Multimodal Subjectivity Analysis of Multiparty Conversation

Stephan Raaijmakers
TNO Information and Communication Technology
Delft, The Netherlands
stephan.raaijmakers@tno.nl

Khiet Truong
TNO Defense, Security and Safety
Soesterberg, The Netherlands
khiet.truong@tno.nl

Theresa Wilson
School of Informatics
University of Edinburgh
Edingburgh, UK
twilson@inf.ed.ac.uk

Abstract

We investigate the combination of several sources of information for the purpose of subjectivity recognition and polarity classification in meetings. We focus on features from two modalities, transcribed words and acoustics, and we compare the performance of three different textual representations: words, characters, and phonemes. Our experiments show that character-level features outperform word-level features for these tasks, and that a careful fusion of all features yields the best performance. [1]

1 Introduction

Opinions, sentiments and other types of subjective content are an important part of any meeting. Meeting participants express pros and cons about ideas, they support or oppose decisions, and they make suggestions that may or may not be adopted. When recorded and archived, meetings become a part of the organizational knowledge, but their value is limited by the ability of tools to search and summarize meeting content, including subjective content. While progress has been made on recognizing primarily objective meeting content, for example, information about the topics that are discussed (Hsueh and Moore, 2006) and who is assigned to work on given tasks (Purver et al., 2006), there has been fairly little work specifically directed toward recognizing subjective content.

In contrast, there has been a wealth of research over the past several years on automatic subjectivity and sentiment analysis in text, including on-line media. Partly inspired by the rapid growth of social media, such as blogs, as well as on-line news and reviews, researchers are now actively addressing a wide variety of new tasks, ranging from blog mining (e.g., finding opinion leaders in an on-line community), to reputation management (e.g. finding negative opinions about a company on the web), to opinion-oriented summarization and question answering. Yet many challenges remain, including how best to represent and combine linguistic information for subjectivity analysis. With the additional modalities that are present when working with face-to-face spoken communication, these challenges are even more pronounced.

The work in this paper focuses on two tasks: (1) recognizing subjective utterances and (2) discriminating between positive and negative subjective utterances. An utterance may be subjective because the speaker is expressing an opinion, because the speaker is discussing someone else's opinion, or because the speaker is eliciting the opinion of someone else with a question.

We approach the above tasks as supervised machine learning problems, with the specific goal of finding answers to the following research questions:

- Given a variety of information sources, such as text arising from (transcribed) speech, phoneme representations of the words in an utterance, and acoustic features extracted from

[1] This work was supported by the Dutch BSIK-project MultimediaN, and the European IST Programme Project FP6-0033812. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 466–474,
Honolulu, October 2008.

the audio layer, which of these sources are particularly valuable for subjectivity analysis in multiparty conversation?

- Does the combination of these sources lead to further improvement?

- What are the optimal representations of these information sources in terms of feature design for a machine learning component?

A central tenet of our approach is that subword representations, such as *character* and *phoneme n-grams*, are beneficial for the tasks at hand.

2 Subword Features

Previous work has demonstrated that textual units below the word level, such as character n-grams, are valuable sources of information for various text classification tasks. An example of character n-grams is the set of 3-grams {`#se, sen, ent, nti, tim, ime, men, ent, nt#, t#a, #an, ana, nal, aly, lys, ysi, sis, is#`} for the two-word phrase *sentiment analysis*. The special symbol # represents a word boundary. While it is not directly obvious that there is much information in these truncated substrings, character n-grams have successfully been used for fine-grained classification tasks, such as named-entity recognition (Klein et al., 2003) and subjective sentence recognition (Raaijmakers and Kraaij, 2008), as well as a variety of document-level tasks (Stamatatos, 2006; Zhang and Lee, 2006; Kanaris and Stamatatos, 2007).

