Extractive Summarization Method for Contact Center Dialogues based on Call Logs

Akihiro Tamura[1*] , Kai Ishikawa[2], Masahiro Saikou[2], and Masaaki Tsuchida[2]
[1]Multilingual Translation Laboratory, MASTAR Project,
National Institute of Information and Communications Technology, Kyoto, Japan
[2] Information and Media Processing Laboratories, NEC Corporation, Nara, Japan
akihiro.tamura@nict.go.jp, k-ishikawa@dq.jp.nec.com,
m-saikou@ax.jp.nec.com , m-tsuchida@cq.jp.nec.com

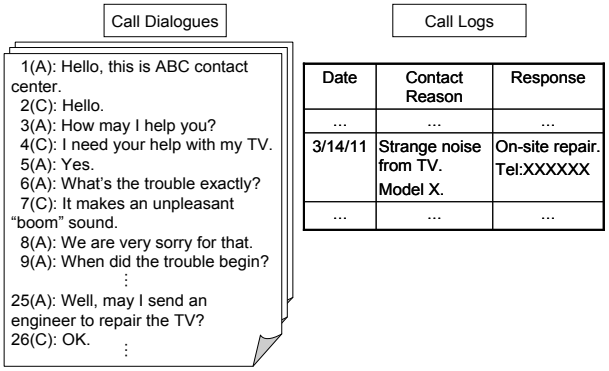
Abstract

This paper proposes a novel extractive summarization method for speech dialogues between agents and customers in contact centers. The proposed method does not require any extra cost for applying the method such as preparing rules or creating training data. Conventional methods such as the tf*idf method, which gives importance to characteristic words in an input text, can miss the essential points for contact center work. Our proposed method evaluates the importance of each utterance from the standpoint of call agents who report calls for managing or analyzing calls. Specifically, the proposed method includes information frequently reported by call agents in summaries using past call logs commonly recorded in the contact center. Evaluation using real data (call dialogues and call logs) shows that the proposed method can extract essential points in terms of contact center work and outperforms the conventional method.

1 Introduction

In recent years, the role of contact centers has become more important in many companies. This is because each contact center is a main channel for individual customers to directly access a company and the voice of the customers can be used to improve the quality of the company's products and services.

Many contact centers have two problems of human cost. The first one is that agents spend much time documenting logs. Logs are usually generated by agents based on their memories or handwritten memos after a call. Such logs are used for reporting, managing, and analyzing their calls. Figure 1 shows an example of a call dialogue and its corresponding call log (We simply refer to these as a dialogue and a log hereafter).



※leftmost number: utterance ID, A: agent, C: customer

Figure 1: Example of a call dialogue and a call log

The other cost is that managers also spend much time understanding the details of calls. This is because managers must listen to speech data or browse dialogue texts generated by automatic speech recognition (ASR), which are lengthy and include many uninformative parts.

Accordingly, automatic summarization of dialogues is required as an effective solution to the above problems. For the first problem, the summary can be an alternative to the logs or draft for agents. Byrd et al. (2008) have shown that automatic summarization helps to reduce the time for documenting calls. For the second one, the summary can help to reduce the time spent listening to the speech data or browsing the dialogue texts.

In previous works (Zechner, 1996; Edmundson, 2004; Orasan et al., 2004; Murray et al., 2007; Higashinaka et al., 2010), the methods summarize an input text without considering the

*This work was done when the author was at NEC Corporation.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 500–508,
Chiang Mai, Thailand, November 8 – 13, 2011.

requirements for contact center work such as reporting, managing, and analyzing calls. Some other works (Hirao et al., 2002; Iwasaki et al., 2005; Murray et al., 2005; Shen et al., 2007; Byrd et al., 2008; Fujii et al., 2008) can generate summaries satisfying such requirements. However, these methods generally require manual work of creating rules or training data.

In this paper, we propose a novel method that can summarize dialogues satisfying the requirements for contact center work without incurring the cost of manual works. The proposed method preferentially extracts sentences (or utterances) consisting of phrases described in logs of past calls. Such logs have been recorded through daily operation and are readily available in most contact centers. Moreover, we propose a method that bridges the gaps between the expressions in dialogues and logs to improve performance.

The main contributions of this paper are as follows.

1. We propose a method that preferentially extracts sentences consisting of phrases frequently written in logs of past calls. In our experiment using real data, we confirm that the proposed method outperforms the conventional methods ($\mathrm{tf} * \mathrm{idf}$ method and ridf method), baseline methods without considering the requirements for contact center work.
2. We propose a method that extracts sentences based on association strength with contents of the past logs so as to bridge gaps between the expressions in dialogues and logs, and confirm experimentally its effectiveness in summarization to improve performance.

This paper is organized as follows. In Section 2, we explain previous works and their problems. In Section 3, we propose our method. In Section 4, we describe the experiment using real contact center data. In Section 5, we discuss the experimental results. In Section 6, we summarize this paper.

2 Related Work

We describe conventional summarization methods for contact center dialogues and their problems. To reduce the burden of agents, Iwasaki et al. (2005) have proposed a method that generates a log of the input dialogue automatically. The method selects important sentences using different approaches according to sentence type (e.g. "contact reason", "response"), where the $\mathrm{tf} * \mathrm{idf}$ method or lead method is used. However, the method requires training data to identify the sentence type of each sentence. Hence, it takes effort to prepare the training data.

Byrd et al. (2008) built a system, Contact-Center Agent Buddies, which generates a candidate log from dialogues, and demonstrated its effectiveness in an actual contact center environment. The system normalizes the dialogue text generated by ASR, calculates the importance of each normalized sentence using a number of heuristic rules, and then extracts important sentences. However, some of the heuristic rules depend on the individual contact center (e.g. rules for annotating *cues* typically found in questions by agents). Hence, it takes effort to create rules for each contact center.

On the other hand, there are methods that do not require preparing training data or creating rules. Higashinaka et al. (2010) have proposed an extractive summarization method for contact center dialogues categorized into multiple domains (e.g. finance, mail order, PC support). The method extracts utterances that are characteristic of the input dialogue domain in relation to other domains using a particular type of hidden Markov model. However, the method cannot be applied to contact centers that do not deal with multiple domains. Additionally, the method seems to have difficulty extracting important information occurring in any domain (e.g., "a customer urgently requests support").

Other notable methods are the lead method proposed by Edmundson (1969) and the $\mathrm{tf} * \mathrm{idf}$ method proposed by Zechner (1996). The lead method extracts sentences from the top in order under the assumption that the important points are described first. However, important parts such as the customer's requirements and agent's responses can be located anywhere because the customer's requirements are identified through conversation interactions, which differ according to customers. Therefore, the lead method is not suitable for summarizing contact center dialogues.

The $\mathrm{tf} * \mathrm{idf}$ method uses word frequencies. First, the method calculates the following weight for each word w in the input text:

$$\mathrm{tf} * \mathrm{idf}(\mathrm{w}) = \mathrm{tf}(\mathrm{w}) \times \log\left(\mathrm{N}_{\mathrm{all}}/\mathrm{N}(\mathrm{w})\right),$$

where $\mathrm{tf}(\mathrm{w})$ is the frequency of w in the input, $\mathrm{N}_{\mathrm{all}}$ is the total number of texts, and $\mathrm{N}(\mathrm{w})$ is the number of texts containing w. Next, the method calculates the average of the weights for words in each sentence as importance of the sentence. Finally, the method extracts sentences in the order

of the importance from the highest. The method extracts utterances including characteristic words that are frequent in the input text and infrequent in other texts. However, there are essential words for contact center work, although the words are frequent in other texts (e.g. words in frequent customer requests). The method does not include the words in summaries. Moreover, there are uninformative sentences including characteristic words such as the customer's speaking habits (e.g. "kind of" being frequently used by the customer). The method can extract these uninformative sentences.

Orasan et al. (2004) and Murray et al. (2007) experimentally showed that ridf(w) defined as the following function is most effective for summarizing texts in several term weighting functions including the $\mathrm{tf} * \mathrm{idf}(w)$.

$\mathrm{ridf}(w) = \log(N_{all}/N(w)) - \log(1 - p(0;\lambda_w))$,

where N_{all} and $N(w)$ have the same definitions as those in the $\mathrm{tf} * \mathrm{idf}(w)$, and p is the Poisson distribution with parameter λ_w, the average number of occurrence of w per text, and $1 - p(0;\lambda_w)$ is the probability of w appearing in a text at least once. However, the method using the ridf(w) also cannot generate summaries from the viewpoint of contact center work.

