What are the Challenges faced in the processing of Arabic language

Nasria Bouhyaoui

Randa Benkhelifa

Department of Computer Science and Information Technologies

Université Kasdi Marbah Ouargla

تاريخ الإرسال: 2019/05/18	تاريخ القبول:2019/11/27	تاريخ النشر: 2019/12/01

Abstract

Arabic is considered a Semitic language closely related to Hebrew and Aramaic. It ranks in fifth place of the most spoken languages in the world. Arabic is official Language in 26 countries, and it is the mother tongue of approximately 295 million speakers (McCarthy, 2018). There is a rapid growth of the content of Arabic language on the internet. This calls for the necessity to develop Arabic Natural Language Processing applications. The process of automatic understanding of texts is what we call Natural Language Processing. There is a lack of NLP applications in Arabic language in comparison with other languages in addition the results of processing of the existed application is considered poor because of the complexity of Arabic language nature. The chief purpose of the current study is to address the features of Arabic language that make the task of Arabic NLP a hard task.

Key words : Arabic language, Arabic texts, Natural Language Processing, Artificial Intelligence.

مُلخّص البحث

تعتبر اللغة العربية لغة سامية مرتبطة ارتباطًا وثيقًا بالعبرية والآرامية. وهي تحتل المرتبة الخامسة بين أكثر اللغات تحدثاً في العالم. اللغة العربية هي اللغة الرسمية في 26 دولة ، وهي اللغة الأم لحوالي 295 مليون شخص (McCarthy, 2018). هناك نمو سريع في محتوى اللغة العربية على شبكة الإنترنت. وهذا يستدعي ضرورة تطوير تطبيقات معالجة اللغة العربية الطبيعية. عملية الفهم التلقائي للنصوص هي ما نسميه معالجة اللغة الطبيعية. هناك نقص في تطبيقات البرمجة اللغوية العصبية في اللغة العربية بالمقارنة مع اللغات الأخرى بالإضافة

إلى أن نتائج معالجة التطبيق الموجود تعتبر سيئة بسبب طبيعة اللغة العربية. الغرض الرئيسي من الدراسة الحالية هو دراسة ميزات اللغة العربية التي تجعل مهمة معالجة اللغة العربية مهمة صعبة.

الكلمات المفتاحية: اللغة العربية ، النصوص العربية ، معالجة اللغة الطبيعية ، الذكاء الاصطناعي.

Introduction

Arabic is considered a Semitic language closely related to Hebrew and Aramaic. It ranks in fifth place of the most spoken languages in the world. Arabic is official Language in 26 countries, and it is the mother tongue of approximately 295 million speakers (McCarthy, 2018) as illustrated in Fig. 1.

