<table>
<thead>
<tr>
<th>項目</th>
<th>内容</th>
</tr>
</thead>
<tbody>
<tr>
<td>タイトル</td>
<td>論文：連続音声認識に基づくラジオ番組のトピック抽出技術の研究</td>
</tr>
<tr>
<td>作者</td>
<td>Katsutoshi Ohtsuki, Tatsuo Matsuoka, Shoichi Matsunaga, Sadaoki Furui</td>
</tr>
<tr>
<td>発行日</td>
<td>2002, 7</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://search.ieice.org/">http://search.ieice.org/</a></td>
</tr>
<tr>
<td>権利情報</td>
<td>本著作物の著作権は電子情報通信学会に帰属します。</td>
</tr>
</tbody>
</table>

権利情報：本文の著作権は電子情報通信学会に帰属します。
PAPER

Topic Extraction Based on Continuous Speech Recognition in Broadcast News Speech

Katsutoshi OHTSUKI, Tatsuo MATSUOKA, Shoichi MATSUNAGA, Regular Members, and Sadaoki FURUI, Fellow

SUMMARY In this paper, we propose topic extraction models based on statistical relevance scores between topic words and words in articles, and report results obtained in topic extraction experiments using continuous speech recognition for Japanese broadcast news utterances. We attempt to represent a topic of news speech using a combination of multiple topic words, which are important words in the news article or words relevant to the news. We assume a topic of news is represented by a combination of words. We statistically model mapping from words in an article to topic words. Using the mapping, the topic extraction model can extract topic words even if they do not appear in the article. We can then compute the degree of relevance between a topic word and a word in the article by using newspaper text covering a five-year period. The degree of relevance between these words is calculated based on measures such as mutual information or the $\chi^2$-method. In experiments extracting five topic words using a $\chi^2$-based model, we achieve 72% precision and 12% recall for speech recognition results. Speech recognition results generally include a number of recognition errors, which degrades topic extraction performance. To avoid this, we employ N-best candidates and likelihood given by acoustic and language models. In experiments, we find that extracting five topic words using N-best candidate and likelihood values achieves significantly improved precision.

key words: topic extraction, topic word, relevance score, continuous speech recognition, broadcast news

1. Introduction

Recent advances in digital technology have made it possible to store and access large amounts of digital text, image, movie, speech, and music data. However, to use such data effectively, it is necessary that the data be well-pigeonholed as a database. That is to say, classifying or indexing for data is essential if the data are to serve as useful 'information.' For example, without a database, a user who wants to know the contents of a movie, speech or music data files, has to play the files one by one and this takes considerable time. If instead information about the contents is available, the user can classify and retrieve the data without having to play the files beforehand.

Keyword spotting methods for topic identification (TID) of conversational speech or news speech often use selected discriminating keywords [1]–[4]. However, too many key-

Manuscript received August 16, 2001.
Manuscript revised January 7, 2002.
† The authors are with NTT Cyber Space Laboratories, NTT Corporation, Yokosuka-shi, 239-0847 Japan.
‡‡ The author is with Broadband Business Department, NTT East Corporation, Tokyo, 100-0004 Japan.
++ The author is with the Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552 Japan.
words tend to be selected if the vocabulary of keywords is large. In contrast, some TID approaches employing continuous speech recognition (CSR) [5]–[7], classify text or speech into at most 10–20 topics, and there are few application areas for data classified into such a small number of topics. Even with a large number of topics, it is not usable if the classification does not match users' demands.

In our work, we attempt to represent a topic of news speech using a combination of multiple words. We define a topic word as an important word in the news article or a word relevant to the news. We assume a topic is represented by a combination of multiple topic words. By using several topic words which express the news with various levels of interpretation from rough categories to details, a kind of a summary of the news can be expressed.

In the process of extracting a topic, we deal with only content words, such as nouns and verbs, which communicate semantic information from the point of view of information retrieval field, instead of with all words in the speech. We trained a topic extraction model capable of computing the degree of relevance between a topic word and a word in an article by using newspaper text covering a five-year period. We then evaluated the model in experiments using manual transcriptions of broadcast news speech and speech recognition results.

