual information (which should be extracted) and non-factual information (which should be discarded or labeled as uncertain). Question answering systems should distinguish between factual and speculative answers. Multi-perspective question answering aims to present multiple answers to the user based upon speculation or opinions derived from different sources (Carbonell, 1979; Wiebe et al., 2003). Multi-document summarization systems should summarize different opinions and perspectives. Automatic subjectivity analysis would also be useful to perform flame recognition (Spertus, 1997; Kaufer, 2000), email classification (Aone, Ramos-Santacruz, and Niehaus, 2000), intellectual attribution in text (Teufel and Moens, 2000), recognizing speaker role in radio broadcasts (Barzilay et al., 2000), review mining (Terveen et al., 1997), review classification (Turney, 2002; Pang, Lee, and Vaithyanathan, 2002), style in generation (Hovy, 1987), and clustering documents by ideological point of view (Sack, 1995). In general, nearly any information seeking system could benefit from knowledge of how opinionated a text is, and whether or not the writer purports to objectively present factual material.

To perform automatic subjectivity analysis, good clues must be found. A huge variety of words and phrases have subjective usages and, while some manually developed

resources exist, such as dictionaries of affective language (General-Inquirer, 2000; Heise, 2000) and subjective features in general purpose lexicons (e.g., the *attitude* adverb features in Comlex (Macleod, Grishman, and Meyers, 1998)) there is no comprehensive dictionary of subjective language. In addition, many expressions with subjective usages have objective usages as well, so a dictionary alone would not suffice. An NLP system must disambiguate these expressions in context.

The goal of our work is learning subjective language from corpora. In this paper, we generate and test subjectivity clues and contextual features, and use the knowledge we gain to recognize subjective sentences and opinionated documents.

Two kinds of data are available to us: a relatively small amount of data manually annotated at the expression level (i.e., labels on individual words and phrases) of Wall Street Journal and newsgroup data, and a large amount of data with existing document-level annotations from the Wall Street Journal (*opinion pieces*, such as editorials and reviews, versus *non-opinion pieces*). Both are used as training data to identify clues of subjectivity. In addition, we cross validate the results between the two types of annotation: the clues learned from the expression-level data are evaluated against the document-level annotations, and those learned using the document-level annotations are evaluated against the expression-level annotations.

There were a number of motivations behind our decision to use document-level annotations, in addition to our manual annotations, to identify and evaluate clues of subjectivity. These annotations were not produced according to our annotation scheme, and were not produced for the purpose of training and evaluating an NLP system. Thus, they are an external influence from outside the laboratory. In addition, there is a great deal of this data, enabling us to evaluate the results on a larger scale, using multiple large test sets. This and cross training between the two types of annotations allows us to assess consistency in performance of the various identification procedures. Good performance

in cross validation experiments between different types of annotations is evidence that the results are not brittle.

We focus on three types of subjectivity clues. The first are *hapax legomena*, the set of words that appear just once in the corpus. We refer to them here as *unique words*. The set of all unique words is a feature with high frequency and significantly higher precision than baseline (Section 3.2).

The second are collocations (Section 3.3). We demonstrate a straightforward method for automatically identifying collocational clues of subjectivity in texts. The method is first used to identify fixed n-grams, such as *of the century* and *get out of here*. Interestingly, many include non-content words that are typically on stop lists of NLP systems (e.g., *of, the, get, out, here* in the above examples). The method is then used to identify an unusual form of collocation: one or more positions in the collocation may be filled by any word (of an appropriate part of speech) that is unique in the test data.

The third are adjective and verb features identified using the results of a method for clustering words according to distributional similarity (Lin, 1998) (Section 3.4). We hypothesized that words may be distributionally similar because they are both potentially subjective (e.g., *tragic*, *sad*, and *poignant* are identified from *bizarre*). In addition, we use distributional similarity to improve estimates of unseen events: words are selected or discarded based on the precision of it together with its N most similar neighbors.

We show that the various subjectivity clues perform better and worse on the same datasets, exhibiting an important consistency in performance (Section 4.2).

In addition to learning and evaluating clues associated with subjectivity, we address disambiguating them in context, i.e., identifying instances of clues that are subjective in context (Sections 4.3 and 4.4). We find that the density of clues in the surrounding context is an important influence. Using two types of annotations serves us well here, too. It enables us to use manual judgments to identify parameters for disambiguating

instances of automatically identified clues. High-density clues are high precision on both the expression-level and document-level data. In addition, we give the results of a new annotation study showing that most high-density clues are in subjective text spans (Section 4.5).

Finally, we use the clues together to perform document-level classification, to further demonstrate the utility of the acquired knowledge (Section 4.6).

At the end of the paper, we discuss related work (Section 5) and conclusions and (Section 6).

2 Subjectivity

Subjective language is language used to express *private states* in the context of a text or conversation. *Private state* is a general covering term for opinions, evaluations, emotions, and speculations (Quirk et al., 1985). Following are examples of subjective sentences from a variety of document types.

The first two examples are from Usenet newsgroup messages.

(1) I had in mind your facts, buddy, not hers.

(2) Nice touch. "Alleges" whenever facts posted are not in your persona of what is "real".

The next one is from an editorial:

(3) We stand in awe of the Woodstock generation's ability to be unceasingly fascinated by the subject of itself. ["Bad Acid", *Wall Street Journal*, 8/17/89]

The next example is from a book review:

(4) At several different layers, it's a fascinating tale. ["Whose Spying on Our Computers", George Melloan, *Wall Street Journal*, 11/1/89]

The last one is from a news story:

> (5) "The cost of health care is eroding our standard of living and sapping industrial strength," complains Walter Maher, a Chrysler health-and-benefits specialist. ["Business and Labor Reach a Consensus on Need to Overhaul Health-Care System", Kenneth H. Bacon, *Wall Street Journal*, 11/1/89]

In contrast, following are examples of *objective sentences*, sentences without significant expressions of subjectivity:

> (6) Bell Industries Inc. increased its quarterly to 10 cents from 7 cents a share.
>
> (7) Northwest Airlines settled the remaining lawsuits filed on behalf of 156 people killed in a 1987 crash, but claims against the jetliner's maker are being pursued, a federal judge said. ["Northwest Airlines Settles Rest of Suits", *Wall Street Journal*, 11/1/89]

A particular model of linguistic subjectivity underlies the current and past research in this area by Wiebe and colleagues. It is most fully presented in (Wiebe and Rapaport, 1986; Wiebe and Rapaport, 1988; Wiebe, 1990; Wiebe and Rapaport, 1991; Wiebe, 1994). It was developed to support NLP research, and combines ideas from several sources in fields outside NLP, especially linguistics and literary theory. The most direct influences were Doležel (1973) (types of subjectivity clues), Uspensky (1973) (types of point of view), Kuroda (1973; 1976) (pragmatics of point of view), Chatman (1978) (story versus discourse), Cohn (1978) (linguistic styles for presenting consciousness), J. Fodor (1979) (linguistic description of opaque contexts), and especially Banfield (1982) (theory of subjectivity versus communication).[1]

1 For additional citations to relevant work from outside NLP, please see (Banfield, 1982; Fludernik,

The remainder of this section sketches our conceptualization of subjectivity, and describes the annotation projects it underlies.

Subjective elements are linguistic expressions of private states in context. Subjective elements are often lexical (examples are *stand in awe*, *unceasingly*, *fascinated* in (3) and *eroding, sapping*, and *complains* in (5)). They may be single words (e.g., *complains*) or more complex expressions (e.g., *stand in awe*, *What a* NP). Purely syntactic or morphological devices may also be subjective elements (e.g., fronting, parallelism, changes in aspect).

A subjective element expresses the subjectivity of a *source*, who may be the writer or someone mentioned in the text. For example, the source of *fascinating* in (4) is the writer, while the source of the subjective elements in (5) is Maher (according to the writer). In addition, a subjective element usually has a *target*, i.e., what the subjectivity is about or directed toward. In (4), the target is a tale; in (5), the target of Maher's subjectivity is the cost of health care.

Note our parenthetical above – "according to the writer" – concerning Maher's subjectivity. Maher is not directly speaking to us, but is being quoted by the writer. Thus, the source is a *nested source*, which we notate (writer, Maher); this represents the fact that the subjectivity is being attributed to Maher, by the writer. Since sources are not directly addressed by the experiments presented in this paper, we merely illustrate the idea here with an example, to give the reader an idea.

> The Foreign Ministry said Thursday that it was "surprised, to put it mildly" by the U.S. State Department's criticism of Russia's human rights record and objected in particular to the "odious" section on Chechnya. [*Moscow Times*, 03/08/2002]

1993; Wiebe, 1994; Stein and Wright, 1995).

Let us consider some of the subjective elements in this sentence, along with their sources:

surprised, to put it mildly: (writer, Foreign Ministry, Foreign Ministry)

to put it mildly: (writer, Foreign Ministry)

criticism: (writer, Foreign Ministry, Foreign Ministry, U.S. State Dept.)

objected: (writer, Foreign Ministry)

odious: (writer, Foreign Ministry)

Consider *surprised, to put it mildly.* This refers to a private state of the Foreign Ministry (i.e., it is very surprised). This is in the context of "The Foreign Minstry said," which is in a sentence written by the writer. This gives us the 3-level source (writer, Foreign Ministry, Foreign Ministry). The phrase *to put it mildly*, which expresses sarcasm, is attributed to the Foreign Ministry by the writer (i.e., according to the writer, the Foreign Ministry said this). So its source is (writer, Foreign Ministry). The subjective element **criticism** has a deeply nested source: according to the **writer**, the **Foreign Ministry** said **it** is surprised by the **U.S. State Dept**'s criticism.

The nested-source representation allows us to pinpoint the subjectivity in a sentence. For example, there is no subjectivity attributed directly to the writer in the above sentence: at the level of the writer, the sentence merely says that someone said something and objected to something (without evaluating or questioning this). If the sentence started "The magnificent Foreign Ministry said..." then we would have an additional subjective element, **magnificent**, with source (writer).

Note that *subjective* does not mean *not true.* Consider the sentence "John criticized Mary for smoking." The verb *criticized* is a subjective element, expressing negative evaluation, with nested source (writer, John). But this does not mean that John does *not* believe that Mary smokes. (In addition, the fact that John criticized Mary is being presented as true by the writer.)

Similarly, *objective* does not mean *true.* A sentence is objective if the language used

to convey the information suggests that facts are being presented; in the context of the discourse, material is objectively presented as if it were true. Whether or not the source truly believes the information, and whether or not the information is in fact true, are considerations outside the purview of a theory of linguistic subjectivity.

An aspect of subjectivity highlighted when working with NLP applications is ambiguity. Many words with subjective usages may be used objectively. Examples are *sapping* and *eroding*. In (5), they are used subjectively, but one can easily imagine objective usages, in a scientific domain, for example. Thus, an NLP system may not merely consult a list of lexical items to accurately identify subjective language, but must disambiguate words, phrases, and sentences in context. In our terminology, a *potential subjective element* (*PSE*) is a linguistic element that may be used to express subjectivity. A *subjective element* is an instance of a potential subjective element, in a particular context, that is indeed subjective in that context (Wiebe, 1994).

Other attributes include strength and polarity of private states, and different varieties of subjective elements.

In this paper, we focus on learning lexical items that are associated with subjectivity (i.e., PSEs) and then using them in concert to disambiguate instances of them (i.e., to determine if the instances are subjective elements).

2.1 Manual Annotations

In our subjectivity annotation projects, we do not give the annotators lists of particular words and phrases to look for. Rather, we ask them to label sentences according to their interpretations in context. As a result, the annotators consider a large variety of expressions when performing annotations.

We use data that has been manually annotated at the expression level, the sentence level, and the document level. For diversity, we use data from the Wall Street Journal

Name	Source	# Words	Annotators	Type of Annotation
WSJ-SE	Wall Street Journal	18341	D,M	subjective elements
NG-SE	Newsgroup	15413	M	subjective elements
NG-FE	Newsgroup	88210	MM,R	flame elements
OP1	Wall Street Journal	640975	M,T	documents
Composed of 4 datasets: W9-4, W9-10, W9-22, W-33				
OP2	Wall Street Journal	629690	M,T	documents
Composed of 4 datasets: W9-2, W9-20, W9-21, W-23				

Table 1

Data Sets and Annotations used in Experiments. Annotators M, MM, and T are co-authors of this paper. D and R are not.

as well as data from a corpus of Usenet newsgroup messages. Table 1 summarizes the datasets and annotations used in this paper. None of the datasets overlap. The annotation types listed in the table are those used in the experiments presented in this paper.

In our first subjectivity annotation project (Wiebe, Bruce, and O'Hara, 1999; Bruce and Wiebe, 1999), a corpus of sentences from the Wall Street Journal Treebank Corpus (Marcus, Santorini, and Marcinkiewicz, 1993) (corpus WSJ-SE in Table 1) was annotated at the sentence-level by multiple judges. The judges were instructed to classify a sentence as subjective if it contains any significant expressions of subjectivity, attributed to either the writer or someone mentioned in the text, and to classify the sentence as objective, otherwise. After multiple rounds of training, the annotators independently annotated a fresh test set of 500 sentences from WSJ-SE. They achieved an average pairwise κ score of 0.70 over the entire test set; an average pairwise κ score of 0.80 for the 85% of the test set for which the annotators were somewhat sure of their judgments; and an average pairwise κ score of 0.88 for the 70% of the test set for which the annotators were very sure.

We later asked the same annotators to identify the subjective elements in WSJ-SE. Specifically, each annotator was given the subjective sentences he identified in the previ-

ous study, and asked to put brackets around the words he believed caused the sentence to be classified as subjective.[2] For example (subjective elements are in parentheses):

They paid (yet) more for (really good stuff).

(Perhaps you'll forgive me) for reposting his response.

No other instructions were given to the annotators and no training was performed for the expression-level task. A single round of tagging was performed, with no communication between annotators. There are techniques for analyzing agreement when annotations involve segment boundaries (Litman and Passonneau, 1995; Marcu, Romera, and Amorrortu, 1999), but our focus in this paper is on words. Thus, our analyses are at the word level: each word is classified as either appearing in a subjective element or not. Punctuation and numbers are excluded from the analyses. The κ value for word agreement in this study is 0.42.

Another two-level annotation project was performed in (Wiebe et al., 2001), this time involving document-level and expression-level annotations of newsgroup data (NG-FE in Table 1). In that project, we were interested in annotating *flames*, inflammatory messages in newsgroups or listservs. Note that inflammatory language is a kind of subjective language. The annotators were instructed to mark a message as a flame if the main intention of the message is a personal attack, and the message contains insulting or abusive language.

After multiple rounds of training, three annotators, MM, R, and L, independently annotated a fresh test set of 88 messages from NG-FE. The average pairwise percentage agreement is 92% and the average pairwise κ value is 0.78. These results are comparable to those of Spertus (1997), who reports 98 percent agreement on non-inflammatory messages and 64 percent agreement on inflammatory messages.

2 We are grateful to Aravind Joshi for suggesting this level of annotation.

Two of the annotators, R and MM were then asked to identify the *flame elements* in the entire corpus NG-FE. Flame elements are the subset of subjective elements that are perceived to be inflammatory. R and MM were asked to do this in the entire corpus, even those messages not identified as flames, because messages that were not judged to be flames at the document level may contain some individual inflammatory phrases. As above, no training was performed for the expression-level task, and a single round of tagging was performed, without communication between annotators. Agreement was measured in the same way as in the subjective element study above. The κ value for flame-element annotations in corpus NG-FE is 0.46.

An additional annotation project involved a single annotator: M performed subjective-element annotations on the newsgroup corpus NG-SE.

The agreement results above suggest that good agreement can be achieved at higher levels of classification (sentence and document), but agreement at the expression level is more challenging. The agreement values are lower for the expression-level annotations, but are still much higher than that expected by chance.

Note that our word-based analysis of agreement is a tough measure, because it requires that exactly the same words are identified by both annotators. Consider the following example from WSJ-SE:

> *D*: (played the role well) (obligatory ragged jeans a thicket of long hair and rejection of all things conventional)
>
> *M*: played the role (well) (obligatory) (ragged) jeans a (thicket) of long hair and (rejection) of (all things conventional)

Judge D consistently identifies entire phrases as subjective, while judge M prefers to select discrete lexical items.

Despite such differences between annotators, the expression-level annotations proved

very useful for exploring hypotheses and generating features, as described below.

Since this paper was written, a new annotation project has been completed. A 10,000 sentence corpus of English-language versions of world news articles has been annotated with detailed subjectivity information as part of a project investigating multiple perspective question answering (Wiebe et al., 2003). These annotations are much more detailed than the annotations used in this paper (including, for example, the source of each private state). The inter-annotator agreement scores for the new corpus are high, and are improvements over the results of the studies described above (Wilson and Wiebe, 2003).

The current paper uses existing document-level subjective classes, namely *Editorials, Letters to the Editor*, *Arts & Leisure Reviews*, and *Viewpoints* in the Wall Street Journal. These are subjective classes in the sense that they are text categories for which subjectivity is a key aspect. We refer to them collectively as *opinion pieces.* All other types of documents in the Wall Street Journal are collectively referred to as *non-opinion* pieces.

Note that opinion pieces are not 100% subjective. For example, editorials contain objective sentences presenting facts supporting the writer's argument, and reviews contain sentences objectively presenting facts about the product. Similarly, non-opinion pieces are not 100% objective. News reports present opinions and reactions to reported events (van Dijk 1988); they often contain segments starting with expressions such as *critics claim* and *supporters argue.* In addition, quoted-speech sentences in which individuals express their subjectivity are often included (Barzilay et al., 2000). For concreteness, let us consider WSJ-SE, which, recall, has been manually annotated at the sentence level. In WSJ-SE, 70% of the sentences in opinion pieces are subjective and 30% are objective. In non-opinion pieces, 44% of the sentences are subjective and only 56% are objective. Thus, while there is a higher concentration of subjective sentences in opinion versus non-opinion pieces, there are many subjective sentences in non-opinion pieces and objective

sentences in opinion pieces.

An inspection of some data revealed that some editorial and review articles are not marked as such by the Wall Street Journal. For example, there are articles whose purpose is to present an argument rather than cover a news story, but they are not explicitly labeled as editorials by the Wall Street Journal. Thus, the opinion-piece annotations of OP1 and OP2 have been manually refined. The annotation instructions are simply to identify any additional opinion pieces that are not marked as such. To test the reliability of this annotation, two judges independently annotated two Wall Street Journal files, W9-22 and W9-33, each approximately 160K words. This is an "annotation lite" task: with no training, the annotators achieved κ values of 0.94 and 0.95, and each spent an average of three hours per WSJ file.

3 Generating and Testing Subjective Features

3.1 Introduction

The goal in this section is to learn lexical subjectivity clues of various types, both single words as well as collocations. Some require no training data, some are learned using the expression-level subjective-element annotations as training data and some are learned using the document-level opinion-piece annotations as training data (i.e., *opinion piece* versus *non-opinion piece*). All of the clues are evaluated with respect to the document-level opinion-piece annotations. While these evaluations are our focus, because much more opinion-piece than subjective-element data exists, we do evaluate the clues learned from the opinion-piece data on the subjective-element data as well. Thus, we cross validate the results both ways between the two types of annotations.

Throughout this section, we evaluate sets of clues directly, by measuring the proportion of clues that appear in subjective documents or expressions, seeking those that appear more often than expected. In later sections, the clues are used together to find

subjective sentences and to perform text categorization.

The following paragraphs give details of the evaluation and experimental design used in this section.

The proportion of clues in subjective documents or expressions is their precision. Specifically:

The precision of a set S with respect to opinion pieces is:

$$prec(S) = \frac{\text{number of instances of members of S in opinion pieces}}{\text{total number of instances of members of S in the data}}$$

The precision of a set S with respect to subjective elements is:

$$prec(S) = \frac{\text{number of instances of members of S in subjective elements}}{\text{total number of instances of members of S in the data}}$$

In the above, S is a set of types (not tokens). The counts are of tokens (i.e., instances or occurrences) of members of S.

Why is S a set? Many good clues of subjectivity are low frequency (Wiebe, McKeever, and Bruce, 1998). In fact, as we shall see below, uniqueness in the corpus is an informative feature for subjectivity classification. Thus, we do not want to discard low-frequency clues, because they are a valuable source of information, and we do not want to evaluate individual low-frequency lexical items, because the results would be unreliable. Our strategy is thus to identify and evaluate sets of words and phrases, rather than individual items.

What kinds of results may we expect? We cannot expect absolutely high precision with respect to the opinion-piece classifications, even for strong clues, for three reasons. First, for our purposes, the data is noisy. As mentioned above, while the proportion of subjective sentences is higher in opinion versus non-opinion pieces, the proportions are not 100 and 0: opinion pieces contain objective sentences and non-opinion pieces contain subjective sentences.

Second, we are trying to learn lexical items associated with subjectivity, i.e., potential subjective elements (PSEs). As discussed above, many words and phrases with subjective usages have objective usages as well. Thus, even in perfect data with no noise, we would not expect 100% precision. (This is the motivation for the work on density presented below in section 4.4.)

