Learning Subjective Nouns using Extraction Pattern Bootstrapping*

2003 Conference on Natural Language Learning (CoNLL-03), ACL SIGNLL.

Ellen Riloff
School of Computing
University of Utah
Salt Lake City, UT 84112
riloff@cs.utah.edu

Janyce Wiebe and Theresa Wilson
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
wiebe,twilson@cs.pitt.edu

Abstract

We explore the idea of creating a subjectivity classifier that uses lists of subjective nouns learned by bootstrapping algorithms. The goal of our research is to develop a system that can distinguish subjective sentences from objective sentences. First, we use two bootstrapping algorithms that exploit extraction patterns to learn sets of subjective nouns. Then we train a Naive Bayes classifier using the subjective nouns, discourse features, and subjectivity clues identified in prior research. The bootstrapping algorithms learned over 1000 subjective nouns, and the subjectivity classifier performed well, achieving 77% recall with 81% precision.

1 Introduction

Many natural language processing applications could benefit from being able to distinguish between factual and subjective information. Subjective remarks come in a variety of forms, including opinions, rants, allegations, accusations, suspicions, and speculation. Ideally, 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. Multi-document summarization systems need to summarize different opinions and perspectives. Spam filtering systems must recognize rants and emotional tirades, among other things. In general, nearly any system that seeks to identify information could benefit from being able to separate factual and subjective information.

Subjective language has been previously studied in fields such as linguistics, literary theory, psychology, and content analysis. Some manually-developed knowledge resources exist, but there is no comprehensive dictionary of subjective language.

Meta-Bootstrapping (Riloff and Jones, 1999) and Basilisk (Thelen and Riloff, 2002) are bootstrapping algorithms that use automatically generated extraction patterns to identify words belonging to a semantic category. We hypothesized that extraction patterns could also identify subjective words. For example, the pattern *"expressed <direct_object>"* often extracts subjective nouns, such as "concern", "hope", and "support". Furthermore, these bootstrapping algorithms require only a handful of seed words and unannotated texts for training; no annotated data is needed at all.

In this paper, we use the Meta-Bootstrapping and Basilisk algorithms to learn lists of subjective nouns from a large collection of unannotated texts. Then we train a subjectivity classifier on a small set of annotated data, using the subjective nouns as features along with some other previously identified subjectivity features. Our experimental results show that the subjectivity classifier performs well (77% recall with 81% precision) and that the learned nouns improve upon previous state-of-the-art subjectivity results (Wiebe et al., 1999).

2 Subjectivity Data

2.1 The Annotation Scheme

In 2002, an annotation scheme was developed for a U.S. government-sponsored project with a team of 10 researchers (the annotation instructions and project reports are available on the Web at http://www.cs.pitt.edu/˜wiebe/pubs/ardasummer02/).

*This work was supported in part by the National Science Foundation under grants IIS-0208798 and IRI-9704240. The data preparation was performed in support of the Northeast Regional Reseach Center (NRRC) which is sponsored by the Advanced Research and Development Activity (ARDA), a U.S. Government entity which sponsors and promotes research of import to the Intelligence Community which includes but is not limited to the CIA, DIA, NSA, NIMA, and NRO.

The scheme was inspired by work in linguistics and literary theory on *subjectivity*, which focuses on how opinions, emotions, etc. are expressed linguistically in context (Banfield, 1982). The scheme is more detailed and comprehensive than previous ones. We mention only those aspects of the annotation scheme relevant to this paper.

The goal of the annotation scheme is to identify and characterize expressions of *private states* in a sentence. *Private state* is a general covering term for opinions, evaluations, emotions, and speculations (Quirk et al., 1985). For example, in sentence (1) the writer is expressing a negative evaluation.

(1) *"The time has come, gentlemen, for Sharon, the assassin, to realize that injustice cannot last long."*

Sentence (2) reflects the private state of Western countries. Mugabe's use of "overwhelmingly" also reflects a private state, his positive reaction to and characterization of his victory.

(2) *"Western countries were left frustrated and impotent after Robert Mugabe formally declared that he had overwhelmingly won Zimbabwe's presidential election."*

Annotators are also asked to judge the strength of each private state. A private state can have *low*, *medium*, *high* or *extreme* strength.

