DISCOURSE SEGMENTATION FOR RADIO NEWS

Yoshimi Suzuki, Fumiyo Fukumoto, Yoshihiro Sekiguchi
Dept. of Electrical Engineering and Computer Science, Yamanashi Univ.
{ysuzuki@suwa, fukumoto@skye, sekiguti@saiko}.esi.yamanashi.ac.jp

1 INTRODUCTION

One of the typical problems in speech recognition of discourse which consists of different topics is to identify what kinds of topics are included in the discourse. There have been many studies for topic identification in information retrieval and speech recognition research [1, 2]. Yokoi reported a method of topic identification of news speech based on keyword spotting [3]. However, he used speech data which are segmented by each topic. Itoh reported a method for speech understanding and speech retrieval for TV program based on spotting algorithms [4]. However, he selected keyword manually.

We focus radio news for topic identification, and propose a method for news story segmentation into appropriate topics. While Ito's method selects keyword manually, our method used a method of automatic extraction of keywords with appropriate weights, and can reduce candidates of recognized sentences as pre-processing of speech recognition.

2 FEATURE VECTOR FOR TOPICS

Every corpus has a topic. In general, the topic in every corpus is characterized by words with high frequencies. For automatic extraction of keywords, frequencies of keywords are very useful factors.

However, the words which are frequently used do not always improve precision. Because frequent words tend to appear in many corpus, such words have a harmful influence for characterizing topic. To solve this problem, term weighting by using χ^2 [5] is introduced in our method.
$\boldsymbol{\chi}_i^2$ vector of $word_i$ is shown

$$\boldsymbol{\chi}_i^2 = (\chi_{i1}^2, \chi_{i2}^2, \cdots, \chi_{in}^2) \tag{1}$$

where,

$$\chi^2_{ij} = \begin{cases} \frac{(x_{ij}-m_{ij})^2}{m_{ij}} & x_{ij} > m_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$m_{ij} \;=\; \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \times \sum_{i=1}^{m} x_{ij} \tag{3}$$

(m : the number of nouns, n : the number of corpus, x_{ij} : frequency of $word_i$ in corpus j, m_{ij} : ideal frequency of $word_i$ in corpus j).
Ideal frequency means the frequency when the word appears in every corpus with the same frequency.

Table 1 shows examples of χ^2 vector. POL, ECO, INT, SPR and ACC means politics, economy, international, sports, and accident, respectively. In Table 1, feature vector of politics is characterized by the noun “Prime Minister”.

Table 1: Examples of χ^2 vectors

noun	feature vector				
	POL	ECO	INT	SPR	ACC
Prime Minister	142.26	0.00	0.87	0.00	0.00
President	0.00	0.00	254.42	0.00	0.00

The system recognizes topic of news story by using feature vector. Each topic is characterized by a feature vector whose coordinate is an m-dimensional Euclidean space, where m is the number of noun in a corpus. Each element of feature vectors is χ^2 value. Figure 1 shows χ^2 vector of $word_i$ $(1 \leq i \leq m)$ and feature vector of each topic in χ^2 matrix. The dotted circle shows feature vector of INT and circle shows χ^2 vector of $word_i$.

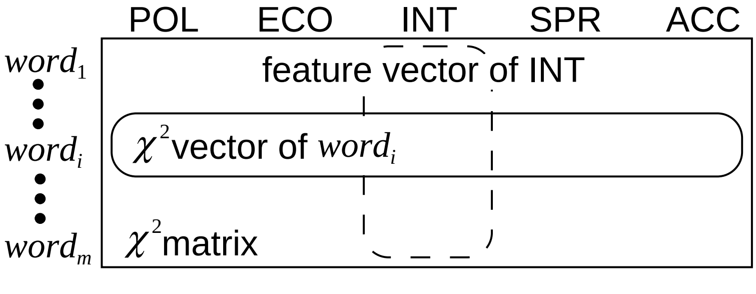

Figure 1: χ^2 vector and feature vector in χ^2 matrix

3 TOPIC IDENTIFICATION

Radio news stories which were used in our experiments were written in phoneme. It was divided by pause which are longer than 0.5 second. We call the phoneme sequence between pauses **unit**. For each **unit**, the system recognizes topic and when $unit_i$ and $unit_{i+1}$ have different topics, the system divide news story between $unit_i$ and $unit_{i+1}$. Input news stories are phoneme sequence without space and word boundary are ambiguous. Then the system uses word spotting technique, and select five word candidates in length of phoneme sequence order at each start point. Figure 2 shows example of word candidates. Then, it selects word combinations in order to make similarity value maximum at each topic.

