Using Meeting Artifacts to Learn About Unobserved Group Dynamics

Gabriel Murray

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

Abstract. It can be useful to model small-group interactions in meetings, but in many cases we do not have audio or video recordings of the meeting itself. However, we very often have meeting artifacts in the form of notes, summaries, slides, and emails. In this work, we investigate whether analysis of some of these artifacts can shed light on aspects of the unobserved group interaction that are not explicitly mentioned within the meeting artifact documents. We present some early positive results on the feasibility of this task.

Keywords: multi-modal interaction, group dynamics, sentiment detection

1 Introduction

When we possess rich multi-modal data streams that capture different aspects of group interaction in a meeting, there are many different interesting analyses we can carry out, such as identifying disagreements, decisions, action items, opinions, dominance relations, and other meeting phenomena. In recent years, multi-modal corpora have been collected and annotated, making this type of analysis feasible. However, most meetings are not recorded. What survives after the meeting is a collection of artifacts such as notes taken during the meeting, short summaries written after the meeting, follow-up emails, slides, and shared documents.

In this work, we consider a small number of meeting artifacts and a small number of in-meeting phenomena, and explore whether we can use those artifacts to predict the unobserved and unrecorded in-meeting phenomena. We can motivate this work using the scenario of a manager who was unable to attend a meeting and wants to know what the group dynamics were like during the meeting. If the manager has only limited artifacts from the meeting and some basic knowledge about the group, such as who was in the meeting and whether they have met before, can the manager get insight into aspects of the group interaction that are not explicitly mentioned in the artifacts? We report some early positive results on the feasibility of this task.

We first describe related work on meeting analysis in Section 2. We then describe the prediction system in Section 3, which includes descriptions of the

meeting corpus, prediction models, and evaluation methods. Results are presented in Section 4 and we conclude in Section 5.

2 Related Work

One of the in-meeting phenomena we try to predict is a sentiment score for the meeting. Previous work on sentiment detection in meetings has attempted to classify individual dialogue act units as being positive or negative [1, 2]. In our case, if the meeting is not recorded then we do not have access to the individual dialogue acts. We instead try to make a prediction of an overall sentiment score for the meeting. This scoring is described in Section 3. There are several recent surveys on sentiment and opinion detection in general [3–5].

Much work has been done on studying multi-modal interaction in meetings more generally [6], including the use of machine learning models to learn about and improve group efficiency and productivity in meetings [7, 8]. There has also been a rich vein of research on modelling group interaction and small group dynamics, including phenomena such as dominance and influence [9–14]. Much of that work has focused on non-verbal cues, while we incorporate both verbal and non-verbal features in these experiments.

Kim and Shah [15] use self-reported summaries to assess whether a group has achieved "consensus of understanding," while Murray [16] shows that a participant's sentiment can greatly differ in their meeting summary vs. in their meeting comments.

3 Leveraging Meeting Artifacts

In these experiments, we use meeting artifacts to try to predict in-meeting phenomena that are not explicitly addressed in the artifacts. We consider each meeting participant individually: that is, given a particular participant's meeting artifacts, we aim to predict in-meeting phenomena relevant to that participant.

Some of these phenomena include the type and intensity of sentiment that was expressed during the meeting. This entails scoring words in the meeting according to their sentiment, and for this we rely on the sentiment lexicon supplied by Taboada et. al [17] as part of their SO-Cal sentiment detection system. The lexicon contains lists of sentiment-bearing adjectives, adverbs, nouns and verbs, each of which is associated with a positive or negative score. Positive scores range from 1 to 5, and negative scores range from -1 to -5.

Taboada et. al, citing Boucher and Osgood [18], note that language tends to have a positive bias, with positive sentiment being more frequently expressed. They hypothesize that negative sentiment carries a larger cognitive weight, and found that increasing their negative word sentiment scores by 50% improved their sentiment detection system in accordance with human judgments of sentiment. We carry out the same 50% increase of the negative word scores.

We then experimented with predicting the following in-meeting phenomena:

1. The sum of positive sentiment word scores for a particular participant in the meeting.
2. The sum of negative sentiment word scores for a particular participant in the meeting.
3. The dominance of the participant, in terms of the fraction of the total meeting time in which they were speaking.
4. The total speaking time of participant.
5. The rate-of-speech of the participant, in words per second.

The AMI Meeting Corpus The meeting data and associated participant summaries are from the AMI meeting corpus [19]. We use the scenario portion of the corpus, where participants are role-playing as members of a company designing a remote control. Each group consists of four members, assigned the roles of project manager, user interface designer, industrial design expert, and marketing expert. Each group goes through a series of four meetings, wherein they discuss different phases of design, finance, and production. After each meeting, the participants were asked to write individual summaries of what happened during the meeting, including any problems that occurred.

