The Japanese Lexical Transducer Based on Stem-Suffix Style Forms

Masakazu Tateno, Hiroshi Masuichi, and Hiroshi Umemoto
Corporate Research Laboratories
Fuji Xerox Company Limited
430 Sakai, Nakai-machi, Ashigarakami-gun, Kanagawa 259-01
Japan
{tateno,masuichi,umemoto}@rsl.crl.fujixerox.co.jp

Abstract. This paper depicts an optimal method to construct a lexical transducer for Japanese by describing the stems and suffixes in different lexicons separately and adding an extra level of the transducers for transducing between canonical citation forms and stem-suffix style forms. This method makes it possible to reduce the complexity of rule descriptions and the computational load of intersecting compared with other methods. We made the full-size lexical transducer for Japanese. The number of states is about 60 thousand and the number of arcs is about 300 thousands. The physical size is from 800KB to 1.5MB depending on compaction methods.

1 INTRODUCTION

A Lexical Transducer (LT) is defined by Karttunen, Kaplan, Zaenen[3]. It is a specialized finite-state transducer (FST) that maps canonical citation forms of words and their morphological categories to inflected surface forms, and vice versa. An LT has the following advantages:

1. The same structure and algorithms can be used for morphological analysis (stemming) as well as generation.
2. The computational speed of the morphological process (analysis and generation) is faster and the data for the process can be compacted more tightly, compared with other methods.

The standard way to construct an LT shown by [3] consists of three steps:

1. Constructing a simple finite-state source lexicon (LA) which defines all valid canonical citation forms of the language.
2. Describing morphological alternations by means of two-level rules[5], compiling the rules to FST's, and intersecting them to form a single rule transducer (RT).
3. Merging the LA built in the step 1. with the RT built in 2. by composing.

Karttunen proposed the following two methods in constructing an LT [2]. We discuss these methods in Section 2 and 3.

1. Moving descriptions of idiosyncratic alternation rules into the source lexicons. It reduces the complexity of the descriptions of morphological alternations.
2. Performing intersection and composition in single operation. It reduces the size of intermediate rule transducers and the computational load.

A Japanese sentence is not delimited between phrases, the elements of a sentence. A phrase in Japanese consists of a word and its subsequent auxiliary verbs or postpositional particles. We made an LT which represents all the possible phrases allowed in Japanese. The LT needs to be used iteratively to analyze a Japanese sentence and outputs some different sequences of analysis results for a sentence. Therefore the algorithms over the LT that can select a correct sequence from the syntactic or semantic viewpoint are necessary. In this paper, we will focus on an optimal constructing method of the LT for Japanese phrases and omit the mention of the algorithms to select a correct sequence of analysis results.

This paper describes our recent progress by: (1) describing the stem-suffix style lexicons, (2) describing the mapping between canonical citation forms and stem-suffix style forms in the extended lexicon, and (3) describing phonological rules, (4) composing (1), (2) and (3) to produce a single LT.

2 JAPANESE LEXICAL TRANSDUCER

There are three types of characters in Japanese: Kanji, Hiragana and Katakana. Kanji characters are ideograms, while Hiragana and Katakana are phonograms.

One of the functions of Hiragana characters in Japanese orthography is to represent the inflectional suffixes, while Kanji and Katakana characters are never used to represent inflections but used to represent uninflectional words or the stems of inflectional words. Kanji and Katakana characters can be replaced with corresponding Hiragana characters according to their pronunciations. Therefore we will discuss Japanese orthography only using Hiragana characters.

