JOINT SEGMENTATION AND CLASSIFICATION OF DIALOG ACTS IN MULTIPARTY MEETINGS

Matthias Zimmermann[1], *Andreas Stolcke*[1,2], *Elizabeth Shriberg*[1,2]

[1]International Computer Science Institute, [2]SRI International, USA
{zimmerma,stolcke,ees}@icsi.berkeley.edu

ABSTRACT

This paper investigates a scheme for joint segmentation and classification of dialog acts (DAs) of the ICSI Meeting Corpus based on hidden-event language models and a maximum entropy classifier for the modeling of word boundary types. Specifically, the modeling of the boundary types takes into account dependencies between the duration of a pause and its surrounding words. Results for the proposed method compare favorably with our previous work on the same task.

1. INTRODUCTION

To support higher-level tasks such as information retrieval and summarization [1, 2], an input speech signal must be segmented into meaningful units, such as dialog acts (DAs). The DA types considered here are statements, questions, backchannels, floorgrabbers, and disruptions. The task we investigate is how to split a stream of words into nonoverlapping segments of text and assign mutually exclusive DA types 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 segmentation of speech [3, 4] or classification of already segmented text into various sets of DA types [5, 6, 7]. 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 is to be expected. Of course, for fully automatic processing of the speech stream both tasks need to be addressed. On the ICSI (MRDA) Corpus [8] a sequential approach based on both word and prosody features is described in [9]. A simple extension performing joint segmentation and classification of DAs was considered in [10], but did not include prosody.

In this paper prosodic features (pause durations) are used to extend [10]. Specifically, we use maximum entropy to model word boundary types based on both words and pause durations for joint segmentation and classification. A similar concept was used in [11] to annotate punctuation from speech. The performance of the proposed method is then evaluated and compared to the results reported in [9].

2. METHODOLOGY

2.1. Word based Segmentation and Classification

For joint segmentation and classification of DAs we are using a technique based on a Hidden-Event Language Model (HE-LM) [12] presented in [10]. After each word, the HE-LM predicts either a non-boundary event or the boundary event corresponding to any of the five DA types under consideration (i.e. backchannel, disruption, floorgrabber, question, or statement) resulting in a six-way classification of word boundaries. To produce the final output, the probability for the non-boundary event is compared to a fixed threshold θ at each word boundary. If its value is lower than the threshold we insert the boundary event associated with the highest a posteriori probability, otherwise the non-boundary event is chosen. The threshold θ therefor controls the insertion rate for DA boundaries.

2.2. Pause Duration and Word Context

The use of the duration of pauses between consecutive words of a given speaker have been found to substantially improve performance of systems segmenting speech into utterances [4] or dialog acts [9]. This finding is supported by Fig. 1 that shows how pause lengths are related to word boundaries or DA boundaries.
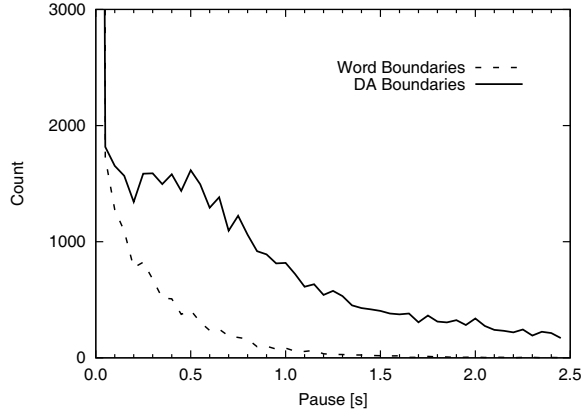
Fig. 1. Distribution of pause duration for pauses at dialog acts (DA) boundaries and words boundaries within a DA.

Word Context	Pause [s]	Count
you know	0.006	2876
the the	0.176	988
um um	0.935	55
because um	0.399	34
um yeah	0.746	30

Fig. 2. Average pause durations for some sample word pairs. The two words correspond to the word before and after the pause. The examples are sorted according to their observed frequency (Count).

However, Fig. 1 also indicates that any classification relying on the duration of the pause alone is prone to high error rates for short pauses. Inspection of such pauses that are not related to DA boundaries shows that such pauses are not distributed randomly but are often caused by hesitations. Very frequently, longer pauses occur in the neighborhood of filled pauses, such as "*uh*" and "*um*" or between word repetitions, as in "*the the*". See Fig. 2 for some examples. In contrast, the most frequently observed word pair "*you know*" does not have a pause between the word "*you*" and the word "*know*" in most cases, which further indicates that pause duration is closely related to the word context.