Third, the distribution of opinions and non-opinions is highly skewed in favor of non-opinions: only 9% of the articles in the combination of OP1 and OP2 are opinion pieces.

In this work, increases in precision over a baseline precision are used as evidence that promising sets of PSEs have been found. Our main baseline for comparison is the number of word instances in opinion pieces, divided by the total number of word instances:

$$Baseline\ Precision = \frac{\text{number of word instances in opinion pieces}}{\text{total number of word instances}}$$

Words and phrases with higher proportions than this appear more than expected in opinion pieces.

To further evaluate the quality of a set of PSEs, we also perform the following significance test. For a set of PSEs in a given dataset, we tested the significance of the difference between (1) the proportion of words in opinion pieces that are PSEs and (2) the proportion of words in non-opinion pieces that are PSEs, using the z-significance test for two proportions.

Before continuing, there are a few more technical items to cover about the data preparation and experimental design:

- All of the datasets are stemmed using Karp's morphological analyzer (1992) and part-of-speech tagged using Brill's tagger (1992).
- When the opinion-piece classifications are used for training, the existing classifications, assigned by the Wall Street Journal, are used. Thus, the

	WSJ-SE		
		D	M
	freq	+prec	+prec
unique words	2615	+.07	+.12
baseline	18341	.07	.08

Table 2

Frequencies and increases in precision of unique words in subjective-element data. Baseline frequency is the total number of words, and baseline precision is the proportion of words in subjective elements.

processes using them as training data may be applied to more data to learn more clues, without requiring additional manual annotation.

- When the opinion-piece data is used for testing, the manually refined classifications (described at the end of section 2.1) are used.

- OP1 and OP2 together consist of eight Treebank files. Below, we often give results separately for the component files, allowing us to assess the consistency of results for the various types of clues.

3.2 Unique Words

In this section, we show that low-frequency words are associated with subjectivity in both the subjective-element and opinion-piece data. Apparently, people are creative when they are being opinionated.

Table 2 gives results for unique words in subjective-element data. Recall that *unique words* are those that appear just once in the corpus, i.e., *hapax legomena*. The first row of Table 2 gives the frequency of unique words in WSJ-SE, followed by the percentage-point improvements in precision over baseline for unique words in subjective elements marked by annotators D and M, respectively. The second row gives baseline frequency and precisions. Baseline frequency is the total number of words in WSJ-SE. Baseline precision for an annotator is the proportion of words included in subjective elements

	W9-04		W9-10		W9-22		W9-33	
	freq	+prec	freq	+prec	freq	+prec	freq	+prec
unique words	4794	+.15	4763	+.16	4274	+.11	4567	+.11
baseline	156421	.19	156334	.18	155135	.13	153634	.14

Table 3

Frequencies and increases in precision for words that appear exactly once in the datasets composing OP1. For each dataset, baseline frequency is the total number of words, and baseline precision is the proportion of words in opinion pieces.

by that annotator. Specifically, consider annotator M. The baseline precision of words in subjective elements marked by M is 0.08, but the precision of unique words in these same annotations is 0.20, .12 points higher than the baseline. This is a 150% improvement over the baseline.

The number of unique words in opinion pieces is also higher than expected. Table 3 compares the precision of the set of unique words to the baseline precision (i.e., the precision of the set of all words that appear in the corpus) in the four WSJ files composing OP1. Before this analysis was performed, numbers were removed from the data (we are not interested in the fact that, say, the number 163,213.01 appears just once in the corpus). The number of words in each dataset and baseline precisions are listed at the bottom of the table. The *freq* columns give total frequencies. The *+prec* columns show the percentage-point improvements in precision over baseline. For example, in W9-10, unique words have precision 0.34: 0.18 baseline plus an improvement over baseline of 0.16. The difference in the proportion of words that are unique in opinion pieces and the proportion of words are are unique in non-opinion pieces is highly significant with $p < 0.001$ ($z \geq 22$) for all of the datasets. Note that, not only does the set of unique words have higher than baseline precision, the set is a frequent feature!

The question arises, how does corpus size affect the precision of the set of unique words? Presumably, uniqueness in a larger corpus is more meaningful than uniqueness in

a smaller one. The results in Figure 1 provide evidence that it is. The Y axis in Figure 1 represents increase in precision over baseline and the X axis represents corpus size. Five graphs are plotted, one for the set of words that appear exactly once, one for the set of words that appear exactly twice ($freq2$), one for the set of words that appear exactly three times ($freq3$), etc.

In Figure 1, increases in precision are given for corpora of size n, where $n = 20, 40, \ldots, 2420, 2440$ documents. Each data point is an average over 25 sample corpora of size n. The sample corpora were chosen from the concatenation of OP1 and OP2, in which 9% of the documents are opinion pieces. The sample corpora are created by randomly selecting documents from the large corpus, preserving the 9% distribution of opinion pieces. At the smallest corpus size (containing 20 documents), the average number of words is 9,617. At the largest corpus size (containing 2440 documents), the average is 1,225,186 words.

As can be seen, the precision of unique and other low-frequency words increases with corpus size, with increases tapering off at the largest corpus size tested. Although not as dramatic, words with frequency 2 also realize a nice increase in precision over baseline. Even words of frequency 3, 4, and 5 show modest increases.

To help us understand low-frequency words in large as opposed to small datasets, we can consider the following analogy. With collectible trading cards, rare cards are the most valuable. However, looking in only a few packs of cards will not tell us if any of our cards are valuable. It is only by looking at many packs of cards that we realize which are the rare ones. It is only in samples of sufficient size that uniqueness is informative.

The results in this section suggest that, an NLP system using uniqueness features to recognize subjectivity should determine uniqueness with respect to the test data augmented with an additional store of (unannotated) data.

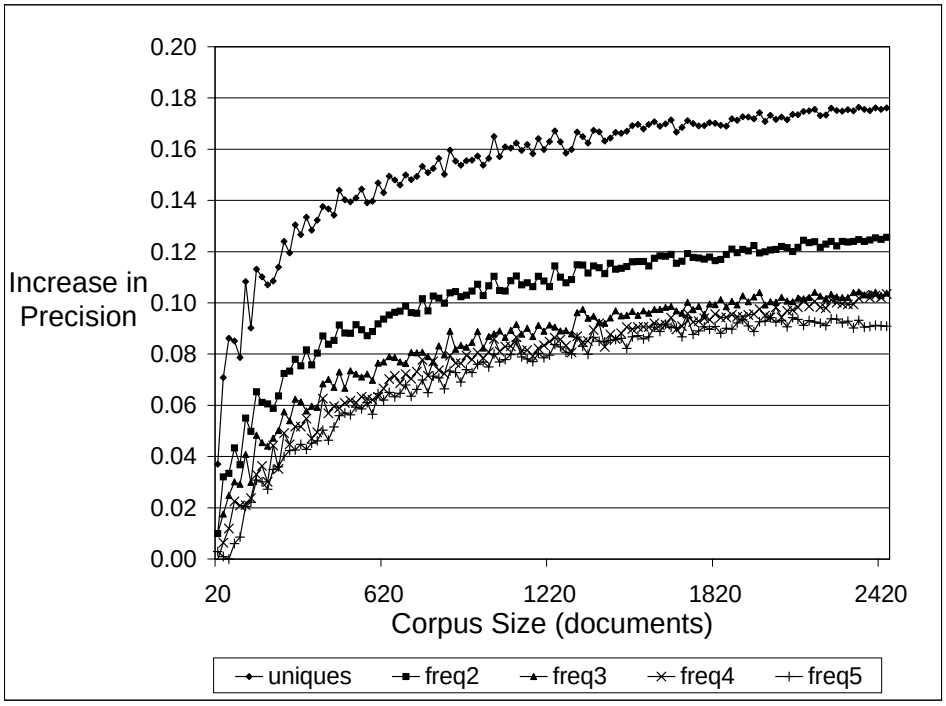

Figure 1

Precision of low-frequency words as corpus size increases

3.3 Identifying potentially subjective collocations from subjective-element (SE) and flame-element (FE) annotations

In this section, we describe experiments in identifying potentially subjective collocations.

Collocations are selected from the subjective element data (i.e., NG-SE, NG-FE, and WSJ-SE), using the union of the annotators' tags for the datasets tagged by multiple taggers. The results are then evaluated on opinion piece data.

The selection procedure is as follows. First, all 1-grams, 2-grams, 3-grams, and 4-grams are extracted from the data. In this work, each constituent of an n-gram is a word-stem, part-of-speech pair. For example, (in-*prep* the-*det* can-*noun*) is a 3-gram that matches trigrams consisting of preposition *in*, followed by determiner *the*, and ending with noun *can*.

A subset of the n-grams are then selected based on precision. The precision of an n-gram is the number of subjective instances of that n-gram divided by the total number of instances of that n-gram. An instance of an n-gram is subjective if each word occurs in a subjective element.

N-grams are selected based on two criteria. First, the precision of the n-gram must be greater than the baseline precision (i.e., the proportion of all word instances that are in subjective elements). Second, the precision of the n-gram must be greater than the maximum precision of its constituents. This criterion is used to avoid selecting unnecessarily long collocations. For example, *scumbag* is a strongly subjective clue. If *be a scumbag* does not have higher precision than *scumbag* alone, we do not want to select it.

Specifically, let $(W1, W2)$ be a bigram consisting of consecutive words $W1$ and $W2$. $(W1,W2)$ is identified to be a potential subjective element if $prec(W1, W2) >= 0.1$ and:

$$prec(W1, W2) > max(prec(W1), prec(W2))$$

For trigrams, we extend the second condition as follows. Let $(W1, W2, W3)$ be a trigram

consisting of consecutive words $W1$, $W2$, and $W3$. The condition is then:

$$prec(W1, W2, W3) > max(prec(W1, W2), prec(W3)) \text{ or}$$

$$prec(W1, W2, W3) > max(prec(W1), prec(W2, W3))$$

The selection of 4-grams is similar to the selection of 3-grams, comparing the 4-gram first with the maximum of the precisions of word $W1$ and trigram ($W2$, $W3$, $W4$) and then with the maximum of the precisions of trigram ($W1$,$W2$,$W3$) and word $W4$. We call the n-gram collocations identified as above *fixed-n-grams*.

We also define a type of collocation called a *unique generalized n-gram (ugen-n-gram)*. Such collocations have placeholders for unique words. As will be seen below, these are our highest precision features.

