2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA) | 978-1-6654-0969-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/AICCSA53542.2021.9686866

A language independent approach to multilingual document representation including Arabic

Souhila Boucham
Faculty of Electronics and Computer Science
USTHB University
Algeria
sbouchem@yahoo.fr

Hassina Aliane
Research Center on Scientific and Technical Information
University of Huaguoshan
CERIST,Algeriaa
haliane@mail.cerist.dz

***Abstract* - Arabic language is of increasing interest in the field of Multilingual Information Retrieval (MIR). We deal in this work with the problem of multilingual document representation including Arabic. The proposed approach combines a surface analysis and a Latent Semantic Analysis (LSA) algorithm in a new way to break down the terms of LSA into units which correspond more closely to morphemes. These morphemes are the variable length character N-gram candidates extracted from different fragments separated by borders. The length of the character N-gram candidates is variable because each language has its own properties. This strategy brings an interesting performance for languages such as Arabic in which the words are not explicitly defined and different words are not separated by spaces. The obtained results are encouraging and variability shows that they are perfectible.**

Keywords - multilingual document representation, multilingual information retrieval (MIR), virtual document, principle of border, variable length character N-grams, concept types, pivot language.

I. INTRODUCTION

Several research projects are investigating and exploring various techniques in Information Retrieval (IR) systems for the English, European and Asian languages. However, in Arabic language, there is little ongoing research in Arabic IR or MIR systems including Arabic.

Arabic language is one of the most widely spoken languages. It has a complex morphological structure and is considered as one of the most prolific languages in terms of linguistic articles. Therefore, Arabic IR models need specific techniques to deal with its complex morphological structure [1].

The Core of a MIR system is the indexing process and the retrieval model. In this study, we will focus on models which use an indexing process to store data and to determine how multilingual documents (Arabic, French and English) are represented.

When processing a large corpus with a statistical tool, the first step typically consists of subdividing the text into information units called tokens. These tokens usually correspond to words. This tokenization process may appear to be quite simple, if not to say trivial-tokenization.

However, from an automated processing point of view, the implementation of this process constitutes a challenge. Indeed, how to reliably recognize words? What are the unambiguous formal surface markers that can delineate words, i.e. their boundaries? These questions are relatively easy to answer for languages such as French or English: basically, any string of characters delimited by a beginning space and an ending space is a simple word. But for many other languages, such as Arabic, the answer is much more complicated. In Arabic, subject pronouns and complements are sometimes attached to the verb. In this case, a token like katabtuhu " كتبته"corresponds in fact to a sentence (here, “I wrote it” or “I’ve written it”) (“je l’ai écrit” in French). Obviously, the simple notion of tokens defined as character strings separated by spaces is an oversimplification that is highly inadequate for many situations and languages.

Considering the above, what then could constitute a reasonable atomic unit of information for the segmentation of a text, independently of the specific language it is written in?

The proposed approach avoids the use of tokenizers, stemmers or other language-dependent tools which are complex and may bring noise to representation especially for the high morphologically complex languages including Arabic. The approach is characterized by:

- Language -independent.
- Easy to apply and does not require Natural Language Processing (NLP) tools.
- Construction of the feature terms is based on the N-grams of characters.

The rest of the paper is organized as follows: Section II introduces existing related work. Section III presents the proposed approach to multilingual document representation. Experimental results and analysis are reported in Section IV. Discussion in Section V, and the last section concludes this paper.

II. RELATED WORK

Several approaches have been proposed to solve the document representation problem: