Natural Language Engineering **2** (4) 329–330. Printed in the United Kingdom

© 1996 Cambridge University Press

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
e-mail: {tateno,masuichi,umemoto}@rsl.crl.fujixerox.co.jp

(*Received 24 December 1996; Revised 6 February 1997*)

A Lexical Transducer (LT) as defined by Karttunen, Kaplan, Zaenen 1992 is a specialized finite state transducer (FST) that relates citation forms of words and their morphological categories to inflected surface forms. Using LTs is advantageous because the same structure and algorithms can be used for morphological analysis (stemming) and generation. Morphological processing (analysis and generation) is computationally faster, and the data for the process can be compacted more tightly than with other methods. The standard way to construct an LT 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, compiling the rules to FSTs, and intersecting them to form a single rule transducer *RT*; and (3) composing *LA* and *RT*.

A Japanese lexical phrase consists of a word and its subsequent auxiliary verbs or postpositional particles. In the full paper we describe the construction of an LT which represents all the possible lexical phrases allowed in Japanese. The key problem addressed by our method is that in Japanese, the canonical citation form of a word always contains an inflectional suffix that represents the present tense, and therefore each inflectional form shows alternation within the suffix. The LT for all the possible Japanese lexical phrases requires more than eighty suffix alternations defined by means of two-level rules, and significant computational power is required to intersect these rules.

Another problem is the complexity of describing the rules. Almost all rules for suffix alternations are conditioned morphologically, in terms of continuation categories, rather than phnologically, in terms of a sequence of phonemes. Such rules are difficult to describe, because their condition parts must be applied to all the finite state source lexicon structure and errors or inconsistencies between the lexicon and the rules often occur.

To reduce the enormous computational load and the description complexity of the rules, we adopt the stem-suffix style instead of the canonical citation style to describe Japanese. In stem-suffix style lexicons, a suffix of an inflected word is separated from its stem and they are stored in different lexicons. The stem-suffix

https://doi.org/10.1017/S1351324997001575 Published online by Cambridge University Press

style lexicons can store the realizations determined by the specific morphological category continuation, so rules for such alternations are not necessary.

The stem-suffix style reduces the number of rules drastically and also simplifies each rule that does not include the suffix in the canonical citation form. Therefore, intersecting the rules becomes computationally inexpensive. 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. Since Japanese is an agglutinative language, our method exploits the fact that there are only a small number of phonological rules, while there are a large number of continuations in the lexicons.

To make the LT for Japanese, another group of transducers is required between the canonical citation forms and stem-suffix style forms. There are about 150 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 1994 pointed out, intersecting the rules may involve unnecessary computation that can be avoided by performing the composition with the finite state source lexicon in a single step of 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: canonical to stem-suffix style transducers are not intersected, but composed with the stem-suffix lexicons in a single step.

The method of constructing the LT for Japanese described in the full paper involves the following steps:

- Producing a finite state machine L based on the lexicon of the stem-suffix style. Idiosyncratic (non-phonological) alternations are included here.
- Compiling the two-level rules which represent phonological alternations to finite state transducers, and intersect them to form a single transducer R.
- Producing a set of transducers T which relate stem-suffix forms to canonical citation forms.
- Composing T, L and R in cascade to produce the LT for Japanese.

We made a full-size LT for Japanese. The number of states is about 60k, the number of arcs is about 300k. The physical size is from 800KB to 1.5MB, depending on compaction methods. The success of our method using transducers between the canonical citation forms and the stem-suffix forms in Japanese lends further support to the idea Karttunen *et al.* 1992 that an extra level of transducers is both practical and linguistically motivated.

References

Karttunen, L., Kaplan, R. M. and Zaenen, A. 1992. Two-Level Morphology with Composition. *Proceedings of the fifteenth International Conference on Computational Linguistics COLING-92* **I** pp. 141–148

Karttunen, L. 1994. Constructing Lexical Transducers. *Proceedings of the sixteenth International Conference on Computational Linguistics COLING-94* **I** pp. 406–411.

https://doi.org/10.1017/S1351324997001575 Published online by Cambridge University Press