In previous researches, text or news speech has been classified into several topics based on relevance of a keyword to each category or topic [1]–[7]. Each keyword has a relevance score with respect to each category. To classify articles, relevance scores of all keywords in an article are summed for each category. The article is then classified into the category having the highest relevance score. In our study, we replace those categories with topic words and extract several topic words of high relevance to an article [9], [10]. The topic extraction model has two lists of words, topic words and words in articles, and each mapping from words in articles to topic words is estimated statistically from training data. As the two lists are different, the topic extraction model can extract topic words even if they do not appear in the article.

The topic mixture model [8] extracts multiple topics similarly. The model is an HMM in which each state of a topic emits words with probabilities. In our approach, the topic extraction model has relevance scores between each topic word and each word in an article, which are calculated on the basis of mutual information or $\chi^2$-value. This model is
much simpler than the HMM-based method, since it can be trained by simply counting word frequency in the training data and the extraction is conducted by accumulating relevance scores to each topic word through an article and picking up topic words with high relevance score. The topic extraction is carried out for results of speech recognition. However speech recognition results generally include a number of recognition errors, which degrades topic extraction performance. To avoid this, we employed N-best candidates and likelihood given by acoustic and language models into the topic extraction process.

The remainder of this paper is organized as follows. Section 2 describes the topic extraction model and the relevance scores used in the model. Section 3 describes the topic extraction methods used for the speech recognition results. Section 4 describes training data for topic extraction models and the evaluation data. Section 5 describes our large-vocabulary continuous-speech recognition (LVCSR) system and experimental results, and Sect. 6 is a conclusion.

2. Topic Extraction

2.1 Topic Extraction Model

Topic words are extracted on the basis of relevance scores between a topic word and an article. The relevance score $R(a, t_j)$ between topic word $t_j$ and article $a$ is calculated as

$$R(a, t_j) = \sum_{k=1}^{N_a} s(w_k) r(w_k, t_j),$$  \hspace{1cm} (1)

where $N_a$ is the number of words in the article $a$, and $r(w_k, t_j)$ is the relevance score between the topic word $t_j$ and $w_k$, the $k$-th word in the article $a$. $s(w_k)$ is a weighting factor for $w_k$ and it is either constant or variable depending on each word. By varying $s(w_k)$ on the basis of $w_k$'s recognition score, the relevance score can become more robust against recognition errors. The topic extraction model contains all relevance scores for every word in the article and every topic word. For each article, only the topic words having high relevance scores are extracted.

2.2 Relevance Scores

The relevance score is statistically calculated using the frequencies of words and topic words in the training data. We calculated the relevance score with mutual information defined as the degree of dependence between random variables and the $\chi^2$-value in the $\chi^2$-test.

2.2.1 The Mutual-Information-Based Method

The mutual-information-based relevance score between word $w_i$ and topic word $t_j$ is expressed as

$$l(w_i, t_j) = \log \frac{P(w_i, t_j)}{P(w_i)P(t_j)}.$$  \hspace{1cm} (2)

If there is no co-occurrence of $w_i$ and $t_j$ in the training data, the numerator in Eq. (2), i.e., $P(w_i, t_j)$ becomes zero, and this causes a problem when adding up the scores. In this case, we consider that no information is obtained and calculate the relevance score based on the mutual information as

$$l'(w_i, t_j) = \begin{cases} l(w_i, t_j), & \text{if } P(w_i, t_j) \neq 0 \\ 0, & \text{if } P(w_i, t_j) = 0 \end{cases}.$$  \hspace{1cm} (3)

The mutual information is based on conditional probability, and the frequencies of $w_i$ and $t_j$ are not considered. We also consider the relevance score based on the mutual information weighted with a joint probability, which is expressed as

$$l''(w_i, t_j) = P(w_i) \cdot l'(w_i, t_j).$$  \hspace{1cm} (4)

