The Kyutech corpus and topic segmentation using a combined method

Takashi Yamamura, Kazutaka Shimada and Shintaro Kawahara
Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka Japan
{t_yamamura, shimada}@pluto.ai.kyutech.ac.jp

Abstract

Summarization of multi-party conversation is one of the important tasks in natural language processing. In this paper, we explain a Japanese corpus and a topic segmentation task. To the best of our knowledge, the corpus is the first Japanese corpus annotated for summarization tasks and freely available to anyone. We call it "the Kyutech corpus." The task of the corpus is a decision-making task with four participants and it contains utterances with time information, topic segmentation and reference summaries. As a case study for the corpus, we describe a method combined with LCSeg and TopicTiling for a topic segmentation task. We discuss the effectiveness and the problems of the combined method through the experiment with the Kyutech corpus.

1 Introduction

In collaborative work, people share information, discuss it, and then make decisions through multi-party conversations, such as meetings. Therefore, understanding such conversations and meetings is one of the most important tasks in natural language processing. Conversation summarization is useful to understand the content of conversations for both participants and non-participants. Many researchers have studied meeting and conversation summarization (Banerjee et al., 2015; Mehdad et al., 2014; Oya et al., 2014).

For the summarization tasks, corpora are very important to analyze characteristics of conversations and to construct a method for summary generation. There are some corpora in English, such as the AMI corpus (Carletta, 2007) and the ICSI corpus (Janin et al., 2003). In contrast, there is no corpus for conversation summarization tasks in Japanese. In this study, we construct a Japanese conversation corpus about a decision-making task with four participants. We call it "the Kyutech corpus." To the best of our knowledge, the Kyutech corpus is the first Japanese corpus annotated for summarization tasks and freely available to anyone[1].

The final goal of our study is to generate a summary from a multi-party conversation. Topic segmentation has often been used as the first process in summarization (Banerjee et al., 2015; Oya et al., 2014). In a similar way, we apply topic segmentation to the Kyutech corpus. In this paper, we combine two different text segmentation methods; LCSeg (Galley et al., 2003) and TopicTiling (Riedl and Biemann, 2012). We evaluate the effectiveness of the methods on the Kyutech corpus.

The contributions of this paper are as follows:

- We open the Kyutech corpus, a freely available Japanese conversation corpus for a decision-making task, on the web. This is the first Japanese corpus for summarization.
- As a case study, we examine a combined method based on LCSeg and TopicTiling for topic segmentation with the Kyutech corpus. This is the first step of our conversation summarization.

[1] http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html

Proceedings of the 12th Workshop on Asian Language Resources,
pages 95–104, Osaka, Japan, December 12 2016.

2 Related work

The AMI (Carletta, 2007) and the ICSI (Janin et al., 2003) are very famous meeting corpora and contain numerous annotations, such as dialogue acts and summaries. These corpora are useful and freely available. In addition, they contain a variety of resources, such as speech information in the AMI and ICSI and Powerpoint slides in the AMI corpus. In this paper, we, however, focus on Japanese corpora. Some discussion and conversation corpora in Japanese have been collected on the basis of different task settings; a chat corpus for a detection task of dialogue breakdown (Higashinaka and Funakoshi, 2014) and a multi-modal corpus for three discussion tasks, such as travel planning for foreign friends (Nihei et al., 2014). On the other hand, our task is summarization and our corpus is annotated for the task. The current version contains topic tags of each utterance and reference summaries. In addition, the corpus is freely available to anyone.

For the topic segmentation, some methods have been proposed. The methods were generally based on lexical cohesion for the topic segmentation. TextTiling proposed by (Hearst, 1994) is one of the most famous approaches using a cosine similarity in word vector space. Galley et al. (2003) have proposed a topic segmentation method, LCSeg. It is also a domain-independent discourse segmentation method based on lexical cohesion. It considered the more sophisticated notion of lexical chains as compared with TextTiling. Eisenstein and Barzilay (2008) have proposed an unsupervised approach to topic segmentation based on lexical cohesion modeled by a Bayesian framework. Banerjee et al. (2015) reported that LCSeg tended to outperform the Bayesian segmentation in the summarization. Therefore, we employ LCSeg as a segmentation method. Riedl and Biemann (2012) have proposed a topic segmentation method using the Latent Dirichlet Allocation (LDA) topic model. It was not based on words, but on the topic IDs assigned by the Bayesian Inference method of LDA. Since the topic model alleviated the problem of the sparsity of word vectors, it led to the improvement of the segmentation accuracy. TopicTiling is essentially different from LCSeg because of the use of the topic model. Therefore, we also employ TopicTiling as another method for the topic segmentation. Since the characteristics of the two methods are different, they have a potential to improve the accuracy by a complementary style. Therefore, in this paper, we combine the two methods with a weight factor.

