Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features

Lai, C., Carletta, J., & Renals, S. (2013). Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features. In Proceedings of Interspeech 2013. (pp. 2723-2727). ISCA.

Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features

Catherine Lai, Jean Carletta, Steve Renals

Centre for Speech Technology Research,
School of Informatics, University of Edinburgh, United Kingdom
clai@inf.ed.ac.uk, j.carletta@ed.ac.uk, s.renals@ed.ac.uk

Abstract

In this paper we investigate how participant involvement and turn-taking features relate to extractive summarization of meeting dialogues. In particular, we examine whether automatically derived measures of group level involvement, like participation equality and turn-taking freedom, can help detect where summarization relevant meeting segments will be. Results show that classification using turn-taking features performed better than the majority class baseline for data from both AMI and ICSI meeting corpora in identifying whether meeting segments contain extractive summary dialogue acts. The feature based approach also provided better recall than using manual ICSI involvement hot spot annotations. Turn-taking features were additionally found to be predictive of the amount of extractive summary content in a segment. In general, we find that summary content decreases with higher participation equality and overlap, while it increases with the number of very short utterances. Differences in results between the AMI and ICSI data sets suggest how group participatory structure can be used to understand what makes meetings easy or difficult to summarize.
Index Terms: Turn-taking, involvement, hot spots, summarization, meetings, dialogue

1. Introduction

Several studies have been motivated by the idea that *involvement detection* should help pinpoint important events in multiparty dialogue, e.g. [1, 2, 3, 4]. The assumption is that parts of the meeting where participants are highly involved will be of interest to external viewers, and hence can act as cues in tasks such as meeting summarization. However, manual annotation of involvement is time consuming and costly. In order to deal with increasing amounts of multiparty meeting recordings becoming available, we would like to know whether automatically derived features can be used directly in detecting those noteworthy segments of meetings to include in a summary.

Currently, involvement detection is generally treated as a supervised learning problem. However, studies differ in how involvement is annotated as ground truth, varying in domain (dialogue act vs conversation vs interval) and in who is actually involved (group vs individual). For example, involvement 'hot spots' in the ICSI Meeting Recorder corpus were initially identified as regions 'about half a minute to one minute in the meeting where more than one participant had a high level of involvement' [1]. In contrast, subsequent annotations labelled specific dialogue acts as hot spots, i.e. attributed to individual speakers [5]. Individual turn-based annotations of involvement can be readily found in work aiming to improve strategies for human computer interaction [6, 7, 8], while multiparty dialogue studies tend towards group oriented notions of involvement or interest, leading to interval- or segment-based labelling schemes [2, 9, 10]. In the latter, annotations vary in terms of what scale is used and again in how group involvement is characterized. In [2] annotators rated 15 second intervals on a scale of 1-5 (low to high interest) based on the perceived degree of interest of the group majority. In [9], the level of active participation is explicitly encoded into a 1-10 scale so that the highest involvement ratings require all participants to be actively participating in a single conversation.

Understanding what is represented as ground truth across these studies is not straightforward. However, it seems clear that group level involvement centers around participation and speaker activation. In this vein, annotators report greater participation across speakers and use of backchannels as guiding heuristics in [2]. Studies relating low level features to involvement annotations similarly reflect this. For example, ICSI hot spots were found to be more prevalent in regions of overlap [11], while features capturing amount of speech/laughter activity from different participants were found to be discriminatory in [12]. Given that we know that affect labelling has relatively low agreement rates [13], we would like to know if automatically derived measures of participation can be used in place of human involvement labels when it comes to tasks like meeting summarization.

Work on automatic meeting summarization has generally focused on extractive summarization, i.e. creating summaries by selecting individual speaker segments/dialogue acts from the transcript. However, in multiparty dialogue, information may be distributed over several turns from different speakers. This diffusion of information causes difficulty when extraction is done with respect to single speaker segments. Including longer distance dependencies has proven to be useful, e.g. question-answer pairs [14, 15]. However, existing studies have not really attempted to leverage high level turn-taking features like participation equality or patterns of speaker switches examined in studies of group decision processes from social psychology, e.g. [16, 17].

