

Morpheme-based Lexical Modeling for Korean Broadcast News Transcription

Young-Hee Park, Dong-Hoon Ahn, Minhwa Chung.

Department of Computer Science, Sogang University
Sinsu-Dong, Mapo-Ku, Seoul 121-742, Korea
{younghp, drahn, mchung}@sogang.ac.kr

Abstract

In this paper, we describe our LVCSR system for Korean broadcast news transcription. The main focus here is to find the most proper morpheme-based lexical model for Korean broadcast news recognition to deal with the inflectional flexibilities in Korean. Since there are trade-offs between lexicon size and lexical coverage, and between the length of lexical unit and WER, in our system we analyzed the training corpus to obtain a compact 24k-morpheme-based lexicon with 98.8% coverage. Then, the lexicon is optimized by combining morphemes using statistics of training corpus under monosyllable constraint or maximum length constraint. In experiments, our system reduced the number of monosyllable morphemes which are the most error-prone, from 52% to 29% of the lexicon and obtained 13.24% WER for anchor and 24.97% for reporter.

1. Introduction

It is well known that radio and television broadcast news contain many different topics, various types of speech, background noises and music. These varieties of sources increase the difficulty of automatic broadcast news transcription [1]. In addition to these, especially for inflectional languages such as Korean, how to define a good recognition unit affects the difficulty significantly. Although morphemes are generally considered as basic recognition units, strict Korean morphemes used for text processing are not suitable for Korean LVCSR tasks for the following reasons. Firstly, the concatenated pronunciation of morphemes in the morpheme sequence obtained by a morphological analysis of an *eojeol*¹ often does not match the pronunciation of the *eojeol* itself. Secondly, a sentence can be morphologically analyzed in more than one way. For example, a compound noun can be one morpheme or a sequence of noun morphemes. This morphological flexibility and mismatch in pronunciations have been major obstacles to advances in Korean LVCSR. How to define recognition units from morphemes has significant effect on lexical and language models. Different definitions give different lexical entries. This in turn results in a variety of OOV rates, lexicon sizes, and language models which would influence the performance of LVCSR systems. The definition of recognition units also affects WER, since most Korean morphemes consists of 1 or 2 syllables² (see Figure 1), which indicates short durations of most morpheme-based recognition units and hence increases possibility of misrecognition.

In the two literatures for Korean LVCSR [2][3], recognition units are defined in terms of morphemes modified so that the concatenated pronunciation of morphemes can match the pronunciation of the *eojeol*. However, their unit selections are not fully optimized. In [2], they selected the morpheme itself as the recognition unit and used a cross-morpheme phone variation lexicon to match the pronunciation and reduce the OOV rate. In [3], they used the likelihood increase measure to merge some “units” into longer ones, but in fact those basic “units” are ones merged from simple morphemes by applying heuristic rules. This approach, however, has the following problems. (1) The units are composed in a rather complicated way since optimal merging rules are not easily obtained across domains. (2) The units lose orthogonal properties of original morphemes since rule applications force most of stems and endings, and a series of endings in verbal inflections to be merged into one lexical entry, where a lexicon has many inflected forms of a verb. (3) Since the (suboptimal) rule-based merging is done prior to statistical merging, the final results cannot be made to be fully optimal.

In our approach, we start with simple morphemes with modifications for compatible pronunciations to reduce the OOV rate effectively; but their short durations, especially those of monosyllable morphemes cause frequent misrecognitions [4]. So we merge those short morphemes into new units which we call *concatenate morphemes*. To merge them optimally, we have tested and compared various merging criteria based on morphological knowledge, frequencies, mutual information and unigram log-likelihood. From these comparisons, we have identified 1,000 additional concatenate morphemes which reduced the number of monosyllable morphemes from 52% to 29% of the lexicon and obtained 13.24% WER for anchor and 24.97% WER for reporter.

This paper is organized as follows. Section 2 summarizes the broadcast news corpus used for our experiment. Section 3 and 4 describe our concatenate-morpheme-based lexical model. Section 5 describes our baseline transcription system, and Section 6 presents experimental results.

