

Automatic Extraction of Cue Phrases for Important Sentences in Lecture Speech and Automatic Lecture Speech Summarization

Yasuhisa Fujii¹, Norihide Kitaoka², and Seiichi Nakagawa¹

¹Department of Information and Computer Sciences, Toyohashi University of Technology

²Department of Media Science Graduate School of Information Science, Nagoya University

fujii@slp.ics.tut.ac.jp, kitaoka@nagoya-u.jp, nakagawa@slp.ics.tut.ac.jp

Abstract

We automatically extract the summaries of spoken class lectures. This paper presents a novel method for sentence extraction-based automatic speech summarization.

We propose a technique that extracts “cue phrases for important sentences (CPs)” that often appear in important sentences. We formulate CP extraction as a labeling problem of word sequences and use Conditional Random Fields (CRF) [1] for labeling. Automatic summarization using CP extraction results as features yields precisions of 0.603 and 0.556 when using manual transcriptions and Automatic Speech Recognition (ASR) results, respectively.

Combining the features derived from the CPs and traditional features (including repeated words, words repeated in a slide text, and term frequency (tf), which are surface linguistic information, and speech power and duration, which are prosodic features) [2, 3], we obtained better summarization performance with a κ -value of 0.380, a F -measure of 0.539, and a *Rouge-4* of 0.709.

Index Terms: automatic speech summarization, sentence extraction, speech synthesis

1. Introduction

If we can index and summarize audio recordings, it becomes much easier to refer to them. Thus studies of automatic speech summarization have been focused on more than ever [4].

Waibel et al. reported a summarization method for meetings based on the detection and deletion of disfluency, the detection of sentence boundaries, and question-answering pairs [5]. Reithinger et al. reported automatic summarization on such tasks as hotel reservations based on statistical dialog act estimation [6]. Koumips et al. reported the importance of using words related to specified topics, proper nouns, dates, and times [7].

Hori et al. presented a summarizer based on such linguistic information as location in a parse tree, and their search used a dynamic programming (DP) technique. When summarizing Japanese news broadcasts, about 70% of the summarized sentences preserved the meanings of the original speech under the 60-70% summarization condition [8]. Their method, however, needs to use such higher-level knowledge as semantics to correct incorrectly summarized sentences.

Zechner presented a summarization that uses machine-learning techniques to identify and remove speech disfluencies [?, 9]. Their experiments showed that their system significantly outperformed the two baselines: the leading part of a speech (LEAD) and Maximum Marginal Relevance (MMR) [10].

Recently, it has become possible to use a lecture speech corpus from conferences and classes [11] that readily leads to spontaneous speech research from read speech.

Kikuchi et al. reported the summarization of lecture speech for conferences based on linguistic features. At a 70% summarization rate, summarization accuracy was almost identical as humans. At a 50% summarization rate, however, the summarizer’s accuracy did not approach humans [12].

Class lectures often last more than an hour, which is long, and therefore automatic summarization, indexing, and segmentation of lecture videos/speeches have been focused on. We summarized speech by automatically extracting important sentences from the transcription of lecture speech segmented to sentence-like units [2, 3]. We easily summarized speech using the results of important sentence extraction since each sentence-like unit was associated with a particular part of the speech.

In our previous works [2, 3], summarization was performed using prosodic and surface linguistic information. Prosodic information includes rate of speech, pitch, and power, and surface linguistic information includes frequency of words (TF) and inclusion of words in slides. Summarization that combined these features outperformed that using only one feature. But the result was still inferior to human summarization. Automatic summarization was also inferior to human efforts in subjective evaluations, in which subjects evaluated whether the summary contained the important points of the original lecture.

In this paper we propose a new technique that automatically extracts cue phrases (CPs) for important sentences, which often appear in important sentences, to improve performance. CPs are extracted word sequences by labeling in sentences. In this paper we use Conditional Random Fields (CRF) [1] for labeling and also show the results of combining the features obtained by the proposed method with the traditional features in a 25% summarization rate condition.

The remainder of this paper is organized as follows: Section 2 introduces sentence extraction methods investigated in our previous works using surface linguistic information and prosodic features. Section 3 presents a method for the extraction of CPs, and Section 4 describes experimental conditions and results. Finally, Section 5 concludes this paper.

2. Sentence Extraction Method

A set of relatively important sentences are extracted, as described in Fig. 1. First, features that determine the sentence’s importance are extracted. After that, sentences are classified as important or unimportant.

Details of feature extraction and classification are described in Sections 2.1 and 2.2, respectively.

