KEYWORD EXTRACTION OF RADIO NEWS USING DOMAIN IDENTIFICATION BASED ON CATEGORIES OF AN ENCYCLOPEDIA

Yoshimi Suzuki, Fumiyo Fukumoto and Yoshihiro Sekiguchi

Department of Computer Science and Media Engineering
Yamanashi University
4-3-11 Takeda, Kofu 400 Japan

ABSTRACT

In this paper, we propose a 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 show 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 effective for keyword extraction 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 [1, 2]. 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, in a short discourse the system might extract few keywords. Yokoi also proposed a topic identification method using co-occurrence of words for topic identification [2]. He classified each dictated sentence of news into 8 topics. In TV or Radio news, however, it is difficult to segment each sentence automatically.

There are some studies of transcription of broadcast news [3]. However there are some remaining problems, e.g. speaking styles and domain identification. We conducted domain identification and keyword extraction experiment [4] for radio news. In the experiments, 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 we used large domains, for example "politics" in stead of "politic party", "election" and so on.

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, namely calculation of term-domain interdependence and keyword extraction in radio news. In calculation of term-domain interdependence, 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, the system extracts keywords in radio news using calculated term-domain interdependence.

3. TERM-DOMAIN INTERDEPENDENCE

In the procedure calculation of term-domain interdependence, 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 method, firstly, feature vectors $FeaVe$ are calculated using an encyclopedia of current term. Then, the system classifies newspaper articles into domains according to the similarity between each article and each domain. Finally, term-domain

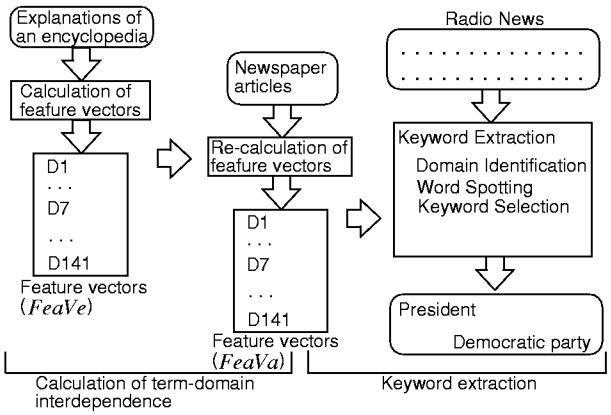

Figure 1: An overview of our method

interdependence $FeaVe$ represented by feature vectors was calculated.

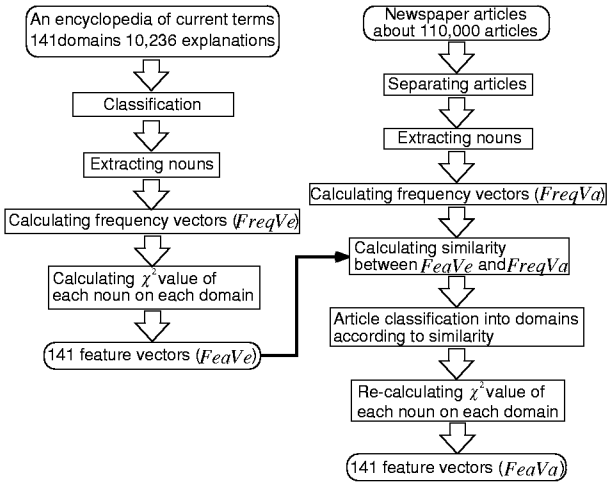

Figure 2: Calculation of term-domain interdependence

3.1. Calculating feature vectors $FeaVe$

Firstly, all sentences in the encyclopedia are analyzed morpheme by Chasen [5] and nouns which frequently appear are extracted. The system calculates word frequency vectors, namely $FreqVe$. Each element of $FreqVe$ is frequency of each noun in each categories of the encyclopedia. The number of $FreqVe$ shows the number of categories of the encyclopedia. Feature vector $FeaVe$ is calculated using $FreqVe$. Each element of $FeaVe$ is a χ^2 value [4].

3.2. Newspaper articles classification into domains according to similarity

Nouns are extracted from newspaper articles by a morphological analysis system [5], and frequency of each noun is counted. Next, similarity between $FeaVe$ of each domain and each newspaper article are calculated by using formula (1). Finally, some suitable domains of each newspaper article are selected by using formula (2).

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

$$Domains_i = \arg Sim(i,j) \times 0.8 > \max_{1 \le j \le N} Sim(i,j) \quad (2)$$

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

3.3. 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 calculated according to similarity value and frequency of each noun in the article. For example, if an article whose selected domains are D20(political party) and D19 (election), and similarity between the article and D20 and similarity between the article and D19 are 100 and 60 respectively, each frequency vector is calculated by formula (3) and formula (4).

$$FreqVa_{D20} = FreqVa'_{D20} + FreqVa_i \times \frac{100}{100} \quad (3)$$

$$FreqVa_{D19} = FreqVa'_{D19} + FreqVa_i \times \frac{60}{100} \quad (4)$$

where i means a newspaper article.

Then, we calculate feature vectors $FeaVa$ using the method mentioned in our previous paper [4]. Each element of feature vectors shows χ^2 value of the domain and $word_k$. All $word_k$ $(1 \le k \le M : M$ means the number of elements of a feature vector) are put into the keyword dictionary.

4. KEYWORD EXTRACTION

Figure 3 shows how to extract keywords. Input news stories are represented by phoneme lattice. There are no marks for word boundaries in phoneme lattice. Phoneme lattices are segmented by pauses which are longer than 0.5 second in recorded radio news. We call a set of phoneme lattice **unit**. The system selects a domain of each unit. At each frame of phoneme lattice, the system selects maximum 20 words whose phoneme sequences start from the frame from keyword dictionary. In keyword extraction, firstly we calculate similarity between a domain and an unit. Then the system performs word spotting, and finally the system selects keyword in the spotted words.