3 Kyutech corpus

In this section, we explain the Kyutech corpus and the annotation for summarization.

3.1 Task

The Kyutech corpus contains multi-party conversations with four participants. The conversations are a decision-making task. The participants pretend managers of a virtual shopping mall in a virtual city, and then determine a new restaurant from three candidates, as an alternative to a closed restaurant. Before the discussion, the participants read a 10-pages document including information about the three candidates, the closed restaurant and the existing restaurants in the mall, the city information, statistics information about the shopping mall, and so on. Figure 1 is a part of the document for the discussion[2].

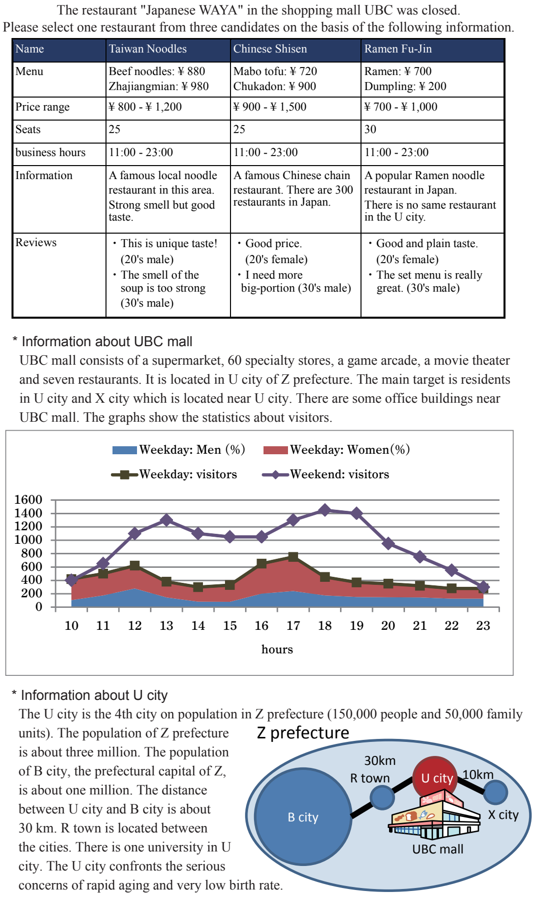
The environment of the discussion is shown in Figure 2. The participants are seated around a 1.8m × 1.8m table in a meeting room. We record the discussion by using a four-direction camera[3] and a video camera. They read the document for 10 minutes, then discuss the candidates for 20 minutes and finally determine one restaurant as a new restaurant opening. We prepared four scenarios with different settings, e.g., different candidates. The participants for each discussion were selected from 20 students consisting of 16 males and 4 females. The current Kyutech corpus consists of nine conversations. After discussion, the participants answer a questionnaire about the satisfaction for the decision, and so on.

3.2 Annotation

We transcribe and annotate the conversations. We annotate topic tags for each utterance and generate summaries for each conversation. The working time for the topic annotation was two hours on average

[2] The original document is written in Japanese because the corpus is Japanese. This is English translation of the document.

[3] KingJim MR360. `http://www.kingjim.co.jp/sp/mr360/`



The restaurant "Japanese WAYA" in the shopping mall UBC was closed.
Please select one restaurant from three candidates on the basis of the following information.

Name	Taiwan Noodles	Chinese Shisen	Ramen Fu-Jin
Menu	Beef noodles: ¥ 880 Zhajiangmian: ¥ 980	Mabo tofu: ¥ 720 Chukadon: ¥ 900	Ramen: ¥ 700 Dumpling: ¥ 200
Price range	¥ 800 - ¥ 1,200	¥ 900 - ¥ 1,500	¥ 700 - ¥ 1,000
Seats	25	25	30
business hours	11:00 - 23:00	11:00 - 23:00	11:00 - 23:00
Information	A famous local noodle restaurant in this area. Strong smell but good taste.	A famous Chinese chain restaurant. There are 300 restaurants in Japan.	A popular Ramen noodle restaurant in Japan. There is no same restaurant in the U city.
Reviews	・This is unique taste! (20's male) ・The smell of the soup is too strong (30's male)	・Good price. (20's female) ・I need more big-portion (30's male)	・Good and plain taste. (20's female) ・The set menu is really great. (30's male)

* Information about UBC mall

UBC mall consists of a supermarket, 60 specialty stores, a game arcade, a movie theater and seven restaurants. It is located in U city of Z prefecture. The main target is residents in U city and X city which is located near U city. There are some office buildings near UBC mall. The graphs show the statistics about visitors.

