Predicting Subjectivity in Multimodal Conversations

Gabriel Murray and **Giuseppe Carenini**
University of British Columbia
Vancouver, Canada
(gabrielm, carenini)@cs.ubc.ca

Abstract

In this research we aim to detect subjective sentences in multimodal conversations. We introduce a novel technique wherein subjective patterns are learned from both labeled and unlabeled data, using n-gram word sequences with varying levels of lexical instantiation. Applying this technique to meeting speech and email conversations, we gain significant improvement over state-of-the-art approaches. Furthermore, we show that coupling the pattern-based approach with features that capture characteristics of general conversation structure yields additional improvement.

1 Introduction

Conversations are rich in subjectivity. Conversation participants agree and disagree with one other, argue for and against various proposals, and generally take turns expressing their private states. Being able to separate these subjective utterances from more objective utterances would greatly facilitate the analysis, mining and summarization of a large number of conversations.

Two of the most prevalent conversational media are meetings and emails. Face-to-face meetings enable numerous people to exchange a large amount of information and opinions in a short period of time, while emails allow for concise exchanges between potentially far-flung participants. Meetings and emails can also feed into one another, with face-to-face meetings occurring at regular intervals and emails continuing the conversations in the interim. This poses several interesting questions, such as whether subjective utterances are more or less likely to be found in email exchanges compared with meetings, and whether the ratios of positive and negative subjective utterances differ between the two modalities.

In this paper we describe a novel approach for predicting subjectivity, and test it in two sets of experiments on meetings and emails. Our approach combines a new general purpose method for learning subjective patterns, with features that capture basic characteristics of conversation structure across modalities. The subjective patterns are essentially n-gram sequences with varying levels of lexical instantiation, and we demonstrate how they can be learned from both labeled and unlabeled data. The conversation features capture structural characteristics of multimodal conversations as well as participant information.

We test our approach in two sets of experiments. The goal of the first set of experiments is to discriminate subjective from non-subjective utterances, comparing the novel approach to existing state-of-the-art techniques. In the second set of experiments, the goal is to discriminate positive-subjective and negative-subjective utterances, establishing their polarity. In both sets of experiments, we assess the impact of features relating to conversation structure.

2 Related Research

Raaijmakers et al. (2008) have approached the problem of detecting subjectivity in meeting speech by using a variety of multimodal features such as prosodic features, word n-grams, character n-grams and phoneme n-grams. For subjectivity detection, they found that a combination of all features was best, while prosodic features were less useful for discriminating between positive and negative utterances. They found character n-grams to be particularly useful.

Riloff and Wiebe (2004) presented a method for learning subjective extraction patterns from a large amount of data, which takes subjective and non-subjective text as input, and outputs significant lexico-syntactic patterns. These patterns are based on syntactic structure output by the Sundance shal-

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1348–1357,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

low dependency parser (Riloff and Phillips, 2004). They are extracted by exhaustively applying syntactic templates such as $< subj >$ *passive-verb* and *active-verb* $< dobj >$ to a training corpus, with an extracted pattern for every instantiation of the syntactic template. These patterns are scored according to probability of relevance given the pattern and frequency of the pattern. Because these patterns are based on syntactic structure, they can represent subjective expressions that are not fixed word sequences and would therefore be missed by a simple n-gram approach.

Riloff et al. (2006) explore feature subsumption for opinion detection, where a given feature may subsume another feature representationally if the strings matched by the first feature include all of the strings matched by the second feature. To give their own example, the unigram *happy* subsumes the bigram *very happy*. The first feature will *behaviorally* subsume the second if it representationally subsumes the second and has roughly the same information gain, within an acceptable margin. They show that they can improve opinion analysis results by modeling these relations and reducing the feature set.

Our approach for learning subjective patterns like Raaijmakers et al. relies on n-grams, but like Riloff et al. moves beyond fixed sequences of words by varying levels of lexical instantiation.

Yu and Hatzivassiloglou (2003) addressed three challenges in the news article domain: discriminating between objective documents and subjective documents such as editorials, detecting subjectivity at the sentence level, and determining polarity at the sentence level. They found that the latter two tasks were substantially more difficult than classification at the document level. Of particular relevance here is that they found that part-of-speech (POS) features were especially useful for assigning polarity scores, with adjectives, adverbs and verbs comprising the best set of POS tags. This work inspired us to look at generalization of n-grams based on POS.

