Implicit Proposal Filtering in Multi-Party Consensus-Building Conversations

Yasuhiro Katagiri
Future University – Hakodate
katagiri@fun.ac.jp

Yosuke Matsusaka
National Institute of Advanced Industrial Science and Technology
yosuke.matsusaka@aist.go.jp

Yasuharu Den
Chiba University
den@cogsci.l.chiba-u.ac.jp

Mika Enomoto
Tokyo University of Technology
menomoto@media.teu.ac.jp

Masato Ishizaki
The University of Tokyo
ishizaki@iii.u-tokyo.ac.jp

Katsuya Takanashi
Kyoto University
takanasi@ar.media.kyoto-u.ac.jp

Abstract

An attempt was made to statistically estimate proposals which survived the discussion to be incorporated in the final agreement in an instance of a Japanese design conversation. Low level speech and vision features of hearer behaviors corresponding to aiduti, noddings and gaze were found to be a positive predictor of survival. The result suggests that non-linguistic hearer responses work as implicit proposal filters in consensus building, and could provide promising candidate features for the purpose of recognition and summarization of meeting events.

1 Introduction

Non-verbal signals, such as gaze, head nods, facial expressions and bodily gestures, play significant roles in the conversation organization functions. Several projects have been collecting multimodal conversation data (Carletta et al., 2006) for multi-party dialogues in order to develop techniques for meeting event recognitions from non-verbal as well as verbal signals. We investigate, in this paper, hearer response functions in multi-party consensus-building conversations. We focus particularly on the evaluative aspect of verbal and non-verbal hearer responses. During the course of a consensus-building discussion meeting, a series of proposals are put on the table, examined, evaluated and accepted or rejected. The examinations of proposals can take the form of explicit verbal exchanges, but they can also be implicit through accumulations of hearer responses. Hearers would express, mostly unconsciously for non-verbal signals, their interest and positive appraisals toward a proposal when it is introduced and is being discussed, and that these hearer responses would collectively contribute to the determination of final consensus making. The question we address is whether and in what degree it is possible and effective to filter proposals and estimate agreement by using verbal and non-verbal hearer responses in consensus-building discussion meetings.

2 Multi-Party Design Conversation Data

2.1 Data collection

We chose multi-party design conversations for the domain of our investigation. Different from a fixed problem solving task with a 'correct' solution, participants are given partially specified design goals and engage in a discussion to come up with an agreement on the final design plan. The condition of our data collection was as follows:

Number of participants: six for each session
Arrangement: face-to-face conversation
Task: Proposal for a new mobile phone business
Role: No pre-determined role was imposed

A compact meeting archiver equipment, AIST-MARC (Asano and Ogata, 2006), which can capture panoramic video and speaker-separated speech streams, was used to record conversations (Fig. 1). The data we examined consist of one 30 minutes conversation conducted by 5 males and 1 female. Even though we did not assign any roles, a chairperson and a clerk were spontaneously elected by the participants at the beginning of the session.

Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pages 100–103,
Columbus, June 2008. ©2008 Association for Computational Linguistics

Figure 1: AIST-MARC and a recording scene

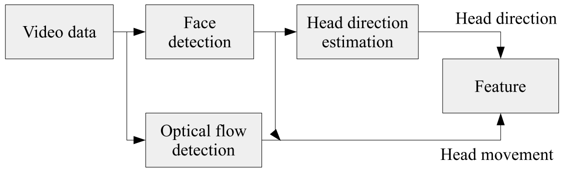

Figure 2: Image processing algorithm

2.2 Data Annotation

2.2.1 Clause units

In order to provide a clause level segmentation of a multi-channel speech stream, we extended the notion of 'clause units (CUs)', originally developed for analyzing spoken monologues in the Corpus of Spontaneous Japanese (Takanashi et al., 2003), to include reactive tokens (Clancy et al., 1996) and other responses in spoken conversations. Two of the authors who worked on the Corpus of Spontaneous Japanese independently worked on the data and resolved the differences, which created 1403 CUs consisting of 469 complete utterances, 857 reactive tokens, and 77 incomplete or fragmental utterances.

2.2.2 Proposal units

We developed a simple classification scheme of discourse segments for multi-party consensus building conversations based on the idea of 'interaction process analysis' (Bales, 1950).

