Arabic Natural Language Processing Laboratory serving Islamic Sciences

Moath M. Najeeb
Computer Science Department at Al-Gunfudah Umm Al-Qura University
Al-Qunfudah, Saudi Arabia

Abdelkarim A. Abdelkader
Computer Science Department at Al-Gunfudah Umm Al-Qura University
Al-Qunfudah, Saudi Arabia

Musab B. Al-Zghoul
Information System Department at Al-Gunfudah Umm Al-Qura University Al-Qunfudah,
Saudi Arabia

***Abstract*— Arabic Natural Language Processing (ANLP) has a great attention as a new research topic in the last few years. In this paper an ANLP laboratory has been created to serve the Islamic Sciences, especially the science of Hadith. The main tasks of this laboratory are creating and using the necessary linguistic resources (Corpora, Lexicon, etc) in developing or adapting the basic tools (Parser, POS-tagger, etc), developing Arabic Natural Language Processing system's evaluation framework and defining research areas and services for universities. The laboratory can also adapt the important theories, resources, tools and applications of other natural language processing such as English and French.**

Keywords—Natural Language Processing; Arabic Language; Islamic Sciences; Framework; Laboratory

I. INTRODUCTION

Natural Language Processing (NLP) is an artificial intelligence branch which has the ultimate goal to invent theories, discover techniques and build software that can understand, analyze and generate the nature human languages in order to interface with computers in both written and spoken contexts using natural human languages, so NLP gives computers the ability to understand the way humans learn and use language is the most challenge inherent in natural language processing [1]. The NLP techniques parse linguistic input (word, sentence, text, dialogue) according to the rules (derivational rules, inflectional rules, grammatical rules, etc.) and resources (like lexicon, corpus, dictionary) of the target language. At the present time, this is at the advanced stages of development especially for the English language. We expect that the current century will focus on NLP.

After several decades of heavy research activity on English NLP and other languages, Arabic Natural Language Processing (ANLP) have become a popular area of research, and some ANLP laboratory have been created [2]. There are some efforts to create ANLP tools [3], but these efforts always face two main challenges: the agglutination in Arabic language and dispensability of vowel diacritics [4].

This enthuses us to work on our language, so we establish a laboratory for Arabic Natural Language Processing to serve Islamic Sciences in Computer College at Al-Gunfudah at Umm Al-Qura University in 2014, this laboratory aims to contribute and involve in research on Islamic Ssciences and Arabic Natural Language Processing and corpus linguistics. Our work focuses on tools and corpus resources for analysis and modeling of Arabic language, especially Classical Arabic (which is the language of the Quran, and it is used primarily for reading and reciting Islamic holy text).

The remainder of this paper is organized as follows: The second section gives the main features of the Arabic language. The third section describes the Arabic Natural Language Processing Framework. The fourth section defines the proposed plan for NLP laboratory, and finally the conclusion and the future work explained in section five.

II. MAIN CHARACTERISTICS OF THE ARABIC LANGUAGE

As its name implies, the Arabic language is the language spoken at the origin by the Arabs. It is a Semitic language (like Hebrew, Armenian and Acadian). Strategically, it is the native language of more than 330 million speakers [2] living in an important region with huge oil reserves crucial to the world economy and home as well to the sacred sites of the world three monotheistic religions. It is also the language in which 1.4 billion [5] Muslims perform their prayers five times daily.

There are two main types of Arabic language: Classical Arabic and Modern Standard Arabic:

A. Classical Arabic

The language of the Qur'an and classical literature. It differs from Modern Standard Arabic mainly in style and vocabulary, some of which is archaic. All Muslims are expected to recite the Qur'an in the original language, however many rely on translations in order to understand the text.

B. Modern Standard Arabic (اللغة العربية الفصحى / al-luġatu l-ʿarabiyyatu l-fuṣḥā)

Universal language of the Arabic-speaking world which is understood by all Arabic speakers. It is the language of the vast majority of written material and of formal TV shows, lectures, etc.

The Arabic alphabet counts twenty-eight letters (or 29 if we add the "hamza" that can be considered a letter). There is no difference between the handwritten letters and the printed letters; the notions of capital letters and lower case letters don't exist. On the other hand, the letters have, most of the time, a different shape depending on their position in words: isolated, in the beginning, in the middle or in the end [6].