On the slightly different task of classifying the intensity of opinions, Wilson et al. (2006) employed several types of features including dependency structures in which words can be backed off to POS tags. They found that this feature class improved the overall accuracy of their system.

Somasundaran et al. (2007) investigated subjectivity classification in meetings. Their findings indicate that both lexical features (list of words and expressions) and discourse features (dialogue acts and adjacency pairs) can be beneficial. In the same spirit, we effectively combine lexical patterns and conversational features.

The approach to predicting subjectivity we present in this paper is a novel contribution to the field of opinion and sentiment analysis. Pang and Lee (2008) give an overview of the state of the art, discussing motivation, features, approaches and available resources.

3 Subjectivity Detection

In this section we describe our approach to subjectivity detection. We begin by describing how to learn subjective n-gram patterns with varying levels of lexical instantiation. We then describe a set of features characterizing multimodal conversation structure which can be used to supplement the n-gram approach. Finally, we describe the baseline subjectivity detection approaches used for comparison.

3.1 Partially Instantiated N-Grams

Our approach to subjectivity detection and polarity detection is to learn significant patterns that correlate with the subjective and polar utterances. These patterns are word trigrams, but with varying levels of lexical instantiation, so that each unit of the n-gram can be either a word or the word's part-of-speech (POS) tag. This contrasts, then, with work such as that of Raaijmakers et al. (2008) who include trigram features in their experiments, but where their learned trigrams are fully instantiated. As an example, while they may learn that a trigram *really great idea* is positive, we may additionally find that *really great NN* and *RB great NN* are informative patterns, and these patterns may sometimes be better cues than the fully instantiated trigrams. To differentiate this approach from the typical use of trigrams, we will refer to it as the VIN (*varying instantiation n-grams*) method.

In some respects, our approach to subjectivity detection is similar to Riloff and Wiebe's work cited above, in the sense that their extraction patterns are partly instantiated. However, the AutoSlog-TS approach relies on deriving syntactic structure with the Sundance shallow parser (Riloff and Phillips, 2004). We hypothesize that our trigram approach may be more robust to disfluent and fragmented meeting speech and emails

1	2	3
really	great	idea
really	great	NN
really	JJ	idea
RB	great	idea
really	JJ	NN
RB	great	NN
RB	JJ	idea
RB	JJ	NN

Table 1: Sample Instantiation Set

on which syntactic parsers may perform poorly. Also, our learned trigram patterns range from fully instantiated to completely uninstantiated. For example, we might find that the pattern *RB JJ NN* is a very good indicator of subjective utterances because it matches a variety of scenarios where people are ascribing qualities to things, e.g. *really bad movie, horribly overcooked steak.* Notice that we do not see our approach and AutoSlog-TS as mutually exclusive, and indeed we demonstrate through these experiments that they can be effectively combined.

Our approach begins by running the Brill POS tagger (Brill, 1992) over all sentences in a document. We then extract all of the word trigrams from the document, and represent each trigram using every possible instantiation. Because we are working at the trigram level, and each unit of the trigram can be a word or its POS tag there are $2^3 = 8$ representations in each trigram's instantiation set. To continue the example from above, the instantiation set for the trigram *really great idea* is given in Table 1. As we scan down the instantiation set, we can see that the level of abstraction increases until it is completely uninstantiated. It is this multilevel abstraction that we are hypothesizing will be useful for learning new subjective and polar cues.

All trigrams are then scored according to their prevalence in relevant versus irrelevant documents (e.g. subjective vs. non-subjective sentences), following the scoring methodology of Riloff and Wiebe (2003). We calculate the conditional probability $p(relevance|trigram)$ using the actual trigram counts in relevant and irrelevant text. For learning negative-subjective patterns, we treat all negative sentences as the relevant text and the remainder of the sentences as irrelevant text, and conduct the same process for learning positive-subjective patterns. We consider significant patterns to be those where the conditional probability is greater than 0.65 and the pattern occurs more than five times in the entire document set (slightly higher than $probability >= 0.60$ and $frequency >= 2$ used by Riloff and Wiebe (2003)).

