ARABIC LANGUAGE CHALLENGES IN TEXT BASED CONVERSATIONAL AGENTS COMPARED TO THE ENGLISH LANGUAGE

Mohammad Hijjawi[1] and Yousef Elsheikh [2]

[1] Department of Computer Science, Applied Science University, Amman, Jordan
[2]Department of Computer Information Systems, Applied Science University, Amman, Jordan

ABSTRACT

This paper is not to compare between the Arabic language and the English language as natural languages. Instead, it focuses on the comparison among them in terms of their challenges in building text based Conversational Agents (CAs). A CA is an intelligent computer program that used to handle conversations among the user and the machine. Nowadays, CAs can play an important role in many aspects as this work figured. In this paper, different approaches that can be used to build a CA will be differentiated. In each approach, the comparison aspects among the Arabic and English languages will be debated with the respect to the Arabic language.

KEYWORDS

Conversational Agents, Chatterbots, Arabic Language and English Language.

1. INTRODUCTION

A Conversational Agent is an intelligent computer program has been developed to handle utterances/conversations among the user and the machine. CAs has been started with the idea of Alan Turing introduced which is Turing test [1]. Turing test or Turing game modelled to test the human's ability of differentiating among the second conversation party character. The human will chat with the second party which might be another human or a machine via textual conversations. The game consider the first human fail if he/she failed to judge that the second party in the conversation was a human or a machine. From Turing's time, many of CAs has been proposed and developed until these days such as ELIZA [2], ALICE [3], ADAM [4], Museum Guide[5], InfoChat[6], InCA[7], Abdullah [8]andArabChat[9]. Most of the previous CAs was for English language. A manual listing that the authors built for the current CAs appeared that more than 97% of current CAs is not for the Arabic language and most of them was for the English. To best of our knowledge, this is due to three main factors; firstly, Arabs has adopt the Internet in general in 1990s which might be considered late comparing to the English based countries. This late affected the research and adopting the technology in general. Secondly, the English language has the first rank in terms of number of speakers worldwide [10] which lead the English based countries and the worldwide companies to focus building their CAs for the English language. Thirdly, the difficulties and challenges that the Arabic language facing.

This paper discusses the challenges that the Arabic based CAs face comparing to the English based CAs.

DOI:10.5121/ijcsit.2015.7301

The Arabic language is a Semitic language that originated in the period between the seventh century BC and the third century AD in the Middle East region [11]. Arabic, like other languages, affects and is affected by other languages in terms of accepting new vocabulary in its dictionary especially from the technology field. Arabic spread significantly when Islam spread to Asia, Africa and Europe. This consequently increased the number of Arabic speakers in those regions, especially because it is the spiritual language of Islam [11].Presently, different types of Arabic language are used in daily conversations between people. These are known as Classical, Modern, and Colloquial Arabic [11]. Arab speakers usually use these different types of Arabic depending on the nature of conversation and sometimes in the same conversation which caused a diglossia. The Diglossia is a phenomenon whereby two or more kinds of the same language exists in the same speech community [12].

Classical Arabic, as used in the eras before Islam and in the Quran, is more complex in its grammar and vocabulary than modern Arabic. Classical Arabic has a large number of diacritics that facilitate word pronunciation and detection in their grammatical cases (for instance, noun or verb). In the Modern Arabic, these diacritics were omitted to accelerate the reading and writing processes. Modern Arabic is formally recognised as the official language of the Arab countries. It is used in everyday language, in the media, education, and literature. Most of computational Arabic-based research use Modern Arabic [13]. The third type, known as Colloquial Arabic, is less sophisticated than Modern Arabic in its grammar and vocabulary. However, due to its simplicity, most people use it in their everyday spoken conversations and in informally written letters [14].

Arab people have weaknesses in using Modern Arabic which make them generate incorrect sentences in terms of grammar. In addition, they might mix between Modern and Colloquial Arabic, which increases the challenge for a CA to understand or recognise the user’s utterance. Moreover, different Arab countries have different Colloquial Arabic or “dialects”, but most these countries can understand each other. This will might increase the challenge for an Arabic CA to understand or recognise user utterances from various Arab countries when they used their Colloquial Arabic. In contrast, this problem will be smaller in the English language as it nowadays has mainly two types which are the standard English and colloquial English. However, the two types are closed to each other in grammar and the sentence’s structure. In addition, the standard English is the same in all countries that speak it such as America, United Kingdom and Canada. Although, the colloquial English is differ from country to country but this difference is in the accent (the spoken language) not in the written text language.