* Information about U city

The U city is the 4th city on population in Z prefecture (150,000 people and 50,000 family units). The population of Z prefecture is about three million. The population of B city, the prefectural capital of Z, is about one million. The distance between U city and B city is about 30 km. R town is located between the cities. There is one university in U city. The U city confronts the serious concerns of rapid aging and very low birth rate.

Figure 1: A part of a document in the decision-making task.

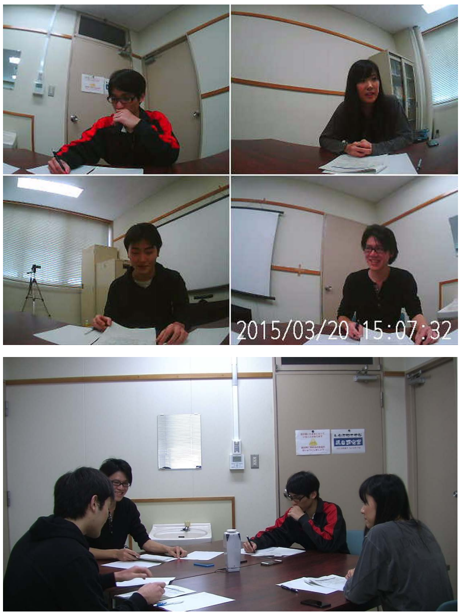


Figure 2: The discussion environment.

Tag	Description
(F) tag	Filler
(D) tag	Falter and Repair
(Q) tag	Question: based on the intonation
(?) tag	Low confidence by inaudibleness
(L) tag	Whispering voice and Monologue
<laugh>	Laughing

Table 1: Tags in transcription.

for one conversation. Besides, the time for the summary generation by an annotator was 30 minutes on average for one conversation. In this sub-section, we explain the way for the corpus construction and report the results.

3.2.1 Transcription

We transcribed the conversations by using ELAN[4]. The transcription rules were based on the construction manual of Corpus of Spontaneous Japanese (CSJ) by (National Institute for Japanese Language and Linguistics, 2006). More properly, we separated utterances by 0.2-sec interval on the basis of the manual and annotated some tags shown in Table 1. As a result, the corpus consists of 4509 utterances.

Each utterance is not always sentence-level because it depends on the 0.2-sec interval rule. Other researchers that want to use this corpus might need sentence-level segmentation for their purpose. Therefore, we added another tags, $+$, $/$ and $*$, to the end of each utterance for sentence-level identification[5]. Here "$+$" denotes that the current utterance links to the next utterance. "$/$" denotes the actual end of a sentence. "$*$" has an intermediate meaning between $+$ and $/$.

[4] https://tla.mpi.nl/tools/tla-tools/elan/

[5] This is just a subjective annotation for other users. Note that we do not use this annotation in the latter part of this paper, namely topic segmentation.

Topic	Description
CandX	Topic about the candidate 1
CandY	Topic about the candidate 2
CandZ	Topic about the candidate 3
CandS	Topic about the candidates
Closed	Topic about the closed restaurant
Exist1	Topic about the existing restaurant 1
Exist2	Topic about the existing restaurant 2
Exist3	Topic about the existing restaurant 3
Exist4	Topic about the existing restaurant 4
Exist5	Topic about the existing restaurant 5
Exist6	Topic about the existing restaurant 6
Exists	Topic about the existing restaurants
ClEx	Topic about the existing restaurants and the closed restaurant
Mall	Topic about the shopping mall
OtherMall	Topic about other shopping malls

Topic	Description
Location	Topic about the positional relation among restaurants
Area	Topic about areas and cities
People	Topic about the target customers
Price	Topic about the price
Menu	Topic about the menu
Atomos	Topic about the atmosphere
Time	Topic about the business hours
Seat	Topic about the number of seats
Sell	Topic about the sales
Access	Topic about the access to the shopping mall
Meeting	Topic about the proceedings and final decision
Chat	Chats that not related to the task
Vague	Others and unknown

Table 2: Topic tags in the Kyutech corpus.

