Resources for Analyzing Productivity in Group Interactions

Gabriel Murray

University of the Fraser Valley
Abbotsford, BC, Canada
gabriel.murray@ufv.ca

Abstract

Productivity can vary both within and across meetings. In this work, we consider the question of how to measure productivity, and survey some of the available and potential resources that correspond to productivity. We then describe an initial experiment in which we define productivity in terms of the percentage of sentences from a meeting that are considered summary-worthy. Given that simple definition of productivity, we fit a logistic regression model to predict productivity levels of meetings using linguistic and structural features.

Keywords: productivity, multimodal interaction, extractive summarization

1. Introduction

How can we measure *productivity* in group interactions? In the absence of gold-standard annotations for productivity, 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 acts 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.

Starting with that simple definition of productivity, we can see that productivity is indeed a critical issue in meetings, and that meetings differ in how productive they are. Using gold-standard extractive summaries of 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. Figure 2 shows that productivity also varies *between* meetings, e.g. longer meetings tend to have a smaller percentage of summary dialogue acts.

This paper discusses our corpora and initial experiments for analyzing meeting productivity. In Section 2. we discuss related work. In Section 3. we describe the two corpora we are currently using in terms of their available resources that relate to productivity, as well as potential new resources. In Section 4. we describe an experiment where productivity is defined in relation to an extractive summarization task. Section 5. gives the results of that first experiment, and we conclude in Section 6.

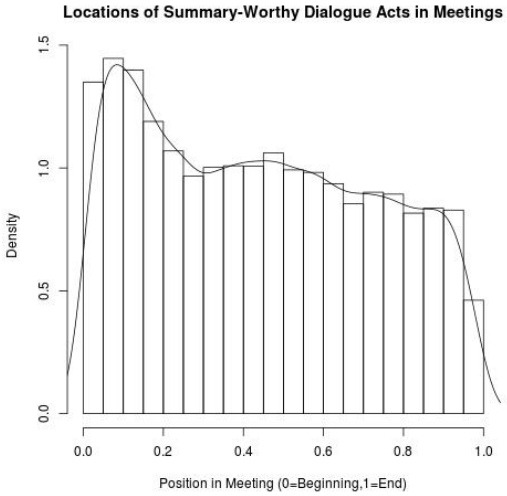

Figure 1: Histogram/KDE of Extractive Locations

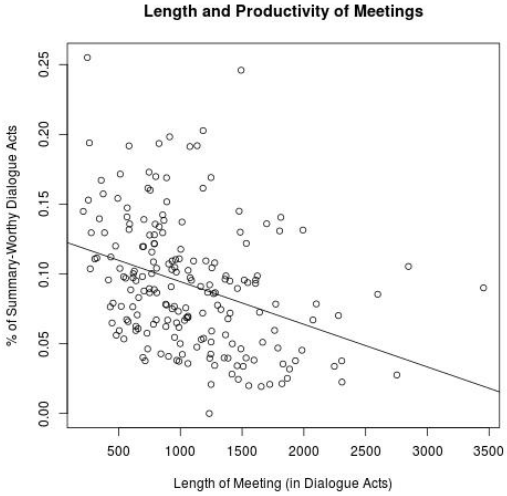

Figure 2: Length and Productivity of Meetings

2. Related Work

This work closely relates to meeting summarization, including *extractive* (Zechner, 2002; Murray et al., 2005; Galley, 2006) and *abstractive* (Kleinbauer et al., 2007; Murray et al., 2010) approaches. Carenini et al (2011) provide a survey of techniques for summarizing conversational data. This work also relates to the task of identify-

ing action items in meetings (Purver et al., 2007; Murray and Renals, 2008; Morgan et al., 2006), detecting decision points (Hsueh et al., 2007; Fernández et al., 2008; Bui et al., 2009), and detecting speaker dominance (Rienks and Heylen, 2005; Rienks et al., 2006; Jayagopi et al., 2009). Renals et al (2012) provide a survey of various work that has been done analyzing multimodal interactions, while op den Akker et al (2012) give a survey of research investigating conversational dynamics in meetings.

3. Corpora

In analyzing meeting productivity, we use both the AMI (Carletta et al., 2005; Carletta, 2006) and ICSI (Janin et al., 2003) 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.

3.1. Available Resources

As part of the AMI project on studying multi-modal interaction (Renals et al., 2012), both meeting corpora were annotated with extractive and abstractive summaries, including many-to-many links between abstractive sentences and extractive dialogue acts. We use these gold-standard summary annotations in the following experiment.

Available resources that are *not* used in the experiment described here, but which may end up being useful in follow-up work, include:

- **Decision items:** Some dialogue acts are linked to the "decisions" portion of the abstractive summary; these constitute a specific subset of the summary dialogue acts we use. For example, a meeting may be considered more productive if it contains more decision items.
- **Action items:** Similarly, some dialogue acts are linked to the "action items" portion of the abstractive summary.
- **Dominance** annotation of Rienks and Heylen (2005). As described in Section 4., we do utilize simple features relating to dominance.
- **Participant Summaries:** In the AMI corpus, meeting participants individually authored short summaries after each meeting. These may yield clues on how productive or efficient they perceived the meetings to be.
- **Subjectivity Annotation:** AMI meetings have been annotated for various subjective criteria (Wilson, 2008). This could be useful if it turned out that less productive meetings are more associated with negative subjectivity, to give one example.

