

A Transformation-Based Learning Method on Generating Korean Standard Pronunciation*

Kim Dong-Sung^a and Chang-Hwa Roh^a

^a Department of Linguistics and Cognitive Science
Hankuk University of Foreign Studies
San89 Wansanri Mohyunmeon Yonginsi, Kyunggido Korea
{dsk202, rayr}@hufs.ac.kr

Abstract. In this paper, we propose a Transformation-Based Learning (TBL) method on generating the Korean standard pronunciation. Previous studies on the phonological processing have been focused on the phonological rule applications and the finite state automata (Johnson 1984; Kaplan and Kay 1994; Koskenniemi 1983; Bird 1995). In case of Korean computational phonology, some former researches have approached the phonological rule based pronunciation generation system (Lee et al. 2005; Lee 1998). This study suggests a corpus-based and data-oriented rule learning method on generating Korean standard pronunciation. In order to substituting rule-based generation with corpus-based one, an aligned corpus between an input and its pronunciation counterpart has been devised. We conducted an experiment on generating the standard pronunciation with the TBL algorithm, based on this aligned corpus.

Keywords: Transformation-Based Learning, Computational Phonology, Data-oriented Processing, Corpus-based Learning, Pronunciation Generation

1. Introduction

This paper presents a Transformation-Based Learning (TBL) method on generating Korean standard pronunciation. Previous studies on the phonological processing have been focused on the computation of the phonological rule application and the representation of the finite state automata (Johnson 1984; Kaplan and Kay 1994; Koskenniemi 1983; Bird 1995). In case of Korean computational phonology, some former researches have approached the pronunciation generation based on the phonological rules (Lee et al. 2005; Lee 1998)¹. Unlike previous works, this study suggests a standard Korean pronunciation generation method on the basis of corpus-based and data-oriented TBL learning.

The role of the computational phonology is to generate a legitimate output counterpart of the underlying phonological input. Phonological rules are involved in the process of phonological generation. The SPE style operations on the computational phonology have used the rewriting rule ordering or the finite state transducer (Bird 1995; Bird and Ellison 1994; Gildea and Jurafsky 1996; Kaplan and Kay 1994). Those approaches, however, should reduce complicated

* This paper was supported by the Second Brain Korea 21.

Copyright 2007 by Kim Dong-Sung and Chang-Hwa Roh

¹ Anyone can visit the website of Lee et al. (2005) and generate standard pronunciation at <http://urimal.cs.pusan.ac.kr>.

orderings because of huge amount of rewriting rules and rule orderings among themselves (Gildea and Jurafsky 1996). Other differently motivated approaches have suggested the data-oriented models, using a pronunciation corpus to derive legitimate outputs (Daelemans, Gillis and Durieux 1994; Johnson 1984).

In this study, we use the learning method of TBL that was proposed by Brill (1995). We design a set of templates and abstract transformations of possible pronunciations. For the experiments, we set up an aligned corpus between the text based on the Korean standard orthography and the text based on the Korean standard pronunciation. We conducted an experiment on generating the standard pronunciation with the TBL algorithm, using this corpus.

We use the phonotactic constraints to reduce the complexity of TBL process. As noticed in Hayes and Wilson (forthcoming), the phonological feature constraints can reduce the complication of phonotactics. We set up a list of constraints on the phonotactics, which is derived from the phonological features.

The rest of the paper is composed of three parts: Section 2 is to introduce the TBL method into the phonological operation. Section 3 describes the experiment on Korean pronunciation. Section 4 deals with the experiment discussions.

2. TBL Application on the Pronunciation Handling

Rule-oriented processing in phonology has been represented with context-sensitive rewrite rules. For example, Korean underlying stops are realized as unreleased voiceless stops in the word final position. The following example shows the rule application on the voiceless stop /t/.

$$(1) t \rightarrow t^{\neg} / _ \# ^2$$

The most popular way of formalizing the phonological rule is to induce two-level formalism in Koskeniemi (1984) and Karttunen (1993), or finite state transducer of Kaplan and Kay (1994). The basic intuition on these operations is that a rule rewrites an underlying string as a surface string, which can be implemented as a transducer that reads a lexical input and writes to a surface tape. [Figure 1] shows an example of this operation using the rule in (1).

