Minimum Bayes-Risk Decoding considering Word Significance for Information Retrieval System

Hiroaki Nanjo[†], *Teruhisa Misu*[‡] *and Tatsuya Kawahara*[‡]

†Faculty of Science and Technology, Ryukoku University
Seta, Otsu 520-2194, Japan
nanjo@ryukoku-u.jp

‡School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
misu@ar.media.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp

Abstract

The paper addresses a new evaluation measure of automatic speech recognition (ASR) and a decoding strategy oriented for speech-based information retrieval (IR). Although word error rate (WER), which treats all words in a uniform manner, has been widely used as an evaluation measure of ASR, significance of words are different in speech understanding or IR. In this paper, we define a new ASR evaluation measure, namely, weighted word error rate (WWER) that gives a weight on errors from a viewpoint of IR. Then, we formulate a decoding method to minimize WWER based on Minimum Bayes-Risk (MBR) framework, and show that the decoding method improves WWER and IR accuracy.

1. Introduction

According to the progress of large vocabulary continuous speech recognition, the target of spoken language systems covers not only simple database queries such as flight information[1] but also general information retrieval (IR) tasks[2][3]. The IR typically searches for appropriate documents such as newspaper articles or Web pages using a statistical matching for a given query. To define similarity between a query and documents, the word vector space model or "bag-of-words" model is widely adopted, and some statistics such as *tf·idf* measure are introduced to take significance of the words into account in the matching. Therefore, when using automatic speech recognition (ASR) as a front-end of the IR systems, the significance of the words should be considered in ASR; such words that greatly affects IR performance should be detected with higher priority.

Conventionally, speech recognition aims at perfect transcription of the utterances, and the recognition accuracy is evaluated by the sentence error rate or word error rate (WER). WER is the most widely used evaluation measure of ASR accuracy, and it is defined as a minimum string edit distance (Levenshtein distance) between the correct transcript and the recognition hypothesis. By its definition, content words and functional words, even fillers, are treated in a same manner. Apparently, this is not consistent with the treatment of words in IR as mentioned above. For IR, "keywords" are more significant than other words. Therefore, WER is not an appropriate evaluation measure when we want to use ASR systems for IR.

In previous studies, keyword recognition accuracy was introduced only for well-defined tasks such as relational database query, where a set of keywords can be determined by the back-end system. In a general IR system, however, a definite set of keywords is not given in a deterministic way. Instead, all words have some numerical weights. Therefore, we introduce a new evaluation measure of ASR, that is, weighted word error rate (WWER) which considers a significance of words from a viewpoint of IR. Then, ASR is designed to minimize WWER based on the Minimum Bayes-Risk (MBR) framework[4]. In [5], we have shown the effect of the decoding method on the task of key-sentence indexing of oral presentations. In this paper, we demonstrate that the decoding method works well for more general IR.

2. Baseline Information Retrieval (IR) System

2.1. Dialogue Navigator

Dialog Navigator[6] has been developed at University of Tokyo as a document retrieval system for a large-scale software support knowledge base, which is provided by Microsoft Corporation. The knowledge base consists of following three kinds: glossary, frequently asked questions (FAQ), and database of support articles. All of them are described with a natural language text. The specification is shown in Table 1.

Table 1: Document set (Knowledge base)

Text collection	# documents	text size
glossary	4,707	1.4M byte
FAQ	11,306	12M byte
DB of support articles	23,323	44M byte

Table 2: Success rate of retrieval using ASR result

	IR success rate (%)			ASR rate
	set-1	set-2	total	(WER (%))
Manual transcript	69.91	63.89	67.74	0
ASR result	62.07	57.78	60.52	19.27

We have developed a speech recognition front-end by introducing an efficient confirmation strategies[7].

2.2. Baseline ASR system

Acoustic model is a gender-independent PTM (phonetic tied-mixture) triphone model which is trained with the JNAS corpus (newspaper reading: 40 hours). For language model training, we used several corpora: knowledge base (Table 1), actual query texts to Dialogue Navigator, and transcripts of simulated dialogue for software support. The text size in total is about 6.9M words. A trigram language model is trained with a vocabulary of 18K words. Speech recognition engine is Julius rev.3.4[8].

2.3. Baseline IR Result

In this paper, we evaluate our retrieval system with 499 utterances by 30 subjects. Here, for cross validation, we set up two utterance sets: set-1 and set-2 consists of 319 and 180 utterances of 30 subjects, respectively.

Table 2 lists the success rate of retrieval. We regard a successful retrieval if the correct document is included in the 10-best retrieval result. The result using the manual transcript indicates an utmost performance achievable by improvement in speech recognition. The retrieval accuracy of 60.52% is obtained on the average when the ASR result (WER: 19.27%) is used.

