Text Mining Using Natural Language Processing

Dr. Emad S. Othman

Abstract— The World Wide Web today has a massive amount of widely distributed, interconnected, rich and dynamic hypertext data. One of the Text mining objectives is to extract knowledge from unstructured textual data. The contribution in this research is to design and implement 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 indexed database. A suitable indexing method and an appropriate searching mechanism are used to extract the required information. The proposed model is evaluated and it showed Average recall between 85% and 100% depending on the selected mode. It also showed average precision between 73.2% and 90.33% and average F-measure of 87.03.

Index Terms—Arabic, Text Mining, Natural Language Processing, Information Retrieval, Information Extraction, computational linguistics

1 Introduction

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 [1] estimated that text documents account about 85% of organizations' knowledge stores. The main structure of text mining is composed 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 [2]. 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 research is a document based one. Information retrieval and information extraction represent the greatest part of the knowledge distillation. Text mining has many target applications [3-5]. In this research, attention is given to Information extraction and knowledge discovery from Arabic text.

2 Related Work

The majority of the work of Arabic text mining falls into the area of automatic Arabic text classification and categorization using different approaches. Examples of these efforts are:

- Al-diabat [6] investigates the problem of Arabic text categorization using different rule-based classification approaches in data mining.
- Froud et. al. [7] studied the semantic dependencies between words expressed by the co-occurrence frequencies of these words. They found that the Stem-based approach outperformed the Root-based one because the latter affects the words meanings.
- Al-Harbi research [8] has evaluated two popular classification algorithms (SVM and C5.0) on classifying Arabic corpora based on text words. Other features selections should be employed.
- El-Kourdi et. al. [9] use Naïve Bayes algorithm for automatic Arabic document classification. The average accuracy reported was about 68.78%.
- Sawaf et. al. [10] 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 [11] present 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. [12] 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.

Rare Arabic text mining researches make use of Arabic natural language processing beside the statistical methods like the research done by Fouzi [13], which is based on using vector space research model and Arabic roots as indexing terms to build a text mining system. In another research [14] light stemmers based on heuristics and a statistical stemmer

• ***Dr. Emad S. Othman**, Senior Member IEEE - Region 8, lecturer in High Institute for Computers and Information Systems, AL-Shorouk Academy, Cairo – Egypt, PH- 002-01025830256. E-mail: emad_osman67@yahoo.com*

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 [15] offers text mining tools and text analysis to work with Arabic text. Also, Sakhr [16] 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 Arabic Computational Linguistics

Arabic is the largest member of the Semitic language family and is spoken by nearly 500 million people worldwide. It is one of the six official UN languages. Despite its cultural, religious, and political significance, Arabic has received comparatively little attention in modern computational linguistics. Arabic Morphology has a complicated structure, and has a great effect on Arabic syntax and semantic. To deal with Arabic Morphology a good stemming algorithm is required for any effective computational linguistic system and an Arabic morphological analysis/generation tool, to extract roots of the words and generates derivatives from roots. There are many researches [17-20], [23] and commercial products [15-16], [21] which handled Arabic morphology.

For example, Keok[23] presented a model that prove that words with common affixes are likely to be in the same syntactic category and uses learned syntactic categories to refine the segmentation boundaries of words. Yet no standard approach to stemming has emerged [12]. Keskes [24] investigates the feasibility of Arabic discourse segmentation into elementary discourse units within the segmented discourse representation theory framework. The need for an electronic Arabic lexicon supported with semantic features and an Arabic synonym dictionary which are not available and they are essential, are perceived.

4 The 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, rare of these researches combine many 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. It can be called the preprocessing phase. Figure (1) describes this phase where 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, each Keyword, in turn, points to the related paragraphs. Semi-structured documents are built in a form of indexed database as in figure (2).

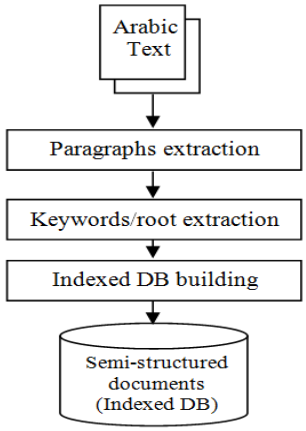


Figure (1) Preprocessing phase

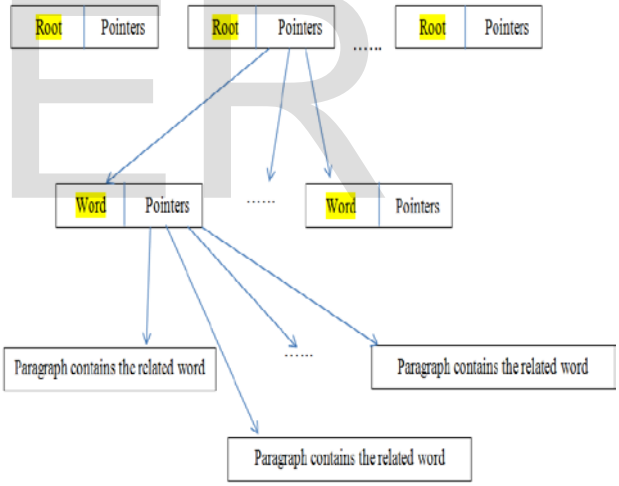


Figure (2) Semi-structured documents built in a form of indexed database.

The structure of the second phase of the proposed system, which concerns with extracting information as a response for a user query, is described in Figure (3).

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

- **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.
- **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 [22]. 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 [20] where PROLOG language is used to build that module.

- **Search with query keywords, synonyms and derivatives with the 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.

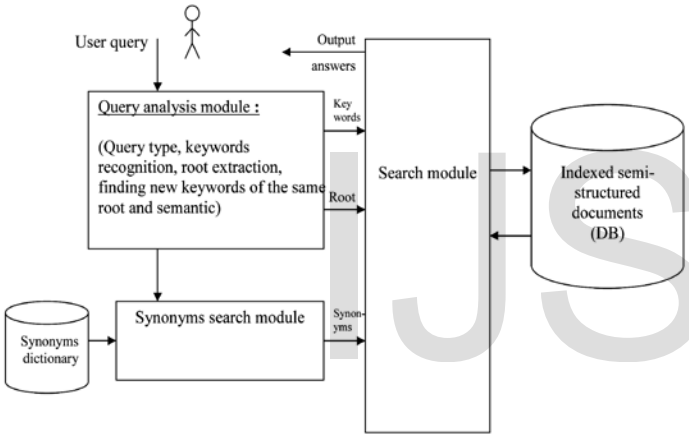


Figure (3) Information extraction phase.

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) [22]. Recall and Precision can be defined as:

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

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.

The F-measure refers to the combination between precision and recall, represented in equation (3).

$$F-measure = \frac{2*Recall*Precision}{Recall+Precision} \quad \ldots. \quad (3)$$

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 are used to evaluate the system results where:

- Recall1, precision1 and F1 refer to the recall, precision and F-measure of the system using the exact keyword(s) of the user's query mode.
- Recall2, precision2 and F2 refer to the recall, precision and F-measure of the system using the query keywords and derivatives with same semantic features mode.
- Recall3, precision3 and F3 refer to the recall, precision and F-measure 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 (4) and figure (5) show comparing results of the Recall and Precision of the three modes respectively. It is noticed from figure (4) 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 (4) and figure (8) that Recall3 is near to 100% because the synonyms and derivatives of the keywords are extracted and are used.

Figure (6), figure (7) and figure (8) show results of the Recall and Precision for each of the three modes correspondingly. It is noticed in figure (7) 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 (8) 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. The comparison between the proposed system outputs, of different modes, was done based on the same data set.