2. Korean Broadcast News Corpus

For Korean broadcast news transcription, we used KBS News 9 speech from February 1997 to December 1998, which included various kinds of speech with human noises, background noises, or music. Particularly, reporter’s speech is of very low quality due to noises. The whole speech database amounts to 6.5 hours for anchors and 19 hours for reporters. As a test set, we used speech of October 1998.

¹ *Eojeol* is a spacing unit in Korean like word in English. Typically, an *eojeol* consists of more than one morpheme.

² A Korean syllable has the form of V, CV or CVC. Thus it includes at most 3 phonemes.

Table 1. Examples of morphological analysis of Korean *eojeols* (“-” stands for syllable boundary and “+” for morpheme boundary)

Example I		Example II		Meaning
Romanized Korean	Morpheme Category	Romanized Korean	Morpheme Category	
bang-song-nju-seu	compound noun	bang-song + nju-seu	noun + noun	broadcast news
beom-gug-min-jeog	compound noun	beom + gug-min + jeog	prefix + noun + suffix	nation-wide
gga-zi-oa-neun	particle	gga-zi + oa + neun	particle + particle + particle	even up to
ha-si-eoss-seub-ni-da	auxiliary verb phrase	ha + si + eoss + seub-ni-da	root + ending + ending + ending	did (respectful word)

The text corpus consists of articles mainly from broadcast news and partly from newspapers. Broadcast news texts are collected from KBS News 9 between January 1997 and February 1999 and between July 2000 and May 2002 obtained from web sites. Newspaper texts are collected from Korean Information Base System II database [5] produced by KAIST, and additionally from Chosun, Donga and Hankyoreh daily newspapers of 1994 and 1997. All the raw texts were preprocessed via conversion of non-Korean alphabets such as numeric characters, dates and foreign words into Korean alphabets, and removal or conversion of extra symbols such as \$ (dollar) and ~ (from ... to ...), etc. Finally we performed a morphological analysis to obtain the text corpus of 16M morphemes.

3. Morpheme-based Lexical Model

As stated in Section 1, Korean sentences can be morphologically analyzed in more than one way. Table 1 shows two examples of morphological analysis of Korean *eojeols*. Example I shows that an *eojeol* can be analyzed as a single morpheme. In Example II, where an *eojeol* is split into as small units as possible, an *eojeol* is analyzed up to 4 morphemes. As far as pronunciation is concerned, morphemes in Example I have at least 8 phonemes, while the morpheme “ha” in Example II has only two phonemes. Obviously morphemes in Example I are more suitable units for recognition since their durations are reasonably long. However, such long recognition units increase the lexicon size and result in poor lexical coverage, especially in case of compound nouns where most individual nouns belonging to them are frequently used in other contexts. On the contrary, morphemes with shorter durations in Example II are not as good as morphemes in Example I for recognition but can give a smaller lexicon and quite a good lexical coverage.

In addition, we learned during experiments that over the half of recognition errors are from monosyllable morphemes. Therefore, we need to reduce the rate of monosyllable morphemes. The Baseline plot of Figure 1 shows the distribution of morpheme lengths obtained from the text

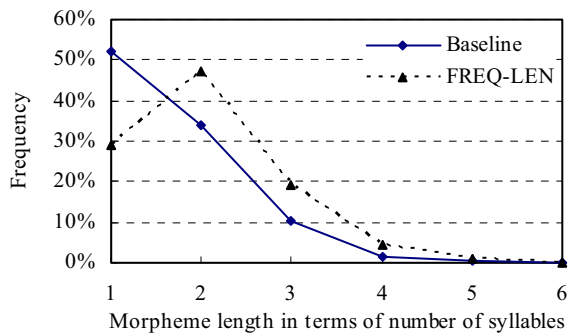


Figure 1. Distribution of morpheme lengths in morphemes (Baseline) and concatenate morphemes (FREQ-LEN).

corpus. We can see that 52% of total morphemes are monosyllables and most of morphemes have at most four syllables. Lexicons based on these morphemes would be very unlikely to be well recognized unless an appropriate language model is supported.

To obtain recognition units which are not only as *orthogonal* and *consistent* with original morphemes as possible, but also *cooperative* with a recognition system, we automatically selected some consecutive morphemes using several measures [3][4][6], concatenated them to make concatenate morphemes, and then added them to the lexicon. The dotted line in Figure 1 shows the situation after 1,000 concatenate morphemes are added using frequency measure. The rate of monosyllable morphemes is dropped from 52% to 29% and many of short morphemes are merged into longer ones, increasing the rate of morphemes with two or more syllables. To decide the size of a baseline lexicon, we plot the relation between lexicon size and lexical coverage in Figure 2. From this graph, we have selected a 24k-morpheme lexicon with lexical coverage of 98.8%. Concatenate morphemes are not yet included in this lexicon. We believe this is a high enough coverage for the lexicon and the result is due to morphologically consistent design. Accordingly, in a situation such as broadcast news where new words are to be repeatedly introduced and they are most likely to be nouns or compound nouns, their additions would not significantly increase the size of the lexicon.

4. Automatic Generation of Concatenate Morphemes

In order to generate concatenate morphemes automatically, we have tested and compared various methods that are broadly categorized into two classes: (1) knowledge-based merging which uses morphological (part-of-speech) information and (2) statistical measure-based merging which uses an optimization criterion.

4.1. Knowledge-based merging

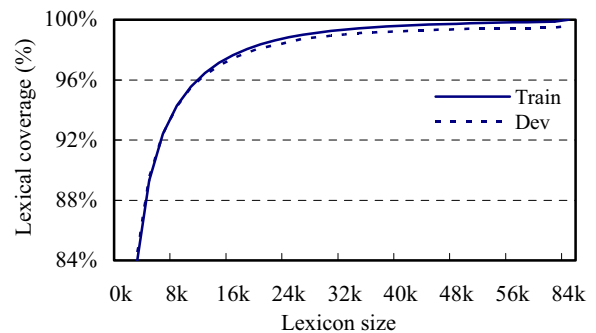


Figure 2. Lexicon size vs. lexical coverage.

Knowledge-based merging strategy is in fact our previous approach [7] and is similar to the rule-based merging step in [3]. In this strategy, merging rules are generated simply by using part-of-speech (POS) of consecutive morphemes and linguistic constraints on them. Most of rules are used to lengthen function words (morphemes), auxiliary verbs and adjectives which are typically of short length and high frequency. This approach, however, have showed several problems. Firstly, it needs an exactly POS tagged corpus, which may not be provided when processing large amount of text corpus. Secondly, “short length” and “high frequency” are heuristic and subjective measures, because they can be different across several domains. For example, this strategy performed well in our initial work [7] for a small read speech dictation task, but didn’t in the current broadcast news task. To summarize, knowledge-based merging has inflexibility and hard-to-find optimality, which is implied in the small amount of decrease in perplexity, as shown in Figure 3.

4.2. Statistical measure-based merging

To test the statistical measure-based merging strategy, we have applied the following criteria [3][4][6], where $N(v)$ is the count of v and N is the total count.

- Frequencies of morpheme pair (FREQ): $N(v, w)$
- Mutual information between morpheme pair (MI):

$$N(v, w) \log \frac{N(v, w)N}{N(v)N(w)}$$

- Changes in unigram log-likelihood (ULL):

$$\sum_{w \in V} \tilde{N}(w) \log \frac{\tilde{N}(w)}{\tilde{N}} - \sum_{w \in V} N(w) \log \frac{N(w)}{N},$$

where tildes denote statistics after morphemes are merged.

Any two consecutive morphemes from the corpus ranked in high positions according to each criterion are merged into new concatenate morphemes, which are again pooled into the text corpus so that they can be merged with other (concatenate) morphemes. We repeat this procedure until the number of new morphemes introduced reaches to the pre-conditioned count or the decrease in LM perplexity is below a threshold value. These criterion-derived concatenate morphemes are very cooperative with a recognition system in that when a lexicon includes these additional concatenate morphemes, bigram perplexities monotonically decrease as shown in Figure 3 and their word error rates (actually *morpheme* error rates) are all less than that of the morpheme-only (baseline) lexicon as shown in Table 2. Of these three criteria, the FREQ-derived lexicon shows both the lowest perplexity and the lowest word error rate while the ULL-derived one is the worst in both the perplexity and the performance.

