

Annotation and Analysis of Extractive Summaries for the Kyutech Corpus

Takashi Yamamura and Kazutaka Shimada

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. For conversation summarization tasks, corpora have an important role to analyze characteristics of conversations and to construct a method for summary generation. We are developing a freely available Japanese conversation corpus for a decision-making task. We call it the Kyutech corpus. The current version of the Kyutech corpus contains topic tags of each utterance and reference summaries of each conversation. In this paper, we explain an annotation task of extractive summaries. In the annotation task, we annotate an importance tag for each utterance and link utterances with sentences in reference summaries that already exist in the Kyutech corpus. By using the annotated extractive summaries, we can evaluate extractive summarization methods on the Kyutech corpus. In the experiment, we compare some methods based on machine learning techniques with some features.

Keywords: multi-party conversation, the Kyutech corpus, summarization, annotation

1. Introduction

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, summarization systems need to recognize significant content from each utterance to cover all important information in the conversation. Therefore, 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). These meeting corpora contain meeting record data with many annotations, such as dialogue acts and summaries and researchers have effectively used such annotations in the summarization task.

We have constructed the Kyutech corpus (Yamamura et al., 2016); it is a Japanese conversation corpus about a decision-making task with four participants. To the best of our knowledge, the Kyutech corpus is the first Japanese corpus annotated for summarization tasks and freely available to anyone¹. The current Kyutech corpus consists of nine conversations² with four scenarios of which discussion settings differ from each other; topic tags of each utterance, and reference summaries of each conversation. On the other hand, the AMI corpus contains numerous annotations, such as extractive summaries and dialogue acts.

In this paper, we focus on an annotation task of extractive summaries. The purpose of extractive summarization is to select important utterances automatically. We annotate an importance tag to each utterance by linking sentences in reference summaries. The annotated utterances are extractive summaries for each conversation. By using the extractive summaries, we can apply extractive summarization methods that have already been proposed by sev-

eral researchers to the Kyutech corpus because there has been a lot of works on extractive techniques in conversation summarization (Murray et al., 2005, Hirohata et al., 2006). Therefore, construction of extractive summaries is effective and leads to a deeper analysis of the Kyutech corpus.

The final goal of our study is to generate an abstractive summary from a multi-party conversation. This annotation is also useful to apply abstractive summarization methods to the Kyutech corpus because there have been some studies on abstractive models by using extractive techniques (Mehdad et al., 2013).

The contributions of this paper are as follows:

- We extend the Kyutech corpus by annotating extractive summaries.
- As a case study, we examine extractive summarization methods using supervised approaches.

2. Related Work

Extractive summarization has been studied in various domains, such as news articles (Nallapati et al., 2016) and meeting records (Tixier et al., 2017). In multi-party conversation, extractive summarization is a difficult challenging task because the meeting transcripts are composed informal and disfluency utterances with overlapping speakers (McKeown et al., 2005). Therefore, extractive approaches for conversation summarization often differ from techniques of other domains, such as document summarization. Feature-based approaches are commonly used for meeting summarization. Xie et al. (2008) have evaluated the effectiveness of different types of features, such as lexical, structural, discourse and topic features.

In this paper, we focus on a summarization task on a Japanese conversation corpus. The Corpus of Spontaneous Japanese (CSJ) (National Institute for Japanese Language and Linguistics, 2006) is a famous speech corpus with audio data, and most of the speech materials are spontaneous monologues. Hirohata et al. (2006) have been proposed sentence extractive speech summarization on the corpus.

¹ <http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html>

² The discussion time is 20 minutes per each conversation.

While the CSJ is a spontaneous monologues corpus, the Kyutech corpus is a multi-party conversation corpus for a decision-making task. Therefore, the technique of Hirohata et al. (2006) to summarize a monologue is not necessarily suitable for the Kyutech corpus.

