Enhancing Arabic Text Mining Using Linguistic Factors

Mohammed Mahmoud Sakre
High Institute for Computers and Information Systems
Al Shorouk Academy
Cairo – Egypt
m_sakre2001@yahoo.com

Abstract

The World Wide Web overwhelms people with immense amount of widely distributed, interconnected, rich and dynamic hypertext information. Text mining concerns extracting knowledge from unstructured textual data. The most important task to achieve this mission is finding the rules that relate specific words and phrases. This research presents how Arabic morphology and Arabic synonymous, as linguistic factors, can be used to extract the required knowledge from Arabic texts.

The contribution in this research is based on the design and implementation of a system combining morphology, synonyms, indexing and databases for Text Mining and Information Retrieval with different modes regarding morphology and synonyms.

The used approach is based on preprocessing the Arabic text to convert it into semi-structured database. A suitable indexing method and an appropriate searching mechanism are used to extract the required information. The proposed model is tested and it showed a promising success. Shortage in Arabic Computational linguistics tools such as Arabic lexicon tagged with semantic features appeared.

Keywords: *Data Mining, Arabic Text Mining, Arabic Natural Language Processing, Information Retrieval, Information Extraction, Database, Indexing.*

1. Introduction

Text mining, also known as text data mining or knowledge discovery from textual databases, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. Regarded by many as the next wave of knowledge discovery, text mining has very high commercial values [1], [2], [9].

Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning and data mining [9].

Text mining has a growing importance as the volume of unstructured text in web pages, digital libraries and community wide intranets continue to increase. Robb [3] estimated that text documents account about 85% of organizations' knowledge stores.

The general framework for text mining consists of two components: Text refining that transforms free-form text documents into an intermediate form (IF); and knowledge distillation that deduces patterns or knowledge from the intermediate form. Tan [2] described a framework of text mining as in Figure (1). IF can be document-based or concept-based. Knowledge distillation from a document-based IF deduces patterns or knowledge across documents. A document-based IF can be projected onto a concept-based IF by extracting object information relevant to a domain. Knowledge distillation from a concept-based IF deduces patterns or knowledge across objects or concepts. The intermediate form used in this article is a document based one. Information retrieval and information extraction represent the greatest part of the knowledge distillation.

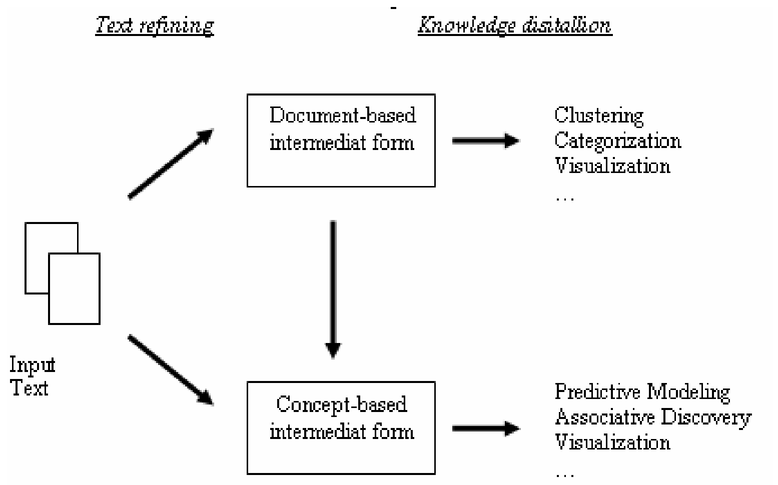


Figure 1: A Text Mining Framework.

Text mining may be used for one or more of the following targets [13],[8],[9] :

- Content analysis
- Business intelligence & competitive analysis of Web sites
- Automatic tagging and classification of documents.
- Taxonomy development and validation.
- Fraud detection, authorship attribution, patent analysis.
- Information extraction and knowledge discovery from texts.

In this article, attention is given to Information extraction and knowledge discovery from Arabic text.

2. Related Work for Arabic Text Mining

Despite of there is few related work for Arabic text mining, the majority of these work falls into the area of automatic Arabic text classification and categorization using different approaches. Examples of these efforts are:

- Al-Harbi research [7]: In this research two popular classification algorithms (SVM and C5.0) has been evaluated on classifying Arabic corpora based on text words. Other features selections should be employed.
- El-Kourdiet et. al. [14] use Naïve Bayes algorithm for automatic Arabic document classification. The average accuracy reported was about 68.78%.
- Sawaf et. al. [15] used statistical classification methods such as maximum entropy to classify and cluster news articles. The best classification accuracy they reported was 62.7% with precision of 50% which is a very low precision in this field.
- El-Halees [1] presents a system for Arabic Text Classification Using Maximum Entropy. This system preprocesses data using natural language processing techniques such as tokenizing, stemming and part of-speech. Then, maximum entropy method to classify Arabic documents is used. The classification accuracy using F-measure reaches 80.41% due to using NLP features like stemming and part of speech.
- Abdulsamad e.t al. [4] presents a research which focused on text mining multilingual datasets including Arabic-English corpus. This work is based on Self-Organizing Map (SOM) and uses Arabic/English corpus as the

test-bed. Issues related to Arabic/English text mining, stemming and clustering are discussed in this research.

Few Arabic text mining researches make use of Arabic natural language processing beside the statistical methods like the research done by Fouzi [10], which is based on using vector space research model and Arabic roots as indexing terms to build a text mining system, called Authentique, for knowledge extraction from a databases of Prophetic Traditions “Hadiths”. This system provides a list of classified hadiths according to their degrees of similarity with respect to user’s query. The Precision and Recall measures of the results of Authentique system is presented as 0.66 and 0.80 respectively. In another research [16] light stemmers based on heuristics and a statistical stemmer based on co-occurrence for Arabic language were developed. Authors claim that the best light stemmer was more effective for cross-language retrieval than a morphological analyzer which tried to find the root for each word.

On the commercial side, Rosette® Base linguistics [12] offers text mining tools and text analysis to work with Arabic text. Also, Sakhr [11] Software Company has developed text mining tools which are based on Arabic natural language processing and which can be used in Arabic texts categorization and summarization. As usual there is no detailed documentation, which explains how these tools work, available for researchers.

3. Challenges in Arabic Computational Linguistics

Arabic is a major and a highly inflected language, and thus requires good stemming for effective text mining. Yet no standard approach to stemming has emerged [4], [6].

In this research, the need for the following Arabic linguistics tools appeared:

a- An Arabic morphological analysis/generation tool, to extract keywords roots and generates derivatives. There are many research and commercial systems [5], [11], [17] dealing with Arabic morphology. A previous work done by the author [5] is considered as a base to build the morphological analysis/generation module of the proposed system in this article.

b- An Arabic lexicon supported with semantic features, to support in selecting the proper words generated as derivatives. Such lexicon is not available tell now. So, a small Arabic dictionary supported with semantic tags is built in the proposed system.

c- An Arabic synonym dictionary. A limited synonym dictionary is used.

Regarding the second requirement, up to the author knowledge there is no complete Arabic lexicon tagged with semantic features which is available in the market till the time of publishing this paper.

4. A Proposed System for Arabic Text Mining

The contribution in this research is based on the design and implementation of a system combining morphology, synonyms, indexing and databases for Text Mining and Information Retrieval with different modes regarding morphology and synonyms. There are many researches covering each of the mentioned modules in a separate way but, up to the author's knowledge, there is no research combing all of them in such a manner like the one presented in this research.

The proposed system consists of two main phases. The first phase corresponds to the "text refining" part in Figure (1). It can be called the preprocessing phase. Figure (2) describes this phase. In this phase, Arabic documents are divided into paragraphs; each paragraph is analyzed to extract its keywords. A copy of these paragraphs is kept out of the system to be used in the evaluation as described later. The system contains a morphological module which extracts the root of each keyword. The system builds a two level index. The first level uses the extracted roots to point to the extracted keywords of that root, in the second level index. Keywords, of the second level index, in turn point to the extracted paragraphs. Semi-structured documents are built in a form of indexed database as in figure (3).