2.2 Corpus and Agreement Results

Our data consists of English-language versions of foreign news documents from FBIS, the U.S. Foreign Broadcast Information Service. The data is from a variety of publications and countries. The annotated corpus used to train and test our subjectivity classifiers (the *experiment corpus*) consists of 109 documents with a total of 2197 sentences. We used a separate, annotated *tuning corpus* of 33 documents with a total of 698 sentences to establish some experimental parameters.[1]

Each document was annotated by one or both of two annotators, A and T. To allow us to measure interannotator agreement, the annotators independently annotated the same 12 documents with a total of 178 sentences. We began with a strict measure of agreement at the sentence level by first considering whether the annotator marked any private-state expression, of any strength, anywhere in the sentence. If so, the sentence should be subjective. Otherwise, it is objective. Table 1 shows the contingency table. The percentage agreement is 88%, and the κ value is 0.71.

[1]The annotated data will be available to U.S. government contractors this summer. We are working to resolve copyright issues to make it available to the wider research community.

		Tagger T: *Subj*	Tagger T: *Obj*
Tagger A	*Subj*	$n_{yy} = 112$	$n_{yn} = 16$
	Obj	$n_{ny} = 6$	$n_{nn} = 44$

Table 1: Agreement for sentence-level annotations

		Tagger T: *Subj*	Tagger T: *Obj*
Tagger A	*Subj*	$n_{yy} = 106$	$n_{yn} = 9$
	Obj	$n_{ny} = 0$	$n_{nn} = 44$

Table 2: Agreement for sentence-level annotations, low strength cases removed

One would expect that there are clear cases of objective sentences, clear cases of subjective sentences, and borderline sentences in between. The agreement study supports this. In terms of our annotations, we define a sentence as borderline if it has at least one private-state expression identified by at least one annotator, and all strength ratings of private-state expressions are *low*. Table 2 shows the agreement results when such borderline sentences are removed (19 sentences, or 11% of the agreement test corpus). The percentage agreement increases to 94% and the κ value increases to 0.87.

As expected, the majority of disagreement cases involve low-strength subjectivity. The annotators consistently agree about which are the clear cases of subjective sentences. This leads us to define the gold-standard that we use in our experiments. A sentence is *subjective* if it contains at least one private-state expression of medium or higher strength. The second class, which we call *objective*, consists of everything else. Thus, sentences with only mild traces of subjectivity are tossed into the objective category, making the system's goal to find the clearly subjective sentences.

3 Using Extraction Patterns to Learn Subjective Nouns

In the last few years, two bootstrapping algorithms have been developed to create semantic dictionaries by exploiting extraction patterns: Meta-Bootstrapping (Riloff and Jones, 1999) and Basilisk (Thelen and Riloff, 2002). *Extraction patterns* were originally developed for information extraction tasks (Cardie, 1997). They represent lexico-syntactic expressions that typically rely on shallow parsing and syntactic role assignment. For example, the pattern "<subject> was hired" would apply to sentences that contain the verb "hired" in the passive voice. The subject would be extracted as the hiree.

Meta-Bootstrapping and Basilisk were designed to learn words that belong to a semantic category (e.g., "truck" is a VEHICLE and "seashore" is a LOCATION). Both algorithms begin with unannotated texts and *seed* words that represent a semantic category. A bootstrapping process looks for words that appear in the same extraction patterns as the seeds and hypothesizes that those words belong to the same semantic class. The principle behind this approach is that words of the same semantic class appear in similar pattern contexts. For example, the phrases "lived in" and "traveled to" will co-occur with many noun phrases that represent LOCATIONS.

In our research, we want to automatically identify words that are subjective. Subjective terms have many different semantic meanings, but we believe that the same contextual principle applies to subjectivity. In this section, we briefly overview these bootstrapping algorithms and explain how we used them to generate lists of subjective nouns.

