Case based Arabic Morphological Analysis with Compositional Non-deterministic Automata

Hamza Harkous **Jad Makhlouta** **Fadi Zaraket**
American University of Beirut
{hhh20, jem04, fz11}@aub.edu.lb

Abstract

Morphological analysis is key in current automated analysis techniques for Arabic text. Current morphological analyzers take an Arabic word as input and enumerate all possible morphological solutions (stems or derivations) using concatenation based analysis. The solutions suffer from accuracy due to the inherent difficulties of morphological analysis of the Arabic language. The enumeration of all possible solutions may also hurt the performance in cases where not all solutions are needed. In this paper, we present a novel efficient morphological analyzer that uses parallel compositional non-deterministic Automata driven by a case based controller. The analyzer keeps alive possible analyses of the text for the controller to intervene with a decision and thus reduce false positives on case basis.

Our concatenative analysis uses recursive affixes in order to retain better part of speech information and enhance the efficiency of affix matching. The analyzer takes as input a character from a full text stream rather than a sequence of words and thus performs tokenization on the fly and based on morphological correctness. This relieves the user from preprocessing the text and provides a more exact tokenization than delimiter based tokenization. This is important for Arabic since two words can occur in a text without a delimiter in between.

We used our analyzer to successfully automate the analysis of three books of Islamic literature where we segmented each book into several narrations. In each narration, we detected a chain of narrators, where each narrator is a complex sequence of proper names separated with name connector phrases. This is a collection of chain of narrator structures where each chain is a structure with three levels of hierarchy. Our results show better accuracy and higher efficiency than current morphological analyzers.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.