A* BASED JOINT SEGMENTATION AND CLASSIFICATION OF DIALOG ACTS IN MULTIPARTY MEETINGS

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

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

ABSTRACT

We investigate the use of the A* algorithm for joint segmentation and classification of dialog acts (DAs) of the ICSI Meeting Corpus. For the heuristic search a probabilistic framework is used that is based on DA-specific N-gram language models. Furthermore, two new metrics for performance evaluation are motivated and described and the influence of different metrics for performance evaluation is demonstrated. The proposed method is evaluated on both traditional and new metrics, and compared 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). Typical DA types are statements, questions, and backchannels. The task we investigate in this paper 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. 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. This is in contrast to sequential approaches that do not take advantage of information produced by the classification of DAs for the segmentation step.

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]. For automatic segmentation of speech, it remains unclear how well a subsequent component handles 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. An integrated approach to segmentation and the classification of DAs based on the A* algorithm was used in the context of the Verbmobil project [8]. On the ICSI (MRDA) Corpus [9] a sequential approach is described in [10], while a simple extension of the segmentation scheme performing joint segmentation and classification of DAs is considered in [11]. In the text below we investigate joint segmentation and classification of DAs following the lines of [8]. The performance of this approach is then evaluated and compared to the results reported in [10, 11].

2. METHODOLOGY

For the sequence of n input words $W = (w_1, \ldots w_n)$, we try to find a segmentation $S = (s_1, \ldots s_m)$ with corresponding DA labels $D = (d_1, \ldots d_m)$. The variables s_i define the number of consecutive words in the ith DA $W_i = (w_k, \ldots w_l)$ where $k = 1 + \sum_{j=1}^{i-1} s_j$ and $l = k + s_i - 1$.

Joint segmentation and classification of DAs can now be formulated as $(\hat{S}, \hat{D}) = arg\ max_{S,D}\, P(S, D|W)$. Instead of maximizing $P(S, D|W)$, we invoke Bayes' Rule to maximize $P(W|S, D)\ P(S|D)P(D)$, which can be decomposed into the product given below, assuming independence of the W_i, given both the s_i, and the d_i, as well as independence of the s_i given the d_i. Finally, we assume that $P(D)$ can be decomposed into a product of bigram probabilities $P(d_i|d_{i-1})$.

$$\prod_{i=1}^{m} P(W_i|s_i, d_i)P(s_i|d_i)P(d_i|d_{i-1}) \qquad (1)$$

In our case, the probabilities $P(W_i|s_i, d_i)$ are estimated by DA type-specific trigram language models (LMs). The $P(s_i|d_i)$

We thank Barbara Peskin for her valuable comments. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication), by DARPA Contract NBCHD030010 through the SRI CALO project (approved for public release, distribution unlimited), NSF Awards IIS-0121396 and IRI-9619921, and the Swiss National Science Foundation through the research network IM2.

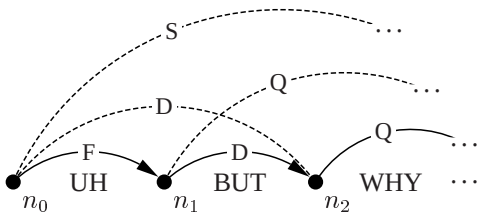

Fig. 1. A* search graph for joint segmentation and classification of DAs. The true segmentation and classification is indicated by the solid edges. Dashed edges correspond to alternative segmentation paths.

are computed using the observed DA type-specific length distributions, and the $P(d_i|d_{i-1})$ are provided by a bigram DA grammar. Expression (1) can then be integrated into the A* search algorithm in a straightforward way.

