Lexical Analysis of Agglutinative Languages Using a Dictionary of Lemmas and Lexical Transducers

Sun-Mee Bae and Key-Sun Choi

Division of Computer Science, Dept. of EECS, Korea Advanced Institute of Science and Technology
373-1 Guseong-dong Yuseong-gu Daejeon 305-701 Korea
{sbae, kschoi}@world.kaist.ac.kr

Abstract

This paper presents a simple method for performing a lexical analysis of agglutinative languages like Korean, which have a heavy morphology. Especially, for nouns and adverbs with regular morphological modifications and/or high productivity, we do not need to artificially construct huge dictionaries of all inflected forms of lemmas. To construct a dictionary of lemmas and lexical transducers, first, we construct automatically a dictionary of all inflected forms from KAIST POS-Tagged Corpus. Secondly, we separate the party of lemmas and one of sequences of inflectional suffixes. Thirdly, we describe their lexical transducers (i.e., morphological rules) to recognize all inflected forms of lemmas for nouns and adverbs according to the combinatorial restrictions between lemmas and their inflectional suffixes. Finally, we evaluate the advantages of this method.

1. Introduction

For inflectional languages like English, French, etc., it is easy to automatically construct electronic dictionaries of all possible inflected forms from dictionaries of lemmas and their inflectional transducers through the aid of an inflection module in an automatic system such as INTEX (Silberztein, 2000). That is to say, two separately constructed data sets (i.e., dictionaries of lemmas and their inflectional transducers) are merged into a single dictionary by inflectional codes attached to each entry. The name of the transducer is the same as the one of the code associated with each entry of dictionary of lemmas. The dictionary of all inflected forms is in turn compressed into a Minimal Finite-State Automaton associated with a linear-time lookup procedure. Lexical analysis of texts is then performed with a mere lookup of compressed dictionaries. While the uncompressed size of the French dictionary with 700,000 inflected forms is more than 30 Mb, the size of compressed dictionaries is only several Mb.

However, for agglutinative languages like Korean, it is almost impossible to generate and compress all possible inflected forms of lemmas using the inflectional module, since the number of inflected forms would be enormous. We estimate that the number of all inflected forms for all grammar categories (verbs, adjectives, nouns and even adverbs) in Korean would be at least over 723 million and the size over 55 Gb.

Another reason, which makes it difficult to use all possible inflected forms, comes from the absence of a table or a list of nominal or verbal inflectional suffixes. Since Korean is an agglutinative language (that is to say, a rich inflectional language), it is very hard to manually generate all possible sequences of inflectional suffixes for each lemma. In traditional grammar, there have been studies on the function or phonological conditions of each inflectional suffix, not of inflectional sequences.

For these reasons, in the traditional lexical analysis of Korean, two separate dictionaries (one for lemmas and another for each inflectional suffix) are used with several statistical theories. However, these approaches often make segmentation errors and over-analyses by segmentation processing with syllabic or character units. To resolve these problems, we propose to perform the lexical analysis using a dictionary of inflected forms for verbs and adjectives with irregular morphological modifications. For nouns and adverbs with regular morphological modifications and/or high productivities, we propose to use a dictionary of lemmas in conjunction with a morphological parser in lexical analysis.

2. Choice of Dictionary Types

If we can use a dictionary of all inflected forms, it would be the best choice for lexical analysis of Korean as it reduces the number of ambiguous segmentation and one of over-analyses. However, we cannot use it because of its huge size in bytes. There are two major points under consideration in using dictionaries for an efficient lexical analysis: size of inflected forms and degree of morphological modifications.

2-1. Quantity and Size of Inflected Forms of Korean

In Korean, adjectives are inflected such as verbs. According to Jee-Sun Nam (1997), the number of inflectional suffixes for one adjective is about 6,000. There is not a great difference between inflected forms of verbs and those of adjectives except some special inflected forms for verbs. Therefore, we may assume that one adjective or verb can be attached to at least 6,000 verbal suffixes.

