Keyword Extraction of Radio News using Term Weighting for Speech Recognition

Yoshimi Suzuki, Fumiyo Fukumoto†and Yoshihiro Sekiguchi‡
Department of Electrical Engineering and Computer Science,
Yamanashi University
4-3-11 Takeda, Kofu 400 Japan
{ysuzuki@suwa, fukumoto@skye†, sekiguti@saiko‡}.esi.yamanashi.ac.jp

Abstract

In this paper, we propose a method for keyword extraction using term weighting in radio news. In our method, firstly, keywords which characterize each domain are automatically extracted from newspaper articles. Next, feature vectors whose elements are χ^2 values between each keyword and each domain are calculated. 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 the proposed 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 until now, most of speech recognition systems are fixed for certain tasks, for example, a tourist information and a hamburger shop (Kanazawa *et al.* 94). 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, a minutes writing system for meetings and interactive information retrieval system for large area).

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.

In this paper, we propose a method for keyword extraction using term weighting for radio news. In our method, term weighting in regard to each domain is represented by the feature vector of the domain. Each element of the feature vectors is χ^2 value. The feature vector of each domain is automatically calculated by using newspaper articles which are classified into each domain. 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 our keyword extracting method uses the most suitable keyword path which is produced in the procedure of domain identification. There are many correct keywords on the most suitable keyword path of the most suitable domain. Therefore, the similarity between the unit and the feature vector of the most suitable domain is larger than those of any other domains. Using our method, even if there are many words whose phoneme sequence are similar to correct keyword in the keyword dictionary, keywords are selected correctly. Our method is robust to partial phoneme misunderstanding, because a domain of each unit is considered for selecting a keyword.

We have conducted the domain identification experiments and keyword extracting experiments using the result of phoneme recognition. The results of the experiments demonstrate the effectiveness of our method for speech recognition.

2 Related Work

there have been many studies of domain identification which use statistical information of words in written language (Yamamoto *et al.* 95) and spoken language (J.McDonough *et al.* 94), (Yokoi *et al.* 97), (Itoh *et al.* 95), (Suzuki *et al.* 96).

McDonough proposed a topic identification method on switch board corpus. The switch board corpora which they used have 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, the problem of his method is that 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 (Matsuoka *et al.* 96) (Kubala *et al.* 96) (Bakis *et al.* 97) (J.L.Gauvain *et al.* 97). However there are few studies which apply domain identification method for extracting keywords.

In this paper, we propose a method for keyword extraction with term weighting which are calculated using newspaper articles. Using our method, good performance is obtained when we used spoken news.

3 Feature Vector of Each Domain

3.1 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.

3.2 χ^2 Vectors

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 an automatic 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 contribute to characterize the domain. In order to cope with this problem, term weighting by χ^2 value (Suzuki *et al.* 97) is used in our method.

Our system identifies the domain of news story by using feature vectors. 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. Figure 1 shows χ^2 vector of word$_k$ $(1 \leq k \leq m)$ and feature vector of each domain in χ^2 matrix. POL, ECO, INT, SPR and ACC means politics, economy, international, sports and accident, respectively. The dotted circle shows feature vector of INT and circle shows χ^2 vector of word "President" and "Prime Minister". The number of element of a feature vector is the number of words of which χ^2 vector was calculated. Figure 1 shows that domain "politics" is characterized by "Prime Minister", and domain "international" is characterized by "President".

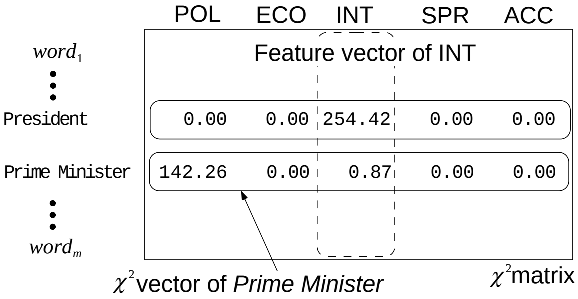


Figure 1: χ^2 vector and feature vector in χ^2 matrix

3.3 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 (Suzuki *et al.* 97).

