The Challenges of Arabic Language Use on the Internet

Noora Albaloshi, Nader Mohamed, and Jameela Al-Jarooodi
Faculty of Information Technology, UAEU
P.O. Box 17551, Al Ain, UAE
{noora.abbas, nader.m}@uaeu.ac.ae, jaljaroodi@gmail.com

Abstract. The Arabic language is used by a significant percentage of the world’s population. Furthermore, beyond personal and public communication within the Arab countries, the Arabic language is becoming very important on the Internet. Numerous documents, resources and applications rely on the language to reach the Arabic speaking audiences. Thus the Arabic language is striving to take its place among the languages supported on Internet such as English, French and Chinese. Efforts are being taken to increase Arabic content, Arabic-based search tools and other models and applications on the Internet. In this paper we offer an overview of the evolution of the Arabic language use on the Internet. There are several obstacles and challenges that the language faces and those need to be resolved efficiently to make the language more usable. Furthermore, we explore how the Arabic language evolution and use on the Internet affects the business world.

Keywords – Arabic Language; Arabic Web content; Arabic Internet support; language challenges.

I. INTRODUCTION

The Arabic Language is spoken by a significant percentage of the world’s population. It is being used as means of communication among Arabs and non-Arab Muslims as well. As a result Arabic is becoming an important language on the Internet due to the increasing number of Arabic speaking online users seeking Arabic content and applications. Recent figures from the Internet World Statistics show that there are 60 million Internet users from the Arab World [1].

The Internet, being an international mean of communication, should support the diverse languages of the world [11]. It has seen world-wide increase in the number of the users and the utilization of its tools and applications among people of various backgrounds and levels including professionals, home makers and students at universities [8]. The Internet is a heterogeneous environment composed of different configurations of hardware and software that should be customized for the purpose of supporting used languages [11]. The development of the Internet content and services to cater for the various users is important to enhance the Internet usage. Part of these enhancements involves the capability to support multiple languages and provide tools to offer and use multi-lingual content. The Arabic language is one of those languages that are growing significantly on the Internet and the research in the Arab world is increasing to further enhance its presence and usage on the Internet.

Several challenges face building, making available, finding, and using Arabic content on the Internet. At the current state, it is usually hard to offer Arabic material and also provide suitable search tools. In addition, many popular current applications cannot fully support the language. The authors in [5] describe the relation between the availability of content in Arabic and the challenges of increasing the awareness of Middle East and North Africa audiences is the relevance of Information Literacy. They endeavor to explain what this inter-relationship is, and why and how the increase of Arabic material on the Web could lead to an increase in awareness of Information Literacy.

Several major companies like Microsoft, Google and IBM are starting to invest in this area of research to enhance their offerings to the Arab world consumers and meet their expectations. There are many obstacles for Arabic language on the Internet including standards, character sets, Arabic content and programs and applications. The main purpose of this paper is to highlight the obstacles facing the better utilization of Arabic language on the Internet. Our aim is to provide an overview of this rapidly developing field of research, and which in particular will be used as a guideline for the implementers to support and implement Arabic content and applications on the Internet. An earlier study [7], attempted to highlight some of the issues. However, the main concern was the security and trust. Our work concentrates directly on the language challenges.

In this paper, we first start by discussing the challenges facing the Arabic language on the Internet in Section II. Then we follow it by a discussion of the current approaches in the field of Arabic language on the Internet in Section III. Next, in Section IV, we provide a discussion of the current issues and the possible approaches to solve them. Finally in Section V we conclude the paper and provide some pointers of possible future work in the field.
II. CHALLENGES

In this section we focus on the challenges and obstacles that face Arabic language on the Internet. We cover three areas including the technologies and applications, the content, and the education systems.

A. Technologies and Applications

1) Morphological Analyzers: The lack of work on morphological analyzers that could identify the different features and structure of the language or the incomplete work in this field makes it difficult to transfer the Arabic language on the Internet. Arabic language and scripts are unique in the sense that it has its representation style that eases the reading and understanding of the meanings of the sentences. However, this representation style is the main cause of difficulties for the morphological analyzers. The analyzers on the Internet would mean getting the right meanings to the readers. Implementing and embedding the suitable analyzers within web browsers also requires a lot of resources and preparations. The question who would handle this task and who will control and manage the standards and resources required for that?

