Multi-Pass ASR using Vocabulary Expansion

Katsutoshi Ohtsuki, Nobuaki Hiroshima, Shoichi Matsunaga† and Yoshihiko Hayashi‡

NTT Cyber Space Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa 239-0847 Japan
ohtsuki.katsutoshi@lab.ntt.co.jp

Abstract

Current automatic speech recognition (ASR) systems have to limit their vocabulary size depending on available memory size, expected processing time, and available text data for building a vocabulary and a language model. Although the vocabularies of ASR systems are designed to achieve high coverage for the expected input data, it cannot be avoided that input data includes out-of-vocabulary (OOV) words. This is called the OOV problem. We propose dynamic vocabulary expansion using a conceptual base and multi-pass speech recognition using an expanded vocabulary. Relevant words to content of input speech are extracted based on a speech recognition result obtained using a reference vocabulary. An expanded vocabulary that includes fewer OOV words is built by adding the extracted words to the reference vocabulary. The second recognition process is performed using the new vocabulary. The experimental results for broadcast news speech show our method achieves a 30% reduction in OOV rate and improves speech recognition accuracy.

1. Introduction

Out-of-vocabulary (OOV) words that are not included in a recognition vocabulary not only cannot be recognized when they appear in input speech but also affect their surrounding words and cause them to be misrecognized. Although the vocabularies of automatic speech recognition (ASR) systems are generally designed to cover as many expected words in input speech as possible, vocabulary sizes are limited depending on available memory size, expected latency of speech recognition processes, and the quantity and variety of available training text data. Therefore, OOV problems cannot be avoided by current ASR technologies, and more or fewer OOV words can be included in input speech. In some kinds of speech recognition applications such as broadcast news indexing [1,2] and meeting-speech transcription [3,4], newly appearing words and infrequent words specific to a certain topic, which tend to be OOV, are critical and therefore need to be recognized accurately.

Modeling OOV words using sub-word units has been shown to have an effect on OOV detection and estimating sub-word sequences [5, 6]. For retrieving contents by keyword queries in indexing applications, notations of OOV words, especially names of persons, places and products, need to be obtained instead of sub-word sequences. Information retrieval (IR) techniques were applied to the OOV problem in [7] and [8]. They dynamically adapted a vocabulary and a language model to the topic of input speech using relevant articles obtained from a database or the Web. Dynamic adaptation of vocabulary using morphological knowledge was also studied in [9] for Serbo-Croatian and German broadcast news speech.

Our approach directly estimates relevant words to input speech based on the conceptual base that models word co-occurrence patterns [10]. The conceptual base enables one to measure the distance between words in word co-occurrence pattern space, and direct estimation of relevant words can reduce OOV words more effectively than estimation of relevant words via relevant documents or sub-word sequences. An expanded vocabulary is built by adding the relevant words to a reference vocabulary and is used in the second recognition process. Since the vocabulary expansion process just adds relevant words to a reference vocabulary, the second recognition process runs just after the first recognition. We refer to this approach of multiple recognition processes as MASSIVE (Multi-pass Automatic Speech recognition Using Vocabulary Expansion).

The rest of the paper is organized as follows. Section 2 presents an overview of multi-pass speech recognition using vocabulary expansion. Section 3 describes vocabulary expansion based on the conceptual base. Section 4 presents the evaluation of MASSIVE on broadcast news speech. Section 5 concludes the paper.

2. Multi-Pass Speech Recognition using Vocabulary Expansion

The speech recognition process using vocabulary expansion is executed according to the following procedure.

1. First run: recognize input speech using a reference vocabulary.
2. Extract relevant words: estimate relevant words to input speech using the recognition result of the first run.
3. Rebuild vocabulary: build an expanded vocabulary by adding relevant words to the reference vocabulary.
4. Second run: recognize input speech using the expanded vocabulary.

