State of the Art in Semantic Web Search Techniques for Arabic Language

[1] Ruqaia Jwad, [2] Dr.Norita Md Norwawi , [3] Bala Musa

[1, 2, 3] Faculty of Science and Technology, Universiti Sains Islam Malaysia, Bander Baru Nilai, Malaysia

Abstract

Arabic language has many differences from English language in terms of morphology and semantic. These areas of difference make it somehow difficult when it comes to web search in Arabic. Unlike Arabic language, other languages including Latin have substantiated amount of research in the use of semantic technologies in processing text and information retrieval. Despite the complexity in Arabic script, some significantly contribution has been made in the area of retrieval algorithms and semantic web techniques which can be measured in terms of the accuracy. This paper therefore, examines the state of the art in the use of semantic web search techniques for the retrieval of Arabic text.

Keywords: *Semantic Web, Arabic Ontology, Natural Language Processing, Arabic Search.*

1. Introduction

There are an increasing number of electronic documents in Arabic language on the web daily. This is due to the awareness for information technology that is conversely growing among the people of predominantly Arabic origin with the growing population This document is set in 10-point Times New Roman. If absolutely necessary, we suggest the use of condensed line spacing rather than smaller point sizes. Some technical formatting software print mathematical formulas in italic type, with subscripts and superscripts in a slightly smaller font size. This is acceptable. Despite the growing trend, processing Arabic language is at the premature stage compared to other languages in terms of significant in the domain. Some of the reasons responsible are complexity, derivational, and inflectional as highlighted by (Abu-Hamdiyyah and ebrary 2000). Other reasons as identified by (Koivunen 2001) include the ambiguity associated with Arabic script such as vowels omission (Zaidi, Laskri et al. 2005), replacement of some characters with others. Also, because of the absence of capitalization that separate names from other verbs as in English language, the Arabic script retrieval is posing a big challenge to sharing of knowledge. Therefore, there is need for data to be shared, retrieved, understood, and manipulated by a tool using various techniques. Semantic Technique according to (Koivunen 2001) is "an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

2. Related Work

In this review, we identify the various works relating to techniques employed in searching and handling Arabic text. We categorized the techniques based on machine learning, non-machine learning and combine hybrid approach (machine and non-machine).

2.1 Machine Learning Technique

The use of machine learning technique has been employed in the past years to handle Arabic text. Some of these works include the proposed hybrid approach by (Selamat and Ng 2011) where decision tree and ARTMAP techniques are used to identify Arabic web page. In their approach, a decision tree was first deployed to retrieve the web page regardless of the language, then an ARTMAP approach is used afterwards to classify the Arabic page from non-Arabic page. The proposed identification approach DT-ARTMAP which represents the combination of both Decision Tree and ARTMAP was experimented and based on the authors' conclusion; there is increase reliability apart from accuracy and noise reduction and precision of recall rate. Similarly, (El-Beltagy and Rafea 2009) uses KP-miner approach without the need for training of any document in order to extract key phrases from Arabic and English with the ability to perform some configurations as rules. The KP-Miner system was evaluated with other key phrase extracting system such as (Kanungo, Marton et al. 1999) system and from the result of the evaluation it shows that the number of times an article title was recognized as the highest ranking key phrase is significantly higher than the number of times the (Kanungo, Marton et al. 1999) system recognized the title as a key phrase. In a similar vein, (El Kourdi, Bensaid et al. 2004) uses the Naïve Bayes machine learning technique to extract Arabic web

documents by reducing the Arabic web text to its canonical form known as root form and then classify it to a predefined category. According to the authors, the result from the experiment conducted has shown that Naïve Bayes algorithm classification of Arabic documents is not directly sensitive to the Arabic root extraction algorithms due to the variance obtained during the cross validation experiment.

2.2 Non Machine Learning Technique

In the similar vein, (Al-Shammari and Lin 2008) proposes an Arabic Lemmatization Algorithm for better word normalization method for Arabic text. In the proposed lemmatization algorithm which points to the fact that Arabic neglected stop words can be highly important and can provide a significant improvement to processing Arabic documents. The algorithms was evaluated with other stemming algorithms such as (Khoja 2001) stemming algorithms, it was found that (Khoja 2001) algorithms may have some stemming error than the lemmatization approach according to the authors.

Further to the use of non-machine learning, (Goweder, Poesio et al. 2004) proposes a basic light stemmer that removes suffixes and prefixes from Arabic word to reduce the original stem of the word thereby making it easier and efficient to identify broken plurals. The basic light stemmer test result according to the authors has shown that reducing broken plural to their original stem results in significant improvement in information retrieval.

