Automatic Extraction of Key Sentences from Oral Presentations using Statistical Measure based on Discourse Markers

Tasuku Kitade† *Hiroaki Nanjo*‡ *Tatsuya Kawahara*†

†School of Informatics, Kyoto University
Sakyo-ku, Kyoto, 606-8501, Japan
‡Faculty of Science and Technology, Ryukoku University
Otsu, 520-2194, Japan

Abstract

Automatic extraction of key sentences from academic presentation speeches 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 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. Comprehensive evaluation using the Corpus of Spontaneous Japanese and a variety of experimental setups is presented in this paper. We carefully designed the evaluation scheme to be compared to human performance. The proposed method using the discourse markers shows consistent effectiveness in the key sentence extraction. Based on the indexing, we realize efficient browsing of lecture audio archives.

1. Introduction

Recent progress of large-volume storage devices and high-speed networks has enabled digital archiving and streaming of audio and video materials. In academic societies and universities, multi-media archives of lectures will be technically feasible. Such archives would help students audit lectures at their convenient time and places with their own paces. In these kinds of audio archives, appropriate indices are necessary for efficient browsing and searching portions of specific topics or speakers.

We have studied automatic indexing of presentation audio archives by detecting section boundaries and extracting key sentences in a statistical framework[1]. Unlike conventional approaches, we focus on discourse markers, which are rather topic independent. We define discourse markers as expressions frequently used at the beginning of sections in presentations. Then, we define the importance of the sentences based on discourse markers, and combine this measure with the measure based on topic words. In the earlier report[1], we demonstrated the effectiveness of the proposed approach using the preliminary data set.

In this paper, we present comprehensive evaluation using the *Corpus of Spontaneous Speech (CSJ)*[2], which is complete and released to public this spring. As the extraction of key sentences is somehow subjective, we carefully designed the evaluation scheme, so that the results can be compared to human performance without depending on the ratio of selected sentences.

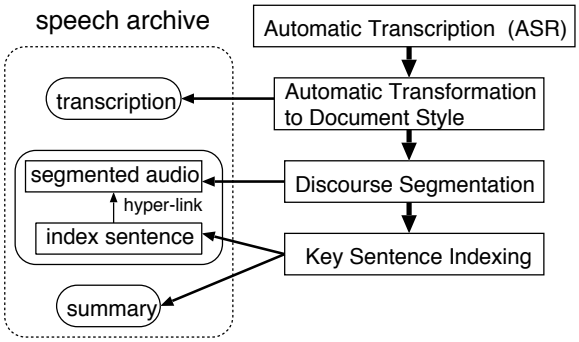

Figure 1: System overview of lecture archiving

The proposed technique constitute a key component in the automatic lecture archiving system we are developing.

2. Overview of Lecture Archiving

We approach the problem of indexing lecture audio archives by assuming a discourse structure of 'sections' and automatically detecting their boundaries. We focus on 'discourse markers', which are rather topic independent and defined as expressions characteristic of the beginning of new sections. Then, from each section we extract key sentences that can be used as content-based tags for the corresponding audio segments. The alignment of audio segments and transcription is also obtained as the result of automatic speech recognition.

Based on the approach, we are developing an intelligent lecture archiving system. An overview of the system is depicted in Figure 1. First, whole speech is automatically transcribed by an automatic speech recognition (ASR) system. The transcription is automatically transformed to document-style sentences for improved readability. Then, the discourse segmentation into section units is performed and key sentences are indexed for each section. Collection of these sentences might also suffice a summary of the talk[3]. In the generated archive, the index sentences are hyper-linked with the segmented audio for easy browsing. The example of browsing is shown at the end of this

paper (Figure 2).

We have taken part in the project of "Spontaneous Speech Corpus and Processing Technology" sponsored by the Science and Technology Agency Priority Program in Japan[4]. The CSJ developed by the project consists of roughly 7M words or 500 hours, which is the largest in scale and provided us with an infrastructure for our ASR system as well as this study.

3. Automatic Indexing of Key Sentences

Next, we describe automatic extraction of key sentences, which will be useful indices in oral presentations. The framework extracts a set of natural sentences, which can be aligned with audio segments for alternative 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 a summary based on the ASR result[5].

3.1. Discourse Modeling of Oral Presentation

In this work, we mainly deal with oral presentations at technical conferences. There is a relatively clear prototype in the flow of presentation, which is similarly observed in technical papers[6]. When using slides for presentation, one or a couple of slides constitute a topic discourse unit we call 'section'. 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, let's move on to experimental evaluation". This phenomenon also suggests that key sentences in presentations 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 derives a set of discourse markers by automatic training without any manual tags. We have shown its effectiveness in segmentation of the presentation audio[7].