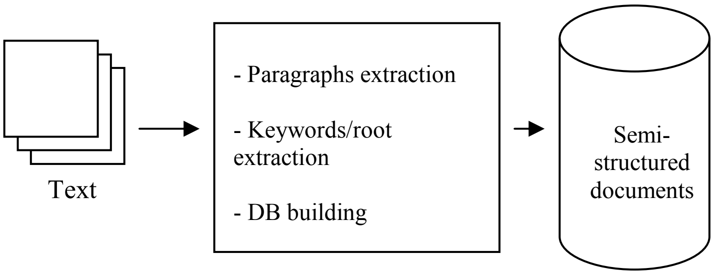
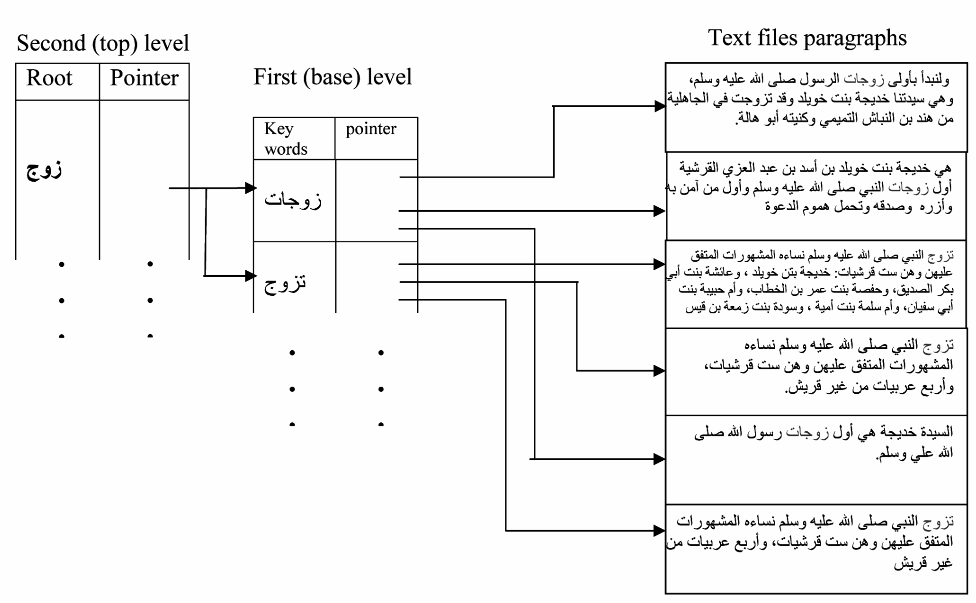


Figure 2 : Preprocessing Phase



ولنبدأ بأولى زوجات الرسول صلى الله عليه وسلم،
وهي سيدتنا خديجة بنت خويلد وقد تزوجت في الجاهلية
من هند بن النباش التميمي وكنيته أبو هالة.

هي خديجة بنت خويلد بن أسد بن عبد العزي القرشية
أول زوجات النبي صلى الله عليه وسلم وأول من آمن به
وأزره وصدقه وتحمل هموم الدعوة

تزوج النبي صلى الله عليه وسلم نساءه المشهورات المتفق
عليهن وهن ست قرشيات: خديجة بتن خويلد ، وعائشة بنت أبي
بكر الصديق، وحفصة بنت عمر بن الخطاب، وأم حبيبة بنت
أبي سفيان، وأم سلمة بنت أمية ، وسودة بنت زمعة بن قيس

تزوج النبي صلى الله عليه وسلم نساءه
المشهورات المتفق عليهن وهن ست قرشيات،
وأربع عربيات من غير قريش.

السيدة خديجة هي أول زوجات رسول الله صلى
الله علي وسلم.

تزوج النبي صلى الله عليه وسلم نساءه المشهورات
المتفق عليهن وهن ست قرشيات، وأربع عربيات من
غير قريش

Figure 3 : Semi-structured Documents Built in a Form of Indexed Database.