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.

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

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

4 Domain Identification

In our method, a domain of each unit of radio news story is identified by using feature vectors which were extracted by the method which was mentioned in Section 3. Radio news stories which were used in our experiments were written in phonemes, and 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.

4.1 Extraction of Word Candidates

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. Figure 2 shows the example of word candidates. In each square frame, the number of word candidates does not exceed 20.

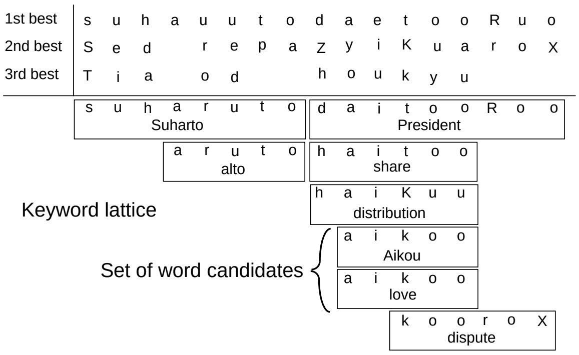


Figure 2: Example of word candidates

4.2 Similarity between Domain and Unit

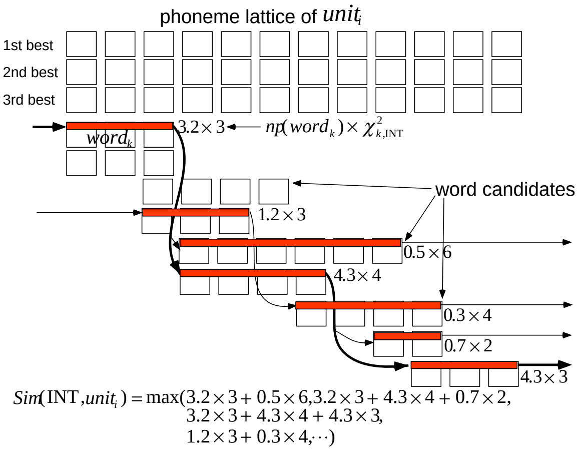
Most of words which appear frequently in newspaper articles about domain "POL", tend to appear in the unit about politics. If word$_k$ appears frequently in the domain$_j$, χ^2 value of word$_k$ in domain$_j$ is large. For example, in a unit about POL, sum of $\chi^2_{w,\mathrm{POL}}$ tends to be large (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 the unit$_i$ and domain$_j$ is calculated using formula (2).

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

In formula (2), word$_k$ is a word which is in word candidates obtained by Section 4.1, and each selected word does not share any phonemes with any other selected words. $np(\mathrm{word}_k)$ 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 path for domain$_j$.

Figure 3 shows the method of calculating similarity between unit$_i$ and domain$_{\mathrm{INT}}$. In Figure 3, there are many word paths from left to right. The system selects a path whose $Sim'(\mathrm{INT},\mathrm{unit}_i)$ is larger than those of any other paths.



$$Sim(\mathrm{INT},unit_i) = \max(3.2\times3+0.5\times6, 3.2\times3+4.3\times4+0.7\times2, \\ 3.2\times3+4.3\times4+4.3\times3, \\ 1.2\times3+0.3\times4, \cdots)$$

Figure 3: Calculating $Sim(\mathrm{INT},\mathrm{unit}_i)$ similarity between unit$_i$ and domain$_{\mathrm{INT}}$

4.3 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. Figure 4 shows domain identification method at each unit. In Figure 4, a similarity for INT (international) is the largest among all domains, and a domain of this unit is identified to "international".

