Dialogue Act Tagging on Spoken Documents: Towards End-To-End Spoken Language Understanding

Ankita Pasad

Toyota Technological Institute at Chicago

ankitap@ttic.edu

Abstract

A dialogue act (DA) is a segment-level tag which describes the function an uttered segment serves in a spoken document. Conventional spoken language understanding systems consist of two main components: an automatic speech recognition (ASR) module that converts audio to text transcript and a natural language understanding module. These modules are typically optimized independently. ASR, being a complex classification task will naturally require much more labelled data when compared to the amount required to train a DA tagger on text. We hypothesize that an audio dialogue act tagger can be trained using very little transcribed speech as compared to a traditional ASR system. While investigating this hypothesis we discover some major discrepancies in the DA tags in a very widely used conversational speech dataset - Switchboard corpus [1].

Index Terms: dialogue act tagging, speech processing

1. Introduction

Identifying dialogue acts is a very important part of understanding human conversations and intelligent human-computer dialogue systems (either written or spoken dialogues). One specific example application could be an autonomous meeting summarizer: knowing aspects such as what opinions were shared, what open-ended questions were asked, could lead to a better quality summary. This information can be obtained using a DA tagger.

Spoken language understanding (SLU) derived applications, DA classification or topic ID for instance, are commonly solved as downstream tasks on word or subword units produced by an ASR [2]. An ASR component is used to transform a speech signal to text tokens, often using a Language Model (LM) alongside, followed by a multi-class classifier that operates on ASR output tokens. This thus requires independently training every component with non-abundant audio (learning the acoustics) and text (learning the LM) data, with supervision on multiple levels in the form of speech segmentation, transcripts for spoken utterances on word and/or phone level, and the downstream task class labels. The downstream tasks, on the other hand, do not typically require as much labeled data for text classification. For instance, [2] empirically shows that degrading ASR performance does not necessarily affect the topic ID accuracy. No such study exists for DA tags.

We hypothesize that it is possible to relax the need of a ASR and thus the strong supervision in the form of speech transcriptions. This is possible if the trained model learns to look for the words/phrases in the spoken segments relevant to particular DA tags. The pipeline approach will also cascade errors made by the ASR and affect the output of the text processing module. This further motivates the advantages of an end-to-end model for language understanding tasks on speech.

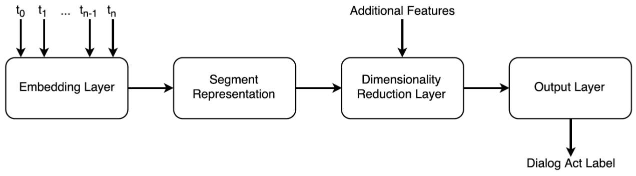

Figure 1: A generic architecture illustrating the individual modules involved in a dialogue act tagger on text [7]

Although DA tagging has been a widely studied task, to the best of our knowledge, no work but [3] attempts this on spoken documents. Stolcke et al. use a pipeline approach using the text-based methods on the output of an ASR system. Recently, efforts have been made to solve such spoken language understanding tasks in an end-to-end (E2E) fashion [4, 5, 6]. We intend to do E2E semantic understanding, specifically, dialogue act tagging on speech data.

1.1. Dialogue Act Tagging on Text

Most models for DA tagging can be broken down into individual blocks as shown 1. These blocks are generic enough to capture changes across models.

The embedding layer encodes the tokensized segment at the input. The weights of the embedding layer can be initialized using word embeddings pretrained on a much larger corpus. *The sentence representation module* combines the token embeddings to generate a vectorial representation of the segment. *The dimensionality reduction layer* is an optional module where more layers could be added to the model to get an even lower dimensional representation while also incorporating any additional features such as speech prosody. Finally, the *output layer* maps the low-dimensional representation into a dialogue act label.