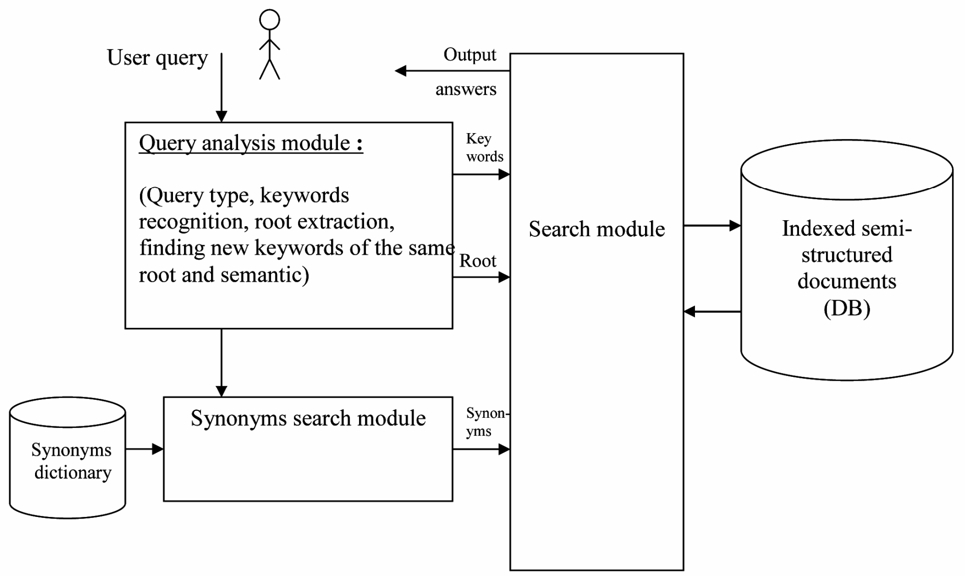

User query
Output
answers
Query analysis module :
(Query type, keywords recognition, root extraction, finding new keywords of the same root and semantic)
Key words
Root
Search module
Indexed semi-structured documents (DB)
Synonyms search module
Synonyms
Synonyms dictionary

Figure 4: Information Extraction Phase.

The structure of the second phase of the proposed system, which concerns with extracting Knowledge/ Information as a response for a user query, is described in Figure (4).

The searching algorithm enables the user to find the related information to his query in many searching modes, these are:

i. **Normal search mode:** In this mode, the searching algorithm seeks for the needed information which contains the exact keyword(s) of the user's query.

ii. **Morphology based search mode:** The system contains a morphological module which extract the root of the each keywords list of the query and then generates all the possible words which have the same root; these are called the derivatives of that root. A group of these derivatives, which have the same semantic features as the original keyword, are selected to be added to the keywords list. So, the number of keywords will increase which yield to an increase in the number of resultant paragraphs which means expecting better Recall measure [6]. To achieve this searching mode an Arabic dictionary supported with semantic tags is used. In the proposed model a small dictionary is built. The morphological module is based on the work done by Ibrahim [5] where PROLOG language is used to build that module.

iii. **Search with query keywords, synonyms and derivatives with same semantic features:**

The searching algorithm searches a Synonym Dictionary, looking for synonyms of the keywords list of the query, and adds them to the list. Derivatives of the original list, like the same result of the previous search mode, are added to the list as well. The resultant list is used to find the related paragraphs in the database.

5. Results and Evaluations:

Since there is no authority supporting research within the Arabic Information Retrieval and Text Mining community by providing an infrastructure necessary for large-scale evaluation of text retrieval methodologies, the proposed model is tested using an Arabic text book about the history of the prophet Mohammed (PBUH) named (الرحيق المختوم) and limited size dictionaries .

Information retrieval systems are usually compared on the basis of the "quality" of the retrieved document sets. This "quality" is traditionally quantified using two metrics, Recall (R) and Precision (P) [6]. Recall and Precision can be defined as:

$$R = r / K \quad \ldots\ldots(1) \qquad P = r / N \quad \ldots\ldots(2)$$

Where : r = The number of relevant and retrieved paragraphs.

N = The total number of retrieved paragraphs.

K = The total number of paragraphs in the answer key.

For testing and evaluation purposes, the answer key for each submitted question to the system is extracted manually from the saved copy of the paragraphs of the original text.

The measures of Recall and Precision [6] are used to evaluate the system results where:

- Recall1 & precision1 refers to the recall and precision of the system using the exact keyword(s) of the user's query mode.
- Recall2 & precision2 refers to the recall and precision of the system using query keywords and derivatives with same semantic features mode.
- Recall3 & precision3 refers to the recall and precision of the system using query keywords, synonyms and derivatives with same semantic features mode.

Many questions are used to evaluate the proposed system. The respond of the system for each question is compared with the corresponding answer key. The values of Recall (R) and Precision (P) are calculated and explained in the next figures.

Figure (5) and figure(6) show comparing results of the Recall and Precision of the three modes respectively.

