

Arabic Language Technologies: a Survey

Abdulelah Ibrahim Almishal

National Center for Electronics, Communications and Photonics
King Abdulaziz City for Sciences and Technology
Riyadh, Saudi Arabia
aalmishal@kacst.edu.sa

Abstract— Currently, Arabic language is one of the most languages used in the world after English and Chinese. This popularity comes from the hundreds of millions of people speaking Arabic as a mother language. However, from technology perspective and compared to other languages, there is lack of technology, tools and applications have been implemented to help and enhance Arabic language and Arabic speakers. This research will look over the recent research work and the literature and explore what have been done in Arabic technologies such as Optical Character Recognition, Arabic translation technologies and Arabic speech Recognition.

Keywords—Arabic Language, OCR, Arabic Speech Recognition, Arabic Translation

1. Introduction

The research will review all the literature written to serve Arabic language including theories, innovations, tools and techniques. Then classify this work into categories such as Optical Character recognition (OCR), Arabic speech recognition, Arabic Translation, Arabic text mining and Arabic machine learning. This research will fill the big gap for researchers who are seeking for any topics related to Arabic language. I've chose the most important areas that rapidly used and taken in the research field.

There are a significant work has been done to enhance Arabic languages technologies. However, this work was scattered. In this research we aim to group this huge work into one single research to be as reference for researchers who might be interest in this kind of topics. In addition, it will include some discussion, evaluation and recommendations of some effort has been done. Also it will enhance some ideas might be interesting and value added.

At the end of this research, we are expected to introduce a full map of Arabic Technologies. This map will improve the awareness about Arabic languages technologies as well as it will guide the researchers in feature.

There is a lack of comprehensive research that cover all work related to Arabic language. As of my knowledge, nobody takes the lead and tries to group what have been done for Arabic language technologies field. However, there some papers focusing on specific topic or specific application related in Arabic such as mentioned in [1, 2, 3 and 4]. We can categorize this work into four main categories. Arabic Optical character recognition, Arabic translation, Speech recognition.

2. Arabic Optical Character Recognition

This field seems to be a mature field; huge work has been done in this area. Before we start in details, it is better to mention some other papers that also reviewed the state of art regarding Arabic Optical Character recognition (AOOCR). In [1], Pervez Ahmed and Yousef Al-Ohali present a comprehensive review of ACR (stands for Arabic Character Recognition) techniques and evaluate the state of art. But this review is out dated as

it published on 2000. However, the challenges might be applicable until these days and the basic algorithms and technologies still used. The authors start with Arabic text characteristics and the factors affecting character design. For both off-line and on-line OCR, The detailed design of OCR systems has been discussed and compared among different papers.

Later on, exactly on 2006, the authors of [2] introduced a full survey about off line handwritten recognition and discussed it from different perspectives such as the pre-recognition methods, segmentation method, Recognition Engine and the main features. They made a full comparison among all OCR systems with deep technical description for each process in OCR system.

Also, Haikal El Abed and Volker Margner [3] present a brief review about the classification techniques in Arabic OCR systems with evaluation of each system. Based on results of this evolution, the authors show how to build an efficient OCR system.

In [4], the authors present a comprehensive survey of recent development in Arabic handwriting recognition. Starting with summary of Arabic text characteristics. Followed by the different models used for recognition, classification and segmentation process, databases used for text, extraction techniques discussed as well. Authors conclude with comparison between different work in terms of database used, extraction method, performance and accuracy rates.

All of the above research work provides survey-based witch similar to our work. However, they are more technical and some of them are out dated.

In this research, we are going to mention the latest research work has been done, Highlight the main technologies, methods and techniques being used. Summarize the advantage and disadvantage for each research work in the field, comparing different work in terms of technologies, algorithm and so on.

We have found about 25 research papers has published in last five years. They are tackling same problem from different application and purposes. Some of them are focused on online recognition while others discussing offline recognition. Some of them try to use a novel method and try to include some Artificial intelligence technologies such as neural networks, machine learning, harmony and genetic algorithms.

