DOMAIN IDENTIFICATION AND KEYWORD EXTRACTION OF RADIO NEWS USING TERM WEIGHTING

Yoshimi Suzuki Fumiyo Fukumoto Yoshihiro Sekiguchi
Dept. of E.E. and C.S, Yamanashi University
Kofu, Japan

Abstract - In this paper, we propose a method for domain identification and keyword extraction using term weighting for radio news. In our method, feature vectors whose elements are χ^2 values between each keyword and each domain are calculated from newspaper articles automatically. Using the feature vectors, a domain of each part of radio news is selected. Then keywords are extracted by using the selected domain. The results of experiments show that our methods are robust and effective for the speech recognition system.

1 INTRODUCTION

Recently, many speech recognition systems are designed for various kinds of tasks. However, most of them are fixed for certain tasks, for example, a tourist information and a hamburger shop. The task which consists of various kinds of domains seems to be in demand for speech recognition systems (e.g., a dictation system for news and a minutes writing system for meetings).

In order to recognize spoken discourse which has several kinds of domains, the system has to have large vocabulary. However, the system can not achieve good word accuracy, since there are many words which have similar phoneme sequences with each other. In order to cope with this problem, N-gram models have been utilized for word selection from large vocabulary. However one of the problems using N-gram models is that very large corpus are necessary for recognizing discourse which consists of various domains.

We think that keyword extraction using term weighting is a breakthrough for speech recognition of discourse. Because it is robust in regard to phoneme misunderstanding and it is not necessary to train by large corpus.

There have been many studies of domain identification which use term weighting [1, 2]. McDonough proposed a topic identification method on switch board corpus. In his switch board corpora, there are some sentences. In the sentences, there are many keywords which characterize a certain topic. He reported the best number of words in keyword dictionary is about 800. However, for a very short part of discourse, the number of keywords is not enough. Yokoi proposed a topic identification method which uses the keywords based on statistical information. He used word cooccurrence for topic identification.

However, he did not use spoken news, and the system conduct topic identification for each sentence in the topic identification experiment. In real news, it is difficult to segment each sentence automatically. Because there are many pairs of sentence-pause-sentence whose pause is quite short.

Some studies for transcription of broadcast news are going to be carried out [3, 4, 5]. However there are few studies which apply domain identification method for extracting keywords.

In this paper, we propose a method for domain identification and keyword extraction using term weighting. In our method, term weighting in regard to each domain is represented by the feature vector of the domain. The feature vector of each domain is automatically calculated by using newspaper articles. The domain which has the largest similarity between the unit of news and the feature vector of each domain is selected as domain of the unit. Then the system picks up keywords from the most suitable keyword path of the selected domain. Using our method, even if there are many words whose phoneme sequence are similar to correct keyword in the keyword dictionary, many keywords are selected correctly. The results of the experiments demonstrate the effectiveness of our method for speech recognition.

2 FEATURE VECTOR OF EACH DOMAIN

2.1 Term Weighting Represented by Feature Vectors

In our method, each part of radio news story is classified into a domain using the feature vectors from newspaper articles which are classified into domains. Each domain is characterized by a feature vector. Each element of feature vectors was based on the frequency of each noun in newspaper articles which are classified into each domain.

In general, a domain in each discourse is characterized by words which are appeared frequently in the discourse. Frequency of each keyword are often used for domain identification. However, all words which frequently appear do not always characterize the domain. If a word appears frequently in many domains, the word does not characterize the domain. Then, term weighting by χ^2 value is used in our method. Our system identifies the domain of news story by using feature vectors [6]. Each domain is characterized by a feature vector whose coordinate is an m-dimensional Euclidean space, where m is the number of nouns which are selected from newspaper articles. Each element of feature vectors is χ^2 value.

2.2 Estimation of χ^2 Vector Using Mutual Information

One of major problems of the method which is based on word frequency is data sparseness problem, i.e., the system can not identify domain, when there are no words which are in the keyword dictionary in the unit. To cope with

this problem, we estimate χ^2 vector of the word by using mutual information, and increase the number of words in feature vectors [6].

First, we calculated mutual information value between each noun pair in the all articles of Mainichi Shimbun '94 CD-ROM. Then, 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.

$$\boldsymbol{\chi}^2_\alpha = \frac{\sum_{k=1}^{m} \boldsymbol{\chi}^2_{\beta_k} * f(\alpha, \beta_k)}{\sum_{k=1}^{m} f(\alpha, \beta_k)} \quad (1)$$

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

3 KEYWORD EXTRACTION

In our method, a domain of each unit of radio news story is identified by using the feature vectors which were extracted by the method mentioned in Section 2. Input news stories are represented by phoneme sequence without space and word boundary does not appear. At each phoneme the system selects maximum 20 word candidates whose start point is the phoneme. phoneme sequences are segmented by pauses which are longer than 0.5 second in recorded radio news. We call a part between pauses a **unit**. The system selects a domain of each unit.