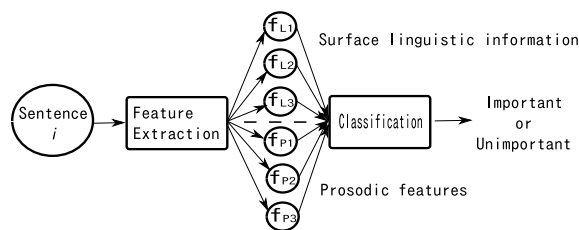


Figure 1 Process of important sentence extraction

10.21437/Interspeech.2007-722

2.1. Feature extraction

Features extracted from sentences consist of two types of groups: surface linguistic features (Section 2.1.1) and prosodic features (Section 2.1.2).

2.1.1. Surface linguistic information

Surface linguistic information means such ‘shallow’ information as cue words or phrases, high frequency words, and sentence location. We examined the following features extracted from manual or ASR transcriptions:

- **Repeated words**
Repeated words denote those that appear frequently in transcriptions and are only nouns except fillers or unimportant words. To find repeated words, we used *ChaSen* [13], a Japanese morphological analyzer, and extracted sentences that included two or more repeated words.
- **Words repeated in slide texts (PPT)**
Class lectures were usually performed using PowerPoint slides (PPT). Therefore, words repeated in slide texts ($WORD_{srpt}$) and words in slide captions ($WORD_s$) may be good cues. We used the inclusion of such words as a feature.
- **tf (term frequency)**
The 1/4 sentences of a whole lecture are also extracted by a simple tf-based text summarizer, Posum [14]. In the experiment, we only counted nouns except for unimportant words and fillers (TF). A TF -based method was used as the baseline for automatic summarization.

2.1.2. Prosodic features

Summarization accuracy can be improved by using prosodic features. We used F_0 , power, duration, and pauses. The following features were comparatively useful among a variety of feature parameters [2]:

- **Power (POW_{avg})**
Sentences having a higher POW than ‘Average $\overline{POW} + S.D.$ ’, where S.D. denotes *standard deviation*.
- **Duration (LEN)**
Sentences having longer duration times
- **Rate of utterance ($SPEED$)**
Sentences having higher rates of utterance

2.2. Classification

The score of sentence i is calculated as follows:

$$Score(S_i) = \mathbf{w}\mathbf{x} + b, \quad (1)$$

where \mathbf{w} is a vector of the weight of each feature, \mathbf{x} is a vector of the value of each feature, and b expresses bias. \mathbf{w} is estimated using SVM, which classifies \mathbf{x} based on margin maximization¹. Each element of \mathbf{x} is a binary or continuous value. When a feature expresses the inclusion of words (e.g., repeated words), it is represented by a binary value, whereas the feature has a continuous value with 0 mean and 1 standard deviation by normalization when representing the others (e.g., tf, Power).

3. Cue Phrase Extraction

If we can find expressions, or word sequences, called Cue Phrases (CPs) that often appear in important sentences and rarely in unimportant ones, they may be good cues to extract important sentences. In other words, if a CP appears in a sentence, the sentence is probably important. Since there are individual

¹We used linear kernel for SVM. Theoretically we can use more complex kernels for our purpose, but we do not due to a lack of training data

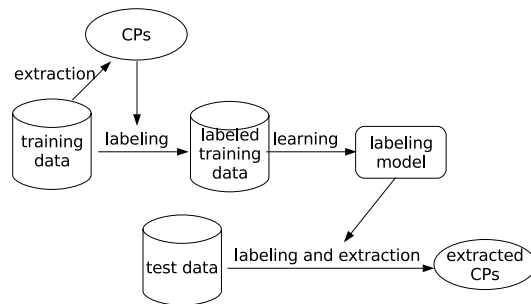


Figure 2 Procedure of cue phrases for important sentences extraction

differences in phrasing, CPs may differ among speakers or lecturers. So CPs found in lectures by a speaker may not be useful for others. There may be speaker-independent/domain word-independent rules for such CPs (in other words, abstracted POS sequences) even though the word sequences are different from each other. In this paper, we propose to acquire the general rules of CPs using a machine learning technique. CPs are extracted by a labeling problem of word sequences in sentences. The labeling rules generating CPs are learned using Conditional Random Fields (CRF).

3.1. Conditional Random Fields (CRF)

CRF [1] is a framework for building probabilistic models to segment and label sequence data. The probability of label sequence \mathbf{y} for input sequence \mathbf{x} is written as:

$$P(\mathbf{y}|\mathbf{x}) = \frac{\exp(\Theta, \Phi(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y} \in Y} \exp(\Theta, \Phi(\mathbf{x}, \mathbf{y}))}, \quad (2)$$

where Θ denotes the importance of each feature and $\Phi(\mathbf{x}, \mathbf{y})$ is a vector consisting of the counts of each feature.