In [5], Iping Supriana and Albadr Nasution develop a system called (AOOCR) stands for Arabic Optical Character Recognition, this system has five stages. Preprocessing, segmentation, thinning, feature extraction, and classification. While the system shows a good performance in segmentation stage and classification, unfortunately, the overall performance only reach 48%.

Another OCR system proposed in [6]. Safwa Taha, Yusra Babiker and Mohamed Abbas present an Arabic OCR system based on Time New Roman font. As usual, the paper started with introduction about Arabic text characteristics and brief about OCR systems. During segmentation stage, three levels of segmentation have been made: lines segmentation, sub-words segmentation and character segmentation. Before testing phase, the database of characters generated for different fonts. Also, the authors compare different font family and conclude that Time New Roman font has the best accuracy rate.

Related to font family comparison, authors of [7] present a study on font family and font size applied to Arabic words recognition with ultra-low resolution. It is difficult to identify the word with ultra-low quality scanners. To tackle this problem, authors used Gaussian Mixture models for each popular font with different font size. The result shows high potential of using this method with 99% accuracy rate.

In [8], the authors discuss the segmentation which is the main challenging issue in recognition processes, the paper present a new approach for segmentation. Starting with simplifying the character into single-pixel-thin images. Then, normalizing the images into horizontal and vertical lines. Finally coding this lines as vectors each vector represent a letter. This approach has been tested using 2 different datasets. It can be clearly seen that the accuracy rate exceeding 91%.

In relation to segmentation, in [9], the authors presented another segmentation method touching connected components in unconstrained Arabic handwritten text. The method presented has been tested in large database of contacted text. The results prove the efficiency of this method.

From the above literature that has been mentioned, we can summarize that different work has been conducted for different proposes. We tried to cover all the literature. However, we focused on novel papers that

present new approaches without digging in details. Moreover, we have mentioned some supporting efforts has been done to facilitate the AOCR such as databases. Hope this review was helpful, and value added.

3. Arabic Translation

In this area, a research effort is limited compared with the AOCR. However, the dictionaries either paper-based or electronic-based, translation applications and website are meeting the needs to translation. Also, the space of improvements is limited. In the section, we are reviewing the latest novel ideas related to translation techniques, covering all approaches available from the literature.

As usually, we are starting with the surveys that talking same topic. In [10], authors present a full survey about Arabic machine translation. Comparing different types of translation, different approaches used and discussing their strengths and weaknesses.

Another similar work has been done by T. Hailat and co-authors in [11]. They present an evaluation of effectiveness two popular machine translation systems (Google Translate and Babylon machine translation systems). By using method called Bilingual Evaluation Understudy (BLEU), the results showing more effectiveness to Google translate rather than Babylon.

Machine translation is an automated process. It could be an interesting application area to Artificial intelligent such as Machine learning, neural networks, rule-based algorithms and inductive learning. These AI concepts are used in order to enhance the translation process and improve quality.

In [12], authors introduced an approach for translating structured English sentences to structured Arabic sentences using new rules in order to tackle ordering problem. This approach is flexible and scalable, the main strengths are: firstly, it is an intelligent rule-based approach, and secondly, it's applicable to other languages small changes. The system is designed to be used as a stand-alone tool separated from other systems.

In other hand, authors in [13] presenting an interesting solution to face some challenges attached with rule-based machine translation. One of the main challenges is the difficulty to rule-based to understand words as a human understood. The suggested model (from Arabic to English) tackles the mentioned problem of building translation model. This model employs Inductive Logic Programming (ILP) to learn the language model from a set of example pairs acquired from parallel corpora and represent the language model in a rule-based format that maps Arabic sentence pattern to English sentence pattern. After testing the model on a small piece of data, it generated translation rules with logarithmic growing rate and with word error rate 11%.

Other Artificial intelligent techniques are used in [14], authors are introducing a Transfer Module for an English-to-Arabic Machine Translation System (MTS) using an English-to-Arabic Bilingual Corpus and Artificial Neural Networks (ANN). The idea is to allow the ANN-based transfer module to automatically learn correspondences between source and target language structures using a large set of English sentences and their Arabic translations. After testing the developed module, the result was very encouraging.

