Journal of King Saud University – Computer and Information Sciences (2016) **28**, 157–169

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Rational kernels for Arabic Root Extraction and Text Classification☆

Attia Nehar [a,*], **Djelloul Ziadi** [b], **Hadda Cherroun** [a]

[a] *Laboratoire d'informatique et Mathématiques, Université A.T. Laghouat, Algeria*
[b] *Laboratoire LITIS – EA 4108, Normandie Université, Rouen, France*

Received 9 April 2015; revised 8 November 2015; accepted 8 November 2015
Available online 19 December 2015

KEYWORDS

N-gram;
Arabic;
Classification;
Rational kernels;
Automata;
Transducers

Abstract In this paper, we address the problems of Arabic Text Classification and root extraction using transducers and rational kernels. We introduce a new root extraction approach on the basis of the use of Arabic patterns (Pattern Based Stemmer). Transducers are used to model these patterns and root extraction is done without relying on any dictionary. Using transducers for extracting roots, documents are transformed into finite state transducers. This document representation allows us to use and explore rational kernels as a framework for Arabic Text Classification. Root extraction experiments are conducted on three word collections and yield 75.6% of accuracy. Classification experiments are done on the Saudi Press Agency dataset and N-gram kernels are tested with different values of *N*. Accuracy and F1 report 90.79% and 62.93% respectively. These results show that our approach, when compared with other approaches, is promising specially in terms of accuracy and F1.

1. Introduction

Text Classification (TC) is a machine learning-based task. It aims to automatically sort a set of documents into one or more classes from a predetermined set (Sebastiani and Ricerche, 2002). Applications of TC include many domains, such as article indexing, Web information searching, mail spam detection, and even automatic assessment systems.

In this work, we enhance the root extraction technique, introduced by authors in a previous paper (Nehar et al., 2012), and we asses its performance in the context of Arabic Text Classification (ATC). Indeed, root extraction method introduced in Nehar et al. (2012) gives a set of possible roots. Our new root extraction approach chooses the best root based on a statistical study of character occurrences in the Arabic roots corpus علي حلمى موسى (1978). Experiment and comparison are conducted to assess performances against standard stemmers. Our root extraction technique transforms documents into finite state transducers. Then, rational kernels (Cortes et al., 2004), which

☆ This work is supported by the MESRS – Algeria under project 8/U03/7015.
* Corresponding author.
E-mail addresses: a.nehar@mail.lagh-univ.dz (A. Nehar), djelloul.ziadi@univ-rouen.fr (D. Ziadi), h.cherroun@mail.lagh-univ.dz (H. Cherroun).
Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2015.11.004

are language/task independent methods, are used as a framework to do ATC. This allows the use of different distance measures or kernels, like N-grams kernels, with the aim of identifying the suitable value for N.

The rest of this paper is organized as follows. Section 4 presents the two main types of stemming, namely: light stemming and root extraction (or heavy stemming) techniques. In Section 5, we recall some notions on weighted transducers and rational kernels. We present, in Section 6 our new root extraction approach, then we explain how to use rational kernels as a framework for ATC. Experiments and results are reported and interpreted in Section 7.

2. Related work

Due to the increased availability of Arabic documents in digital form and the complexity of the Arabic language, Arabic Text Classification (ATC) has increasingly begun to receive attention. Significant work has been conducted to improve performance of ATC systems (El Kourdi et al., 2004; Syiam et al., 2006; Duwairi, 2007; Althubaity et al., 2008; Hadi et al., 2008; Mesleh, 2008; Gharib et al., 2009; Kanaan et al., 2009; Khreisat, 2009; Alsaleem, 2011; Hadni et al., 2013; Hmeidi et al., 2014).

