Multimodal dialogue segmentation with gesture post-processing

Kodai Takahashi, Masashi Inoue
Yamagata University
3-16, 4 Jyonan, Yonezawa, Japan
mi@yz.yamagata-u.ac.jp

Abstract

We investigate an automatic dialogue segmentation method using both verbal and non-verbal modalities. Dialogue contents are used for the initial segmentation of dialogue; then, gesture occurrences are used to remove the incorrect segment boundaries. A unique characteristic of our method is to use verbal and non-verbal information separately. We use a three-party dialogue that is rich in gesture as data. The transcription of the dialogue is segmented into topics without prior training by using the TextTiling and U00 algorithm. Some candidates for segment boundaries – where the topic continues – are irrelevant. Those boundaries can be found and removed by locating gestures that stretch over the boundary candidates. This filtering improves the segmentation accuracy of text-only segmentation.

Keywords: Multimodal, Dialogue, Segmentation

1. Introduction

In this paper, we address two issues of automatic multimodal dialogue segmentation. The first issue is the lack of explicitness in the segmentation algorithm. Assuming that multimodality plays an important but complex role in structuring dialogues, different modalities are often exploited simultaneously in topic segmentation. For example, words, prosody, motion, interaction cues, and speaker intention and role were jointly used for segmentation and achieved better results (Hsueh and Moore, 2007). We explicitly divide the segmentation process into speech-content-based topic segmentation and gesture-spanning-based post-processing to avoid blackboxing. The second issue is the lack of copious amounts of training data for the algorithms used for statistical learning. Often, access to large amounts of training data enables models that are learnt by statistical algorithms to outperform the on-the-fly algorithms. However, many multimodal dialogues are unique in themselves and the problem of training and test data mismatch cannot be avoided. To overcome this limitation, the proposed method is entirely unsupervised and there is no necessity to train any model prior to the segmentation of incoming dialogue data. We combined the gesture-based post-processing with two text segmentation algorithms, namely, TextTiling (Hearst, 1997) and U00 (Utiyama and Isahara, 2001), and experimentally compared the segmentation accuracy obtained in both cases. The initial results are promising; the gestural information succeeded in filtering out the irrelevant segmentations and yielded higher precision scores without compromising on recall scores.

2. Background

Multimodality has been of interest to researchers working on dialogue analysis. Various dialogue units have been targeted for automatic identification, especially around local phenomena of dialogue, such as utterances and turns. Multimodal annotation for dialogue acts are bundled into functional segments (Bunt et al., 2012). However, there is no established definition for larger dialogue units. In the field of natural language processing, topic segmentation of text has been developed attempting to achieve better automatic summarization and information retrieval. We can use the concept of topic segmentation in multimodal dialogue as well. However, direct application of existing segmentation algorithms is insufficient because of the irregularities of spoken text, such as fragmentation and lack of semantic information. We have to alleviate the differences between the written text and the spoken words by leveraging multimodality. Although there have been attempts to exploit multimodalities, gestures in particular, in localized phenomena such as sentence unit identification (Chen et al., 2006), turn-taking, anaphora resolution, and discourse segments (Xiong and Quek, 2006), relatively little has been done to exploit them in global dialogue phenomena such as dialogue topics and dialogue flows.

3. Method

3.1. Topic Segmentation Algorithms

We first apply existing text topic segmentation algorithms to the dialogue data. The first algorithm is TextTiling; it finds the gaps in similarities between word distributions in two text windows. Within fixed-size windows sliding over textual documents, words occur differently if the topics of the respective windows are different. The similarity gaps bigger than a threshold indicate the existence of topic boundaries. The second algorithm is U00; it finds the maximum probabilities of segmentation. The probability model assumes that words are generated from a certain segment according to the word occurrence probabilities within the segment. Although these text segmentation algorithms work well on written text, their performances are generally lower on dialogue transcripts. This could be attributed to the conversational texts often being quite sparse and containing many irregular expressions. As

Figure 1: Sample screenshot from dialogue video.

a result, there are many irrelevant topic boundary candidates generated by the algorithms that must be removed. However, it is difficult to eliminate such incorrect boundaries using textual features alone, since the candidates have already been determined on the basis of the textual features. Then, we considered that the process can be facilitated by using multimodal information, especially gesture information.