The informativeness of these low-level features comes in part from a form of *attenuation* (Eisner, 1996): a slight abstraction of the underlying data that leads to the formation of string equivalence classes. For instance, words in a sentence will invariably share many character n-grams. Since every unique character n-gram in an utterance constitutes a separate feature, this leads to the formation of string classes, which is a form of abstraction. For example, Zhang and Lee (2006) investigate similar subword representations, called key substring group features. By compressing substrings in a corpus in a trie (a prefix tree), and labeling entire sets of distributionally equivalent substrings with one group label, an attenuation effect is obtained that proves very beneficial for a number of text classification tasks.

Aside from attenuation effects, character n-grams, especially those that represent word boundaries, have additional benefits. Treating word boundaries as characters captures micro-phrasal information: short strings that express the transition of one word to another. Stemming occurs naturally within the set of initial character n-grams of a word, where the suffix is left out. Also, some part-of-speech information is captured. For example, the modals *could*, *would*, *should* can be represented by the 4-gram, `ould`, and the set of adverbs ending in *-ly* can be represented by the 3-gram `ly#`.

A challenging thought is to extend the use of n-grams to the level of phonemes, which comprise the first symbolic level in the process of sound to grapheme conversion. If n-grams of phonemes compare favorably to word n-grams for the purpose of sentiment classification, then significant speedups can be obtained for online sentiment classification, since tokenization of the raw speech signal can make a halt at the phoneme level.

3 Data

For this work we use 13 meetings from the AMI Meeting Corpus (Carletta et al., 2005). Each meeting has four participants and is approximately 30 minutes long. The participants play specific roles (e.g., Project Manager, Marketing Expert) and together function as a design team. Within the set of 13 meetings, there are a total of 20 participants, with each participant taking part in two or three meetings as part of the same design team. Meetings with the same set of participants represent different stages in the design process (e.g., Conceptual Design, Detailed Design).

The meetings used in the experiments have been annotated for subjective content using the AMIDA annotation scheme (Wilson, 2008). Table 1 lists the types of annotations that are marked in the data. There are three main categories of annotations, *subjective utterances*, *subjective questions*, and *objective polar utterances*. A subjective utterance is a span of words (or possibly sounds) where a *private state* is being expressed either through choice of words or prosody. A private state (Quirk et al., 1985)

is an internal mental or emotional state, including opinions, beliefs, sentiments, emotions, evaluations, uncertainties, and speculations, among others. Although typically when a private state is expressed it is the private state of the speaker, as in example (1) below, an utterance may also be subjective because the speaker is talking about the private state of someone else. For example, in (2) the negative opinion attributed to the company is what makes the utterance subjective.

(1) Finding them is really a pain, you know
(2) The company's decided that teletext is outdated

Subjective questions are questions in which the speaker is eliciting the private state of someone else. In other words, the speaker is asking about what someone else thinks, feels, wants, likes, etc., and the speaker is expecting a response in which the other person expresses what he or she thinks, feels, wants, or likes. For example, both (3) and (4) below are subjective questions.

(3) Do you like the large buttons?
(4) What do you think about the large buttons?

Objective polar utterances are statements or phrases that describe positive or negative factual information about something without conveying a private state. The sentence *The camera broke the first time I used it* gives an example of negative factual information; generally, something breaking the first time it is used is not good.

For the work in this paper, we focus on recognizing subjectivity in general and distinguishing between positive and negative subjective utterances. *Positive subjective* utterances are those in which any of the following types of private states are expressed: agreements, positive sentiments, positive suggestions, arguing for something, beliefs from which positive sentiments can be inferred, and positive responses to subjective questions. *Negative subjective* utterances express private states that are the opposite of those represented by the positive subjective category: disagreements, negative sentiments, negative suggestions, arguing against something, beliefs from which negative sentiments can be inferred, and negative responses to subjective questions. Example (5) below contains two positive subjective utterances and one negative subjective utterance. Each annotation is indicated by a pair of angle brackets.