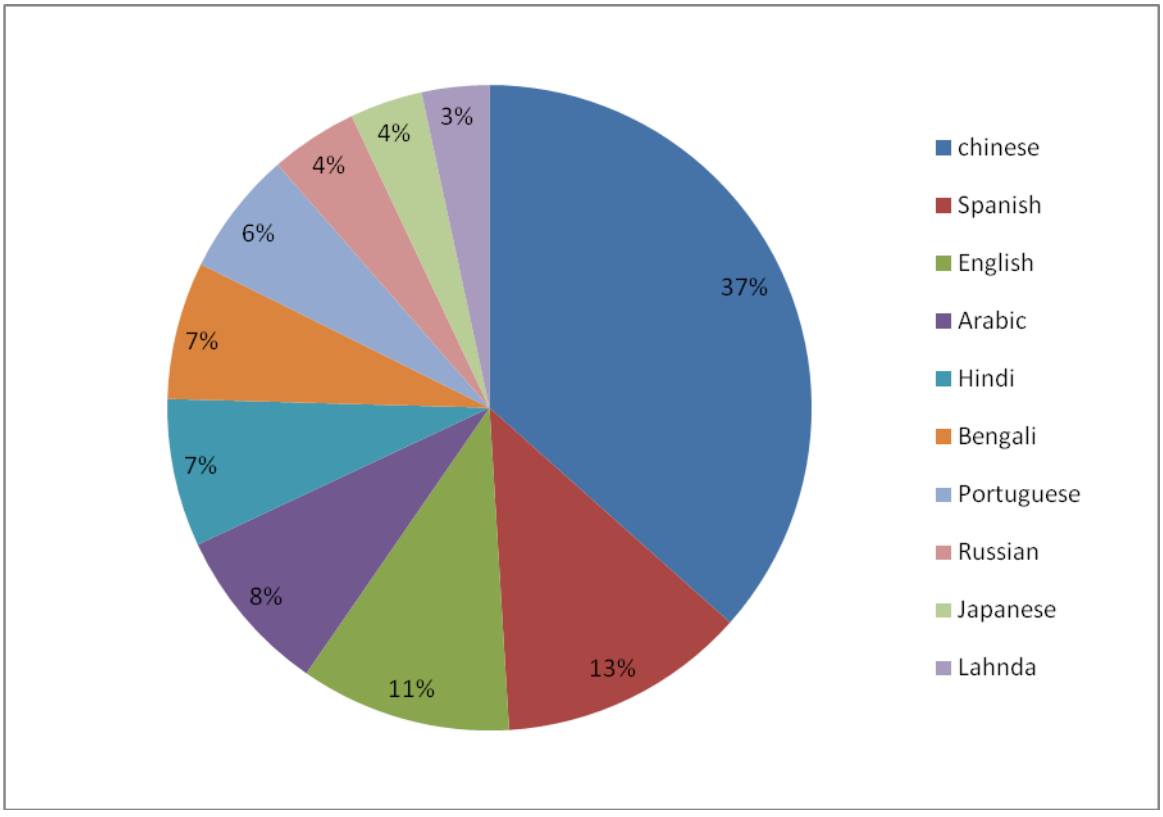

Fig. 1 The Top 10 spoken languages in the World

The Arabic alphabet consists of 28 letters, these letters are the consonants and long vowels .the letters can be extended to ninety by added shapes, marks, and vowels

(Khorsheed, 2002). The form of the letter differs depending on whether it occurs at the beginning, middle, end or alone. In contrary to Latin-based alphabets, the Arabic writing is from right to left (Amin, 1998).

Modern Standard Arabic is different from the classical Arabic, where the latter is the regular version of Arabic which is the language of the Holy Qur'an. The Modern Standard Arabic is written without short vowels (Boudelaa, 2010), and this is one of the main features that distinguish it, where it is used in the official documents, Books, newspapers, etc.

Arabic is characterized by a rich and complex morphology (Farghaly and Shaalan, 2009). Therefore, Arabic language "presents significant challenges to many natural language processing (NLP) applications (Farghaly and Shaalan 2009)." (Zitouni, 2011).

Natural Language Processing (NLP) is a subject of interest of many fields like computer science, artificial intelligence and linguistics, robotics, etc ((Liddy, 2001); (Chowdhury, 2003)). NLP has been investigated in many tasks including: question answering, text summarization, machine translation, speech recognition, opinion mining, and text categorization. The process of automatic understanding of texts is what we call natural language processing. So, NLP techniques are developed for the aim of understanding human language (Chowdhury, 2003); (Liddy, 2001)) by machines.

The author Liddy (2001) provided the following definition for NLP "Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications".

NLP starts as the intersection of Artificial Intelligence and Linguistics. "Linguistics is the science of language which includes Phonology that refers to sound, Morphology word formation, Syntax sentence structure, Semantics syntax and Pragmatics which refers to understanding" (Plisson et al., 2004).

The development of NLP applications is challenging task. Where, the difficulties of processing language differ from one language to other language. In Arabic language, there is a lack of NLP applications in comparison with other languages. The result that is obtained from the processing of the existed application is considered poor. The complexities related to Arabic language nature make the task of processing Arabic language a hard task. So the question asked is: What are the features that make Arabic language processing a complex task?

In this paper we aim to present the main levels of NLP and to address the particularities of Arabic language.

Natural Language Processing

Natural Language Processing (NLP) is a subject of interest of many fields like computer science, artificial intelligence and linguistics, robotics, etc ((Liddy, 2001); (Chowdhury, 2003)). NLP has been investigated in many tasks including: question answering, text summarization, machine translation, speech recognition, opinion mining, and text categorization. The process of automatic understanding of texts is what we call natural language processing. So, NLP techniques are developed for the aim of understanding human language (Chowdhury, 2003); (Liddy, 2001)) by machines.

The author Liddy (2001) provided the following definition for NLP "Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications".

NLP starts as the intersection of Artificial Intelligence and Linguistics. "Linguistics is the science of language which includes Phonology that refers to sound, Morphology word formation, Syntax sentence structure, Semantics syntax and Pragmatics which refers to understanding" (Plisson et al., 2004).

