Analyzing Productivity Shifts in Meetings

Gabriel Murray

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

Abstract. Group productivity can vary between and within meetings, and here we consider the case of productivity shifting within meetings. We divide a meeting into intervals and measure the productivity of each interval using the number of summary-worthy sentences contained therein. We evaluate the relationship between productivity and a variety of linguistic and structural features, using correlation and regression analysis. We then attempt to identify the point at which productivity shifts in meetings, using Bayesian changepoint analysis.

1 Introduction

A motivating intuition of this work is that extractive summarization can operate as a proxy for assessing productivity in meetings. If a meeting is highly productive, there should correspondingly be a high number of extracted sentences, relating to phenomena such as decisions, action items, and generally active, on-task discussion. But even a productive meeting may not be consistently productive throughout. For example, a meeting may have many extracted sentences from the first half of the meeting and then very few summary-worthy sentences in the second half, perhaps because participants were becoming tired or simply because the conversation continued long after all the decision items were addressed. Throughout this work, we assume that the number of extracted sentences (or more properly, extracted *dialogue act* units) reflects the group productivity level, both for the meeting as a whole and for intervals of a meeting.

Give that measurement of productivity, it is clear that productivity is not consistent within meetings. Figure 1 shows that extracted dialogue acts are more frequent at the beginnings of meetings and less frequent at the ends of meetings. This suggests that many meetings begin productively but become less productive as they go on.

In this paper, we analyze how productivity can shift *within* meetings. We divide meetings into intervals and perform correlation and regression analyses to determine which linguistic and structural features are closely associated with rising and falling productivity. We are also interested in why a meeting suddenly becomes less productive or suddenly becomes more productive. We use Bayesian changepoint analysis to find the spot in the meeting where productivity shifts, and then aim to learn the traits that characterize the less-productive portion of the meeting from the more-productive portion, using a logistic regression model.

The central contributions of this work are as follows:

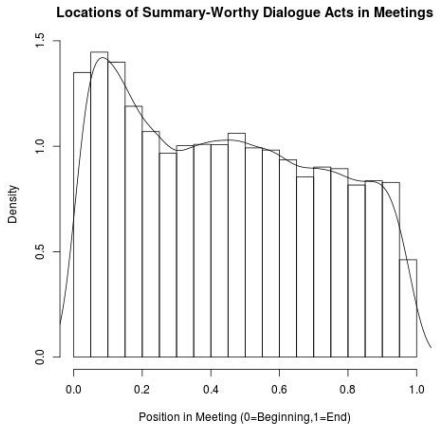

Fig. 1. Histogram + KDE Showing Meeting Position for Extracted Sentences

1. This is the first in-depth investigation of how productivity shifts within meetings.
2. We include detailed analysis of linguistic and structural features relating to productivity.
3. We introduce a novel application of Bayesian changepoint analysis for detecting changes in productivity.

The structure of the paper is as follows. Section 2 describes related work on extractive summarization and changepoint analysis. Section 3 introduces the features used in these experiments, with a correlation and regression analysis. Section 4 describes the application of Bayesian changepoint analysis to our particular problem, while Section 5 describes the logistic regression model. The experimental setup for the changepoint experiments is described in Section 6 while the main results are presented in Section 7. Finally, we discuss future work in Section 8 and offer our main conclusions in Section 9.

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 [1–3] where the goal is to classify dialogue acts as important or not important, to *abstractive* systems [4–7] 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] and detecting decision points [10]. Renals et al.

[11] provide a survey of various work that has been done analyzing meeting interactions.

Other research [12, 13] has looked at productivity *across* meetings, i.e., determining how productive and unproductive meetings differ in terms of linguistic and structural features. However, that work does not examine how productivity shifts *within* meetings. For example, even meetings that are mostly productive will have periods where participants get off task and are less productive.

The classic example of applying Bayesian changepoint analysis involves a much-studied dataset of historical coal-mining accidents in Britain [14, 15]. The changepoint analysis reveals that such accidents had a marked decrease after the introduction of new safety regulations in the late 1880's.

