TOPIC RECOGNITION FOR NEWS SPEECH BASED ON KEYWORD SPOTTING

Yoichi Yamashita

Dep. of Computer Science, Ritsumeikan University
1-1-1, Noji-Higashi, Kusatsu-shi, Shiga, 525-8577 Japan
yama@cs.ritsumei.ac.jp

Toshikatsu Tsunekawa *Riichiro Mizoguchi*

I.S.I.R., Osaka University
8-1, Mihogaoka, Ibaraki-shi, Osaka, 567-0047 Japan
{tsune,miz}@ei.sanken.osaka-u.ac.jp

ABSTRACT

This paper describes topic identification for Japanese TV news speech based on the keyword spotting technique. Three thousands of nouns are selected as keywords which contribute to topic identification, based on criterion of mutual information and a length of the word. This set of the keywords identified the correct topic for 76.3% of articles from newspaper text data. Further, we performed keyword spotting for TV news speech and identified the topics of the spoken message by calculating possibilities of the topics in terms of an acoustic score of the spotted word and a topic probability of the word. In order to neutralize effect of false alarms, bias of the topics in the keyword set is removed. Topic identification rate is 66.5% assuming that identification is correct if the correct topic is included in the top three topics. The removal of the bias improved the identification rate by 6.1%.

1. INTRODUCTION

Enormous speech data, such as TV news, can be stored in the data base as computer technology progresses. Automatic topic identification (TID) is an important technique to fast and easy information retrieval. The TID task is a classification problem which assigns the topic label to a message. In the TID for speech data, a word sequence must be derived from a spoken message by speech recognition before it is mapped into a topic.

Two different approaches, continuous speech recognition (CSR)[1][2] and keyword spotting (KS) [3][5][4][6], have been tried to obtained the word sequence. The performance of CSR is dependent on the fitness of the language model to input speech data. The training of the language model is a crucial issue in the CSR-based TID, particularly for a new domain including many unknown words. On the other hand, the KS-based TID is free from the problem of unknown words, but have to consider false alarms which are incorrect detections of the word and are inevitable in KS.

In order to minimize the bad effect of false alarms, this paper proposed a new idea of unbiasing the topics in the keyword set by equalizing expectations of topic probabilities for all keywords. This procedure cancels the effect of false alarms if the false alarms arise randomly.

2. METHOD

Topic boundaries are unknown for TV news speech. A topic is identified for a short window of M seconds, called a topic analysis window (TAW), assuming that one topic continues in a TAW. The process of topic identification is repeated shifting the TAW, illustrated in Figure 1. In each TAW, the keyword spotting generates a keyword sequence $\boldsymbol{w}$. The most possible topic is determined by scoring topics in terms of topic probability based on the keyword sequence, recognition score of the keywords, and bias of the topics in the keyword set. The score of i-th topic T_i, $F(T_i)$, is defined as

$$F(T_i) = \frac{\log P(T_i|\boldsymbol{w}) + k \times R(\boldsymbol{w})}{N} - B(T_i), \qquad (1)$$

where $R(\boldsymbol{w})$ is total recognition score of keywords in the spotting, k is a weighting factor and 0.5 in this paper, and N is a number of keywords detected in a TAW. In the keyword spotting, false alarms, that are detection of non-keywords, are inevitable. It is important to minimize the bad effect of false alarms. If keyword sequences are generated randomly, the topic probability $P(T_i|\boldsymbol{w})$ should be equal for every topic. When a keyword set has bias of topic probabilities and false alarms boost some topics probabilistically, it is necessary to remove the bias of the topics from the keyword set. $B(T_i)$ denotes bias of the i-th topic in the keyword set.

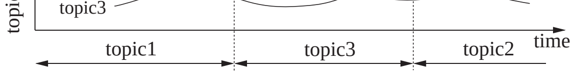

Figure 1. Topic identification for news speech.

The probability of the i-th topic, $P(T_i|\boldsymbol{w})$, is calculated by

$$P(T_i|\boldsymbol{w}) = \frac{P(\boldsymbol{w}|T_i)P(T_i)}{P(\boldsymbol{w})}. \qquad (2)$$

We can ignore the denominator $P(\boldsymbol{w})$ because it is independent of T_i, and make all prior probabilities of topics equal. On the assumption that appearance of a word in the $\boldsymbol{w}$ is independent each other, $P(T_i|\boldsymbol{w})$ is approximated by

$$\begin{aligned} P(T_i|\boldsymbol{w}) &\simeq P(\boldsymbol{w}|T_i) \\ &\simeq \prod_{j=1}^{N} P(kw_j|T_i), \end{aligned} \qquad (3)$$

where kw_j is a keyword found in the $\boldsymbol{w}$. Keywords should be words with strong relation to topics and are a priori selected based on the analysis of text corpus. Non-keywords in the $\boldsymbol{w}$ are ignored in the topic identification.