3.1 Similarity between Domain and Unit

Most of words which appear frequently in newspaper articles about domain "POL", tend to appear in the unit about politics. If word_k appears frequently in the domain_j, χ^2 value of word_k in domain_j is large. For example, in a unit about POL, sum of $\chi^2_{w,\text{POL}}$ tends to be large (w : a word in the unit). Then, the system selects a word sequence whose sum of $\chi^2_{k,j}$ is maximum among other word sequences at domain_j. The similarity between unit_i and domain_j is calculated using formula (2).

$$\begin{aligned} Sim(j,i) &= \max_{all\ paths} Sim'(j,i) \\ &= \max_{all\ paths} \sum_k np(\text{word}_k) \times \chi^2_{k,j} \end{aligned} \quad (2)$$

In formula (2), word_k is a word which is in word candidates in the word lattice, and each selected word does not share any phonemes with any other selected words. $np(\text{word}_k)$ is the number of phonemes of word_k. $\chi^2_{k,j}$ is χ^2value of word_k for domain_j. The system determines a keyword path whose $Sim'(j,i)$ is the largest among all keyword paths for domain_j.

Figure 1 shows the method of calculating similarity between unit_i and $\text{domain}_{\text{INT}}$. The system selects a path whose $Sim'(\text{INT}, \text{unit}_i)$ is larger than those of any other paths.

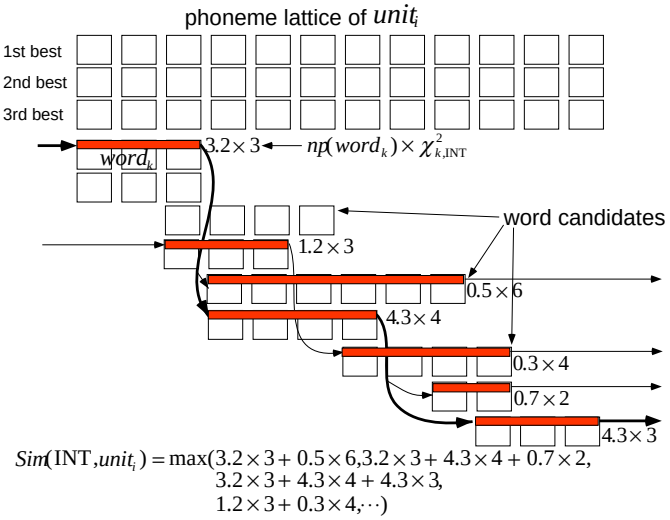

Figure 1: Calculating similarity between unit_i and INT (international)

3.2 Domain Identification and Keyword Extraction

In the domain identification process, the system identifies domain of each unit by using $Sim(\mathrm{domain}, \mathrm{unit}_i)$ of all domains. If a similarity between a unit and a domain is larger than similarities between a unit and any other domains, the domain seem to be the domain of the unit. Therefore, the system selects the domain which is the largest of all similarities in N of domains as the domain of the unit (formula (3)) . The words in the suitable path for selected topic are picked up as keywords.

$$domain_i = \arg \max_{1 \le j \le N} Sim(j, i) \tag{3}$$

4 EXPERIMENTS

We have conducted domain identification experiments and keyword extraction experiments.

4.1 Data

The test data we have used is a radio news which is selected from NHK 6 o'clock radio news in August and September of 1995. Some news stories are hard to be classified into one domain in radio news. Therefore we select news stories which two persons classified into the same domain are selected for the experiments. The **units** which are used as test data are segmented by pauses which are longer than 0.5 second. We selected 50 units for the experiments. The 50 units consist of 10 units of each domain. We used two kinds of test data. One is described with correct phoneme sequence. The other is written

in phoneme lattice which is the results of phoneme recognition [7]. In each segment of phoneme lattice, the number of phoneme candidates did not exceed 3. The following equations show the results of phoneme recognition.

$$\frac{\text{correct phonemes in phoneme lattice}}{\text{uttered phonemes}} = 95.6\%$$

$$\frac{\text{correct phonemes in phoneme lattice}}{\text{phoneme segments in phoneme lattice}} = 81.2\%$$

243 articles of Mainichi Shimbun in 1994 from CD-ROM were used in order to calculate feature vectors. We classified these articles into 5 domains. i.e., "politics", "economy", "international", "sports" and "accident". In order to calculate feature vectors of each domain, 243 articles are tagged by parts-of-speech using JUMAN [8]. Then we selected the 784 nouns, and 5 feature vectors whose dimensions are 784.