In general, Extracted Dialogue Acts (EDAs) have been found to have longer than average duration (or word count) and more non-overlapping speech [18]. In fact, [18] found that duration based features, such as non-overlap duration and next/previous DA latency, give results competitive with classifiers based on feature sets including prosodic, positional and term-weighting information on AMI and ICSI corpora. However, looking only at local overlap and latency misses other aspects of turn-taking structure that may be relevant to summarization. For example, we would expect high levels of backchannel feedback to be important for detecting when de-

Lai, C., Carletta, J., & Renals, S. (2013). Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features. In Proceedings of Interspeech 2013. (pp. 2723-2727). ISCA.

cisions are made. So, we would like to differentiate regions with overlapping backchannel content from those characterized by smooth speaker switches. Similarly, we would like to know what happens in regions where turn-taking structure is less predictable. On the one hand, increased equality of participation, overlap and floor-grabbing freedom, suggests higher participant interest and hence greater relevance for post-meeting browsing. On the other hand, more chaotic turn-taking also suggests greater sharing and diffusion of information [19, 20], which would make selecting specific DAs harder. As such, we would expect that regions that are harder to summarize should contain more EDA material in order to represent what happened.

In the following we show that turn-taking based involvement measures are predictive of whether meeting segments contain extractive summary content (i.e. *summarization* hot spots) in the AMI and ICSI meeting corpora [21]. For the ICSI data, these automatically derived features work as well or better than using manual involvement annotations. We present results of experiments on detecting EDA bearing meeting segments using turn-taking features, and investigate the relation between turn-taking features and the amount of EDA content in topic segments. In general, we find that the probability of finding summary worthy content decreases with higher participation equality and overlap, while it increases with the number of very short utterances, and we discuss the implications of these results for understanding meeting summarizability.

2. Experimental Setup

2.1. Meeting Data and Extractive Summaries

The experiments described in the following were carried out on the ICSI [22] and AMI [23] meeting corpora. The ICSI corpus contains recordings of 75 naturally occuring meetings drawn from 8 different ongoing research groups (3-9 speaker per meeting). As mentioned above, the ICSI dialogue acts have been annotated for involvement hot spots [5]. We use the scenario data from the AMI meetings corpus (140 meetings). Each of these meetings involved 4 speakers who worked on designing a remote control given various informational and budget constraints. Each group participated in a series of 4 meetings focusing on different stages of the design process. Participants were given specific roles within a ficticious company, e.g. project manager, user interface designer.

2.1.1. Extractive Summaries

Manual extractive summaries are available for all of the ICSI meetings and for 131 of the AMI scenario meetings. There were 6 annotators involved in creating summaries, with 2 contributing to both corpora. Annotators selected dialogue acts with the aim of helping an external stakeholder (e.g. department head) understand what happening in the meeting. There was no upper limit on how many dialogue acts annotators could select for the extractive summary, although a rough guideline of 10% was given.

2.1.2. Prediction Domain

In the following experiments we look at predicting the presence of EDA material in two types of segments (i) 15 second windows (drawn every 5 seconds) and (ii) manually annotated topic segments. The length of the former is based on previous work suggesting that extracted segments correspond to about 3-12 seconds in real time, regardless of DA boundaries [14]. With respect to topics, we use subtopics as regions with greater topical coherence and that are more likely to require abstraction over for summarization. These topic segmentations have much more variable duration (ICSI, AMI: μ=127, 158 seconds; σ=145s, 163 seconds).

2.2. Turn-Taking Features

The turn-taking measures used in the following are calculated using *spurts*: segments separated by at least 500ms silence [24], where we use word alignments to mark silence. We look at the following turn-taking features to capture different aspects of participation. Participation equality P_{eq} is defined as:

$$P_{eq} = 1 - \frac{\sum (T_i - T)^2 / T}{E}, \quad (1)$$

where N is the number of participants, T_i total spurt time for participant i, $T = (\sum_i^N T_i)/N$. E represents the maximum possible value of the term under the sum: the average distance from equal participation (so E represents the case when only one participant speaks for the entire segment). Values closer to 1 indicate greater equality [17]. Similarly, let $H(Y|X)$ be the conditional entropy of speaker Y being the next participant to speak after X begins their spurt, with $H_{\text{max}}(Y|X)$ representing the maximal possible value for this. Turn taking freedom F_{cond} is defined as

$$F_{\text{cond}} = 1 - \frac{H_{\text{max}}(Y|X) - H(Y|X)}{H_{\text{max}}(Y|X)}. \quad (2)$$

So, F_{cond} is 0 when turn-taking follows a strict order (i.e. only speaker y follows x) and is 1 when every speaker follows everyone else in equal proportion.

