Topic Segmentation of News Speech Using Word Similarity

Seiichi Takao
Dept. of Electronics and Informatics
Ryukoku University
Seta, Otsu-shi, 520-2194, Japan
tail@arikilab.elec.ryukoku.ac.jp

Jun Ogata
Dept. of Electronics and Informatics
Ryukoku University
Seta, Otsu-shi, 520-2194, Japan
ogata@arikilab.elec.ryukoku.ac.jp

Yasuo Ariki
Dept. of Electronics and Informatics
Ryukoku University
Seta, Otsu-shi, 520-2194, Japan
ariki@rins.ryukoku.ac.jp

ABSTRACT

Conventional topic segmentation utilizes *cosine* measure as the similarity between consecutive passages. However, the *cosine* measure has a problem that it can not reflect the similarity unless exactly the same words are included in the passages. To solve this problem, in this paper, we propose a method to acquire the word similarity between different words from the input data directly and automatically by managing to collect the same topic sections. Further more, we propose a method to compute the passage similarity based on the word similarity. Finally we propose a method of topic segmentation based on the passage similarity in an unsupervised mode.

1. INTRODUCTION

Recently, many news programs are broadcast owing to its digitization. In this situation, viewers require to quickly select and watch his interesting news stories. From this viewpoint, news on demand systems have been developed[1][2]. In the systems, it is required to segment news programs into individual news story automatically using mainly news speech data, because manual segmentation is almost impossible due to a large amount of news programs.

In automatic topic segmentation [3]-[4], a small passage is compared with the successive passage and if the similarity between them is smaller than some threshold, the topic boundary is found between them. As the similarity measure, *cosine* is usually utilized between two vectors which are produced from the consecutive passages by counting the frequencies of the important words as the vector elements. However, the *cosine* measure has a problem that it can not reflect the similarity unless exactly the same words are included in the passages. To solve this problem, word similarity between different words is required before computing the *cosine* similarity.

In order to acquire the word similarity automatically, a large amount of training data is usually required. However, newly input data (evaluation data or test data) is different from the training data so that the word similarity acquired from the training data is sometimes useless. To solve this problem, it is effective to acquire the word similarity directly from input data, not from the training data. For that purpose, time sections, where the same topic continues, have to be collected from the input data. We define this time sections as *topic sections*.

In the context of news program, the topic sections correspond with the time sections where the same video caption continues. From this viewpoint, in this paper, we propose a method to acquire the word similarity from the input data directly by collecting the topic sections. Further more, we propose a method to compute the passage similarity based on the word similarity. Finally a method of topic segmentation is proposed based on the passage similarity in an unsupervised mode.

In this paper, we describe the conventional unsupervised topic segmentation in section 2. A proposed method of topic segmentation on the basis of passage similarity and word similarity is described in 3. Speech transcription and video caption detection are described in section 4 and 5. Finally the experimental results will be described in 6.

2. CONVENTIONAL SEGMENTATION

2.1 Outline

In conventional methods for topic segmentation, a small passage is compared with the successive passage and if their similarity is less than some threshold, the topic is segmented at the passage. This method is called an unsupervised topic segmentation. In the context of news program, the process is summarized as follows;

1. Speech transcription is carried out for Japanese broadcast continuous news speech.
2. Topic vectors in each passage (analytical window) is constructed by using term weighting method and the analytical window is shifted.
3. The similarity between consecutive topic vectors, constructed from the analytical windows, is computed.
4. Topic boundary is detected where similarity is lower than some threshold.

2.2 Topic Vectors for Passages

Word frequencies in each analytical window are counted. Then weighting degree is computed for each word by using

a term weighting method. We describe this method later. Finally, a vector for each analytical window is constructed by extracting words showing higher weighting degree than a threshold. We call this vector as *topic vector*.

The term weighting method we employed is mutual information incorporating TF-IDF due to the following reason. The conventional mutual information shows high value even when occurrences of word w_i are low, because mutual information is computed based on probability, not on word frequency. This is a weak point of the conventional mutual information.