In another similar approach of trying to handle Arabic text using non machine learning technique, (Al-Radaideh and Masri 2011) proposes a remapping and bi-gram approach to Arabic mobile multi-tap texting entry. The remapping approach distributes Arabic letters on the keypad according to the frequency of letters while the bi-gram based method was used to predict the next letter to be typed on the screen automatically after the user enters the first letter. A letter bi-gram based model is used to make text entry more efficient and faster by predicting the next letter to be typed during writing an SMS. According to the authors, the result of the test has shown a good improvement by limiting the keystroke required to input in a message.

2.3 Hybrid Approach

In terms of the combining both machine learning and other non-machine learning technique, (Isbaitan and Al-Wahidi 2011) proposed a Web Ontology Language OWL which is an extended graph model from RDF model that aids in the building of Arabic vocabularies and software logics. According to the authors, the hybrid of the two techniques provides for the creation of an object-oriented model which connects both RDF triples to classes, associations, and other complex relationships. Similarly, using a rule based machine learning approach and linguistic grammar base technique in combination, (Shaalan and Raza 2007) developed a system for recognizing person names entities in Arabic language. The system which the author refers to as Person Name Entity Recognition for Arabic (PERA) provides flexibility and adaptability features that can be easily configured to work with different languages. The system was evaluated with some corpus data which according to the author, the results achieved were satisfactory and confirm to the targets set forth for the precision, recall, and f-measure.

3. Discussion

This work identified the various technique used to classify Arabic documents ranging from machine learning, non machine learning and hybrid techniques. Many of the works on machine learning techniques are centered on the use of decision tree, KP-miner and Naïve Bayes algorithms to classify Arabic text. Similarly, works on non machine learning focuses on lemmatization algorithms, light stemming and remapping and bi-gram approach. Other works combined some machine and non-machine effort such as web ontology language that combines with RDF to classify Arabic text and Person Name Entity Recognition for Arabic (PERA) which is used for feasibility and ease of classification.

4. Conclusion

Enormous effort has been put in place to facilitate efficient and effective use of techniques for Arabic language text retrieval. These efforts are still not standardized and still huge gap is yet to be filled in terms of Arabic retrieval. Therefore, there is need for an efficient and effective approach that will incorporate the best techniques and eliminate the current impediments associated with Arabic text retrieval.

References

.

[1] M.Abu-Hamdiyyah, and I. ebrary. The Qur'an: an introduction, 2000,Taylor & Francis.

[2] Q. A Al-Radaideh,. and K. H. Masri. Improving mobile multi-tap text entry for Arabic language. 2011 Computer Standards & Interfaces 33(1): 108-113. Al-Shammari, E. and J. Lin 2008. A novel Arabic lemmatization algorithm, ACM.

[3] S. R.El-Beltagy, and A. Rafea. KP-Miner: A keyphrase extraction system for English and Arabic documents. 2009. Information Systems 34(1): 132-144.

[4] M.El Kourdi, , A. Bensaid, et al. Automatic Arabic document categorization based on the Naïve Bayes algorithm. 2004.

[5] A.Goweder, , M. Poesio, et al. Broken plural detection for arabic information retrieval. 2004. ACM.

[6] O.Isbaitan, and H. Al-Wahidi Arabic model for semantic web 3.0. Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications. 2011.Amman, Jordan, ACM: 1-6.

[7] T.Kanungo, , G. A. Marton, et al. OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products, 1999.SPIE INTERNATIONAL SOCIETY FOR OPTICAL.

[8] S.Khoja, APT: Arabic part-of-speech tagger. 2001.

[9] M. R.Koivunen. W3C semantic web activity. Semantic Web .2001.KickOff in Finland: 27-41.

[10] Nana Yaw Asabere, Nana Kwame Gyamfi, AIDSS-HR: An Automated Intelligent Decision Support System for Enhancing the Performance of Employees, arXiv:1307.8335

[11] A.Selamat, and C. C. Ng. Arabic script web page language identifications using decision tree neural networks. 2011. Pattern Recognition 44(1): 133-144.

[12] K.Shaalan, and H. Raza . Person name entity recognition for Arabic, Association for Computational Linguistics. 2007.

[13] S Zaidi,., M. Laskri, et al. A cross-language information retrieval based on an Arabic ontology in the legal domain. 2005.