Confidence estimation and keyword extraction from speech recognition result based on Web information

Hara Kensuke, Sekiya Hideki, Kawase Tetsuya, Tamura Satoshi, and Hayamizu Satoru
Department of Information Science, Gifu University, Japan
E-mail:{hara@asr.info.,sekiya@asr.info.,tetsuya@asr.info.,tamura@info.,hayamizu@}gifu-u.ac.jp

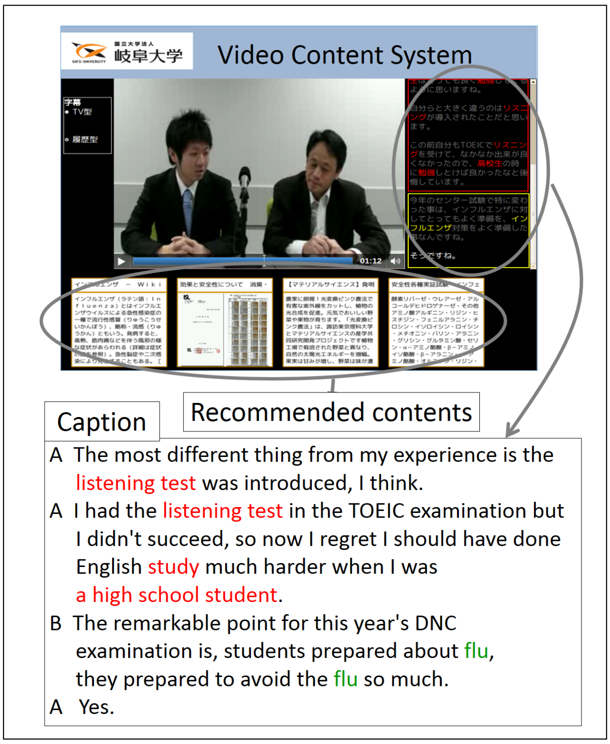
***Abstract*—This paper proposes to use Web information for confidence measure and to extract keywords for speech recognition results. Spoken document processing has been attracting attention particularly for information retrieval and video (audio-visual) content systems. For example, measuring a confidence score which indicates how likely a document or a segmented document includes recognition errors has been studied. It is well known keyword extraction from recognition results is also an important issue. For these purposes, in this paper, pointwise mutual information (PMI) between two words is employed. PMI has been used to calculate a confidence measure of speech recognition, as a coherence measure by co-occurrence of words. We propose to further improve the method by using a Web query expansion technique with term triplets which consist of nouns in the same document. We also apply PMI to keyword estimation by summing a co-occurrence score ($sumPMI$) between a targeting keyword candidate and each term. The proposed methods were tested with 10 lectures in Corpus of Spontaneous Japanese (CSJ) and 2 simulated movie dialogues. In the experiments it is shown that the estimated confidence score has high relationship with recognition accuracy, indicating the effectiveness of our method. And $sumPMI$ scores for keywords have higher values in the subjective tests.**

I. Introduction

As speech recognition technology has been developed, it becomes very important to determine how we should utilize and apply recognition results. Nowadays, spoken document processing has been attracting attention particularly for information retrieval and video (audio-visual) content systems. Here, spoken document means recorded speeches or recognition results, e.g. news and lectures.

We have worked to assist users to understand media content, and proposed a video content system which displays keywords and related information obtained from speech recognition results [1]. Fig. 1 shows an sample use of the system. For example, when there is a topic about flu in the video, our system recommends news about flu using the extracted keywords. History-type captions are represented in the right-hand area of the system. Recognition results are added in the area as the dialogue proceeds, and the user can check the history of the dialogue. Keywords in the captions are highlighted in different colors according to topics. In the bottom, recommended notes related with the topics to complement the contents are automatically displayed.