Arabic has 28 letters, each of which has many written forms depending on their position in the word (initial, middle, or end) such as the letter ‘ب’ ‘b’ which is in the independent shape has three forms in Arabic which are ‘بـ’ when it is written in first of a word, and ‘ـبـ’ when it is written in the middle and ‘ـب’ when it is written at the end of a word. In contrast, the letter ‘b’ has the same form wherever written in a word. These different forms are due to the Arabic writing style (concatenative and cursive). Arabic differs from English in terms of the direction of writing, it iswritten from right to left [16]. Where the English written from the left to right. Moreover, Arabic does not have capital or small letters, and does not support capitalisation features. These missed features in Arabic will increase the processing challenges as will discuss in this paper.

Generally, linguistic textual CAs has three main approaches to build which are; Natural Language Processing based approach, sentence similarity measures based approach and Pattern Matching approach. These approaches can be used to build the Arabic based and English based CAs. However, all of them has different research history for Arabic and English but the English language research amount has the majority of this research history. The next section describes the variances among these approaches for the Arabic and English.

2. Conversational Agent approaches challenges for the Arabic and English

As mentioned above, three main approaches can be adopted to develop CAs, namely, Pattern Matching (PM), Natural Language Processing (NLP), and Sentence Similarity Measures (SSM).Researchers should take care when they are deciding which approach they are going to choose in order to build their CA in terms of the compatibility between the used approach and the targeted natural language. In general, each approach has advantages and disadvantages. These advantages and disadvantages might be increased or decreased depending on the used approach with a specific natural language. In other words, an approach that could be useful when used with one natural language might not be useful enough when used with another natural language for number of reasons as the following subsections investigate.

2.1 Natural Language Processing based CA

NLP is defined in computational linguistics as “the computational processing of textual materials in natural human languages”[17]. NLP might also be defined as studying the constructing and meanings of a natural language through applying its rules to process information enclosed within its sentences [18]. The general aim of NLP is to develop computational techniques, which will analyse huge number of spoken or written texts in the manner of human carrying out the same task [17]. NLP emerged in the 1950s with machine translation, and it had further focusing by research in artificial intelligence field. More concentration has been placed on expert systems or intelligent programs to simulate human behaviour and human’s knowledge in order to make inferences and interpretations and then reach conclusions like human.

NLP seems to be the most effective way to build CAs especially when its techniques try to deal with utterances semantically, which means understanding the utterance’s subject and its content. Understanding a sentence needs to understand the sentence’s structure, its components (words) and the relations among those components.

Arabic and English has a linguistic morphology features. According to [13], a “Morphology is the study of internal word structure”. It is usually focuses on two fundamental issues: firstly, derivational morphology, which concerns how words are formed; and secondly, inflectional morphology, which concerns how words interact with the syntax [11]. However, derivational morphology governs the principles of a word’s transformation from its root. For instance, from the root “كتب”, many different words might be generated such as “يكتب” “he is writing”, “تكتب” “she is writing”, “مكتوب” “written or a letter”, and “مكتبة” “library”. This richness of generating words with different semantic meanings presents a challenge to any kind of Arabic based computational processing.Morphological features work upon integrated dependencies among several linguistic factors, such as vowels, affixes and root-based systems, as the rest of this section will describe.

Affixes and root base system

Semitic languages such as Arabic have a very rich derivational and inflectional features that are based on a root to generate the language’s words. A root is the initial form of a word that cannot be analysed further [15]. For instance, from the root “كتب”, many different words might be generated such as “يكتب” “he is writing”, “تكتب” “she is writing”, “مكتوب” “written or a letter”, and “مكتبة” “library”.

The Arabic language is based on a root system to generate its words. Arabic has more than 10,000 roots [15]. According to Al-Fedaghi and Al-Anzi[19], “85% of words are derived from tri-literal roots”; the rest are derived from quad-literal, pent-literal and hex-literal roots [15].

Affixes such as prefixes, infixes, and suffixes can be added before, inside, or after a root respectively, to generate more meaningful words. Affixes in Arabic such as "ا", "ال", "وال", "ب", "بال", "ن", "ون" and "والا" can be generated from the combination of letters of the word "سألتمونيها" "You asked her" [13].

