ARABIC LANGUAGE SCRIPT AND ENCODING IDENTIFICATION WITH SUPPORT VECTOR MACHINES AND ROUGH SET THEORY

MOHAMED OULD MOHAMED SIDYA

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

UNIVERSITI TEKNOLOGI MALAYSIA

NOVEMBER, 2007

DEDICATION

This thesis is dedicated to my beloved family.

ACKNOWLEDGEMENT

First and foremost, I would like to thank ALLAH S.W.T. for guiding me throughout my lifetime. Next, I wish to extend my grateful appreciation to my beloved family for all the supports and prayers that they make for me. I would like to take this opportunity to thank my supervisor, Dr. Ali Selamat for the attention and guidance throughout the length of this study. Not forgetting also, my examiners Associate Professor Dr. Mohd Nor bin Mohd Sap and Associate Professor Dr.naomi salim for many helpful suggestions. My sincere appreciation goes to all those who have contributed directly and indirectly to the preparation of this study. I am grateful to all my colleagues, friends, staff, and lecturers in Faculty of Computer Science and Information System, Universiti Teknologi Malaysia for their help and support at every step during this course of study.

Abstract

Arabic is ranking sixth among the world's spoken languages with more than 230 million speakers around the Arabic world. There are different flavors and dialects of Arabic; the most common one is the Egyptian Arabic which has the largest number of users (more than 50 millions). Although, only a small number Arabic speakers use the internet, still it constitutes a considerable share to the internet community. Unfortunately, so far, there has been no research to automatically distinguish between the Arabic language and the other languages that use the same script. This project deals with identifying the Arabic language from the Persian language; both languages are written in the Arabic script. The data for this project has been collected from the internet, the BBC website in particular. Many operations have been applied to this data, including stop word removal and stemming. This project is established to compare the performance of Support Vector Machines with Rough Set Theory in Identifying the Arabic language. The results show that both methods perform well but the Support Vector Machines outperform the Rough Set Theory.

Abstrak

Bahasa Arab merupakan tangga ke-enam bahasa pertuturan dunia dengan melebihi 230 juta jumbahasa di selunih dunia arab. Terdapat pelbagai citarasa dan dialek bahsa arab; bahasa Arab Egypt menupakan bahasa yang paling banyak dituturkan di mana ianya mempunyao bilangan pengguna terbesar (melebihi 50 junta). Walaupun hanya sebilangan kecil pengguna arab di internet, tetapi ianya boleh dipertimbangkan sebagai perkongsian komuniti internet. Malangya, setakat ini tiada penyelidikan yang dapat membezakan antara bahasa arab dan bahasa lain yang menggunakan skrip bahasa yang sama. Projek ini melibatkan pengecaman antara bahasa arab dan bahasa parsi; di mana ke dua-dua bahasa di tulis dalam skrip arab. Data untul projek ini di ambil daripada internet, terutama di laman web BBC. Terdapat pelbagai operasi dilaksanakan terhadap data ini, termasuk membuang perkataan imbuhan dan penapisan. Projek ini disempurnakan dengan membandingkan prestasi antara Support Vector Machines dan Teori Rough Set dalam pengecaman bahasa Arab. Hasil menunjukkan kedua-dua teknik menghasilkan prestasi yang baik tetapi Support Vector Machines menunjukkan prestasi yang lebih baik daripada Teori Rough Set.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SVM	Support Vector Machines
KM	Kernel Methods
SK	String Kernels
PCA	Principal Component Analysis
RS	Rough Set Theory
LSE	Language Script and Encoding

CHAPTER 1

OVERVIEW

1.1 Introduction

Arabic is ranking sixth among the world's spoken languages with more than 230 million speakers around the Arabic world. There are different flavors and dialects of Arabic; the most common one is the Egyptian Arabic which has the largest number of users (more than 50 millions). Not only is Arabic spoken in the countries whose official language is Arabic, but also some significant communities in western countries such as Netherlands, France, United kingdom and others, and some African countries such as Central African Republic. All dialects of Arabic use the same writing script. In addition this writing script is used by many other languages around the world, in the Middle East, south East Asia and the Indian continent.

Although, only a small number of this population uses the internet, still it constitutes a considerable share to the internet community. Unfortunately, so far, there has been no research to automatically distinguish between the Arabic language and the other languages that use the same script. That is what this project is all about. This work

proposes a combination of Rough Set Theory and Kernel String Computation to identify Arabic script and encoding from other languages using the same script.

Language script and encoding is a growing area of research. Although continually increasing, not much research has been done in this area. The work that has been done in LSE so far concerns languages with Roman alphabets and some other languages such as Chinese, Japanese and Thai.

This project tries to approach the Arabic LSE identification using a combination of one common approach in LSE and another not very common one. The former is the String kernel computation, which has been used frequently, with all of its flavors that will be highlighted later, in the LSE identification. The String Kernels are being used to compare the set of all common substring between two strings. In String kernels, each document is encoded as a feature vector with substring frequencies as elements. The latter is the Rough set theory, which will also be explained in brief, and will be used for classification of patterns.

Standard learning systems operate on their input data after it has been transformed into an appropriate form, most commonly a feature vector d1,…dn in an m dimensional space. However, this form usually requires the system to perform some additional and very heavy computation in order to convert the data into the desired feature vector. Besides, it is very common that the data itself cannot be readily converted into feature vectors; either because of the data complexity or because the conversion into a feature vector might result in losing some important information because of the conversion process. Kernel methods are an alternative to feature extraction that avoids these pitfalls.

