Conversational Topic Modelling in First Encounter Dialogues

Trung Ngo Trong and Kristiina Jokinen[1]

University of Helsinki, Finland
University of Tartu, Estonia

trung.ngotrong@helsinki.fi kristiina.jokinen@helsinki.fi

Abstract

This paper explores the possibility of modelling conversational topics for multi-modal interactive dialogues. However, representing the conversational topics is an extremely difficult task due to its dynamic characteristic, the issue involves aligning a vast amount of dialogue responses to unlimited combination of words to represent a certain topic. We analyse the diversity of these topic representation in the first encounter situation. The experiments are conducted using Finnish and Estonian corpora to acquire a robust perspective for the communicative behaviours. We further improve the topic understanding using multi-modal features, and our experimental findings are corroborated by outperforming single modality approach in topic recognition task.

Index Terms: topic model, conversation management, first-encounter dialogues, deep learning

1. Introduction

An important feature for automatic interactive systems is to manage dialogues in a natural manner, and for this the ability to handle smooth topic shifts in different situations and different contexts, i.e. be able to provide relevant continuation in a given dialogue state, is important. The actual speakers rely on both visual and auditory information from their partners to infer the context of discussion, and an intelligent system thus also need to acquire deeper understanding of the dialogue context and track the conversational states from multiple perspective. Many relevant rule-based and statistical mechanisms have been proposed for dialogue management, but an exploration of the topic and its correlation with multi-modal features remains an open and challenging issue.

Conventional topic modelling and text classification have mainly focused on static documents, i.e. documents collected from archives, journals, logs, on-line chats and so on, in a single modality environment (Alvarez-Melis and Saveski, 2016; Blei, 2012). Text is usually formalized and edited to express a clear and coherent topic, the sentences are long and elaborated, and enhanced with details. Talking, on the other hand, is spontaneous and depending on the speaker, it encapsulates a high level of diversity and ambiguity in a continuous speech. Topic modelling for task-based dialogues has been studied e.g. by Jokinen et al. (1998), while recently Nguyen et al. (2014) and Yeh et al. (2014) studied topics in spoken dialogues using statistical models. However, these studies do not concern multi-modal aspects of the dialogues.

In the context of human-robot interaction, Jokinen and Wilcock (2014) describe a model to enable human-robot interaction based on Wikipedia information, while Bohus and Horvitz (2009) demonstrated the use of multi-modal signalling in multi-party conversations where the participants enter and leave the interactive situation freely (the interaction is with animated agent, not with a robot agent). Brethes et al. (2004) propose a method to extract information from visual modality that supports richer human-robot interaction. In these contexts, topic models are not taken into account explicitly.

In this paper, we study topic modelling in a special social context, namely in first encounter dialogues, and examine if the participants' gesturing and movement correlates with topic changes. In particular, we tackle the following practical issues for smooth dialogue management:

- Investigating topic flows in first-encounter dialogues,
- Improving topic recognition using multi-modal features.

2. First-encounter Corpora

We use two multi-modal first encounter corpora: the Finnish corpus collected in the Nordic NOMCO project (Navarretta et al. 2012) and the Estonian corpus collected in the MINT project (Jokinen and Tenjes, 2012). The first encounter dialogues consists of interactions between two participants who do not know each other in advance. They are engaged in a chatting interaction, with no other particular task but to get to know each other in a short interaction. The video recordings consist of the two standing individuals recorded separately as well as jointly in a centre view.

The Estonian first encounter dataset consists of a total of 23 encounters, and each encounter is about 8 minutes long. The participants (12 male and 11 female) are native speakers of Estonian, and they are students or university employees with the age ranging between 21 and 61 years. The corpus contains have 8 female-female encounters, 7 female-male encounters, and 8 male-male encounters.

Finnish dataset consists of 16 conversations, with the average length of the conversations is 6 minutes 25 seconds (the shortest conversation is 3 minutes 49 seconds and the longest 8 minutes 2 seconds). There are 14 participants, 4 males and 10 females, all native speakers of Finnish, and the conversation pairs are 2 male-male conversations, 6 male-female conversations, and 8 female-female conversations.

3. Topic changing

We first investigate how conversational topics evolve in the two datasets, given that the interactions concern the same communicative activity. We segmented each dialogue based on the manual independent topic annotations of the corpora. Table 1 gives an example segmentation of one of the Estonian dialogues with details of the annotated dialogue topics. Start and end times are given in seconds and measured from the start of the first utterance in the beginning of the dialogue.

Start	End	Segmentation into topics
13	18	greetings, introductions
18	27	occupations, studying language technology
27	64	language technology at the university
64	200	specializing fields in LT
200	278	spoken dialogue systems
278	328	morphological analysis and synthesis, topics in LT

Table 1 Summarization of dialogue topics in C-16-MM-15-16.

