A Corpus Study of Evaluative and Speculative Language

Janyce Wiebe[*], **Rebecca Bruce**[†], **Matthew Bell**[*], **Melanie Martin**[‡], **Theresa Wilson**[*]
University of Pittsburgh[*], University of North Carolina at Asheville[†], New Mexico State University[‡]
wiebe,mbell,twilson@cs.pitt.edu, bruce@cs.unca.edu, mmartin@cs.nmsu.edu

Abstract

This paper presents a corpus study of evaluative and speculative language. Knowledge of such language would be useful in many applications, such as text categorization and summarization. Analyses of annotator agreement and of characteristics of subjective language are performed. This study yields knowledge needed to design effective machine learning systems for identifying subjective language.

1 Introduction

Subjectivity in natural language refers to aspects of language used to express opinions and evaluations (Banfield, 1982; Wiebe, 1994). *Subjectivity tagging* is distinguishing sentences used to present opinions and other forms of subjectivity (*subjective sentences*) from sentences used to objectively present factual information (*objective sentences*). This task is especially relevant for news reporting and Internet forums, in which opinions of various agents are expressed. There are numerous applications for which subjectivity tagging is relevant. Two are information retrieval and information extraction. Current extraction and retrieval technology focuses almost exclusively on the subject matter of documents. However, additional aspects of a document influence its relevance, including, e.g., the evidential status of the material presented, and the attitudes expressed about the topic (Kessler et al., 1997). Knowledge of subjective language would also be useful in flame recognition (Spertus, 1997; Kaufer, 2000), email classification (Aone et al., 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), generation and style (Hovy, 1987), clustering documents by ideological point of view (Sack, 1995), and any other application that would benefit from knowledge of how opinionated the language is, and whether or not the writer purports to objectively present factual material.

To use subjectivity tagging in applications, good linguistic clues must be found. As with many pragmatic and discourse distinctions, existing lexical resources are not comprehensively coded for subjectivity. The goal of our current work is learning subjectivity clues from corpora. This paper contributes to this goal by empirically examining subjectivity. We explore annotating subjectivity at different levels (expression, sentence, document) and produce corpora annotated at different levels. Annotator agreement is analyzed to understand and assess the viability of such annotations. In addition, because expression-level annotations are fine-grained and thus very informative, these annotations are examined to gain knowledge about subjectivity.

We also use our annotations and existing editorial annotations to generate and test features of subjectivity. Altogether, the observations and results of these studies provide valuable information that will facilitate designing effective machine learning systems for recognizing subjectivity.

The remainder of this paper first provides background about subjectivity, then presents results for document-level annotations, followed by an analysis of expression-level annotations. Results for features generated using document-level annotations are next, ending with conclusions.

2 Subjectivity

Sentence (1) is an example of a simple subjective sentence, and (2) is an example of a simple objective sentence:[1]

(1) At several different layers, it's a fascinating tale.

[1] The term *subjectivity* is due to Ann Banfield (1982). For references to work on subjectivity, please see (Banfield, 1982; Fludernik, 1993; Wiebe, 1994; Stein and Wright, 1995).

(2) Bell Industries Inc. increased its quarterly to 10 cents from 7 cents a share.

The main types of subjectivity are:

1. *Evaluation.* This category includes emotions such as hope and hatred as well as evaluations, judgements, and opinions. Examples of expressions involving positive evaluation are *enthused*, *wonderful*, and *great product!*. Examples involving negative evaluation are *complained*, *you idiot!*, and *terrible product*.

2. *Speculation.* This category includes anything that removes the presupposition of events occurring or states holding, such as speculation and uncertainty. Examples of speculative expressions are *speculated*, and *maybe*.

Following are examples of strong negative evaluative language from a corpus of Usenet newsgroup messages:

(3a) I had in mind your facts, buddy, not hers.

(3b) Nice touch. "Alleges" whenever facts posted are not in your persona of what is "real".

Following is an example of opinionated, editorial language, taken from an editorial in the Wall Street Journal:

(4) We stand in awe of the Woodstock generation's ability to be unceasingly fascinated by the subject of itself.

