Simultaneous Summarization of Japanese Spoken Monologue for Real-time Captioning

Tomohiro OHNO[1,a)], Shigeki MATSUBARA[2,3], Hideki KASHIOKA[3,4], Yasuyoshi INAGAKI[5]
[1]Graduate School of Information Science, Nagoya University, Japan
[2]Information Technology Center, Nagoya University, Japan
[3]National Institute of Information and Communications Technology, Japan
[4]ATR Spoken Language Communication Research Laboratories, Japan
[5]Faculty of Management and Information Science, Aichi Institute of Technology, Japan
a) ohno@el.itc.nagoya-u.ac.jp

Abstract

The development of a captioning system that supports the real-time understanding of monologue speech such as lectures and commentary is now in demand. In a real-time captioning system, it is necessary to summarize speech so that the audience can understand it within the display time and to output the caption simultaneously with the monologue speech input. This paper proposes a technique for simultaneous summarization of Japanese spoken monologue toward real-time captioning. Our technique identifies a unit for which the summarization is executed each time a clause boundary is detected. Then our technique summarizes it based on the dependency structure. An experiment using Japanese monologues has shown the feasibility of our technique.

1 Introduction

The development of a captioning system that supports the real-time understanding of monologue speech such as lectures and commentary is now in demand. In a real-time captioning system, it is necessary to output the caption simultaneously with the monologue speech input. This is why the display time cannot be extended needlessly. Therefore, considering the speed at which an audience reads a caption, it is necessary to summarize the speech by deleting redundancies.

There exists some research about captioning. Holter et al. (2000) developed a system of captioning in real time by using the recognition result of the speech paraphrased by another person. Moreover, captioning methods by summarizing news manuscript copy automatically also have been proposed (Mikami et al., 1999; Monma et al., 2003; Daelemans et al., 2004). However, in the former system, since the summarization is performed by a person, the system cannot work fully automatically. In the latter, on the assumption that manuscript copy exists beforehand, the method generates captions by automatically summarizing the copy therefore, real-time summarization has not been achieved. In summarization of spoken monologue for real-time captioning, it is also necessary to consider the time limitation. Furthermore, these conventional methods have adopted the sentence as the basic unit of summarization. However, since there is not an evincive sentence boundary in a spoken monologue, it is not easy to divide a monologue into sentences beforehand. In addition, even if the sentence boundaries can be detected, since sentences in monologues tend to be long, the simultaneity of outputting the caption following the speech is impaired. Thus, in the case of summarizing the speech in real time, what kind of language unit is defined as the basic unit for which the summarization is executed (hereafter called the **summary unit**) becomes an important point.

This paper proposes a technique for simultaneous summarization of Japanese spoken monologues toward real-time captioning. The technique identifies a summary unit using the results of incremental dependency parsing based on clause boundaries (Ohno et al., 2005). That is, a summary unit is identified based on the dependency structure that is determined each time a clause boundary is detected. After that, the technique summarizes the summary unit. This enables us to simultaneously summarize the speech input and to caption it in real time. Additionally, since this unit is a sequence of clauses connected by dependency relations of which the modifier *bunsetsus*[1] are the final ones in a clause, it constitutes a syntactically sufficient and semantically meaningful language unit. Therefore, it can be concluded that this unit is suitable for summarization. Furthermore, considering the dependency structure prevents our summarization technique from generating unnatural captions. An experiment using Japanese spoken monologues has shown the feasibility of our technique.

2 Real-time Captioning

Generally speaking, in order to achieve real-time captioning by summarizing speech, it is necessary to perform the following process.

(1) Speech recognition

(2) Detection of a summary unit

(3) Summarization

(4) Decision of layout of caption in a screen

(5) Decision of display time

(6) Display of a caption

Among these, (2) ～ (5) are treated in our research and (2) and (3) are mainly described in this paper. In (4), we have to decide the number of lines in a screen and the number of characters in a line on each screen. In this research, all characters of the summarizing results for one summary unit are displayed on one screen not taking into account the line feeds and the limitation of the number of characters. In (5), our method displays a caption only in the speech time of the corresponding summary unit.

