Natural Language Engineering **17** (3): 397–418. © Cambridge University Press 2010
doi:10.1017/S1351324910000264

Subjectivity detection in spoken and written conversations

GABRIEL MURRAY and GIUSEPPE CARENINI
Department of Computer Science, 201-2366 Main Mall, University of British Columbia Vancouver, Canada V6T 1Z4
emails: gabrielm@cs.ubc.ca, carenini@cs.ubc.ca

(*Received 16 February 2010; revised 18 June 2010; accepted 23 August 2010; first published online 9 December 2010*)

Abstract

In this work we investigate four subjectivity and polarity tasks on spoken and written conversations. We implement and compare several pattern-based subjectivity detection approaches, including 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. We compare the use of these learned patterns with an alternative approach of using a very large set of raw pattern features. We also investigate how these pattern-based approaches can be supplemented and improved with features relating to conversation structure. Experimenting with meeting speech and email threads, we find that our novel systems incorporating varying instantiation patterns and conversation features outperform state-of-the-art systems despite having no recourse to domain-specific features such as prosodic cues and email headers. In some cases, such as when working with noisy speech recognizer output, a small set of well-motivated conversation features performs as well as a very large set of raw patterns.

1 Introduction

Conversations typically exhibit a large amount of subjective content. 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, and to determine the polarity of subjective sentences, would greatly facilitate the analysis, mining, and summarization of a large number of conversations. For example, knowing whether an individual meeting participant was arguing for or against a proposal could be valuable for an organization conducting a decision audit (Murray *et al.* 2008). An automated meeting assistant could summarize discussions while indicating the viewpoints of the participants. The growing prevalence of public conversations on weblogs and micro-blogs such as Twitter opens the possibility of large-scale opinion mining on the conversational web, which could allow companies and other organizations to track evolving public sentiment regarding their products or services.

Two of the most prevalent conversational media are meetings and emails. Face-to-face meetings enable numerous people to synchronously exchange a large amount of information and opinions in a short period of time, while emails allow for concise

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

exchanges between potentially far-flung participants over an extended period of time. Meetings and emails can also feed into one another, generating extended multi-modal conversations, 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. More generally, we investigate what subjectivity detection techniques are more suitable for spoken and written conversations (and combinations of the two), which tend to be more informal and less fluent than edited documents (e.g., news) (Baron 2000; Germesin *et al.* 2008).

In this paper, we propose two novel approaches for predicting subjectivity, and test them in experiments using meetings and emails. Our approaches combine lexico-syntactic features with features that capture basic characteristics of conversation structure across modalities. The first approach relies on a new general purpose method for learning subjective patterns, which essentially are n-gram sequences with varying levels of lexical instantiation, and we demonstrate how they can be learned from both labeled and unlabeled data. In contrast, our second approach relies on a very large raw feature set (200,000+) incorporating varying instantiation n-grams, POS tags, and character trigrams. For both approaches, we hypothesize that they may be more robust on disfluent and fragmented meeting speech and emails on which syntactic parsers may perform poorly.

We address four tasks relating to subjectivity detection and polarity weighting, which are as follows:

- Classification of subjective utterances (the union of positive- and negative-subjective utterances)
- Classification of all subjective phenomena including subjective questions[1]
- Classification of just positive-subjective utterances
- Classification of just negative-subjective utterances.

In all sets of experiments, we assess the impact of features relating to structural characteristics of multimodal conversations as well as participant information.

We show that using a large feature set of lexico-syntactic features combined with conversational features gives the best (or equivalent) results on most subjectivity detection tasks on our corpora. In contrast to many existing systems, we achieve this high performance even without using features specific to a particular conversational modalities such as prosodic cues for meetings or email headers for emails. This is a key advantage of our approach because it can be easily applied to multimodal conversation and it can also support domain-adaptation across multiple conversational modalities.

In certain cases, particularly on positive-subjective classification of emails and when working with ASR output on meetings, conversation features alone are either superior to or competitive with much larger raw feature sets incorporating lexical

[1] A comparison with previous work; see Section 4.3.

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

patterns and character trigrams. This demonstrates that it can be worthwhile to develop a small but well-motivated set of features for such tasks.

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 nonsubjective text as input, and outputs significant lexico-syntactic patterns. These patterns are based on shallow syntactic structure output by the Sundance 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. By experimenting with a corpus of news

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

documents, 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. But in our approach we consider a much richer set of such 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 compare this relatively small feature set of learned patterns with a very large raw feature set incorporating word n-grams, character trigrams, and other patterns. We then describe a set of features characterizing multimodal conversation structure, which can be used to supplement the n-gram approaches. Finally, we describe the baseline subjectivity detection approaches used for comparison.