1.2. Dialogue Act Tagging on Speech

The speech signal encodes information about signal properties such as intonation, energy, pauses as well as the word content. The former characteristics are collectively known as speech prosody. Capturing prosodic features from the speech signal has been found useful for DA tagging [8, 9]. For instance, whether an utterance is a declarative question or a statement (e.g: "We are out of watermelons") cannot be determined without listening to the audio. So, given the nature of this task, speech is a more complete source of data than text.

One can argue that previous context, that is easily available for most applications, might help disambiguate such confusions. But how much past context is relevant for the classification of the current utterance might be ambiguous. Answering this will not only help us strengthen the motivation for the task and but also help us analyse the weightage different blocks of the model should get.

We device human subject evaluation experiments to analyze the effects of context and prosody individually on this task. Interesting contradictory observations regarding the dataset were made in the process which have been discussed further in section 5.1.

1.2.1. DA Classification without ASR

As discussed before, we hypothesize that the task of DA tagging does not call for a very accurate ASR module. Instead, it might suffice to have a speech processing unit that (i) infers relevant prosodic features and (ii) detects only those words that help the task. Our goal is, thus, to find an "optimal" ASR regime for this task, where we could get away with a smaller amount of labeled speech data and still achieve a good performance on DA tagging.

In a nutshell, in this work -

1. We try a few vanilla baseline models but they do not give us the expected results, more details in section 5.2. We also propose a set of sophisticated models for the task of DA tagging on spoken documents in section 4.
2. While trying to understand the importance of prosody in the presence of past context, we observe interesting contradictions in the dataset itself. The DA label types on which these ambiguities were observed constitute about 70% of the dataset. Details on the human-subject evaluation experiments designed are in section 5.1 and the observations made are detailed with examples in section 6.

2. Dataset

Switchboard DAMSL [1] is one of the widely used datasets for the task. The dialogue acts, along with other annotations done on switchboard by different groups were all combined together in the NITE XML format [10] to form switchboard NXT [11]. The dataset is manually segmented into utterances suitable for dialog act tags. Every segment has a single tag. The annotators had access to the complete textual conversation. Even though speech is a more natural source of data for this task the annotators did not have any access to the speech input. Inter-annotator agreement for the dataset is 84% using kappa statistics [3]. Table 1 gives data statistics for the dialogue acts.

	train	dev	test
# conversation sides	1,148	114	22
# utterances	105,450	9,192	2,708
# of hours	56.5	5	1.3

Table 1: Dataset statistics; data-split adopted from [12]

Another peculiar thing about this dataset is that there is a large variance in the frequency of occurrence of different labels. The graph in figure 2 shows the long tail effect. Top 3 frequent labels themselves constitute 70 % of the dataset.

3. Previous Work

Dialogue act recognition on switchboard dialogue act corpus has been widely explored. Although switchboard NXT is a shorter version of the original switchboard corpus, this section does not differentiate between the two. Both have the same flavor of data - spontaneous conversational speech corpus. Various types of machine learning based methods have been used to solve the problem. In the recent few years neural network based approaches have surfaced. Thus the discussion on related works has been broadly divided in 2 categories followed by a brief discussion on other datasets that have been used for this task. All these methods perform classification on text (ground truth transcriptions for the switchboard corpus). To the best of our knowledge, Stolcke et al. [3] are the only ones who analyze the performance on speech data. They achieved an accuracy of 71% when applied to natural transcriptions and 64.8% when applied to automatic transcriptions with 41% word error rate on the test set.