3 Proposed Method

We propose a novel summarization method for contact center dialogues without the problems described in Section 2. In this work, we assume that essential information for contact center work should be frequently written in past logs. First, the proposed method calculates the likelihood of being reported by agents for each utterance using past logs recorded in the contact center. Hereafter, we refer to the likelihood as *report score*. Next, the proposed method extracts utterances in the order of the report score from the highest and then outputs the extracted utterances in the order of appearance in the input dialogue as its summary.

The proposed method does not require any extra cost for applying the method because the method exploits past logs automatically accumulated in the contact center through daily operation. Additionally, the proposed method includes the essential points in terms of contact center work in summaries because the logs are described to report, manage, and analyze the contacts. Note that there is no corresponding log of an input dialogue when the method summarizes the dialogue.

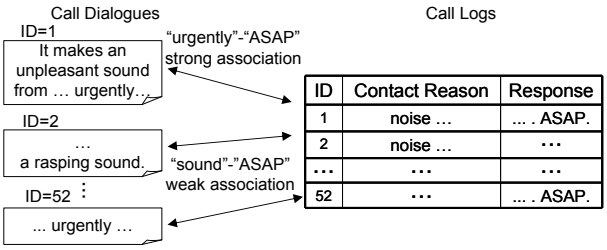


Figure 2: Association strength between words

The rest of this section is organized as follows. In Section 3.1, we first describe a method that simply calculates the report score based on frequency in past logs, and explain the problem with the simple method. In Section 3.2, we propose a method that calculates the report score based on past pairs of a dialogue and its corresponding log so as to handle the problem with the simple method.

3.1 Call Log Frequency Method (LF Method)

The LF method regards high frequency words in past logs as important, and then extracts utterances that have the high frequency words. The LF method proceeds as follows.

Step 1. $\mathrm{R\_Score}(w)$, the report score of a word w, for each word in an input dialogue is calculated by the following equation (1). Here, equation (1) is the document frequency of the word w, where a set of documents is a set of past logs (Call Log Frequency; LF).

$$\mathrm{R\_Score}(w) = L(w)/L_{all} \cdots (1).$$

Let $L(w)$ be the number of past logs containing the word w, and let L_{all} be the total number of past logs.

Step 2. $\mathrm{R\_Score}(U)$, the report score of an utterance U, for each utterance in the input dialogue is calculated as the average[1] of the report scores of the words in the utterance U as follows:

$$\mathrm{R\_Score}(U) = \frac{\sum_{w \in W(U)} \mathrm{R\_Score}(w)}{|W(U)|} \cdots (2).$$

Let $W(U)$ be a set of all words in an utterance U.

Step 3. Top-K utterances in the order of the report score from the highest are extracted.

3.1.1 Problem with LF Method

In logs, concise expressions, abbreviations, and specialized terminology tend to be frequently used because the logs are generated so as to report to managers or persons involved. On the

[1] We regard sentences that express content compactly as important to summarization. Hence, we do not use the total but the average.

other hand, an agent usually uses multiple expressions in a dialogue according to the situation or customer's level of understanding. Therefore, there is a gap between the expressions in the dialogue and in its corresponding log even though those are the same in meaning.

In Figure 1, "an unpleasant 'boom' sound" in the dialogue is concisely described as "strange noise" in its corresponding log, and in Figure 2, "urgently" in a dialogue is described with the abbreviation "ASAP" in the log. However, the LF method cannot handle the differences between the expressions in dialogues and logs. Concretely, the LF method cannot select important information that is not described with the same expression in the logs.[2] Consider Figure 1, for example; the LF method cannot regard as important the words not appearing in the log such as "unpleasant", "boom" and "sound". Therefore, the LF method may not include ID 7, which is important for contact center work, in the summary.

3.2 Proposed Method

In Section 3.2.1, we propose a method that can handle the problem described in Section 3.1.1. In Section 3.2.2, we introduce a component that removes frequent utterances from the summary to improve performance of our method.

3.2.1 Association Strength Method (AS Method)

To handle differences between the expressions in dialogues and logs, we have based our proposed method on the following assumption: association strength (AS) between each occurrence of two words in past pairs of a dialogue and its corresponding log indicates semantic similarity between the two words.

Consider, for example, Figure 2. "Urgently" in dialogues and "ASAP" in logs have the same meaning. When "urgently" appears in a dialogue, "ASAP" also appears in its corresponding log (ID=1,52). Moreover, when "urgently" does not appear in a dialogue, "ASAP" also does not appear in its corresponding log (ID=2). In short, Figure 2 shows that association strength between "urgently" and "ASAP" is strong. On the other hand, "sound" and "ASAP" have different meanings. Figure 2 shows that the association strength between "sound" and "ASAP" is weak.

[2] Not only the LF method but any supervised methods using logs as training data have the same problem.

Under the above assumption, we propose the AS method, which estimates semantic similarity between a word w in dialogues and a word v in logs as $\mathrm{AS}(w, v)$, the association strength between each occurrence of w and v. $\mathrm{AS}(w, v)$ is calculated by association measures such as mutual information, chi-square value, and z-value.

We calculate the likelihood that a word w in a dialogue is reported in past logs as a word v, by the following expression (3). The expression is the product of $\mathrm{AS}(w, v)$, the semantic similarity between w and v, and the likelihood that the word v is reported in past logs.

$$\mathrm{AS}(w, v) \cdot \mathrm{L}(v)/\mathrm{L}_{\mathrm{all}} \cdots (3).$$

$\mathrm{L}(v)$ and $\mathrm{L}_{\mathrm{all}}$ in the expression (3) have the same definitions as $\mathrm{L}(w)$ and $\mathrm{L}_{\mathrm{all}}$ in equation (1) respectively.

Some of the words in dialogues have multiple synonymous expressions in logs. For example, "noise" in dialogues can sometimes be described as "noise" and other times as "sound" in logs. Additionally, the meaning of a word in a dialogue is often expressed with multiple words in its corresponding log. For example, the meaning of "boom" in the dialogue in Figure 1 is expressed with "strange noise" in its log. To deal with the above paraphrases, we expand expression (3). Specifically, the likelihood that the meaning of a word w in a dialogue is reported in past logs as a set of words V is calculated as the sum of expression (3) for each word v in V as follows:

$$\sum_{v \in V} \mathrm{AS}(w, v) \cdot \mathrm{L}(v)/\mathrm{L}_{\mathrm{all}} \cdots (4).$$

In conclusion, $\mathrm{R\_Score}(w)$, the report score of a word w, is estimated by expression (4), where V is the set of all words in past logs. Here, the amount of calculation is enormous. Therefore, we limit V in expression (4) to N words in order of $\mathrm{AS}(w, v)$ from the highest. We denote the set of such N words for a word w by $\mathrm{V_N}(w)$. The AS method estimates $\mathrm{R\_Score}(w)$ by the following equation (5).

$$\mathrm{R}_{\mathrm{Score}(w)} = \sum_{v \in \mathrm{V_N}(w)} \mathrm{AS}(w, v) \cdot \frac{\mathrm{L}(v)}{\mathrm{L}_{\mathrm{all}}} \cdots (5).$$

Here, association strength between unrelated words should be close to zero, and the number of association words for one word is limited. Accordingly, equation (5) is not sensitive to larger N, and we assume that the AS method is effective with little or no tuning of N. We examine this point in Section 4. Hereafter, the AS method performs steps 2 and 3 in Section 3.1 in sequence.

Table 1: Performance of ASR

Speaker	Agent	Customer
P.C. (%)	91.5	89.9
W.A. (%)	87.7	84.8

Table 2: Test data

Data type	SR	MT	Log
Avg. number of utterances	175	111	-
Avg. number of characters	1,207	1,320	166

3.2.2 Removal of Frequent Utterances (RFU)

Contact center dialogues include many uninformative utterances that do not have important contents such as back-channel feedback (e.g. "Yes", "Well"), set phrases (e.g. "This is XX contact center."), and greetings (e.g. "Hello", "Good morning"). These uninformative utterances must not be included in a summary. We aim to improve performance of our method by directly detecting these uninformative utterances.

We assume that such uninformative utterances frequently occur in any dialogue. Hence, the proposed method calculates the occurrence rate for each utterance U defined as "the number of dialogues containing utterance U / total number of dialogues", and then identifies the utterances whose occurrence rates are higher than threshold θ as uninformative utterances. Finally, the proposed method removes the identified utterances from a summary.

4 Experiment

In this section, we describe the experiments using dialogues and logs in a real Japanese contact center and show their results, where we examine the following effects of our proposal.

- By preferentially including information frequently reported in past logs, the performance of automatic text summarizers for contact center works can be improved.
- Association strength between two words enables our method to handle differences between the expressions in dialogues and logs, and improves performance.
- RFU improves performance of our method.