2) Language Identification Systems: Language identification requires the recognition and classification of the different character sets and codes. Since Arabic uses a unique character set, when searches are carried out in different search engines to find papers in Arabic it results in few or sometimes no results for the intended search. This is considered a limitation in the Arabic language on the Internet. Even when you go to libraries and type the search word in Arabic the results might be different than intended or it might not find any matches. Having methods and systems for correctly identifying the Unicode and character set of Arabic is important in retrieving the right results and information, which by default will mean easiness in transferring the Arabic language on the Internet.

3) Search Engines: Another one of the challenges is the awareness of users. The users of The Internet especially people using the Arabic language should be made aware of the available search engines for their research. It is important to know that there are search engines other than Google and Yahoo that can give better search results for the Arabic Language. The question is, how and who will educate the users on Arabic search engines and what is the role of advertising in this matter.

4) Translation Systems: Translation is communicating the meaning of some data and information from the original language into another intended language. As an example having some text in English and translating it into Arabic for the understanding of certain receivers. Translation systems and applications have been widely spread on the Internet to provide online users with information in their native language. Due to the unique structure and character set of the Arabic language, it is very hard to implement accurate translators. Therefore, the challenge is whether the translation systems and applications provide the correct meaning and information as intended by the original language.

B. Arabic Content and Material

The most common language of use on the Internet is English. Various other languages started to become more visible recently sue the efforts made to increase the content in those languages. The Arabic content is also starting to be made more available based on original Arabic documentation and information being added. However, one of the main sources for Arabic content is through translation. However, as we discussed earlier, automatic translations are extremely inaccurate and in many cases cannot convey the original intended meanings in the content. We face the same thing in everyday life, when we try to translate English sentences into Arabic using the translation software available, we notice that it requires a lot of adjustments and modification to get the exact meaning. As a result, for uses who do not speak the original language, there is no way of knowing if the translated content matches the original one. In addition, due to the complex structure of the Arabic language, it is hard to build the documents in a correct and readable form. The challenge here is how to increase the content available in Arabic language on the Internet and who will be doing that.

C. Education System

This is another important factor in transferring the Arabic language to the Internet. The education system in the Arab world mainly focuses on reading, writing and speaking in English which generated educated people using the English language fluently. They might be native Arabic speakers, but writing in Arabic is a difficult task. The education system in the Arab world might not be supporting the advancements in technology that support the Arabic language. Since the education system (the higher education in particular) in the Arab world supports the English language where most of its graduates are more comfortable surfing the Internet using their second language instead of Arabic language. As a result, we are becoming more and more of Arabic language speakers who are capable of authoring and generating content for the Internet. Furthermore, it may be necessary to encourage Arabic scientists to use the Arabic language along with the English language to write their research which will enrich the Arabic language content on the Internet.
Going over the challenges we discussed above it is clear now that the Arabic language on The Internet needs a lot of research and focus on the implementation efforts to support and further develop this area.

III. CURRENT APPROACHES

Much research and numerous studies have been conducted in the field of Arabic Language on the Internet. Many started with studies focusing on character set analyzers, identification systems, search engines and Arabic content online. Many of these efforts had contributed significantly to the field. Here we try to cover some of the technologies and applications that support Arabic language on the Internet.

In [2] the authors developed a morphological analyzer that consists of the analyzer, running on a network server, and Java applets that run on the user’s machine and render words in standard Arabic orthography both for input and output. The developed system accepts on-line orthographical words of Arabic and returns morphological analyses. The system is considered a valid contribution to the Arabic language support but the implementation would require resources and online standards to be embedded in all web browsers.

To continue with the efforts in promoting the Arabic language, researchers focused also on handwriting recognition systems. In [3] the authors proposed a handwriting recognition system that handles the diacritic strokes of the Arabic scripts which differ among writers separately to add a layer for recognizing scripts. The suggested system proposes a template matching scheme to the recognition of Arabic script with an algorithm for dynamically treating the diacritical marks. The limitation of the proposed system is the practicality of usage and resources required for the actual implementation and support on the Internet.

Following handwriting recognition, language identification and script recognition was also an important field. The authors approach in [6] focused on a hybrid approach in Arabic script identification. They have combined decision trees and the ARTMAP approach to improve the performance of the Arabic script language identification on web documents in a variety of languages. Their result showed that the proposed approach has outperformed both the decision tree and the default ARTMAP approaches when used separately. The advantage of the proposed system is to effectively retrieve online information written in Arabic scripts in which words are similar to other languages.

