УДК 004.91

A. Glybovets, Zaid Envigi Mohammed Musbag

ARABIC NATURAL LANGUAGE PROCESSING

Challenges imposed by Arabic language nature push NLP to the extreme, motivating creativity and exhaustive exploitation of every single bit of already available techniques and linguistic resources. Our article is a first step to understanding problems and development of natural language processing for Arabic language.

Keywords: Arabic language, natural language processing.

Introduction

The Arabic language is both challenging and interesting. It is interesting due to its history [15], the strategic importance of its people and the region they occupy, and its cultural and literary heritage [4]. It is also challenging because of its complex linguistic structure [2].

At the historical level, Classical Arabic has remained unchanged, intelligible and functional for more than fifteen centuries. Culturally, the Arabic language is closely associated with Islam and with a highly esteemed body of literature. Strategically, it is the native language of more than 500 million speakers 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 Muslims perform their prayers five times daily. Linguistically, it is characterized by a complex Diglossia situation [7; 9; 11; 12]. Chronologically Classical Arabic represents the language spoken by the Arabs more than fourteen centuries ago, while Modern Standard Arabic is an evolving variety of Arabic with constant borrowings and innovations proving that Arabic reinvents itself to meet the changing needs of its speakers. At the regional level there are as many Arab dialects as there are members of the Arab league. Arabic is a Semitic language spoken by more than 500 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. Arabic is a highly structured and derivational language where morphology plays a very important role [1; 4; 6; 9; 13; 14]. Arabic Natural language processing (ANLP) applications deals 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 there is no capitalization in Arabic. In addition, Arabic letters change shape according to their position in the word. Modern Standard Arabic (MSA) 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 subject to recoverability of deletion [8].

As a natural language, Arabic has much in common with other languages such as English. However, it also is unique in terms of its history, diglossic nature, internal structure, inseparable link with Islam, and the Arabic culture and identity. Any Arabic NLP system that does not take the specific features of the Arabic language into account is certain to be inadequate. The challenge the Arabic language poses to researchers is not limited to the social aspects of the language, but also extends to its inherent linguistic structure which will be elaborated.

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, including machine translation, information retrieval and extraction, speech synthesis and recognition, localization and multilingual information retrieval systems, text to speech, and tutoring systems. These applications had to deal with several complex problems pertinent to the nature and structure of the Arabic language.

Significance of ANLP for the Arabic-Speaking Population

Funding for the development of ANLP applications has surged in the U.S. since September 11, 2001. The U.S. Department of Homeland Security was confronted with very difficult tasks ranging from identifying Arabic names correctly at airport security and in Arabic documents seized by the American authorities in the U.S. and abroad. They also had accumulated an enormous volume of Arabic texts that they had no clue as to whether they were relevant or not. They had neither the human

expertise needed to perform the task nor the time to wait for human translators to complete the task. ANLP tools that could scan such documents to recognize names, places, dates, etc., of interest soon became essential. As a result, funding became available for companies and research centers to develop tools such as named entity recognition, machine translation, especially spoken machine translation, document categorization, etc.

On the other hand, ANLP applications developed in the Arab World have different objectives and usually employ both rule-based and machine-learning approaches. The following are some of the objectives of ANLP for the Arab World:

1. Transfer of knowledge and technology to the Arab World. Most recent publications in science and technology are published in the English language and are not accessible to Arab readers which have little or no competence in English. To use human translators to translate such an enormous amount of data to Arabic is very costly and time consuming. So Arabic NLP could help reduce the time and cost of translating, summarizing, and retrieving information in Arabic for Arab speakers.

2. Modernize and fertilize the Arabic language. This follows from (1) above. Translating new concepts and terminology into Arabic involves coinage, arabization, and making use of lexical gaps in the Arabic language. This will positively effect the revitalization of the Arabic language and enable it to fulfill the essential needs for its speakers.

3. Improve and modernize Arabic linguistics. Arabic NLP needs a more formal and precise grammar of Arabic than the traditional grammar so widely employed today. Innovation is needed as well to preserve the valuable heritage of traditional Arab grammarians.

4. Make information retrieval, extraction, summarization, and translation available to the Arab user. The hope is to bridge the gap between peoples of the Arab world and their peers in more technically advanced countries. By making information available to Arabic speakers in their native language, Arabic NLP tools empower the present generation of educated Arabs. Thus Arabic NLP tools are indispensable in the struggle of Arabic speakers to attain parity with the rest of the world which is, in turn a matter of national security to the Arab World [10].

