IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 10, October 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

Towards a Context-Free Grammar for Arabic Words

Zeinab E. E. Mahmoud[1] and Hamdah A. Alotaibi[2]

[1] Assistant Professor, Faculty of Computing and Information Technology/ King Abdulaziz University/ Jeddah/ Saudi Arabia
Zmahmoud@kau.edu.sa

[2] Faculty of Computing and Information Technology/ King Abdulaziz University/ Jeddah/ Saudi Arabia
Halotaibi0136@stu.kau.edu.sa

Abstract

Solving the language related problems is the primary concern of the field known as natural language processing. Machine translation, grammar checking, information retrieval and Spelling correction are some applications of language processing. The Arabic language is the fifth most widely spoken language in the world and it the official language in 26 countries. For such reasons most of the research take care of it and design a morphological, syntactic and semantic analysis of it. Lexical semantics is the knowledge of meanings of the components words. The procedure of defining the right meaning of a specific word called word sense disambiguation. More than one solution was initiate to have the correct meaning for a particular word. In this article, an Arabic parser through using Context-Free-Grammar is presented for a word level to reduce the word sense ambiguation. A recommendation should be take in consideration after trying the suggested solution.

Keywords: *parsing, natural language processing, Arabic language, context-free grammar.*

1. Introduction

Natural language processing abbreviate as NLP refers to the computer being able to process in a natural language like English as opposed to a computer programming language. NLP allows computers to read and understand the information we input into them. NLP is essential because it helps in managing the vast and massive information in the current time and turn it into useful data. Most applications of NLP understand text using part of speech tagging (going through a sentence and labeling each word as noun, verb, adjective). There are three fundamental components in NLP: lexicon, syntax, and morphology [1].

The Arabic language is a beautiful language that is hugely derivational and complex because of its fantastic structure. Even if there are many articles on designing English parser The amount of Arabic-based parser studies and tools are still limited.

There are three exciting stages in Natural Language Processing: Morphological, Syntactic and Semantic analysis. The morphological analysis takes care of word structure (the different forms of the root word). The lexicon is an essential part of any natural language processing analysis. It can be considered as a dictionary that lists all the words clearly. The syntactic analysis takes care of the building of the sentence by both (grammar and lexicon). The semantic analysis takes care of the word meaning. Typically by taking each word in the phrase individually and extract the possible definitions for it. Good to mention that most of the studies integrate both syntactic and semantic analysis in one stage. Context-Free-Grammar is a formal grammar which is used to present the structure of a sentence. It is used in this article because it shows promising results in many research papers. A parse tree can be seen as a graphical representation of a derivation.

Although there are a well amount of research papers that discuss the Arabic language processing, this amount cannot be compared to some research papers that are interested in the English language for example. For this reason, more investigation should be applied to Arabic language processing. Also, most of the research papers are focused on the "Arabic sentence" either in morphological and parser analysis. An exciting property in the Arabic language is that some word without being in a sentence can have more than one meaning. In this paper, a proposed parser is presented for definition the words that carry multiple meaning based on diacritics trough Context-Free Grammar.

This research paper organized as follows: related work is shown in section two, an overview of Arabic language is presented in section three, section four explain the Context-Free Grammar, part five represent the proposed solution. Finally, the conclusion will take place in part six.

2. Related Work

It is interesting to see all the research papers that have been published and also the ones that are still working on it. In this section we will try our best to cover most of these papers:

In [1] the authors describe their experience in developing a parser for modern standard Arabic. A definite clause grammar was used for the parser to be a part of a machine translation system. The procedure is done in two phases: the first phase aims to have the rules that create a grammar for Arabic. The second phase is merely applying the parser that gives grammatical structure on input sentence. The structure of the system is described in figure 4. Their grammar covers the simple sentence (that not connected to another sentence) and a compound sentence (that is more than sentence connected to each other). Then, they categorize the simple statements into nominal, verbal and special sentences. The authors used the Definite Clause Grammer (DCG) because it balances between computational efficiency and linguistic expression and the flexibility in syntax and semantic integration. Finally, it allows writing the grammar rules directly in Prolog programming language. Good to mention that authors work with the issues that could appear because of the language ambiguity and they emphasized that ambiguity can be semantic. They test the parser at the end on 212 sentences. 146 sentences were parsed correctly, 44 were incorrect, and 22 sentences were not parsed.

