Arabic documents classification using N-gram

Majed SANAN
Paris 8 university
Paris – France
Sinane80@hotmail.com

Mahmoud RAMMAL
Lebanese university
Beirut-Lebanon
mrammal@ul.edu.lb

Khaldoun ZREIK
Paris 8 university
Paris - France
zreik@univ-paris8

Abstract- ***This paper is aimed to study and explain the useful classification of Arabic text using N-gram as indexing method. The Lebanese official journal documents can be categorized into several classes. Supposing that we know the class(es) of some documents(called learning texts), this can help us to know the candidate words of each class by segmenting the documents.***
In our approach we have used N-gram as a representation method in the case of 3 characters.
In this paper, we study the effect of using n-grams (sequences of words of length n) for text categorization.

Keywords: Classification, learning method, N-gram, Arabic, Categorization.

I. INTRODUCTION

The rapid growth of the Internet has increased the number of online documents available. This has led to the development of automated text and document classification systems that are capable of automatically organizing and classifying documents. Text classification (or categorization) is the process of structuring a set of documents according to a group structure that is known in advance. There are several different methods for text classification, including statistical-based algorithms, Bayesian classification, distance-based algorithms, knearest neighbors, decision tree-based methods....
Text classification techniques are used in many applications, including e-mail filtering, mail routing, spam filtering, news monitoring, sorting through digitized paper archives, automated indexing of scientific articles, classification of news stories and searching for interesting information on the web
The majority of these systems is designed to handle documents written in non Arabic language, Developing text classification systems for Arabic documents is a challenging task due to the complex and rich nature of the Arabic language. The Arabic language consists of 28 letters. The language is written from right to left. It has very complex morphology, and the majority of words have a tri-letter root. The rest have either a quadletter root, penta-letter root or hexa-letter root.
In our approach we will use only the similarity measures and compare the results in order to know the convenient measure in classification using N-grams. And because that classification is one method of text mining we will explain in the following paragraph the steps of text mining, then we will see the preprocessing and indexing of texts before to be classified. At the next paragraph, we will explain the different similarity measures that we will use in our approach, and then the effectiveness measure used to calculate the precision and recall of each class. At paragraph 6 we will explain our approach and experiments, and finally we will see the conclusion and future approaches.

II. TEXT MINING

A. Definition

Text mining is defined [1] as the non trivial extraction of implicit, previously unknown, and potentially useful information from (large amount of) textual data.
Text Mining is the process of applying automatic methods to analyze and structure textual data in order to create useable knowledge from previously unstructured information.

B. Text mining methods

There are many text mining applications or methods. Four of these methods are the following:

- Information Retrieval

This method consists of indexing and retrieval of

textual documents

- Information Extraction

It means extraction of partial knowledge in the text

- Web Mining

It consists on indexing and retrieval of textual documents and extraction of partial knowledge using the web

- Classification

Given: a collection of labelled documents (*training set*), the goal is to find a model for the class as a function of the values of the features

C. *Text mining steps*

These steps concern principally the manner in which a text is represented (or structured), the choice of predicted algorithm to use, and then how to evaluate the obtained results to guarantee a good generalization of the model applied.

1) Representation of the information

In this step we have to segment the unstructured information and put the units segmented into a table. But we have to choose the descriptors (important terms in documents) which can be chosen as words , lemmas, stemmas, or n-grams (characters or words or phrases).
And finally in some cases we have to think how to reduce the dimension of this textual space.

2) Automatic categorization-classification of documents

This is the second step, the text categorization can be defined as the process that permit to associate a category (ies) or class(es) to a text (or document), in function of information contained in this text .
This association is very long and expensive then we think about the automation of this process.
The functional link between a class and a document, that is called 'a prediction model', is estimated by a machine learning method.
The categorization of documents comports a choice of a learning technique (or classifier). The main classifiers used are the following:
- Discriminated factorial analysis [2].
- Neuronal network [3].
- K-Neighbors [4].
- Decision Tree [5].
- Bayesian network [6].

...

3) Validation method

In this final step we have to evaluate the obtained results to guarantee a good generalization of the model applied.

III. TEXT PREPROCESSING AND INDEXING

All text documents went through a preprocessing stage. This was necessary due to the variations in the way text can be represented in Arabic. The preprocessing was performed for the documents to be classified and the training classes themselves. Preprocessing consisted of the following steps:

1) Convert text files to UTF-8 encoding.
2) Remove punctuation marks, diacritics, non letters, stop words. The definitions of thesewere obtained from the Khoja stemmer.
3) Replace initial إ أ آ with ا
4) Replace final ى followed by ء with

A. Spelling Normalization and Mapping

Arabic orthography is highly variable. For instance, changing the letter YEH
(ي) to ALEF MAKSURA (ى) at the end of a word is very common. (Not surprisingly, the shapes of the two letters are very similar.) Since variations of this kind usually result in an "invalid" word, in our experiments we detected such "errors" using a stemmer (the Buckwalter Stemmer) and restored the correct word ending.
A more problematic type of spelling variation is that certain glyphs combining HAMZA or MADDA with ALEF (e.g. أ ,إ and آ) are sometimes written as a plain ALEF (ا), possibly because of their similarity in appearance. Often, both the intended word and what is actually written are valid words.
This is much like confusing "résumé" with "resume" in English. Since both the intended word and the written form are correct words, it is impossible to correct the spellings without the use of context.
We explored two techniques to address the problem.
1) with normalization technique, we replace all occurrences of the diacritical ALEFs by the plain ALEF.
2) with the mapping technique, we map a word with the plain ALEF to a set of words that can potentially be written as that word by changing diacritical ALEFs to the plain ALEF. In this