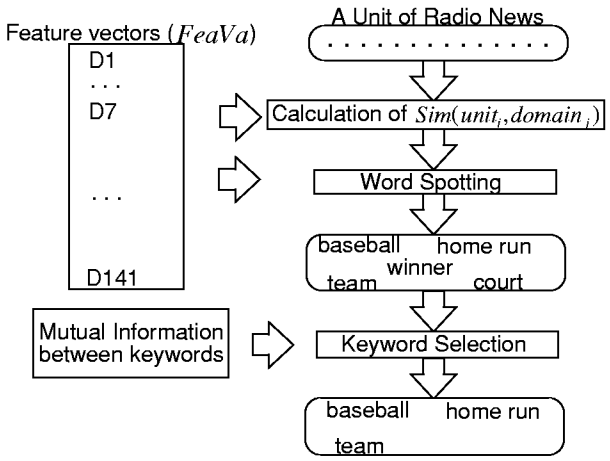

Figure 3: Keyword extraction method

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 $domain_j$, there are many keywords of $domain_j$ and the χ^2 value of keywords in the feature vector of $domain_j$ is large. Therefore, sum of $\chi^2_{w,j}$ tends to be large (w : a uttered word in the unit). In our method, the system selects a word path whose sum of $\chi^2_{k,j}$ is maximum in the word lattice at domain$_j$. The similarity between unit$_i$ and domain$_j$ is calculated by formula (5).

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

In formula (5), word$_k$ is a word in the word lattice, and each selected word doesn't 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^2 value of word$_k$ on domain$_j$.

The system selects a word path whose $Sim'(i,j)$ is the largest among all word paths for domain$_j$.

4.2. Word spotting and keyword selection

After calculating similarity between the unit and each domain, the system spots words in a word path whose $Sim(i,j)$ is the largest among all domains.

Finally, the system selects keywords in the word path using mutual information between words in the word path.

5. EXPERIMENTS

5.1. Training data

In order to classify newspaper articles into small domains, we used an encyclopedia of current terms "Chiezo"[6]. In the encyclopedia, there are 141 categories in 9 large categories. We used the 141 categories for 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 [5]. 9,805 nouns which appeared more than 5 times in the same explanations of domains were selected and a feature vector of each domain was calculated. Using 141 feature vectors which were calculated using the encyclopedia, we classified about 110,000 newspaper articles into suitable domains. Using the classified newspaper articles, the system calculated 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.

5.2. 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 even for a 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 our phoneme recognition system [7]. In each frame of phoneme lattice, the number of phoneme candidates did not exceed 3. The following equations show the results of phoneme recognition.

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

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

5.3. Result of domain identification

Table 1 shows the result of domain identification. In Table 1, "number of domains" 9 of our method means that we classified 141 domains into 9 large domains. "number of domains" 5 of our method means that we selected 5 domains (accident, economy, international, politics and

sports) from 9 large domains. When the number of domains was 5, the result using our method was better than the result of our previous method.

Table 1: The result of domain identification

method	number of domains	input data	
		correct phoneme	phoneme lattice
our method	141	62%	40%
	9	78%	54%
	5	90%	82%
previous method	5	86%	78%

5.4. Result of keyword extraction

Table 2 shows the result of keyword extraction. Using mutual information, recall value rised by 8.4% and precision value fell by 1.1%. The result using mutual information was better than the result without mutual information.

Table 2: The result of keyword extraction

method	recall	precision
with mutual information	42.5%	39.1%
without mutual information	34.1%	40.2%

6. DISCUSSION

As the result of classification of newspaper articles and stories of radio news, about 85% of data are classified into suitable domains using our method. In the encyclopedia, the number of domains which are about accident is only one. For better result, we have to divide domain accident into some domains, e.g. traffic accident, murder case and bribery case.

As the result of domain identification, 61% of units are classified into suitable domains. If the length of each unit is longer than unit which we decided, the result is better than the result. For better result, we have to segmentation of radio news stories.

The result of keyword extraction without mutual information shows that recall and precision was 34.1% and 40.2%, respectively. In a similar way, the result with mutual information shows that recall and precision was 42.5% and 39.1%, respectively. We confirmed that the result using mutual information of extracted keyword is better than the result without mutual information.

7. CONCLUSIONS

When recall and precision of keyword selection is high, the speech recognition system can recognize discourse by filling other words. Using our method, recall value of keyword selection was better than the result without mutual information. The results of experiments demonstrate the applicability of the method for keyword extraction of broadcast news.

In this study, we handled radio news. However, our method can also be used in other application, e.g. interviews, talk shows or meetings. In future, we will apply our method to other application and demonstrate the applicability of the method.

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.

9. REFERENCES

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

[2] Kentaro Yokoi and Tatsuya Kawahara and Shuji Doshita "Topic Identification of News Speech using Word Cooccurrence Statistics" Technical Report of IEICE SP96-105, 1997

[3] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk and S.J. Young "Experiments in Broadcast News Transcription" ICASSP'98, SP27.4, 1998

[4] Yoshimi Suzuki and Fumiyo Fukumoto and Yoshihiro Sekiguchi "Domain Identification and Keyword Extraction of Radio News Using Term Weighting" NLPRS97, pp.301–306, 1997

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

[6] Shin Yamamoto "The Asahi Encyclopedia of Current Terms 'Chiezo"' Asahi Shimbun, 1995

[7] Yoshimi Suzuki and Chieko Furuichi and Satoshi Imai "Spoken Japanese Sentence Recognition Using Dependency Relationship with Systematical Semantic Category" Trans. of IEICE J76 D-II,Vol.11,pp.2264-2273,1997