A Prototype For Automatic Grammatical Rule Extraction For Arabic Language

Ghadeer Al Natshah, Manal Tamimi*
Palestine Polytechnic University
ghadeer_natshah@student.ppu.edu, manalt@ppu.edu

Abstract

In this research we describe a technique used for extracting grammatical rules of standard Arabic language using machine learning. The proposed technique is supported with special dictionaries built for this purpose. In addition to dictionaries, we employed a corpus with data obtained from Open Source Arabic Corpus (OSAc) . This research will form a core for more comprehensive researches that support the standard Arabic. Our system is adaptive, it has the ability to benefit from the feedback taken from its users in improving its performance.

Introduction

To be able to detect and correct grammatical errors that may exist in a Natural language's sentences, it is important to bring together the set of grammatical rules of that language. But, in the case of the Arabic language, obtaining concrete rules is tedious, because the Arabic language is complex, and has a huge number of words and phrases that are considered as anomalous base, which makes it important for us to use machine learning in rule extraction.

Using learning in rule extraction is very much similar to the way humans learn languages. This is achieved by observing the structure of a massive number of grammatically correct sentences, through which the system concludes the way words may be arranged inside the language sentences.

Methodology

To achieve our purpose, we have developed a dictionary with more than 120000 Arabic verbs. This dictionary is populated with the passive and active verbs in present, past, and imperative forms, for singular, dual, plural of both masculine and

* corresponding author

feminine**.** Also a complete morphological analysis of each verb could be gained out of this dictionary. A nouns' dictionary is also employed. It contains Arabic nouns in single, dual, and plural forms.

The vast number of the sentences used in the research were extracted from Open Source Arabic Corpus (OSAc). We have chosen this corpus because it has a wide diversity of Arabic subjects. In addition, the sentences in it are written in standard Arabic, far from dialects.

In order to formulate the sentence pattern, each sentence was broken into separate words. Each word is given a special code that is extracted from its description in the dictionary, and that reflects its type. The combination of codes of all words of a sentence forms the sentence's pattern. This pattern is saved in a database.

The set of all patterns are used in inferring a set of rules, using machine learning. These rules will be used in the judgment of correctness of the Arabic sentences.

Results

The results obtained from this research are a database that contains a huge number of patterns, and a set of rules extracted from these set of patterns. These will be further used in more valuable researches in the future.

References

1) Habash N .(2010)Introduction to Arabic Natural Language Processing. A Publication in the Morgan & Claypool Publishers.ISBN: 9781598297959. 2010.
2) Hoseini A.(2011) Semantic processing of Arabic language. Maryam. Journal of American Science, 2011.
3) Graham K.(2011) Introduction to the Special Issue on Arabic Computational Linguistics. ACM Transactions on Asian Language Information Processing, Vol. 10, No. 1, Article 1, Publication date: March 2011.
4) Altantawy M & Habash N.(2011) .Fast Yet Rich Morphological Analysis.. Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, pages 116–124, Blois (France), July 12-15, 2011. c 2011 Association for Computational Linguistics.