3.1 Learned patterns: Varying instantiation 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 earlier, in the sense that their extraction patterns are partly instantiated. However, the AutoSlog-TS approach relies on deriving shallow syntactic structure with the Sundance shallow parser (Riloff and Phillips 2004). We hypothesize that our trigram approach may be more robust on disfluent and fragmented meeting speech and emails 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* (adverb–adjective–noun) 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*

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

Table 1. *Sample instantiation set*

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

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 versus nonsubjective 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 $probability \geq 0.60$ and $frequency \geq 2$ used by Riloff and Wiebe (2003), in order to identify a moderately sized set of high-precision patterns).

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, one supervised and one unsupervised. 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.

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

Table 2. *Example pos. and neg. patterns*

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	Doesn't really VB	0.833
It could VBP	0.83	Be DT problem	0.71
We could VBP	0.78	A bit JJ	0.66
NNS should VBP	0.71	Think PRP might	0.66

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 already 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. Table 2 shows how two patterns in the same instantiation set can have substantially different probabilities, as evidenced by *doesn't RB VB* with probability of 0.875 and *doesn't really VB* with probability of 0.833.

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

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

Table 3. *Example subjective patterns (BLOG06)*

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

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[2] 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 (versus 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. As noted, this work is also similar to work with AutoSlog-TS (Riloff 1996) which requires only document sets labeled as relevant and irrelevant in order to identify relevant patterns; in our case, there is no manual labeling of documents as relevant or irrelevant. We simply hypothesize that one dataset (weblogs) will inherently be more subjective than the other (newswire).

3.1.3 Deriving VIN features

There is some evidence in the subjectivity detection literature (Wilson *et al.* 2006) that it is beneficial to aggregate a large number of features into sets and create a new feature for each set. So, while we explore the usage of raw features in the next section, here we derive, for each sentence, features indicating the presence of

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

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

the significant VIN patterns. Patterns are binned according to their conditional probability range (i.e., $0.65 \leq p < 0.75$, $0.75 \leq p < 0.85$, $0.85 \leq p < 0.95$, and $0.95 \leq p$). 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. For instance, the sentence *I completely agree with the proposal, but I can not really see how we would afford it* would have a feature representation of $< 0, 0, 0, 2 >$ according to the patterns in Table 3. 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.

3.2 Raw feature set

The feature set described earlier is fairly small, with the VIN patterns grouped into features according to probability range. As a comparison approach and an assessment on the impact of feature set size, we utilize a much larger raw feature set that includes varying instantiation n-grams, character n-grams, word n-grams, and other patterns as described as follows:

- *Character trigrams*: We derive all of the character trigrams in the collected corpora and include features indicating the presence or absence of each trigram in a given sentence.
- *Word bigrams*: We derive all of the word bigrams in the collected corpora and include features indicating the presence or absence of each bigram in a given sentence.
- *POS bigrams*: We derive all of the POS-tag bigrams in the collected corpora and include features indicating the presence or absence of each bigram in a given sentence.
- *Word pairs*: We consider w_1, w_2 to be a word pair if they occur in the same sentence and w_1 precedes w_2. We derive all of the word pairs in the collected corpora and includes features indicating the presence or absence of each word pair in the given sentence. This is essentially a skip bigram where any amount of intervening material is allowed as long as the words occur in the same sentence.
- *POS pairs*: We consider p_1, p_2 to be a POS pair if they occur in the same sentence and p_1 precedes p_2. We derive all of the POS pairs in the collected corpora and includes features indicating the presence or absence of each POS pair in the given sentence. This is essentially skip bigrams for POS tags.
- *Varying instantiation ngrams*: This corresponds to a set of raw VIN features. For each word bigram w_1, w_2, we further represent the bigram as p_1, w_2

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

and w_1, p_2 so that each pattern consists of a word and a POS tag. We include a feature indicating the presence or absence of each of these varying instantiation bigrams.

After eliminating patterns that occur five or fewer times in the collected corpora, we end up with 200,000+ raw features. Note that unlike the previously described VIN patterns that were first filtered based on their probabilities, here we simply feed all of the raw features into the machine learning algorithm and attempt to learn the best ones.

3.3 Conversational features

While we hypothesize that the general purpose pattern-based approaches described above will greatly aid our various subjectivity tasks, we also recognize that there are many additional features specific for characterizing multimodal conversations that may correlate well with subjective phenomena. 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.

The conversational features used in these experiments are listed and briefly described in Table 4. These features have previously been used with success on the task of automatically summarizing conversations (Murray and Carenini 2008). We treat emails and meetings as conversations comprised of turns between multiple participants. We follow Carenini *et al.* (2007) in working at the finer granularity of email fragments, so that for an email thread, a turn consists of a single email fragment in the exchange. For meetings, a turn is a sequence of dialogue acts by one speaker, with the turn boundaries delimited by dialogue acts from other meeting participants. For example, if speaker A speaks two dialogue acts, followed by speaker B speaking one dialogue act, then speaker A again responding with four dialogue acts, there is a total of three turns. The features we derive for subjectivity detection are based on this view of the conversational structure.