To model the boundary type between two words based on pause duration and its word context we use a maximum entropy based classifier. As the maximum entropy framework typically uses binary functions that mark the presence or absence of a particular feature, a given pause duration is represented by the corresponding bin of the histogram. Durations from zero up to three seconds were partitioned into ten bins in such a way that each bin received the same amount of training samples. For pause durations longer than three seconds a special bin was used. In addition, up to four surrounding words

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	Rate
NIST-SU	3 FA, 1 miss	5 boundaries	80%
DSER	3 match errors	5 DAs	60%

Fig. 3. The NIST-SU, and the DSER metrics for the assessment of segmentation error rates. Both the reference and the system line represent a sequence of words tagged with corresponding DA types, with statement (S), question (Q), backchannel (B), and disruption (D).

are used as textual features. For the baseline case word based features are omitted completely. Then, word w_i right before the pause and the joint feature (w_i, p_i) is included where p_i identifies the bin associated with the pause. For larger contexts the word w_{i+1} right after the pause is first added before words w_{i-1} and w_{i-2} are taken into account. Depending on the word context size, additional joint features (w_{i-1}, w_i), (w_i, w_{i+1}), and (p_i, w_{i+1}) are added. To limit the amount of features, we only consider features that have been observed at least five times in the training data. As in the case of the HE-LM the task of the maximum entropy classifier is to determine the DA type corresponding to the boundary between two words. We therefor end up with probabilities for the same six event types. These probabilities can then be integrated in the HE-LM conveniently, producing the most likely event for each word boundary according to [4]. These probabilities are then weighted against the probabilities produced by the HE-LM with a log likelihood weight α. For $\alpha = 0$ the probabilities from the maximum entropy classifier are not taken into account. Increasing values of α lead to a final result that is more and more influenced by the maximum entropy classifier (e.g. $\alpha = 1$ assigns the same weight to the n-gram HE-LM probabilities as to the probabilities from the maximum entropy classifier).

2.3. Performance Metrics

To assess the performance of segmentation or classification of DAs, a number of metrics have been proposed. For the case of joint segmentation and classification most available metrics do not directly fit. For instance, metrics evaluating segmentation performance do not consider the correctness of the classification task while metrics for the classification of DAs assume perfect segmentation. Since tuning of system parameters is inherent to most systems, it is important to tune to metrics that are appropriate to the task at hand.

We first describe two metrics for the measurement of the segmentation performance and then define metrics for the joint

```
Reference   S|Q.Q.Q.Q|S.S.S|B|S.S|
System      S|Q|S|Q.Q|D.D.D|S.S|S|
Strict      C E E E E E E E E E E
DER         C|   E   |   E  |E| E  |
```

Metric	Errors	Reference	Rate
Strict	10 match errors	11 words	91%
DER	4 match errors	5 DAs	80%

Fig. 4. Comparison of the Strict and the DER metrics to measure joint performance of segmentation and classification of DAs.

Words	α	θ	NIST-SU	DSER	Strict	DER
0	0.6	0.90	39.4	41.6	68.2	58.4
1	0.8	0.90	41.0	40.2	68.1	55.3
2	1.6	0.95	38.6	38.3	65.9	54.0
3	1.6	0.95	38.3	38.0	65.6	53.3
4	1.6	0.95	38.3	38.2	65.5	53.3

Table 1. The effect of increasing word context (Words) to performance metrics. These measurements were made on the validation set under reference conditions. Parameters α and θ were tuned to optimize the Strict and the DER metric jointly.

segmentation and classification of DAs. The NIST-SU metric [13] was used to report the segmentation performance in previous work [9]. To provide a more intuitive metric that is directly related to DAs, we introduced the DA Segmentation Error Rate (DSER) in [10]. The DSER measures the percentage of wrongly segmented reference DA segments, where a DA is considered to be mis-segmented if and only if its left or right boundary (or both) does not exactly correspond to the reference segmentation. This implies that for the DSER metric missed cases are penalized more than false alarms (FA) compared to the NIST-SU metric. Also, for the DSER metric the maximum error rate is 100% (e.g. not putting boundaries anywhere) while for the NIST-SU metric the error rate can easily exceed 100% (e.g. 500% when we assume that we put a DA boundary between all words and a DA contains 6 words on average). See Fig. 3 for an illustration.

For the assessment of the joint performance of the segmentation and classification of DAs, a word-based and a DA-based metric are used in the experiments described below [1]. The word-based Strict metric has been introduced in [9] while the DA-based DER metric was proposed in [10] as an analog to the DSER segmentation metric. 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. The DA Error Rate (DER) not only requires a DA candidate to have exactly matching boundaries but also to be tagged with the correct DA type. The DER thus measures the percentage of the misrecognized reference DAs and can be seen as a length-normalized version of the Strict metric. See Fig. 4 for an illustration.

3. EXPERIMENTS AND DISCUSSION

For all experiments reported here, the experimental setup as described in [9] is used. Of the 75 available meetings of the ICSI MRDA corpus, two meetings of a different nature are

---

[1] Two additional metrics found in the literature, the "recognition accuracy" as defined in [14], and the "lenient" metric [9] are not considered here, since they do not take into account segmentation errors.

excluded (Btr001, and Btr002). From the remaining meetings, we use 51 for training, 11 for validation, and 11 for evaluation. For the segmentation and classification of the DA types, the available speech is first sorted according to speakers, and then by time. The available DA types are mapped to the following five distinct types: backchannels (B), disruptions (D), floorgrabbers (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 the 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 two major aspects from the system investigated in this paper. While the previous system has the potential drawback of working in a sequential fashion, it can take advantage of the fact that the full sequence of words of a DA is available for classification. This is contrast to the method presented here that has the potential weakness of making decisions relying on local evidence only.