Natural Language Processing layers

NLP systems mission requires the use of several methods. NLP methods can consist of different representation levels including Phonology, Morphology, Lexical, Syntactic, Semantic, Discourse and Pragmatic (Liddy, 2001).

Generally, NLP methods fall in three levels: syntactic, semantic and pragmatics.

Syntactic level

The syntactic level focuses on the definition of sentence structure which is the constituent words of the sentence. In addition, it is interested in the detection of the sentence grammatical structure (Liddy, 2001). So, it concerns in the grammar and sentence structure. This level produces as a result a range of structural relationships between the sentence words. Syntactic layer has several modules including tokenization, Sentence Boundary Disambiguation, Lemmatization, Part-of- Speech tagging. In the following we describe some of the modules:

- Sentence Boundary Disambiguation

Sentence Boundary Disambiguation is the task of text segmentation into sentences. This deconstruction is an important phase in order to deal with more granular portion of text for the aim of achieving a particular treatment. In English and many other languages, the punctuation marks are used for recognizing boundaries of sentences. However, Finding Boundary of sentences is a challenging task in some languages (e.g. Arabic, Chinese). For example, the Arabic language does not follow a strict punctuation rules. Where, Arabic paragraphs can be written with only one period at the end of the paragraph. Also, it is characterized by the use of coordination, subordination, etc. these features made the task of **Sentence Boundary Disambiguation** so complex.

- **Lemmatization**

Lemmatization aims for the construction of a normalized form (Plisson et al., 2004). Lemmatization is the process of converting words into their basic word form, which is called the lemma. In another words, lemmatization is the grouping together of different forms of the same word. Lemmatization process mapping all verb forms to infinite tense and converting nouns to a single form (e.g. The words help, helps, helped, helping are mapped to the verb help). Lemmatization is considered a hard task when it comes to process highly inflected languages (Toman et al., 2006).

- Part-of-Speech Tagging

Part-of-Speech Tagging (POS) tagging is considered one of the important task in NLP. POS tagging consists of defining a speech part of a particular sentence through the association of labels such as noun, pronoun, verb, adverb, adjective, etc. to each word or token in the given sentence. The definition provided in Wikipedia is "*In corpus linguistics, part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context— i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph*."[1]

Semantic level

The Semantics level input is the output of the syntactic layer. Semantics level is applied when it comes to extract meaningful information. This level concerns the definition of the meaning of words and sentences (Chowdhury, 2003). The author Liddy (2001) said that:"Semantic processing determines the possible meanings of a

[1] https://en.wikipedia.org/wiki/Part-of-speech_tagging] 10/11/2018

sentence by focusing on the interactions among word-level meanings in the sentence". Accordingly, the sentence meaning is derived relying on words meaning (Cambria et al., 2017). Semantic layer has several modules including Named Entity Recognition, Concept extraction, Word sense disambiguation and Anaphora resolution, etc. In the following we briefly describe some of modules that belong to this level:

- Named Entity Recognition

  Named Entity Recognition (NER) also known as (*entity extraction* and entity identification) is a popular technique applied in order to extract relevant information or entities from unstructured text. NER is an important method used for identifying and classifying named entities in free text into pre-defined categories or classes such as names of persons, organizations, locations, times expressions, numerical quantities, etc.

- Word sense disambiguation

  Word sense disambiguation (WSD) task consists of recognizing the meaning of an ambiguous word depending on its context (Stevenson and Wilks, 2003). Some words can convey multiple meanings; the correct word sense is detected according of the particular context. Identifying the correct sense is an important task in NLP. WSD is an open problem that needs to be solved in order to improve understanding of natural language (Cambria et al., 2017). Where, it is so difficult to define the right meaning of the word in different context.

Pragmatic level

The Pragmatic level requires in its task both syntactic and semantic layers. The pragmatic level is interested in the meaning of the text. That meaning extracted depending on the context of the text. The authors Cambria and White (2014) say that "Pragmatics deals with how meaning changes in the presence of a specific

context and how the contexts affect the meaning of the sentences". Pragmatic layer has several modules including Sarcasm Detection, Aspect extraction, Polarity detection, etc.