In [2] the authors record their experience in designing a chart parser for analyzing modern standard Arabic sentences. They did find that the parser successfully reduces the ambiguity with the lexical semantic features. They used the prolog also for implementing both the parser and the analyzer. The system description is shown in figure 5. Their morphological analyzer was made up of three units: analyzer, lexical disambiguation, and feature extraction unit. As it is known, the lexicon is used to figure out the word categories. It has two kinds of features: syntactic to remove the syntactic ambiguity and lexical to remove lexical ambiguity. The unification-based grammar was used to describe the Arabic grammar rule. It consists of 170 rules that divided to 22 groups of rules which define the different grammatical category.

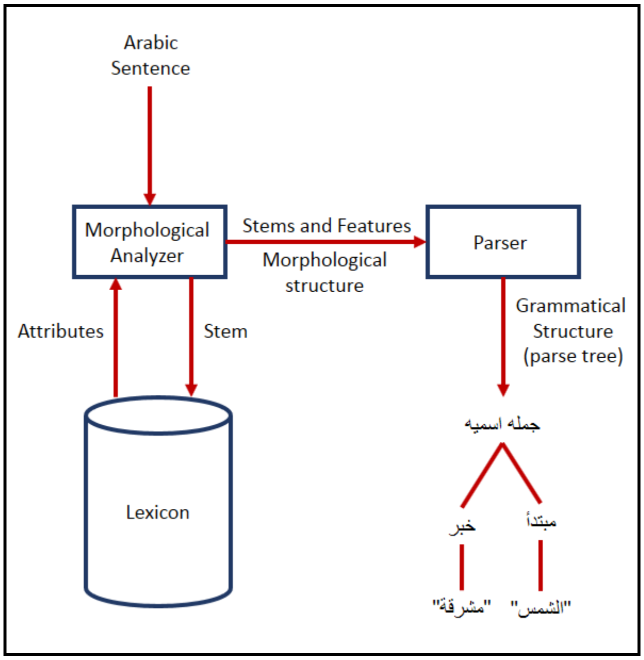


Figure 1

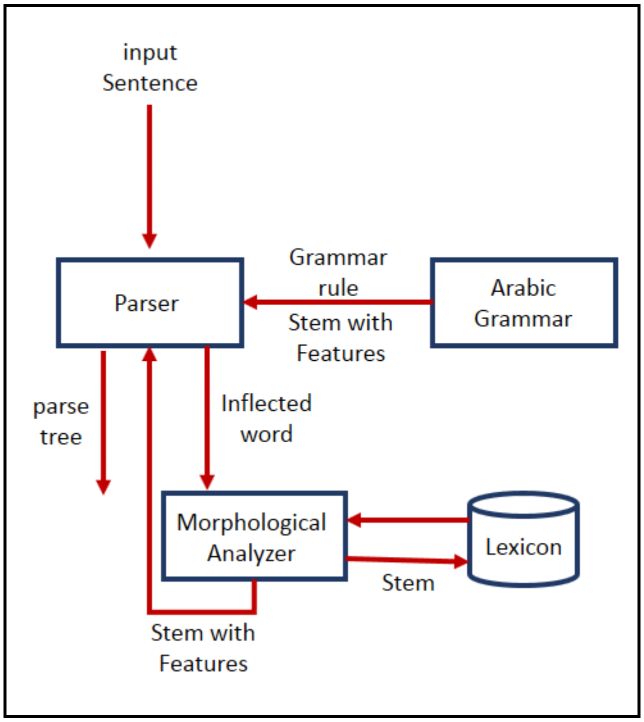


Figure 2

In [3] the authors were encouraged to search about the lexical functional grammar because they did not find many resources that speak about it in depth. They describe the background and the architecture (constituent-structure, functional-structure) of the lexical functional grammar and how it was used to solve the NLP problems like ambiguity, defining clause grammar and parsing Arabic sentence (here they had a training set and testing set). Their work can be considered as an interesting review of many research articles.

