The Arabic Natural Language Processing: Introduction and Challenges

Boukhatem Nadera
El Tarf University
Algeria

Arabic is a Semitic language spoken by more than 330 million people as a native language, in an area extending from the Arabian/Persian Gulf in the East to the Atlantic Ocean in the West. Moreover, it is the language in which 1.4 billion Muslims around the world perform their daily prayers. Over the last few years, Arabic natural language processing (ANLP) has gained increasing importance, and several state-of-the-art systems have been developed for a wide range of applications.

Keywords: Arabic, natural language processing, computer
1. 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. 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 immense research 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. There are some efforts to create ANLP tools, but these efforts always face two main challenges: the agglutination in Arabic language and dispensability of vowel diacritics. Here it becomes essential to present some background to CALL in order to make the ideas simpler.

2. CALL

Computer-assisted language learning (CALL) addresses the use of computers for language teaching and learning. CALL emerged in the early days of computers. Since the early 1960s, CALL software was designed and implemented. The effectiveness of CALL systems has been made obvious by many researchers (Lam & Pennington, 1995; Mc Enery, Baker, & Wilson, 1995). Recently, though, computers have become so widespread in schools and homes and their uses have expanded so dramatically that the majority of language teachers must now begin to think about the implications of computers for language learning. As rightly pointed out by Warschauer (1996), using computers provides a number of advantages for language learning.

3. Modern Standard Arabic (MSA)

A comprehensive description is given in Maamouri and Bies (2004) of ‘Modern Standard Arabic’ (MSA) as the ‘language’ mostly targeted by Arabic NLP research and, therefore, by the Penn Arabic Treebank annotation which has so far only focused on Arabic newswire text. The term MSA is commonly used among linguists and computational linguists, although there is often little agreement on its definition. MSA, nobody’s native or first language, though there exists a ‘living’ writing and reading MSA community, is mainly the language of written discourse and is used in formal communication both written and oral with a well-defined range of stylistic registers. A more convenient term than MSA would have been ‘Modern Written Arabic’ if it were not for the ambiguity of mixing together written MSA with written dialectal occurrences, though this mix is more and more evident mainly in MSA broadcast newswire text (mostly in Egypt, Lebanon and a few other Middle-Eastern countries). Another term which is also appearing on the Arabic NLP scene is ‘Modern Conversational Arabic.’

4. Aspects of the Arabic Language

Arabic is rooted in the Classical or Qur’anic Arabic, but over the centuries, the language has developed to what is now accepted as MSA. MSA is a simplified form of Classical Arabic, and follows its grammar. The main differences between Classical Arabic and MSA are that MSA has a larger (more modern) vocabulary, and does not use some of the more complicated forms of grammar found in Classical Arabic. For example, short vowels are
omitted in MSA such that letters of the
Arabic text are written without diacritic
signs.

The Arabic language is written from
right to left. It has 28 letters, some of which
have one form (like "د"), while others have
two forms ("س" "س"), three forms ("١" "١" "١") or four forms ("١٤" "١٤" "١٤" "١٤"). Arabic words are generally classified
into three main categories [19]: noun, verb,
and particle. Arabic is a language of rich
and complex morphology, both derivational
and inflectional. Word derivation in Arabic
involves three concepts: root, pattern, and
form. Word forms (e.g. verbs, verbal nouns,
agent nouns, etc.) are obtained from roots
by applying derivational rules to obtain
responding patterns. Generally, each
pattern carries a meaning which, when
combined with the meaning inherent in the
root, gives the target meaning of the lexical
form.

For example, the meaning of the
word form "كاتب" (writer) is the
combination of the meaning inherent in the
root "كاتب" (write) and the meaning carried
by the pattern (or 'template') "فعل" (fa'il)
which is the pattern of the doer of the root.
Arabic also has some more morphological
peculiarities. For example, an indefinite
word can be made definite by attaching the
prefix definite article "ال" (the) to it, but
there is no indefinite article. As another
example, a verb can take affix pronouns
such as "ساعطيكما" (will-I-give-you); this
also shows that the verb is conjugated with
the dual suffix pronoun "كما" (you). An
Arabic inflected verb can form a complete
sentence, e.g. the verb "سمعتك" (heard-I-
you) contains a complete syntactic structure
in just a one-word sentence. Moreover, the
rich morphology of Arabic allows the
dropping of the subject pronoun ('pro-
drop'), i.e. to have a null subject when the
inflected verb includes subject affixes
(Shaalan, 2010).