4.3. Statistical measure-based merging with constraints

We also tested the use of morphological constraints to further control the merging step and to let new units as morphologically consistent with original morphemes as possible. These constraints are tested on the FREQ-derived lexicon.

The first constraint considered is a monosyllable constraint (FREQ-MS) which focuses on reducing small units since monosyllable morphemes are the most error-prone and takes the biggest portion of the entire corpus. Under this constraint, at least one candidate for concatenation should be

monosyllable. The second constraint considered is a maximum length constraint (FREQ-LEN). As more morphemes are concatenated, the length of the concatenate morpheme becomes longer. While longer units are good for recognition, they may be inflexible and inconsistent from the morphological aspect. Only when concatenated lengths do not exceed a specified threshold, their concatenations are permitted. From experiments, we found that three or four are good thresholds. These constraints improved both recognition rate and perplexities than those of unconstrained ones, respectively.

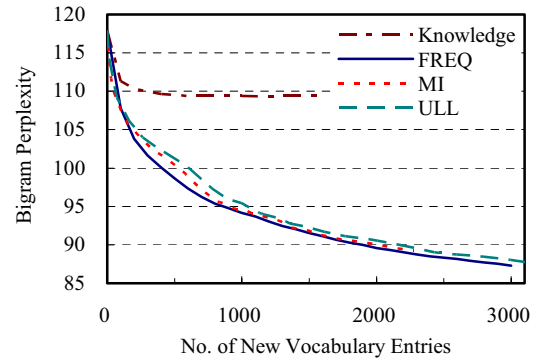


Figure 3. Bigram perplexity vs. the number of concatenate morphemes added to the lexicon.

5. Korean LVCSR System for Broadcast News Transcription

5.1. Acoustic and language models

The prototype system used in this experiment is based on the 1-pass semi-dynamic trigram network decoder, which was originally developed for Korean read speech dictation task[8]. Its frontend generates MFCC-based 39th order coefficients and its acoustic models are trained from the database described in Section 2, to get a set of continuous HMMs with each state of 8 Gaussian mixtures. The lexical entries are morphologically obtained from the corpus in Section 2 with some additions of concatenate morphemes described in Section 3 and 4. The language model is a trigram estimated with modified Kneser-Ney discounting method (cutoff 2 for trigram) and an entropy pruning. This trigram is fed into the semi-dynamic network decoder to build a pre-compiled compact static network.

5.2. 1-pass semi-dynamic trigram network decoder

Our decoder is a one-pass Viterbi decoder which runs on a static network. The static network is precompiled from a language model (LM) network representing LMs with a finite-state machine and then made compact by constructing *aligned shared tails*[8]. These aligned shared tails exploit the indistinguishable relations among linear tails which is essentially the same idea as the classical automata minimization algorithm. In the last implementation[8], however, LM entries or successor trees that are cut off by pruning has formed empty trees with only one root node, just making *epsilon* transitions to ones with backed off histories. This prohibited sharing linear tails among successor trees with histories of different sizes and consequently resulted in poor compaction ratios for higher order N-grams.

This limitation has been resolved in developing the prototype system by applying the epsilon removal algorithm [9] when a node has only back-off arcs. For example, Figure 4(a) shows the original network structure where the word *W* is a rare event and thus LM entries with history including *W* are cut off to give their trees just one root nodes for back-off transitions. In this case, linear tails from *W* in the unigram tree cannot be shared with those from *W* in trees with one- or two-word histories because they do not reach the same root. However, if the back-off arcs are removed and their weights are accumulated to their preceding arcs' weight then we can get more back-off-free (epsilon-free) network structure in Figure 4(b). Now linear tails from any trees including history *W* are gathered into shared tails. With this improvement, we can get more compact networks, whose sizes are reduced to 30~40% of the original networks in case of trigram. For comparison, the last implementation gave at best about 60% of compaction [8].