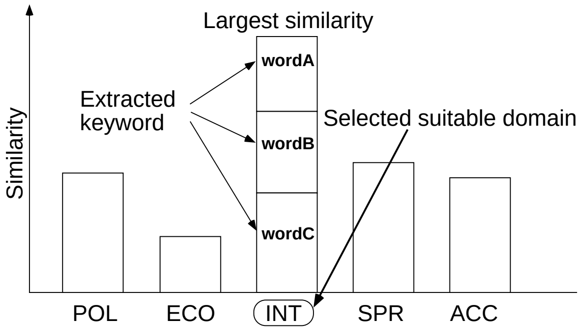


Figure 4: Domain identification method

5 Experiment

We have conducted domain identification experiments and keyword extraction experiments with correct phonemes and phoneme lattices which are the results of phoneme recognition.

5.1 Test 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. One day news consists of about 15 stories on average. There are some articles which are hard to be classified into one domain in news stories. 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 of the experiments 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 (Suzuki *et al.* 93). In each segment of phoneme lattice, the number of phoneme candidates did not exceed 3. The following equations show the results of phoneme recognition.

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

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

5.2 Training Data

243 articles of Mainichi Shimbun in 1994 from CD-ROM were used in order to calculate feature vectors. There are 427 characters par an article on average. We classified these articles into 5 domains. i.e., "politics", "economy", "international", "sports" and "accident" by using classification code of CD-Mainichi Shimbun 1994. There are about 20,000 characters in each domain of the corpus (in Table 1).

Table 1: Domain names

Domain	the number of articles	total number of characters
politics(POL)	34	20,840
economy(ECO)	63	21,050
international(INT)	59	20,750
sports(SPR)	37	20,133
accident(ACC)	50	21,014

In order to calculate feature vectors of each domain, 243 articles are tagged by parts-of-speech using JUMAN (Nag93). As a result, there are 534,932 nouns in the articles. From these articles, we selected the 796 nouns of which the frequency is larger than 5.

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_{α}^{2} (word$_\alpha$ does not appear in the 243 newspaper articles). Using mutual information, we selected the words which co-occur with the 796 keywords which have been selected for feature vectors from the articles of newspaper in CD-ROM, and increased keywords to 9,637 by the method which was mentioned in Section 3.3. The total number of words in the news units which belong to the keyword dictionary (9,637 words) was 77% larger than that of the original keyword dictionary (796 words).

5.3 Domain Identification Experiment

In the experiments of domain identification with phoneme lattice which is the results of phoneme recognition, word candidates are extracted by using DP matching between a part of the phoneme lattice and each phoneme sequence of word in the dictionary. The domain identification method is the same as the method with correct phoneme sequence. Figure 5 shows the results of the experiments of domain identification. This shows the results of domain identification using keywords of which the minimum value of mutual information is 10,11,12,13,14,15,16,17,18,19 and 20 respectively. When the minimum value of mutual information was slid from 10 to 20, the number of words was changed from 9,637 to 796. The best performance was obtained when the system used the keyword dictionary has 4,212 words (78%).

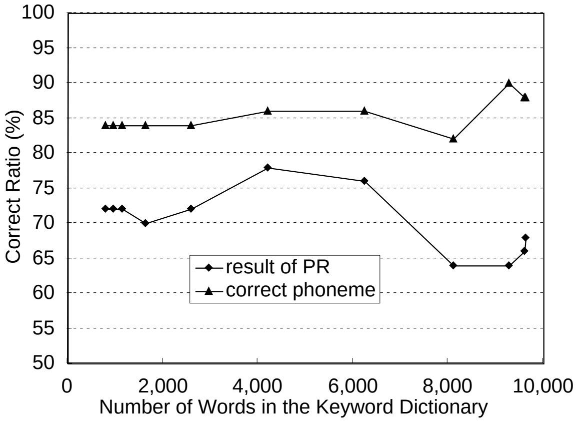


Figure 5: Domain identification results

5.4 *Keyword Extracting Experiment*

We have conducted keyword extracting experiments. Figure 6 shows the result of the experiments. In Figure 6, "with TW" means the method by using term weighting. "without TW" means the method without term weighting. CP means correct phonemes and PR means the result of phoneme recognition. At each experiment, the number of keywords was slid from 796 to 9,637.

