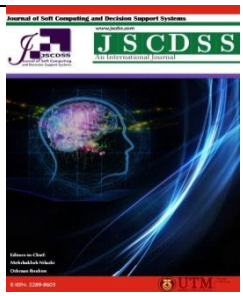
An International Journal
JSCDSS
http://www.jscdss.com
Vol.3 No.2 April 2016: 45-52
Article history:
Accepted 23 February 2016
Published online 23 February 2016

Journal of Soft Computing and Decision Support Systems



E-ISSN: 2289-8603

Arabic Part-of-Speech Tagging

Rabab Ali Abumalloh [a,*], Hassan Maudi Al-Sarhan [b], Othman Bin Ibrahim [c], Waheeb Abu-Ulbeh [c]
[a] University of Dammam, Department of Computer Science, Dammam, Saudia Arabia
[b] Ajloun National University, Faculty of Information Technology, Ajloun, Jordan
[c] Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia

*** Corresponding author email address**: ramolloh@uod.edu.sa

Abstract

The study described in this paper belongs to the area of computational linguistics. Computational linguistics is a field of artificial intelligence dealing with the logical modeling of natural language from a computational perspective. It unites two areas that are quite different in appearance, computer science and natural languages. Computational linguistics might be considered as a synonym of automatic processing of natural language, since the main task of computational linguistics is just the construction of computer programs to process words and texts in natural language. There are many areas that may be considered as properly included within the discipline of computational linguistics. One of these areas is part-of-speech tagging (POS-tagging). POS-tagging is considered as a process for automatically assigning the proper grammatical tag to each word of a written text according to its appearance on the text. Thus, the task of POS-tagging is attaching appropriate grammatical or morpho-syntactical category labels to each word, token, symbol, abbreviation and even punctuation mark in a corpus. POS-tagging is usually the first step in linguistic analysis. Also, it is very important intermediate step to build many natural language processing applications. It could be used in spell checking and correcting systems, speech recognition systems, information retrieval systems and text-to-speech synthesis systems.

Keywords: Tagging, Natural language processing, Arabic language

1. Introduction

Part-of-Speech tagging belongs to the field of computational linguistics. Computational linguistics is a discipline of artificial intelligence that deals with the logical modelling of natural language from a computational perspective (Das et al., 2015). It combines two fields: computer science and natural language processing (NLP). Computational linguistics focuses on proving linguistics theories using computer (Muaidi and Ayesh, 2006). These days the computational linguistics is concerned as one of the hottest topic research areas. Some of the areas that are studied by computational linguistics are: design of machine translation, design of parsers, design of spell-checkers and correctors.

In literature, there are many definitions for the term Part-of-Speech tagging". The well-known definition is the one that is introduced by Jurafsky and Martin (2000) in their valuable book: "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition" This definition is as follows: "Part-of-speech tagging (or just tagging for short) is the process of assigning a part-of-speech or other syntactic class marker to each word in a corpus". POS-tagging is considered as a base stage for many of Natural Language Processing (NLP) applications. Some of these applications are:

i. Parsing System: For any parsing system there is a need for lexical information to be added to the words, this process is done using POS tagging.
i. Information Retrieval: Enhancing information retrieval applications using POS tagging process can be done by removing the lexical ambiguity and adding the syntactic class of the words.
ii. Building Dictionaries: The tagged text will be very useful by providing the required information needed for building dictionaries.
iii. Speech Synthesis Systems: Increasing the accuracy of speech recognition systems by producing more natural pronunciations in the speech synthesis systems.
iv. Word Processing: If the class of the misspelled word is known, then the process of correcting the word will be easier.

Computer corpus linguistics concerned with using large quantities of texts in machine-readable form. In the study of linguistic phenomena a corpus is defined by Leech (1992) as: "A Corpus is a large collection of natural language material stored in machine readable form that can be easily accessed, automatically searched, manipulated, copied and transferred". Corpus provides POS systems with the needed