UNSUPERVISED VOCABULARY EXPANSION FOR AUTOMATIC TRANSCRIPTION OF BROADCAST NEWS

Katsutoshi Ohtsuki[†], *Nobuaki Hiroshima*[‡], *Masahiro Oku*[‡], *and Akihiro Imamura*[†]

[†]NTT Cyber Space Laboratories and [‡]NTT Cyber Solution Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka, Kanagawa 239-0847, Japan

ABSTRACT

In this paper, we present an unsupervised vocabulary adaptation method for large vocabulary continuous speech recognition based on relevant word extraction. This method addresses the out-of-vocabulary (OOV) problem, which is one of the most challenging problems in current automatic speech recognition (ASR) systems. Words relevant to the content of input speech are extracted from a vocabulary database, based on speech recognition results obtained in the first recognition process using a reference vocabulary. The relevance between words is calculated based on concept vectors, which are trained using word co-occurrence statistics. An expanded vocabulary that includes fewer OOV words is built by adding the extracted words to the reference vocabulary and used for the second recognition process. The experimental results for broadcast news speech show that our method achieves a 30% reduction in the OOV rate and also improves speech recognition accuracy.

1. INTRODUCTION

Out-of-vocabulary (OOV) words that are not included in a recognition vocabulary, not only are wrongly recognized when they appear in input speech, but also affect their surrounding words and make them be wrongly recognized. Although the vocabularies of automatic speech recognition (ASR) systems are generally designed to cover as many expected words in input speech as possible, their vocabulary sizes are limited, depending on the 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 are desirable 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]. However, to retrieve content by keyword queries in indexing applications, OOV word notations, 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 topics of input speech using relevant articles obtained from a database or the Web. Mahajan et al. [9] and Chen et al. [10] also updated language models using retrieved articles. Dynamic adaptation of vocabulary using morphological knowledge was also studied in [11] for Serbo-Croatian and German broadcast news speech.

Our approach directly estimates relevant words to input speech based on the concept base that models word co-occurrence patterns [12]. Word co-occurrence is also effective to adaptive language modeling [13, 14]. The concept 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 used in the second recognition process. Since the vocabulary expansion process just adds relevant words to a reference vocabulary, the second recognition process is able to run just after the first. We refer to this approach of multiple recognition processes as Multi-pass Automatic Speech recognition uSIng Vocabulary Expansion (MASSIVE).

The rest of the paper is organized as follows. Section 2 presents an overview of our approach to unsupervised vocabulary expansion. Section 3 describes relevant word extraction based on the concept base. Sections 4 and 5 present an evaluation of MASSIVE on broadcast news speech. Section 6 concludes the paper.

2. UNSUPERVISED VOCABULARY EXPANSION

Figure 1 shows a block diagram of the speech recognition system using unsupervised vocabulary expansion. The input speech is recognized by a reference vocabulary in the first run. Relevant words are extracted from the vocabulary database, based on the hypotheses obtained in the preceding run and are added to the reference vocabulary to build an expanded vocabulary. In the second run, the input speech is re-recognized with the expanded vocabulary to obtain recognition results.

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

It takes far fewer processes to add words to a vocabulary than to rebuild a language model. It also does not need the text corpus or word sequence frequency counts necessary for updating language models. In addition, the second run can be executed much faster than the first run if the same acoustic model is used for both runs, and the acoustic likelihood calculation result in the first run is kept for the second run.