To find and select such generalized collocations, we first find every word that appears just once in the corpus and replace it with a new word, 'UNIQUE' (but remembering the part of speech of the original word). In essence, we treat the set of single-instance words as a single, frequently-occurring word (which occurs with various parts of speech). Precisely the same method for extracting and selecting n-grams above is used to obtain the potentially subjective collocations with 1 or more positions filled by a 'UNIQUE', part-of-speech pair.

To test the ugen-n-grams extracted from the subjective-element training data using the method outlined above, we assess their precision with respect to opinion piece data. As with the training data, all unique words in the test data are replaced by 'UNIQUE'. When a ugen-n-gram is matched against the test data, the 'UNIQUE' fillers match words (of the appropriate parts of speech) that are unique in the test data.

Table 4 shows the results of testing the fixed-n-gram and the ugen-n-gram patterns identified as described above on the four data sets composing OP1. The *freq* columns give total frequencies, and the *+prec* columns show the improvements in precision from the

	W9-04		W9-10		W9-22		W9-33	
	freq	+prec	freq	+prec	freq	+prec	freq	+prec
fixed-2-grams	1840	+.07	1972	+.07	1933	+.04	1839	+.05
ugen-2-grams	281	+.21	256	+.26	261	+.17	254	+.17
fixed-3-grams	213	+.08	243	+.09	214	+.05	238	+.05
ugen-3-grams	148	+.29	133	+.27	147	+.16	133	+.15
fixed-4-grams	18	+.15	17	+.06	12	+.29	14	-.07
ugen-4-grams	13	+.12	3	+.82	15	+.27	13	+.25
baseline	156421	.19	156334	.18	155135	.13	153634	.14

Table 4

Frequencies and increases in precision of fixed-n-gram and ugen-n-gram collocations learned from the subjective-element data. For each dataset, baseline frequency is the total number of words, and baseline precision is the proportion of words in opinion pieces.

one-*noun* of-*prep* his-*det*	worst-*adj* of-*prep* all-*det*
quality-*noun* of-*prep* the-*det*	to-*prep* do-*verb* so-*adverb*
in-*prep* the-*det* company-*noun*	you-*pronoun* and-*conj* your-*pronoun*
have-*verb* taken-*verb* the-*det*	rest-*noun* of-*prep* us-*pronoun*
are-*verb* at-*prep* least-*adj*	but-*conj* if-*prep* you-*pronoun*
as-*prep* a-*det* weapon-*noun*	continue-*verb* to-*to* do-*verb*
purpose-*noun* of-*prep* the-*det*	could-*modal* have-*verb* be-*verb*
it-*pronoun* seem-*verb* to-*prep*	to-*pronoun* continue-*verb* to-*prep*
have-*verb* be-*verb* the-*det*	do-*verb* something-*noun* about-*prep*
cause-*verb* you-*pronoun* to-*to*	evidence-*noun* to-*to* back-*adverb*
that-*prep* you-*pronoun* are-*verb*	i-*pronoun* be-*verb* not-*adverb*
of-*prep* the-*det* century-*noun*	of-*prep* money-*noun* be-*prep*

Table 5
Random sample of Fixed-3-gram collocations in OP1

baseline. The number of words in each dataset and baseline precisions are given at the bottom of the table. For all n-gram features besides the fixed-4-grams and ugen-4-grams, the proportion of features in opinion pieces is significantly greater than the proportion of features in non-opinion pieces.[3]

The question arises, how much overlap is there between instances of fixed-n-grams and instances of ugen-n-grams? In the test data of Table 4, there are a total of 8,577 fixed-n-grams instances. Only 59 of these, fewer than 1%, are contained (wholly or in part) in ugen-n-gram instances. This small intersection set shows that two different types of potentially subjective collocations are being recognized.

Randomly selected examples of our learned collocations that appear in the test data are given in Tables 5 and 6. It is interesting to note that the ugen collocations were learned from the training data by matching different unique words from the ones they match in the test data.

3 Specifically, the difference between (1) the number of feature instances in opinion pieces divided by the number of words in opinion pieces and (2) the number of feature instances in non-opinion pieces divided by the number of words in non-opinion pieces is significant ($p < 0.05$) for all datasets.

Pattern	Instances
U-*adj* as-*prep:*	drastic as; perverse as; predatory as
U-*adj* in-*prep:*	perk in; unsatisfying in; unwise in
U-*adverb* U-*verb:*	adroitly dodge; crossly butter; unceasingly fascinate
U-*noun* back-*adverb:*	cutting back; hearken back
U-*verb* U-*adverb:*	coexist harmoniously; flouncing tiresomely
ad-*noun* U-*noun:*	ad hoc; ad valorem
any-*det* U-*noun:*	any over-payment; any tapings; any write-off
are-*verb* U-*noun:*	are escapist; are lowbrow; are resonance
but-*conj* U-*noun:*	but belch; but cirrus; but ssa
different-*adj* U-*noun:*	different ambience; different subconferences
like-*prep* U-*noun:*	like hoffmann; like manute; like woodchuck
national-*adj* U-*noun:*	national commonplace; national yonhap
particularly-*adverb* U-*adj:*	particularly galling; particularly noteworthy
so-*adverb* U-*adj:*	so monochromatic; so overbroad; so permissive
this-*det* U-*adj:*	this biennial; this inexcusable; this scurrilous
your-*pronoun* U-*noun:*	your forehead; your manuscript; your popcorn
U-*adj* and-*conj* U-*adj:*	arduous and raucous; obstreperous and abstemious
U-*noun* be-*verb* a-*det:*	acyclovir be a; siberia be a
U-*noun* of-*prep* its-*pronoun:*	outgrowth of its; repulsion of its
U-*verb* and-*conj* U-*verb:*	wax and brushed; womanize and booze
U-*verb* to-*to* a-*det:*	cling to a; trek to a
are-*verb* U-*adj* to-*to:*	are opaque to; are subject to
a-*det* U-*noun* and-*conj:*	a blindfold and; a rhododendron and
a-*det* U-*verb* U-*noun:*	a jaundice ipo; a smoulder sofa
it-*pronoun* be-*verb* U-*adverb:*	it be humanly; it be sooo
than-*prep* a-*det* U-*noun:*	than a boob; than a menace
the-*det* U-*adj* and-*conj:*	the convoluted and; the secretive and
the-*det* U-*noun* that-*prep:*	the baloney that; the cachet that
to-*to* a-*det* U-*adj:*	to a gory; to a trappist
to-*to* their-*pronoun* U-*noun:*	to their arsenal; to their subsistence
with-*prep* an-*det* U-*noun:*	with an alias; with an avalanche

Table 6
Random sample of Ugen Collocations in OP1. 'U' means unique.

3.4 Generating features from Document-Level Annotations Using Distributional Similarity

In this section, we identify adjective and verb PSEs using *distributional similarity*. Opinion-piece data is used for training, and (a different set of) opinion-piece data and the subjective-element data are used for testing.

With *distributional similarity*, words are judged to be more or less similar based on their distributional patterning in text (Lee, 1999; Lee and Pereira, 1999). Our motivation for experimenting with it to identify PSEs was two-fold. First, we hypothesized that words might be distributionally similar because they share pragmatic usages, such as expressing subjectivity, even if they are not close synonyms. Second, as shown above, low-frequency words appear more often in subjective texts than expected. We did not want to discard all low-frequency words from consideration, but cannot effectively judge the suitability of individual words. Thus, to decide whether to retain a word as a PSE, we consider the precision not of the individual word, but of the word together with its cluster of similar words.

Many variants of distributional similarity have been used in NLP (Lee, 1999; Lee and Pereira, 1999). Dekang Lin's (1998) method is used here. In contrast to many implementations, which focus exclusively on verb-noun relationships, Lin's method incorporates a variety of syntactic relations. This is important for subjectivity recognition, because PSEs are not limited to verb-noun relationships. In addition, Lin's results are freely available.

A set of seed words begins the process. For each seed s_i, the precision of the set $\{s_i\} \cup C_{i,n}$ in the training data is calculated, where $C_{i,n}$ is the set of n words most similar to s_i, according to Lin's method (1998). If the precision of $\{s_i\} \cup C_{i,n}$ is greater than a threshold T, then the words in this set are retained as PSEs. If it is not, neither s_i nor the words in $C_{i,n}$ are retained. The union of the retained sets will be notated $R_{T,n}$,

$trainingPrec(s)$ is the precision of s in the training data
$validationPrec(s)$ is the precision of s in the validation data
$testPrec(s)$ is the precision of s in the test data
(similarly for $trainingFreq$,$validationFreq$, and $testFreq$)
S = the set of all adjectives (verbs) in the training data
for T in [0.01,0.04,...,0.70]:
    for n in [2,3,...,40]:
        $retained = \{\}$
        For s_i in S:
            if $trainingPrec(\{s_i\} \cup C_{i,n}) > T$:
                $retained = retained \cup \{s_i\} \cup C_{i,n}$
        $R_{T,n} = retained$
$ADJ_{pses} = \{\}$ ($VERB_{pses} = \{\}$)
for T in [0.01,0.04,...,0.70]:
    for n in [2,3,...,40]:
        if $validationPrec(R_{T,n}) \geq 0.28$ (0.23 for verbs)
        and $validationFreq(R_{T,n}) \geq 100$:
            $ADJ_{pses} = ADJ_{pses} \cup R_{T,n}$ ($VERB_{pses} = VERB_{pses} \cup R_{T,n}$)
Results in Table 7 show $testPrec(ADJ_{pses})$ and $testFreq(ADJ_{pses})$.

Figure 2

Algorithm for selecting adjective and verb features using distributional similarity.

that is, the union of all sets $\{s_i\} \cup C_{i,n}$ with precision on the training set $> T$.

In (Wiebe, 2000), the seeds (the s_i's) were extracted from the subjective-element annotations in corpus WSJ-SE. Specifically, the seeds were the adjectives that appear at least once in a subjective element in WSJ-SE. In this paper, the opinion-piece corpus is used to move beyond the manual annotations and small corpus of the earlier work, and a much looser criterion is used to choose the initial seeds: all of the adjectives (verbs) in the training data are used.

The algorithm for the process is given in Figure 2. There is one small difference for adjectives and verbs noted in the figure, i.e., the precision thresholds of 0.28 versus 0.23. These thresholds were determined using validation data.