The number of inflectional suffixes for a noun ranges from 1,650 to 3,400 depending on their classes (Bae, 2002). We classified all Korean nouns into 9 classes according to the combinational restrictions based on phonological context or semantic features between nouns and nominal inflectional suffixes. In our experiment, 200 million inflected forms were generated for 45,000 Korean compound nouns with optional spacing. The size of uncompressed dictionary including all lexical information was 12 Gb. However, the size of 200 million inflected forms was too large to be compressed into one transducer. Therefore, we used 30 transducers for all inflected forms for compound nouns and then, we compressed 30 transducers. The compressed size was about 1 Gb with lexical information.

In Korean, there are three types of nominal inflectional suffixes for Korean nouns: casual postpositions, auxiliary postpositions, and conjunctive postpositions. Among them, some auxiliary postpositions are used to add certain semantic meanings to sentence, and they can be also attached to certain adverbs to do the same function. We can estimate all inflected forms as below[1]:

Grammar Category	Number of Lemma	Number of Inflectional Suffixes	Number of Inflected forms	Size of Inflected forms
Noun	150,000	2,000	300×10^6	18 Gb
PropN	50,000	2,000	100×10^6	6 Gb
TechN	100,000	2,000	200×10^6	12 Gb
Adverb	2,500	100	0.25×10^6	12 Mb
V/Adj	20,500	6,000	123×10^6	7.4 Gb
Total	222,500	8,100	723×10^6	55 Gb

Table 1. Number and size of inflected forms

2-2. Morphological Modifications in Concatenation

In addition to the size of inflected forms, we have to consider morphological modifications in concatenation according to the grammar categories for the efficient lexical analysis of Korean.

Morphological modification in inflected forms of adjectives and verbs are too complicate to process the segmentations between the part of lemmas and one of inflectional verbal suffixes. We have to consider not only 11 types of irregular conjugations but also vowel harmony [2] and contraction between vowels, when two morphemes are concatenated. There would be modification of only stems or one of both stems and suffixes.

However, morphological modification of inflected forms for nouns or adverbs does not exist except some contraction between nouns ending with vowels and non-syllable forms (*n*, *l*). Therefore, it is easy to segment inflected forms of nouns into lemmas and their inflectional suffixes. Moreover, there are three combinational restrictions between nouns and inflectional suffixes according to the phonological conditions: nouns ending with consonants except *l*, nouns ending with any vowel and nouns ending with the consonant *l*[3].

2-3. Choice of Dictionary Type

Through the Section 2-1 and 2-2, we know the productivity of adjectives and verbs is not so high such as nouns, but morphological modification of these inflected forms are too irregular. To avoid segmentation errors between lemma and inflectional suffixes, we propose to use dictionary of inflected forms for adjectives and verbs. However, for nouns and adverbs, there exist only two simple contractions between lemmas and their inflectional suffixes. Moreover, especially nouns are productive categories to generate all inflected forms. Therefore, we propose to perform the lexical analysis using a dictionary of lemmas in conjunction with a morphological parser for nouns and adverbs with regular morphological modifications and/or high productivity.

3. Construction of Dictionary

3.1. Construction of Dictionary of Inflected Forms

[1] *PropN* means proper nouns and *TechN* does technical nouns. We estimate the number of lemma through editorial dictionaries. For the number of sequences of inflectional suffixes, we follow dictionaries which are manually constructed by Jee-Sun Nam (1997) and Sun-Mee Bae (2002), since there does not exist a list of sequences of inflectional suffixes in editorial dictionaries.

[2] Vowels such as *a* or *o* are named clear vowel, and they are considered as requiring one of clear vowels in suffix forms; vowels such as *e* or *u*, *eu* or *i* are called dark vowels, which are considered as requiring one of dark vowels in suffix forms. This phenomenon is called vowel harmony in Korean grammar.

[3] For details, see Sun-Mee Bae (2002).

