Learning How Productive and Unproductive Meetings Differ

Gabriel Murray

University of the Fraser Valley, Abbotsford, BC, Canada
gabriel.murray@ufv.ca
http://www.ufv.ca/cis/gabriel-murray/

Abstract. In this work, we analyze the *productivity* of meetings and predict productivity levels using linguistic and structural features. This task relates to the task of automatic extractive summarization, as we define productivity in terms of the number (or percentage) of sentences from a meeting that are considered summary-worthy. We describe the traits that differentiate productive and unproductive meetings. We additionally explore how meetings begin and end, and why many meetings are slow to get going and last longer than necessary.

Keywords: productivity, automatic summarization, extractive summarization, meetings

1 Introduction

How can we quantify the intuition that some meetings are less productive than others? We can begin by defining *productivity* within the context of an automatic summarization task. If we employ *extractive* techniques to summarize a meeting by labeling a subset of dialogue act segments (sentence-like speech units) from the meeting as important, then productive meetings would seem to be ones that have a high percentage of important, summary-worthy dialogue acts, while unproductive meetings would have a low percentage of such important dialogue acts.

Given that simple definition of productivity, we can see that productivity (or lack of it) is indeed a critical issue in meetings, and that meetings differ in how productive they are. Using gold-standard extractive summaries generated by human judges on the AMI and ICSI corpora (to be described later), we can index the extracted dialogue acts by their position in the meeting and see from Figure 1 that important dialogue acts are more likely to occur at the beginning of meetings and are less likely at the end of meetings. This suggests that many meetings decrease in productivity as they go on, and may be longer than necessary.

We can also see from Figure 2 that meetings overall have a low percentage of summary-worthy dialogue acts, with an average of only 9% of dialogue acts being marked as summary-worthy in the combined AMI and ICSI corpora. Meetings also greatly differ from each other, with some having 20-25% of dialogue acts

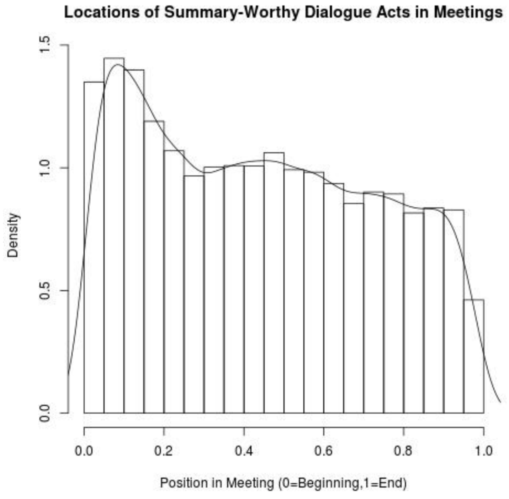

Fig. 1. Histogram/KDE of Extractive Locations

extracted, and others having only 3-4% extracted. We also see that there is a relationship between the length of a meeting and the percentage of summary-worthy dialogue acts, with longer meetings tending to have a smaller percentage of summary dialogue acts.

Another intuition is that meetings often are slow to get going (e.g. featuring idle chit-chat at first) and last longer than necessary (e.g. continuing long after the last decision items and action items have been made). Again viewing this issue from the vantage of extractive summarization, we are interested in predicting the number of dialogue acts that occur *before* the first summary-worthy dialogue act, and the number that occur *after* the last summary-worthy dialogue act. We call these *buffer* dialogue acts because they occur at the beginnings and ends of meetings.

These observations motivate us to explore meeting productivity further. Specifically, in this paper we will carry out two tasks:

- Predict the overall productivity levels of meetings, using linguistic and structural features of meetings.
- Predict the number of *buffer* dialogue acts in meetings, using the same linguistic and structural features.

We use generalized linear models (GLM's) for both tasks. Specifically, we fit a Logistic regression model for the first task and a Poisson regression for the second.