3. Evaluation Measure of ASR for IR – Weighted Word Error Rate (WWER)

The conventional word error rate (WER) is defined as equation (1). Here, *N* is the number of words in the correct transcript, *I* is the number of incorrectly inserted words (insertion errors), *D* is the number of deletion errors, and *S* is the number of substitution errors.

$$\text{WER} = \frac{I+D+S}{N} * 100 \tag{1}$$

For each utterance, DP matching of the ASR result and the correct transcript is performed to identify the correct words and calculate WER.

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

$$\text{WWER} = (V_I + V_D + V_S)/V_N * 100$$

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

v_i: weight of word *i*

Figure 1: Example of weighted word error rate (WWER) calculation

Apparently, in WER, all words are treated in a uniform manner. However, there must be a difference in the weight of errors, since several “keywords” have more impact on IR than functional words.

Based on the background, we generalize WER and introduce Weighted Word Error Rate (WWER), in which each word has a different weight according to its influence on IR. WWER is defined as follows.

$$\text{WWER} = \frac{V_I + V_D + V_S}{V_N} * 100 \tag{2}$$

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

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

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

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

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

Here, v_{w_i} is a weight of word w_i, which is the i-th word of the correct transcript, and $v_{\hat{w}_i}$ is a weight of word $\hat{w}_i$, which is the i-th word of the ASR result. And seg_j represents the j-th substituted segment and v_{seg_j} is a weight of segment seg_j. For the segment seg_j, the total weight of the correct words and total weight of the recognized words are calculated, and then a 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 is equivalent to WER if all word weights are set to 1. In Fig. 1, an example of WWER calculation is shown.

4. Minimum Bayes-Risk Decoding

In this section, we present a decoding strategy to minimize WWER. It is based on the Minimum Bayes-Risk (MBR) framework[4].

The orthodox statistical ASR is formulated as finding the most probable word sequence $\hat{W}$ for an input speech X, which is described in equation (7).

$$\hat{W} = \underset{W'}{\text{argmax}}\ P(W'|X) \tag{7}$$

In the 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[4].

$$\delta(X) = \underset{W}{\text{argmin}} \sum_{W'} l(W, W') \cdot P(W'|X) \tag{8}$$

It is equivalent to the orthodox ASR described in equation (7) when the 0/1 loss function is used in equation (8). In our baseline ASR system, this decoding is used.

In order to minimize WER, Levenshtein distance, which is equivalent to WER, is conventionally used as a loss function $l(W, W')$[4][9]. In this work, we want to minimize the weighted word error rate (WWER) to improve IR accuracy, thus we define the loss function based on WWER as described in equation (9).

$$\delta(X) = \underset{W}{\text{argmin}} \sum_{W'} \text{WWER}(W, W') \cdot P(W'|X) \tag{9}$$

Since $P(W'|X)$ can be rewritten as $P(W', X)/P(X)$ and $P(X)$ does not affect the minimization, equation (9) is rewritten as below.

$$\delta(X) = \underset{W}{\text{argmin}} \sum_{W'} \text{WWER}(W, W') \cdot P(W', X) \tag{10}$$

Moreover, a normalizing parameter λ is also introduced[4], so the decision rule is finally described as follows.

$$\delta(X) = \underset{W}{\text{argmin}} \sum_{W'} \text{WWER}(W, W')^{\lambda_1} \cdot P(W', X)^{\frac{1}{\lambda_2}} \tag{11}$$

To find the best word sequence W in a practical computation, an N-best list is generated by the baseline ASR system, and then N-best rescoring is performed.

5. Word Weight for IR System

5.1. Weight based on tf·idf Measure

There are apparently significant words that are potentially influential in IR. If they are not correctly recognized, retrieval result would be severely damaged. Thus, such words should have large weights.

From this point of view, word weights are defined using *tf·idf* measures which are typically used in IR. While inverse document frequency $idf(w)$ is an inverse of the number of documents that contain word w, term frequency $tf(w, d)$ is an occurrence count of word w in a specific document d and is not defined for the entire set of the documents. Averaging tf over all documents would weaken the characteristics of the words for IR.

In this work, we define a word weight based on the *tf·idf* measure in an indirect way as follows: First, we select five words having high *tf·idf* values in each document as its representative and extract the words that are representatives of many documents as potential keywords. For each word, we count the number of documents in which the word is a representative, to define a weight of the word. By this procedure, some words have a weight of 0. For such words, the weight is set to 1.

WWER using these weights is referred to as "WWER$_{\text{tf·idf}}$".

5.2. Weight based on LM of Target Documents

In[7], we proposed a relevance score to measure a potential degree of matching with the document set. It is computed phrase by phrase based on perplexity by the language model of the target document set. If the perplexity of a phrase is small, it means that the phrase is matching well the document and is more likely to be used in IR. Unlike *tf·idf* measure, the measure captures characteristics of word sequences. Actually, the relevance score is calculated by converting the perplexity via a sigmoid function. In this work, we define weights of words in the phrase by its relevance score.