Sentences (5) and (6) illustrate the fact that sentences about speech events may be subjective or objective:

(5) 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.

(6) "The cost of health care is eroding our standard of living and sapping industrial strength," complains Walter Maher, a Chrysler health-and-benefits specialist.

In (5), the material about lawsuits and claims is presented as factual information, and a federal judge is given as the source of information. In (6), in contrast, a complaint is presented. An NLP system performing information extraction on (6) should not treat the material in the quoted string as factual information, with the complainer as a source of information, whereas a corresponding treatment of sentence (5) would be appropriate.

Subjective sentences often contain individual expressions of subjectivity. Examples are *fascinating* in (1), and *eroding*, *sapping*, and *complains* in (6). The following paragraphs mention aspects of subjectivity expressions that are relevant for NLP applications.

First, although some expressions, such as *!*, are subjective in all contexts, many, such as *sapping* and *eroding*, may or may not be subjective, depending on the context in which they appear. 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).

Second, 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 (1) is the writer, while the source of the subjective elements in (6) is Maher. In addition, a subjective element has a *target*, i.e., what the subjectivity is about or directed toward. In (1), the target is a tale; in (6), the target of Maher's subjectivity is the cost of health care. These are examples of *object-centric subjectivity*, which is about an object mentioned in the text (other examples: "I love this project"; "The software is horrible"). Subjectivity may also be *addressee-oriented*, i.e., directed toward the listener or reader (e.g., "You are an idiot").

Third, there may be multiple subjective elements in a sentence, possibly of different types and attributed to different sources and targets. For example, in (4), subjectivity of the Woodstock generation is described (specifically, its fascination with itself). In addition, subjectivity of the writer is expressed (e.g., 'we stand in awe'). As described below, individual subjective elements were annotated as part of this work, refining previous work on sentence-level annotations. Finally, PSEs may be complex expressions such as 'village idiot', 'powers that be', 'You' *NP*, and 'What a' *NP*. There is a great variety of such expressions, including many studied under the rubric of idioms (see, for example, (Nunberg et al., 1994)). We address learning such expressions in another project.

3 Previous Work on Subjectivity Tagging

In previous work (Wiebe et al., 1999; Bruce and Wiebe, 1999), a corpus of sentences from the Wall Street Journal Treebank Corpus (Marcus et al., 1993) was manually annotated with subjectivity classifications by multiple judges. The judges were instructed to consider a sentence to be subjective if they perceived any significant expression of subjectivity (of any source) in the sentence, and to consider the sentence to be objective, otherwise. Agreement was summarized in terms of Cohen's κ (Cohen, 1960), which compares the total probability of agreement to that expected if the taggers' classifications were statistically independent (i.e., "chance agreement"). After two rounds of tagging by three judges, an average pairwise κ value of .69 was achieved on a test set. The EM learning algorithm was used to produce corrected tags representing the consensus opinions of the taggers (Goodman, 1974; Dawid and Skene, 1979). An automatic system to perform subjectivity tagging was developed using the new tags as training and testing data. In 10-fold cross validation experiments, a probabilistic classifier obtained an average accuracy on subjectivity tagging of 72.17%, more than 20 percentage points higher than a baseline accuracy obtained by always choosing the more frequent class. Five part-of-speech features, two lexical features, and a paragraph feature were used.

To identify richer features, (Wiebe, 2000) used Lin's (1998) method for clustering words according to distributional similarity, seeded by a small amount of detailed manual annotation, to automatically identify adjective PSEs. There are two parameters of this process, neither of which was varied in (Wiebe, 2000): C, the cluster size considered, and FT, a filtering threshold, such that, if the seed word and the words in its cluster have, as a set, lower precision than the filtering threshold on the training data, the entire cluster, including the seed word, is filtered out. This process is adapted for use in the current paper, as described in section 7.