We calculate two *length* features. For each sentence, we derive a word-count feature normalized by the longest sentence in the conversation (*SLEN*) and a word-count feature normalized by the longest sentence in the turn (*SLEN2*).

There are several *structural* features used, including position of the sentence in the turn (*TLOC*) and position of the sentence in the conversation (*CLOC*). We also include the time from the beginning of the conversation to the current turn (*TPOS1*) and from the current turn to the end of the conversation (*TPOS2*). Conversations in both modalities can be well structured, with introductory turns, general discussion, and ultimate resolution or closure, and sentence informativeness might significantly correlate with this structure. We calculate two pause-style features: the time between the following turn and the current turn (*SPAU*), and the time between the current turn and previous turn (*PPAU*), both normalized by the overall length of the conversation. These features are based on the email and meeting transcript

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

Table 4. *Features key*

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*

timestamps. We hypothesize that pause features may be useful if informative turns tend to elicit a large number of responses in a short period of time, or if they tend to quickly follow a preceding turn, to give two examples.

There are two features related to the conversation *participants* directly. One measures how dominant the current participant is in terms of words in the conversation (*DOM*), and the other is a binary feature indicating whether the current participant initiated the conversation (*BEGAUTH*), based simply on whether they were the first contributor. It is hypothesized that informative sentences may more often belong to participants who lead the conversation or have a good deal of dominance in the discussion.

There are several *lexical* features used in these experiments. For each unique word, we calculate two conditional probabilities. For each conversation participant, we calculate the probability of the participant given the word, estimating the probability from the actual term counts, and take the maximum of these conditional probabilities as our first term score, which we will call *Sprob*.

$$Sprob(t) = \max_{S} p(S|t)$$

where t is the word and S is a participant. For example, if the word *budget* is used 10 times in total, with seven uses by participant A, three uses by participant B,

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

and no uses by the other participants, then the *Sprob* score for this term is 0.70. The intuition is that certain words will tend to be associated with one conversation participant more than the others, owing to varying interests and expertise between the people involved.

Using the same procedure, we calculate a score called *Tprob* based on the probability of each turn given the word.

$$Tprob(t) = \max_{T} p(T|t)$$

The motivating factor for this metric is that certain words will tend to cluster into a small number of turns, owing to shifting topics within a conversation.

Having derived *Sprob* and *Tprob*, we then calculate several sentence-level features based on these term scores. Each sentence has features related to *max*, *mean*, and *sum* of the term scores for the words in that sentence (*MXS*, *MNS*, and *SMS* for *Sprob*, and *MXT*, *MNT*, and *SMT* for *Tprob*). Using a vector representation, we calculate the cosine between the conversation preceding the given sentence and the conversation subsequent to the sentence, first using *Sprob* as the vector weights (*COS1*) and then using *Tprob* as the vector weights (*COS2*). This is motivated by the hypothesis that informative sentences might change the conversation in some fashion, leading to a low cosine between the preceding and subsequent portions. We similarly calculate two scores measuring the cosine between the current sentence and the rest of the conversation, using each term-weight metric as vector weights (*CENT1* for *Sprob* and *CENT2* for *Tprob*). This measures whether the candidate sentence is generally similar to the conversation overall.

There are three word entropy features, calculated using the formula

$$went(s) = \frac{\sum_{i=1}^{N} p(x_i) - \log(p(x_i))}{\left(\frac{1}{N} - \log\left(\frac{1}{N}\right)\right) M}$$

where s is a string of words, x_i is a word type in that string, $p(x_i)$ is the probability of the word based on its normalized frequency in the string, N is the number of word types in the string, and M is the number of word tokens in the string.

Note that word entropy essentially captures information about type-token ratios. For example, if each word token in the string was a unique type then the word entropy score would be 1. We calculate the word entropy of the current sentence (*THISENT*), as well as the word entropy for the conversation up until the current sentence (*PENT*) and the word entropy for the conversation subsequent to the current sentence (*SENT*). We hypothesize that informative sentences themselves may have a diversity of word types, and that if they represent turning points in the conversation they may affect the entropy of the subsequent conversation.

Finally, we include a feature that is a rough approximation of the ClueWordScore (CWS) used by Carenini *et al.* (2007). For each sentence, we remove stopwords and count the number of words that occur in other turns besides the current turn. The CWS is therefore a measure of conversation cohesion.

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

3.4 Baselines and proposed approaches

There are two baselines in particular to which we are interested in comparing our more sophisticated approaches. 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 utterance detection experiment 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 1,200 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 use the AutoSlog-TS algorithm 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 described 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 $p \geq 0.65$ and $frequency \geq 5$ 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 3,000 significant patterns are learned. Among significant patterns learned from the AMI corpus are *<subj> BE good*, *change <dobj>*, *<subj> agree*, and *problem with <NP>*.