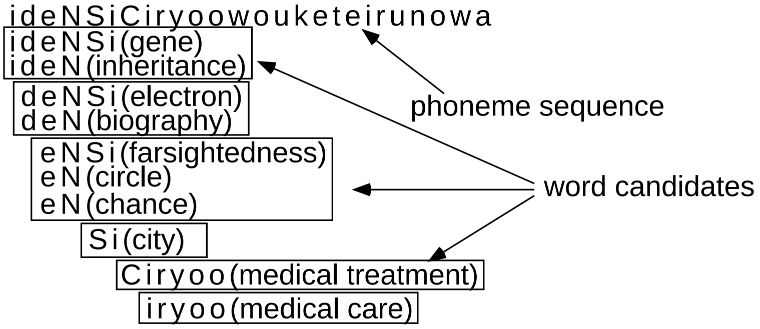

Figure 2: Example of word candidates

Let x_j be one of the topics x_1, $\cdots$, x_t and y be an **unit**. Let also the feature vector of x_j $(1 \leq j \leq t)$ be $X_j = (\chi^2_{1j}, \cdots, \chi^2_{mj})$, and the word frequency vector of y be Y. The similarity Sim between x_j and y is defined by formula (4).

$$Sim(x_j, y) \quad = \quad \frac{X_j * Y}{W} \tag{4}$$

W means the total number of intersection between $word_1$, $\cdots$, $word_m$, and the words in y. The system selects x_j if the value of $Sim(x_j, y)$ is maximum among all topics x_j $(1 \leq j \leq t)$. Every **unit** is assigned the topic. The system doesn't calculate similarity value for overlapped word combinations. Figure 3 shows topic identification method. In Figure 3, the value of similarity between the **unit** and POL is maximum among all topics. Then, the system selects POL as the topic of the **unit**.

4 EXPERIMENT

4.1 Data

The data used in the experiments is 13 days' radio news stories which are selected from NHK 6 o'clock radio news in August and September. One day news consist of about 15 stories on average.

243 articles of Mainichi Shinbun in 1994 from CD-ROM were collected to make feature vectors. There are 427 characters par an article on average. We manually sort the articles to five topics (POL,ECO,INT,SPR, and ACC). There are about 20,000 characters in each collected topic corpus.

In order to make feature vectors of each news story, 243 articles are tagged by part of speech using JUMAN [6]. As a result, we have got 534,932 nouns in all. From these, we selected 1,038 nouns of which frequency is larger than 5.

For data sparseness problem, the system can't calculate similarity value if there are no words which are selected from newspaper in the **unit** of news story. In order to cope with

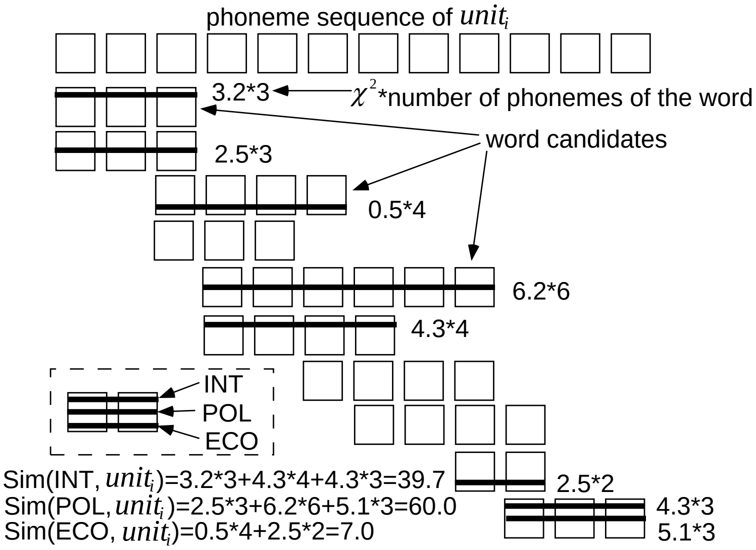
phoneme sequence of *unit_i*
3.2*3
χ^2*number of phonemes of the word
word candidates
2.5*3
0.5*4
6.2*6
4.3*4
INT
POL
ECO
Sim(INT, *unit_i*)=3.2*3+4.3*4+4.3*3=39.7
Sim(POL, *unit_i*)=2.5*3+6.2*6+5.1*3=60.0
Sim(ECO, *unit_i*)=0.5*4+2.5*2=7.0
2.5*2
4.3*3
5.1*3