Proposal: Presentation of new ideas and their evaluation. Substructure are often realized through elaboration and clarification.

Summary: Sum up multiple proposals possibly with their assessment

Orientation: Lay out a topic to be discussed and signal a transition of conversation phases, initiated mostly by the facilitator of the discussion

Miscellaneous: Other categories including opening and closing segments

The connectivity between clause units, the content of the discussion, interactional roles, relationship with adjacent segments and discourse markers were considered in the identification of proposal units. Two of the authors, one worked on the Corpus of Spontaneous Japanese and the other worked for the project of standardization of discourse tagging, independently worked on the data and resolved the differences, which resulted in 19 proposals, 8 summaries, 19 orientations and 2 miscellaneouses.

2.3 Core clause units and survived proposal units

Core clause units (CUs) were selected, out of all the clause units, based on whether the CUs have substantial content as a proposal. A CU was judged as a core CU, when the annotator would find it appropriate to express, upon hearing the CU, either an approval or a disapproval to its content if she were in the position of a participant of the conversation. Three of the authors worked on the text data excluding the reactive tokens, and the final selection was settled by majority decision. 35 core CUs were selected from 235 CUs in the total of 19 proposal PUs. Cohen's kappa agreement rate was 0.894.

Survived proposal units (PUs) were similarly selected, out of all the proposal units, based on whether the PUs were incorporated in the final agreement among all the participants. 9 survived PUs were selected from 19 proposal PUs.

3 Feature Extraction of Hearer's Behavior

For each clause unit (CU), verbal and non-verbal features concerning hearer's behavior were extracted from the audio and the video data.

3.1 Non-Verbal Features

We focused on nodding and gaze, which were approximated by vertical and horizontal head movements of participants.

An image processing algorithm (Figure 2) was applied to estimate head directions and motions (Matsusaka, 2005). Figure 3 shows a sample scene and the results of applying head direction estimation algorithm.

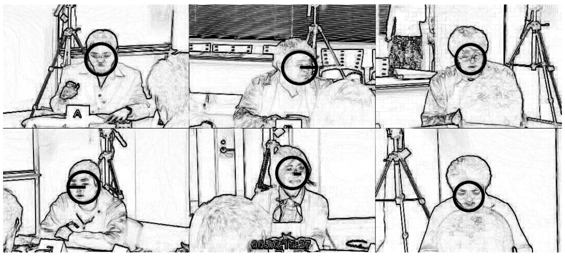

Figure 3: Sample scene with image processing results. The circles represent detected face areas, and the lines in the circles represent head directions.

For each CU, the vertical and horizontal components of head movements of 5 hearers were calculated for two regions, the region inside the CU and the 1-sec region immediately after the CU. For each of the two regions, the mean and the peak values and the relative location, in the region, of the peak were computed. These 12 non-verbal features were used for the statistical modeling.

3.2 Verbal Features

Verbal features were extracted from the audio data. For each CU, power values of 5 hearers were extracted for two regions, 'within' and 'after' CU, and for each of the two regions, the mean and the peak values and the relative location, in the region, of the peak were computed. In addition to these verbal features, we also used aiduti features of reactive tokens (RTs). The percentage of the total duration of RTs, the total number of RTs, and the number of participants who produced an RT were computed in 'within' and 'after' regions for each of the CUs. A total of 12 CU verbal features were used for the statistical modeling.

4 Experiments

4.1 Overview of the Algorithm

Statistical modeling was employed to see if it is possible to identify the proposal units (PUs) that are survived in the participants' final consensus. To this end, we, first, find the dominant clause unit (CU) in each PU, and, then, based on the verbal and non-verbal features of these CUs, we classify PUs into 'survived' and 'non-survived.'

Table 1: The optimal model for finding core-CUs

	Estimate
(Intercept)	-1.72
within/speech power/mean	-11.54
after/vertical motion/peak loc.	-4.25
after/speech power/mean	3.91
after/aiduti/percent	3.02

Table 2: Confusion matrix of core-CU prediction experiment (precision = 0.50, recall = 0.086)

	Predicted	
Observed	Non-core	Core
Non-core	431	3
Core	32	3

4.2 Finding Dominant CUs

A logistic regression model was used to model the coreness of CUs. A total of 24 verbal and non-verbal features were used as explanatory variables. Since the number of non-core CUs was much larger than that of core CUs, down-sampling of negative instances was performed. To obtain a reliable estimation, a sort of Monte Carlo simulation was adopted.

