GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

Enhancing THE ARABIC INFORMATION RETRIEVAL SYSTEM PERFORMANCES ON THE BASIS OF NAMED ENTITIES

Sabri El Fraoui*1, Noura Aknin2, Téba El Asri3 and Yacine El Younoussi*1

*1, 2, 3 Research Unit 3ITMS Faculty of sciences, 5AEU Tetouan, Morocco

*1(4SIGL research team National School of Applied Sciences, 5AEU Tetouan, Morocco)

ABSTRACT

The detection of named entities (NEs) in the Arabic language is a potentially beneficial pre-treatment for numerous applications of natural language treatment, especially for information retrieval. This task is a major challenge, given the specificities of the Arabic language. In this paper, we will present an approach of named entities extraction and classification in order to improve the preferences of our information retrieval system dedicated to the Arabic language.

Keywords- *Arabic Information Retrieval System; Named-Entities; Arabic Language Processing; ARABIRS.*

I. INTRODUCTION

The detection of named entities (NEs) is important for many tasks in Natural Language treatment (NLP), especially in the Information Retrieval (IR). In fact, the Information Retrieval Systems (IRS) has become a principal obligation respond to rapid changes in technology and boosting the remarkable amount of information available on the web. Hence, text documents in Arabic on the Web have multiplied in number, content, and quantity.

In this article, we will focus on the treatment of NEs in the context of developing Intelligent Information Retrieval System dedicated to Arabic entitled ARABIRS. The main objective of this system is to be able to find the most relevant outcomes to a user query.

Our approach is based on the analysis of query terms of the user to determine which will undergo a lexical process. Stemming has already begun in previous studies [1] [2]. On the other hand, we will try to detect the named entities that do not accept stemming, and then require a particular classification and extraction, which is the purpose of this article.

We will continue, in the second section, with a description of the problem followed by a presentation of the general architecture of the ARABIRS system, we will present later, in the third section, the different techniques of named entities extraction, with a summary of the challenges presented by the characteristics of the Arabic language. Finally, in the fourth section, we will present our approach for the detection and classification of named entities, to enrich user query automatically.

II. CONTEXTE AND MOTIVATION

A) Problematic

According to an internal study by Microsoft, at least 20- 30% of queries submitted to its search engine Bing are simply: named entities, and 71% of queries contain named entities. [3] The main goal of any IRS is to find the relevant results from the set of documents in the corpus, following a user query. This process generally involves three main steps:

- Indexing: it applies together on documents and user query.
- Pairing document query: to find the relevant documents.
- Reformulation of query: to increase the relevance of the IRS rate.

Indexing is to build a structured representation called INDEX containing discriminatory terms (concepts or descriptors) in a document (resp query), usable in the matching document query stage. In other words, the search for documents relevant to the user query.

Two main factors influence the effectiveness of the IRS, the size of the index and the relevance of the descriptors that it contains. The problem we are trying to address in this article is how to benefit from named entities recognition (simple or compound) to decrease the size of the index and improve the relevance of descriptors.

B) ARABIRS Presentation

ARABIRS is a modular system, which consists of two main modules:

- Analysis Module (see Fig. 1): it is responsible for making all necessary treatments before the search for information. It consists of four sub-modules :
  - ✓ Segmentation (tokenization): segment the text into multiple lexical unit (tokens)
  - ✓ Pre-treatment: Apply pre-treatment on words such as elimination of stop words.
  - ✓ Detection of named entities: Recognize and extract named entities.

- ✓ Stemming: extract the roots of words.
- ✓ Indexing: to create the index.

➢ Information Retrieval Module: the second module is responsible for searching relevant documents vis-à-vis the user query. It consists of two sub-modules responsible for:

- ✓ Matching document queries: the mapping between the relevant documents and the user query.
- ✓ Reformulation of queries: reformulating the user query by adding new words in order to increase the relevance of the results.