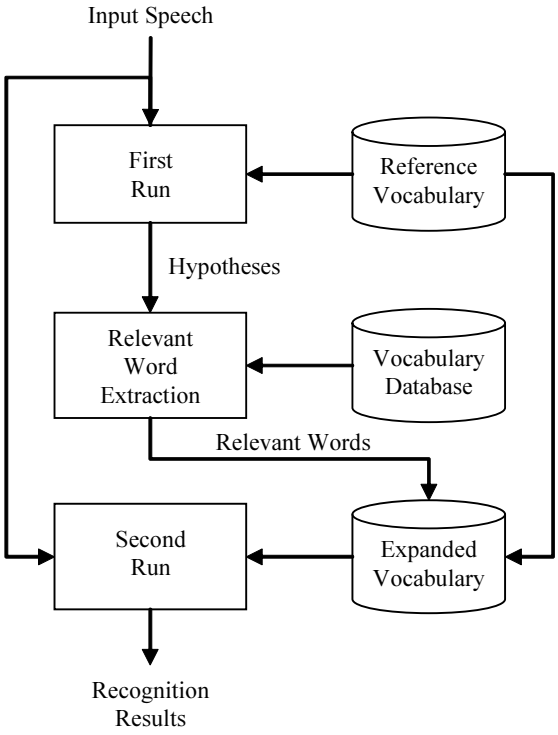

Fig. 1. Overview of the system

3. RELEVANT WORD EXTRACTION BASED ON CONCEPT VECTORS

Relevant word extraction is carried out based on inter-word distance, measured by word concept vectors that model word co-occurrence patterns. We calculate a representative vector of the hypotheses obtained in the first run and extract relevant words with similar vectors from a vocabulary database.

3.1. concept base

A concept base is a database that consists of concept words and the corresponding concept vectors. Concept words are mid- and high-frequent content words picked up from a text corpus. To build a concept base, first a word co-occurrence matrix is created by collecting word co-occurrence statistics for each concept word within one sentence in a training text corpus. Each row on the matrix is a word co-occurrence vector for a particular concept 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 concept vectors. The word concept vector is a vector representation of a word co-occurrence pattern, and, if a pair of words has similar concept vectors, they tend to appear in the same sentence and are relevant to each other.

3.2. Vocabulary database

The vocabulary database consists of words for vocabulary expansion, which are not in the reference vocabulary, and each word comes with concept vectors for calculating their relevance to words in the hypotheses obtained from the first run. However, as mentioned above, word concept 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 to reduce the number of OOV words. We assign smoothed concept vectors to all words, regardless of their frequency. A smoothed concept vector $\mathbf{g}_j$ for target word j is calculated as

$$\mathbf{g}_j = \frac{\sum_s \delta(s,j)\mathbf{v}_s}{\sum_s \delta(s,j)}, \tag{1}$$

where $\mathbf{v}_s$ is the mean vector of sentence s, and $\delta(s,j)$ is an auxiliary function that returns 1 if target word j is included in sentence s, otherwise it returns 0. The denominator on the right side of equation (1) is the number of sentences that include word j. The sentence concept vector $\mathbf{v}_s$ and the auxiliary function $\delta(s,j)$ are calculated as

$$\mathbf{v}_s = \frac{1}{N_s}\sum_{k=1}^{N_s} \mathbf{c}_{w_k}, \tag{2}$$

$$\delta(s,j) = \begin{cases} 1 & \text{if } j \in s \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $\mathbf{c}_{w_k}$ is the word concept vector of concept word w_k, and N_s is the number of concept words in sentence s.

The smoothed concept vector is based on the word concept vectors of words that appear in the same sentence as a target word and can be derived even for infrequent words if they appear with concept words in a sentence. Regardless of the frequency of words, the distance between words depends on the distance between the word concept vectors of the concept words that appear along with each word. The vocabulary database consists of all the words that appear in training text data and their corresponding smoothed concept vectors.

3.3. Relevance score

The distance between a document and words in the vocabulary database is measured using relevance scores. The relevance score $r(j,d)$ between a word $j = 1, 2, \ldots, J$, which is a word in the vocabulary database, and a document d is calculated by a cosine measure as

$$r(j,d) = \frac{\mathbf{g}_j \cdot \mathbf{v}_d}{|\mathbf{g}_j| \cdot |\mathbf{v}_d|}, \tag{4}$$

where $\mathbf{g}_j$ is a smoothed concept vector of word j, and $\mathbf{v}_d$ is a mean vector of the concept vectors that appear in a document d.

By calculating the relevance score to an input document, or a hypothesis from the first run, for all the words in the vocabulary database, words with high relevance scores are extracted as words relevant to the input speech.

3.4. Building expanded vocabulary

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,

the existing probabilities can be shared for the new words. Since n-gram probabilities can be used without any changes in this case, the second run can be executed just after vocabulary expansion. For the following experiments, we used class n-gram language models to distribute n-gram probabilities to the added words without adapting or re-training the models.

