

Future Model for Information Retrieval Systems Based on Arabic Language



Tiruvedula GopiKrishna¹, Mohamed A.Abdelhadi², DavuluriNageswara Rao³

¹Sirt University, Research Scholar, Faculty of Computer Science, Hoon, Al-Jufra, Libya,(Indian), gktiruvedula@gmail.com

²Sirt University, Faculty of Computer Science, Hoon, Al-Jufra, Libyan, mabdeldaiem2009@gmail.com

³Sirt University, Lecturer, Faculty of Computer Science, Sirt, Libya(Indian),dnrao99@gmail.com

Abstract: In This research paper we introduced a fully developed classification for Information Retrieval System based on Arabic Language; and also, in this paper the effects of the Arabic text classification on Information Retrieval have been investigated. It has indeed led to an improved computational framework as well as provided excellent solutions for Arabic-IR system by means of emerging both Arabic Linguistics and Information Retrieval Systems. And also his article investigates the impact on retrieval of prefixes in Arabic, which are far more common than in English.

Keywords : Classifications, Taxonomy, Information Retrieval, Framework

INTRODUCTION

The Arabic language is somewhat difficult to deal with due to its right-to-left orientation, the discretization of scripts, vowels which may or may not be included, and its morphological structure. The grammatical system of the Arabic language is based on a root-and-pattern structure and considered as a root-based language with not more than 10,000 roots [8]. This research explains the philosophy behind the whole idea of our scientific work. This work has attempted to link three areas: Information Retrieval, Arabic Language and Knowledge Management Systems in a professional way to join together in one system, belong to a relatively new subject Area, some call it the "Computer Linguistics" and/or "Software Linguistics" or even "Language Engineering". In this research; we have studied the problems of the Arabic language from the standpoint of "Information Science" in general and "Information Retrieval Systems" in particular and to identify its potential importance in Information Retrieval System based on Arabic Language by modern Information Retrieval Systems. Most research centers in universities around the world have dealt directly within Arabic language morphologies; cross-languages or computational linguistics; NLP as well, rather than other important issues like the language generality [9], [10] and its special characteristics. We have considered the Arabic Language in the sense of its simplicity and convergent instead of its complexity and Ambiguous to other spoken Languages such as English language. In Arabic Information Retrieval Systems literatures there were no such any Taxonomy that has tried to concern about the Arabic Language as Information Retrieval System as it should be. We have had almost migrated the models been introduced, and at best way we had adopt our Arabic-IR models to the best existed one. Instead; we should have our own

Information Retrieval System model based on Arabic Language. This new developed classification method should bring new great addition as state-of-the-art research for the Science of Information Retrieval Systems.

2-Related Works

Despite the abundance of computational Arabic studies, there was no specific study on Arabic Retrieval Systems which dealt directly with such an overall classification for computational Methods used to enhance the Arabic Retrieval Systems optimality. [4] Have studied the problem of restoring vocals in Arabic based on linguistic analysis. Vocalization of an Arabic text requires a full understanding of the text. Although this is not at all easy for computers, some semantic analysis can help improving the performance. [5] Were interested in the Arabic cross-language information retrieval track (limited to monolingual Arabic retrieval) and also in both named page and topic distillation searches. [6] studied the techniques for combining evidence for cross language retrieval, searching Arabic documents using English queries. Evidence from multiple sources of translation knowledge was combined to estimate translation probabilities, and four techniques for estimating query language term weights from document language evidence were tried. A new technique that exploits translation probability information was found to outperform a comparable technique in which that information was not used. [7] Have introduced a new Arabic Stemmer for the cross language retrieval which was to translate the English topics into Arabic using on-line English-Arabic machine translation systems. [13], they have investigated the use of Hierarchical Multi Label Classification to tackle text filtering challenges.

3-RESEARCH METHODOLOGY

In our newly introduced model of Arabic-IRs-Classification Method there are a lot of new issues as state-of-the-art. We are concerning with all new issues in the Arabic Information Retrieval System such as for instance, Dialects and its affect on Modern Arabic Language. This can be as new aspects for realization by studying the new concepts and methods for how to optimize the Arabic IR within its Retrieval performance [12]. The developed Classification Taxonomy is totally clear to understand and is as organized as shown below in Figure 1.