4.1 Experimental Settings

4.1.1 Experimental data

We collected 4,596 call speech data and their corresponding logs in a real Japanese contact center, and generated the following two types of texts from the call speech data. We used the texts as dialogue data in the experiments.

1. Speech Recognition Text (SR):

The texts were generated by ASR from the read speech data. Table 1 shows the accuracy of ASR. In Table 1, P.C. is percent correct calculated by (T-S-D)/T, and W.A. is word accuracy calculated by (T-S-D-I)/T. Let T be the total number of words, S be the number of substitutions, I be the number of insertions, and D be the number of deletions.

2. Manual Transcription Text (MT):

The texts were transcribed manually from the speech data and divided manually into utterances.

4.1.2 Test Data

We used 40 pairs of dialogue data and their corresponding logs as test data, which were selected randomly from a total of 4,596 data. The average length of the dialogue data and the logs are shown in Table 2, which shows that the log corresponds to 12.6% (=166/1,320) of compressed text of the call speech data.

We used the following two types of summaries with different compression rates as the references in the experiments so as to examine the effectiveness for various types of contact center work such as reporting, managing, and analyzing contacts. The references were manually generated from the MT of the test data. Note that logs themselves are not suitable for references because there are differences between the expressions in dialogues and logs.

1. Indicative Summary:

We generated summaries with a 30% compression rate by manually extracting utterances on the assumption that these are used as alternatives to the logs or drafts for agents, and for managers to grasp the gist of calls at a glance. Here, the compression rate is defined as "number of characters in a summary / number of characters in a dialogue data". Note that we set the compression rate to 30%, which is higher than that of the logs (12.6%), because the expressions in dialogues tend to be lengthy compared to those in logs.

2. Informative Summary:

We generated summaries that are sufficient to obtain all contents by manually extracting utterances without thinking of compression rate on the assumption that these are used for managers to grasp the details of calls, and as cleansed texts for analysis of calls such as information retrieval and text mining. The average compression rate is 65.2%.

Table 3: Experiments with different N in equation (5)

N	1	2	4	8	16
F-measure	0.526	0.528	0.527	0.529	0.519
ROUGE-1	0.565	0.572	0.571	0.576	0.567
ROUGE-2	0.567	0.574	0.578	0.584	0.575
N	32	64	128	LF Method	
F-measure	0.504	0.497	0.494	0.478	
ROUGE-1	0.563	0.549	0.546	0.527	
ROUGE-2	0.568	0.550	0.548	0.543	

4.1.3 Competing Methods

We evaluate the following five methods in the experiments. For MT, each method judges each utterance manually detected as to whether it is necessary for the summary or not, and for SR, each method judges each utterance detected by the ASR engine.

1. $\mathrm{tf} * \mathrm{idf}$ method and ridf method:

The methods are described in Section 2. Using a total of 4,556 dialogue data, the $\mathrm{tf} * \mathrm{idf}$ method calculated the $\mathrm{tf} * \mathrm{idf}$ score for each word and the ridf method calculated the ridf score.

2. LF method:

The method is described in Section 3.1. In calculating the report score for each word, the method used a total of 4,556 logs.

3. AS method:

The method is described in Section 3.2.1. Note that the method does not introduce RFU. In calculating the report score for each word, the method used a total of 4,556 pairs of the dialogue data and its corresponding logs, and used z-value as the association measure. We used various numbers as N in equation (5) to investigate the dependency of N on the performance of the AS method.

4. AS with RFU method:

The method introduces RFU into the AS method. The settings in the AS method are the same as the above. In RFU, we used 0.5 as a threshold, which was determined by the preliminary experiment. Additionally, RFU calculated the occurrence rate using a total of 4,556 dialogue data. Here, RFU regards an utterance as the bag of content-word bigrams so as to relieve differences of the expressions between utterances (e.g. "This is XX speaking." and "This is XX."). RFU calculates the average of the occurrence rate of the bigrams in the utterance as the occurrence rate of the utterance.

4.1.4 Evaluation Measure

We used the following evaluation measures.