The boundary of section 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 discourse segmentation together with extraction of key sentences.

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 of each section 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. Word frequency is used to represent the former property and sentence frequency is used for the latter. For a word w_j, the word frequency wf_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 presentations that contain the word. We adopt the following evaluation function.

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

Here, N_s is the total number of sentences in all presentations. 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 sentence 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 presentation.

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 presentation. The most orthodox statistical measure to define and extract such keywords is the following tf-idf criterion.

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

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

Then, we introduce a new measure of importance by combining $S_{KW}(s_i)$ with $S_{DM}(s_i)$. These two scores are combined by taking their geometric mean. Though a value of the weight α is chosen empirically, the final performance is not so sensitive unless extreme values are used.

$$S_{final}(s_i) = S_{DM}(s_i)^{\alpha} * S_{KW}(s_i)^{(1-\alpha)}$$

Table 1: Agreement among subjects in key sentence extraction

	by 2 persons	by 3 persons
50% extraction	75.5%(0.463)	62.7%
10% extraction	46.6%(0.314)	30.8%

* The figures inside represent the κ-value.

Table 2: Human performance of key sentence extraction

answer set	recall	precision	F-measure	κ-value
50-2AND	81.5%	60.1%	0.692	0.463
10-2OR	38.2%	60.1%	0.467	0.391

4. Evaluation using the CSJ

4.1. Evaluation Scheme

We make use of the set of key sentences included as a part of the CSJ. In this work, we picked up 21 academic presentation speeches that have been also used for ASR evaluation[8]. Each presentation lasts approximately 15 minutes. The key sentences were labeled by three human subjects. 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 20% from those 50% (= 10% of all).

First, we investigate the agreement among the three subjects in selecting key sentences (Table 1). The agreement by two persons is the average of all combinations of the three. We also compute the κ-value, which is often used to measure agreement by considering the chance rate. Thus, it enables comparison between 10% and 50% cases. While a relatively higher agreement is observed in the 50% extraction, it is harder to get agreement in the 10% extraction, and the number of agreed sentences becomes very small. Apparently, the task of selecting 10% is difficult and the annotation is subjective.

Since the key sentence indexing is subjective as observed above, it is desirable that the results of system's performance can be compared with human performance exactly in the same condition. Therefore, we design an evaluation scheme by defining an answer set using one or two subject's selections and estimating human performance by matching it with the left-out person's selection. For that scheme, we can consider following six variations for deriving answer sets: three sets each for 50% and 10% cases.

50-2AND : set of sentences agreed upon by two subjects in 50% extraction (37.0% of all).

50-2OR : set of sentences picked up by either of two subjects in 50% extraction (63.4%).

50-1 : set of sentences picked up by one subject in 50% extraction (50.2%).

10-2AND : set of sentences agreed upon by two subjects in 10% extraction (4.4%).

10-2OR : set of sentences picked up by either of two subjects in 10% extraction (16.1%).

10-1 : set of sentences picked up by one subject in 10% extraction (10.3%).

The set 50-2OR has so many sentences (63.4% of all), so it is not appropriate for indexing. On the other hand, the set 10-2AND has too few answers (4.4%), so it is not adequate, either. As we consider the appropriate ratio of indexed sentences for browsing audio archives is 20-40%, we adopt 50-2AND and 10-2OR sets, which realizes compression ratio of 37.0% and 16.1%, respectively.

Since three combinations exist for picking up two subjects out of three, the performance is evaluated by averaging for these three sets for each case. We 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 listed in Table 2. F-measure is a normalized mean of recall and precision rates. Table 2 also shows κ-value, which indicates the agreement of the selection with the answer set. These figures are regarded as a target for the proposed system.

Table 3: Results of key sentence extraction from manual transcription (answer set: 50-2AND)

method	recall	precision	F-measure	κ-value
DM	69.9%	51.7%	0.594	0.297
KW	70.4%	52.0%	0.598	0.304
DM+KW	72.4%	53.5%	0.615	0.333
human	81.5%	60.1%	0.692	0.463

DM: discourse marker (proposed), KW: keyword

Table 4: Results of key sentence extraction from manual transcription (answer set: 10-2OR, 20% extraction)

method	recall	precision	F-measure	κ-value
DM	35.2%	28.0%	0.312	0.162
KW	37.7%	30.0%	0.334	0.189
DM+KW	39.2%	31.1%	0.347	0.205
human *	38.2%	60.1%	0.467	0.391

DM: discourse marker (proposed), KW: keyword
*: 10% extraction

4.2. Evaluation with Manual Transcriptions

The proposed method based on the discourse markers (DM) and its combination with the keyword-based method (KW) were evaluated on this scheme. The indexing performance of key sentences for manual transcriptions is listed in Table 3 in the case of 50% extraction (50-2AND). Although the method using the discourse marker alone was comparable to the keyword-based method, the combination effect was clearly observed. When we compare the system performance against human judgement, the accuracy (F-measure) by the system is lower by about 10%. The proposed method performs reasonably, but it still has room for improvement.