3.2. Cue phrase extraction method

CPs are extracted as shown in Fig. 2. First, they are determined from the training data set with labels. Then, labeling rules are acquired by using CRF. In the test phase, CPs are extracted from the data by labeling using the CRF.

3.2.1. CP extraction from training data and labeling

To train the CRF, we need a training data set with important/unimportant labels for each sentence. For each word sequence that consists of more than two words unnecessary to be continuous, and for constituents within an eight-word length window in a sentence, we count the appearances in important (N_i) and unimportant sentences (N_u). If the word sequence appears in important sentences at least Th_N times (that is, $N_i > Th_N$) and probability R of the sentence in which the word sequence appears is more than Th_R (that is, $R = N_i / (N_i + N_u) > Th_R$), the word sequence is determined to be a *cue phrase*.

After all the cue phrases determined by the above process are listed, the phrase including the most words is detected as the final CP, and it is labeled using the labels shown in Table 1. Labeling CPs in training data are shown in Figure 3.

3.2.2. Training CP labeler

A CRF is trained using the training data labeled by the procedure described in Section 3.2.1. Figure 4 shows the graphical representation of CRF. The state-state feature between neighboring states (transition features) and three state-observation features between a state and each of previous, current, and next words (observation features) are used as features in CRF. Only

label	meaning
0	unimportant words
1	head words of CP
2	middle words of CP
3	tail words of CP
-1	unimportant words in CP (skipped words)

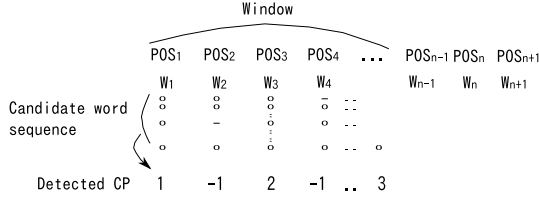


Figure 3 Labeling CPs in training data

POS information is used as an observation feature when the word is a noun. Using POS rather than the word itself, we expect CRF generalization to treat unknown sequences that do not appear in the training data, that is, for other topics/domains.

3.2.3. Extracting CPs using CRF

By applying CRF to the test data, the word sequences determined to be CPs are labeled by the sequences of 1, 2, and 3, optionally including the skipped words, -1. Figure 5 shows an example of extracted CPs.

4. Experiments

4.1. Experimental conditions

In our experiments, we used eight lectures performed by four teachers. These lectures were related to spoken language processing, multi-modal interface, pattern recognition, and natural language processing. Table 2 (a) shows the statistics of the data. Each lecture was about 70 minutes long containing about 1000 sentences.

All speech data were transcribed by humans and an ASR system. SPOJUS, a Japanese ASR system, was used [15], and a two-pass decoder with syllable HMMs and a trigram language model was trained using the CSJ Corpus [16].

The recognition performance of the lecture speech ranged from 31.0 to 57.5% in word accuracy, as shown in Table 2 (b).

4-fold cross variation was used to evaluate the CP extraction method. When using lectures by a speaker, the lectures given by the other three were used as training data.

We set Th_R and Th_N in 3.2.1 to 0.75 and 10, respectively.

4.2. Reference of automatic summarization

Important sentence extraction was conducted by six speech research experts who understood the extremely high-level content of the targeted graduate.

All subjects were instructed to mark the important sentences in the speech transcriptions as “important” 1/4 of the sentences of the whole speech (rate of summary, 25%) in each lecture. In other words, every sentence was classified as important

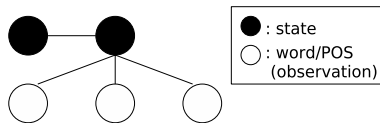


Figure 4 Graphical representation of CRF

aruiha ... “noun” ga	(Japanese phrase)
Alternatively, “noun” is ...	(English phrase)

Figure 5 An example of extracted CPs

or not. Here “sentence” is defined as a portion between pauses longer than about 200 msec. We also prepared reference data called **man3/6**, which corresponds to the sentences extracted by equal or more than three subjects out of six. Man3/6, which reflects the consensus of the subjects, was used as a reference. The target value of summarization shown in Table 2 is the averages of the scores obtained between each subject and **man3/5** (sentences extracted by three or more subjects out of all but the evaluated subject).