1. Sentence recall, precision, and F-measure:

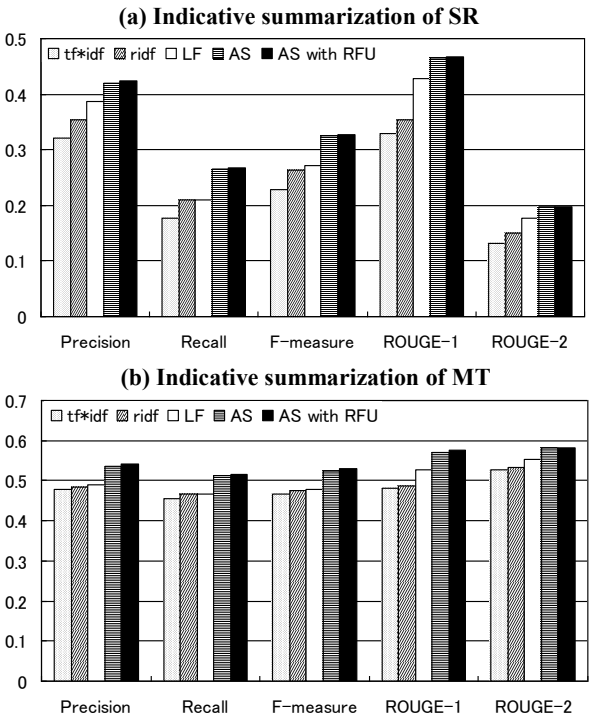


Fig 3: Result of indicative summarization

Sentence recall is the number of sentences correctly extracted by a method over the number of sentences in the reference. Sentence precision is the number of sentences correctly extracted over the number of sentences extracted by the method. F-measure is defined as follows;

$$2 \times \mathrm{Recall} \times \mathrm{Precision} / (\mathrm{Recall} + \mathrm{Precision}).$$

When SR is summarized, estimated sentence (or utterance) boundaries based on ASR results do not always agree with those in the references. In this paper, extraction of a sentence in the SR is considered as extraction of one or multiple sentences in the reference with an overlap of 50% or more words as in Hirohata et al. (2005).

2. $\mathrm{ROUGE_N}$ (Lin et al., 2003):

$\mathrm{ROUGE_N}$ is an N-gram recall between reference (R) and the summary generated by a method (C), which indicates similarity between them. $\mathrm{ROUGE_N}$ is calculated as follows:

$$\mathrm{ROUGE_N}(\mathrm{C}, \mathrm{R}) = \frac{\mathrm{C_m}(\mathrm{N\text{-}gram})(\mathrm{C}, \mathrm{R})}{\mathrm{C}(\mathrm{N\text{-}gram} \in \mathrm{R})},$$

where $\mathrm{C_m}(\mathrm{N\text{-}gram})(\mathrm{C}, \mathrm{R})$ is the number of co-occurrences of N-gram in C and R, and $\mathrm{C}(\mathrm{N\text{-}gram} \in \mathrm{R})$ is the number of N-grams in R. In our experiments, 1-grams (ROUGE-1) and 2-grams (ROUGE-2) are used, where the words are only content words.

4.2 Experimental Results

First, we investigated the dependency of the number of association strength in estimating the

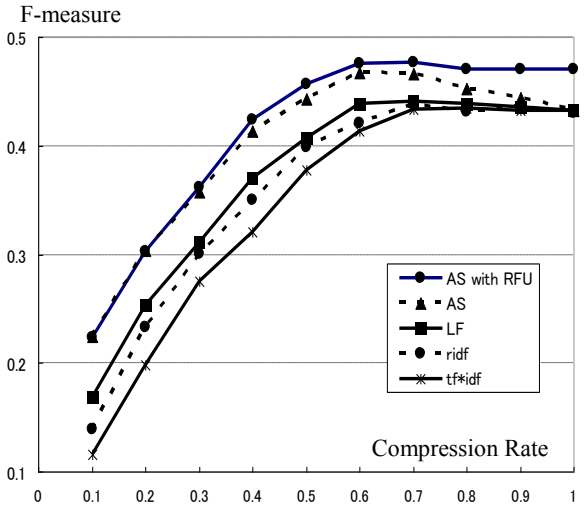


Fig 4: F-measure of informative summarization

report score (N in equation (5)) on the performance of the AS method. Table 3 shows the performance of indicative summarization with different N for MT. Table 3 shows that there is not much difference between N=64 and N=128. The result indicates that the number of related words for one word is limited and association strength between unrelated words is close to zero. Accordingly, the performance of the AS method is not sensitive to larger N. Table 3 also shows that the performance with N = 8 is best. Hereafter, let N be 8 with regard to the AS method.

