Korma 2003: Newly Improved Korean Morpheme Analysis Module for Reducing Terminological and Spacing Errors in Document Analysis*

Ho-cheol Choi and Sang-yong Han

Graduate School of Information, Chungang University, Korea
hansy@cau.ac.kr

Abstract. The paper describes the newly improved Korean morpheme analysis module KorMa 2003. This new module applies the custom user dictionary for analyzing new and unknown words and special terms and operates an automatic word spacing module during post-processing to prevent failures of sentence analysis due to incorrect spacing between words. KorMa 2003 has accuracy enhanced by 15% in comparison with the previously reported version.

1 Introduction

With the emergence of the Internet industry, the amount of documents produced and distributed online is increasing tremendously. One notable fact is that such documents are generally ungrammatical and also contain many newly coined words. What is more, in Korean language most words are composed of several roots, each root corresponding to one syllable, i.e., one Korean glyph, e.g., *dehanminguk*—the official name of South Korea: *de* 'great', *han* 'Korean', *min* 'democracy', *guk* 'country'. This causes many errors consisting in incorrect spacing between words, such as **dehanmin guk* or **de hanminguk*. Thus, there is a need for a morpheme analysis module with improved analysis of newly coined words, special terms used in special fields, and words with spacing errors.

Recently, much effort has been devoted to correcting such word spacing errors. Many of such proposals use various heuristics or correct word spacing errors during preprocessing before morpheme analysis is executed. However, if word spacing corrections are made through heuristics, it is difficult to handle every single error among the countless mistakes the writer makes. In other words, corrections are restricted to the common mistakes made by a relatively large number of people [1, 3].

Our previous morpheme analysis module KorMa2000 [5] generated a list of candidate morphemes and then formed a final list of the most appropriate morphemes according to the probability of joints within or between phrasal units (*eojeols*—Korean phrasal units composed of one or more words) according to the equations:

* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the Chung-Ang University HNRC-ITRC (Home Network Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 198 – 201, 2005.

$$\mathbf{P}\tau(P) \approx \prod \mathbf{P}\tau(P_i \mid P_{i-1}),$$
$$\mathbf{P}\tau(W \mid P) \approx \prod \mathbf{P}\tau(W_i \mid P_i),$$
$$p' = \arg\max \prod \mathbf{P}\tau(P_i \mid P_{i-1}) \prod \mathbf{P}\tau(W_i \mid P_i) = \arg\max \sum \log \mathbf{P}\tau(P_i \mid P_{i-1}) + \prod \mathbf{P}\tau(W_i \mid P_i).$$

Such analysis-through-generation architecture (successfully applied by other authors to inflective languages [1, 2]) made our Korean analysis module very sensitive to word spacing. In this paper we describe a newly revised morpheme analysis module, which applies a custom user dictionary for newly coined words, special terminology, and automatic word spacing. With this dictionary, the user can classify special words such as compound nouns and words adopted from foreign language as correct words beforehand in order to reduce errors in analysis of certain words.

2 The Korean Morpheme Analysis Module KorMA 2003

The design of the Korean morpheme analysis module KorMa2003 is shown in Fig. 1.