Since there does not exist the table or list of all inflected forms in Korean like English or French, first of all, we have to construct a dictionary of all possible inflected forms for all grammar categories. We used 1 million KAIST POS-Tagged corpuses to automatically construct a dictionary of inflected forms. As the first step, we segmented by separators all typographic units that would be inflected forms or invariables[4]. If we store all typographic units in a dictionary, it would be automatically the dictionary of inflected forms. However, it would not be sufficient to recognize all typographic units in Korean texts. Therefore, to complete all possible inflected forms, we need to prepare a dictionary of lemmas and one of sequences of inflection suffixes to generate or to recognize them. As the second step, we automatically segment the part of lemmas and one of inflectional suffixes using tags, which indicate inflectional suffixes. Then, we describe inflectional codes according to the combinational restriction rules in dictionary of lemmas and one of sequences of inflection suffixes. In this step, we can use editorial dictionaries to complete all possible inflected forms if necessary. Finally, we can automatically generate all possible inflected forms for verbs and adjectives. For nouns and adverbs, we propose to use lexical transducers (cf. See Section 3.2). We can extend these dictionaries by adding more POS-Tagged corpus and by repeating this dictionary construction processing. The principle of dictionary construction is shown below:

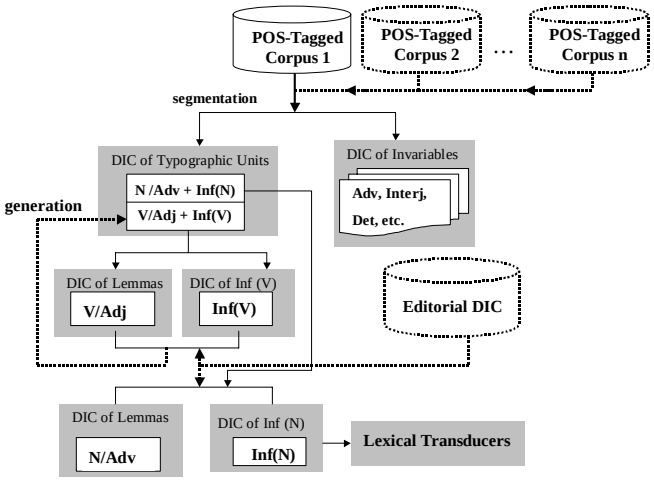

Figure 1. Construction of Dictionaries From KAIT POS-Tagged Corpus[5]

By using only 1 million POS-Tagged Corpus with 64,572 sentences (23 Mb), we obtained dictionaries in Table 2:

Grammar Category	Number of forms	Number of	Number of suffixes
N + Inf(N)	95,160	40,861	3,387
CN + Inf(N)	42,673	37,175	678
Adv + Inf(N)	371	170	122
V/Adj + Inf(V)	101,724	11,001	7,687
Vaux + Inf(V)	11,209	3,509	1,727
Invariables	58,743	-	-
Punctuations	154	-	-
Total	**309,880**	**92,716**	**13,601**

Table 2. Number of Inflected Forms Constructed From POS-Tagged Corpus[6]

3.2. Construction of Lexical Transducers

Morphological rules are implemented in the form of lexical transducers. They are the graphs whose input is used to recognize the word forms, and whose output to compute the corresponding lemma and lexical information. Lexical transducers, which are stored in the dictionary directory, function just like electronic dictionaries. The results of the morphological parsing are stored in the vocabulary of the text, exactly the same way as in the dictionaries of inflected forms.

Before construction of lexical transducers, we have to classify nouns according to the combinatorial restrictions between lemmas and their inflectional suffixes, and then describe inflectional codes in the dictionary of lemmas. For lexical transducers of nouns, we use the list of nominal inflectional suffixes, which is automatically constructed from tagged corpus (cf. Section 3.1). Then, we manually describe their lexical transducers, which allow us to recognize inflected forms with correct segmentations not only between nouns and nominal inflectional suffixes but also between inflectional suffixes. We used the basic classification of Sun-Mee Bae (2002). The example of a lexical transducer extracted from <Flex31.grf> [7] is shown below (Figure 2).

4. Lexical Analysis with Dictionary of Lemmas and Lexical Transducers

We can describe useful information in the dictionary of lemmas for the natural language processing applications. Here is an example of

[4] Invariables means the forms without inflectional suffixes.

[5] *Inf(N)* means nominal inflectional suffixes and *Inf(V)* does verbal inflectional suffixes.

[6] *CN* means compound nouns with at least one space, *Vaux* auxiliary verbs with a space.

[7] *L* stands for any letter.

