Multistream Recognition of Dialogue Acts in Meetings

Multistream Recognition of Dialogue Acts in Meetings

Alfred Dielmann and Steve Renals

Centre for Speech Technology Research
University of Edinburgh,
Edinburgh EH8 9LW, UK
Email:{a.dielmann,s.renals}@ed.ac.uk

Abstract. We propose a joint segmentation and classification approach for the dialogue act recognition task on natural multi-party meetings (ICSI Meeting Corpus). Five broad DA categories are automatically recognised using a generative Dynamic Bayesian Network based infrastructure. Prosodic features and a switching graphical model are used to estimate DA boundaries, in conjunction with a factored language model which is used to relate words and DA categories. This easily generalizable and extensible system promotes a rational approach to the joint DA segmentation and recognition task, and is capable of good recognition performance.

1 Introduction

This paper is concerned with automatically extracting some facets of the discourse structure of multiparty meetings. In particular we are concerned with the automatic recognition of *dialogue acts* (DAs). Each utterance in a transcription of a meeting can be associated to a dialogue act (or several dialogue acts) describing the function that the utterance serves in the conversation. This generic definition leaves space for several different DA coding schemes, that may be targeted on different aspects of the conversational process or simply characterised by a different number of sub-categories.

In this work we are interested in a DA dictionary composed of a few generic DA categories [1]. Classes of dialogue act in this scheme, which was obtained from the richer Meeting Recorder Dialogue Act (MRDA) annotation scheme [2], consisted of *statements*, *questions*, *fillers*, *back-channel* and *disruptions*. Those broad DA categories can be seen as the basic building blocks of a conversation, and thus they may be employed in modelling more complex meeting behaviours, such as meeting phases, or to enhance processes such as language modelling for automatic speech recognition or topic detection.

The DA recognition process is composed of two main steps: segmentation and tagging. The first step consists of subdividing the sequence of transcribed words in terms of DA segments. The goal is to segment the text into utterances that have approximately similar temporal boundaries to the annotated DA units. The second step of DA tagging takes DA segments as input and classifies them as one of the five DA classes listed above. These two steps may be performed either sequentially (segmentation followed by classification) or jointly (both tasks carried out simultaneously by an integrated system). In this paper we focus on the joint segmentation and classification approach, using

trainable statistical models: dynamic Bayesian networks (DBNs). We note that the full DA recogniser can be forced to operate on pre-segmented data, hence acting as a simpler DA tagger. Alternatively, by discarding the DA tags the system may be employed for the segmentation task alone.

The paper is organised as follows. The next section reviews some DA recognition works carried out on natural multi-party meetings, with a particular focus on the ICSI meeting corpus, described in section 3. Section 4 outlines our DA recognition framework and its components: the feature extraction process (section 5), the DA factored language model (section 6), and the generative DBN-based infrastructure (section 7). Experiments using this framework and five different setups are reported in section 8. Finally, section 9 proposes a brief summary and concludes with some final notes.

2 Related Work

Stolcke et al. [3] provide a good introduction to dialogue act modelling in conversational telephone speech, a domain with some similarities to multiparty meetings. Dialogue acts may be modelled using a generative hidden Markov model [4], in which observable feature streams are generated by hidden state DA sequences. Most DA recognizers are based on statistical language models evaluated from transcribed words, or on prosodic features extracted directly from audio recordings. Various language models have been tried, including factored language models [5], although any kind of trainable language model can be adopted. Prosodic features provide a large range of opportunities, with entities such as duration, pitch, energy, rate of speech and pauses being measured using different approaches and techniques [6, 7]. Other features, such as speaker sex, have also been usefully integrated into the processing framework.

The most likely sequence of dialogue acts is inferred from the lexical and prosodic data, and from a discourse model. The discourse or dialogue act grammar could be estimated using a simple n-gram model based on DA labels or more exotic language models evaluated from the distribution of DA-tags. Note that precise utterance and dialogue act boundaries are often assumed to be known a priori as part of the DA annotation (tagging task). When this information is not available (recognition task), it is estimated by employing automatic segmentation algorithms.