4 Choices in Annotation

In expression-level annotation, the judges first identify the sentences they believe are subjective. They next identify the subjective elements in the sentence, i.e., the expressions they feel are responsible for the subjective classification. For example (subjective elements are in parentheses):

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

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

Subjective-element (expression-level) annotations are probably the most natural. Ultimately, we would like to recognize the subjective elements in a text, and their types, targets, and sources. However, both manual and automatic tagging at this level are difficult because the tags are very fine-grained, and there is no predetermined classification unit; a subjective element may be a single word or a large expression. Thus, in the short term, it is probably best to use subjective-element annotations for knowledge acquisition (analysis, training, feature generation) alone, and not target automatic classification of subjective elements.

In this work, document-level subjectivity annotations are text categories of which subjectivity is a key aspect. We use three text categories: editorials (Kessler et al., 1997), reviews, and "flames", i.e., hostile messages (Spertus, 1997; Kaufer, 2000). For ease of discussion, we group editorials and reviews together under the term *opinion pieces.*

There are benefits to using such document-level annotations. First, they are more directly related to applications (e.g., filtering hostile messages and mining reviews from Internet forums). Second, there are existing annotations to be exploited, such as editorials and arts reviews marked as such by newspapers, as well as on-line product reviews accompanied by formal numerical ratings (e.g., 4 on a scale from 1 to 5).

However, a challenging aspect of such data is that opinion pieces and flames contain objective sentences, while documents in other text categories contain subjective sentences. News reports present reactions to and attitudes toward 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). On the other hand, editorials contain objective sentences presenting facts supporting the writer's argument, and reviews contain sentences objectively presenting facts about the product. This "impure" aspect of opinionated text categories must be considered when such data is used for training and testing. Some specific results are given below in section 7.

We believe that sentence-level classifications will continue to provide an important level of analysis. The sentence provides a prespecified classification unit[2] and, while sentence-level judgements are not as fine-grained as subjective-

[2] While sentence boundaries are not always unambiguous in unedited text or spoken language, the data can always be segmented into sentence-like units before subjectivity tagging is performed.

element judgements, they do not involve the large amount of noise we face with document-level annotations.

5 Document-Level Annotation Results

5.1 Flame Annotations

In this study, newsgroup messages were assigned the tags *flame* or *not-flame*. The corpus consists of 1140 Usenet newsgroup messages, balanced among the categories alt, sci, comp, and rec in the Usenet hierarchy. The corpus was divided, preserving the category balance, into a training set of 778 messages and a test set of 362 messages.

The annotators were instructed to mark a message as a flame if the "main intention of the message is a personal attack, containing insulting or abusive language." A number of policy decisions were made in the instructions, dealing, primarily, with included messages (part or all of a previous message, included in the current message as part of a reply). Some additional issues addressed in the instructions were who the attack was directed at, nonsense, sarcasm, humor, rants, and raves.

During the training phase, two annotators, MM and R, participated in multiple rounds of tagging, revising the annotation instructions as they proceeded. During the testing phase, MM and R independently annotated the test set, achieving a κ value on these messages of 0.69. A third annotator, L, trained on 492 messages from the training set, and then annotated 88 of the messages in the test set. The pairwise κ values on this set of 88 are: MM & R: 0.80; MM & L: 0.75; R & MM: 0.80; for an average pairwise κ of .78.

This study provides evidence for the viability of document-level flame annotation. We plan to build a flame-recognition system in the future. As will be seen below, MM and R also tagged this data at the subjective-element level.

5.2 Opinion-Piece Classifications

Our opinion-piece classifications are built on existing annotations in the Wall Street Journal. Specifically, there are articles explicitly identified to be *Editorials, Letters to the Editor, Arts & Leisure*, and *Viewpoints*; together, we call these *opinion pieces*. This data is a good resource for subjectivity recognition. However, an inspection of some data revealed that some editorials and reviews are not marked as such. For example, there are articles written in the first person, and the purpose of the article is to present an argument rather than cover a news story, but there is no explicit indication that they are editorials. To create high quality test data, two judges manually annotated WSJ data for opinion pieces. The instructions were to find any additional opinion pieces that are not marked as such. The annotators also had the option of disagreeing with the existing annotations, but did not opt to do so in any instances.