Arabic belongs to the Semitic family of languages. One of the characteristic features of Semitic languages is their system

of roots and patterns. Most (but not all) Arabic words have trilateral roots—in other words, there are three consonant letters in these words that connect them to a "root" meaning and to other words that share the same root. In Arabic, we can manipulate roots by varying the internal (short) vowel-ling between the root letters, by adding suffixes and prefixes, or placing other consonants and long vowels between the root letters [7].

The root word s-l-m is a common example. From the basic verb "salima", (to be safe), we can derive other verbs such as sallama, "to hand over or deliver;" aslama, "to submit;" and istaslama, "to surrender." The nouns salaam, "peace;" salaama, "health or safety;" and muslim, "a Muslim," derive from the same roots.

Arabic tends to prefer the word order VSO (verb before subject) rather than SVO (subject before verb) [8]. However, the word order is fairly flexible, since words are tagged by case endings. Subject pronouns are normally omitted except for emphasis or when using a participle as a verb (participles are not marked for person). Auxiliary verbs precede main verbs, and prepositions precede their objects.

Adjectives follow the noun they are modifying, and agree with the noun in case, gender, number, and state: For example, "bintun jamīlatun" "a beautiful girl" but "al-bintu l-jamīlatu" "the beautiful girl", "al-bintu jamīlatun" "the girl is beautiful". Elative adjectives, however, precede their modifying noun, do not agree with it, and require that the noun be in the genitive case.

Arabic has three grammatical cases roughly corresponding to: nominative, genitive and accusative, and three numbers: singular, dual and plural. Normally, singular nouns take the ending -u(n) in the nominative, -i(n) in the genitive and -a(n) in the accusative. Some exceptional nouns, known as dip-totes, never take the final n, and have the suffix -a in the genitive except when the dip-totic noun is in the definite state (preceded by al- or is in the construct state).

However, case is not shown in standard orthography, with the exception of indefinite accusative nouns ending in any letter but ta marbuta or hamza, where the -a(n) "sits" upon an alif added to the end of the word (the alif still shows up in unvowelled texts). When speaking or reading aloud, articulating the case ending is optional. Technically, every noun has such an ending, although at the end of a sentence, no inflection is pronounced, even in formal speech, because of the rules of 'pause'[9].

III. Arabic Natural Language Processing Framework

There are some projects in Arabic Nature Language Processing (ANLP) that can lead to extract the main specification and requirements to build a general framework for ANLP, one of the most important issues is to draw a "road map" of this framework; which means to determine the goals of the framework, the tasks that must be achieved, and to determine the resources and the tools that could be used to serve the ANLP projects.

The second issue that must be considered is the expertise and the group works involved in the ANLP framework, these groups must be consistent with the main goals of the ANLP framework. The groups must contain experts in different areas such as: Arabic Linguistics, Software Engineering, Programming and Algorithms, Scientific Research and Artificial Intelligence, also the groups must contain administrative staff from the national agencies concerned with the Arabic language research.

The third issue that must be considered into ANLP framework is the "Projects developments"; which determine and explain the resources, tools and applications that must be created or improved, these projects include: Arabic Summarization, Automatic Arabic Answering System, Automatic Arabic Translation, Arabic Transliteration, Arabic Spell Checker, etc.

Arabic language has many specific features that distinct it from other languages, so any ANLP framework must consider these features or it will be unfortunate framework. The framework must take into account the Arab world people's aspirations such as: the knowledge translation from other languages, reforming Arabic grammars or create new ones to facilitate the building of the ANLP tools, create variant applications for summarization, Information Retrieval, etc.

There are many obstacles in creating ANLP framework; one of them is the lake of the ANLP resources and tools due to the difficulties in transforming and adapting resources and tools from other languages.

Another obstacle is the regional dialects; these dialects are used by Arab people on daily basis beside the MSA and Classical Arabic, so it is a challenge to build one framework for all these types.

There are some obstacles related to Arabic scripts; as example, Arabic language does not have a specific characters to show the short vowels, sometimes, short vowels can be presented by diacritics in Classical Arabic, but MSA and dialects rarely use the diacritics, also Arabic letter has many shapes depending on its position in the word, so the tokenization process represents a challenge in Arabic language because one word may contain many tokens.