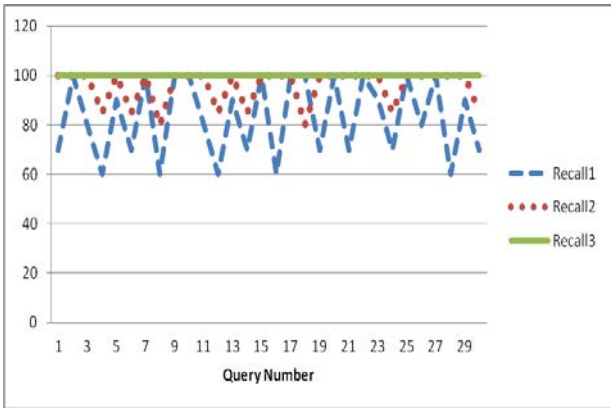


Figure (4): Recall of the three modes

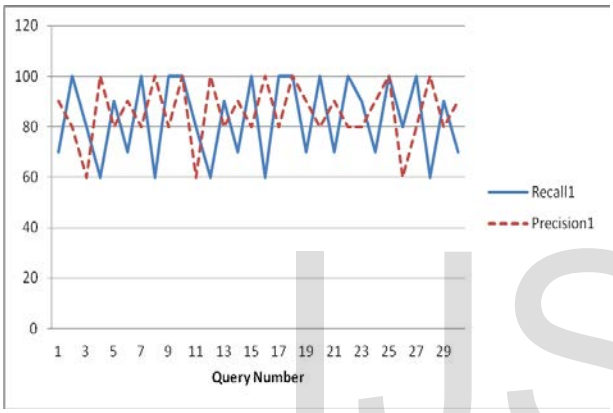


Figure (5) Precision of the three modes

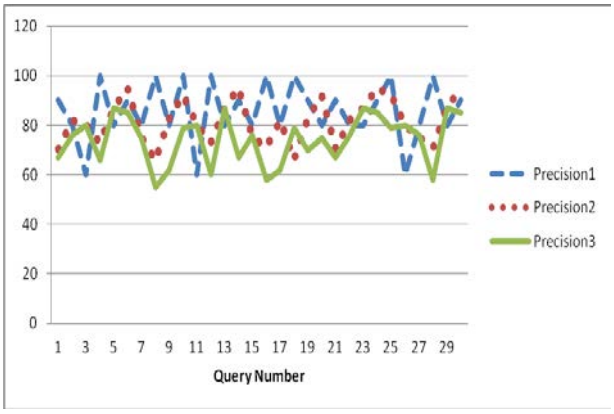


Figure (6) Precision and Recall of the keywords only mode.

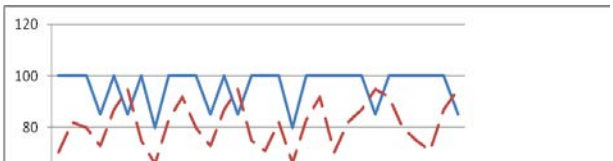


Figure (7) Precision and Recall of the keywords and derivatives with same semantic features mode

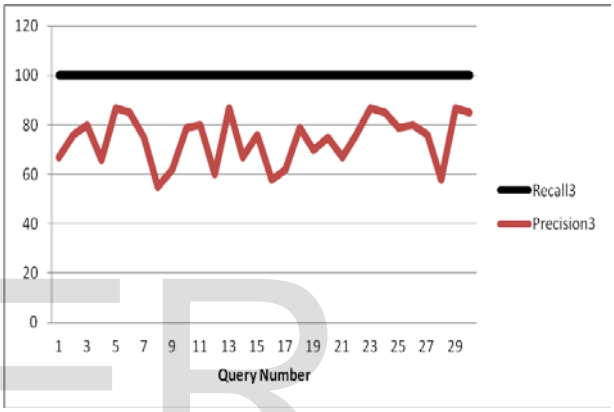
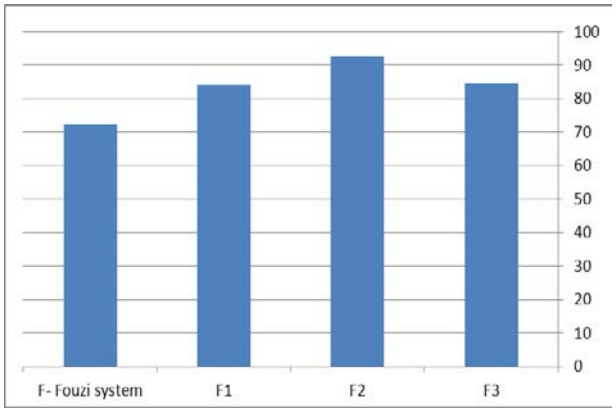


Figure (8) Precision and Recall of the keywords, synonyms and derivatives with same semantic features mode.



