Keyword Extraction of Radio News using Term Weighting with an Encyclopedia and Newspaper Articles

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

Abstract In this paper, we propose a method for keyword extraction of radio news. Using our method, data sparseness problem and false alarm problem was lightened even for short discourse or document. Also, our method is robust for partial errors of phoneme recognition. In our method, there are two procedures: i.e. term weighting and keyword extraction. In procedure of term weighting, a feature vector of each domain is calculated using an encyclopedia and newspaper articles. In procedure of keyword extraction, keywords are extracted using feature vectors and result of domain identification. The results of experiments demonstrate the applicability of the method.

1 Introduction

Keyword extraction is one of crucial themes of natural language processing and spoken language processing. Many studies of keyword extraction were based on keyword density method [2]. However, the method has a problem that the word which is not a key, is judged to be a keyword. Suzuki [4] proposed keyword extraction methods using text data. Suzuki proposed a method [4] for an automatic keyword extraction using cohesion chart and thesaurus. In his method which uses labels of categories of a thesaurus, a thesaurus has ordinary classification: e.g. the category 'Party' has the word 'Team' and 'Political Party', however, 'Team' is a keyword of 'sports' and 'Political Party' is a keyword of 'politics'.

In spoken language processing, keyword extraction is useful for domain identification and speech understanding. Although, speech data is more difficult to search than text data. The difficulties of reference of speech data are the following three points:

1. Acoustic ambiguity
2. Ambiguity of boundary between words
3. Ambiguity of boundary between domains

For speech data, keyword density method [2] can not be used, because speech data is usually shorter than text data and there are few keywords which are frequently appeared. Furthermore, some data has no words which are in a category because the thesaurus has few proper nouns (the problem of data sparseness).

Some studies proposed keyword extraction methods for topic identification of Radio News and TV News.

Wright [6] proposed a method of topic spotting using keyword extraction. However, his experiment was only one topic : weather forecast. Imai [1] proposed a method for topic spotting using topic-mixture model which was based on Hidden Markov Model. He classified news stories into many different topics (9,062 topics). However, his test data is already divided between news stories and the data is rather long. Also, his method seems to extract many incorrect keywords. For keyword extraction with his method, the system has to select correct keywords from many extracted words.

In this paper, we propose a keyword extraction method which is suitable for speech data, and we report on the experiment of keyword extraction using radio news. Our method consists of two procedures; term weighting and keyword extraction. In procedure of term weighting, we calculate feature vectors twice, in order to obtain accurate feature vectors efficiently. In procedure of keyword extraction, the system selects suitable domain and extracts keywords using the feature vectors and word lattice which is generated by phoneme recognition. The system compares each feature vector with word lattice of radio news. It selects the suitable category of part of radio news. In the part of radio news, it regards the words whose element of feature vector of the domain as keywords of the part.

We have conducted two experiments. In the first experiment, we used correct phonemes of 8 days' radio news (643 parts of radio news). In the second experiment, we used results of phoneme recognition of 50 parts of radio news.

2 Term Weighting

Figure 1 shows how to calculate feature vectors in the procedure of term weighting. In Figure 1, the system calculates feature vectors twice using an encyclopedia [7] and newspaper articles. It calculates draft feature vectors using an encyclopedia and calculates final feature vectors using draft feature vectors and newspaper articles.

Firstly, the system counts frequency of each word which is in explanation of a category of an encyclopedia. We used categories of an encyclopedia as domains. The number of domain is 141. For calculating draft feature vectors, all sentences in an encyclopedia are analysed morpheme by JUMAN [3] and nouns which frequently appear are extracted. In each category, the system calculates feature vector [5] whose each element is χ^2 value.

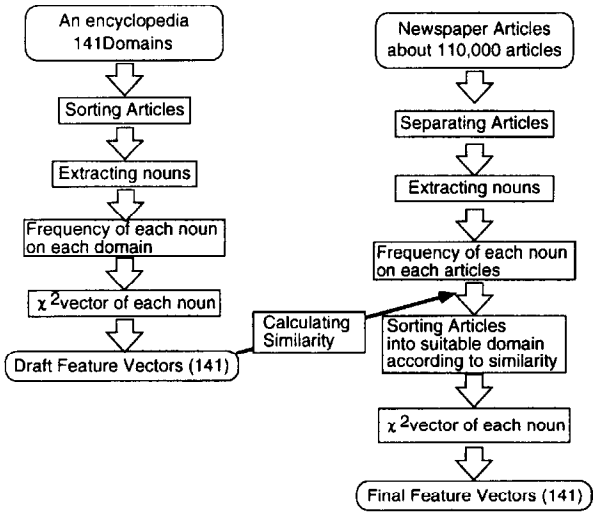

Figure 1: Calculating feature vectors

Then, we analyse morpheme of newspaper articles, extract nouns, and count frequency of each noun. We calculate similarity between feature vector of each domain and each article. Formula (1) shows similarity between $domain_i$ and $article_j$.

