AUTOMATIC INDEXING OF KEY SENTENCES FOR LECTURE ARCHIVES

Tatsuya Kawahara Kazuya Shitaoka Tasuku Kitade Hiroaki Nanjo

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

ABSTRACT

Automatic extraction of key sentences from lecture audio archives is addressed. The method makes use of the characteristic expressions used in initial utterances of sections, which are defined as discourse markers and derived in an unsupervised manner based on word statistics. The statistics of the discourse markers is then used to define the importance of the sentences. It is also combined with the conventional tf-idf measure for content words. Experimental results confirm the effectiveness of the method using the discourse markers and its combination with the keyword-based method. We also present a statistical method for inserting periods into raw speech transcriptions for improving the readability.

1. INTRODUCTION

Automatic indexing of audio archives is a promising application of large vocabulary continuous speech recognition. Even if recognition performance is not so high, it is often possible to detect their topics or segment into topic boundaries so as to help users efficiently find desired portions. There have been studies on topic classification of broadcast news and voice mails. Most of them extract a set of keywords that characterize topics for classification. 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. The characteristic keywords appear throughout the speech, but broad classification based on such keywords is not meaningful. Instead, browsing function is needed for these kinds of long materials[1]. Specifically, exact time index for boundaries of sub-topics or 'sections' is highly required. More preferable form will be index attached to key sentences of these section units.

The structure of sections and paragraphs is known as useful for extracting key sentences from text materials, because most of key sentences appear at the beginning of the articles or sections. In audio materials, however, there is not explicit definition of sections and paragraphs such as line-breaking and indentation in the text.

In this paper, we approach the problem of indexing lecture audio archives by detecting section boundaries and extracting key sentences in a statistical framework. Unlike conventional studies, we focus on discourse markers, which are rather topic independent. We define discourse markers as expressions frequently used at the beginning of sections in lectures. The proposed method extracts them without any manually tagged information such as topics and boundaries, namely realizes unsupervised training.

2. AUTOMATIC TRANSCRIPTION SYSTEM

We take part in the project of "Spontaneous Speech Corpus and Processing Technology" sponsored by the Science and Technology Agency Priority Program in Japan[2][3]. The *Corpus of Spontaneous Japanese (CSJ)*[4] developed by the project consists of a variety of academic presentation speeches at technical conferences and extemporaneous public speeches on given topics. They are manually given orthographic and phonetic transcriptions, but they are not segmented into sentences both in audio and text forms.

For language model training, all transcribed data (as of January 2003) are used. There are 2592 presentations and talks by distinct speakers. The text size in total is 6.7M words (=Japanese morphemes). A trigram language model is trained for the vocabulary of 24K words. As for acoustic model training, only male speakers of academic presentations are used in this work. Using 781 presentations that amount to 106 hour speech, we set up a gender-dependent phonetic tied-mixture triphone model that consists of 25K Gaussian components and 576K mixture weights. We also revised our recognition engine Julius so that very long speech can be handled without prior segmentation[2].

With the baseline system, the word error rate is 30.9% for the test-set of 15 academic presentation speeches[3]. Adaptation of acoustic and language models based on the initial recognition result together with the speaking-rate dependent decoding strategy[5] improves it to 21.9%, which is the best figure for this test-set ever reported.

0-7803-7980-2/03/$17.00 © 2003 IEEE

3. AUTOMATIC INSERTION OF PERIODS IN SPEECH TRANSCRIPTIONS

Transcriptions of spontaneous speech include many disfluency phenomena and do not have linguistic punctuation such as periods. In read speech, a long pause is regarded as a mark of the end of utterances, thus can be converted to a period or comma. In spontaneous speech, however, the assumption does not hold. Speakers put pauses at arbitrary places, and disfluency causes irregular pauses. Therefore, we make use of linguistic information as well as pause information in order to insert periods. The period insertion procedure is necessary for segmenting speech into appropriate units and to define sentences to be indexed.

N-gram language model is used to judge whether a period should be inserted at the position of a pause. We made use of another language model trained with punctuated texts of lecture text archives consisting of 1.7M words. As the texts had been edited for public readability, the model is not matched to spontaneous lectures. For a word sequence around a pause $X = (w_{-2}, w_{-1}, pause, w_1, w_2)$, a period is inserted at the place of the pause if $Y_1 = P(w_{-2}, w_{-1}, period, w_1, w_2)$ is larger than $Y_2 = P(w_{-2}, w_{-1}, w_1, w_2)$ by some margin. This is referred to as a baseline method.

Then, we introduce a more elaborate model to convert pauses to periods selectively considering pause duration information and the adjacent words. A pause duration threshold with which pauses can be converted to periods is set up depending on the contextual words. Specifically, if $P(w_{-1}, period, *)$ or $P(*, period, w_1)$ is significantly large, we allow periods to be inserted regardless of the pause duration. Otherwise, only long pauses (longer than the average in a talk) can be converted to periods. In any cases, the final judgement is done by computing the language model score $P(Y)$.