Table 1: AMIDA Subjectivity Annotation Types

Subjective Utterances
positive subjective
negative subjective
positive and negative subjective
uncertainty
other subjective
subjective fragment
Subjective Questions
positive subjective question
negative subjective question
general subjective question
Objective Polar Utterances
positive objective
negative objective

(5) Um ⟨**POS-SUBJ** it's very easy to use⟩. Um ⟨**NEG-SUBJ** but unfortunately it does lack the advanced functions⟩ ⟨**POS-SUBJ** which I I quite like having on the controls⟩.

The *positive and negative subjective* category is for marking cases of positive and negative subjectivity that are so closely interconnected that it is difficult or impossible to separate the two. For example, (6) below is marked as both positive and negative subjective.

(6) Um ⟨**POS-AND-NEG-SUBJ** they've also suggested that we um we only use the remote control to control the television, not the VCR, DVD or anything else⟩.

In (Wilson, 2008), agreement is measured for each class separately at the level of dialogue act segments. If a dialogue act overlaps with an annotation of a particular type, then the segment is considered to be labelled with that type. Table 2 gives the Kappa (Cohen, 1960) and % agreement for subjective segments, positive and negative subjective segments,[2] and subjective questions.

[2] A positive subjective segment is any dialogue act segment that overlaps with a positive subjective utterance or a positive-and-negative subjective utterance. The negative subjective segments are defined similarly.

Table 2: Interannotator agreement for the AMIDA subjectivity annotations

	Kappa	% Agree
Subjective	0.56	79
Pos Subjective	0.58	84
Neg Subjective	0.62	92
Subjective Question	0.56	95

4 Experiments

We conduct two sets of classification experiments. For the first set of experiments (Task 1), we automatically distinguish between subjective and non-subjective utterances. For the second set of experiments (Task 2), we focus on distinguishing between positive and negative subjective utterances. For both tasks, we use the manual dialogue act segments available as part of the AMI Corpus as the unit of classification. For Task 1, a segment is considered subjective if it overlaps with either a subjective utterance or subjective question annotation. For Task 2, the segments being classified are those that overlap with positive or negative subjective utterances. For this task, we exclude segments that are both positive and negative. Although limiting the set of segments to be classified to just those that are positive or negative makes the task somewhat artificial, it also allows us to focus in on the performance of features specifically for this task.[3] We use 6226 subjective and 8707 non-subjective dialog acts for Task 1 (with an average duration of 1.9s, standard deviation of 2.0s), and 3157 positive subjective and 1052 negative subjective dialog acts for Task 2 (average duration of 2.6s, standard deviation of 2.3s).

The experiments are performed using 13-fold cross validation. Each meeting constitutes a separate fold for testing, e.g., all the segments from meeting 1 make up the test set for fold 1. Then, for a given fold, the segments from the remaining 12 meetings are used for training and parameter tuning, with roughly a 85%, 7%, and 8% split between training, tuning, and testing sets for each fold. The assignment to training versus tuning set was random, with the only constraint being that a segment could only be in the tuning set for one fold of the data.

[3]In practice, this excludes about 7% of the positive/negative segments.

The experiments we perform involve two steps. First, we train and optimize a classifier for each type of feature using BoosTexter (Schapire and Singer, 2000) AdaBoost.MH. Then, we investigate the performance of all possible combinations of features using linear combinations of the individual feature classifiers.

4.1 Features

The two modalities that are investigated, prosodic, and textual, are represented by four different sets of features: prosody (PROS), word n-grams (WORDS), character n-grams (CHARS), and phoneme n-grams (PHONES).

Based on previous research on prosody modelling in a meeting context (Wrede and Shriberg, 2003) and on the literature in emotion research (Banse and Scherer, 1996) we extract PROS features that are mainly based on pitch, energy and the distribution of energy in the long-term averaged spectrum (LTAS) (see Table 3). These features are extracted at the word level and aggregated to the dialogue-act level by taking the average over the words per dialogue act. We then normalize the features per speaker per meeting by converting the raw feature values to z-scores ($z = (x - \mu)/\sigma$).