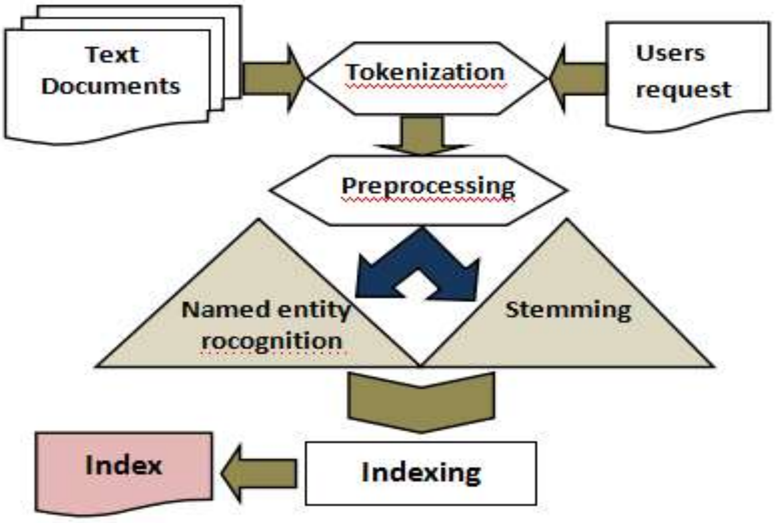

Fig.1: Indexing process of ARABIRS

III. The NEs IN ARABIC LANGUAGE

A) Specificities of the Arabic language

The Arabic language is morphologically rich and complex. The automatic analysis of Arabic words is complicated by the absence of vowels out of context; it is difficult to distinguish the meaning and function of words. This feature introduces considerable ambiguity in the writing of a part text [4], and secondly, by the existence of many spelling variations including proper names in the Arabic alphabet, each letter has four allographs, or how to write, with the exception of a small number of letters whose layout remains unchanged. Each variant is used in a dependent place in the specific context word, which multiplies the unknown forms in the texts. Labelling NE Arabic represents many interesting challenges: Arabic is characterized by the lack of dictionary resources and especially the lack of distinction capital / lower case is a very useful indicator to identify proper names in languages using the Latin alphabet [5] [6].

B) Detection and extraction of NE

The detection of named entities in Arabic, despite many years of research, remains a difficult problem. Several methods have been proposed and validated to improve the detection and extraction of NE. These works are mainly classified into three types of approaches:

➢ The linguistic approach is based on handwritten generic rules. The first work on the recognition of NE in the Arabic language dating from 1998 [7], see also the most recent work [8] [9] and [10] use a parallel corpus to extract NE in Arabic.

➢ The numerical approach is based on a supervised learning mechanism, which is weakly supervised or unsupervised from a large corpus of pre-labeled texts. [11] Use of automatic learning techniques (the Maximum Entropy Markov Models) considering a set of appropriate descriptors and come to very good results. This work was extended in particular [12] and [13]

➢ The hybrid approach presents a combination of the two previous approaches [14] [15] [16].

IV. OUR APPROACH

A) Presentation

Our approach is based on the design of a system for detection and classification of named entities in Arabic in order to reduce the size of the index by classifying entities classes or categories (person, location, organization, date, ...) on the one hand, to enrich user queries by semantic relationships between entities so as to provide the most relevant results, on the other hand. Taking advantage of the encouraging results of some previous work in the field of automatic treatment of named entities in Arabic, we aim to develop a treatment cycle in two consecutive steps:

- Detection and extraction of named entities contained in the corpus and user query.
- Grouping of entities under categories or classes of entities.

This system will be implemented in our IRS ARABIRS between the pre-treatment stage and indexing (see Fig.2). Our system must react in real time, as queries are instantaneous and the terms they contain are constantly changing [17].