3.1 Meta-Bootstrapping

The Meta-Bootstrapping ("MetaBoot") process (Riloff and Jones, 1999) begins with a small set of seed words that represent a targeted semantic category (e.g., 10 words that represent LOCATIONS) and an unannotated corpus. First, MetaBoot automatically creates a set of extraction patterns for the corpus by applying and instantiating syntactic templates. This process literally produces thousands of extraction patterns that, collectively, will extract every noun phrase in the corpus. Next, MetaBoot computes a score for each pattern based upon the number of seed words among its extractions. The best pattern is saved and *all* of its extracted noun phrases are automatically labeled as the targeted semantic category.[2] MetaBoot then re-scores the extraction patterns, using the original seed words as well as the newly labeled words, and the process repeats. This procedure is called *mutual bootstrapping*.

A second level of bootstrapping (the "meta-" bootstrapping part) makes the algorithm more robust. When the mutual bootstrapping process is finished, all nouns that were put into the semantic dictionary are re-evaluated. Each noun is assigned a score based on how many different patterns extracted it. Only the five best nouns are allowed to remain in the dictionary. The other entries are discarded, and the mutual bootstrapping process starts over again using the revised semantic dictionary.

3.2 Basilisk

Basilisk (Thelen and Riloff, 2002) is a more recent bootstrapping algorithm that also utilizes extraction patterns to create a semantic dictionary. Similarly, Basilisk begins with an unannotated text corpus and a small set of seed words for a semantic category. The bootstrapping process involves three steps. (1) Basilisk automatically generates a set of extraction patterns for the corpus and scores each pattern based upon the number of seed words among its extractions. This step is identical to the first step of Meta-Bootstrapping. Basilisk then puts the best patterns into a Pattern Pool. (2) All nouns[3] extracted by a pattern in the Pattern Pool are put into a Candidate Word Pool. Basilisk scores each noun based upon the set of patterns that extracted it and their collective association with the seed words. (3) The top 10 nouns are labeled as the targeted semantic class and are added to the dictionary. The bootstrapping process then repeats, using the original seeds and the newly labeled words.

The main difference between Basilisk and Meta-Bootstrapping is that Basilisk scores each noun based on *collective* information gathered from *all* patterns that extracted it. In contrast, Meta-Bootstrapping identifies a single best pattern and assumes that everything it extracted belongs to the same semantic class. The second level of bootstrapping smoothes over some of the problems caused by this assumption. In comparative experiments (Thelen and Riloff, 2002), Basilisk outperformed Meta-Bootstrapping. But since our goal of learning subjective nouns is different from the original intent of the algorithms, we tried them both. We also suspected they might learn different words, in which case using both algorithms could be worthwhile.

3.3 Experimental Results

The Meta-Bootstrapping and Basilisk algorithms need seed words and an unannotated text corpus as input. Since we did not need annotated texts, we created a much larger training corpus, the *bootstrapping corpus*, by gathering 950 new texts from the FBIS source mentioned in Section 2.2. To find candidate seed words, we automatically identified 850 nouns that were positively correlated with subjective sentences in another data set. However, it is crucial that the seed words occur frequently in our FBIS texts or the bootstrapping process will not get off the ground. So we searched for each of the 850 nouns in the bootstrapping corpus, sorted them by frequency, and manually selected 20 high-frequency words that we judged to be strongly subjective. Table 3 shows the 20 seed words used for both Meta-Bootstrapping and Basilisk.

We ran each bootstrapping algorithm for 400 iterations, generating 5 words per iteration. Basilisk generated 2000 nouns and Meta-Bootstrapping generated 1996 nouns.[4] Table 4 shows some examples of extraction pat-

[2] Our implementation of Meta-Bootstrapping learns individual nouns (vs. noun phrases) and discards capitalized words.

[3] Technically, each head noun of an extracted noun phrase.

[4] Meta-Bootstrapping will sometimes produce fewer than 5 words per iteration if it has low confidence in its judgements.

cowardice	embarrassment	hatred	outrage
crap	fool	hell	slander
delight	gloom	hypocrisy	sigh
disdain	grievance	love	twit
dismay	happiness	nonsense	virtue

Table 3: Subjective Seed Words

Extraction Patterns	Examples of Extracted Nouns
expressed <dobj>	condolences, hope, grief, views, worries, recognition
indicative of <np>	compromise, desire, thinking
inject <dobj>	vitality, hatred
reaffirmed <dobj>	resolve, position, commitment
voiced <dobj>	outrage, support, skepticism, disagreement, opposition, concerns, gratitude, indignation
show of <np>	support, strength, goodwill, solidarity, feeling
<subject> was shared	anxiety, view, niceties, feeling

Table 4: Extraction Pattern Examples

terns that were discovered to be associated with subjective nouns.