The structure of this paper is as follows. In Section 2, we discuss related work, particularly in the area of extractive meeting summarization. In Section 3, we address the first task above, while in Section 4 we address the second

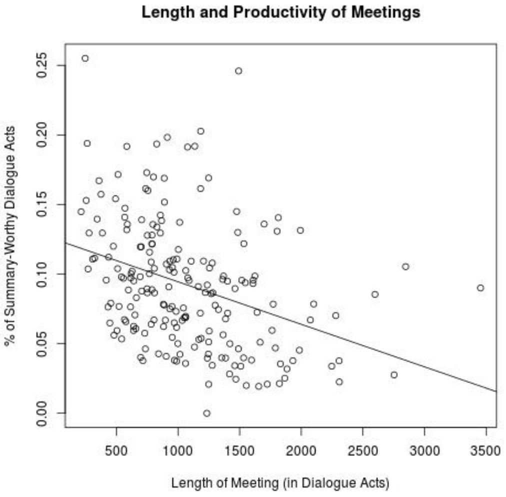

Fig. 2. Length and Productivity of Meetings

task. In Section 5 we describe the experimental setup, including the corpora and evaluation metrics used. Section 6 gives the experimental results, further discussed in Section 7. Finally, we summarize and conclude in Section 8.

2 Related Work

The most closely related work to ours is on meeting summarization, an area that has seen increased attention in the past ten years, particularly as automatic speech recognition (ASR) technology has improved. These range from *extractive* (cut-and-paste) approaches [1–4] where the goal is to classify dialogue acts as important or not important, to *abstractive* systems [5–7] that include natural language generation (NLG) components intended to describe the meeting from a high-level perspective. Carenini et al [8] provide a survey of techniques for summarizing conversational data.

This work also relates to the task of identifying action items in meetings [9, 10] and detecting decision points [11–13]. Renals et al [14] provide a survey of various work that has been done analyzing meeting interactions. We are not aware of other work that has specifically looked in-depth at meeting productivity as we have in this paper.

3 Predicting the Overall Productivity Levels of Meetings: Task 1

In Task 1, the goal is to predict the overall productivity of a meeting, given some linguistic and structural features of the meeting. The productivity is measured as the percentage of meeting dialogue acts labeled as summary-worthy. That is, we are predicting a value between 0 and 1. For that reason, we employ Logistic regression for this task.

Logistic regression is well-known in natural language processing, but is usually used in cases where there are dichotomous (0/1) outcomes, e.g. in classifying dialogue acts as extractive or non-extractive [15]. Unfortunately, we do not have gold-standard labeling of meetings indicating that they were productive or non-productive. However, Logistic regression can also be used in cases where each record has some associated numbers of successes and failures, and the dependent variable is then a proportion or percentage of successes. That is our case here, where each meeting has some number of extractive dialogue acts ("successes") and some remaining non-extractive dialogue acts ("failures").

The Logistic regression model is straight-forward. If we have features (or *predictors*) X and parameters (or *coefficients*) θ, then $\theta^T X$ is a linear predictor. Generalized linear models include some function $g()$ that transforms the predictions. In the case of Logistic regression, the sigmoid function is used:

$$g = \frac{1}{1 + e^{-\theta^T X}}$$

Thus, the predictions are constrained to fall between 0 and 1.

For this task, the meeting-level features we use are described below, with abbreviations for later reference. We group them into feature categories, beginning with **term-weight (tf.idf)** features:

- **tfidfSum** The sum of $tf.idf$ term scores in the meeting.
- **tfidfAve** The average of $tf.idf$ term scores in the meeting.
- **conCoh** The conversation cohesion, as measured by calculating the cosine similarity between all adjacent pairs of dialogue acts, and averaging. Each dialogue act is represented as a vector of $tf.idf$ scores.

Next are the structural features relating to meeting and dialogue act **length**:

- **aveDALength** The average length of dialogue acts in the meeting.
- **shortDAs** The number of dialogue acts in the meeting shorter than 6 words.
- **longDAs** The number of dialogue acts in the meeting longer than 15 words.
- **countDA** The number of dialogue acts in the meeting.
- **wordTypes** The number of unique word types in the meeting (as opposed to word tokens).

