AUTOMATIC INDEXING OF LECTURE SPEECH BY EXTRACTING TOPIC-INDEPENDENT DISCOURSE MARKERS

Tatsuya Kawahara *Masahiro Hasegawa*

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

Automatic detection of section (sub-topic) boundaries in lecture speech is addressed. The method makes use of the characteristic expressions used in initial utterances of sections defined as discourse markers, as well as pause and language model information. The discourse markers are derived in a totally unsupervised manner based on word statistics used in the information retrieval technique. The statistics is used to select candidates picked up by other information. Experimental results show that the proposed method realizes better indexing performance (better precision at high recall rates) than the simple baseline method using pause information only. Moreover, it is shown to be robust against speech recognition errors.

1. INTRODUCTION

Automatic indexing of speech materials is one of the applications of large vocabulary continuous speech recognition. Even if recognition performance is not so high, it is often possible to detect their topics or segment them into topic boundaries. There have been studies on topic classification of broadcast news[1] and voice mails[2]. Most of them extract a set of keywords that characterize topics for classification[3]. The approach is effective when there are a lot of short speech materials such as news clips and voice messages.

It is not easily applicable to indexing of long speech materials such as lectures and discussions, where one broad topic is unchanged and small issues come along with close relation. Moreover, for such spontaneous speech, recognition performance is much lower to spot a lot of keywords. On the other hand, browsing function is needed for this kinds of long materials.[4]. Specifically, exact time index for boundaries of sub-topics or sections is highly required, since such indexes are used for skipping and searching desired segments to be replayed.

In this paper, we approach the problem of indexing lecture speech by detecting the boundaries of sections. Unlike previous studies, we focus on discourse markers, which are rather topic independent. We define discourse markers as expressions frequently used at the beginning of new sections in lectures and oral presentations. The proposed method extracts them without any manually tagged information such as topics and boundaries, namely realizes unsupervised training.

2. INDEXING LECTURE SPEECH

2.1. Database

We take part in the project of "Spontaneous Speech Corpus and Processing Technology" sponsored by the Science and Technology Agency Priority Program in Japan[5]. The *Corpus of Spontaneous Japanese (CSJ)* currently developed by the project consists of a variety of oral presentations at technical conferences and informal monologue talks on given topics. They are manually given orthographic and phonetic transcription, but they are not segmented at all, i.e. one large file corresponds to a lecture.

For language model training, all transcribed data (as of June 2001) are used. There are 612 presentations and talks by distinct speakers. The text size in total is 1.48M words (=Japanese morphemes). As for acoustic model training, only male speakers are used in this work. We use 224 presentations that amount to 37.9 hour speech.

2.2. Problem and Approach

In this work, we deal with oral presentations at technical conferences. There is a relatively clear prototype in the flow of presentation: First, background of the work is introduced. Next, the problem and approach are described. Then comes explanation of specific algorithms and systems, followed by experimental evaluation. When using slides for presentation, a couple of slides corresponds to these sections and sub-sections.

Our goal is to segment lecture speech material into these units, or to find the boundaries between them. The index is useful for skipping and searching segments. If they are aligned with slides, though it is not done here, multi-media browsing is realized.

0-7803-7402-9/02/$17.00

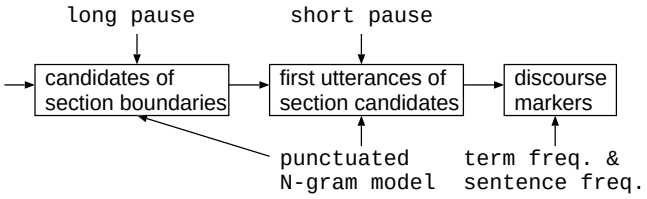

Fig. 1. Flow of extracting discourse markers for indexing

Prosodic information such as pause and pitch may be useful[6]. Preliminary analysis suggests however, prosody alone is not sufficient for the purpose, since speakers often put arbitrary long pauses during talks and sometimes keep talking even when changing slides.