Meta-Bootstrapping and Basilisk are semi-automatic lexicon generation tools because, although the bootstrapping process is 100% automatic, the resulting lexicons need to be reviewed by a human.[5] So we manually reviewed the 3996 words proposed by the algorithms. This process is very fast; it takes only a few seconds to classify each word. The entire review process took approximately 3-4 hours. One author did this labeling; this person did not look at or run tests on the experiment corpus.

Strong Subjective		Weak Subjective	
tyranny	scum	aberration	plague
smokescreen	bully	allusion	risk
apologist	devil	apprehensions	drama
barbarian	liar	beneficiary	trick
belligerence	pariah	resistant	promise
condemnation	venom	credence	intrigue
sanctimonious	diatribe	distortion	unity
exaggeration	mockery	eyebrows	failures
repudiation	anguish	inclination	tolerance
insinuation	fallacies	liability	persistent
antagonism	evil	assault	trust
atrocities	genius	benefit	success
denunciation	goodwill	blood	spirit
exploitation	injustice	controversy	slump
humiliation	innuendo	likelihood	sincerity
ill-treatment	revenge	peaceful	eternity
sympathy	rogue	pressure	rejection

Table 5: Examples of Learned Subjective Nouns

[5]This is because NLP systems expect dictionaries to have high integrity. Even if the algorithms could achieve 90% accuracy, a dictionary in which 1 of every 10 words is defined incorrectly would probably not be desirable.

	B	M	B ∩ M	B ∪ M
StrongSubj	372	192	110	454
WeakSubj	453	330	185	598
Total	825	522	295	1052

Table 6: Subjective Word Lexicons after Manual Review (B=Basilisk, M=MetaBootstrapping)

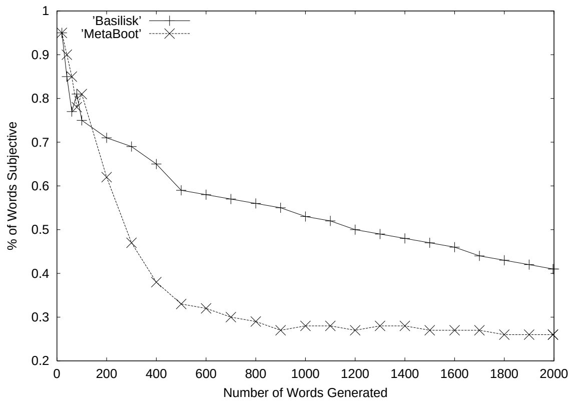

Figure 1: Accuracy during Bootstrapping

We classified the words as *StrongSubjective*, *WeakSubjective*, or *Objective*. *Objective* terms are not subjective at all (e.g., “chair” or “city”). *StrongSubjective* terms have strong, unambiguously subjective connotations, such as “bully” or “barbarian”. *WeakSubjective* was used for three situations: (1) words that have weak subjective connotations, such as “aberration” which implies something out of the ordinary but does not evoke a strong sense of judgement, (2) words that have multiple senses or uses, where one is subjective but the other is not. For example, the word “plague” can refer to a disease (objective) or an onslaught of something negative (subjective), (3) words that are objective by themselves but appear in idiomatic expressions that are subjective. For example, the word “eyebrows” was labeled *WeakSubjective* because the expression “raised eyebrows” probably occurs more often in our corpus than literal references to “eyebrows”. Table 5 shows examples of learned words that were classified as *StrongSubjective* or *WeakSubjective*.

Once the words had been manually classified, we could go back and measure the effectiveness of the algorithms. The graph in Figure 1 tracks their accuracy as the bootstrapping progressed. The X-axis shows the number of words generated so far. The Y-axis shows the percentage of those words that were manually classified as subjective. As is typical of bootstrapping algorithms, accuracy was high during the initial iterations but tapered off as the bootstrapping continued. After 20 words, both algorithms were 95% accurate. After 100 words Basilisk was 75% accurate and MetaBoot was 81% accurate. After 1000 words, accuracy dropped to about 28% for MetaBoot, but Basilisk was still performing reasonably well at 53%. Although 53% accuracy is not high for a fully automatic process, Basilisk depends on a human to review the words so 53% accuracy means that the human is accepting every other word, on average. Thus, the reviewer's time was still being spent productively even after 1000 words had been hypothesized.

