Evaluating the Performance of the Survey Parser with the NIST Scheme

Alex Chengyu Fang

Department of Chinese, Translation and Linguistics
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
acfang@cityu.edu.hk

Abstract. Different metrics have been proposed for the estimation of how good a parser-produced syntactic tree is when judged by a correct tree from the treebank. The emphasis of measurement has been on the number of correct constituents in terms of constituent labels and bracketing accuracy. This article proposes the use of the NIST scheme as a better alternative for the evaluation of parser output in terms of *correct match*, *substitution*, *deletion*, and *insertion*. It describes an experiment to measure the performance of the Survey Parser that was used to complete the syntactic annotation of the International Corpus of English. This article will finally report empirical scores for the performance of the parser and outline some future research.

1 Introduction

Different metrics have been proposed and all aim at the estimation of how good a parse tree is when judged by a correct tree from the Treebank (see [1], [2], [3], [4], and [5]). The emphasis of measurement has been on the number of correct constituents either in terms of constituent labels, such as *labelled match*, *precision*, and *recall*, or in terms of bracketing such as *bracketed match*. Together with *crossing brackets*, these measures indicate the number of correct and wrong matches in the parse tree. However, these measures outlined above do not constitute a satisfactory assessment. We may well imagine a parse tree with only two correct constituents scoring a high rate in terms of labelled and bracketed matches, crossing brackets, precision, and recall while deletions and insertions of nodes and associated labels could render the parse tree totally different from the correct one.

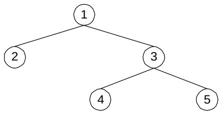

Fig. 1. *A correct tree*

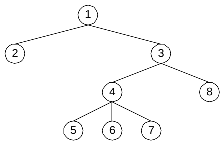

Fig. 2. *A parser-produced tree*