This paper studies two problems: confidence estimation and keyword extraction from speech recognition results. The confidence measure estimates how likely the recognition results are to be correct [2], [3]. In the keyword extraction method, terms which are likely to be important in a document are extracted [4], [5]. These two problems have been studied; however, they have still been challenging tasks due to recognition errors. In the conventional keyword extraction methods, TF-IDF is widely used to estimate an importance of each word. However, in the recognition results, mis-recognized terms sometimes have large TF-IDF scores, and are extracted as keywords. Table I shows top-five TF-IDF values and corresponding words in the recognition results. The words with asterisk are mis-recognized words. In the document used for Table I, a speaker mentioned EEG (Electroencephalogram). From Table I, the mis-recognized words ”consideration” and ”authority” are extracted as keywords because they have high TF-IDF values than the other words which must be suitable for keywords. This degrades the performance of whole spoken document system.



A The most different thing from my experience is the listening test was introduced, I think.
A I had the listening test in the TOEIC examination but I didn't succeed, so now I regret I should have done English study much harder when I was a high school student.
B The remarkable point for this year's DNC examination is, students prepared about flu, they prepared to avoid the flu so much.
A Yes.

Fig. 1. An sample use of video content system.

There is another issue; confidence estimation and keyword extraction have been investigated for whole spoken documents or long video contents, on the other hand, it is also necessary to study the methods for short video clips or segmented

TABLE I
AN EXAMPLE OF TOP-FIVE TF-IDF VALUES

word		TF-IDF
電位	electrical potential	10.587
整合	matching	5.682
***検討**	***consideration**	**3.546**
誘発	induction	3.478
***権威**	***authority**	**3.131**

documents. Some contents have several topics; for example, on the video in Fig. 1, two persons talked about university entrance examination, before the topic was changed to flu. For these contents, it is crucial to trace the topic according to keywords. And in order to do so, keyword extraction from short documents is expected.

Pointwise mutual information (PMI) between two words has been used for several purposes in text processing and speech recognition literatures [2], [6]. For example, PMI is used to estimate how likely a set of words coheres about a topic. On the other hand, we use PMI to give confidence measures for recognition results. Confidence measure for speech recognition using PMI was proposed as contextual coherence measure by co-occurrence of words, and a smoothing technique using statistical validation test was also proposed [2]. In the work, PMI was used based on the assumption that correctly recognized words are consistent and have high coherence values; in contrast, mis-recognized words are different from the other words and have small PMI scores. It is expected that the performance can be further improved by using the data related with the target document, e.g. retrieved information from Web.

In this paper, we propose a new method for the confidence estimation by using PMI and a Web query expansion technique. We also apply PMI to keyword estimation by summing a co-occurrence score ($sumPMI$) between a targeted keyword candidate and each term. In order to evaluate the proposed schemes, experiments were conducted using 10 lectures and 2 simulated movie dialogues. Here, movie dialogue means a conversation extracted from a video. A spoken document for each content was obtained by speech recognizer, and subsequently segmented into several frames. In each segment, a confidence score and keywords are computed and tested.

This paper is organized as follows. Section 2 shows how to estimate our recognition confidence for spoken document frames. Section 3 describes the keyword extraction method using PMI. Experimental conditions and results are mentioned in Section 4, and Section 5 concludes this paper.

II. CONFIDENCE ESTIMATION

A. PMI

PMI (Pointwise mutual information) is a measure that represents strength of relationship between two events. PMI is calculated as:

$$
\begin{aligned}
PMI(x,y) &= \log \frac{P(x,y)}{P(x)P(y)} && (1)\\
&= \log \frac{f(x,y)\cdot K}{f(x)f(y)} && (2)
\end{aligned}
$$

where $P(x)$ is an occurrence probability of word x, $f(x)$ is the number of occurrences of word x. $P(x,y)$ is a co-occurrence probability of words x and y, $f(x,y)$ is the number of co-occurrences of words x and y. $P(x)$, $P(y)$, and $P(x,y)$ are calculated using a data set for PMI computation; sometimes the data set corresponds to the document itself. The stronger relationship between x and y is, the larger $PMI(x,y)$ is. In this research, we use nouns in the same document as words x and y. When PMI is applied to a document, it is needed to divide the document into frames (windowing). Frames each including N content words (e.g. nouns) are extracted from the document every M content words, that is, window size and window shift are N and M, respectively. Note that in Eq.(2) K is the number of frames in the data set.