Arabic derivational and inflectional morphology features are based on affixes. For instance, the derivational morphology feature can generate words such as "مدرس, مدروس, دارس" from the root "درس" by mapping the root with Arabic word-based patterns. Inflectional morphology, which are generated by adding prefixes or suffixes to a root, may reflect a tense (past/present), a gender (masculine, feminine) and/or a number (singular, dual, plural). For instance, from the root "درس" many words might be generated, such as "مدرسون "teachers"درس "I taught", مدرستان"two teachers or two schools", مدرسة"a school" ".

Arabic grammar has a number of powerful structures to govern word transformation, for instance words patterns. Words patterns are words formed based on affixation and vocalisation processes on the Arabic roots. They can be classified into noun patterns and verb patterns [15] and they formed based upon the essential pattern "فعل". For instance, the word "عامل" "Worker" belongs to pattern (فاعل), and "يعمل" "He works" belongs to pattern (يفعل). The generated words from the same root might not be related to each other semantically, which increases the ambiguity in Arabic semantic processing [20]. For instance, the root "شهر" can generate the word "شَهر" which means "month", and "مَشهور" which means "famous". Table 1 shows an example of how derivational and inflectional morphology features can reform the root ("طلب"). In Table 1, adding the infix "ا" to the mentioned root, a new word will be generated which is "طالب" "a male student", while adding the prefix "ي" to the root, it will generate the word "يطلب" "He requests". These rich morphological features will increase the challenge for a CA to determine the grammatical case for a user's utterance words. This might leads to an inability to understand or recognise the user utterance for CAs based on NLP.

Table 1: Examples of how derivational and inflectional morphology can reform a root.

Arabic word	**Prefix**	**Infix**	**Suffix**	**Stem**	**Root**	**English Translation**
الطالبين	ال	ا	ين	طالب	طلب	Students (dual, masculine)
الطالبتين	ال	ا	تين	طالب	طلب	Students (dual, feminine)
الطالبان	ال	ا	ان	طالب	طلب	Students (dual, masculine)
الطالبتان	ال	ا	تان	طالب	طلب	Students (dual, feminine)
الطلاب	ال	ا	---	طلاب	طلب	Students (plural, masculine)
الطالبات	ال	ا	ات	طالب	طلب	Students (plural, feminine)
الطالب	ال	ا	---	طالب	طلب	Student (Singular, masculine)
الطالبه	ال	ا	ه	طالب	طلب	Student (Singular, feminine)
يطلب	ي	---	---	طلب	طلب	He requests (present tense, singular, masculine)
تطلب	ت	---	---	طلب	طلب	She requests (present tense, singular, feminine)

Vowels

English, like the Arabic language, uses vowels and consonants. However, there are two types of vowels in Arabic, short vowels (symbols) and long vowels (letters). Short vowels are symbols added above or below the letters (also known as diacritics). These symbols have grammatical functions to express a word's grammatical case [13]. Therefore, these symbols can change the whole meaning of the word. For instance, the word "حُب" means "love", but changing its diacritics to "حَب" changes the meaning to "seed".

Arab people use these diacritics when they speak but to increase the speed of reading and writing, they often omit them from Modern Arabic written forms. However, omitting diacritics from words might confuse the reader when these words are read in isolation or out of context. For instance, the word "عين" alone and without diacritics might have many meanings, such as "eye", "spy", "assistance", "a spring of water" or "ع" (one of the Arabic letters). When the word "عين" has its diacritics, or is in context without diacritics, this confusion will be resolved. In rare cases, the reader will not be able to differentiate between un-vocalised words even if they are put in their context. For instance, in the sentence "عمان جميلة" the reader will not be able to know if the writer intended "Amman is nice" or "Oman is nice", because the word "عمان" without diacritics might mean "Amman" or "Oman". To resolve this confusion, the writer has to put the suitable diacritics on the word (to be "عَمّان", which means "Amman", or "عُمَان", which means "Oman").

Keeping the diacritics might be useful in applications such as translation and transliteration in order to overcome the aforementioned confusion problems and to detect words' grammatical cases. In contrast, keeping the diacritics in other applications might not be necessary, such as in stemming applications that return a word to its original root and in CAs that use Modern Arabic. Existing the diacritics will increase the processing/analysing complexity to the maximum in Arabic. In contrast, there is no diacritics in English which leads to reduce the analysing complexity and save the computational processing time and effort.