Figure 3: Topic identification method

this problem, we estimate $\boldsymbol{\chi}^2_\alpha$ which α does not appear in the news story. Our estimating method is as follows: First, we collect 13,445 keywords using mutual information from the all articles of Mainichi Shinbun '94 CD-ROM. Each feature vector is extended to 13,445 dimensions. Then, we calculated mutual information value between each noun pair in the all articles of Mainichi Shinbun '94 CD-ROM. Next, we collected pairs (α,β) which β is stored in χ^2 matrix and α is not. For each (α,β),$\boldsymbol{\chi}^2_\alpha$ is estimated by using the following formula.

$$\boldsymbol{\chi}^2_\alpha \quad = \quad \frac{\sum_{i=1}^{m} \boldsymbol{\chi}^2_{\beta_i} * f(\alpha,\beta_i)}{\sum_{i=1}^{m} f(\alpha,\beta_i)} \qquad (5)$$

Here, m is the number of β. $f(\alpha,\beta_i)$ is co-occurrence between α and β_i in this order.

4.2 Results

Table 2 shows results of topic identification based on an **unit**. The number shows the number of **unit** of specific topic which the system identifies for radio news story. In Table 2, column shows news stories and rank shows topic name. For example, the system identifies 124 out of 154 POL as POL, and the correct ratio attained at 80.5%. There are four **units** which the system identifies POL as INT, and the incorrect ratio which the system recognized the POL stories as INT is 2.6%.

Table 3 shows results of topic identification based on phoneme. The numeral numbers are total length of phoneme of specific topic which the system identifies for radio news story. In Table 3, column shows news stories, and rank shows topic name.

The relationship between the similarity for each topic and the **unit** number is shown in Figure 4. The vertical lines are boundaries between stories. The topics of stories are ECO, SPR, SPR, ECO, and ECO, respectively.

Table 2: Results of topic identification (**unit**)

	radio news story										
	POL	(%)	ECO	(%)	INT	(%)	SPR	(%)	ACC	(%)	Total
POL	124	(80.5)	6	(4.3)	5	(5.2)	0	(0.0)	2	(1.3)	137
ECO	1	(0.6)	115	(82.7)	3	(3.1)	3	(2.9)	6	(4.0)	128
INT	4	(2.6)	2	(1.4)	75	(78.1)	0	(0.0)	2	(1.3)	83
SPR	15	(9.7)	13	(9.4)	6	(6.3)	101	(96.2)	26	(17.2)	161
ACC	10	(6.5)	3	(2.2)	7	(7.3)	1	(1.0)	115	(76.2)	136
Total	154		139		96		105		151		645

Table 3: Results of topic identification (phoneme)

	radio news story										
	POL	(%)	ECO	(%)	INT	(%)	SPR	(%)	ACC	(%)	Total
POL	16,203	(84.1)	565	(3.4)	571	(4.1)	0	(0.0)	228	(1.2)	17,567
ECO	93	(0.5)	14,406	(86.2)	424	(3.1)	324	(2.8)	531	(2.7)	15,778
INT	371	(1.9)	207	(1.2)	11,225	(81.0)	0	(0.0)	232	(1.2)	12,035
SPR	1,539	(8.0)	1,283	(7.7)	663	(4.8)	11,319	(96.9)	2,704	(14.0)	17508
ACC	1,062	(5.5)	259	(1.5)	968	(7.0)	37	(0.3)	15,636	(80.9)	17962
Total	19,268		16,720		13,851		11,680		19,331		80,850

