Automatic Estimation of Word Significance oriented for Speech-based Information Retrieval

Takashi Shichiri
Graduate School of Science and Tech.
Ryukoku University
Seta, Otsu 520-2194, Japan
shichiri@nlp.i.ryukoku.ac.jp

Hiroaki Nanjo
Faculty of Science and Tech.
Ryukoku University
Seta, Otsu 520-2194, Japan
nanjo@nlp.i.ryukoku.ac.jp

Takehiko Yoshimi
Faculty of Science and Tech.
Ryukoku University
Seta, Otsu 520-2194, Japan
yoshimi@nlp.i.ryukoku.ac.jp

Abstract

Automatic estimation of word significance oriented for speech-based Information Retrieval (IR) is addressed. Since the significance of words differs in IR, automatic speech recognition (ASR) performance has been evaluated based on weighted word error rate (WWER), which gives a weight on errors from the viewpoint of IR, instead of word error rate (WER), which treats all words uniformly. A decoding strategy that minimizes WWER based on a Minimum Bayes-Risk framework has been shown, and the reduction of errors on both ASR and IR has been reported. In this paper, we propose an automatic estimation method for word significance (weights) based on its influence on IR. Specifically, weights are estimated so that evaluation measures of ASR and IR are equivalent. We apply the proposed method to a speech-based information retrieval system, which is a typical IR system, and show that the method works well.

1 Introduction

Based on the progress of spoken language processing, the main target of speech processing has shifted from speech recognition to speech understanding. Since speech-based information retrieval (IR) must extract user intention from speech queries, it is thus a typical speech understanding task. IR typically searches for appropriate documents such as newspaper articles or Web pages using statistical matching for a given query. To define the similarity between a query and documents, the word vector space model or "bag-of-words" model is widely adopted, and such statistics as the TF-IDF measure are introduced to consider the significance of words in the matching. Therefore, when using automatic speech recognition (ASR) as a front-end of such IR systems, the significance of the words should be considered in ASR; words that greatly affect IR performance must be detected with higher priority.

Based on such a background, ASR evaluation should be done from the viewpoint of the quality of mis-recognized words instead of quantity. From this point of view, word error rate (WER), which is the most widely used evaluation measure of ASR accuracy, is not an appropriate evaluation measure when we want to use ASR systems for IR because all words are treated identically in WER. Instead of WER, weighted WER (WWER), which considers the significance of words from a viewpoint of IR, has been proposed as an evaluation measure for ASR. Nanjo et.al showed that the ASR based on the Minimum Bayes-Risk framework could reduce WWER and the WWER reduction was effective for key-sentence indexing and IR (H.Nanjo et al., 2005).

To exploit ASR which minimizes WWER for IR, we should appropriately define weights of words. Ideal weights would give a WWER equivalent to IR performance degradation when a corresponding ASR result is used as a query for the IR system. After obtaining such weights, we can predict IR degradation by simply evaluating ASR accuracy, and thus, minimum WWER decoding (ASR) will be the most effective for IR.

For well-defined IRs such as relational database retrieval (E.Levin et al., 2000), significant words (=keywords) are obvious. On the contrary, determining significant words for more general IR task (T.Misu et al., 2004) (C.Hori et al., 2003) is not easy. Moreover, even if significant words are given, the weight of each word is not clear. To properly and easily integrate the ASR system into an IR system, the weights of words should be determined automatically. Conventionally, they are determined by an experienced system designer. Actually, in conventional studies of minimum WWER decoding for key-sentence indexing (H.Nanjo and T.Kawahara, 2005) and IR (H.Nanjo et al., 2005), weights were defined based on TF-IDF values used in back-end indexing or IR systems. These values reflect word significance for IR, but are used without having been proven suitable for IR-oriented ASR. In this paper, we propose an automatic estimation method of word weights based on the influences on IR.