Rich Morphology

Arabic language has a very rich morphology. Words in Arabic are constructed out of prefixes, stem, infixes and suffixes. The stem itself is composed of two basic elements: the root and the morphological pattern (or diacritical pattern). The root could be a "tri root", consisting of three characters, or a "quad root" composed of four characters. The application of a morphological pattern to a root generates a stem, which is the basic form of an Arabic word token. Arabic language consists of about 6000 roots and 700 morphological patterns. Not all patterns could be applied to a given root. The actual valid root-pattern combinations in Arabic generate around 150,000 stems. Two thirds such stems are considered classical Arabic and only the remaining 50,000 stems are those actually used in daily life.

Arabic stems allow attachment of a multitude of prefixes and suffixes, hence constructing final form word tokens. Just to name some, prefixes in Arabic could be:

- Prepositions: ل → to / for; ب → with; ك → as
- Conjunctions: و → and; ف → then
- Adverbs: ف → so
- Auxiliary verb: س → will
- Interrogative: أ → did-have

On the other side of the word, suffixes have a much wider range of values, such as:

- Case ending suffixes (nominative/accusative/genitive)
- Number suffixes (dual/proper plural)
- Gender suffixes (masculine/feminine)
- Personal pronouns
- Object pronouns
- Genitive pronouns
- Possessive pronouns

By catenating prefixes and suffixes to a stem, a whole English sentence could be represented in one single Arabic word. Example:

Arabic: امكْلبباقأسف

English: Then I will meet both of you

The rich morphological nature of Arabic words represents an additional major obstacle for computational processing, especially on the morphological level. Example:

لضف → (لضْفَ)
(لَضَّف)
(لَّضَ + فَ)

In the above example, among the seven possible morphological interpretations for the input non-diacritized word (لضف), two will consider the first character an adverbial prefix, assuming a root totally different from the other five alternatives.

While the highly inflectional nature of Arabic poses many complexities in morphological analysis and disambiguation, it does however, in combination with the syntactic constraints of verb-subject and noun-adjective agreement, provide on the syntactic level, many useful clues serving structural analysis and disambiguation [3].

Free Word-Order

Arabic has a relatively free word-order syntax. Tokens constituting an Arabic sentence could be freely moved, without affecting the syntactic validity or the semantic interpretation of the sentence. Example:

Ate the man the apple أكل الرجل التفاحة
Ate the apple the man أكل التفاحة الرجل
The man ate the apple الرجل أكل التفاحة

One major problem arising from such flexible word-order, is the ability to swap subjects and objects, which would enable interpretations such as (the apple ate the man) that are syntactically valid but semantically wrong. Therefore, processing Arabic sentences requires syntactic analysis to permanently work with semantic analysis. Any attempt to attach a noun phrase to a verb as a subject, object, complement or even adverb, has to consult the semantic analyzer prior to attachment.

Arabic free word-order requires also a much more complex formal grammar, compared with its Latin counterpart, in order to reach a comprehensive coverage of valid Arabic structures [3].

Elliptic Personal Pronoun

One main characteristic of Arabic Language is the presence of an elliptical personal pronoun, which is always single but may be masculine or feminine. In combination with the omitted diacritics, elliptical pronouns create one of the most complex problems to Arabic computational processing. Consider the word «أكل» , without diacritics, this word could be one of eight possible words. Two such possibilities are

(i) أَكَلَ (eat),
(ii) أَكَّلَ (feed).

Although both word forms have a common root (أ ك ل), making them belonging to the same semantic cluster, different morphological patterns have resulted in totally different meanings and lexicon-syntactic features. One such feature, relevant to our context, is "transitivity". While (أَكَلَ) could be intransitive or transitive; (أَكَّلَ) can be transitive or intransitive. Based on the above, the sentence:

أكل الولد تفاحة Could have the following two interpretations:

(i) the boy ate an apple أَكَلَ الوَلَدُ تُفاحَةً
(ii) (He) fed the boy an apple أَكَّلَ الوَلَدَ تُفاحَةً

Where (i) has assumed the transitive alternative for the verb (أكل) and (ii) the intransitive one. Resolving this issue necessitates robust and intelligent syntactic analyzer, supported by a mandatory pronominal reference resolver [3].

Conclusion

Challenges imposed by Arabic language nature push NLP to the extreme, motivating creativity and exhaustive exploitation of every single bit of already available techniques and linguistic resources. The recent emergence of Arabic electronic texts (newspapers, magazines, books, Web sites,…etc.) is paving the road to the implementation and integration of new statistical-based modules within the originally rule-based system resulting in a more powerful and accurate hybrid system.

There are Arabic language features that are inherently challenging for ANL Researchers and developers. These morphology, the absence of the orthographic representation of Arabic short vowels from contemporary Arabic texts, the need for an explicit grammar of MSA that defines linguistic constituency in the absence of case marking. The new grammar also must describe important aspects such as anaphoric relations, the subjectless, sentences, and discourse analysis.

