Keyword Extraction using Term-Domain Interdependence for Dictation of Radio News

Yoshimi Suzuki **Fumiyo Fukumoto** **Yoshihiro Sekiguchi**
Dept. of Computer Science and Media Engineering
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 keyword extraction method for dictation of radio news which consists of several domains. In our method, newspaper articles which are automatically classified into suitable domains are used in order to calculate feature vectors. The feature vectors shows term-domain interdependence and are used for selecting a suitable domain of each part of radio news. Keywords are extracted by using the selected domain. The results of keyword extraction experiments showed that our methods are robust and effective for dictation of radio news.

1 Introduction

Recently, many speech recognition systems are designed for various tasks. However, most of them are restricted to certain tasks, for example, a tourist information and a hamburger shop. Speech recognition systems for the task which consists of various domains seems to be required for some tasks, e.g. a closed caption system for TV and a transcription system of public proceedings. In order to recognize spoken discourse which has several domains, the speech recognition system has to have large vocabulary. Therefore, it is necessary to limit word search space using linguistic restricts, e.g. domain identification.

There have been many studies of domain identification which used term weighting (J.McDonough et al., 1994; Yokoi et al., 1997). McDonough proposed a topic identification method on switch board corpus. He reported that the result was best when the number of words in keyword dictionary was about 800. In his method, duration of discourses of switch board corpora is rather long and there are many keywords in the discourse. However, for a short discourse, there are few keywords in a short discourse. Yokoi also proposed a topic identification method using co-occurrence of words for topic identification (Yokoi et al., 1997). He classified each dictated sentence of news into 8 topics. In TV or Radio news, however, it is difficult to segment each sentence automatically. Sekine proposed a method for selecting a suitable sentence from sentences which were extracted by a speech recognition system using statistical language model (Sekine, 1996). However, if the statistical model is used for extraction of sentence candidates, we will obtain higher recognition accuracy.

Some initial studies of transcription of broadcast news proceed (Bakis et al., 1997). However there are some remaining problems, e.g. speaking styles and domain identification.

We conducted domain identification and keyword extraction experiment (Suzuki et al., 1997) for radio news. In the experiment, we classified radio news into 5 domains (i.e. accident, economy, international, politics and sports). The problems which we faced with are;

1. Classification of newspaper articles into suitable domains could not be performed automatically.
2. Many incorrect keywords are extracted, because the number of domains was few.

In this paper, we propose a method for keyword extraction using term-domain interdependence in order to cope with these two problems. The results of the experiments demonstrated the effectiveness of our method.

2 An overview of our method

Figure 1 shows an overview of our method. Our method consists of two procedures. In the procedure of term-domain interdependence calculation, the system calculates feature vectors

of term-domain interdependence using an encyclopedia of current term and newspaper articles. In the procedure of keyword extraction in radio news, firstly, the system divides radio news into segments according to the length of pauses. We call the segments **units**. 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. Finally, the system extracts keywords in each unit using the feature vector of selected domain which is selected by domain identification.

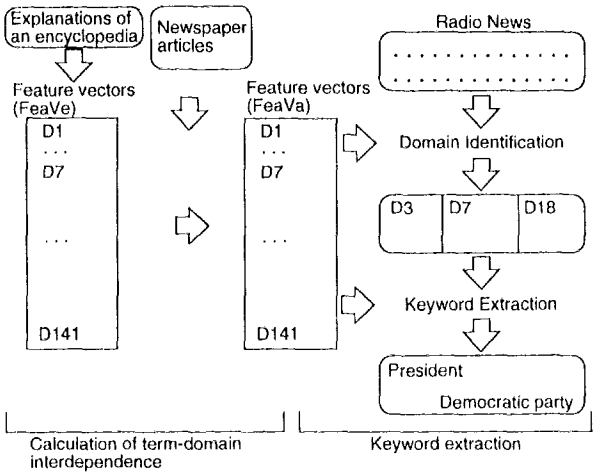

Figure 1: An overview of our method

3 Calculating feature vectors

In the procedure of term-domain interdependence calculation, We calculate likelihood of appearance of each noun in each domain. Figure 2 shows how to calculate feature vectors of term-domain interdependence.

