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).