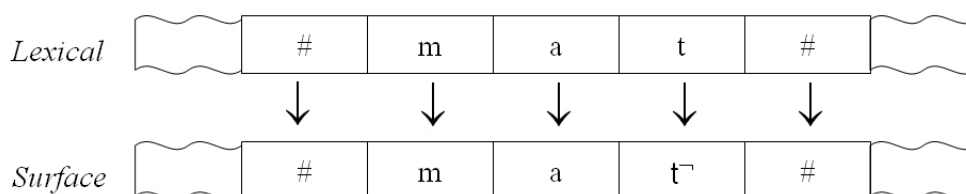


Figure 1: Rule based operation on the phonology

Phonological derivation-based method ought to have the complicated rule ordering systems. A phonological input has the chance of different output realization(s), depending on rule orderings. Computation based on the finite state transducer is so complicated that the processing mechanisms are varied among the researchers. Gildea and Jurafsky (1996) suggest a method to reduce complicated rule ordering.

Another different method is to use the data-oriented approach. Daelemans, Gillis and Durieux (1994) suggest a stochastic method to assign stress, the supra-segmental feature. This approach utilizes stochastic gain of information from a corpus.

TBL is known as learning the most approximate tagging rules from the corpus. TBL is a data-oriented method. It considers every possible transformations of the tagging, using a limited set

² For Korean sound and feature system, see Figure 4 and Figure 7 in the next section.

of transformations. The algorithm of TBL needs a small set of templates, abstracted transformations. A phonological input can be transformed into a phonological output. In Korean, the voiceless stop /t/ is varied among [t], [d], and [t̚], depending environments. Consider the following templates that transform the phonological input.

If the preceding phonological environment is #, then /t/ becomes [t].
 If the preceding phonological environment is Vowel, then /t/ becomes [d].
 If the following phonological environment is Consonant, then /t/ becomes [t̚].
 If the following phonological environment is #, then /t/ becomes [t̚].

Figure 2: TBL application on the phonological change

TBL method learns the phonological environment, by instantiating the incoming items in the templates. Every possible phonological environment in the template is iteratively tested by filling in every specific phonological input. This method transforms an input into an output, following the list in the template. In some sense, this approach is similar with one in two-level formalism, matching an input and an output. However, TBL needs a learning text (corpus). As Brill (1995) notes, a small amount of training data can resolve a large amount of processing data.

Templates in TBL method have the list of environment which the phonological change must follow. The environment is conceptually the same as a context window in the KeyWord In Context (KWIC). In Figure 3, an example of context window is given.

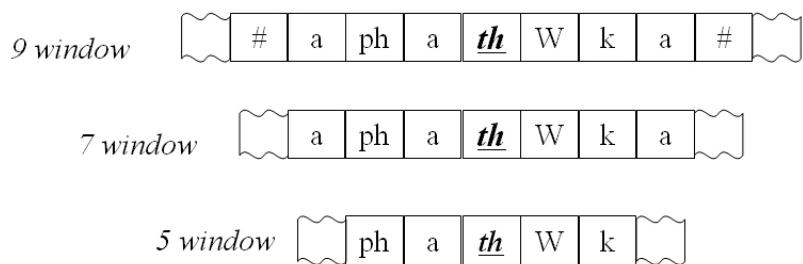


Figure 3: Context Windows in TBL

The phonological features are inter-related with the phonotactic constraints. As Hayes and Wilson (forthcoming) insist, phonological features reduce the phonotactic constraints. Following such idea, we set up constraints on phonotactics, combining the phonological feature systems. This simplifies the search mechanism of TBL processing.

3. Experiments

For the experiment, we set up the corpus which aligns the spoken data from the Sejong corpus and its standard pronunciation. The spoken data has 14,500 *ejeols*³ (approximately 60,000 morphemes), which is composed of the transcription in the Korean standard orthography. We converted the data into the standard pronunciation, using Korean standard IPA converter of Lee et al. (2006). For instance, (2a) is converted into (2b) with Korean standard IPA converter.