There are several **entropy** features. If s is a string of words, and N is the number of words types in s, M is the number of word tokens in s, and x_i is a word type in s, then the word entropy $went$ of s is:

$$went(s) = \frac{\sum_{i=1}^{N} p(x_i) \cdot - \log(p(x_i))}{(\frac{1}{N} \cdot - \log(\frac{1}{N})) \cdot M}$$

where $p(x_i)$ is the probability of the word based on its normalized frequency in the string. Note that word entropy essentially captures information about type-token ratios. For example, if each word token in the string was a unique type then the word entropy score would be 1. Given that definition of entropy, the derived **entropy** features are:

- **docEnt** The word entropy of the entire meeting.
- **speakEnt** This is the speaker entropy, essentially using speaker ID's instead of words. The speaker entropy would be 1 if every dialogue act were uttered by a unique speaker. It would be close to 0 if one speaker were very dominant.
- **speakEntF100** The speaker entropy for the first 100 dialogue acts of the meeting, measuring whether one person was dominant at the start of the meeting.
- **speakEntL100** The speaker entropy for the last 100 dialogue acts of the meeting, measuring whether one person was dominant at the end of the meeting.
- **domSpeak** Another measure of speaker dominance, this is calculated as the percentage of total meeting DA's uttered by the most dominant speaker.

We have one feature relating to **disfluencies**:

- **filledPauses** The number of filled pauses in the meeting, as a percentage of the total word tokens. A filled pause is a word such as *um*, *uh*, *erm* or $mm - hmm$.

Finally, we use two features relating to **subjectivity / sentiment**. These features rely on a sentiment lexicon provided by the SO-Cal sentiment tool [16].

- **posWords** The number of positive words in the meeting.
- **negWords** The number of negative words in the meeting.

4 Predicting the Number of *Buffer* Dialogue Acts in Meetings: Task 2

In Section 1, we introduced the term *buffer* dialogue acts to describe the dialogue acts that occur *before* the first summary-worthy dialogue act and *after* the last summary-worthy dialogue act in the meeting. Intuitively, a high total number of buffer dialogue acts can indicate an unproductive meeting, e.g. a meeting that was either slow to get going or continued longer than necessary, or both. In Task 2, we want to predict the number of buffer dialogue acts for each meeting. For this experiment, we predict the *total* number of buffer dialogue acts, combined from both the beginning and end of the meeting, though we could alternatively predict those separately.

Since our task now is to predict non-negative count data, we use Poisson regression. Like Logistic regression, Poisson regression is a type of generalized linear model. If we have our linear predictor $\theta^T X$, then the transformation function $g()$ for Poisson regression is:

$$g = e^{\theta^T X}$$

Thus, the output is constrained to be between 0 and ∞. For this task, we use the same features/predictors as described in Section 3.

5 Experimental Setup

In this section we briefly describe the corpora and evaluation methods used in these experiments.

5.1 Corpora

In analyzing meeting productivity, we use both the AMI [17] and ICSI [18] meeting corpora. These corpora each include audio-video records of multi-party meetings, as well as both manual and speech recognition transcripts of the meeting discussions. The main difference between the two corpora is that the AMI meetings are scenario-based, with participants who are role-playing as members of a fictitious company, while the ICSI corpora features natural meetings of real research groups.

As part of the AMI project on studying multi-modal interaction [14], both meeting corpora were annotated with extractive and abstractive summaries, including many-to-many links between abstractive sentences and extractive dialogue act segments. We use these gold-standard summary annotations in the following experiments.[1]

5.2 Evaluation

For the following regression experiments, we evaluate the fitted models primarily in terms of the *deviance*. The deviance is -2 times the log likelihood:

$$Deviance(\theta) = -2\ log[\ p(y|\theta)\ ]$$

A lower deviance indicates a better-fitting model. Adding a random noise predictor should decrease the deviance by about 1, on average, and so adding an informative predictor should decrease the deviance by more than 1. And adding k informative predictors should decrease the deviance by more than k.

For both tasks, we perform an in-depth analysis of the individual features used and report the θ parameters of the fitted models. We report a parameter

[1] While we utilize the dialogue act segmentation from those annotations, in this work we make no attempt to classify dialogue act *types* (e.g. *inform*, *question*, *backchannel*) [19].

estimate to be significant if it is at least two standard errors from zero. For the Logistic regression model, the θ parameters can be interpreted in terms of the *log odds*. For a given parameter value θ_n, a one-unit increase in the relevant predictor is associated with a change of θ_n in the log odds. For the Poisson model, given a parameter value θ_n, a one-unit increase in the relevant predictor is associated with the output being multiplied by e^{θ_n}.

