Conversation summarization using machine learning and scoring method

Kazutaka Shimada, Shinpei Toyodome and Tsutomu Endo
Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Iizuka Fukuoka 820-8502 Japan
{*shimada, s_toyodome, endo*}*@pluto.ai.kyutech.ac.jp*

***Abstract*—In this paper, we propose a method for conversation summarization. For the method, we combine two approaches; a scoring method and a machine learning technique. We extract important utterances with high confidence from a conversation by using the scoring method. However, the number of extracted utterances is not enough as a summary. To solve this problem, we incorporate utterances extracted by SVMs to the summary. For the integration process, we compare some approaches. In the experiment, our method generated balanced summaries in terms of the summary length and readability, as compared with a method with SVMs only.**

***Keywords*-Conversation summarization, Scoring, SVMs, Integration.**

I. Introduction

Multi-party conversation is a communication that involves three or more participants with utterances. There are many types of multi-party conversation such as spontaneous dialogues, meetings and chats on the Web. To understand the content of a conversation easily, the summarization has an important role.

In this paper, we propose a method for conversation summarization. Traditional summarization studies have handled a single document or multi-documents as the target [5]. Many studies in the summarization are based on extraction approaches [2], [13]. In these approaches, the systems extract sentences on the basis of term frequency, location, cue words and so on. Our method is also based on an extraction approach.

The target data in this paper is multi-party conversation. For conversation summarization, relations between utterances are more important, as compared with document summarization such as news papers. Higashinaka et al. [4] have proposed an improved HMM-based summarization method for contact center dialogues. The dialogue in a contact center consists of utterances between two persons. We handle conversations with four persons as the target data. The data are free conversation about a topic. In other words, our multi-party conversations are more spontaneous and not well-structured. Therefore, relations between utterances are more complicated. Xie et al. [11] have evaluated the effectiveness of different types of features. They compared lexical, structural, discourse and topic features for machine learning. However, utterances in conversations usually contain anaphoric relations. Lack of these relations in a summary leads to decrease of readability. To solve this problem, we introduce features about anaphora.

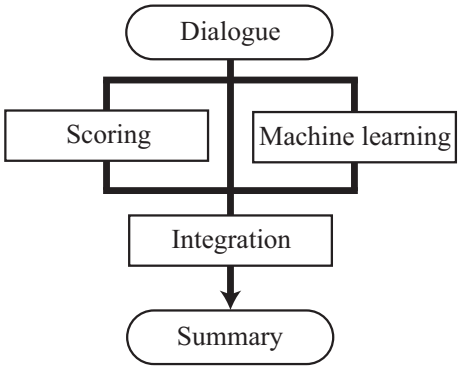

Figure 1. The outline of our method.

For the conversation summarization, we combine two approaches; a scoring method and a machine learning technique. We extract important utterances with high confidence from a conversation by using the scoring method. However, the number of utterances extracted by the scoring method is not enough as a summary. To solve this problem, we incorporate utterances extracted by SVMs to the summary. We apply three integration approaches to summary generation. In the experiment, we compare the proposed method with a baseline method based on SVMs only.

II. Method

In this section, we explain our proposed method.

A. Outline

Our method extracts utterances in a conversation on the basis of the importance and relations between utterances in the summarization process. We use two approaches for the process; a scoring method and a machine learning technique. Figure 1 shows the outline of our method.

The purpose of the scoring process is to extract the most important utterances in each conversation with high precision. We extract utterances exceeding a threshold. Although the number of utterances is small, the extracted utterances are the core of a summary.

The second method is a machine learning technique. We use Support Vector Machines (SVM) [9]. The purpose of the

method based on SVMs is to extract important utterances that are not extracted by the scoring method. The output of SVMs contributes to the improvement of the recall rate about a generated summary.

In the integration process, we combine the outputs from the two approaches. In this paper, we propose three patterns for the integration.

B. Scoring

We compute a score of each utterance in a conversation. We apply the panoramic view system proposed by Sunayama and Yachida [8] to the scoring process. The method computed a score by word frequencies and conditional probabilities based on the co-occurrence frequency of words in sentences. They defined three types of keywords, (1) basic keywords, (2) topic keywords and (3) feature keywords, and then computed a score of each sentence by using the scores of the three keywords.