Ang et al. [1] addressed the automatic dialog act recognition problem using a sequential approach, in which DA segmentation was followed by classification of the candidate segments. Promising results were achieved by integrating a boundary detector based on *vocal pauses* with a hidden-event language model HE-LM (a language model including dialogue act boundaries as pseudo-words). The dialogue act classification task was carried out using a maximum entropy classifier, together with a relevant set of textual and prosodic features. This system segmented and and tagged DAs in the ICSI Meeting Corpus, with relatively good levels of accuracy. However results comparing manual with automatic ASR transcriptions indicated that the ASR error rate resulted in a substantial reduction in accuracy.

Using the same experimental setup, Zimmermann et al. [8] proposed an integrated framework to perform joint DA segmentation and classification. Two lexical based approaches were investigated, based on an extended HE-LM (able to predict not only the

DA boundaries but also the DA type), and a HMM part of speech inspired approach. Both these approaches provided slightly lower accuracy when compared with the two-step framework [1], but this may be accounted by the lack of prosodic features.

Ji et al. [9] propose a switching-DBN based implementation of the HMM approach outlined above, which they applied to dialogue act tagging on ICSI meeting data. They also investigated a conditional model, in which the words of the current sentence generate the current dialog act (instead of having dialogue acts which generate sequence of words). Since this work used only lexical features, and a large number of DA categories (62), a direct comparison with the results provided by [1] is not possible.

Venkataraman et al. in [10] proposed an approach to bootstrap a HMM-based dialogue act tagger from a small amount of labeled data followed by an iterative retraining on unlabeled data. This procedure enables a tagger to be trained on an annotated corpus, then adapted using similar, but unlabeled, data. The proposed tagger makes use of the standard HMM framework, together with dialogue act specific language models (3-grams) and a decision tree based prosodic model. The authors also advance the idea of a completely unsupervised DA tagger in which DA classes are directly inferred from data.

3 Annotated Data

The experiments reported in this paper use the ICSI Meeting Corpus [11]. This corpus consists of 75 multiparty meetings recorded with multiple microphones: one head-mounted microphone per participant and four tabletop microphones. Each meeting lasts about one hour and involves an average of six participants, resulting in about 72 hours of multichannel audio data. The corpus contains human-to-human interactions recorded from naturally occurring meetings. Moreover, having different meeting topics and meeting types, the data set is heterogeneous both in terms of content and structure.

Orthographic transcriptions are available for the entire corpus, and each meeting has been manually segmented and annotated in terms of Dialogue Acts, using the ICSI MRDA scheme [2]. The MRDA scheme is based on a hierarchy of DA types and subtypes (11 generic tags and 39 specific sub-tags), and allows multiple sub-categorizations for a single DA unit. This extremely rich annotation scheme results in more than a thousand unique DAs, although many are observed infrequently. To reduce the number of sparsely observed categories, we have adopted a reduced set of five broad DA categories [1, 8]. Unique DAs were manually grouped into five generic categories: statements, questions, backchannels, fillers and disruptions. The distribution of these categories across the corpus is shown in table 1. Note that statements are the most frequently occurring unit, and also the longest, having an average length of 2.3 seconds (9 words). All the other categories (except backchannels which usually last only a tenth of a second) share an average length of 1.6 seconds (6 words). An average meeting contains about 1500 DA units.

The corpus has been subdivided into three data sets: training set (51 meetings), development set (11 meetings) and test set (11 meetings). All our experiments were conducted on the same dataset subdivision proposed by Ang et al.[1] in order to have directly comparable results.

Category	% of total DA units	% of corpus length
Statement	58.2	74.5
Disruption	12.9	10.1
Backchannel	12.3	0.9
Filler	10.3	8.7
Question	6.2	5.8

Table 1. Distribution of DA categories by % of the total number of DA units and by % of corpus length.

4 Methodology

Our framework for the integrated DA recogniser uses a generative approach composed of four main blocks: a Factored Language Model (FLM, section 6), a feature extraction component (section 5), a trigram discourse model, and a Dynamic Bayesian Network (section 7). The FLM is used to map sequences of words into DA units, and is the main component of the tagger. The discourse model consists of a standard trigram language model over DA label sequences[1]. Note that our DA tagger uses only lexical information and a discourse model. Experiments using both the reference orthographic transcription and the output of automatic speech recognition (ASR) have been carried out. The automatic transcription was provided by the AMIASR team and generated through an ASR system similar to the one outlined in [12] (word error rate of about 29%). A set of six continuous features are used for DA segmentation purposes, together with part of a DBN model. This graphical model also plays a crucial role in the tagging process and acts as the master control unit for the entire recognition process.

5 Features

A vector of six continuous word related features was extracted from audio recordings and orthographic transcriptions.

Mean and variance of F0 Fundamental frequency (F0) was estimated using the ESPS pitch tracking algorithm get_f0[2] and sampled every 10 msec. The word temporal boundaries provided by the transcription [3] were then used to estimate the mean and variance of F0 for each word. Mean F0 was subsequently normalised against the speaker average pitch in order to have a participant independent feature.

RMS energy Average root mean square energy was estimated for each word W_i and then normalised by both the average channel energy (in order to compensate for factors such as channel gain and microphone position) and the mean energy for all tokens of word W_i.

[1] Estimated using the SRILM toolkit, available from http://www.speech.sri.com/projects/srilm/

[2] Available from http://www.speech.kth.se/snack/

[3] Note that word boundaries are estimated automatically through forced alignment between acoustic models and orthographic transcriptions, thus are characterised by a relevant amount of uncertainty.

Word length This is the word duration normalised by the mean duration for that word computed on the entire dataset. Therefore the resulting entity is inversely proportional to the rate of speech, neglecting estimation errors.

Word relevance The word relevance was computed to be the ratio between local term frequency within the current document and absolute term frequency across the whole meetings collection. Terms which are more relevant for the current meeting will assume scores well above the unity.

Pause duration Interword pauses were estimated using word boundary times obtained from aligning the transcription with the acoustic signal, and re-scaled in order to have a unitary range. Note that long pauses between words may highlight sentence boundaries and thus be a strong cue to DA segmentation. In fact pause related features have already been successfully employed in several DA segmentation frameworks (section 2).

6 Factored Language Models

Factored Language Models (FLMs) [13] are a generalisation of class-based language models in which words and word-related features are bundled together. The factors in an FLM may include word-related features such as part of speech, relative position in the sentence, stem, and morphological class. Indeed, there is no limit to the number of possible factors. In the FLM perspective even the words themselves, are usually considered one of the factors. Class based language models may be interpreted as a 2-factor FLM, in which words are bundled with classes.

Given a word f_t^0 and $k-1$ features $f_t^1, f_t^2, \ldots, f_t^{k-1}$, a sentence can be seen as sequence of these factor vectors $v_t \equiv \left\{ f_t^0, f_t^1, \ldots, f_t^k \right\}$. As for standard language models, the goal of FLMs is to factorise the joint distribution $p(v_1, v_2, \ldots, v_n)$ as a chain product of conditional probabilities in the form $p(v_t \mid v_{t-1}, ..., v_{t-n})$. Since words have been replaced by vectors of factors, each conditional probability is now a function of these factors: $p(f_t^0, f_t^1, \ldots, f_t^k \mid f_{t-1}^0, f_{t-1}^1, \ldots, f_{t-1}^k, f_{t-2}^0, \ldots, f_{t-2}^k, \ldots, f_{t-n}^{0:k})$.

In order to build a good FLM it is necessary to choose the optimal factorisation (analogous to the structure learning problem in graphical models) and a backoff strategy to cope with data sparsity. Note that backoff is usually operated by dropping one or more factors from a Conditional Probability Table (CPT) in favour of a simpler conditional distribution (and smaller CPT), reiterating this procedure several times. Often multiple backoff paths (strategies) are feasible and it is even possible to concurrently follow all of them by adopting a generalized parallel backoff [5].

In order to model the relationship between words and DAs we have adopted a FLM based on three factors: words, DAs and the position of each word in the DA unit. Each word w_t is part of a DA unit and is characterised by the DA label d_t. Moreover each DA segment has been subdivided in blocks of five words: if w_t is one of the first five words the position factor n_t will be equal to one, if w_t belongs to the second block $n_t = 2$, and so on. The adopted language model is defined by a product of conditional probabilities $p(w_t \mid w_{t-1}, n_t, d_t)$. Note that considering only the word factor w_t the proposed FLM could be compared to a bigram since only the relation between w_t and w_{t-1} is taken into account. When backoff is required the first term to be dropped is the previous word

w_{t-1}, leading to the backoff model $p(w_t \mid n_t, d_t)$. If a further backoff is required, the DA tag d_t will be dropped resulting in the simpler model: $p(w_t \mid n_t)$. We use Kneser-Ney discounting to smooth both the backoff steps.

In order to compare different FLM candidates, instead of comparing their perplexities, we have defined a simplified *DA tagging* task. We compare FLMs by measuring their ability to assign the correct DA label to unseen DA units. This preliminary evaluation was conduced by enhancing the FLM section of the SRILM toolkit [14] with a simple decoder, able to label each DA unit (sentence) with the most likely DA tag (factor label from a list of possible options).

The above described FLM, after training on the 51 meeting training set, was able to perform DA labeling on the 11 development set meetings with an accuracy of 69.7% using reference transcriptions and 63.4% using automatic transcriprions (70.9% and 63.6% on the 11 meetings from the test set). Replacing for example the word position factor n_t with part-of-speech tags p_t (automatically labeled by using a POS tagger trained on Broadcast News data) the accuracy on manual transcriptions fell to 61.7% (63.5% on the test set). Building the model $p(w_t \mid w_{t-1}, m_t, d_t)$, where m_t represents the information about the meeting type, the recognition rate rose to 68.2% (68.8% on the test set). A model including each of n_t, p_t and m_t with three backoff steps had slightly lower recognition rates of 67.7% on the development set and 68.2% on the test set.

7 Generative DBN model

Bayesian Networks (BNs) are examples of directed acyclic Graphical Models (GMs). GMs represent a unifying concept in which probability theory is encapsulated inside the formalism of graph theory. Random variables are associated to nodes, and statistical independence between two random variables is represented by the lack of a connecting arc between the corresponding nodes. To model time series or data sequences, the BN formalism has been generalised into the Dynamic Bayesian Network (DBN) concept. A DBN is a collection of BNs where a single BN, with private intra-frame relations among variables, is instantiated for each temporal frame, and a set of inter-frame arcs is defined. Those connections between nodes of adjacent BNs explicitly describe the flow of time and help highlighting the temporal structure of each time-series.

A DBN is a modular and intuitive representation which provides a common underlying formalism [15] for models including Kalman filters, Hidden Markov Models, coupled HMMs and hierarchical HMMs among the others. Note that since the DBN formalism is dual to a well defined mathematical theory, a unique set of tools and techniques can be developed to perform inference, model learning and decoding of any DBN model. The Graphical Model ToolKit (GMTK) [16], for example, provides a formal language to describe DBNs and a common set of tools to experiment with them. Thus this toolkit has been adopted as the main development package for all the DBN related experiments described in this work. As anticipated in section 4 the DA recognition process is coordinated by a generative DBN based model. The overall model is depicted in figure 1. The node Y_t represents the continuous observable feature vector outlined in section 5 (associated to the word W_t). E is a binary variable that switches from zero to one when a DA boundary is detected. Since the node W_t represents a word,

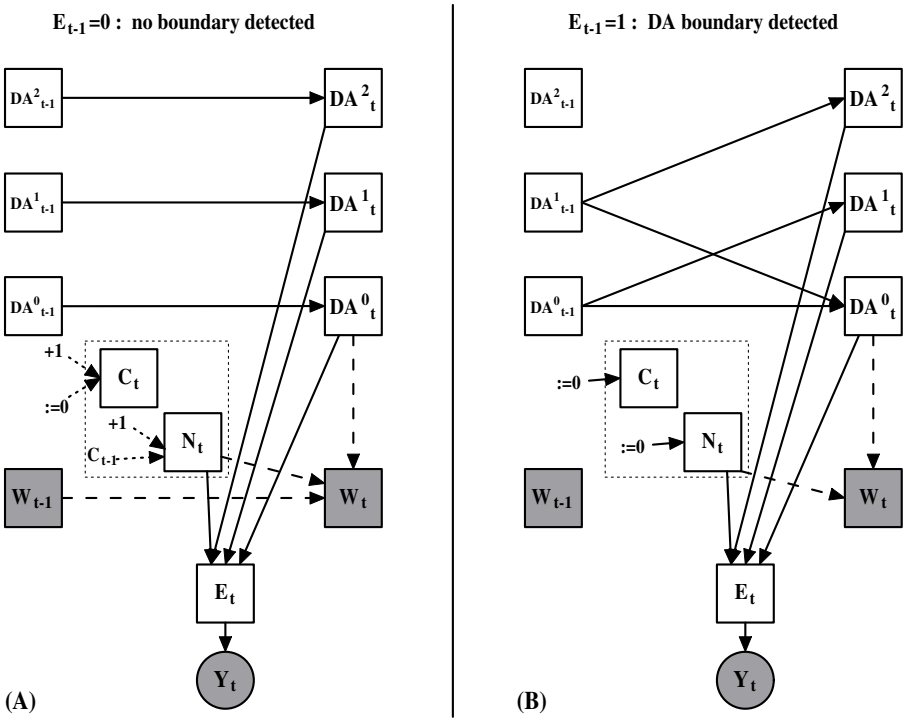

Fig. 1. Overview of the DBN model for the integrated Dialogue Act recogniser. The model's topology depends on the state of the boundary detector E_{t-1} during the previous frame: the model's graph within a DA segment has been depicted on the left(A). The right side of the picture (B) shows the new topology immediately after a DA boundary detection. Shaded square nodes represent observable discrete variables, unshaded squares correspond to hidden discrete variables, and shaded circles are associated with continuous observations. Dotted arcs are not really part of the DBN: they symbolise relationships implied by the FLM.

a DA unit can be interpreted as a sequence of words $W_{t-k},\ldots,W_{t-2},W_{t-1},W_t$ with a DA label DA^0 ($DA^0_{t-j} = DA^0$, $\forall j \in [0,k]$). DA^1 will contain the label of the previous DA unit, and DA^2 will go one more step back on the DA recognition history. C is a cyclical counter (from 0 to 5 and back to 0,1,2,...) which is used to count blocks of five words, and N accumulates the encountered word-blocks. Note that since the model's topology changes according to the state of the switching variable E_{t-1}, this is an example of a Bayesian multi-net [17].

Figure 1(A) shows the model's topology when a DA boundary has not been detected (intra-segment phase: $E_{t-1} = 0$). The current DA label DA^0 is responsible for the current sentence $W_t, W_{t-1},\ldots,W_{t-k}$ and the joint sentence probability is estimated through the FLM $p(W_t \mid W_{t-1}, N_t, DA^0_t)$ introduced in section 6. Note that FLMs are fully supported by GMTK, which will automatically take care of the backoff procedure whenever required. The word block counter N needed by the FLM is automatically incremented

whenever the cyclical word counter C reach the fifth word (word block dimension defined in section 6). All the DA label related nodes DA_t^k are simply copied from the previous temporal slice ($DA_t^k = DA_{t-1}^k$ with $k = 0, 1, 2$) since a new DA segment has not yet been recognised.

The state of the end boundary detector E is directly related to the word block counter N and the DA label history DA_t^k through a discrete CPT which is learned during training. The two states of E are linked to continuous feature vectors Y by two sets of Gaussian Mixture Models. Nodes E and Y (together with the associated CPT and GMMs) are fully responsible for the DA segmentation process. If the DA boundaries are known a priori, they can be injected into the model by making E an observable node, and the resulting system will operate as a DA tagger.

If during the previous frame $t-1$ a DA boundary has been detected, the model will be switched to the topology shown in figure 1(B) (inter-segment phase: $E_{t-1} = 1$). Since a new DA unit has been detected at the end of the previous frame $t-1$, both the counters C and N will be set to zero, and the FLM is forced to restart with a new set of estimations. The DA recognition history is updated by copying DA_{t-1}^1 into DA_t^2 and DA_{t-1}^0 into DA_t^1. The new DA hypotheses will be generated by taking in account the current DA label DA_{t-1}^0 and the previous one DA_{t-1}^1 through a trigram language model $p(DA_t^0 \mid DA_{t-1}^0, DA_{t-1}^1)$ (section 4).

The graphs in figure 1 show only the BN slices that are actually duplicated for $t > 1$. During $t = 0$ all the hidden states are properly initialised and the FLM is forced to backoff to $p(W_0 \mid N_0, DA_0^0)$ since W_0 is the first word. During the second frame $t = 1$, DA_1^2 is set to zero and the discourse language model is eventually forced to backoff to a bigram.

8 Experimental setup and performance measures

All the experiments have been performed on the ICSI corpus using the five DA categories and the data sets described in section 3. The system outlined in the previous sections is primarily targeted on the DA recognition task intended as joint segmentation and classification, but as explained in section 7, it is possible to provide the ground truth segmentation and evaluate the DA tagger alone.

The percentage of correctly labeled units is about 76% on reference transcriptions and about 66% on ASR output. The classification procedure is exclusively based on the lexical information (through the FLM) and on the DA language model; prosodic related features are used only for segmentation purposes. Comparing these results with those shown in section 6, we can deduce that the introduction of a trigram discourse model has resulted in an absolute improvement included between 2% (on automatic transcriptions) and 5% (on manual transcriptions).

If performance evaluation is straightforward for the DA tagging task, the same cannot be said about DA segmentation or recognition tasks. Several evaluation metrics have been proposed, but the debate on this topic is still open. In our experiments we have adopted all the performances metrics proposed by Ang et al. [1] and subsequently extended by Zimmermann et al. [8], together with a new recognition metric inherited from

		Metric	LEXICAL	PROSODY	PAUSE	ALL (REF)	ALL (ASR)
T	S	NIST-SU	93.7	83.4	48.0	**35.6**	**43.6**
E	E	DSER	83.6	90.7	51.2	48.9	58.2
S	G	STRICT	87.4	85.8	66.4	56.5	63.5
T	M	BOUNDARY	14.5	12.9	7.4	5.5	7.3
	R	SCLITE	52.7	60.7	48.8	44.6	53.5
S	E	NIST-SU	104.1	93.8	68.5	56.8	69.6
E	C	DER	86.7	92.1	62.9	61.4	72.1
T	O	STRICT	89.1	87.6	72.5	64.7	**72.5**
	G.	LENIENT	20.7	22.0	19.5	**19.7**	**22.0**

Table 2. Segmentation and recognition error rates (%) of five different system configurations.

the speech research community. A detailed description of these metrics (NIST "Sentence like Unit" (SU) derived metrics, strict, lenient and boundary based metrics) can be found in [1]. The DA Error Rate (DER) and DA Segmentation Error Rate (DSER) are discussed in [8].