Next, we evaluate the proposed method using the answer set 10-2OR. Since the ratio of answer key sentences is 16% and correct sentences should not be missed, we extract 20% instead of 10% in this case. The results are listed in Table 4. Same tendency as Table 3 is observed and combination effect is verified. Compared with human judgement of the 10% extraction, the recall rate is much the same, while the precision rate by the system is about half. That means correct indexing is realized with much the same number of redundant ones. Still, it will be useful for the indexing purpose.

Table 5: Results of key sentence extraction from ASR result (answer set: 50-2AND)

	transcript / segment	recall	precision	F-measure	κ-value
(1)	manual / manual	72.4%	53.5%	0.615	0.333
(2)	manual / auto	72.7%	46.5%	0.567	0.216
(3)	auto / auto	76.1%	45.5%	0.569	0.186

Table 6: Results of key sentence extraction from ASR result (answer set: 10-2OR, 20% extraction)

	transcript / segment	recall	precision	F-measure	κ-value
(1)	manual / manual	39.2%	31.1%	0.347	0.205
(2)	manual / auto	42.3%	26.4%	0.325	0.160
(3)	auto / auto	43.7%	24.1%	0.311	0.124

4.3. Evaluation with ASR Result

Then, we evaluate the indexing method using the transcriptions generated by the automatic speech recognition (ASR) system. The ASR system is set up using the CSJ. The acoustic model is a gender and style dependent PTM triphone model consisting of 25K Gaussian components and 576K mixture weights. A trigram language model is trained for the vocabulary of 24K words. The word error rate for the test-set is 30.3% with the baseline system.

Then, we applied our indexing method based on the discourse marker and keyword combination (DM+KW).

Table 5 and Table 6 lists the recall, precision rates, F-measure and κ-value in comparison with the case of manual transcription in the 50% and 20% extraction, respectively. Here, we also tested the case where the sentence segmentation (period insertion) is done automatically on the manual transcription to see individual effects. Since the derived sets of sentences by 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 5 and Table 6, it is observed that the automatic sentence 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 30.3%. These results demonstrate that the statistical evaluation of the importance of the sentences is robust.

5. Conclusions

We have evaluated an automatic key sentence extraction method for audio archives of oral presentations. 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 presentation. It is also shown that the method is robust against ASR errors. The proposed indexing method constitutes a key component of our automatic lecture archiving system. The evaluation results presented in this paper show that the indexing performance is reasonable.

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.

6. References

[1] T.Kawahara, K.Shitaoka, T.Kitade, and H.Nanjo. Automatic indexing of key sentences for lecture archives. In *In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.

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

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

[4] 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*, Vol. 3, pp. 518–521, 2000.

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

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

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

[8] 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*, pp. 135–138, 2003.

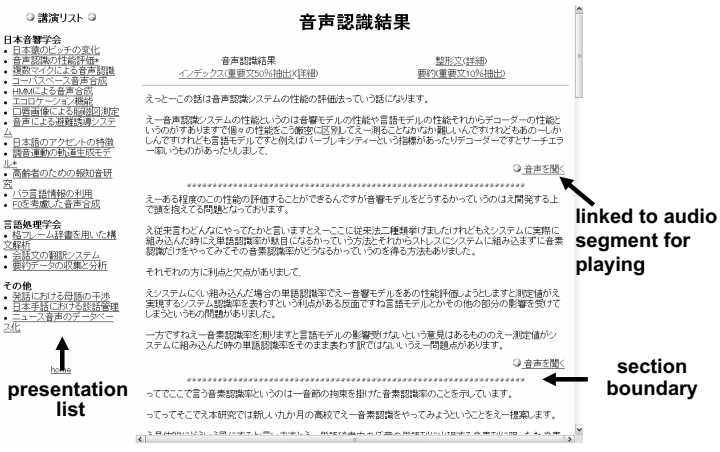

Figure 2: Archiving system of oral presentations