PMI has two problems. One is that PMI cannot measure the relationship of words which do not co-occur, and another is that PMI has an excessively large value when x and y rarely occur. In order to solve these problems, we employ the smoothed PMI [2]. This method uses t-test to examine whether $f(x)$ and $f(y)$ are large enough or not. The t-score $t(x,y)$ tests whether the difference between $P(x,y)$ and $P(x)P(y)$ is significant or not. The smoothed PMI is calculated as:

$$
\begin{aligned}
PMI(x,y) &= \begin{cases} \log \frac{\hat{f}(x,y)\cdot K}{f(x)f(y)} & \text{if } t(x,y) > \theta \\ 0 & \text{otherwise} \end{cases} && (3)\\
\hat{f}(x,y) &= \begin{cases} f(x,y) & \text{if } f(x,y) > 0 \\ \frac{N_1}{N_0} & \text{otherwise} \end{cases} && (4)\\
t(x,y) &= \frac{|\hat{f}(x,y) - \frac{f(x)f(y)}{K}|}{\sqrt{\hat{f}(x,y)}} && (5)
\end{aligned}
$$

where N_0 is the number of word pairs which do not co-occur in any frames in a document, N_1 is the number of word pairs which co-occur only once in the document. θ is a threshold of t-test, which is determined according to the significance level. In this paper, significance level is set to 5% ($\theta = 1.65$).

B. PMI-based confidence estimation

PMI is used to measure contextual coherence and finally to estimate confidence for spoken document [2]; in each document or bag of words, a contextual coherence value is calculated by summing PMI scores for all the word pairs up. Subsequently, a confidence score for the document is obtained as a mean value of the contextual coherence scores. This can be accomplished under the assumption that a mis-recognized word has weak relationships with the other words in the same document. A conventional confidence is calculated by speech recognizer based on 3-gram. The PMI-based method can estimate the confidence for longer-range frames, e.g. several utterances, than the decoder-based schemes.

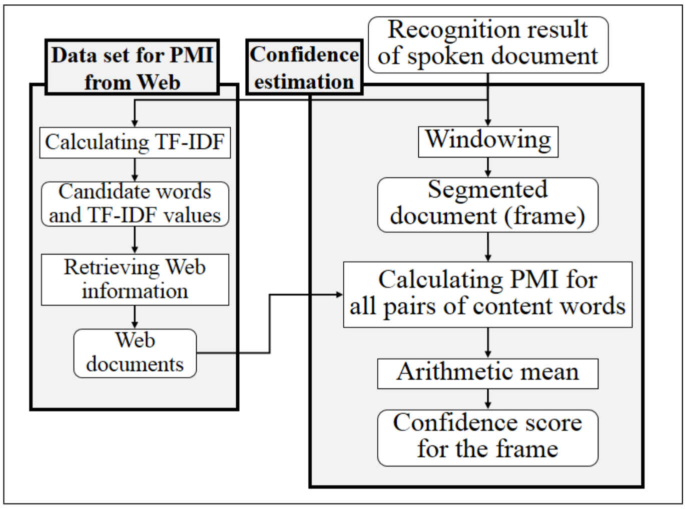

Data set for PMI from Web
Confidence estimation
Recognition result of spoken document
Calculating TF-IDF
Candidate words and TF-IDF values
Retrieving Web information
Web documents
Windowing
Segmented document (frame)
Calculating PMI for all pairs of content words
Arithmetic mean
Confidence score for the frame

Fig. 2. A flow of our proposed confidence estimation.

In the above previous work, it is shown that PMI can be used as a confidence score. In this paper, we propose to apply the PMI-based confidence estimation method to document frames, by utilizing Web retrieval results related with the frames. Employing Web retrieval techniques, we can obtain more related data to each frame for PMI computation than the general corpus. Thereby, our proposed method can estimate a recognition confidence score suitable not only for whole spoken document but also for its frames. The flow of our proposed method is illustrated in Fig. 2.

1) Building a data set for PMI computation from Web: A data set for PMI computation is built using the query expansion technique [7]. At first, for each word, a TF-IDF score is calculated using speech recognition results. Then, candidate words are chosen based on the TF-IDF values. Secondly, Web documents related with candidate words are obtained. A retrieval query consisting of three candidate words is prepared. Using the query, Web retrieval is conducted to obtain Web documents. The query is generated and the retrieval is done for all the possible combinations of candidate words. If we denote the number of candidate words by L, then the number of queries corresponds to ${}_LC_3$. In this paper, we studied two methods for Web retrieval.