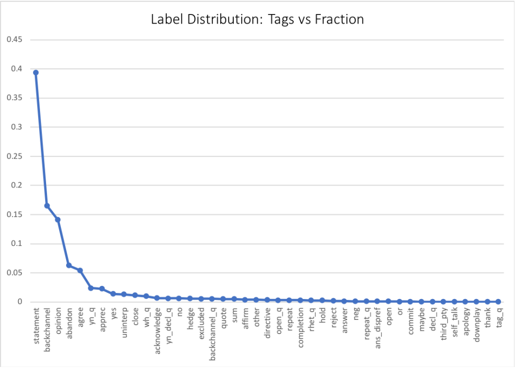

Figure 2: Data distribution

3.1. Neural Network Based Approaches

Kalchbrenner et al. [13] were the first ones to implement a deep neural network architecture for solving this task. Their approach uses a hierarchical convolutional neural network to generate segment representations from randomly initialized word embeddings. They also utilize speaker information in the model which outputs a sequence of dialogue acts. Having a context information of 2 past utterances, they achieved an accuracy of 73.9%. Lee et al. [12] tried both, CNN as well as RNN based approaches to generate sentence representation from GloVe pretrained word embeddings [14]. They observe that CNN based representations (71.4%) perform slightly better than the RNN based ones (69.5%) when passed through a 2 layer feed forward neural network. The context of 2 preceding segments has been used. Ji et al. [15] combined RNNLM with a latent variable model to get the best of both the neural network architectures and the probabilistic graphical models. This model achieved an accuracy of 77%.

Khanpour et al. [16] use an stacked LSTM units to get a segment representation by concatenating the last unit output for 10 LSTM units in a stack. The model is thus able to capture long term dependencies. Liu et al. [17] used 3 parallel CNNs with different context lengths in order to capture different functional patterns. They made an observation that providing the context information in terms of the automatically obtained classification labels gives better results than using the words themselves. Their best model achieves an accuracy of 79.6 %. Ribeiro et al. [7] do a very comprehensive study, experiments and analysis showing the effect each and every module shown in figure 1 has on the final performance. They report the accuracy of

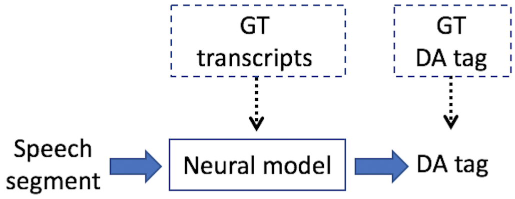

Figure 3: Block diagram for multitask setting for training a dialogue act tagger on speech. Dashed blocks and the arrows indicate modules used only during the training phase.

77.81% using a recurrent convolutional neural network model without using any past context. Lastly, the most recent work by Li et al. [18], uses topic modeling as auxiliary task and observes improvement in the performance of a dialog act tagger.

3.2. Other Methods

Traditionally, multiple approaches have been used for the task of DA tagging. Stolcke et al. [3] relied on hidden markov models using word n-grams as features. Since then multiple attempts have been made to get a good accuracy on this task. These methods include those based on k-nearest neighbors [19], support vector machines [20], maximum entropy model [21]. All these use different feature set including word n-grams, wh-words, punctuations, as with traditional machine learning approaches we have a freedom to design the specific input features derived from the dataset.

3.3. Other Datasets

Switchboard is one of the widely used datasets for the dialogue act tagging task. Apart from switchboard, there are two meeting corpus which have been used for DA tagging - Augmented Multiparty Interaction (AMI) [22] and ICSI [23]. As these are meeting interactions, they are more versatile then the conversational speech corpus. AMI also has visual features recorded and is annotated for gestures, difluencies, summaries along with dialogue acts. These datasets have been used for a multitude of applications, a few researchers have tried to solve dialogue acts on these datasets [24] [25], but again on textual inputs.

4. Proposed Methods

The task here is to perform dialogue act tagging on spoken documents without using an ASR system. So the model has to be designed and trained such that the text transcripts are not required during test time. A very recent work on intent classification attempts the task on speech data in similar settings [5]. They propose multiple models which we can directly adapt in our work - a ‘direct model’ which tries to achieve the task directly on speech without using any annotations, a ‘joint model’ which tries to decode the words and the semantics from speech jointly, ‘multitask model’ which learns a common encoder but separate decoders for decoding words and semantics separately, ‘multistage model’ which learns two encoder-decoder models in series where the first one is trained to decode words and the second one is for learning semantics. The following characteristics about these models make it suitable for adaptation to our tasks -

- They are designed for segment-level classification task on speech data
- None of these models require decoding the speech to get the word sequence during test time
- They are all E2E models and thus no pipeline approach is being used which might lead to cascading errors or necessitate availability of huge amount of annotated spoken data

In the above mentioned study, multitask model works slightly better than the rest for intent classification.

