# 한국어 방송 뉴스 인식 시스템을 위한 OOV update module

정의정, 윤 승

한국전자통신연구원 음성언어팀

# Korean broadcast news transcription system with out-of-vocabulary(OOV) update module

Eui_Jung, Jung , Yun Seung

Spoken Language Processing Team, ETRI

{euijoung, yunseung}@etri.re.kr

## Abstract

We implemented a robust Korean broadcast news transcription system for out-of-vocabulary (OOV), tested its performance. The occurrence of OOV words in the input speech is inevitable in large vocabulary continuous speech recognition (LVCSR). The known vocabulary will never be complete due to the existence of for instance neologisms, proper names, and compounds in some languages. The fixed vocabulary and language model of LVCSR system directly face with these OOV words. Therefore our Broadcast news recognition system has an offline OOV update module of language model and vocabulary to solve OOV problem and selects morpheme-based recognition unit (so called, pseudo-morpheme) for OOV robustness.

## 1. Introduction

In LVCSR, the occurrence of OOV words in the input speech is inevitable and the known vocabulary will never be complete. Broadcast news deals with many topics and has the property that unseen words appears continuously including many proper nouns like the names of persons, areas, organizations, etc.. The fixed vocabulary and language model directly face with these OOV words. Therefore our Broadcast news recognition system has an offline update module of language model and vocabulary to solve OOV problem. First, it gathers recent newspaper articles or broadcast news transcriptions from Internet. Next, it extracts and registers OOV words into the vocabulary. To prevent from the continuous increase of dictionary size, words that was newly registered but not used in the latest one month are removed. Language model is also updated by taking a copy of UNK's or the representative word's language model probability. This module does this operation everyday. Also we selects morpheme-based recognition unit (so called, pseudo-morpheme) for OOV robustness. For Korean speech recognition, we can consider eojeol (word phrase), morpheme, or syllable as recognition units. An eojeol is a spacing unit in text. Therefore it might be a reasonable approach for small vocabulary continuous speech recognition systems. if we use eojeol as a dictionary and language modeling unit for LVCSR, the number of distinct entries and the number of OOV words overflow as the running text becomes large [1]. This is because a particle can be attached to a noun and a verb is heavily inflected depending on its syntactic role in Korean as in Japanese. Therefore, we should devise a smaller unit than eojeol (e.g., morpheme and syllable) for LVCSR. If we select syllable as the base unit for dictionary and language modeling, we have a small number of distinct words in dictionary but should do much work to recover eojeol sequence. In case when we select morpheme, we should reflect pronunciation variation between two morphemes in the same eojeol according to the pronunciation rule. Therefore, we selected pseudo-morpheme unit as the vocabulary and language modeling unit to cope with high OOV rate in Korean LVCSR, whose pronunciation is unchanged with adjacent morphemes and hence the original word can be obtained by concatenating morphemes. Then we concatenate a few short or frequent morphemes according to some rules and then define it as a new morpheme [2]. More discussion on unit selection for Korean speech recognition is found in [3-5].

This paper is organized as follows. In section 2 we describe pseudo-morpheme recognition unit. In section 3 we discuss the Broadcast news recognition system with OOV update module. In Section 4 we present experimental results. Finally conclusions are drawn in Section 5.

## 2. Pseudo-morpheme recognition unit

We have modified a Korean morphology analyzer [7] to produce a sequence of pseudo-morphemes from a sentence. The pseudo-morpheme is characterized as its pronunciation is maintained after decomposition. Usually a pseudo-morpheme is equal to a morpheme. Two or more consecutive endings and/or particles are merged into a new pseudo-morpheme unit. We also merged a suffix and the following particle into a new pseudo-morpheme. This is because suffixes are usually very short and thus are error-prone. Auxiliary verbs contribute to a large portion of recognition errors because they occur frequently in text and have short syllable length. We do not split eojeols of inflected auxiliary verbs because their frequency was very high. In addition to segmenting word-endings and particles, the analyzer decomposes compound nouns. The resulting pseudo-morpheme sequence obtained by segmenting an eojeol can be concatenated back to produce the original eojeol. We converted the eojeol-based text corpus into pseudo-morpheme based corpus. An eojeol was converted into average 1.8 pseudo-morphemes.

## 3. Offline OOV update module of LM and Vocabulary

Korean broadcast news transcription system has an offline update module of LM and dictionary to solve OOV problem. Reducing the number of OOV is one of the problems for LVCSR to solve. Broadcast news deals with many topics and has the property that unseen words appears continuously including many proper nouns like the names of persons, areas, organizations, etc.. The fixed vocabulary and LM directly faces with these OOV words. The offline update module gives the flexibility to vocabulary dictionary and LM.

