INDEXING AND CLASSIFICATION OF TV NEWS ARTICLES BASED ON SPEECH DICTATION USING WORD BIGRAM

J.Ogata and Y.Ariki

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

ABSTRACT

In order to construct a news database with a function of video on demand (VOD), it is required to classify news articles into topics. In this paper, we propose a method to automatically index and classify TV news articles into 10 topics based on a speech dictation techniques using speaker independent triphone HMMs and word bigram.

1. INTRODUCTION

Recently, TV news programs are broadcasted from all over the world owing to the broadcast digitization. In this situation, TV viewers require to select and watch the most interesting news. For that purpose, word indexing and article classification are key techniques.

The word indexing is a technique to put the discriminative keywords on the news speech articles in order to retrieve the interesting articles. On the other hand, the article classification is a technique to classify the news speech articles into groups (topics) based on their contents such as politics, economy, science and sports in order to retrieve the same kind of articles[1]. These two techniques are strongly required because manual indexing and classification are almost impossible.

From this viewpoint, we propose in this paper a method to automatically index and classify TV news articles into 10 topics based on a speech dictation technique using speaker independent triphone HMMs and word bigram. After the dictation of the spoken news articles, pre-defined keywords are searched and given to the new articles as indices. Then the articles are classified based on the indices.

The keywords are selected as the words which has strong association with the topics. In order to compute the association between the keywords and the topics, we used χ^2 value and selected the words with high χ^2 value as the keywords.

Each keyword has a topic contribution probability indicating what to degree this keyword contributes to classify the article into the topics. It is computed using "classification indices for ASAHI newspaper article database". By multiplying this topic contribution probability with the acoustic probability obtained from the new speech dictation and by summing them over all the keywords within the article, the topic probability is computed. The article is classified into the topic with the highest topic probability.

2. SPEECH DICTATION

2.1. Experimental Condition

We carried out speech dictation for the 55 NHK news articles 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 and 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 is applied to each sentence to remove the difference of input circumstances. Table1 shows the experimental conditions for acoustic analysis (AA) and H-MM.

In the dictation experiment, we used HTK (HMM Toolkit)[2] as the decoder which can perform Viterbi decoding with beam search using above mentioned language model and acoustic model. The acoustic probability of the keyword is computed by back tracking the Viterbi forward-computation and then extracting the time section of the keywords.

Table 1: Acoustic Analysis(AA) and HMM

	Sampling frequency	12kHz
	High-pass filter	$1 - 0.97z^{-1}$
A	Feature parameter	MFCC,Pow,Δ, ΔΔ (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 mixtures	8

2.2. Dictation Result

The dictation was carried out for 55 NHK news articles extracted from the broadcasted articles during 1993 and 1994. The total number of time duration was 1.13 hours and each article took about 71 seconds at average. The sentence pauses were automatically extracted by detecting the silence lasting more than 1 second. Table2 shows the property of the 55 news articles. These news articles are closed to the training data of the language model in terms of the collection time. The dictation result is shown in table3. In the table, word correct rate and word accuracy are defined as follows;

$$\text{Word correct rate} \;=\; \frac{N - S - D}{N} \cdot 100 \qquad (1)$$

$$\text{Word accuracy} \;=\; \frac{N - S - D - I}{N} \cdot 100 \qquad (2)$$

S : The number of substituted words

D : The number of deleted words

I : The number of inserted words

N : Total number of words

The word error rate is defined as $(100 - word\ accuracy)$. The perplexity is rather than low so that the word accuracy is 80.3%. This dictation result is used for keyword indexing and also topic classification in the successive process.

Table 2: News articles used for dictation

20K unknown word ratio	0.8%
test-set perplexity	78.3

Table 3: Dictation result(%)

Word error rate	19.7
Word correct rate	85.6
Word accuracy	80.3

3. INDEXING OF NEWS ARTICLES

3.1. Keyword Selection

Keywords play an important role in topic classification and affect the classification rate so that they have to be selected based on some theory. Keywords have strong association with topics. In order to compute the association between the keywords and the topics, we used χ^2 value shown in Eq.(3) and selected the words with high χ^2 value as the keywords.

In computing the χ^2 value, a part (1993 and 1994) of RWC text database was used which was produced by morphologically analyzing the MAINICHI Japanese newspaper of 45 months from 1991 and 1994. At first, each article of RWC text database was given a correct topic name by the method described in 5.. Using the text articles and their correct topic names, χ^2 values were computed for all the nouns included in the articles. Then the nouns higher than some threshold were selected as the keywords.

$$\begin{aligned} \chi^2_{i,j} &= \frac{(x_{i,j} - m_{i,j})^2}{m_{i,j}} \\ &= \frac{\sum_{j=1}^{n} x_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{i,j}} \times \sum_{i=1}^{m} x_{i,j} \qquad (3) \end{aligned}$$

m : The number of different words

n : The number of topics

$x_{i,j}$: Frequency of word i in topic j

$m_{i,j}$: Predicted frequency of word i in topic j

3.2. Result of Keyword Selection

Table4 shows the result of keyword selection by changing the threshold (T) to χ^2 value. In the table, "Key1" indicates the number of keywords whose χ^2 value is grater than threshold T among the nouns included in the training RWC text database. On the other hand, "Key2" indicates the number of keywords included in the "classification indices for ASAHI newspaper article database" among the keywords selected as "Key1". The "Correct" indicates the ratio of the number of correctly extracted keywords after the dictation to the total number of keywords included in the 55 NHK news speech articles. The number of keywords is mentioned within the parenthesis in Table4.

4. CLASSIFICATION OF NEWS ARTICLES

4.1. Classification Flow

Fig.1 shows the classification flow of news speech articles by speech dictation technique. Before the article classifi-

Table 4: Result of keyword selection

T	Key1	Key2	Correct
0	26998	6230	83.2%(2061/2476)
10	14648	4265	83.7%(1860/2223)
50	4299	2176	85.4%(1460/1709)
100	2140	1303	87.2%(1120/1284)
200	1043	702	86.8%(781/ 900)
500	346	264	85.4%(363/ 425)

cation, news articles in the news program are automatically separated each other using the algorithm mentioned in [3]. In the flow, there are following two phases;

(1) Speech dictation phase: word sequence and their probabilities $Ps(w)$ are computed by applying the speech dictation technique.

(2) Article classification phase: articles are classified by integrating the keyword probability $Ps(w)$ and the topic contribution probability of each word $P(n|w)$.

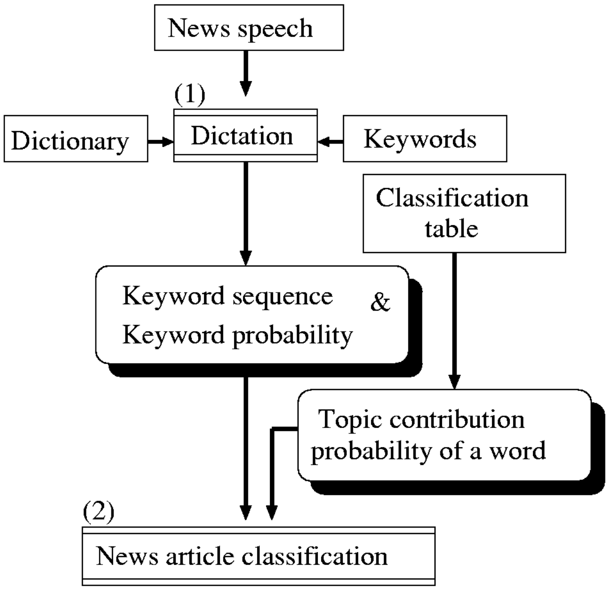

Figure 1: Flow of news article classification

4.2. Word Probability

Speech dictation by HTK described in 2. can produce the word probability as well as the word sequence. After speech dictation for the 55 news articles, keywords w_i are searched together with their probability $P_s(w_i)$. The keywords were determined in advance as the words included in both "classification indices for ASAHI newspaper article database" and 55 news articles.

4.3. Topic Contribution Probability

In article classification using a keyword sequence, topic contribution of a keyword (TCKW) $P(n|w)$ is computed in advance as shown in Eq.(4).

The TCKW indicates how the keyword contributes to identify the topic and is defined as the ratio of the occurrence of the keyword w included in the topic n to the occurrence of the keyword w included in all the topics. This definition is *a posteriori* probability of the topic n conditioned by the word w.

$$P(n|w) = \frac{\left\{\begin{array}{l}\text{The number of occurrence of the}\\ \text{keyword } w \text{ included in the topic}\\ n\end{array}\right\}}{\left\{\begin{array}{l}\text{The number of occurrence of the}\\ \text{keyword } w \text{ included in all the}\\ \text{topics}\end{array}\right\}} \tag{4}$$

In this study, we used "classification indices for ASAHI newspaper article database" in computing the TCKW.

