Abstract

Telegraphic messages with numerous instances of omission pose a new challenge to parsing in that a sentence with omission causes a higher degree of ambiguity than a sentence without omission. Misparsing induced by omissions has a far-reaching consequence in machine translation. Namely, a misparse of the input often leads to a translation into the target language that has incoherent meaning in the given context. This is more frequently the case if the structures of the source and target languages are quite different, as in English and Korean. Thus, the question of how we parse telegraphic messages accurately and efficiently becomes a critical issue in machine translation. In this paper, we describe a technical solution for the issue, and present the performance evaluation of a machine translation system on telegraphic messages before and after adopting the proposed solution. The solution lies in a grammar design in which lexicalized grammar rules defined in terms of semantic categories and syntactic rules defined in terms of part-of-speech are utilized together. The proposed grammar achieves a higher parsing coverage without increasing the amount of ambiguity/misparsing when compared with a purely lexicalized semantic grammar, and achieves a lower degree of ambiguity/misparses without decreasing the parsing coverage when compared with a purely syntactic grammar.