3.2. Gesture-based Filtering

We prepared multiple rules to filter out irrelevant topic boundary candidates. The basic idea was that when a gesture spans over the topic boundary candidates, it is unlikely that the topic ends there since a single gesture can be associated with multiple utterances or turns but not with multiple dialogue topics. On the basis of this framework, depicted in Figure 2, we removed irrelevant boundary candidates and combined the over-segmented topics into one single topic. There was an investigation on the human recognition of discourse boundaries using lecture data focusing on the presence of particular gestures around the boundaries (Chandlee and Veilleux, 2010). In contrast, we utilize the absence of gestures spanning over the true boundaries.

4. Experiment

4.1. Data

For our experiments, we used NII Grand Challenge dialogue data that consisted of video and audio recording of three-party conversation in Japanese. The data is provided as a sample dialogue for the multimodal corpus that will be distributed by The Informatics Research Data Repository (IRD)[1]. The dialogue contained an explanation of an animation film to a participant who had not viewed the film, by two other participants who had watched it. An overview of the data is listed in Table 1. Utterances and gestures were manually annotated in the ELAN format for all three participants[2]. Utterances were segmented at pauses

Table 1: Data overview

Total length	10 min 17 sec
Number of participants	3
Annotated modalities	Utterance
	Hand gesture
	Head nodding
	Eye gaze

Table 2: Individual and total action statistics from dialogue. S1 represents listener. S2 and S3 represent explainers. Durations are represented in seconds.

Annotation	S1	S2	S3	Total
Number of utterance	80	206	198	484
Mean utterance duration	0.97	1.08	1.07	1.06
Mean silence duration	6.09	1.71	1.98	2.54
Number of morphemes	292	749	704	1745
Number of nouns	71	171	167	409
Number of unknown	11	36	49	96
Number of gestures	6	59	64	129
Mean gesture duration	19.66	7.77	7.80	7.73

longer than 200 ms as defined in Corpus of Spontaneous Japanese (CSJ) provided by the National Institute for Japanese Language and Linguistics (NINJAL)[3]. Transcripts were segmented into words using the Mecab morphological analyser. As a textual feature of the transcript, we used words whose parts of speech were either noun or unknown (most of them are proper nouns that are not registered in the dictionary) after morphological analysis. This concentration is introduced to reduce the adverse effect of data sparseness by focusing on the morphemes that carry meanings. The starting and ending times of the gestures were extracted from the annotations. The information used in the experiments is summarized in Table 2. The number of morphemes were reduced from 1745 to 505 (nouns and unknowns) and they were used as the basis for text segmentation. The dialogue consisted of 11 topics and the goal of automatic segmentation was to identify the boundaries between these topics. The 11 topic segments are listed in Table 4. Although the segment boundaries were assigned by the authors, because the topics basically correspond to the episodes in the animation film, the segments can be identified firmly by using the film structure.

4.2. Experimental Conditions

The process for our method is as follows. First, textual features are extracted from the transcript, and text segmentation algorithms are applied. The output from each segmentation algorithm is considered a list of boundary candidates. Boundary candidates are defined for words or morphemes. The boundary candidates need to be mapped onto the temporal axis in order to be used with gesture information. To perform