We can formulate our problem in a multitask setting. The idea has been illustrated in Figure 3. In the DA classification module, we can use representations learned at one of the intermediate hidden layers and train a separate decoder on the ASR loss to output the text tokens. Along with this, adding a separate module that embeds prosody features explicitly could help us understand the role of prosody for the task. Using such a learning framework we could answer questions such as, but not limited to,

- How many hours of transcribed speech data would suffice to train a good DA tagger?
- How does the traditional pipeline approach perform, in presence of limited annotated speech data, when compared to the E2E models?
- What words does the “ASR layer” in the multitask framework learn to decode? What does the auxiliary ASR later learn to ignore?
- What common knowledge about the dialogue acts can be derived from these auxiliary ASR errors?

5. Experimentation

5.1. Understanding the Task: Human-Subject Evaluation

One of the motivations for this work is that speech is a more complete source of information as compared to text for the task of dialogue act tagging. As discussed before, the additional information prosody adds over the presence of context is unclear. So we decided to set up human subject evaluation experiments where we will have the following four groups of dataset styles.

1. Textual segments provided in random order, not necessarily from the same conversation
2. Conversation is broken down into segments and is presented in the original order
3. Spoken counterpart of the first setting
4. Spoken counterpart of the second setting

The second and the forth settings have access to the complete past context. The idea is to crowd source these different groups and then get an average human performance on these groups. The effect of having access to the context over having none can then be analyzed by comparing the performance between group 1 and 2 and that between group 3 and 4. The effect of prosody can be judged by the improvement seen on group 3 wrt group 1 and on group 4 wrt group 2. Finally the effect of prosody vs the effect of context can be quantified by the relative performance improvement observed on group 2 over 1 and group 4 over 2.

For the sake of simplicity and disambiguity, the tagset of 43 tags was boiled down to 7. 14 of the tags were clubbed together into 6 tags and the rest were put into “other” category. These 6 groups, which constituted 78.5% of the dataset, are broadly classified as - opinion statement, non-opinion statement, yes-no question, wh question, backchannel, appreciation. The final set up along with the manual having tag descriptions and examples has been made publicly available [1]

5.2. Baseline Models

In order to set up the baselines, I trained vanilla feed forward neural network on text transcripts to output the utterance-level dialogue acts. But due to the nature of the imbalance present in the data distribution, it was hard to optimize the network. The results were stuck to the sub-optimal local minima which predicted the most frequent tag for all the utterances.

Various different settings were tried to alleviate the issue of sub-optimal local minima - training on only frequent labels, training on only non-frequent labels, restricting the vocabulary, different optimization algorithms, using pre-trained word embeddings as the embedding layer, introducing dropouts. But none of these worked and the natural next step is to use more complex neural network models. This has been left to future work.

6. Discussion

Once the test bed for human subject evaluation was set-up, I evaluated 2 conversations from group 1 and 2 each. The average accuracy for group 1 was 66% and that for group 2 was 75.6%. In the process, certain issues with the ground truth labels were detected and have been presented below along with the other observations. Even though I did not evaluate any experiments involving audio from group 3 and 4, I listened to the audio files for certain segments on a case-by-case basis.

The backchannel tag was especially confusing in the absence of context. The manual [1] defines this as a continuer in a conversation, common phrases used are “uh-huh”, “i see”, “yeah”, “um”, “oh”, “really”. The tag was most confused for the utterances such as “yeah” and “really”. These utterances were very commonly used with other intents such as an agreement/answer or a yes-no question respectively. Figuring this out was impossible without access to the context. The access to the audio helped in only a few cases as utterances used as a continuer (backchannel) in a conversation generally sound less confident than the others.

