Abstract

A lot of work was done in the statistical parsing area aiming to improve the parsing process. Most of this work was done on English language and used by applications like machine translation, speech recognition and others for English. Little work has been done on statistical parsing for Arabic. The motivation behind the work in this thesis is to develop a statistical Arabic parser.

The objective is to investigate the development of an Arabic statistical parser using Arabic Treebank and a statistical parsing engine. We investigated and searched for a suitable parser that can handle and deal with the special features of Arabic and also that can handle the selected treebank. We selected the Bikel Parsing Engine, as it satisfied our requirement, and the LDC Arabic treebank.

The different steps followed to develop and test the parser have been described. These steps include dividing the LDC2005T20 Arabic Treebank into training and testing sets. 90 % of the treebank was used to train the Bikel parser package while 10% of it was randomly selected to test the developed parser. The testing data set annotations were removed to convert it into pure text to be introduced to the trained parser. The gold testing data set was prepared, by mapping its tags, to the tags produced by the trained parser. This mapping was necessary to evaluate the parser results using a standard evaluation tool. The metrics widely applied for parsers evaluation were computed for the developed parser results. The evaluation metrics of the developed parser were comparable to evaluation metrics results of well known English parsers. The Developed Arabic parser achieved 84.6% labeled precision, 82.74% labeled recall and 99.11% POS tagging accuracy on the selected test set.

In addition, we built an Editor for the annotated sentences in order to be used to convert it into a reference or a treebank annotation.

Table of Contents

Table of Abbreviations

CB	Cross Brackets
CCG	Combinatory Categorial Grammar.
CYK	Cocke-Younger-Kasami algorithm, also Known as (CKY).
HBG	History-Based Grammar.
LP	Labeled Precision
LR	Labeled Recall
PCFG	Probabilistic Context Free Grammar.
POS	Part Of Speech.
TAG	Tree-Adjoining Grammar.
WSJ	Wall Street Journal.