Proceedings of HLT-NAACL 2003
Short Papers , pp. 34-36
Edmonton, May-June 2003

DETECTION OF AGREEMENT vs. DISAGREEMENT IN MEETINGS: TRAINING WITH UNLABELED DATA

Dustin Hillard and **Mari Ostendorf**
University of Washington, EE
{hillard,mo}@ee.washington.edu

Elizabeth Shriberg
SRI International and ICSI
ees@speech.sri.com

Abstract

To support summarization of automatically transcribed meetings, we introduce a classifier to recognize agreement or disagreement utterances, utilizing both word-based and prosodic cues. We show that hand-labeling efforts can be minimized by using unsupervised training on a large unlabeled data set combined with supervised training on a small amount of data. For ASR transcripts with over 45% WER, the system recovers nearly 80% of agree/disagree utterances with a confusion rate of only 3%.

1 Introduction

Meetings are an integral component of life in most organizations, and records of meetings are important for helping people recall (or learn for the first time) what took place in a meeting. Audio (or audio-visual) recordings of meetings offer a complete record of the interactions, but listening to the complete recording is impractical. To facilitate browsing and summarization of meeting recordings, it is useful to automatically annotate topic and participant interaction characteristics. Here, we focus on interactions, specifically identifying agreement and disagreement. These categories are particularly important for identifying decisions in meetings and inferring whether the decisions are controversial, which can be useful for automatic summarization. In addition, detecting agreement is important for associating action items with meeting participants and for understanding social dynamics. In this study, we focus on detection using both prosodic and language cues, contrasting results for hand-transcribed and automatically transcribed data.

The agreement/disagreement labels can be thought of as a sort of speech act categorization. Automatic classification of speech acts has been the subject of several studies. Our work builds on (Shriberg et al., 1998), which showed that prosodic features are useful for classifying speech acts and lead to increased accuracy when combined with word based cues. Other studies look at prediction of speech acts primarily from word-based cues, using language models or syntactic structure and discourse history (Chu-Carroll, 1998; Reithinger and Klesen, 1997). Our work is informed by these studies, but departs significantly by exploring unsupervised training techniques.

2 Approach

Our experiments are based on a subset of meeting recordings collected and transcribed by ICSI (Morgan et al., 2001). Seven meetings were segmented (automatically, but with human adjustment) into 9854 total spurts. We define a 'spurt' as a period of speech by one speaker that has no pauses of greater than one half second (Shriberg et al., 2001). Spurts are used here, rather than sentences, because our goal is to use ASR outputs and unsupervised training paradigms, where hand-labeled sentence segmentations are not available.

We define four categories: positive, backchannel, negative, and other. Frequent single-word spurts (specifically, *yeah, right, yep, uh-huh,* and *ok*) are separated out from the 'positive' category as backchannels because of the trivial nature of their detection and because they may reflect encouragement for the speaker to continue more than actual agreement. Examples include:

Neg: (6%)	*"This doesn't answer the question."*
Pos: (9%)	*"Yeah, that sounds great."*
Back: (23%)	*"Uh-huh."*
Other: (62%)	*"Let's move on to the next topic."*

The first 450 spurts in each of four meetings were hand-labeled with these four categories based on listening to speech while viewing transcripts (so a sarcastic "yeah, right" is labeled as a disagreement despite the positive wording). Comparing tags on 250 spurts from two labelers produced a kappa coefficient (Siegel and Castellan, 1988) of .6, which is generally considered acceptable. Additionally, unlabeled spurts from six hand-transcribed training meetings are used in unsupervised training experiments, as described later. The total number of automatically labeled spurts (8094) is about five times the amount of hand-labeled data.

For system development and as a control, we use hand-transcripts in learning word-based cues and in training. We then evaluate the model with both hand-transcribed words and ASR output. The category labels from the

hand transcriptions are mapped to the ASR transcripts, assigning an ASR spurt to a hand-labeled reference if more than half (time wise) of the ASR spurt overlaps the reference spurt.

Feature Extraction. The features used in classification include heuristic word types and counts, word-based features derived from n-gram scores, and prosodic features.

Simple word-based features include: the total number of words in a spurt, the number of "positive" and "negative" keywords, and the class (positive, negative, backchannel, discourse marker, other) of the first word based on the keywords. The keywords were chosen based on an "effectiveness ratio," defined as the frequency of a word (or word pair) in the desired class divided by the frequency over all dissimilar classes combined. A minimum of five occurrences was required and then all instances with a ratio greater than .6 were selected as keywords.

Other word-based features are found by computing the perplexity (average log probability) of the sequence of words in a spurt using a bigram language model (LM) for each of the four classes. The perplexity indicates the goodness of fit of a spurt to each class. We used both word and class LMs (with part-of-speech classes for all words except keywords). In addition, the word-based LM is used to score the first two words of the spurt, which often contain the most information about agreement and disagreement. The label of the most likely class for each type of LM is a categorical feature, and we also compute the posterior probability for each class.

Prosodic features include pause, fundamental frequency (F0), and duration (Baron et al., 2002). Features are derived for the first word alone and for the entire spurt. Average, maximum and initial pause duration features are used. The F0 average and maximum features are computed using different methods for normalizing F0 relative to a speaker-dependent baseline, mean and max. For duration, the average and maximum vowel duration from a forced alignment are used, both unnormalized and normalized for vowel identity and phone context. Spurt length in terms of number of words is also used.

Classifier design and feature selection. The overall approach to classifying spurts uses a decision tree classifier (Breiman et al., 1984) to combine the word based and prosodic cues. In order to facilitate learning of cues for the less frequent classes, the data was upsampled (duplicated) so that there were the same number of training points per class. The decision tree size was determined using error-based cost-complexity pruning with 4-fold cross validation. To reduce our initial candidate feature set, we used an iterative feature selection algorithm that involved running multiple decision trees (Shriberg et al., 2000). The algorithm combines elements of brute-force search (in a leave-one-out paradigm) with previously determined heuristics for narrowing the search space. We used entropy reduction of the tree after cross-validation as a criterion for selecting the best subtree.

