2014 IIAI 3rd International Conference on Advanced Applied Informatics

Multi-party conversation summarization based on sentence selection using verbal and nonverbal information

Yo Tokunaga
Department of Artificial Intelligence,
Kyushu Institute of Technology,
680-4 Kawazu Iizuka Fukuoka 820-8502 Japan
Email: y_tokunaga@pluto.ai.kyutech.ac.jp

Kazutaka Shimada
Department of Artificial Intelligence,
Kyushu Institute of Technology,
680-4 Kawazu Iizuka Fukuoka 820-8502 Japan
Email: shimada@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 (SVMs). First we compare important utterance extraction by the scoring method and SVMs. In the machine learning technique, we introduce verbal features, such as relations between utterances and anaphora features, and nonverbal features. Next we generate a summary from the outputs of the scoring method and SVMs. In our approach, a basic summary consists of utterances with high confidence extracted from the scoring method. Utterances from SVMs are used as supplementary information. In the experiment, we compare a combination method and a method with only SVMs. The output of our method was suitable in terms of readability and correctness as a summary of original conversation.**

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. The summarization has an important role to understand the content of a conversation easily.

In this paper, we propose a method for conversation summarization. Traditional summarization studies have handled a single document or multi-documents as the target [8]. Many studies in the summarization are based on extraction approaches [4], [14]. 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 a multi-party conversation. Shen et al. [9] have proposed a summarization method using conditional random fields to handle a relation between utterances. For conversation summarization, relations between utterances are more important, as compared with document summarization such as news papers. Higashinaka et al. [6] 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. [13] 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. There are many methods with Rhetorical Structure Theory (RTS) for document summarization tasks [3], [7]. Such discourse information is generally effective for summarization. However, the task that we handle in this paper is more complicated; many sub topics and many noise utterances, such as nods, in a conversation. Therefore, the discourse information, RST, is not always suitable for our task. Moreover, free conversations contain many nonverbal information. We focus on hot spots and laughing in conversations for the summarization 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. We call it " a basic summary". However, the number of utterances in the basic summary is not enough as a final summary. To solve this problem, we incorporate utterances extracted by SVMs to the basic summary. Furthermore, we add some utterances which is located near utterances in the basic summary to the combination summary from the scoring and SVMs. In the experiment, we compare the proposed method with a baseline and a method based on only SVMs.

II. Corpus

In this paper, we construct a conversation corpus consisting of 8 spontaneous conversations with 1295 utterances[1]. The number of participants in each conversation is 4 persons. The participants had a free talk about "Movies", "Games", "SNS" and so on.

For the machine learning and evaluation, we need a tagged corpus with an importance degree of each utterance. Three annotators judged the importance degree of each utterance in a phased manner. We regard all utterances in each conversation as level-1. First, the annotator selected three quarters

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

978-1-4799-4173-5/14 $31.00 © 2014 IEEE
DOI 10.1109/IIAI-AAI.2014.99

of utterances from all utterances (level-2). 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). In other words, the annotator classified all utterances into four classes on the basis of the importance of each utterance. In this annotation process, the annotators paid attention to keep the meaning and context of the original conversation. 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 of the importance level (level-1 to 4) 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 [2] are approximately 0.3 and not high[2]. 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.

In addition, the tagged corpus needs to include nonverbal information for each utterance. The targets of nonverbal information in this paper are (1) hot spot value and (2) laughing information. The hot spot value denotes a degree of an excited situation in a conversation. Two annotators judged the degree of each utterance. The degree is rated on a scale from 1 (low) to 5 (high). For laughing, we detect laughing in each conversation manually first, and then classify the trigger of laughing into two classes; internal and external. Here the internal laughing denotes that a participant laughs at his/her own utterance or laughs out of embarrassment. The external laughing denotes that a participant laughs at other participant's utterance or behavior.

III. 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. This result is a basic summary for the summarization process.

The second method is a machine learning technique. We use Support Vector Machines (SVM) [11]. 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 final summary. We add the output with high confidence from SVMs to the basic summary. We call it "a combination summary".