On the other hand, authors in [9] implemented a prototype of a web page translation system that assists in automatically translating English into Arabic for users. The proposed system supports online Arabic speakers in retrieving information in the Arabic language. The system also assists in avoiding the problem of having the knowledge of English as a prerequisite for using the Web efficiently and benefiting from the multitude of services the Internet offers.

Search engines also play a vital role in retrieving information using the Arabic language. Information retrieval, as a language-dependent operation, is greatly affected by the language of documents and how a search engine handles the character set of the language. Lot of researchers focused on designing and implementing portals that support searching for non-English Web content. Authors in [10] suggested an online portal, which is a language-independent approach that uses meta-searching, summarization, categorization, visualization, and statistical language processing techniques to identify high-quality domain-specific collections and to support searching and browsing of non-English information. The authors developed SBizPort and AMedPort portals for the Spanish business and Arabic medical domains respectively.

The work in [12] also focused on developing a distributed multilingual search engine with focus on Arabic language. In their paper, the authors presented the architecture of “Barq” system and how the system handles Arabic language processing. The system indexes all Web documents and utilizes Arabic word root extraction, information retrieval methods, automatic categorization, focused crawling, distributed database systems, and performance tuning to process Arabic language and ease the searching of online Arabic information.

As stated in [13], searching for Arabic content and information is a neglected area within search engines. The authors experimented using a set of eight Arabic search terms and run in three general search engines (AlltheWeb1, AltaVista2, and Google3) and three Arabic engines (Al bahar4, Ayna5, and Morfix). The experiment results showed the importance of making users aware of the limitations of general search engines in retrieving Arabic documents, and of the high number of documents that will be lost when only the exact forms of Arabic words are entered as search terms on the Web. The limitation of this work is that it was focused on a limited number of search engines neglecting the applications and other tools offered to retrieve information online.

The problems of the Arabic language search engines in information retrieval were discussed in [20]. The authors used a Lebanese official newspaper for their experiment and used three “keyword matching” in three search engines: Google, Yahoo, and Idrisi. They found that this method is not good because of the uniqueness of the syntax of the Arabic language and suggested to use the character n-grams for better search strategies.

The authors in [21] presented and discussed major Information Retrieval (IR) tools and techniques of Arabic content and it highlighted a few challenges in this regard. They discussed the information retrieval tools such as literal search, stemming, n-gram, query expansion and many more. They highlighted one of the main challenges as not getting the documents related to the intended search when a user enters a search word or phrase, other related words/phrases are not retrieved, which gives the user an impression that only a few documents exist on the Web. The challenge is that Arabic is root-based and highly derivational with rich vocabulary and it is not easy to get to the intended search phrase.

Arabic content and material is another area of focus in promoting Arabic language on the Internet. The results of the
work in [4] suggested that as the Internet users from the Arab world is increasing; the online Arabic content related to health should increase as well. It also mentions that the Internet can serve as an important vehicle to distribute important health information. Through their surveys and observing user profiles, access patterns and requested pages, they reached a generalization that it is important to make health information available in Arabic. On the other hand, [5] suggested through their personal experience that the availability of the Arabic online content highly depends on the education systems and transforming human knowledge to online form.

Furthermore, the work in [7] conducted a survey on a sample of 70 companies in Bahrain to gather information using questionnaires on services used online. The usage of the Internet included e-mails, financial news, market research, and information gathering. The study highlighted that the shortage of Arabic sites was not considered a major problem in accessing the Internet by Bahraini’s. The major problem was security and trust of the available online content.

On the important technical aspects of using Arabic over the Internet is the classification of Arabic text and documents based on linguistic features. This classification is an important process in a number of applications over the Internet. Examples of these applications are e-mail spam detection, web content filtering, web mining, and automatic message routing. The classification process involves assigning documents to one or more predefined categories based on its content. There is some work in developing and evaluating algorithms for automatic Arabic text classification. Sawaf uses statistical classification methods such as maximum entropy to classify and cluster news articles [14]. El-Halees developed a method based on association rules to classify Arabic documents [15]. Al-Fedaghi and Al-Anzi’s developed an algorithm that tries to find the root of the word by matching the word with all possible patterns including all of its possible affixes [16]. El-Kourdi used Naïve Bayes algorithm to classify Arabic documents [17]. Al-harbi evaluated document classifications achieved on seven different Arabic corpora using statistical methodology [18]. Khreisat studied Arabic text classification using N-Gram frequency statistics [19]. These developed algorithms and techniques provide different accuracy levels of Arabic documents classification.