We possess a fairly small amount of conversational data annotated for subjectivity and polarity. The AMI meeting corpus and BC3 email corpus are described in more detail in Section 4.1. To address this shortfall in annotated data, we take two approaches to learning patterns. In the first, we learn a set of patterns from the annotated conversation data. In the second approach, we complement those patterns by learning additional patterns from unannotated data that are typically overwhelmingly subjective or objective in nature. We describe these two approaches here in turn.

3.1.1 Supervised Learning of Patterns from Conversation Data

The first learning strategy is to apply the above-described methods to the annotated conversation data, learning the positive patterns by comparing *positive-subjective* utterances to all other utterances, and learning the negative patterns by comparing the *negative-subjective* utterances to all other utterances, using the described methods. This results in 759 significant positive patterns and 67 significant negative patterns. This difference in pattern numbers can be explained by negative utterances being less common in the AMI meetings, as noted by Wilson (2008). It may be that people are less comfortable in expressing negative sentiments in face-to-face conversations, particularly when the meeting participants do not know each other well (in the AMI scenario meetings, many participants were meeting each other for the first time). But there may be a further explanation for why we learn many more positive than negative patterns. When conversation participants *do* express negative sentiments, they may couch those sentiments in more euphemistic or guarded terms compared with positive sentiments. Table 2 gives examples of significant positive and negative patterns learned from the labeled meeting data. The last two rows in Table 2 show how two patterns in the same instantiation set can have substantially different probabilities.

POS	$p(r\|t)$	NEG	$p(r\|t)$
you MD change	1.0	VBD not RB	1.0
should VBP DT	1.0	doesn't RB VB	0.875
very easy to	0.88	a bit JJ	0.66
we could VBP	0.78	think PRP might	0.66
NNS should VBP	0.71	be DT problem	0.71
PRP could do	0.66	doesn't really VB	0.833
it could VBP	83	doesn't RB VB	0.875

Table 2: Example Pos. and Neg. Patterns (AMI)

Pattern	$p(r\|t)$
can not VB	0.99
i can RB	0.99
i have not	0.98
do RB think	0.97
RB think that	0.95
RB agree with	0.95
IN PRP opinion	0.95

Table 3: Example Subjective Patterns (BLOG06)

3.1.2 Unsupervised Learning of Patterns from Blog Data

The second pattern learning strategy we take to learning subjective patterns is to use a relevant, but unannotated corpus. We focus on weblog (blog) data for several reasons. First, blog posts share many characteristics with both meetings and emails: they are conversational, informal and the language can be very ungrammatical. Second, blog posts are known for being subjective; bloggers post on issues that are passionate to them, offering arguments, opinions and invective. Third, there is a huge amount of available blog data. But because we do not possess blog data annotated for subjectivity, we take the following approach to learning subjective patterns from this data. We work on the assumption that a great many blog posts are inherently subjective, and that comparing this data to inherently *objective* text such as newswire articles, treating the latter as our irrelevant text, should lead to the detection of many new subjective patterns and greatly increase our coverage. While the patterns learned will be noisy, we hypothesize that the increased coverage will improve our subjectivity detection overall.

For our blog data, we use the BLOG06 Corpus[1] that was featured as training and testing data for the Text Analysis Conference (TAC) 2008 track on summarizing blog opinions. The portion used totals approximately 4,000 documents on all manner of topics. Treating that dataset as our relevant, subjective data, we then learn the subjective trigrams by comparing with the *irrelevant* TAC/DUC newswire data from the 2007 and 2008 update summarization tasks. To try to reduce the amount of noise in our learned patterns, we set the conditional probability threshold at 0.75 (vs. 0.65 for annotated data), and stipulate that all significant patterns must occur at least once in the irrelevant text. This last rule is meant to prevent us from learning completely blog-specific patterns such as *posted by NN* or *linked to DT*. In the end, more than 20,000 patterns were learned from the blog data. While manual inspection does show that many undesirable patterns were extracted, among the highest-scoring patterns are many sensible subjective trigrams such as those indicated in Table 3.