Seeds and their clusters are assessed on a training set for many parameter settings (cluster size n from 2 through 40, and precision threshold T from 0.01 through 0.70 by 3). As mentioned above, each n, T parameter pair yields a set of adjectives $R_{T,n}$, that

Training	Validation	Test	Baseline freq	Baseline prec	ADJ_{pses} freq	ADJ_{pses} +prec	$VERB_{pses}$ freq	$VERB_{pses}$ +prec
W9-10	W9-22							
W9-22	W9-10	W9-33	153634	.14	1576	+.12	1490	+.11
W9-10	W9-33							
W9-33	W9-10	W9-22	155135	.13	859	+.15	535	+.11
W9-22	W9-33							
W9-33	W9-22	W9-10	156334	.18	249	+.22	224	+.10
All pairings of W9-10, W9-22,W9-33		W9-4	156421	.19	1872	+.17	1777	+.15

Table 7

Frequencies and increases in precision for adjective and verb features identified using distributional similarity with filtering. For each test dataset, baseline frequency is the total number of words, and baseline precision is the proportion of words in opinion pieces.

is, the union of all sets $\{s_i\} \cup C_{i,n}$ with precision on the training set $> T$. A subset, ADJ_{pses}, of those sets is chosen based on precision and frequency in a validation set. Finally, the ADJ_{pses} are tested on the test set.

Table 7 shows the results for four opinion-piece test sets. Multiple training-validation dataset pairs are used for each test set, as given in Table 7. The results are for the union of the adjectives (verbs) chosen for each pair. The *freq* columns give total frequencies, and the *+prec* columns show the improvements in precision from the baseline. For each dataset, the difference in the proportion of instances of ADJ_{pses} in opinion pieces and the proportion in non-opinion pieces is significant ($p < 0.001$, $z \geq 9.2$). The same is true for $VERB_{pses}$ ($p < 0.001$, $z \geq 4.1$).

In the interests of testing consistency, Table 8 shows the results of assessing the adjective and verb features generated from opinion-piece data (ADJ_{pses} and $VERB_{pses}$ in Table 7) on the subjective-element data. The left side of the table gives baseline figures for each set of subjective-element annotations. The right side of the table gives the average frequencies and increases in precision over baseline, for the ADJ_{pses} and $VERB_{pses}$ sets on the subjective-element data. The baseline figures in the table are the

	Adj Baseline		Verb Baseline		ADJ_{pses}		$VERB_{pses}$	
	freq	prec	freq	prec	freq	+prec	freq	+prec
WSJ-SE-D	1632	.13	2980	.15	136	+.16	151	+.10
WSJ-SE-M	1632	.19	2980	.12	136	+.24	151	+.13
NG-SE	1104	.37	2629	.15	185	+.25	275	+.08

Table 8

Average frequencies and increases in precision in subjective-element data of the sets tested in Table 7. The baselines are the precisions of adjectives/verbs that appear in subjective elements in the subjective-element data.

frequencies and precisions of the sets of adjectives and verbs that appear at least once in a subjective element. Since these sets include words that appear just once in the corpus (and thus have 100% precision), the baseline precision is a challenging one.

Testing the $VERB_{pses}$ and ADJ_{pses} on the subjective-element reveals some interesting consistencies for these subjectivity clues. Precision increases of the $VERB_{pses}$ on the subjective-element data are comparable to their increases on the opinion-piece data. Similarly, the precision increases of the ADJ_{pses} on the subjective-element data is as good or better than the performance of this set of PSEs on the opinion-piece data. Finally, precisions increases for the ADJ_{pses} are higher than for the $VERB_{pses}$ on all datasets. This is again consistent with the higher performance of the ADJ_{pses} sets in the opinion-piece datasets.

4 Features Used in Concert

4.1 Introduction

In this section, we examine the various types of clues used together. In preparation for this work, all instances in OP1 and OP2 of all of the PSEs identified as described in Section 3 have been automatically identified. All training to define the PSE instances in OP1 was performed on data separate from OP1, and all training to define the PSE instances in OP2 was performed on data separate from OP2.

	W9-04		W9-10		W9-22		W9-33	
	freq	+prec	freq	+prec	freq	+prec	freq	+prec
unique words	4794	+.15	4763	+.16	4274	+.11	4567	+.11
fixed-2-grams	1840	+.07	1972	+.07	1933	+.04	1839	+.05
ugen-2-grams	281	+.21	256	+.26	261	+.17	254	+.17
fixed-3-grams	213	+.08	243	+.09	214	+.05	238	+.05
ugen-3-grams	148	+.29	133	+.27	147	+.16	133	+.15
fixed-4-grams	18	+.15	17	+.06	12	+.29	14	-.07
ugen-4-grams	13	+.12	3	+.82	15	+.27	13	+.25
adjectives	1872	+.17	249	+.22	859	+.15	1576	+.12
verbs	1777	+.15	224	+.10	535	+.11	1490	+.11
baseline	156421	.19	156334	.18	155135	.13	153634	.14

freq: total frequency. **+prec:** increase in precision over baseline.

Table 9

Frequencies and Increases in Precision for All Features. For each dataset, baseline frequency is the total number of words, and baseline precision is the proportion of words in opinion pieces.

4.2 Consistency in Precision Among Datasets

Table 9 summarizes the results from previous sections in which the opinion-piece data is used for testing. The performance of the various features are consistently good or bad on the same datasets: the performance is better for all features on W9-10 and W9-04 than on W9-22 and W9-33 (except for the ugen-4-grams, which are very low frequency, and the verbs, which are low on W9-10). This is so despite the fact that the features were generated using different procedures and data: the adjectives and verbs were generated from WSJ document-level opinion-piece classifications; the n-gram features were generated from newsgroup and WSJ expressions-level subjective-element classifications; and the unique unigram feature requires no training. This consistency in performance suggests that the results are not brittle.

4.3 Choosing Density Parameters from Subjective Element Data

In (Wiebe, 1994), whether a PSE is interpreted to be subjective depends, in part, on how subjective the surrounding context is. We explore this idea in the current work, assessing whether PSEs are more likely to be subjective if they are surrounded by subjective

0. $PSEs$ = all adjs, verbs, modals, nouns, and adverbs that appear at least once in an SE (except *not*, *will*, *be*, *have*).
1. $PSEinsts$ = the set of all instances of $PSEs$
2. $HiDensity = \{\}$
3. For P in $PSEinsts$:
    4. leftWin(P) = the W words before P
    5. rightWin(P) = the W words after P
    6. density(P) = # of SEs whose first or last word is in leftWin(P) or rightWin(P)
    7. if density(P) $\geq T$:
        $HiDensity = HiDensity \cup \{P\}$

8. $prec(PSEinsts) = \frac{\text{\# of } PSEinsts \text{ in } SEs}{|PSEinsts|}$

9. $prec(HiDensity) = \frac{\text{\# of } HiDensity \text{ in } SEs}{|HiDensity|}$

Figure 3

Algorithm for calculating density in subjective element (SE) data

elements. In particular, we experiment with a density feature to decide whether or not a PSE instance is subjective: if a sufficient number of subjective elements are nearby, then the PSE instance is considered to be subjective; otherwise, it is discarded. The density parameters are a window size W and a frequency threshold T.

In this section, we explore the density of manually-annotated PSEs in subjective-element data, and choose density parameters to use later in Section 4.4, where we apply them to automatically identified PSEs in opinion-piece data.

The process for calculating density in the subjective-element data is given in Figure 3. The PSEs are defined to be all adjectives, verbs, modals, nouns, and adverbs that appear at least once in a subjective element, with the exception of some stop words (line 0 of Figure 3). Note that these PSEs depend only on the subjective-element manual annotations, not on the automatically identified features used elsewhere in the paper, nor on the document-level opinion-piece classes. $PSEinsts$ is the set of PSE instances to be disambiguated (line 1). $HiDensity$ (initialized on line 2) will be the subset of $PSEinsts$

that are retained. In the loop, the density of each PSE instance P is calculated, which is the number of subjective elements that begin or end in the W words preceding or following P (line 6). P is retained if its density is at least T (line 7).

Lines 8-9 assess the precision of the original ($PSEinsts$) and new ($HiDensity$) sets of PSE instances. If $prec(HiDensity)$ is greater than $prec(PSEinsts)$, then there is evidence that the number of subjective elements near a PSE instance is related to its subjectivity in context.

To create more data points for this analysis, WSJ-SE was split into two (WSJ-SE1 and WSJ-SE2) and M and D's annotations are considered separately. WSJ-SE2-D, for example, refers to D's annotations of WSJ-SE2. The process in Figure 3 was repeated for different parameter settings (T in $[1, 2, 4, \ldots, 48]$ and W in $[1, 10, 20, \ldots, 490]$) on each of the SE datasets. To find good parameter settings, the results for each dataset were sorted into 5-point precision intervals, and then sorted by frequency within each interval. Information for the top three precision intervals for each dataset are shown in Table 10, specifically the parameter values (i.e., T and W) and the frequency and precision of the most frequent result in each interval. The intervals are in the rows labeled "Range". For example, the top three precision intervals for WSJ-SE1-M, 0.77-0.82, 0.82-0.87, and 0.87-0.92 (no parameter values yield higher precision than 0.92).

The top of Table 10 gives baseline frequencies and precisions, which are $|PSEinsts|$ and $prec(PSEinsts)$, respectively, in line 8 of Figure 3.

The parameter values exhibit a range of frequencies and precisions, with the expected tradeoff between precision and frequency. We choose the following parameters to test in Section 4.4 below: for each dataset, for each precision interval whose lower bound is at least 10 percentage points higher than the baseline for that dataset, the top two T,W pairs yielding the highest frequencies in that interval are chosen. Among the five datasets, a total of 45 parameter pairs were so selected.

	WSJ-SE1-M	WSJ-SE1-D	WSJ-SE2-M	WSJ-SE2-D	NG-SE
Baseline freq	1566	1245	1167	1108	3303
Baseline prec	.49	.47	.41	.36	.51
Range	.87-.92	.95-1.0	.95-1.0	.95-1.0	.95-1.0
T,W	10,20	12,50	20,50	14,100	10,10
freq	76	12	1	1	3
prec	.89	1.0	1.0	1.0	1.0
Range	.82-.87	.90-.95	.73-.78	.51-.56	.67-.72
T,W	6,10	12,60	46,190	22,370	26,90
freq	63	22	53	221	664
prec	.84	.91	.78	.51	.67
Range	.77-.82	.84-.89	.66-.71	.46-.51	.63-.67
T,W	12,40	12,80	18,60	16,310	8,30
freq	292	42	53	358	1504
prec	.78	.88	.68	.47	.63

Table 10

Most frequent entry in the top 3 precision intervals for each subjective element dataset

0. $PSEinsts$ = the set of instances in the test data of all PSEs described in Section 3
1. $HiDensity = \{\}$
2. For P in $PSEinsts$:
   3. leftWin(P) = the W words before P
   4. rightWin(P) = the W words after P
   5. density(P) = # of $PSEinsts$ whose first or last word is in leftWin(P) or rightWin(P)
   6. if density(P) $\geq$ T:
      $HiDensity = HiDensity \cup \{P\}$

7. $prec(PSEinsts) = \frac{\text{\# of } PSEinsts \text{ in } OPs}{|PSEinsts|}$

8. $prec(HiDensity) = \frac{\text{\# of } HiDensity \text{ in } OPs}{|HiDensity|}$

Figure 4

Algorithm for calculating density in opinion piece (OP) data

This exercise was done once, without experimenting with different parameter settings.