6 Results

In this section we present the results on both tasks, first using a Logistic regression model to predict the productivity of each meeting, and then using a Poisson regression model to predict the number of buffer dialogue acts for each meeting.

6.1 Logistic Regression: Task 1 Results

Feature	**Deviance**
null (intercept)	4029.7
tfidfSum	3680.3
tfidfAve	3792.8
conCoh	3825.1
aveDALength	4029.7
shortDAs	3690.7
longDAs	3705.9
countDA	3637.8
wordTypes	3599.4
docEnt	3652.3
domSpeak	3575.2
speakEnt	3882.6
speakEntF100	3758.9
speakEntL100	3825.8
filledPauses	3986.9
posWords	3679.2
negWords	3612.5
COMBINED-FEAS	**2843.7**

Table 1. Logistic Regression: Deviance Using Single and Combined Predictors

For the productivity prediction task, Table 1 shows the deviance scores when using a baseline model (the “null” deviance, using just a constant intercept term), when using individual predictor models, and when using a combined predictor model. We see that the combined model has a much lower deviance (2843.7) compared with the null deviance (4029.7). Using 16 predictors, we expected a decrease of greater than 16 in the deviance, and in fact the decrease is 1186. We can see that the individual predictors with the largest decreases in deviance are *wordTypes*, *docEnt*, *domSpeak* and *negWords*.

Table 2 gives the parameter estimates for the predictors. For completeness, we report parameter estimates using each predictor in a univariate (single predictor) and multivariate (combined predictors) model. Note that the signs, values and significance of the parameter estimates can change between the univariate and multivariate models, e.g. due to correlations between predictors. We restrict most of our discussion to the univariate models. In the univariate models, all parameter estimates except for *aveDALength* are significant (at least two standard errors from zero). Of the three features giving the largest decreases in deviance, *docEnt* and *domSpeak* have positive parameter values while *speakEnt* has a negative value. That is, an increase in the word entropy of the document is associated with an increase in the log odds of productivity, as is an increase in the dominance of the most dominant speaker. This latter fact is also reflected in the negative value of the *speakEnt* parameter. The greater the dominance of the most dominant speaker, the lower the speaker entropy, and the greater the log odds of productivity.

Feature	**θ (Univariate)**	**θ (Multivariate)**
null (intercept)	-	-4.644e+00
tfidfSum	**-3.172e-05**	**2.686e-04**
tfidfAve	**-0.062929**	**-1.139e-01**
conCoh	**8.62097**	-6.542e-02
aveDALength	0.0001451	**2.908e-01**
shortDAs	**-4.896e-04**	**7.970e-04**
longDAs	**-1.460e-03**	**-8.122e-03**
countDA	**-2.865e-04**	**-5.551e-04**
wordTypes	**-5.563e-04**	**-1.689e-03**
docEnt	**5.2647**	**1.912e+00**
domSpeak	**2.3336**	**1.391e+00**
speakEnt	**-4.3420**	-1.801e-01
speakEntF100	**-2.4884**	-4.008e-01
speakEntL100	**-2.79522**	5.711e-01
filledPauses	**3.59026**	-1.590e-01
posWords	**-0.0092129**	7.038e-04
negWords	**-0.007798**	-3.103e-04

Table 2. Logistic Regression: Parameter Estimates (significance indicated by boldface)

We use the trained model to predict on 26 held-out test meetings, and the results are shown in Figure 3, which plots predicted values (the x-axis) against the observed-predicted values (the y-axis). This shows that our model tends to under-predict on the held-out meetings.

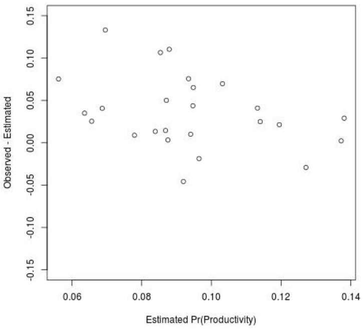

Fig. 3. Productivity Prediction on Held-Out Test Meetings

6.2 Poisson Regression: Task 2 Results

For the buffer dialogue act prediction task, Table 3 shows the null deviance baseline, the deviance scores for individual predictors, and the deviance of the combined model. We can see that the combined model exhibits drastically lower deviance over the null baseline. The decrease is 6722.5, where a decrease of 16 would be expected by adding random noise predictors. The best three predictors in terms of lowering the deviance are *countDA*, *wordTypes* and *docEnt*. The *wordTypes* predictor is likely an effective predictor because, as with *countDA*, it is a correlate of meeting length. Similarly, *docEnt* is likely effective because shorter meetings tend to have higher word entropy.

