SYLLABLE-BASED MODEL FOR THE KOREAN MORPHOLOGY

Seung-Shik Kang

Dept. of Computer Science & Statistics
Hansung University
Seoul 136-792, Korea

Yung Taek Kim

Dept. of Computer Engineering
Seoul National University
Seoul 151-742, Korea

Abstract

This paper describes a syllable-based computational model for the Korean morphology. In this model, morphological analysis is considered as a process of candidate generation and candidate selection. In order to increase the performance of the system, the number of candidates is highly reduced and the system requires small number of dictionary accesses. Idiosynchratic features of a syllable, formalized as a characteristic function, make it possible to reject implausible candidates before dictionary confirmation. Instead of a letter, syllable is a basic processing unit for the practical implementation of the morphological analyzer.

1. Introduction

There are two linguistic phenomena that are interested in the processing of computational morphology. They are morphological transformation and morpheme identification. Two-level model and syllable-based formalism focussed on the problem of morphological transformation [Bear88, Cahi90, Kosk83]. Morpheme identification is an important issue in some languages where two or more morphemes are combined to make a word, a compound word, or a sentence without any delimiters between morphemes [Abe86, Chen92, Pach92].

The goal of morphological analysis is to find the base form of morphemes in a word. It consists of a generation of analysis candidates and the selection of correct candidates. Analysis candidates are generated as a reverse process of word formation rules: morpheme isolation and morphological transformation. Then, correct candidates are selected by the coherence restrictions among adjacent morphemes and dictionary confirmation. The morphological analyzer tries to generate all the possible candidates only to accept the correct candidates.

2. The problem

Two-level model is widely known to be a computationally efficient method for the practical system on the condition that the number of rules is small[Bart86, Kosk88]. However, when the size of the rulebase is large it causes an exponential problem. In case of the Korean language, it is common that a stem is succeeded by grammatical morphemes. If we use the two-level model for a practical system, a small set of phonological rules and a large set of morpheme isolation rules are required because there are several thousand combinations of grammatical morphemes[Zhan90].

In order to solve the problem, we can try a 2-pass algorithm. All the possible morphemes are isolated, and then do a phonological processing. It is also possible to do a phonological processing first and morphemes are isolated at the second pass. However, this kind of solution causes another serious problem that occurs from the conditional restrictions: (1) some morphological transformation occurs not only at a stem but also at a functional morpheme, (2) there are cooccurrence restrictions between two morphemes, (3) morphological transformation occurs only for the special word group.

3. Syllable-based writing system

The writing system for most languages is based on the letter set called as alphabet. Instead of a letter set, Chinese writing

system is based on the set of characters that consists of one or more letters. Each character is a meaning unit and words are represented by the combination of characters. In case of Korean, words are represented by one or more characters as in Chinese. The difference is that Korean character is a well-formed written syllable, which is a sound unit rather than a meaning unit as in Chinese. A written syllable is a combination of two or three sound symbols, which corresponds to a spoken syllable in a one-to-one fashion[Chun90]. Korean words are constructed as follows based on the syllable unit.

word ::= { syllable }*
syllable ::= open_syll | closed_syll
open_syll ::= initial + medial
closed_syll ::= initial + medial + final

4. Idiosynchratic features of syllable

There are 11,172 syllables in the modern Korean language(= 19 initials * 21 medials * 27 finals plus one for null). However, it is interesting to investigate the usage of syllables to make a word. About 2,350 syllables cover more than 99.9% of the modern Korean words. Furthermore, 267 syllables(11.36% of 2,350 syllables) are only used for the surface form of verbs, and grammatical morphemes are combinations of 151 syllables(6.43% of 2,350 syllables). In addition, only a very small set of syllables, 1 to 46 syllables for each type of irregular verbs, are tied to the morphological transformation[Kang93]. This kind of information is very useful to improve the efficiency of the morphological analyzer. For example, if a syllable used only for the surface form of verb is found in a word, we can easily guess that the word is a verb, the string before that syllable is a stem, and the rest is a grammatical morpheme. There is no other chance for the different result except typographic errors.

Suppose that X is a set of syllables that are used at the first position of grammatical morphemes. We can easily guess the syllable boundary position of grammatical morpheme in an n-syllable word at syllable x_i, where $x_j \in X$ and $i \leq j \leq n$. There is no the possibility at other positions. It is based on the fact that only 48 syllables are used for the first position of postpositions and 72 syllables for the first position of final endings in the Korean language.

