

# Toward Joint Segmentation and Classification of Dialog Acts in Multiparty Meetings

Matthias Zimmermann<sup>1</sup>, Yang Liu<sup>1</sup>, Elizabeth Shriberg<sup>1,2</sup>, Andreas Stolcke<sup>1,2</sup>

<sup>1</sup> International Computer Science Institute

<sup>2</sup> SRI International, USA

{zimmerma,yangl,ees,stolcke}@icsi.berkeley.edu

**Abstract.** We present baseline results for the joint segmentation and classification of dialog acts (DAs) of the ICSI Meeting Corpus. Two simple approaches based on word information are investigated and compared with previous work on the same task. We also describe several metrics to assess the quality of the segmentation alone as well as the joint performance of segmentation and classification of DAs.

## 1 Introduction

As spoken language technology research moves toward more complex domains, further processing of the stream of words provided by a recognizer is often necessary. To support higher-level tasks such as information retrieval and summarization [1,2], the input speech signal must be segmented into meaningful units, for example dialog acts (DAs). The five DA types used in this work are statements, questions, backchannels, floorgrabbers, and disruptions. The task we investigate here is how to split a stream of words into non-overlapping segments of text and assigning mutually exclusive DA types mentioned above to these segments. While this task description suggests a sequential solution, an approach based on joint segmentation and classification most likely performs best because knowledge of the classification might also improve the segmentation. We use the term *joint segmentation and classification* for systems that do not implement this task in the form of two independent modules running in sequence but produce their final result by taking into account information from both the segmentation and the classification.

Previous work mainly concentrated on either the segmentation of speech into sentences [3,4] or the classification of already segmented text into various sets of DA types [5,6,7,8]. For automatic segmentation of speech it remains unclear how well a subsequent component can handle segmentation errors. For the latter case, the classification of DAs, it is typically assumed that the true segmentation boundaries are provided. As a consequence, a degradation of the performance due to imperfect segmentation boundaries must be expected. To provide more realistic results for the task of automatic segmentation and classification of DAs, a sequential approach is described in [9]. Results for the related task of subtype detection for sentence-like units (statements, backchannels, questions, or

incomplete) for broadcast news and spontaneous telephone conversations were reported in [10]. In this paper we make a first attempt toward joint segmentation and classification of DAs on the ICSI (MRDA) Corpus [11].

## 2 Methodology and Performance Metrics

For the joint segmentation and classification of DAs, two simple techniques are investigated in this paper. The first technique is based on a hidden-event language model (HE-LM) described in [12], and the second relies on a hidden Markov model (HMM) based tagger. The HE-LM is frequently used for detection of sentence boundaries [9,4], where after each word the model predicts a nonboundary or a sentence boundary event. In contrast, we use the HE-LM to predict not only a DA boundary or a nonboundary event, but the type of the DA boundary at the same time. This extension to [12] was also used in [3] to detect sentence boundaries and 5 different types of disfluencies. In our case the DA-specific boundary posterior probabilities are computed using forward-backward dynamic programming. The model can be seen as an  $n^{\text{th}}$  order HMM in which the word/event pairs correspond to states and the words to observations, with the transition probabilities given by the  $n$ -gram LM.

The second technique relies on the concept of disambiguation of words, which is widely used in the form of HMM-based part of speech (POS) taggers. In our case a conventional  $n$ -gram LM is used to model the priors of sequences  $((w_1, d_1), (w_2, d_2), \dots (w_n, d_n))$ . The  $w_i$  are the words from the lexicon provided by the speech to text (STT) system and the  $d_i$  represent specific DAs, such as statements, questions, etc. To model segmentation boundaries between words of the same DA type, the lexicon of the DA types also includes special symbols indicating the first word of a new DA (e.g. the symbol  $S+$  tags the first word of a statement, while the other words of a statement are tagged with an  $S$ ). Mapping probabilities  $p(w|w, d)$  are then estimated from the LM training corpus. Note that compared to the conventional way of POS tagging based on HMMs, our model states do not correspond to the tags only, but to joint events of words and tags. Simple add-1 smoothing is applied to account for unseen word-DA combinations. Finally, the sequence  $((w_1, d_1), (w_2, d_2), \dots)$  with the highest posterior probability is computed for a provided input sequence  $(w_1, w_2, \dots)$ .

To assess the performance of joint segmentation and classification of DAs, a number of measures have been proposed. We first describe two metrics for the measurement of the segmentation performance before metrics for the joint segmentation and classification of DAs are explained. The NIST-SU metric was used to report the segmentation performance in previous work [9] and has been defined by NIST for the EARS MDE evaluations [13]. As this measure takes into account only the local correspondence of reference boundaries and boundaries computed by the system, a direct interpretation of the resulting error rates is not always easy. To provide a more intuitive metric we propose the DA segmentation error rate (DSER), which measures the percentage of wrongly segmented DA segments. A DA is considered to be mis-segmented if and only if its left and/or

Reference	S Q.Q.Q.Q S.S.S B S.S
System	S Q S Q.Q D.D.D S.S S
NIST-SU	C E E C C E E C
DSER	C  E   C  E  E

Metric	Errors	Reference Units	Error Rate
NIST-SU	3 FA, 1 miss	5 boundaries	80%
DSER	3 match errors	5 DAs	60%

**Fig. 1.** Two metrics for the assessment of segmentation performance. S, Q, B, and D represent words of statements, questions, backchannels, and disruptions. DA boundaries are indicated using the symbol ‘|’, while ‘.’ is used for nonboundaries. Errors and correct cases are indicated using letters E and C.