Unsupervised training. In order to train the models with as much data as possible, we used an unsupervised clustering strategy for incorporating unlabeled data. Four bigram models, one for each class, were initialized by dividing the hand transcribed training data into the four classes based upon keywords. First, all spurts which contain the negative keywords are assigned to the negative class. Backchannels are then pulled out when a spurt contains only one word and it falls in the backchannel word list. Next, spurts are selected as agreements if they contain positive keywords. Finally, the remaining spurts are associated with the "other" class.

The keyword separation gives an initial grouping; further regrouping involves unsupervised clustering using a maximum likelihood criterion. A preliminary language model is trained for each of the initial groups. Then, by evaluating each spurt in the corpus against each of the four language models, new groups are formed by associating spurts with the language model that produces the lowest perplexity. New language models are then trained for the reorganized groups and the process is iterated until there is no movement between groups. The final class assignments are used as "truth" for unsupervised training of language and prosodic models, as well as contributing features to decision trees.

3 Results and Discussion

Hand-labeled data from one meeting is held out for test data, and the hand-labeled subset of three other meetings are used for training decision trees. Unlabeled spurts taken from six meetings, different from the test meeting, are used for unsupervised training. Performance is measured in terms of overall 3-way classification accuracy, merging the backchannel and agreement classes. The overall accuracy results can be compared to the "chance" rate of 50%, since testing is on 4-way upsampled data. In addition, we report the confusion rate between agreements and disagreements and their recovery (recall) rate, since these two classes are most important for our application.

Results are presented in Table 1 for models using only word-based cues. The simple keyword indicators used in a decision tree give the best performance on hand-transcribed speech, but performance degrades dramatically on ASR output (with WER $> 45\%$). For all other training conditions, the degradation in performance for the system based on ASR transcripts is not as large, though still significant. The system using unsupervised training clearly outperforms the system trained only on a small amount of hand-labeled data. Interestingly, when

Features	Hand Transcriptions			ASR Transcriptions		
	Overall Accuracy	A/D Confusion	A/D Recovery	Overall Accuracy	A/D Confusion	A/D Recovery
Keywords	82%	2%	87%	61%	7%	53%
Hand Trained LM	71%	13%	74%	64%	10%	67%
Unsupervised LM	78%	10%	81%	67%	14%	70%
All word based	79%	8%	83%	71%	3%	78%

Table 1: Results for detection with different classifiers using word based features.

the keywords are used in combination with the language model, they do provide some benefit in the case where the system uses ASR transcripts.

The results in Table 2 correspond to models using only prosodic cues. When these models are trained on only a small amount of hand-labeled data, the overall accuracy is similar to the system using keywords when operating on the ASR transcript. Performance is somewhat better than chance, and use of hand vs. ASR transcripts (and associated word alignments) has little impact. There is a small gain in accuracy but a large gain in agree/disagree recovery from using the data that was labeled via the unsupervised language model clustering technique. Unfortunately, when the prosody features are combined with the word-based features, there is no performance gain, even for the case of errorful ASR transcripts.

Transcripts Train/Test	Overall Accuracy	A/D Confusion	A/D Recovery
Hand/Hand	62%	17%	62%
Unsup./Hand	66%	13%	72%
Hand/ASR	62%	16%	61%
Unsup./ASR	64%	14%	75%

Table 2: Results for classifiers using prosodic features.

4 Conclusion

In summary, we have described an approach for automatic recognition of agreement and disagreement in meeting data, using both prosodic and word-based features. The methods can be implemented with a small amount of hand-labeled data by using unsupervised LM clustering to label additional data, which leads to significant gains in both word-based and prosody-based classifiers. The approach is extensible to other types of speech acts, and is especially important for domains in which very little annotated data exists. Even operating on ASR transcripts with high WERs (45%), we obtain a 78% rate of recovery of agreements and disagreements, with a very low rate of confusion between these classes. Prosodic features alone provide results almost as good as the word-based models on ASR transcripts, but no additional benefit when used with word-based features. However, the good performance from prosody alone offers hope for performance gains given a richer set of speech acts with more lexically ambiguous cases (Bhagat et al., 2003).

Acknowledgments

This work is supported in part by the NSF under grants 0121396 and 0619921, DARPA grant N660019928924, and NASA grant NCC 2-1256. Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these agencies.

References

D. Baron et al. 2002. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In *Proc. ICSLP*, pages 949–952.

S. Bhagat, H. Carvey, and E. Shriberg. 2003. Automatically generated prosodic cues to lexically ambiguous dialog acts in multi-party meetings. In *ICPhS*.

L. Breiman et al. 1984. *Classification And Regression Trees*. Wadsworth International Group, Belmont, CA.

J. Chu-Carroll. 1998. A statistical model for discourse act recognition in dialogue interactions. In *Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium*, pages 12–17.

N. Morgan et al. 2001. The meeting project at ICSI. In *Proc. Conf. on Human Language Technology*, pages 246–252, March.

N. Reithinger and M. Klesen. 1997. Dialogue act classification using language models. In *Proc. Eurospeech*, pages 2235–2238, September.

E. Shriberg et al. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech, 41(3–4)*, pages 439–487.

E. Shriberg et al. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154, September.

E. Shriberg et al. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proc. Eurospeech*, pages 1359–1362.

S. Siegel and J. Castellan. 1988. *Nonparametric Statistics For the Behavioral Sciences*. McGraw-Hill Inc., New York, NY, second edition edition.