Three kinds of syllable features are defined from where the features are extracted. 'Unit feature' is a syllable feature defined on the syllable itself. If a syllable x_i itself has an idiosynchratic feature f_j, then x_i has a unit feature f_j. 'Partial feature' is defined by the component of a syllable. A syllable x_i is called to have a partial feature p_k, if x_i includes a component p_k as an initial, a medial, or a final letter. 'Successive feature' is a meta-level feature defined for the adjacent two syllable features. For example, if there is a set of two successive syllables x_ix_{i+1} that construct grammatical morphemes and that cannot construct any noun/verb, then the boundary position of a grammatical morpheme is possible only at syllable x_i or x_{i+1}.

5. Characteristic function

Idiosynchratic features of syllables are represented using a characteristic set of syllables. Suppose that a part of speech(i), morpheme length(j), and the position of syllable in a word(k) are discriminating features of a characteristic set. Let $\mathbb{P}_i$ be a set of syllables that are used for a part of speech i, $\mathbb{Q}_j$ be a set of syllables that are used for the morpheme length j, and $\mathbb{R}_k$ be a set of syllables that are used for the k-th position of syllable in the word. Then, a characteristic set of syllables A<i,j,k> is an intersection of $\mathbb{P}_i$, $\mathbb{Q}_j$, and $\mathbb{R}_k$.

$$A<i,j,k> = \mathbb{P}_i \cap \mathbb{Q}_j \cap \mathbb{R}_k$$

For the characteristic set of syllables A<i,j,k>, characteristic function $C_{A<i,j,k>}$ is defined from A<i,j,k> to {0,1}.

[Definition] characteristic function
Let X be a set of Korean syllables and A<i,j,k> be a characteristic set of syllables

where $A<i,j,k> \subseteq X$ for part of speech i, morpheme length j, and the k-th position of morpheme. Define the function

$$C_{A<i,j,k>} : X \longrightarrow \{ 0, 1 \}$$

$$C_{A<i,j,k>}(x) = \begin{cases} 1, & \text{if } x \in A<i,j,k> \\ 0, & \text{otherwise} \end{cases}$$

A lot of characteristic functions are possible by the arguments i, j, and k. However, some of them are chosen for the morpheme isolation or morphological transformation, and they are reorganized as syllable information function(f) in order to find out the characteristics of a specific syllable. The value of $f(x)$ on a syllable x is defined by the characteristic function $C_{A<i,j,k>}(x)$. Suppose that α be the number of parts of speech, β be the maximum number of syllables in a word, then a triple $A_{<i,j,k>}$ can be transformed into A_t by the following expression.

$$t = (k-1)*\alpha*\beta + (j-1)*\alpha + i$$
$$(1 \le i \le \alpha,\ 1 \le j \le \beta,\ 1 \le k \le \beta)$$

Let g be a function from a set of syllables to a Cartesian product of characteristic functions and h be a function from a Cartesian product of characteristic functions to an integer. Then, function g and h are defined as follows.

$$g : X \longrightarrow C_{A1} \times C_{A2} \times \ldots \times C_{An}$$
$$g(x) = (C_{A1}(x), C_{A2}(x), \ldots, C_{An}(x))$$

$$h : C_{A1} \times C_{A2} \times \ldots \times C_{An} \longrightarrow N$$
$$h(C_{A1}(x), C_{A2}(x), \ldots, C_{An}(x)) = \Sigma (C_{Ai}(x)*W(i)), \text{ where } W(i)=2^{i-1}$$

Now, syllable information function f is defined as a combination of h and g. Domain of the function f is a set of syllable and the range is a bit string of integer where bit position t is used for the specific feature and the value of the t-th bit means whether the syllable has the corresponding feature or not.

$$f : X \longrightarrow N$$
$$f(x) = \sum_i (C_{Ai}(x)*W(i)), \text{ where } W(i)=2^{i-1}$$

6. Syllable-based formalism

Morphological analysis system is formalized as a function F. The domain of function F is a set of words and the range of F is a Cartesian product of a set of morphemes and their morpho-syntactic features.

$$y = F(x)$$
$$F : W \longrightarrow W'$$
W : a set of words
$W' = M \times F$
M : a set of morphemes
F : a set of morpho-syntactic features

Suppose that m_i be a root form of lexical morpheme, f_j be a combination of features and r_k be a two-level rule. Then, function F is defined as follows. Function p is to check the condition of two-level rules. Function q generates a combination of morpho-syntactic features of a word.

$$F(\text{word}) = \begin{cases} \text{a set of } (m_i, f_j), \\ \quad \text{if } m_i = p(\text{word}, r_k) \text{ and} \\ \quad\quad f_j = q(\text{word}) \\ \Phi, \text{ otherwise} \end{cases}$$