This approach is similar in spirit to the work of Biadsy et al. (2008) on unsupervised biography production. Without access to labeled biographical data, the authors chose to use sentences from Wikipedia biographies as their positive set and sentences from newswire articles as their negative set, on the assumption that most of the Wikipedia sentences would be relevant to biographies and most of the newswire sentences would not.

3.2 Deriving VIN Features

For our machine learning experiments, we derive, for each sentence, features indicating the presence of the significant VIN patterns. Patterns are binned according to their conditional probability range (i.e., $0.65 <= p < 0.75$, $0.75 <= p < 0.85$, $0.85 <= p < 0.95$, and $0.95 <= p$). There are three bins for the blog patterns, since the probability cutoff is 0.75 For each bin, there is a feature indicating the count of its patterns in the given sentence. When attempting to match these trigram patterns to sentences, we allow up to two wildcard lexical items between the trigram units. In this way a sentence can match a learned pattern even if the units of the n-gram are not contiguous (Raaijmakers et al. (2008) similarly include an n-gram feature allowing such intervening material).

A key reason for counting the number of matched patterns for each probability range as just described, rather than including a feature for each individual pattern, is to maintain the same level of dimensionality in our machine learning experiments when comparing the VIN approach to the baseline approaches described in Section 3.4.

[1] http://ir.dcs.gla.ac.uk/test_collections/blog06info.html

3.3 Conversational Features

While we hypothesize that the general purpose pattern-based approach described above will greatly aid subjectivity and polarity detection, we also recognize that there are many additional features specific for characterizing multimodal conversations that may correlate well with subjectivity and polarity. Such features include structural characteristics like the position of a sentence in a turn and the position of a turn in the conversation, and participant features relating to dominance or leadership. For example, it may be that subjective sentences are more likely to come at the end of a conversation, or that a person who dominates the conversation may utter more negative sentences.

We use the feature set provided by Murray and Carenini (2008), which they used for automatic summarization of conversations and which are shown in Table 4. Many of the features are based on so-called $Sprob$ and $Tprob$ term-weights, the former of which weights words based on their distributions across conversation participants and the latter of which similarly weights words based on their distributions across conversation turns. Other features include word entropy of the candidate sentence, lexical cohesion of the sentence with the greater conversation, and structural features indicating position of the candidate sentence in the turn and in the conversation, such as the elapsed time since the beginning of the conversation.

3.4 Baseline Approaches

There are two baselines in particular to which we are interested in comparing the VIN approach. As stated earlier, we are hypothesizing that the increasing levels of abstraction found with partially instantiated trigrams will lead to improved classification compared with using only fully instantiated trigrams. To test this, we also run the subjective/non-subjective and positive/negative experiments using *only* fully instantiated trigrams. There are 71 such positive trigrams and 5 such negative trigrams learned from the AMI data. There are just over 1200 fully instantiated trigrams learned from the unannotated BLOG06 data.

Believing that the current approach may offer benefits over state-of-the-art pattern-based subjectivity detection, we also implement the AutoSlog-TS method of Riloff and Wiebe (2003) for extracting subjective extraction patterns. In AutoSlog-TS, once all of the patterns are extracted using the Sundance parser, the scoring methodology is much the same as desribed in Section 3.1. Conditional probabilities are calculated by comparing pattern occurrences in the relevant text with occurrences in all text, and we again use a threshold of $p >= 0.65$ and $frequency >= 5$ for significant patterns. For the BLOG06 data, we use a probability cutoff of 0.75 as before. For deriving the features used in our machine learning experiments, the patterns are similarly grouped according to conditional probability. From the annotated data, 48 patterns are learned in total, 46 positive and only 2 negative. From the BLOG06 data, more than 3000 significant patterns are learned. Among significant patterns learned from the AMI corpus are $< subj >$ *BE good*, *change* $< dobj >$, $< subj >$ *agree* and *problem with* $< NP >$.