Another obstacle facing ANLP tools is the syntax form of the Arabic sentence; in Classical Arabic and MSA the VSO (Verb, Subject, Object) form is used, whereas in dialects the SVO (Subject, Verb, Object) and OVS (Object, Verb, Subject) forms can also be used beside VSO form, this will change the sentences structure and sometimes the meaning of the sentence. These issues must be considered in the ANLP framework creation.

IV. Proposed Laboratory For Islamic Sciences

As we mentioned earlier, our work focuses on tools and corpus resources for analysis and modeling of Arabic language, especially classical Arabic (which is the language of the Quran, and is used primarily for reading and reciting Islamic holy text).

Services to be provided by the members of the laboratory can be organized around the following axes:

- *Development of content processing tools* such as bibliographic study (related works on ANLP, and Islamic sciences)
- *linguistic studies*: investigate the characteristics of Arabic language especially classical Arabic, the classical Arabic morphology, the grammar, some Islamic studies like "mustalah Alhadith"
- *Research and development*: which concentrate on the implementation and development of the theories and algorithms those were required in the laboratory , the defining of research areas and issues and presenting them to research centers, the identification, definition and development of ANLP tools and resources such as: corpus, lexical semantic network, Treebank, parallel corpus, parser, tagger, syntactic analyzer, the application of this tools on Islamic sciences especially the science of Hadith.
- *Integration and evaluation*: Integrating existing and new components for evaluation and providing services and Evaluating systems and tools developed in the laboratory or outside it.
- *Standardization*: Identification and development of standards for the laboratory's various products and activities such as test dataset, interfaces, evaluation procedures and the application of e-learning and m-learning standards like SCORM and LOM.

The laboratory will be formed in a three-layer model, the first layer represent ANLP (Linguistic) resources, ANLP (Linguistic) tools and algorithms and theories in the second layer and ANLP applications in the third layer. The scheme of this model is shown in figure 1.

Until now, many linguistic studies and discussions with linguistics experts have been achieved, the objective of these studies is to understand the Arabic language morphology and grammar and linguistic phenomena like the coordination, the anaphor, the ellipse, etc.

Since Arabic morphological analysis techniques have become a popular area of research, several systems are known in the Morphological Analysis domain [7], for example, the Khoja stemmer [10], the Buckwalter Morphological Analyzer [11], and AlKhalil Morpho System [12]. AlKhalil (AlKhalil Morpho Sys) could be considered as the best Arabic morphological system, it won the first position, among 13 Arabic morphological systems around the world, at a competition held by the Arab League Educational, Cultural and Scientific Organization (ALECSO) (المنظمة العربية للتربية و الثقافة و العلوم) and King Abdul Aziz City for Science and Technology (KACST). So, we had put a special effort on understanding and testing it and used its open source database as part of our linguistics resources. For a given word, AlKhalil identifies all possible solutions with their morphosyntactic features: vowelizations proclitics and enclitics, nature of the word Voweled patterns, stems, Roots Syntactic [12].

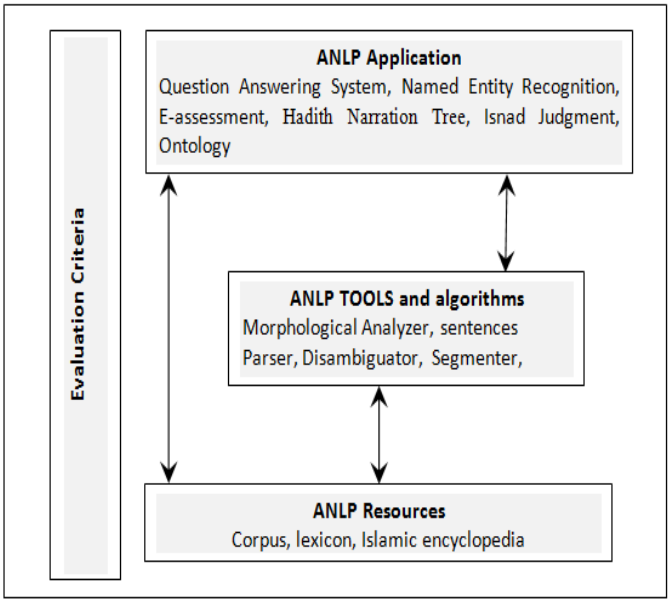

Fig. 1. ANLP Model

As we mentioned in the introduction, we will develop a new tools and algorithms, and also using the existing ones such as AlKhalil to serve the Islamic Sciences, especially the science of Hadith (the second fundamental source of Islamic legislation after the holy Quran). Each Hadith consists of two parts: the narration, known as matn and the chain of narrators through whom the narration has transmitted, traditionally known as Isnad.