An example

A: ahh, in this condition +
A: which one is suitable (Q) /
C: I think the ramen is better /
B: me too /

In this example, the first and the second utterances by the participant A are connected by the tag +. The process is as follows:

Step1: The worker of the transcription subjectively judges whether the end of each utterance should be + or /.

Step2: After that, we check the worker's results with some conditions. If a condition is satisfied, replace + with /. The following is a condition.

Condition: the next utterance begins with "conjunction", "filter" or "adverb".

Step3: Replace + with ∗ if we subjectively judge that the current utterance links to the next one although the condition in **Step2** is not satisfied.

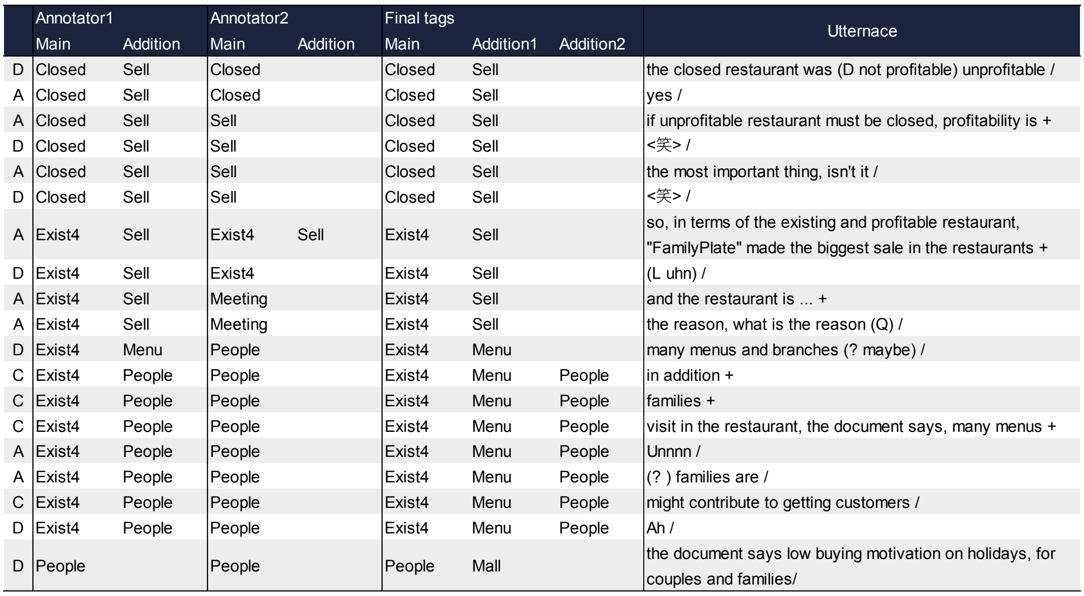
3.2.2 Topic annotation

There are a wide variety of tags that should be annotated to utterances; e.g., communicative functions such as INFORM and REQUEST. Here we focus on a summarization task. In general, topic segmentation has an important role as the first step in the meeting summarization (Banerjee et al., 2015; Oya et al., 2014). Therefore, we manually annotated the topics of each utterance in the Kyutech corpus, as the first annotation[6].

First, we examined the conversations in the Kyutech corpus by four annotators including the authors. We repeated this process twice, and then created a topic tag set consisting of 28 tags. Table 2 shows the tag names and the descriptions.

Next, six annotators who included persons not related to this study annotated topic tags to each utterance, on the basis of the tag set. We applied two annotators into one conversation and the annotation was independently executed. In this process, each annotator annotated at least one tag to one utterance as the main tag of the utterance. In addition, we allowed adding the second-tag if an annotator wanted to add it. The annotators checked the document in Section 3.1 during the annotation process and considered the context in the conversation to select suitable topic tags. Although we allowed creating a new tag if an annotator wanted to create it, no new tags were generated in this process. After the annotation with two