To gauge the effectiveness of the various feature types, for our initial subjective utterance detection experiment 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), conversation features plus all the learned patterns such as VIN and AutoSlog-TS patterns(CONV + LEARNED), and conversation features plus the large raw feature set (CONV+ RAW).

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 min.

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

The AMI corpus contains ASR output in addition to manual meeting transcripts, and we report results on both transcript types. The ASR output was provided by the AMI–ASR team (Hain *et al.* 2007), and the word error rate for the AMI corpus is 38.9%. The AMI automatic transcription system uses the standard framework of hidden Markov model (HMM) acoustic modeling and n-gram language models, in this case tri-grams. To achieve fair recognition output, the corpus is divided into five parts, employing a leave-one-out procedure of training the language and acoustic models on four portions of the data and testing on the fifth, rotating to obtain recognition results for the entire corpus.

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. We have made this corpus freely available for download.[3]

4.2 Subjectivity annotation

Wilson (2008) has annotated 20 AMI meetings for a variety of subjective phenomena at the dialogue act level which fall into the broad classes of *subjective utterances*, *objective polar utterances*, and *subjective questions*. Two subclasses of subjective utterances are *positive subjective* and *negative subjective* utterances. Such subjective utterances involve the expression of a private state (Quirk *et al.* 1985) (an emotion or state of mind that is not always observable), such as a positive/negative opinion, positive/negative argument, and agreement/disagreement. An objective polar utterance is one that conveys positive or negative facts without indicating any private state, e.g., *The camera broke the first time I used it* is a negative polar utterance (Wilson 2008). 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 4,000 are labeled as *positive-subjective* and nearly 1,300 as *negative-subjective*. ASR output is missing for one of these 20 meetings (IS1003b) and so all ASR results are based on a 19 meeting test set. For the first experimental task, we consider the subjective class to be the union of positive-subjective and negative-subjective dialogue acts. We initially concentrate only on positive subjective and negative subjective utterances (ignoring, for example, subjective questions and uncertainties) because we want to make a direction comparison with the BC3 email corpus which only includes annotation for positive and negative subjective sentences. However, in Section 5.2, we include these other phenomena in our subjectivity classification task and compare our results with Raaijmakers *et al.* (2008).

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

[3] http://www.cs.ubc.ca/nest/lci/bc3.html

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

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 training classifiers, 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 1,800 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 ambiguously.

4.3 Classifier and experimental setup

For these experiments, we employ a maximum entropy classifier using the *liblinear* toolkit[4] (Fan *et al.* 2008).[5] Because the annotated portions of our corpora are fairly small (20 meetings, 40 email threads), we employ 10-fold cross-validation for training and testing rather than using dedicated training and test sets.

We carry out four tasks in total, which are as follows:

(1) *Subjective Utterance Detection*: In these experiments we aim to discern the union of positive-subjective and negative-subjective utterances from the remainder of the sentences.
(2) *Subjective Question and Utterance Detection*: To compare with previous work, for this task we aim to identify all subjective utterances and questions. This class consists of sentences labeled as *positive-subjective*, *negative-subjective*, *positive-and-negative subjective*, *uncertainty*, *other subjective*, *subjective fragment*, *positive-subjective question*, *negative-subjective question* and *general-subjective question*. This is equivalent to Task 1 of Raiijmakers *et al.* (2008), to which we compare our system. We report results only on the AMI corpus because the BC3 corpus is limited in its subjectivity annotation.
(3) *Positive-Subjective Utterance Detection*: In this task, we attempt to detect all positive-subjective utterances.
(4) *Negative-Subjective Utterance Detection*: In this task, we attempt to detect all negative-subjective utterances.

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

[4] http://www.csie.ntu.edu.tw/ cjlin/liblinear/
[5] In previous experiments we found comparable performance between *libsvm* and *liblinear*.

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

plots the true-positive/false-positive ratio while the posterior threshold is varied, and 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 scores, but our discussions focus primarily on the AUROC comparisons.

5 Results

In this section, we describe the experimental results, beginning with the subjective utterance detection task.

5.1 Task 1: Subjective utterance detection

For the subjective utterance detection task, the results on the AMI and BC3 data closely mirror each other, with the best results found by our two novel systems combining conversational features with more general lexico-syntactic patterns (i.e., CONV+LEARNED and CONV+RAW). More specifically our approaches consistently outperform standard trigrams and AutoSlog-TS patterns. We report the results on meetings and emails in turn.

5.1.1 AMI corpus

Figure 1 shows the AUROC scores for all approaches on the subjective utterance detection task, applied to manual transcripts. A first key finding is that 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 second key finding is that the conversation features plus raw features gives the best overall performance with an AUROC of 0.76 on manual transcripts, compared with 0.70 for each of CONV and CONV+LEARNED, a statistical improvement according to t-test ($p<0.05$). However, on ASR transcripts these top three systems are statistically indistinguishable. Transcript type does not have a significant effect on AUROC scores according to analysis of variance, and in general ASR has a modest impact on most of the systems, with the exception of CONV+RAW where the AUROC decreases from 0.76 to 0.70. The interaction of the system and transcript factors can be seen in Figure 2.