By using the results of domain identification, when the keyword dictionary has 4,212 words, the number of selected correct keywords using our method was 125, and the number of selected correct keywords using the method without term weighting was 67 in the experiments with the result of phoneme recognition.

6 Discussion

6.1 *Domain Identification*

Figure 5 shows the correct ratios of domain identification with the result of phoneme recognition by using various number of keywords. This shows that 78% of units are identified with the most suitable domains by using the keyword dictionary which has 4,212 words.

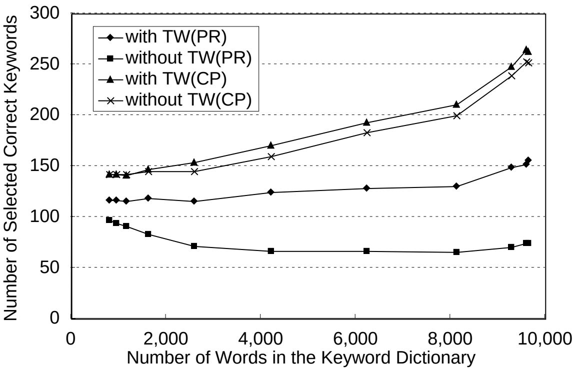


Figure 6: Performance comparison using two methods of keyword extracting: using term weighting (with TW(PR) and (with TW(CP)) and without term weighting (without TW(PR) and without TW(CP))

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.

6.2 *Keyword Extracting*

Figure 6 shows the number of selected correct keywords. It shows that the number of extracted keywords with large keyword dictionary is larger than that of small keyword dictionary. When the input data was correct phonemes, gaps of the number of keywords between "with TW" and "without TW" was small. However, when the input data was the result of phoneme recognition, 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 4,212 words in the experiments with the result of phoneme recognition.

Figure 7 shows recall and precision which are shown in formula (3), and formula (4), respectively, when the input data was phoneme lattice.

$$\text{recall} = \frac{\text{number of correct words in MSKP}}{\text{number of selected words in MSKP}} \quad (3)$$

$$\text{precision} = \frac{\text{number of correct words in MSKP}}{\text{number of correct nouns in the unit}} \quad (4)$$

MSKP : the most suitable keyword path for selected domain

Using the keyword dictionary which has 796 words, precision was about 31%, and recall was about 41%. Using the keyword dictionary which has 9,637 words, precision was about 41%, and recall was about 24%. The result shows that the system extracted many incorrect keywords, because the system tries to find keywords for all parts of the units and extracts incorrect keywords. In order to extract only correct keywords, the system has to use co-occurrent frequency between keywords in the most suitable keyword path.

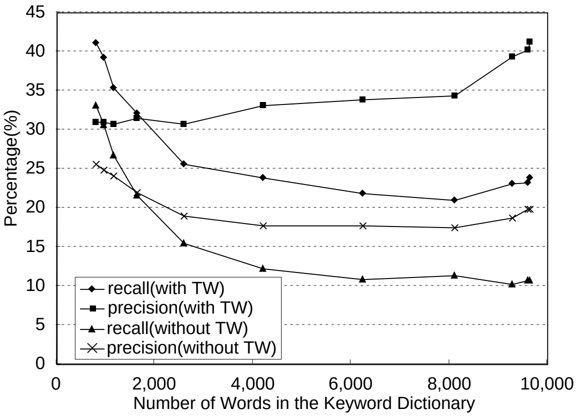


Figure 7: Recall and precision using two methods: proposed method which used term weighting (with TW) and a method which did not use term weighting (without TW)

7 Conclusions

In this paper, we have proposed a method for extracting keywords by using term weighting in radio news. When we used the dictionary which has 4,212 words, 78% of spoken units could be identified correctly (Figure 5). Using the results of domain identification, the number of selected correct keywords was larger than the number without the results (Figure 6). Using the result of domain identification, about 33% of correct nouns could be extracted (Figure 7). We are now conducting an experiment of domain identification and keyword extracting with other news stories. In our current experiment, we used χ^2 method for term weighting. 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 nouns from extracted keywords in order to use our method for speech recognition. Also, we will classify newspaper articles into certain domain automatically.