Because of data sparseness problem, the system could not calculate similarity value at some units which have no words in the newspaper articles. Therefore, we estimated χ^2_α (word_α does not appear in the 243 newspaper articles). We selected the words which co-occur with the 784 keywords which have been selected for feature vectors from the articles of newspaper in CD-ROM, and increased words in the keyword dictionary to 9,186 by the method which was mentioned in Section 2.2. We omitted words which are first names of persons in the keyword dictionary. The total number of words in the news units which belong to the keyword dictionary (9,186 words) was 77% larger than that of the original keyword dictionary (784 words).

4.2 Domain Identification Experiment

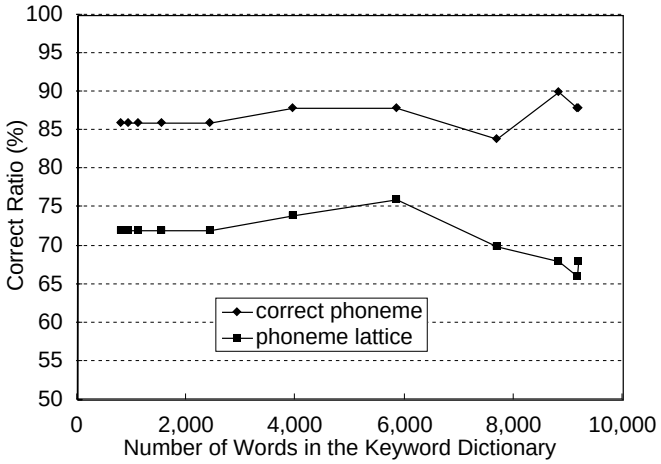

Figure 2: Domain identification results

In the experiments of domain identification with phoneme lattice, word candidates are extracted by using DP matching between a part of the phoneme

lattice and each phoneme sequence of word in the dictionary. Figure 2 shows the results of the experiments of domain identification. When the minimum value of mutual information was slid from 10 to 20, the number of words was changed from 9,186 to 784. The best performance was obtained when the keyword dictionary has 5,860 words (76%).

4.3 Keyword Extracting Experiment

We have conducted keyword extracting experiments. Figure 3 shows the result of the experiments. CP means correct phonemes and PL means the phoneme lattice. At each experiment, the number of words in the keyword dictionary was slid from 784 to 9,186. When the keyword dictionary has 5,860 words and input data was phoneme lattice, the number of selected correct keywords using our method was 132, and the number of selected correct keywords using the method without term weighting was 67.

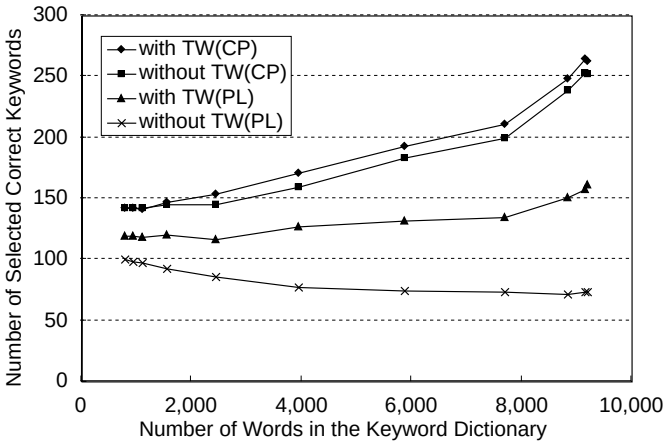

Figure 3: Performance comparison using two methods of keyword extracting: using term weighting (with TW(PL) and (with TW(CP)) and without term weighting (without TW(PL) and without TW(CP))

5 DISCUSSION

Figure 2 shows that when we used the keyword dictionary which has 5,860 words and phoneme lattice, 76% of units are identified with the most suitable domains by using the proposed method. For further improvement of domain identification, it is necessary to use larger corpus in order to calculate feature vectors precisely and have to improve phoneme recognition.