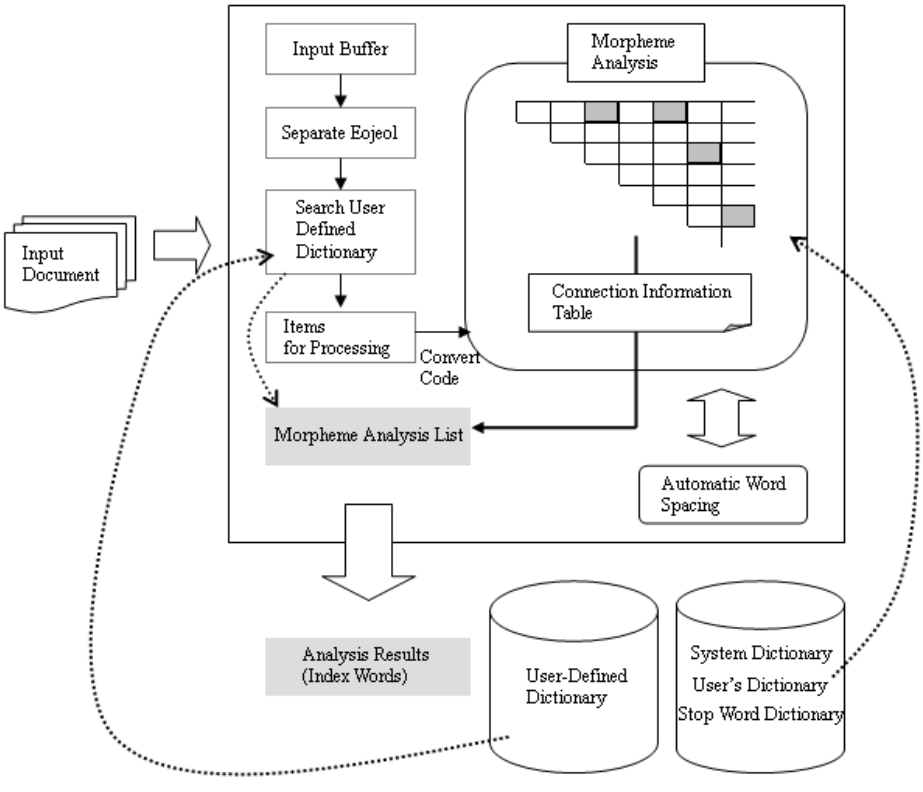

Fig. 1. Design of User-Defined Dictionary and Automatic Spacing Suppor

2.1 User-Defined Dictionary

Among the many kinds of dictionaries, the present study applies the user-defined dictionary to allow the user to define words themselves. We use to enhance accuracy in analyzing certain words or newly coined words. Although there are a variety of ways to deal with unregistered words, newly coined words, special terminology, etc., the most widely used methods are the user dictionary and special processing of suffixes and postpositions. However, these methods are insufficient in handling the matter. Table 1 exemplifies the difficulties in analyzing unregistered and special words using the former morpheme analysis module [5].

Table 1. Incorrect Results of Analysis of Unregistered Words and Special Words by the Former Morpheme Analysis Module

Compound Nouns	▪ 한국(Korea)사회(social)보장(security)제도(policy) *vs.* 한국(Korea)/사회보장제(social security system)/도(island) ▪ 한국(Korea)사회(Social)보장(Security)법(Law)론(discourse) *vs.* 한국사(Korean History)/회보(Bulletin)/장법/(law against stolen goods)/론(theory)
Unknown Words	▪ 공동정범(common/criminal) *vs.* (ball)/동정(sympathy,virginity)/범(tiger) *공동/common , 정범/crime, criminal
Special Words	▪ 뇌사자 *vs.* 뇌(brain)/사자(lion) *뇌 brain/사 dead /자 person ▪ 대어음 *vs.* 대어(big fish)/음(negative) *대 fictitious/ 어음 bill

2.2 Automatic Word Spacing

The automatic word spacing checking function looks for incorrect spacing between each syllable of an incorrect phrase using right-to-left and left-to-right search and the longest and shortest match methods [5]. Additionally, a system has been implemented to correct any mistakes in spacing, for both space insertion and space omission errors.

3 Experimental Results

We used three document collections for our tests. Table 2 and Figure 3 show the comparison of the number of index words and correct words extracted by each system.

Table 2. Percentage of Correct Words

Text collection	Correct set	KorMa2003	KorMa2000
Chamber of Commerce (unregistered)	20132	97.9%	83.6%
Military Terms (special words)	6891	97.9%	83.0%
Court (special words)	4958	97.9%	84.3%

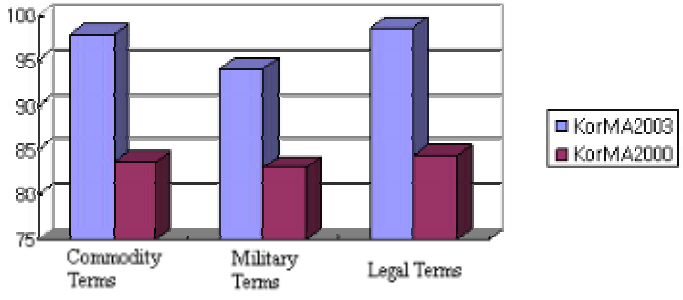

Fig. 2. Percentage of correct words after application of user-defined dictionary

4 Conclusion

To prevent errors in analysis of grammatically correct sentences as well as sentences with incorrect word spacing, we implemented an independent word spacing function for making spacing corrections when an analysis error has occurred after morpheme analysis. The Korean morpheme analysis module KorMa2003 is 15% to 17% more accurate than the former module supporting only user dictionary and post-processing (presuming pure Korean words, postposition processing).

The described morpheme analysis module—an indexing module for information retrieval systems—can be used more efficiently in cases such as documents of special fields using special terms, such as National Assembly legislation proposals that have no spaces between words, or Internet message boards where incorrect spacing occurs frequently, especially when a massive amount of documents must be indexed.

For the future work, for efficient operation of the new morpheme analysis module in indexing a vast amount of documents a method for automation should be developed to efficiently implement the user correct word set.

References

1. A. Gelbukh, G. Sidorov. Morphological Analysis of Inflective Languages through Generation. *Procesamiento de Lenguaje Natural*, No 29, Spain, p. 105–112.
2. A. Gelbukh, G. Sidorov. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. In: A. Gelbukh (ed.), Computational Linguistics and Intelligent Text Processing (CICLing-2003). Lecture Notes in Computer Science, N 2588, Springer-Verlag, pp. 215–220.
3. DoSam Hwang, KiSun Choi, and TaeSuk Kim, *Natural Language Processing*.
4. JongKeun Kwak, J.J. Eun, Y.S. Kang. Structure and Characteristics of LGKNA. In: *1st workshop on Evaluation of Morpheme Analysis and Tagging*.
5. SungJin Lee, DukBong Kim, JungYun Seo, KiSun Choi, and GilChang Kim, Construction and Analysis of Korean Morpheme based on two-level model. In: Proc. of KISS Fall Conference, Volume 19.