2.2.2 The $\chi^2$-Value-Based Method

The $\chi^2$-value-based relevance score between word $w_i$ and topic word $t_j$ is expressed as

$$\chi^2_i = \frac{(f_{ij} - F_i)^2}{F_i},$$  \hspace{1cm} (5)

where $f_{ij}$ is the frequency of $w_i$ for $t_j$, and $F_i$ is the theoretical frequency of $w_i$ when it appears with equal probability for all topic words. It is given by

$$F_i = \frac{\sum_{i=1}^{N} f_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{M} f_{ij}},$$  \hspace{1cm} (6)

where $N$ is a distinct number of words and $M$ is a distinct number of topic words. If the actual frequency is much larger than the theoretical frequency, the word should appear as a topic word. However Eq. (5) has the same value regardless of whether $f_{ij} - F_i$ is positive or negative. Thus, we calculate the $\chi^2$-value based relevance score as

$$\chi^2_i' = \frac{(f_{ij} - F_i) \cdot |f_{ij} - F_i|}{F_i}.$$  \hspace{1cm} (7)

3. Topic Extraction from Speech Recognition Results

Topic extraction from speech recognition results is different from topic extraction from texts. This is mainly because speech recognition results often contain recognition errors. However, speech recognition results have useful information related to recognition processes such as N-best candidates and likelihood values. We employed this information into the topic extraction process in order to make the process robust against recognition errors.

3.1 N-Best Candidates

In N-best candidates, correctly recognized words frequently
appear in the candidates with high likelihood values. On the other hand, incorrectly recognized words have low likelihood values and are often replaced by other words in the candidates. Correct words are more frequent than incorrect words in N-best candidates. Therefore, topic extraction from N-best candidates should be more robust against speech recognition errors than that from a single best candidate.

3.2 Weighting with Likelihood

Continuous speech recognition is based on the acoustic likelihood and language likelihood of each word. Given a word sequence of speech recognition results, words with high values for both acoustic and language likelihood are likely correct, but those with either a low acoustic likelihood or language likelihood are possibly recognized incorrectly. Therefore, by employing these likelihood values obtained in recognition processes, topic extraction becomes robust against recognition errors. For example, the weighting factor $s(w_i)$ for $w_i$ with likelihood $l(w_i)$ can be calculated as

$$s(w_i) = \frac{l(w_i) - L_{\text{min}}}{L_{\text{max}} - L_{\text{min}}} (S_{\text{max}} - S_{\text{min}}) + S_{\text{min}},$$

where $L_{\text{max}}$ and $L_{\text{min}}$ are maximum and minimum likelihood values, and $S_{\text{max}}$ and $S_{\text{min}}$ are maximum and minimum values of weighting factors.

4. Data

4.1 Training Data

The topic extraction model contained all relevance scores between each word of a sentence and each topic word. We trained the topic extraction model with newspaper articles and headlines covering about five years from January 1990 to September 1994. The data contained about 900k articles (6.8M sentences). Since Japanese sentences have no spaces between words, we segmented the articles and the headlines into words with a morphological analyzer. The training data consisted of 623k unique words and the total number of words was 180M. Words in headlines were treated as topic words. To reduce the enormous number of word and topic word combinations, low-frequency words and function words, such as postpositional particles, were omitted. The number of distinct topic words was then about 70k words.

4.2 Evaluation Data

4.2.1 Speech Data

The evaluation data was broadcast news speech segmented into sentences. The evaluation data sets consisted of 29 articles totalling 142 utterances. Each article had from two to 14 utterances (average of five utterances). Fifteen male speakers (eight anchor speakers and seven other speakers) read the text. Their speech contained errors such as 'uh' at the beginning of a sentence or the correction of slips, and the recordings also included background noise or music.