In a Japanese multi-party conversation, Tokunaga and Shimada (2015) introduced machine learning approaches with verbal and nonverbal features at the sentence extraction step. In our past research, we reported that additional features about time information were relatively variable (Yamamura et al., 2015). In this paper, we compare three machine learning approaches with these features; Support Vector Machines (SVMs) (Vapnik, 1999), Conditional Random Fields (CRFs) (Lafferty et al., 2001), and Random Forests (Breiman, 2001).

3. The Kyutech corpus

In this section, we explain the current Kyutech corpus. It contains topic tags for each utterance and three reference summaries for each conversation (Yamamura et al., 2016). The Kyutech corpus contains multi-party conversations with four participants randomly selected from sixteen male students and four female students. The participants pretended managers of a virtual shopping mall in a virtual city and then determined a new restaurant, as an alternative to a closed restaurant, from three candidates. 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, statistical information about the shopping mall, and so on. They read the document for 10 minutes, then discussed the candidates for 20 minutes and finally determined one restaurant as a new restaurant opening. The current Kyutech corpus consists of nine conversations based on four scenarios of which task settings differ from each other.

The transcription rules were based on the construction manual of the Corpus of Spontaneous Japanese (CSJ) by (National Institute for Japanese Language and Linguistics, 2006). All utterances in the corpus were separated by 0.2-second interval by the guideline and annotated some tags such as filler, question, and so on. Each utterance was not always sentence-level because it depended on the 0.2-second interval rule. Therefore, other tags were appended to the end of each utterance for sentence-level identification. The corpus consists of 4,509 utterances in nine conversations, with a total of 2,810 sentences.

The Kyutech corpus contains the annotations for conversation summarization. In general, topic segmentation (Galley et al., 2003) has an important role as the first step in the meeting summarization (Banerjee et al., 2015, Oya et al., 2014). By dividing the utterances into topic units by topic segmentation, it is possible to take into account of topics of the discussion in conversation summarization. Each utterance has topic tags³ representing its topics to analyze topic sequences. Table 1 shows the tag names and the descriptions.

³ At least one tag is annotated to one utterance, and up to two additional tags are also allowed.

Topic	Description
CandX	Topic about the candidate 1
CandY	Topic about the candidate 1
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
CIEx	Topic about the existing restaurants and the closed restaurant
Mall	Topic about the shopping mall
OtherMall	Topic about other shopping malls
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 1: Topic tags in the Kyutech corpus.

At the summary generation steps, we complied with the guideline of abstractive hand summaries of the AMI corpus⁴. Based on the manual, two abstractive summaries for one conversation were generated, and the size of each summary was 250 characters to 500 characters⁵. When each annotator generates a summary, they receive the following message for the summary generation: “Write a summary that is understandable for somebody who was not present at the meeting. We also generated third reference summaries from the two summaries of annotators as the consensus summary.

4. Extractive summarization

In this section, we explain an extractive summarization task in the Kyutech corpus. First, we detect important utterances in each conversation. Here the important utterances denote utterances that relate to sentences in reference summaries. Next, we explain some extractive summarization models based on machine learning techniques.

4.1. Annotation

The current Kyutech corpus contains the three reference summaries of each conversation. We created extractive

⁴ <http://groups.inf.ed.ac.uk/ami/corpus/guidelines.shtml>

⁵ The number of words was approximately 150 content words on average. The number of unique words was 80 words on average.

- (a) At the meeting, participants discussed which of candidate restaurants is suitable as an alternative to the closed restaurant *Japanese WAYA*.
 (b) They held a meeting to select a new restaurant from *Ethnic Food Okinawa*, *Taiwan Noodle*, and *BonoPasta* as an alternative to the closed restaurant *Yakiniku Mr.K* that was unprofitable.
 (c) They discussed a new restaurant from *The Ramen Kaibutsu*, *The Tsukemen Fujin*, and *BonoPasta* to replace the closed restaurant *Japanese WAYA*.
 (d) They discussed the selection of a new restaurant which is most suitable in *Ethnic food Okinawa*, *Taiwan Noodles*, and *BonoPasta* to make up for the closed restaurant *Yakiniku Mr.K*.
 (e) After that, they discussed the merits and demerits of the remaining candidates, *The Tsukemen Fujin* and *BonoPasta*.

