MINIMUM BAYES-RISK DECODING WITH PRESUMED WORD SIGNIFICANCE FOR SPEECH BASED INFORMATION RETRIEVAL

Takashi Shichiri, Hiroaki Nanjo, Takehiko Yoshimi

Graduate School of Science and Technology, Ryukoku University
Seta, Otsu 520-2194, Japan
{shichiri, nanjo, yoshimi}@nlp.i.ryukoku.ac.jp

ABSTRACT

This paper addresses automatic speech recognition (ASR) oriented for speech based information retrieval (IR). Since the significance of words differs in IR, in ASR for IR, ASR performance should be evaluated based on weighted word error rate (WWER), which gives a different weight on each word recognition error from the viewpoint of IR, instead of word error rate (WER), which treats all words uniformly. In this paper, we firstly discuss an automatic estimation method of word significance (weights), and then, we perform ASR based on Minimum Bayes-Risk framework using the presumed word significance, and show that the ASR approach that minimizes WWER calculated from the presumed word weighs is effective for speech based IR.

Index Terms— Speech recognition, Information retrieval, Speech processing

1. INTRODUCTION

Speech based information retrieval (IR) is addressed. 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, some word statistics such as TF-IDF ("Term Frequency" and "Inverse Document Frequency") measure are introduced to consider the significance of words in the matching. Therefore, in speech based IR system which uses automatic speech recognition (ASR) as a front-end of text based IR systems, the significance of the words should be considered in ASR; words that greatly affect IR performance must be detected with higher priority. Moreover, ASR evaluation should be done from the viewpoint of the quality of misrecognized words instead of quantity. From this point of view, word error rate (WER) is not an appropriate evaluation measure for ASR 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. One of the authors has presented an ASR that minimizes WWER based on the Minimum Bayes-Risk (MBR) framework and showed that WWER reduction was effective for key-sentence indexing [1] and IR [2].

To exploit minimum 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, significant words (=keywords) are obvious. On the contrary, determining significant words for more general IR task [3] [4] 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, when we tested a minimum WWER decoding for key-sentence indexing and IR, weights were defined based on the word statistics 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.

Based on the background, we have proposed an automatic estimation method of word weights for speech based IR [5]. In this paper, we firstly describe our estimation method, and then, we show that minimum WWER decoding using the presumed word weights improves both ASR and IR performances.

2. ASR STRATEGY FOR INFORMATION RETRIEVAL

2.1. Evaluation Measure of ASR

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

$$\mathrm{WER} = (I + D + S)/N \tag{1}$$

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, we generalize WER and introduce weighted WER (WWER), in which each word has a different weight that reflects its influence on IR. WWER is defined as follows.

$$\text{WWER} = \frac{V_I + V_D + V_S}{V_N} \quad (2)$$

$$V_N = \Sigma_{w_i} \quad v_{w_i} \quad (3)$$

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

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

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

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

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.

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 [6] 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) when a 0/1 loss function is used.

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

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

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

In order to minimize WER, Levenshtein distance or WER is used as a loss function $l(W, W')$[6][7]. The minimization of WWER is realized using WWER as a loss function. We have already shown the minimization of WWER based on the framework [1] [2]. To find the best word sequence W in a practical way, we perform N-best list rescoring.

2.3. 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 [8]. As for acoustic model, gender independent monophone model (129 stats, 16 mixtures) trained with JNAS corpus are used. Speech analysis is performed every 10 msec. and a 25 dimensional parameter is computed (12 MFCC + 12 Δ MFCC + Δ Power). For language model, a word trigram model with the vocabulary of 60K words trained with WEB texts is used.

3. INFORMATION RETRIEVAL – WEB PAGE RETRIEVAL

3.1. Retrieval using Word Statistics

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.

$$\text{TF-IDF}(t, i) = \frac{tf_{t,i}}{\frac{\text{DL}_i}{\text{avglen}} + tf_{t,i}} \cdot \log \frac{N}{df_t} \quad (9)$$

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.

3.2. Task

For evaluation data, web retrieval task distributed by NTCIR [9] (NTCIR-3 WEB task) 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.

3.3. Evaluation Measure of IR

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

$$\text{DCG}(i) = \begin{cases} g(1) & \text{if } i = 1 \\ \text{DCG}(i-1) + \frac{g(i)}{\log(i)} & otherwise \end{cases} \quad (10)$$

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

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", and "b" are the gains, and in this work, $(h, a, b) = (3, 2, 1)$ 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.

$$\text{IRDR} = 1 - \frac{H}{R} \tag{11}$$

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. ESTIMATION OF WORD WEIGHTS

4.1. Algorithm

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 (IRDR). We have proposed the estimation method [5], which 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 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.

$$F(\mathbf{x}) = \sum_m \left(\frac{E_m(\mathbf{x})}{C_m(\mathbf{x})} - \text{IRDR}_m \right)^2 \rightarrow \min \tag{12}$$

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. The steepest decent method is adopted 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 until the mean square error between WWER and IRDR converges [5].

Table 1. Correlation between IR and ASR evaluation measures

	correlation coef. with IRDR
WER	0.386
KER	0.465
$\text{WKER}_{\text{sup.}}$	0.991
$\text{WKER}_{\text{semi}}$	0.693

The method enables us to extend text-based IR systems to speech-based IR systems with 1) typical text queries for the IR system, 2) ASR results of the queries, and 3) 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 semi-supervised estimation of word weights, is also tested. In semi-supervised 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.