To examine the role of overlaps, we measure barge-in rate (*barge*) as the number of times any spurt in the segment is overlapped by a later starting spurt in the segment. We also measure total amount of overlapping speech (*ovl*) and count the number of Very Short Utterances (*n.vsu*): spurts that have duration less than 500 ms. The latter are likely to represent backchannels or other forms of short feedback [25]. Finally, we measure the proportion of the interval that is silence (*sil*), and for topic segments we record the segment duration (*dur*). All features are converted to z-scores to center and scale the data.

2.3. Regression Models

We use multilevel logistic regression to test whether turn-taking features are predictive of EDA bearing segments. For individual level predictors, we look at the summed duration of hot spots in the segment (*HS Time*), as well as the turn-taking features described above. To account for differences between annotators, meeting types, and corpora (when we combine data from both data sets), as well as the unbalanced nature of the data, we include indicators for these at the group level. That is, we model different annotators as being drawn from a normal distribution (for example) so that the effect of annotators who label fewer meetings is drawn towards the mean [26]. We use a similarly multilevel linear regression model to predict the sum duration of EDAs intersected with the segment in question (converted to log scale). For this part we only consider segments that are know to contain at least some overlap with EDAs. In both cases, model parameters were fit using `lme4` package in R. The following sections report 10-fold cross-validation results. When a group level is not included in the training folds, the effect is taken to be zero as this represents the expected effect of a completely new annotator or meeting type.

Lai, C., Carletta, J., & Renals, S. (2013). Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features. In Proceedings of Interspeech 2013. (pp. 2723-2727). ISCA.

	¬ Hot spot	Hot spot
¬ EDA	0.52	0.06
EDA	0.39	0.04

Table 1: *Proportion of 15 second segments that contain annotated hot spots and EDAs.*

	Accuracy	Precision	Recall	F1
ICSI				
Baseline	0.580	0	0	0
HS Time	0.587	0.533	0.134	0.214
Turn-taking	0.595	0.527	0.355	0.424
AMI				
Baseline	0.718	0.718	1	0.836
Turn-taking	0.774	0.782	0.950	0.858
Combined				
Baseline	0.554	0.554	1	0.713
Turn-taking	0.673	0.696	0.726	0.711
- AMI	0.774	0.777	0.961	0.859
- ICSI	0.590	0.515	0.397	0.448

Table 2: *Detecting 15 second windows containing EDAs: Logistic regression 10-fold cross-validation results.*

3. Results

3.1. Detecting EDA bearing segments

3.1.1. EDAs and Involvement Hot Spots

Table 1 shows the proportion of 15 second segments that overlap with EDAs and/or hot spots in the the ICSI data. We can immediately see that the proportion of hot spots is very small compared to EDAs and majority of the dialogue acts annotated as hot spots are in non-EDA regions. The cross-validation results using sum hot spot time to predict EDA presence on 15 second windows (Table 2, top) are a little better than the majority baseline (no EDAs), however this approach has very low recall. In fact, after fitting logistic regression parameters we see an overall negative effect of log hot spot time (estimate = -0.07, standard error = 0.02, for segments that include hot spots) in predicting whether a segment includes EDAs. So, it seems that hot spots are in fact more indicative of non-summary content.

3.1.2. Turn-Taking Features

Cross-validation results for the binary EDA detection task on 15 second windows for ICSI and AMI corpora are shown in Table 2. The table shows results for classifiers trained on each of the corpora separately, but also for the combined set. Within corpus, turn-taking features perform better than the baseline for both data sets. However, the improvement for the AMI data is much greater. For the ICSI data, nevertheless, using turn-taking features improves the recall quite substantially compared with using only sum hot spot duration. So, turn-taking features are predictive of whether segments contain EDAs.