3 Correlation and Regression Analysis

To analyze productivity shifts within meetings, we divide each meeting into one-minute intervals. We then count the number of extracted dialogue acts (a feature *numSum*) for each interval. We also extract a variety of linguistic, structural and speaker-related features for each interval. We group them into feature categories, beginning with **term-weight (tf.idf)** features:

- *tfidfSum* The sum of $tf.idf$ term scores in the meeting portion.
- *tfidfAve* The average of $tf.idf$ term scores in the meeting portion.
- *conCoh* The conversation cohesion in the meeting portion, 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**:

- *DALength* The average length of dialogue acts in the meeting portion.
- *countDA* The number of dialogue acts in the meeting portion.
- *wTypes* The number of unique word types in the meeting portion (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 portion.

- *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.
- *domSpeak* Another measure of speaker dominance, this is calculated as the percentage of total meeting portion DA's uttered by the most dominant speaker.

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

- *fPauses* The number of filled pauses in the meeting portion, 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 portion.
- *negWords* The number of negative words in the meeting portion.

Figure 2 depicts a correlogram for the features, illustrating the correlations between all features. In the portion below the diagonal, the lines and colouring indicate a positive or negative correlation (blue=positive,red=negative), and the shading indicates the strength of the correlation. The portion above the diagonal shows the confidence ellipses and smoothed lines. Of particular interest is the first row and first column *numSum*, corresponding to the number of extracted dialogue acts in each interval, our dependent variable in the regression analysis below. We can see that most of the features have a positive correlation with *numSum*, with $tfidfSum$ having the strongest positive correlation. Only document entropy (*docEnt*) and speaker entropy (*speakEnt*) have negative correlations with the number of summary-worthy dialogue acts in each interval. Interestingly, the total number of dialogue acts in an interval does not strongly correlate with the number of summary-worthy dialogue acts in the interval.

We construct a Poisson regression model using *numSum* as the dependent variable and the other features as predictors. We can evaluate the fitted model using the *deviance* measure. 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. The deviance of our fitted model is 18049, compared with a baseline intercept-only deviance of 19046. Since we added 12 predictors, we should expect the deviance to decrease by at least 12 over the baseline, and in fact it decreased by about 1000.

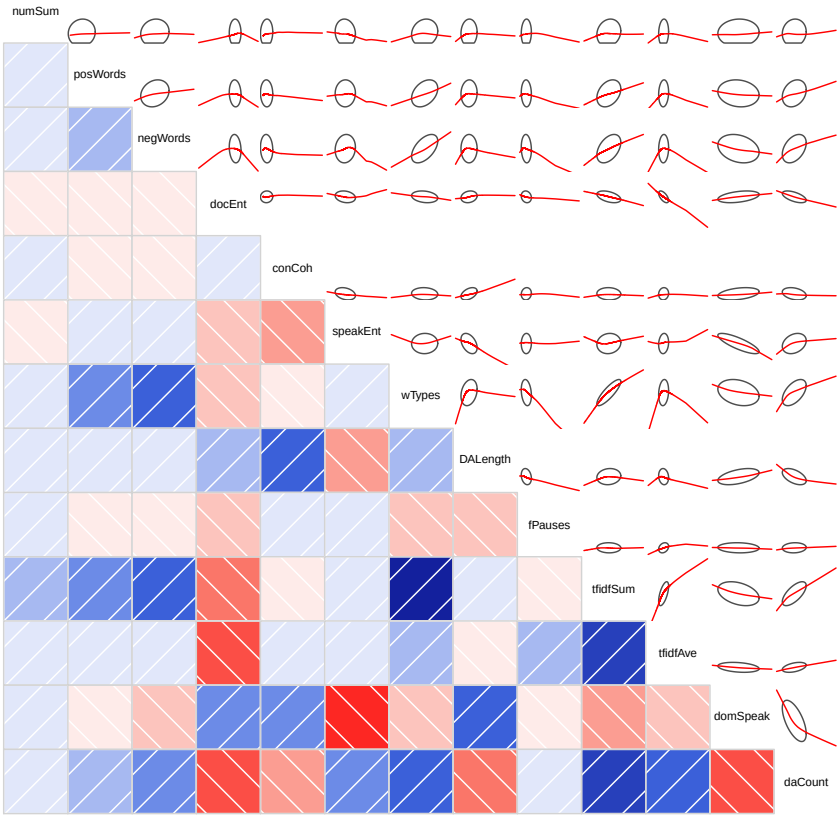

Fig. 2. Feature Correlogram

4 Bayesian Changepoint Analysis for Meeting Productivity

For changepoint estimation, our time-series data consists of the number of extracted (summary-worthy) dialogue acts in each one-minute interval of a meeting. So for a given meeting, our number of datapoints n will be equal to the length of the meeting in minutes. We hypothesize that in many meetings there will be a shift where the number of extracted sentences per minute markedly increases or decreases. Our first task is to locate the changepoint, if there is one, for each meeting.