Figure 1: An example of reference summary sentences not related to any utterance. Italic fonts denote the restaurant names in the Kyutech corpus.

Dialog ID	Utterances	Links	Ratio (%)
0313_C1	759	240	31.6
0320_C1	505	124	24.6
0326_C1	502	76	15.1
0326_C2	566	160	28.3
0327_C2	284	52	18.3
0323_C3	324	102	31.5
0327_C3	445	118	26.5
0320_C4	637	69	10.8
0326_C4	487	98	20.1

Table 2: The number of utterances and links with reference summary of each conversation in the Kyutech corpus

summaries using the consensus summaries for each conversation. First, the annotator compared all utterances with each sentence in the consensus summary for each conversation. Next, if an utterance implies the content of a sentence in a reference summary, the annotator links it with the sentence in the reference. More precisely, we annotate the sentence number in the reference summary to the utterance. In other words, we link important utterances with sentences in reference summaries that already exist in the Kyutech corpus. We regard them as extractive summaries for each conversation.

Table 2 shows the annotation results. “Utterances” in the table denotes the total number of utterances of the conversation and “Links” denotes the total number of utterances linked with reference summaries as extractive summaries of the conversation. “Ratio” in the table denotes the percentage of “Links” to “Utterances”. The ratio on all conversations was 23 percent on average.

In this annotation, there were 19 percent of the sentences not related to any utterances although most sentences in reference summaries were linked with several utterances. Figure 1 shows examples of these non-linked sentences. The sentences (a) to (d) were the lead sentences of the reference summaries and described the task of the discussion, e.g., theme and setting. It is hard to link the description and utterances in the conversation. Moreover, as shown in the sentence (e), it is also difficult to uniquely determine an utterance related to the reference summary because there are many utterances related to this description in the conversation. The annotation for these sentences is one important future work.

4.2. Extractive summarization method

One of the main purposes of this study is to summarize a multi-party conversation. In this section, we explain some

Criteria	Random Forests	SVMs	CRFs
Precision	0.422	0.301	0.411
Recall	0.311	0.308	0.294
F-measure	0.346	0.294	0.326

Table 3: Macro-averaged precision, recall, and F-measure scores.

extractive summarization methods.

Some abstractive conversation summarization approaches utilize extractive summarization techniques in the process (Banerjee et al., 2015). In other words, extractive summarization has an important role in summary generation. Therefore, we evaluate the extractive summarization task as the first step for our abstractive summarization.

We have studied extractive summarization for another conversation corpus (Tokunaga and Shimada, 2015, Yamamura et al., 2015). In this paper, we introduce our previous approaches to the Kyutech corpus.

We examine extractive summarization by using three machine learning techniques; SVMs, CRFs, and Random Forests. As features of these machine learning techniques, we use 20 types of verbal and nonverbal information that utilized in (Tokunaga and Shimada, 2015, Yamamura et al., 2015). The following are simple lists of the featurese:

Features in an utterance

Speaker information (Speaker ID), Utterance position in a conversation, Topic tags in the Kyutech corpus, The number of morphemes in an utterance, Length of an utterance, Presence of declinable words and Presence of interrogatives.

Features between utterances

Difference of lengths of current and previous utterances, Word frequency, Presence of same words, Presence of consecutive utterances of one person.

Nonverbal features

Utterance speed, Utterance timing, Overlap of speech.

5. Experiment

The Kyutech corpus contains nine conversations. We evaluated our methods on the Kyutech corpus with nine-fold cross validation for nine conversations. In other words, we evaluated one test conversation with the model that was generated from the other conversations and repeated this process for all conversations. We computed the precision, recall rates, and F-measure for each conversation and took an average of the overall scores (macro-averaging).