Although this kind of procedure needs to run recognition processes for input speech at least twice and cannot be executed in real-time, some applications such as transcription or indexing of archived speech data does not have to obtain recognition results in real-time. In view of this, multi-pass recognition approaches combined with unsupervised adaptation techniques for improving recognition accuracy benefit such applications.

It takes much fewer processes to add words to a vocabulary than to rebuild a language model. In addition the second run can be executed much faster than the first run if the acoustic model does not change and acoustic likelihood calculation of

---

† Currently at Department of Computer and Information Sciences, Nagasaki University
‡ Currently at Graduate School of Language and Culture, Osaka University
the first run can be used for the second run.

3. Vocabulary Expansion using Conceptual Base

Relevant words for vocabulary expansion are extracted using word conceptual vectors that model word co-occurrence patterns. Words that appear in a recognition result of the first run are clustered based on word conceptual vectors, and words with similar vectors to the centroid vectors of the obtained clusters are extracted from a vocabulary database as relevant words.

3.1. Conceptual Base

A conceptual base is a database that consists of concept words and corresponding conceptual vectors. To build a conceptual base, first a word co-occurrence matrix is created by collecting word (content word) co-occurrence frequencies within one sentence in a training text corpus. Each row on the matrix is a word co-occurrence vector for a particular word. Since the matrix is quite sparse even with a huge corpus, it is transformed into a lower order matrix by singular value decomposition (SVD). The reduced matrix is composed of word conceptual vectors. The word conceptual vector is a vector representation of a word co-occurrence pattern, and if a pair of words has similar conceptual vectors, they tend to appear in the same sentence and are relevant to each other.

3.2. Vocabulary Database

As mentioned, word conceptual vectors are derived based on statistical word co-occurrence patterns so that they cannot be derived properly for infrequent words, which are to be extracted in vocabulary expansion for reducing OOV words. To assign conceptual vectors to all words regardless of their frequency, we calculated smoothed conceptual vectors as follows.

1. Obtain sentence conceptual vectors for each sentence that includes a target word by calculating centroid vectors of all concept words appearing in the sentence. A sentence conceptual vector \( \mathbf{v}_s \) for a sentence \( s \) which includes \( N_s \) concept words is calculated as
   \[
   \mathbf{v}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{c}_{w_i},
   \]
   where \( \mathbf{c}_{w_i} \) is a word conceptual vector of concept word \( w_i \).

2. Obtain a smoothed conceptual vector of the target word by calculating a centroid vector for the sentence conceptual vectors of all sentences that include the target word. A smoothed conceptual vector \( \mathbf{g}_j \) for a target word \( j \) is calculated as
   \[
   \mathbf{g}_j = \frac{\sum_s \delta(s,j)\mathbf{v}_s}{\sum_s \delta(s,j)},
   \]
   where \( \delta(s,j) \) is an auxiliary function that returns 1 if a target word \( j \) is included in a sentence \( s \) and returns 0 otherwise. The denominator of the right side of equation (2) is the number of sentences that includes word \( j \).

The smoothed conceptual vector is based on word conceptual vectors of co-occurrence words with a target word and can be derived even for infrequent words if they co-occur with concept words in a sentence. Regardless of the frequency of words, the distance between words depends on the distance between word conceptual vectors of co-occurred concept words for each word.

The vocabulary database consists of all the words appearing in training text data and their corresponding smoothed conceptual vectors and is used for vocabulary expansion.

3.3. Relevant Word Extraction

Relevant words to input speech are extracted from the vocabulary database based on the distance between input speech and the words in the vocabulary database. To measure the distance, a result of the first run for input speech is represented as conceptual vectors through the clustering process.

3.3.1. Clustering

All the concept words appearing in the result of the first run are clustered based on their word conceptual vectors using the centroid method. The clusters are considered to consist of a lot of concept words representing the content of input speech, and the clusters with small number of words are irrelevant to the input speech or include recognition error words. As centroid vectors of large clusters are considered to represent the content of the input speech, the relevant words are extracted based on them.