Table 4 shows the parameter estimates for the predictors, again using both univariate and multivariate models. Of the three best features mentioned above, *countDA* has a positive parameter estimate, meaning that longer meetings tend to have more buffer dialogue acts at the beginnings and ends of meetings. So not only do longer meetings take up more of your time, but that time is not necessarily well-spent. Meetings with a large number of word types (the *wordTypes* predictor) also tend to have more buffer dialogue acts, while higher word entropy *docEnt* tends to indicate fewer buffer dialogue acts.

Examining the parameter estimate for *speakEnt* (as well as *domSpeak*), we see that meetings with a dominant participant tend to have fewer buffer dialogue acts. And meetings that contain many sentiment words (both *posWords* and *negWords*) tend to have more buffer dialogue acts. Together these findings suggest that meetings with many active participants who are expressing opinions do not always make best use of the time.

Figure 4 shows prediction on the 26 held-out test meetings. Here we see a tendency to over-predict the number of buffer dialogue acts in the test set.

Feature	**Deviance**
null (intercept)	12951.8
tfidfSum	8695.1
tfidfAve	9710.2
conCoh	12457.0
aveDALength	12274.0
shortDAs	9377.7
longDAs	8795.5
countDA	8660.4
wordTypes	8309.1
docEnt	8514.2
domSpeak	11873.0
speakEnt	12540.0
speakEntF100	11014.0
speakEntL100	12542.0
filledPauses	12493.0
posWords	8701.6
negWords	9339.1
COMBINED-FEAS	**6229.3**

Table 3. Poisson Regression: Deviance Using Single and Combined Predictors

Feature	θ **(Univariate)**	θ **(Multivariate)**
null (intercept)	-	4.535
tfidfSum	**9.421e-05**	**-6.071e-04**
tfidfAve	**0.201454**	**2.466e-01**
conCoh	**-11.8006**	**-3.092e+00**
aveDALength	**0.145172**	**-1.600e-01**
shortDAs	**1.319e-03**	**-1.080e-03**
longDAs	**4.499e-03**	**1.215e-02**
countDA	**7.849e-04**	**2.428e-03**
wordTypes	**1.596e-03**	**-1.689e-03**
docEnt	**-16.4528**	**-3.810e+00**
domSpeak	**-3.70093**	1.659e-01
speakEnt	**7.11037**	**-1.801e-01**
speakEntF100	**6.1627**	**1.622e+00**
speakEntL100	**4.07004**	**-3.535e+00**
filledPauses	**-10.63467**	-1.283e+00
posWords	**0.0265242**	**7.185e-03**
negWords	**0.0198312**	**-6.018e-03**

Table 4. Poisson Regression: Parameter Estimates (significance indicated by boldface)

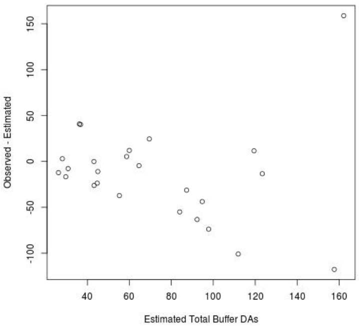

Fig. 4. Buffer DA Prediction on Held-Out Test Meetings

7 Discussion

If we accept that examining meeting summaries is a good proxy for examining meeting productivity, then looking at our two tasks overall it seems we can draw the following conclusions:

- Shorter meetings tend to be more productive.
- Meetings with a dominant participant, i.e. a leader, are more productive.
- Meetings with a large number of sentiment words tend to be less productive.