8 Acknowledgment

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.

References

(Bakis *et al.* 97) Baimo Bakis, Scott Chen, Ponani Gopalakrishnan, Ramesh Gopinath, Stephane Maes, and Lazaros Pllymenakos. Transcription of broadcast news - system robustness issues and adaptation techniques. In *Proc. ICASSP'97*, pages 711–714, 1997.

(Itoh *et al.* 95) Yoshiaki Itoh, Jiro Kiyama, and Ryuichi Oka. Speech understanding and speech retrieval for tv program based on spotting algorithms. In *Proc. of ASJ Spring Meeting*, pages 3–P–22, 1995.

(J.L.Gauvain *et al.* 97) J.L.Gauvain, G. Adda, L. Lamel, and M. Adda-Decker. Transcribing broadcast news shows. In *Proc. ICASSP'97*, pages 715–718, 1997.

(J.McDonough *et al.* 94) J.McDonough, K.Ng, P.Jeanrenaud, H.Gish, and J.R.Rohlicek. Approaches to topic identification on the switchboard corpus. In *Proc. IEEE ICASSP'94*, volume 1, pages 385–388, 1994.

(Kanazawa *et al.* 94) Hiroshi Kanazawa, Shigenobu Seto, Hideki Hashimoto, Hideaki Shinchi, and Yoichi Takebayashi. A user-initiated dialogue model and its implementation for spontaneous human-computer interaction. In *Proc. ICSLP'94*, volume 1, pages 111–114, 1994.

(Kubala *et al.* 96) Francis Kubala, Tasos Anastasakos, Hubert Jin, Long Nguyen, and Richard Schwartz. Transcribing radio news. In *Proc. ICSLP'96*, volume 1, page FrA1L1.5, 1996.

(Matsuoka *et al.* 96) Tatsuo Matsuoka, Katsutoshi Ohtsuki, Takeshi Mori, Sadaoki Furui, and Katsuhiko Shirai. Japanese large-vocabulary continuous-speech recognition using a busines-newspaper corpus. In *Proc. ICSLP'96*, volume 1, page ThA2L1.6, 1996.

(Nag93) Nagao Lab. Kyoto University. *Japanese Morphological Analysis System JUMAN Manual*, 1993.

(Suzuki *et al.* 93) Yoshimi Suzuki, Chieko Furuichi, and Satoshi Imai. Spoken japanese sentence recognition using dependency relationship with systematical semantic category. *Trans. of IEICE J76 D-II*, 11:2264–2273, 1993.

(Suzuki *et al.* 96) Yoshimi Suzuki, Fumiyo Fukumoto, and Yoshihiro Sekiguchi. Discourse segmentation for radio news. In *ASA and ASJ Third Joint Meeting, PROCEEDINGS of the papers submitted to the ASJ*, pages 1009–1014, 1996.

(Suzuki *et al.* 97) Yoshimi Suzuki, Fumiyo Fukumoto, and Yoshihiro Sekiguchi. Word spotting of radio news based on topic identification for speech recognition. In *RANLP97*, pages 306–311, 1997.

(Yamamoto *et al.* 95) Kazuhide Yamamoto, Shigeru Masuyama, and Naito Shozo. An automatic classification method for japanese texts using mutual category relations. In *SIG-IPS Japan 106-2*, pages 7–12, 1995.

(Yokoi *et al.* 97) Kentaro Yokoi, Tatsuya Kawahara, and Shuji Doshita. Topic identification of news speech using word cooccurrence statistics. In *Technical Report of IEICE SP96-105*, pages 71–78, 1997.