By ignoring small clusters, negative effects of speech recognition errors are reduced. Also, by using centroid vectors of multiple large clusters, input that includes multiple topics can be handled.

3.3.2. Relevance Score

The distance between the input speech and words in the vocabulary database is measured by using relevance score \( r(j,D) \) between a word \( j \) and a document \( D \) calculated as follows,

\[
r(j,D) = \max_k \frac{\mathbf{g}_j \cdot \mathbf{d}_k}{\| \mathbf{g}_j \| \| \mathbf{d}_k \|},
\]

where \( j = 1, \ldots, J \) is a word in the vocabulary database, \( \mathbf{g}_j \) is a smoothed conceptual vector of word \( j \), \( k = 1, \ldots, K \) is a cluster of concept words appearing in the results of the first run, and \( \mathbf{d}_k \) is a centroid vector of cluster \( k \).

By calculating the relevance score to the input document, or the recognition results of the first run, for all the words in the vocabulary database, the words with high relevance scores are extracted as relevant words to input speech.

3.4. Rebuilding Vocabulary for Second Run

The extracted relevant words are added to a reference vocabulary to build an expanded vocabulary. Adaptation techniques for n-gram language models can be applied to assign n-gram
probabilities to the new words. In case of class n-gram language models, existing probabilities can be shared for the new words. Since n-gram probabilities can be used without any changes in the case, the second run can be executed just after the vocabulary expansion.

4. Evaluations

We evaluated OOV reduction in proposed dynamic vocabulary expansion and ASR performance of MASSIVE using broadcast news speech data.

4.1. Evaluation Data

We used 30 Japanese broadcast news programs, which were aired on December 2002, as evaluation data. The programs, which vary from 5 to 30 minutes in length, include 265 news stories in total. The number of utterances is 2,898, and the number of words is 69,068 in the evaluation data. An evaluation was carried out for each news story.

4.2. Speech Recognition

A speech recognition engine called VoiceRex, which is being developed at NTT, was used for the speech recognition experiments. The acoustic models were 3-state 12-mixture state-tied triphone HMMs (male, female, and gender-independent) trained using approximately 300 hours of speech (150 hours each for male and female models). There were approximately 5,000 states for each model. The beginning of each utterance was evaluated with 96-mixture GMMS, each one representing one of the three acoustic models, and the model used for recognition was selected automatically [2].

The reference vocabularies and the trigram language models were trained using 450 thousand sentences (15 million words) of broadcast news transcription and newspaper text collected before December 2002. The vocabulary sizes were 25 thousands (25k) and 50 thousands (50k), and include words appearing ten or more times and words appearing twice or more, respectively. The 25k vocabulary covered 99.18% of the training text and 97.90% of the evaluation data (2.10% OOV). The 50k covered 99.87% and 98.98% (1.02% OOV).

4.3. Vocabulary Expansion

The conceptual base was trained using one year (2002) of newspaper text: approximately 100 thousand articles. The rows of the word co-occurrence matrix, or concept words, had 47 thousand frequent words in the training data, and the columns had 1 thousand frequent words except the 50 most frequent. The columns were compressed to 100 by SVD, that is to say, the word conceptual vectors had a hundred dimensions.

The vocabulary database consisted of 160 thousand words that were all the content words appearing in the training data, and smoothed conceptual vectors, as described in Section 3.2, were assigned to all of them. The clustering started from the status of each word constituting a single cluster and stopped when the number of clusters was less than one fifth from the start. The centroid vector of the single largest cluster was used for calculating the relevance score.

The top 100 and 1000 words in terms of relevance score were extracted and added to the reference vocabulary for each news story. Class n-gram probability of the OOV word class was 0.01 when adding 100 words and was 0.001 when adding 1000 words.