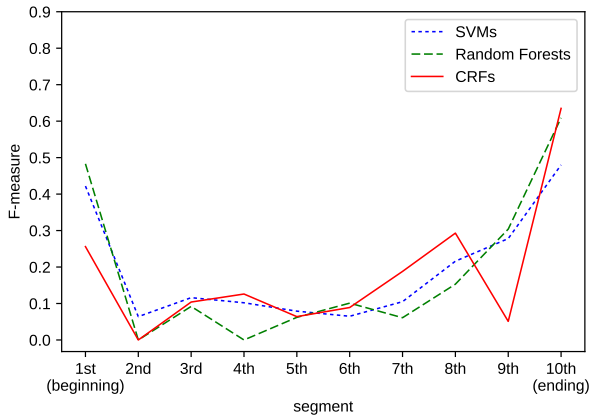


Figure 2: Macro-averaged F-measure scores’ comparison for each segment of the meetings. The x-axis, shown as segment, denotes the number of each divided utterances group. The 1st segment means the beginning part of a conversation and the 10th segment means the ending part of a conversation.

Table 3 shows the experimental results. The F-measures of each method were not high on the whole. We also evaluated our methods with some combination patterns of the features to confirm the effectiveness of each feature. As the result of the analysis, the utterance position feature, namely utterance numbers normalized by the total number of utterances, was most effective in our features. Moreover, there was almost no difference between a method using only the utterance position feature and others. In other words, our models heavily depended on the utterance position feature. In order to understand why the feature is effective, we analyzed utterances extracted by our models. The result showed that the models with the utterance position feature mainly extracted utterances in the beginning and the ending parts of conversations as important utterances.

As you can see from Table 3, the performance was not enough. We analyzed the results more deeply. The current models tended to extract important utterances from the beginning and the ending parts in conversations. Therefore, we compared the F-measure scores in each segment in conversations. We divided each conversation into ten segments equally. Figure 2 shows the F-measure scores of each segment for each model. All models obtained high scores in the last segment of the conversation, as compared with the other segments. The last segment of each conversation tends to contain important utterances because it usually contains the final decision of the conversation. Therefore, our models could extract the most important points, namely the decision, in each conversation although the overall scores shown in Table 3 were not enough. The F-measure scores of the 1st segment were relatively higher than the other segments. The beginning part of the conversation usually contains a trigger of the discussion. Therefore, the higher accuracy in the segment shows potential effectiveness of the models.

6. Discussion

In this section, we discuss the following points: the experiment, another extractive summaries, and some summariza-

tion tasks.

As mentioned in Section 4.1., there were 19 percent of the reference summary sentences not related to any utterances. To generate these summary sentences, we need to capture not only the contents of the conversation but also discussion settings. Another problem in the experiment was that the size of the Kyutech corpus was not always sufficient for the statistical methods. 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. Future work should also focus on applying unsupervised techniques, such as submodularity framework (Tixier et al., 2017) and relation extraction (Wang and Cardie, 2012).

In this work, we designed an extractive summary annotation task for linking utterances in a conversation with sentences in a reference summary. However, there are cases when someone often wants to know the course of discussion in a meeting. For this demand, our extractive summaries are not suitable. However, we have already studied the annotation process of extractive summaries that keep the meaning and context of the original conversation in our past research (Tokunaga and Shimada, 2015). It is also important to apply this process to the Kyutech corpus.

Although we handled a generic summarization task, some researchers have studied other aspects of summarization: focused meeting summarization that creates abstract summaries of specific aspects of meeting such as decisions, actions, and problems (Murray et al., 2010, Wang and Cardie, 2013) and query-based summarization that generates abstract summaries based on users phrasal queries (Mehdad et al., 2014). Introducing these summarization systems is also the future work.

7. Conclusion

In this paper, we explained the annotation task for extractive summarization on the Kyutech corpus. We annotated importance utterances that already exist in the Kyutech corpus, on the basis of reference summaries. As a result, we linked 23 percent of utterances in conversations with sentences in reference summaries as extractive summaries.