Below we show a sample of the types of comments participants make in these post-meeting summaries:

– "A lack of direction in the meetings."
– "I was not convinced myself that some of the trends were desirable to incorporate, and the group confirmed this."
– "Industrial Designer, Alima, who was originally frustrated because she could not find enough information, presented a very coherent explanation of how the remote works."

3.1 Prediction Features and Models

In this prediction experiments, we utilize two types of features: those relating to meeting artifacts, and those conveying basic available information about the group.

Meeting Artifact Features For the meeting artifacts, we limit ourselves in these experiments to participant-authored summaries. The participant summaries were authored by each participant after each meeting, and were not shared with the other participants. The provided template for these summaries encouraged the participants to note any decisions and action items from the meeting, as well as any problems that were encountered during the meeting. The artifact features are as follows:

– **numProblems,probWords**: Respectively, the number of problems noted in the summary and the number of words used to describe those problems.
– **summWords**: The word count of the participant summary.

- **posWordSumm,posScoreSumm**: Respectively, the number of positive sentiment words used in the participant's summary, and the sum of the positive word scores, using the lexicon scoring method described earlier.
- **negWordSumm,negScoreSumm**: Respectively, the number of negative sentiment words used in the participant's summary, and the sum of the negative word scores, using the lexicon scoring method described earlier.
- **normSummScore**: The average sentiment score for all sentiment-bearing words in the participant summary.

Basic Group Features We also consider features conveying basic information about the group and the members' roles:

- **meetA,meetB,meetC**: There are four meetings in the series, A-D. The position in the series is encoded using three binary features.
- **PM,UI,ME** There are four assigned roles in the meeting, encoded by three binary features.

A central question of this work is whether the artifact features can complement the basic group features and improve prediction of the in-meeting phenomena.

3.2 Experimental Setup

Each meeting yields four datapoints, one for each participant. However, not all AMI meetings contain participant summaries. We ultimately ended up with 302 datapoints. For the multiple regression predictions, we report results using 10-fold cross-validation. For the random forest regression, we report out-of-bag prediction results. The evaluation metric used is mean-squared error (MSE).

4 Results

The MSE scores for all of the prediction tasks are shown in Table 1. For each task, we include a baseline MSE score corresponding to a constant intercept prediction. Overall, the random forests and multiple regression methods performed comparably. Those two models consistently outperformed the baseline predictions.

SYS	Speech Rt.	Dom.	Time	+ Sent.	- Sent.
(Intercept)	0.140	0.008	13.23	18868	19878
Regression	0.126	0.005	7.18	12726	11486
Rand. Forest	0.127	0.004	6.70	12939	10574

Table 1. MSE Scores

For predicting all in-meeting phenomena, the best results are found by combining the artifact features with the basic group features. In most cases the basic group features are by far the most useful. For example, the project manager tends to be most dominant, and the first meeting in the series tends to be shorter and to have fewer sentiment-bearing words. Figure 1 shows two measures of variable importance for the negative sentiment prediction task using random forests. The "%IncMSE" score indicates the percentage increase in the MSE when removing that variable, and "IncNodePurity" indicates the node purity when splitting on that variable.

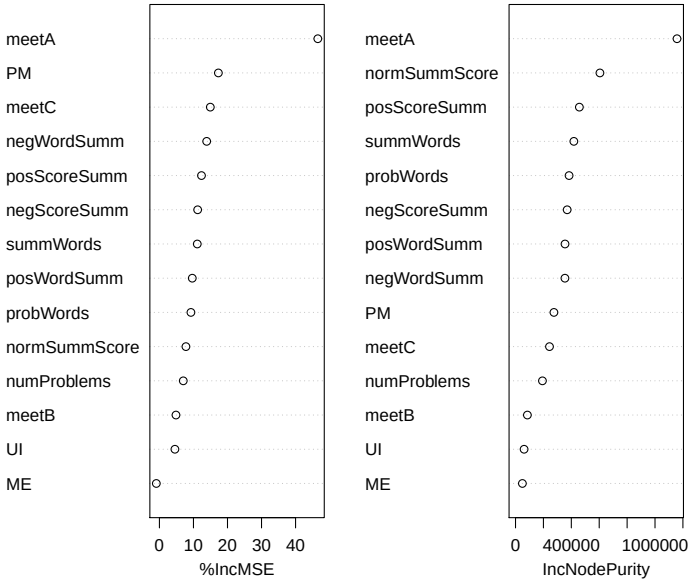

Fig. 1. Variable Importance for Negative Sentiment

In contrast, for the task of predicting speech rate, the artifact features are most useful, as shown in Figure 2. The length of the participant summary and the amount of sentiment in their notes are very good predictors of their speaking rate in the meeting.