Web_rank: for each query, the same number of documents are obtained. For example, if we get D_r documents for each query, we finally obtain $D_r \cdot {}_LC_3$ documents.

Web_weight: The number of total Web documents is fixed, and the number of retrieval results for each query is determined in proportion to TF-IDF values. For example, when we obtain D_w documents in total, the number of documents for a query $q = (w_{q,1}, w_{q,2}, w_{q,3})$ is:

$$D_w \cdot \frac{s(q)}{\sum_{q \in Q} s(q)} \tag{6}$$

where Q is a set of queries, and $s(q)$ is a sum of TF-IDF scores of $w_{q,1}$, $w_{q,2}$ and $w_{q,3}$.

2) Computing confidence measure for each frame: The spoken document is divided into several frames (windowing). In each frame F, a PMI value for each pair of content words is computed using the data set. Finally a recognition confidence score $C(F)$ for the frame is obtained as:

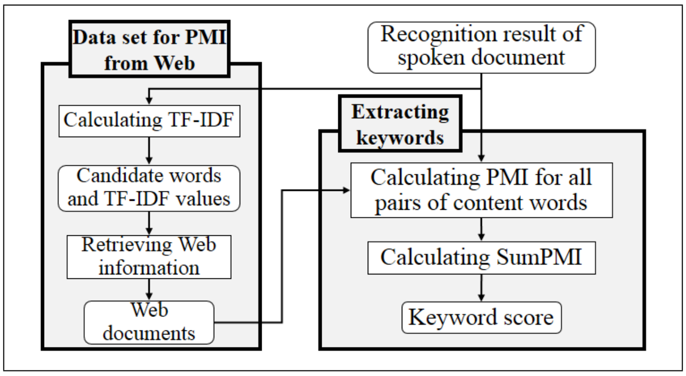

Data set for PMI from Web
Recognition result of spoken document
Calculating TF-IDF
Candidate words and TF-IDF values
Retrieving Web information
Web documents
Extracting keywords
Calculating PMI for all pairs of content words
Calculating SumPMI
Keyword score

Fig. 3. A flow of our proposed keyword extraction by PMI.

$$C(F) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} PMI(w_i, w_j) \tag{7}$$

where w_i, w_j are i-th and j-th content words, respectively, in the frame F ($F = \{w_1, w_2, ..., w_N\}$). The confidence score is normalized into $0 - 1$.

III. Keyword Extraction

We propose a keyword extraction method which uses PMI as well as Web query expansion. As mentioned, conventional TF-IDF is not suitable for keyword extraction from spoken documents, since mis-recognized terms often have high TF-IDF scores. In our proposed method, a keyword score for each term is calculated by summing up PMI between the term and the other words in a document. Terms in a recognition result, which do not or rarely co-occur with the neighbor words, have small PMI values. And actually these terms are often mis-recognized words. In contrast, terms co-occurring with the neighbor words have large scores. Such terms might characterize the document and become keywords. Therefore, we can extract keywords having large summed PMI values, which have strong relationships with the document.

A flow of our keyword extraction using PMI and Web query expansion is shown in Fig. 3. As same as the confidence measure scheme in the previous section, a data set for PMI computation is at first obtained from Web. Secondly, PMI scores for all the content word pairs are calculated using the data set. Thirdly, our proposed keyword score, that is $sumPMI$, is computed for each word. For a word w_i, $sumPMI(w_i)$ is calculated as:

$$sumPMI(w_i) = \sum_{j=1, j \neq i}^{N} PMI(w_i, w_j) \tag{8}$$

Finally keywords are obtained by choosing the terms having relatively high $sumPMI$ scores. Note that we can extract keywords either from the whole document, or segmented documents.