In our previous experiments, we used 5 domains which were sorted manually and calculated 5 feature vectors for classifying domains of each unit of radio news and for extracting keywords. Our previous system could not extract some keywords because of many noisy keywords. In our method, newspaper articles and units of radio news are classified into many domains. At each domain, a feature vector is calculated by an encyclopedia of current terms and newspaper articles.

3.1 Sorting newspaper articles according to their domains

Firstly, all sentences in the encyclopedia are analyzed morpheme by Chasen (Matsumoto et al., 1997) and nouns which frequently appear are extracted. A feature vector is calculated by frequency of each noun at each domain. We call the feature vector $FeaVe$. Each element of $FeaVe$ is a χ^2 value (Suzuki et al., 1997). Then, nouns are extracted from newspaper articles by a morphological analysis system (Matsumoto et al., 1997), and frequency of each noun are counted. Next, similarity between $FeaVe$ of each domain and each newspaper article are calculated by using formula (1). Finally, a suitable domain of each newspaper article are selected by using formula (2).

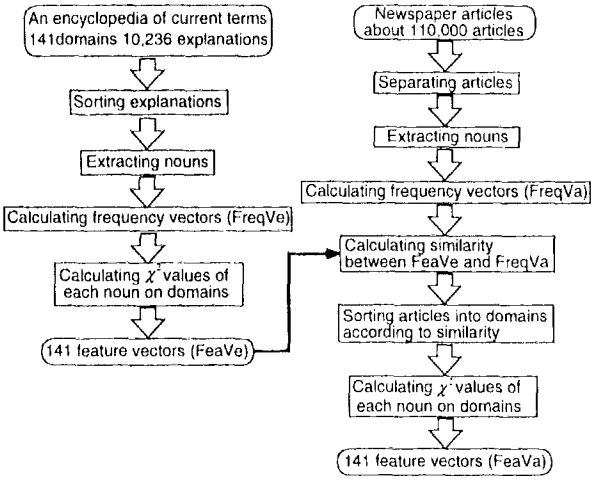

Figure 2: Calculating feature vectors

$$Sim(i,j) = FeaVe_j \cdot FreqVa_i \quad (1)$$

$$Domain_i = \arg \max_{1 \le j \le N} Sim(i,j) \quad (2)$$

where i means a newspaper article and j means a domain. $(\cdot)$ means operation of inner vector.

3.2 Term-domain interdependence represented by feature vectors

Firstly, at each newspaper articles, less than 5 domains whose similarities between each article and each domain are large are selected. Then, at each selected domain, the frequency vector is modified according to similarity value and frequency of each noun in the article. For example, If an article whose selected domains are "political party" and "election", and similarity between the article and "political party"

and similarity between the article and "election" are 100 and 60 respectively, each frequency vector is calculated by formula (3) and formula (4).

$$FreqV_{pp} = FreqV'_{pp} + FreqVa_i \times \frac{100}{100} \quad (3)$$

$$FreqV_{el} = FreqV'_{el} + FreqVa_i \times \frac{60}{100} \quad (4)$$

where i means a newspaper article.

Then, we calculate feature vectors $FeaVa$ using $FreqV$ using the method mentioned in our previous paper (Suzuki et al., 1997). Each element of feature vectors shows χ^2 value of the domain and word$_k$. All word$_k$ ($1 \leq k \leq M$:M means the number of elements of a feature vector) are put into the keyword dictionary.

4 Keyword extraction

Input news stories are represented by phoneme lattice. There are no marks for word boundaries in input news stories. Phoneme lattices are segmented by pauses which are longer than 0.5 second in recorded radio news. The system selects a domain of each unit which is a segmented phoneme lattice. At each frame of phoneme lattice, the system selects maximum 20 words from keyword dictionary.