Preliminary evaluation is done using four lectures. A professional editor cleaned the transcriptions and inserted periods. Following three methods are compared.

1) zero threshold (=baseline): any pause can be converted to a period based on the $P(Y)$.

2) threshold by the average of pause duration in a talk, which was most effective as a fixed threshold

3) proposed method that uses different threshold values (zero or the average) depending on the context

The recall, precision rates and F-measure for these methods are listed in Table 1. When we use the zero threshold, a large number of erroneous insertions are caused and the precision rate is degraded. In contrast, setting the threshold to the average value degrades the recall rate. Using the context dependent threshold, both high recall and precision rates are obtained.

Table 1: Result of period insertion

pause threshold	recall	precision	F-measure
1) zero	83.2%	75.4%	0.791
2) average	64.4%	93.7%	0.763
3) proposed (variable)	76.3%	92.3%	0.835

4. AUTOMATIC INDEXING OF KEY-SENTENCES

Next, we address automatic extraction of key sentences, which will be useful indices in lectures. Collection of these sentences may suffice summarization of the talk[6]. The framework extracts a set of natural sentences, which can be aligned with audio segments for alternative summary output. It is considered as a more practical solution in spontaneous speech, in which ASR accuracy is around 70-80%, as opposed to the approach of generating summarization based on the ASR results[7].

4.1. Discourse Modeling of Lecture Presentations

In this work, we mainly deal with lecture presentations at technical conferences. There is a relatively clear prototype in the flow of presentation, which is similarly observed in technical papers[8]. When using slides for presentation, one or a couple of slides constitute a topic discourse unit we call 'section' in this paper. The unit in turn usually corresponds to the (sub-)sections in the proceedings paper.

It is also 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. We have proposed a method to automatically train a set of discourse markers without any manual tags, and shown the effectiveness in segmentation of the lecture speech[9].

The boundary of sections is known as useful for extracting key sentences in the text-based natural language processing. However, the methodology cannot be simply applied to spoken language because the boundary of sections is not explicit in speech. Thus, the goal of the study is to apply the discourse segmentation to extraction of key sentences from the lectures.

4.2. Statistical Derivation of Discourse Markers

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 on pause duration to pick up the boundary candidates, 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 talk as the threshold.

From the candidates of the first sentences picked up by the pause information, we extract characteristic expressions, namely select discourse markers useful for indexing. 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 is used for the latter. For a word w_j, the term frequency $tf1_j$ is defined as its occurrence count in the set of first sentences. The sentence frequency sf_j is the number of sentences in all lectures that contain the word. We adopt the following evaluation function.

$$S_{DM}(w_j) = tf1_j * \log(\frac{N_s}{sf_j}) \quad (1)$$

Here, N_s is the total number of sentences in all lectures. A set of discourse markers are selected by the order of $S_{DM}(w_j)$. In this work, we selected 75 words.

4.3. Measure of Importance based on Discourse Markers

In the text-based natural language processing, a well-known heuristics for key sentence extraction is to pick up initial sentences of the articles or paragraphs. Using the automatically-derived discourse markers that characterize the beginning of sections, the heuristics is now applicable to speech materials.

The importance of sentences is evaluated using the same function (equation (1)) that was used as appropriateness of discourse markers. For each sentence s_i of possible section beginning picked up by the pause information, we compute a sum score $S_{DM}(s_i) = \sum_{w_j \in s_i} S_{DM}(w_j)$.

Then, key sentences are selected based on the score up to the specified number (or ratio) of sentences from the whole lecture.

4.4. Combination with Keyword-based Method

The other approach to extraction of key sentence is to focus on keywords that are characteristic to the lecture. The most orthodox statistical measure to define and extract such keywords is the following tf-idf criterion.

$$S_{KW}(w_j) = tf2_j * \log(\frac{N_d}{df_j}) \quad (2)$$

Here, term frequency $tf2_j$ is the occurrence count of a word w_j in the lecture, and document frequency df_j is the number of lectures (=documents) in which the word w_j appears. N_d is the number of lectures used for normalization. Here, we regard a sequence of nouns that appear more than twice in a talk as individual compound entries. For each sentence s_i, we compute $S_{KW}(s_i) = \sum_{w_j \in s_i} S_{KW}(w_j)$.

Then, we introduce a new measure of importance that combines it with the discourse marker-based method. The two are linearly interpolated with a weight w. Though a value of the weight is chosen empirically, the final performance is not so sensitive unless extreme values are used.

$$S_{final}(s_i) = w \cdot S_{DM}(s_i) + (1 - w) \cdot S_{KW}(s_i)$$

4.5. Experimental Results

We use the evaluation set of 14 presentations. Duration of the lectures is 11-15 minutes. We had human subjects select key sentences. The ratio of the key sentences among the overall sentences is 21.6% (=233/1077). The evaluation measures used are recall, precision rates and the F-measure.