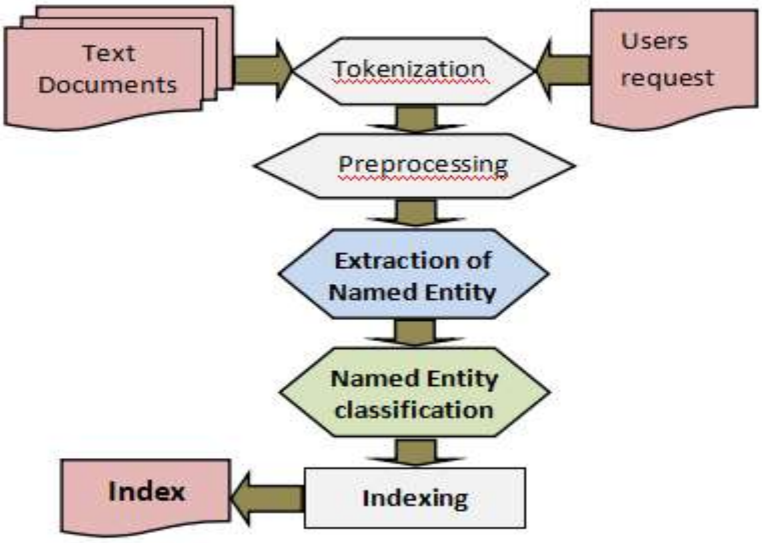

Fig.2: Modified indexing process

B) Implementation

The implementation of our approach consists of two consecutive phases, the output which represents the input of the other.

Detection and extraction of Named Entities

In this phase, we opted for a hybrid approach that is based on:

- A method of automatic learning rules able to detect words that make up the NE according to its type (person, location, organization, number, ...) generated by using an algorithm running on a learning corpus, tested and validated.
- A set of rules extracted manually, based on proper names glossaries and lists of EN previously known (see Fig. 3).

This adjustment will allow us to have two complementary systems, one favoring the recall (numerical approach) and the other precision (symbolic approach), without requiring any additional annotation showing the complementarity of numerical and symbolic methods for the resolution of linguistic tasks. These results are confirmed by the work of [14] [15] [16].

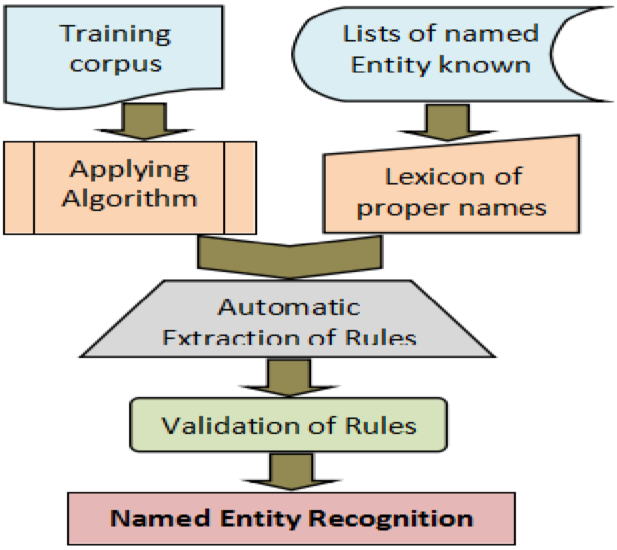

Fig.3: Extraction of Named Entity

Classification of Named Entities

After identifying named entities, we will try to group them into feature classes in a hierarchical tree-like structure (see Figure 4), [18] [19]. This will reduce the size of INDEX on the one hand, and may participate in enriching reformulated queries taking advantage of semantic links between entities of the same class or classes of the same domain. [20] Building links between entities requires learning algorithms and linguistic resources dedicated to the Arabic language, which we will try to develop according to our need.

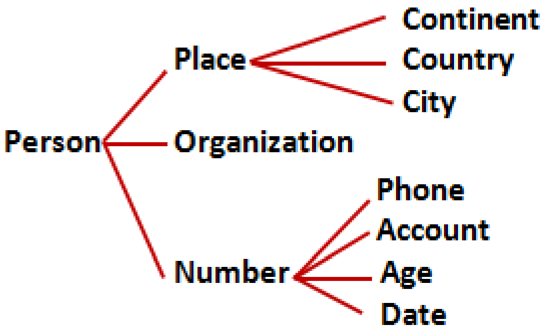

Fig.4: Example of relations between classes of named entity

V. CONCLUSION AND PERSPECTIVES

In this paper, we proposed an approach for the detection and classification of named entities in the context of developing an information retrieval system dedicated to the Arabic language ARABIRS. This approach will reduce the search time by decreasing, on the one hand, the INDEX size; on the other hand, it will improve the relevance of results for the named entities taking advantage of reformulation of queries based semantic links between entities with common relations.