One judge annotated all articles in four datasets of the Wall Street Journal Treebank corpus (Marcus et al., 1993) (W9-4, W9-10, W9-22, and W9-33, each approximately 160K words) as well as the corpus of Wall Street Journal articles used in (Wiebe et al., 1999) (called *WSJ-SE* below). Another judge annotated all articles in two of the datasets (W9-22 and W9-33).

This annotation task appears to be relatively easy. With no training at all, the κ values are very high: .94 for dataset W9-33 and .95 for dataset W9-22.

The agreement data for W9-22 is given in Table 1 in the form of a contingency table. In section 7, this data is used to generate and test candidate potential subjective elements (PSEs).

6 Subjective-Element Annotation Results and Analyses

6.1 Annotations and Data

These subsections analyze subjective element annotations performed on three datasets, *WSJ-SE*, *NG-FE*, and *NG-SE*.

WSJ-SE is the corpus of 1001 sentences of the Wall Street Journal Treebank Corpus referred to above in section 3. Recall that the sentences of this corpus were manually annotated with subjectivity classifications as described in (Wiebe et al., 1999; Bruce and Wiebe, 1999).

For this paper, two annotators (D and M) were asked to identify the subjective elements in WSJ-SE. Specifically, the taggers were given the subjective sentences identified in the previous study, and asked to put brackets around the words they believe cause the sentence to be classified as subjective.

Note that inflammatory language is a kind of subjective language. NG-FE is a subset of the Usenet newsgroup corpus used in the document-level flame-annotation study described in section 5.1. Specifically, NG-FE consists of the 362-message test set for taggers R and MM. For this study, R and MM were asked to identify the *flame elements* in 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 all 362 messages, because some messages that were not judged to be flames at the message level do contain individual inflammatory phrases

		Tagger 2		
		Op	*Not Op*	
Tagger 1	*Op*	$n_{11} = 23$	$n_{12} = 0$	$n_{1+} = 23$
	Not Op	$n_{21} = 2$	$n_{22} = 268$	$n_{2+} = 270$
		$n_{+1} = 25$	$n_{+2} = 268$	$n_{++} = 293$

Table 1: Contingency Table for Opinion Piece Agreement in W9-22

(in these cases, the tagger does not believe that these phrases express the main intent of the message).

In addition to the above annotations, tagger M performed subjective-element tagging on a different set of Usenet newsgroup messages, corpus *NG-SE*. The size of this corpus is 15413 words.

In datasets WSJ-SE and NG-SE, the taggers were also asked to specify one of five subjective element types: $e+$ (positive evaluative), $e-$ (negative evaluative), $e?$ (some other type of evaluation), u (uncertainty), and o (none of the above), with the option to assign multiple types to an instance. All corpora were stemmed (Karp et al., 1992) and part-of-speech tagged (Brill, 1992).

6.2 Agreement Among Taggers

There are techniques for analyzing agreement when annotations involve segment boundaries (Litman and Passonneau, 1995; Marcu et al., 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 is excluded from our analyses. The WSJ data is divided into two subsets in this section, *Exp1* and *Exp2*.

As mentioned above, in WSJ-SE Exp1 and Exp2, the taggers also classified subjective elements with respect to the type of subjectivity being expressed. Subjectivity type agreement is again analyzed at the word level, but, in this analysis, only the words classified as belonging to subjective elements by both taggers are considered.

Table 2 provides κ values for word agreement in NG-FE (the flame data) as well as for WSJ-SE Exp1 and Exp2. The task of identifying subjective elements in a body of text is difficult, and the agreement results reflect this fact; agreement is much stronger than that expected by chance, but less than what we would like to see when verifying a new classification. Further refinement of the coding manual is required. Additionally, it may be possible to refine the classifications automatically using methods such as those described in (Wiebe et al., 1999). In this analysis, we explore the patterns of agreement exhibited by the taggers in an effort to better understand the classification.