It is observed that there is a typical pattern in the first utterances of the units. Speakers try to briefly tell what comes next and attract audiences' attention. For example, "Next, I will explain how it works." and "Now, move on to experimental evaluation". We define such characteristic expressions that appear at the beginning of section units as discourse markers.

A set of discourse markers are automatically trained without any manual tags. Moreover, we do not assume correct segmentations are given for training because it costs too much to manually tag the large database. Thus, the discourse markers are derived automatically from a set of transcriptions of lecture speech. They are used for indexing of lecture speech data through automatic speech recognition.

3. AUTOMATIC DERIVATION OF DISCOURSE MARKERS

The procedure of extracting discourse markers (training) is illustrated in Figure 3. At first, candidates of section boundaries and their first utterances are picked out. Then, we compute statistics of the term frequency and sentence frequency, based on which discourse markers are selected. In these processes, various information sources of pause, N-gram language model and statistics of word occurrences are utilized.

3.1. Use of Pause Information

Pause information is used for pre-selection, namely picking up candidates of section boundaries and sentence ending.

It is expected that speakers put relatively long pauses in shifting topics or changing slides, although a long pause does not always mean a section boundary. Here, we set a threshold not to lose correct hypotheses, which will be selected by the following process. The threshold value is different from person to person, depending mainly on the speaking rate. Therefore, we use the average of pause length during a lecture as the threshold.

3.2. Use of N-gram Language Model

N-gram language model is used to judge whether the detected pauses are actually end of utterances. The training texts are not punctuated with periods, nor speech recognition results using a language model trained with the corpus. Thus, we use another language model trained with punctuated texts of lecture archives (different from the corpus mentioned). As the texts are edited for public readability, the model is not matched to spontaneous lecture speech[7].

We also assume existence of a short pause at the end of utterances, though it is not critical to incorrectly concatenate next utterances for our purpose. When the short pause model is detected as the result of decoding, we test whether it is a period or not using the neighboring word sequences (w_{-2}, w_{-1}, pause, w_1, w_2). Specifically, it is judged as a period if $P(w_{-2}, w_{-1}, period, w_1, w_2)$ is higher than $P(w_{-2}, w_{-1}, w_1, w_2)$ by some margin. The margin is empirically set and the decision realizes a recall rate of 98% with precision of 75% for test samples.

By combining pause information and the language model, we obtain a candidate set of first utterances of sections.

3.3. Use of Term Frequency and Sentence Frequency

From the candidates, we extract characteristic expressions, namely select discourse markers useful for indexing. As pre-processing, we exclude functional words and proper nouns because the functional words appear in any utterances and proper nouns appear only in limited lectures.

Discourse markers should frequently appear in the first utterances, but should not appear in other utterances so often. Term frequency is used to represent the former property and sentence frequency (referred to as document frequency in information retrieval) is used for the latter. For a word w_i, the term frequency tf_i is defined as its occurrence count in the set of first sentences. The sentence frequency df_i is the number of sentences of all lectures that contain the word. The larger tf_i and the smaller df_i, the word is more appropriate as a discourse marker for indexing. We adopt the following evaluation function.

$$F(w_i) = tf_i * \log(\frac{1}{df_i}) \quad (1)$$

A set of discourse markers are selected by the order of $F(w_i)$.

3.4. Indexing using Discourse Markers

For a given new lecture, automatic indexing using the defined discourse markers is done by almost same procedure as in Figure 3. At first, candidates are chosen based on long pauses. Next, their initial utterances are cut out based on short pauses and the N-gram language model with punctuation. Suppose a sentence contains several discourse markers w_j, it is indexed if $\sum_j F(w_j)$ is larger than a threshold θ. More precisely, an index is attached at the starting time of the utterance.

4. EXPERIMENTAL EVALUATION

We use a portion of CSJ database, specifically 72 oral presentations for training the discourse markers, although much more data are used for training acoustic and language models for speech recognition. About half of them are collected at the annual meetings of the Acoustical Society of Japan and the others are from a variety of conferences.

We also set up an evaluation set of 17 presentations that are not included in the training set. Duration of lectures is 11-15 minute. The correct section boundaries for the test-set are given by human observation. The number of boundaries are 7 to 16 for each lecture.