Table 6 shows the size of the final lexicons created by the bootstrapping algorithms. The first two columns show the number of subjective terms learned by Basilisk and Meta-Bootstrapping. Basilisk was more prolific, generating 825 subjective terms compared to 522 for Meta-Bootstrapping. The third column shows the intersection between their word lists. There was substantial overlap, but both algorithms produced many words that the other did not. The last column shows the results of merging their lists. In total, the bootstrapping algorithms produced 1052 subjective nouns.

4 Creating Subjectivity Classifiers

To evaluate the subjective nouns, we trained a Naive Bayes classifier using the nouns as features. We also incorporated previously established subjectivity clues, and added some new discourse features. In this section, we describe all the feature sets and present performance results for subjectivity classifiers trained on different combinations of these features. The threshold values and feature representations used in this section are the ones that produced the best results on our separate tuning corpus.

4.1 Subjective Noun Features

We defined four features to represent the sets of subjective nouns produced by the bootstrapping algorithms.

BA-Strong: the set of *StrongSubjective* nouns generated by Basilisk

BA-Weak: the set of *WeakSubjective* nouns generated by Basilisk

MB-Strong: the set of *StrongSubjective* nouns generated by Meta-Bootstrapping

MB-Weak: the set of *WeakSubjective* nouns generated by Meta-Bootstrapping

For each set, we created a three-valued feature based on the presence of 0, 1, or ≥ 2 words from that set. We used the nouns as feature sets, rather than define a separate feature for each word, so the classifier could generalize over the set to minimize sparse data problems. We will refer to these as the **SubjNoun** features.

4.2 Previously Established Features

Wiebe, Bruce, & O'Hara (1999) developed a machine learning system to classify subjective sentences. We experimented with the features that they used, both to compare their results to ours and to see if we could benefit from their features. We will refer to these as the **WBO** features.

WBO includes a set of stems positively correlated with the subjective training examples ($subjStems$) and a set of stems positively correlated with the objective training examples ($objStems$). We defined a three-valued feature for the presence of 0, 1, or ≥ 2 members of $subjStems$ in a sentence, and likewise for $objStems$. For our experiments, $subjStems$ includes stems that appear ≥ 7 times in the training set, and for which the precision is 1.25 times the baseline word precision for that training set. $objStems$ contains the stems that appear ≥ 7 times and for which at least 50% of their occurrences in the training set are in objective sentences. WBO also include a binary feature for each of the following: the presence in the sentence of a pronoun, an adjective, a cardinal number, a modal other than *will*, and an adverb other than *not*.

We also added manually-developed features found by other researchers. We created 14 feature sets representing some classes from (Levin, 1993; Ballmer and Brennenstuhl, 1981), some Framenet lemmas with frame element *experiencer* (Baker et al., 1998), adjectives manually annotated for polarity (Hatzivassiloglou and McKeown, 1997), and some subjectivity clues listed in (Wiebe, 1990). We represented each set as a three-valued feature based on the presence of 0, 1, or ≥ 2 members of the set. We will refer to these as the **manual** features.

4.3 Discourse Features

We created **discourse** features to capture the density of clues in the text surrounding a sentence. First, we computed the average number of subjective clues and objective clues per sentence, normalized by sentence length. The subjective clues, $subjClues$, are all sets for which 3-valued features were defined above (except $objStems$). The objective clues consist only of $objStems$. For sentence S, let $ClueRate_{subj}(S) = \frac{|subjClues\ in\ S|}{|S|}$ and $ClueRate_{obj}(S) = \frac{|objStems\ in\ S|}{|S|}$. Then we define $AvgClueRate_{subj}$ to be the average of $ClueRate(S)$ over all sentences S and similarly for $AvgClueRate_{obj}$. Next, we characterize the number of subjective and objective clues in the previous and next sentences as: higher-than-expected (*high*), lower-than-expected (*low*), or expected (*medium*). The value for $ClueRate_{subj}(S)$ is *high* if $ClueRate_{subj}(S) \geq AvgClueRate_{subj} * 1.3$; *low* if $ClueRate_{subj}(S) \leq AvgClueRate_{subj}/1.3$; otherwise it is *medium*. The values for $ClueRate_{obj}(S)$ are defined similarly.