[1] http://www.nii.ac.jp/cscenter/idr/en/index.html

[2] http://www.lat-mpi.eu/tools/elan/

[3] http://www.ninjal.ac.jp/corpus_center/csj/misc/preliminary/index_e.html

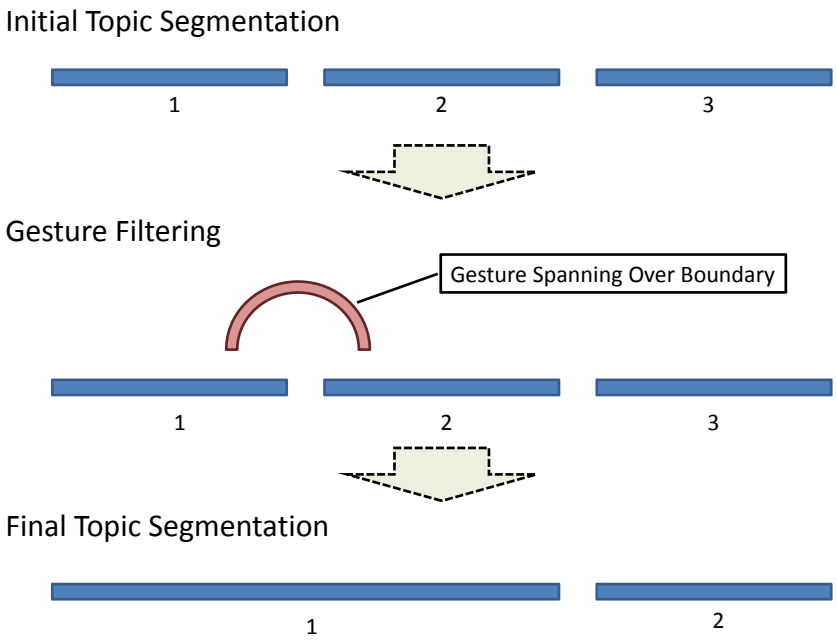

Figure 2: Schematic procedure of post-processing after dialogue segmentation.

such mapping, we identify the utterance that contains the word associated with a boundary candidate. Then, the end time of the utterance is used as the temporal information for the boundary candidate. The temporal boundary candidates are sent to the gesture-based filter as defined in Table 3.

Experimentally, we compared four settings: 1) TextTiling only, 2) TextTiling and gesture combined, 3) U00 only, and 4) U00 and gesture combined. The output of the TextTiling algorithm depends on the window size parameter. We examined window sizes that generate boundary candidates larger than the true boundary numbers.

In evaluating the correspondence between estimated boundaries and human-annotated boundaries, it should be noted that the time information might not match exactly because the boundaries have ranges. Therefore, we relaxed the matching condition: if the boundary candidate falls within the 10-second time range of a boundary (five seconds before and after the human-defined boundary), the boundary candidate is considered successful.

4.3. Experimental Results

When the TextTiling algorithm was used, the effect of gesture-based post processing was mixed as shown in Table 5. For window sizes 11 and 13, the use of gesture post-processing decreased the F-score because of the drop in the recall scores. In the case of window size 12, the F-score increased after post-processing and achieved the highest precision of 0.22 as shown in TT-4 in Table 5. When U00 algorithm was used, gesture-based post processing could improve the segmentation results as shown in Table 6. Although the U00 algorithm comes with model selection functionalities to determine the optimal number of segments, for the purpose of comparison, we specified the number of segment candidates. After applying gesture filters, we obtained better precision rates and lower recall rates except in the case when the number of candidates was 34. Despite the worsened recalls, the overall performance measured in terms of F-scores improved. It should be noted that the precision value could be higher if we expand the current 10-second matching range .

5. Extension

Our method requires utterance transcripts and gesture annotations unless we apply automatic annotation technologies. To test the generalities, we attempted to use a different fully annotated multimodal corpus in a separate domain. We used the Referring Expression (REX) corpus that contains cooperative puzzle dialogues in computer-mediated communication settings(Tokunaga et al., 2012). REX contains verbal transcripts and mouse movements for moving puzzle pieces on a computer screen. In order to use REX with gestures and topic transitions, we considered mouse movements as gestures in face-to-face dialogue and the completion of a puzzle piece as a topic transition. Unfortunately, prior to testing the utility of gestural filtering, we discovered that the base text segmentation algorithms did not work well on the REX corpus. We

Table 3: Relationship between topic boundary candidates and gesture occurrences. The notation follows Allen's definition of temporal relationships and is extended to include ternary relationships(Allen, 1983). Utterances are represented by 'U' characters, gestures are represented by 'G' characters, and boundary candidates are represented by 'B' characters that always come after 'U's.

Relation	Symbol	Inverse	Pictoral Example	Action
U and B before G	<	>	UUUB GGG	Do nothing
U equal G	=	=	UUUB GGG	Do nothing
U meets G	m	mi	UUUB GGG	Do nothing
U overlaps G	o	oi	UUUB GGG	Do nothing
U and B during G	d	di	UUUB GGGGGG	Do nothing
U starts G	s	si	UUUB GGGGG	Do nothing
U finishes G	f	fi	UUUB GGGGG	Do nothing
G bridges Us	b	-	UUUBUUU GGGG	Remove candidate
G within B	w	-	UUUBBB GG	Remove candidate
G enters B	en	-	UUUBB GG	Do nothing
G exits B	en	-	UUUBB GG	Do nothing

Table 4: Segment content and length (sec.).