4. Topic clustering

The pre-processing of the dialogue transcriptions included the removal of stop-words, punctuations, out-of-dictionary words, and stemming. As we wanted the most compact representation for the topics, we chose term frequency inverse document frequency (*tf-idf*) to extract the vector representation for each word. Each of the topic is then represented as a collection of ranked words from the extracted vocabulary. This task can be interpreted as clustering of the word vectors into N most distinguishable clusters, where N is a heuristic number for predefined number of topics. We compare the performance of traditional K-mean clustering and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) for this task.

Figure 1 shows the difference between K-mean and LDA clustering. K-mean produces more overlapping topics in both semantic and temporal aspects (e.g. Topic 3 and 5). The topics detected by LDA are often discussed for a specific period of time during the conversation, and each topic has a specific concentrated region in the dialogue (e.g. topic 1 is often discussed at the end of the conversation and topic 4 after greetings and introductions).

Comparison of Figure 1 and Figure 2 shows that the discussion on some topics is often narrow in the Finnish corpus and topic 2 is overwhelmed. Furthermore, time distribution is different. E.g. topic 6 deals with introductions and recent activities, but these topics often occur earlier in the Estonian corpus than in the Finnish corpus.

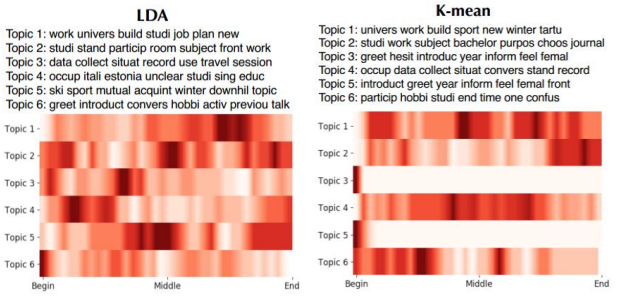

Figure 1 Comparison of LDA and K-mean for topic extraction on the Estonian corpus. The image illustrates the time spectrum of the topics in the dialogues, with each topic represented by the top 7 terms.

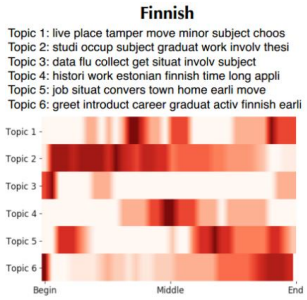

Figure 2 LDA topic extraction from Finnish first-encounter corpus.

5. Multimodal Topic modelling

Our dialogues have been recorded in both video and audio, but at different sampling rate. The videos are recorded at 25 frames per second, while the wave file is captured at the rate of 44100 Hz. It is important to scale the two types of signals into a universal time scale for synchronization. As a result, we divide each conversation into discrete units u = 0.01(second). This number was carefully selected to be sufficiently small so that all the video actions and speech events would be longer. In order to extract visual features, we used bounding boxes which keep tracking the marginal movements of individual speakers, see more in Vels and Jokinen (2015).

Figure 3 (next page) illustrates the relation between multi-modal and conversational topic. Strong movement is seen at the beginning of each dialogue, since most of the participants start with a handshake. The movements are also more intense in the area where the speakers change topics frequently as indicated by the red ellipses. It can also be noticed that the participants often laugh when talking about topic 3 (data collection and present situation), but the events rarely happen for topic 6 (greetings, today's activities, previous interview). Correlation between the topic and multi-modal features shows that the differences are significant to distinguish certain topics from the others. E.g. the participants tend to move a lot at the beginning of the conversations for greetings and handshake, and as a result, topic 6 correlates strongly with movements. Moreover, many speakers also move when they discuss topics 1 and 2, here are samples from two given topics:

• Topic 1: occupations, teaching music to children, working on Saturdays, a trip to Copenhagen, equipment in the university classrooms, girlfriends, relationships, a movie theatre Athena.
• Topic 2: studying, how long it takes to graduate, studying politics and government, how to talk and stand in the recording situations, confusion about participants of the study.

6. Discussion and Conclusions

The paper presented preliminary work on topic modelling for the first encounter dialogues and included multimodal features (head, body, leg movement) to improve topic recognition. The findings are visualised in Figures 1-3, and the full paper will discuss them in more detail.

Acknowledgements

The authors would like to thank Graham Wilcock for extracting movement data from the videos and for useful discussions related to multimodal features in general, and Katri Hiovain for topic annotations.

The second author would also like to thank the NEDO project at AIRC AIST Tokyo Waterfront Japan, for supporting her research.