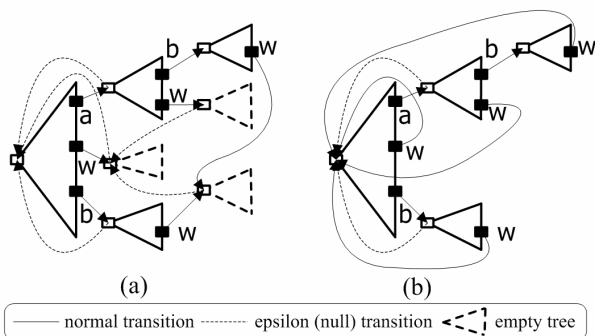


Figure 4. Improved network structure.

6. Experimental Results

We evaluated our system using speech from October 1998. The test set consists of 181 utterances from 4 anchors and 180 utterances from 45 reporters without OOVs. The average numbers of words of utterances for anchors and reporters are 21 and 23, respectively, and the total numbers of morphemes uttered are 3,594 and 4,290, respectively.

Table 2 shows the perplexities normalized by the original number of words and WER in case of adding 1,000 concatenate morphemes. In our system, the best results are absolute reductions of WER 2.68% (FREQ-MS) for anchor and 2.21% (FREQ-LEN) for reporter, showing that they are both superior to the morpheme-only lexicon (Baseline).

Table 3 presents the breakdown of errors resulting from morpheme-only and FREQ-MS derived lexicons. This shows that our approach effectively reduced the number of errors in monosyllable morphemes which dominate errors originally; the monosyllable error reduction accounts for 72.7% of all errors reductions.

7. Conclusions

We have described a morpheme-based lexical modeling for our LVCSR system for Korean broadcast news transcription. Starting from a small 24k morpheme-based lexicon, it is optimized by combining morphemes using statistics of training corpus under monosyllable constraint or maximum length constraint. In experiments, our system reduces the number of

Table 2. Perplexity vs. WER

Measure	Perplexity	WER for anchor (%)	WER for reporter (%)
Baseline	68.2	15.92	27.18
MI	64.0	13.93	26.06
FREQ-MS	63.6	13.24	25.73
FREQ-LEN	63.7	14.22	24.97

Table 3. Breakdown of errors according to the morpheme length in terms of the number of syllables

Morpheme length	No. of errors Baseline	No. of errors FREQ-MS	No. of error reductions
1	813	693	120
≥ 2	639	594	45
total	1,452	1,287	165

monosyllable morphemes from 52% to 29% of the lexicon and produces absolute reductions of WER 2.68% (FREQ-MS) for anchor and 2.21% (FREQ-LEN) for reporter.

Acknowledgement

This work has been supported by Ministry of Science & Technology's Brain Neuroinformatics Research Program (Project No. M1-0107-01-0003). We would like thank ETRI for allowing us to use their broadcast news database.

References

- [1] Langzhou Chen, L. Lamel, G. Adda and J.L. Gauvain, "Broadcast news transcription in Mandarin," *Proc. of ICSLP*, 2000.
- [2] H.-J. Yu, H. Kim, J.S. Choi, J.M. Hong, K.S. Park, J.S. Lee, H.Y. Lee, "Automatic recognition of Korean broadcast news speech," *Proc. of ICSLP*, 1998.
- [3] Oh-Wook Kwon and Jun Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," *Speech Communication*, v.39, n.3-4, 2003.
- [4] George Saon and Mukund Padmanabhan, "Data-driven approach to designing compound words for continuous speech recognition," *IEEE Trans. on ASSP*, v.9, n.4, 2001.
- [5] Korean Information Base System II CD-ROM, KAIST, <http://kibs.kaist.ac.kr>.
- [6] Dietrich Klakow, "Language-model optimization by mapping of corpora," *Proc. of ICASSP*, 1998.
- [7] Kyong-Nim Lee and Minhwa Chung, "Pseudo-Morpheme-Based Continuous Speech Recognition," *Proc. of Workshop on Speech Communication and Signal processing*, 1998. (in Korean)
- [8] Dong-Hoon Ahn and Minhwa Chung, "Compact Subnetwork-based Large Vocabulary Continuous Speech Recognition," *Proc. of ICSLP*, 2002.
- [9] Mehryar Mohri, "Generic Epsilon-Removal Algorithm for Weighted Automata," In Sheng Yu and Andrei Paun, ed., 5th International Conference, CIAA 2000, v.2088 of Lecture Notes in Computer Science, Springer-Verlag, 2001.