We begin by looking at word agreement. Word agreement is higher in the flame experiment (NG-FE) than it is in either WSJ experiment (WSJ-SE Exp1 and Exp2). Looking at the WSJ data provides one plausible explanation for the lower word agreement in the WSJ experiments. As exhibited in the subjective elements identified for the single clause below,

D: (e+ played the role well) (e? obligatory ragged jeans a thicket of long hair and rejection of all things conventional)

M: (e+ well) (e? obligatory) (e- ragged) (e? thicket) (e- rejection) (e- all things conventional)

tagger D consistently identifies entire phrases as subjective, while Tagger M prefers to select discrete lexical items. This difference in interpretation of the tagging instructions does not occur in the flame experiment. Nonetheless, even within the flame data, there are many instances where both taggers identify the same segment of a sentence as forming a subjective element but disagree on the boundaries of that segment, as in the example below.

R: (classic case of you deliberately misinterpreting my comments)

MM: (you deliberately misinterpreting my comments)

These patterns of partial agreement are also evident in the κ values for words from specific syntactic categories (see Table 2 again). In the WSJ data, agreement on determiners is particularly low because they are often included as part of a phrase by tagger D but typically not included in the specific lexical items chosen by tagger M. Interestingly, in the WSJ experiments, the taggers most frequently agreed on the selection of modals and adjectives, while in the flame experiment, agreement was highest on nouns and adjectives. The high agreement on adjectives in both genres is con-

	All Words	Nouns	Verbs	Modals	Adj's	Adverbs	Det's
NG-FE	0.4657	0.5213	0.4571	0.4008	0.5011	0.3576	0.4286
WSJ-SE, Exp1	0.4228	0.3999	0.4235	0.6992	0.6000	0.4328	0.2661
WSJ-SE, Exp2	0.3703	0.3705	0.4261	0.4298	0.4294	0.2256	0.1234

Table 2: κ Values for Word Agreement

sistent with results from other work (Bruce and Wiebe, 1999; Wiebe et al., 1999), but high agreement on nouns in the flame data verses high agreement on modals in the WSJ data suggests a genre specific usage of these categories. This would be the case if, for example, modals were most frequently used to express uncertainty, a type of subjectivity that would be relatively rare in flames.

Turning to subjective-element type, in both WSJ experiments, the κ values for type agreement are comparable to those for word agreement. Recall that multiple types may be assigned to a single subjective instance. All such instances in the WSJ data are *u* in combination with an evaluative tag (i.e., *e+*, *e-* and *e?*), and they are not common: each tagger assigned multiple tags to fewer than 7% of the subjective instances. However, if partial matches between type tags are recognized, i.e., if they share a common tag, then the κ values improve significantly. Table 3 shows both types of results.

It is interesting to note the variation in type agreement for words of different syntactic categories. Agreement on adjectives is consistently high while the agreement on the type of subjectivity expressed by modals and adverbs is consistently low. This contrasts with the fact that word agreement for modals, in particular, and, to a lesser extent, adverbs was high. This lack of agreement suggests that the type of subjectivity expressed by adjectives is more easily distinguished than that of modals or adverbs. This is particularly important because the number of adjectives included in subjective elements is high. In contrast, the numbers of modals and adverbs are relatively low.

Additional insight can be gained by combining the 3 evaluative classifications (i.e., *e+*, *e-* and *e?*) to form a single tag, *e*, representing any form of evaluative expression. Table 4 presents type agreement results for the tag set *e, u, o*. In contrasting Tables 3 and 4, it is surprising to note that most of the κ values decrease when the distinction among the evaluative types is removed. This suggests that the three evaluative types are natural classifications. Only for adverbs does type agreement improve with the smaller tag set; this indicates that it is difficult to distinguish the evaluative nature of adverbs. Note also that agreement for modals is not impacted by the change in tag sets. This fact supports the hypothesis that modals are used primary to express uncertainty. As a final point, we look at patterns of agreement in type classification using the models of symmetry, marginal homogeneity, quasi-independence, and quasi-symmetry. Each model tests for a specific pattern of agreement: symmetry tests the interchangeability of taggers, marginal homogeneity verifies the absence of bias among taggers, quasi-independence verifies that the taggers act independently when they disagree, and quasi-symmetry tests for the presence of any pattern in their disagreements. For a more complete description of these models and their use in analyzing intercoder reliability see (Bruce and Wiebe, 1999). In short, the results presented in Table 5 indicate that the taggers are not interchangeable: they exhibit biases in their type classifications, and there is a pattern of correlated disagreement in the assignment of the original type tags. Surprisingly, the taggers appear to act independently when they disagree in assigning the compressed type tags (i.e., tags *e*, *u* and *o*). This shift in the pattern of disagreement between taggers again suggests that the compression of the evaluative tags was inappropriate. Additionally, these findings suggest that it may be possible to automatically correct the type biases expressed by the taggers using the technique described in (Bruce and Wiebe, 1999), a topic that will be investigated in future work.