In [4] the author aimed to display a modest parser for Arabic sentences and determine the correctness of it. They use the Context Free Grammar with a top-down procedure. The result was pleasing because the accuracy average was 95% when they test their work on 150 Arabic sentences through the NLTK toolkit.

In [5] the authors present a technique to label the structure of a sentence in modern standard Arabic. This way was to assume a hypothesis and to test it on actual data. Like most of the authors, here the sentence were two types: nominal and constituents (for example adverb, adjective and so forth).

In [6] the authors design a system that takes care of both the structure and syntactic functions in the Arabic language to reduce ambiguities. Their work can be considered in three steps: morphological analysis, extracting phrases by using Context Free Grammar and using unification grammar to collect phrases and getting the final result.

In [7] the authors present an Arabic parsing that consists of two sections: lexical and syntax analysis. The lexical analysis takes care of the words (find the root for it and assigns it to the proper tokens). Syntax analysis receives the token and sees the best grammar for all the tokens by using the Context Free Grammar.

In [8] the authors develop a top-down technique chart parser for parsing the Arabic language. They represent the Arabic grammar by using Context Free Grammar. The work will be done like the following: the grammar will give an accurate description of the sentence. Then the parser will allocate the structure to the input sentence.

3. Arabic language

The Arabic language is the fifth most widely spoken language in the world with 293 million native speakers and 422 million speakers in total. It is an official language in 26 countries (that does not mean it is the majority language in all of those countries, but it is one of the official languages). It is also one of the six official languages of the united nations. Nice to mention that the Arabic language is the language of Quran (the holy book of Islam) and the liturgical of 1.7 billion Muslims around the world (most of those people do not speak Arabic, but many have some knowledge of Arabic for reading and for reciting prayers and religious study).

Speaking about Arabic can be confusing because there are many different varieties of the language. One of the main varieties is the classical Arabic of the Quran; this is considered by many to be a perfect form of Arabic because it is the language in which God revealed the Quran to the prophet Mohammed peace be upon him. Then there is modern standard Arabic which is the form of Arabic used as an official language today. It is the modern form of literary Arabic which was based on the classical Arabic of the Quran but with some adaptations and a greatly expanded vocabulary to make it more suitable for modern times. It is precisely the same as classical Arabic, but both of them are referred to by Arabs as Al-Fusha "الفصحى" meaning "eloquent speech." Modern standard Arabic is the language of books, media, education and formal situations, but not as the language of everyday speech. For everyday speech, Arabic speakers use their local dialects or ('amiya) "العاميه" which can differ quite significantly from country to country and even from one place to another within a single country.

Arabic is well known for its state of diglossia, Arabic speakers used two distinctly different forms of the language in parallel for various purpose: modern standard Arabic that it is not learned by anyone as a native language, but it is used in reading and writing in media, on children's TV shows, and informal speeches while the dialects are used almost universally for daily conversations.

Let's take a look at some script features of modern standard Arabic language:

- The Arabic script is written from right and consists of letters that imitate handwriting.
- Most letters join to the letter that comes after them. However, a few letters remain disjoint (see figure 3).
- The letters that join have two forms, one short form at the beginning or in the middle of the

words and another long form at the end of words or when the letter is byself (see figure 4).

- The Arabic script is an abjad "أبجد" meaning that each letter represents a consonant and that short vowels are not really and that long vowels can be ambiguous (see figure 5).

An interesting question could appear here about how can we read Arabic without vowels? Will if we can read this word (see figure 5) we can conclude that the short vowels are not written, and the others seem somewhat incomplete, but we have a hint about what the vowels are.

Arabic has more predictable vowel patterns than English, so it is easier to guess, also Arabic can be written with (Harakat) "حركات" which extra diacritic markings that indicate the short vowel sounds, these are generally only used in texts that are really important to pronounce correctly like the Quran or poetry or children's materials.