It includes 12,000 keywords and they have links to the related topics. There are three levels in grouping of the topics; coarse, middle and fine level. They have about 10, 92 and 737 kinds of topics respectively. In the fine topics, about 16 indices are prepared at average in each topic. We selected coarse level of 10 topics classification in this study. The 10 topics for classification are Politics, Economy, Labor, Culture, Science, Society, Accidents, Sports, Internationality and Others.

Table5 shows an example of how to compute the TCKW. In the table, there are three complex words which include "Japan-U.S." word in the topic of politics. In the same way, there are two and zero in the topic of economy and society respectively. In total the number of complex words including "Japan-U.S." is five. In this example, the TCKW of the "Japan-U.S." word is computed as follows;

$$P(\text{Politics}|\text{Japan-U.S.}) = \frac{3}{5} = 0.6$$
$$P(\text{Economy}|\text{Japan-U.S.}) = \frac{2}{5} = 0.4$$
$$P(\text{Society}|\text{Japan-U.S.}) = \frac{0}{5} = 0.0$$

The TCKW is computed for all the keywords in advance.

4.4. Topic Probability

The topic probability $P(n|w_1, \cdots, w_k)$ that the article is classified into the topic n after the extraction of the keywords $w_1, \cdots, w_k$ is shown in Eq.(5).

$$P(n|w_1, \cdots, w_k) = \sum_{i=1,\cdots,k} P(w_i) \times P(n|w_i) \tag{5}$$

Table 5: Example of topics and keywords

Topic	Japan-U.S.	total
Politics	Japan-U.S. security treaty	3
	Japan-U.S. administrative agreement	
	Japan-U.S. relation	
Economy	Japan-U.S. economic friction	2
	Japan-U.S. trade friction	
Society		0

where $P(n|w_i)$ is the topic contribution probability of the keyword w_i. The probability $P(w_i)$ is replaced by the normalized word probability as follows;

$$P(w_i) = \frac{Ps(w_i)}{\sum\limits_{j=1,\cdots,k} Ps(w_j)} \tag{6}$$

This topic probability is the integration of acoustic word probability $Ps(w)$ and *apriori* knowledge probability TCK-W. The news article can be classified into a topic with the highest topic probability $P(n|w_1, \cdots, w_k)$.

5. CLASSIFICATION RESULT

Table6 shows the classification result of the 55 news articles. In the table, the number of keywords "Key1" and "Key2" selected by thresholding χ^2 value are mentioned. The classification rate is the ratio of the number of correctly classified articles to the total number of articles. The article is judged to be correctly classified if it is classified into the correct topic. The correct topic is determined by setting the word probability $Ps(w_i) = 1$ for the true keywords which are obtained from the text data.

From the table, it can be seen that the highest classification rate 81.8% was obtained at the threshold 100. When the threshold is going high, the classification rate becomes high. This means that the keywords with high χ^2 value contribute to the article classification. However when the threshold goes too high, the classification rate decreases. This indicates that the number of keywords is too reduced. The classification rate in Table6 is well associated with the "correct" which indicates the ratio of the correctly extracted keywords mentioned in Table4.

6. CONCLUSION

We have described the automatic classification system of TV news articles. Keywords were extracted from news speech articles after their dictation using word bigram and speaker independent triphone HMMs. The acoustic probabilities of the keywords were multiplied with the topic contribution probabilities which are computed from "classification indices for ASAHI newspaper article database" and the topic probability of the article is produced. The news speech articles were classified based on this topic probability.

Table 6: Result of article classification

T	Key1	Key2	Classification rate
0	26998	6230	78.2%
10	14648	4265	80.0%
50	4299	2176	80.0%
100	2140	1303	81.8%
200	1043	702	78.2%
500	346	264	69.1%

The word accuracy of news speech dictation was 80.3% and article classification rate was 81.8%. We are planning to improve the classification rate by using more effective method of keyword selection than we used. Further, we apply the proposed topic classification technique to the topic segmentation task which automatically divide the continuous news speech into topic segments.

7. REFERENCES

1. Y.Ariki, M.Sakurai and Y.Sugiyama : " Article Extraction and Classification of TV News Using Image and Speech Processing", CODAS96 (International Symposium on Cooperative Database Systems for Advanced Applications), pp.247-254, 1996.

2. Cambridge University Engineering Department Speech Group and Entropic Research Laboratory Inc.:"HTK Hidden Markov Model Toolkit V2.0"

3. Y.Ariki and Y.Saito : " Extraction of TV News Articles based on Scene Cut Detection", ICIP96, pp.III847-III850, 1996.