2.1. A* Search

During A* search, an optimal path through the input word sequence W is found. This is achieved by defining the nodes of the search graph as the positions between the words. The start node n_0 corresponds to the position before the first word w_1 and the final node comes directly after the last word w_n. Edges of the search graph carry a label indicating the DA type and span one or more consecutive words. Each edge therefore hypothesizes a potential DA within the input sequence of words. See Fig. 1 for an illustration. The costs $\mathcal{C}_i$ of DA candidates are directly derived from Expression (1) as shown below.

$$\begin{aligned} \mathcal{C}_i \quad = \quad & -\lambda_1 \log P(W_i|s_i, d_i) - \lambda_2 \log P(s_i|d_i) \\ & -\lambda_3 \log P(d_i|d_{i-1}) \end{aligned} \quad (2)$$

By taking the negative logarithm of the probabilities related to the DA candidates we can replace the product of Expression (1) by the sum over the cost of subsequent edges in the search graph. Parameters λ_1, λ_2, and λ_3 (with $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$) are introduced to reduce negative side effects of the imperfect modeling of the different probabilities and will be optimized experimentally on heldout data.

In complex search graphs, a problem-specific heuristic function helps the A* algorithm to find the optimal solution efficiently. To achieve this, the heuristic function must provide a lower bound of the costs to reach the final node from the current node. Since in our case the search graph just consists in a linear sequence of nodes (i.e the size of the search graph only grows linearly with the number of words), we can use a trivial heuristic function that always returns zero. If the input would consist of a word lattice instead of the single best output of a speech-to-text (STT, i.e. automatic speech recognition) system, a more sophisticated approach might be needed to keep search times within reasonable bounds.

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. 2. The NIST-SU, and the new 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 S=Statement, Q=Question, B=Backchannel, and D=Disruption.

2.2. 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 segmentation and classification of DAs. The NIST-SU metric was used to report the segmentation performance in previous work [10] and has been provided by NIST in the EARS MDE evaluations [12]. 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 that is directly related to DAs, we introduced the DA Segmentation Error Rate (DSER) in [11]. The DSER measures the percentage of wrongly segmented 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

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. 3. Comparison of the Strict and the new DER metric to assess joint performance of segmentation and classification of DAs.

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. 2 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 in Sec. 3 [1]. The word-based strict metric has been introduced in [10] while the DA-based DER metric was proposed in [11] 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 DAs and can be seen as a length-normalized version of the strict metric. See Fig. 3 for an illustration.

3. EXPERIMENTS AND DISCUSSION

For all experiments reported here, the experimental setup used is as described in [10]. Of the 75 available meetings of 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 the spoken words, while under the STT condition the recognizer's top-choice sequence of words is provided.

[1]Two additional metrics found in the literature, the "recognition accuracy" as defined in [8], and the "lenient" metric [10] are not considered here, since they do not take into account segmentation errors.

Cond.	System	NIST-SU	DSER	Strict	DER
Ref	[10]	34.5	40.8	64.4	54.4
	[10] np[1]	46.0	53.0	72.4	64.1
	[11]	46.3	55.3	74.3	66.5
	A*	51.0	48.9	73.1	62.3
STT	[10]	45.5	49.4	75.4	64.3
	[10] np[1]	59.5	62.0	82.9	73.2
	[11]	59.6	62.4	83.8	73.9
	A*	71.1	55.8	83.9	71.4

[1]: reduced system, no prosody

Table 1. Comparison of the NIST-SU and the DSER segmentation error rates, and the Strict and DER joint segmentation and classification error rates for both reference and STT conditions.

The sequential approach to segmentation and classification of DAs described in [10] differs in a number of aspects from the systems investigated in this paper. While this system has the potential drawback of working in a sequential fashion, it is taking advantage of prosody in the segmentation step. To better compare the performance of the proposed approaches, a reduced version of [10] that does not make any use of prosody is included in the experiments. The reduced system uses a hidden-event LM (HE-LM) for segmentation, and classification of DAs is based on the maximum entropy framework. See [10] for details. Our first attempts at joint segmentation and classification of DAs included an extended version of a HE-LM which not only predicted the presence of a DA boundary (as in [10]) but the type of the DA boundary at the same time, as described in [11].