Some morpho-syntactic features are defined for the morphological analysis. Parts of speech, irregular types and other features are defined as follows.

pos = { N, V, ADJ, ADV, DET, ... }
irtype = { B, D, G, H, L, N, R, S, U }
prefix = (prefix-1, prefix-2, ... , prefix-n)
suffix = (suffix-1, suffix-2, ... , suffix-n)
pres, past, fut, pp, hon, ... = { +, - }

A syllable-based rule consists of left-hand side(LHS) and right-hand side(RHS). They are described by the following primitive functions.

syllable(word, i)
subsyl(word, i, j)
$C_{A<i,j,k>}(x)$
irreg_type(word)

initial(x), medial(x), final(x)
noun(word), verb(word), adv(word),
det(word), impr(word)
change(x, y, z, INITIAL/MEDIAL/FINAL)
insert(x, word, i):
    insert syllable *x* at *i*-th position
delete(word, i): delete *i*-th syllable

'syllable(word,i)' fetches *i*-th syllable of word and 'subsyl(word,i,j)' is to get *j* syllables starting from *i*-th syllable of word. $C_{A<i,j,k>}$ is to check whether a syllable *x* belongs to a syllable characteristic function or not. For example, *b*-irregular rule in Korean is described as follows. Set 'A_7' is supposed to be a characteristic set of the last syllables of b-irregular verbs.

C_{A7}(s[i]) = 1,
head <— subsyl(word, 1, i-1),
change(head[i-1], null, 'p(ㅂ)', FINAL),
verb(head) <— IRREG_B

↓

tail <— subsyl(word, i, n-i-1),
change(tail[1], 'we(ㅝ)', 'e(ㅓ)', MEDIAL)

The *b*-irregular rule is described as a syllable-based formalism and it is applied after the isolation of stem parts. So, stem and ending candidates should be identified first.

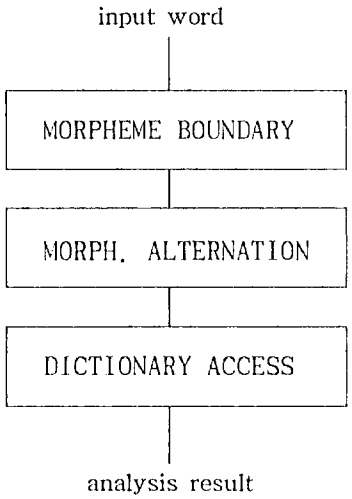

Fig. morphological analysis

Overall view of the morphological analyzer is shown in the figure. The first step is to find the morpheme boundaries using characteristic function for syllables. Stem candidates are generated at the second step by the phonological rules. Phonological rules are only applied at a syllable w[i] if and only if w[i-1] is an element of a required characteristic set, and w[i+1] is the beginning syllable of other morpheme.

Following algorithm is to guess the beginning position of grammatical morpheme. In the algorithm, GM_SET1 and GM_SET2 are characteristic sets for the first and the rest syllables of grammatical morphemes, respectively.

```
algorithm boundary_syllable(word)
syllable word[]; /* input word */
begin
    n = nsyl(word);
    for (i = 1; i < n; i = i+1) {
        if (word[i] ∈ GM_SET1) {
            if (word[i+1] ∈ GM_SET2)
                return(i);
        }
    }
    return(n);
end
```

Algorithm. morpheme boundary

7. Evaluation of the model

There are two types of candidates for a word. The first type is generated by the morpheme isolation at all the syllable boundary and the second type is generated for each morpheme candidate by the phonological rules. We can count the number of candidates as follows. Suppose that α be the maximum number of syllables that causes an inflexion, β be the candidates for prefinal endings, and γ be the maximum number of inflexions for one syllable. In case of Korean, α is less than n, β is 2, and γ is 3. If a word consists of n syllables, then the maximum number of candidates is $10n+8\alpha+2$.

- candidates for 1-morpheme word
  and (noun+postposition)

① 1-morpheme word: 1
② noun + postposition: n-1
③ noun + suffix + postposition: n-2

- candidates for irregular verbs and (verb+ending)

④ verb + ending: $n-1+\alpha$
⑤ verb + prefinal_ending + ending: β
⑥ verb inflexion: $\gamma(n-1+\alpha+\beta)$
⑦ verb + suffix + ending:
$(n-2+\alpha+\beta) + \gamma(n-2+\alpha+\beta)$

$$C(n) = ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦$$
$$= 1 + (n-1) + (n-2) + (n-1+\alpha) + \beta + \gamma(n-1+\alpha+\beta) + (n-2+\alpha+\beta) + \gamma(n-2+\alpha+\beta)$$
$$= (4+2\gamma)n + (2\alpha+2\alpha\gamma+2\beta+2\beta\gamma-3\gamma-5)$$
$$= 10n + 8\alpha + 2 \quad \leftarrow\!\!- \quad \beta=2,\ \gamma=3$$

It is very inefficient to look up the dictionary for all the implausible stems and grammatical morphemes. Only plausible candidates are generated using the idiosynchratic features of syllable. Now, maximum number of candidates is counted as a constant and the number of dictionary accesses is highly reduced.