[6] Currently we are also developing the corpus with communicative functions



	Annotator1		Annotator2		Final tags			Utternace
	Main	Addition	Main	Addition	Main	Addition1	Addition2	
D	Closed	Sell	Closed		Closed	Sell		the closed restaurant was (D not profitable) unprofitable /
A	Closed	Sell	Closed		Closed	Sell		yes /
A	Closed	Sell	Sell		Closed	Sell		if unprofitable restaurant must be closed, profitability is +
D	Closed	Sell	Sell		Closed	Sell		<笑> /
A	Closed	Sell	Sell		Closed	Sell		the most important thing, isn't it /
D	Closed	Sell	Sell		Closed	Sell		<笑> /
A	Exist4	Sell	Exist4	Sell	Exist4	Sell		so, in terms of the existing and profitable restaurant, "FamilyPlate" made the biggest sale in the restaurants +
D	Exist4	Sell	Exist4		Exist4	Sell		(L uhn) /
A	Exist4	Sell	Meeting		Exist4	Sell		and the restaurant is ... +
A	Exist4	Sell	Meeting		Exist4	Sell		the reason, what is the reason (Q) /
D	Exist4	Menu	People		Exist4	Menu		many menus and branches (? maybe) /
C	Exist4	People	People		Exist4	Menu	People	in addition +
C	Exist4	People	People		Exist4	Menu	People	families +
C	Exist4	People	People		Exist4	Menu	People	visit in the restaurant, the document says, many menus +
A	Exist4	People	People		Exist4	Menu	People	Unnnn /
A	Exist4	People	People		Exist4	Menu	People	(?) families are /
C	Exist4	People	People		Exist4	Menu	People	might contribute to getting customers /
D	Exist4	People	People		Exist4	Menu	People	Ah /
D	People		People		People	Mall		the document says low buying motivation on holidays, for couples and families/

Figure 3: Topic tags by two annotators and final tags with utterances.

annotators, we computed an agreement between tags of the annotators. The agreement score was based on a partial match scheme (AS_p) as follows:

$$AS_p(A_1, A_2) = \frac{\sum_{i \in U} PM_i(A_1, A_2)}{U_N} \tag{1}$$

where PM_i is the partial match scheme between tag sets of annotators, A_1 and A_2, for an utterance i. In other words, PM_i is true if a tag of an annotator for an utterance is the same as at least one tag of another annotator. For example, $PM_i(A_1, A_2)$ is 1 in the case that $A_1 = \{CandX, People\}$ and $A_2 = \{People\}$ for an utterance i. U is the set of utterances and U_N is the number of utterances, namely 4509. The agreement score AS_p was 0.879.

After that, we checked the tags of two annotators in each conversation. Here we extended the number of tags for one utterance; 2 to 3, namely one main tag and two additional tags. We discussed each tag from annotators, and then determined the final tags of each utterance. After the discussion and the determination of the final tags[7], we also computed an agreement score of them. Here the agreement score was also based on a partial match scheme between the final tag that the authors created (F) and the tag set from two annotators (A_{all}). For example, assume $F = \{People\}$, $A_1 = \{People, Mall\}$ and $A_2 = \{Mall, Menu\}$. Here A_{all} is $\{People, Mall, Menu\}$ and A_{all} contains $F = \{People\}$. Therefore, $PM_i(F, A_{all})$ in this situation is 1. The partial agreement score between the final tags and the tags by two annotators, namely $AS_p(F, A_{all})$, was 0.965. Thus, we obtained a corpus with the high agreement topic tag set. Figure 3 shows an example of the annotation result. In the Kyutech corpus, assuming that the main tag sequence is one topic, one topic sequence usually consists of approximately 10 utterances.

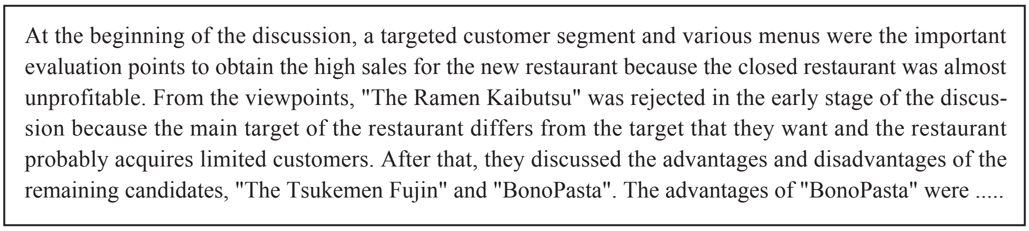
3.2.3 Reference summary

Next, each annotator generated a summary of the conversation. The size of a summary is from 250 characters to 500 characters[8]. The summary generation complied with the guideline of abstractive hand summaries of the AMI corpus[9]. Based on the guideline, the generation carried out after the process in

[7] The working time for the final tag determination was approximately two hours for each conversation.