4.4 Density for Disambiguation

In this section, density is exploited to find subjective instances of automatically identified PSEs. The process is shown in Figure 4. There are only two differences between the algorithms in Figures 3 and 4. First, in Figure 3, density is defined in terms of the number of subjective elements nearby. However, subjective-element annotations are not available in test data. In Figure 4, density is defined in terms of the number of other PSE instances nearby, where $PSEinsts$ consists of all instances of the automatically identified PSEs described in Section 3 and for which results are given in Table 9.

Second, in Figure 4, we assess precision with respect to the document-level classes (lines 7-8).

The test data is OP1.

An interesting question arose when defining the PSE instances: what should be done

	WSJ-SE1-M	WSJ-SE1-D	WSJ-SE2-M	WSJ-SE2-D	NG-SE
T,W	10,20	12,50	20,50	14,100	10,10
freq	237	3176	170	10510	8
prec	.87	.72	.97	.57	1.0
T,W	6,10	12,60	46,190	22,370	26,90
freq	459	5289	1323	21916	787
prec	.68	.68	.95	.37	.92
T,W	12,40	12,80	18,60	16,310	8,30
freq	1398	9662	906	24454	3239
prec	.79	.58	.87	.34	.67
PSE Baseline: Freq=30938, Prec=.28					

Table 11

Results for high-density PSEs in test data $OP1$ using parameters chosen from subjective-element data

with words that are identified to be PSEs (or parts of PSEs) according to multiple criteria? For example, *sunny*, *radiant*, and *exhilarating* are all unique in corpus OP1, and are all members of the adjective PSE feature defined for testing on OP1. Collocations add additional complexity. For example, consider the sequence *and splendidly*, which appears in the test data. The sequence *and splendidly* matches the ugen-2-gram (and-*conj* U-*adj*), and the word *splendidly* is unique. In addition, more than one n-gram feature may be matched by a sequence. For example, *is it that* matches three fixed-n-gram features: *is it*, *is it that*, and *it that*.

In the current experiments, the more PSEs a word matches, the more weight it is given. The hypothesis behind this treatment is that additional matches represent additional evidence that a PSE instance is subjective. This hypothesis is realized as follows: each match of each member of each type of PSE is considered to be a PSE instance. Thus, among them, there are 11 members in $PSEinsts$ for the 5 phrases *sunny*, *radiant*, *exhilarating*, *and splendidly*, and *is it that*, one for each of the matches mentioned above.

The process in Figure 4 was performed with the 45 parameter-pair values (T and W) chosen from the subjective-element data as described in Section 4.3. Table 11 shows results for a subset of the 45 parameters, namely the most frequent parameter pair

chosen from the top three precision intervals for each training set. The bottom of the table gives a baseline frequency and a baseline precision in OP1, defined as $|PSEinsts|$ and $prec(PSEinsts)$, respectively, on line 7 of Figure 4.

The density features result in substantial increases in precision. Of the 45 parameter pairs, the minimum percentage increase over baseline is 22%. Fully 24% of the 45 parameters pairs yield increases of 200% or more; 38% yield increases between 100% and 199%, and 38% yield increases between 22%-99%. In addition, the increases are significant. Using the set of high-density PSEs defined by the parameter pair with the least increase over baseline, we tested the difference in the proportion of PSEs in opinion pieces that are high-density and the proportion of PSEs in non-opinion pieces that are high-density. The difference between these two proportions is highly significant ($z = 46.2$, $p < 0.0001$).

Notice that, except for one blip ($T, W = 6, 10$ under WSJ-SE-M), the precisions decrease and the frequencies increase as we go down each column in Table 11. The same pattern can be observed with all 45 parameter pairs (not included due to space). But the parameter pairs are ordered in Table 11 based on performance in the manually-annotated subjective-element data, not based on performance in the test data! For example, the entry in the first row, first column ($T, W = 10, 20$) is the parameter pair giving the highest frequency in the top precision interval of WSJ-SE-M (frequency and precision in WSJ-SE-M, using the process of Figure 3). Thus, the relative precisions and frequencies of the parameter pairs are carried over from the training to the test data. This is quite a strong result, given that the PSEs in the training data are from manual annotations, while the PSES in the test data are our automatically identified features.

4.5 High-Density Sentence Annotations

To assess the subjectivity of sentences with high-density PSEs, we extracted the sentences in corpus OP2 that contain at least one high-density PSE, and manually annotated them.

There are 133 such sentences. We refer to them as the *system-identified* sentences.

We chose the density parameter pair T,W=12,30, based on its precision and frequency in OP1. This parameter setting yields results that are relatively high precision and low frequency. We chose a low-frequency setting to make the annotation study feasible.

The extracted sentences were independently annotated by two judges. One is a co-author of this paper (judge 1), and the other has performed subjectivity annotation before, but is not otherwise involved in this research (judge 2). Sentences were annotated according to the coding instructions of (Wiebe, Bruce, and O'Hara, 1999), which, recall, are to classify a sentence as subjective if there is a significant expression of subjectivity in the sentence, of either the writer or someone mentioned in the text. In addition to the subjective and objective classes, a judge could tag a sentence "unsure" if he or she is unsure of his or her rating or considers the sentence to be borderline.

An equal number (133) of other sentences were randomly selected from the corpus to serve as controls. The 133 system-identified sentences and the 133 control sentences were randomly mixed together. The judges were asked to annotate all 266 sentences, not knowing which are system-identified and which are control. Each sentence was presented with the sentence that precedes it and the sentence that follows it in the corpus, to provide some context for interpretation.

Table 12 shows examples of the system-identified sentences. Sentences classified by both judges as objective are marked "oo" and those classified by both judges as subjective are marked "ss".

Judge 1 classified 103 of the system-identified sentences as subjective; 16 as objective; and 14 as unsure. Judge 2 classified 102 of the system-identified sentences as subjective;

(1)	The outburst of shooting came nearly two weeks after clashes between Moslem worshippers and Somali soldiers.	oo
(2.a)	But now the refugees are streaming across the border and alarming the world.	ss
(2.b)	In the middle of the crisis, Erich Honecker was hospitalized with a gall stone operation.	oo
(2.c)	It is becoming more and more obvious that his gallstone-age communism is dying with him: ...	ss
(3.a)	Not brilliantly, because, after all, this was a performer who was collecting paychecks from lounges at Hiltons and Holiday Inns, but creditably and with the air of someone for whom "Ten Cents a Dance" was more than a bit autobiographical.	ss
(3.b)	"It was an exercise of blending Michelle's singing with Susie's singing," explained Ms. Stevens.	oo
(4)	Enlisted men and lower-grade officers were meat thrown into a grinder.	ss
(5)	"If you believe in God and you believe in miracles, there's nothing particularly crazy about that."	ss
(6)	He was much too eager to create "something very weird and dynamic," "catastrophic and jolly" like "this great and coily thing" "Lolita."	ss
(7)	The Bush approach of mixing confrontation with conciliation strikes some people as sensible, perhaps even inevitable, because Mr. Bush faces a Congress firmly in the hands of the opposition.	ss
(8)	Still, despite their efforts to convince the world that we are indeed alone, the visitors do seem to keep coming and, like the recent sightings, there's often a detail or two that suggests they may actually be a little on the dumb side.	ss
(9)	As for the women, they're pathetic.	ss
(10)	At this point, the truce between feminism and sensationalism gets might uneasy.	ss
(11)	MMPI's publishers say the test shouldn't be used alone to diagnose psychological problems or in hiring; it should be given in conjunction with other tests	ss
(12)	While recognizing that professional environmentalists may feel threatened, I intend to urge that UV-B be monitored whenever I can.	ss

Table 12

Examples of System-Identified Sentences

Bathed in cold sweat, I watched these Dantesque scenes, holding tightly the damp hand of Edek or Waldeck who, like me, were convinced that there was no God.
"The Japanese are amazed that a company like this exists in Japan," says Kimindo Kusaka, head of the Softnomics Center, a Japanese management-research organization.
And even if drugs were legal, what evidence do you have that the habitual drug user wouldn't continue to rob and steal to get money for clothes, food or shelter?
The moral cost of legalizing drugs is great, but it is a cost that apparently lies outside the narrow scope of libertarian policy prescriptions.
I doubt that one exists.
They were upset at his committee's attempt to pacify the program critics by cutting the surtax paid by the more afluent elderly and making up the loss by shifting more of the burden to the elderly poor and by delaying some benefits by a year.

Table 13

Examples of Subjective Sentences Adjacent to System-Identified Sentences

	S	O	U
S	98	2	3
O	2	14	0
U	2	11	1

Table 14

Sentence annotation contingency table; judge 1 counts are in rows and judge 2 counts are in columns.

27 as objective; and 4 as unsure. The contingency table is given in Table 14.[4]

The κ value using all three classes is 0.60, reflecting the highly skewed distribution in favor of subjective sentences, and the disagreement on the lower frequency classes ("unsure" and "objective"). Consistent with the findings in (Wiebe, Bruce, and O'Hara, 1999), the κ value for agreement on the sentences for which neither judge is unsure is very high: 0.86.

A different breakdown of the sentences is illuminating. For 98 of the sentences, call them SS, judges 1 and 2 tag the sentence as subjective. Among the other sentences, 20 appear in a block of contiguous system-identified sentences that includes a member of SS. For example, in Table 12, (2.a) and (2.c) are in SS and (2.b) is in the same block of subjective sentences as they are. Similarly, (3.a) is in SS and (3.b) is in the same block.

Among the remaining 15 sentences, 6 are adjacent to subjective sentences that were not identified by our system (so were not annotated by the judges). All of those sentences contain significant expressions of subjectivity of the writer or someone mentioned in the text, the criterion used in this work for classifying a sentence as subjective. Samples are shown in Table 13.

Thus, 93% of the sentences identified by the system are subjective or are near subjective sentences. All the sentences, together with their tags and the sentences adjacent

4 In contrast, Judge 1 classified only 53 (45%) of the control sentences as subjective and Judge 2 classified only 47 (36%) of them as subjective.

to them, are available on the Web at www.cs.pitt.edu/~wiebe.