Table 3: Prosodic features used in experiments.

pitch	mean, standard deviation, minimum, maximum, range, mean absolute slope
intensity (energy)	mean, standard deviation, minimum, maximum, range, RMS energy
distribution energy in LTAS	slope, Hammerberg index, centre of gravity, skewness

The textual features, WORDS and CHARS, and the PHONES features are based on a manual transcription of the speech. The PHONES were produced through dictionary lookup on the words in the reference transcription. Both CHARS and PHONES representations include word boundaries as informative tokens. The textual features for a given segment are simply all the WORDS/CHARS/PHONES in that segment. Selection of n-grams is performed by the learning algorithm.

4.2 Single Source Classifiers

We train four single source classifiers using BoosTexter, one for each type of feature. For the WORDS, CHARS, and PHONES, we optimize the classifier by performing a grid search over the parameter space, varying the number of rounds of boosting (100, 500, 1000, 2000, 5000), the length of the n-gram (1, 2, 3, 4, 5), and the type of n-gram. BoosTexter can be run with three different n-gram configurations: n-gram, s-gram, and f-gram. For the default configuration (n-gram), BoosTexter searches for n-grams up to length n. For example, if $n = 3$, BoosTexter will consider 1-grams, 2-grams, and 3-grams. For the s-gram configuration, BoosTexter will in addition consider sparse n-grams (i.e., n-grams containing wildcards), such as *the * idea*. For the f-gram configuration, BoosTexter will only consider n-grams of a maximum fixed length, e.g., if $n = 3$ BoosTexter will only consider 3-grams. For the PROS classifier, only the number of rounds of boosting was varied. The parameters are selected for each fold separately; the parameter set that produces the highest subjective F_1 score on the tuning set for Task 1, and the highest positive subjective F_1 score for Task 2, is used to train the final classifier for that fold.

4.3 Classifier combination

After the single source classifiers have been trained, they have to be combined into an aggregate classifier. To this end, we decided to apply a simple linear interpolation strategy. Linear interpolation of models is the weighted combination of simple models to form complex models, and has its roots in generative language models (Jelinek and Mercer, 1980). (Raaijmakers, 2007) has demonstrated its use for discriminative machine learning.

In the present binary class setting, BoosTexter produces two decision values, one for every class. For every individual single-source classifier (i.e., PROS, WORDS, CHARS and PHONES), separate weights are estimated that are applied to the decision values for the two classes produced by these classifiers. These weights express the relative importance of the single-source classifiers.

The prediction of an aggregate classifier for a class c is then simply the sum of all weights for all participating single-source classifiers applied to the decision values these classifiers produce for this class. The class with the maximum score wins, just as in the simple non-aggregate case.

Formally, then, this linear interpolation strategy finds for n single-source classifiers n interpolation weights $\lambda_1, \ldots \lambda_n$ that minimize the empirical loss (measured by a loss function $\mathcal{L}$), with λ_j the weight of classifier j ($\lambda \in [0, 1]$), and $C_c^j(x_i)$ the decision value of class c produced by classifier j for datum x_i (a feature vector). The two classes are denoted with $0, 1$. The true class for datum x_i is denoted with $\hat{x}_i$. The loss function is in our case based on subjective F-measure (Task 1) or positive subjective F-measure (Task 2) measured on heldout development training and test data.

The aggregate prediction $\tilde{x_i}$ for datum x_i on the basis of n single-source classifiers then becomes

$$\tilde{x_i} = \arg\max_c (\sum_{j=1}^{n} \lambda_j \cdot C_{c=0}^j(x_i), \sum_{j=1}^{n} \lambda_j \cdot C_{c=1}^j(x_i)) \quad (1)$$

and the lambdas are defined as

$$\lambda_j^n = \arg \min_{\lambda_j^n \subset [0,1]} \sum_{i}^{k} \mathcal{L}(\hat{x}_i, \tilde{x}_i; \lambda_j, \ldots, \lambda_n) \quad (2)$$

The search process for these weights can easily be implemented with a simple grid search over admissible ranges.