First, we divide each utterance in a conversation to morphemes by using a morphological analyzer[1]. In this paper, we handle nouns[2], verbs and adjectives in each utterance. The score of a basic keyword is based on the frequency.

$$key1(w) \quad = \quad frequency(w) \tag{1}$$

Next, we compute the second score. They defined words which often appear together with the basic keywords as topic keywords. The score of a topic keyword is computed as follows:

$$key2(w) \quad = \quad \sum_{g \subset G} \frac{n(w \cap g)}{n(g)} \tag{2}$$

where G is the set of basic keywords and $n(g)$ is the number of utterances containing a basic keyword g. $n(w \cap g)$ is the co-occurrence of w and g. We use words in the top 10 % of all as topic keywords.

Finally, we detect feature keywords and compute the score.

$$key3(w) \quad = \quad \sum_{s \subset S} \frac{n(w \cap s)}{n(w)} \tag{3}$$

where S is the set of topic keywords. The purpose of this score is to detect words that appear only in sentences containing the topic keywords. It is based on the idea that words consistent with the flow of topic keywords forming the main topic of the text and not appearing in other sentences are given high evaluations.

[1]We used Mecab. http://mecab.sourceforge.net/

[2]We use nouns of which the frequencies are more than 2.

Then, we compute the score of each utterance by using three scores on word-level, namely $key1$, $key2$ and $key3$.

$$sent1(U) \quad = \quad \sum_{w \subset T} key1(w) \tag{4}$$

$$sent2(U) \quad = \quad \sum_{w \subset T} key2(w) \tag{5}$$

$$sent3(U) \quad = \quad \sum_{w \subset T} key3(w) \tag{6}$$

By using these equations, we obtain three scores on utterance-level. The score of an utterance is computed by the summation of these scores.

$$score(U) \quad = \quad sent1(U) + sent2(U) + sent3(U) \tag{7}$$

Here we add a new factor to the scoring. We assume that a long utterance has an important role in a conversation. Therefore, we introduce a weighting factor basing on the number of morphemes in each utterance.

$$finalScore(U) \quad = \quad score(U) \times \frac{MorpT}{AveMorp} \tag{8}$$

where $MorpT$ is the number of morphemes in the utterance U and $AveMorp$ is the average number of morphemes in one utterance in the conversation.

The purpose of the scoring process is to extract the most important utterances in each conversation with high precision. Therefore we set a strong limitation on the extraction. In this paper, we regard utterances with the top 10 % score of all as the important utterances in this scoring method[3].

C. Machine learning

One approach to extract important information is to utilize machine learning techniques. We apply SVMs to this summarization process. We use 19 features for SVMs. They are classified as (1) features in an utterance, (2) features between utterances, (3) features about anaphora and (4) features based on scores.

1) Features in an utterance: The first feature category consists of four features focusing on each utterance itself.

- Length:
  Long utterances include much information and often contain high potential as important utterances. Therefore, we use the number of morphemes as the feature.
- Presence of word:
  We set TRUE to the feature if an utterance contains a word that occurs twice or more in a conversation.
- Presence of verbs and adjectives:
  In spontaneous conversations, utterances without any verbs and adjectives often exist. We apply the presences of verbs and adjectives to the feature.

[3]Here we use another limitation. Our method does not extract three or more consecutive utterances by one person. This is a heuristic rule.

- Interrogative:
  Interrogative utterances often have an important role in a conversation because they are a turning point of a topic and contain strong context for previous and next utterances. Therefore, we use the presence of interrogative as the feature.

2) Features between utterances: The second feature category consists of six features about relations between utterances.