We can summarize the results as showing that when we have only some basic information about the small group, we can make a good baseline prediction about these particular unobserved aspects of group dynamics, but that the predictions can be improved by adding information from the meeting artifacts. And for one task in particular, the meeting artifacts were the best predictors.

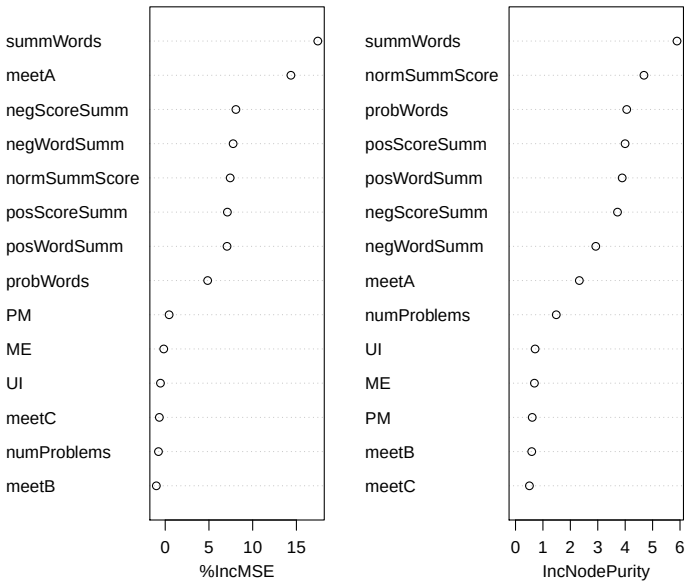

Fig. 2. Variable Importance for Rate of Speech

5 Conclusion

By combining basic knowledge about a group with a small number of features derived from their meeting artifacts, we have shown that we can make predictions about in-meeting phenomena that are consistently and substantially better than baseline predictions.

There are other meeting artifacts that we have not yet exploited. For the AMI corpus, these include whiteboard content, PowerPoint slides, and in-meeting notes. We have also focused on predicting a relatively small number of in-meeting phenomena, relating to sentiment, dominance, speaking time, and speaking rate. In future work, we will use additional meeting artifacts to make predictions about a wider range of in-meeting phenomena. We believe that this vein of research can be valuable in understanding group dynamics in cases where the actual interaction was not recorded.

References

1. Raaijmakers, S., Truong, K., Wilson, T.: Multimodal subjectivity analysis of multiparty conversation. In: Proc. of EMNLP 2008, Honolulu, HI, USA. (2008)
2. Murray, G., Carenini, G.: Subjectivity detection in spoken and written conversations. Natural Language Engineering **17**(03) (2011) 397–418
3. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Now Publishers (2008)
4. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)
5. Liu, B.: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge (2015)
6. 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)
7. Murray, G.: Analyzing productivity shifts in meetings. In: Advances in Artificial Intelligence. Springer (2015) 141–154
8. Kim, B., Rudin, C.: Learning about meetings. Data Mining and Knowledge Discovery **28**(5-6) (2014) 1134–1157
9. Rienks, R., Zhang, D., Gatica-Perez, D., Post, W.: Detection and application of influence rankings in small group meetings. In: Proc. of ICMI 2006, Banff, Canada. (2006)
10. Pentland, A., Heibeck, T.: Honest signals. MIT press (2008)
11. Jayagopi, D., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations from non-verbal activity cues. IEEE Transactions on Audio, Speech and Language Processing **17**(3) (2009) 501–513
12. op den Akker, R., Gatica-Perez, D., Heylen, D.: Multi-modal analysis of small-group conversational dynamics. In Renals, S., Bourlard, H., Carletta, J., Popescu-Belis, A., eds.: Multimodal Signal Processing. Cambridge University Press, New York (June 2012) 155–169
13. Dong, W., Lepri, B., Pianesi, F., Pentland, A.: Modeling functional roles dynamics in small group interactions. IEEE Transactions on Multimedia **15**(1) (2013) 83–95
14. Frauendorfer, D., Mast, M.S., Sanchez-Cortes, D., Gatica-Perez, D.: Emergent power hierarchies and group performance. International Journal of Psychology (2014)
15. Kim, J.H., Shah, J.: Automatic prediction of consistency among team members' understanding of group decisions in meetings. In: Proc. of IEEE SMC. (2014) 3702–3708
16. Murray, G.: Uncovering hidden sentiment in meetings. In: Proc. of Canadian AI 2016, Victoria, BC, Canada. (2016)
17. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics **37**(2) (June 2011) 267–307
18. Boucher, J., Osgood, C.: The pollyanna hypothesis. Journal of Verbal Learning and Verbal Behavior **8**(1) (1969) 1–8
19. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. In: Proc. of LREC 2006, Genoa, Italy. (2006) 181–190