6.3 Uniqueness

Based on previous work (Wiebe et al., 1998), we hypothesized that low-frequency words are associated with subjectivity. Table 6 provides evidence that the number of unique words (words that appear just once) in subjective elements is higher than expected. The first row gives information for all words and the second gives information for words that appear just once. The figures in the *Num* columns are total counts, and the figures in the *P* columns give the proportion that appear in subjective elements. The *Agree* columns give in-

		All Words	Nouns	Verbs	Modals	Adj's	Adverbs	Det's
Exp1	Full Match	0.4216	0.4228	0.2933	0.1422	0.5919	0.1207	0.5000
	Partial Match	0.5156	0.4570	0.4447	0.3011	0.6607	0.3305	0.5000
Exp2	Full Match	0.3041	0.2353	0.2765	0.1429	0.5794	0.1207	0.0000
	Partial Match	0.4209	0.2353	0.3994	0.3494	0.6719	0.4439	0.1429

Table 3: κ Values for Type Agreement Using All Types in the WSJ Data

		All Words	Nouns	Verbs	Modals	Adj's	Adverbs	Det's
Exp1	Full Match	0.3377	0.0440	0.1648	0.1968	0.5443	0.3810	0.0000
	Partial Match	0.5287	0.1637	0.3765	0.4903	0.8125	0.3810	0.0000
Exp2	Full Match	0.2569	0.0000	0.1923	0.1509	0.4783	0.1707	0.1429
	Partial Match	0.4789	0.0000	0.4167	0.4000	0.8056	0.7671	0.4000

Table 4: κ Values for Type Agreement Using E,O,U in the WSJ Data

			Sym.	M.H.	Q.S.	Q.I.
Exp1	All Types	G^2	112.351	92.447	19.904	66.771
		Sig.	0.000	0.000	0.527	0.007
	e,o,u	G^2	85.478	84.142	1.336	12.576
		Sig.	0.000	0.000	0.248	0.027
Exp2	All Types	G^2	94.669	76.247	18.422	58.892
		Sig.	0.000	0.000	0.241	0.001
	e,o,u	G^2	66.822	66.819	0.003	0.0003
		Sig.	0.000	0.000	0.986	0.987

Table 5: Tests for Patterns of Agreement in WSJ Type-Tagged Data

	WSJ-SE						NG-FE					
	D		M		Agree		Agree		R		MM	
	Num	P	Num	P	Num	P	Num	P	Num	P	Num	P
All words	18341	.07	18341	.08	16857	.04	15413	.15	86279	.01	88210	.02
unique	2615	.14	2615	.20	2522	.15	2348	.17	5060	.07	4836	.03

Table 6: Proportions of Unique Words in Subjective Elements

formation for the subset of the corresponding data set upon which the two annotators agree.

Comparison of rows 1 and 2 across columns shows that the proportion of unique words that are subjective is higher than the proportion of all words that are subjective. In all cases, this difference in proportions is highly statistically significant.

6.4 Types and Context

An interesting question is, when a word appears in multiple subjective elements, are those subjective elements all the same type? Table 7 shows that a significant portion are used in more than one type. Each item considered in the table is a word-POS pair that appears more than once in the corpus. The figures shown are the total number of word-POS items that appear more than once (the columns labeled *MultInst*) and the proportion of those items that appear in more than one type of subjective element (the columns labeled *MultType*). These results highlight the need for contextual disambiguation. For example, one thinks of *great* as a positive evaluative term, but its polarity depends on the context; it can be used negatively evaluatively in a context such as "Just great." A goal of performing subjective-element annotations is to support learning such local contextual influences.