In general, an ATC system consists of three steps:

1. *Preprocessing step*: text is normalized by removing diacritics, punctuation marks, stopwords, special characters, numbers and all non-letter characters.
2. *Features extraction*: text is transformed into a vectorial form by extracting a set of features. For example, (Khreisat, 2009) performed features extraction by using N-gram technique, while (Syiam et al., 2006) relied on stemming. In addition, terms weighting and feature reduction techniques could be used to enhance performance.
3. *Learning step*: in this step, the goal is to teach the system how to classify Arabic text documents. Many supervised algorithms were used: Support Vector Machines (Alsaleem, 2011; Gharib et al., 2009; Mesleh, 2008), K-Nearest Neighbours (Hadi et al., 2008; Syiam et al., 2006, Naive Bayes Alsaleem, 2011; El Kourdi et al., 2004; Hadi et al., 2008). Most techniques rely on similarity measures over extracted features to determine whether two documents are similar.

In the first step, elements that add nothing to the meaning of documents are removed from the text. The second step aims to represent documents in a vectorial form by extracting a set of features from the documents. The Bag of Words (BOW) was by far the simplest way to represent documents in a vectorial form. All the words are used to index a vector representing occurrences of words in a document. Many improvements of the BOW were proposed to enhance ATC systems, including feature selection, dimension reduction, terms weighting and stemming. The BOW is a word level representation. In Khreisat (2009), N-gram technique with a character level is used. Stemming is used to reduce dimensionality. Several stemming techniques are proposed (Buckwalter, 2004; Al-Nashashibi et al., 2010). Authors of Khoja and Garside (1999) designed a dictionary based root extraction technique that reports good results, but the dictionary have to be kept up-to-date. The root extraction technique developed in Al-Serhan et al. (2003) gets the three-letter roots for Arabic words without relying on any language resources such as pattern files or roots dictionary.

In Arabic language, surface words could be classified to root family classes, i.e, words that are derived from the same root but do not have the same sense. Reducing semantically distinct words to the same root can lead to classification performance decrease. In order to avoid this, light stemming is applied in TC systems (Aljlayl and Frieder, 2002). Light stemming consists of removing a small set of prefixes and/or suffixes, without trying to consider infixes, or detecting patterns. Due to this strategy, light stemming results in a large number of features compared with root extraction.

In the third step, most algorithms depend on distance metrics to evaluate the similarity (or dissimilarity) between documents using feature vectors. The quality of the classification system is related to the used distance measure.

3. Challenges and linguistic issues in ATC and root extraction

Arabic Text Classification faces many challenges. The first important challenge is related to Arabic morphological analysis, which is a crucial tool for ATC systems. Indeed, the process of Text Classification depends on the content of documents, a massive number of features can lead to poor performance in terms of accuracy and time. Arabic is lexically a very rich language, important number of surface words can be generated from one stem or root. Treating all surface words will end up with a very large number of features, one solution is to use morphological tasks, like stemming and root extraction. The second challenge concerns the semantical level of Arabic language, Text Classification is sensitive to expressions meaning. The morphological richness and orthographic ambiguity, due to optional diacritization, can lead to a large number of homographs and homonyms (Habash, 2010). Synonyms are also widespread in Arabic language. The third challenge is the lack of publicly available free Arabic corpora for evaluating ATC systems. Much work was done on manually obtained datasets. This lack should be fulfilled over time with standard and benchmark corpora. In the next paragraphs we will give more details about these challenges.

Morphological analysis is the study of internal word structure (Habash, 2010). Morphologically, the Arabic language is the most complicated and rich language. Many words can be formed using the same root, a few patterns, and a few affixes. One of the challenges in a root extraction is that words in Modern Standard Arabic are free of diacritics, which makes them more ambiguous. For instance, the two words (كَتَبَ, He wrote) and (كُتُبٌ, Books) are originated from the same root but has different meanings when vocalized. Furthermore, unvowelled words can lead to more important ambiguity. Lets take the two words: (يَنَعَتْ, Ripen) and (يَنْعِتُ, Qualify). These words originated from two different roots (ينع and نعت respectively) though have the same orthography.

Multiple affixes and clitics can appear in a word, due to agglutinative nature of Arabic, sometimes giving word-forms