Figure(9) Comparing results with Fouzi system

The average results of the proposed model is compared with the results of another Arabic text mining system Fouzi [13] which is based on using vector space research model and Arabic roots for knowledge extraction from a database of Prophetic Traditions "Hadiths". The Precision and Recall measures of the results of Fouzi system is presented as 0.66 and 0.80 respectively. Figure (9) shows the comparing results of the average F-measures of Fouzi system and the three modes of the proposed system. The figure shows better results for the three modes than Fouzi system. This comparison with "Fouzi 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 TREC1 is recommended to help in comparing information retrieval systems.

6 Conclusions and Future Work:

This paper presents 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 composed of a preprocessing phase and a run time phase. During the first phase the Arabic text is processed to convert it into semi-structured database. A two level indexing method is used. In the second phase three modes searching mechanism is used to extract the required information.

The results of the proposed system show F-measure values of (83.98, 92.60, and 84.52) for the three system modes and compared with Fouzi system (F-measure 72.32). The comparison results can be considered as a promising success in the field of Arabic text mining. However, more efforts are still required to build Arabic lexicon tagged with semantic features and to be available for scientific researches.

7 Acknowledgment

The author wish to thank Prof. Dr. Mohammed M. Sakre in AL-Shorouk Academy, Cairo – Egypt for his effective support.

8 References

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

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

[3] http://provalisresearch.com/ Visited in 10/6/2015.

[4] 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.

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

[6] Mofleh Al-diabat, " Arabic Text Categorization Using Classification Rule Mining ", Applied Mathematical Sciences, Vol. 6, 2012, no. 81, 4033 – 4046.

[7] Hanane Froud et al. "A comparative study of root-based and stem-based approaches for measuring the similarity between arabic words for arabic text mining applications", Advanced Computing An International Journal (ACIJ), November 2012, Volume 3, Number 6.

[8] 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.

[9] El-Kourdi, M., Bensaid, A., Rachidi, T., Automatic Arabic DocumentCategorization Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics. August 28th. Geneva (2004).

[10] Sawaf, H., Zaplo, J., Ney, H., Statistical Classification Methods forArabic News Articles. Arabic Natural Language Processing, Workshop on the ACL'2001. Toulouse, France, July (2001).

[11] 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).

[12] 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.

[13] 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.

[14] 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

[15] http://www.basistech.com/ text-analytics/ rosette/ base-linguistics. Visited in 10/6/2015.

[16] http://www.sakhr.com/index.php/en/technology/arabic-resources. Visited in 10/6/2015.

[17] Sarah Alkuhlani et al. "Automatic Morphological Enrichment of a Morphologically Underspecified Treebank", Proceedings of NAACL-HLT 2013, pages 460–470, Atlanta, Georgia, 9–14 June 2013. Association for Computational Linguistics.

[18] Mourad Gridach et al. " Developing a New System for Arabic Morphological Analysis and Generation" Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, pages 52–57.

[19] Michelle A. Fullwood et al. "Learning non-concatenative morphology", Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 21–27, Sofia, Bulgaria, August 8, 2013.

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

[21] http://www.rdi-eg.com/technologies/arabic_nlp.htm Visited in 10/6/2015.

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

[23] Yoong Keok et al. "Modeling Syntactic Context Improves Morphological Segmentation", Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 1–9, Portland, Oregon, USA, 2011.

[24] Iskander Keskes et al. " Splitting Arabic Texts into Elementary Discourse Units". ACM Transactions on Asian Language Information Processing, Association for Computing Machinery (ACM), 2014, vol. 13 , pp. 1-23.

[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.*