In the final process, we add an utterance between important utterances by the scoring method to the combination summary to keep the coherency of the final summary. By using this process, the readability of each generated summary is drastically improved.

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

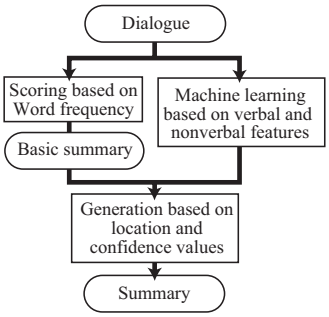

Fig. 1. The outline of our method. It combines a scoring method and machine learning to generate a summary.

B. Scoring

We compute a score of each utterance in a conversation. We apply the panoramic view system proposed by [10] 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[3]. In this paper, we handle nouns[4], 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.

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

[4]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 $key3$.

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

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

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

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

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

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.

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

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[5].

C. Machine learning

One approach to extract important information is to utilize machine learning techniques. We apply SVMs to this summarization process. The features for SVMs are classified as (f1) features in an utterance, (f2) features between utterances, (f3) features about anaphora, (f4) features based on scores and (f5) nonverbal features.

(f1) 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.
- 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.

(f2) 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[6].
- 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.

(f3) Features about anaphora
Anaphoric relation is one of the most considerable points in conversation summarization tasks. 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)”.

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

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

- 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)".

(f4) Features based on scores
The fourth feature category consists of three features based on the scores in Section III-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.

(f5) Nonverbal features
Nonverbal information has an important role of the utterance extraction for conversation summarization. Here we focus on two nonverbal information; hot spots and laughing points in each conversation. The hot spot value, which is annotated by two annotators, denotes a degree of an excited situation in a conversation. The internal laughing and external laughing denote that a participant laughs at his/her own utterance or laughs out of embarrassment and a participant laughs at other participant's utterance or behavior, respectively.
- Hot spot value: the annotated value (1-5)
- Presence of laughing: 0 or 1
- Timing of laughing: we classify the timing into some classes, such as laughing in an utterance and an utterance in laughing.
- Trigger of laughing: internal or external
- Distance from external laughing: we compute the distance between an utterance and the nearest external laughing.

D. Summary generation

In our method, we regard the extracted utterances by the scoring process as a basic summary. However, the basic summary is not enough in terms of the size because they are just the top 10 % score of all. In other words, the recall rate of the scoring method is extremely low as a summary. Therefore we add the outputs from SVMs to the basic summary. Our method selects one utterance with the high output score of SVMs between utterances in the basic summary.

The one of the most important points for conversation summarization is the coherency of a generated summary. The coherency of the combination of a basic summary and outputs of SVMs is often insufficient because the selection approaches of the scoring method and SVMs are independent and do not always play a complementary role. The most intuitive solution is to add the previous and next utterances of each important utterance to the summary. It might lead to the improvement of the coherency. However, the method is too naive and insufficient because there is a possibility that previous and next utterances often are noise utterances as a summary.

Our method adds an utterance between important utterances by the scoring method into the combination summary. The process is as follows:

Utterance ID	Score of SVM
Utterance 1	0.15
Utterance 2	0.82 << selected by SVM
Utterance 3	0.19
Utterance 4	0.75 << selected by Step 2
Utterance 5	0.45
Utterance 6	<< selected by Score based
Utterance 7	0.75 << selected by Step 1
Utterance 8	0.45
Utterance 9	0.85 << selected by SVM
Utterance 10	0.27
Utterance 11	0.55 << selected by Step 2
Utterance 12	0.44
Utterance 13	0.32
Utterance 14	<< selected by Score based
.....	.....

Fig. 2. An example of the summary generation. Our method fleshes a basic summary from the scoring method out with some more utterances by the output of SVMs, Step 1 and Step 2 in Section III-D as the final summary.

Step 1 : we focus on the previous and next utterances of each important utterance in a basic summary. If the score from SVMs of the utterance is more than a threshold value, we select the utterance for the summary. If the score of the utterance is less than the threshold value, we proceed to the next step.

Step 2 : we select the utterance with the second high score of SVMs between utterances by the scoring method[7].

Here we set 0.7 as the threshold.