The speech recognition inspired metric derives from Word Error Rate but having words replaced by DA units. Recognised DA segments are firstly time-aligned against the ground truth annotation, and then the sum of substitution, deletion and insertions errors is scored against the number of reference DA units. This error metric is estimated using the publicly available tool SCLITE (part of the NIST Speech Recognition Scoring Toolkit [4]) which also provides detailed statistics on erroneous segments and significance tests. The SCLITE metric, compared with all the other recognition metrics (except the lenient one), is more focused on a correct DA classification rather than on an extremely accurate segmentation.

Table 2 shows the segmentation and recognition results on five different setups. Results are reported using all the evaluation metrics cited above. Note that all the nine adopted metrics are "error rates", thus lower numbers correspond to better performances. The proposed setups differ only in the information used to detect DA boundaries: the *Lexical* setup makes no use of continuous features (node *Y* has been removed from the DBN), the *Prosody* setup uses only five out of six features (excluding pauses), the *Pause* setup uses the pause information but not the other continuous features, the *All (REF)* and *All (ASR)* configurations exploit the full feature set. *All (REF)* reports the results achieved by training and evaluating the DA recogniser on manually annotated orthographic transcriptions, whenever in *All (ASR)* the system has been developed and tested on automatic transcriptions. Therefore in the later experiment the combination of ASR and DA recogniser constitutes a fully automatic approach, since manual annotations are not needed. Note that the *Lexical* setup makes use of the lexical information just for DA classification purposes. Boundary detection is estimated from the current DA label, the DA history and the word block counter. Therefore this setup and the lexically based systems investigated in [8] cannot be directly compared.

The adoption of prosodic and word related features made in the *Prosody* setup presents a conflicting behaviour: NIST-SU, strict and boundary metrics show an im-

[4] SCTK available from http://www.nist.gov/speech/tools/

provement over the baseline setup; while DSER, DER, lenient and SCLITE based metrics move toward higher error rates. The *Pause* setup shows a clear improvement over the baseline approach under all the evaluation metrics, and proves its strength over the *Prosody* setup highlighting the importance of pause related information on the segmentation task.

The fully integrated approach (*All-REF*) is the most accurate model. The error rates are similar to the NIST-SU segmentation error rate (34.4%) and the lenient recognition error rate (19.6%) of the two step recogniser presented by Ang et al. [1] (section 2). This result suggests that, even if the two competing systems have similar segmentation performances, and the maximum entropy based DA classifier (about 80% correct classification [1]) seems to be more powerful than our generative approach, the joint segmenter+classifier framework is potentially able to outperform a sequential framework. This is even more evident with the fully automatic ASR based system (*All-ASR*) which provides a relevant improvement if compared to the sequential approach outlined in [1] (lenient recognition error rate of 25.1%). In the sequential approach the DA classifier will be able to process only one segmentation hypothesis, whereas in the joint approach multiple segmentation hypotheses are taken in account by the DA tagger. The final choice between multiple candidates will be carried out by taking the most likely sequence of DA units, intended as the optimal combination of DA boundaries and DA labels.

9 Summary and discussion

We have investigated the dialogue act recognition task in multiparty conversational speech, by applying a joint segmentation and tagging approach on natural meetings (ICSI meeting recordings). The proposed system makes use of a heterogeneous set of technologies: a graphical model, a factored language model and some continuous features. The graphical model, implemented as a DBN-based multi-net, oversees the whole recognition process. The proposed model adopts a generative paradigm for the DA tagging task and performs DA segmentation through a feature based architecture. DA tagging is performed using a factored language model over DA labels and word positions, together with a discourse language model. DA segmentation is obtained by exploiting both the DA discourse model and a set of six continuous features extracted from audio recordings and orthographic transcriptions.