Kernel methods (KMs) are an effective alternative to explicit feature extraction. The building block of Kernel-based learning methods (KMs) is a function known as the

kernel function, i.e. a function returning the inner product between the mapped data points in a higher dimensional space. The learning then takes place in the feature space, provided the learning algorithm can be entirely rewritten so that the data points only appear inside dot products with other data points. Several linear algorithms can be formulated in this way, for clustering, classification and regression. The most well known example of a kernel-based system is the Support Vector Machine (SVM), but also the Perceptron, PCA, Nearest Neighbour, and many other algorithms have this property. The non-dependence of KMs on dimensionality of the feature space and flexibility of using any kernel function makes them a good choice for different classification tasks especially for text classification.

One technique that has been used in text categorization makes use of the classical text representation technique that maps a document to a high dimensional feature vector, where each entry of a vector represents the presence or absence of a feature. Another approach to text categorization is to consider documents as symbol sequences and makes use of specific kernels. This approach does not require any domain specific knowledge, as it considers documents just as long sequences, and nevertheless is capable of capturing topic information.

1.2 Problem Background

Rough Set theory is a formal mathematical tool that can be applied, among other things, to reducing the dimensionality of datasets by providing a measure of the information content of datasets with respect to a certain classification. Its peculiarity is a well understood formal model, which allows finding several kinds of information, such as relevant features or classification rules, using minimal model assumptions.

Advantages offered by the theory are: the implicit inclusion of Boolean logic; term weighting; and the ability to rank retrieved documents. The main advantage of

rough set theory in data analysis is that it does not need any preliminary or additional information about data like probability in statistics, or basic probability assignment in Dempster-Shafer theory, grade of membership or the value of possibility in fuzzy set theory.

Many problems in machine learning require a data classification algorithm to work with discrete data. Common examples include biological sequence analysis and Natural Language Processing (NLP). In these applications data is given as a string, an annotated sequence, or a combination of a string and its parse tree.

The basic Idea of the String Kernels that we are going to be using is that documents are mapped into some higher dimensional feature space and each document is encoded as a feature vector with substring frequencies as elements.

So far, many techniques have been proposed for language identification. Some of these techniques are String Kernels using Suffix Arrays, Unsupervised Learning with the Spherical Gaussian Expectation Maximization in combination with the Principal Component Analysis. Also, the Rough Set Theory has been used in Text Classification. However, the combination of Rough Set Theory and String Kernel Computation has not been proposed in Arabic Text identification.

[Padmini Das-Gupta] used the Rough set theory in Information Retrieval. Specifically, they applied it to the design of information retrieval systems accessing collections of documents.

Rough set theory has been used in Text Classification by Alexios Chouchoulas et al. They proposed a dimensionality reduction approach based on Rough Set Theory.

The main motivation behind this project is to contribute to the computerization of the Arabic Language. The reason behind choosing Arabic as a subject to this research is the wide spread of Arabic characters and encoding, which is used by many other languages such as Persian, Jawi, Turkish, and others. Another reason is my being an Arab and would like to help developing the language.

With no doubt, Arabic is spoken by quite a number of people in the world. Considered to be in the top five languages in the world in terms of number of speakers as well as spread world wide. Arabic is used as an official language in the countries of the Middle East and North Africa and this region, besides being a political hot point, is also an attractive economical region because it contains some of the world's largest oil reserves. Arabic has been and is still gradually increasing its presence in the World Wide Web. With the number of web pages written in Arabic increasing dramatically everyday, it is crucial to find a way to identify Arabic text.

1.3 Problem Statement

Language, script, and encoding (LSE) identification is one of major challenges in the area of natural language computerization research. Many techniques have been proposed to process text over the web. Although more than 6,900 languages reported by the Ethnologue are currently used all over the world, only a small number of them have been used in the cyberspace. In addition, only a few of these languages have been given priority in active research related to LSE identification.

Many efforts have been made to prevent the fall-off in using minority languages in the online community and less-computerized languages. With the increasing number of Arabic pages on the web, it has become a necessity to provide some techniques to identify and retrieve Arabic encoded information.

This research is concerned with implementing Rough set theory to identify Arabic text using String Kernel methods to classify the text. Even though there have been few researches in this area, this project is unique because it targets the Arabic text specifically, which has never been done before.

1.4 Objectives

a) Determine the applicability of Rough Set Theory for Arabic identification problems.

b) Test Support Vector Machines to Identify Arabic.

c) Test the Rough Set Theory to identify Arabic scripts from languages with similar encoding such as Persian.

d) Analyze the effectiveness of the two methods.

1.5 Project Scope

This project will be limited to the following points:

a) Collection of textual data from the BBC website. There will be around 400 webpage for each language.

b) Preprocessing of the collected.

c) Testing of Support Vector Machines and Rough Set Theory.

d) Analysis of results of the tests.

1.6 Expected Contribution of this work

This study will be a detailed comparison of the results obtained from Support Vector Machines and Rough Set Theory when used to identify and separate languages written in Arabic scripts and encoding.

1.7 Organization of the Report

Chapter one introduces the project and identifies important points such as the objectives and scope. Chapter two explains the literature review and background of the project, discusses few related work and their pros and cons. Chapter three illustrates the methodology followed here and explains a flow chart of the project.