Detecting Subjectivity in Multiparty Speech

Gabriel Murray and Giuseppe Carenini

Department of Computer Science, University of British Columbia, Vancouver, Canada
gabrielm@cs.ubc.ca, carenini@cs.ubc.ca

Abstract

In this research we aim to detect subjective sentences in spontaneous speech and label them for polarity. We introduce a novel technique wherein subjective patterns are learned from both labeled and unlabeled data, using n-grams with varying levels of lexical instantiation. Applying this technique to meeting speech, we gain significant improvement over state-of-the-art approaches and demonstrate the method's robustness to ASR errors. We also show that coupling the pattern-based approach with structural and lexical features of meetings yields additional improvement.

1. Introduction

Face-to-face meetings are rich in subjectivity. Being able to separate objective utterances (e.g. *We made a prototype*) from subjective utterances (e.g. *Looks good* or *Yeah, I'd have to agree*) would allow a person reviewing the conversation to see *what was discussed* and *what attitudes existed* towards the items under discussion.

In this paper we describe a novel approach for predicting subjectivity, and apply that method to meeting speech using both manual and automatically generated transcripts. Our approach combines a new general purpose method for learning subjective patterns, with features that capture essential characteristics of multiparty conversations. 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, lexical and participant information.

We run two sets of experiments, first assessing our approach on the task of discriminating subjective and non-subjective utterances, and secondly establishing the polarity of the utterances (i.e., discriminating positive and negative subjectivity). In both sets of experiments, we assess the impact of automatic transcription errors. The results indicate that our approach consistently outperforms existing state-of-the-art lexico-syntactic approaches. We hypothesize that the key advantage of our approach is to be more robust to disfluent and ungrammatical speech.

2. Related Research

Raaijmakers et al. [1] 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. They found character n-grams to be particularly useful.

Riloff and Wiebe [2] presented the AutoSlog-TS method for learning subjective extraction patterns from a large amount of data, which takes relevant and irrelevant text as input (e.g. subjective and non-subjective sentences), and outputs significant lexico-syntactic patterns. These patterns are based on syntactic structure output by the Sundance shallow dependency parser [3]. They are extracted by exhaustively applying syntactic templates to a training corpus, with an extracted pattern for every instantiation of the syntactic template. These patterns are scored according to probability of relevance (i.e. subjectivity) 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.

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, in our case by considering n-grams of varying levels of lexical instantiation.

3. Corpus and Annotation

The AMI corpus [4] consists of meetings in which participants take part in role-playing exercises concerning the design and development of a remote control. The corpus contains automatic speech recognition (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 [5], and the word error rate for the AMI corpus is 38.9%.

Wilson [6] 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 [6] did conduct a study of annotator agreement on two meetings, finding a κ of 0.56 for subjectivity labeling. Of the roughly 20,000 utterances total in the 20 AMI meetings, 36.6% are labeled as subjective, 22% are labeled as *positive subjective*, and 8.5% are labeled as *negative subjective*. We ultimately discarded one meeting (IS1003b) because there was no ASR output available.

4. 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 briefly describe a set of features characterizing multiparty conversation structure which can be used to supplement the n-gram approach. Finally, we describe the baseline subjectivity detection approaches used for comparison.

4.1. Partially Instantiated Language Models

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. [1] 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 [2, 3], 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 [3]. We hypothesize that the VIN approach may be more robust to disfluent and fragmented meeting speech. 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.

VIN begins by running the Brill POS tagger 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 {*really great idea*, *really great NN*, *really JJ idea*, ..., *RB JJ NN*}. As we scan through 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, following the scoring methodology of Riloff and Wiebe [2]. We calculate the conditional probability $p(relevance|trigram)$ using the actual trigram counts in relevant and irrelevant text. For learning negative patterns, we treat all negative sentences as the relevant text and the remainder of the sentences as irrelevant text, and proceed similarly for learning positive 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 [2]).

We possess a fairly small amount of meeting data annotated for subjectivity and polarity. To address this data shortfall, we take both a supervised and an unsupervised approach to learning patterns, described in turn below.