On the other hand, TF counts the occurrences of word w_i and compensates for a weak point of the conventional mutual information. IDF shows the degree how word w_i depends on topic t_k by counting the number of topics which include word w_i. Therefore it can be said that IDF shows the co-occurrence of word w_i and topic t_k in different viewpoint from mutual information. This is the reason we employed here the combination of the mutual information and TF-IDF as shown in Eq.(1) to extract important words. We call this method as mutual information incorporating TF-IDF.

$$\begin{aligned} & i(t_k; w_i) \times (TF - IDF) \\ &= (i(t_k) - i(t_k|w_i)) \cdot TF(w_i, t_k) \cdot IDF(w_i) \\ &= (\log \frac{P(t_k, w_i)}{P(t_k)P(w_i)}) \cdot TF(w_i, t_k) \cdot IDF(w_i) \end{aligned} \quad (1)$$

2.3 Similarity Between Topic Vectors

A topic vector can be created from each analytical window. The elements of the topic vector are important words selected using the term weighting method (mutual information incorporating TF-IDF) already mentioned. Therefore an inner-product of the topic vectors can now be used to find the vocabulary overlap between any two topic vectors. Eq.(2) shows the similarity between two topic vectors X_k and X_l.

$$\begin{aligned} \cos\theta &= (X_k, X_l) \\ &= (x_{1k}, x_{2k}, \cdots, x_{nk})(x_{1l}, x_{2l}, \cdots, x_{nl})^T \\ &= (x_{1kc}, x_{2kc}, \cdots, x_{nkc})(x_{1lc}, x_{2lc}, \cdots, x_{nlc})^T \\ &= \sum_i x_{ikc} \cdot x_{ilc} \end{aligned} \quad (2)$$

where x_{ikc} and x_{ilc} are the normalized frequency of words appearing in both X_k and X_l. If $\cos\theta$ nearly equals 1, the similarity between two passages (analytical windows) is regarded as high.

After the similarity between consecutive passages is computed by using Eq.(2), topic boundary is detected as the point where the similarity is lower than some threshold.

3. PROPOSED SEGMENTATION

3.1 Problem of Conventional Methods

The *cosine* measure used in the conventional topic segmentation has a problem that it can not reflect the similarity unless exactly the same words are included in the passages. To solve this problem, similarity between different words has to be computed before the *cosine* similarity computation.

In order to obtain this type of word similarity automatically, a large amount of training data is usually required. However, newly input data (evaluation data or test data) is usually different from the training data so that the word similarity obtained from the training data is sometimes useless. To solve this problem, it is effective to compute the word similarity directly from input data, not from the training data. For that purpose, time sections, where the same topic continues, have to be collected from the input data. We define this time sections as *topic sections*.

In the context of news program, the topic sections correspond with the time sections where the same video caption continues. From this viewpoint, in this paper, we propose a method to compute the word similarity from the input data by collecting the *topic sections*.

3.2 Word Distance and Word Similarity

We propose, in this paper, word distance in a word space to compute the word similarity. The word space is three dimensional, constructed by values of mutual information, TF and IDF. The word distance between word w_i and w_j is computed as follows;

$$\begin{aligned} WD(w_i, w_j) &= \frac{1}{m} \sum_m ((TF(w_i, t_m) - TF(w_j, t_m))^2 \\ &\quad + (IDF(w_i) - IDF(w_j))^2 \\ &\quad + (i(t_m; w_i) - i(t_m; w_j))^2)^{\frac{1}{2}} \end{aligned} \quad (3)$$

In Eq.(3), $TF(w_i, t_m)$ shows the term frequency that word w_i occurs in *topic section* t_m, $IDF(w_i)$ shows inverse document frequency of the word w_i, $i(t_m; w_i)$ shows mutual information of word w_i in *topic section* t_m. The m also shows the number of *topic sections*. Then the word distance shows the distance between word w_i and w_j in all *topic sections* in the word space (Mutual-TF-IDF). The word similarity is computed as the inverse of the word distance.

3.3 Passage Similarity

Eq.(2) counts only the number of overlapping words, but it doesn't take into consideration of the word similarity. This causes the decreasing of topic segmentation performance in the case where the analytical window is short, because similarity between consecutive analytical windows can not be correctly computed by using only word overlapping shown in Eq.(2). If the analytical window is long, this problem may be solved. But it becomes difficult to detect the topic boundaries precisely.

To solve this problem, similarity between consecutive analytical windows (passage similarities) has to be computed by using not only word overlapping but also word similarity. The passage similarity based on the word similarity can be computed using Eq.(4).

$$(X_k, X_l) = \sum_i \sum_j x_{ik} \times x_{jl} \times \frac{1}{WD(w_i, w_j)} \quad (4)$$

where x_{ik} and x_{jl} are the normalized frequency of word w_i and w_j in passage t_k and t_l respectively as shown in Eq.(2).

4. SPEECH TRANSCRIPTION

4.1 Experimental Condition

We carried out automatic speech transcription[5] for the NHK Japanese broadcast continuous news speech, using a

language model and an acoustic model. The language model is the word bigram constructed from RWC text database which was produced by morphologically analyzing the MAINICHI Japanese newspaper of 45 months from 1991 to 1994. The number of the words in the dictionary is 20,000. The word bigram was back-off smoothed after cutting off at 1 word.