4.6 Using Features for Opinion-piece Recognition

In this section, we assess the usefulness of the PSEs identified in section 3 and listed in Table 9 by using them to perform document-level classification of opinion pieces. Opinion-piece classification is a difficult task for two reasons. First, as discussed in section 3.1, both opinionated and factual documents tend to be composed of a mixture of subjective and objective language. Second, the natural distribution of documents in our data is heavily skewed toward non-opinion pieces. Despite these hurdles, using only our PSEs, we achieve positive results in opinion-piece classification using the basic k-nearest-neighbor (KNN) algorithm with leave-one-out cross-validation (Mitchell, 1997).

Given a document, the basic KNN algorithm classifies the document according to the majority classification of the document's k closest neighbors. For our purposes, each document is characterized by one feature, the count of all PSE instances (regardless of type) in the document, normalized by document length in words. The distance between two documents is simply the absolute value of the difference in the normalized PSE count for each document.

With leave-one-out cross-validation, the set of n documents to be classified is divided into a training set of size n-1 and a validation set of size 1. The one document in the validation set is then classified according to the majority classification of its k closest-neighbor documents in the training set. This process is repeated until every document is classified.

Which value to use for k is chosen during a preprocessing phase. During the pre-processing phase, we run KNN with leave-one-out cross-validation on a separate training set, for odd values of k from 1 to 15. The value of k that results in the best classification during the preprocessing phase is the one used for later KNN classification.

For the classification experiment, the dataset OP1 was used in the preprocessing phase to select the value of k, and then classification was performed on the 1222 documents in OP2. During training on OP1, k equal to 15 resulted in the best classification. On the test set, OP2, we achieved a classification accuracy of 0.939; the baseline accuracy for choosing the most frequent class (non-opinion pieces) is 0.915. Our classification accuracy is a 28% reduction in error and is significantly better than baseline according to McNemar's test (Everitt, 1977).

The positive results from the opinion-piece classification show the usefulness of the various PSE features when used together.

5 Relation to Other Work

There has been much work in other fields involving subjective language, including linguistics, literary theory, psychology, philosophy, and content analysis. As mentioned above in Section 2, the conceptualization underlying our manual annotations is based on work in literary theory and linguistics, most directly (Doležel, 1973; Uspensky, 1973; Kuroda, 1973; Kuroda, 1976; Chatman, 1978; Cohn, 1978; Fodor, 1979; Banfield, 1982). We also mentioned existing knowledge resources such as affective lexicons (General-Inquirer, 2000; Heise, 2000) and annotations in more general purpose lexicons (e.g., the *attitude* adverb features in Comlex (Macleod, Grishman, and Meyers, 1998)). Such knowledge may be used in future work to complement the work presented in this paper, for example to seed the distributional similarity process described in Section 3.4.

There is also work in fields such as content analysis and psychology on statistically characterizing texts in terms of word lists manually developed for distinctions related to subjectivity. For example, (Hart, 1984) performs counts of a manually-developed list of words and rhetorical devices (e.g., "sacred" terms such as *freedom*) in political speeches to explore potential reasons for public reactions. Anderson and McMasters (1989) use

fixed sets of high-frequency words to assign connotative scores to documents and sections of documents along dimensions such as how pleasant, acrimonious, pious, confident, etc. the text is.

What distinguishes our work from work on subjectivity in other fields is that we focus on (1) automatically learning knowledge from corpora, (2) automatically performing contextual disambiguation, and (3) using knowledge of subjectivity in NLP applications.

This paper expands and integrates the work reported in (Wiebe and Wilson, 2002; Wiebe, Wilson, and Bell, 2001; Wiebe et al., 2001; Wiebe, 2000).

Previous work in NLP on the same or related tasks includes sentence-level and document-level subjectivity classifications. At the sentence level, (Wiebe, Bruce, and O'Hara, 1999) developed a machine learning system to classify sentences as subjective or objective. The accuracy of the system was more than 20 percentage points higher than a baseline accuracy. Five part-of-speech features, two lexical features, and a paragraph feature were used. These results suggested to us that there are clues of subjectivity that might be learned automatically from text, and motivated the work reported in the current paper. The system was tested in 10-fold cross validation experiments using corpus WSJ-SE, a small corpus of only 1001 sentences. As discussed in Section 1, a main goal of our current work is to exploit existing document-level annotations, because they enable us to use much larger datasets, they were created outside our research group, and they allow us to assess consistency of performance by cross validating between our manual annotations and the existing document level annotations. Because the document-level data is not annotated at the sentence level, sentence-level classification is not highlighted in this paper. The new sentence annotation study to evaluate sentences with high-density features (Section 4.5) uses different data from WSJ-SE, because some of the features (n-grams and density parameters) were identified using WSJ-SE as training data.

Other previous work in NLP addressed related document-level classifications. Spertus

(1997) developed a system for recognizing inflammatory messages. As mentioned earlier in the paper, inflammatory language is a type of subjective language, so the task she addresses is closely related to ours. She uses machine learning to select among manually developed features. In contrast, the focus in our work is automatically identifying features from the data.

A number of projects investigating genre detection include editorials as one of the targeted genres. For example, in Karlgren and Cutting (1994), editorials are one of 15 categories, and in Kessler et al. (1997), editorials are one of six. Given their goal to perform genre detection in general, they use low-level features that are not specific to editorials. Neither shows significant improvements for editorial recognition. Argamon et al. (1998) address a slightly different task, though it does involve editorials. Their goal is to distinguish not only, e.g., news from editorials, but also these categories in different publications. Their best results are distinguishing among the news categories of different publications; their lowest results involve editorials. Because we focus specifically on distinguishing opinion pieces from non-opinion pieces, our results are better than theirs for those categories. In addition, in contrast to the above studies, the focus of our work is learning features of subjectivity. We perform opinion-piece recognition in order to assess the usefulness of the various features when used together.

Other previous NLP research used features similar to ours for other NLP tasks.

Low-frequency words were used as features in information extraction (Weeber, Vos, and Baayen, 2000) and text categorization (Copeck et al., 2000).

A number of researchers worked on mining collocations from text to extend lexicographic resources for machine translation and word sense disambiguation (e.g., (Smajda, 1993; Lin, 1999; Biber, 1993)).

In Samuel et al.'s (1998) work on identifying collocations for dialog-act recognition, a similar filter as ours was used to eliminate redundant n-gram features: n-grams were

eliminated if they contain substrings with the same or better entropy score.

While it is common in studies of collocations to omit low-frequency words and expressions from analysis, because they give rise to invalid or unrealistic statistical measures (Church and Hanks, 1990), we are able to identify higher-precision collocations by including placeholders for unique words (i.e., the ugen-n-grams). We are not aware of other work that uses such collocations as we do.

Features identified using distributional similarity were used for syntactic and semantic disambiguation (Hindle, 1990; Dagan, Pereira, and Lee, 1994) and to develop lexical resources from corpora (Lin, 1998; Riloff and Jones, 1999).

We are not aware of other work identifying and using density parameters as described in this paper.

Since our experiments, other related work in NLP has been performed. Some address related but different classification tasks. Three studies classify reviews as positive or negative (Turney, 2002; Pang, Lee, and Vaithyanathan, 2002; Dave, Lawrence, and Pennock, 2003). The input is assumed to be a review, so this task does not include finding subjective documents in the first place. The first study listed above (Turney, 2002) uses a variation of the semantic similarity procedure presented in (Wiebe, 2000) (Section 3.4). The third (Dave, Lawrence, and Pennock, 2003) uses n-gram features identified with a variation of the procedure presented in (Wiebe, Wilson, and Bell, 2001) (Section 3.3). Tong (2001) addresses finding *sentiment timelines*, i.e., tracking sentiments over time in multiple documents. For clues of subjectivity, he uses manually developed lexical rules, rather than automatically learning them from corpora. Similarly, Gordon et al. (2003) use manually developed grammars to detect some types of subjective language. Agrawal et al. (2003) partition newsgroup authors into camps based on quotation links. They do not attempt to recognize subjective language.

The most closely related new work are (Riloff, Wiebe, and Wilson, 2003; Riloff and

Wiebe, 2003; Yu and Hatzivassiloglou, 2003). The first two focus on finding additional types of subjective clues (nouns and extraction patterns identified using extraction pattern bootstrapping). Yu and Hatzivassiloglou (2003) perform opinion text classification. They also use existing WSJ document classes for training and testing, but they do not include the entire corpus in their experiments, as we do. Their opinion-piece class consists only of *Editorials* and *Letters to the Editor*, and their non-opinion class consists only of *Business* and *News*. They report an average F-measure of 96.5%. Our result of 94% accuracy on document level classification is almost comparable. They also perform sentence-level classification, including polarity.

We anticipate that knowledge of subjective language may be usefully exploited in a number of NLP application areas, and hope that the work presented in this paper will encourage others to experiment with subjective language in their applications. More generally, there are many types of AI systems for which state-of-affairs types such as beliefs and desires are central, including systems that perform plan recognition to understand narratives (Dyer, 1982; Lehnert et al., 1983), argument understanding (Alvarado, Dyer, and Flowers, 1986), understanding stories from different perspectives (Carbonell, 1979), and generating language under different pragmatic constraints (Hovy, 1987). Knowledge of linguistic subjectivity could enhance the abilities of such systems to recognize and generate expressions referring to such states of affairs in natural text.

6 Conclusions

Knowledge of subjective language promises to be beneficial for many NLP applications including information extraction, question answering, text categorization, and summarization. This paper presents the results of an empirical study in learning knowledge of subjective language from corpora, in which a number of feature types are learned and evaluated on different types of data with positive results.

We showed that unique words are more often subjective than expected and that unique words are valuable clues of subjectivity. We also presented for automatically identifying potentially subjective collocations, including fixed collocations and collocations with placeholders for unique words. In addition, we used the results of a method for clustering words according to distributional similarity (Lin, 1998) to identify adjectival and verbal clues of subjectivity.

Table 9 summarizes the results of testing all of the above types of potentially subjective expressions ($PSEs$). All show increased precision in the evaluations. Together, they show consistency in performance. In almost all cases they perform better or worse on the same datasets. This is so despite the fact that different kinds of data and procedures are used to learn them. In addition, PSEs learned using expression level subjective-element data have precisions higher than baseline on document-level opinion-piece data, and vice versa.

Having a large stable of PSEs, it was important to disambiguate whether or not PSE instances are subjective in the contexts in which they appear. We discovered that the density of other potentially subjective expressions in the surrounding context is important. If a sufficient number of other clues are nearby, a clue is more likely to be subjective than if there are not. Parameter values are selected using training data manually annotated at the expression level for subjective elements, and then tested on data annotated at the document level for opinion pieces. All of the selected parameters lead to increases in precision on the test data, most leading to increases over 100%. Once again we found consistency between expression-level and document-level annotations. PSE sets defined by density have high precision in both the subjective-element data and the opinion-piece data. The large differences between training and testing suggest that our results are not brittle.