[8] The number of words was approximately 150 content words on average. The number of unique words was 80 words on average.

[9] `http://groups.inf.ed.ac.uk/ami/corpus/guidelines.shtml`



At the beginning of the discussion, a targeted customer segment and various menus were the important evaluation points to obtain the high sales for the new restaurant because the closed restaurant was almost unprofitable. From the viewpoints, "The Ramen Kaibutsu" was rejected in the early stage of the discussion because the main target of the restaurant differs from the target that they want and the restaurant probably acquires limited customers. After that, they discussed the advantages and disadvantages of the remaining candidates, "The Tsukemen Fujin" and "BonoPasta". The advantages of "BonoPasta" were .....

Figure 4: The abstractive summary in the Kyutech corpus.

Section 3.2.2. Each annotator received the following message for the summary generation: “Write a summary that is understandable for somebody who was not present during the meeting.”

We obtained two abstractive summaries from two annotators for one conversation. We computed an agreement rate between the two summaries by using ROUGE-N (Lin and Hovy, 2003). ROUGE-N is an n-gram recall between a reference summary and a system summary and widely used in automatic evaluation of summaries. ROUGE-N is computed as follows:

$$ROUGE\text{-}N(S,R) \quad = \quad \frac{\sum_{e \in n\text{-}gram(S)} Count_{match}(e)}{\sum_{e \in n\text{-}gram(R)} Count(e)} \tag{2}$$

where n stands for the length of the n-gram, e and $Count_{match}(e)$ is the maximum number of n-grams co-occurring in a system summary and a reference summary. We used ROUGE-1 in this paper. The ROUGE-1 between the two summaries was 0.527 on average; one is a summary from an annotator as a reference summary and the other is a summary from the other annotator as a system summary. In general, the score, 0.527, is qualitatively reasonable in summarization tasks although it is difficult to evaluate whether the score is quantitatively adequate. In a similar way to the topic annotation, we generated a summary from the two summaries of annotators. For generating the third summary, we scanned not only the two summaries but also the transcription of each conversation. Thus, the third summary we made is sort of a consensus summary of two annotators. Figure 4 shows an example of the consensus summary. The ROUGE-1 between each consensus summary and two annotators' summaries was 0.564. We also regard each consensus summary and each annotator's summary as a reference summary and a system summary, respectively, in the ROUGE calculation. The ROUGE score of consensus summaries was higher than that between two annotators' summaries (0.564 vs. 0.527). This result shows that the third summaries are appropriate as consensus summaries.

4 Topic segmentation

In this section, we explain topic segmentation for the Kyutech corpus. There are two types of methods for topic segmentation; supervised and unsupervised methods. In this paper, we focus on unsupervised methods. We describe three topic segmentation methods, LCSeg, TopicTiling and the combined method, and then evaluate the methods on the Kyutech corpus, as a case study.

4.1 LCSeg

LCSeg is an unsupervised cohesion-based technique proposed by (Galley et al., 2003) to topic modeling for meeting transcripts. We compute the $tfidf$ score for LCSeg.

$$tfidf(R_i) \quad = \quad freq(t_i) \cdot log(\frac{L}{L_i}) \tag{3}$$

where R_i denotes a repetition score of a term t_i. $freq(t_i)$ is the frequency of t_i in a chain. L_i and L are the respective length and the length of the text, respectively. Then, we compute a lexical cohesion by using the cosine similarity at the transition between two windows. For the calculation, LCSeg uses lexical chains that overlap with the two windows. The similarity cos_L between windows (A and B) is

ConvID	Utterances	Segments
Conv1	505	52
Conv2	637	77
Conv3	324	33
Conv4	502	36
Conv5	566	48
Conv6	487	51
Conv7	284	31
Conv8	445	42
Conv9（dev）	759	48

Table 3: The number of utterances and segments of each conversation in the Kyutech corpus.

computed with

$$cos_L(A,B) = \frac{\sum_i w_{i,A} \cdot w_{i,B}}{\sqrt{\sum_i w_{i,A}^2 \sum_i w_{i,B}^2}} \tag{4}$$

$$where$$

$$w_{i,\Gamma} = \begin{cases} tfidf(R_i) & if\ R_i\ overlaps\ \Gamma \in \{A,B\} \\ 0 & otherwise \end{cases}$$

