Developed Taxonomy for Information Retrieval Systems Based on Arabic Language

Mohamed Abdelhadi, Tiruveedula Gopi Krishna, Ghassan Kanann

Abstract— This research has introduced a fully developed Taxonomy for Information Retrieval System based on Arabic Language; and also studied the rationale behind Information Retrieval from Arabic Linguistics with respect to new prospective for Arabic-IRs in such well-done Data organization. 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.

Index Terms— Arabic Language, Information Retrieval Systems, Knowledge Management Systems.

I. INTRODUCTION

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 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 has dealt directly within Arabic language morphologies; cross-languages or computational linguistics; NLP as well, rather than other important issues like the language generality [Elkhafaifi, H.M. 2002] 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
Taxonomy should bring new great addition as state-of-the-art research for the Science of Information Retrieval Systems.

II. 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 taxonomy for computational Methods used to enhance the Arabic Retrieval Systems optimality. Oumayma Dakkak and et.al [2006] has 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. Jacques Savoy and Yves Rasolofo [2002] 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. Kareem Darwish and Douglas Oard, [2002] 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. Aitao Chen and Fredric Gey [2002] 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. Paul McNamee, Christine Piatko, and James Mayfield [2002] they have investigated the use of Support Vector Machines (SVMs) to tackle text filtering challenges.

III. A DEVELOPED TAXONOMY FOR ARABIC-IRS

In our developed Arabic-IRs- Taxonomy 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 [Beesley, Kenneth R. 1996]. The developed Taxonomy is totally clear to understand and is as organized as shown below in Fig1.

Taxonomy for Arabic-IR-System

Manuscript received January 15, 2014.

Mohamed Abdelhadi, Computer Department, Sirt University, Faculty of Arts and Science, Hoon, Al-jufra, Libya.
Tiruveedula Gopi Krishna, Computer Department, Sirt University, Faculty of Arts and Science, Hoon, Al-jufra, Libya.
Ghassan Kanann, Prof. Dr. Ghassan Kanann, Yarmouk University at Irbed in Jordan.

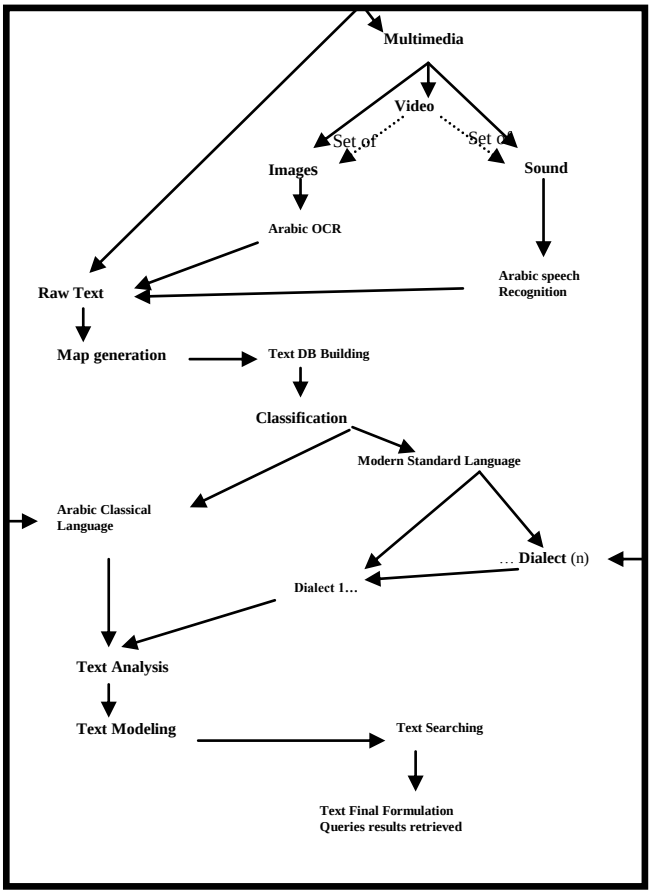


Fig. 1. A developed Taxonomy for Arabic Information Retrieval System

IV. A FRAMEWORK FOR ARABIC-IRS

The proposed Framework is used to define the new developed Taxonomy 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), [Suleiman, and Y. 2003] and [Elkhafaifi, H.M. 2002]. In this phase we have used the Hierarchal– Multi-Label-Classification technique, [Juho Rousu, Craig Saunders, Sándor Szedmák, John Shawe-Taylor 2006]; to classify the entire Arabic Words (Filtered Documents) to enhance the System Optimality.

A. Text-Preprocessing

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) [Khoja, S. and Garside, R, 1999]. 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- [Beesley, Kenneth R.1998; Berry, G.1999]; and [Baxter I, Pidgeon, C., & Mehlich, M. 2004]. 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.