Table 5 gives the average F measures for the top three systems compared with a lower bound (LB) in which the positive class is always predicted, leading to perfect recall. We again see that CONV+RAW yields the best results overall but that this

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

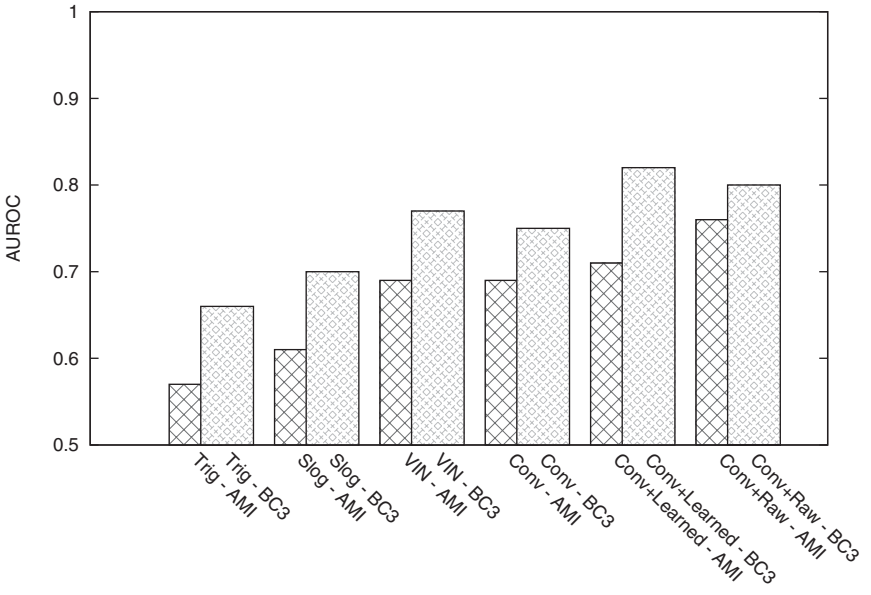


Fig. 1. AUROCs on subjectivity task for AMI and BC3 corpora.

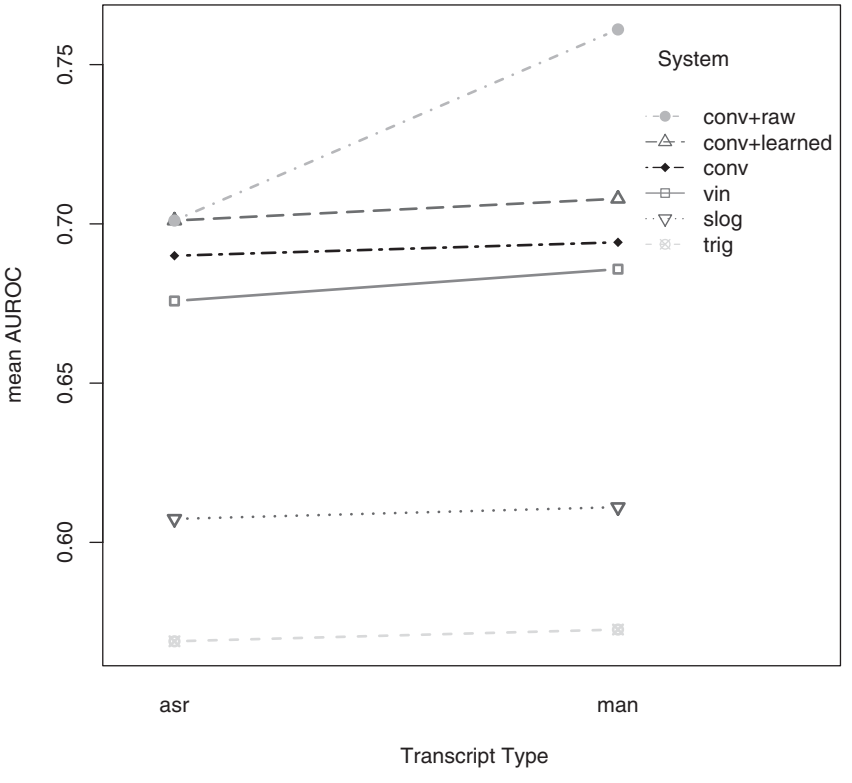


Fig. 2. Effect of transcript and system, AMI subjectivity task.

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

Table 5. *F-Measures on subjective utterance detection task*

Sys	LB	Conversation	Conv+Learned	Conv+Raw
F Measures				
AMI – Manual	41	47	49	52
AMI – ASR	41	47	47	47
BC3	17	27	33	27

advantage disappears on ASR transcripts. For conducting this task on noisy recognition output, a set of 24 conversation features is as good as 200,000+ raw patterns.