3. KEYWORD SELECTION

3.1. Measures of Keyword Selection

Mutual information is used as a measure of relation between a word and a topic. Random variables describing a topic of a message and a word in the message are denoted by T and W, respectively. The information obtained by knowing the word for topic identification, that is the mutual information $I(T;W)$, is defined as

$$\begin{aligned} I(T;W) &= I(W;T) \\ &= H(W) - H(W|T) \\ &= -\sum_j p(w_j)\log_2 p(w_j) \\ &\quad + \sum_i p(T_i) \sum_j p(w_j|T_i)\log_2 p(w_j|T_i) \\ &= \sum_j [\ -p(w_j)\log_2 p(w_j) \\ &\quad + \sum_i p(T_i)p(w_j|T_i)\log_2 p(w_j|T_i)\ ] \\ &\equiv \sum_j [G(w_j)]. \end{aligned} \qquad (4)$$

$G(w)$ is contribution from a word w to identifying T. Large $G(w)$ means that a word w have much information for identifying a topic. We used $G(w)$ as a measure of selecting keywords.

The easiness of keyword detection is an important factor for selecting keywords as well as the relation with topics. Because short words are difficult to spot correctly, short words less than 6 phonemes are omitted from the keywords.

We selected 3,000 nouns as keywords out of *Mainichi-shinbun* newspaper text corpus based on these two criteria and obtained $P(w_j|T_i)$. The corpus contains newspaper text during 48 months, 385K articles and 103M words. The topic set is composed of 8 categories: "international", "economy", "family", "culture", "science", entertainment", "sports", and "society". The articles in this corpus are manually divided into 16 categories, a half of which is the same as the topics. Another half of the categories, such as "top news", "general", and so on, are not used because they are difficult to be associated by words in sentences. The 45 month data in the corpus which contains 216K articles of 8 topics and 51M words is used for the keyword selection, other 3 month data which contains 16K articles and 4M words is reserved for evaluation mentioned in 3.3.

Table 1. Bias of logarithmic topic probability for each topic.

T_i	$B(T_i)$	T_i	$B(T_i)$
international	-4.556	science	-5.443
economy	-4.500	entertainment	-5.084
family	-4.792	sports	-5.105
culture	-4.955	society	-4.322

3.2. Bias of The Keyword Set

The $\log P(T_i|\boldsymbol{w})$ in the equation (1) should be equal if the keyword sequence $\boldsymbol{w}$ is composed of randomly generated words. To this end, the score $F(T_i)$ is subtracted by $B(T_i)$, which is bias of each topic in the keyword set. Introduction of $B(T_i)$ neutralizes the effect of false alarms which are inevitable in the word spotting. $B(T_i)$ is defined as

$$B(T_i) = \frac{\sum_{kw_j \in K} P(kw_j)\log P(kw_j|T_i)}{\sum_{kw_j \in K} P(kw_j)}, \qquad (5)$$

where K is the keyword set. Table 1 lists the bias of each topic.

3.3. Evaluation of the Keyword Set for Text Data

Before the topic identification for speech data, we tried topic identification for text data in order to evaluate the selected keywords. Test texts are newspaper articles longer than 50 words and are not used to obtain $P(w_j|T_i)$ mentioned in 3.1. The topic boundaries of the text is given. The possibilities of the topic are calculated by the equation (1) except $B(T_i) = 0$ because there are no false alarms in keyword matching of the text. The topic was correctly identified for 76.3% of articles.

4. KEYWORD SPOTTING

The HMM-based recognizer carries out the keyword spotting using two linguistic constraints illustrated in Figure 2. The first constraint (a) allows any sequences of phonemes, and the second one (b) supposes that the last word is a keyword. The existence possibility of a

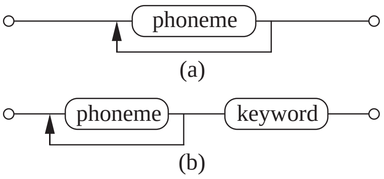

Figure 2. Linguistic constraints for keyword spotting.

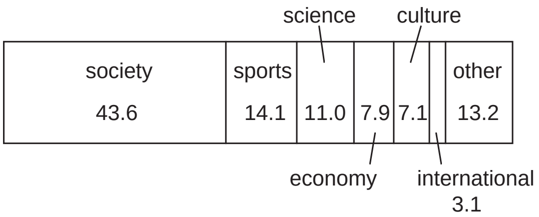

Figure 3. Duration ratio[%] of topics.