4.2.2 Topic Data

We transcribed the evaluation speech data manually and assigned topic words to each transcribed article to use as correct topic words in the evaluation. We asked three subjects to select topic words as keys for retrieving the article and evaluated the extracted topics with them. Each of the subjects provided at least four topic words (average of 10 topic words) for each article. The set of topic words selected by one subject does not agree exactly with the set selected by another subject. Considering the variation of topic words, we made two topic evaluation sets, an AND set containing topic words for all the subjects selected, and an OR set containing all topic words selected by at least one subject. The topic words selected by the subjects were generally longer than the topic words in the training data, which were segmented by the morphological analyzer, and were not consistent with them. We segmented selected topic words with a morphological analyzer to enable evaluation. The average number of segmented topic words per article was 10.4 for the AND set and 35.7 for the OR set.

5. Experiments

5.1 Large Vocabulary CSR

Our LVCSR system comprises context-dependent phoneme HMMs and statistical n-gram language models [11]–[13]. The state-tied phoneme HMMs are triphone models designed using tree-based clustering [14] and were trained using phonetically-balanced sentences and dialogues read by 53 male speakers. The total number of utterances was 13,270 and the total volume of training data was approximately 20 hours. The total number of states was 2,106 and each state has four diagonal Gaussian mixtures. The bigram and trigram language models were trained using broadcast news manuscripts covering a five-year period and containing about 500k sentences. We estimated unseen n-gram probabilities using Katz's backoff smoothing method [15]. The vocabulary size of the system was 20k.

We used a multi-pass decoding strategy to apply trigram language models at less computational cost. In the first pass, the N-best hypotheses for an utterance were computed with the Viterbi algorithm using bigram language models on a simple word-loop network. Then, in the second pass, the hypotheses were rescored using trigram language models and the most likely hypotheses were chosen as the recognition results. In the experiments described below, the 300-best hypotheses were generated in the first pass and the same acoustic models were used in both passes.

Table 1 shows the test-set perplexity and LVCSR results for the evaluation speech data. The improvement in recognition performance with the trigram language model was rather small due to insufficient training data.
Table I LVCSR results for broadcast news speech data.

<table>
<thead>
<tr>
<th>Language model</th>
<th>Test-set perplexity</th>
<th>Word error rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>bigram</td>
<td>98</td>
<td>28.2</td>
</tr>
<tr>
<td>trigram</td>
<td>56</td>
<td>24.6</td>
</tr>
</tbody>
</table>

5.2 Topic Extraction Experiments

Figure 1 shows the results obtained for 50 topic words that were extracted from transcribed news speech and evaluated using the OR set. The experiments were conducted for each article (the boundaries between articles were given). For each article, 50 topic words were extracted in descending order of the relevance score to the article. The uppermost left point in the figure means the 1-topic-word extraction and the fifth point from the upper left means the 5-topic-word extraction result. The relevance score was calculated with a constant weighting factor $s(w_k) = 1$. Recall and precision are defined as

\[
\text{recall} = \frac{C}{T} \cdot 100, \quad \text{(9)}
\]

\[
\text{precision} = \frac{C}{H} \cdot 100, \quad \text{(10)}
\]

where $C$ is the number of correct topic words retrieved, $T$ is the total number of topic words in the evaluation topic sets, and $H$ is the total number of topic words retrieved. Recall and precision tend to have a trade-off relationship; however, achieving both high recall and high precision (i.e., plotted points are directed toward the upper right corner of the charts) is required for an information retrieval system. Figure 1 shows that the $\chi^2$-value based method (CHI: Eq. (7)) achieved better performance than the mutual information method (MI: Eq. (3)). It also shows that although the mutual information method weighted joint probability (wMI: Eq. (4)) achieved better results than the MI method, the results were not as good as those obtained using CHI. This is because the MI and the wMI methods are based on the conditional probabilities of topic words and are too sensitive to infrequent topic words, where the CHI method relatively depends on the frequency of the topic words.