Figure 3 shows that the number of extracted keywords with large keyword dictionary is larger than that of small keyword dictionary. When the input data was the phoneme lattice, the larger the number of keywords, the larger

the gaps of the number of keywords between "with TW" and "without TW". The number of selected correct keywords using term weighting was twice as many as the number without term weighting with the keyword dictionary which has larger than 5,860 words in the experiments with the phoneme lattice. Figure 4 shows recall and precision which are shown in formula (4), and formula (5), respectively, when the input data was phoneme lattice.

$$\text{recall} = \frac{\text{number of correct words in MSKP}}{\text{number of selected words in MSKP}} \quad (4)$$

$$\text{precision} = \frac{\text{number of correct words in MSKP}}{\text{number of correct nouns in the unit}} \quad (5)$$

MSKP : the most suitable keyword path for selected domain

When the keyword dictionary had 784 words, precision was 31%, and recall was 42%. When the keyword dictionary had 9,186 words, precision was 43%, and recall was 25%. The result shows that the system extracted many incorrect keywords, because the system tries to find keywords for all parts of the units. In order to extract only correct keywords, the system has to use co-occurrent frequency between keywords in the most suitable keyword path.

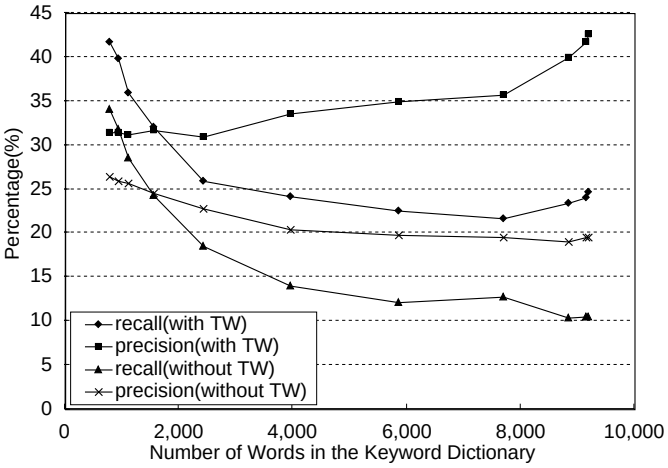

Figure 4: Recall and precision using two methods: proposed method which used term weighting (with TW) and a method which did not use term weighting (without TW)

6 CONCLUSIONS

In this paper, we have proposed a method for domain identification and extracting keywords by using term weighting in radio news. We are now conducting an experiment of domain identification and keyword extracting

with other news stories. We have to compare χ^2 method and other term weighting method in order to examine how χ^2 method is effective for domain identification and keyword extracting. In future, we will study how to remove incorrect words from extracted keywords in order to use our method for speech recognition. Also, we will classify newspaper articles into certain domain automatically.

7 ACKNOWLEDGMENTS

The authors would like to thank for permission to use newspaper articles Mainichi Shimbun and to use radio news Japan Broadcasting Corporation (NHK). The authors would also like to thank the anonymous reviewers for their valuable comments. This work was partially supported by the Telecommunications Advancement Foundation of Japan.

References

[1] J.McDonough, K.Ng, P.Jeanrenaud, H.Gish and J.R.Rohlicek, "Approaches to Topic Identification on the Switchboard Corpus", in *Proc. of ICASSP'94 volume1*, pp.385-388, 1994.

[2] K.Yokoi and T.Kawahara and S.Doshita, "Topic Identification of News Speech using Word Cooccurrence Statistics", in *Technical Report of IEICE SP96-105*, pp.71-78, 1997.

[3] T.Matsuoka, K.Ohtsuki, T.Mori, S.Furui and K.Shirai, "Japanese Large-Vocabulary Continuous-Speech Recognition Using a Business-Newspaper Corpus", in *Proc. of ICSLP'96 Volume1*, ThA2L1.6, 1996.

[4] B.Bakis, S.Chen, P.Gopalakrishnan, R.Gopinath, S.Maes and L.Pllymenakos, "Transcription of Broadcast News - System Robustness Issues and Adaptation Techniques", in *Proc. of ICASSP'97*, pp.96–105,1997.

[5] F.Kubala, T.Anastasakos, H.Jin, L.Nguyen and R.Schwartz, "Transcribing Radio News", in *Proc. of ICSLP'96 Volume1*, FrA1L1.5, 1996.

[6] Y.Suzuki, F.Fukumoto and Y.Sekiguchi, "Word Spotting of Radio News based on Topic Identification for Speech Recognition", in *Proc. of RANLP97*, pp.306–311, 1997.

[7] Y.Suzuki and C.Furuichi and S.Imai, "Spoken Japanese Sentence Recognition Using Dependency Relationship with Systematical Semantic Category", in *Trans. of IEICE J76 D-II, volume 11*, pp.2264-2273, 1993.

[8] Y.Matsumoto, S.Kurohashi, T.Utsuro, H.Taeki and M.Nagao, "Japanese Morphological Analysis System JUMAN Manual", "Nagao Lab. Kyoto University", 1993.