In more general terms, we have count data $x_1, \cdots, x_n$ and a possible changepoint at some point k in the series. If it is determined that $k = n$, then there is no changepoint. If there *is* a changepoint k, we would then have two Poisson data-generating processes:

$$\begin{aligned} x_i \mid \lambda &\sim \mathcal{P}(\lambda) \quad i = 1, \cdots, k \\ x_i \mid \phi &\sim \mathcal{P}(\phi) \quad i = k+1, \cdots, n \end{aligned} \tag{1}$$

The parameters that we want to estimate are λ, ϕ and k. The priors for each of these are:

$$\begin{aligned} &\lambda \sim \mathcal{G}(\alpha, \beta) \\ &\phi \sim \mathcal{G}(\gamma, \delta) \\ &k \sim \text{discrete uniform on } [1, 2, \cdots, n] \end{aligned} \tag{2}$$

We use Gibbs sampling to estimate these parameters, and each step of Gibbs sampling uses the following conditional probabilities:

$$\begin{aligned} \lambda \mid \phi, k &\sim \mathcal{G}(\alpha + \sum_{i=1}^{k} x_i, \beta + k) \\ \phi \mid \lambda, k &\sim \mathcal{G}(\gamma + \sum_{i=k+1}^{n} x_i, \delta + n - k) \\ p(k \mid x, \lambda, \phi) &= \frac{L(x, \lambda, \phi \mid k)\, p(k)}{\sum_{l=1}^{n} L(x, \lambda, \phi \mid k_l)\, p(k_l)} \end{aligned} \tag{3}$$

where $p(k)$ in the last line is a uniform prior. For these experiments we set the prior parameters directly ($\alpha = \beta = \gamma = \delta = 10$), though they could be estimated as well. In any case, we found that the final results were not particularly sensitive to the choice of hyperparameter.We run the Gibbs sampler for 50 replications of 10,000 iterations each, with a burn-in of 1000. We estimate k using the posterior mean over the 50 replications. Figure 3 shows an example of the Gibbs sampling updates for k, λ and ϕ for one particular meeting.

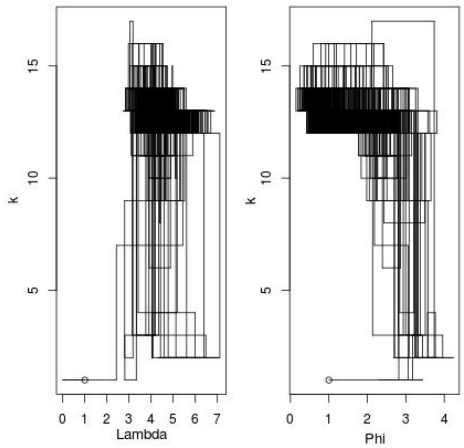

Fig. 3. Gibbs Sampling Updates

5 Classification of Productive/Unproductive Meeting Portions

Once the changepoint has been determined, we want to learn the differences between the two portions of the meeting: the portion before the changepoint, and the portion after. To do so, we differentiate between the following two cases:

- **Case 1 – Increasing Productivity:** These are meetings where the productivity increases after the changepoint. In such meetings, we want to learn the differences between the less-productive beginning of the meeting and more-productive end.
- **Case 2 – Decreasing Productivity:** These are meetings where the productivity decreases after the changepoint. In such meetings, we want to learn the differences between the more-productive beginning of the meeting and the less-productive end.

In both cases, we want to learn the linguistic and structural features that characterize the less-productive from the more-productive portion. But we treat these cases above as two separate machine-learning tasks because, for example, an unproductive beginning of a meeting may have much different characteristics than an unproductive end of a meeting. In fact, we will later see that this is the case.

In addition to the features described in Section 3, we also use:

- *speakEntF100* The speaker entropy for the first 100 dialogue acts of the meeting portion.