4.1 Similarity between a domain and an unit

We define the words whose χ^2 values in the feature vector of domain$_j$ are large as keywords of the domain$_j$. In an unit of radio news about "political party", there are many keywords of "political party" and the χ^2 value of keywords in the feature vector of "political party" is large. Therefore, sum of $\chi^2_{w,\text{politicalparty}}$ tends to be large (w : a word in the unit). In our method, the system selects a word path whose sum of $\chi^2_{k,j}$ is maximized in the word lattice at domain$_j$. The similarity between unit$_i$ and domain$_j$ is calculated by formula (5).

$$Sim(i,j) = \max_{all\ paths} Sim'(i,j) = \max_{all\ paths} \sum_k np(\text{word}_k) \times \chi^2_{k,j} \quad (5)$$

In formula (5), word$_k$ is a word in the word lattice, and each selected word does not share any frames with any other selected words. np(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 selects a word path whose $Sim'(i,j)$ is the largest among all word paths for domain$_j$.

Figure 3 shows the method of calculating similarity between unit$_i$ and domain$_{\text{D1}}$. The system selects a word path whose Sim'(unit$_i$, D1) is larger than those of any other word paths.

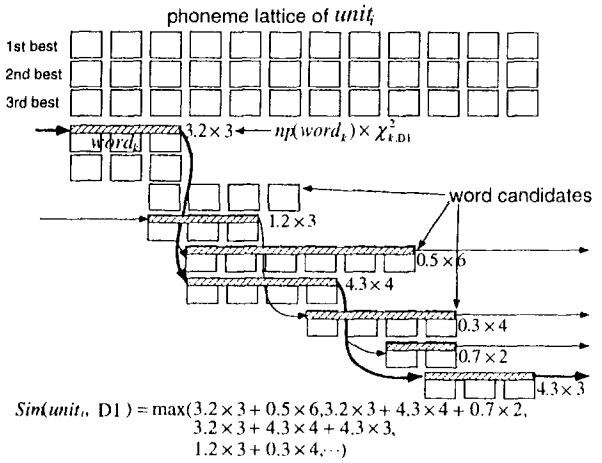

Figure 3: Calculating similarity between unit$_i$ and D1

4.2 Domain identification and keyword extraction

In the domain identification process, the system identifies each unit to a domain by formula (5). If $Sim(i,j)$ is larger than similarities between an unit and any other domains, domain$_j$ seems to be the domain of unit$_i$. The system selects the domain which is the largest of all similarities in N of domains as the domain of the unit (formula (6)) . The words in the selected word path for selected domain are selected as keywords of the unit.

$$Domain_i = \arg \max_{1 \leq j \leq N} Sim(i,j) \quad (6)$$

5 Experiments

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. Some news stories are hard to be classified into one domain in radio news by human. For evaluation of domain identification experiments, we

selected news stories which two persons classified into the same domains are selected. The **units** which were used as test data are segmented by pauses which are longer than 0.5 second. We selected 50 units of radio news for the experiments. The 50 units consisted 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 obtained by a phoneme recognition system (Suzuki et al., 1993). In each frame of phoneme lattice, the number of phoneme candidates did not exceed 3. The following equations show the results of phoneme recognition.

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

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

5.2 Training data

In order to classify newspaper articles into small domain, we used an encyclopedia of current terms "Chiezo"(Yamamoto, 1995). In the encyclopedia, there are 141 domains in 9 large domains. There are 10,236 head-words and those explanations in the encyclopedia. In order to calculate feature vectors of domains, all explanations in the encyclopedia are performed morphological analysis by Chasen (Matsumoto et al., 1997). 9,805 nouns which appeared more than 5 times in the same domains were selected and a feature vector of each domain was calculated. Using 141 feature vectors which were calculated in the encyclopedia, we identified domains of newspaper articles. We identified domains of 110,000 articles of newspaper for calculating feature vectors automatically. We selected 61,727 nouns which appeared at least 5 times in the newspaper articles of same domains and calculated 141 feature vectors.

5.3 Domain identification experiment

The system selects suitable domain of each unit for keyword extraction. Table 1 shows the results of domain identification. We conducted domain identification experiments using two kinds of input data, i.e. correct phoneme sequence and phoneme lattice and two kinds of domains, i.e. 141 domains and 9 large domains. We also compared the results and the result using previous method (Suzuki et al., 1997). For comparison, we selected 5 domains which are used by previous method in our method. In previous method, we used a keyword dictionary which has 4,212 words.