First, we verified the effect of heuristics on the section structure and its automatic detection using correct transcriptions. In this experiment, the baseline period insertion algorithm is used to segment sentences. The proposed method using the discourse markers was implemented and evaluated when 30% of sentences are extracted based on the score $S_{DM}(s_i)$. The recall rate of the correct key sentences was 48.5%. For reference, when the same number of sentences were extracted from both the beginning and end of the whole lecture, which corresponds to introduction and conclusion parts respectively, the recall rate was only 27.5%. When the section structure was segmented by a human expert and the initial sentences of the sections were extracted by the same number, the recall rate was 54.2%. These results show that the heuristics on the section structure is useful and that automatic detection of section boundaries realizes sufficient performance with only a little degradation.

For comparison, we also tested a method that detects section boundaries based on the pause length only. For each sentence, duration of a longer pause among preceding and following pauses is computed and converted to a measure of importance after $N(0, 1)$ normalization. The recall rate was only 31.3%. The proposed method is shown to be more effective in detecting section boundaries and extracting key sentences.

Next, the proposed method based on the discourse markers (DM) is compared and combined with the conventional method that focuses on topic-dependent keywords (KW). The results are shown for respective methods and the combined case in Figure 1, where the F-measure is plotted by changing the extraction rate of sentences from 10 to 40%. The proposed method (DM) achieves better performance than the keyword-based method (KW). Moreover, combination of both achieves significantly higher performance. It means that the features the two methods capture are quite different and have synergetic effect when combined.

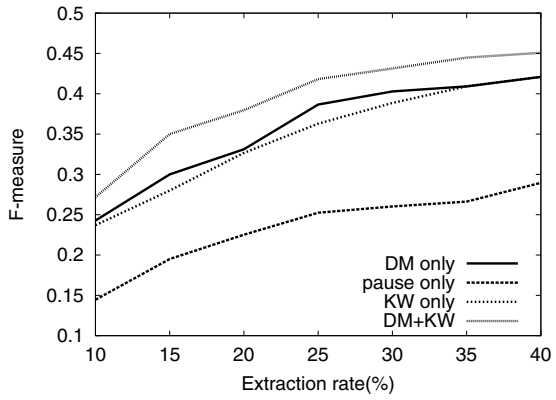

Figure 1: Extraction performance of key sentences using discourse markers (DM) and keywords (KW)

Table 2: Comparison of methods in extracting key sentences (input: correct transcription; extraction rate: 30%)

method	recall	precision	F-measure
DM	48.5%	34.5%	0.403
pause	31.3%	22.3%	0.260
KW	46.8%	33.2%	0.389
DM+pause	45.5%	32.3%	0.378
DM+KW	**51.9%**	**36.9%**	**0.431**
pause+KW	45.5%	32.1%	0.378
DM+pause+KW	51.5%	36.6%	0.428

DM: discourse marker (proposed), KW: keyword

The extraction performance in case of the extraction rate of 30% is summarized in Table 2. Among individual methods, the proposed method (DM) is most effective, which is slightly better than the conventional keyword-based method (KW). In combination of these, the DM+KW achieves the best result. Use of the pause duration information in the score of sentence importance has apparently no effect.

Finally, we made evaluation with the ASR result. Eight lectures are chosen among the test-set. Here, the improved period insertion algorithm that was proposed in Section 3 is incorporated. It actually improves the extraction rate. Table 3 shows the comparison of results for the cases of correct transcription and the ASR result as an input. Although some degradation due to speech recognition errors is observed, it is relatively small considering the word error rate of 30%. Thus, the indexing method is robust.

5. CONCLUSIONS

We have presented an automatic indexing method for lecture audio archives. It assumes the slide-based discourse structure and focuses on the characteristic expressions of the initial utterances of section units defined as discourse markers. A set of discourse markers are statistically trained in a completely unsupervised manner, which does not need any manual tags. The statistics is then used at the task of extraction of key sentences. It realizes better performance than the conventional keyword-based method. Moreover, combination of the two methods further improves the accuracy because they focus on different characteristics in a talk.

Table 3: Comparison of text and speech input

input	recall	precision	F-measure
correct transcription	55.7%	53.1%	0.544
ASR result	50.4%	51.0%	0.507

method: DM+KW

Acknowledgments: 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 for the collaboration in this fruitful project.

References

[1] 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.

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

[3] T.Kawahara, H.Nanjo, T.Shinozaki, and S.Furui. Benchmark test for speech recognition using the Corpus of Spontaneous Japanese. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 135–138, 2003.

[4] K.Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, 2003.

[5] H.Nanjo and T.Kawahara. Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. In *Proc. IEEE-ICASSP*, pages 725–728, 2002.

[6] I.Mani and M.Maybury, editors. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, 1999.

[7] C.Hori, S.Furui, R.Malkin, H.Yu, and A.Waibel. Automatic speech summarization applied to English broadcast news speech. In *Proc. IEEE-ICASSP*, volume 1, pages 9–12, 2002.

[8] S.Teufel and M.Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[9] T.Kawahara and M.Hasegawa. Automatic indexing of lecture speech by extracting topic-independent discourse markers. In *Proc. IEEE-ICASSP*, pages 1–4, 2002.