Speaker independent cross-word triphone HMMs were constructed. They were trained using 21,782 sentences spoken by 137 Japanese males. These speech data is taken from the database of acoustic society of Japan. The acoustic parameters are 39 MFCCs with 12 Mel cepstrum, log energy and their first and second order derivatives. Cepstrum mean normalization was applied to each sentence to remove the difference of input circumstances. Table1 shows the experimental conditions for acoustic analysis (AA) and HMM.

In the transcription experiment, we used HTK (HMM Toolkit) as the decoder which can perform Viterbi decoding with beam search using above mentioned language model and acoustic model.

Table 1: Acoustic Analysis(AA) and HMM

A	Sampling frequency	12kHz
	High-pass filter	$1 - 0.97z^{-1}$
A	Feature parameter	MFCC,Pow,Δ,$\Delta\Delta$ (39th)
A	Frame length	20ms
	Frame shift	5ms
	Window type	Hamming window
H	Learning method	Concatenated training
M	Type	Left to right continuous HMM
M	Number of states	5 states with 3 loops
M	Number of mixtures	8

4.2 Transcription Result

The automatic transcription was carried out for the NHK Japanese continuous news speech for 20 minutes in 1998. We show the transcription result in Table 2. In the table, the "Corr" indicates the correctness and the "Acc" indicates the accuracy of the speech transcription.

The reason why the transcription result is a little lower is explained as follows. The language model was constructed from the MAINICHI Japanese newspaper published from 1991 to 1994. On the other hand, the test data was NHK spoken news broadcast in 1998. This time difference caused the lower transcription result. This transcription result is used for topic segmentation.

Table 2: Transcription result(%)

	Corr	Acc
19980820-12:00NHK	77.83	75.57

5. DETECTION OF TOPIC SECTION

In this section, we describe a method [6] to correctly detect the *topic sections* where the same video caption continues. When the characters of video captions appear, edge correspondence between present frame and next frame shows high value. Therefore edge ratio defined by Eq.(5) changes along time at the following three sections; the character appearing section, character stable section and character disappearing section.

Edge ratio equals almost 1 in character stable section. Consequently, the *topic section* is determined by extracting the character stable section. Experimental results of topic section detection is shown in Table.3.

$$Edge\ ratio = \frac{Edge\ correspondence(present, next)}{Edge\ correspondence(present, previous)} \quad (5)$$

Table 3: Detection of topic sections

	Correct	Accuracy
19980820-12:00NHK	70.1% (47/67)	75.8% (47/62)

6. EXPERIMENTAL RESULTS

We compared two topic segmentation techniques for the transcription results shown in Table2, conventional unsupervised method and proposed topic segmentation method using the word similarity. Experimental results are shown in Table.4. The measure to evaluate the topic segmentation is recall, precision and F-measure. F-measure is the combination of recall and precision as defined in Eq.(6). If evaluated by F-measure, the proposed topic segmentation on the basis of word similarity showed about 10% superiority to conventional topic segmentation. From Table4, it can be said that the computation of word similarity and passage similarity using the *topic sections* in the test data is effective in topic segmentation.

Table 4: Topic Segmentation Results

	recall	precision	F-measure
Conventional	80.00%	41.37%	54.54%
Proposed	73.33%	57.89%	64.70%

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (6)$$

7. CONCLUSION

In this paper, we proposed a new topic segmentation method based on passage similarity and word similarity in a word space which is constructed from the *topic sections* in the test data itself. The proposal method showed about 10% superiority to the conventional topic segmentation method.

8. REFERENCES

[1] F.Kubala, A.Colbath, D.Lie, A.Srivastava and J.Makhoul: "Integrated technologies for indexing spoken language", Communications of the ACM, Vol.43, No.2, pp.48-56, 2000.

[2] SRI MAESTRO Team: "MAESTRO:conductor of multimedia analysis technologies", Communications of the ACM, Vol.43, No.2, pp.57-63, 2000.

[3] J.P.Yamron, I.Carp, L.Gillick, S.Lowe, and P.van Mulbregt: "A Hidden Markov Model Approach to Text Segmentation and Event Tracking", ICASSP98, Volume I, pp.333-336, 1998.

[4] P.van Mulbregt, I.Carp, L.Gillick, S.Lowe and J.Yamron: "Text Segmentation and Topic Tracking on Broadcast News Via A Hidden Markov Model Approach", ICSLP98, Volume VI, pp.2519-2522, 1998.

[5] Y.Ariki and J.Ogata, "Indexing and Classification of TV News Articles Based on Speech Dictation Using Word Bigram", ICSLP98, Volume 7, pp.3265-3268.

[6] Y.Ariki and S.Takao, "Extraction and Recognition of Open Captions Superimposed on TV News Articles", to appear in ACCV00, 2000.