As an evaluation measure, we use the F-measure that is a combination of the recall rate of correct boundaries and the precision rate of the detected boundaries.

$$\text{F-measure}(\alpha) = \frac{(1+\frac{1}{\alpha}) * recall * precision}{\frac{1}{\alpha} * recall + precision} \quad (2)$$

Here, the recall rate is more weighted since the correct boundaries should not be missed in indexing, while false alarms can be simply skipped in searching. We set α=10, which put 10 times larger weight on the recall rate.

4.1. Effect of Discourse Markers

We first evaluated the proposed indexing method with manual transcriptions of lectures. Based on the evaluation function (equation (1)), 75 discourse markers are selected. The recall rate, precision rate and F-measures (α=1 and 10) are plotted in Figure 2 by changing the threshold θ.

For comparison, we also tried a simple indexing method using pause length only, where pauses longer than a threshold are indexed. It corresponds to the method current tape recorders adopt, and, in [6], a longer pause is derived as the distinct feature for paragraph boundaries. The operation curve by varying the length threshold is plotted in Figure 3.

By comparing the two graphs, it is confirmed that the proposed method gets better indexing performance. Especially, in the area of high recall rates (left-most region of the graph), it gets much higher precision rates. By the F-measure (10) that puts priority on the recall rate, its advantage is more significant. Therefore, the use of discourse markers that are statistically derived is effective.

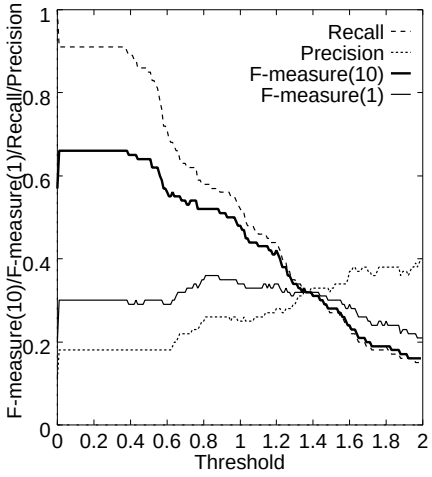

Fig. 2. Indexing performance using discourse markers

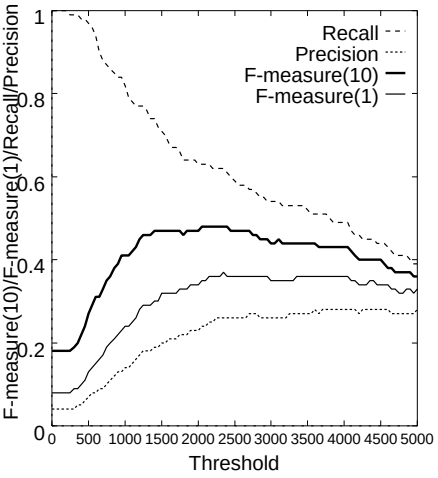

Fig. 3. Indexing performance using pause length only

4.2. Number of Discourse Markers

Next, we compare the effect of discourse markers by changing their number to 25, 75 and 125. The F-measure (10) is plotted in Figure 4. It is observed that we need some sufficient markers, but too many markers increase false alarms. Thus, we derived 75 markers.

4.3. Evaluation on ASR results

Finally, we apply the indexing method to automatic speech recognition results.