³ *Ejeol* is the similar with *bunsetsu* in Japanese. *Ejeol* is the terminology for the chunks between spaces in a sentence. For more information, see Sohn (1999).

- (2) a. Na-nun cip-e ka-n-ta.
 I-Top house-Loc go-Asp-End⁴
 b. Na-nWn tsi-be ka-n-da⁵

Following this process, we gathered the aligned corpus as follows.

- (3) Na-nun {N.a-n.W.n} cip-e {ts.i-p.e} kan-ta {k.a.n-d.a}

In (3) the convention ‘-’ and ‘.’ split intra syllables and inner syllable structures (onset-rhyme-coda), respectively. The ejeol initial position is marked with ‘{’ and the ejeol final one with ‘}’. The statistics of the standard pronunciation corpus is that the total phonemes are 106,478 with 7 phonemes per an ejeol and 1.78 phonemes per a morpheme.

For the phonetic purpose, 19 consonants and 10 vowels are used for the Korean pronunciation as follows:

IPA	ARPAbet	IPA	ARPAbet
p	p	i	i
p ^h	ph	ɛ	E
p [*]	p [*]	ɯ	W
t	t	ʌ	A
t ^h	th	u	u
t [*]	t [*]	o	o
k	k	a	a
k ^h	kh	ɥ	Y
k [*]	k [*]	J	y
s	s	w	w
s [*]	s [*]		
tɕ	ts		
tɕ ^h	tsh		
tɕ [*]	ts [*]		
m	m		
n	n		
ŋ	ŋ		
l	l		
h	H		

Figure 4: ARPAbet for Korean pronunciation

Depending on the word positions (either ejeol initial or ejeol final) and syllable positions (onset-rhyme-coda), we gather 600 different phonemic types for the context window. These types are used to induce the template of TBL. The following is an example of TBL template with the very first preceding and the very next following.

If the preceding phonological environment is #, then /t/ becomes [t].
 If the preceding phonological environment is V, the /t/ becomes [d].
 If the following phonological environment is #, the /t/ becomes [t̚].

Figure 5: TBL templates

⁴ Top: Topical marker, Loc: Locative marker, Asp: Aspectual, End: Ending

⁵ For the font and other convenient issue, we adopt convention of ARPAbet phoneset transcription. IPA fonts have the trouble when one can deal with the text processing.

As noted in Figure 3, the phonological environment is similar with a context window. If we enlarge the context window by 4, Figure 5 is changed into Figure 6.

If the phonological environment with 4 context window is {#, #, _, Vowel_{Rhyme}}, then /t/ becomes /t/.

If the phonological environment with 4 context window is {Vowel_{Rhyme}, -, _, Vowel_{Rhyme}}, then /t/ becomes /d/.

.....

Figure 6: Example of 4 context windows TBL

We randomly gathered 1,000 ejeols from the Sejong corpus for the test purpose. Using the aligned corpus, we converted the test material into its pronunciation. We have tested 20, 10, 5, 4, 3, or 2 context windows to see if there is any difference in accuracy. Brill (1995) suggests that TBL also reduce the training size of tagging. We also checked the total training, increasing training size by 1,000 until we reached the total size of the aligned corpora.

Hayes and Wilson (forthcoming) claim that English phonotactics is explainable with 24 different constraints based on phonological features. Phonotactic constraints can reduce the search space of the TBL templates. Because the phonotactic constraint stops to search ill-formed phonotactics in the templates, the wrongly-predicted pronunciation is eradicated.

In Korean, 29 phonemes in Figure 4 have the constraints on the phonotactic placements. The moderate phonological feature set of Korean is as follows.