Reference	S Q.Q.Q.Q S.S.S B S.S
System	S Q S Q.Q D.D.D S.S S
NIST-SU	C E E C E E E C
Lenient	C C E C C E E E C C
Strict	C E E E E E E E E E
DER	C  E   E  E  E

Metric	Errors	Reference	Error Rate
NIST-SU	1 sub., 3 FAs, 1 miss	5 boundaries	100%
Lenient	5 match errors	11 words	45%
Strict	10 match errors	11 words	91%
DER	4 match errors	5 DAs	80%

**Fig. 2.** Comparison of metrics to measure joint performance of segmentation and classification of DAs.

right boundary does not correspond to the reference segmentation exactly. This implies that the DSER metric penalizes missed cases more than false alarm (FA) cases, compared to the NIST-SU metric. See Fig. 1 for an illustration.

For the assessment of the joint performance of the segmentation and classification of DAs, four different metrics are used in the experiments described in Sec. 3. These metrics are illustrated in Fig. 2. First, the NIST-SU error metric is adapted to also include substitutions, not only missed boundaries or false alarms. Substitutions occur when the system outputs a DA boundary at the correct position, but the reference and the system disagree on the DA type on the left side of the boundary. The word-based “lenient” and “strict” metrics have been introduced in [9]. The lenient metric does not take into account the segmentation boundaries and only compares the DA types assigned to corresponding words. For the strict metric, a word is considered to be correctly classified if and only if it has been assigned the correct DA type and it lies in exactly the same DA segment as the corresponding word of the reference.

As a metric for the joint segmentation and classification of DAs that is easy to interpret, we propose the DA error rate (DER). This metric is derived from the DSER and not only requires a DA to have exactly matching boundaries but also to be tagged with the correct DA type. The DER thus measures the percentage of the misrecognized DA and can be seen as a length-normalized version of the strict metric.

For completeness we also mention the recognition accuracy as described in [14], which corresponds to the classical word error rate. As in the case of the word error rate, the accuracy metric of [14] only relies on the sequence of symbols (DA types in our case) and does not consider the actual segmentation boundaries. Scoring is then based on the string edit distance. This metric is not used in the experiments below.

### 3 Experiments and Discussion

For all experiments reported here the experimental setup used is as described in [9]. Of the 75 available meetings in the ICSI MRDA corpus, two meetings of a different nature are excluded (Btr001 and Btr002). From the remaining meetings we use 51 for training, 11 for development, and 11 for evaluation. For the segmentation and classification of the DA types, the available speech is first sorted according to the speaker, and then by time. The available DA types are mapped to the following five distinct types: backchannels (B), disruptions (D), floor grabbers (F), questions (Q), and statements (S). Each system is then optimized and evaluated under both reference and STT conditions. Under the reference condition it is assumed that we have access to the true sequence of spoken words, while under the STT condition the recognizer’s top-choice sequence of words is provided. The sequential approach to segmentation and classification of DAs described in [9] differs in a number of aspects from the systems investigated in this paper. Major differences lie in its sequential nature and the usage of prosodic and word-based information for both segmentation and classification of DAs. Prosody has been shown to help both the segmentation [4] and the classification of DAs [7]. While this system has the potential drawback of working in a sequential fashion, it is taking advantage of prosody in the segmentation step and requires access to the complete DA segment for classification. The potential advantage of the systems described in this paper lies in their ability to produce segmentation boundaries that are based on the estimation of the previous DA type for the last  $n$  words. However, both the HE-LM and the HMM tagger approach decide to segment and classify DAs based on local information only. Since the classification of the DA is implicitly done by predicting a corresponding DA boundary, valuable information is lost when the beginning of the current DA has fallen out of the current  $n$ -gram context.

Segmentation performance results of the different systems are provided in Table 1. To better compare the integrated approaches with the previous results, we report the segmentation error rate for [9] using the HE-LM alone without taking into account the prosodic pause feature. Note that, due to a minor dif-

Condition	System	NIST-SU	DSER
Ref	[9]	34.5	40.8
	[9] np <sup>1</sup>	46.0	53.0
	HE-LM	46.3	55.3
	Tagger	51.1	61.7
STT	[9]	45.5	49.4
	[9] np <sup>1</sup>	59.5	62.0
	HE-LM	59.6	62.4
	Tagger	62.8	66.9

<sup>1</sup> reduced system, no prosody features

**Table 1.** Comparison of the segmentation error rates of the different systems under both reference and STT conditions.

