AUTOMATIC INDEXING OF KEY SENTENCES FOR LECTURE ARCHIVES USING STATISTICS OF PRESUMED DISCOURSE MARKERS

Hiroaki Nanjo, Tasuku Kitade and Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{nanjo, kitade}@ar.media.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp

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 a totally unsupervised manner based on word statistics. The statistics of the presumed discourse markers are then used to define the importance of the sentences. It is also combined with the conventional tf-idf measure of content words. Experimental results using a large corpus of lectures confirm the effectiveness of the method based on the discourse markers and its combination with the keyword-based method. It is also shown that the method is robust against ASR errors and sentence segmentation accuracy is more vital. Thus, we also enhance the segmentation by incorporating prosodic information.

1. INTRODUCTION

Automatic indexing of audio materials is one application of large vocabulary continuous speech recognition. Even if recognition performance is not very high, it is often possible to detect topics and segment speech based on topic boundaries allowing users to efficiently find the desired portions. 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. 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 remains unchanged and closely-related small issues come along. The characteristic keywords appear throughout the speech, but a broad classification based on such keywords is meaningless. Instead, a browsing function is needed for this kind of long materials[3][4]. Specifically, exact time index for boundaries of sub-topics or 'sections' is highly required, since such indices are used to skip and search for the segments to be replayed. More preferable form will be index attached to key sentences of these section units.

The structure of sections and paragraphs is known to be useful for extracting key sentences from text materials, because most of the key sentences appear at the beginning of the articles or sections. In audio materials, however, there is no explicit definition of sections and paragraphs such as the line-breaks and indentation of 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 presumptively extracts them without any manually tagged information such as topics and boundaries.

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[5]. The *Corpus of Spontaneous Japanese (CSJ)*[6] 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, we use 2592 presentations whose 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, we use 2496 presentations that amount to 486 hour speech. We set up a gender-independent triphone model that has 3000 shared states with 16 Gaussian mixture components. We also revised our recognition engine Julius so that very long speech can be handled without prior segmentation[7].

With the baseline system, the word error rate is 30.9% for the test-set of 15 academic presentation speeches[5]. Adaptation of acoustic and language models based on the initial recognition result together with the speaking-rate dependent decoding strategy[8] improves it to 21.9%, which is the best figure for this test-set ever reported.

The recognition system does not output periods as the language model is trained with the CSJ transcriptions which are not segmented into sentences and contain no periods. Thus, the recognition results need to be segmented into sentences. In read speech, a long pause is regarded as a mark of the end of utterances, thus, it can be converted to a period or comma. In spontaneous speech, however, this assumption does not hold. Speakers put pauses at arbitrary places. Therefore, we proposed a statistical translation framework that converts pauses to periods selectively[9]. The method achieved higher segmentation accuracy than the conventional methods.

3. AUTOMATIC INDEXING OF KEY-SENTENCES

We address automatic extraction of key sentences, which will be useful indices in lectures. Collection of these sentences may suffice summarization of the talk[10]. The framework extracts a set of natural sentences, which can be aligned with audio segments for

0-7803-8484-9/04/$20.00 ©2004 IEEE

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[11].

3.1. 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[12]. 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 numbered (sub-)sections in the proceedings paper.

It is also observed that there is a typical pattern in the initial 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". This phenomenon also reflects that key sentences in lectures often appear at the beginning of sections. We define such characteristic expressions that appear at the beginning of section units as discourse markers. Unlike previous studies, where discourse markers are manually defined based on linguistic analysis, our method automatically derives a set of discourse markers without any manual tags. We have shown its effectiveness in segmentation of the lecture audio[13].

The boundary of sections is known as useful for extracting key sentences in the text-based natural language processing[14]. 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.

3.2. 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. This threshold value differs 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 (similar to document frequency in information retrieval) 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.

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

Here, N_s is the total number of sentences in all lectures. A set of discourse markers is statistically selected according to $S_{DM}(w_j)$.

3.3. 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, 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 a specified number (or ratio) of sentences from the whole lecture.

3.4. Combination with Keyword-based Method

The other approach to extraction of key sentences 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.

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

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 combining $S_{KW}(s_i)$ with $S_{DM}(s_i)$ of 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. EXPERIMENTAL RESULTS

4.1. Preliminary Evaluation

For the preliminary evaluation, we used a set of fourteen presentations and had one human subject selected key sentences. The ratio of the key sentences among the overall sentences is 21.6% (=233/1077). For the evaluation measures, we use 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. The proposed method using the discourse markers was evaluated when 30% of the 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 was extracted from both the beginning and end of the whole lecture, which corresponds to the introduction and conclusion, 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 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 between the preceding and following pauses is calculated and converted into a measure of importance after the $N(0, 1)$ normalization. The recall rate was only 31.3%. Thus, the proposed method is shown to be more effective in detecting section boundaries and extracting key sentences.

