Morphological Analysis of Arabic Words Using Parsing

Majed AbuSafiya#1, Mohammed Awad *2
# 1 Software Engineering Department
Al-Ahliyya Amman University, Amman, Jordan
majedabusafiya@gmail.com
*2 Dept. of Computer Science and Engineering
American University of Ras Al Khaimah, Ras Al Khaimah, UAE
mohammed.awad@aurak.ac.ae

Abstract— **In this paper, a new approach for Arabic language morphological analysis is presented. Unlike other commonly used approaches that are based on pattern matching, we propose defining the structure of Arabic language words using context free grammar (instead of regular expressions). Thus, parsers can be generated to do morphological analysis. This approach considers the annotation marks that are usually neglected by current morphological analyzers, which results in a more detailed analysis.**

Keywords: morphological analysis, context free grammar, parsing, Arabic language

I. Introduction

Morphological Analysis is the process of recognizing the internal structure of the natural language words [1]. For example, the word يُقاتِل (fights) can be viewed as a present four-letter verb (فعل مضارع رباعي). Morphological analysis is one important step towards natural language understanding and processing. Morphological analysis of Arabic languages is a rather complex process. The reason is the richness of Arabic words in prefixes, suffixes and the high multiplicity of words that can be derived from the same root. For example the word (يُقاتِلون) has prefix (ي) and suffix (ون) and the verb (يُقاتِل) that is derived from the root (قَتَل). Another complexity of Arabic morphological analysis is that the meaning of a word varies according to the annotation signs used (e.g. fatha, dhamma). For example, we can see a clear difference between a word like (قاتِلْ) and (قاتَل) although they seem to have the same internal structure.

One of the most known Arabic Morphological Analyzers is Buck Walter Arabic Morphological Analyzer [2]. This analyzer is based on pattern matching. The main idea of this analyzer is to try to transliterate the Arabic word into a Latin word. This is carried out through a pattern matching were every Arabic letter is mapped into a corresponding Latin letter using a special table. For example the word (يُقاتِلون) is transliterated into the word (yqAtlwn). Once found, this transliteration word is used as a key to look up in a huge table that contains the details about every possible transliteration.

This approach has two main shortcomings. First, an entry (in the look up table) is required for every possible transliteration. So words with similar structures that belong to different roots map to two different entries. For example, although the two words (يقاتلون , يسافرون) both have the same structure, yet they are transliterated to different keys (yqAtlwn, ysAfrwn). This means that look up table has to have an entry for every possible transliteration even if words have the same structure, which means a rather large look up table (not to mention that the richness of the Arabic language set of roots). Second, neglecting the annotation signs causes a loss of important information where the transliteration will map into the same entry in the look up table. For example both (قاتِلْ) and (قاتَل) will be transliterated into (qAtl) which will map to the same entry in the look up table.

II. Proposed Solution

Our solution for morphological analysis is based on using more complex structures to capture the structure of Arabic words than pattern matching using regular expressions. We propose using context free grammar [3] to capture the structure of Arabic words. A given Arabic word is assumed to belong to one structure. For example, all the four letter present verbs of the form fa'al (فاعَل) for absent multiple male subject (e.g. يُقاتِلون) can be represented by the context free grammar rules shown in Fig. 1.

```
X →  ‘ن’ ‘و’ L ‘ ◌ِ ’ L ‘ا’ L  ‘ ◌ُ ’ ‘ي’
L → ‘ا’ | ‘ب’ | ‘ت’ | ‘ث’ | ……| ‘ي’
```

Fig. 1. Example of context free grammar rules

Note that what is captured in the rule is the structure regardless of the root to which the word belongs (e.g. L can be substituted by any letter of the Arabic alphabet). Note also that the annotation signs are taken into consideration. Once the context free grammar rules of the possible structures are defined, parsers can be built to identify the internal structure of a given word. The main advantage of defining context free grammar (to define the word structure) is that it allows us to define a detailed structure and better morphological analysis for a wide set of words that follow the same structure. That is, it allows us to define structures for Arabic words such that words of different roots but belong to the same structure can be identified by a single rule. It also allows us to consider annotation signs in the morphological analysis process.

From implementation point of view, the availability of automated parser generators [4, 5] simplifies the effort of building the morphological analyzer. These tools can be used to automatically generate the parsers by only defining the rules of the different possible structures of the Arabic language words.

III. Conclusion

In this paper we have proposed a new approach for morphological analysis for Arabic words. It is based on defining context free grammar rules for the different internal structures of Arabic words and generating parsers based on these rules. A context free grammar rule defines a structure that is followed by a set of words that follow that structure while these words may belong to wide range of roots. However, in other approaches a pattern matching is required to recognize a word (and may be a small set of words) that belongs to the same root, which means a large lookup table. Also, a context free grammar allows us to define more complex structures (which include annotation signs) and hence better and more detailed morphological analysis than using pattern matching that is based on regular expressions. As future work, a quantitative comparison is needed to show how the proposed approach may outperform earlier approaches in terms of time and accuracy.

References

[1] C. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, Massachusetts. 1999.
[2] Lingusitic Data Consortium. Buckwalter Arabic Morphological Analyzer. http://www.ldc.upenn.edu/
[3] J. Hopcroft, J. Ullman, D. Jeffrey. Introduction to Automata Theory. Languages, and Computation. Addison-Wesley. 1979. pp. 77–106.
[4] T. Parr. The definitive ANTLR reference. Building Domain Specific Languages. 2007.
[5] D. Brown, J. Levine, T. Mason. Lex & Yacc. O'Rielly. 1995.

AUTHOR PROFILE

Majed AbuSafiya earned his PhD in Computer Science from the New Mexico Tech in the United States in 2008. He earned a MSc in Computer Science from the Middle East Technical University in Turkey. Dr. AbuSafiya's' research interests include document engineering and business process management.

Mohammed Awad earned his PhD in Computer Science from the University of Houston in the United States in 2011. He earned a MSc in Computer Science at the same university in 2006. Dr. Awad's research interests include security, more specifically in E-voting and I-voting security. An additional area of interest concerns safeguarding the transmission of biometric data and integrating captured biometric data (iris scans or any other biometric data) into the electoral process in order to achieve more reliable systems.