In [15], authors adopting Naïve Bayesian Classifier as an approach. This approach consists of two main steps: First, a natural language processing method that deals with reach morphology of Arabic language and second, the translation is including word sense disambiguation. The most important competitive advantage for this approach is the adoption of Naïve Bayesian Classifier method. Moreover, a large parallel corpus has been generated from training corpus. As a result, the system will be tackling translation ambiguity between Arabic and other languages.

Another approach introduced by [16], authors present Injected Tags (ITs) approach that improves the phrase based statistical machine translation (PBSMT) approach. This Injected Tags approach has been applied to "English into Arabic translation". This approach is language independent and can be used with any language pair. It has shown considerable improvement of the translation quality of at least 13% increase of BLEU score. The approach has been evaluated and has been compared with several online Machine Translation (MT) services. The experiments reveal that the results achieved by this approach considered significant enhancements over PBSMT.

In [17], authors propose a hybrid-based system for noun phrase translation. This system combined between rule-based transfer techniques with statistical language model. Targeting noun phrase helps to increase effectiveness of this approach. Firstly, words for each noun phrase are reordered according to a set of rules. Secondly, they

lookup Arabic words in the dictionary and generate a set of English phrase translation candidates, according to how many possible word translations exist for an Arabic word. Finally, the set of translation candidates are ranked using a statistical language model and the translation candidates with the best score is output as the best translation. As a result, the proposed system shows an improvement of translation process in terms of quality and performance.

Also, and related to hybrid approaches, authors in [18] present a hybrid approach for translating from Moroccan Arabic dialect to standard Arabic. As previous paper, this system combining a rule-based approach and a statistical approach, using tools designed for Arabic standard and adapting these tools to Moroccan dialect. According to authors, this is the first translation work concerning the Moroccan dialect.

In [19], authors present English to Arabic approach for translating well-structured English sentences into well-structured Arabic sentences, using a grammar-based and example-translation techniques to tackle the problems of ordering and agreement. This technique combines rule-based machine translation (RBMT) and example-based machine translation (EBMT) which is called hybrid-based MT (HERBMT). This methodology is flexible and scalable. The main advantages of HERBMT are that it combines the advantages of RBMT and EBMT, and it can be applied to other languages with minor modifications. EBMT extracts an example of target language sentences that are similar to input source language sentences. The extraction of appropriate translated sentences is preceded by an analysis stage for the decomposition of input sentences into appropriate fragments. RBMT is used when examples of the Source language to be translated into the target language is not found in the machine database. The OAK Parser is used to analyze the input English text to get the part of speech (POS) for each word in the text as a pre-translation process. The evaluation is done on 250 independent test suites, and the analysis shows that HERBMT achieved good performance with an average of 97.2% precision.

As the previous section discussing AOCR, the authors in [20] introduce an image-based automatic Arabic translation system. This system automatically translates Arabic text embedded in the image into English. The system consists of three main components: text detection from images, character recognition, and machine translation. The detected text images are processed by off-the-shelf optical character recognition (OCR) software. They employ an error correction model to post-process the noisy OCR output, and apply a bigram language model to reduce word segmentation errors. The translation module is tailored with compact data structure for hand-held devices. The experimental results show substantial improvements in terms of word recognition accuracy and translation quality. For example, in the experiment of Arabic transparent font, the BLEU score increases from 18.70 to 33.47 with use of the error correction module.

From the above literature that has been discovered, we can summarize that the Arabic translation field still need improvement in terms of technologies, methods and applications. We attempted to cover all the literature. However, we focused on novel papers that present new approaches without digging in details or rewriting similar work. Hopefully it will be helpful and value-added.

4. Arabic Speech Recognition

One of the most relevant areas to Arabic language is speech recognition. It is enable devices to recognize and understand spoken words. We are going to discover any topic related to Arabic Speech Recognition without digging in the technical signal processing. It's just an overview about techniques, approaches.

As always, we are starting to discuss any survey paper or any comparison that reviews the literature.

