ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Automatic Vocabulary Adaptation Based on Semantic Similarity and Speech Recognition Confidence Measure

Shoko Yamahata[1], *Yoshikazu Yamaguchi*[1], *Atsunori Ogawa*[2], *Hirokazu Masataki*[1], *Osamu Yoshioka*[1], *Satoshi Takahashi*[1]

[1]NTT Cyber Space Laboratories, [2]NTT Communication Science Laboratories, NTT Corporation,
yamahata.shoko@lab.ntt.co.jp

Abstract

Out-Of-Vocabulary (OOV) word utterances are unavoidable in speech recognition since the vocabulary size of a recognition dictionary is limited. And therefore, automatic vocabulary adaptation, which selects unregistered (i.e. OOV) words from relevant documents and registers them to a dictionary with their proper probability values, is an important technique. To improve recognition accuracy, a vocabulary adaptation method is required to register only relevant words that will actually be spoken in target utterances and not to register words that will not be spoken (i.e. redundant word entries). In this paper, we propose a novel automatic vocabulary adaptation method that satisfies these requirements based on semantic and acoustic similarities. Acoustic similarity is represented in speech recognition confidence measure. Experiments show that, with our method, the word selection accuracy is improved twice and the recognition accuracy focused on newly registered words is improved 15.1% in F-measure, compared with conventional methods.

Index Terms: out-of-vocabulary, vocabulary adaptation, semantic similarity, confidence measure

1. Introduction

Large vocabulary continuous speech recognition (LVCSR) systems are being applied to various tasks, such as lecture speech recognition [1], automatic transcription of conference minutes [2], and analysis of call center dialogs [3]. Although task specified keywords or key phrases (e.g. product names or technical term) are spoken frequently in such tasks, and they are usually important for understanding the recognition results, they will be out-of-vocabulary (OOV) since the vocabulary size of a recognition dictionary is limited. Consequently, a lot of researches that aim at automatic vocabulary adaptation, which select unregistered (i.e. OOV) words from relevant documents and register them to a dictionary with their proper probability values, have been directed. The major approach is collecting text documents which are relevant to target task from some corpus such as Web documents, and linearly interpolating their N-gram counts with the base language model [4, 5]. However these methods register all OOV words in relevant documents, words that are not spoken in the target utterances (i.e. redundant word entries) will also be registered and may displace the correct words, which are in-vocabulary words or other OOV words. In particular, spontaneous speech conversations like call center dialogs are rough conversation so that the recognition errors by redundant word entries are significant. Furthermore, since many of relevant documents are written in literary style, these documents have difficulty to improve recognition accuracy for conversation style speech data.

To improve recognition accuracy, we focus on the approach with selecting only OOV words that will actually be spoken in target spoken documents, since we aim to decrease recognition errors by suppressing registration of the redundant words. There are several studies that focused on selecting relevant OOV words. These studies selected words in terms of semantic similarity or acoustic similarity. The methods based on semantic similarity intend to select words that have similar meanings to the target spoken documents [6, 7, 8]. This approach can select the OOV words that semantically suit the target spoken document, but they include words only used in literary style. On the other hand, the methods based on acoustic similarity detects phones or syllables of an OOV word in the target spoken documents [9]. However, OOV words which are not relevant semantically are also likely to be included. Furthermore, these studies have not discussed as well as estimating proper probability values for each selected OOV word.

Hence, we propose a novel OOV selection method that uses both semantic and acoustic similarities. By combining of both measures, we can expect that only relevant OOV words are extracted with high-accuracy. We also propose a method of estimating proper probability for each OOV word based on semantic similarity under the condition of class-based language model, and show that our method achieves better recognition accuracy for registered words than conventional methods.

This paper is organized as follows. Section 2 details the proposed method of OOV selection and probability estimation. Section 3 describes experimental conditions and results, and Section 4 concludes this paper.

