Halabi D., Fayyoumi E., Awajan, A. 2022. I3rab: A new Arabic Dependency Treebank Based on Arabic Grammatical Theory. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Volume 21, Issue 2, March 2022, Article No.: 23, pp 1–32, https://doi.org/10.1145/3472295, (ISI, Scopus)

Abstract

Treebanks are valuable linguistic resources that include the syntactic structure of a language sentence in addition to part-of-speech tags and morphological features. They are mainly utilized in modeling statistical parsers. Although the statistical natural language parser has recently become more accurate for languages such as English, those for the Arabic language still have low accuracy. The purpose of this article is to construct a new Arabic dependency treebank based on the traditional Arabic grammatical theory and the characteristics of the Arabic language, to investigate their effects on the accuracy of statistical parsers. The proposed Arabic dependency treebank, called I3rab, contrasts with existing Arabic dependency treebanks in two main concepts. The first concept is the approach of determining the main word of the sentence, and the second concept is the representation of the joined and covert pronouns. To evaluate I3rab, we compared its performance against a subset of Prague Arabic Dependency Treebank that shares a comparable level of details. The conducted experiments show that the percentage improvement reached up to 10.24% in UAS and 18.42% in LAS.