4. EXPERIMENTAL SETUPS

We evaluated the OOV and word error reduction of our method using broadcast news speech data.

4.1. Evaluation data

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

4.2. Vocabulary expansion setup

The concept base was trained using one year (2002) of newspaper text: approximately 100 thousand articles. The concept words were 47 thousand frequent-content words from the training data and the word concept vectors had 100 dimensions, obtained by compressing the co-occurrence statistics of 1000 frequent words.

The vocabulary database consisted of 160 thousand words that were all the content words that appeared in the training data, and smoothed concept vectors were assigned to all of them as described in Section 3.2.

The top 100 and 1000 words in terms of relevance scores were extracted and added to the reference vocabulary for each news story. The added words equally shares the n-gram probabilities of the OOV word class. That is to say, the unigram probability of the added words in the OOV word class was 0.01 when 100 words were added and 0.001 when 1000 were added.

4.3. Speech recognition setup

A speech recognition engine called VoiceRex, currently 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). 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 based on likelihood of GMMs [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 language model is a class-based model and it has an OOV word class. The recognition process using the expanded vocabulary is executed with the same language model as the first recognition process. The added words uses the OOV word class probabilities as mentioned above. We prepared two vocabularies whose sizes were 25 (25k) and 50 thousands words (50k). 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).

5. 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 concept vector (SCV) with the results of a relevant document retrieval approach using the Okapi similarity measure (OKAPI) [7]. The proposed method greatly reduced the number of OOV words (#oov: number of OOV, %oov: OOV rate, %red.: OOV reduction rate) more than the conventional method for both numbers (100 or 1000) of additional words (#add) for each news story. The word error rates (%wer) were also reduced by the proposed method, where the conventional method increased the error rate by adding 100 words to the reference vocabulary. The results for the 50k vocabulary in Table 2 show a similar trend to the 25k. It should be noted that the 25k vocabulary using vocabulary expansion yielded better speech recognition performance than the 50k reference vocabulary. Statistical hypothesis testing showed that the 0.5% improvement in word error rate for the 25k vocabulary from 27.50% (REF) to 27.00% (SCV,1000) was highly significant ($p < 0.01$) and the 0.33% improvement for the 50k vocabulary from 27.32% (REF) to 26.99% (SCV,1000) was significant ($p < 0.05$).

Table 1. Experimental results (25k)

	#add	#oov	%oov	%red.	%wer
REF	-	1471	2.10	-	27.50
OKAPI	100	1440	2.06	2.1	27.71
(conventional)	1000	1159	1.66	21.2	27.38
SCV	100	1206	1.72	18.0	27.17
(proposed)	1000	1002	1.43	31.9	27.00

Table 2. Experimental results (50k)

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

Table 3 shows the number of OOV word reductions (#oov red.) and word error reductions (#word error red.). The efficiency (%eff.) is the ratio of the reduction in word errors to the reduction in OOV words. It represents the contribution of the additional words to word error reduction. More than a 100% efficiency for a 50k vocabulary means that more words were correctly recognized than the obtained OOV words.

Table 3. Reduction of OOV and word errors

	#add	#oov red.	#word error red.	%eff.
25k	100	265	228	86.0
	1000	469	349	74.4
50k	100	126	164	130.2
	1000	206	224	108.7

Tables 4 and 5 show the experimental results for each range of OOV rates for the reference vocabularies. The 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 for news stories with less than a 1% OOV rate for the 25k vocabulary. We checked data with less than a 1% OOV rate and found that the word error rate increased when a result from the reference vocabulary had relatively high word error rate because of noise, speaker, or speaking style. Table 6 shows word error rates without vocabulary expansion if the story has a confidence score below the thresholds. The confidence score is calculated based on the frame-by-frame differences in the acoustic scores between the hypotheses and competing candidates. The thresholds were varied from -1σ to -3σ of the confidence score. The word error rate was improved with a threshold of -2σ and -2.5σ over the vocabulary expansion at any confidence score ($-\infty$).