REFERENCES

1. *El younoussi Yacine, Doukkali Sdigui Abdelaziz, Ben Lahmer El habib, 2008. An hybrid approch of arabic language stemming with finite state automata. The 4th International Conference on Computer Science Practice in Arabic (CSPA'08), co-located with the ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'08), April 2008, Doha (Qatar).*
2. *El younoussi Yacine, Doukkali Sdigui Abdelaziz, Ben Lahmer El habib,2009. Arabic language stemming using finite state automata: one word, several roots. Journal of Computer Science and Engineering in Arabic, ISSN 1936-0525, Volume 3, issue 1. Phillips Publishing Company, November, 2009*

3. J. Guo, G. Xu, X. Cheng, and H. Li. Named entity recognition in query. SIGIR'09.
4. HABASH, N. (2010). Introduction to Arabic Natural Language Processing. Morgan Claypool.
5. Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, Thomas Lavergne. Repérage des entités nommées pour l'arabe : adaptation non-supervisée et combinaison de systèmes. JEP-TALN-RECITAL 2012, volume 2: TALN, pages 487–494
6. Debili F., Achour H., Soussi E. : La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique, Correspondances de l'IRMC, N° 71, juillet-août 2002, pp 10-28
7. MALONEY, J. et NIV, M. (1998). TAGARAB : a fast, accurate Arabic name recognizer using high-precision morphological analysis. In Proc. of the Workshop on Computational Approaches to Semitic Languages, Semitic '98, pages 8–15, Stroudsburg, PA, USA.
8. SHAALAN, K. et RAZA, H. (2009). NERA : Named entity recognition for arabic. Journal of the American Society for Information Science and Technology, 60(9):1652–1663.
9. ZAGHOUANI, W., POULIQUEN, B., EBRAHIM, M. et STEINBERGER, R. (2010). Adapting a resource-light highly multilingual named entity recognition system to arabic. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pages 563–567.
10. SAMY, D., MORENO, A. et MA GUIRAO, J. (2005). A proposal for an Arabic named entity tagger leveraging a parallel corpus. RANLP '05.
11. ZITOUNI, I., SORENSEN, J., LUO, X. et FLORIAN, R. (2005). The impact of morphological stemming on Arabic mention detection and coreference resolution. In Proc. of Workshop on Computational Approaches to Semitic Languages, pages 63–70, Ann Arbor, Michigan.
12. BENAJIBA, Y. et ROSSO, P. (2008). Arabic named entity recognition using conditional random fields. In Proceedings of the Conference on Language Resources and Evaluation.
13. BENAJIBA, Y., DIAB, M. et ROSSO, P. (2008). Arabic named entity recognition using optimized feature sets. In Proc. of EMNLP, EMNLP, pages 284–293.
14. MANSOURI A, SURIANI AFFENDEY L, MAMAT A. (2008). Named Entity Recognition Approaches. IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2, February 2008. 339-344.
15. Inès Zribi, Souha Mezghani Hammami, Lamia Hadrich Belguith (2010). L'apport d'une approche hybride pour la reconnaissance des entités nommées en langue arabe. TALN 2010, Montréal, 19-23 juillet 2010.
16. Frédéric Béchet, Benoît Sagot, Rosa Stern (2011). Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées. TALN 2011, Montpellier, 27 juin – 1er juillet 2011
17. Harith Al-Jumaily, Paloma Martínez, José L. Martínez-Fernández, Erik Van der Goot (2011). A real time Named Entity Recognition system for Arabic text mining. Lang Resources & Evaluation (2012) 46:543–563
18. Xiaoxin Yin, Sarthak Shah (2010). Building Taxonomy of Web Search Intents for Name Entity Queries. International World Wide Web Conference Committee (IW3C2), WWW 2010, April 26-30, 2010, Raleigh, North Carolina, USA.
19. Houda Saadane (2013). Une approche linguistique pour l'extraction des connaissances dans un texte arabe. TALN-RECITAL 2013, 17-21 Juin, Les Sables d'Olonne.
20. R. Baeza-Yates, A. Tiberi. Extracting semantic relations from query logs. KDD'07.