In the experiments described below, we investigate all possible combinations of the four different sets of features (PROS, WORDS, CHARS, and PHONES) to determine which combination yields the best performance for subjectivity and subjective polarity recognition.

5 Results and Discussion

Results for the two tasks are given in Tables 4 and 5 and in Figures 1 and 2. We use two baselines, listed at the top of each table. The bullets in a given row indicate the features that are being evaluated for a given experiment. In Table 4, subjective F_1, recall, and precision are reported as well as overall accuracy. In Table 4, the F_1, recall, and precision scores are for the positive subjective class. All values in the tables are averages over the 13 folds.

Table 4: Results Task 1: Subjective vs. Non-Subjective.

	PROS	WORDS	CHARS	PHONES	F_1	PREC	REC	ACC
BASE-SUBJ	always chooses subjective class				60.3	43.4	100	43.4
BASE-RAND	randomly chooses a class based on priors				41.8	42.9	41.3	50.6
single	•				54.6	55.3	54.5	63.1
		•			60.5	68.5	54.5	71.0
			•		61.7	67.5	57.2	71.1
				•	60.3	66.4	55.5	70.2
double	•	•			63.9	72.1	57.6	73.4
	•		•		65.6	71.9	60.3	74.0
	•			•	64.6	72.3	58.4	73.7
		•	•		66.2	73.8	60.1	74.9
		•		•	65.2	73.2	58.8	74.3
			•	•	66.1	72.8	60.7	74.5
triple	•	•	•		66.5	74.3	60.3	75.1
	•	•		•	65.5	73.5	59.0	74.5
	•		•	•	66.5	73.3	60.8	74.8
		•	•	•	66.9	74.3	60.9	75.3
quartet	•	•	•	•	67.1	74.5	61.2	75.4

Table 5: Results Task 2: Positive Subjective vs. Negative Subjective.

	PROS	WORDS	CHARS	PHONES	F_1	PREC	REC	ACC
BASE-POS-SUBJ	always chooses positive subjective class				85.6	75.0	100	75.0
BASE-RAND	randomly chooses a class based on priors				75.1	74.4	76.1	62.4
single	•				84.8	74.8	98.1	73.9
		•			85.6	79.6	93.1	76.8
			•		85.9	81.9	90.5	78.0
				•	85.5	80.5	91.3	77.0
double	•	•			88.7	83.0	95.4	81.9
	•		•		88.7	83.1	95.1	81.8
	•			•	88.5	83.3	94.4	81.6
		•	•		89.5	84.2	95.7	83.3
		•		•	89.2	83.7	95.5	82.8
			•	•	89.0	84.2	94.6	82.6
triple	•	•	•		89.6	84.0	96.1	83.4
	•	•		•	89.3	83.6	95.8	82.8
	•		•	•	89.2	83.7	95.5	82.7
		•	•	•	89.8	84.4	96.0	83.8
quartet	•	•	•	•	89.9	84.4	96.2	83.8

It is quite obvious that the combination of different sources of information is beneficial, and in general, the more information the better the results. The best performing classifier for Task 1 uses all the features, achieving a subjective F_1 of 67.1. For Task 2, the best performing classifier also uses all the features, although it does not perform significantly better than the classifier using only WORDS, CHARS, and PHONES.[4] This classifier achieves a positive-subjective F_1 of 89.9.

We measured the effects of adding more information to the single source classifiers. These results are listed in Table 6. Of the various feature types, prosody seems to be the least informative for both subjectivity and polarity classification. In addition to producing the single-source classifier with the lowest performance for both tasks, Table 6 shows that when prosody is added, of all the features it is least likely to yield significant improvements.

[4]We measured significance with the non-parametric Wilcoxon signed rank test, $p < 0.05$.

