Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 240–247

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017)

The Use of Hidden Markov Model in Natural ARABIC Language Processing: a survey

Dima Suleiman[a,b,*], Arafat Awajan[a], Wael Al Etaiwi[a]

[a] COMPUTER SCIENCE DEPARTMENT KING HUSSEIN FACULTY OF COMPUTING SCIENCES PRINCESS SUMAYA UNIVERSITY FOR TECHNOLOGY AMMAN, JORDAN

[b] BUSINESS INFORMATION TECHNOLOGU DEPARTMENT THE UNIVERSITY OF JORDAN AMMAN, JORDAN

Abstract

Hidden Markov Model is an empirical tool that can be used in many applications related to natural language processing. In this paper a comparative study was conducted between different applications in natural Arabic language processing that uses Hidden Markov Model such as morphological analysis, part of speech tagging, text classification, and name entity recognition. Comparative results showed that HMM can be used in different layers of natural language processing, but mainly in pre-processing phase such as: part of speech tagging, morphological analysis and syntactic structure; however in high level applications text classification their use is limited to certain number of researches.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: **Hidden Markov Model; Arabic Natural Language Processing; Part-of-Speech Tagger; Morphology; Statistical Language Model;Trigram; Bigram, first order logic; second order logic; Text classification; Name Entity Recognition**

* Corresponding author. Tel.: +962795016922
E-mail address: dimah_1999@yahoo.com

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.363

1. INTRODUCTION

Natural Language Processing (NLP) applications that utilize statistical approach, has been increased in recent years. One of the most important models of machine learning used for the purpose of processing natural language is Hidden Markov Model (HMM) [1]. Markov Model is a probabilistic model that are considered as sequence classifier such as letters classifier; it calculates the probability of label sequence and chooses the best sequence according to the best possible labels probability distributions. Moreover, Hidden Markov Model is a model that contains a set of state and transitions where transition from one state to another state is determined according to certain input. Each transition contains a value or weight that is determined according to certain probability distribution. Therefore, if certain input causes transmission from state x to state y then the overall weight will be augmented by the weight w that is the value of transition or transition probability between state x and state y. The probability distribution of a certain transition determines the observation or outcome of a certain state. However, Hidden Markov model is called hidden since the states are not visible and only outcomes can be seen. In our case the input is a sequence of words or letters, so the sequence of words will determine the sequence of states; this sequence represents a chain called Markov chain.

Hidden Markov Model was used in many applications of statistical NLP such as morphological analysis, part of speech tagging (PoST) and text classification. This research provides a comparative study between different applications using Hidden Markov Model in statistical language processing of Arabic language.

The remainder of this paper is structured as follows: Section 2 will explain some Arabic language features. Terminology about Hidden Markov Model, tagging and Markov chain will be introduced in section 3. Section 4 will discuss the related work and finally conclusion will be presented in section 5.

2. ARABIC LANGUAGE FEATURES

Arabic Language has many features that make processing harder than other Languages. Most of Arabic language roots consist of three or four characters, however few have five or more. Arabic language is full of morphology that can be divided into templatic and concatenative. There are no templates for foreign languages words which considered as nonderivative words. Moreover morphemes that are concatenative consist of stem in addition to affixes and clitics. There are three types of affixes: prefix, circumfixes and suffix, also there are two types of clitics: proclitics and enclitics. Generally stem may be preceded by prefix and followed by suffixes; however circumfixes may occurs at the middle of the stem, proclitics at the beginning of the word and enclitic at the end. All these are called morphemes and all morphemes except the stem are optional. The General structure of morphemes can be represented as follows, where a character that represents the optional morpheme is [ ]:

[Proclitic(s)+[Prefix(es)]] + stem + [Suffix(es) + [Enclitic]].

Arabic language also consists of stop or functional words such as pronouns, prepositions, conjunctions and many others, in most of NLP applications functional words were removed in preprocessing phase. Clitics and affixes can be used with stop words, derivative and nonderivative words.

3. TERMINOLOGY

3.1 Hidden Markov Model

Hidden Markov is one of the models that can be used as classifier; Markov consists of a set of state where transition from one state to another depends on certain input. Therefore the transition between states continues until reaching the output state or observation. Moreover, the probability of certain transition depends on the probability of transition from the previous state to current state. The probability model consists of three main elements: experiments with well-defined results, sample space (Ω) which consists of all possible events and finally the event