- *speakEntL100* The speaker entropy for the last 100 dialogue acts of the meeting portion.
- *shortDAs* The number of dialogue acts in the meeting portion shorter than 6 words.
- *longDAs* The number of dialogue acts in the meeting portion longer than 15 words.

For this binary classification task, we use a standard logistic regression model where our predictions $\theta^T X$ are constrained by the sigmoid function to fall between 0 and 1: $g = \frac{1}{1+e^{-\theta^T X}}$ We employ a leave-one-out procedure to maximize our training data, testing on each meeting individually after training on the rest. For both Case 1 and Case 2 meetings, we consider the more-productive portion to be the "positive" class.

6 Experimental Setup

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

Corpora In analyzing meeting productivity, we use both the AMI [17] and ICSI [18] meeting corpora. As part of the AMI project on studying multi-modal interaction [11], 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 experiments. There are 197 meetings in total. Since we are performing classification at the level of meeting sub-portions, each meeting contributes two datapoints: a more-productive portion and a less-productive portion. This gives us a total of 394 training examples, requiring the leave-one-out procedure.

Evaluation For the logistic regression experiments, we evaluate our predictions on test data using precision/recall/F-score. We also evaluate the fitted models primarily in terms of the *deviance*, described earlier in Section 3. We also present the θ parameters of the fitted logistic regression models. 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.

7 Results

We first present key results of the Bayesian changepoint analysis, followed by results of the classification task.

7.1 Bayesian Changepoint Results

Of the 197 meetings used in these experiments, only one meeting was determined to have a changepoint very close (within two minutes) to the end of the meeting, and three meetings had changepoints very close to the beginning of the meeting. We excluded those meetings from the remainder of the experiments and focused on the 193 meetings that featured a clear shift in productivity. Of those, 75 meetings fall into Case 1 (increasing productivity after the changepoint) and 118 fall into Case 2 (decreasing productivity after the changepoint).

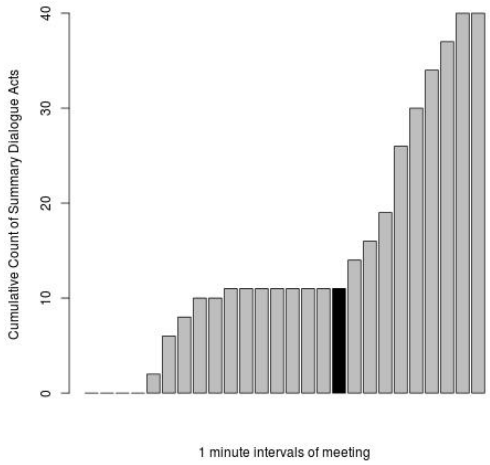

Fig. 4. Example of Meeting Changepoint (Increasing Productivity)

Figure 4 shows an example of a changepoint in a Case 1 meeting, where the x-axis is divided into 1-minute intervals and the y-axis shows the cumulative count of extractive dialogue acts seen up to that interval of the meeting. In this particular meeting, the first 15 minutes featured only about 12 summary-worthy dialogue acts and there was a period of 8 minutes where nothing summary-worthy was said. The meeting then became much more productive after the changepoint, with about 30 extracted dialogue acts in the final 9 minutes.

In contrast, Figure 5 shows a Case 2 meeting changepoint where there is a dramatic decrease in productivity near the end of the meeting. There are no summary-worthy dialogue acts for nearly 20 minutes before the meeting is finally wrapped up.

Once the changepoint for each meeting has been determined, we classify meetings into Case 1 or Case 2 based on whether the meeting portion before the changepoint has a higher or lower number of extracted dialogue acts per minute than the portion after the changepoint.