Throughout the experiments, adding an additional type of textual feature always yields higher results. In all cases but two, these improvements are significant. The best performing of the features are the character n-grams. Of the single-source experiments, the character n-grams achieve the best performance, with significant improvements in F_1 over the other single-source classifiers for both Task 1 and Task 2. Also, adding character n-grams to other feature combinations always gives significant improvements in performance.

An obvious question that remains is what the effect is of classifier interpolation on the results. To answer this question, we conducted two additional experiments for both tasks. First, we investigated the performance of an uninterpolated combination of the four single-source classifiers. In essence, this combines the separate feature spaces without explicitly weighting them. Second, we investigated the results of training a single BoosTexter model using all the features, essentially merging all feature spaces

Table 6: Addition of features separately (for Task 1 and 2): '+' for a row-column pair (r, c) means that the addition of column feature c to the row features r significantly improved r's F_1; '-' indicates no significant improvement; 'X' means 'not applicable'

	+PROS		+ WORDS		+CHARS		+PHONES	
Task	1	2	1	2	1	2	1	2
PROS	X	X	+	+	+	+	+	+
WORDS	-	+	X	X	+	+	+	+
CHARS	-	+	-	+	X	X	-	+
PHONES	-	+	+	+	+	+	X	X
PROS+WORDS	X	X	X	X	+	+	+	+
PROS+CHARS	X	X	+	+	X	X	+	+
PROS+PHONES	X	X	+	+	+	+	X	X
WORDS+CHARS	+	-	X	X	X	X	+	+
WORDS+PHONES	+	-	X	X	+	+	X	X
CHARS+PHONES	+	+	+	+	X	X	X	X
PROS+WORDS+CHARS	X	X	X	X	X	X	+	+
PROS+WORDS+PHONES	X	X	X	X	+	+	X	X
PROS+CHARS+PHONES	X	X	+	+	X	X	X	X
WORDS+CHARS+PHONES	+	-	X	X	X	X	X	X

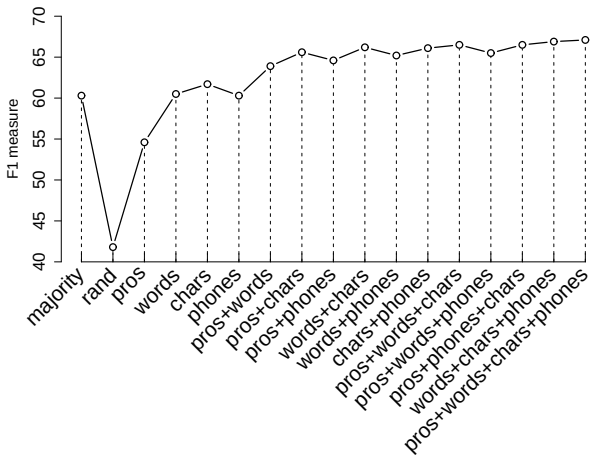

Figure 1: *Results (F_1) experiment 1: subjective vs. non-subjective.*

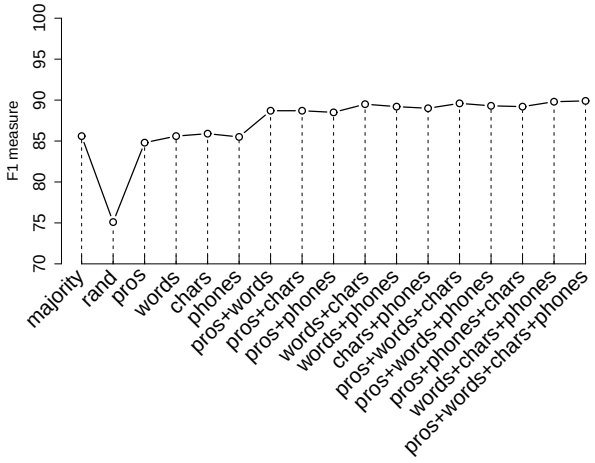

Figure 2: *Results (F_1) experiment 2: positive subjective vs. negative subjective.*

into one agglomerate feature space. The results for these experiments are given in Table 7, along with the results from the all-feature interpolated classification for comparison.