4.2. Results

We analyzed the correlations of ASR evaluation measures with IRDR by selecting appropriate test data as follows. First, 13 queries with which no IR results are retrieved is eliminated from 47 queries. Then, ASR is performed for 340 spoken queries (34 queries x 10 speakers), and queries whose ASR results do not contain recognition errors are eliminated. Finally, we select 287 pairs of query transcript and its ASR result as test data.

Firstly, we investigate the correlation between conventional ASR measures and IRDR. Here, error rate of ASR result whose error rate is more than 100% is regarded as 100% because IRDR will not be greater than 100% according to its definition. IRDR of ASR result whose IRDR is less than 0 is regarded as 0 because ASR error rate will not be less than 0.

Table 1 lists the correlation between WER and IRDR. Correlation coefficient between both is 0.386. Since our IR system only uses the statistics of keywords (=nouns), keyword error rate (KER), which is calculated by setting all keyword weights to 1 and all weights of the other words to 0 in WWER calculation, is one of the most popular evaluation measures. Table 1 also lists the correlations between KER and IRDR. Although IRDR is more correlated with KER than WER, KER is not highly correlated with IRDR (correlation coefficient: 0.465).

Next, we investigate the correlation between WWER and

IRDR. Here, we assume that each keyword has a different 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 (287 queries), keyword weights were estimated with the our estimation method described in section 4. The correlation between IRDR and WKER calculated with the estimated word weights is also listed in Table 1. A high correlation is confirmed in supervised estimation case ($WKER_{sup.}$: 0.991 of correlation coefficient). Without hand-labeled answer sets (semi-supervised estimation), we obtained higher correlation ($WKER_{semi}$: 0.693 of correlation coefficient). The result shows that our estimation method works well, and proves that giving a different weight to each word is significant.

These results show that minimum WWER approach is more effective than WER or KER minimization approaches.

5. MINIMUM BAYES-RISK DECODING WITH WORD SIGNIFICANCE AND ITS EFFECT

In this section, we show that the ASR approach to minimize such WWER is effective for IR. Here, WWER minimization is realized with decoding based on a MBR framework described in section 2.2.

Table 2 lists the results. Each MBR decoding improved its minimization target. For example, WER minimization reduces WER from 21.25% to 20.87%, and $WKER_{sup.}$ minimization improved $WKER_{sup.}$ from 38.65% to 38.21%. Although WER and KER improvement were achieved by MBR, we did not obtain an improvement of IR accuracy. On the other hands, according to the minimization of $WKER_{sup.}$ and $WKER_{semi}$, which are defined with estimated word weights, we achieved an improvement of IR performance by 0.21% and 0.11%, respectively.

We investigated the ASR results in detail, and then, found that there were many queries whose ASR result by MBR is identical with the ASR result by conventional decoding (likelihood maximization). MBR decoding minimizing WER and KER generated different results for 55 and 50 queries, respectively. WKER minimization generated different results for 68 and 71 queries. WKER minimization yields more different hypothesises in ASR than WER or KER minimization. We found that for these queries ASR performance was relatively lower. ASR and IR improvement for these queries are also listed in Table 2 (lower part). According to the improvement of $WKER_{sup.}$ and $WKER_{semi}$, IR improvement of 1.06% and 0.58% has been achieved, respectively. The results show that presumed word weights are significant for speech based IR.

Table 2. Effect of MBR decoding with word weights

Results for whole test-set queries

minimization target in MBR (# of queries)	ASR error rate (%) 1-best → MBR	IRDR (%) 1-best → MBR
WER (287)	21.25 → 20.87	42.67 → 42.65
KER (287)	33.02 → 32.23	42.67 → 42.88
$WKER_{sup.}$ (287)	38.65 → 38.21	42.67 → **42.46**
$WKER_{semi}$ (287)	46.43 → 45.97	42.67 → **42.55**

Results for queries whose MBR results differ from 1-best results

minimization target in MBR (# of queries)	ASR error rate (%) 1-best → MBR	IRDR (%) 1-best → MBR
WER (55)	27.24 → 24.86	50.82 → 50.72
KER (50)	40.82 → 35.58	48.13 → 49.59
$WKER_{sup.}$ (68)	47.96 → 45.43	53.12 → **52.06**
$WKER_{semi}$ (71)	48.69 → 46.55	48.40 → **47.82**

6. CONCLUSION

We described the IR-oriented ASR based on Minimum Bayes-Risk framework using presumed word significance. For each word, its significance was estimated so that ASR performance will be equivalent to IR performance. We performed MBR decoding with the ASR evaluation measure "WWER" based on the presumed weights, and showed that the minimization of the WWER achieved the improvement of IR performance.

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

7. REFERENCES

[1] 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, pp. 1053–1056.

[2] H.Nanjo, T.Misu, and T.Kawahara, "Minimum Bayes-risk decoding considering word significance for information retrieval system," in *Proc. INTERSPEECH*, 2005, pp. 561–564.

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

[4] 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.

[5] T. Shichiri, H. Nanjo, and T. Yoshimi, "Automatic estimation of word significance oriented for speech-based information retrieval," in *Proc. IJCNLP*, 2008, pp. 204–209.

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

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

[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] "NTCIR project web page," `http://research.nii.ac.jp/ntcir/`.