The joint DA recognition approach, if compared to a sequential one, provides a clearer view of the addressed problem and an intuitive strategy to its solution. The integrated approach encourages the reuse of common resources such as features and model parts. For example our graphical model shares the DA discourse model between the two subtasks (segmentation and classification), and makes the word block counter required by the FLM available for segmentation purposes (duration model). Furthermore the joint approach operates on a wider search space (producing joint sequences of segmentation boundaries and DA labels based on a trigram discourse model), and thus it is potentially capable of better recognition results. For example the results achieved in our reference transcription based experiments are similar to the sequential DA recognition approach proposed by Ang et al. [1], even though the maximum entropy DA classi-

fication approach chosen by the former work provides a 5% higher tagging accuracy. The advantage of a joint approach is substantial when manual orthographic transcriptions are replaced by imperfect automatic transcriptions. The lenient DA recognition error rate is degraded by only 2.3% and the comparison between sequential and joint approach is in favour of the latter one.

In the near future it is our intention to evaluate the present system on the new AMI meeting corpus [18] and on a richer DA annotation scheme. Moreover we would like to improve both DA classification and DA segmentation by improving the factored language model and by adopting a wider set of multimodal features.

Acknowledgment

We thank Matthias Zimmermann and Elizabeth Shriberg for advice on broad DA categories. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-162).

References

1. J. Ang, Y. Liu, and E. Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. *Proc. of the IEEE ICASSP*, March 2005.
2. E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. *Proc. HLT-NAACL SIGDIAL Workshop*, April–May 2004.
3. A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, (26):339–373, 2000.
4. M. Nagata and T. Morimoto. An experimental statistical dialogue model to predict the speech act type of the next utterance. *Proc. of the International Symposium on Spoken Dialogue*, pages 83–86, November 1993.
5. J. Bilmes and K. Kirchhoff. Factored language models and generalized parallel backoff. *Proceedings of HLT/NAACL 2003*, May 2003.
6. E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, (41):439–487, 1998.
7. H. Hastie, M. Poesio, and S. Isard. Automatically predicting dialogue structure using prosodic features. *Speech Communication*, (36):63–79, 2002.
8. M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke. Toward joint segmentation and classification of dialog acts in multiparty meetings. *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–05)*, 2006.
9. G. Ji and J. Bilmes. Dialog act tagging using graphical models. *Proc. of the IEEE ICASSP*, March 2005.
10. A. Venkataraman, L. Ferrer, A. Stolcke, and E. Shriberg. Training a prosody-based dialog act tagger from unlabeled data. *Proc. of the IEEE ICASSP*, April 2003.
11. A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C.Wooters. The ICSI meeting corpus. *Proc. IEEE ICASSP*, April 2003.
12. T. Hain, M. Karafit, G. Garau, D. Moore, V. Wan, R. Ordelman, and S. Renals. Transcription of conference room meetings: an investigation. *Proc. Interspeech 2005 - Eurospeech, Lisbon*, September 2005.

13. K. Kirchhoff, J. Bilmes, J. Henderson, R. Schwartz, M. Noamany, P. Schone, G. Ji, S. Das, M. Egan, F. He, D. Vergyri, D. Liu, and N. Duta. Novel approaches to arabic speech recognition - final report from the jhu summer workshop 2002. *Tech. Rep., John-Hopkins University*, 2002.
14. A. Stolcke. SRILM an extensible language modeling toolkit. *Proc. Int. Conf. on Spoken Language Processing*, September 2002.
15. K. P. Murphy. Dynamic Bayesian networks: Representation, inference and learning. *Ph.D. Thesis, UC Berkeley, Computer Science Division*, July 2002.
16. J. Bilmes and G. Zweig. The Graphical Model ToolKit: an open source software system for speech and time-series processing. *Proc. IEEE ICASSP*, Jun. 2002.
17. J.A. Bilmes. Dynamic bayesian multinets. *Proc. Int. Conf. on Uncertainty in Artificial Intelligence*, 2000.
18. J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus: A pre-announcement. *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–05)*, 2006.