Using these definitions we created four features: $ClueRate_{subj}$ for the previous and following sentences, and $ClueRate_{obj}$ for the previous and following sentences. We also defined a feature for sentence length. Let $AvgSentLen$ be the average sentence length. $SentLen(S)$ is *high* if $length(S) \geq AvgSentLen * 1.3$; *low* if $length(S) \leq AvgSentLen/1.3$; and *medium* otherwise.

4.4 Classification Results

We conducted experiments to evaluate the performance of the feature sets, both individually and in various combinations. Unless otherwise noted, all experiments involved training a Naive Bayes classifier using a particular set of features. We evaluated each classifier using 25-fold cross validation on the experiment corpus and used paired t-tests to measure significance at the 95% confidence level. As our evaluation metrics, we computed accuracy (Acc) as the percentage of the system's classifications that match the gold-standard, and precision (Prec) and recall (Rec) with respect to subjective sentences.

	Acc	Prec	Rec
(1) Bag-Of-Words	73.3	81.7	70.9
(2) WBO	72.1	76.0	77.4
(3) Most-Frequent	59.0	59.0	100.0

Table 7: Baselines for Comparison

Table 7 shows three baseline experiments. Row (3) represents the common baseline of assigning every sentence to the most frequent class. The Most-Frequent baseline achieves 59% accuracy because 59% of the sentences in the gold-standard are subjective. Row (2) is a Naive Bayes classifier that uses the **WBO** features, which performed well in prior research on sentence-level subjectivity classification (Wiebe et al., 1999). Row (1) shows a Naive Bayes classifier that uses unigram bag-of-words features, with one binary feature for the absence or presence in the sentence of each word that appeared during training. Pang et al. (2002) reported that a similar experiment produced their best results on a related classification task. The difference in accuracy between Rows (1) and (2) is not statistically significant (Bag-of-Word's higher precision is balanced by WBO's higher recall).

Next, we trained a Naive Bayes classifier using only the **SubjNoun** features. This classifier achieved good precision (77%) but only moderate recall (64%). Upon further inspection, we discovered that the subjective nouns are good subjectivity indicators when they appear, but not every subjective sentence contains one of them. And, relatively few sentences contain more than one, making it difficult to recognize contextual effects (i.e., multiple clues in a region). We concluded that the appropriate way to benefit from the subjective nouns is to use them in tandem with other subjectivity clues.

	Acc	Prec	Rec	
(1)	76.1	81.3	77.4	WBO+SubjNoun+ manual+discourse
(2)	74.3	78.6	77.8	WBO+SubjNoun
(3)	72.1	76.0	77.4	WBO

Table 8: Results with New Features

Table 8 shows the results of Naive Bayes classifiers trained with different combinations of features. The accuracy differences between all pairs of experiments in Table 8 are statistically significant. Row (3) uses only the **WBO** features (also shown in Table 7 as a baseline). Row (2) uses the **WBO** features as well as the **SubjNoun** features. There is a synergy between these feature sets: using both types of features achieves better performance than either one alone. The difference is mainly precision, presumably because the classifier found more and better combinations of features. In Row (1), we also added the **manual** and **discourse** features. The **discourse** features explicitly identify contexts in which multiple clues are found. This classifier produced even better performance, achieving 81.3% precision with 77.4% recall. The 76.1% accuracy result is significantly higher than the accuracy results for all of the other classifiers (in both Table 8 and Table 7).

Finally, higher precision classification can be obtained by simply classifying a sentence as subjective if it contains any of the *StrongSubjective* nouns. On our data, this method produces 87% precision with 26% recall. This approach could support applications for which precision is paramount.