Feature ID	Description
MXS	max *Sprob* score
MNS	mean *Sprob* score
SMS	sum of *Sprob* scores
MXT	max *Tprob* score
MNT	mean *Tprob* score
SMT	sum of *Tprob* scores
TLOC	position in turn
CLOC	position in conv.
SLEN	word count, globally normalized
SLEN2	word count, locally normalized
TPOS1	time from beg. of conv. to turn
TPOS2	time from turn to end of conv.
DOM	participant dominance in words
COS1	cosine of conv. splits, w/ *Sprob*
COS2	cosine of conv. splits, w/ *Tprob*
PENT	entropy of conv. up to sentence
SENT	entropy of conv. after the sentence
THISENT	entropy of current sentence
PPAU	time btwn. current and prior turn
SPAU	time btwn. current and next turn
BEGAUTH	is first participant (0/1)
CWS	rough ClueWordScore (cohesion)
CENT1	cos. of sentence & conv., w/ *Sprob*
CENT2	cos. of sentence & conv., w/ *Tprob*

Table 4: Features Key

To gauge the effectiveness of the various feature types, for both sets of experiments we build multiple models on a variety of feature combinations: fully instantiated trigrams (TRIG), varying instantiation n-grams (VIN), AutoSlog-TS (SLOG), conversational structure features (CONV), and the set of all features.

4 Experimental Setup

In this section we describe the corpora used, the relevant subjectivity annotation, and the statistical classifiers employed.

4.1 Corpora

We use two annotated corpora for these experiments. The AMI corpus (Carletta et al., 2005) consists of meetings in which participants take part in role-playing exercises concerning the design and development of a remote control. Participants are grouped in fours, and each group takes part in a sequence of four meetings, bringing the remote control from design to market. The four members of the group are assigned roles of project manager, industrial designer, user interface designer, and marketing expert. In total there are 140 such scenario meetings, with individual meetings ranging from approximately 15 to 45 minutes.

The BC3 corpus (Ulrich et al., 2008) contains email threads from the World Wide Web Consortium (W3C) mailing list. The threads feature a variety of topics such as web accessibility and planning face-to-face meetings. The annotated portion of the mailing list consists of 40 threads.

4.2 Subjectivity Annotation

Wilson (2008) has annotated 20 AMI meetings for a variety of subjective phenomena which fall into the broad classes of *subjective utterances*, *objective polar utterances* and *subjective questions*. It is this first class in which we are primarily interested here. Two subclasses of subjective utterances are *positive subjective* and *negative subjective* utterances. Such subjective utterances involve the expression of a private state, such as a positive/negative opinion, positive/negative argument, and agreement/disagreement. The 20 meetings were labeled by a single annotator, though Wilson (2008) did conduct a study of annotator agreement on two meetings, reporting a κ of 0.56 for detecting subjective utterances. Of the roughly 20,000 dialogue acts total in the 20 AMI meetings, nearly 4000 are labeled as *positive-subjective* and nearly 1300 as *negative-subjective*. For the first experimental task, we consider the subjective class to be the union of positive-subjective and negative-subjective dialogue acts. For the second experimental task, the goal is to discriminate positive-subjective from negative-subjective.

For the BC3 emails, annotators were initially asked to create extractive and abstractive summaries of each thread, in addition to labeling a variety of sentence-level phenomena, including whether each sentence was subjective. In a second round of annotations, three different annotators were asked to go through all of the sentences previously labeled as subjective and indicate whether each sentence was *positive*, *negative*, *positive-negative*, or *other*. The definitions for positive and negative subjectivity mirrored those given by Wilson (2008). For the purpose of these experiments, we consider a sentence to be subjective if at least two of the annotators labeled it as subjective, and similarly consider a subjective sentence to be positive or negative if at least two annotators label it as such. Using this majority vote labeling, 172 of 1800 sentences are considered subjective, with 44% of those labeled as *positive-subjective* and 37% as *negative-subjective*, showing that there is much more of a balance between positive and negative sentiment in these email threads compared with meeting speech (note that some subjective sentences are not positive or negative). The κ for labeling subjective sentences in the email corpus is 0.32. The lower annotator agreement on emails compared with meetings suggests that subjectivity in email text may be manifested more subtly or conveyed somewhat amibiguously.

4.3 Classifier and Experimental Setup

For these experiments we use a maximum entropy classifier using the *liblinear* toolkit[2] (Fan et al., 2008). Feature subset selection is carried out by calculating the F-statistic for each feature, ranking the features according to the statistic, and training on increasingly smaller subsets of feature in a cross-validation procedure, ultimately choosing the feature set with the highest balanced accuracy during cross-validation.