Classification hierarchy for Arabic-IR-System

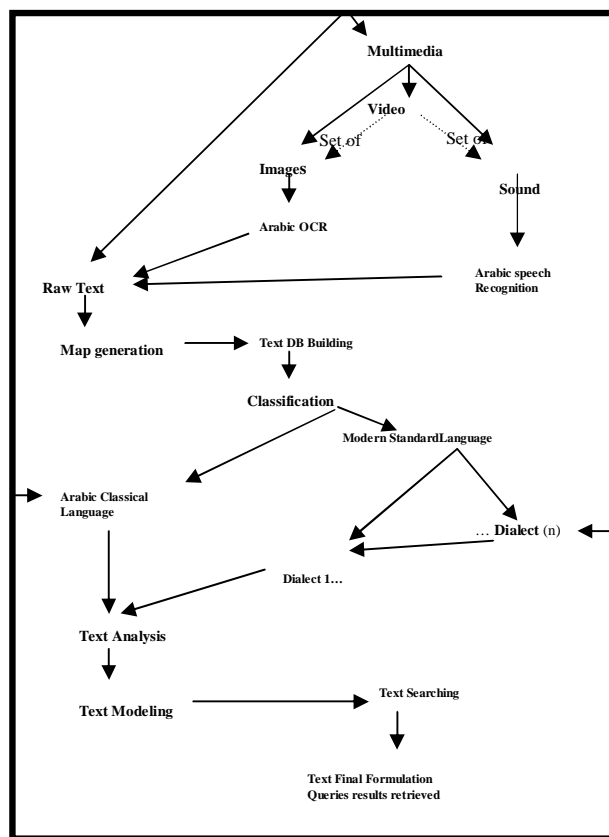


Figure1: A hierarchical Classification for Arabic Information Retrieval System

3.1-A FRAMEWORK FOR ARABIC-IRS

The proposed Framework is used to define the new developed Classification method for Arabic-IRs and also to setup the link between three areas: Information Retrieval; Arabic Linguistic and Knowledge Management in professional way to join together in one System. . The Framework consists of three phases, each of which has its own functional tasks to be implemented so far. We have in the first phase; the preprocessing phase for Arabic documents preprocessing which was the main process depending on the Arabic language characteristics. We have started within the documents as row material as to determine if its Text document or Multi-Media document. In the second phase we have done some processes on documents which have been in the first phase preprocessed. The preprocessed documents which were classified into two categories, one was for Classical Arabic Language, and the other is for Modern Standard Arabic Language including (Arabic-Dialects), [9] and [10]. In thisphase we have used theHierarchalMultilevel-Classification technique[12]; to classify the entire Arabic Words (Filtered Documents) to enhance the System Optimality.

3.2-Arabic Text-Analysis and Processing

The original Arabic Words are divided in turn into two sub Categories; Derived Arabic Words, which are the Verbs and Nouns that are built according to the Arabic derivation Rules,

and Fixed Arabic Words which are a set of Words molded by Arabs, in ancient times, and do not obey the Arabic derivation Rules. Most Vowels in written Arabic are represented by Diacritic marks. Most modern Text is printed in DE vowelized form without these Diacritic marks. While most Words in a Text are traditional Arabic Words, some Words are “Arabized” loan Words from other Languages (perhaps with some phonetic adjustments to facilitate pronunciation) [2]. To filter the Arabic Row-Text we have used Reg-Ex-parser with some modification for extracting Arabic Word-Roots and also to remove the Arabic stop-words from the collected Arabic Row- [11]; and[8]. We have used the benefit of that technique in our Arabic Retrieval System to remove the Arabic stop-word and also to stem the Arabic words, fortunately we have found that there were possibilities to improve our system by using the new Stemmer in extracting Arabic Words Roots too. After having especial studies on the Arabic Roots such as (Subantative, Acusative, Dative) and how can we extract the root from it; we found that there are a specific 3 or 4 letters and the other letters are boundaries Letters, so we can make a Reg-Ex pattern for each Root then compare the Root with the Text to remove the boundaries Letters.

3.3-PerformanceEvaluation of Arabic-Text preprocessing

Many researches focused on achieving those objectives with less regard to storage overhead or performance. In this work we have evaluated the stop-words removing; as shown in Table1, and tested out some Arabic Text collection by performing some queries operation as shown in Figure2, which evaluated in Figure3 to improve the first phase in system performance.