It is noticed from figure (5) that Recall of the first mode (Recall1) is less than the second and third modes (Recall2 and Recall3). This is due to the existence of paragraphs, in the database, which include synonyms or derivatives of the keywords and which are not extracted in the first mode. Also, it is noticed from figure (5) and figure (9) that Recall3 is near to 100% because the synonyms and derivatives of the keywords are extracted and are used.

Figure (7), figure(8) and figure(9) show comparing results of the Recall and Precision for each of the three modes correspondingly.

It is noticed in Figure (8) that using derivatives enhanced recall while precision became worse. This is because the used derivatives of keywords existing in relevant and extracted paragraphs enhance the Recall, while those exist in irrelevant and extracted paragraphs negatively affect precision.

It is noticed in figure(9) that recall is greatly enhanced while precision became worse. This is because the used derivatives and synonyms of keywords sometimes exist in relevant paragraphs, so recall is enhanced, and other times exist in irrelevant paragraphs which when extracted negatively affect precision.

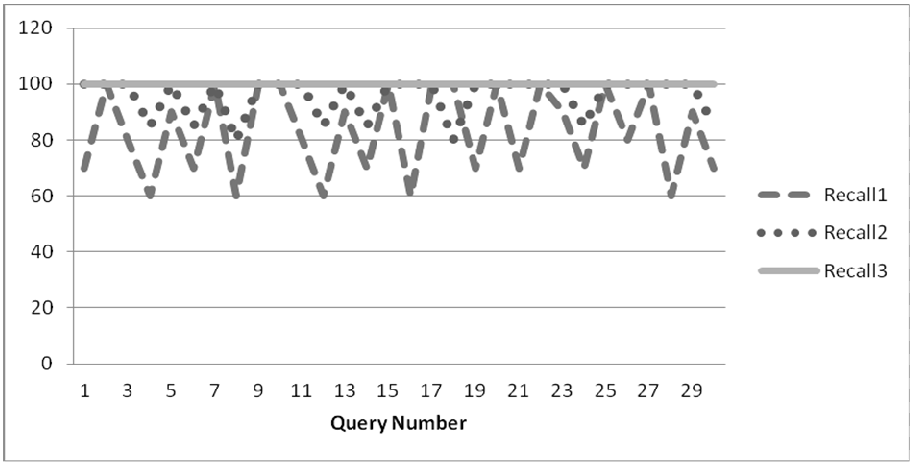


Figure 5 : Recall of the Three Modes

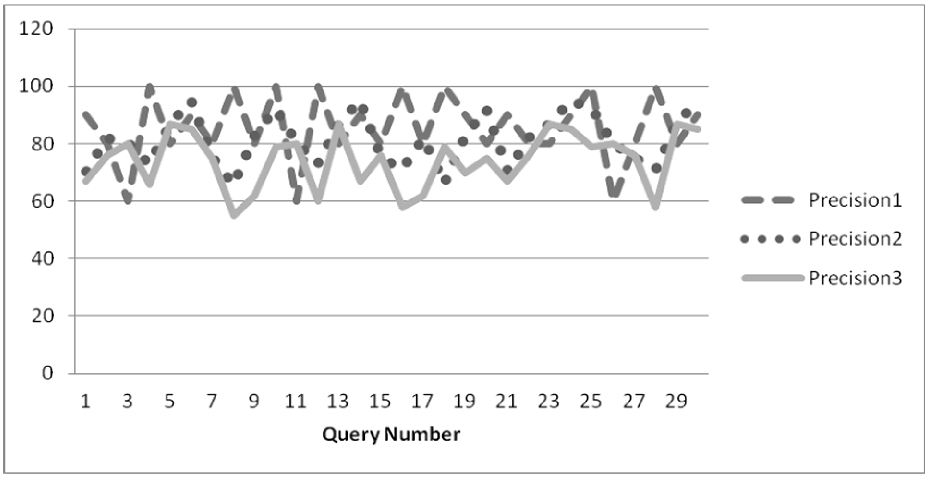


Figure 6 : Precision of the Three Modes

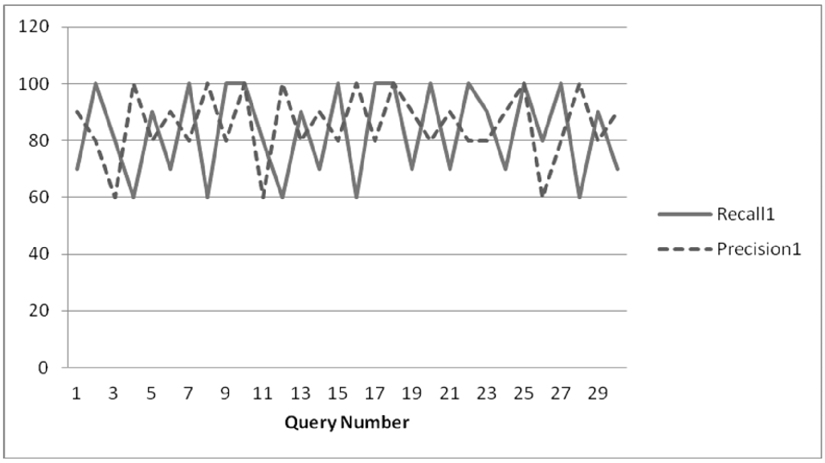


Figure 7 : Precision and Recall of the Keywords Only Mode

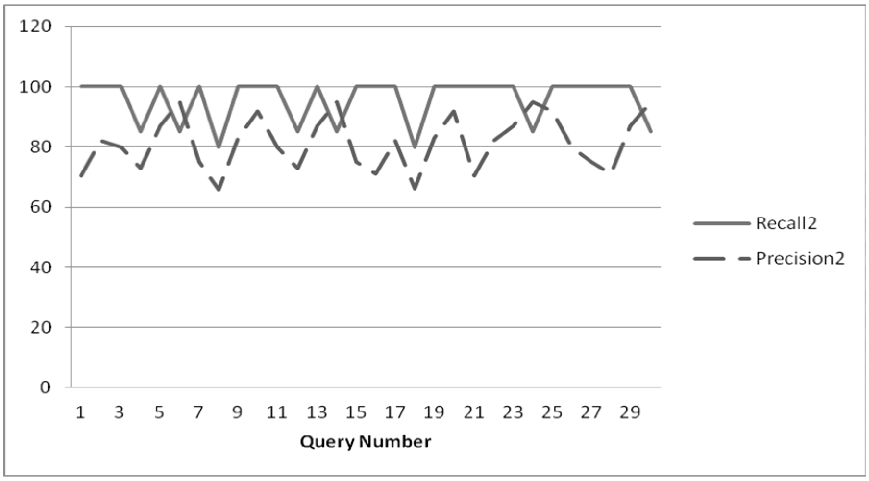


Figure 8: Precision and Recall of the Keywords and Derivatives with Same Semantic Features Mode.

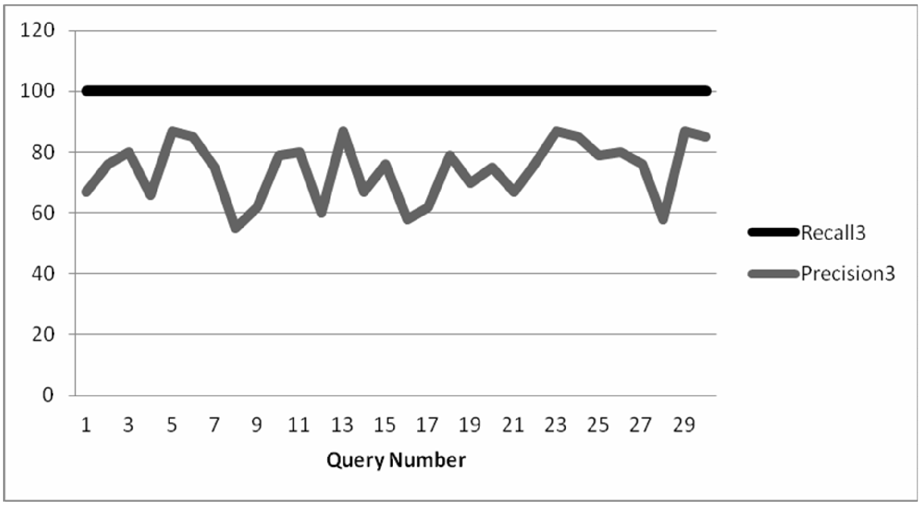


Figure 9 : Precision and Recall of the Keywords , Synonyms and Derivatives with Same Semantic Features Mode.