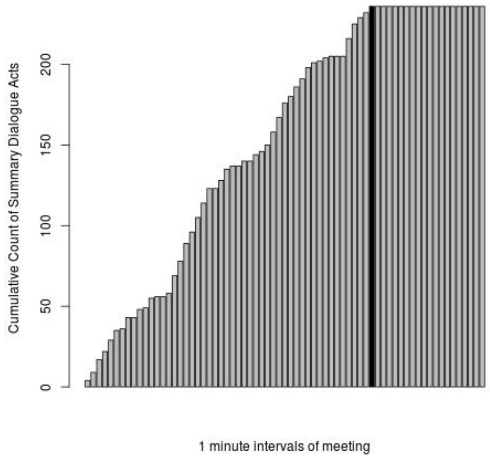

Fig. 5. Example of Meeting Changepoint (Decreasing Productivity)

For meetings that fall into Case 1 (increasing productivity), the changepoint tends to fall relatively early in the meeting. Specifically, the less productive early part of the meeting has about 270 dialogue acts on average, while the more productive later portion of the meeting has more than 700 dialogue acts on average. In contrast, for meetings that fall into Case 2 (decreasing productivity), the changepoint is closer to the middle of the meeting on average. Specifically, the more productive early portion of the meeting has about 588 dialogue acts on average, while the less productive later portion has 557 on average.

7.2 Classification Results

Table 1 shows the precision, recall and F-scores for Cases 1 and 2. The most interesting result here is that classification is easier for Case 1 than Case 2. In other words, when a meeting has a changepoint after which it becomes more productive, it is easier to discriminate the two sub-portions of the meeting in terms of linguistic and structural features, relative to the alternative case where the changepoint leads to the meeting becoming less productive.

Table 2 shows the deviance scores for individual predictor models (each trained using a single feature) as well as the combined predictor model, compared with the null (intercept-only) baselines. For Case 1, we see that the best single predictor is the word entropy of the meeting portion, and that the number of word types in the meeting portion and the tf.idf average are also very useful. The combined model for Case 1 shows a marked decrease in deviance from the null baseline, from 206.6 to 135.4, much more than would be expected by adding random noise predictors.

Productivity Case	Precision	Recall	F-Score
Case 1: Increasing	0.73	0.77	0.75
Case 2: Decreasing	0.60	0.59	0.60

Table 1. Prediction Results

For Case 2, none of the individual predictor models are particularly effective on their own, though conversation cohesion and number of short dialogue acts are the most useful features. Despite the individual predictor models not being particularly effective, the combined model for Case 2 shows a substantial decrease in deviance over the null baseline, from 325.9 to 287.4.

Feature	Case 1	Case 2
null (intercept)	206.6	325.9
tfidfSum	171.2	325.9
tfidfAve	161.4	326.0
conCoh	205.4	**315.6**
DALength	204.5	325.2
shortDAs	207.7	**315.2**
longDAs	203.3	326.1
countDA	167.9	326.8
wTypes	162.4	325.5
docEnt	**156.9**	327.1
domSpeak	180.6	326.9
speakEnt	199.1	327.2
speakEntF100	207.9	326.8
speakEntL100	207.9	326.4
fPauses	207.4	327.0
posWords	193.8	327.1
negWords	207.1	326.8
COMBINED-FEAS	**135.4**	**287.4**

Table 2. Deviance Using Single and Combined Predictors

Our final presented results are the parameter (coefficient) estimates of the logistic regression model, shown in Table 3 for both Cases 1 and 2. These are the parameters of the individual predictor models, each trained using a single feature. We are more interested here in the sign of the parameter than the magnitude of the parameter (though we note that all features were normalized to fall within the 0-1 range). It is also interesting to note cases where the sign of the parameter is flipped between Case 1 and Case 2.

Feature	θ (Case 1)	θ (Case 2)
tfidfSum	3.74	0.43
tfidfAve	5.53	0.64
conCoh	-7.76	21.28
DALength	1.92	1.20
shortDAs	-0.93	-6.48
longDAs	5.75	2.26
countDA	3.48	0.19
wTypes	4.53	0.59
docEnt	-24.16	0.59
domSpeak	-8.29	-0.66
speakEnt	-10.26	-0.20
speakEntF100	0.02	-1.65
speakEntL100	-0.41	-1.88
fPauses	5.79	2.75
posWords	63.569	-2.14
negWords	10.873	6.99

Table 3. Single Predictor Parameter Estimates

As mentioned above, the most useful feature for discriminating the more-productive portion from the less-productive portion in Case 1 was the document entropy, and the θ coefficient for word entropy is negative, meaning that an increase in entropy is associated with a decrease in the log-odds of productivity. This is flipped in Case 2, where θ is positive for the entropy feature. Conversation cohesion was the most useful feature for classification within Case 2 meetings, and its sign is positive for Case 2, meaning that the more-productive portions tend to have higher conversational cohesion. This is flipped in Case 1, there θ is negative for the conversation cohesion feature.