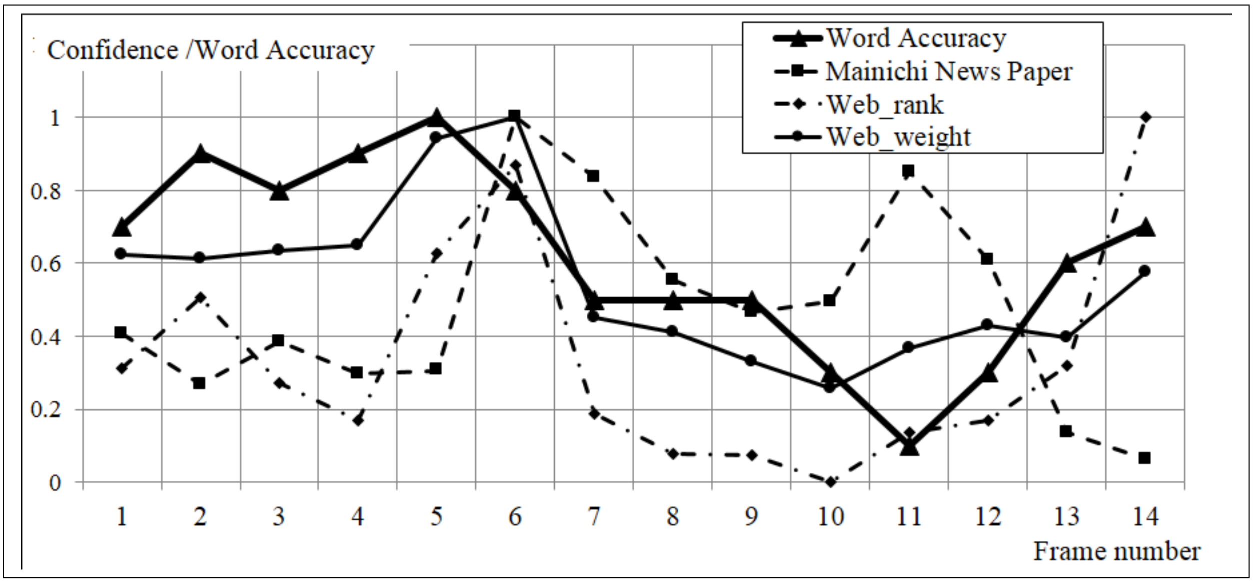


Fig. 4. Estimated confidence scores and word accuracy in each frame where $(N, M) = (10, 5)$.

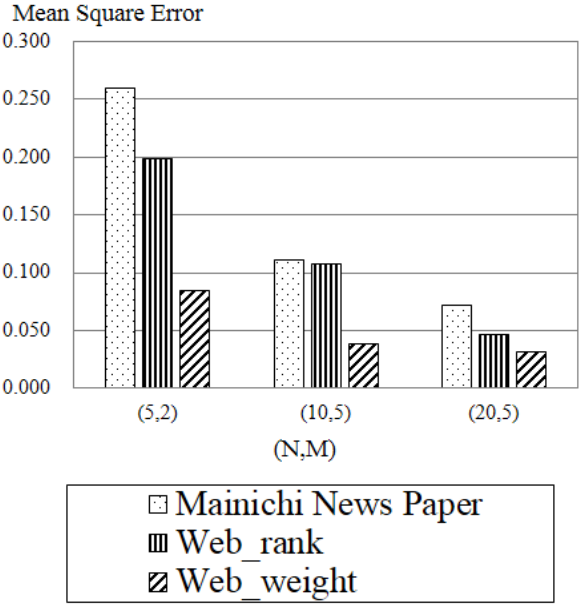


Fig. 5. Mean square errors for estimated confidence scores.

IV. Evaluation experiment

A. Experiments for confidence estimation

We evaluated our proposed confidence measure method using segmented spoken documents. In this paper, we regarded a recognition rate (word accuracy) in each frame as a correct confidence measure. Then in every frames, a mean square error between an estimated confidence score of each method and the word accuracy is computed and compared.

The experimental condition is described as follows; recognition results of 10 lectures in CSJ [8] and 2 simulated movie dialogues like Fig.1 were employed as test data. Recognition results of CSJ were used which have been distributed in the SpokenDoc of NTCIR9 [9]. Recognition results of two simulated dialogues were obtained using Julius [10] with its standard setup. The average accuracy is 69.2%. We tested three window size and window shift parameters as: $(N, M) = (5, 2), (10, 5), (20, 5)$. Only nouns are used as content words. Transcription documents of 2702 CSJ lectures were used to calculate IDF. In order to evaluate the effectiveness of our Web expansion technique, we tested three data sets for PMI computation:

1) A data set of *Mainichi* newspaper [11].
2) A data set of Web retrieved data (**Web_rank**), where $L = 8$, $D_r = 100$.
3) A data set of Web retrieved data (**Web_weight**), where $L = 8$, $D_w = 10000$.