The comparison between the proposed system outputs, of different modes, was done based on the same data set.

Table(4) presents a comparison between the average results of the proposed model and the results of another Arabic text mining system "Authentique" [10], which is based on using vector space research model and Arabic roots as indexing terms. Due to the unavailability of the "Authentique" system, the scores of its Recall and Precision are obtained from the published article describing the system [10],

Table 4: Comparing Results

The other system "Authentique"		**The proposed model**					
		Keywords only		Keywords + derivatives		Keywords + derivatives + synonyms	
Recall	Precision	Average recall	Average precision	Average recall	Average precision	Average recall	Average precision
80%	66%	85%	83%	95%	90.33%	100%	73.2%

The author believes that this comparison with "Authentique" system is not fair for both models, since different test beds and different queries are used while they should be similar. A sort of standardization like the approach used by TREC[1] is recommended to help in comparing information retrieval systems.

6. Conclusions and Future Work

The results of the proposed model show a promising success in the field of Arabic text mining by using Arabic synonyms and derivatives, with same semantic, as linguistic features, supported with a suitable indexing method. The used approach is based on preprocessing the Arabic text to convert it into semi-structured database. A two level indexing method and a three modes searching mechanism are used to extract the required information.

However, more efforts are still required to build Arabic lexicon tagged with semantic features and to be available for scientific researches. It is recommended for the Arab community who have interest to achieve advances in the field of text mining and information retrieval to establish a standardization like the one supported by TREC[1] for Latin languages to help in evaluating researches' results.

References :

[1] Alaa M.El-Halees, "*Arabic Text Classification Using Maximum Entropy*", The Islamic University Journal (series of natural studies and engineering) Vol. 15, No.1,pp 157-167, (2007).

[2] Ah- Hwee Tan, "Text mining : *The state of the art and the challenges*", In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (1999).

[3] Robb, D., *Text mining tools take on unstructured information. Computerworld*, 21 June (2004).

[1] *Since 1999, the TREC (Text REtrieval Conference) series organized by the US National Institute of Standards and Technology (NIST) has provided a forum for comparative evaluation of question answering (QA) technology.*

[4] Abdulsamad Al-marghilani, Husien Zedan and Aladdin Ayesh , *"A general framework for multilingual text mining using self-organizing maps"*, Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications, Innsbruck, Austria, Pages: 520 – 525, Year of Publication: 2007.

[5] Ibrahim M. M., "*Information Retrieval of Arabic text using A.I. Techniques*", M.Sc. thesis, Military Technical College, Cairo, Egypt, 1987.

[6] Chowdhury, G. "*Introduction to modern information retrieval*", second edition, Facet publishing, 2004.

[7] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity ,M. S. Khorsheed and A. Al-Rajeh, "*Automatic Arabic Text Classification*", JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles, Pages 77-83.

[8] Raymond J. Mooney and Un Yong Nahm, "*Text Mining with Information Extraction*", Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005.

[9] Ronen Feldman and James Sanger, " *The Text Mining Handbook*: Advanced Approaches in Analyzing Unstructured Data ", Cambridge University Press,2006

[10] Fouzi Harrag and Aboubekeur Hamdi-Cherif, "*UML modeling of text mining in Arabic language application to the prophetic traditions*", The first international smposium on computers and Arabic language, Riyadh, 2007.

[11]http://www.sakhr.com/products/Mining/Default.aspx?sec=Product&item=Mining", 1-12-2008.

[12] http://www.basistech.com/base-linguistics/, 5-12-2008.

[13] "http://www.provalisresearch.com/wordstat/wordstat.html" 6-12-2008.

[14] El-Kourdi, M., Bensaid, A., Rachidi, T., *Automatic Arabic Document*

Categorization Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics. August 28th. Geneva (2004).

[15] Sawaf, H., Zaplo, J., Ney, H., *Statistical Classification Methods for*

Arabic News Articles. Arabic Natural Language Processing, Workshop

on the ACL'2001. Toulouse, France, July (2001).

[16] Leah S. Larkev et al. "*Improving stemming for Arabic information retrieval*: light stemming and co-occurrence analysis", SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval Pages 275 - 282, New York, NY, USA, 2002

[17] “http://www.rdi-eg.com/technologies/Morpho.aspx”, 1/3/2012