As described in Section 1, we made an LT which represents Japanese phrases. An LT in Figure 1 shows the following four different Japanese phrases. A thick arrow shows that there is an alternation in the transition. The lexical forms are written over the surface forms. [CV] represents one of Hiragana

© 1996 M. Tateno, H. Masuichi and H. Umemoto
Proceedings of the ECAI 96 Workshop
Extended Finite State Models of Language
Edited by A. Kornai.

characters whose pronunciation is "CV" where C is one of the consonants in {k, g, s, z, t, d, n, h, b, p, m, y, r, w} or null and V is one of the vowels in {a, i, u, e, o}. There are some exceptions: (1)[yi], [ye], [wi], [wu] and [we] do not exist. (2)[n] which has no vowel exists.

```
L:  [ki][ru]{V1}   [na][i]{auxV}   [so][u][da]{auxV}
S:  [ki][ra]       [na][i]         [so][u][da]

L:  [ki][ru]{V1}   [na][i]{auxV}   [so][u][da]{auxV'}
S:  [ki][ra]       [na]            [so][u][da]

L:  [ki][ru]{V2}   [na][i]{auxV}   [so][u][da]{auxV}
S:  [ki]           [na][i]         [so][u][da]

L:  [ki][ru]{V2}   [na][i]{auxV}   [so][u][da]{auxV'}
S:  [ki]           [na]            [so][u][da]
```

This LT can be described by the lexicons and two-level rules in Figure 2.

Lexicons

```
LEXICON  Verb                        LEXICON  hearsayAux
[ki][ru]{V1}       NegativeAux;      [so][u][da]{auxV}      #;
[ki][ru]{V2}       NegativeAux;

LEXICON  NegativeAux                 LEXICON  guessAux
[na][i]{auxV}      hearsayAux;       [so][u][da]{auxV'}     #;
[na][i]{auxV}      guessAux;
```

Two-level rules

```
[ru]:[ra]   <=>   _ {V1}: [na]: [i]: {auxV}: ;
[ru]:0      <=>   _ {V2}: [na]: [i]: {auxV}: ;
[i]:0       <=>   _ {auxV}: [so] [u] [da]: {auxV'}: ;
```

Figure 2. Lexicons and two-level rules for an LT of Figure 1

A Japanese canonical citation form of an inflectional word always contains an inflectional suffix that represents the present tense. Therefore each of inflectional forms needs to have an alternation within a suffix (right before the morphological category) such as [ru]:[ra] for {V1} or [ru]:0 for {V2}. The LT for all the possible Japanese phrases needs more than eighty suffix alternations defined by means of two-level rules. A lot of computational power is required to intersect these rules. As Karttunen pointed out[2], the rule intersection may involve unnecessary computation that is proved when composing with the finite-state source lexicon.

Another problem is the complexity of descriptions of the rules. In the example of Figure 1, [ki] [ru] {V1}, which means 'cut', and [ki] [ru] {V2}, which means 'wear', have the same canonical citation forms. Similarly, [so] [u] [da] {auxV}, which means 'hearsay', and [so] [u] [da] {auxV'}, which means 'guess', have the same canonical citation forms. The rules in Figure 2 and almost all the other rules for suffix alternations are never phonological at all because they specify the alternations mainly determined by the specific morphological category continuations instead of the specific sequence of the phonemes. Such rules are difficult to describe, because the condition parts of them are to be applied to all the finite-state source lexicon structure according to only morphological category descriptions and the errors or inconsistencies between the lexicon and the rules often occur.

3 THE OPTIMAL WAY TO DESCRIBE THE JAPANESE LEXICAL TRANSDUCER

To reduce the enormous computational load described in Section 2, we adopt the stem-suffix style to describe the lexicons for Japanese.

In the stem-suffix style lexicons, a suffix of an inflectional word is separated from its stem and they are described in different lexicons. The various realizations of an inflectional suffix are described not in rules but in lexicons. The transducers which maps stem-suffix forms to canonical citation forms and vice versa are necessary. The transducers have extremely simple structures and are almost independent of each other. This property of the transducers is advantageous to execute the Intersecting Composition described in [2].

As described in Section 1, Karttunen proposed moving descriptions of idiosyncratic alternation rules into the source lexicons in order to reduce the complexity of the descriptions of morphological alternations. In the case of Japanese it is important to move description of non-phonological alternation (suffix alternation) rules into the source lexicons, because they are difficult to describe and the number of them is large. To adopt the stem-suffix style, the realizations of the suffixes can be described in the lexicons directly, so the description complexity of non-phonological alternations as mentioned in Section 2 can be reduced.