- Difference of length:
  The next utterance of an utterance with high importance sometimes consists of a small number of words, such as expressions about agreement. Therefore we use the difference of the utterance length between the current utterance and the next three utterances[4].
- Presence of word in the previous utterance:
  We also set TRUE to this feature if the previous utterance contains a word that occurs twice or more in a conversation.
- Presence of interrogative in the previous utterance:
  We also use the presence of interrogative in the previous utterance as the feature.
- Same word in the previous utterance:
  If a word in the current utterance is included in the previous utterance, the current utterance contains high potential as important utterances because it indicates that the two utterances contain strong context. We handle nouns, verbs and adjectives for the feature.
- Same word in the next three utterances:
  We also use the same word feature for the next three utterances.
- Consecutive utterance:
  If one person utters continuously, the second utterance sometimes is supplemental information. It often boosts the importance of the utterance. Therefore we use the presence of consecutive utterances of one person as the feature.

3) Features about anaphora: Anaphoric relation is one of the most considerable points in conversation summarization tasks. Nishikawa et al. [6] have reported the significance of handling of anaphoric relations. If a summary contains an utterance with an anaphora and does not contain an utterance with the antecedent, the readability of the summary dramatically decreases. Therefore, the features about anaphoric relations are of extreme importance. The third feature category consists of six features from three pairs.

- Referring expression in the current utterance and referring expression in the next three utterances:
  These features are based on the presence of referring expressions such as "kore (this)" and "sotti (there)".
- Connective expression in the current utterance and connective expression in the next three utterances:
  These features are based on the presence of connective expressions such as "demo (but)", "shikamo (furthermore)" and "tadasi (unless)".
- Response expression in the current utterance and response expression in the next three utterances:
  These features are based on the presence of response expressions such as "hee (heh)" and "un (Yes)".

4) Features based on scores: The fourth feature category consists of three features based on the scores in Section II-B.

- Score of basic keywords:
  We use the score of $sent1$ (Eq. 4).
- Score of topic keywords:
  We use the score of $sent2$ (Eq. 5).
- Score of feature keywords:
  We use the score of $sent3$ (Eq. 6).

Here, these scores are normalized in 0 to 1.

D. Summary generation

The previous and next utterances of the important utterance extracted by our method are usually significant in the conversation summarization. They have an impact on the readability of the generated summary.

In this paper, we apply three approaches to the summary generation process. The basic summary consists of utterances by the scoring method including Eq. 8.

Method$_{PN}$:
: This method adds the previous and next utterances of each utterance in the basic summary to the summary.

Method$_{ML}$:
: This method adds utterances by SVMs to the summary. It selects one utterance with the high output score of SVMs between utterances in the basic summary.

Method$_{PNML}$:
: This method is the combination of Method$_{PN}$ and Method$_{ML}$

Figure 2 shows an example of each approach.

III. Dataset

For the machine learning and evaluation, we need a tagged corpus with an importance degree of each utterance. In this paper, we use the conversation corpus used in [7]. It consists of 10 spontaneous conversations with 1615 utterances[5]. The number of participants in each conversation is 4 persons. The participants had a talk about "Movies", "Games", "SNS" and so on.

For the conversations, three annotators judged the importance degree of each utterance in a phased manner. Figure 3 shows the process. We regard all utterances in each conversation as level-1. First, the annotator selected three quarters of utterances from all utterances (level-2).

[4]Actually, the difference is based on the number of morphemes.

[5]All utterances in the conversations are in Japanese.

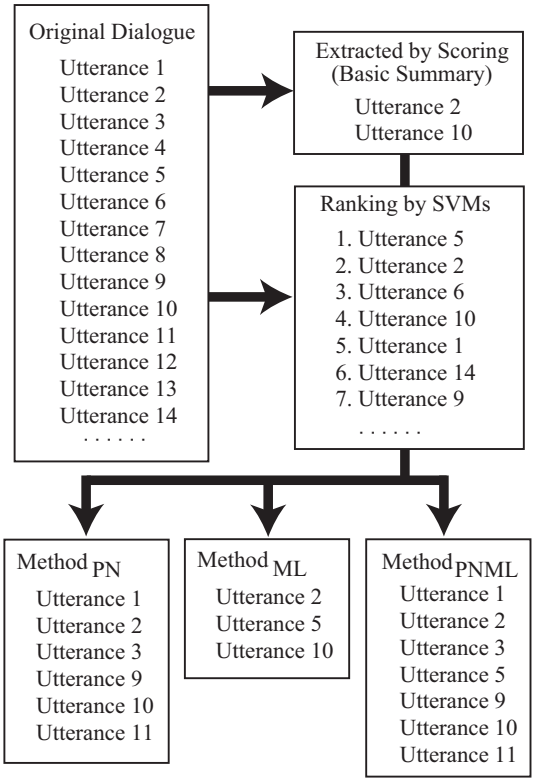

Figure 2. An example of the summary generation.