2 Evaluation Measure of ASR for IR

2.1 Weighted Word Error Rate (WWER)

The conventional ASR evaluation measure, namely, word error rate (WER), is defined as Equation (1).

$$\text{WER} = \frac{I + D + S}{N} \tag{1}$$

Here, *N* is the number of words in the correct transcript, *I* is the number of incorrectly inserted words, *D* is the number of deletion errors, and *S* is the number of substitution errors. For each utterance, DP matching of the ASR result and the correct transcript is performed to identify the correct words and calculate WER.

Apparently in WER, all words are treated uniformly or with the same weight. However, there must be a difference in the weight of errors, since several keywords have more impact on IR or the understanding of the speech than trivial functional words. Based on the background, WER is generalize and weighted WER (WWER), in which each word has a different weight that reflects its influence on IR, is introduced. WWER is defined as follows.

ASR result	:	a	b	c	d	e	f	
Correct transcript	:	a		c	d'		f	g
DP result	:	*C*	*I*	*C*	*S*		*C*	*D*

$$\text{WWER} = (V_I + V_D + V_S)/V_N$$

$$V_N = v_a + v_c + v_{d'} + v_f + v_g,\ V_I = v_b$$
$$V_D = v_g,\ V_S = \max(v_d + v_e, v_{d'})$$

v_i: weight of word *i*

Figure 1: Example of WWER calculation

$$\text{WWER} = \frac{V_I + V_D + V_S}{V_N} \tag{2}$$

$$V_N = \Sigma_{w_i} \quad v_{w_i} \tag{3}$$

$$V_I = \Sigma_{\hat{w}_i \in I} \quad v_{\hat{w}_i} \tag{4}$$

$$V_D = \Sigma_{w_i \in D} \quad v_{w_i} \tag{5}$$

$$V_S = \Sigma_{seg_j \in S} \quad v_{seg_j} \tag{6}$$

$$v_{seg_j} = \max(\Sigma_{\hat{w}_i \in seg_j} v_{\hat{w}_i}, \Sigma_{w_i \in seg_j} v_{w_i})$$

Here, v_{w_i} is the weight of word w_i, which is the i-th word of the correct transcript, and $v_{\hat{w}_i}$ is the weight of word $\hat{w}_i$, which is the i-th word of the ASR result. seg_j represents the j-th substituted segment, and v_{seg_j} is the weight of segment seg_j. For segment seg_j, the total weight of the correct words and the recognized words are calculated, and then the larger one is used as v_{seg_j}. In this work, we use alignment for WER to identify the correct words and calculate WWER. Thus, WWER equals WER if all word weights are set to 1. In Fig. 1, an example of a WWER calculation is shown.

WWER calculated based on ideal word weights represents IR performance degradation when the ASR result is used as a query for IR. Thus, we must perform ASR to minimize WWER for speech-based IR.

2.2 Minimum Bayes-Risk Decoding

Next, a decoding strategy to minimize WWER based on the Minimum Bayes-Risk framework (V.Goel et al., 1998) is described.

In Bayesian decision theory, ASR is described with a decision rule $\delta(X)$: $X \rightarrow \hat{W}$. Using a real-valued loss function $l(W, \delta(X)) = l(W, W')$, the

decision rule minimizing Bayes-risk is given as follows. It is equivalent to the orthodox ASR (maximum likelihood ASR) when a 0/1 loss function is used.

$$\delta(X) = \underset{W}{\text{argmin}} \sum_{W'} l(W, W') \cdot P(W'|X) \quad (7)$$

The minimization of WWER is realized using WWER as a loss function (H.Nanjo and T.Kawahara, 2005) (H.Nanjo et al., 2005).

3 Estimation of Word Weights

A word weight should be defined based on its influence on IR. Specifically, weights are estimated so that WWER will be equivalent to an IR performance degradation. For an evaluation measure of IR performance degradation, IR score degradation ratio (IRDR), which is described in detail in Section 4.2, is introduced in this work. The estimation of weights is performed as follows.