3.1 DESCRIBING THE STEM-SUFFIX STYLE LEXICONS

In the stem-suffix style lexicons, an inflectional suffix is removed from a canonical citation form of an inflectional word and the realizations of the suffix with tags are added in front of the continuations of the stem lexicon. It is a straightforward way to describe the Japanese lexicons. The stem-suffix style lexicons already contain the realizations determined by the specific continuation, so rules for such alternations are not necessary. Figure 3 shows that no rules remain in the examples in Figure 2. There are about twenty auxiliary verbs in Japanese. Each of them has very specific (or idiosyncratic) inflections that are written as the stem-suffix style forms in the lexicons.

3.2 DESCRIBING THE PHONOLOGICAL RULES

One of the inflections of adjectives needs the phonological rules. The stem of [a] [ri] [ga] [ta] [i] {Adj}, which means 'grateful', is [a] [ri] [ga] [ta] and its suffix is [i]. Most of the inflections are specified with stem-suffix style lexicons, but if the stem continues to the suffix [u] to mean politeness, the last syllable [ta] of the stem should be realized as [to]. So the rule for this alternation is as follows:

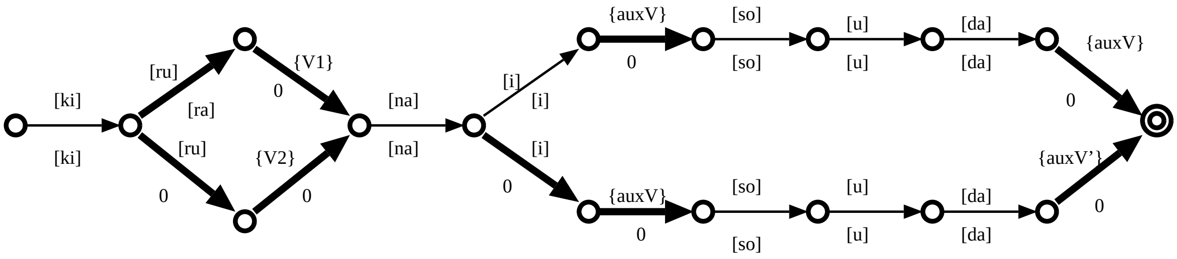


Figure 1. A lexical transducer

```
LEXICON  Verb                      LEXICON  NegAuxSuffix
[ki]{V1}        Verb1Suffix;       [i]{suff}          HearsayAux;
[ki]{V2}        Verb2Suffix;       {suff}             GuessAux;

LEXICON  Verb1Suffix               LEXICON  HearsayAux
[ra]{suff}      NegAux;            [so][u]{auxV}      soudaSuff;

LEXICON  Verb2Suffix               LEXICON  GuessAux
{suff}          NegAux;            [so][u]{auxV'}     soudaSuff;

LEXICON  NegAux                    LEXICON  soudaSuff
[na]{auxV}   NegAuxSuffix;         [da]{suff}         #;
```

Figure 3. Stem-suffix style lexicons

```
[ta]:[to] <=> _ {Adj}: [u] {suff}: ;
```

There are several other rules for inflectional alternations other than the adjectives. The canonical citation forms are never considered in the rules. It reduces the number of rules drastically and also simplifies each rule, which does not include the suffix in the canonical citation form. So the cost for intersecting the rules is very cheap. By composing the stem-suffix style lexicons and the intersected rules, the transducer between the stem-suffix style string as the lexical side and the surface string as the surface side is produced. [ki] [ra] [na] [i] [so] [u] [da] is analyzed as [ki] {V1} [ra] {suff} [na] {auxV} [i] {suff} [so] [u] {aux} [da] {suff}. Japanese language is one of the agglutinative languages. It coincides with our result that there are small number of the phonological rules, while there are a lot of continuations in the lexicons.

3.3 DESCRIBING THE TRANSDUCERS BETWEEN CANONICAL CITATION FORMS AND STEM-SUFFIX FORMS

Another group of transducers is necessary to transduce between the canonical citation forms and stem-suffix style forms such as [ki] [ru] {V1} and [ki] {V1} [ra] {suff} to make the LT for Japanese. Figure 4 shows the example of the transducers corresponding to the lexicons in Figure 3. There are about one hundred and fifty transducers in this group. A canonical citation form with the morphological category is transduced into its stem-suffix style form with the morphological category and a tag for the suffix by one of the transducers.

