Report
Our current activity and the benefits from the attendance of the Medar conference.

Arabic Information Extraction Problems

Abd El Salam AL HAJJAR

PHD Student
Institute University of Technology
Lebanese University
Lebanon
Paragraph Laboratory
University of Paris 8- Vincennes- Saint-Denis
France
abdsalamhajjar@hotmail.com

Arabic language is used by more than 330 million arabic speakers that are spread over 22 countries. However, the performance of information retrieval in arabic language is very problematic due to the specific morphological and structural changes in the language: polysemy, irregular and inflected derived forms, various spelling of certain words, various writing of certain combination character, short (diacritics) and long vowels, most of the arabic words contain affixes.

Our current work activity specifies in the arabic information extraction and contributes to the enhancement of the arabic information retrieval system performance.

In the Medar conference we presented our paper “Classification of Arabic Information Extraction methods”, in this paper we propose a preliminary classification of arabic information extraction methods. Several methods are proposed to address the problems of information extraction from arabic documents. We have found that these methods can be classified into two main categories. The first one is called "Stemmer". This category includes the following subcategories: Stemmer based on affixes, Stemmer based on translation and Stemmer based on pattern and affixes. The second is called "N-gram". This category regroups the subcategories: N-gram based on Dice's similarity coefficient and N-gram based on “Manhattan distance” dissimilarity coefficient. However, we find methods which implement the two approaches "Stemmer" and "N-gram".

Every method has its advantages and disadvantages. For that, in the next step, we will present of a detailed comparative study of the early described categories. This comparative study will cover mainly the following topics: performances, stabilities, usability, advantages, and disadvantages. Another possible extension of the present work is to test these categories in similar conditions. To accomplish this comparative study, an application must be implemented allowing the evaluation of the performance of each method. In this application we must develop a method for each category. The application implementation requires the method algorithms, the Arabic resources (dictionary, affixes list…), and a corpus that contains a number of Arabic words used in the Arabic world. To evaluate the technical development performance of each method, we must compare it with the method application developed by their authors (if exist).

We can find an application for each method that is developed by its author, but these applications are closed, and we do not have an access on their codes. We can use these applications as an end user only (entry a word and get a root as output), for that, our self method development allows us to access on the code and to apply it automatically on many corpuses.

Our attendance to the Medar conference allowed us to contact and discuss with many people working with the arabic NLP, and Arabic information extraction, to get many items that are required in the application implementation, to find missing methods algorithms, to find some Arabic resources, and to get versions of many systems of methods which are developed. Also, after viewing the work presented in the conference, we evaluated our work in the Arabic information extraction problems, to determine if these problems are resolve by using the information extraction root based, stem based or word based.