4.4. Experimental Results

The experimental results for the 25k vocabulary are shown in Table 1. For the reference vocabulary (REF), we compared our proposed method using the smoothed conceptual vector (SCV) with the results of a relevant document retrieval approach using the Okapi similarity measure (OKAPI) [7]. The proposed method reduced many more OOV words (\#oov: number of OOV, \%oov: OOV rate, \%red.: OOV reduction rate) than the conventional method for both number (100 or 1000) of additional words for each news story (\#add). Though the word error rate (%wer) of speech recognition increased when 100 words were added using the conventional method, the word error rates were improved when the vocabulary was expanded by the proposed method. Table 2 shows the results for the 50k vocabulary and a similar trend to the 25k. It is notable that the 25k vocabulary using vocabulary expansion yielded better speech recognition performance than the 50k reference vocabulary.

Table 1: Experimental results (25k vocabulary)

<table>
<thead>
<tr>
<th></th>
<th>#add</th>
<th>#oov</th>
<th>%oov</th>
<th>%red.</th>
<th>%wer</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF</td>
<td>-</td>
<td>1471</td>
<td>2.10</td>
<td>-</td>
<td>27.50</td>
</tr>
<tr>
<td>OKAPI</td>
<td>100</td>
<td>1440</td>
<td>2.06</td>
<td>2.1</td>
<td>27.71</td>
</tr>
<tr>
<td>(conventional)</td>
<td>1000</td>
<td>1159</td>
<td>1.66</td>
<td>21.2</td>
<td>27.38</td>
</tr>
<tr>
<td>SCV</td>
<td>100</td>
<td>1002</td>
<td>1.43</td>
<td>31.9</td>
<td>27.00</td>
</tr>
<tr>
<td>(proposed)</td>
<td>1000</td>
<td>506</td>
<td>0.72</td>
<td>28.9</td>
<td>26.99</td>
</tr>
</tbody>
</table>

Table 2: Experimental results (50k vocabulary)

<table>
<thead>
<tr>
<th></th>
<th>#add</th>
<th>#oov</th>
<th>%oov</th>
<th>%red.</th>
<th>%wer</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF</td>
<td>-</td>
<td>712</td>
<td>1.02</td>
<td>-</td>
<td>27.32</td>
</tr>
<tr>
<td>OKAPI</td>
<td>100</td>
<td>710</td>
<td>1.01</td>
<td>0.3</td>
<td>27.39</td>
</tr>
<tr>
<td>(conventional)</td>
<td>1000</td>
<td>580</td>
<td>0.83</td>
<td>18.5</td>
<td>27.22</td>
</tr>
<tr>
<td>SCV</td>
<td>100</td>
<td>586</td>
<td>0.84</td>
<td>17.7</td>
<td>27.08</td>
</tr>
<tr>
<td>(proposed)</td>
<td>1000</td>
<td>506</td>
<td>0.72</td>
<td>28.9</td>
<td>26.99</td>
</tr>
</tbody>
</table>

Table 3 shows the number of OOV words (\#oov) and the number of word errors (\#word error) with their reduction (\#red.) when the 25k reference vocabulary (REF) was expanded using the proposed method. The efficiency (%eff.), which is the ratio of the reduction in word error to the reduction in OOV words, was quite high, and the additional words contributed greatly to word error reduction. Table 4 shows the same numbers for the 50k vocabulary. In this case, the efficiency was more than 100%. That is to say more words came to be recognized than the obtained OOV words.

Tables 5 and 6 show the experimental results using the proposed method by breaking the news stories for evaluation down by their OOV rate with the 25k and 50k reference vocabularies. For both, OOV rates (%oov) were reduced for the news stories in every range of OOV rate (%oov range). The word error rates were also reduced for every range except the
news stories with less than 1% OOV rate for the 25k vocabulary in Table 5.