4.3. Evaluation measures

We used κ -value [17], precision, F -measure, and a Rouge [18] metric to measure summarization performance. These are defined as follows:

- **Precision and Recall**

$$Precision = \frac{|M \cap H|}{|M|}, \quad Recall = \frac{|M \cap H|}{|H|}.$$

Here, H and M are the sets of sentences extracted by humans and machines, respectively. In this paper, we used precision.

- **F -measure**

F -measure is defined as the harmonic average of precision and recall:

$$F\text{-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- **κ -value [17]**

The κ -value is a measure that adjusts the agreement frequency of judgment by two subjects in consideration of a chance agreement.

- **ROUGE-N [18]**

This includes measures that automatically determine the quality of an automatically generated summary by comparing it to reference summaries created by humans. ROUGE has some variations, but in this paper, we use the basic ROUGE-N metric computed as follows with 4-gram statistics:

$$ROUGE - N = \frac{\sum_{S \in \{Ref-Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Ref-Summaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Table 2 Details of speech materials

Lecture	Duration	No. of Sent.	Target value of summarization		
			κ	F	Rouge-4
NS-1	67'56"	742	0.462	0.595	0.632
NS-2	54'59"	719	0.491	0.613	0.633
KN-1	65'49"	680	0.474	0.599	0.547
KN-2	71'14"	1099	0.450	0.579	0.638
NT-1	69'28"	582	0.493	0.617	0.618
NT-2	78'30"	648	0.320	0.447	0.527
AT-1	70'02"	1749	0.454	0.586	0.615
AT-2	65'23"	1571	0.477	0.605	0.592
average	67'55"	974	0.453	0.580	0.600

Lecture	Accuracy [%]	Correct [%]	
NS-1	47.4	55.6	$Accuracy = 100 -$
NS-2	31.0	37.0	
KN-1	54.9	60.8	$Substitution - Deletion$
KN-2	50.7	58.9	
NT-1	48.8	54.8	$Correct = 100 -$
NT-2	45.0	55.2	
AT-1	57.1	61.4	$Substitution - Deletion$
AT-2	57.5	62.5	

Table 3 Important sentence extraction results based on CP extraction (precision)

Lecture	Using transcriptions by human			Using ASR transcriptions		
	Ext.	Imp.	Prec.	Ext.	Imp.	Prec.
NS-1	112	58	0.518	82	40	0.488
NS-2	85	48	0.565	59	32	0.542
KN-1	46	34	0.739	43	26	0.605
KN-2	24	17	0.708	29	20	0.690
NT-1	81	40	0.494	47	23	0.489
NT-2	95	44	0.463	85	37	0.435
AT-1	57	33	0.579	33	18	0.545
AT-2	37	28	0.757	35	23	0.657
average	67.1	37.8	0.603	51.63	27.38	0.556

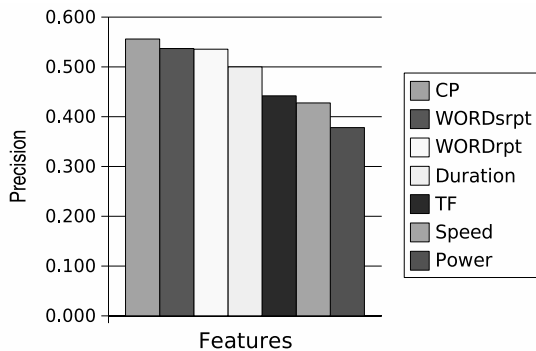


Figure 6 Summarization precision based on each feature (ASR transcriptions, summarization rate=25%).

4.4. Important sentence extraction results based on CP extraction results

In this section, we extract the sentences, which include word sequences, determined to be important CPs by our CP labeler. Table 3 shows the summarization results, where average precision using human transcriptions was 0.603 and using ASR transcriptions was 0.556. Figure 6 compares the summarization precision using individual features. The CP-based method using ASR transcriptions outperformed the others.

4.5. Important sentence extraction results

We combine our new CP-based feature with the others using SVM. Table 4 shows the results with features explained in Section 2.1 (prev.) and with those and CP features (with CP). We obtained improvement for many lectures by adding CP features, and the average results are $\kappa = 0.380$, F -measure = 0.539, and $Rouge-4 = 0.709$ with ASR transcriptions.

Table 5 summarizes the ASR transcription results. The summarization results using previous features outperformed TF (baseline), and the summarization results using the new feature outperformed summarization using previous features. But the final result is still inferior to humans when evaluated by κ -value and F -measure, while the result outperformed one of the humans when evaluated by $Rouge-4$.

5. Conclusion

In this paper, we proposed a novel feature for automatic summarization based on Cue Phrase (CP) extraction with CRF for important sentences. Summarization with this technique yields precisions of 0.603 and 0.556 for human transcription and ASR results, respectively. We also combined this CP-based feature with previous features and obtained a better result with a κ -value of 0.380, a F -measure of 0.539, and a $Rouge-4$ of 0.709.