The results in Table 7 show that interpolation outperforms both the unweighted and single-model combinations for both tasks. For Task 1, the effect of interpolation compared to a single model is marginal (a .03 point difference in F_1). However, compared to the uninterpolated combination, interpolation gives a clear 3.1 points improvement of F_1. For Task 2, interpolation outperforms both the uninterpolated and single-model classifiers, with 2 and 3 points improvements in F_1, respectively.

6 Related Work

Previous work has demonstrated that textual units below the word level, such as character n-grams, are valuable sources of information. Character-level models have successfully been used for named-entity recognition (Klein et al., 2003), predicting authorship (Keselj et al., 2003; Stamatatos, 2006), text categorization (Zhang and Lee, 2006), web page genre identification (Kanaris and Stamatatos, 2007), and sentence-level subjectivity recognition (Raaijmakers and Kraaij, 2008) In spoken-language data, Hsueh (2008) achieves good results using chains of phonemes to automatically segment meetings according to topic. However, to the best of our knowledge there has been no investigation to date on the

Table 7: Results of interpolated classifiers compared to uninterpolated and single-model classifiers for all features.

Task	Combination	ACC	REC	PREC	F_1
1	interpolated	75.4	61.2	74.5	67.1
	uninterpolated	73.0	58.7	70.6	64.0
	single model	74.7	62.1	72.7	66.8
2	interpolated	83.8	96.2	84.4	89.9
	uninterpolated	79.8	98.0	79.7	87.9
	single model	79.5	91.0	83.3	86.9

combination of character-level, phoneme-level, and word-level models for any natural language classification tasks.

In text, there has been a significant amount of research on subjectivity and sentiment recognition, ranging from work at the phrase level to work on classifying sentences and documents. Sentence-level subjectivity classification (e.g., (Riloff and Wiebe, 2003; Yu and Hatzivassiloglou, 2003)) and sentiment classification (e.g., (Yu and Hatzivassiloglou, 2003; Kim and Hovy, 2004; Hu and Liu, 2004; Popescu and Etzioni, 2005)) is the research in text most closely related to our work. Of the sentence-level research, the most similar is work by Raaijmakers and Kraaij (2008) comparing word-spanning character n-grams to word-internal character n-grams for subjectivity classification in news data. They found that character n-grams spanning words perform the best.

Research on recognizing subjective content in multiparty conversation includes work by Somasundaran et al. (2007) on recognizing sentiments and arguing in meetings, work by Neiberg el al. (2006) on recognizing positive, negative, and neutral emotions in meetings, work on recognizing agreements and disagreements in meetings (Hillard et al., 2003; Galley et al., 2004; Hahn et al., 2006), and work by Wrede and Shriberg (2003) on recognizing meeting hotspots. Somasundaran et al. use lexical and discourse features to recognize sentences and turns where meeting participants express sentiments or arguing. They also use the AMI corpus in their work; however, the use of different annotations and task definitions makes it impossible to directly compare their results and ours. Neiberg et al. use acoustic–prosodic features (Mel-frequency Cepstral Coefficients (MFCCs) and pitch features) and lexical n-grams for recognizing emotions in the ISL Meeting Corpus (Laskowski and Burger, 2006).

Agreements and disagreements are a subset of the private states represented by the positive and negative subjective categories used in this work. To recognise agreements and disagreements automatically, Hillard et al. train 3-way decision tree classifiers (agreement, disagreement, other) using both word-based and prosodic features. Galley et al. model this task as a sequence tagging problem, and investigate whether features capturing speaker interactions are useful for recognizing agreements and disagreements. Hahn et al. investigate the use of contrast classifiers (Peng et al., 2003) for the task, using only lexical features.