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 1: Example Pos. and Neg. Patterns

4.1.1. 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 [6]. 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. It may also be the case that when conversation participants *do* express negative sentiments, they couch those sentiments in more euphemistic or guarded terms compared with positive sentiments. Table 1 gives examples of significant positive and negative patterns learned from the labeled meeting data. The last two rows in Table 1 show how two patterns in the same instantiation set can have substantially different probabilities.

4.1.2. 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 meeting speech: 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 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. Newswire artices may contain subjective content such as reported sentiment, but generally will not contain directly stated sentiment or opinions as found in meeting speech. 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 a portion of 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 rel-

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

evant, 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 *IN PRP opinion*, *RB think that* and *RB agree with*.

4.1.3. 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. 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. [1] similarly include an n-gram feature allowing such intervening material).

4.2. 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 multiparty speech 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.

We use the feature set described by Murray and Carenini [7], which they used for automatic summarization of meetings and emails. 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 meeting participants and the latter of which weights words based on their distributions across conversation turns.

4.3. Baseline Approaches

There are two baselines in particular to which we are interested in comparing the VIN approach. To test the hypothesis that the increasing levels of abstraction found with partially instantiated trigrams will lead to improved classification, 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, and 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 [2] 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 4.1, using the same probability and frequency thresholds, and patterns are similarly binned to create multiple features. 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* $< NP >$.

To gauge the effectiveness of the various feature types, for both sets of experiments we build multiple systems 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.

5. Experimental Setup

For these experiments we use maximum entropy classifiers with the *liblinear* toolkit[2], which incorporates feature subset selection based on ranking individual features according to the F-statistic and choosing the feature set with the highest balanced accuracy during cross-validation.

Because the annotated portions of our corpora are fairly small, we employ a leave-on-out method for training and testing rather than using dedicated training and test sets. We evaluate each classifier by plotting the receiver operator characteristic (ROC) curve and finding the area under the ROC curve (AUROC). The ROC curve plots the true-positive/false-positive ratio while the posterior threshold is varied, giving us an indication of the classifier performance across all thresholds.

6. 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.

6.1. Subjective / Non-Subjective Classification

For the subjectivity task, the choice of system has a significant effect according to analysis of variance ($p<0.001$), while the transcript type has no significant effect. Figure 1 shows the performance of each system on both manual and ASR transcripts and illustrates how all approaches show little or no decline when applied to recognition output. To further investigate the significant effect of system on AUROC scores, we conduct a post-hoc Tukey test. The top three approaches (VIN, conversational features, and the full feature set) are each significantly better than the AutoSlog-TS and trigram approaches (all $p<0.001$), while we find that the full feature set can bring significant improvement over the VIN-only approach ($p<0.05$). The AutoSlog-TS approach is significantly better than the standard trigram method ($p<0.001$). The fact that the VIN approach is significantly better than the standard fully instantiated trigram pattern approach suggests that the increased level of abstraction found in the varying instantiation n-grams does improve performance.

An interesting question is whether our use of the BLOG06 data was worthwhile. We can measure this by comparing the VIN results reported above with the VIN results 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 0.63 on the AMI data when the blog patterns are *not* used, a significantly lower result ($p<0.01$). 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.