Since the early centuries, Muslims are interested of Isnad science because it helps differentiate between the sahih (accepted) and da'ief (rejected) Hadith. The scholar of hadith judges it based on the narration chain and the individuals involved in the chain.

Through this laboratory, we want to develop this science using the new technologies. We will build a domain-dependent ontology that helps to automatically generate a suggested judgment of Hadith Isnad and share the common understanding in this science. For this purpose we will use the e-Narrator application. This application parses a plain Hadith text and automatically generates the full narration tree [13].

V. Conclusion and Future Work

Arabic Natural Language Processing framework was discussed and a laboratory for Islamic sciences has been initiated in Umm Al-Qura university, the laboratory aims to create new theories, resources and tools for Arabic Natural Language Processing and adapt the existing one to serve the Islamic sciences especially the science of Hadith.

In the future work, ontology for all concepts of Hadith will be created. So, we will begin by studying Hadith sciences to extract the components of this ontology (objects, properties, relations and rules). Next, we will develop an approach based on natural language processing techniques to the automatic generation of ontology instances from a collection of hadith documents.

ACKNOWLEDGMENT

The authors would like to greatly acknowledge the help and support of Dr. Yassine A. Alzubedi the dean of the Computer College at Al-Gunfoudah at Umm Al-Qura University, Makka, Saudi Arabia..

REFERENCES

[1] A. Abdelkader, D. Souilem Boumiza and R. Braham, "A categorization algorithm for the Arabic language," International Conference on Communication, Computer and Power (ICCCP'09), Muscat, February 2009.

[2] NLP4Arabic. https://sites.google.com/site/nlp4arabic/ 2014.

[3] A. Farghaly, K. Shaalan, "Arabic natural language processing: challenges and solutions,". ACM Trans. Asian Lang. Inform. Process. 8, 4, Article 14, 22 pages, December 2009.

[4] Khaled Shaalan, "Rule-based approach in Arabic natural language processing," International Journal on Information and Communication Technologies, Vol. 3, No. 3, June 2010.

[5] CIA. 2008. CIA Word Fact Book. Central Intelligence Agency, Washington, D.C.

[6] Y.A. Alotaibi, A.H. Meftah, "Comparative evaluation of two Arabic speech corpora," Natural Language Processing and Knowledge Engineering International Conference, pp. 1-5, 2010.

[7] Al-Sughaiyer, I. Al-Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," Journal of The American Society for Information Science and Technology (JASIST), John Wiley & Sons, Inc., NJ, USA,55, 3, 189-213, 2004.

[8] A. Abdelkader, D. Souilem Boumiza and R. Braham, " An oline Arabic learning environment based on IMS-QTI," 10-th IEEE International Conference on Advanced Learning Technologies, Sousse Tunisia, July 2010.

[9] M. Altabbaa, A. Al-Zaraee and M. Shukairy "An Arabic Morphological Analyzer and Part-Of-Speech Tagger," A Thesis Presented to the Faculty of Informatics Engineering, Arab International University, Damascus, Syria.

[10] Al-Sughaiyer, Imad A. and Ibrahim A. Al-Kharashi. "Arabic Morphological Analysis Techniques: A Comprehensive Survey". Journal of the American Society for Information Science and Technology 55(3):189–213. 2004.

[11] LDC, Linguistic Data Consortium. Buckwalter Morphological Analyzer Version 1.0, LDC2002L49, 2002. http://www.ldc.upenn.edu/Catalog/.

[12] Alkhalil Morpho Sys: A Morphosyntactic analysis system for Arabic texts (2010, April 16). Retrieved February, 2014, from ALECSO: http://www.alecso.org.tn/index.php?option=com_content&task=view&id=13 02&Itemid=998&lang=ar.

[13] Aqil M. Azmi and Nawaf bin Badia, "e-NARRATOR – AN APPLICATION FOR CREATING AN ONTOLOGY OF HADITHS NARRATION TREE SEMANTICALLY AND GRAPHICALLY Aqil M. Azmi and Nawaf bin Badia, Department of Computer Science, King Saud University, Riyadh, Saudi Arabia, 2011.