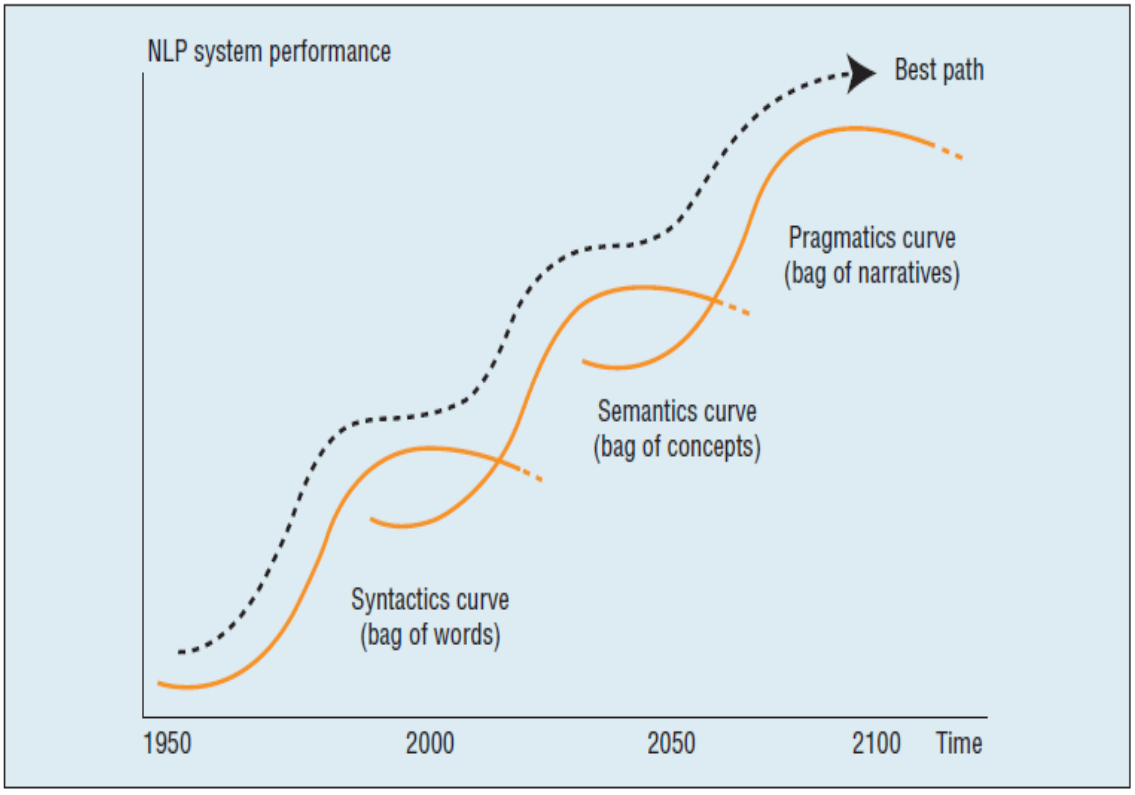

Fig. 1 Jumping NLP curves (Cambria et al., 2017).

The process of understanding natural language is a complex task, where, many challenging tasks can be faced. The authors Cambria et al. (2017) illustrated the estimation of the evolution of NLP research through the three layers (syntactic, semantic and pragmatic) that lead the NLP research to arrive at the natural language understanding. The illustration is represented in the Fig.1. The difficulty of NLP processing varies from language to language, following the nature or the characteristics of the language.

Arabic language processing challenges

Despite the efforts that are exerted in the field of automatic Arabic language processing, there are modest results. Several problems arise when dealing with the automatic processing of text. The difficulty of treatment lies in nature of the language treated (Harmanani et al., 2006). This particularity is illustrated in the Arabic language, in which the automatic processing of text is a difficult task due to their complex morphology (Al-Kharashi and Al-Sughaiyer, 2004; Alrahabi et al., 2006).

In te following we depict some of characteristics of Arabic language:

1. The vocalization is so important in Arabic language (Al-Kharashi and Al-Sughaiyer, 2004), the same word can express several meanings, according to the vocalization (Al-Kharashi and Al-Sughaiyer, 2004; Alrahabi et al., 2006; Bousmaha et al., 2013). Words are indicated by diacritics, or short vowels that are written above or below the letters, for example, the non-vowel word "**ذهب**" may take a the meaning of "golden" with the diacritical marks: "**ذَّهبَ**". And the meaning of "to go" with the diacritical marks: "**ذَّهبَ**".However, recently vocalization is almost unused; we find only some vowels used in some words.
2. The agglutinative nature of Arabic language, for example, we take the word '**بإمكانهم**': -which is resulted from the agglutination of the prefix "**ب**" which is a preposition, the stem "**إمكان**", and the suffix "**هم**" which is a possessive pronoun. The composition of the parts (prefix, stem, and suffix) can create a complex morphology.
3. Unlike other languages that depend on the capitalization feature in order to distinguish for example the name of persons, Arabic language does not use capitalization. That makes the Task of NER a hard task.
4. The same sentence in Arabic can be expressed with different structures (El Kassas and Kahane, 2004); where the word in the sentence can take different orders, but the meaning remains the same. We mention for example:
    - "**ذهب أحمد إلى المدرسة**".

- "أحمد ذهب إلى المدرسة".

Conclusion

Improving Arabic language application is so essential in order to improve the research and access to the Arabic information. In this paper we present a … on some of the main features or characteristics of Arabic language that make their processing a complex task. We starts by addressing some levels of NLP and we discussed the points that are considered as a challenging task in NLP in general. After that we present some features of Arabic language.

References

Al-Kharashi, I. A., Al-Sughaiyer, I. A. (2004) 'Arabic morphological analysis techniques: A comprehensive survey'. *Journal of the American Society for Information Science and Technology,* Vol. 55, No.3, pp.189–213.

Alrahabi, M., Ibrahim, A. H., Descles, J. P. (2006) 'Semantic Annotation of Reported Information in Arabic'. *Proceedings of FLAIRS 06,* Melbourne Beach, Florida, 11-13 May, pp. 263-268, aaai press.

Amin, A. (1998) 'Off-line Arabic character recognition: the state of the art'. *Pattern Recognition,* Vol. 31, No. 5, pp. 517–530.

Boudelaa, S., Marslen-Wilson, WD. (2010) 'Aralex: a lexical database for Modern Standard Arabic'. *Behavior Research Methods*, Vol. 42, No. 2, pp. 481-487.

Bousmaha, K.Z , Charef_Abdoun,S., Hadrich_Belguith, L., Rahmouni, M.K. (2013) 'Une approche de désambiguïsation morpho_lexicale évaluée sur l'analyseur morphologique Alkhalil'. *La revue de l'Information Scientifique et Technique (RIST),* Vol. 20, No. 2, pp.32-46.

Bousmaha, K.Z , Charef_Abdoun,S., Hadrich_Belguith, L., Rahmouni, M.K. (2013) 'Une approche de désambiguïsation morpho_lexicale évaluée sur l'analyseur morphologique Alkhalil'. *La revue de l'Information Scientifique et Technique (RIST),* Vol. 20, No. 2, pp.32-46.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational intelligence magazine, 9(2), 48-57.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. IEEE Intelligent Systems, 32(6), 74-80.

Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.

El Kassas, D., Kahane, S. (2004) 'Modélisation de l'ordre des mots en arabe standard'. Atelier sur le Traitement Automatique de la Langue Arabe Ecrite et Parlée, pp.259-264.

Farghaly, A., Shaalan, K. (2009) 'Arabic Natural Language Processing: Challenges and Solutions'. *ACM Transactions on Asian Language Information Processing (TALIP),* Vol. 8, No. 4 ,pp.14.

Harmanani, H. M., Keirouz, W., Raheel, S. (2006) 'A Rule-Based Extensible Stemmer for Information Retrieval with Application to Arabic'. *The international Arab journal of Information Technology,* Vol. *3*, No.3 , pp. 265-272.

Khorsheed, M. S. (2002) 'Off-line Arabic character recognition--a review'. Pattern analysis & applications, Vol. 5, No. 1, pp. 31-45.

Liddy, E. D. (2001). Natural language processing

McCarthy, N. (2018) The worlds most spoken languages, https://www.statista.com/chart/12868/the-worlds-most-spoken-languages/. mai 10, 2019

Plisson, J., Lavrac, N., & Mladenić, D. (2004). A rule based approach to word lemmatization.

Stevenson, M., & Wilks, Y. (2003). Word sense disambiguation. The Oxford Handbook of Comp. Linguistics, 249-265

Toman, M., Tesar, R., & Jezek, K. (2006). Influence of word normalization on text classification. Proceedings of InSciT, 4, 354-358

Zitouni, I. (2011). Book Reviews: Introduction to Arabic Natural Language Processing by Nizar Y. Habash. Computational Linguistics, 37(3).