The ground truth labels marked the “yeah” which is followed by other utterances from the listener’s side as a backchannel too. This does not seem quite correct to us as the “yeah” here serves the purpose of an agreement rather than as a continuer in a conversation since the first speaker did not continue the conversation further. Table 2 gives an example for this observation. The snippet of the left is where “yeah” as backchannel was expected since speaker A continues talking. This is not true for the snippet of the right, the “‘yeah” here is clearly intended as an agreement since the listener (side B) supports the previous conversation with more information. All the examples of the form of snippet on the right labeled “yeah” as backchannel, thus hinting at a discrepancy in the instructions conveyed to the annotators.

Opinion vs non-opinion statement was another set of ambiguous tags. The opinion tag has been formally defined as an utterance where the listener has a basis to dispute and non-opinion tag is the one where the listener does not have a basis to dispute. Examples have been shown in table 3. Since these are complete sentences in themselves, context had no effect on the tag for these utterances as far as we are concerned with opinion vs non-opinion confusion. So the examples presented are standalone utterances with their ground truth tag. Example 1 definitely does not present something which could be disputed, so the tag is clearly wrong. In example 2 the listener, in my opinion, could definitely dispute saying that they disagree that those parks are nice. So it should have been labeled as an opinion. Examples 3 and 4 are a clear contradiction, since example 3 was also spoken in the context of mosquitoes.

side	tokenized utterance	side	tokenized utterance
A	Well my husband and I have n't done much camping	A	I I do n't like really camping in the rough
A	we but we bought a van last year	A	I like the the the little necessities
B	yeah	B	yeah
A	and we were hoping uh to do some camping in the van	B	we went we went once to a lean-to

Table 2: Example where discrepancy with backchannel tag was observed. Blue: backchannel

sr. no.	tokenized utterance	tag
1	I do n't know how you can really keep uh the inside of a tent clean	opinion
2	um up here some of the state parks are really nice	statement
3	they 're really terrible	opinion
4	and its its its mosquitoes are terrible	statement

Table 3: Example where discrepancy with opinion/non-opinion tags was observed.

These discrepancies are a major issue because backchannel, opinion and non-opinion statements together constitute 70% of the dataset. This implies that these issues will be prevalent in a large fraction. Quantifying these numbers has been left to future work since we need to crowd-source these tasks and get many more examples evaluated in order to get a good estimate.

Prosody helped despite having access to past context in certain cases. Table 4 provides examples for this observation. The snippet on the right is a case where a question is followed by an agreement. This can be easily confused with statement followed by backchannel. The snippet on the left is an example of a declarative yes-no question. Only after we see the next utterance, “no”, it becomes clear that the previous phrase is a question and not a statement. But we do not assume access to the future context, so based on the just the past context it is hard to guess that “Wuthering Heights” is in fact a question.

side	tokenized utterance	side	tokenized utterance
A	it was n't JANE EYRE	A	They have them right at the campsites
B	WUTHERING HEIGHTS	B	yeah
A	No		

Table 4: Examples where listening to the audio helped over having access to just the past context in textual form. Blue: yes-no question, Red: answer

Although we have some definitive examples where speech prosody is necessary to make a correct decision, such cases are

[1] https://github.com/ankitapasad/daTagging/tree/master/human-subject-evaluation

very limited. So the question still remains whether or not the prosody significantly helps.

7. Conclusion and Future Work

From the observations made on the dataset we conclude that a simple models don't give a good baseline. We need more advanced models - either CNN or RNN architectures. Also, the switchboard nxt corpus, despite its popularity, seems to have some serious ambiguities with important labels. This has also been validated by Li at al. in [18] where the authors present a confusion matrix showing that huge number of confusions occur between opinion and non-opinion statement tags, and between backchannel and acknowledge/agree/answer tags.