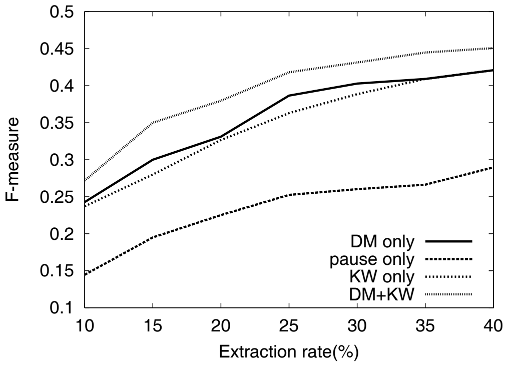

Fig. 1. Extraction performance of key sentences using discourse markers (DM) and keywords (KW)

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 a synergetic effect when combined.

4.2. Evaluation with the CSJ Key Sentence Set

Then, we did an evaluation by using another set of data. For part of the CSJ, key sentences labeled by human subjects will be included in the final corpus. In this work, we made use of those available as of August 2003. Specifically, key sentences were labeled by three human subjects for nineteen academic presentation speeches. The subjects were researchers in linguistics, thus they were familiar with the academic presentation style, but were not professionals in the area of most of the test-set. They were instructed to select sentences which seemed important by 50% of all, and then 10% from those 50%.

First, we investigated the agreement among the three subjects in indexing. Here, the agreement by two persons is the average of all combinations of the three. While relatively a higher agreement is observed in the 50% extraction, it is much harder to get agreement in the 10% extraction. Apparently, the task of selecting 10% is more difficult and the annotation is more subjective. As a result, the number of agreed sentences becomes very small (3-4%). Therefore, we set up experiments based on the agreed portion of the 50% extraction data for reliable and meaningful evaluation. Specifically, we picked up sets of sentences agreed upon by two subjects. Since three combinations exist for picking up two subjects out of three, we derived three answer sets. The performance is evaluated by averaging for these three sets. They amount to 37.5% of all sentences on the average. Using this scheme, we can also estimate the human performance by matching one subject's selection with the answer set derived from the other two. The recall, precision and F-measure are 83.2%, 62.7% and 0.715, respectively. These figures are regarded as a target for the proposed system.

The proposed method based on the discourse markers (DM) and its combination with the keyword-based method (KW) were evaluated on this test-set. The indexing performance of key sentences for the correct transcriptions is listed in Table 1. We confirmed much the same tendency as in Fig. 1. Although the method using the discourse marker alone was comparable to the keyword-based method, the synergetic effect of their combination was clearly verified.

Table 1. Results of key sentence extraction from manual transcription

method	recall	precision	F-measure
DM	71.0%	53.3%	0.609
KW	71.7%	53.8%	0.614
DM+KW	74.0%	55.5%	0.635
human	83.2%	62.7%	0.715

DM: discourse marker (proposed), KW: keyword

Table 2. Results of key sentence extraction from ASR results

	transcript.	segment.	recall	precision	F-measure
(1)	manual	manual	74.0%	55.5%	0.635
(2)	manual	auto	73.1%	45.8%	0.563
(3)	auto	auto	72.7%	45.6%	0.561

transcript.: transcription, segment.: segmentation

When we compare the system performance against human judgment, the accuracy by the system is lower by about 10%. The proposed method performs reasonably, but it still has room for improvement.

4.3. Evaluation with ASR Results

We also made an evaluation using the transcriptions generated by the automatic speech recognition (ASR) system. The indexing method is based on the discourse marker and keyword combination (DM+KW).

Table 2 lists the recall, precision rates and F-measure in comparison with the case of manual transcription. Here, we also tested the case where the sentence segmentation or period insertion is done automatically on the manual transcription to see individual effects. Since the derived sets of sentences for automatic and manual segmentation are different, we automatically align the hypothesized sentences with the correct ones, and calculate accuracy based on the alignment.

Comparing the cases (1) and (2) in Table 2, it is observed that the automatic segmentation has a bad effect on accuracy, especially on the precision. On the other hand, no degradation is observed by adopting automatic speech recognition regardless of the word error rate of 23%. These results demonstrate that the statistical evaluation of the importance of the sentences is robust.

The detailed results for the individual lectures in the test-set are listed in Table 3. Here, the indexing accuracy (F-measure) of the key sentences is shown with the word recognition accuracy and the segmentation accuracy (=F-measure of period insertion).

5. INCORPORATION OF PROSODIC FEATURES FOR KEY SENTENCE EXTRACTION

As we observed that sentence segmentation is more critical in indexing, we investigate its enhancement by using prosodic features. The features are also used for section boundary detection.

