AUTOMATIC DIALOG ACT SEGMENTATION AND CLASSIFICATION IN MULTIPARTY MEETINGS

Jeremy Ang[1], *Yang Liu*[1,2], *Elizabeth Shriberg*[1,3]

[1]International Computer Science Institute, [2]Purdue University, [3]SRI International, USA
{jca,yangl,ees}@icsi.berkeley.edu

ABSTRACT

We explore the two related tasks of dialog act (DA) *segmentation* and DA *classification* for speech from the ICSI Meeting Corpus. We employ simple lexical and prosodic knowledge sources, and compare results for human-transcribed versus automatically recognized words. Since there is little previous work on DA segmentation and classification in the meeting domain, our study provides baseline performance rates for both tasks. We introduce a range of metrics for use in evaluation, each of which measures different aspects of interest. Results show that both tasks are difficult, particularly for a fully automatic system. We find that a very simple prosodic model aids performance over lexical information alone, especially for segmentation. Both tasks, but particularly word-based segmentation, are degraded by word recognition errors. Finally, while classification results for meeting data show some similarities to previous results for telephone conversations, findings also suggest a potential difference with respect to the effect of modeling DA context.

1. INTRODUCTION

A growing interest in spoken language technology research is the automatic processing of multiparty meetings. Common goals include automatic browsing, retrieval, question answering, and summarization [1,6,19]. Such tasks require more than just a stream of recognized words. Parsing continuous speech into dialog acts (DAs), such as statements, questions, backchannels, and so on, is a useful step in answering questions like "Who asked what to whom?, "Where did participants disagree?", or "Who interrupted whom?".

Past work has addressed automatic dialog act classification in various domains, but has tended to "cheat" by using a segmentation (into dialog acts) that is given by human labelers [3,5,8,13,16,17,18]. In this paper, we explore both segmentation and dialog act classification for audio recordings from the publicly available ICSI Meeting Corpus [7]. Work on automatic processing of multiparty meetings has only recently begun (e.g., [3,8]), and has proven challenging because of the presence of multiple speakers, frequent speaker overlap, and high rates of both self- and other-interruptions. For each task, we compare results using different knowledge sources and using human-generated versus fully automatic processing. We compare results using various metrics, some newly proposed in this work. Each metric is intended to convey different but useful information.

2. METHOD

2.1. Data

The ICSI Meeting Recorder corpus [7] includes 75 naturally occurring meetings containing roughly 72 hours of multitalker speech data and associated human-generated word-level transcripts. The audio is recorded by close-talking microphones (used here) as well as table-top microphones. The corpus was hand-annotated for dialog acts (and their boundaries) as described in detail in [4,15]. For this paper we grouped labels into five broad categories: statements, questions, backchannels, fillers, and disruptions. The ICSI DA annotations also provide an option for breaking versus not breaking at boundaries like that at the "|" in "yeah | I agree", when produced as one prosodic unit. In this work we chose to break at these boundaries, but results for a non-breaking analysis were fairly similar.

2.2. Speech recognition and alignment

The meeting data was automatically recognized by a version of the system that SRI used in the DARPA 2003 Rich Transcription Evaluation [11]. The recognizer models used here were trained on conversational telephone speech (with some added broadcast and Web data for language modeling); they were not trained on any meeting data. The system was simplified to run quickly, and yielded an average word error rate of 39% on the entire corpus, and 32% on native speakers of American English.

2.3. Features

We extracted various lexical and prosodic features (based on recognition and forced alignment information) to use in the segmentation and classification tasks. As lexical features, we used word n-gram information for segmentation, and various lexical cues for classification. As prosodic features, we used pause information for segmentation, and a larger set of features for classification. The latter included pause, duration, pitch, energy, and spectral tilt features, many normalized by speaker-specific statistics and/or phonetic context.

3. DIALOG ACT SEGMENTATION

3.1. Segmentation metrics

To investigate the segmentation performance of different models, we used three different metrics. The first metric (NIST-SU)

```
Reference: A|B C D|E F G|H|I J|
   System: A|B C D E F|G H|I J|
  NIST-SU:  C     E   E E C   C
 Boundary:  C C C E C E E C C C
 "Strict": C E E E E E E E C C
```

Metric	Errors	Ref. Units	Error
NIST-SU	2 misses, 1 FA	5 DAs	60%
Boundary	3 boundary err.	10 boundaries	30%
"Strict"	7 match errors	10 words	70%

Figure 1: Comparison of segmentation metrics. Top: Example reference and system pair and their errors using the different metrics. 'E' = error/incorrect decision. 'C' = correct decision. '|' = boundary 'A'-'J' = words. Table: Summary of errors according to the different metrics.

parallels the SU (sentence-like unit) evaluation metric given by NIST in the EARS MDE evaluations [12]. This associates an end marker for the unit in question (DA units here, SUs in MDE) with a word in the word stream. Segmentation error is determined by finding the misses and false alarms and dividing by the total number of reference units. This is the primary metric we use in our comparisons. A similar metric (Boundary-based) divides instead by the total number of reference words rather than the number of reference units. This results in a much lower error value. A third metric ("Strict") measures the percentage of words that were placed in a segment perfectly identical to that in the reference. In other words, if an output segment perfectly matches a corresponding reference segment on the word level, each word in that segment is counted as correct. All other words are counted as incorrect. We call this the "Strict" metric because of its stringent requirement for DA segmentation.