Figure 2 shows an example of the summary generation. In the example, assume that the utterance 6 and 14 are extracted by the scoring method. They are a basic summary. Then, the utterances with the highest value from SVMs, namely 2 and 9, between the utterances in the basic summary are added. They are a combination summary. In the Step 1, the utterance 7 is selected on the basis of the threshold. Finally, the utterance 4 and 11 are selected in the Step 2. They are the final summary of our method.

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. The precision rate of the scoring was 0.959 although the recall rate was 0.190. The purpose of the scoring method is to construct the basic summary for the summarization process. Therefore, the scoring method with the high precision rate was suitable as the first step of the summarization process.

Next, we evaluated our method based on SVMs with 10-fold cross validation. We used the data mining tool WEKA

[7]Note that the utterance with the first high score is already selected in the combination summary.

TABLE I. RESULT OF UTTERANCE EXTRACTION BY SVMS

Feature	Precision	Recall	F
ALL	0.801	0.740	0.769
ALL−(f1)	0.777	0.677	0.724
ALL−(f2)	0.790	0.719	0.753
ALL−(f3)	0.803	0.722	0.760
ALL−(f4)	**0.804**	**0.752**	**0.777**
ALL−(f5)	**0.804**	0.740	0.771

[5] for the implementation. In the experiment, we compared the effectiveness of each feature category; (f1) features in an utterance, (f2) features between utterances, (f3) features about anaphora, (f4) features based on scores and (f5) nonverbal features.

Table I shows the experimental result. In the table, "ALL−(f1)" denotes the method without the feature set (f1). In other words, the features of the method consists of the feature (f2), (f3), (f4) and (f5). The method without the feature (f4) produced the best performance. The most effective feature category was (f1) features in an utterance because the accuracy decreased in the case that the feature set was removed. In particular, the length feature was effective for the accuracy. The nonverbal feature (f5) did not work well. The average hot spot values of important utterances and non-important utterances differed only slightly (3.2 and 3.0). The laughing occurred in a mere 25% of all utterances.

In this paper, we focused on surface linguistic features and nonverbal features, namely hot spot and laughing, for the method. However, conversations contain many characteristics, such as prosodic features [12]. In addition, conversations contain discourse information such as dialogue acts [1]. These characteristics are useful for the summarization method. Incorporating them to our method is the important future work.

B. Evaluation of summary

Next, we evaluated the outputs of the summarization process. We compared the proposed method with two method; a baseline and the top n-utterances from SVMs. The baseline was a naive approach. It added the previous and next utterance to the combination summary consisting of a basic summary from the scoring method and utterances with the high output score of SVMs between utterances in the basic summary. Therefore, the size of the summary is the same as the output of our method. The second method, the top n-utterances from SVMs, generated a summary from only the outputs of SVMs ($n\text{Utter}_{SVM}$). In other words, the method did not use the outputs from the scoring method. The value of n was the number of utterances in the summary by the proposed method.

Figure 3 shows an example of the output of our method. In the figure, the utterances with a rectangle are the output from the scoring method, that is a basic summary. The utterances with "**" are utterances with the high output score of SVMs between utterances in the basic summary. In other words, the combination summary consists of the utterances with a rectangle and the utterances with the "**" mark. The utterances with "++" are the utterances selected by the Step 1 and 2 in Section III-D. Therefore, the utterances with a rectangle, "**" and "++" were the output of the proposed method.

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.
.....

Fig. 3. An example of the output from the proposed method. The combination of utterances with a rectangle, "**" and "++" is the output of the proposed method.

TABLE II. THE EXTRACTION ACCURACY AS A SUMMARY.

Method	Precision	Recall	F
Proposed	0.732	0.509	0.600
Baseline	0.628	0.440	0.517
$n\text{Utter}_{SVM}$	**0.859**	**0.605**	**0.710**

Here, the baseline selected the utterance "Ah! Ah! I remember now." instead of "The passwords were the date of birth or something..." because of the next utterance of the output of the scoring method.

First, we evaluated the extraction accuracy as the summarization. Table II shows the precision, recall and F-values of each method. The baseline essentially contained noise utterances because it is a naive approach. Therefore the accuracy became low. The proposed method outperformed the baseline. This result shows the effectiveness of the sentence selection by the Step 1 and 2 in Section III-D. It's only natural that the $n\text{Utter}_{SVM}$ produced the best accuracy because the summary essentially tended to not include noise utterances as compared with the baseline and proposed methods[8].

The extraction accuracy of the proposed methods was lower than the $n\text{Utter}_{SVM}$. The reason was that the proposed method sacrificed the extraction accuracy to keep the coherency in the summary. The result shown in Table II is just the extraction accuracy. It might not always link to the correctness of the generated summary. On the other hand, the most important point of a summary is the readability and quality of the summary. Therefore, we evaluated "readability" and "correctness" of each output.

[8]Note that $n\text{Utter}_{SVM}$ did not select utterances for the coherency.

TABLE III. THE READABILITY AND CORRECTNESS OF EACH METHOD.

Method	readability	correctness
Proposed	**3.52 (0.79)**	**3.47 (0.91)**
Baseline	2.66 (0.99)	2.57 (0.90)
$n\text{Utter}_{SVM}$	3.09 (0.97)	3.04 (1.17)

- readability: Is the document readable?
- correctness: Does the document contain the meaning and correct context of the original document?

In the experiment, seven test subjects, who were not related to this research, evaluated the outputs from three methods. The score range from the test subjects was 1 (Bad) to 5 (Good) points. The evaluation way of the two criteria, "readability" and "correctness" , is as follows:

- First step: The test subjects received three summaries generated from each method simultaneously. In this step, they did not know that these documents were summaries. They evaluated the summaries in terms of a pure readability.
- Second step: After evaluation of the first step, we told them that the documents were summaries which were automatically generated from computers. Next, they received the original conversation, namely the non-summarized document. Then, they evaluated each summary in terms of the correctness as a summary of the original conversation.

In these two step, test subjects can appropriately evaluate "readability" and "correctness" of each summary.

Table III shows the result. The values in the table are the average values of seven test subjects. The values in parentheses are the standard variation of each method. The proposed method outperformed the baseline method and the $n\text{Utter}_{SVM}$. This result shows that the outputs from $n\text{Utter}_{SVM}$ was insufficient because of lack of coherency as a summary. The proposed method generated better summaries from utterances with high precision from the scoring method and the outputs from SVMs. The summaries from the proposed method contained the meaning and context of the original conversations and retained the conversational interaction. Furthermore, the selection method consisting of Step 1 and 2 for additional utterances was better than the baseline. It led to the improvement of the "readability" and "correctness" as compared with the baseline. However, even the proposed method sometimes lost the coherency such as lack of an utterance about the beginning of a new topic in the conversation. Free conversations generally contain many sub topics. Detection of each topic in the conversations is important to keep the coherency of the summary generation. In addition, discourse information is important for the summary generation process to select appropriate utterances. We need to incorporate these features, namely topic detection and discourse information, to the summary generation process.

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. For the extraction process by SVMs, we handled five types of features; (f1) features in an utterance, (f2) features between utterances, (f3) features about anaphora, (f4) features based on scores and (f5) nonverbal features. The most effective feature category was (f1) features in an utterance. The nonverbal features did not work well in the experiment. Introducing other nonverbal features is the important future work.

For the generated summaries, the method with only SVMs outperformed the naive baseline and the proposed method in quantitative evaluation, namely recall, precision and F-value. However, this result was not always appropriate for the evaluation of the generated summaries. Therefore, we evaluated the three methods in qualitative evaluation. We computed "readability" and "correctness" of the summaries from each method by seven test subjects. In the evaluation, the proposed method outperformed the method with only SVMs. This result shows the essential effectiveness of our method as generation of readable summaries. Introducing additional utterance selection approaches, such as detection of the beginning of a topic, is future work to improve the readability of the generated summary.

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] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[3] Hal Daume III and Daniel Marcu. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 449–456, 2002.

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

[5] 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.

[6] 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.

[7] Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. Single-document summarization as a tree knapsack problem. In *The Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

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

[9] Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. Document summarization using conditional random fields. In *the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.