3 Detection of Summary Units

In our research, summary units are decided based on the result of incremental dependency parsing on a clause-by-clause basis (Ohno et al., 2005). Figure 1 shows the flow of detecting summary units. In what follows, first, we outline incremental dependency parsing based on clause boundaries. Next we describe the algorithm for deleting summary units, and finally we show an example of detecting summary units.

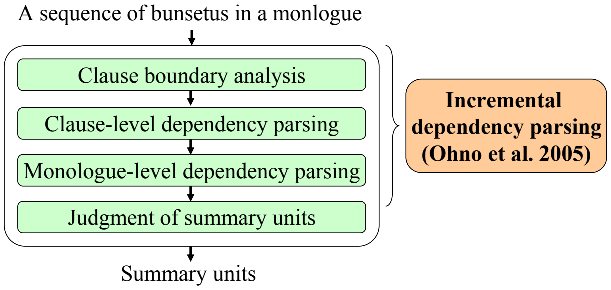

A sequence of bunsetus in a monlogue
Clause boundary analysis
Clause-level dependency parsing
Monologue-level dependency parsing
Judgment of summary units
Summary units
Incremental dependency parsing (Ohno et al. 2005)

Figure 1: Flow of detecting summary units

3.1 Incremental Dependency Parsing Based on Clause Boundaries

In this method, we adopt a clause as a parsing unit and perform the incremental dependency parsing, which can output the dependency structure of a clause simultaneously with the monologue speech input. In Japanese, a clause basically contains one verb phrase. Since a clause constitutes a syntactically sufficient and semantically meaningful language unit, it can be used as an alternative parsing unit to a sentence. Our proposed method assumes that a monologue is a sequence of one or more clauses, and every bunsetsu in a clause, except the final bunsetsu, depends on another bunsetsu in the same clause. As an example, the dependency structure of a part of a Japanese spoken monologue is presented in Figure 2:

> "先日総理府が発表いたしました世論調査によりますと死刑を支持するという人が八十パーセント近くになっております(A public opinion poll announced by the Prime Minister's Office the other day indicates that the ratio of people supporting capital punishment is nearly 80%.)"

Here, although it is essentially impossible to divide a monologue into clauses on one dimension, a monologue can be approximately segmented into clauses by a clause boundary annotation program (CBAP)[2] (Kashioka and Maruyama, 2004). In our research, we call the unit sandwiched between two clause boundaries detected by the clause boundary analysis the **clause boundary unit** and adopt it as an alternative parsing unit.

[1] A bunsetsu is one of the linguistic units in Japanese, and roughly corresponds to a basic phrase in English. A bunsetsu consists of one independent word and more than zero ancillary words. A dependency is a modification relation in which a modifier bunsetsu depends on a modified bunsetsu.

[2] This program can specify the positions and types of clause boundaries simply from a local morphological analysis. There exist 147 types such as "連体節 (adnominal clause)."

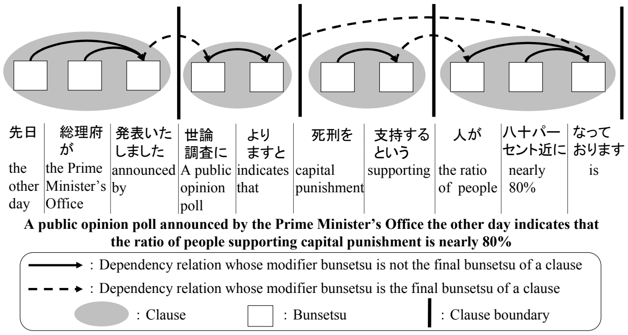

先日
the other day
総理府が
the Prime Minister's Office
発表いたしました
announced by
世論調査に
A public opinion poll
よりますと
indicates that
死刑を
capital punishment
支持するという
supporting
人が
the ratio of people
八十パーセント近に
nearly 80%
なっております
is
A public opinion poll announced by the Prime Minister's Office the other day indicates that the ratio of people supporting capital punishment is nearly 80%
: Dependency relation whose modifier bunsetsu is not the final bunsetsu of a clause
: Dependency relation whose modifier bunsetsu is the final bunsetsu of a clause
: Clause
: Bunsetsu
: Clause boundary

Figure 2: Relation between clause boundary and dependency structure

In this dependency parsing method, the transcribed sentence for which a morphological analysis, clause boundary detection, and bunsetsu segmentation are provided is considered as an input. Note that our method can parse the input data in which sentence boundaries are not detected. Our dependency parsing method is a two-stage model. That is, the following two methods are executed by turns.

1. **Clause-level parsing**
   The dependency relations of a clause boundary unit inside are identified for every clause boundary unit in a monologue.
2. **Monologue-level parsing**
   The dependency relations of which the modifier bunsetsus are the final bunsetsus of the clause boundary units in a monologue are identified.

In those two parsings, the structure that maximizes the likelihood that is calculated by using the dependency probability provided from the corpus is regarded as the dependency structure and calculated by dynamic programming (DP). Refer to the literature (Ohno et al., 2006b) for details of the statistical model.

In monologue-level dependency parsing, since it is not clear when their modified bunsetsus are provided, the timing on which the dependency relation is decided is important. In our method, each time a clause boundary unit is provided, the maximum likelihood dependency structure of that point is parsed and if a dependency relation for the final bunsetsu of a clause boundary unit does not change during a fixed input time (hereafter referred to as a **fixed value**), the dependency relation is decided as having a modified bunsetsu.

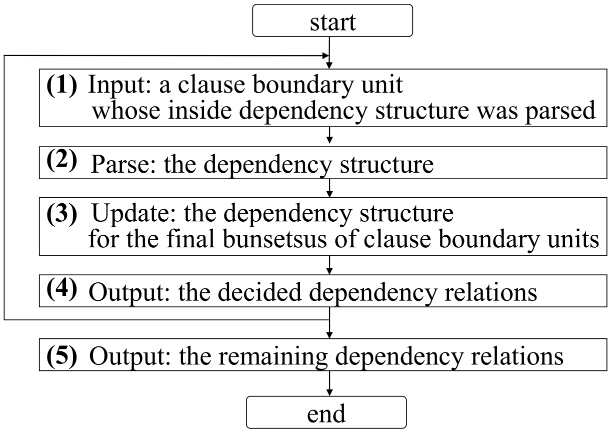

start
(1) Input: a clause boundary unit whose inside dependency structure was parsed
(2) Parse: the dependency structure
(3) Update: the dependency structure for the final bunsetsus of clause boundary units
(4) Output: the decided dependency relations
(5) Output: the remaining dependency relations
end

Figure 3: Flow of incremental dependency parsing

Figure 3 shows the flow of incremental dependency parsing for the final bunsetsus of clause boundary units. This algorithm executes incremental parsing by updating the dependency structure $D = \{(dep(b^{j}_{n_j}), k) \mid 1 \leq j \leq i-1\}$ for the final bunsetsus $b^{1}_{n_1}, \cdots, b^{i-1}_{n_{i-1}}$ of the clause boundary units $C_1, \cdots, C_{i-1}$, which are already provided each time a new clause boundary unit C_i is provided. k is a number called continuation, that is, the number of times into which $dep(b^{j}_{n_j})$ does not change. The following indicates the algorithm of dependency parsing. Here, we describe the fixed value as σ.

(1) The clause boundary unit C_i, whose inside dependency structure was decided, is provided.

(2) The dependency relations containing a modifier bunsetsu whose modified bunsetsu is not identified and which is the final bunsetsu of a clause boundary unit are parsed by the monologue level dependency method.

(3) Based on the dependency relations $dep(b^{j}_{n_j})$ $(1 \leq j \leq i-1)$, which were generated in (2), the dependency structure D for the final bunsetsus is updated. Here, if $dep(b^{j}_{n_j})$ does not change, continuation k is updated into $k+1$, and if it does change, it is updated into 1.

(4) Assuming that the dependency relations $(dep(b^{j}_{n_j}), k) \in D$ that satisfy $k = \sigma$ are reliable enough to be decided, the dependency relations are generated.

(5) After all clause boundary units were provided, the dependency relations that are undecided, that is, satisfy $k < \sigma$ in $(dep(b^{j}_{n_j}), k) \in D$ are generated.

3.2 Judgment of Summary Units

In order to detect summary units, the following judgment is performed every time dependency relations whose modifier bunsetsus are the final bunsetsus of clause boundary units are decided.

- If one of the modifier bunsetsus of the decided dependency relations is
  - the earliest spoken bunsetsu among the undecided bunsetsus, a sequence of clause boundary units connected by the identified dependency relations is decided as a summary unit and is handed off into the next summarization process.
  - not the earliest spoken bunsetsu among the undecided bunsetsus, a summary unit is not decided.

3.3 Example of Detecting Summary Units

Figure 4 shows the process of detecting a summary unit in the following sequence of bunsetsus in a part of a monologue.

1: その[the]
2: 一方で[on the other hand]
3: 廃止を[it]
4: した[abolishing]
5: 後に[after]
6: 再開を[restarted capital punishment]
7: したという[that]
8: 国も[some countries]
9: ございます[there are]
10: 先日[the other day]
11: 総理府が[the Prime Minister's Office]
12: 発表いたしました[announced by]
13: 世論調査に[a public opinion poll]
14: よりますと[indicates that]
15: 死刑を[capital punishment]
16: 支持するという[supporting]
17: 人が[the ratio of people]
18: 八十パーセント近くに[nearly 80%]
19: なっております[is]

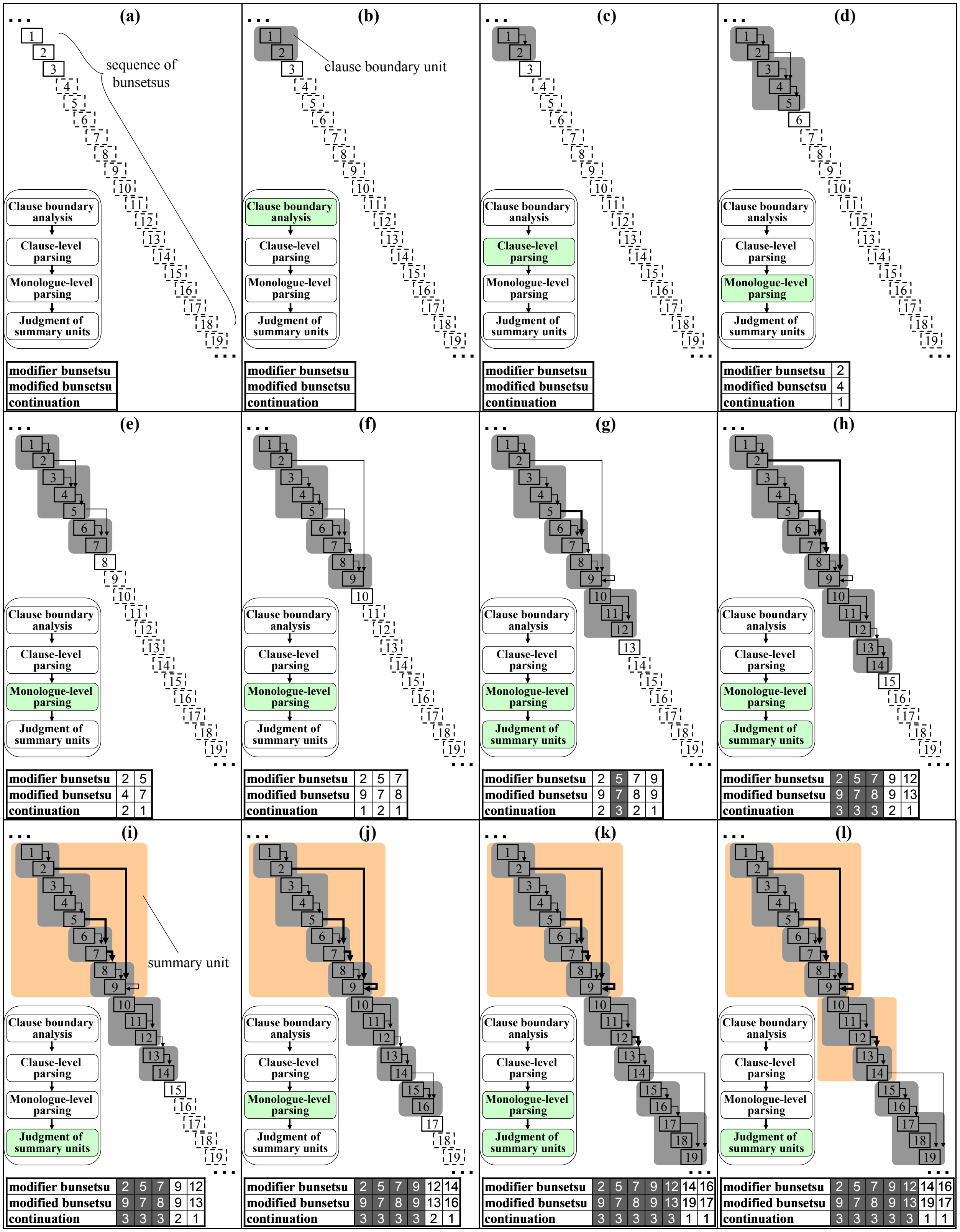
[On the other hand, there are some countries that restarted capital punishment after abolishing it. A public opinion poll announced by the Prime Minister's Office the other day indicates that the ratio of people supporting capital punishment is nearly 80%.]

This figure consists of 12 processes **(a)**～**(l)**. Each subfigure shows an input sequence of bunsetsus from the top left to the middle right, and the current procedure in the middle left, the dependency relations whose modifier bunsetsus are the final bunsetsus of the clause boundary units in the bottom table. $dep(b^{j}_{n_j})$ and k of $(dep(b^{j}_{n_j}), k) \in D$ respectively correspond to "modifier bunsetsu and modified bunsetsu" and "continuation" in the bottom table. Here, we explain the process based on the assumption that the fixed value is 3.

(a) shows the state in which the 3rd bunsetsu was provided. Here, the solid and dotted quadrangles respectively mean the bunsetsus that were provided and the bunsetsus that have not been provided yet. **(b)** shows the state in which the clause boundary between the 2nd and 3rd bunsetsus was detected by clause boundary analysis. If the clause boundaries are detected, the clause boundary unit is identified and the clause-level parsing is executed. **(c)** shows the state in which the clause-level parsing was executed for the 1st clause boundary unit.

(d)～**(l)** show the process of monologue-level parsing and detecting of summary units. **(d)** shows the state in which the 2nd clause boundary unit was identified and the dependency structure $\{dep(2)\}$ was parsed. $dep(2)$ corresponds to the arrow between the 2nd bunsetsu and the 4th bunsetsu. 1 is recorded to the continuation in the bottom table. Similarly, **(e)** and **(f)** respectively show the state in which each maximum likelihood dependency structure $\{dep(2), dep(5)\}$, $\{dep(2), dep(5), dep(7)\}$ was parsed when the 3rd and 4th clause boundary units respectively were identified.

(g) shows the state in which the 5th clause boundary unit was identified and the maximum likelihood dependency structure $\{dep(2), dep(5), dep(7), dep(9)\}$ was parsed. In this time, since the continuation of the dependency relation $dep(5)$ reaches to 3, the dependency relation is decided, and then the judgment of summary units is executed. Since the modifier bunsetsu of the decided dependency relation is not the earliest



(a)

modifier bunsetsu	
modified bunsetsu	
continuation	

(b)

modifier bunsetsu	
modified bunsetsu	
continuation	

(c)

modifier bunsetsu	
modified bunsetsu	
continuation	

(d)

modifier bunsetsu	2
modified bunsetsu	4
continuation	1

(e)

modifier bunsetsu	2	5
modified bunsetsu	4	7
continuation	2	1

(f)

modifier bunsetsu	2	5	7
modified bunsetsu	9	7	8
continuation	1	2	1

(g)

modifier bunsetsu	2	5	7	9
modified bunsetsu	9	7	8	9
continuation	2	3	2	1

(h)

modifier bunsetsu	2	5	7	9	12
modified bunsetsu	9	7	8	9	13
continuation	3	3	3	2	1

(i)

modifier bunsetsu	2	5	7	9	12
modified bunsetsu	9	7	8	9	13
continuation	3	3	3	2	1

(j)

modifier bunsetsu	2	5	7	9	12	14
modified bunsetsu	9	7	8	9	13	16
continuation	3	3	3	3	2	1

(k)

modifier bunsetsu	2	5	7	9	12	14	16
modified bunsetsu	9	7	8	9	13	19	17
continuation	3	3	3	3	3	1	1

(l)

modifier bunsetsu	2	5	7	9	12	14	16
modified bunsetsu	9	7	8	9	13	19	17
continuation	3	3	3	3	3	1	1

Figure 4: Example of detecting summary units (in case where fixed value is 3)

spoken bunsetsu among the undecided bunsetsus, a summary unit is not decided. **(h)** shows the state in which the *dep*(2) and *dep*(7) were decided in monologue-level parsing. In this time, similar to **(g)**, the judgment of summary units is executed. **(i)** shows the state in which the 1st summary unit was identified because one of the modifier bunsetsus of the dependency relations which were decided on **(h)** is the earliest spoken bunsetsu among the undecided bunsetsus. **(j)**〜**(l)** show the state similar to the previous sub figures. In the state of **(l)**, the 2nd summary unit was identified.

4 Summarization Based on Dependency Structure

Our method deletes redundancies in a summary unit identified in the previous section. This deletion is executed in consideration of the dependency structure not generating ungrammatical language.

Here, it is necessary to set the standard of the length of the summarized text (hereafter, target number of characters), in consideration of the trade-off between the following two demands for the captioning system.

- The captioning system should summarize the speech briefly so that the audience can read the captions within its display time.
- The captioning system should display the transcription of the speech without summarization to preserve information.

We adopted a criterion of “4 characters per second” established by JAPAN TELETEXT Co. Ltd. (JAPAN TELETEXT Co., Ltd., 2006) as the display speech at which the audience can read the closed-caption. The target number of characters of each summary unit can be decided automatically based on the above criterion and the display time described in Section 2. In short, the transcribed text exceeding the target number of characters is summarized and shortened as much as possible. However, attaching too high a value to the target number of characters causes deletion of even important information, altering the original meaning of the speech and interfering with the audience’s understanding. Then, to avoid the deletion of an important part of the speech, the above criterion is not assumed to be an absolute limitation.

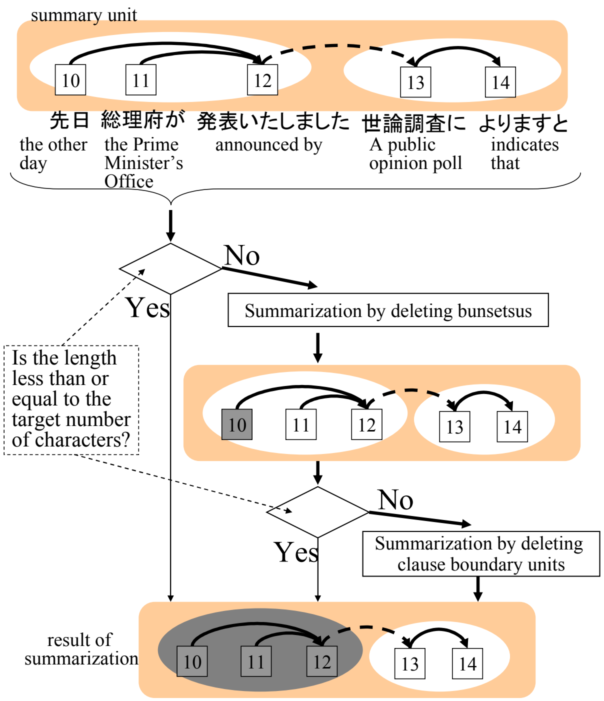


Figure 5: Flow of summarization

Figure 5 shows the flow of summarization with the example of summarizing one of the summary units, which were identified in Figure 4. In summarization our method first deletes bunsetsus that fulfill each condition described in Section 4.1 in order until the length reaches the target number of characters. Second, our method deletes clause boundary units that fulfill each condition described in Section 4.2 in order until the length reaches the target number of characters. Here, since the speech time of the summary unit in Figure 5 is 3.577 seconds, the target number of characters is 14.308 (= 3.577 × 4).

4.1 Summarization by Deleting Bunsetsus

In this process, our method summarizes a summary unit by deleting bunsetsus in the following order until the length reaches the target number of characters. The order was decided by taking into account the dependency structure, the part-of-speech of the self-sufficient word, the modified bunsetsu and so on based on previous work (Mikami et al., 1999).

(1) Delete bunsetsus whose self-sufficient word is an adverb and whose modified bunsetsu’s self-sufficient word is an adverb or adjective.

(2) Delete bunsetsus whose self-sufficient word is an adverb except the above.

(3) Delete bunsetsus whose modified bunsetsu's self-sufficient word is a noun except a formal noun.

In Figure 5, the 10th bunsetsu is deleted because the 10th bunsetsu is an adverb. Since the length of the deleting result is more than 14.308, the process of the next section is executed.

4.2 Summarization by Deleting Clause Boundary Units

In this process, our method summarizes a summary unit by deleting clause boundary units which is very subordinative like "連体節 (adnominal clause)" "ナガラ節 (nagara-clause)," "ツツ節 (tsutsu-clause)" and so on (Kashioka and Maruyama, 2004) until the length reaches the target number of characters. The order in which the clause boundary units are deleted depends on the depth of the dependency structure. The depth means the number of paths from the final bunsetsu of the summary unit to the final bunsetsu of the clause boundary unit.

In Figure 5, the 1st clause boundary unit, which consist of the 10th, 11th, and 12th bunsetsus, is deleted because the type of clause boundary unit is "連体節(adnominal clause)." Consequently, the length of the deleting result becomes 10, which is less than the target number of characters (14.308).

5 Experiment

In order to evaluate the effectiveness of our summarization method for real-time captioning of spoken monologue, we conducted an experiment.

5.1 Outline of Experiment

We used 7 programs (470 sentences) in the syntactically annotated corpus of spoken monologue "Asu-Wo-Yomu[3] (Ohno et al., 2006a)." In incremental dependency parsing, we set the fixed value as 3 and used 95 programs (5,532 sentences) as the learning data. We evaluated the results by the following evaluation index:

(1) The display speed of a caption (the number of characters displayed in a second)

(2) The quality of the summarization.

Table 1: Criterion for evaluation of quality

value	criterion
4	a caption that can be read naturally (naturalness) and in which all important parts were preserved (fidelity)
3	a caption of which the naturalness and fidelity were slightly impaired
2	a caption of which either the naturalness or the fidelity was very impaired but which is narrowly understandable
1	a caption of which both the naturalness and the fidelity were very impaired and which is not understandable

Table 2: Number of summary units, number of characters and the summary rate

summary unit	character	summary rate (%)
802	21,462	77.4

In the evaluation by the index (1), we compared our method with the transcript method, in which the transcription of each summary unit is displayed without summarizing only in its speech time. In the evaluation by the index (2), the transcription and the summarizing result of each summary unit were presented to two estimators. Then the estimators judged the caption for each summary unit on a scale of 1 to 4 based on the degree of naturalness and fidelity (Mikami et al., 1999). The criterion for this judgment is shown in Table 1.

5.2 Experimental Results

Table 2 shows the number of our identified summary units, the number of characters in all the summary units and the summary rate for the test data. The summary rate is calculated by the equation: (the number of characters in all captions) / (the number of characters in all summary units). The summary rate was 77.4% (16,612/21,462). The number of identified summary units was 802, which was about 1.7 times the number of sentences. Here, the precision and recall of the clause boundary analysis were 99.1% and 97.6%, respectively. The dependency accuracy was 76.2%.

Figure 6 shows the frequency distribution of the caption by the display speed in the case of our method and the transcript method. The average number of characters displayed in a second was 5.3 (characters/sec.) in our method and 6.5 (characters/sec.) in the transcript method. Furthermore, the

[3] Asu-Wo-Yomu is a TV commentary program of the Japan Broadcasting Corporation (NHK). The commentator speaks on some current social issue for 10 minutes.

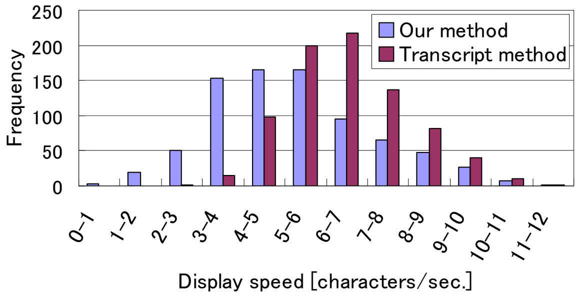


Figure 6: Frequency distribution of captions by display speed

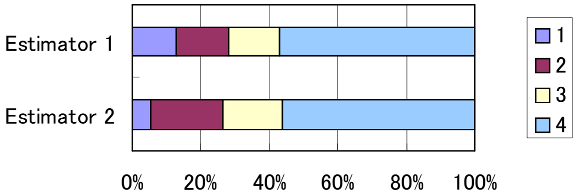


Figure 7: Result of evaluation of the quality of captions

rate of the caption, which fulfills the criterion that the number of characters displayed in a second should be less than or equal to 4, is 28.4% (228/802) in our method and 2.0% (16/802) in the transcript method.

From these results, we have confirmed that our method's captions can be read more slowly.

Next, Figure 7 shows the evaluation result of the summarization quality. The graph indicates the ratio of the number of summary units for which each evaluated value was assigned. We obtained the good result that the ratio of summary units for which the evaluated value 1 was assigned by the two estimators is about 10%. Furthermore, in both results of the two estimators, summary units of which the evaluated value was more than or equal to 3 occupy about 70%. From the above result, we could confirm most of the captions summarized by our method have admissible quality.

6 Conclusion

In this paper, we have proposed a summarization technique for real-time captioning of Japanese spoken monologues. Our technique has achieved the real-time summarization of speech input based on the dependency structure that is parsed each time clause boundaries are detected. As a result of the experiment using spoken monologues, we have confirmed the feasibility of our technique.

Future research will involve improving the summarization method and evaluating the quality and simultaneity of the captions in more detail.

Acknowledgements: This research was supported in part by the Grant-in-Aid for Young Scientists of JSPS. The first author is partially supported by JSPS Research Fellowships for Young Scientists.

References

W. Daelemans, A. Hothker, and E. T. K. Sang. 2004. Automatic sentence simplification for subtitling in Dutch and English. In *Proc. of 4th LREC*, pages 1045–1048.

T. Holter, E. Harborg, M. H. Johnsen, and T. Svendsen. 2000. Asr-based subtitling of live TV-programs for the hearing impaired. In *Proc. of 6th ICSLP*, volume 3, pages 570–573.

JAPAN TELETEXT Co., Ltd. 2006. FAQ about captioning. http://www.telemo.co.jp/jimaku/jimaku-main6.html. [Online; accessed 10-June-2007] (In Japanese).

T. Kashioka and T. Maruyama. 2004. Segmentation of semantic unit in Japanese monologue. In *Proc. of ICSLT-O-COCOSDA2004*, pages 87–92.

M. Mikami, S. Masuyama, and S. Nakagawa. 1999. A summarization method by reducing redundancy of each sentence for making captions of newscasting. *Journal of NLP*, 6(6): 65–81. (In Japanese).

T. Monma, E. Sawamura, T. Fukushima, I. Maruyama, T. Ehara, and K. Shirai. 2003. Automatic closed caption production system on TV programs for hearing-impaired people. *Systems and Computers in Japan*, 34(13): 71–82.

T. Ohno, S. Matsubara, H. Kashioka, N. Kato, and Y. Inagaki. 2005. Incremental dependency parsing of Japanese spoken monologue based on clause boundaries. In *Proc. of 9th EUROSPEECH*, pages 3449–3452.

T. Ohno, S. Matsubara, H. Kashioka, N. Kato, and Y. Inagaki. 2006a. A syntactically annotated corpus of Japanese spoken monologue. In *Proc. of 5th LREC*, pages 1590–1595.

T. Ohno, S. Matsubara, H. Kashioka, T. Maruyama, and Y. Inagaki. 2006b. Dependency parsing of Japanese spoken monologue based on clause boundaries. In *Proc. of ACL/COLING2006*, pages 169–176.