A model selection by using AIC was applied for the 35 core CUs and another 35 non-core CUs that were re-sampled from among the set of 434 complete and non-core CUs. This process was repeated 100 times, and the features frequently selected in this simulation were used to construct the optimal model. Table 1 shows the estimated coefficient for the optimal model, and Table 2 shows the accuracy based on a leave-1-out cross validation. The dominant CU in each PU was identified as the CU

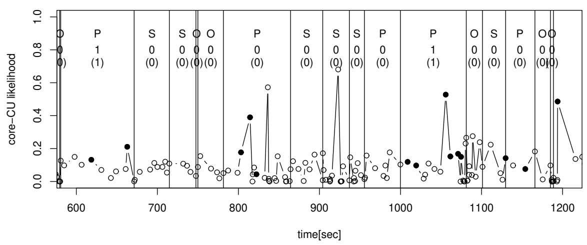

Figure 4: The predicted coreness of CUs. Dominant CUs were defined to be CUs with the highest coreness in each of the PUs. Black and white dots are CUs labeled as core and non-core.

Table 3: The optimal model for finding survived-PUs

	Estimate
within/vertical motion/peak val.	3.96
within/speech power/mean	−27.76
after/speech power/peak val.	1.49

Table 4: Result of the survived-PU prediction (precision = 0.83, recall = 0.44)

	Predicted	
Observed	Non-survived	Survived
Non-survived	37	1
Survived	4	5

with the highest predicted value in that PU. Figure 4 shows the predicted values for coreness.

4.3 Finding Survived PUs

The verbal and non-verbal features of the dominant CUs of each of the PUs were used for the modeling of the survived-PU prediction. Discriminant analysis was utilized and a model selection was applied for the 47 PUs. Table 3 shows the estimated coefficient for the optimal model, and Table 4 shows the accuracy based on a leave-1-out cross validation.

5 Discussions

The results of our estimation experiments indicate that the final agreement outcome of the discussion can be approximately estimated at the proposal level. Though it may not be easy to identify actual utterances contributing to the agreement (core-CUs), the dominant CUs in PUs were found to be effective in the identification of survived-PUs. The prediction accuracy of survived-PUs was about 89%, with the chance level of 69%, whereas that of core-CUs was about 92%, with the chance level of 86%.

In terms of hearer response features, intensity of verbal responses (*within/speech power/mean*, *after/speech power/mean*), and immediate nodding responses (*after/vertical motion/peak loc.*) were the most common contributing features in core-CU estimation. In contrast, occurrence of a strong aiduti immediately after, rather than within, the core-CU (*after/speech power/peak val.*), and a strong nodding within the core-CU (*within/vertical motion/peak val.*) appear to be signaling support from hearers to the proposal. It should be noted that identification of target hearer behaviors must be validated against manual annotations before these generalizations are established. Nevertheless, the results are mostly coherent with our intuitions on the workings of hearer responses in conversations.

6 Conclusions

We have shown that approximate identification of the proposal units incorporated into the final agreement can be obtained through the use of statistical pattern recognition techniques on low level speech and vision features of hearer behaviors. The result provides a support for the idea that hearer responses convey information on hearers' affective and evaluative attitudes toward conversation topics, which effectively functions as implicit filters for the proposals in the consensus building process.

Acknowledgments

The work reported in this paper was supported by Japan Society for the Promotion of Science Grants-in-aid for Scientific Research (B) 18300052.

References

F. Asano and J. Ogata. 2006. Detection and separation of speech events in meeting recordings. In *Proc. Interspeech*, pages 2586–2589.

R. F. Bales. 1950. A set of categories for the analysis of small group interaction. *American Sociological Review*, 15:257–263.

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. 2006. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39.

P. M. Clancy, S. A. Thompson, R. Suzuki, and H. Tao. 1996. The conversational use of reactive tokens in English, Japanese and Mandarin. *Journal of Pragmatics*, 26:355–387.

Y. Matsusaka. 2005. Recognition of 3 party conversation using prosody and gaze. In *Proc. Interspeech*, pages 1205–1208.

K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara. 2003. Identification of 'sentence' in. spontaneous Japanese: detection and modification of clause boundaries. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 183–186.