1. Query pairs of a spoken-query recognition result and its correct transcript are set as training data. For each query pair m, do procedures 2 to 5.
2. Perform IR with a correct transcript and calculate IR score R_m.
3. Perform IR with a spoken-query ASR result and calculate IR score H_m.
4. Calculate IR score degradation ratio (IRDR$_m = 1 - \frac{H_m}{R_m}$).
5. Calculate WWER$_m$.
6. Estimate word weights so that WWER$_m$ and IRDR$_m$ are equivalent for all queries.

Practically, procedure 6 is defined to minimize the mean square error between both evaluation measures (WWER and IRDR) as follows.

$$F(\mathbf{x}) = \sum_m \left(\frac{E_m(\mathbf{x})}{C_m(\mathbf{x})} - \text{IRDR}_m \right)^2 \rightarrow \min \quad (8)$$

Here, $\mathbf{x}$ is a vector that consists of the weights of words. $E_m(\mathbf{x})$ is a function that determines the sum of the weights of mis-recognized words. $C_m(\mathbf{x})$ is a function that determines the sum of the weights of the correct transcript. $E_m(\mathbf{x})$ and $C_m(\mathbf{x})$ correspond to the numerator and denominator of Equation (2), respectively.

In this work, we adopt the steepest decent method to determine the weights that give minimal $F(\mathbf{x})$. Initially, all weights are set to 1, and then each word weight (x_k) is iteratively updated based on Equation (9) until the mean square error between WWER and IRDR is converged.

$$x_k{}' = \begin{cases} x_k - \alpha & \text{if } \dfrac{\partial F}{\partial x_k} > 0 \\ x_k + \alpha & \text{else if } \dfrac{\partial F}{\partial x_k} < 0 \\ x_k & \text{otherwise} \end{cases} \quad (9)$$

where

$$\begin{aligned} \frac{\partial F}{\partial x_k} &= \sum_m 2 \left(\frac{E_m}{C_m} - \text{IRDR}_m \right) \cdot \left(\frac{E_m}{C_m} - \text{IRDR}_m \right)' \\ &= \sum_m 2 \left(\frac{E_m}{C_m} - \text{IRDR}_m \right) \cdot \frac{E_m' \cdot C_m - E_m \cdot C_m'}{C_m^2} \\ &= \sum_m 2 \left(\frac{E_m}{C_m} - \text{IRDR}_m \right) \cdot \frac{1}{C_m} \left(E_m' - C_m' \cdot \frac{E_m}{C_m} \right) \\ &= \sum_m \frac{2}{C_m} \left(\text{WWER}_m - \text{IRDR}_m \right) \left(E_m' - C_m' \cdot \text{WWER}_m \right) \end{aligned}$$

4 Weight Estimation on Orthodox IR

4.1 WEB Page Retrieval

In this paper, weight estimation is evaluated with an orthodox IR system that searches for appropriate documents using statistical matching for a given query. The similarity between a query and documents is defined by the inner product of the feature vectors of the query and the specific document. In this work, a feature vector that consists of TF-IDF values is used. The TF-IDF value is calculated for each word t and document (query) i as follows.

$$\text{TF-IDF}(t, i) = \frac{tf_{t,i}}{\frac{\text{DL}_i}{\text{avglen}} + tf_{t,i}} \cdot \log \frac{N}{df_t} \quad (10)$$

Here, term frequency $tf_{t,i}$ represents the occurrence counts of word t in a specific document i, and document frequency df_t represents the total number

of documents that contain word t. A word that occurs frequently in a specific document and rarely occurs in other documents has a large TF-IDF value. We normalize TF values using length of the document (DL_i) and average document lengths over all documents (avglen) because longer document have more words and TF values tend to be larger.