Figure 4 depicts frame-based confidence scores of the proposed methods. The horizontal axis represents a frame number, and the vertical axis indicates a confidence score (word accuracy).

From Fig. 4, it is found that the result using co-occurrence information obtained by **Web_weight** is the most similar to the word accuracy; the method using the newspaper corpus was not acceptable, in contrast, the PMI-based confidence measure using the Web expansion technique was successful for short-range spoken documents. Fig. 5 shows mean square errors between estimated speech recognition confidence and the word accuracy. The same tendency was also obtained in Fig. 5. Regarding the windows size and shift, the method using **Web_weight** was not strongly affected by these parameters, on the other hand, the results using the newspaper corpus was drastically degraded as N and M became small. These facts suggest that we can estimate the confidence of short-range

TABLE II
An example of extracted keywords

A. TF-IDF		B. Mainichi News Paper		C. Web_rank		D. Web_weight	
周波数	frequency	日本	Japan	日本	Japan	周波数	frequency
キロヘルツ	kilohertz	確認	confirmation	周波数	frequency	コウモリ	bat
*思い	***thought**	*子供	***child**	パルス	pulse	生物	living
パルス	pulse	*工場	***factory**	生物	living	日本	Japan
生物	living	*思い	***thought**	細胞	cell	パルス	pulse
人工	atrificial	開始	start	人工	artificial	時間	time
*字	***letter**	基本	fundamental	機能	function	生息	inhabitation
*精神	***spirit**	機能	function	システム	system	システム	system

recognition result by using PMI as well as the Web expansion technology.

B. *Experiments for keyword extraction*

In order to evaluate the effectiveness of proposed keyword extraction, we did subjective experiments. After representing 10 keywords for each keyword extraction method, 20 subjects were asked to put preference scores of 1 to 4 (1 for the best, 4 for the worst) for the following four cases (A-D):

A. TF-IDF.
B. $sumPMI$ using the data set 1.
C. $sumPMI$ using the data set 2.
D. $sumPMI$ using the data set 3.

The subjects are asked to answer the following two questions.

i. lower misrecognition words

> Focus on mis-recognized keywords, and put higher preference scores if lower rank are assigned for mis-recognized keywords. In other words, evaluate whether mis-recognized words could be dropped or not.

ii. overall

> Put preference scores considering both of correctly recognized keywords and mis-recognized keywords.

The data used for the evaluation and the other experimental conditions were the same as the previous confidence estimation experiment.

Table II shows an example of extracted keywords using TF-IDF and our proposed methods. The words with asterisk show that the words are mis-recognized words. The topic for Table II is artificial sonar referring to an echo location function of bat living in Japan. From Table II, keywords extracted from TF-IDF contains mis-recognized words. On the other hand, our proposed methods using Web query expansion (C. and D.) didn't extract mis-recognized words. Fig. 6 shows the average preference scores for every keyword extraction methods. In the first evaluation, it is obvious that the methods using Web query expansion and $sumPMI$ (C. and D.) are better than the conventional TF-IDF method. In fact, the TF-IDF method contained several mis-recognized nouns, but $sumPMI$ methods had few or no recognition errors. And in the second evaluation, $sumPMI$ schemes using **Web_rank** and **Web_weight** are superior to the others. These results show that the proposed method has the effectiveness for keyword extraction.