Hotspots are places in a meeting in which the participants are highly involved in the discussion. Although high involvement does not necessarily equate subjective content, in practice, we expect more sentiments, opinions, and arguments to be expressed when participants are highly involved in the discussion. In their work on recognizing meeting hotspots, Wrede and Shriberg focus on evaluating the contribution of various prosodic features, ignoring lexical features completely. The results of their study helped to inform our choice of prosodic features for the experiments in this paper.

7 Conclusions

In this paper, we investigated the use of prosodic features, word n-grams, character n-grams, and phoneme n-grams for subjectivity recognition and polarity classification of dialog acts in multiparty conversation. We show that character n-grams outperform prosodic features, word n-grams and phoneme n-grams in subjectiviy recognition and polarity classification. Combining these features significantly improves performance. Comparing the additive value of the four information sources available, prosodic information seem to be least informative while character-level information indeed proves to be a very valuable source. For subjectivity recognition, a combination of prosodic, word-level, character-level, and phoneme-level information yields the best performance. For polarity classification, the best performance is achieved with a

combination of words, characters and phonemes.

References

R. Banse and K. R. Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, pages 614–636.

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. In *Proceedings of the Measuring Behavior Symposium on "Annotating and Measuring Meeting Behavior"*.

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

J. Eisner. 1996. An empirical comparison of probability models for dependency grammar. In *Technical Report IRCS-96-11, Institute for Research in Cognitive Science, University of Pennsylvania*.

M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of ACL*.

S. Hahn, R. Ladner, and M. Ostendorf. 2006. Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In *Proceedings of HLT/NAACL*.

D. Hillard, M. Ostendorf, and E. Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of HLT/NAACL*.

P. Hsueh and J. Moore. 2006. Automatic topic segmentation and lablelling in multiparty dialogue. In *Proceedings of IEEE/ACM Workshop on Spoken Language Technology*.

P. Hsueh. 2008. Audio-based unsupervised segmentation of meeting dialogue. In *Proceedings of ICASSP*.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD*.

F. Jelinek and R. L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings, Workshop on Pattern Recognition in Practice*, pages 381–397.

I. Kanaris and E. Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Proceedings of ICTAI*.

V. Keselj, F. Peng, N. Cercone, and C. Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of PACLING*.

S. Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of Coling*.

D. Klein, J. Smarr, H. Nguyen, and C.D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of CoNLL*.

K. Laskowski and S. Burger. 2006. Annotation and analysis of emotionally relevant behavior in the ISL meeting corpus. In *Proceedings of LREC 2006*.

D. Neiberg, K. Elenius, and K. Laskowski. 2006. Emotion recognition in spontaneous speech using GMMs. In *Proceedings of INTERSPEECH*.

K. Peng, S. Vucetic, B. Han, H. Xie, and Z Obradovic. 2003. Exploiting unlabeled data for improving accuracy of predictive data mining. In *Proceedings of ICDM*.

A. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP*.

M. Purver, P. Ehlen, and J. Niekrasz. 2006. Detecting action items in multi-party meetings: Annotation and initial experiments. In *Proceedings of MLMI*.

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

S. Raaijmakers and W. Kraaij. 2008. A shallow approach to subjectivity classification. In *Proceedings of ICWSM*.

S. Raaijmakers. 2007. Sentiment classification with interpolated information diffusion kernels. In *Proceedings of the First International Workshop on Data Mining and Audience Intelligence for Advertising (ADKDD'07)*.

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP*.

R. E. Schapire and Y. Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

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

E. Stamatatos. 2006. Ensemble-based author identification using character n-grams. In *Proceedings of TIR*.

T. Wilson. 2008. Annotating subjective content in meetings. In *Proceedings of LREC*.

B. Wrede and E. Shriberg. 2003. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proceedings of EUROSPEECH*.

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

D. Zhang and W. S. Lee. 2006. Extracting key-substring-group features for text classification. In *Proceedings of KDD*.