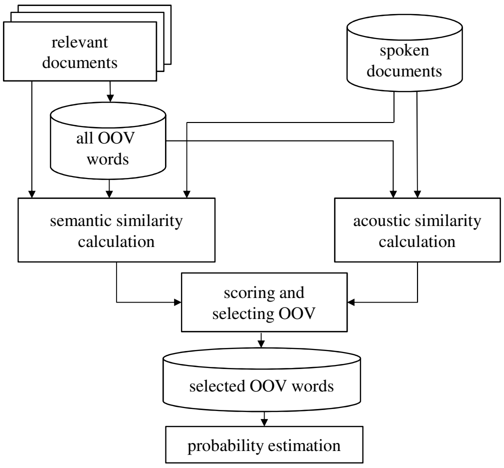
relevant documents
spoken documents
all OOV words
semantic similarity calculation
acoustic similarity calculation
scoring and selecting OOV
selected OOV words
probability estimation

Figure 1: *Overview of our method.*

2. Proposed method

This section describes the method of OOV selection and probability estimation using target spoken documents and text documents which are relevant to target spoken documents. Figure 1 shows an overview of our method. First, OOV words extracted from relevant documents are scored by semantic similarity and acoustic similarity, and highly ranked OOV words are selected. Second, the proper probability for each selected OOV word is estimated. In following subsections, we describe details of each processing.

2.1. OOV selection

The intention is to improve selection accuracy through the combination of semantic and acoustic similarities. The words not included in the base recognition dictionary are selected from relevant documents as OOV words. For each OOV word, semantic similarity and acoustic similarity with spoken documents are calculated, and we select OOV words by registration score calculated as the linear combination of both measures.

2.1.1. Semantic similarity

It is assumed that semantics of an OOV word is represented as co-occurrence word set with the OOV word in relevant documents, and also semantics of a spoken document is represented as word set taken from the spoken document. Therefore, semantic similarity is described as the consistency of words between these word set. We formulate word consistency by cosine similarity between the semantic vector of each OOV word and the semantic vector of each spoken document. Each semantic vector is represented as bag-of-words of each word set.

Semantic vector of each OOV word is calculated as follows. Let I be the number of all OOV words, and o_i be the i-th OOV word (1$\leq i \leq I$). In relevant documents, a sentence including o_i and preceding/following n sentences are extracted by the co-occurrence window of o_i. Then, for each sentence including o_i in relevant documents, the co-occurrence window is extracted. These windows are defined as $\mathrm{win_{all}}(o_i)$. The semantic vector $\mathbf{v}(o_i)$ consists of $\mathrm{tfidf}(w_t)$ of each word w_t in $\mathrm{win_{all}}(o_i)$. $\mathrm{tfidf}(w_t)$ is defined as follows:

$$\mathrm{tfidf}(w_t) = \mathrm{tf}(w_t) \times \log\left(\frac{N}{\mathrm{df}(w_t)} + 1\right) \quad (1)$$

where, $\mathrm{tf}(w_t)$ is frequency of w_t in $\mathrm{win_{all}}(o_i)$, $\mathrm{df}(w_t)$ is the number of relevant documents including w_t, N is the total number of relevant documents.

The semantic vector of each spoken documents is calculated as follows. Let S be a set of J spoken documents and let s_j be the recognition result of j-th spoken document (1$\leq j \leq J$). The semantic vector $\mathbf{v}(s_j)$ consists of $\mathrm{tfidf}(w_t)$ of each word w_t in s_j. $\mathrm{tfidf}(w_t)$ is defined as follows:

$$\mathrm{tfidf}(w_t) = \mathrm{tf}(w_t) \times \log\left(\frac{J}{\mathrm{df}(w_t)} + 1\right) \quad (2)$$

where, $\mathrm{tf}(w_t)$ is the frequency of w_t in s_j, $\mathrm{df}(w_t)$ is the number of spoken documents including w_t.

Then, the semantic similarity $r(o_i, S)$ is obtained by calculating the average of cosine similarity of $\mathbf{v}(o_i)$ and $\mathbf{v}(s_j)$ as follows:

$$r\left(o_i, S\right) = \frac{1}{J}\sum_j \frac{\mathbf{v}(o_i)\cdot\mathbf{v}(s_j)}{|\mathbf{v}(o_i)|\,|\mathbf{v}(s_j)|} \quad (3)$$

2.1.2. Acoustic similarity

To introduce an acoustic similarity, we detect OOV words in spoken documents. We assume the case that an OOV word is temporarily registered to the base recognition dictionary (i.e. a temporarily registered new word). In this case, if the newly registered word is not recognized, it is not probably uttered. If it is recognized, then, we have to confirm the reliability of the recognition result. If the recognition result is correct, the newly registered word is actually uttered. On the other hand, if the recognition result is incorrect, the newly registered word is not uttered. Hence, we use a confidence measure based on a posterior probability [10] to determine whether the word is correctly recognized or not.