3.2. Proposed Resources

While the experiment described below in Section 4. shows promising results using extractive annotations as a proxy for productivity, we may need to more directly address the issue by directly annotating for the phenomenon. This could include:

- **Meeting-Level Annotation:** Meetings could be either categorized (e.g. unproductive, productive) or rated on a scale of productivity (e.g. 1-10). This would be the least costly and time-consuming annotation, and least difficult for the human judges.
- **Dialogue Act-Level Annotation:** Individual dialogue acts could be rated on how much they contributed to meeting productivity. We expect that this could be difficult for many ambiguous dialogue acts, and would likely be a challenging, time-consuming task for human judges.
- **Turn-Level Annotation:** If we define a turn as a sequence of dialogue acts by the same speaker, then each turn could be rated on how much it contributed to meeting productivity. This would be less costly and time-consuming than labeling of every dialogue act.
- **Participant-Level Annotation:** Each participant in the meeting could be rated on how much of a contribution they made to meeting productivity. In conjunction with one or more of the other proposed annotations above, we could learn how individual participants affect the productivity outcome of a meeting.

4. Predicting Productivity of Meetings

In this initial experiment, the task 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 (Murray and Carenini, 2008). 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").

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 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 (Taboada et al., 2011).

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

5. Experimental Results

For this experiment, we evaluate the fitted model 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.

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: Deviance Using Single and Combined Predictors

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*, *domSpeak*, and *negWords*.

6. Conclusion

Using the percentage of extracted dialogue acts as a proxy for a meeting's productivity, we have shown that a logistic regression model can predict productivity effectively based on linguistic and structural features. In our ensuing work, we plan to leverage available resources such as *dominance* and *sentiment* annotations, as well as participant summaries. We will also begin meeting-level annotations of productivity in order to more directly study this phenomenon.

7. References

T. Bui, M. Frampton, J. Dowding, and S. Peters. 2009. Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proceedings of the SIGDIAL 2009, London, UK*.

G. Carenini, G. Murray, and R. Ng. 2011. *Methods for Mining and Summarizing Text Conversations*. Morgan Claypool, San Rafael, CA, USA, 1st edition.

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Proc. of MLMI 2005, Edinburgh, UK*, pages 28–39.

J. Carletta. 2006. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. In *Proc. of LREC 2006, Genoa, Italy*, pages 181–190.

R. Fernández, M. Frampton, P. Ehlen, M. Purver, and S. Peters. 2008. Modelling and detecting decisions in multiparty dialogue. In *Proc. of the 2008 SIGdial Workshop on Discourse and Dialogue, Columbus, OH, USA*.

M. Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proc. of EMNLP 2006, Sydney, Australia*, pages 364–372.

P-Y. Hsueh, J. Kilgour, J. Carletta, J. Moore, and S. Renals. 2007. Automatic decision detection in meeting speech. In *Proc. of MLMI 2007, Brno, Czech Republic*.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *Proc. of IEEE ICASSP 2003, Hong Kong, China*, pages 364–367.

D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. 2009. Modeling dominance in group conversations from non-verbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing*, 17(3):501–513.

T. Kleinbauer, S. Becker, and T. Becker. 2007. Indicative abstractive summaries of meetings. In *Proc. of MLMI 2007, Brno, Czech Republic*, page poster.

W. Morgan, P-C. Chang, S. Gupta, and J. Brenier. 2006. Automatically detecting action items in audio meeting recordings. In *Proc. of the 7th SIGdial Workshop on Discourse and Dialogue*.

G. Murray and G. Carenini. 2008. Summarizing spoken and written conversations. In *Proc. of EMNLP 2008, Honolulu, HI, USA*.

G. Murray and S. Renals. 2008. Detecting action items in meetings. In *Proc. of MLMI 2008, Utrecht, the Netherlands*.

G. Murray, S. Renals, and J. Carletta. 2005. Extractive summarization of meeting recordings. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 593–596.

G. Murray, G. Carenini, and R. Ng. 2010. Generating and validating abstracts of meeting conversations: a user study. In *Proc. of INLG 2010, Dublin, Ireland*.

R. op den Akker, D. Gatica-Perez, and D. Heylen. 2012. Multi-modal analysis of small-group conversational dynamics. In S. Renals, H. Bourlard, J. Carletta, and A. Popescu-Belis, editors, *Multimodal Signal Processing*, pages 155–169. Cambridge University Press, New York, June.

M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, and S. Noorbaloochi. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium*.

S. Renals, H. Bourlard, J. Carletta, and A. Popescu-Belis. 2012. *Multimodal Signal Processing: Human Interactions in Meetings*. Cambridge University Press, New York, NY, USA, 1st edition.

R. Rienks and D. Heylen. 2005. Automatic dominance detection in meetings using easily obtainable features. In *Proc. of MLMI 2005, Edinburgh, UK*.

R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post. 2006. Detection and application of influence rankings in small group meetings. In *Proc. of ICMI 2006, Banff, Canada*.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, June.

T. Wilson. 2008. Annotating subjective content in meetings. In *Proc. of LREC 2008, Marrakech, Morocco*.

K. Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.