① 1-morpheme word: 1
② noun + postposition: 2
③ noun + suffix + postposition: 2
④ verb + ending: 2
⑤ verb + prefinal_ending + ending: 2β
⑥ verb inflexion: $\gamma(2+2\beta)$
⑦ verb + suffix + ending: $(2+2\beta)+\gamma(2+2\beta)$

$$C(n) = ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦$$
$$= 2\beta + 4\gamma + 4\beta\gamma + 9$$

The previous algorithm has O(n) complexity because it tries to isolate function word at all the syllable positions. However, if syllable features are used then the worst-time complexity of the Korean morphological analysis becomes a constant. In this case, we should use the fact that there is no stem that includes two successive syllables 'xy' such that 'xy' is a substring of grammatical morpheme.

8. Conclusion

Syllable-based formalism is proposed to solve the problem of morphological alternation with morpheme isolation where many candidates are generated by the phonological rules. It improved the worst-time complexity O(n) to a constant, and the number of dictionary accesses is highly reduced using the syllable features that are extracted from words and formalized to be available for a morphological analyzer. They are very useful for the isolation of morphemes, which make it possible to guess the boundary position of a stem without accessing the dictionary. They are also useful to reject the implausible base forms from a verb.

Characteristic set of syllables and syllable-based formalism may be applied for the languages whose words consists of syllables and morphological operation is described as a syllable-to-syllable transformation to increase the performance of the morphological analyzer. In addition, idiosynchratic features of syllable may be used for the analysis and recognition of natural languages such as spelling check, phonological representation of words, and character recognition.

Korean morphological analyzer was implemented at IBM-PC 486 using C language. The system analyzed Korean text at a speed of about 100 words/sec.

REFERENCES

[Abe86] M. Abe, Y. Ooshima, K. Yuura and N. Takeichi, "A Kana-Kanji Translation System for Non-Segmented Input Sentences Based on Syntactic and Semantic Analysis," Proceedings of the 11th International Conference on Computational Linguistics, pp.280-285, 1986.

[Bart86] E. Barton, "Computational Complexity in Two-Level Morphology," 24th Annual Meeting of the Association for Computational Linguistics, 1986.

[Bear88] J. Bear, "Morphology and Two-level Rules and Negative Rule Features," Proceedings of the 12th International Conference on Computational Linguistics,

vol.3, pp.28-31, 1988.

[Cahi90] L.J. Cahill, "Syllable-based Morphology," Proceedings of the 13th International Conference on Computational Linguistics, vol.3, pp.48-53, 1990.

[Chen92] K.J. Chen and S.H. Liu, "Word Identification for Mandarin Chinese Sentences," Proceedings of the 14th Internatioanl Conference on Computational Linguistics, Vol.1, pp.101-107, 1992.

[Chun90] H.S. Chung, "A Phonological Knowledge Base System Using Unification-based Formalism - A Case Study of Korean Phonology -," Proceedings of the 13th International Conference on Computational Linguistics, pp.76-78, 1990.

[Kang93] S.S. Kang, *Korean Morphological Analysis using Syllable Information and Multi-word unit Information*, PhD dissertation, Seoul National University, 1993.

[Kosk83] K. Koskenniemi, "Two-level Model for Morphological Analysis," Proc. of the 8th International Joint Conference on Artificial Intelligence, pp.683-685, 1983.

[Kosk88] K. Koskenniemi, "Complexity, Two-Level Morphology and Finnish," Proceedings of the 12th International Conference on Computational Linguistics, pp.335-339, 1988.

[Pach92] T. Pachunke, O. Mertineit, K. Wothke and R. Schmidt, "Broad Coverage Automatic Morphological Segmentation of German Words," Proceedings of the 14th Conference on Computational Linguistics, pp.1219-1222, 1992.

[Zhan90] B.T. Zhang and Y.T. Kim, "Morphological Analysis and Synthesis by Automated Discovery and Acquisition of Linguistic Rules," Proceedings of the 13th International Conference on Computational Linguistics, pp.431-436, 1990.