Figure 1 shows coefficient estimates for turn-taking features for the separate corpus models. Note that the majority class differs between corpora. Accordingly, the AMI intercept estimate is positive, while negative for the ICSI set (similarly for the combined model where corpus is included as a group level effect). However, after accounting for the effects of this, as well as annotators and meeting types we see consistent behaviour across corpora with respect to turn-taking features. Negative effects are estimated for participation equality and barge-in rate. This suggest that we are unlikely to find extractable material in regions with more chaotic turn-taking. However, positive effects for number of VSUs and turn-taking freedom suggest that more short feedback and less strict turn-taking structure are indicative of regions of summary noteworthiness. Similarly, the negative effect of silence proportion indicates that regions where too little is happening are not likely to be summary relevant either.

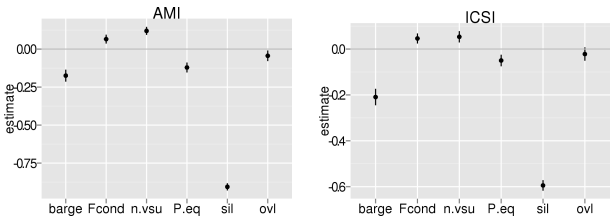

Figure 1: *Coefficient estimates from the two data sets: fixed window size. Vertical bars indicate estimate confidence intervals (2 standard errors).*

	Accuracy	Precision	Recall	F1
ICSI				
Baseline	0.728	0.728	1	0.843
Turn-taking	0.790	0.803	0.94	0.867
AMI				
Baseline	0.889	0.889	1	0.941
Turn-taking	0.913	0.935	0.970	0.952

Table 3: *Detecting topic regions with EDAs, cross-validation results.*

Table 3 shows classification results for classification of topic segments. Since these segments vary in duration we include z-scored log duration of the segment as an individual level predictor. Note, in this case the baseline is high for both corpora, but we still see improvements using the turn-taking features over the majority class baseline for precision and overall accuracy. Figure 2 shows similar estimates as the fixed window sized results for turn-taking features for AMI. For ICSI data, the role of participation equality and turn-taking freedom are less clear, although we see overall negative estimates for the overlap features which again suggest that rougher turn-taking leads to less summary time. Again, we see a positive effect for VSU number for both corpora.

3.2. Amount of EDA Content in Topic Segments

We would like to know if turn-taking features also give us an indication of how much EDA material is in a topic segment,

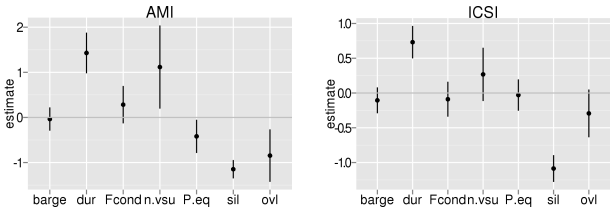

Figure 2: *Logistic regression coefficient estimates for topic segments.*

Lai, C., Carletta, J., & Renals, S. (2013). Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features. In Proceedings of Interspeech 2013. (pp. 2723-2727). ISCA.

	RMSE	95% CI
ICSI: Duration	29.4	(29.2, 29.6)
ICSI: Turn-Taking	28.3	(27.9, 28.6)
AMI: Duration	33.1	(33.0, 33.3)
AMI: Turn-taking	28.6	(28.4, 29.1)

Table 4: *Median and 95% confidence intervals of RMSE of predicted EDA times (in seconds) in topic segments from 10-fold cross-validation with 100 different randomization of the data.*

since we assume that segments containing more EDA material are more noteworthy for summarization. Unsurprisingly, the estimate for topic duration for the binary classification task shows that the probability of a segment containing EDAs increases with topic duration (Figure 2). So, we would like to see if turn-taking features can make better predictions than a duration-only model. Table 4 shows RMSE for linear models predicting sum EDA duration in topic segments (converted back from log scale). As expected there is a positive effect for duration, however we see that turn-taking features do improve on the duration-only model. Estimated coefficients again suggest that there is less summary material in topics with more equal participation and overlap. Estimated effects are negative for all turn-taking features except overlap duration for the ICSI model. For the AMI data, however, we again see positive effects for turn-taking freedom and number of VSUs. This points to fundamental differences in the structure of meetings in these corpora.