Table 1: Korean phonological feature system⁶

		p	p*	ph	t	t*	th	k	k*	kh	s	s*	ts	ts*	tsh	m	n	G	l	H	
Major Class Features		sonornat	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	
		consonant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
		syllable	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Consonant Features	Manner Features	continuent	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	+	
		delayed release	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	
		lateral				-	-	-				-	-	-	-	-	-	-	+		
	Place Features	coronal	-	-	-	+	+	+	-	-	-	+	+	+	+	+	-	+	-	+	-
		anterior	+	+	+	+	+	+	-	-	-	+	+	-	-	-	+	+	-	+	-
	Subsidiary Features	tense	-	+	+	-	+	+	-	+	+	-	+	-	+	+	-	-	-	-	-
aspiraion		-	-	+	-	-	+	-	-	+	-	-	-	-	+	-	-	-	-	+	
			i	E	W	A	u	o	a	y	w	Y									
Major Class Features		sonornat	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
		consonant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
		syllable	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
	Tongue Body	high	+	-	+	-	+	-	+	-	+	-	+	-	+	+	+	+	+		

⁶ The feature map is from Shin and Cha (2004).

Features	Features	low	-	-	-	-	-	-	+	-	-	-
		back	-	-	+	+	+	+	+	-	+	+
	Rounded	round	-	-	-	-	+	+	-	-	+	-

What the feature map in Figure 5 specifies is any clusters with consonant and diphthong [yE] cannot be placed next to each other. This cluster is predictable with the phonological feature of *[+cons][-back,-rnd,-syl][+syl]. The constraint restricts under the system in Figure 7. This restriction stops the searching mechanism of TBL since any restricted item is found. We build up the restriction list of 20 constraints.⁷

Generally, the morphological information is pre-requisite for the phonological handling. The phonological change depends on the morphological information; such as irregular verbs, grammatical functions, word classes, etc. Our assumption on the morphological issue is that larger context windows in TBL include more morphological information. We doubted that such information can be replaceable with the size of context window. If the context window is larger, such morphological information is possibly included. We considered two groups of experiments; one with morphological information and the other is without morphological information. We compared the accuracy rate of two groups.

4. Discussion

We use 20, 10, 5, 4, 3, or 2 context windows in the template to see the change in the precision. This test did not contain the morphological information, but only aligned corpus was used.

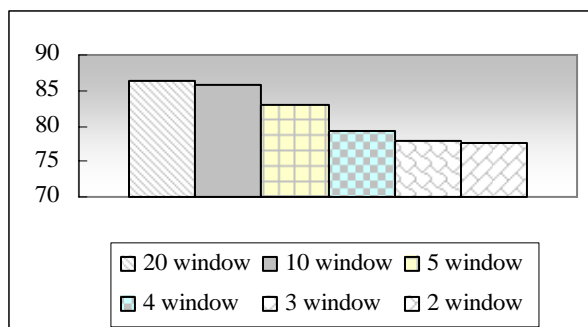


Figure 7: Difference in precision rate without phonological information

The result shows that the context window becomes larger and the precision rate goes up. Consider that an ejeol contains an average of 7 phonemes and each phoneme has an average of 1.78 morpheme(s). 10 and 20 context window contains more than 2 ejeols, which show the better precision rate. This reflects that the morphological information across an ejeol is reflected in the larger windows.

With the morphological information, the precision rate of the experiment is as follows.

⁷ We used the hand-written constraints. To handle with the phonological feature system, very different computational mechanism is required. Bird (1995) and Bird and Ellison (1994) present a way to compute such features in the logical way. The problem for handling the feature system is to cope with the very complexity of feature systems. Gildea and Jurafsky (1996) use a decision tree to simply handle the feature geometry in phonology, as a way of simplifying the feature systems. We kept it for the future study.

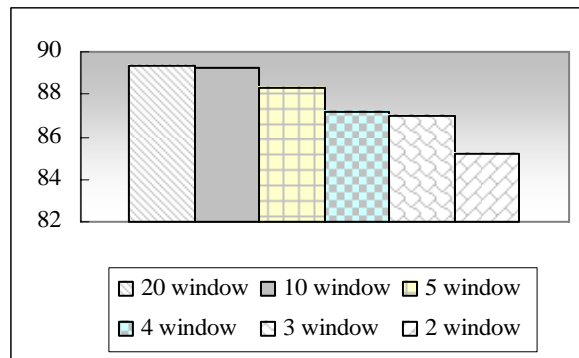


Figure 8: Difference in precision rate with the morphological information

The morphological information provides the phonological process with more information. Thus, there is a rise in the precision rate in case of smaller context windows. Morphological information seems to appropriately contribute on the phonological processing.