Next, we examined the performance of indicative summarization for each competing method. Specifically, we summarized the dialogues in the test data by each method at a 30% compression rate, and then evaluated their results. Figure 3 shows the results. Moreover, we examined the performance of informative summarization for each competing method. Unfortunately, we cannot get the compression rate suitable for the informative summarization in each competing method. In the examination, we summarized the dialogues in the test data by each method, where the compression rate was changed from 0.1 to 1.0 at 0.1 intervals, and then evaluated their results. Figure 4 shows the F-measure for SR. Other evaluation measures and the results for MT are omitted in this paper because the results indicate a similar tendency to Figure 4, and the conclusions in this paper do not change.

Figures 3 and 4 show that the LF method, the AS method, and the AS with RFU method outperform the tf ∗ idf method and the ridf method. The results show that preferentially selecting information frequently reported in past logs is effective for summarizing dialogues from the viewpoint of contact center work regardless of compression rate.

Figures 3 and 4 also show that the AS method outperforms the LF method, and Table 3 shows that the performance of the AS method with N = 128 is better than that of the LF method. These results show that using association strength between two words in the past pairs of a dialogue and its corresponding log improves the performance of our method. This means the association measure helps to handle differences between the expressions in dialogues and logs. We discuss the point in Section 5.1 in detail. Additionally, Table 3 shows that the AS method is effective with little or no tuning of N.

Figure 4 shows that the AS with RFU method outperforms the AS method when the compression rate is high. The result shows that RFU is effective for summarization with a high compression rate. However, Figures 3 and 4 also show that when compression rate is low, there is not much difference between performance of the AS method and that of the AS with RFU method. We discuss this point in Section 5.2 in detail.

5 Discussion

5.1 Effectiveness of Association Strength between Two Words

We discuss whether the assumption described in Section 3.2.1, that the association strength between two words indicates semantic similarity, is correct or not. We examined relationships of the 500 pairs of two words with strong association in the experimental data. The result is that 67.8% are synonymous, 15.2% are *related*, and 17% are *unrelated*. Here, *related* is when the two words are associated with one another including is-a relations and part-of relations (e.g. "freeze" and "break", "ATM" and "cash"). *Unrelated* is when the two words are irrelevant to one another. The result shows that accuracy of identifying two words with the same meaning using the word pairs with high association strength is 67.8%.

Table 3 shows that the performance with N = 8 is best although the number of synonymous expressions for one word is supposed to be smaller than 8. We suppose that *related* helps to find utterances reported in the logs and 83% (total of *synonymous* and *related*) strong association is an efficient clue for summarization.

Most *unrelated* words belong to different types of information (e.g. "CUSTOMER X" and "MODEL Y", which should be respectively described in the "customer's name" and "contact

Table 4: Rate of eliminated utterances by RFU

Data Type	Compression Rate	Rate of eliminated utterances by RFU
SR	30%	1.9% (23/1,213)
	50%	2.4% (51/2,098)
	65%	9.3% (269/2,879)
	80%	35.9% (1,500/4,182)
MT	30%	0.6% (4/715)
	50%	2.3% (27/1,168)
	65%	8.0% (129/1,617)
	80%	53.8% (1,239/2,305)

reason" parts of the logs). By calculating the association measure of two words in corresponding parts of a dialogue and its log after topic segmentation of the dialogue, we can further improve our method.

5.2 Effectiveness of RFU

We examined whether the frequent utterances identified by RFU in the test set are unnecessary for the summaries (uninformative) or not. In a total of 6,985 utterances in SR, 2,102 utterances (161 varieties) are identified by RFU, and 99.4% (=2,090/2,102) are uninformative. In a total of 4,436 utterances in MT, 1,985 utterances (156 varieties) are identified by RFU, and 99.7% (=1,979/1,985) are uninformative. The results show that RFU can eliminate uninformative utterances from the summaries with high accuracy.

Additionally, we examined the rate of eliminated utterances by introducing RFU into the AS method, which is the number of frequent utterances in the summary generated by the AS method over the total number of utterances in the summary. Table 4 shows the result. Table 4 shows that there are few frequent utterances in the summary generated by the AS method when the compression rate is low. In other words, the result indicates that the AS method (without RFU) can eliminate uninformative frequent utterances when generating an indicative summary or informative summary with a low compression rate. As a result, there is not much difference between the performance of the AS method and that of the AS with RFU method in Figures 3 and 4 with a low compression rate.