Total Docs in Collection	Total Words in Does Collection	Que ry in Words	Word Frequency	Relevant Docs Retrieved	Non Relevant Docs Retrieved	Recall	Precision
39	27128	History	13	9	30	0.23	0.30
39	27128	World	7	5	34	0.12	0.15
39	27128	Arabic	16	8	31	0.20	0.29

Table 1: Evaluation for Arabic-Text-Preprocessing

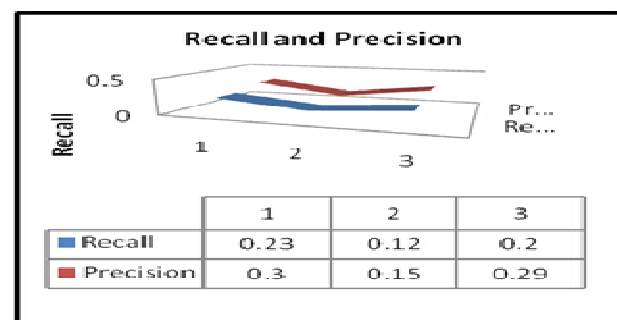


Figure 2: Preprocessing Evaluation

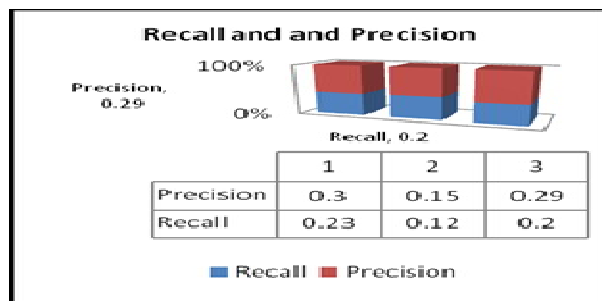


Figure 3: Recall and precision evaluation

CONCLUSION AND FUTURE WORK

As it has been stated in the proposed Framework which has explained our Classification for Arabic-IRs; this research has introduced the first phase in our research work. We will implement the next second and third phases of our work which will be as research contribution for future work. Finally, it is so planned that to complete the other new Framework components such as: Map-Generation, Text Data Base, Modern Standard Arabic Language, Arabic Dialects, Text-Classification, Text Analysis, Text-Modeling and finally Text-Searching.

ACKNOWLEDGEMENT

We would like to thank all of our co-others who helped us in the practical sessions to fulfill this research work.

REFERENCES

- [1] S. E. Robertson, S. Walker, and M. Beaulieu, Okapi at TREC-7: automatic ad hoc, filtering, VLC and filtering tracks. In Proceedings of 7th Text Retrieval Conference (TREC-7), pages 253–264. NIST special publication, 1999.
- [2] Khoja, S. and Garside, R. "Stemming Arabic Text", Computing Department Lancaster University, Lancaster, 1999.
- [3] A. Singhal, C. Buckley, and M. Mitra," Pivoted document length normalization", In Proceedings of 19th annual international ACM SIGIR conference on Research and development in information retrieval, pages 21–29, Zurich, Switzerland, 1996, ACM Press.
- [4] Hani Safadi. Dr. OumaymaDakkak, Dr. Nada Ghneim, Computational Methods to Vocalize Arabic Texts, 2006.
- [5] Jacques Savoy, Yves Rasolofo, and Report on the TREC-11 Experiment: "Arabic, Named Page and Topic Distillation Searches", Jacques Savoy, Yves Rasolofo, 2002.
- [6] Kareem Darwish and Douglas W. Oard, "CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval", 2002.J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.
- [7] Aitao Chen and Fredric Gey," Building an Arabic Stemmer for Information Retrieval", 2002.
- [8] Abdelali, A," Localization in modern standard Arabic, Journal of the American Society for Information Science and Technology, 55, 1, 23-28,2004.

- [9] Suleiman, Y," The Arabic Language and National Identity", Washington, D.C. Georgetown University Press, 2003.
- [10] Elkhafaifi, H.M," Arabic language planning in the age of globalization, Language Problems and Language Planning", Vol. 26, No. 3, 253-269,2002.J. Jones. (1991, May 10). Networks (2nd Ed.) [Online]. Available: <http://www.atm.com>
- [11] Beesley, Kenneth R. "Arabic Finite-State Morphological Analysis and Generation", In COLING the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, pp. 89-94, 1996.
- [12] Cesa-Bianchi, C. Gentile, and L. Zaniboni Incremental Algorithms for Hierarchical Classification. Journal of Machine Learning Research, 7:31–54, 2006.
- [13] PurviPrajapati, Amit Thakkar, Amit Ganatra,"A Comprehensive and Comparative Study on Hierarchical Multi Label Classification", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-2, February 2012.