Because the annotated portions of our corpora are fairly small (20 meetings, 40 email threads), we employ a leave-one-out method for training and testing rather than using dedicated training and test sets. For the polarity labeling task applied to the BC3 corpus, we pool all of the sentences and perform 10-fold cross-validation at the sentence level.

4.4 Evaluation Metrics

We employ two sets of metrics for evaluating all classifiers: precision/recall/f-measure and the receiver operator characteristic (ROC) curve. The ROC curve plots the true-positive/false-positive ratio while the posterior threshold is varied, and

[2] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

we report the area under the curve (AUROC) as the measure of interest. Random performance would feature an AUROC of approximately 0.5, while perfect classification would yield an AUROC of 1. The advantage of the AUROC score compared with precision/recall/f-measure is that it evaluates a given classifier across all thresholds, indicating the classifier's overall discriminating power. This metric is also known to be appropriate when class distributions are skewed (Fawcett, 2003), as is our case. For completeness we report both AUROC and p/r/f, but our discussions focus primarily on the AUROC comparisons.

5 Results

In this section we describe the experimental results, first for the subjective/non-subjective classification task, and subsequently for the positive-negative classification task.

5.1 Subjective / Non-Subjective Classification

For the subjectivity detection task, the results on the AMI and BC3 data closely mirrored each other, with the VIN approach constituting a very effective feature set, outperforming both baselines. We report the results on meeting and emails in turn.

5.1.1 AMI corpus

For the subjectivity task with the AMI corpus, we first report the precision, recall and f-measure results in Table 5 where the various classifiers are compared with a lower bound (LB) in which the positive class is always predicted, leading to perfect recall. It can be seen that the novel systems exhibit substantial improvement in precision and f-measure over this lower-bound. While the VIN approach yields the best precision scores, the full feature set achieves the highest f-measure.

As shown in Figure 1, the average AUROC with the VIN approach is 0.69, compared with 0.61 for AutoSlog-TS, a significant difference according to paired t-test (p<0.01). The VIN approach is also significantly better than the standard fully instantiated trigram pattern approach (p<0.01). This latter result suggests that the increased level of abstraction found in the varying instantiation n-grams does improve performance.

The conversational features alone give comparable performance to the VIN method (no significant difference), and the best results are found using the full feature set, which gives an average AUROC of 0.71, a significant improvement over VIN only (p<0.05).

Sys	**Precision**	**Recall**	**F-Measure**
AMI Corpus			
LB	26	100	41
Trig	25	63	36
Slog	39	48	43
VIN	41	58	48
Conv	36	73	49
All Feas	38	70	49
BC3 Corpus			
LB	10	100	17
Trig	27	10	14
Slog	24	13	17
VIN	27	22	24
Conv	25	29	27
All Feas	33	34	33

Table 5: P/R/F Results, Subjectivity Task

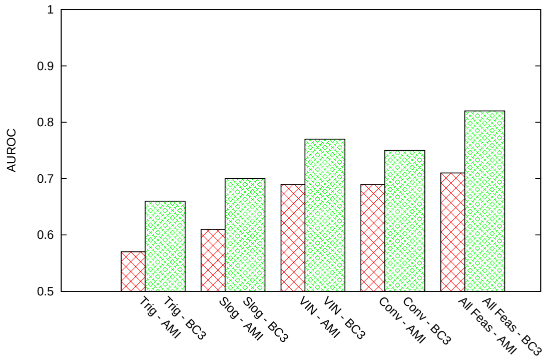

Figure 1: AUROCs on Subjectivity Task for AMI and BC3 corpora

5.1.2 BC3 corpus

For the subjectivity task with the BC3 corpus, the best precision and f-measure scores are found by combining all features, as displayed in Table 5. The f-measure for the VIN approach is ten points higher than for the standard trigram approach.

The average AUROC with the VIN approach is 0.77, compared with 0.70 for AutoSlog-TS (significant at p<0.05). The varying instantiation approach is significantly better than the standard trigram pattern approach (p<0.01), where the average AUROC is 0.66. We again find that conversational features are very useful for this task, and that the best overall results utilize the entire feature set. These results are displayed in Figure 1.