$$Sim(i,j) = \frac{\boldsymbol{D}_i \cdot \boldsymbol{A}_j}{|\boldsymbol{D}_i||\boldsymbol{A}_j|} \quad (1)$$

In formula (1), $\boldsymbol{D}_i$ and $\boldsymbol{A}_j$ indicate feature vector of $domain_i$ and word frequency vector of $article_j$, respectively. $(\cdot)$ shows operation of inner vector.

Using the calculated similarities, final feature vectors are calculated.

3 Keyword Extraction

The other procedure is keyword extraction. In this procedure, the system calculates similarity between each unit and each domain. The system selects a path whose similarity is larger than those of any other paths. Then, the system selects suitable domain using similarity between the unit and each domain. The system extracts keywords using the result of domain identification and suitable keyword path which is created by domain identification. Figure 2 shows how to identify domain and extract keywords. In Figure 2, a similarity for D45 (one of the domains "economy") is the largest among all domains, and a domain of the unit is identified to "economy". The system selects word A, word B and word C as keywords in the unit.

4 Experiments

Firstly, we conducted the experiment using correct phonemes of 8 days' radio news (643 parts of radio news). As a result, recall was 58.7% and precision was 74.0%, where recall and precision is defined by formula (2) and formula (3), respectively.

$$\text{recall} = \frac{\text{The number of keywords in MSKP}}{\text{The number of selected words in MSKP}} \quad (2)$$

$$\text{precision} = \frac{\text{The number of keywords in MSKP}}{\text{The number of keywords in the unit}} \quad (3)$$

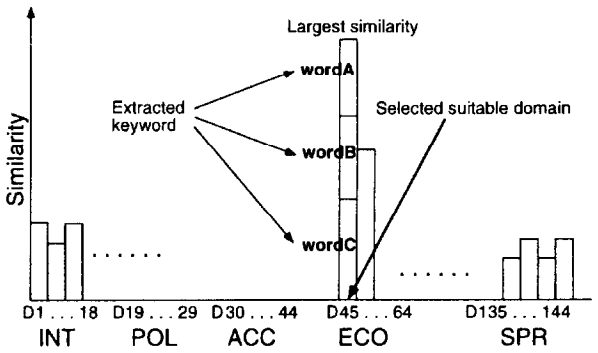

Figure 2: Domain Identification and keyword extraction method

MSKP : the most suitable keyword path for selected domain

Then we conducted the experiment using results of phoneme recognition of 50 parts of radio news. As the result, recall was 34.1% and precision was 42.5%.

5 Conclusions

In this paper, we proposed a method for domain identification and keyword extraction by using term weighting in radio news. In our current experiment, we used χ^2 method for term weighting. We have to compare χ^2 method with other term weighting method in order to examine how χ^2 method is effective for domain identification and keyword extracting.

Acknowledgements We are grateful to Mainichi Shimbun and Asahi Shimbun for allowing us to use the newspaper text data base and the text of a encyclopedia "Chiezo", respectively. We are also grateful to Japan Broadcasting Corporation (NHK) for allowing us to use radio news.

References

[1] Toru Imai, Richard Schwartz, Francis Kubala, and Long Nguyen. Improved topic discrimination of broadcast news using a model of multiple simultaneous topics. In *Proc. ICASSP'97*, pages 727–730, 1997.

[2] H.P. Luhn. The automatic creation of literature abstracts. *IBM journal*, 2(1):159–165, 1968.

[3] Yuji Matsumoto, Sadao Kurohashi, Takehito Utsuro, Hiroshi Taeki, and Makoto Nagao. *Japanese Morphological Analysis System JUMAN Manual.* Nagao Lab. Kyoto University, 1993.

[4] Hitoshi Suzuki, Shigeru Masuyama, and Shozo Naito. Examination of keyword extraction using thesaurus in japanese text. In *Information Processing Society of Japan Technical paper NL98-10*, pages 73–78, 1993.

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

[6] Jerry H. Wright, Michael J. Carey, and Eluned S. Parris. Improved topic spotting through statistical modelling of keyword dependencies. In *Proc. ICASSP'95*, volume 1, pages 313–316, 1995.

[7] Shin Yamamoto, editor. *Chiezo.* Asahi Shimbun, 1995.