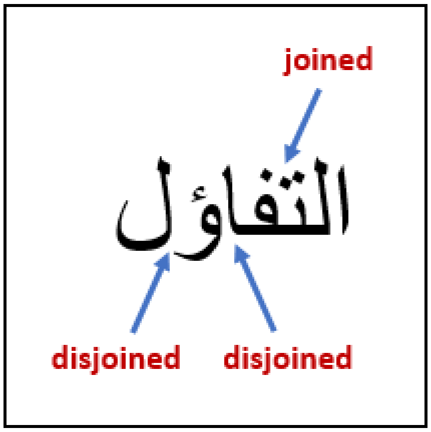


Figure 3

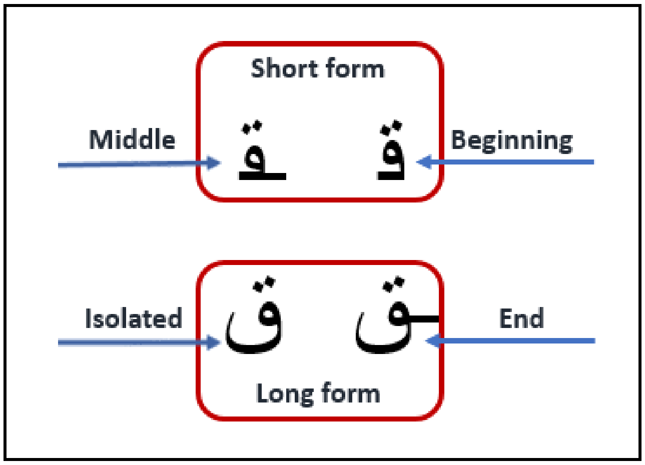


Figure 4

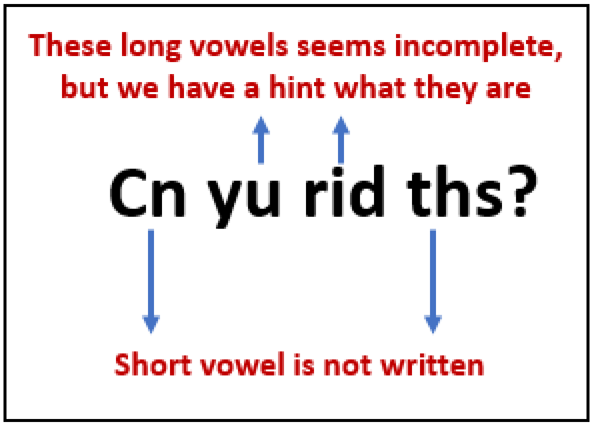

These long vowels seems incomplete, but we have a hint what they are
Cn yu rid ths?
Short vowel is not written

Figure 5

4. Context-Free Grammar and Parse Tree

Context-free grammar has a central importance in both theoretical computer science and programming languages. Also, it is a very important formalism used within the parsing problem. A context-free grammar has the following four tuples as G = {V, Σ, S, P} where:

- V represents a set of variables or non-terminal symbols.
- Σ represents a set of terminal symbols.
- S represents the start symbol (one of the non-terminal symbols).
- P is a set of production rules. Each rule has:
  - Left part (non-terminal).
  - Arrow.
  - Right part (consist of non-terminal symbols).

Parse tree shows how the start symbol of grammar can drive a string in the language. A parse tree has the following properties:

- The root is the start symbol.
- Each internal node is a non-terminal.
- Each leaf is a token or ϵ.

Parsing is known as the process of determining if a string of tokens can be generated by a grammar. So, it determines that the sentence is constructed in correct way according to the grammar rules. It is essential in many applications like grammar checkers, information extraction, and machine translation.

5. Proposed solution

First, it is good to explain the difference between Lexical, Syntactic and Semantic analysis. Lexical analysis reads the input word as a character stream (character by character) and produces a stream of lexemes and token strings as an output. It will search the word in a particular table. Syntactic analysis (parser) directly takes the output from the lexical analyzer (lexeme and tokens) and produce the parse tree. The semantic analysis determines if the meaning is respected. Semantic analysis is considered as a part of the syntactic analysis. Each of these analyzers has particular kind of problems that are related to its phase.