4.2 TopicTiling

TopicTiling is a text segmentation method with the Latent Dirichlet Allocation (LDA) topic model (Riedl and Biemann, 2012). It uses topic IDs obtained from the LDA inference method, instead of words. The method first estimates a topic distribution from the Kyutech corpus. Then, it generates a vector space based on topic IDs in the LDA model. The calculation of the similarity is similar to LCSeg. The similarity cos_T between windows (A and B) is also computed as follows:

$$cos_T(A,B) = \frac{\sum_n tp_{n,A} \cdot tp_{n,B}}{\sqrt{\sum_n tp_{n,A}^2 \sum_n tp_{n,B}^2}} \tag{5}$$

where tp denotes the probabilistic distribution from LDA.

4.3 Combined method

Since the characteristics of the two methods are different, they have a potential to improve the accuracy by a complementary style. Therefore, in this paper, we combine the two methods with a weight factor wf. The similarity cos_C between windows (A and B) is computed as follows:

$$cos_C(A,B) = wf \times cos_L(A,B) + (1 - wf) \times cos_T(A,B) \tag{6}$$

The weight factor wf is a trade-off parameter.

4.4 Experiment for topic segmentation

We evaluated these methods with the Kyutech corpus. The details of the Kyutech corpus are shown in Table 3. In the experiment, we used the main tags as the topic sequence. In other words, a changing point of the main tags is a border of two topics, e.g., the 7th utterance in Figure 3.

We used one conversation (Conv9) as the development data for the method. Hence we evaluated the methods with eight conversations without Conv9. In the experiment, we compared two weight factors $wf = 0.3$ and $wf = 0.7$. For the LDA, we compared three types of the number of topics, 10, 20 and 30. Parameters on LCSeg, such as the window size, were based on (Galley et al., 2003).

Method	Comp	Partial
LCseg	0.195	0.396
Topic(10)	0.142	0.394
Topic(20)	0.148	0.345
Topic(30)	0.100	0.299
Comb(10,0.3)	0.155	0.401
Comb(10,0.7)	0.182	0.399
Comb(20,0.3)	0.168	0.367
Comb(20,0.7)	0.184	0.391
Comb(30,0.3)	0.132	0.308
Comb(30,0.7)	0.172	0.362

Table 4: The F-measure on complete match and partial match.

We evaluated these methods with two criteria; complete matching and partial matching that were used in (Tajima, 2013). We computed the F-measure from the recall and precision rates for the complete and partial matching. The values are computed as follows:

$$p_{comp} = \frac{|B_r \cap B_h|}{[B_h|}, \; r_{comp} = \frac{|(B_r \cap B_h)|}{|B_r|} \tag{7}$$

where B_r is the set of the sentence IDs before each topic change. B_h is the set of the outputs from each method.

$$p_{part} = \frac{|B'_r \cap B_h|}{[B_h|}, \; r_{part} = \frac{|(B_r \cap B'_h)|}{|B_r|} \tag{8}$$

where $B'_r = \bigcup_{i \in B_r} i-1, i, i+1$ and $B'_h = \bigcup_{i \in B_h} i-1, i, i+1$. The F-measure is the harmonic mean between the recall and precision rates.

Table 4 shows the experimental result about the complete match and the partial match. Topic and Comb are the methods with TopicTiling and the combined methods, respectively. Topic(β) in the table denotes the number of topics in LDA and $\beta = \{10, 20, 30\}$. β and wf in Comb(β, wf) denote the number of topics and the value of the weight factor ($wf \in \{0.3, 0.7\}$). For the complete matching, LCSeg produced the best performance. For the partial matching, Comb(10,0.3) obtained the highest F-measure value although there is no dramatic improvement as compared with the single methods, LCSeg and TopicTiling. TopicTiling-based methods were low accuracy on the whole. This is one reason that the combined methods did not improve the accuracy. The size of the Kyutech corpus is not always sufficient for the statistical methods, as compared with the AMI corpus. For the TopicTiling-based methods, we need a larger dataset. Moreover, the values on the F-measure were not high (0.401 even on the partial match scheme). Galley et al. (2003) reported that a feature-based segmentation method outperformed LCSeg. Applying a supervised method into our task leads to the improvement of the accuracy of the topic segmentation. In general, machine learning methods need a large dataset to generate a strong classifier. Therefore, scaling up the Kyutech corpus is the most important future work.