5.1.3 Impact of Blog Data

An interesting question is whether our use of the BLOG06 data was worthwhile. We can measure this by comparing the VIN AUROC results re-

Sys	Precision	Recall	F-Measure
AMI Corpus			
LB	76	100	86
Trig	87	8	14
Slog	75	46	57
VIN	83	60	70
Conv	82	47	60
All Feas	83	56	67
BC3 Corpus			
LB	54	100	70
Trig	50	84	63
Slog	58	56	57
VIN	53	84	65
Conv	63	80	71
All Feas	60	76	67

Table 6: P/R/F Results, Polarity Task

ported above with the VIN AUROC scores using only the annotated data for learning the significant patterns. The finding is that the blog data was very helpful, as the VIN approach averages only 0.66 on the BC3 data and 0.63 on the AMI data when the blog patterns are *not* used, both significantly lower ($p<0.01$). Figure 2 shows the ROC curves for the VIN approach with and without blog patterns applied to the AMI subjectivity detection task, illustrating the impact of the unsupervised pattern-learning strategy.

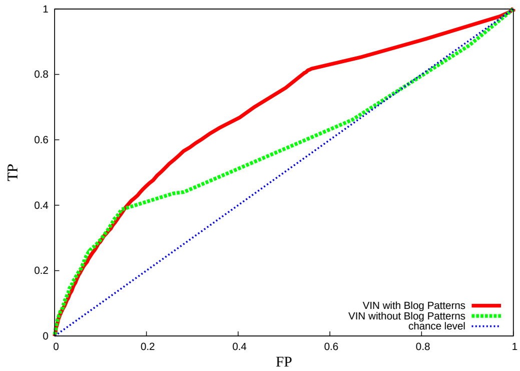

Figure 2: Effect of Blog Patterns on AMI Subjectivity Task

5.2 Positive / Negative Classification

For the polarity classification task, the results differ between the two corpora. We describe the results on meetings and emails in turn.

5.2.1 AMI corpus

The p/r/f results for the AMI polarity task are presented in Table 6, with the scores pertaining to the positive-subjective class. The VIN classifier and full features classifier achieve the highest precision, but the f-measures are below the lower-bound.

Comparing AUROC results, the VIN approach is again significantly better than AutoSlog-TS, averaging 0.65 compared with 0.56, and significantly better than the standard trigram approach ($p<0.01$ in both cases). The results are displayed in Figure 3. The conversational features are significantly less effective than the VIN features ($p<0.05$), and the best overall results are found by utilizing all features, with significant improvement over VIN only at $p<0.05$ and significant improvement over AutoSlog-TS only at $p<0.01$.

5.2.2 BC3 corpus

The results of the polarity task on the BC3 corpus are markedly different from the other experimental results. In this case, neither VIN nor AutoSlog-TS are particularly good for discriminating between positive and negative sentences, and the best strategy is to use features relating to conversational structure. According to p/r/f (Table 6), the only method outperforming the lower-bound in terms of f-measure is the conversational features classifier. According to AUROC scores shown in Figure 3, the conversational features by themselves are significantly better than the VIN approach ($p<0.01$ for non-paired t-test). So for emails, we are more likely to correctly classify positive and negative sentence by looking at features such as position in the turn and participant dominance than by matching our learned patterns. While we showed previously that pattern-based approaches perform well for the subjectivity task on this dataset, there was less success in using the patterns to discern the polarity of email sentences.

We are again interested in whether the use of the BLOG06 data was beneficial. For the BC3 data, there is very little difference between the VIN approach with and without the blog patterns, as they both perform poorly, but with the AMI corpus, the blog patterns yield significant improvement in polarity classification, increasing from an average of 0.56 without the blog patterns to 0.65 with them ($p<0.01$).