As mention in the previous sections, most of the research papers present morphological, syntactic and semantic analysis on a sentence level (bear in mind that most of the authors make the syntactic and semantic in one stage). In this article, we will consider applying the syntactic and semantic stage for a word level, not a sentence level. In most of the languages, the process of determining the correct meaning of an individual word is done by adding the context information to each word in the lexicon. Respectively every word in a sentence can help in knowing the meaning of other words in the same sentence. This procedure is not needed in our case (when dealing on a word level) because with the use of diacritics (vowel) in written Arabic "التشكيل" we can determine the correct meaning of the words that have the same structure.

Diacritics can be considered as zero-width characters (see figure 6). It is optional in normal text expect in holy Quran. It is used to give the correct meaning for any word.

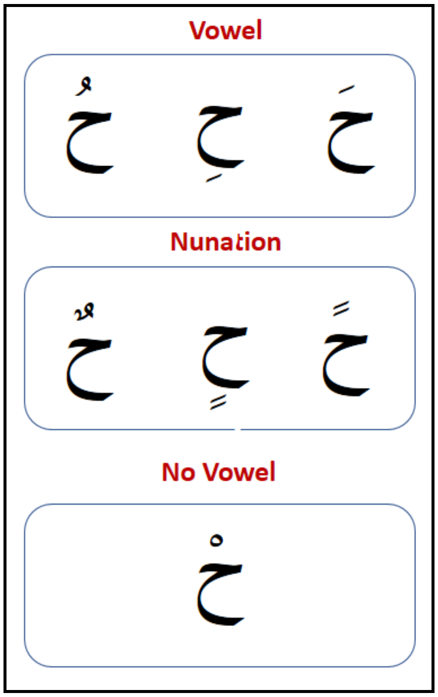
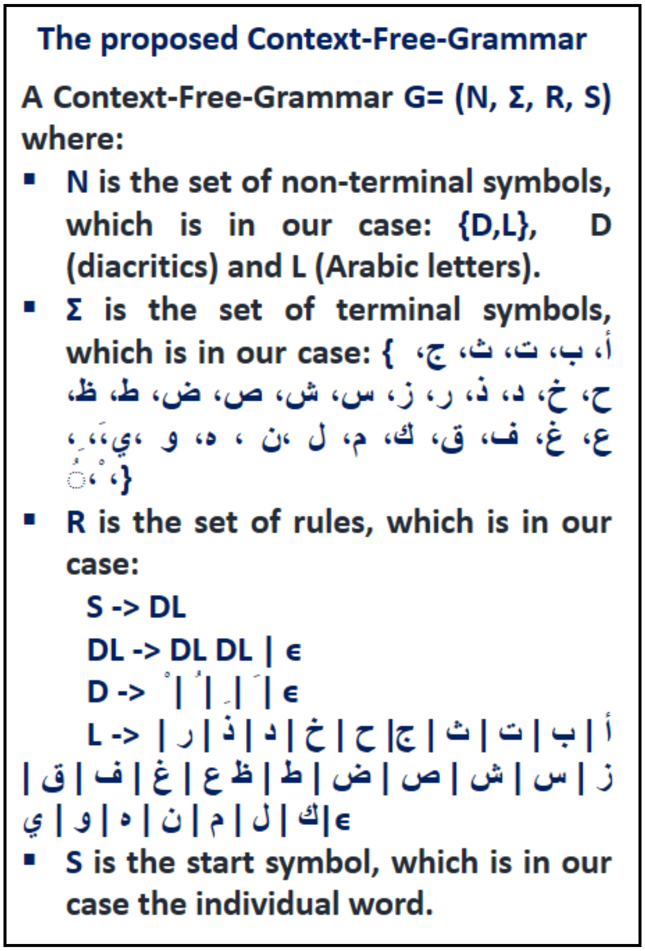


Figure 6

In this article, we proposed a context-free-grammar that will reduce the word sense disambiguation in the Arabic language. Figure 7 shows the proposed solution.