Using a density feature selected from a training set, sentences containing high-density

PSEs were extracted from a separate test set, and manually annotated by two judges. Fully 93% of the sentences are subjective or are near subjective sentences. Admittedly, the chosen density feature is a high-precision, low-frequency one. But since the process is fully automatic, the feature could be applied to more unannotated text to identify regions containing subjective sentences. In addition, because the precision and frequency of the density features is stable across datasets, lower precision but higher frequency options are available.

Finally, the value of the various types of PSEs was demonstrated with the task of opinion-piece classification. Using the k-nearest neighbor classification algorithm with leave-one-out cross-validation, a classification accuracy of 94% was achieved on a large test set, with a reduction in error of 28% from the baseline.

Future work is required to determine how to exploit density features to improve the performance of text categorization algorithms. Another area of future work is searching for clues of objectivity, such as the politeness features used by (Spertus, 1997). Another is identifying the type of a subjective expression (e.g., positive or negative evaluative), extending work such as (Hatzivassiloglou and McKeown, 1997) on classifying lexemes to the classification of instances in context (compare, e.g., "great!" and "oh great.")

In addition, it would be illuminating to apply our system to data annotated with discourse trees (Carlson, Marcu, and Okurowski, 2001). We hypothesize that most objective sentences identified by our system are dominated in the discourse by subjective sentences, and that we are moving toward identifying subjective discourse segments.

Acknowledgments

We thank the anonymous reviewers for their helpful and constructive comments. This research was supported in part by the Office of Naval Research under grants N00014-95-1-0776 and N00014-01-1-0381.

References

Agrawal, Rakesh, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International World Wide Web Conference (WWW2003)*. Web Proceedings.

Alvarado, Sergio J., Michael. G. Dyer, and Margot Flowers. 1986. Editorial comprehension in oped through argument units. In *Proc. of the 1986 National Conference on Artificial Intelligence (AAAI-86)*, pages 250–256.

Anderson, Clifford W. and George C. McMaster. 1989. Quantification of rewriting by the brothers grimm: A comparison of successive versions of three tales. *Computers and the Humanities*, 23(4-5):341–346.

Aone, Chinatsu, Mila Ramos-Santacruz, and William J. Niehaus. 2000. Assentor: An nlp-based solution to e-mail monitoring. In *Proceedings of the Eleventh Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-2000)*, pages 945–950.

Argamon, Shlomo, Moshe Koppel, and Galit Avneri. 1998. Routing documents according to style. In *Proceedings of the First International Workshop on Innovative Internet Information Systems (IIIS-98)*.

Banfield, Ann. 1982. *Unspeakable Sentences*. Routledge and Kegan Paul, Boston.

Barzilay, Regina, Michael Collins, Julia Hirschberg, and Steve Whittaker. 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 679–684.

Biber, Douglas. 1993. Co-occurrrence patterns among collocations: A tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, 19(3):531–538.

Brill, Eric. 1992. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP-92)*, pages 152–155.

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

Carbonell, J. G. 1979. *Subjective Understanding: Computer Models of Belief Systems*. Ph.D. thesis, Tech. Rept. 150, Department of Computer Science, Yale University, New Haven, CT.

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged

corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd ACL Workshop on Discourse and Dialogue (SIGdial-2001)*, pages 30–39.

Chatman, Seymour. 1978. *Story and Discourse: Narrative Structure in Fiction and Film.* Cornell University Press, Ithaca, NY.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 1:22–29.

Cohn, Dorrit. 1978. *Transparent Minds: Narrative Modes for Representing Consciousness in Fiction.* Princeton University Press, Princeton, NJ.

Copeck, Terry, Kim Barker, Sylvain Delisle, and Stan Szpakowicz. 2000. Automating the measurement of linguistic features to help classify texts as technical. In *Proceedings 7th Conference on Automatic NLP (TALN-2000)*, pages 101–110.

Dagan, Ido, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pages 272–278.

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of produce reviews. In *Proceedings of the 12th International World Wide Web Conference (WWW2003)*. Web Proceedings.

Doležel, Lubomir. 1973. *Narrative Modes in Czech Literature.* University of Toronto Press, Toronto, Canada.

Dyer, Michael G. 1982. Affect processing for narratives. In *Proceedings of the Second National Conference on Artificial Intelligence (AAAI-82)*, pages 265–268.

Everitt, Brian S. 1977. *The Analysis of Contingency Tables.* Chapman and Hall, London.

Fludernik, Monika. 1993. *The Fictions of Language and the Languages of Fiction.* Routledge, London.

Fodor, Janet Dean. 1979. *The Linguistic Description of Opaque Contexts.* Outstanding dissertations in linguistics 13. Garland, New York & London.

General-Inquirer, The. 2000. *http://www.wjh.harvard.edu/~inquirer/homecat.htm.*

Gordon, Andrew, Abe Kazemzadeh, Anish Nair, and Milena Petrova. 2003. Recognizing expressions of commonsense psychology in english text. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, pages 208–215.

Hart, Roderick P. 1984. Systematic analysis of political discourse: The development of diction. In K. Sanders et al., editor, *Political Communication Yearbook: 1984*. Southern Illinois University Press, pages 97–134.

Hatzivassiloglou, Vasileios and Kathy McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 174–181.

Heise, David. 2000. *http://www.indiana.edu/socpsy/ACT/index.htm*. Affect Control Theory.

Hindle, Don. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL-90)*, pages 268–275.

Hovy, Eduard. 1987. *Generating Natural Language under Pragmatic Constraints*. Ph.D. thesis, Yale University.

Karlgren, Jussi and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94)*, pages 1071–1075.

Karp, Daniel, Yves Schabes, Martin Zaidel, and Dania Egedi. 1992. A freely available wide coverage morphological analyzer for English. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING-92)*, pages 922–928.

Kaufer, David. 2000. *Flaming: A White Paper*. www.eudora.com.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 32–38.

Kuroda, S.-Y. 1973. Where epistemology, style and grammar meet: A case study from the japanese. In P. Kiparsky and S. Anderson, editors, *A Festschrift for Morris Halle*. Holt, Rinehart & Winston, New York, NY, pages 377–391.

Kuroda, S.-Y. 1976. Reflections on the foundations of narrative theory—from a linguistic point of view. In T.A. van Dijk, editor, *Pragmatics of Language and Literature*. North-Holland, Amsterdam, pages 107–140.

Lee, Lillian. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 25–32.

Lee, Lillian and Fernando Pereira. 1999. Distributional similarity models: Clustering vs. nearest neighbors. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 33–40.

Lehnert, Wendy G., Michael Dyer, Peter Johnson, C.J. Yang, and Steve Harley. 1983. BORIS: An Experiment in In-Depth Understanding of Narratives. *Artificial Intelligence*, 20:15–62.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-98)*, pages 768–773.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 317–324.

Litman, Diane J. and Rebecca J. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, pages 108–115.

Macleod, Catherine, Ralph Grishman, and Adam Meyers. 1998. Comlex syntax reference manual. Technical report, New York University.

Marcu, Daniel, Magdalena Romera, and Estibaliz Amorrortu. 1999. Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. In *Proceedings of the International Workshop on Levels of Representation in Discourse (LORID-99)*, pages 71–78.

Marcus, Mitch, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Mitchell, Tom. 1997. *Machine Learning*. McGraw-Hill.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 79–86.

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

Riloff, Ellen and Rosie Jones. 1999. Learning Dictionaries for Information Extraction by

Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-1999)*, pages 474–479.

Riloff, Ellen 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.

Riloff, Ellen, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pages 25–32.

Sack, Warren. 1995. Representing and recognizing point of view. In *Proceedings AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, page 152.

Samuel, Ken, Sandra Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-98)*, pages 1150–1156.

Smajda, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

Spertus, Ellen. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of the Eighth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-97)*, pages 1058–1065.

Stein, Dieter and Susan Wright, editors. 1995. *Subjectivity and Subjectivisation.* Cambridge University Press, Cambridge.

Terveen, Loren, Will Hill, Brian Amento, David McDonald, and Josh Creter. 1997. Building task-specific interfaces to high volume conversational data. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI-97)*, pages 226–233.

Teufel, Simone and Marc Moens. 2000. What's yours and what's mine: Determining intellectual attribution in scientific texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the Workshop on Very Large Corpora (EMNLP/VLC-2000)*, pages 9–17.

Tong, Richard. 2001. An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the SIGIR Workshop on Operational Text Classification*, pages 1–6.

Turney, Peter. 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-2000)*, pages 417–424.

Uspensky, Boris. 1973. *A Poetics of Composition*. University of California Press, Berkeley, CA.

Weeber, Marc, Rein Vos, and R. Harald Baayen. 2000. Extracting the lowest-frequency words: Pitfalls and possibilities. *Computational Linguistics*, 26(3):301–317.

Wiebe, J. and W. Rapaport. 1986. Representing *de re* and *de dicto* belief reports in discourse and narrative. *Proceedings of the IEEE*, 74:1405–1413.

Wiebe, J. and T. Wilson. 2002. Learning to disambiguate potentially subjective expressions. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 112–118.

Wiebe, Janyce. 1990. *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text*. Ph.D. thesis, State University of New York at Buffalo.

Wiebe, Janyce. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Wiebe, Janyce. 2000. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 735–740.

Wiebe, Janyce, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. 2003. Recognizing and organizing opinions expressed in the world press. In *Working Notes of the AAAI Spring Symposium in New Directions in Question Answering*, pages 12–19.

Wiebe, Janyce, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. 2001. A corpus study of evaluative and speculative language. In *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-2001)*, pages 186–195.

Wiebe, Janyce, Rebecca Bruce, and Thomas O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 246–253.

Wiebe, Janyce, Kenneth McKeever, and Rebecca Bruce. 1998. Mapping collocational properties into machine learning features. In *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-98)*, pages 225–233.

Wiebe, Janyce and William J. Rapaport. 1988. A computational theory of perspective and reference in narrative. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL-88)*, pages 131–138.

Wiebe, Janyce, Theresa Wilson, and Matthew Bell. 2001. Identifying collocations for recognizing opinions. In *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, pages 24–31.

Wiebe, Janyce M. and William J. Rapaport. 1991. References in narrative text. *Noûs*, 25(4):457–486.

Wilson, Theresa and Janyce Wiebe. 2003. Annotating opinions in the world press. In *Proceedings of the 4th ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*, pages 13–22.

Yu, Hong 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.