Arabic Tag Sets: Review

Marwah Alian[1,2] and Arafat Awajan[2]
[1] Hashemite University, Zarqa, Jordan
Marwah2001@yahoo.com
[2] Princess Sumaya University for Technology, Amman, Jordan
Awajan@psut.edu.jo

Abstract. Labeling a word with a suitable tag based on its context and its grammatical category is a major step in many applications of natural language processing. Constantly, there is an effort for inventing a set of these tags for Arabic language. In this research, a review for the existing Arabic tag sets is presented. A description for their features and limitations is also introduced.

Keywords: Tag Tag set Arabic tag set

1 Introduction

Part of Speech Tagging is the process of assigning proper tag for each word in a text representing its grammatical and morphological syntactic feature for a word [1]. Further, a tag is a code that holds simple or complex information that represent a word features and it labels the word in a text [2]. The development of a tag set that consist of representative tags at early stages is important for diacritical based tagging system. The need for such a tag set comes from the fact that Arabic language does not have a standard or complete tag set [3].

The approaches used for Part Of Speech (POS) tagging are classified into three main approaches; the first approach is the Rule -based Approach w hich sometimes called linguistic approach or Knowledge -Based Approach [4, 5], this approach use a set of linguistic rules during the process of tagging. The second approach is the Statistical Approach tha t is also called Probabilistic Approach or Stochastic Approach [6, 7]. This approach depends on building a statistical language model by gathering statistics from existing tagged corpora. The third approa ch is the hybrid approach in which rule -based and statistical approaches are involved [8, 10]. In the hybrid approach, both rule-based and statistical approaches are combined. On the other hand, some syst ems use other approaches, like machine learning algorithms, neural networks and decision trees [5, 9]. The existing Arabic tag sets vary in size from 6 tags to 2,000 detailed tags.

Some of these tag sets follow the same standards adopted in the tag set design for English, but these tag sets may be inappropriate for Arabic. Also, there are some morphological features that are common between Arabic tag sets like number, gender, case, person, definiteness and mood. However, the attributes are not uniformed among the morphological features [15]. In this research, a review for the existing Arabic tag sets is presented with their features and limitations.

K. Arai et al. (Eds.): IntelliSys 2018, AISC 868, pp. 592-606, 2019.