keyword, $L(kw_j)$, is defined as

$$L(kw_j) = \frac{L_{kw}(kw_j) - L_{ph}(kw_j)}{length\_of\_keyword} \quad , \qquad (6)$$

where $L_{kw}(kw_j)$ and $L_{ph}(kw_j)$ are logarithmic recognition score from the beginning of the speech to an arbitrary time under constraints (b) and (a), respectively. The *length_of_keyword* is defined as the number of phoneme of the keyword. A keyword is detected at the time when $L(kw_j)$ is larger than the threshold. The $R(\boldsymbol{w})$ in the equation (1) is defined as

$$R(\boldsymbol{w}) = \sum_{j=1}^{N} L(kw_j). \qquad (7)$$

We used the HTK tools provided by Entropic to obtain the phoneme models. The feature vector is composed of 12 melcepstrum, 12 delta-melcepstrum, and delta-energy. Phonemes are divided into 26 categories. Each phoneme model has 5 states with 4 mixtures and 3 loops. The phoneme models were trained with 13.6 hour phonetically balanced sentences by 64 speakers in the ASJ Continuous Speech Corpus for Research[7].

5. TOPIC RECOGNITION

5.1. Speech Data

Speech data is news speech of 40 minutes which was recorded from a TV news program. Six topics except "family" and "entertainment" are appeared in this data. The duration ratio of the topics is shown in Figure 3. "Other" in this figure includes "weather", background music, noise, and so on. The coverage of the keyword set is 55.3% against 85 nouns which appear in the beginning part of the news and are longer than 5 phonemes.

5.2. Results of Topic Identification

The topic identification was tried with two kinds of the TAW length, M = 30 or 60[sec]. The interval of the window shift is 10[sec] for both the window length. Table 2 summarizes the topic identification results.

Topic identification rate is 66.5% when $M = 30$ assuming that identification is correct if the correct topic is included in the top three. Removal of bias $B(T_i)$ improves the identification rate by 6.1%. The identification rate is slightly higher with a shorter window.

Figure 4 shows the transition pattern of the F score of 8 topics for a segment of test speech. In this figure, (a) is a results without the unbiasing, that is $B(T_i) = 0$ in the equation (1), and (b) is with the unbiasing. In the non-unbiasing case, the topic "society" got the highest F score in most TAWs, shown in Figure 4(a). The frequency of "society" in the test data is very high shown in Figure 3. It resulted in the high rate of rank 1 in the non-unbiasing case.

After 1,470 seconds where the correct topic is "sports", in Figure 4(b), F score successfully becomes large, although the score of "sports" is vary low in most segments. The speech around 1,465 seconds is on-the-spot broadcasting of *sumo* wrestling overlapped by spectators' cheer. Only about ten keywords were detected in these TAWs, while average number of detected keywords in a TAW was 280 when $M = 30$. The F scores of most topics became very low because of insufficiency of keywords.

6. SUMMARY

It is important to minimize bad effects of false alarms in the spotting-based topic identification. Unbiasing topic probabilities in a keyword set is efficient to cancel the bad effect of false alarms. It improved topic identification rate by 6.1%.

The performance of the spotting-based topic identification is much dependent on keyword selection. In the keyword selection, the goodness of a word is measured not only by the relationship with topics but also by easiness of spotting. In this paper, we used the length of words in terms of a phoneme count as the easiness of spotting. Introduction of characteristics of phoneme recognition errors into the keyword selection will be a future work.

Acknowledgment

This paper used CD *Mainichi-shinbun* '91-'94, ASJ Continuous Speech Corpus for Research, RWC text data base (RWC-DB-TEXT-95-1).

REFERENCES

[1] B. Peskin, S. Connolly, L. Gillick, S. Lowe, D. McAllaster, V. Nagesha, P. Mulbregt, and S. Wegmann : "Improvements in switchboard recognition and topic identification", Proc. of ICASSP '96, pp.303-306 (1996).

[2] K. Ohtsuki, T. Matsuoka, S. Matsunaga, and S. Furui : "Topic Extraction with multiple topic-words in broadcast-news speech", Proc. of ICASSP '98, pp.329-336 (1998).

(b) Unbiasing

Figure 4. An example of topic identification result.

[3] J.McDonough and H.Gish : "Issues in topic identification on the switchboard corpus", Proc. of ICSLP '94, pp.2163-2166 (1994).

[4] Y. Itoh, J. Kiyama, and R. Oka : "Speech Understanding and Speech Retrieval for TV News by Using Connected word Spotting", Proc. of Eurospeech '95, pp.2141-2144 (1995).

[5] J.T. Foote, G.J.F. Jones, K. Sparck Jones, and S.J. Young : "Talker-Independent Keyword Spotting for Information Retrieval", Proc. of Eurospeech '95, pp.2145-2148 (1995).

[6] D.A.James : A system unrestricted topic retrieval from radio news broadcasts", Proc. of ICASSP '96, pp.279-282 (1996).

[7] "Continuous Speech Corpus for Research", CD-ROM, Vol.1-3, Acoustical Society of Japan (1993).