8 Future Work

For future work, we plan to adopt methods for detecting multiple changepoints [19]. Some meetings may, for example, have an unproductive portion at the beginning and another unproductive portion at the end, with a productive section in the middle. For example, the meeting shown in Figure 6 seems to have at least two changepoints, if not more, as evidenced by the multiple plateaus in the cumulative plot. Our current model cannot handle multiple shifts in productivity. On the other hand, this type of plot was a rarity among the meetings; the assumption that a meeting would have zero or one changepoints was correct in the majority of cases, based upon inspection of these visualizations.

Ongoing work would also be aided by the creation of gold-standard annotations for productivity in meetings. In this work we have simply exploited existing

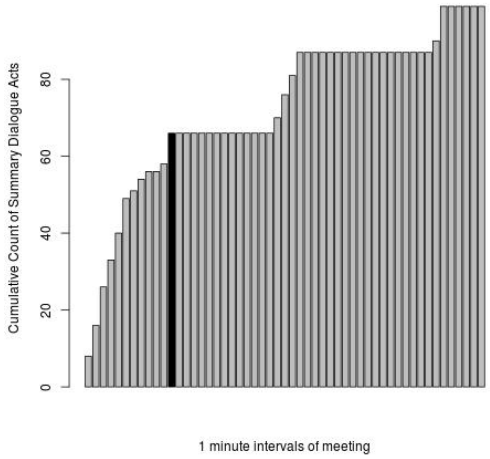

Fig. 6. Potentially Multiple Changepoints

resources [13] that act as a proxy for measuring productivity. Some of these are reliable indicators of productivity or metrics of whether participants are "on-task": for example, meeting dialogue acts linked to the *decision* or *action item* portions of the abstractive summary are by definition on-task as pertains to the goals of the group. However, it would be useful to have annotations indicating the degree to which each dialogue act is productive or contributes to the stated purpose of the meeting.

9 Conclusion

In this work we have analyzed how productivity shifts within meetings, using extractive summarization as a proxy task. We first carried out correlation and regression analyses to investigate which linguistic and structural features relate to productivity. We then developed a novel application of Bayesian changepoint analysis to determine where in the meeting productivity shifts, subsequently learning the linguistic and structural features that discriminate the more-productive portion of the meeting from the less-productive portion.

References

1. Galley, M.: A skip-chain conditional random field for ranking meeting utterances by importance. In: Proc. of EMNLP 2006, Sydney, Australia. (2006) 364–372
2. 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)
3. Murray, G., Carenini, G., Ng, R.: The impact of asr on abstractive vs. extractive meeting summaries. In: Proc. of Interspeech 2010, Tokyo, Japan. (2010) 1688–1691
4. Murray, G., Carenini, G., Ng, R.: Generating and validating abstracts of meeting conversations: a user study. In: Proc. of INLG 2010, Dublin, Ireland. (2010) 105–113
5. Mehdad, Y., Carenini, G., Tompa, F., Ng, R.: Abstractive meeting summarization with entailment and fusion. In: Proc. of ENLG 2013, Sofia, Bulgaria. (2013) 136–146
6. 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
7. Wang, L., Cardie, C.: Domain-independent abstract generation for focused meeting summarization. In: Proc. of ACL 2013, Sofia, Bulgaria. (2013) 1395–1405
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. 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)
11. 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)
12. Murray, G.: Learning how productive and unproductive meetings differ. In: Advances in Artificial Intelligence. Springer (2014) 191–202
13. Murray, G.: Resources for analyzing productivity in group interactions. In: Proc. of LREC 2014 Workshop on Multi-Modal Corpora, Reykjavik, Iceland. (2014) 39–42
14. Jarrett, R.: A Note on the Intervals Between Coal-Mining Disasters. Biometrika **66**(1) (1979) 191–193
15. Gill, J.: Bayesian Methods: A Social and Behavioral Sciences Approach. 2nd edn. Chapman & Hall, London, GB (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. Fearnhead, P.: Exact and efficient bayesian inference for multiple changepoint problems. Statistics and computing **16**(2) (2006) 203–213