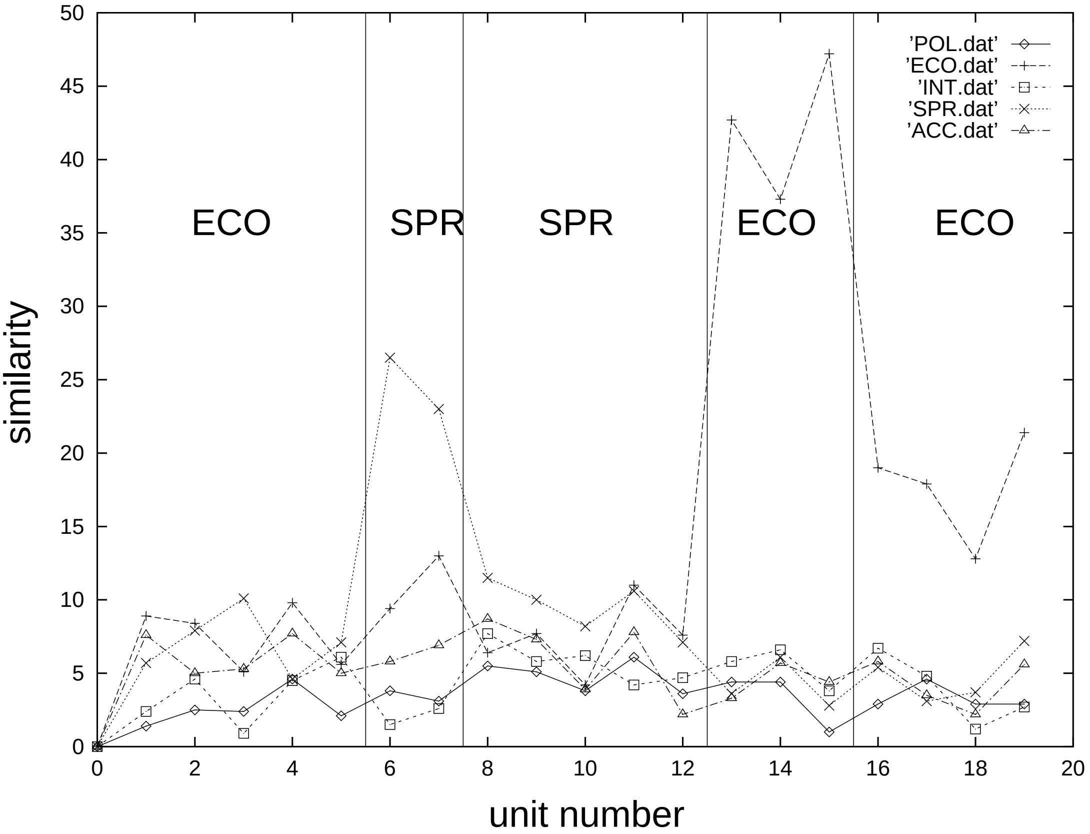

Figure 4: The relationship between the similarity for each topic and the **unit** number

5 DISCUSSION

In Table 2, the topic identification ratio based on **unit** was 82.2% (A/B = 82.2%, where A is sum of underlines in the table (530) and B is Total (645). In Table 3, the topic identification ratio based on phoneme was 85.1% (A/B = 85.1%, where A is sum of underlines in the table (68,789) and B is Total (80,850). Both ratios could achieve higher than 82%. It demonstrates the effect of our method. However, for example, **unit** based experiment, there were 115 **units** which could not be identified correctly.

The causes of the error are shown in the following. (1) The system frequently could not recognize the news stories which are not about sports as the sports news stories. This is because the word of tomorrow is frequently used in sports story of newspaper and in the feature vector, the word is a keyword of sports topic. On the other hand, in radio news the announcers frequently uttered " ました.". "ました." is polite auxiliary verb and it frequently used by announcers. "ました." was pronounced as "maSita". and the word of tomorrow was pronounced as "aSita". Tomorrow was frequently matched in the most of stories of radio news.
(2) In newspaper, United State are called as "米" and its phoneme sequence is "bei". And radio news said it as "amerika". Word of United State is frequently used in both newspaper and radio news. However, phoneme sequence of United State didn't match each other. In order to deal with this problem, we will investigate the method of linking of the word which has same meanings used in newspaper and used in radio news.

6 CONCLUSIONS

We have proposed a segmentation method for radio news. The results of experiment based on **unit** showed that 530 out of 645 topics could be identified correctly, and the percentage attained at 82%. In our current experiment, the input is all phenome sequences which are recognised correctly in advance. We will investigate how our method can use effectively for speech recognition system.

7 ACKNOWLEDGMENT

The authors thank for permission to use newspaper articles Mainichi Shinbun and to use radio news Japan Broadcasting Corporation (NHK).

References

[1] J.McDonough and H.Gish, *Proceedings of ICSLP* S36–10 (1994).

[2] S.Sekine, *Proceedings of COLING* page 913-918 (1996).

[3] K.Yokoi, T.Kawahara and S.Doshita, IPSJ Technical Report, SLP 6-3, (1995.05).

[4] Y.Itoh, J.Kiyama and R.Oka, ASJ Spring Meeting, 3-P-22(1995.03).

[5] Y.Watanabe, M.Murata, M.Takeuchi and M.Nagao, *Proceedings of COLING* page 794-799 (1996).

[6] Y.Matsumoto, S.Kurohashi, T.Utsuro, Y.Taeki and M.Nagao, "User's Guide for the Juman System Version 1.0" Nagao Lab. Electrical Engineering Kyoto Univ. (1993.04).