References

1. Attia M. A large scale computational processor of Arabic morphology and applications : Master's Dissertation, Computer Engineering / M. Attia ; Cairo University. – Cairo, Egypt, 1999.
2. Attia M. Handling Arabic morphological and syntactic ambiguities within the LFG framework with a view to machine translation : PhD Dissertation / M. Attia ; University of Manchester. – Manchester, 2008.
3. Achraf Chalabi. Challenges in Arabic NLP / Achraf Chalabi ; Sakhr Software Co. Sakhr Building. – Nasr City, Free Zone, Cairo, Egypt.
4. Bakkala M. H. Arabic Language : Through Its Language and Literature / M. H. Bakkala. – London : Kegan Paul, 2002.
5. Beesley K. Finite-state morphological analysis of Arabic at Xerox Research: Status and plans in 2001 [Electronic resource] / K. Beesley // Proceedings of the Workshop on Arabic Natural Language Processing at the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01). – 2001. – P. 1–8. – Mode of access: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.2.3703&rep=rep1&type=pdf. – Title from the screen.
6. Buckwalter T. Issues in Arabic orthography and morphology analysis / T. Buckwalter // Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (CAASL'04). – 2004.
7. Diab M. Arabic dialect tutorial / M. Diab, N. Habash // Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'07). – 2007.
8. Farghaly A. Subject pronoun deletion rule / A. Farghaly // Proceedings of the 2nd English Language Symposium on Discourse Analysis (LSDA'82). – 1982. – P. 110–117.
9. Farghaly A. Three level morphology for Arabic / A. Farghaly // Proceedings of the Arabic Morphology Workshop (AMW'87). – 1987.
10. Farghaly A. Arabic diglossia and Arabic identity in the information age / A. Farghaly. – Al-Fikr Al-Arabi, 1999.

11. Ferguson C. Diglossia. Arabic Natural Language Processing: Challenges and Solutions / C. Ferguson. – 1959.
12. Ferguson C. Epilogue: Diglossia revisited / C. Ferguson // Contemporary Arabic Linguistics in Honor of El-Said Badawi / The American University in Cairo. – Cairo, 1996.
13. McCarthy J. A prosodic theory of nonconcatenative morphology / J. McCarthy. – Linguistic Inquiry. – 1981. – Vol. 12, No. 3. – P. 373–418.
14. Soudi A. Arabic Computational Morphology: Knowledge-Based and Empirical Methods / A. Soudi, A. van den Bosch, G. Neumann. – Springer, 2007. – 308 p. – (Text, Speech, and Language Technology).
15. Versteegh K. The Arabic Language / K. Versteegh. – NY : Columbia University Press, 1997.

Глибовець А. М., Мусбаг Заід Енвігі Мохаммед

ОСОБЛИВОСТІ АВТОМАТИЧНОЇ ОБРОБКИ АРАБСЬКОЇ МОВИ

Складність арабської мови ставить перед методами обробки природної мови великі виклики і вимагає докладних досліджень. Ця стаття є першим кроком до розуміння проблем та спробою дати поштовх до пошуку їх вирішення в автоматичній обробці арабської мови.

Ключові слова: арабська мова, обробка природної мови, NLP.

Матеріал надійшов 02.09.2015

УДК 681.3

Олецький О. В.

ПРО ПІДХІД ДО АВТОМАТИЧНОГО ФОРМУВАННЯ РЕКОМЕНДАЦІЙ ДЛЯ ВІДВІДУВАЧІВ ВЕБ-ПОРТАЛУ НА ОСНОВІ ТЕОРІЇ НЕЧІТКИХ МНОЖИН

Розглянуто задачу автоматичного формування рекомендацій для відвідувачів тематичного порталу щодо того, які сторінки видаються найбільш перспективними для подальшого перегляду. При цьому взято до уваги, що рекомендовані матеріали не повинні бути ні надто схожими на поточну сторінку, ні надто віддаленими від неї. Розглянуто функцію залежності між мірами релевантності та відстанями між документами, для опису якої використовується апарат теорії нечітких множин. Запропоновано методику розрахунку мір релевантності на основі відповідного нечіткого правила, наведено конкретний приклад такого розрахунку.

Ключові слова: тематичний портал, рекомендаційна система, міри релевантності, нечітке правило.

Вступ

У роботах [1–5] розвивається напрям, пов’язаний з автоматичним формуванням рекомендацій щодо добору найбільш релевантних навчальних матеріалів на тематичному порталі. Мова йде про ситуацію, в якій основною метою відвідувача порталу є отримання якомога більш повної та всебічної інформації з певного питання, інакше кажучи, максимізація рівня своїх знань з певної