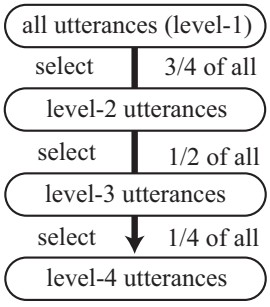

Figure 3. The annotation process.

Next, the annotator selected a half of utterances from the selected utterances (level-3). Then, the annotator selected a quarter of utterances from the level 3 utterances (level-4). Finally, we selected utterances obtaining the average level of three annotators which was more than 3, as the important utterances for the summarization.

The agreement between annotators is as follows: 0.59 for the Annotator 1 and 2 and 0.57 for the Annotator 1 and 3. Both the κ values are approximately 0.3 and not high[6]. We also compute the agreement as a two-class problem. In other words, we integrate the level 3 and 4 to "important" and 1 and 2 to "not important". In this situation, the agreements are 0.75 and 0.73 and the κ values are 0.50 and 0.46.

[6]The average value of the mean squared errors between them, namely the annotator 1, 2 and 3, is approximately 1.2.

Table I
RESULT OF THE SCORING

	Precision	Recall	F
Our method	**0.948**	0.194	0.323
without Eq. 8	0.873	0.197	0.321

Table II
RESULT OF SVM

Feature	Precision	Recall	F
(1)	0.789	0.682	0.732
(2)	0.754	0.626	0.684
(3)	0.539	0.650	0.589
(4)	0.728	0.561	0.634
ALL	**0.800**	**0.740**	**0.769**

IV. EXPERIMENT

In this experiment, first, we evaluated each utterance extraction method; Scoring and SVMs. Then, we evaluated the readability of the generated summaries.

A. Accuracy of each method

First, we evaluated the scoring method. Table I shows the experimental result. The scoring method with Eq. 8 outperformed the method without that in terms of the precision rate. The purpose of the scoring method is to construct the basic summary for the summarization process. Therefore, the scoring method with Eq. 8 was suitable, as compared with the method without that.

Next, we evaluated our method based on SVMs with 10-fold cross validation. We used the data mining tool WEKA [3] for the implementation. In the experiment, we compared the effectiveness of each feature category; (1) features in an utterance, (2) features between utterances, (3) features about anaphora and (4) features based on scores. Table II shows the experimental result. The most effective feature category was (1) features in an utterance. In particular, the length feature was effective for the accuracy. The best accuracy was produced by the combination of all features.

B. Evaluation of summary

Next, we evaluated the outputs of the summarization process. We generated summaries from 5 conversations in a qualitative manner. We compared our three methods with a baseline. The baseline summary consisted of all outputs from SVMs. Figure 4 shows an example of the output of our method. In the figure, the utterances with a rectangle are the output from the scoring method. The utterances with "**" are the outputs from SVMs and they are selected by

D: Do you do SNS? **
B: Ya.
A: Facebook and ...
C: Yup. I do.
D: I use Twitter.++
D: Ameba Pigg is a kind of SNS?
B: Well, I suppose it's a SNS.++
C: Meh, maybe.
A: There is something about Ameba Pigg in recent days.**
B: Huh? ... Happening?
C: I don't know.
A: You know, elementary school kid and junior high-school student were .!.++
A: They cracked passwords of some persons.
C: Ah! Ah! I remember now.++
C: The passwords were the date of birth or something...
B: Ya, ya!
B: Anyway, Mixi is SNS, isn't it? My first SNS is Mixi. **
C: I agree.
B: Me too.
C: That is the encounter with SNS for many people.++
B: I think a number of people do Mixi only.
D: Yup.++
D: Oh, Mobage, Mobage might be my first SNS. **
B: I've not played Mobage.
A: Me too.
.....