5 Discussion and Conclusions

In this paper, we explained the Kyutech corpus and a topic segmentation task for the corpus as the first step of multi-party conversation summarization. The Kyutech corpus consists of conversations about a decision-making task with four participants. The corpus contained utterances with time information, topic annotation and reference summaries.

For the topic annotation, we prepared 28 topic tags, and generated the annotated corpus in the two steps; (1) annotation by two annotators and (2) final judgment of each tag by three annotators. The partial agreement score AS_p between annotators was 0.879. In addition, the AS_p between final tags that the authors created and tag sets from two annotators was 0.965. In a similar way, we generated three summaries; two summaries by annotators and a consensus summary of the two summaries. The ROUGE-1 score among them was 0.564 on average. To the best of our knowledge, the Kyutech corpus is the first Japanese corpus annotated for summarization tasks and freely available to anyone.

As a case study of the corpus, we evaluated some topic segmentation methods. We compared LCSeg, TopicTiling and a combined method on the Kyutech corpus. However, there is no dramatic improvement of the accuracy. One reason was that TopicTiling was not effective in our experiment. It was caused by the size of the Kyutech corpus. Therefore, scaling up the Kyutech corpus is the most important future work.

The Kyutech corpus contains the topic tags and summaries. On the other hand, the AMI corpus contains numerous annotations, such as extractive summaries and dialogue-acts. Our topic tags focused on semantic contents of each utterance because of our purpose, namely summarization. However, communicative functions (Bunt, 2000), such as `INFORM` and `Auto-Feedback`, are also an important role as a conversation corpus. We are currently developing the Kyutech corpus with communicative functions, and then are going to open the new corpus in the next phase. In addition, hierarchical topic tag definition, such as (Ohtake et al., 2009), might be appropriate for our summarization task because each utterance often contained some topic tags. Other annotation to the Kyutech corpus is also future work. In addition, an extension of the Kyutech corpus to a multi-modal corpus with audio-visual data, such as (Sanchez-Cortes et al., 2013) and (Nihei et al., 2014), is important future work. In this paper, we just dealt with a topic segmentation task. However, the main purpose is to summarize a multi-party conversation. Abstractive summarization using the segmented topics is also the important future work.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 26730176. A part of this work was carried out during the second author's stay at the University of British Columbia. We would like to express our gratitude to Giuseppe Carenini and Raymond T. Ng.

References

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Generating abstractive summaries from meeting transcripts. In *Proceedings of ACM Symposium on Document Engineering (DocEng '15)*, pages 51–60.

Harry Bunt, 2000. *Abduction, Belief and Context in Dialogue: Studies in computational pragmatics*, chapter Dialogue pragmatics and context specification, pages 81–150. John Benjamins Publishing Company.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2008)*, pages 334–343.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, pages 562–569.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceeding of the 32nd Annual Meeting on Association for Computational Linguistics (ACL 1994)*, pages 9–16.

Ryuichiro Higashinaka and Kotaro Funakoshi. 2014. Chat dialogue collection and dialogue breakdown annotation in the dialogue task of project next nlp (in Japanese). In *JSAI, SIG-SLUD-B402-08*, pages 45–50.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *Proceedings of the IEEE ICASSP*, pages 364–367.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language*, pages 71–78.

Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1220–1230.

Fumio Nihei, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. 2014. Predicting influential statements in group discussions using speech and head motion information. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 136–143.

Kiyonori Ohtake, Teruhisa Misu, Chiori Hori, Hideki Kashioka, and Satoshi Nakamura. 2009. Annotating dialogue acts to construct dialogue systems for consulting. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 32–39.

Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of INLG 2014*, pages 45–53.

Martin Riedl and Chris Biemann. 2012. Topictiling: A text segmentation algorithm based on LDA. In *Proceedings of the 50th Annual Meeting on Association for Computational Linguistics (ACL 2012)*, pages 37–42.

Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. 2013. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 7(1):39–53.

Yasuhiro Tajima. 2013. Performance comparison between different language models on a text segmentation problem via hmm (in Japanese). *Information Processing Society of Japan. Transactions on mathematical modeling and its applications*, 6(1):38–46.

National Institute for Japanese Language and Linguistics. 2006. Construction manual of Corpus of Spontaneous Japanese (CSJ) (in Japanese).