Table 3: Reduction of OOV and word errors
(25k vocabulary)

<table>
<thead>
<tr>
<th>#oov</th>
<th>#red.</th>
<th>#word error</th>
<th>#red.</th>
<th>%eff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF</td>
<td>1471</td>
<td>-</td>
<td>18995</td>
<td>-</td>
</tr>
<tr>
<td>100</td>
<td>1206</td>
<td>265</td>
<td>18767</td>
<td>228</td>
</tr>
<tr>
<td>1000</td>
<td>1002</td>
<td>469</td>
<td>18646</td>
<td>349</td>
</tr>
</tbody>
</table>

Table 4: Reduction of OOV and word errors
(50k vocabulary)

<table>
<thead>
<tr>
<th>#oov</th>
<th>#red.</th>
<th>#word error</th>
<th>#red.</th>
<th>%eff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF</td>
<td>712</td>
<td>-</td>
<td>18867</td>
<td>-</td>
</tr>
<tr>
<td>100</td>
<td>586</td>
<td>126</td>
<td>18703</td>
<td>164</td>
</tr>
<tr>
<td>1000</td>
<td>506</td>
<td>206</td>
<td>18643</td>
<td>224</td>
</tr>
</tbody>
</table>

Table 5: Classification by OOV rate (25k vocabulary)

<table>
<thead>
<tr>
<th>%oov range</th>
<th>#story</th>
<th>%oov</th>
<th>%wer</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF 100</td>
<td>100</td>
<td>0.53</td>
<td>0.35</td>
</tr>
<tr>
<td>0.5% ≤ %&lt;2%</td>
<td>57</td>
<td>1.39</td>
<td>1.00</td>
</tr>
<tr>
<td>2% ≤ %&lt;4%</td>
<td>46</td>
<td>2.48</td>
<td>1.57</td>
</tr>
<tr>
<td>3% ≤ %&lt;5%</td>
<td>31</td>
<td>3.49</td>
<td>2.39</td>
</tr>
<tr>
<td>4% ≤ %&lt;5%</td>
<td>15</td>
<td>4.56</td>
<td>2.45</td>
</tr>
<tr>
<td>5% ≤ %</td>
<td>27</td>
<td>8.55</td>
<td>6.55</td>
</tr>
</tbody>
</table>

Table 6: Classification by OOV rate (50k vocabulary)

<table>
<thead>
<tr>
<th>%oov range</th>
<th>#story</th>
<th>%oov</th>
<th>%wer</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF 100</td>
<td>100</td>
<td>0.48</td>
<td>0.29</td>
</tr>
<tr>
<td>0.5% ≤ %&lt;2%</td>
<td>53</td>
<td>1.41</td>
<td>0.64</td>
</tr>
<tr>
<td>2% ≤ %&lt;4%</td>
<td>30</td>
<td>2.42</td>
<td>1.31</td>
</tr>
<tr>
<td>3% ≤ %&lt;5%</td>
<td>20</td>
<td>3.42</td>
<td>2.07</td>
</tr>
<tr>
<td>4% ≤ %&lt;5%</td>
<td>5</td>
<td>4.59</td>
<td>3.53</td>
</tr>
<tr>
<td>5% ≤ %</td>
<td>15</td>
<td>7.54</td>
<td>5.59</td>
</tr>
</tbody>
</table>

5. Conclusions

This paper described a multi-pass approach of speech recognition using dynamic vocabulary expansion based on the conceptual base and its experimental results for broadcast news speech. Relevant words to input speech were extracted from the vocabulary database based on the relevance score that was calculated by using word conceptual vectors. The extracted words were added to the reference vocabulary to build an expanded vocabulary that is used for the second recognition process. The vocabulary expansion process is simple and n-gram probabilities do not have to be changed. Thus, the second run can be executed shortly after the first run. The experimental results show that the proposed method can reduce the number of OOV words effectively and the obtained words contribute to reducing word error rate efficiently.

6. Acknowledgements

The authors would like to thank Hisashi Ohara, Manager of the Speech, Acoustics and Language Laboratory, and Masahiko Hase, Director of Cyber Space Laboratories, for their support and encouragement of this work.

7. References