5 Related Work

Several types of research have involved document-level subjectivity classification. Some work identifies inflammatory texts (e.g., (Spertus, 1997)) or classifies reviews as positive or negative ((Turney, 2002; Pang et al., 2002)). Tong's system (Tong, 2001) generates *sentiment timelines*, tracking online discussions and creating graphs of positive and negative opinion messages over time. Research in genre classification may include recognition of subjective genres such as editorials (e.g., (Karlgren and Cutting, 1994; Kessler et al., 1997; Wiebe et al., 2001)). In contrast, our work classifies individual sentences, as does the research in (Wiebe et al., 1999). Sentence-level subjectivity classification is useful because most documents contain a mix of subjective and objective sentences. For example, newspaper articles are typically thought to be relatively objective, but (Wiebe et al., 2001) reported that, in their corpus, 44% of sentences (in articles that are not editorials or reviews) were subjective.

Some previous work has focused explicitly on learning subjective words and phrases. (Hatzivassiloglou and McKeown, 1997) describes a method for identifying the *semantic orientation* of words, for example that *beautiful* expresses positive sentiments. Researchers have focused on learning adjectives or adjectival phrases (Turney, 2002; Hatzivassiloglou and McKeown, 1997; Wiebe, 2000) and verbs (Wiebe et al., 2001), but no previous work has focused on learning nouns. A unique aspect of our work is the use of bootstrapping methods that exploit extraction patterns. (Turney, 2002) used patterns representing part-of-speech sequences, (Hatzivassiloglou and McKeown, 1997) recognized adjectival phrases, and (Wiebe et al., 2001) learned N-grams. The extraction patterns used in our research are linguistically richer patterns, requiring shallow parsing and syntactic role assignment.

In recent years several techniques have been developed for semantic lexicon creation (e.g., (Hearst, 1992; Riloff and Shepherd, 1997; Roark and Charniak, 1998; Caraballo, 1999)). Semantic word learning is different from subjective word learning, but we have shown that Meta-Bootstrapping and Basilisk could be successfully applied to subjectivity learning. Perhaps some of these other methods could also be used to learn subjective words.

6 Conclusions

This research produced interesting insights as well as performance results. First, we demonstrated that weakly supervised bootstrapping techniques can learn subjective terms from unannotated texts. Subjective features learned from unannotated documents can augment or enhance features learned from annotated training data using more traditional supervised learning techniques. Second, Basilisk and Meta-Bootstrapping proved to be useful for a different task than they were originally intended. By seeding the algorithms with subjective words, the extraction patterns identified expressions that are associated with subjective nouns. This suggests that the bootstrapping algorithms should be able to learn not only general semantic categories, but any category for which words appear in similar linguistic phrases. Third, our best subjectivity classifier used a wide variety of features. Subjectivity is a complex linguistic phenomenon and our evidence suggests that reliable subjectivity classification requires a broad array of features.

7 Acknowledgements

This work was supported in part by the National Science Foundation under grants IIS-0208798 and IRI-9704240. The data preparation was performed in support of the Northeast Regional Reseach Center (NRRC) which is sponsored by the Advanced Research and Development Activity (ARDA), a U.S. Government entity which sponsors and promotes research of import to the Intelligence Community which includes but is not limited to the CIA, DIA, NSA, NIMA, and NRO.

References

C. Baker, C. Fillmore, and J. Lowe. 1998. The berkeley framenet project. In *Proceedings of the COLING-ACL*.

T. Ballmer and W. Brennenstuhl. 1981. *Speech Act Classification: A Study in the Lexical Analysis of English Speech Activity Verbs*. Springer-Verlag.

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

S. Caraballo. 1999. Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 120–126.

C. Cardie. 1997. Empirical Methods in Information Extraction. *AI Magazine*, 18(4):65–79.

V. Hatzivassiloglou and K. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL-EACL 1997*.

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics*.

J. Karlgren and D. Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING-94*.

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

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, New York.

E. Riloff and R. Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence*.

E. Riloff and J. Shepherd. 1997. A Corpus-Based Approach for Building Semantic Lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124.

B. Roark and E. Charniak. 1998. Noun-phrase Co-occurrence Statistics for Semi-automatic Semantic Lexicon Construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1110–1116.

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

M. Thelen and E. Riloff. 2002. A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pa ttern Contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

R. Tong. 2001. An operational system for detecting and tracking opinions in on-line discussion. In *SIGIR 2001 Workshop on Operational Text Classification*.

P. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

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)*.

J. Wiebe, T. Wilson, and M. Bell. 2001. Identifying collocations for recognizing opinions. In *Proc. ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, July.

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

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