The proposed Context-Free-Grammar

A Context-Free-Grammar G= (N, Σ, R, S) where:

- N is the set of non-terminal symbols, which is in our case: {D,L}, D (diacritics) and L (Arabic letters).
- Σ is the set of terminal symbols, which is in our case: { أ، ب، ت، ث، ج، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ك، م، ل ،ن ، ه، و ،ي،َ،ِ، ُ، ْ،}
- R is the set of rules, which is in our case:

  S -> DL
  DL -> DL DL | ϵ
  D -> ْ | ُ | ِ | َ | ϵ
  L -> أ | ب | ت | ث | ج| ح | خ | د | ذ | ر | ز | س | ش | ص | ض | ط | ظ ع | غ | ف | ق | ك | ل | م | ن | ه | و | ي |ϵ
- S is the start symbol, which is in our case the individual word.

Figure 7

The proposed parser will work as the following: in the lexicon, each word will be attached with it diacritics and the meaning of it. So if for example enter the world that has been entered is "حَب" we will figure that it means "seed." But if the entered word is "حُب" we will figure that the word means "love." Figure 8 shows the parse tree for the word ".حُب"

6. Conclusions

This article highlights the effort that has been done in processing the Arabic language. A lack of research was found, and this is expected because the Arabic language is complicated because of its structure. We propose a parser that will decrease the word sense ambiguation using the

Context-Free-Grammar. The next step is to evaluate the proposed solution by testing it.

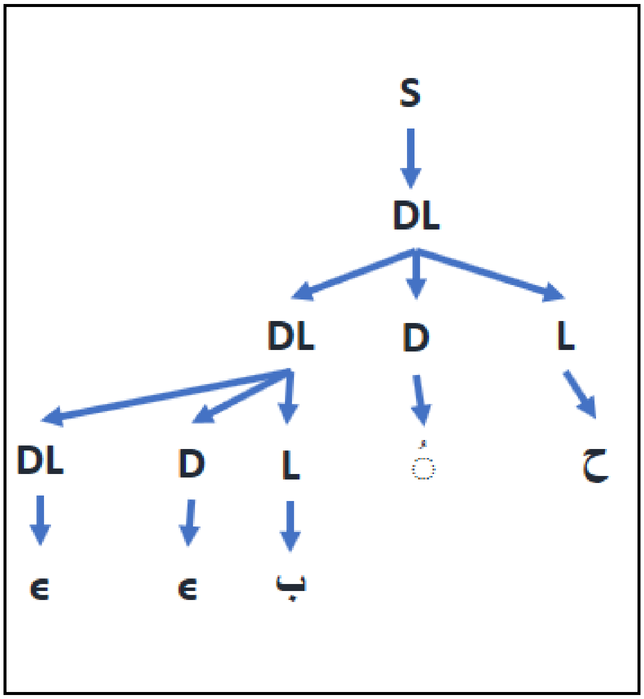


Figure 8

References

[1] A. F. A. R. Khaled Shaalan, "Towards An Arabic Parser for Modern Scientific Text," research gate.

[2] K. S. a. A. R. Eman Othman, "A Chart Parser for Analyzing Modern Standard Arabic Sentence."

[3] M. A.-E. a. K. S. Said A. Salloum, "A Survey of Lexical Functional Grammar in the Arabic Context," International Journal of Computing and Network Technology, 2016.

[4] H. M. a. M. S. A. Shihadeh Alqrainy, "Context-Free Grammar Analysis for Arabic Sentences," International Journal of Computer Applications, Jordan, 2012.

[5] E. Ditters, "A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic," Netherlands.

[6] A. M. a. R. B. Nabil Ababou, "Parsing Arabic Nominal Sentences Using Context-Free Grammar and Fundamental Rules of Classical Grammar," I.J. Intelligent Systems and Applications, Morocco, 2017.

[7] E. A. D. a. A. Basata, "A Framework to Automate the Parsing of Arabic Language Sentences," International Arab Journal of Information Technology, Jordan, 2009.

[8] M. M. a. S. W. Ahmed Al-Taani, "A Top-Down Chart Parse for Analyzing Arabic Sentences," The International Arab Journal of Information Technology, 2012.

Zeinab E. E. Mahmoud Currently, Assistant professor, Faculty of computing and information technology.

Hamdah A. Alotaibi received B. S. degree in computer science in 2015 from Umm Al-Qura University. Currently, study Master in King Abdul-Aziz University.