The context vs prosody experiments led us to other important observations on the dataset. We could not finish those experiments to get any quantitative results on which (context or prosody) helps more. But based on the discrepancies observed in the switchboard corpus, we are currently looking for a different dataset to perform similar tasks, spoken language understanding on speech. AMI [22] and ICSI [23] look promising if not for the dearth of enough previous work. AMI is a very versatile multimodal meeting corpus with both audio as well as visual features collected and annotated. ICSI is again a meeting corpus. Both being meeting corpus, very wide range of work has been published on these two, including summarization, diarization, detecting agreements/arguments.

8. Acknowledgements

I would like to thank Prof. Karen Livescu for her suggestions and involvement throughout the course of the project.

9. References

[1] D. B. Dan Jurafsky, Liz Shriberg, "Switchboard SWBD-DAMSL," https://web.stanford.edu/~jurafsky/ws97/manual.august1.html, 1997.

[2] T. J. Hazen, "Mce training techniques for topic identification of spoken audio documents," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2451–2460, 2011.

[3] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, 2000.

[4] Y.-P. Chen and R. Price, "Spoken language understanding without speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[5] P. Haghani, A. Narayanan, M. Bacchiani, G. Chuang, N. Gaur, P. Moreno, R. Prabhavalkar, Z. Qu, and A. Waters, "From audio to semantics: Approaches to end-to-end spoken language understanding," *arXiv preprint arXiv:1809.09190*, 2018.

[6] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[7] E. Ribeiro, R. Ribeiro, and D. M. de Matos, "Deep dialog act recognition using multiple token, segment, and context information representations," *arXiv preprint arXiv:1807.08587*, 2018.

[8] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and speech*, 1998.

[9] D. Ortega and N. T. Vu, "Lexico-acoustic neural-based models for dialog act classification," *arXiv preprint arXiv:1803.00831*, 2018.

[10] J. Carletta, S. Evert, U. Heid, and J. Kilgour, "The nite xml toolkit: data model and query language," *Language resources and evaluation*, vol. 39, no. 4, pp. 313–334, 2005.

[11] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language Resources and Evaluation*, 2010.

[12] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," *arXiv preprint arXiv:1603.03827*, 2016.

[13] N. Kalchbrenner and P. Blunsom, "Recurrent convolutional neural networks for discourse compositionality," *arXiv preprint arXiv:1306.3584*, 2013.

[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[15] Y. Ji, G. Haffari, and J. Eisenstein, "A latent variable recurrent neural network for discourse relation language models," *arXiv preprint arXiv:1603.01913*, 2016.

[16] H. Khanpour, N. Guntakandla, and R. Nielsen, "Dialogue act classification in domain-independent conversations using a deep recurrent neural network," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016.

[17] Y. Liu, K. Han, Z. Tan, and Y. Lei, "Using context information for dialog act classification in dnn framework," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[18] R. Li, C. Lin, M. Collinson, X. Li, and G. Chen, "A dual-attention hierarchical recurrent neural network for dialogue act classification," *arXiv preprint arXiv:1810.09154*, 2018.

[19] M. Rotaru, "Dialog act tagging using memory-based learning," *Term project, University of Pittsburgh*, 2002.

[20] B. Gambäck, F. Olsson, and O. Täckström, “Active learning for dialogue act classification,” in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, 2011.

[21] V. K. R. Sridhar, S. Bangalore, and S. Narayanan, “Combining lexical, syntactic and prosodic cues for improved online dialog act tagging,” *Computer Speech & Language*, vol. 23, no. 4, pp. 407–422, 2009.

[22] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, “The ami meeting corpus,” in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[23] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, “The icsi meeting recorder dialog act (mrda) corpus,” Tech. Rep., 2004.

[24] K. Laskowski and E. Shriberg, “Comparing the contributions of context and prosody in text-independent dialog act recognition,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010.

[25] A. Dielmann and S. Renals, “Recognition of dialogue acts in multiparty meetings using a switching dbn,” *IEEE transactions on audio, speech, and language processing*, 2008.