The existing work on Arabic language on the Internet adds value in showing the tools and technologies required for promoting Arabic language on the Internet. The work limits the focus on what are the resources required to have successful implementations of these technologies.

According to [1], in 2009 there were 60 million Arabic-speaking people online. This represents 17.5 percent of the penetration rate of the Internet over 344 million worldwide speakers. Between 2000 and 2009, Arabic speaker’s percentage on the Internet had grown by 2,297.7 percent compared to the world average of 399.03 percent. However, this rapid increase is not matched by any rate with the growth in the Arabic content and the supporting Arabic-based applications and tools.

As discussed earlier in section II, there are a number of challenges and obstacles that face the Arabic language on the Internet. It starts from the morphological analyzers and language identification systems to search engines and Arabic content. It all depends on the technologies that can support the character set and preserve the Arabic language uniqueness. The important aspects are implementing and developing the right tools, environments and applications to accept and process the Arabic character set and correctly recognize the different letters, diacritic strokes, scripts and codes. It is important for the Arabic user to understand correctly the translated information in their first language.

Moreover, it is important to make information and Arabic content available to users. The unavailability of Arabic content limits the users from surfing the Internet for the knowledge in Arabic. This is mainly due to several factors including the reluctance of contributors to provide the content in Arabic. This is the result of the education system in the Arab world which concentrates on the English language especially in the science, engineering and technology fields. In addition, the absence of online communities and bodies plays a vital role in the absence of Arabic content online. Limited Arabic content online would mean limiting the accessibility to the actual knowledge available online for Arabic users.

Some researchers tried to focus on the technological issues of the World Wide Web in supporting the Arabic language as mentioned in section III. The technologies will not be of effect if the implementations were limited. The complexity and compatibility of the proposed technologies and applications are important factors in the implementation phase. Nevertheless, the ease of use and awareness of Arabic users plays an important role in the continuance of online usage.

Efforts of big companies and organizations to support Arabic language on the Internet are absent. Organizations such as Google, Yahoo and Hotmail can become one of the main sources of spreading Arabic language and content online. The resources and mechanisms that they can provide in supporting the implementations and research on Arabic language on the Internet are enormous. The challenge would be securing the commitment of these organizations and companies to meet the need of the online Arabic speaking users. However, as in any business, the main target is profit, it will not be a priority unless these ventures will offer these companies an excellent opportunity for good profit.

Yet, some organizations like Google started to implement and invest in its own translation system. Google’s translation
system defines vocabularies and grammars in different languages that it supports to assist users in understanding online information easily. One of the supported languages is Arabic for Arabic speaking users. Their efforts continue to improve their translation system by involving Google users in ranking and assessing the translations that users get to validate and update their system to meet customer requirements.

The Arabic language on the Internet has other challenges. People and organizations lack the knowledge of the top risks of moving Arabic content online such as trust of content, lack of control over Arabic content, raising Arabic consumer expectations, and non-compliance with standards and many more. These challenges are open questions in front of developers and researchers to find the suitable solutions that cater for the different requirements.

V. CONCLUSION AND FUTURE WORK

To conclude, The Arabic language on the Internet is a growing field as lot of major investors in online technologies focus on it. The growth of the number of Arabic speakers in online communities and the focus of government initiatives on the spread of the Arabic language, have a major impact on the availability of the Arabic language on the Internet. The focus on these initiatives and others to support and overcome the challenges of Arabic language on the Internet will ensure meeting customer demands and expectations. More research and implementation efforts should be directed on this field to meet the growing need of online Arabic users.

The existing work on Arabic language on the Internet has limitations. The research shows that there are technologies and applications for processing and using Arabic language online. However, the challenges limit the impact of the Arabic language and impede the spreading of the language and supporting tools among Arabic users. The Arabic language on the Internet is an important part of today’s communication mechanism among people. The future requires addressing various issues and obstacles to provide workable solutions that support not only the Arabic language but any other common language. Some of these issues involve: how to ensure correct interactions and translations among different languages? What standards can or should be introduced to simplify and facilitate the use of Arabic languages on the Internet? Are there possibilities to find more efficient models for morphological analysis and character set handling? Is it possible to build language-independent tools that can operate and provide usable results from the Internet in any language and in particular the Arabic language? Finally, working on finding business and profit opportunities that will encourage companies to invest in supporting the Arabic language on the Internet. Another direction of action is to work with the Arab and Islamic governments to provide support for such activities and provide research and entrepreneurship funds for projects that will increase the use and visibility of Arabic language on the Internet.

REFERENCES