On the other hand, Table 4 shows that there are a lot of frequent utterances in the summary generated by the AS method when the compression rate is high. The results indicate that when the compression rate is high, it is difficult to judge whether an utterance is important or not using only the report score based on the association strength. This is because the input dialogue can include detailed information not described in the logs, and also subjects not occurring in past calls. In the above situation, suitable utterances can be included in a summary by eliminating frequent utterances preferentially. As a result, the AS with RFU method enables maintaining the quality of summarization with a high compression rate.

5.3 Robustness to ASR errors

Figure 3 shows that performance on summarization of SR is lower than that of MT in every method. This is because the words that should be ideally included in the summary are missing in SR due to substitutions or deletions in ASR.

To examine the robustness to ASR errors, we calculated the reduction rate[3] of F-measure by comparing the performance to SR with that to MT. As a result, the reduction rate of the AS method is 37.9% and the rate of the LF method is 43.1%. The result shows that the AS method is more robust to ASR errors than the LF method.

6 Conclusions

We proposed a novel method that can summarize contact center dialogues satisfying the requirements for contact center work without any extra cost for applying the method. We proposed the idea of preferentially selecting information frequently reported in past logs so as to include the essential information for contact center work in summaries. Moreover, we proposed a method that extracts utterances based on association strength between each sentence and the past logs so as to bridge the gaps between the expressions in logs and dialogues.

In the evaluation using real data, experimental results showed that our proposed method outperforms the conventional methods (the $\mathrm{tf} * \mathrm{idf}$ method and the ridf method), and association strength between two words improves the performance of automatic text summarizers for contact center works. Additionally, we improved our method by removing frequent utterances from the summaries.

We are planning to prove the effectiveness of our proposed method for actual contact center work according to the cost reduction effect in the call log documentation process.

[3] (F-measure to MT – F-measure to SR) / F-measure to MT

Reference

Roy J. Byrd, Mary S. Neff, Wilfried Teiken, Youngja Park, Keh-Shin F. Cheng, Stephen C. Gates, and Karthik Visweswariah. 2008. Semi-Automated Logging of Contact Center Telephone Calls. In *Proceedings of the 17th ACM International Conference on Information and Knowledge Management*, pages 133-142.

Harold P. Edmundson. New Methods in Automatic Extracting. 1969. Journal of ACM, pages 264-285.

Yasuhisa Fujii, Kazumasa Yamamoto, Norihide Kitaoka, and Seiichi Nakagawa. 2008. Class Lecture Summarization Taking into Account Consecutiveness of Important Sentences. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, pages 2438-2441.

Ryuichiro Higashinaka, Yasuhiro Minami, Hitoshi Nishikawa, Kohji Dohsaka, Toyomi Meguro, Satoshi Takahashi, and Genichiro Kikui. 2010. Learning to Model Domain-Specific Utterance Sequences for Extractive Summarization of Contact Center Dialogues. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 400-408.

Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda, and Yuji Matsumoto. 2002. Extracting Important Sentences with Support Vector Machines. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 342-348.

Makoto Hirohata, Yousuke Shinnaka, Koji Iwano, and Sadaoki Furui. 2005. Sentence Extraction-Based Presentation Summarization Techniques and Evaluation Metrics. In *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing*, pages 1065-1068.

Reijirou Iwasaki and Kenji Araki. 2005. Important Sentence Extraction Method for Automatic Generation of Business Days Report for Conversation Data of Call Center. In *Proceedings of 19th the Annual Conference of The Japanese Society for Artificial Intelligence*, 1E1-102. [in Japanese].

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics, In *Proceedings of the 4th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technology,* pages 71-78.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive Summarization of Meeting Recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 593-596.

Gabriel Murray and Steve Renals. 2007. Term-Weighting for Summarization of Multi-Party Spoken Dialogues. In A. Popescu-Belis, S. Renals, and H. Bourlard, editors, *Machine Learning for Multimodal Interaction IV, volume 4892 of Lecture Notes in Computer Science,* pages 155-166.

Constantin Orasan, Viktor Pekar, and Laura Hasler. 2004. A comparison of summarisation methods based on term specificity estimation . In *Proceedings of the 4th international conference on Language Resources and Evaluation*, pages 1037-1040.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zhen Chen. 2007. Document Summarization using Conditional Random Fields. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2862-2867.

Klaus Zechner. 1996. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 986-989.