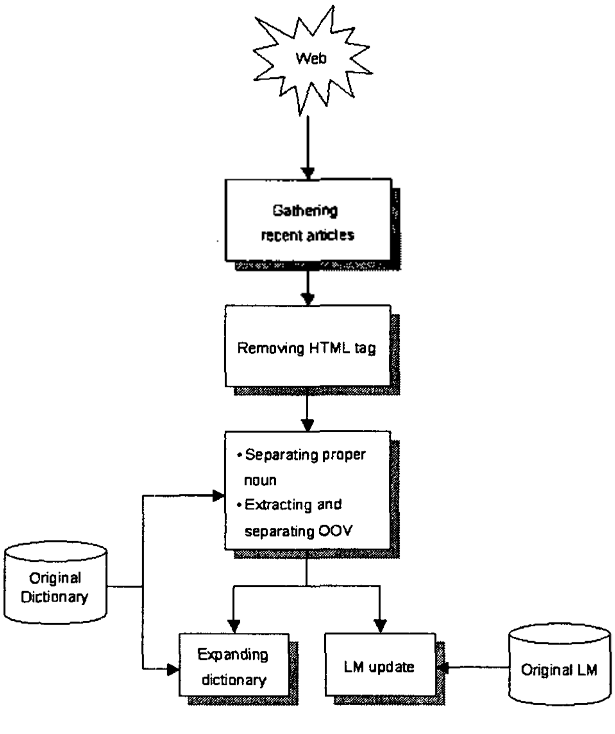

Figure 1. The flow of the offline updating LM and dictionary

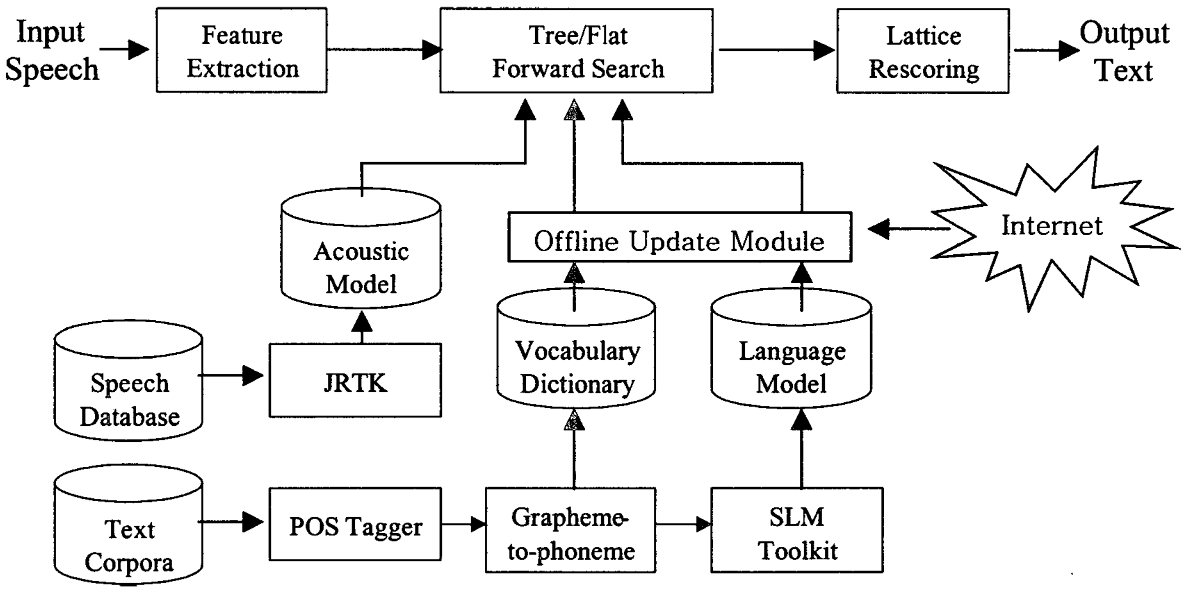

Figure 2. Experimental Setup of Broadcast News Recognition System

Figure 1 shows the flow of the offline updating. First, it gathers recent newspaper articles or broadcast news transcriptions from Internet [6]. Next, it extracts and registers OOV words into the vocabulary dictionary. To prevent from the continuous increase of vocabulary size, words that was newly registered but not used in the latest one month are removed. Language model is also updated by taking a copy of the representative word's language model probability. This module does this operation everyday.