	Content	Duration
1	Overview of the animated film.	84.936
2	Explanation of the scene where the cat enter the hotel through the front door, but is kicked out.	10.28
3	Explanation of the scene where the cat climbs the water pipe to enter the room through the window, but the elderly lady knocks it out with her umbrella.	23.12
4	Explanation of the scene where the cat enters the water pipe and climbs up, but a bowling ball thrown into the pipe knocks it out.	34.04
5	Explanation of the scene where the cat disguises itself as a street performing monkey and creeps into the hotel, but the elderly lady detects it and hit it with her umbrella.	71.56
6	Explanation of the scene where the cat disguises itself as a bellboy and steals the birdcage, but the elderly lady is hiding in the cage and hits the cat hard.	50.59
7	Explanation of the scene where the cat uses a seesaw to jump to the window, but it fails and is crushed by the weight it used on the seesaw.	55.24
8	Explanation of the scene where the cat uses a rope to swing and jump toward the window, but it crashes into the wall.	34.09
9	Explanation of the scene where the cat walks on the electrical wire to reach the window, but the elderly woman chases it on the wire by the train.	45.21
10	Summarisation of the animated film.	22.11
11	Questions and answers on the content of the film.	186.42

could not observe a change in word usage when working on one puzzle piece, and then changing to another piece. Because our method is based on text segmentation, our method is not applicable when the target dialogues do not have clear topic shifts.

6. Conclusion

In this paper, an automatic multimodal dialogue topic segmentation method was proposed. In the proposed method, gestural post-processing is applied to the outputs of textual topic segmentation. The advantages of

Table 5: Results for post-processing where TextTiling was used as the base algorithm.

Method	Window Size	# of Candidates	Precision	Recall	F-score
TT-1) TextTiling	11	28	0.14	0.40	0.21
TT-2) TextTiling + Gesture	11	11	0.09	0.10	0.10
TT-3) TextTiling	12	23	0.09	0.20	0.12
TT-4) TextTiling + Gesture	12	9	**0.22**	0.20	**0.21**
TT-5) TextTiling	13	13	0.15	0.20	0.17
TT-6) TextTiling + Gesture	13	5	**0.20**	0.10	0.13

Table 6: Results for post-processing where U00 was used as the base algorithm.

Method	# of Candidates	Precision	Recall	F-score
UU-1) U00	34	0.15	0.50	0.23
UU-2) U00 + Gesture	13	0.15	0.20	0.17
UU-3) U00	35	0.14	0.50	0.22
UU-4) U00 + Gesture	13	**0.23**	0.30	**0.26**
UU-5) U00	36	0.14	0.50	0.22
UU-6) U00 + Gesture	14	**0.21**	0.30	**0.25**
UU-7) U00	100	0.05	0.50	0.09
UU-8) U00 + Gesture	40	**0.08**	0.30	**0.12**

our method are that it is explicit, not in a black-box, and does not require the preparation of training data. Our initial experiment suggests that the method could improve the precision scores in segmentation given by the text-only methods moderately affecting the recall scores. Further, we found that the effect of post-processing depended on the base text segmentation algorithm.

A disadvantage of our approach may be the lower accuracy in comparison with supervised approaches. The difference in performance should be examined in dialogue domains for which training data are readily available. In addition, other modalities such as gaze and head-nodding would be utilized in either the pre- or post-processing in our framework.

7. Acknowledgements

This research was partially supported by the Grant-in-Aid for Scientific Research 24500321.

8. References

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November.

Bunt, H., Kipp, M., and Petukhova, V. (2012). Using DiAML and ANVIL for multimodal dialogue annotations. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May.

Chandlee, J. and Veilleux, N. (2010). Gestural cues of discourse segmentation. In *Speech Prosody*, Chicago, IL, May.

Chen, L., Harper, M., and Huang, Z. (2006). Using maximum entropy (ME) model to incorporate gesture cues for SU detection. In *International Conference on Multimodal Interaction*, pages 185–192, Banff, Canada, June.

Hearst, M. A. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, March.

Hsueh, P.-Y. and Moore, J. D. (2007). Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Association of Computational Linguistics*, pages 1016–1023, Prague, Czech Republic, June.

Tokunaga, T., Iida, R., Terai, A., and Kuriyama, N. (2012). The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May.

Utiyama, M. and Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 499–506.

Xiong, Y. and Quek, F. (2006). Hand motion oscillatory gestures and multimodal discourse analysis. *International Journal of Human-computer Interaction*, 21(3):285–312.