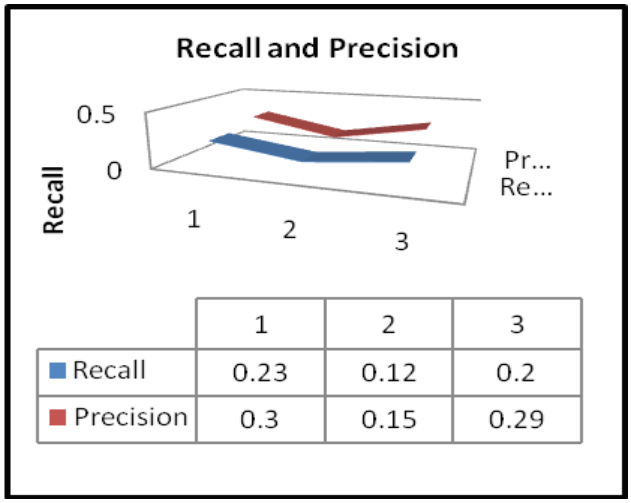
B. Evaluation of Arabic-Text preprocessing

All relevant documents for a user query and only those relevant documents as possible. Many researches [S. E. Robertson, S. Walker, and M. Beaulieu, Okapi 1999], 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 Fig.3 to improve the first phase in system performance.

TABLE- Evaluation of Arabic-Text by Test-Collection

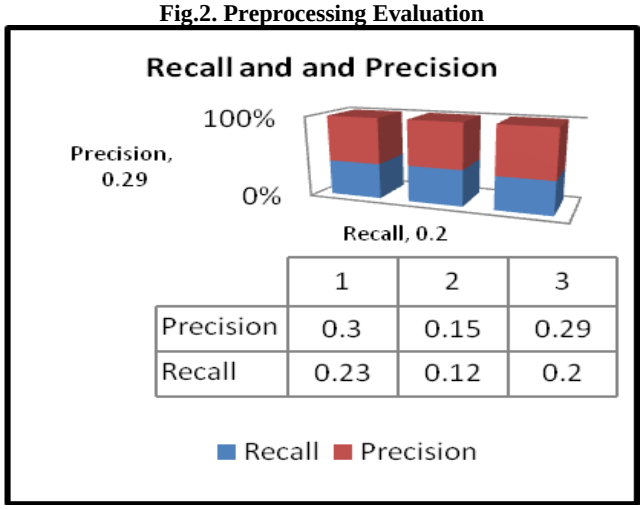
Total Docs in Collection	Total Words in Docs Collection	Query in Words	Word Frequency	Relevant Docs Retrieved	Non Relevant Docs Retrieved	Recall	Precision
39	27128	History	13	9	30	0.23	0.30
39	27128	World	7	5	34	0.12	0.15
39	27128	Arabic	16	8	31	0.20	0.29

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



	1	2	3
Recall	0.23	0.12	0.2
Precision	0.3	0.15	0.29

Fig.2. Preprocessing Evaluation



	1	2	3
Precision	0.3	0.15	0.29
Recall	0.23	0.12	0.2

Fig.3. Recall and Precision Evaluation

V. CONCLUSION

As it has been stated in the proposed Framework which has explained our Taxonomy 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 planed 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.

ACKNOWLEDGMENT

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. Oumayma Dakkak, 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.

AUTHORS PROFILE

Dr. Mohamed Abdeldaiem Abdelhadi received his B.Sc degree in Computer Science from Sebha University at Sebha-Libya in 1988, M.Sc degree in Computer Engineering in 1991 from Humboldt University-Faculty of Teschnische Informatik at Berlin-Germany and PhD degree in Computer Information Systems from the University of Banking and Financial Sciences-Faculty of Information Technology Sciences at Amman-Jordan in 2010; current research interests; Data Mining, Information Retrieval.and 7 Research papers has been published in various reputed and high impact factor cited International Journals.

Tiruveedula GopiKrishna Received B.Sc,M.Sc,M.E,M.Phil from Andhra University,Anna University and Manav bharati Universities in (1997,2001,2004,2010) from India respectively. Currently pursuing Ph.D through Rayalaseema University, India, Registered 2010, and Research interests in Data Mining. Currently working as a faculty of computer science for Sirt University, Libya since 2007.And 12 Research Papers published in various International Reputed high cited impact factor Journals 4 national journal papers published Computer Science and Engineering.

Prof. Dr. Ghassan Kanann, Yarmouk University at Irbed in Jordan, full professor, Faculty Dean, has more than 40 research papers published in many different International Journal in Computer Science.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication