Subtopic Segmentation in the Lecture Speech

Noboru Kanedera∗ Asuka Sumida∗ Takao Ikehata∗ Tetsuo Funada†
∗ Ishikawa National College of Technology, Ishikawa, Japan
† Kanazawa University, Kanazawa, Ishikawa, Japan

Abstract

This paper proposes a method of segmentation that segments lecture video material into subtopics based on speech signals for creation of educational video contents. To represent subtopics of video segments, the text recognized by automatic speech recognition (ASR) from a lecture speech was converted into an index using independent component analysis (ICA) instead of conventional TF-IDF. This research attempted a method of segmentation using dynamic programming that minimizes the sum of cosine measures between adjacent indexes. The validity of the proposed method was evaluated using sample lecture videos. Results indicated that subtopic segmentation using automatic speech recognition performed as well as that using transcription text.

1. Introduction

In recent years, high-speed network environments have been set up which allow students to do preparatory study and revision in their own homes using lecture videos. However, the number of such lecture videos is still limited. One of the reasons for this is, no doubt, that a very large amount of time and effort is required to edit these videos. Editing a video requires searching for subtopic segmentation positions, deletion, transfer and merging of video segments. In particular, when searching subtopic segmentation positions, a large amount time and effort are required to review the video from beginning to end, which is a considerable burden. In this context, we have been studying the development of a system which can perform the automatic segmentation of videos as a means of supporting the creation of lecture video material (Fig.1).

In this report, we describe the results of automatic subtopic segmentation using speech information extracted from a lecture video in a class before editing. Previous studies have been carried out on subtopic segmentation of lectures in academic meetings [1, 2] or lectures after editing [3, 4]. However, almost no studies have been carried out on lectures in classes before editing. When subtopic segmentation is performed on a lecture in a class before editing, as there are a large number of unnecessary sentences, it might be expected that it would be even more difficult to detect a change-over.

2. Indexes for estimating subtopic segmentation positions

In general, a video comprises several subtopic segments. If adjacent video segments are similar, these segments can be considered as a continuous subtopic segment, and if they are not similar, there would be a subtopic segmentation position between the segments. To compare video segments on a computer, topic information in each segment must be converted into some kind of index. As indexes, TF-IDF[1, 3], a mutual information amount which takes TF-IDF into account [4] and the $\chi^2$ value [5] are often used. On the other hand, a method has been proposed wherein sentences are converted to indexes and the number of dimensions does not depend on the number of words using Independent Component Analysis (ICA) [6]. In this report, sentences are converted into indexes using Independent Component Analysis, which are then used as indexes for subtopic segmentation.

The number of elements of an index by TF-IDF is given by the number of words $T \times$ the number of sentences $N$. In general, as the number of words $T$ is extremely large, this number of elements is also extremely large. On the other hand, as the number of elements in the index due to independent component analysis is the number of topics $K \times$ the number of sentences $N$, the size of the index can be reduced compared to TF-IDF.
3. Subtopic segmentation method

Firstly, it shall be assumed that the video can be segmented into several sentence sections. Next, a conversion to the index for each sentence section is performed, and the question of whether or not there is a similarity between adjacent sentence sections is examined. To examine whether or not there is a similarity between adjacent sentence sections, the cosine measure [7] of the index is used. We shall consider that the smaller the cosine measure, the less sentence sections resemble each other, and the larger the cosine measure, the more sentence sections resemble each other. In other words, if the sum of the cosine measures between adjacent sentence sections is a minimum, there is no resemblance between any of the segments, and the overall sentence can be appropriately segmented [4]. Hence, in this report, it shall be assumed that estimation of subtopic segmentation position is the problem of searching for the subtopic combination for which the sum of cosine measures is a minimum, and we now propose a method of resolving this using dynamic programming (DP).

Let sentences $1 - N$ be divided into $P$ segments $1 - b_1, b_1 + 1 - b_2, \ldots, b_{p-1} + 1 - N$ using an index $I$. Each column in the index $I$ corresponds to the sentence $1 - N$. The sum $r_p$ of the index $I$ corresponding to the $p$-th sentence section $b_{p-1} + 1 - b_p$, is defined by the following equation:

$$r_p = \sum_{m=b_{p-1} + 1}^{b_p} I_m$$

where $I_m$ is the $m$-th column of $I$.

A subtopic boundary $\hat{B}_P = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_{p-1})$ such that the sum of cosine measures between adjacent $r_p$ and $r_{p+1}$, is a minimum, is determined by the following equation:

$$\hat{B}_P = \arg \min_{B_P} \sum_{p=1}^{P-1} d(r_p, r_{p+1})$$

(1)

where $B_P = (b_1, b_2, \ldots, b_{p-1})$, and $d(a, b)$ represents the cosine measure between vectors $a$ and $b$. Dynamic programming is used to solve Equation (1). First, a cumulative measure $g(i, j)$ between adjacent sentence sections when a sentence section $1 - i$ is divided into $j$ segments, is defined by the following equation:

$$g(i, j) = \min_{j-1} \sum_{p=1}^{j-1} d(r_p, r_{p+1})$$

(2)

where $B_j = (b_1, b_2, \ldots, b_{j-1})$. Further, if $s(i, j)$ is $r_j$ when the $j$-th sentence section finishes at the sentence $i$, the scene boundary can be calculated as follows:

1. When $j = 1$, for $i = 1, 2, \ldots, N$
   $$g(i, 1) = 0$$
   $$s(i, 1) = \sum_{m=1}^{i} I_m$$

2. When $j \geq 2$, for $i = j, j+1, \ldots, N$
   $$\hat{k}(i, j) = \arg \min_{k=j-1, \ldots, i-1} \left\{ g(k, j - 1) + d \left( s(k, j - 1), \sum_{m=k+1}^{i} I_m \right) \right\}$$
   $$s(i, j) = \sum_{m=k(\hat{i}, j)+1}^{i} I_m$$
   $$g(i, j) = g(\hat{k}(i, j), j - 1) + d \left( s(\hat{k}(i, j), j - 1), s(i, j) \right)$$

3. $\hat{b}_P = N$
   For $p = P - 1, P - 2, \ldots, 1$
   $$\hat{b}_p = \hat{k}(\hat{b}_{p+1}, p + 1)$$

Fig.2 shows an example where the cumulative cosine measure of adjacent sentence sections when sentences $1 - i$ are segmented into three parts, is calculated. Assuming that the third sentence section is the sentence $k + 1 - i$, the third sentence section index is given by $r_3 = \sum_{m=k+1}^{i} I_m$. If the second sentence section finishes at $k$, the second sentence section start point is given by $\hat{k}(k, 2) + 1$, and the second sentence section index is given by $r_2 = s(k, 2) = \sum_{m=k+1}^{k} I_m$. Therefore, $g(i, 3)$ is obtained by adding the cosine measure of the index $r_2$ for the second sentence section and the index $r_3$ for the third sentence section, to the cumulative cosine measure $g(k, 2)$ when sentences $1$ to $k$ are segmented into two parts.

4. Speech recognition for lecture speech

In this section we describe results of speech recognition experiments for lecture speech in a class. The experiments were carried out on videos of the lectures shown in Table 1. These videos were two lectures of approximately 60 minutes. The number of sentences in Table 1, is the number of speech intervals divided by whether or not a silent interval continued for 1 second or more. As the recording was made using a headset, there is little effect due to noise, etc. Speech information alone was extracted from these videos, and down-sampled at 16 kHz. The speech was segmented at each silent interval of 1 second or more, and speech recognition was performed using Japanese dictation toolkit (1998 version) [8]. The
Table 1: Lecture videos used.

<table>
<thead>
<tr>
<th>Lecture video</th>
<th>The number of sentences</th>
<th>The number of correct boundaries</th>
<th>The number of subtopics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>412</td>
<td>61</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>605</td>
<td>99</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 2: Speech recognition results for lecture speech.

<table>
<thead>
<tr>
<th>Lecture video</th>
<th>Linguistic model by newspaper articles</th>
<th>Linguistic model by lecture speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Word correct rate [%]</td>
<td>Word accuracy [%]</td>
</tr>
<tr>
<td>1</td>
<td>49.8</td>
<td>40.9</td>
</tr>
<tr>
<td>2</td>
<td>48.9</td>
<td>41.2</td>
</tr>
<tr>
<td>Mean</td>
<td>49.4</td>
<td>41.1</td>
</tr>
</tbody>
</table>

Table 2 shows the speech recognition results for the lecture speech. In Table 2, by using a language model [9] learnt from lecture speech in academic meetings which are closer to lectures in classes rather than a language model learnt from newspaper articles [8], an improvement of approximately 6% for the word correct rate was obtained, but there was almost no improvement of word accuracy.

5. Subtopic segmentation results

5.1. Experimental conditions

Speech recognition was performed under the conditions of Section 4. Independent words were extracted alone from text obtained from this result, and an index was calculated by independent component analysis (ICA). The number of independent components was taken to be approximately 0.15 times the number of sentences shown in Table 1 from previous experiments [10] (for Video 1, 60, and for Video 2, 90). After calculating the index, scene segmentation was performed by the method shown in Section 3. Smoothing was performed on the word-document matrix by the preceding and succeeding documents.

The lecturer was asked to extract the subtopic boundaries for the lecture videos. If the analysis results matched the subtopic boundary extracted by the lecturer, the result was considered successful. As the number of subtopics is less than the number of correct segment boundaries (Table 1), almost all of them are boundaries of sections to be deleted (unnecessary parts). Therefore, the two ends of a silent interval of 3 seconds or longer were added to the segment boundaries. This is because there is a possibility that long silent intervals may be deleted as unnecessary parts more than required.

The following are the recall and precision used as test criteria:

\[
\text{recall} = \frac{C}{T}, \quad \text{precision} = \frac{C}{R}
\]

where C is the number of correct boundaries in responses, T is the total number of correct boundaries, and R is the total number of boundaries in responses. In this study, recall was given priority. For example, if there are a large number of errors in the estimation of subtopic segmentation position, the user must successively search the subtopic segmentation, which requires a major effort. On the other hand, excess subtopic segmentation points can be disregarded. Therefore, it is desirable that there are few errors in the estimation of subtopic segmentation position (recall is high).

5.2. Test results

Fig.3 shows the subtopic segmentation result obtained by speech recognition. For an index obtained using independent component analysis (ICA), approximately the similar results were obtained as for TF-IDF. The number of dimensions in the index using TF-IDF is depend on the number of words (in general, a large number). On the other hand, the number of dimensions when independent component analysis is used, is depend on the number of independent components specified. Therefore, subtopic segmentation using independent component analysis can be performed rapidly regardless of the number of words when DP is performed.

Fig.4 shows the results of performing subtopic segmentation on Video 1 for the text by automatic speech recognition and the transcription text using independent component analysis. Fig.5 shows the relation between recall and the number of divisions. From these results, it was found that the subtopic segmentation using speech recognition was about the same as that of transcription text. Hence, if the same errors are present at several locations, there is evidently not much effect on subtopic segmentation even if the speech recognition performance is fairly low.

6. Conclusions

We performed automatic subtopic segmentation by using speech information in lecture videos before editing with the aim of providing support for the creation of video educational material. To segment videos, we used topic expression (indexes) using independent component analysis and pause information. Optimization was performed to minimize the sum of the cosine measures of adjacent segments using DP. As a result of experiments, it was found that, if indexes from independent component analysis were used, approximately the same results as those of
TF-IDF could be obtained rapidly. Also, it was confirmed that approximately the same subtopic segmentation performance as that of transcription text was obtained using automatic speech recognition.

7. Acknowledgements
Part of this research was performed with a Scientific Research Grant (No, 14580246) from the Ministry of Education, Culture, Sports, Science and Technology.

8. References