Though both the Logistic and Poisson models for the two tasks show great improvement in comparison with their respective null deviances, the deviances of the combined predictor models are still fairly high. We plan to do further research on other predictors that may indicate productivity or lack of it. It may also be worthwhile to do some gold-standard labeling of meeting productivity, e.g. enlisting human annotators to make judgments about how productive a meeting was, how well the participants managed the time, whether they achieved their desired decision items, etc. We would also like to make use of participant summaries, which are included in the AMI corpus.

8 Conclusion

In this work, we examined the issue of meeting productivity through the lens of extractive summarization. We carried out two tasks. First, we predicted the productivity levels of meetings using linguistic and structural features. Then we introduced the idea of *buffer* dialogue acts that occur before the first summary-worthy dialogue act and after the last summary-worthy dialogue act in the meeting, and we predicted the number of buffer dialogue acts in a meeting using the same linguistic and structural features. For both tasks, we analyzed and interpreted the individual features used, and found that combined predictor models far outperformed the baselines.

References

1. Murray, G., Renals, S., Carletta, J.: Extractive summarization of meeting recordings. In: Proc. of Interspeech 2005, Lisbon, Portugal. (2005) 593–596
2. Galley, M.: A skip-chain conditional random field for ranking meeting utterances by importance. In: Proc. of EMNLP 2006, Sydney, Australia. (2006) 364–372
3. Xie, S., Favre, B., Hakkani-Tür, D., Liu, Y.: Leveraging sentence weights in a concept-based optimization framework for extractive meeting summarization. In: Proc. of Interspeech 2009, Brighton, England. (2009)
4. Gillick, D., Riedhammer, K., Favre, B., Hakkani-Tür, D.: A global optimization framework for meeting summarization. In: Proc. of ICASSP 2009, Taipei, Taiwan. (2009)
5. Kleinbauer, T., Becker, S., Becker, T.: Indicative abstractive summaries of meetings. In: Proc. of MLMI 2007, Brno, Czech Republic. (2007) poster
6. Murray, G., Carenini, G., Ng, R.: Generating and validating abstracts of meeting conversations: a user study. In: Proc. of INLG 2010, Dublin, Ireland. (2010)
7. Liu, F., Liu, Y.: Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression. IEEE Transactions on Audio, Speech and Language Processing **21**(7) (2013) 1469–1480
8. Carenini, G., Murray, G., Ng, R.: Methods for Mining and Summarizing Text Conversations. 1st edn. Morgan Claypool, San Rafael, CA, USA (2011)
9. Purver, M., Dowding, J., Niekrasz, J., Ehlen, P., Noorbaloochi, S.: Detecting and summarizing action items in multi-party dialogue. In: Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium. (2007)
10. Murray, G., Renals, S.: Detecting action items in meetings. In: Proc. of MLMI 2008, Utrecht, the Netherlands. (2008)
11. Hsueh, P.Y., Kilgour, J., Carletta, J., Moore, J., Renals, S.: Automatic decision detection in meeting speech. In: Proc. of MLMI 2007, Brno, Czech Republic. (2007)
12. Fernández, R., Frampton, M., Ehlen, P., Purver, M., Peters, S.: Modelling and detecting decisions in multi-party dialogue. In: Proc. of the 2008 SIGdial Workshop on Discourse and Dialogue, Columbus, OH, USA. (2008)
13. Bui, T., Frampton, M., Dowding, J., Peters, S.: Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In: Proceedings of the SIGDIAL 2009, London, UK. (2009)
14. Renals, S., Bourlard, H., Carletta, J., Popescu-Belis, A.: Multimodal Signal Processing: Human Interactions in Meetings. 1st edn. Cambridge University Press, New York, NY, USA (2012)
15. Murray, G., Carenini, G.: Summarizing spoken and written conversations. In: Proc. of EMNLP 2008, Honolulu, HI, USA. (2008)
16. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics **37**(2) (June 2011) 267–307
17. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The AMI meeting corpus: A pre-announcement. In: Proc. of MLMI 2005, Edinburgh, UK. (2005) 28–39
18. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI meeting corpus. In: Proc. of IEEE ICASSP 2003, Hong Kong, China. (2003) 364–367
19. Dielmann, A., Renals, S.: DBN based joint dialogue act recognition of multiparty meetings. In: Proc. of ICASSP 2007, Honolulu, USA. (2007) 133–136