For evaluation data, web retrieval task "NTCIR-3 WEB task", which is distributed by NTCIR (NTC,), is used. The data include web pages to be searched, queries, and answer sets. For speech-based information retrieval, 470 query utterances by 10 speakers are also included.

4.2 Evaluation Measure of IR

For an evaluation measure of IR, discount cumulative gain (DCG) is used, and described below.

$$\mathrm{DCG}(i) = \begin{cases} g(1) & \text{if } i = 1 \\ \mathrm{DCG}(i-1) + \dfrac{g(i)}{\log(i)} & otherwise \end{cases} \tag{11}$$

$$g(i) = \begin{cases} h & \text{if } d_i \in H \\ a & \text{else if } d_i \in A \\ b & \text{else if } d_i \in B \\ c & \text{otherwise} \end{cases}$$

Here, d_i represents i-th retrieval result (document). H, A, and B represent a degree of relevance; H is labeled to documents that are highly relevant to the query. A and B are labeled to documents that are relevant and partially relevant to the query, respectively. "h", "a", "b", and "c" are the gains, and in this work, $(h, a, b, c) = (3, 2, 1, 0)$ is adopted. When retrieved documents include many relevant documents that are ranked higher, the DCG score increases.

In this work, word weights are estimated so that WWER and IR performance degradation will be equivalent. For an evaluation measure of IR performance degradation, we define IR score degradation ratio (IRDR) as below.

$$\mathrm{IRDR} = 1 - \frac{H}{R} \tag{12}$$

R represents a DCG score calculated with IR results by text query, and H represents a DCG score given by the ASR result of the spoken query. IRDR represents the ratio of DCG score degradation affected by ASR errors.

4.3 Automatic speech recognition system

In this paper, ASR system is set up with following acoustic model, language model and a decoder Julius rev.3.4.2(A.Lee et al., 2001). As for acoustic model, gender independent monophone model (129 states, 16 mixtures) trained with JNAS corpus are used. Speech analysis is performed every 10 msec. and a 25 dimensional parameter is computed (12 MFCC + 12ΔMFCC + ΔPower). For language model, a word trigram model with the vocabulary of 60K words trained with WEB text is used.

Generally, trigram model is used as acoustic model in order to improve the recognition accuracy. However, monophone model is used in this paper, since the proposed estimation method needs recognition error (and IRDR).

4.4 Results

4.4.1 Correlation between Conventional ASR and IR Evaluation Measures

We analyzed the correlations of conventional ASR evaluation measures with IRDR by selecting appropriate test data as follows. First, ASR is performed for 470 spoken queries of an NTCIR-3 web task. Then, queries are eliminated whose ASR results do not contain recognition errors and queries with which no IR results are retrieved. Finally, we selected 107 pairs of query transcripts and their ASR results as test data.

For all 107 pairs, we calculated WER and IRDR using corresponding ASR result. Figure 2 shows the correlations between WER and IRDR. Correlation coefficient between both is 0.119. WER is not correlated with IRDR. Since our IR system only uses the statistics of nouns, WER is not an appropriate evaluation measure for IR. Conventionally, for such tasks, keyword recognition has been performed, and keyword error rate (KER) has been used as an evaluation measure. KER is calculated by setting all keyword weights to 1 and all weights of the other words to 0 in WWER calculation. Figure 3 shows the correlations between KER and IRDR. Although IRDR is more correlated with KER than WER, KER is not significantly correlated with IRDR (correlation coefficient: 0.224). Thus, KER is not a suitable evaluation measure of ASR for IR. This fact shows that each keyword has a different influence on IR and