5. The Challenges of the Arabic Language

Most ANLP systems developed in the
Western world focus on tools to enable
non-Arabic speakers make sense of Arabic
texts. Arabic Tools such as Arabic named
entity recognition; machine translation and
sentiment analysis are very useful to
intelligence and security agencies. Because
the need for such tools was urgent, they
were developed using machine learning
approaches. Machine learning does not
usually require deep linguistic knowledge.
Due to such issues, developers of such tools
had to deal with difficult issues. Of various
problems encountered while so, is when
Arabic texts include many translated and
transliterated named entities whose spelling
in general tends to be inconsistent in Arabic
texts [Shaalan and Raza 2008]. For example
a named entity such as the city of
Washington could be spelled:

واشنطن/واشنطن

Therefore, Arabic NLP applications
must deal with several complex problems
pertinent to the nature and structure of the
Arabic language. For example, Arabic is
written from right to left like Chinese,
Japanese, and Korean. Also, there is no
capitalization in Arabic. In addition, Arabic
letters change shape according to their
position in the word. Modern Standard
Arabic does not have orthographic
representation of short letters which
requires a high degree of homograph
resolution and word sense disambiguation.
Like Italian, Spanish, Chinese, and
Japanese, Arabic is a pro-drop language,
that is, it allows subject pronouns to drop
[Farghaly 1982] subject to recoverability of
deletion [Chomsky 1965].

Although Arabic is a phonetic
language in the sense that there is one-to-
one mapping between the letters in the
language and the sounds they are associated

with, Arabic is far from being an easy language to read due to the lack of dedicated letters to represent short vowels, changes in the form of the letter depending on its place in the word, and the absence of capitalization and minimal punctuation. As an example of the regularity of the association of letters to sounds, the letter (‘ب’) is always pronounced as “baa,” unlike letters in English that have more than one pronunciation. For example, the letter “s” in English may be pronounced as /z/ as in “cause” or /s/ as in “sail” or /sh/ as in “sure.” Further, while English has silent letters such as the “p” in “pneumatic” the “b” in “doubt,” the “k” in “know” and the “gh” in “weight,” Arabic has no silent letters.

Moreover, Arabic does not combine two letters to produce a new sound. For example, combining the letters “t” and “h” in English sometimes produces a voiceless interdental fricative as in “think” or a voiced interdental fricative as in “though.” However, this distinction is highly systematic in English since in most lexical words the “th” is pronounced as it is in “think” while most functional words assume the other pronunciation.

The absence of short vowels from MSA texts makes it even more difficult for non-native speakers of Arabic to learn the language and presents challenges to the automatic processing of Arabic. Scripts such as Arabic, Chinese, Japanese, and Korean have neither capitalization nor strict rules of punctuation and their absence makes the task of preprocessing a text much more difficult. Some other challenges include:

**Derivational ambiguity**
- قاعدة: basis/principle/rule, military base, Qa'ida/Qua'ida/Quda

**Inflectional ambiguity**
- تكتب: /taktub/: you write, she writes

**Segmentation ambiguity**
- جد: جد + لغة: for a language

**Morphological ambiguity in Arabic**
It is a notorious problem that has not been sufficiently addressed (Kiraz 1998). This ambiguity is also increased by the inappropriate application of spelling relaxation rules and by overlooking rules that combine words with clitics and affixes (grammar-lexis specifications). Another source of confusion is whether to allow Arabic verbs to inflect for the imperative mood and the passive voice or not.

To make it less ambiguous, two morphological analyzers were utilised for Modern Standard Arabic (MSA) - Xerox Arabic Finite State Morphology and Buckwalter Arabic. Xerox adopted the overgeneralization that all verbs inflect for the imperative and the passive, leading it to overgenerate. Buckwalter’s morphology, on the other hand allowed only some verbs to have these inflections. Yet, because it did not follow a unique criteria or a systematic approach, the analysis is either underspecified or superfluous.

6. **The Role of Computer**
- Repeated exposure to the same material is beneficial or even essential to learning. A computer is ideal for carrying out repeated drills, since the machine does not get bored with presenting the same material and since it can provide immediate nonjudgmental feedback. A computer can present such material on an individualized basis, allowing students to proceed at their own pace and freeing up class time for other activities. The process of finding the right answer involves a fair amount of student choice, control, and interaction. Also, the computer can create a realistic learning environment, since listening can be combined with seeing, just as in the real world. Also, Multimedia and hypermedia technologies allow a variety of media (text, graphics, sound, animation, and video) to be accessed on a single machine. Hence,
skills are easily integrated, since the variety of media makes it natural to combine reading, writing, speaking and listening in a single activity.