Figure 4. An example of the output from Method$_{PNML}$.

the principle of the method$_{ML}$[7]. The utterances with "++" are the output by the method$_{PN}$. Therefore, Fig. 4 is an output of the method$_{PNML}$ because it is the integration of the method$_{ML}$ and method$_{PN}$[8].

In the experiment, three test subjects evaluated the outputs from three methods and the baseline with the following points; (1) readability, (2) understandable and (3) quantity of the summary. The evaluation point of "understandable" was the suitability of the context of summary. The evaluation point of "quantity" was the adequacy of the size as the summary. First, the test subjects received four summaries by our methods simultaneously. Then, they compared the summaries and judged the scores of each summary. The score range from the test subjects was 1 (Bad) to 5 (Good) points. We rank each method on the basis of the average score of each criterion.

Table III shows the ranking for each conversation. In the ranking, the baseline and Method$_{PNML}$ were ranked in the 1st place for three conversations and two conversations, respectively. Table IV shows the average ranking for each method. In the evaluation score, there was a large difference between the top two methods, namley the baseline and Method$_{PNML}$, and others, namely Method$_{PN}$ and Method$_{ML}$ (1.4 and 1.6 vs. 3.4 and 3.6). Therefore, we compared the top two methods in detail.

[7]Note that utterances with "**" are not all outputs from SVMs.

[8]In other words, the combination of utterances with a rectangle and utterances with "++" is the method$_{PN}$ and the combination of utterances with a rectangle and utterances with "**" is the method$_{ML}$.

Table III
THE RANKING OF EACH DATA.

ID	1st	2nd	3rd	4th
C1	Baseline	Method$_{PNML}$	Method$_{PN}$	Method$_{ML}$
C2	Baseline	Method$_{PNML}$	Method$_{PN}$	Method$_{ML}$
C3	Baseline	Method$_{PNML}$	Method$_{PN}$	Method$_{ML}$
C4	Method$_{PNML}$	Baseline	Method$_{ML}$	Method$_{PN}$
C5	Method$_{PNML}$	Baseline	Method$_{ML}$	Method$_{PN}$

Table IV
THE AVERAGE RANKINGS OF EACH METHOD.

Baseline	Method$_{PNML}$	Method$_{PN}$	Method$_{ML}$
1.4	1.6	3.4	3.6

Table V shows the comparison of the baseline and Method$_{PNML}$. In the table, RED and UND denote the evaluation criteria "readability" and "understandable". The baseline often generated high scores in the criteria as compared with the Method$_{PNML}$. In the C4, the score of UND for the baseline was better than that for the Method$_{PNML}$ although the ranking of the Method$_{PNML}$ was better. The reason was the length of the generated summaries. Table VI shows the summary size of each method. In the table, each cell denotes the number of utterances in the summary and the summarization rates in parentheses. Since the size of the generated summaries by the baseline was the largest in them, they contained much information. In other words, the test subjects comprehensively judged that the generated summaries by the Method$_{PNML}$ were suitable in terms of the quality and the size. In other words, the Method$_{PNML}$ was a balanced approach.

In addition, the summaries generated from the Method$_{PNML}$ contained turning points in a conversation and utterances with anaphoric relations. As a result, the summaries were understandable although the size of them was small as compared with the baseline method. On the other hand, one reason that the baseline method produced good performance was that the method, namely SVMs, included some context features, namely features between utterances and anaphora. As a conclusion, information about context is the most important for the conversation summarization task. Therefore we need to apply other context features to the method for the machine learning method and the integration process of the scoring and machine learning.

In this paper, we focused on surface linguistic features for the method. However, conversations contain many char-

Table V
THE COMPARISON OF BASELINE AND METHOD$_{PNML}$.

ID	Baseline		Method$_{PNML}$	
	RED	UND	RED	UND
C1	4.7	3.7	3.3	3.0
C2	4.3	4.7	3.0	3.0
C3	4.0	3.7	3.3	3.3
C4	4.7	3.7	3.7	3.7
C5	4.7	3.0	4.7	3.7

Table VI
THE SUMMARIZATION RATES. THE NUMBER OF UTTERANCES (THE RATE).