Condition	System	NIST-SU	Lenient	Strict	DER
Ref	[9]	52.6	20.0	64.4	54.4
	[9] np <sup>1</sup>	62.3	21.0	72.4	64.1
	HE-LM	62.2	23.3	74.3	66.5
	Tagger	69.5	22.6	78.6	72.6
STT	[9]	68.3	25.1	75.4	64.3
	[9] np <sup>1</sup>	78.3	25.0	82.9	73.2
	HE-LM	78.0	26.2	83.8	73.9
	Tagger	81.3	22.4	85.4	77.3

<sup>1</sup> reduced system, no prosody features

**Table 2.** Comparison of the segmentation and classification performance of the different systems under both reference and STT conditions.

ference in the counting of errors under STT conditions, the error rates given in Table 1 are slightly lower than those previously reported in [9]. Comparing the HE-LM and the tagger approach of this paper, we notice that the HE-LM consistently outperforms the tagger on both segmentation metrics.

Performance results for the joint segmentation and classification of DAs are provided in Table 2 for the different systems. Again, performance results for the reduced version of [9] (not including prosody) is used for better comparison with the HE-LM and the tagger based methods. Compared with these results, the HE-LM approach shows a comparable performance, which is promising, given the simplicity of the approach. As we would expect, the system described in [9] in its original form outperforms the approaches investigated here. A notable result from these experiments is the observation that the tagger based approach shows the lowest lenient error rates and, at the same time, the highest error rates for the NIST-SU, the strict, and the DER metrics. This observation suggests that the lenient metric is most useful when used in combination with other metrics that take into account the quality of the segmentation as well.

## 4 Conclusion and Outlook

We have investigated two simple approaches based on word information for joint segmentation and classification of DAs in multiparty meetings. Furthermore, with the DSER and the DER we propose additional performance metrics for segmentation and joint segmentation and classification of DAs with a simple semantic interpretation. The DSER measures the percentage of the correctly segmented DAs, while the DER quantifies the percentage of correctly segmented and tagged DAs. Based on the experiments, we suggest that the lenient metric proposed in [9] should not be used alone but in combination with other metrics that also take into account the quality of the segmentation.

The results provided in this paper serve as a baseline against which we will measure the results of future work on joint segmentation and classification. As a next step we will investigate approaches that do not rely only on local evidence, but rather are able to take into account complete DA hypotheses along the lines of [14]. In such a framework it is also possible to integrate prosodic information and to consider word lattices.

## 5 Acknowledgments

We thank Barbara Peskin for her valuable comments. This work was supported by the EU Framework 6 project on Augmented Multiparty Interaction, DARPA Contract NBCHD030010, NSF Awards IIS-0121396 and IRI-9619921, and the Swiss National Science Foundation through the research network IM2.

## References

1. Armstrong, S., et al.: Natural language queries on natural language data. In: Proc. NLDB, Burg, Germany (2003) 14–27
2. Waibel, A., et al.: Advances in automatic meeting record creation and access. In: Proc. ICASSP. Volume 1., Rhodes, Greece (2001) 207–210
3. Stolcke, A., et al.: Automatic detection of sentence boundaries and disfluencies based on recognized words. In: Proc. ICSLP. Volume 5., Sydney, Australia (1998) 2247–2250
4. Shriberg, E., et al.: Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication* **32** (2000) 127–154
5. Ji, G., Bilmes, J.: Dialog act tagging using graphical models. In: Proc. ICASSP. Volume 1., Philadelphia, USA (2005) 33–36
6. Ries, K.: HMM and neural network based speech act detection. In: Proc. ICASSP. Volume 1., Rhodes, Greece (2001) 207–210
7. Stolcke, A., et al.: Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* **26** (2000) 339–371
8. Webb, N., Hepple, M., Wilks, Y.: Dialog act classification based on intra-utterance features. Cs-05-01, Dept. of Comp. Science, University of Sheffield, UK (2005)
9. Ang, J., et al.: Automatic dialog act segmentation and classification in multiparty meetings. In: Proc. ICASSP. Volume 1., Philadelphia, USA (2005) 1061–1064

10. Liu, Y., et al: Structural metadata research in the EARS program. In: Proc. ICASSP. Volume 5., Philadelphia, USA (2005) 957–980
11. Shriberg, E., et al.: The ICSI meeting recorder dialog act (MRDA) corpus. In: Proc. SIGDIAL, Cambridge, USA (2004) 97–100
12. Stolcke, A., Shriberg, E.: Automatic linguistic segmentation of conversational speech. In: Proc. ICSLP. Volume 2., Philadelphia, USA (1996) 1005–1008
13. NIST website: Rt-03 fall rich transcription.  
<http://www.nist.gov/speech/tests/rt/rt2003/fall/> (2003)
14. Warnke, V., et al.: Integrated dialog act segmentation and classification using prosodic features and language models. In: Proc. 5th Europ. Conf. on Speech, Communication, and Technology. Volume 1., Rhodes, Greece (1997) 207–210