For the approach presented in this paper, grid search was applied to find optimal values for parameters α and θ on the validation data for context sizes from zero up to four words. Optimality was defined by the average of the resulting Strict and DER error rates (see Table 1). Based on this optimization step, a context size of four words was chosen for both reference and STT condition. Under STT condition training of the maximum entropy classifier on STT data lead to better results than training on reference data. Looking at the optimal choice for parameters α and θ for different amounts of word context (as shown in Table 1) we can make two interesting observations. First, the average of the Strict and the DER error rates monotonically falls for increasing amounts of word context. Second, the log likelihood weight α for the integration of the pause duration also increases with the number of context words for the pause modeling indicating a growing reliability of the maximum entropy based classifier.

The test set results provided in Table 2 confirm the expected benefit of the use of word context for modeling of the pause duration. A substantial improvement over the experiments that did not include word context is achieved. Depend-

Cond.	System	NIST-SU	DSER	Strict	DER
	[9]	34.5	40.8	64.4	54.4
Ref	No context	35.5	39.7	65.0	55.9
	With context	34.8	36.8	62.8	51.0
	[9]	45.5	49.4	75.4	64.3
STT	No context	49.7	48.0	76.2	65.1
	With context	44.6	47.4	73.6	62.6

Table 2. Test set results for the NIST-SU and the DSER segmentation error rates, and the Strict and DER joint segmentation and classification error rates under both reference and STT conditions.

ing on the experimental conditions and the joint error metric absolute error rate reductions of 2.2% up to 4.9% were found. Furthermore, the proposed integrated segmentation and classification scheme also outperforms our previous results based on the sequential approach investigated in [9] by 1.6% to 3.4% absolute.

When the maximum entropy classifier is used alone for joint segmentation and classification its performance is slightly worse than [9] under reference condition but comparable under STT condition (using a context size of 4 words) indicating a robust performance under noisy conditions.

4. CONCLUSION AND OUTLOOK

We investigated the task of joint segmentation and classification of of DAs by extending a scheme presented in [10] to also make use of prosodic features, namely word pause duration. For this, we propose a modeling of word boundary types based on both the duration of a pause and its surrounding words. This is in contrast to previous work [4, 9] that treated pause durations independently from surrounding words. Resulting error rate reductions of 2.2% up to 4.9% (absolute) over word context independent pause modeling confirm the validity of the approach presented in this paper.

In future work we will integrate the proposed method into the A* based approach presented in [15] that does not rely on local evidence only but is able to take into account complete DA hypotheses.

5. REFERENCES

[1] S. Armstrong and et al., "Natural language queries on natural language data," in *Proc. NLDB*, Burg, Germany, 2003, pp. 14–27.

[2] A. Waibel and et al., "Advances in automatic meeting record creation and access," in *Proc. ICASSP*, Rhodes, Greece, 2001, vol. 1, pp. 207–210.

[3] A. Stolcke and et al., "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *Proc. ICSLP*, Sydney, Australia, 1998, vol. 5, pp. 2247–2250.

[4] E. Shriberg and et al., "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[5] G. Ji and J. Bilmes, "Dialog act tagging using graphical models," in *Proc. ICASSP*, Philadelphia, USA, 2005, vol. 1, pp. 33–36.

[6] K. Ries, "HMM and neural network based speech act detection," in *Proc. ICASSP*, Rhodes, Greece, 2001, vol. 1, pp. 207–210.

[7] A. Stolcke and et al., "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–371, 2000.

[8] E. Shriberg and et al., "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SIGDIAL*, Cambridge, USA, 2004, pp. 97–100.

[9] J. Ang and et al., "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. ICASSP*, Philadelphia, USA, 2005, vol. 1, pp. 1061–1064.

[10] M. Zimmermann and et al., "Toward joint segmentation and classification of dialog acts in multi-party meetings," in *Proc. 2nd MLMI*, Edinburgh, UK, 2005.

[11] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. ICSLP*, Denver, USA, 2002, pp. 917–920.

[12] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Proc. ICSLP*, Philadelphia, USA, 1996, vol. 2, pp. 1005–1008.

[13] NIST website, "Rt-03 fall rich transcription," http://www.nist.gov/speech/tests/rt/rt2003/fall/, 2003.

[14] V. Warnke and et al., "Integrated dialog act segmentation and classification using prosodic features and language models," in *Proc. 5th Europ. Conf. on Speech, Communication, and Technology*, Rhodes, Greece, 1997, vol. 1, pp. 207–210.

[15] M. Zimmermann and et al., "A* based joint segmentation and classification of dialog acts in multi-party meetings," in *Proc. 9th ASRU*, Cancun, Mexico, 2005.