Table 4. Classification by OOV rate (25k)

%oov range	#story	%oov %wer		
		REF	100	1000
- < 1%	89	0.53 20.0	0.44 20.2	0.35 20.2
1% ≤ - < 2%	57	1.39 28.7	1.21 28.6	1.00 28.5
2% ≤ -	119	3.93 32.8	3.16 31.9	2.65 31.5

Table 5. Classification by OOV rate (50k)

%oov range	#story	%oov %wer		
		REF	100	1000
- < 1%	142	0.48 23.0	0.29 22.9	0.26 22.9
1% ≤ - < 2%	53	1.41 33.0	0.76 32.9	0.64 32.9
2% ≤ -	70	3.73 32.6	2.41 32.0	2.06 31.6

Table 6. Vocabulary expansion with confidence score (25k)

threshold	%wer			
	100		1000	
	%oov < 1%	all	%oov < 1%	all
$-\infty$	20.15	27.17	20.24	27.00
-1σ	20.15	27.16	20.25	27.06
-1.5σ	20.11	27.15	20.24	27.02
-2σ	20.11	27.15	20.22	27.00
-2.5σ	20.11	27.16	20.21	27.00
-3σ	20.15	27.17	20.25	27.00

6. CONCLUSIONS

This paper described an unsupervised vocabulary expansion based on a concept base for reducing OOV words and experimental results for broadcast news speech. Words relevant to the input speech were extracted from the vocabulary database, based on the relevance score calculated using word concept vectors. The extracted words were added to the reference vocabulary to build an expanded vocabulary that was used for the second recognition process. The experimental results show that the proposed method can reduce the number of OOV words effectively, and the obtained words contribute to reducing the word error rate for data with any OOV rate. Although the word error rate increased for some of the data, the confidence score could limit the increase. Our future works include vocabulary optimization by combining irrelevant word elimination with the vocabulary expansion, and n-gram probability adaptation.

7. REFERENCES

[1] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proc. IEEE*, vol. 88, no. 8, pp. 1338–1353, 2000.

[2] K. Ohtsuki, K. Bessho, Y. Matsuo, S. Matsunaga, and Y. Hayashi, "Automatic indexing of multimedia content by integration of audio, spoken language, and visual information," in *Proc. ASRU*, 2003, pp. 601–606.

[3] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin abd I. Rogina, R. Stiefelhagen, and J. Yang, "SMaRT: The smart meeting room task at ISL," in *Proc. ICASSP*, 2003, pp. 752–755.

[4] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc. HLT*, 2001, pp. 246–252.

[5] I. Bazzi and J. Glass, "A multi-class approach for modeling out-of-vocabulary words," in *Proc. ICSLP*, 2002, pp. 1613–1616.

[6] K. Tanigaki, H. Yamamoto, and Y. Sagisaka, "A hierarchical language model incorporating class-dependent word models for OOV words recognition," in *Proc. ICSLP*, 2000, vol. 3, pp. 123–126.

[7] T. Kemp and A. Waibel, "Reducing the OOV rate in broadcast news speech recognition," in *Proc. ICSLP*, 1998, pp. 1839–1842.

[8] H. Yu, T. Tomokiyo, Z. Wang, and A. Waibel, "New developments in automatic meeting transcription," in *Proc. ICSLP*, 2000, vol. 4, pp. 310–313.

[9] M. Mahajan, D. Beeferman, and X.D. Huang, "Improved topic-dependent language modeling using information retrieval techniques," in *Proc. ICASSP*, 1999, vol. 1, pp. 541–544.

[10] L. Chen, JL Gauvain, L. Lamel, and G. Adda, "Unsupervised language model adaptation for broadcast news," in *Proc. ICASSP*, 2003, vol. 1, pp. 220–223.

[11] P. Geutner, M. Finke, and P. Scheytt, "Adaptive vocabularies for transcribing multilingual broadcast news," in *Proc. ICASSP*, 1998, pp. 925–928.

[12] T. Kato, S. Shimada, M. Kumamoto, and K. Matsuzawa, "Idea-deriving information retrieval system," in *Proc. of 1st NTCIR Workshop*, 1999, pp. 187–193.

[13] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-based language models: A maximum entropy approach," in *Proc. ICASSP*, 1993, vol. 2, pp. 45–48.

[14] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer Speech and Language*, vol. 10, pp. 187–228, 1996.