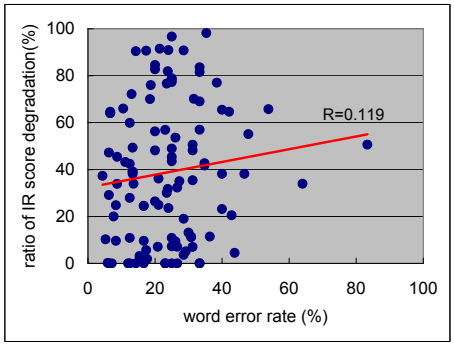

Figure 2: Correlation between ratio of IR score degradation and WER

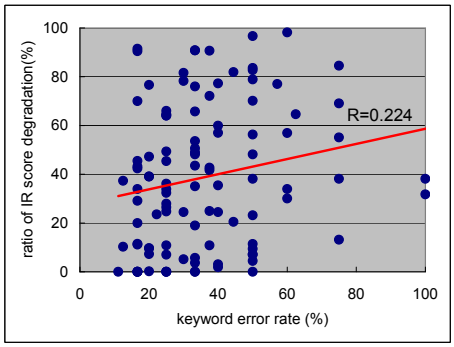

Figure 3: Correlation between ratio of IR score degradation and KER

should be given a different weight based on its influence on IR.

4.4.2 Correlation between WWER and IR Evaluation Measure

In ASR for IR, since some words are significant, each word should have a different weight. Thus, we assume that each keyword has a positive weight, and non-keywords have zero weight. WWER calculated with these assumptions is then defined as weighted keyword error rate (WKER).

Using the same test data (107 queries), keyword weights were estimated with the proposed estimation method. The correlation between IRDR and WKER calculated with the estimated word weights is shown in Figure 4. A high correlation between IRDR and WKER is confirmed (correlation coefficient: 0.969). The result shows that the proposed method works well and proves that giving a different weight to each word is significant.

The proposed method enables us to extend text-based IR systems to speech-based IR systems with typical text queries for the IR system, ASR results of the queries, and answer sets for each query. ASR results are not necessary since they can be substituted with simulated texts that can be automatically generated by replacing some words with others. On the contrary, text queries and answer sets are indispensable and must be prepared. It costs too much to make answer sets manually since we should consider whether each answer is relevant to the query. For these reasons, it is difficult to apply the method to a large-scale speech-based IR system. An estimation method without hand-labeled answer sets is strongly required.

An estimation method without hand-labeled answer sets, namely, the unsupervised estimation of word weights, is also tested. Unsupervised estimation is performed as described in Section 3. In unsupervised estimation, the IR result (document set) with a correct transcript is regarded as an answer set, namely, a presumed answer set, and it is used for IRDR calculation instead of a hand-labeled answer set.

The result (correlation between IRDR and WKER) is shown in Figure 5. Without hand-labeled answer sets, we obtained high correlation (0.712 of correlation coefficient) between IRDR and WKER. The result shows that the proposed estimation method is effective and widely applicable to IR systems since it requires only typical text queries for IR. With the WWER given by the estimated weights, IR performance degradation can be confidently predicted. It is confirmed that the ASR approach to minimize such WWER, which is realized with decoding based on a Minimum Bayes-Risk framework (H.Nanjo and T.Kawahara, 2005)(H.Nanjo et al., 2005), is effective for IR.

4.5 Discussion

In this section, we discuss the problem of word weight estimation. Although we obtained high correlation between IRDR and KWER, the estimation may encounter the over-fitting problem when we use small estimation data. When we want to design a speech-based IR system, a sufficient size of typical queries is often prepared, and thus, our proposed method can estimate appropriate weights for typical significant words. Moreover, this problem will be

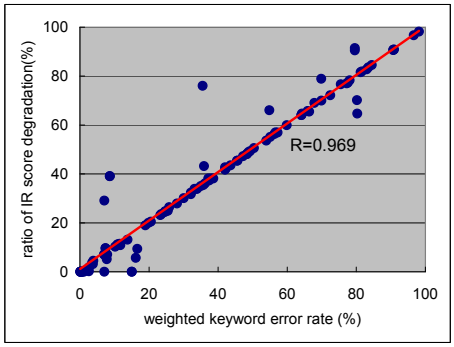

Figure 4: Correlation between ratio of IR score degradation and WKER (supervised estimation)