In [21], authors present a comparative study is performed in two different ways. The first includes two tests. The first test is an objective test where the computer has to recognize the recorded data. The second test is a subjective test where 15 persons judged the recognition process for the tested materials. The second form of comparison is performed for different transmission media, acoustical speech (direct), via telephone (PSTN, PBX), Wireless (cellular), Internet (VoIP). The results showed that the objective test recognition rates for all the Arabic words in the different testing materials concerning the transmission media are lower than those for direct. The objective test recognition rates are the lowest when the used recognizer (neural network) is trained with the direct transmission medium data. It also showed that the subjective test recognition rates are higher.

In [22], authors discuss the development and implementation of an Arabic automatic speech recognition engine. The engine can recognize both continuous speech and isolated words. The proposed system was developed using the Hidden Markov Model Toolkit. First, an Arabic dictionary was built by composing the

words to its phones. Then, Mel Frequency Cepstral Coefficients (MFCC) of the speech samples is derived to extract the speech feature vectors. Then, the training of the engine based on triphones is developed to estimate the parameters for a Hidden Markov Model. To test the engine, the database consisting of speech utterance from thirteen Arabian native speakers is used which is divided into ten speaker-dependent and three speaker-independent samples. The results showed that the overall system performance was 90.62%, 98.01 % and 97.99% for sentence correction, word correction and word accuracy respectively.

Related to HMM model, another research written by [23] the authors present a Viseme-based Visual Speech Recognition (VSR) systems, using Hidden Markov Models (HMM) for phoneme recognition, generally use 3-state left-right HMM for each viseme to recognize. They propose a novel approach introducing a consonant-vowel detector and using two classifiers: an HMM based classifier for the recognition of the “consonant part” of the phoneme and a classifier for the “vowel part”. The benefits of such an approach are first, reducing the number of hidden states second, reducing the number of HMMs. The system tested the method on a limited set of words of the Modern Classic Arabic language and achieved a recognition rate of 81.7%. Moreover, the proposed model is speaker-independent and uses visemes as the basic units, thereby, making it applicable to any set of words of varying size or content.

Authors in [24], introduced hybrid Artificial Neural Network (ANN) and Hidden Markov Model (HMM) models for Arabic speech recognition by using optimal codebook with Self Organizing Maps (SOM). The main innovation in this work is to use an optimal neural network to determine the optimal class. The accuracy rate reaches 86%.

Another proposed system introduced by [25] the system able to identify an individual from sample of his or her speech. The system based on Arabic language speech. Also, this system is word-independent system. Speech features are extracted using MFCC. As a result, the system achieved 96.25% accuracy rate.

Related to individual recognition, authors in [26] present an audiovisual system for speech recognition. This system tackling the noise issue with audio-only recognition system by combining the lip information associated with audio to increase the accuracy rate. The system has been trained for 28 basic Arabic phonemes using the recorded samples of a five different speakers. Results show that accuracy rate reach 94%.

Related to Visual Speech Recognition (VSR), author in [27] presents a novel viseme-based Recognition system using hidden Markov Models (HMM) for phoneme recognition. This system based on 2 classifier. First, an HMM based classifier for consonant part of the phoneme. Second, classifier for vowel part. The method has been tested on a limited set of words of the Modern Classic Arabic language and achieved a recognition rate of 81.7%.

In [28], authors present the design, implementation and evaluation of proposed speech recognition system. This system developed mainly using the Carnegie Mellon University (CMU) Sphinx tools together with the Cambridge HTK tools. HMM and Gaussian mixture techniques has been used to improve efficiency. The system obtained a word recognition accuracy of 92.67%.

Related to Gaussian mixture techniques, authors in [29] introducing a new techniques for automatic speech recognition. This technique employs a full measure of statistical dependence among random variables that is known as copulas. A novel probabilistic classifier that combines finite Gaussian mixture modeling for marginal distribution function and Gaussian copula is developed. Compared with other benchmark Arabic speech recognition, the result demonstrates the improvement and shows an excellent performance.

Arabic Language has its own characteristics hence some speech features may be more suited for Arabic speech recognition than the others.

In [30], some feature extraction techniques are explored to find the features that will give the highest speech recognition rate. The authors showed that Mel-Frequency Cepstral Coefficients (MFCC) gave the best result. Also, they look at using an operator well known in image processing field to modify the way we calculate MFCC, this results in a new feature that we call LBPCC. They propose the way they use this operator. Then they conduct some experiments to test the proposed feature.