Table 1: The result of domain identification

method	number of domains	Correct phoneme	Phoneme lattice
our method	141	62%	40%
	9	78%	54%
	5	90%	82%
previous method	5	86%	78%

5.4 Keyword extraction experiment

We have conducted keyword extraction experiment using the method with 141 feature vectors (our method), 5 feature vectors (previous method) and without domain identification. Table 2 shows recall and precision which are shown in formula (7), and formula (8), respectively, when the input data was phoneme lattice.

$$\text{recall} = \frac{\text{the number of correct words in MSKP}}{\text{the number of selected words in MSKP}} \quad (7)$$

$$\text{precision} = \frac{\text{the number of correct words in MSKP}}{\text{the number of correct nouns in the unit}} \quad (8)$$

MSKP : the most suitable keyword path for selected domain

6 Discussion

6.1 Sorting newspaper articles according to their domains

For using χ^2 values in feature vectors, we have good result of domain identification of newspaper articles. Even if the newspaper articles which are classified into several domains, the suitable domains are selected correctly.

6.2 Domain identification of radio news

Table 1 shows that when we used 141 kinds of domains and phoneme lattice, 40% of units were identified as the most suitable domains by our

Table 2: Recall and precision of keyword extraction

Method	R/P	Correct phoneme	Phoneme lattice
our method	R	88.5%	48.9%
(141 domains)	P	69.0%	38.1%
previous method	R	80.0%	24.0%
(5 domains)	P	63.1%	33.0%
without DI	R	77.0%	12.2%
(1 domain)	P	60.1%	9.5%

R: recall P: precision DI: domain identification

method and shows that when we used 9 kinds of domains and phoneme lattice, 54% of units are identified as the most suitable domains by our method. When the number of domains was 5, the results using our method are better than our previous experiment. The reason is that we use small domains. Using small domains, the number of words whose χ^2 values of a certain domain are high is smaller than when large domains are used.

For further improvement of domain identification, it is necessary to use larger newspaper corpus in order to calculate feature vectors precisely and have to improve phoneme recognition.

6.3 Keyword extraction of radio news

When we used our method to phoneme lattice, recall was 48.9% and precision was 38.1%. We compared the result with the result of our previous experiment (Suzuki et al., 1997). The result of our method is better than the our previous result. The reason is that we used domains which are precisely classified, and we can limit keyword search space. However recall was 48.9% using our method. It shows that about 50% of selected keywords were incorrect words, because the system tries to find keywords for all parts of the units. In order to raise recall value, the system has to use co-occurrence between keywords in the most suitable keyword path.

7 Conclusions

In this paper, we proposed keyword extraction in radio news using term-domain interdependence. In our method, we could obtain sorted large corpus according to domains for keyword extraction automatically. Using our method, the number of incorrect keywords in extracted words was smaller than the previous method.

In future, we will study how to select correct words from extracted keywords in order to apply our method for dictation of radio news.

8 Acknowledgments

The authors would like to thank Mainichi Shimbun for permission to use newspaper articles on CD-Mainichi Shimbun 1994 and 1995, Asahi Shimbun for permission to use the data of the encyclopedia of current terms "Chiezo 1996" and Japan Broadcasting Corporation (NHK) for permission to use radio news. The authors would also like to thank the anonymous reviewers for their valuable comments.

References

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

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

Yuji Matsumoto, Akira Kitauchi, Tatuo Yamashita, Osamu Imaichi, and Tomoaki Imamura, 1997. *Japanese Morphological Analysis System ChaSen Manual.* Matsumoto Lab. Nara Institute of Science and Technology.

Satoshi. Sekine. 1996. Modeling topic coherence for speech recognition. In *Proc. COLING 96*, pages 913–918.

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

Yoshimi Suzuki, Fumiyo Fukumoto, and Yoshihiro Sekiguchi. 1997. Keyword extraction of radio news using term weighting for speech recognition. In *NLPRS97*, pages 301–306.

Shin Yamamoto, editor. 1995. *The Asahi Encyclopedia of Current Terms 'Chiezo'.* Asahi Shimbun.

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