Next, we applied the $\chi^2$-value-based topic extraction models to speech recognition results. Figure 2 shows the results obtained for recognized broadcast news speech (reg) evaluated with both the AND and OR evaluation sets. Results for manual transcription (txt) are also shown for comparison. It is clear that the results for the recognition results are inferior to those for transcription recognition errors are the reason for the difference.

We consider that five or so is a reasonable number of keywords for classifying or browsing news articles using keywords. We therefore extracted five topic words (the first five symbols plotted on a line graph in the charts, beginning from the upper left corner) and evaluated them with the OR set, which contains all topic words selected by the subjects. When extracted topic words are used for classifying or browsing news articles, the precision rate for the OR set is more important than the recall rate for the AND set. The results obtained were 83% precision (14% recall) for the manual transcription (txt) and 72% precision (12% recall) for the recognition results. The given topic words varied according to
the subjects; the overlapping rate of the topic words selected by the subjects was 74% on average. This means that if five topic words are chosen by a subject, in general three or more of them are the same as those selected by another subject. We therefore consider that precision of around 70% is sufficient for practical-use purposes.

5.3 Employing N-Best Candidates and Likelihood

Table 2 shows topic extraction results obtained using N-best candidates with the CHI method. The N-best candidates are the results obtained through second-pass decoding with the trigram language model. The figures in the table are the precision obtained in extracting one, five, or ten topic words and evaluating them with the OR set. We found that improved precision was obtained when topic words were extracted from N-best speech recognition candidates. In our experiments, the best performance for 5-topic-word extraction was achieved with N-values of 5 through 15.

We carried out experiments with relevance scores weighted with a factor calculated from acoustic and language likelihood. Table 3 shows the results of the topic extraction experiments from the 1-best candidates of speech recognition results. The range of the weighting factor is between one ($S_{max}$) and zero ($S_{min}$), and the acoustic likelihood values of words are normalized by the duration of each word because the likelihood values are accumulated within each word. Here, “word likelihood” refers to the product of the acoustic and language weighting factors. The maximum ($L_{max}$) and minimum ($L_{min}$) likelihood values for calculating the weighting factor were obtained from the evaluation set.

From Table 3, it can be seen that the use of language likelihood improved precision for 5- and 10-topic-word extraction, while the use of acoustic likelihood achieved very little improvement. With word likelihood, which as mentioned is the product of acoustic and language likelihood, even more improvement in precision was obtained than with language likelihood.

Higher precision was obtained with a combination of the N-best candidates and the weighting with likelihood than with either one of them separately. Table 4 shows the results obtained for topic extraction from N-best candidates with relevance score weighted with word likelihood in comparison with the results using transcribed speech and other methods. In Table 4, although the 10-best+weight achieved the best performance, it was only slightly better than the 10-best or the 1-best+weight. This suggests that employing the N-best candidates and employing the score weighting depend on similar information of speech recognition results.

Figure 3 shows the results of topic extraction from transcribed speech, 1-best candidate, and 10-best candidates, which achieved one of the best performances for 5-topic-word extraction (Table 2), with likelihood weighting. From the table and the figure it is clear that employing N-best candidates and their likelihood for topic extraction compensates for the degradation of topic extraction performance caused by speech recognition errors. As shown in the figure, the method improved the recall as well as the precision. For 1- to 10-topic-word extraction, the combination of 10-best candidates and word likelihood weighting eliminated more than 50% of the precision degradation caused by speech recognition errors.

6. Conclusion

In this paper we reported results obtained in topic extraction
from broadcast news speech based on continuous speech recognition. A topic represented by multiple topic words was extracted from a news article on the basis of statistical information about topic words. The topic can be used for retrieving or indexing news speech without playing them back.