As Karttunen proposed [2], these transducers are not intersected but composed with the stem-suffix lexicons one by one (Intersecting Composition). The standard Intersecting Composition is executed by referring the surface side of the lexicon to find the places to compose. In our method, the intersecting composition is executed by referring the lexical side.

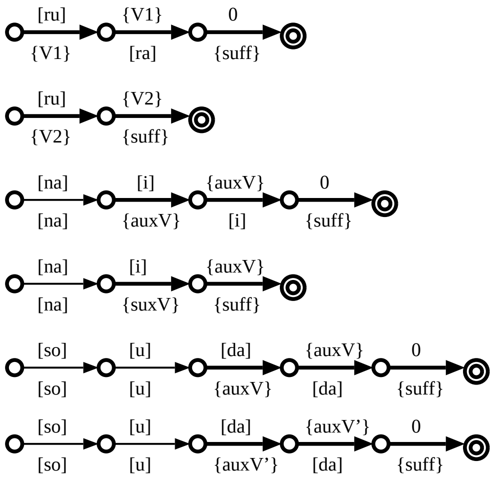


Figure 4. Transducers between canonical citation forms and stem-suffix style forms

Phonological rules can describe the alternations in the specific sequence of the phonemes in the entire lexicons ([1], [4]). It is not necessary to specify the place where the morphological alternation occurs, while each of them just specifies one alternation surrounded by other phonemes that may alter in other rules.

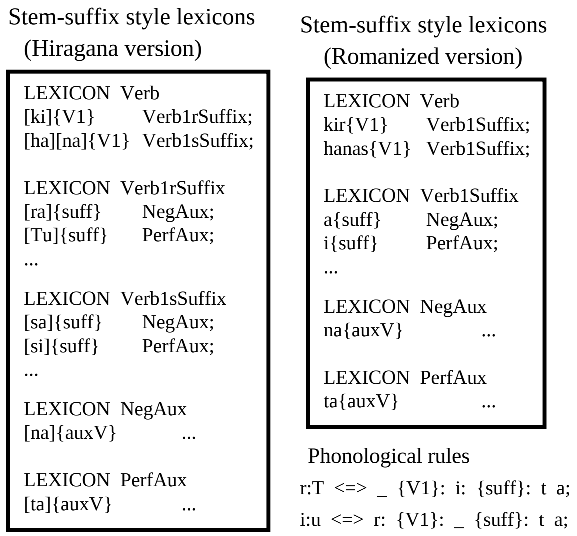
It is possible to write rules for the alternations in a specific sequence in the lexicons. But the same effect can be written as a transducer for a series of the alternations on the lexicons. Such transducers can be represented in the lexicons by writing both lexical forms and corresponding surface forms in the lexicons. This method using transducers is less powerful relative to writing rules but is enough for a series of the alternations.

4 DISCUSSION

If the lexicons are specified with Romanized characters for Hiragana, the definition of stems may not be the same. For example, the stem [ki] {V1} in Hiragana notation (Figure 3) would be kir{V1} in Romanized notation because the number of continuation lexicons can be reduced as shown in Figure 5. By attaching the first consonant of the suffix to the stem, the continuation lexicons are merged according to the first vowel of the suffix. But more phonological rules are necessary for the morphological alternations neglected by the simplification.

Transducers not only for the canonical citation forms and stem-suffix style forms but also for surface forms between Romanized characters and Hiragana are necessary.

As a result, these two notations have no significant difference on the complexity of the rules and the lexicons.