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

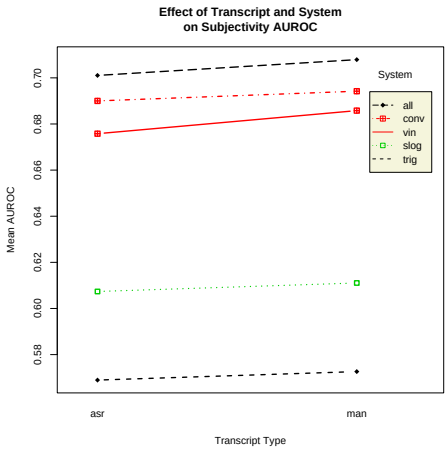

Figure 1: Subjectivity Scores

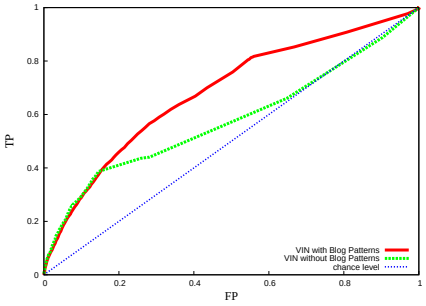

Figure 2: Effect of Blog Patterns on AMI Subjectivity Task

6.2. Positive / Negative Classification

For the polarity task, both the system type and transcript type have a significant effect on AUROC scores (both $p<0.01$). Figure 3 illustrates that all scores are lower on ASR transcripts compared with manual transcripts. The VIN approach applied to ASR is the best of all approaches except for the classifier using all features. Investigating the system effect using the post-hoc Tukey test, we find that the full feature approach is significantly better than all other approaches ($p<0.001$) with the exception of the VIN approach. The VIN approach is significantly better than AutoSlog-TS and the standard trigram approach (both $p<0.001$). There is a wider performance gap between the VIN and conversation features approaches on the polarity task compared with the subjectivity task, with the VIN approach superior at a marginal significance level ($0.05<p<0.1$).

7. Discussion and Conclusion

The novel VIN approach performed very well on both tasks, and significantly better than the standard trigram approach and the AutoSlog-TS method. The conversational features are comparable to VIN in effectiveness, and the best results on both tasks are found by combining all features.

The unsupervised technique for learning patterns from blog data was successful, greatly increasing our coverage and significantly improving results compared with using only the patterns from the annotated meeting data.

The impact of ASR on all systems is more pronounced

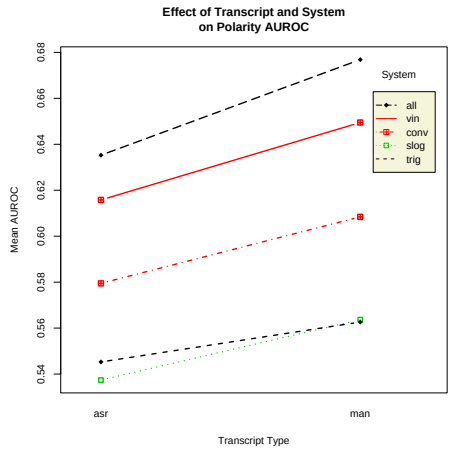

Figure 3: Polarity Scores

on the polarity task compared with the subjectivity task where there was little or no effect. This finding merits further research on identifying features to mitigate that impact for the second task. With the exception of the classifier combining all features, the VIN approach performed best on the noisy recognition output.

We have presented a novel approach for learning subjective patterns in spontaneous speech, significantly outperforming two baseline approaches. We have demonstrated that varying the instantiation level of trigram patterns can improve performance over the standard trigram approach. We also presented a method for unsupervised learning of subjective patterns from unlabeled web data.

8. References

[1] S. Raaijmakers, K. Truong, and T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proc. of EMNLP 2008, Honolulu, HI, USA*, 2008.

[2] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proc. of EMNLP 2003, Sapporo, Japan*, 2003.

[3] E. Riloff and W. Phillips, "An introduction to the sundance and autoslog systems," 2004. [Online]. Available: http://www.cs.utah.edu/ riloff/pdfs/official-sundance-tr.pdf

[4] 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, "The AMI meeting corpus: A pre-announcement," in *Proc. of MLMI 2005, Edinburgh, UK*, 2005, pp. 28–39.

[5] T. Hain, L. Burget, J. Dines, G. Garau, V. Wan, M. Karafiat, J. Vepa, and M. Lincoln, "The AMI system for transcription of speech in meetings," in *Proc. of ICASSP 2007*,, 2007, pp. 357–360.

[6] T. Wilson, "Annotating subjective content in meetings," in *Proc. of LREC 2008, Marrackech, Morocco*, 2008.

[7] G. Murray and G. Carenini, "Summarizing spoken and written conversations," in *Proc. of EMNLP 2008, Honolulu, HI, USA*, 2008.