We proposed topic extraction models that employ mutual information or χ²-values as the degree of relevance between topic words and words in news articles. In extracting five topic words using a χ²-value-based topic extraction model, 72% precision and 12% recall were obtained for speech-recognized news speech that had a word error rate of about 25%. To compensate for the performance degradation caused by speech recognition errors, we incorporated N-best candidates and likelihood weighting into the topic extraction and achieved 77% precision and 13% recall (with the OR set) for 5-topic-word extraction by improving both precision and recall. This is a practical precision level, since it exceeds the 74% overlap rate of topic words chosen by different subjects. The use of topic extraction technology combined with automatic transcription of broadcast news should prove to be a valuable tool for indexing digital news speech archives and retrieving news articles from the archives.

Broadcast news sometimes include new words that have not been seen in the past and those words are possibly important in the news. The topic extraction model represents such news articles by combining words in the article and words that do not appear in the article but have relevance to it. However, a method of dealing with such new unknown words has to be studied in future work, and it will make topic extraction a more powerful tool for the daily updated news.

The topic extraction model proposed in this paper can be estimated from data consisting of text and corresponding topic words. Although we evaluated this method only for the broadcast news task, it can be applied to any other task, such as telephone conversation with an operator, as long as an appropriate data is available.

Acknowledgments

The authors are grateful to Nihon Keizai Shimbun Inc. for allowing us to use the newspaper text database (Nikkei CD-ROM 90–94) for our research. We are also indebted to NHK (Japan Broadcasting Corporation) for providing us with a broadcast-news database for use in our research.

References


Sadaoki Furui received B.S., M.S. and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University, Tokyo, Japan in 1968, 1970, and 1978, respectively. He is currently a Professor at Tokyo Institute of Technology, Department of Computer Science. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction and has authored or coauthored over 350 published articles. From 1978 to 1979, he served on the staff of the Acoustics Research Department of Bell Laboratories, Murray Hill, New Jersey, as a visiting researcher working on speaker verification. He is a Fellow of the IEEE and the Acoustical Society of America. He is President of the Acoustical Society of Japan (ASJ), the International Speech Communication Association (ISCA), and the Permanent Council for International Conferences on Spoken Language Processing (PC-ICSLP). He is a Board of Governor of the IEEE Signal Processing Society. He has served on the IEEE Technical Committee on Speech and MMSP and on numerous IEEE conference organizing committees. He is Editor-in-Chief of the Transaction of the IEICE. He is also an Editorial Board member of Speech Communication, the Journal of Computer Speech and Language and the Journal of Digital Signal Processing. He has received the Yonezawa Prize and the Paper Award from the IEICE (1975, 1988, 93), and the Sato Paper Award from the ASJ (1985, 87). He has received the Senior Award from the IEEE ASSP Society (1989) and the Achievement Award from the Minister of Science and Technology, Japan (1989). He has received the Book Award from the IEICE (1990). He has also received the Mira Paul Memorial Award from the AFECT, India. In 1993 he served as an IFFE SPS Distinguished Lecturer.

Katsutoshi Ohtsuki received his B.E. in 1993 and M.E. in 1996, both from Waseda University, Tokyo, Japan. In 1996, he joined the Human Interface Laboratories of Nippon Telegraph and Telephone (NTT). Since 1999, he has been with NTT Cyber Space Laboratories. He has been working on the research and development of speech recognition technology. He is a member of the Acoustical Society of Japan.

Tatsuo Matsuo received his B.E. and M.E. degrees in electrical engineering from Waseda University in 1982 and 1984, respectively. Since 1984, he has been working at NTT Laboratories, where he has been engaged in speech circuit design for digital telephone equipments and statistical speech recognition. From 1992 to 1993 he was a visiting researcher at Bell Laboratories, AT&T working on speaker adaptation and N-best search methods. Since 1997, he has been with Multimedia Business Department (Broadband Business Department, since 2001) of NTT East Corporation. He is a member of the IEEE and the Acoustical Society of Japan.

Shoichi Matsuura received the B.S., M.S., and Ph.D. degrees in information science engineering from Kyushu University, Fukuoka, Japan in 1979, 1981 and 1992, respectively. Since 1982, he has been associated with NTT, working on automatic speech recognition. Dr. Matsuura is a member of the IEEE and the Acoustical Society of Japan.