Acoustic similarity is evaluated as follows. All OOV words extracted from relevant documents are temporarily registered with the base dictionary, and we recognize spoken documents by the temporary dictionary to obtain temporary recognition results. When o_i is detected K_{o_i} times in the temporary recognition results, we define $c_k(o_i)$ as the confidence measure where o_i is recognized k-th (1$\leq k \leq K_{o_i}$), and define $\bar{c}(o_i)$ as the average of $c_k(o_i)$.

$$\bar{c}(o_i) = \frac{1}{K_{o_i}}\sum_k c_k(o_i) \quad (4)$$

We define $\bar{c}(o_i)$ as acoustic similarity. Here, we use the average value of $c_k(o_i)$ since we statistically determine correctness of each OOV word with K_{o_i} samples.

2.1.3. Registration score

The registration score, $R(o_i)$, is calculated as the linear combination of $r(o_i, S)$ and $\bar{c}(o_i)$ for each o_i as follows:

$$R(o_i) = (1-\alpha)r\left(o_i, S\right) + \alpha\bar{c}(o_i) \quad (5)$$

where, α is the confidence weight. By scoring with this measure, we can select only the relevant words without selecting too many OOV words such as words that are not actually spoken, or words that are not relevant semantically.

2.2. Probability estimation

In our method, probability estimation is calculated from the semantic similarity described in Section 2.1.1. Various studies on the N-gram probability estimation of OOV words used linearly interpolated N-gram count among a base language model and relevant documents [4, 5]. However, these methods require a number of relevant documents that contain sufficient number of N-gram count of each OOV word, and it is not always possible to acquire such documents. Therefore, we use class-based language model to deal with the matter that acquiring sufficient number of N-gram count is very difficult. Also, we estimate class uni-gram probability, $p(o_i|C_{o_i})$, by weighting with semantic similarity, where C_{o_i} is the suitable class for each OOV word. There are two reasons for which weighting semantic similarity can estimate proper class uni-gram probability. First, each OOV semantic similarity correlates with each OOV word frequency in relevant documents. Second, semantic similarity represents the semantic correlation of OOV word in the spoken documents if there are insufficient numbers of relevant documents.

In our method, each OOV word is registered to a suitable class, and assigned a probability proportional to its semantic similarity. We simply use a linear formulation in this paper. To prevent the probability estimates from being completely different from the in-vocabulary (IV) words, $p(o_i|C_{o_i})$ is weighted with semantic similarity based on the average of $p(w_{\mathrm{IV}}|C_{o_i})$, where w_{IV} is an IV word in C_{o_i}, and $\bar{p}(w_{\mathrm{IV}}|C_{o_i})$ is the average of $p(w_{\mathrm{IV}}|C_{o_i})$. Each $p(o_i|C_{o_i})$ is calculated as follows:

$$p(o_i|C_{o_i}) = \frac{r(o_i, S)}{r_{\mathrm{base}}}\bar{p}(w_{\mathrm{IV}}|C_{o_i}) \qquad (6)$$

where r_{base} is a parameter which tunes the overall order of probability $p(o_i|C_{o_i})$. Since the absolute value of $r(o_i, S)$ differs with the relevant documents or spoken documents, weighting is executed appropriately with parameter r_{base}.

3. Experiments

We conducted two experiments to evaluate our method. The first evaluated OOV selection accuracy. The second evaluated the recognition accuracy of OOV words and confirmed that our method decreased the recognition error caused by redundant word entries.

3.1. Experimental setup

We used 2484 phone calls as spoken documents used to calculate acoustic similarity and 187 phone calls as evaluation data. Each phone call consisted of a simulated call center dialog. A total of 3427 Web documents were used as the relevant documents; all were subjected to morphological analysis; noun words not in base recognition dictionary were treated as OOV. The number of obtained OOV words is 3123, 88 of them were spoken in the evaluation data. These 88 OOV words were spoken a total of 546 times in the evaluation data. The acoustic model is triphone HMM and the base language model is class-based 3-grams. Baseline vocabulary size is 30k. Speech recognition decoder is VoiceRex [11]. The window size, n, used in calculating semantic similarity is 3.

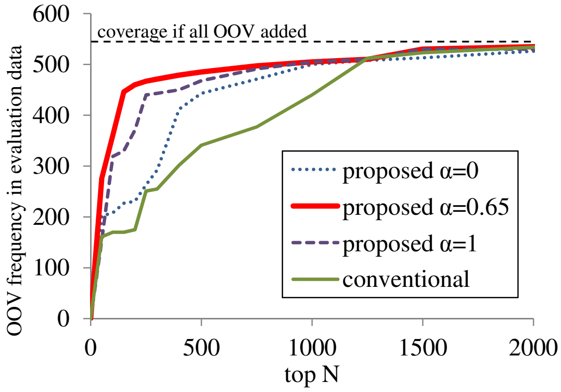

Figure 2: *Results of OOV selection accuracy.*

3.2. OOV selection accuracy

First, we evaluated selection accuracy by whether relevant OOV words were scored at the top. We compared the result of our method to that of the conventional OOV selection method using concept base (CB) [6] since it is similar to our method in using semantic similarity.

Figure 2 shows selection accuracy of OOV words yielded by the conventional method and the proposed method versus α. $\alpha = 0$ means using only semantic similarity, and $\alpha = 1$ means using only acoustic similarity. Our method could score the words in the evaluation data more highly than the conventional method; in particular, at the top of 1000 words. Also, the combination of both measures (proposed $\alpha = 0.65$) yielded the best accuracy. Specifically, the proposed method with $\alpha = 0.65$ achieved max coverage of 90% in the top 599; This coverage was matched by the conventional method only with the top 1128. Thus, the proposed method offers roughly twice the selection accuracy of the conventional method, and we confirmed the effectiveness of combining semantic similarity and acoustic similarities.

We describe our thought about relation between α and dataset features. α indicates the balance of selection accuracy for the two measures. The accuracy of semantic similarity depends on the reliability of semantic vectors for relevant OOV words. Therefore, to formulate reliably vector, relevant OOV words are present sufficiently in relevant documents. On the other hand, the accuracy of acoustic similarity depends on the consistency of OOV words among spoken documents used to calculate $\bar{c}(o_i)$ and evaluation data. From discussion above, the optimal α may very. We need to continue discussion about α features.

3.3. Recognition accuracy of OOV words

Next, we evaluate the recognition accuracy of registered OOV words. We select OOV words based on the result of Section 3.2 and estimate the probability of selected OOV words as described in Section 2.2. OOV words are selected with $\alpha = 0.65$. The number of selected OOV

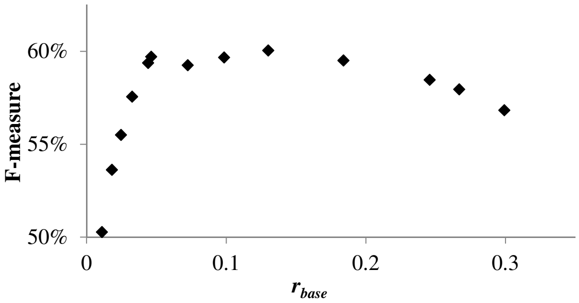

Figure 3: *Relevance between F-measure and* r_{base}

words is the top 599. The OOV words is assigned to the class of language model that has several products or service names that is used for call center dialog processing. We evaluated recognition accuracy by precision, recall, and F-measure of registered OOV words. Here, precision means the ratio of total recognized number to the correctly recognized number of registered OOV words, and recall means the ratio of total number in the evaluation data to the correctly recognized number of registered OOV words.

We first examined the relation between accuracy and r_{base}. Figure 3 plots F-measure versus r_{base}. In following experiment, we use the $r_{\mathrm{base}} = 0.13$ in the optimal value of F-measure.

