Sentence Segmentation Model to Improve Tree Annotation Tool

So-Young Park* Dongha Shin* and Ui-Sung Song**

*College of Computer Software & Media Technology, SangMyung University,
7 Hongji-dong, Jongno-gu, SEOUL, 110-743, KOREA
ssoya@smu.ac.kr, dshin@smu.ac.kr

**Dept. of Computer Science & Engineering, Korea University,
5-ka 1, Anam-dong, Seongbuk-ku, SEOUL, 136-701, KOREA
ussong@disys.korea.ac.kr

Abstract. In this paper, we propose a sentence segmentation model for a semi-automatic tree annotation tool using a parsing model. For the purpose of improving both parsing performance and parsing complexity without any modification of the parsing model, the tree annotation tool performs two-phase parsing for the intra-structure of each segment and the inter-structure of the segments after segmenting a sentence. Experimental results show that it can reduce manual effort about 28.3% by the proposed sentence segmentation model because an annotator's intervention related to cancellation and reconstruction remarkably decrease.

1 Introduction

A treebank is a corpus annotated with syntactic information. In order to reduce manual effort for building a treebank by decreasing the frequency of the human annotators' intervention, several approaches have tried to assign an unambiguous partial syntactic structure to a segment of each sentence. The approaches [1, 2] utilize the reliable heuristic rules written by the grammarians. However, it is too difficult to modify the heuristic rules, and to change the features used for constructing the heuristic rules [3]. One the other hand, the approaches [3, 4] use the rules which are automatically extracted from an already built treebank. Nevertheless, they place a limit on the manual effort reduction and the annotating efficiency improvement because the extracted rules are less credible than the heuristics.

In this paper, we propose a tree annotation tool using an automatic full parsing model for the purpose of shifting the responsibility of extracting the reliable syntactic rules to the parsing model. In order to improve both parsing performance and parsing complexity without any modification of the parsing model, it utilizes a sentence segmentation model so that it performs two-phase parsing for the intra-structure of each segment and the inter-structure of the segments after segmenting a sentence. Next, section 2 will describe the proposed sentence segmentation model for the tree annotation tool, and section 3 shows the experimental results. Finally, we conclude this paper in section 4.