So far we have only used the features for each sentence itself. In the future, we will adopt the features to express relations among sentences.

Table 4 Important sentence extraction results by feature combination (ASR transcriptions)

Lecture	κ -value		F -measure		Rouge-4	
	prev.	with CP	prev.	with CP	prev.	with CP
NS-1	0.294	0.294	0.480	0.480	0.733	0.733
NS-2	0.343	0.350	0.512	0.518	0.720	0.729
KN-1	0.439	0.447	0.587	0.592	0.735	0.735
KN-2	0.472	0.476	0.606	0.610	0.795	0.797
NT-1	0.315	0.324	0.495	0.502	0.650	0.660
NT-2	0.365	0.365	0.510	0.510	0.677	0.675
AT-1	0.343	0.349	0.515	0.519	0.620	0.621
AT-2	0.428	0.431	0.578	0.580	0.730	0.725
average	0.375	0.380	0.535	0.539	0.708	0.709

Table 5 Summarizer results by a combination model (ASR transcriptions)

	κ -value	F -measure	$Rouge-4$
TF (baseline)	0.230	0.427	0.466
using previous features [3]	0.375	0.535	0.708
+ new feature	0.380	0.539	0.709
human	0.453	0.580	0.600

6. References

- [1] F. Pereira, J. Lafferty, and A. McCallum. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.
- [2] S. Kobayashi, N. Yoshikawa, and N. Nakagawa. Extracting summarization of lectures based on linguistic surface and prosodic information. *SSRP*, pp. 211–214, 2003.
- [3] S. Togashi, M. Yamaguchi, and S. Nakagawa. Summarization of spoken lectures based on linguistic surface and prosodic information. *IEEE/ACL Workshop on Spoken Language Technology*, pp. 34–37, 12 2006.
- [4] A. Nenkova. Summarization evaluation for text and speech: issues and approaches. *ICSLP*, pp.1527–1530, 2006.
- [5] A. Waibel, M. Bett, F. Metz, K. Ries, T. Schaaf, H. Soltan, H. Yu, and K. Zechner. Advances in automatic meeting record creation and access. *ICASSP*, pp. 597–600, 2001.
- [6] N. Reithinger, M. Kipp, R. Engel, and Alexandersson. Summarizing multilingual spoken negotiation dialogues. In *ACL*, pp. 310–317, 2000.
- [7] K. Koumpis, S. Renals, and M. Miranjan. Extractive summarization of voicemail using lexical and prosodic feature subset selection. *EuroSpeech*, pp. 2377–2380, 2001.
- [8] C. Hori and S. Furui. Automatic speech summarization based on word significance and linguistic likelihood. *ICASSP*, pp. 1759–1582, 2000.
- [9] K. Zechner. Automatic summarization of spoken dialogues in unrestricted domains. Technical report, CMU-LTI-01-168, 2001.
- [10] J. Carbonell, Y. Geng, and J. Goldstein. Automated query-relevant summarization and diversity-based reranking. *AI and Digital Libraries*, pp. 9–14, 1997.
- [11] J. Glass et al. Analysis and processing of lecture audio data: preliminary investigations. in *HLT-NAACL 2004 Workshop: Interdisciplinary Approaches to Speech Indexing and Retrieval*, pp. 9–12, 2004.
- [12] T. Kikuchi, S. Furui, and C. Hori. Automatic speech summarization based on sentence extraction and compaction. *ICASSP*, pp. 384–387, 2003.
- [13] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. *Morphological Analysis System ChaSen 2.3.3 Users Manual*. Nara Institute of Science and Technology, <http://chasen.naist.jp/hiki/ChaSen/>, 2003.
- [14] H. Mochizuki. *Automatically Text Summarizer Posum, version 1.50.2*. Japan Advanced Institute of Science and Technology <http://www.tufs.ac.jp/ts/personal/motizuki/software/posumcl/>, 2002.
- [15] A. Kai, Y. Hirose, and S. Nakagawa. Continuous speech recognition using segmental unit input hmms with a mixture of probability density functions and context dependency. *ICSLP*, pp. 2935–2938, 1998.
- [16] S. Furui, K. Maekawa, and H. Isahara. A Japanese national project on spontaneous speech corpus and processing technology. *Proc. ASR2000*, pp. 244–248, 9 2000.
- [17] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, pp. 378–382, 1971.
- [18] C. Lyn and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. *The Human Language Technology Conference*, pp. 71–78, 2003.