4. Discussion

The results reported above show that operationalization of specific aspects of participation gives us more insight into what makes a segment likely to contain extractive summary content than manual involvement annotations. In particular, we observed negative effects for participation equality and overlaps, contrasting with positive effects for the number of VSUs and turn-taking freedom (although the latter effect was less consistent). This makes sense when we consider that information contained in a single speaker turn is more context independent, hence more attractive for summarization, than the same information spread out over several speaker turns. However, meetings are still a *group* process, so participation is still a factor for what makes a dialogue fragment noteworthy. That is, for general overview extractive summaries, the selection problem requires a balance between information density and participation which is harder to obtain when participation is very chaotic. So, it may be the case that regions of very high participation are regions relevant - they are just difficult to include in DA-based extractive summaries. Note, chit-chat was available as a topic category however it only occurs in 15 and 5 times in ICSI and AMI segments respectively. Similarly, 6 ICSI segments have 'joke' or 'laughter' in their topic description. Other meeting relevant regions might require abstraction or compression for summarization. So, interval-based extraction over all participants may be more appropriate than individual speaker turn selection for creating meeting summaries.

If high participation regions are difficult to compress but still relevant to the meeting outcomes, we would expect them to contain more EDA material to compensate. This is not what we found in predicting sum EDA duration in topic segments. However, there are several factors not accounted for in this work that may be relevant, e.g. changes in participation and the spread of information across topics. We still might expect that the less structured turn-taking is within a segment, the more difficult it will be to summarize. So, turn-taking features seem highly connected to the notion of summarizability. In fact, being able to measure summarizability should be useful for several other tasks. For example, extrinsic evaluations have also shown that users don't like meeting browsers with too many links from abstractive summaries to EDA anchors because the abstracts tend to be too vague [27], so knowing what regions are likely to be difficult select extracts from would help creation of more useful abstractive summaries.

Beyond this, further investigation of turn-taking features should help understanding of how summarization strategies should vary for different tasks and meeting types (cf. [28]). We hypothesize that differences in results between ICSI and AMI data are due to differences in the task structure. On the one hand, AMI meetings having a specific goal, predefined role assignments, and little time for off-topic discussion. On the other hand, ICSI meetings explore longer term goals, have more open ended questions, have a longer history, and simply have more participants. Better systems for differentiating high participation segments as topic relevant or rapport building should be relevant for selecting regions to abstract or extract from.

Finally, users may still want to summarize or query meetings on a truly affectual basis, which is something the summaries used in these studies do not capture. In this case, it seems reasonable to build a segmentation selecting regions with high participation features. However, it is perfectly possible for a participant to be affectively engaged or interested without vocally participating in discussions. This is particularly true for face-to-face meetings with a large number of participants, where feedback signals may be more visual, e.g. nods. Further work incorporating multimodal features is necessary to pursue this.

5. Conclusion and Future Work

The experiments reported above show that automatically derived measures of group level involvement, like participation equality and turn-taking freedom, can help identify summarization-relevant meeting segments. Turn-taking features were additionally found to be predictive of the amount of extractive summary content in a segment. In general, we find that summary content decreases with higher participation equality and overlap, while it increases with the number of very short utterances.

The current work only examined turn-taking features based on speaker activity. However, a number of other non-lexical features are likely to be helpful for understanding the relationship between involvement and meeting summarization, e.g. prosodic features and visual cues such as nodding and overall participant movement. As far as we know, no studies measuring summarizability of spoken dialogue have been previously presented, and only a small amount of work has been done on text summarizability [29, 30]. However, lexically based topic coherence features like input entropy identified in those studies should also be helpful for understanding what makes meetings hard to summarize.

6. Acknowledgements

This work was supported by the European Union under the FP7 project inEvent (grant agreement 287872).

7. References

[1] B. Wrede and E. Shriberg, "Spotting "hot spots" in meetings: Human judgments and prosodic cues," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[2] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[3] K. Laskowski, "Finding emotionally involved speech using implicitly proximity-annotated laughter," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5226–5229.

