Automatic Conversion of Constituency Trees into Dependency Trees or Manual Annotation?

Katalin Ilona Simkó[1], Veronika Vincze[1,2], Zsolt Szántó[1], Richárd Farkas[1]

[1]University of Szeged, Department of Informatics
Szeged, Árpád tér 2.
kata.simko@gmail.com, szantozs@inf.u-szeged.hu, rfarkas@inf.u-szeged.hu
[2]MTA-SZTE Research Group on Artificial Intelligence
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Nowadays, two popular approaches to data-driven syntactic parsing are based on constituency grammar on the one hand and dependency grammar on the other hand. Hungarian is one of those rare examples where there exist manual annotations for both constituency and dependency syntax on the same bunch of texts, the Szeged (Dependency) Treebank, which makes it possible to evaluate the quality of a rule-based automatic conversion from constituency to dependency trees, to compare the two sets of manual annotations and also the output of constituency and dependency parsers trained on converted and gold standard dependency trees.

We investigate the effect of automatic conversions related to the two parsing paradigms as well. It is well known that for English, the automatic conversion of a constituency parser's output to dependency format can achieve competitive unlabeled attachment scores (ULA) to a dependency parser's output trained on automatically converted data. One of the possible explanations for this is that English is a configurational language, hence constituency parsers have advantages over dependency parsers here. We check whether this hypothesis holds for Hungarian too, which is the prototype of free word order languages.

In this paper, we compare three pairs of dependency analyses in order to evaluate the usefulness of converted trees. First, we examine the errors of the conversion itself by comparing the converted dependency trees with the manually annotated gold standard ones. Second, we argue for the importance of training parsers on gold standard trees by looking at the typical differences between the outputs of dependency parsers trained on converted (silver standard) trees, parsers trained on gold standard trees and the manual annotation itself. Third, we demonstrate that similar to English, training on a constituency treebank and converting the results to dependency format can achieve similar results in terms of ULA to the dependency parser trained on the automatically converted treebank, but the typical errors they make differ in both cases.

We present the details of the results achieved by different parsing methods as well as a linguistic analysis and categorization of the types of errors they made. For instance, analysing multiword names seems to be easier for the constituency parser, while the dependency parser is better at finding the arguments of verbs.