5. Conclusion and Future work

In this paper, scattered research work has been shown, over the three relevant topics in Arabic Language Technologies categorized into three main categories, Optical Character Recognition, Arabic translation technologies and Arabic speech Recognition. Researchers have a chance to expand this effort by adding more categories.

6. References

- [1] Pervez Ahmed and Yousef Al-Ohali, "Arabic Character Recognition: Progress and Challenges," J. King Saud Univ., Vol. 12, Comp. & Info. Sci., pp. 85-116 (A.H. 1420/2000)
- [2] Liana M. Lorigo and Venu Govindaraju, "Offline Arabic Handwriting Recognition : A Survey", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 28, NO. 5, MAY 2006
- [3] Haikal El Abed and Volker M'argner," Arabic Text Recognition Systems - State of the Art and Future Trends", IIT 2008. International Conference on Innovations in Information Technology, Al Ain, 2008, p 692 - 696
<http://dx.doi.org/10.1109/innovations.2008.4781781>
- [4] M. T. Parvez and S. A. Mahmoud,"Offline Arabic Handwritten Text Recognition: A Survey", ACM Computing Surveys, Vol. 45, No. 2, Article 23, February 2013.
<http://dx.doi.org/10.1145/2431211.2431222>
- [5] Iping Supriana and Albadr Nasution, "Arabic Character Recognition System Development", The 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013), p334 – 341.
<http://dx.doi.org/10.1016/j.protcy.2013.12.199>
- [6] Safwa Taha, Yusra Babiker and Mohamed Abbas,"Optical Character Recognition of Arabic Printed Text",2012 IEEE Student Conference on Research and Development, S5-2.
<http://dx.doi.org/10.1109/scored.2012.6518645>
- [7] Fouad Slimane,Slim Kanoun , Jean Hennebert, Adel M. Alimi , Rolf Ingold, "A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution",Pattern Recognition Letters 34 (2013) 209–218
<http://dx.doi.org/10.1016/j.patrec.2012.09.012>
- [8] Jafaar Al Abodi and Xue Li,"An effective approach to offline Arabic handwriting recognition",Computers and Electrical Engineering Vol. 40, p 1883–1901,(2014)
<http://dx.doi.org/10.1016/j.compeleceng.2014.04.014>
- [9] N. Aouadi, S. Amiri and A. Kacem Echi,"Segmentation of Connected Components in Arabic Handwritten Documents", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013, p738 – 746.
<http://dx.doi.org/10.1016/j.protcy.2013.12.417>
- [10] Arwa Alqudsi, Nazlia Omar, Khalid Shaker, "Arabic machine translation: a survey", Artif Intell Rev (2014) Vol.42 p 549–572.
<http://dx.doi.org/10.1007/s10462-012-9351-1>
- [11] Taghreed Hailat, Mohammed N. Al-kabi, Izzat M. Alsmadi and Emad Al-Shawakfa, "Evaluating English To Arabic Machine Translators",2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)
<http://dx.doi.org/10.1109/AEECT.2013.6716439>
- [12] Mohammad M. Abu Shquier, Mohammed M. AlNabhan, Tengku mohammed Sembok, "Adopting new rules in Rule-Based Machine Translation", 2010 12th International Conference on Computer Modelling and Simulation.
- [13] Ahmed Hossny, Khaled Shaalan, Aly Fahmy, "Machine Translation Model using Inductive Logic Programming", NLP-KE 2009 International Conference on Natural Language Processing and Knowledge Engineering, p1-8, 2009.
<http://dx.doi.org/10.1109/nlpke.2009.5313850>
- [14] Rasha Al Dam and Ahmed Guessoum,"Building a Neural Network-Based English-to-Arabic Transfer Module from an Unrestricted Domain", 2010 International Conference on Machine and Web Intelligence (ICMWI), Algiers, 2010
<http://dx.doi.org/10.1109/ICMWI.2010.5648157>
- [15] Farag Ahmed and Andrea Nurnberger, " Arabic/English Word Translation Disambiguation Approach based on Naive Bayesian Classifier", Internation Multiconferece on Computer sciences and information technology, pp. 331 - 338.