Figure 1 illustrates the three metrics. Notice that the reference and system words match exactly, because our focus is segmentation performance: we want to know, given the words (whether true or recognized), how well we can place segment boundaries within that word stream.

When assessing the performance of automatic segmentation and classification jointly, we use a modified "Strict" metric and a "Lenient" metric. The "Strict" metric is modified so that we add the constraint that the DA class must be correct also. This means that only the words that are in a correct segment and are labeled with the correct class are counted as correct. The "Lenient" metric bases classification accuracy on the number of words that are assigned the correct class, regardless of whether the segment boundaries are correct. See Figure 2 for an example.

3.2. Segmentation techniques

We investigated two simple techniques for DA segmentation, using a split of 51 meetings for training, 11 meetings for development, and 11 meetings for testing. First, we used a decision tree (DT) with only a simple pause feature (pause length between contiguous words from the same speaker), to act as a classifier that estimates the conditional probability of the boundary class (DA or not) at each word boundary. The prosodic classifier employs bagging [10] to better estimate posterior probabilities. It is likely, based on results in [9] for

```
Reference: B|S  S  S|D  D  D|B|Q  Q|
   System: S|S  S  S  S  S|B  B|Q  Q|
 "Lenient": E  C  C  C  E  E  E  C  C  C
  "Strict": E  E  E  E  E  E  E  E  C  C
```

Metric	Errors	Ref. Units	Error
"Lenient"	4 match errors	10 words	40%
"Strict"	8 match errors	10 words	80%

Figure 2: Comparison of metrics for segmentation and classification. 'B','S','D','Q' = differently classified words. 'E' = error/incorrect decision. 'C' = correct decision. '|' = boundary.

conversational telephone speech, that using a larger set of prosodic features extracted around each word would benefit performance here, but this requires a further effort and thus awaits future work. Our second approach used a hidden-event language model [14]. For a sequence $W_1\ E_1\ W_2\ E_2 \ldots W_n$, where the DA events E_i are included as pseudo-word tokens, the language model (LM) models the joint probability of the word and event sequence. We also tried an HMM-based combination of the two approaches above, following the framework in [9]. A forward-backward algorithm was used to find the most likely event at each inter-word boundary.

3.3. Segmentation results and discussion

Table 1 summarizes the results of our segmentation experiments. According to human segmentation, about 16.2% of the words (in both conditions) are at the end of a DA. In the ASR condition, the recognition words are dropped into the given segments according to the midpoint of each word. These words then inherit the DA label of the segment that they fall into. Because the alignment is time based, occasionally DA segments have no

	Pause DT	HE-LM	Combination
Ref	56.02	45.92	34.35
ASR	58.25	61.81	48.60

Table 1: DA Segmentation error rate (in %) using NIST-SU metric. 'Pause DT' = decision tree with pause. 'HE-LM' = hidden-event LM. 'Combination' = HMM combination.

recognition words. These segments are included in the errors and the total number of units in the NIST-SU metric. The Boundary-based metric counts these units as errors, too, while the "Strict" metric ignores them (since it focuses on words).

In the reference words condition (Ref), the combination system gives a 10% absolute improvement over the HE-LM and 20% over the Pause DT. Using recognition words (ASR) instead of reference words, the HE-LM suffers much more (16%) than does the prosodic DT (2%). In fact, the pause model alone is better than the language model alone in the ASR condition. Despite this poor performance of the HE-LM alone in the ASR condition, however, a roughly 10% improvement is found over the Pause DT alone by combining the two models, indicating that both make important contributions.

4. DIALOG ACT CLASSIFICATION

We grouped the DA labels given in the MRDA corpus into five classes (see Sect. 2.1), which we use to assess DA classification performance by finding classification error rates, that is, the number of incorrectly classified segments divided by the total number of segments. To develop and test the models described below, we use the same data split presented in Section 3.2.

4.1. Classification techniques

To perform DA classification, we determine the class of each given DA unit by using a maximum entropy (Maxent) classifier. This approach maximizes the conditional likelihood over the training data, and thus explicitly optimizes discrimination of correct from incorrect class labels for each unit. It also provides a principled way to incorporate many correlated features. See [9] for more details.

In the Maxent classifier, we use the following textual features: the length of unit, the first two words (after removing filler words), the final two words, and the initial word of the following DA. To take advantage of prosody, we generated a variety of features and fed them to a decision tree. The features most used include the number of speech phones, same-speaker and different-speaker DA pause gaps, normalized last pitch and average pitch, and DA duration. We added the posterior probabilities given by the tree as features in the Maxent model. Because the Maxent approach uses binary features conveniently, we cumulatively binned the posterior probabilities into multiple binary features.

4.2. Classification results and discussion

Classification results using human segmentations are given in

	Chance	Word	Word + DT posteriors
Ref	44.92	20.47 (0.66)	18.82 (0.70)
ASR	42.93	27.67 (0.51)	26.04 (0.56)

Table 2: DA classification error rate (in %) based on *human segmentations*. Kappa (a measure of agreement) is given in parentheses. All results are significant with a Sign test. Word = word-based features only. Word + DT posteriors = word-based features and DT binned posterior probabilities.

Table 2, which reports error rate for the five-way task. Chance error rate is different in the two conditions because of the segments that contain no ASR words in the ASR condition (see Sect. 3.3). When the binned tree posteriors are added to the word-based features, error rate decreases slightly in both the Reference and ASR conditions. In the ASR condition, performance consistently drops about 7% absolute when compared to the Reference condition. The results show the importance of having correct words for the classification task.

Kappa, a measure of agreement that adjusts for the level of agreement expected by chance, is also included in Table 2 because it is commonly reported. We note however that for skewed class distributions, there is some debate over the measure's interpretation. We find that adding tree posteriors increases Kappa by about 0.05, and that the ASR condition decreases Kappa by about 0.15 from the Reference condition.

4.3. Effect of DA context

We examined whether DA context information would aid performance on the classification task. Starting with a baseline result of 18.82% for Reference and 26.04% for ASR (see last column of Table 2), we added the knowledge of the true DA for the previous and following segments for both the *same speaker* and the *closest previous and following different speaker* (as estimated from an ordering by DA start times in the forced alignment). With the DA context, we obtain error rates of 18.11% (Ref) and 25.46% (ASR). This slight improvement is also seen in the experiment using only word-based features.

4.4. Automatic segmentation and classification

A final experiment involved using our best automatic segmentation as input for DA classification, so that we could investigate its effect on DA classification. The results are summarized in Table 3. We see a 5% error increase (when using the "Lenient" metric) due to automatic segmentation. The main observation, however, is that over 75% of the words from this system have a segmentation or classification error (or both). When the automatic segmentation is correct (using the "Strict" metric), 29.62% of the segments (21.05% of the words) are incorrectly classified.

	"Lenient"	"Strict" with DA label
Using Human Segments	19.60	19.60
Using Automatic Segments	25.13	75.39

Table 3: Classification error rate (in %), using different inputs for segmentation. ("Combination" segmenter and "Word-based features only" classifier used.)

4.5. Comparisons with previous research

Although there is no directly comparable work, it is interesting to draw a comparison to previously reported DA classification results using data from Switchboard (SWB)—both in terms of error rates and in terms of the effect of context on classification. Table 4 summarizes results of the present study along with results reported in [13]. Note that in addition to the difference in data sources, the studies also differ somewhat in terms of the target class breakdown. In both studies, the set of classes includes statement, question, backchannel, and disruption/incomplete. The studies differ on remaining classes, and [13] used 7 rather than 5 classes; thus only general comparisons can be made. The trends of the two approaches

		Chance	Words	Words+DT
Ref	Meetings	80.00	31.55	27.03
	SWB	85.71	29.70	28.86
ASR	Meetings	80.00	44.26	37.89
	SWB	85.71	41.40	39.88

Table 4: Classification error rate (in %) using equal class priors (after downsampling) in this work and for SWB [13].

show similarity in the effect of ASR words over true words. However, the present study finds more gain from prosodic information than did [13], despite the simpler prosodic features used in the current work. Further study should investigate whether this reflects a fundamental corpus difference in speakers' use of prosody, versus an effect of differing class definitions, bandwidth conditions, or modeling improvements.

When we compare the two studies in terms of the effect of adding DA context, we find a major difference. Whereas [16] found a large improvement from modeling surrounding DAs, our results show little improvement (only 0.6-1.0%) from this information. Recent results reported in [8] for the same corpus also find a lack of improvement. The difference between these results for meetings and results for telephone conversations [16] is even more striking if one considers that this study used true DAs while [16] used automatically obtained DAs. Further investigation is necessary to determine with this result reflects a stylistic difference between telephone conversations and meetings, versus a difference in the DA classes compared or in the modeling approach (maximum entropy here; an HMM framework in [16]) used to capture DA context.

5. SUMMARY AND DISCUSSION

We explored both DA segmentation and classification using simple lexical and prosodic knowledge sources. Results show that both tasks are difficult, particularly for a fully automatic system. We find that prosodic information aids performance over lexical information alone. Both tasks are impacted by word recognition errors, with more severe degradation occurring for lexical- than for prosodic-based segmentation models. Rough comparisons of DA classification results for meeting data show some similarities to previous results for telephone conversations. Unlike previous work however, meetings show little gain from the modeling of DA context.

As noted earlier, this study aimed to provide baseline performance rates for a new domain; considerable further work is in order. One long-term goal is to model these two tasks jointly, along the lines of [20], which addressed joint modeling but for a much simpler domain. The present study used only very simple features, particularly for prosody. Results could show considerable improvement from adding in a larger set of prosodic features (e.g., pitch and energy patterns) in the vicinity of each word—a feasible next step. This would allow comparison to the problem of SU (sentence-like unit) detection in the EARS program [9], since DA segments here have close overlap with the SU segments used in that effort [12]. Further research is in order in the future to learn about the effect of domain on the degradation from errorful word recognition and on the contribution from lexical, prosodic, and context-based knowledge sources. Finally, a long-term goal is to assess the contribution of automatic dialog act modeling to downstream tasks in automatic meeting applications.

6. ACKNOWLEDGMENTS

We thank Sonali Bhagat and Raj Dhillon for advice on DA annotations, and Andreas Stolcke for meeting recognition output and alignment. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-34) and by NSF IIS-0121396, NSF IRI-9619921, the Swiss National Science Foundation through IM2, and DARPA under contract MDA972-02-C-0038. The views in this paper are those of the authors and do not represent the views of the funding agencies. Distribution is unlimited.

7. REFERENCES

[1] Armstrong, S., et al., "Natural Language Queries on Natural Language Data: a Database of Meeting Dialogues," *Proc. NLDB*, Burg/Cottbus, Germany, 2003.

[2] Carletta, J., "Assessing Agreement on Classification Tasks: The Kappa Statistic," *Computational Linguistics*, 22(2), 249-254, 1996.

[3] Clark, A. and Popescu-Belis, A., "Multi-level Dialogue Act Tags," *Proc. SIGDIAL*, Cambridge, MA, 163-170, 2004.

[4] Dhillon, R., et al., "Meeting Recorder Project: Dialog Act Labeling Guide," *ICSI Technical Report TR-04-002*, International Computer Science Institute, 2004.

[5] Finke, M., et al., "CLARITY: Inferring Discourse Structure from Speech," *AAAI Spring Symposium Series*, Stanford University, CA, March 23-25, 1998.

[6] Galley, M., et al., "Discourse Segmentation of Multi-Party Conversation," *Proc. ACL*, 562-569, 2003.

[7] Janin, A., et al., "The ICSI Meeting Corpus," *Proc. ICASSP*, 2003.

[8] Ji, G. and Bilmes, J., "Dialog Act Tagging Using Graphical Models," *Proc. ICASSP*, Philadelphia, 2005.

[9] Liu, Y., et al., "The ICSI-SRI-UW Metadata Extraction System," *Proc. ICSLP*, Jeju, Korea, 2004.

[10] Liu, Y., et al., "Using Machine Learning to Cope with Imbalanced Classes in Natural Speech: Evidence from Sentence Boundary and Disfluency Detection," *Proc. ICSLP*, Jeju, Korea, 2004.

[11] NIST Website, *RT-03 Spring Workshop Presentations*, http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/.

[12] NIST Website, *RT-03 Fall Rich Transcription*, http://www.nist.gov/speech/tests/rt/rt2003/fall/.

[13] Shriberg, E., et al., "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech* 41(3-4), 439-487, 1998.

[14] Shriberg, E., et al., "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics," *Speech Communication*, 32(1-2), 127-154, 2000.

[15] Shriberg, E., et al., "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus," *Proc. SIGDIAL*, Cambridge, MA, April-May 2004.

[16] Stolcke, A., et al., "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistics*, 26(3), 339-371, 2000.

[17] Venkataraman, A., et al., "Automatic Dialog Act Labeling with Minimal Supervision," *Proc. 9th Australian International Conference on Speech Science and Technology*, Melbourne, Australia, 2002.

[18] Venkataraman, A., et al., "Training a Prosody Based Dialog Act Tagger from Unlabeled Data," *Proc. ICASSP*, Hong Kong, 2003.

[19] Waibel, A., et al., "Advances in Automatic Meeting Record Creation and Access," *Proc. ICASSP*, 2001.

[20] Warnke, V., et al., "Integrated Dialog Act Segmentation and Classification Using Prosodic Features and Language Models," *Proc. Eurospeech*, 207-210, September 1997.