5.1.2 BC3 corpus

With the BC3 emails, we again find that the VIN approach is significantly better than AutoSlog-TS ($p<0.05$) and the standard trigram approach ($p<0.01$), with respective AUROCs of 0.77, 0.70, and 0.66. We find that conversational features are very useful for this task, and that the best overall results utilize the conversation features plus all learned patterns (CONV+LEARNED). The CONV+RAW approach is only slightly lower (no statistical difference), suggesting that learned patterns are slightly more useful for this email data. The F measures for CONV, CONV+LEARNED, and CONV+RAW are given in Table 5, showing the same general trend. Note that while the F measures on the BC3 data are generally low, they are nonetheless much higher than the lower-bound (LB) in which the positive class is always predicted.

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 reported earlier 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 0.55 on the BC3 data and 0.63 on the AMI data when the blog patterns are *not* used, compared with 0.69 and 0.77 previously, both significantly lower ($p<0.01$). Figures 3 and 4 show the ROC curves for the VIN approach with and without blog patterns applied to the AMI (manual) and BC3 subjectivity detection task, illustrating the impact of the unsupervised pattern-learning strategy.

5.2 Task 2: Subjective question and utterance detection

In the previous section, we found that the conversation features plus raw features approach (CONV+RAW) was in most cases superior to or equal to the best of the other systems on both meetings and emails. For the remaining three tasks, we focus on comparing conversation features with the supplementary raw features. For Task 2 (all subjective sentences versus rest), we report results only on the AMI corpus, as the BC3 corpus does not currently contain annotations for subjective questions, uncertainties and other phenomena beyond positive- and negative-subjective sentences.

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

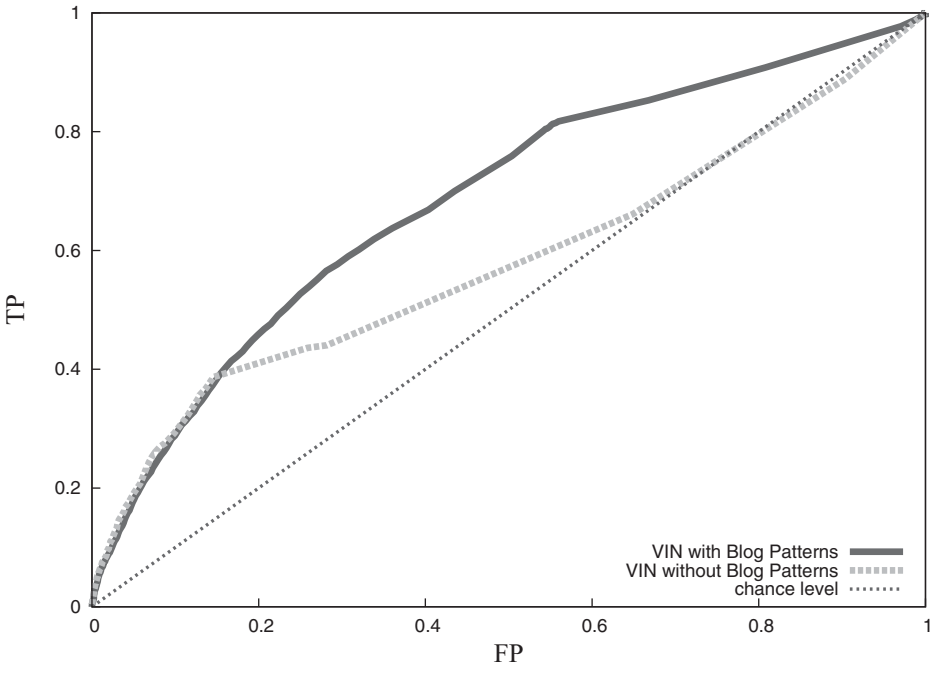


Fig. 3. Effect of blog patterns on AMI subjectivity task.