- [16] Waleed Oransa, Mohamed Kouta, Mohammed Sakre, "Injected Linguistic Tags to Improve Phrase Based SMT", The 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010.
<http://dx.doi.org/10.1109/iccae.2010.5451681>
- [17] Ahmed R. Nabhan, Ahmed Rafea, "A Hybrid Noun Phrase Translation System", The 7th International Conference on Informatics and Systems (INFOS), p1-p7, 2010.
- [18] Ridouane Tachicart and Karim Bouzoubaa, "A hybrid approach to translate Moroccan Arabic dialect", 9th International Conference on Intelligent Systems: Theories and Applications, 2014, p1-p5.
<http://dx.doi.org/10.1109/sita.2014.6847293>
- [19] Mouiad fadiel Alawneh, Tengku Mohd Sembok, Masnizah Mohd, "GRAMMAR-BASED AND EXAMPLE-BASED TECHNIQUES IN MACHINE TRANSLATION FROM ENGLISH TO ARABIC", 2013 5th International Conference on Information and Communication Technology for the Muslim World.
- [20] Yi Chang, Datong Chen, Ying Zhang, Jie Yang, "An image-based automatic Arabic translation system", Pattern Recognition, Vol. 42, p2127-p2134, (2009).
<http://dx.doi.org/10.1016/j.patcog.2008.10.031>
- [21] Onsy Abdel Alim Ali, Mohamed M. Moselhy, Aya Bzeih, "A Comparative Study of Arabic Speech Recognition", 2012 16th IEEE Mediterranean Electrotechnical Conference (MELECON).
- [22] Bassam A. Q. Al-Qatab, Raja N. Ainon, "Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK)", 2010 International Symposium in Information Technology (ITSim).
<http://dx.doi.org/10.1109/itsim.2010.5561391>
- [23] Pascal Damien, "Visual Speech Recognition of Modern Classic Arabic Language", 2011 International Symposium on Humanities, Science and Engineering Research
<http://dx.doi.org/10.1109/SHUSER.2011.6008499>
- [24] Mohamed ETT AOUIL, Mohamed LAZAAR, Zakariae EN-NAIMANIA, "hybrid ANN/HMM models for arabic speech recognition using optimal codebook", 2013 8th International Conference on Intelligent Systems: Theories and Applications (SITA).
- [25] Suliman S. Al-Dahri, Youssaf H. Al-Jassar, Yousef A. Alotaibi, Mansour M. Alsulaiman, Khondaker, Abdullah-Al-Mamun, "A Word-Dependent Automatic Arabic Speaker Identification System", IEEE International Symposium on Signal Processing and Information Technology, 2008. ISSPIT 2008.
<http://dx.doi.org/10.1109/isspit.2008.4775669>
- [26] Fatma zohra Chelali and Amar Djeradi, "Audiovisual speech/speaker recognition, Application to Arabic language", 2011 International Conference on Multimedia Computing and Systems (ICMCS).
- [27] Pascal Damien, "Visual Speech Recognition of Modern Classic Arabic Language", 2011 International Symposium on Humanities, Science and Engineering Research.
<http://dx.doi.org/10.1109/SHUSER.2011.6008499>
- [28] Mohammad A. M. Abushariah, Raja N. Ainon, Roziati Zainuddin, Moustafa Elshafei, Othman O. Khalifa, "Natural Speaker-Independent Arabic Speech Recognition System Based on Hidden Markov Models Using Sphinx Tools", International Conference on Computer and Communication Engineering (ICCCE 2010).
<http://dx.doi.org/10.1109/iccce.2010.5556829>
- [29] Nacereddine Hammami, Mouldi Bedda, Nadir Farahl, "Probabilistic Classification Based on Gaussian Copula for Speech Recognition: Application to Spoken Arabic Digits", Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2013
- [30] Mansour Alsulaiman, Ghulam Muhammad, Zulfiqar Ali, "Comparison of Voice Features for Arabic Speech Recognition", 2011 Sixth International Conference on Digital Information Management (ICDIM).
<http://dx.doi.org/10.1109/ICDIM.2011.6093369>