Second, we compared the proposed method with conventional methods in terms of recognition accuracy of OOV words. We show two results of proposed method here. One is the optimal value in Figure 3 (“proposed estimate”), and the other is the case that each OOV word is assigned constant average probability $\bar{p}(w_{\mathrm{IV}}|C_{o_i})$ (“proposed const”) so as to allow comparison with the probability estimates for each OOV word. Conventional methods are linearly interpolating N-gram counts (LI) of base language model and relevant documents, and concept base (CB) [6]. For LI, we use the interpolation weight found empirically. For CB, we select OOV words that achieved the coverage of 90% when registering all OOV words by this method, and set class as OOV class and class uni-gram probabilities in this class as constant value (0.0001), which was also found empirically. Table 1 shows the precision, recall, and F-measure of each method. “Proposed estimate” improved the F-measure by 11.62% against CB, and 15.13% against LI. Compared with CB, “proposed const” is superior so that our OOV selection method is more effective than CB. Also, compared with LI, “proposed estimate” is superior. This means that the OOV selection and probability estimation by our method are more effective than registering all OOV words and estimating probability by N-gram count. Furthermore, the results of “proposed const” and “proposed estimate” show that probability estimation is effective.

Table 1: *Precision recall and F-measure results of conventional and proposed method.*

system	precision(%)	recall(%)	F-measure(%)
conventional CB	46.15	50.94	48.43
conventional LI	40.56	50.34	44.92
proposed const	52.42	54.16	53.27
proposed estimate	**74.29**	**50.38**	**60.05**

4. Conclusion

In this paper, we proposed a novel method for automatic vocabulary adaptation by OOV selection and probability estimation of each OOV word. OOV words are selected using semantic similarity and acoustic similarity, and the class uni-gram probabiliy of each OOV word is estimated from semantic similarity. Experiments showed that our method could select OOV words in spoken documents effectively, and could recognize OOV words while suppressing the registration of redundant word entries and hence decrease the recognition error rate. This study estimated the probability by simple liner formulation of semantic similarity, and we need more examination of other formulations.

5. References

[1] S. Furui., “Recent advances in spontaneous speech recognition and understanding”, in Proc. IEEE Workshop on Spont. Speech Proc. and Rec., pp. 1-6, 2003.

[2] S, Renals., T, Hain., and H, Bourlard., “Recognition and understanding of meetings: The AMI and AMIDA projects”, in Proc. IEEE Workshop Automatic Speech Recognition & Understanding, 2007.

[3] J, Mamou., D, Carmel., And R, Hoory., “Spoken Document Retrieval from Call-Center Conversations”, in Proc. ACM SIGIR., pp. 51-58, 2006.

[4] T, Kawahara., Y, Nemoto., and Y, Akita., “Automatic Lecture Transcription by Exploiting Presentation Slide Information for Language Model Adaption”, in Proc. ICASSP, pp.4929-4932, 2008.

[5] M, Creutz., S, Virpioja., and A, Kovaleva.,“Web augmentation of language models for continuous speech recognition of SMS text messages”, in Proc. ACL, pp-157-165, 2009. Web Data”, in Proc. MUSP’11, pp.125-131, 2011.

[6] K, Ohtsuki., N, Hiroshima., M, Oku., and A, Imamura., “Unsupervised Vocabulary Expansion for Automatic Transcription of Broadcast News”, in Proc. ICASSP, pp.1021-1024, 2005.

[7] C, E, Liu., K, Thambiratnam., and F, Seide., “Online Vocabulary Adaptation using Limited Adaptation Data”, in Proc. INTERSPEECH, pp.1821-1824, 2007.

[8] A, Ito., T, Meguro., S, Makino., and M, Suzuki., “Discrimination of Task-RelatedWords for Vocabulary Design of Spoken Dialog Systems”, in Proc. INTERSPEECH, pp.207-210, 2008.

[9] M, Hannemann., S, Kombrink., M, Karafiát., and L, Burget., “Similarity Scoring for Recognizing Repeated Out-of-VocabularyWords”, in Proc. INTERSPEECH, pp.897-900, 2010.

[10] F, Wessel., R, Schlüter., K, Macherey., and H, Ney., “Confidence Measures for Large Vocabulary Continuous Speech Recognition”, IEEE trans. speech audio process, Vol.9, No.3, pp.288-298, 2001.

[11] H, Masataki., D, Shibata., Y, Nakazawa., S, Kobashikawa., A, Ogawa., and K, Ohtsuki., “VoiceRex - Spontaneous speech recognition technology for contact-center conversations ” , NTT Tech. Rev., 5(1):22-27, 2007.