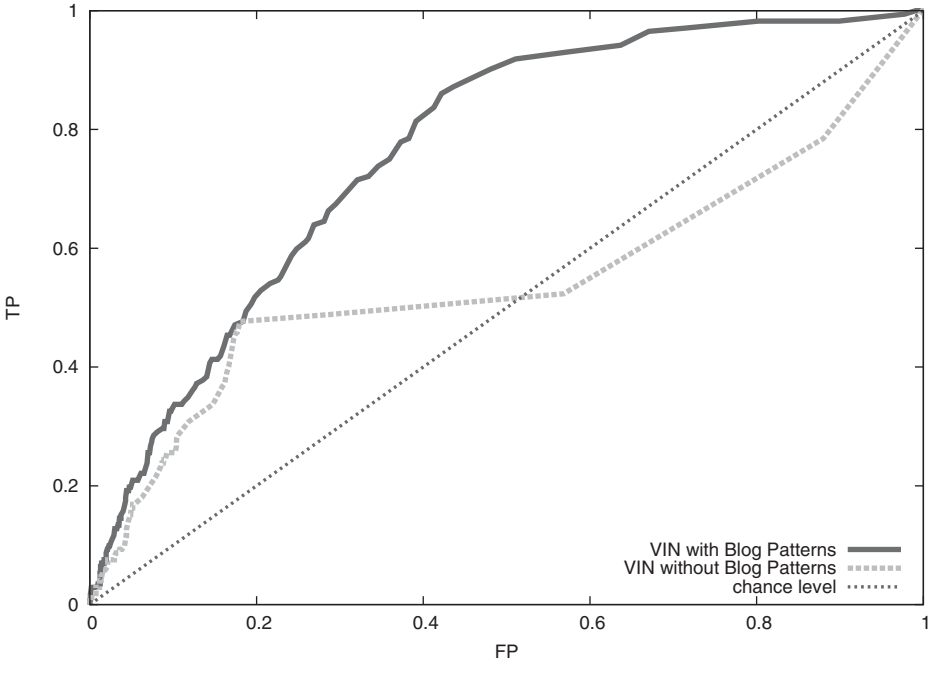


Fig. 4. Effect of blog patterns on BC3 subjectivity task

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

Table 6. *AUROC scores, all tasks*

Sys	Conversation	Conv+Raw
Task 2: Subjective questions and utterances		
AMI - Manual (20 meetings)	0.74	**0.81**
AMI - ASR (19 meetings)	0.75	**0.77**
Task 3: P-Subj utterances		
AMI - Manual	0.66	**0.76**
AMI - ASR	0.65	**0.71**
BC3	*0.77*	0.66
Task 4: N-Subj utterances		
AMI - Manual	0.72	**0.76**
AMI - ASR	0.71	0.72
BC3	0.71	0.75

Table 7. *F measures, all tasks*

Sys	LB	Conversation	Conv+Raw
Task 2: Subjective questions and utterances			
AMI - Manual (13 meetings)	58	65	**68**
AMI - Manual (20 meetings)	54	61	**66**
AMI - ASR (19 meetings)	54	60	**63**
Task 3: P-Subj utterances			
AMI - Manual	32	36	**45**
AMI - ASR	32	34	**41**
BC3	06	15	12
Task 4: N-Subj utterances			
AMI - Manual	12	21	**33**
AMI - ASR	12	21	**28**
BC3	04	11	13

The AUROC scores for Task 2 are shown in Table 6. Significant differences between systems according to t-test are in boldface (all $p<0.05$), while differences of marginal significance ($0.05<p<0.1$) are in italics. It can be seen that the CONV+RAW system is significantly better than CONV on both manual and ASR transcripts. We ran our classifiers on the same 13 meeting test set as Raaijmakers *et al.* (2008) in addition to the full 20 meeting set. Table 7 shows the F measures for Task 2. A key finding is that, for the subjectivity detection on the 13 meeting subset used by Raaijmakers *et al.*, the F measure for our expanded feature set slightly outperforms their best result, at 68.3–67.1. Furthermore, their best result depended on prosodic features and phoneme n-grams, while we only include features that are general to any conversation modality. Regarding the tradeoff between precision and recall, the Raaijmakers *et al.* system achieves higher precision (74.5 versus our 64.1) but lower recall (61.2 versus our 73.7). For all 20 meetings, our average F measure if 0.66. Performance on ASR transcripts dips slightly to 0.63.

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

5.3 Task 3: Positive-subjective utterance detection

For Task 3, positive-subjective utterance detection, the results differ between the meeting and email corpora. For the AMI corpus, the CONV+RAW system is significantly better on both manual and ASR transcripts, with a best AUROC score of 0.76 on manual transcripts. For the BC3 corpus, the conversation features alone give the best classification performance, with an average performance more than ten points higher than CONV+RAW, a marginally significant difference. In the email domain, sentences containing positive-subjective sentiments are more easily detected using features of conversation structure and speaker status than through explicit lexical patterns. These same trends can be seen in the F Measures given in Table 7.

5.4 Task 4: Negative-subjective utterance detection

The results on Task 4, negative-subjective utterance detection, differ from the findings in Task 3. Here CONV+RAW always gives the better performance compared with CONV features alone. This difference is significant on AMI manual transcripts according to AUROC, and on both manual and ASR transcripts according to F measures. Whereas positive-subjective sentences in emails were more easily detected via features relating to conversation structure, we find here that negative-subjective sentences correlate well with lexical patterns and POS n-grams. While the F measures for the BC3 corpus are generally low, they are still much higher than the lower bound (LB).