For the A* based approach presented in this paper, a grid search was applied to find optimal values for parameters λ_1, λ_2, and λ_3 on the development data. Parameter values that minimize the DER metric under reference conditions were found at $\lambda_1 = 0.7$, $\lambda_2 = 0$, and $\lambda_3 = 0.3$. Under STT conditions optimal values were $\lambda_1 = 0.8$, $\lambda_2 = 0$, and $\lambda_3 = 0.2$. The setting of $\lambda_2 = 0$ indicates that the use of the DA-specific length distributions $P(s_i|d_i)$ does not help to improve the joint segmentation and classification performance as measured by the DER (and the strict) metric. A possible reason for this is an implicit modeling of the length by the DA-specific N-gram LMs. It is interesting to note that to optimize the segmentation, very different settings would be selected, e.g., $\lambda_1 = 1$ for the minimization of the DSER metric. To optimize the A* system for the NIST-SU metric, parameters $\lambda_1 = 0.5$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.3$

DA	Count	[10]	[10] np[1]	A*
B	1946	25.2	29.2	15.7
D	2220	72.9	84.2	80.9
F	1918	54.6	68.4	57.3
Q	1159	75.0	80.9	69.0
S	8889	53.4	63.6	68.1

[1]: reduced system, no prosody

Table 2. DA-specific error rates using the DER metric for the different systems under reference conditions. Column "Count" contains the number of corresponding DAs in the test set.

worked best under reference conditions. As a consequence, the test set NIST-SU error rate was reduced from 51.0% (as reported in Table 1) to 49.1%. At the same time, the DER was increased from 62.3% to 68%.

Test set results are provided in Table 1. It can generally be observed that the A* approaches outperform the sequential approach of [10] on the DSER and the DER metrics when prosody is excluded. When compared to our previous results for joint segmentation and classification based on the HE-LM [11], we find a substantial improvement of the A* approach for all conditions and metrics except for NIST-SU, and under STT conditions the strict metric. The low performance achieved using the NIST-SU metric is mainly a result of the tendency of the DA-specific LMs to oversegment the input text[2]. By cutting longer DAs into a sequence of several shorter ones, many false alarms are generated which harms the NIST-SU performance significantly. On the other hand, this leads to fewer errors for single-word DA candidates like "YEAH" or "RIGHT" which boosts the performance for the DSER and the DER metrics. The presented results also indicate that different systems perform better for certain error metrics than for others e.g. the system described in [10] performs better for the NIST-SU and the strict error metrics while the evaluation of the A* approach described in this paper works best for the DSER and the DER metrics.

A DA-specific error analysis for the different systems under reference conditions is provided in Table 2. The A* approach outperforms the sequential approach of [10] significantly for both backchannels and questions. A manual inspection of the differences of the results produced by the different systems suggests that the DA-specific LMs recognize questions and backchannels more often than the maximum entropy based classifier used in [10], which frequently tags questions and backchannels as statements.

[2]The use of a word insertion penalty helped to improve the performance for the NIST-SU metric. As in the case of the DA-specific length modeling, the improvements came at the cost of higher error rates for the strict and the DER metrics.

4. CONCLUSION AND OUTLOOK

We investigated the use of the A* graph search algorithm for joint segmentation and classification of DAs in multiparty meetings. For this, the use of word-based DA-specific LMs is motivated in the context of a probabilistic framework, and experimental results confirm the validity of the chosen approach. Furthermore, two new performance metrics, the DSER for segmentation (measuring the percentage of the correctly segmented DAs), and the DER for joint segmentation and classification of DAs (quantifying the percentage of correctly segmented and classified DAs), are described, and the influence of different metrics for performance evaluation is demonstrated.

Results based on the A* approach outperform our previous work[11] for joint segmentation and classification under all conditions and for all metrics but the NIST-SU metric and, under STT conditions, the strict metric. The original system [10] is outperformed for the proposed DER and DSER metrics when prosodic features are excluded. When compared to the original system [10] including prosody, the proposed approach still does better on questions and backchannels.

Next steps will include the integration of prosody and the processing of word lattices.

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] 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.

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

[10] 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.

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

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