Stem-suffix style lexicons (Hiragana version)

```
LEXICON  Verb
[ki]{V1}          Verb1rSuffix;
[ha][na]{V1}  Verb1sSuffix;

LEXICON  Verb1rSuffix
[ra]{suff}        NegAux;
[Tu]{suff}        PerfAux;
...

LEXICON  Verb1sSuffix
[sa]{suff}        NegAux;
[si]{suff}        PerfAux;
...

LEXICON  NegAux
[na]{auxV}            ...

LEXICON  PerfAux
[ta]{auxV}            ...
```

Stem-suffix style lexicons (Romanized version)

```
LEXICON  Verb
kir{V1}        Verb1Suffix;
hanas{V1}   Verb1Suffix;

LEXICON  Verb1Suffix
a{suff}         NegAux;
i{suff}          PerfAux;
...

LEXICON  NegAux
na{auxV}             ...

LEXICON  PerfAux
ta{auxV}              ...
```

Phonological rules

```
r:T <=> _ {V1}: i: {suff}: t a;
i:u <=> r: {V1}: _ {suff}: t a;
```

Figure 5. Comparison of the lexicons and the rules between two notations

5 CONCLUSION

We adopted the stem-suffix style forms instead of the canonical citation forms to describe the lexicon. It removes the burden to describe the rules for most inflections. To meet the specification of the LT, the transducers between the canonical citation forms and the stem-suffix style forms is added on top of the lexicon. Thus the rules are isolated from the canonical citation forms. They basically specify the alternations within the stems.

The method of constructing the LT for Japanese is summarized as the following steps:

1. producing a finite-state machine (L) based on the lexicon of the stem-suffix style. Idiosyncratic alternations are also included in it.
2. compiling the two-level rules which represent phonological alternations to finite-state transducers, and intersect them to form a single rule transducer (R).
3. producing a set of transducers (T) which maps stem-suffix forms to canonical citation forms and vice versa.
4. merging the L with the R and T by composition in the cascading manner (T o L o R) that produces the LT for Japanese (Figure 6).

We made the full-size LT for Japanese. The number of states is about 60 thousand and the number of arcs is about 300 thousands. The physical size is from 800KB to 1.5MB depending on compaction methods.

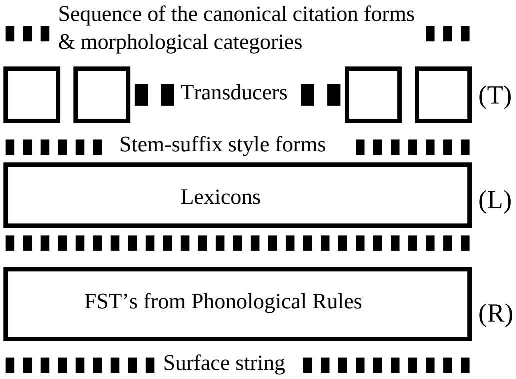

Sequence of the canonical citation forms
& morphological categories
Transducers
(T)
Stem-suffix style forms
Lexicons
(L)
FST's from Phonological Rules
(R)
Surface string

Figure 6. Layered structure of transducers

This success of the optimal way to construct the LT for Japanese emphasizes the idea pointed out by [3] that an extra level of the transducers is both practical and linguistically motivated. Transducers between the canonical citation forms and the stem-suffix forms in Japanese are typical one of such examples.

REFERENCES

[1] R. M. Kaplan and M. Kay, 'Regular models of phonological rule systems', *Computational Linguistics*, **20**, 331–378, (1994).

[2] L. Karttunen, 'Constructing lexical transducers', in *Proceedings of the sixteenth International Conference on Computational Linguistics COLING-94*, volume I, pp. 406–411, Kyoto, Japan, (August 1994). ICCL.

[3] L. Karttunen, R. M. Kaplan, and A. Zaenen, 'Two-level morphology with composition', in *Proceedings of the fifteenth International Conference on Computational Linguistics COLING-92*, volume I, pp. 141–148, Nantes, (August 1992). ICCL.

[4] L. Karttunen, K. Koskenniemi, and R. M. Kaplan, 'A compiler for two-level phonological rules', in *Tools for Morphological Analysis*, eds., M. Dalrymple and et al, Center for the Study of Language and Information, Stanford University, Palo Alto, (1987).

[5] Kimmo K. Koskenniemi, 'A general computational model for word-form recognition and production', in *Proceedings of the 10th International Conference on Computational Linguistics COLING-84*, volume I, pp. 178–181, Stanford University, California, (July 1984). ICCL.