It is interesting to note that on the AMI corpus, there is a larger performance gap between CONV and CONV+RAW on the positive-subjective task (Task 3) than on the negative-subjective task (Task 4). For detecting negative sentiments, the raw lexical and POS features give a comparatively small bonus. It may be the case that negative sentiment in face-to-face meetings is expressed more subtly or couched in a variety of euphemistic terms whereas positive sentiment is often clearly signaled by specific repeated word patterns. This is borne out by the fact that in Sections 3.1 and 3.4 we learned only a handful of negative VIN and AutoSlog-TS patterns compared with the positive patterns that were learned.

6 Summary and discussion

The best classification results over all four tasks, according to AUROC, are on detecting all subjective utterances and questions (Task 2), with an AUROC of 0.81 for manual transcripts and 0.77 for ASR. Scores dip slightly when trying to classifying positive- and negative-subjective sentences versus the rest (Tasks 1, 3, and 4), showing that it is easier to discern all subjective sentences and questions as a single group than it is to identify more specific subjective phenomena.

The meeting and email domains are similar in that they both constitute multiparty conversations, and both are amenable to carrying out various subjectivity classification tasks using a unified set of conversational features and supplementary raw lexico-syntactic features. However, the domains differ somewhat in how these subjective phenomena are manifested, and it may be possible to improve results

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

in one domain by leveraging knowledge from the other domain. For example, the BC3 emails contain much more of a balance between positive-subjective and negative-subjective utterances. It may be possible to improve negative-subjective classification in the meetings domain by examining how negative opinions are expressed in emails. And because we possess much more subjectivity annotation for the meetings domain, in future work we will aim to improve email results by implementing domain adaptation techniques from the source domain of meetings to the target domain of emails. We also plan to apply this classification approach to other conversational domains such as blogs and forums.

While the best results often utilized conversation features plus supplementary lexical and POS features, it is worth noting that not only are conversation features alone sometimes superior (e.g., positive-subjective classification in emails) but conversation features alone are generally competitive on all tasks. It turns out that 24 well-motivated conversation features can be nearly as effective as 200,000+ raw features. And some subjective phenomena, such as negative sentiments in face-to-face conversations, are simply not often signaled by overt lexico-syntactic cues. This provides a great deal of motivation for further research on negative subjectivity classification in particular. On all tasks, the impact of ASR errors on CONV+RAW tends to erase or diminish its advantage over the other approaches.

7 Conclusion

We have presented results from four experimental tasks on both meeting speech and email threads: subjective utterance detection, subjective question and utterance detection, classification of positive-subjective utterances, and classification of negative-subjective utterances. On all these tasks, we found that coupling conversation features with a large raw feature set incorporating varying instantiation n-grams and character trigrams often gives the best performance results. A key finding is that conversation features alone can give credible performance on these tasks, particularly when working with ASR and for classifying positive subjectivity in emails, despite comprising a relatively small feature set. On the task of detecting all subjective utterances and questions, we showed that the current approach surpasses a state-of-the-art subjectivity detection system despite having no recourse to domain-specific features, which makes it not only generally applicable to multi-modal conversations, but also suitable for domain adaptation across conversation modalities.

References

Baron, N. 2000. *Alphabet to Email: How Written English Evolved and Where it's Heading*. New York, NY: Routledge (Taylor & Francis).

Biadsy, F., Hirschberg, J., and Filatova, E. 2008. An unsupervised approach to biography production using wikipedia. In *Proceedings of Acl-Hlt 2008,* Columbus, OH.

Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of Darpa Speech and Natural Language Workshop,* San Mateo, CA, pp. 112–116.

Carenini, G., Ng, R., and Zhou, X. 2007. Summarizing email conversations with clue words. In *Proceedings of Acm www 07,* Banff, Canada.

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press

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

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

Fawcett, T. 2003. Roc graphs: Notes and practical considerations for researchers. Technical Report HP Labs HPL-2003–4.

Germesin, S., Becker, T., and Poller, P. 2008. Hybrid multi-step disfluency detection. In *Proceedings of Mlmi 2008,* Utrecht, The Netherlands, pp. 185–195.

Hain, T., Burget, L., Dines, J., Garau, G., Wan, V., Karafiat, M., Vepa, J., and Lincoln, M. 2007. The AMI system for transcription of speech in meetings. In *Proceedings of Icassp 2007,* pp. 357–360.

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

Murray, G., Kleinbauer, T., Poller, P., Renals, S., Becker, T., and Kilgour, J. 2008. Extrinsic summarization evaluation: A decision audit task. In *Proceedings of Mlmi 2008,* Utrecht, The Netherlands.

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

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

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

Riloff, E. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of Aaai 1996,* Portland, OR, pp. 1044–1049.

Riloff, E., and Phillips, W. 2004. An introduction to the sundance and autoslog systems. Technical Report UUCS-04-015, University of Utah School of Computing.

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

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

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

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

Wilson, T. 2008. Annotating subjective content in meetings. In*: Proceedings of Lrec 2008,* Marrakech, Morocco.

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

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

https://doi.org/10.1017/S1351324910000264 Published online by Cambridge University Press