[4] C. Oertel, S. Scherer, and N. Campbell, "On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[5] B. Wrede, S. Bhagat, R. Dhillon, and E. Shriberg, "Meeting recorder project: Hot spot labeling guide," ICSI, Tech. Rep. TR-05-004, 2005.

[6] W. Y. Wang and J. Hirschberg, "Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning," in *Proceedings of the SIGDIAL 2011*. Association for Computational Linguistics, 2011, pp. 152–161.

[7] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[8] K. Forbes-Riley, D. Litman, H. Friedberg, and J. Drummond, "Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system," in *Proc. NAACL-HLT*, 2012.

[9] C. Oertel, F. Cummins, N. Campbell, J. Edlund, and P. Wagner, "D64: a corpus of richly recorded conversational interaction," in *Proceedings of Language Resources and Evaluation Conference (LREC'10)*, 2010, pp. 2992–2995.

[10] F. Bonin, R. Bock, and N. Campbell, "How do we react to context? annotation of individual and group engagement in a video corpus," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 2012, pp. 899–903.

[11] O. Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition," in *Ninth International Conference on Spoken Language Processing*, 2006. [Online]. Available: ftp://ftp.speech.sri.com/pub/papers/icslp06-cetin-Overlap.pdf

[12] K. Laskowski, "Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings," in *Spoken Language Technology Workshop, 2008*, 2008, pp. 81–84.

[13] S. Afzal and P. Robinson, "Natural affect datacollection & annotation in a learning context," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–7.

[14] K. Zechner, "Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres," *Computational Linguistics*, vol. 28, no. 4, pp. 447–485, Dec. 2002.

[15] M. Galley, "Incorporating discourse and syntactic dependencies into probabilistic models for summarization of multiparty speech," Ph.D. dissertation, Columbia University, 2007.

[16] R. Ruback, J. Dabbs, and C. Hopper, "The process of brainstorming: An analysis with individual and group vocal parameters." *Journal of Personality and Social Psychology*, vol. 47, no. 3, p. 558, 1984.

[17] J. Carletta, S. Garrod, and H. Fraser-Krauss, "Placement of authority and communication patterns in workplace groups the consequences for innovation," *Small Group Research*, vol. 29, no. 5, pp. 531–559, 1998.

[18] G. Murray, "Using speech-specific characteristics for automatic speech summarization," Ph.D. dissertation, University of Edinburgh, 2008.

[19] G. Stasser, L. Taylor, and C. Hanna, "Information sampling in structured and unstructured discussions of three-and six-person groups." *Journal of Personality and Social Psychology*, vol. 57, no. 1, p. 67, 1989.

[20] S. Silver, B. Cohen, and J. Crutchfield, "Status differentiation and information exchange in face-to-face and computer-mediated idea generation," *Social Psychology Quarterly*, pp. 108–123, 1994.

[21] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings." *Interspeech'05*, pp. 593–596, 2005. [Online]. Available: http://lac-repo-live7.is.ed.ac.uk/handle/1842/1040

[22] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The ICSI meeting corpus," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol. 1, 2003, pp. I–364.

[23] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[24] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Association for Computational Linguistics, 2003, pp. 34–36.

[25] J. Edlund, M. Heldner, S. Al Moubayed, A. Gravano, and J. Hirschberg, "Very short utterances in conversation," in *Proceedings of fonetik*, 2010, pp. 11–16.

[26] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press Cambridge, 2007.

[27] G. Murray, G. Carenini, and R. Ng, "Generating and validating abstracts of meeting conversations: a user study," in *Proceedings of INLG 2010*, 2010.

[28] H. Christensen and B. Kolluru, "From text summarisation to style-specific summarisation for broadcast news," in *Advances in Information Retrieval*, S. McDonald and J. Tait, Eds. Springer-Verlag, 2004, pp. 223–237.

[29] A. Nenkova and A. Louis, "Can you summarize this? identifying correlates of input difficulty for generic multi-document summarization," in *Proceedings of ACL-08: HLT*, 2008, pp. 825–833.

[30] A. Louis and A. Nenkova, "Predicting Summary Quality using Limited Human Input," in *Proceedings of Text Analysis Conference*, 2009.