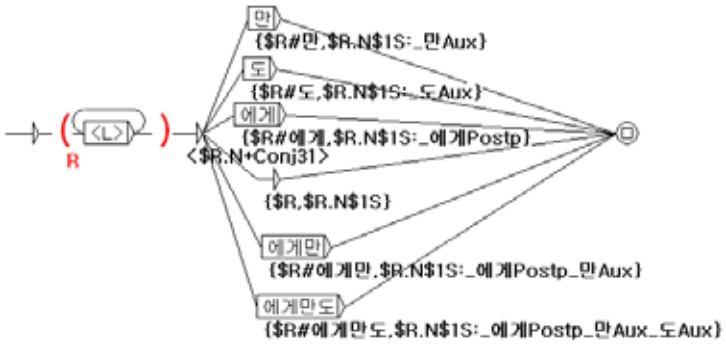

Figure 2. Extracted Example of < Flex31.grf>

Korean dictionary of lemmas[8]:

(1)*gajeong^gyosa*,N+Flex31+sk_sk-PRED+Hum
"a private teacher"

In (1), "^" indicates spacing information. "sk" means sino-Korean vocabulary and "-PRED" non-predicativity. "Hum" indicates human nouns. "Flex31" is a code to recognize all nouns that end with a vowel and that semantically belong to human nouns, collective human nouns or animal nouns. For instance, the morphological inflection graph "Flex31" uses the lexical constraint <$R.N+Flex31> together with a dictionary of lemmas. It insures that nouns are only listed in the dictionary, and that they are inflected in the right way. Finally, all inflected forms of nouns with the code "Flex31" are recognized by the graph <Flex31.grf>. The form without any space *gajeonggyosa-ege-do* and the one with a space *gajeong gyosa-ege-do* (private teacher-to-also = also to a private teacher) are recognized by the graph "Flex31"; the lemma variable "$R" then stores the lemmas *gajeonggyosa* and *gajeong gyosa*; morphological parser checks if *gajeonggyosa* and *gajeong gyosa* are lexical entries associated with the information "N+Flex31" in the dictionary; then it finally produces the resulting tags: {gajeonggyosa-ege-do,gajeonggyosa.N+Flex31+pk-PRED+Hum: *_egePostp*_Aux} and {gajeong gyosa-ege-do,gajeonggyosa.N+Flex31+pk-PRED+Hum: _egePostp _Aux}.

The method shown above has the following advantages:

(i) We can automatically construct a dictionary of inflected forms, one of lemmas and one of sequences of inflectional suffixes from POS-Tagged Corpus.
(ii) Without constructing huge dictionaries of all inflected forms for nouns and adverbs, the dictionary of lemmas in conjunction with the morphological parser allows segmentation of inflected forms into a lemma and an inflectional suffix without errors.
(iii) The lexical transducers of sequences of inflectional suffixes make their segmentation in inflectional suffixes possible without errors.
(iv) Each entry of the dictionary of lemmas comprises useful codes for the natural language processing applications: codes indicating spacing, semantic features, status of predicative noun, origin and grammatical category, etc.

[8] We can add useful codes for the natural language processing applications into each entry of the dictionary of lemmas.

5. Conclusion

In this paper, we proposed to use two different types of dictionaries according to the grammar categories in lexical analysis: for verbs and adjectives a dictionary of inflected forms, and for nouns and adverbs a dictionary of lemmas and lexical transducers, without constructing a huge dictionary in size of all inflected forms of lemmas. Especially, using dictionary of lemmas for nouns and lexical transducers of nominal inflectional suffixes, we can expect the same results as in the dictionaries of inflected forms. In addition to the instructions of uses of dictionaries, we propose a method for automatic construction of dictionaries of inflected forms, lemmas and sequences of inflected suffixes in Korean.

Acknowledgement

This work was supported by the Brain Korea 21 Project, School of Information Technology, KAIST in 2004.

References

Bae, Sun-Mee (2002). Le dictionnaire électronique des séquences nominales figées en coréen et de leurs formes fléchies. *Thèse de doctorat*, IGM. Université de Marne-la-Vallée.

Nam, Jee-Sun (1997). Lexique-grammaire des adjectifs coreens et analyse syntaxique automatique, *Langages* 126, Larousse, Paris.

Silberztein Max (2000). *Le manuel d'INTEX*. LADL, Université de Paris 7.