WWER using the weight is referred to as "WWER$_{\text{PP}}$".

5.3. Combination of Word Weights

We also define WWER by combining the above two. A new measure (WWER) is defined by taking their geometric mean with a weight ϕ.

$$WWER = WWER_{\text{tf·idf}}^{\phi} * WWER_{\text{PP}}^{(1-\phi)} \tag{12}$$

6. Experimental Results

6.1. Comparison of ASR Evaluation Measures

We evaluated WWER minimization decoding and its effect on IR. For each utterance, we generate N-best list with $N = 100$. To estimate the rescoring parameters λ_1 and λ_2 in equation (11) and the weight ϕ in equation (12), we performed 2-fold cross validation, that is, set-1 was used as a development set to estimate parameters for evaluation of set-2, and set-2 was used to estimate parameters for evaluation of set-1. The rescoring parameters were determined to minimize the WWER of the development set.

Table 3: Comparison of evaluation measures

	IR Success Rate (%)
baseline	60.52
MBR (WWER$_{\mathrm{tf \cdot idf}}$)	60.32
MBR (WWER$_{\mathrm{PP}}$)	61.12
MBR (WWER)	61.72

Table 4: Improvement of WER and WWER

	WER (%)	WWER (%)
baseline	19.27	25.88
MBR (WWER)	20.31	25.35

The new ASR evaluation measures are compared to the baseline (MAP) framework by the IR scucess rate in Table 3. When we used the weights based on document LM (WWER$_{\mathrm{PP}}$), the success rate was improved. Although the use of word weights based on *tf·idf* alone (WWER$_{\mathrm{tf \cdot idf}}$) is not effective, the combination with WWER$_{\mathrm{PP}}$ leads to improvement of IR accuracy. It is confirmed that the proposed WWER and its minimization decoding are meaningful for IR.

6.2. Effect of WWER Minimization Decoding

Table 4 shows the change of WER and WWER achieved by WWER minimization decoding. Although WWER minimization decoding degraded WER, it improved WWER from 25.88% to 25.35%. Table 5 lists the improvement of IR rate in comparison with the conventional MBR decoding. According to the WWER improvement, the IR accuracy was improved to 61.72%. On the contrary, when MBR decoding is conducted to minimize WER instead of WWER (conventional MBR decoding), WER reduction was achieved (19.27% → 18.68%), but there is no improvement in IR (0.40% degraded).

These results show the significance of WWER minimization decoding for IR.

7. Conclusion

We proposed a new ASR evaluation measure oriented for IR, and introduced WWER based on word weights that is closely related with IR. Then, we designed a decoding strategy to minimize WWER. It is shown that WWER is an appropriate measure and WWER minimization decoding is effective for improving IR performance.

Acknowledgment: The work was partly supported by MEXT e-Society leading project and also by Ryukoku University Science and Technology fund.

Table 5: Comparison of decoding strategies

ASR method	IR Success Rate (%)		
(loss function)	set-1	set-2	total
baseline (0/1)	62.07	57.78	60.52
MBR (WER)	60.56	59.44	60.12
MBR (WWER)	**62.38**	**60.56**	**61.72**

8. References

[1] 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, "The AT&T-DARPA communicator mixed-initiative spoken dialogue system," in *Proc. ICSLP*, 2000.

[2] S. Harabagiu, D. Moldovan, and J. Picone, "Open-domain voice-activated question answering," in *Proc. COLING*, 2002, pp. 502–508.

[3] C. Hori, T. Hori, H. Isozaki, E. Maeda, S. Katagiri, and S. Furui, "Deriving disambiguous queries in a spoken interactive ODQA system," in *Proc. IEEE-ICASSP*, 2003.

[4] V.Goel, W.Byrne, and S.Khudanpur, "LVCSR rescoring with modified loss functions: A decision theoretic perspective," in *Proc. IEEE-ICASSP*, vol. 1, 1998, pp. 425–428.

[5] H.Nanjo and T.Kawahara, "A new ASR evaluation measure and minimum Bayes-risk decoding for open-domain speech understanding," in *Proc. IEEE-ICASSP*, 2005.

[6] Y. Kiyota, S. Kurohashi, and F. Kido, "Dialog Navigator: A question answering system based on large text knowledge base," in *Proc. COLING*, 2002, pp. 460–466.

[7] T.Misu, K.Komatani, and T.Kawahara, "Confirmation strategy for document retrieval systems with spoken dialog interface," in *Proc. ICSLP*, 2004, pp. 45–48.

[8] A.Lee, T.Kawahara, and K.Shikano, "Julius – an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.

[9] A.Stolcke, Y.Konig, and M.Weintraub, "Explicit word error minimization in N-best list rescoring," in *Proc. EUROSPEECH*, 1997, pp. 163–165.