We examined some extractive summarization methods. The utterance position feature was most effective in our designed features. However, the accuracy was not enough. Improving the accuracy is one of the most important future work. In addition, scaling up the Kyutech corpus is important future work. In this work, we used features about surface information of utterances and topic tags. On the other hand, there are other important features in conversations, such as dialogue acts (Bunt et al., 2012). We are currently developing the Kyutech corpus with dialogue acts (Hino et al., 2016). We will apply the dialogue acts to extractive summarization models in future work. In addition, we have a plan to open the annotated tags; namely, extractive summaries and dialogue acts, shortly.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 17H01840.

References

- Banerjee, S., Mitra, P., and Sugiyama, K. (2015). Generating abstractive summaries from meeting transcripts. In *Proceedings of ACM Symposium on Document Engineering (DocEng '15)*, pages 51–60.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
- Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hosida, K., Petukhova, V., Porescu-Belis, A., and Trauim, D. (2012). Iso 24617-2 : A semantically-based standard for dialogue annotation.
- Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.
- Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, pages 562–569.
- Hino, M., Yamamura, T., and Shimada, K. (2016). Design and analysis of dialogue acts for the kyutech corpus. In *IEICE Technical Report, Natural Language Understanding and Models of Communication (In Japanese)*, volume 116, pages 1–6.
- Hirohata, M., Shinnaka, Y., Iwano, K., and Furui, S. (2006). Sentence-extractive automatic speech summarization and evaluation techniques. In *Speech Communication*, volume 46, pages 1151–1161.
- Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings of the IEEE ICASSP*, pages 364–367.
- Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 282–289.
- McKeown, K., Hirschberg, J., Galley, M., and Maskey, S. (2005). From text to speech summarization. In *Proceedings of the 30th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 997–1000.
- Mehdad, Y., Carenini, G., Tompa, F. W., and Ng, R. T. (2013). Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Natural Language Generation (ENLG - SIGGEN 2013)*, pages 136–146.
- Mehdad, Y., Carenini, G., and Ng, R. T. (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.
- Murray, G., Renals, S., and Carletta, J. (2005). Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 593–596.
- Murray, G., Carenini, G., and Ng, R. T. (2010). Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of INLG 2010*, pages 105–113.
- Nallapati, R., Zhou, B., Santos, C., Gulcehre, C., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 280–290.
- Oya, T., Mehdad, Y., Carenini, G., and Ng, R. (2014). A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of INLG 2014*, pages 45–53.
- National Institute for Japanese Language and Linguistics. (2006). Construction manual of corpus of spontaneous japanese (csj) (in japanese).
- Tixier, A., Meladianos, P., and Vazirgiannis, M. (2017). Combining graph degeneracy and submodularity for unsupervised extractive summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 48–58.
- Tokunaga, Y. and Shimada, K. (2015). Extractive summarization based on a combined method using several features for multi-party conversation. In *ACIS International Journal of Computer and Information Science*, volume 16, pages 12–21.
- Vapnik, V. N. (1999). *Statistical Learning Theory*. Wiley.
- Wang, L. and Cardie, C. (2012). Focused meeting summarization via unsupervised relation extraction. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 304–313.
- Wang, L. and Cardie, C. (2013). Domain independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1395–1405.
- Xie, S., Liu, Y., and Lin, H. (2008). Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *Spoken Language Technology Workshop*, pages 157–160.
- Yamamura, T., Tokunaga, Y., and Shimada, K. (2015). Multi-party conversation summarization using time information and text segmentation. In *IEICE Technical Report, Natural Language Understanding and Models of Communication (In Japanese)*, volume 115, pages 43–48.
- Yamamura, T., Shimada, K., and Kawahara, S. (2016). The Kyutech corpus and the topic segmentation using a combined method. In *Proceedings of the 12th Workshop on Asian Language Resources*, pages 95–104.