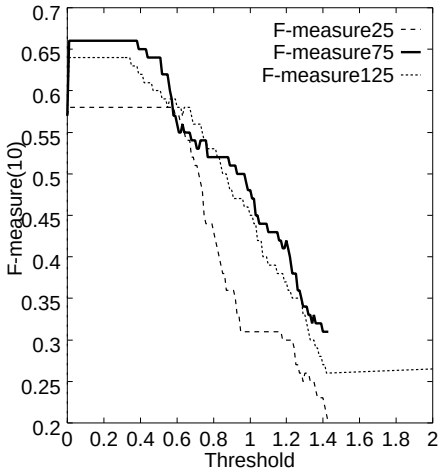

Fig. 4. Indexing performance by changing the number of discourse markers

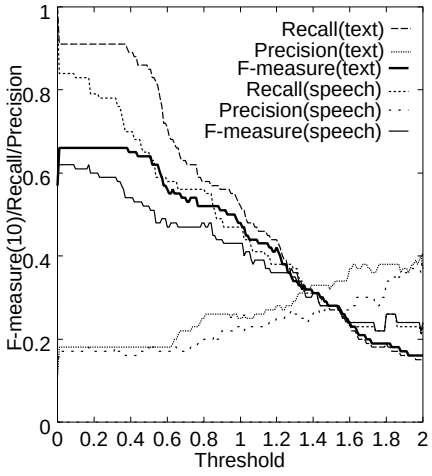

Fig. 5. Indexing performance for speech recognition results

Automatic lecture speech recognition system has been developed using the CSJ database described in Section 2[8]. Specifically, we set up a 19K-vocabulary continuous speech recognition system with cross-word triphone models of 32K Gaussian mixture components. Word accuracy for the test-set lectures is 60-70%.

The F-measure (10) as well as the recall and precision rates are plotted in Figure 5 with comparison of results for the correct transcriptions. Although the recall rate gets lower due to speech recognition errors, the degradation is relatively small considering the word error rates. It is still definitely better than the baseline method of Figure 3. The result shows that statistical evaluation of section boundaries with discourse markers is robust.

5. CONCLUSIONS

We have presented an automatic indexing method for lecture speech materials. It focuses on the characteristic expressions of the first utterances of section units defined as discourse markers. A set of discourse markers are statistically trained in a completely un-supervised manner, which does not need any manual tags. The method achieves a recall rate of 85% and a precision rate of 20%, which suffice practical indexing for fast search in long speech materials. The method is shown to be robust against speech recognition errors of 30-40%.

Ongoing works include application of the method to other domains such as lectures at universities and meeting speech.

Acknowledgment: The work was conducted in the Science and Technology Agency Priority Program on “Spontaneous Speech: Corpus and Processing Technology”. The authors are grateful to Prof. Sadaoki Furui and other members of the fruitful project.

References

[1] T.Imai, R.Schwartz, F.Kubala, and L.Nguyen. Improved topic discrimination of broadcast news using a model of multiple simultaneous topics. In *Proc. IEEE-ICASSP*, pages 727–730, 1997.

[2] G.J.F.Jones, J.T.Foote, K.S.Jones, and S.J.Young. Video mail retrieval: The effect of word spotting accuracy on precision. In *Proc. IEEE-ICASSP*, pages 309–312, 1995.

[3] J.McDonough, K.Ng, P.Jeanrenaud, H.Gish, and J.R.Rohlicek. Approaches to topic identification on the SWITCHBOARD corpus. In *Proc. IEEE-ICASSP*, volume 1, pages 385–388, 1994.

[4] A.Waibel, M.Bett, F.Metze, K.Ries, T.Schaaf, T.Schultz, H.Soltau, H.Yu, and K.Zechner. Advances in automatic meeting record creation and access. In *Proc. IEEE-ICASSP*, volume 1, pages 597–600, 2001.

[5] S.Furui, K.Maekawa, and H.Isahara. Toward the realization of spontaneous speech recognition – introduction of a Japanese priority program and preliminary results –. In *Proc. ICSLP*, volume 3, pages 518–521, 2000.

[6] M.Haase, W.Kriechbaum, G.Mohler, and G.Stenzel. Deriving document structure from prosodic cues. In *Proc. EUROSPEECH*, pages 2157–2160, 2001.

[7] K.Kato, H.Nanjo, and T.Kawahara. Automatic transcription of lecture speech using topic-independent language modeling. In *Proc. ICSLP*, volume 1, pages 162–165, 2000.

[8] T.Kawahara, H.Nanjo, and S.Furui. Automatic transcription of spontaneous lecture speech. In *Proc. IEEE workshop on Automatic Speech Recognition and Understanding*, 2001.