6 Discussion and Future Work

A key difference between the AMI and BC3 data with regards to subjectivity is that negative utterances are much more common in the BC3 email threads. Additionally, the pattern-based approaches fared worst in discriminating between

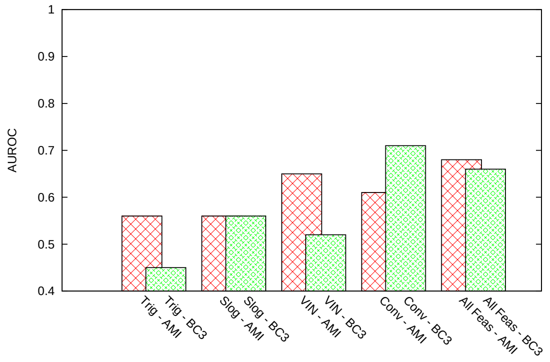

Figure 3: AUROCs on Polarity Task for AMI and BC3 corpora

negative and positive utterances in that corpus. Positive and negative email sentences are more easily recognized via features relating to conversation structure and participant status than through the learned lexical patterns.

The use of patterns learned from unlabeled blog data significantly improved performance. We are currently developing further techniques for learning subjective and polar patterns from such raw, natural text.

A potential area of improvement is to reduce the feature set by eliminating some of the subjective patterns. In Section 2, we briefly described the work of Riloff et al. (2006) on feature subsumption relationships. In our case, in a VIN instantiation set a given trigram instantiation subsumes the less abstract instantiations in the set, so the most abstract instantiation (i.e. completely uninstantiated trigram) representationally subsumes the rest. Eliminating some of the representationally subsumed instantiations when they are also behaviorally subsumed may improve our results.

It is difficult to compare our results directly with those of Raaijmakers et al. (2008) as they used a smaller set of AMI meetings for their experiments, and because for the first experiment we consider the subjective class to be the union of positive-subjective and negative-subjective dialogue acts whereas they additionally include subjective questions and dialogue acts expressing uncertainty. These differences are reflected by the substantially differing scores reported for majority-vote baselines on each task. However, their success with character n-gram features suggests that we could improve our system by incorporating a variety of character features. Character n-grams were the best single feature class in their experiments.

The VIN representation is a general one and may hold promise for learning patterns relevant to other interesting conversation phenomena such as decision-making and action items. We plan to apply the methods described here to these other applications in the near future.

7 Conclusion

In this work we have shown that learning subjective trigrams with varying instantiation levels from both annotated and raw data can improve subjectivity detection and polarity labeling for meeting speech and email threads. The novel pattern-based approach was significantly better than standard trigrams for three of the four tasks, and was significantly better than a state-of-the-art syntactic approach for those same tasks. We also found that features relating to conversational structure were beneficial for all tasks, and particularly for polarity labeling in email data. Interestingly, in three out of four cases combining all the features produced the best performance.

References

F. Biadsy, J. Hirschberg, and E. Filatova. 2008. An unsupervised approach to biography production using wikipedia. In *Proc. of ACL-HLT 2008, Columbus, OH, USA*.

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proc. of DARPA Speech and Natural Language Workshop, San Mateo, CA, USA*, pages 112–116.

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Proc. of MLMI 2005, Edinburgh, UK*, pages 28–39.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

T. Fawcett. 2003. Roc graphs: Notes and practical considerations for researchers.

G. Murray and G. Carenini. 2008. Summarizing spoken and written conversations. In *Proc. of EMNLP 2008, Honolulu, HI, USA*.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 1-2(2):1–135.

S. Raaijmakers, K. Truong, and T. Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. In *Proc. of EMNLP 2008, Honolulu, HI, USA*.

E. Riloff and W. Phillips. 2004. An introduction to the sundance and autoslog systems.

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proc. of EMNLP 2003, Sapporo, Japan*.

E. Riloff, S. Patwardhan, and J. Wiebe. 2006. Feature subsumption for opinion analysis. In *Proc. of EMNLP 2006, Sydney, Australia*.

S. Somasundaran, J. Ruppenhofer, and J. Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proc. of SIGDIAL 2007, Antwerp, Belgium*.

J. Ulrich, G. Murray, and G. Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In *Proc. of AAAI EMAIL-2008 Workshop, Chicago, USA*.

T. Wilson, J. Wiebe, and R. Hwa. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99.

T. Wilson. 2008. Annotating subjective content in meetings. In *Proc. of LREC 2008, Marrakech, Morocco*.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proc. of EMNLP 2003, Sapporo, Japan*.