7 Generating and Testing PSEs using Document-Level Annotations

This section uses the opinion-piece annotations to expand our set of PSEs beyond those that can be derived from the subjective-element annotations.

Precision is used to assess feature quality. The precision of feature F for class C is the number of Fs that occur in units of class C over the total number of Fs that occur anywhere in the data.

An important motivation for using the opinion-piece data is that there is a large amount of it, and manually refining existing annotations as described in section 5.2 is much easier and more reliable than other types of subjectivity annotation. However, we cannot expect absolutely high precisions for two reasons. First, the distribution of opinions and non-opinions is highly skewed in favor of non-opinions. For example, in Table 1, tagger 1 classifies only 23 of 293 articles as opinion pieces. Second, as discussed in section 4, opinion pieces contain objective sentences and non opinion-pieces contain subjective sentences. For example, in WSJ-SE, which has been annotated at the sentence and document levels, 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.

To give an idea of expected precisions, let us consider the precision of subjective sentences with respect to opinion pieces. Suppose that 15% of the sentences in the dataset are in opinions, 85% in non-opinions. Let us assume the proportions of subjective and objective sentences in opinion and non-opinion pieces given just above. Let N be the total number of sentences. The desired precision is the number of subjective sentences in opinions over the total number of subjective sentences. It is .22:
$p = .15 * N * .70 / (.15 * N * .70 + .85 * N * .44)$.

In addition, we are assessing PSEs, which are only potentially subjective; many have objective as well as subjective uses.

Thus, even if precisions are much lower than 1, we use increases in precision over a baseline as evidence of promising PSEs. The baseline for comparison is the number of word instances in opinion pieces, divided by the total number of word instances. Table 8 shows the precisions for three types of PSEs. The *freq* columns give total frequencies, and the *+prec* columns show the improvements in precision from the baseline. The baseline precisions are given at the bottom of the table.

As mentioned above, (Wiebe, 2000) showed success automatically identifying adjective PSEs using Lin's method, seeded by a small amount of detailed manual annotations. Desiring to move away from manually annotated data, for this paper the same process is used, but the seed words are all the adjectives (verbs) in the training data. In addition, in the current setting, there are no a priori values to use for parameters C (cluster size) and FT (filtering threshold), as there were in (Wiebe, 2000), and results vary with different parameter settings. Thus, a train-validate-test process is appropriate. In Table 8, the numbers given under, e.g., W9-10, are the results obtained when W9-10 is used as the test set. One of the other datasets, say W9-22, was used as the training set, meaning that all the adjectives (verbs) in that dataset are the seed words, and all filtering was performed using only that data. The seed-filtering process was repeated with different settings of C and FT, producing a different set of adjectives (verbs) for each setting. A third dataset, say W9-33, was used as a validation set, i.e., among all the sets of adjectives generated from the training set, those with good performance on the validation set were selected as

WSJ-SE-M		WSJ-SE-D		NG-SE-M	
MultInst	MultType	MultInst	MultType	MultInst	MultType
413	.17	378	.16	571	.29

Table 7: Word-POS-Types Used in Multiple Types of Subjective Elements

	W9-10		W9-22		W9-33		W9-04	
	freq	+prec	freq	+prec	freq	+prec	freq	+prec
adjectives	373	.21	1340	.11	2137	.09	2537	.14
verbs	721	.16	1436	.08	3139	.07	3720	.11
unique words	6065	.10	5441	.07	6045	.06	6171	.09
baseline precision		.17		.13		.14		.18

freq: Total frequency **+prec:** Increase in precision over baseline

Table 8: Frequencies and Increases in Precision

the PSEs to test on the test set. A set was considered to have good performance on the validation set if its precision is at least .25 and its frequency is at least 100. Since this process is meant to be a method for mining existing document-level annotations for PSEs, the existing opinion-piece annotations were used for training and validation. Our manual opinion-piece annotations were used for testing.

The row labeled *unique words* shows the precision on the test set of the individual words that are unique in the test set. The increase over baseline precision shows that low-frequency words can be informative for recognizing subjectivity.