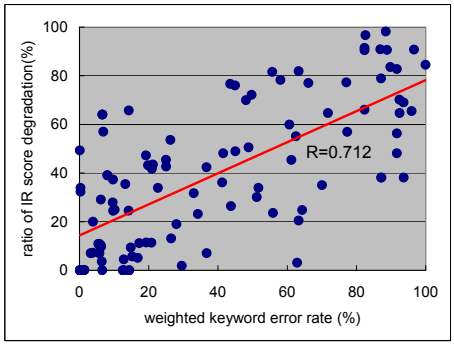

Figure 5: Correlation between ratio of IR score degradation and WKER (unsupervised estimation)

avoided using a large amount of dummy data (pair of query and IRDR) with unsupervised estimation. In this work, although obtained correlation coefficient of 0.712 in unsupervised estimation, it is desirable to obtain much higher correlation. There are much room to improve unsupervised estimation method.

In addition, since typical queries for IR system will change according to the users, current topic, and so on, word weights should be updated accordingly. It is reasonable approach to update word weights with small training data which has been input to the system currently. For such update system, our estimation method, which may encounter the over-fitting problem to the small training data, may work as like as cache model (P.Clarkson and A.J.Robinson, 1997), which gives higher language model probability to currently observed words.

5 Conclusion

We described the automatic estimation of word significance for IR-oriented ASR. The proposed estimation method only requires typical queries for the IR, and estimates weights of words so that WWER, which is an evaluation measure for ASR, will be equivalent to IRDR, which represents a degree of IR degradation when an ASR result is used as a query for IR. The proposed estimation method was evaluated on a web page retrieval task. WWER based on estimated weights is highly correlated with IRDR. It is confirmed that the proposed method is effective and we can predict IR performance confidently with such WWER, which shows the effectiveness of our proposed ASR approach minimizing such WWER for IR.

Acknowledgment: The work was partly supported by KAKENHI WAKATE(B).

References

A.Lee, T.Kawahara, and K.Shikano. 2001. Julius – an open source real-time large vocabulary recognition engine. In *Proc. EUROSPEECH*, pages 1691–1694.

C.Hori, T.Hori, H.Isozaki, E.Maeda, S.Katagiri, and S.Furui. 2003. Deriving disambiguous queries in a spoken interactive ODQA system. In *Proc. IEEE-ICASSP*, pages 624–627.

E.Levin, S.Narayanan, R.Pieraccini, K.Biatov, E.Bocchieri, G.D.Fabbrizio, W.Eckert, S.Lee, A.Pokrovsky, M.Rahim, P.Ruscitti, and M.Walker. 2000. The AT&T-DARPA communicator mixed-initiative spoken dialogue system. In *Proc. ICSLP*.

H.Nanjo and T.Kawahara. 2005. A new ASR evaluation measure and minimum Bayes-risk decoding for open-domain speech understanding. In *Proc. IEEE-ICASSP*, pages 1053–1056.

H.Nanjo, T.Misu, and T.Kawahara. 2005. Minimum Bayes-risk decoding considering word significance for information retrieval system. In *Proc. INTERSPEECH*, pages 561–564.

NTCIR project web page. `http://research.nii.ac.jp/ntcir/`.

P.Clarkson and A.J.Robinson. 1997. Language Model Adaptation using Mixtures and an Exponentially Decaying cache. In *Proc. IEEE-ICASSP*, volume 2, pages 799–802.

T.Misu, K.Komatani, and T.Kawahara. 2004. Confirmation strategy for document retrieval systems with spoken dialog interface. In *Proc. ICSLP*, pages 45–48.

V.Goel, W.Byrne, and S.Khudanpur. 1998. LVCSR rescoring with modified loss functions: A decision theoretic perspective. In *Proc. IEEE-ICASSP*, volume 1, pages 425–428.