A Semi-Automatic Tree Annotating Workbench For Building A Korean Treebank

Joon-Ho Lim, So-Young Park, Yong-Jae Kwak, and Hae-Chang Rim

Department of Computer Science & Engineering Korea University
5-ka, Anam-dong, SEOUL, 136-701, KOREA
{jhlim, ssoya, yjkwak, rim}@nlp.korea.ac.kr

Abstract. In this paper, we propose a semi-automatic tree annotating workbench for building a Korean treebank. Generally, building a treebank requires an enormous effort by the annotator. In order to improve annotating efficiency, decrease the number of intervention required by the annotator, and help maintain consistent annotation in building a treebank, we have developed a semi-automatic tree annotating workbench consisting of following three stages: syntactic pattern extraction, syntactic pattern selection, and syntactic pattern application. The experiment was carried out with 27,966 tree tagged sentences as a training set and 3,108 sentences as a test set. As a result, the burden of manual annotation can be reduced by about 47% with the best selection of the feature set by using the proposed tree annotating workbench.

1 Introduction

A syntactically annotated treebank is a highly useful language resource which can represent syntactic information with the tree structures of the given sentences. However, in building a treebank, a vast amount of time and effort is required by the annotator. Furthermore, maintaining the consistency of the constructed treebank is difficult if the annotation is performed only manually ([4]). Therefore, we need a tree annotating workbench which can improve the annotating efficiency, decrease the number of annotator's intervention, and help maintain consistent annotation in constructing a treebank.

Some approaches to tree annotating workbench, such as [1] for PennTreeBank([4]), [2] for STEP2000([3]), and [8], have been previously developed. In [1] and [2], the heuristic rules written by annotators are used for deterministically attaching a partial syntactic structure. However, the workbenches are limited in terms of efficiency improvement and the manual work reduction. In [8], the limited part of speech tag sequences are extracted from the previously constructed treebank, and the extracted POS sequences are used in building a new treebank. However, this workbench does not allow the annotator to select the context size or the context information.

In this paper, we propose a semi-automatic tree annotating workbench for building a Korean treebank. It extracts various syntactic patterns from the previously constructed treebank based on the selected features, and It automatically applies the extracted syntactic patterns to the appropriate states.