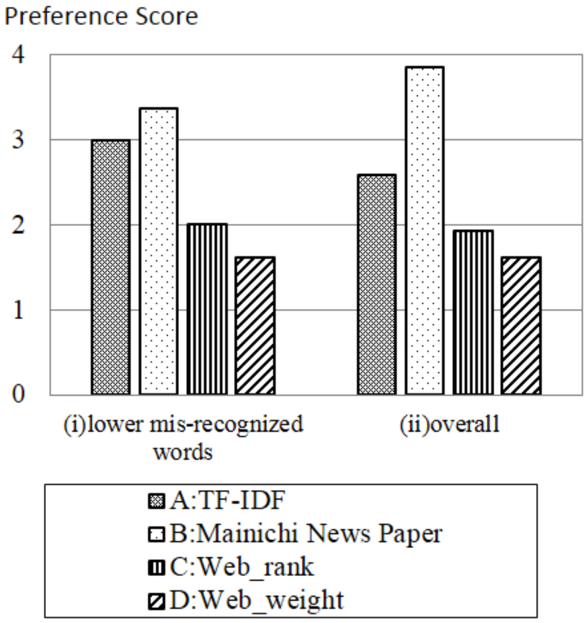


Fig. 6. Preference scores for keyword extraction methods.

V. Conclusion

A. *Summary*

This paper proposes new confidence estimation and keyword extraction methods for spoken documents, based on PMI and our Web query expansion technique. In the confidence estimation, PMI is computed using Web documents related with segmented spoken documents, and its summation is used as a confidence score. In the keyword extraction, $sumPMI$ is proposed and obtained for each content word as a keyword score. Experiments were conducted to compare the proposed methods with conventional methods. As a result, our proposed methods showed the significant performance. We can hence conclude that it is useful to use Web information related with the spoken document, and to employ PMI not only for recognition confidence estimation but also for keyword extraction.

B. *Future work*

While the usefulness in the confidence estimation and keyword extraction using Web information is proven, there are several issues as our future works. The first one is that obtaining Web documents requires retrieving time. To save the time, preparing co-occurrence information for every fields prior to PMI computation may be useful. In the second point, our proposed method uses word occurrence probabilities only. The other information, e.g. N-gram probabilities, might improve the performance. Alternatively, language model adaptation must be considered. In addition, recognition error detection and correction are expected to improve the performance of our keyword extraction. It can be pointed out that the confidence score of our proposed method is different from the conventional posteriori probability based ones. Combining our confidence estimation and the posteriori probability based schemes has a great possibility to get more accurate confidence scores. And applying confidence estimation to N-best recognition results will be useful to get more reliable recognition results.

ACKNOWLEDGMENT

This work was supported by JSPS Grant-in-Aid for Scientific Research Grant Number 22500151.

REFERENCES

[1] Okamono. M., Hasegawa. K., Sobue. M., Nakamura. A., Tamura. S., Hayamizu. S., "Topic based generation of caption and keywords for video content", Proc. APSIPA ASC 2010, pp.605-608, 2010.

[2] Asami. T., Nomoto. N., Kobashikawa. S., Yamagchi. Y., Masataki. H., Takahashi. S., "Spoken document confidence estimation using contextual coherence", Proc. INTERSPEECH2011, pp.1961-1964, 2011.

[3] H. Jiang., "Confidence measures for speech recognition: A survey", Speech Communication, Vol.45, pp.455-470, 2005.

[4] Matsuo. Y., Ishizuka. M., "Keyword extraction from a single document using word co-occurrence statistical information" International Journal on Artificial Intelligence Tools, Vol.13, pp.157-169, 2004.

[5] Matsumura. N., Ohsawa. Y., Ishizuka. M., "PAI: Automatic indexing for extracting asserted keywords from a document", New Generation Computing, Vol.21, Issue 1, pp.37-47, 2003 .

[6] Nikolaos. A., and Mark. S., "Evaluating Topic Coherence Using Distributional Semantics.", Proc. IWCS 2013, pp.13-22, 2013

[7] Hasegawa. K., Sekiya. H., Takehara. M., Niinomi. T., Tamura. S.,Hayamizu. S., "Toward improvement of SDR accuracy using LDA and query expansion for SpokenDoc." Proc. 9th NTCIR Workshop Meeting, pp.261-263, 2011.

[8] Maekawa. K., Koiso. H., Furui. S., Isahara. H., "Spontaneous speech corpus of Japanese" Proc. LREC2000, pp.947-952, 2000.

[9] NTCIR9, http://research.nii.ac.jp/ntcir/ntcir-9/

[10] Julius, http://julius.sourceforge.jp/

[11] *Mainichi* newspaper data, http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html