Also, Internet technology facilitates communications among the teacher and the language learners. It allows a teacher or students to share a message with a small group, the whole class, a partner class, or an international discussion list of hundreds or thousands of people. Thus, Incorporating NLP techniques provide learners with more flexible—indeed, more ‘intelligent’—feedback and guidance in their language learning process. Therefore, it can be used for an educational environment. While doing so, the nature of the use of computer technology needs to be taken into consideration.

6.1 Computer as Tool

Hegazi, Ali, Abed and Hamada (1989) presented a way of representing Arabic syntax in Prolog as production rules. The system can detect some errors concerning Arabic syntax, and Abou Ela (1994) developed an expert system, the Arabic Syntax Analyzer (ESASA), which can be used as a tool to assist Arabic linguists in building Arabic grammar rules. The grammar is expressed using a declarative language called Grammar Writing Language (GWL). This tool is aimed at building Arabic natural language applications including CALL. However, using the Internet for publishing web-based CALL materials, that contain non-Latin alphabets, requires the solution of various technical problems.

6.2 Computer as Tutor

Gheith, Dawa, and Afifty (1996) developed Instructional Software for Teaching the Arabic Language (ISTAL) for grade one prep school. The system presents the curriculum as a simple concept associated with a set of generated sentences. Then, the system generates an exercise for the student and the student’s answer is automatically evaluated by comparing it to the system’s solution. Recently Nielsen (2001) and Nielsen and Carlsen (2003) developed a system for learning Arabic, Arab VISL, at the University of Southern Denmark. The system is an interactive Web-based application. It allows students of Arabic as a foreign language to analyse Arabic sentences by using Arabic script and specific Arabic grammatical terminology. The Interactive Language Learning Project at London Guildhall University has produced course materials for the University’s Arabic classes (Cushion & Hémard, 2003). The system is designed for learning Arabic at the beginner level. Mote, Johnson, Sethy, Silva, and Narayanan (2004) have also developed a speech-enabled computer learning environment designed to teach Arabic spoken communication to American English speakers, called Tactical Language Training System (TLTS). The advantage of this is that this system can detect errors in learner speech.

7. Limitations of Arabic ICALL

Arabic CALL system has some limitations. The system as described is targeted at a particularly well-formed subset of Arabic, which would not extend well to more colloquial dialects. Even standard newswire is likely to frequently include pre-verbal subjects and adverbials which are not considered in this work. This restriction to a well-formed subset might be appropriate for people trying to learn Arabic in a formal style. As vowels are usually omitted in written Arabic, the system does not handle the vowled Arabic text where letters are written with diacritic signs. Although ordering words to form a sentence is a type of question normally classified under the objective test method, it is not included into the system due to the free word order nature of Arabic that is usually dependent on semantics. As the task
of automatic processing of free natural language in ICALL is hard, the objective test method is used such that the expected learner’s answer is relatively short and well-focused. Also, the present system does not diagnose spelling errors. It accepts only answers that are free of typographical errors. We have designed, but only partially implemented due to lack of time and fund, an Arabic Spell Checker (Shaalan, Allam,&Gomah, 2003).

8. Conclusion
In this paper, I have explored number of areas related to the future of CALL with that of language teacher education. The apparent lack of sufficient training in the majority of existing teacher preparation programs and the growing interest among people in CALL proficiency were noted and it is believed that the future of CALL and teacher education is bright. However, there are a number of obstacles, the greatest of these is the limited number of qualified person able to integrate technology into language education effectively, a situation perhaps causally linked to the institutional reluctance to recognize and reward those who choose to devote their professional lives to this field. To sum up, if CALL is to survive and prosper, then we need a dedicated cadre of graduate students especially doctoral students, willing to select CALL as their area of specialization. With such option, it is hoped that the paths of CALL and language teacher education will increasingly be determined by such students and those they will educate in the decades to come.

About the Author:
Dr. BOUKHATEM Nadera is an MCB at the University of El Tarf ALGERIA. She received her M.D. from Tlemcen University and eventually earned her as a Lecturer at the University of El Tarf. In addition to teaching, she is a regular contributor to all the researches about NLP and TEFLE. She has also participated in various conferences in Algeria, France and Germany.

References:


18-Received June 2009; revised September 2009; accepted October 2009.