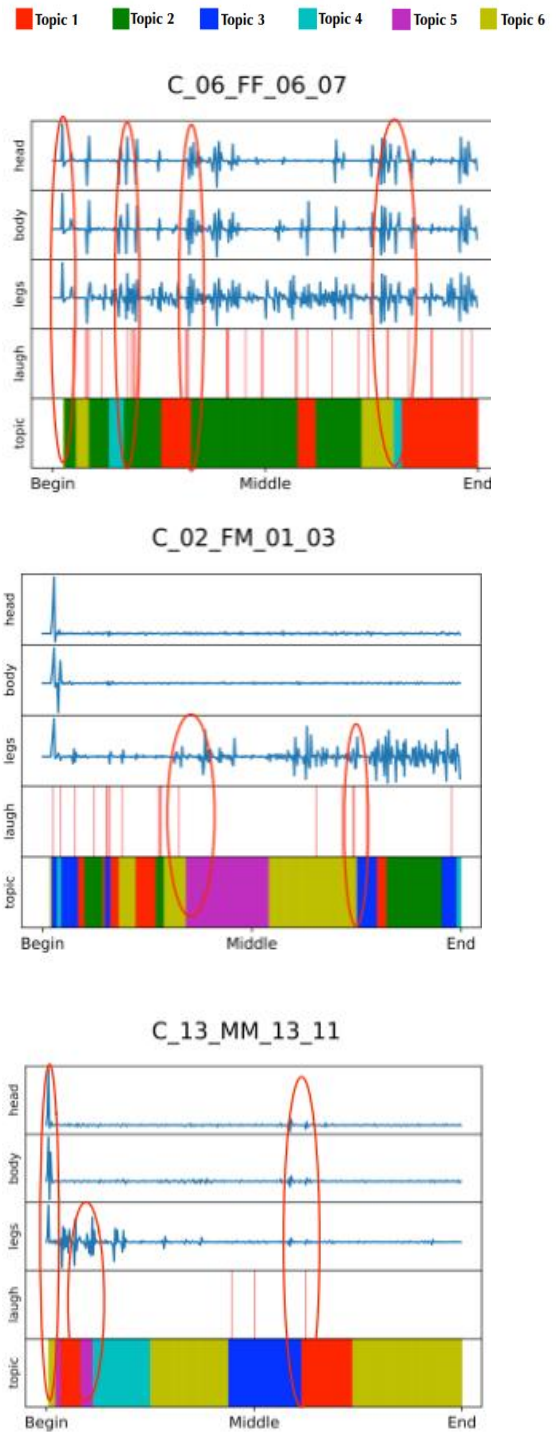

Figure 3 Time synchronized multi-modal features with topics, the fluctuation of the top three figure represents the intensity of both speakers movement in given body part. Three conversation between FF - two females, FM - female and male, MM - two male are selected in order from left to right. The topics are the same as from LDA model in Figure 1.

References

[1] David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. In ICWSM.

[2] David M. Blei. 2012. Probabilistic topic models. Commun. ACM 55(4):77–84. https://doi.org/10.1145/2133806.2133826

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3:993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[4] Dan Bohus and Eric Horvitz. 2009. Models for multiparty engagement in open-world dialog. Proceedings of SIGDIAL '09, pp. 225–234. http://dl.acm.org/citation.cfm?id=1708376.1708409

[5] L. Brethes, P. Menezes, F. Lerasle, and J. Hayet. 2004. Face tracking and hand gesture recognition for human-robot interaction. Proceedings of ICRA '04. Vol. 2, pp. 1901–1906. https://doi.org/10.1109/ROBOT.2004.1308101

[6] Kristiina Jokinen, Hideki Tanaka, and Akio Yokoo. 1998. Context management with topics for spoken dialogue systems. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, pp. 631–637.

[7] Kristiina Jokinen and Silvi Tenjes. 2012. Investigating engagement intercultural and technological aspects of the collection, analysis, and use of Estonian multiparty conversational video data. Proceedings of LREC-2012.

[8] Kristiina Jokinen and Graham Wilcock. 2014. Multimodal Open-Domain Conversations with the Nao Robot, Springer New York, pp. 213–224.

[9] Costanza Navarretta, Elisabeth Ahlsen, Jens Allwood, Kristiina Jokinen, and Patrizia Paggio. 2012. Feedback in Nordic First-Encounters: a Comparative Study. LREC, pp. 2494–2499.

[10] Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah A. Cai, Jennifer E. Midberry, and Yuanxin Wang. 2014. Modeling topic control to detect influence in conversations using non-parametric topic models. Mach. Learn. 95(3):381–421. https://doi.org/10.1007/s10994-013-5417-9.

[11] Martin Vels and Kristiina Jokinen 2015. Detecting Body, Head, and Speech in Engagement. Proceedings of ESIVA.

[12] J. F. Yeh, C. H. Lee, Y. S. Tan, and L. C. Yu. 2014. Topic model allocation of conversational dialogue records by Latent Dirichlet Allocation. Proceedings of Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific. pp. 1–4. https://doi.org/10.1109/APSIPA.2014.7041546.