Our period insertion (=sentence segmentation) algorithm was totally statistical based on lexical and pause information. We first introduce some linguistic heuristics, and then incorporate prosodic features. As prosodic features, we use fundamental frequency

Table 3. List of test-set lectures with speech recognition accuracy, segmentation performance, indexing performance

lecture ID	recognition accuracy	segmentation accuracy	indexing accuracy
A01M0056	85.15%	0.821	0.458
A01M0096	91.21%	0.812	0.567
A01M0151	92.21%	0.920	0.656
A01M0035	64.95%	0.505	0.529
A01M0007	78.32%	0.613	0.533
A01F0001	77.56%	0.851	0.559
A01M0025	92.18%	0.878	0.671
A01M0110	86.15%	0.915	0.598
A01F0132	87.15%	0.794	0.495
A01M0083	91.35%	0.822	0.580
A01M0137	72.74%	0.740	0.561
A01M0074	80.54%	0.745	0.484
A01M0097	84.76%	0.844	0.536
A03M0112	81.41%	0.912	0.630
A03M0106	61.37%	0.720	0.489
A03F0072	71.31%	0.735	0.591
A05M0031	74.68%	0.783	0.629
A06M0134	68.58%	0.643	0.606
YG99JUN001	69.17%	0.512	0.501
total	76.99%	0.740	0.561

(F0), which is relatively higher at the beginning of Japanese sentences. We pre-select sentence boundary candidates with linguistics heuristics without relying on pause information, and then compute prosodic and linguistic scores. The prosodic score is derived from a difference of average F0, which is normalized by $N(0, 1)$, before and after the possible boundary. The linguistic score, which was used in the baseline period insertion algorithm, is defined as a difference of the trigram language model likelihoods with and without a period inserted[9]. These two scores are linearly combined for judgment of sentence boundaries.

For section boundary detection, we use F0 and power onset based on the assumption that speakers emphasize the beginning of sections to attract audiences' attention. For each sentence, F0 and power onset are computed and converted to a measure of importance. This measure is then linearly combined with other measures of importance computed by the discourse markers and keywords.

The results of sentence segmentation and key sentence extraction are shown in Table 4. The indexing accuracy is improved to 0.583 from 0.561 by the enhanced sentence segmentation. Moreover, by incorporating the prosodic features, we achieved the key sentence extraction rate (F-measure) of 0.592 for automatic speech recognition results. The figure is close to the case of manual transcriptions given. The improvement (0.561 → 0.592) is statistically significant.

6. CONCLUSIONS

We have presented an automatic key sentence extraction 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. It realizes comparable performance to the conventional keyword-based method. Moreover, the combination of the two methods significantly improves accuracy because they focus on different characteristics in a lecture. We also investigate the incorporation of prosodic features and confirm its effectiveness.

Table 4. Results of key sentence extraction incorporating prosodic features (ASR results)

indexing method	segmentation method	segmentation accuracy	indexing accuracy
DM+KW	baseline	0.740	0.561
DM+KW	revised	0.759	0.583
DM+KW+PROSODY	revised	0.759	0.592

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 for the collaboration in this fruitful project.

7. REFERENCES

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

[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," *ICASSP*, 1995, pp. 309–312.

[3] S.Whittaker, J.Choi, J.Hirschberg, and C.H.Nakatani, "What you see is (almost) what you hear: Design principles for user interfaces for accessing speech archives," *ICSLP*, 1998, pp. 2355–2358.

[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," *ICASSP*, 2001, pp. 597–600.

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

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

[7] T.Kawahara, H.Nanjo, and S.Furui, "Automatic transcription of spontaneous lecture speech," *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001.

[8] H.Nanjo and T.Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," *ICASSP*, 2002, pp. 725–728.

[9] H.Nanjo, K.Shitaoka, and T.Kawahara, "Automatic transformation of lecture transcription into document style using statistical framework," *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 215–218.

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

[11] C.Hori, S.Furui, R.Malkin, H.Yu, and A.Waibel, "Automatic speech summarization applied to English broadcast news speech," *ICASSP*, 2002, pp. 9–12.

[12] S.Teufel and M.Moens, "Summarizing scientific articles: Experiments with relevance and rhetorical status," *Computational Linguistics*, vol. 28, no. 4, pp. 409–445, 2002.

[13] T.Kawahara and M.Hasegawa, "Automatic indexing of lecture speech by extracting topic-independent discourse markers," *ICASSP*, 2002, pp. 1–4.

[14] C-Y.Lin and E.H.Hovy, "Identifying topics by position," *Applied Natural Language Conference*, 1997, pp. 283–290.