## 4. Experimental Result

### 4.1 Experimental Setup

As shown in Figure 2, we use the JRTk recognition toolkit to train acoustic model and SLM toolkit[9] to get trigram language model. We use a part-of-speech (POS) tagger[7] to segment eojeols into morphemes. Pronunciation dictionary is automatically obtained by using a morpheme-based grapheme-to-phoneme converter [8]. Speech signals are sampled at 16kHz to produce 16 bit data. The window size is 16ms and the frame shift is 10ms. We also use melcepstrum and its differential coefficients as feature. Then a linear discriminant analysis (LDA) is performed to produce 24 dimensional features. We use 40 basic phonemes. For each phoneme, we use a 3-state hidden Markov model without skip transition. For observation probability, we use senone-based acoustic modeling with inter-morpheme coarticulation considered. The maximum context width is 2 for both left and right directions for intraword contexts and 1 for interword contexts. We use 3,000 senones and each senone has its own codebook with 16 Gaussian mixtures. For context clustering we use 47 detailed phoneme categories (e.g., vowel, consonants, fricative, and so on) as context questions in the decision tree. The fixed vocabulary size is 64,014 including human and nonhuman noise. we use the backoff smoothing method to estimate probabilities with small data[10]. The number of the extracted OOV words from two day test-set(1999-12-17, 2000-1-17) is 256 and each OOV word has the category tag. It makes each OOV words map the representative words included in the fixed vocabulary and language model. Each OOV word is registered into the vocabulary, language model is also updated by taking a copy of the representative word's language model probability. Broadcast news deals with many topics and the kinds of unseen words are numerous. Therefore we adapt " UNK" , " NAME" category tag, which are mapped the representative word in the vocabulary and language model.

### 4.2 Result

We evaluate the performance of OOV update module using two cases language models, TD1 and TD2, which TD1 is based on 6 millions size text corpora and TD2 is 14 millions

Table 1. Experimental Result

| Test set | LM | OOV update | | ERR |
|---|---|---|---|---|
| | | Do Not | Do | |
| 1999-12-17 | TD1 | 81.0% | 81.6% | 3.16 |
| | TD2 | 84.0% | 84.5% | 3.13 |
| 2000-01-17 | TD1 | 76.2% | 77.1% | 3.78 |
| | TD2 | 79.6% | 79.8% | 0.98 |

In this experiment, we can find that OOV update module contributes the performance of the recognition system. Now, we considered the 4 parts, that is, proper noun, non-predicative common noun, active-predicative common noun, stative-predicative common noun by frequency information. But the kinds of the OOV' s part are very manifold and these must are considered later.

## 5. Conclusion

In this paper, we implemented the offline OOV update module in Korean broadcast news transcription system to resolve the inevitable OOV problem in large vocabulary continuous speech recognition (LVCSR) and selected morpheme-based recognition unit (so called, pseudo-morpheme) for OOV robustness. It gathers recent newspaper articles or broadcast news transcriptions from Internet and extracts and registers OOV words into the vocabulary. Language model is also updated by taking a copy of UNK's or the representative word's language model probability. We can find that OOV update module contributes the performance of the recognition system by recognition experiments

ACKNOWLEDGEMENTS

This research was funded by the Korea Ministry of Information and Communication.

## REFERENCE

[1] L. Tomokiyo and K. Ries, " What makes a word: Learning base units in Japanese for speech recognition," Proceedings of the Workshop on Natural Language Learning, 1997.

[2] O.W. Kwon, K.Hwang and J. Park, " Korean Large Vocabulary Continuous Speech Recognition of Newspaper Articles," Proc. ICSP99, pp.333-336, 1999.

[3] O.W. Kwon, K. Hwang, and J. Park, " Korean large vocabulary continuous speech recognition using pseudomorpheme units," EUROSPEECH ' 99, Budapest, Hungary, Sept. 1999.

[4] D. Kiecza, T. Schultz, and A. Waibel, " data-Driven determination of appropriate dictionary units for Korean LVCSR," ICSP ' 99, pp. 323-327, Aug. 1999.

[5] O.W. kwon, " Performance of LVCSR with morpheme-based and syllable-based recognition units," ICASSP 2000, pp. 1567-1570, June 2000.

[6] Yun Seung, " Unknown Word Extractor Development for ETRI Broadcast News Caption System," Acoustic Society of Korea 2002, ChangWon, Korea, July. 2002

[7] J.-H. Kim, Lexical disambiguation with error-Driven Learning, Ph.D. dissert. Dept. Computer Science, Korea Advanced Institue of Science and Technology, 1996.

[8] J. Jeon, S. Cha, M. Chung, J. Park, and K. Hwang, " Automatic generation of Korean pronunciation variants by multistage applications of phonological rules," ICSLP ' 98, Sydney, Austrailia, Dec. 1998.

[9] P. Clarkson and R. Rosenfeld, " Statistical language modeling using the CMU-Cambridge toolkit," EUROSPEECH ' 97, pp. 2707-2710, 1997.

[10] S. M. Kalt, " Estimation of probabilities from sparse data for the language model component of a speech recognizer," IEEE Trans. ASSP, Vol. 35, pp. 400-401, 1987.