Note that the features all do better and worse on the same data sets. This shows that the subjectivity is somehow harder to identify in, say, W9-33 than in W9-10; it also shows an important consistency among the features, even though they are identified in different ways.

8 Conclusions

This paper presents the results of an empirical examination of subjectivity at the different levels of a text: the expression level, the sentence level, and the document level. While analysis of subjectivity is perhaps most natural and precise at the expression level, document-level annotations are freely available from a number of sources and are appropriate for many applications. The sentence-level annotation is a workable intermediate level: sentence-level judgments are not as fine-grained as expression-level judgments, and they don't involve the large amount of noise found at the document level.

As part of this examination, we present a study of annotator agreement characterizing the difficulty of identifying subjectivity at the different levels of a text. The results demonstrate that not only can subjectivity be identified at the document level with high reliability, but that it is also possible to identify expression-level subjectivity, albeit with lower reliability.

Using manual annotations, we are able to characterize subjective language. At the expression level, we found that it is natural to distinguish among positively evaluative, negatively evaluative, and speculative uses of a word. We also found that subjective text contains a high proportion of unique word occurrences, much more so than ordinary text. Rather than ignoring or discarding unique words, we demonstrate that the occurrence of a unique word is a PSE. We also found that agreement is higher for some syntactic word classes, e.g., for adjectives in comparison with determiners.

Finally, we are able to mine PSEs from text tagged at the document level. Given the difficulty of evaluating PSEs in document-level subjectivity classification due to the mix of subjective and objective sentences, the PSEs identified in this study exhibit relatively high precision. In future work, we will investigate document-level classification using these PSEs, as well as other methods for extracting PSEs from text tagged at the document level; methods to be investigated include mutual-bootstrapping and/or co-training.

References

C. Aone, M. Ramos-Santacruz, and W. Niehaus. 2000. Assentor: An nlp-based solution to e-mail monitoring. In *Proc. IAAI-2000*, pages 945–950.

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

R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. 2000. The rules behind roles:

Identifying speaker role in radio broadcasts. In *Proc. AAAI*.

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

R. Bruce and J. Wiebe. 1999. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5(2).

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Meas.*, 20:37–46.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer errorrates using the EM algorithm. *Applied Statistics*, 28:20–28.

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

L. Goodman. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:2:215–231.

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

D. Karp, Y. Schabes, M. Zaidel, and D. Egedi. 1992. A freely available wide coverage morphological analyzer for English. In *Proc. of the 14th International Conference on Computational Linguistics (COLING-92)*.

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

B. Kessler, G. Nunberg, and H. Schutze. 1997. Automatic detection of text genre. In *Proc. ACL-EACL-97*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. COLING-ACL '98*, pages 768–773.

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

D. Marcu, M. Romera, and E. Amorrortu. 1999. Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. In *The Workshop on Levels of Representation in Discourse*, pages 71–78.

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

G. Nunberg, I. Sag, and T. Wasow. 1994. Idioms. *Language*, 70:491–538.

W. Sack. 1995. Representing and recognizing point of view. In *Proc. AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*.

E. Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Proc. IAAI*.

D. Stein and S. Wright, editors. 1995. *Subjectivity and Subjectivisation*. Cambridge University Press, Cambridge.

L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter. 1997. Building task-specific interfaces to high volume conversational data. In *Proc. CHI 97*, pages 226–233.

S. Teufel and M. Moens. 2000. What's yours and what's mine: Determining intellectual attribution in scientific texts. In *Proc. Joint SIGDAT Converence on EMNLP and VLC*.

J. Wiebe, K. McKeever, and R. Bruce. 1998. Mapping collocational properties into machine learning features. In *Proc. 6th Workshop on Very Large Corpora (WVLC-98)*, pages 225–233, Montreal, Canada, August. ACL SIGDAT.

J. Wiebe, R. Bruce, and T. O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99)*, pages 246–253, University of Maryland, June. ACL.

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

J. Wiebe. 2000. Learning subjective adjectives from corpora. In *17th National Conference on Artificial Intelligence (AAAI-2000)*.