ID	Baseline	Method$_{PNML}$	Method$_{PN}$	Method$_{ML}$
C1	60 (38.0)	49 (31.0)	38 (24.1)	30 (19.0)
C2	79 (43.4)	56 (30.8)	44 (24.2)	32 (17.6)
C3	63 (38.7)	57 (35.0)	45 (27.6)	36 (22.1)
C4	66 (39.3)	57 (33.9)	45 (26.8)	33 (19.6)
C5	72 (47.4)	52 (34.2)	42 (27.6)	31 (20.4)
Ave	68 (41.3)	54 (33.0)	43 (26.1)	32 (19.7)

acteristics, such as prosodic features [10], hot spots in each conversation [12] and laughing [7]. In addition, conversations contain discourse information such as dialogue acts [1]. These characteristics are useful for the summarization method. In particular, discourse information is important for the summary generation process to select appropriate utterances. Incorporating them to our method is the important future work.

V. CONCLUSION

In this paper, we proposed a method for conversation summarization. For the method, we utilized two approaches, namely a scoring method and a machine learning technique, and integrated them in the final summary generation process.

In the experiment, the scoring method obtained high accuracy. The weighting factor basing on the utterance length was effective for the scoring. The method based on SVMs also generated high performance. The features in the target utterance were the most effective for the method.

For the integration process, we proposed three approaches. We compared our methods with a baseline method based on SVMs only in the experiment. For the overall evaluation, the ranking of our best method sometimes was lower than the baseline method. However, our method was a balanced approach in terms of the quality and the size. In other words, the summaries by our method were small and suitable as compared with the baseline method.

Future work includes (1) introducing other context features, (2) eliminating redundant outputs from SVMs, and (3) applying new features, such as laughing information, to our method.

REFERENCES

[1] James Allen and Mark Core. Draft of DAMSL: Dialog act markup in several layers. Technical report, University of Rochester, Rochester, USA. The Multiparty Discourse Group., 1997.

[2] Harold P Edmundson. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285, 1969.

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update;. In *SIGKDD Explorations*, volume 11, 2009.

[4] Ryuichiro Higashinaka, Yasuhiro Minami, Hitoshi Nishikawa, Kohji Dohsaka, Toyomi Meguro, Satoshi Kobashikawa, Hirokazu Masataki, Osamu Yoshioka, Satoshi Takahashi, and Genichiro Kikui. Improving hmm-based extractive summarization for multi-domain contact center dialogues. In *Spoken Language Technology Workshop*, pages 61–66, 2010.

[5] Inderjeet Mani. *Automatic Summarization (Natural Language Processing, 3)*. John Benjamins Pub Co, 2001.

[6] Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. Improving quality of extractive summarization via detecting intersentential anaphora. In *Proceedings of NLP2011 (in Japanese)*, 2011.

[7] Kazutaka Shimada, Akihiro Kusumoto, Takahiko Yokoyama, and Tsutomu Endo. Hot spot detection in multi-party conversation using laughing feature. In *Technical report of IEICE, NLC2012-7*, pages 25–30, 2012.

[8] Wataru Sunayama and Masahiko Yachida. A panoramic view system for extracting key sentences with discovering keywords featuring a document. *Systems and Computers in Japan*, 34(11):81–90, 2003.

[9] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1999.

[10] Shasha Xie, Dilek Hakkani-Tur, Benoit Favre, and Yang Liu. Integrating prosodic features in extractive meeting summarization. In *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2009.*, pages 387–391, 2009.

[11] Shasha Xie, Yang Liu, and Hui Lin. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *Spoken Language Technology Workshop*, pages 157–160, 2008.

[12] Takahiko Yokoyama, Kazutaka Shimada, and Tsutomu Endo. Hot spot detection in multi-pary conversation using linguistic and non-linguistic information. In *Proceedings of NLP2012 (in Japanese)*, 2012.

[13] Klaus Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Proceedings of the 16th conference on Computational linguistics*, volume 2, pages 986–989, 1996.