In case of phonotactic constraints, there is only 0.2~0.3% rise in the precision rate. The processing time with the phonotactic constraints is shorter than the processing time without it.

Like Brill (1995) experimented on the learning size. We have tested the relationship between the precision rate and the size of the training data. We found out that the precision rate is stable with more than 4,000 training data.

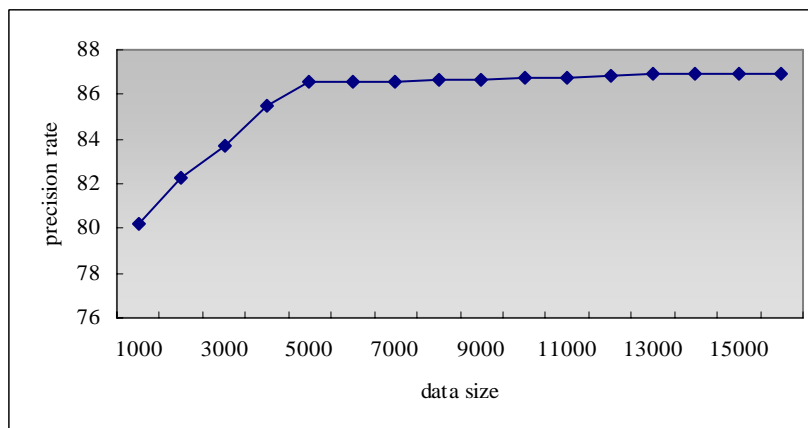


Figure 9: Training data size and precision rate

5. Conclusion

In this paper, we suggest that the TBL method generates the standard Korean pronunciation. We used a corpus and data-oriented transformation method. We found out that the larger context windows in TBL carry more morphological information.

The importance of this study lies in the speech technology. The study of the phonological change is the main topic in the domain of computational phonology. Also, the pronunciation generation is prerequisite to the speech-related technology. In text-to-speech system, the pronunciation generation mechanism provides a system with a better accurate mechanism. Also in speech recognition area, the more advanced pronunciation prediction reveals better recognition results.

The phonological information is related with the morphological encodings; regular vs. irregular, word classes, tagging information of previous word, etc. Such information is essential for the phonological processing. In this study, the concept on context window can cope with morphological information. But this idea needs further exploration.

References

- Bird S. 1995. *Computational Phonology*. Cambridge: Cambridge University Press.
- Bird S. and T. M. Ellison. 1994. One-level phonology. *Computational Linguistics*, 20(1), 55-90.
- Brill E. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, 21(4), 543-565.
- Daelemans W., S. Gillis and G. Durieux. 1994. The acquisition of stress: A data-oriented approach. *Computational Linguistics*, 20(3), 421-451.
- Gildea D. and D. Jurafsky. 1996. Learning bias and phonological-rule induction. *Computational Linguistics*, 22(4), 497-530.
- Johnson M. 1984. A discovery procedure for certain phonological rules. *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pp 344-347.
- Hayes B. and Wilson C. forthcoming. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*.
- Kaplan R. and M. Kay. 1994. Regular models of phonological rule system. *Computational Linguistics*, 20(3), 331-378.
- Karttunen L. 1993. Finite-state constraints. In Goldsmith, ed., *The Last Phonological Rule*, pp 173-194. University of Chicago Press, Chicago.
- Karttunen L. 1998. The Proper Treatment of Optimality in Computational Phonology. *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing*, pp 1-12 .
- Koskenniemi, K. 1983. *Two-level morphology*. Ph.D. thesis, Department of General Linguistics, University of Helsinki.
- Lee G. 1998. Desing and implementation of vocal sound variation rules for Korean Language. *Journal of Korean Informational Society*, 5(3), 851-861.
- Lee E. et al. 2005. IPA Converter of Korean Standard Pronunciation. *Proceedings of the Conference of Korean Cognitive Society*, pp 206-211.
- Sohn Ho-Min. 1999. *The Korean Language*. Cambridge: Cambridge University Press.
- Shin J. and J. Cha. 2005. *Korean Sound System*. Seoul: Hanuk-Munwha-Sa .