Long vowels in Arabic are comprised of three different letters: ا,و and ي. These letters can either be originally part of the root or are infixes in a word. Long vowels might change or replace each other during a word transformation process. For instance, the word "قام" means "he stands" and the word "يقوم" means "he is standing", both of which have the same root "قوم". This example shows that the vowel letter "و" in the root "قوم" has been replaced with the vowel letter "ا" during the transformation process of word "قام". No fixed rules control these changes during a word transformation [21], which increases the difficulty of analysing such words to understand the sentence by an NLP based CA.

Nouns and verbs

A noun is "a word that indicates a meaning by itself without being connected with the notion of time" [22], whilst a verb is a word that has meaning with being connected with notion of time [16].

Unlike English, nouns in Arabic express three number-based cases: singular, dual and plural. For instance, the word "teacher" in English expresses the singular form of the noun and "teachers" expresses dual or plural cases. In contrast, the Arabic word "مدرس" meaning "a teacher" express the singular status of the noun, and the words مدرسان and مدرسين meaning "two teachers" express the dual case; مدرسون and مدرسين meaning "three teachers or more" expresses plural status. This happened by adding some affixes to the mentioned word. In dual and plural nouns, as shown in the previous example, two different words are used to express their cases. The selection process between the two words for each pair is based on complicated grammatical rules or using

diacritics. Therefore, Arab people habitually make mistakes when choosing between them. This will make detect and analyse these words variations for a CA understanding is much difficult.
Most natural languages differentiate between male (masculine) and female (feminine) words. In Arabic, this differentiation is expressed simply by pronouns such as “هو” “he” and “هي” “she”. In addition, this differentiation might be achieved by adding some affixes to words to express their gender. For instance, in English, the word “player” expresses both male and female, but in Arabic the word “لاعب“ is used to express the male player, and the suffix letter “ه“ should be added at the end of the word “لاعب“ to be “لاعبه” for a female player.

In Arabic, masculine and feminine nouns might be related to humans, animals, or even inanimate objects. In English, all inanimate objects use the word “it” to express them, but in Arabic the inanimate object can be verbally masculine or feminine. For instance, the word “table” is considered as feminine “هي طاولة“ and the word “chair” is considered masculine “هو كرسي“. Therefore, detecting such pronouns (“He” “هو”, “She” “هي”) in Arabic does not reflect if the intended something is a human or an inanimate. This will increase the challenge to detect and analyse a sentence in Arabic.

Verbs in Arabic can be generated to express masculine and feminine. For instance, the verb “يدرس” “he is studying” is assigned for a male student, while “تدرس” “she is studying” assigned for a female student. Moreover, unlike the English, verbs can express the number of people who are doing the action. For instance, the verb “يدرس” “he is teaching” is used to express that there is one male person doing the action of teaching, while the verb “يدرسان” “they are teaching” is used to express that there are two male people doing this action. However, the verb “يدرسون” “they are teaching” is used to express that there are more than two male teachers doing the action. This richness of Arabic morphology makes it difficult to an NLP based CA to analyse such verbs for the large number of affixes and then inability of determining if the intended person is single, dual or plural.

Proper nouns and foreign Words

Proper nouns in Arabic include the “name of a specific person, place, organization, thing, idea, event, date, time, or other entity” [23]. Unlike English, Arabic does not support capitalisation of proper nouns such as “محمد” “Mohammad” [24]. Differentiating the proper nouns in English is easy as the word has been written capitalised. In contrast, in Arabic there is no way to differentiate them. In addition, foreign arabised words are words that are borrowed from other languages, mainly from English, such as microwave”مايكرويف”. These words cannot be broken down into roots because originally they are not Arabic words and thus do not have roots. As a result, a Conversational Agent must deal with such a word as it is. Arabised words have no standardised writing form. For instance, the word “microwave” has many written forms, such as “مايكرويف” and “ماكرويف”. Thus the challenge of dealing with these words is recognising them within the conversation. This increases the difficulty of detecting such nouns in textual Arabic conversations.

Unfortunately, NLP suffers from many limitations and it still faces many challenges. For instance, a word might have more than one meaning (Lexical ambiguity) or a sentence might have more than one structure (structure ambiguity). In addition, these ambiguities are different between languages. A language has one phrase to express something while another language might have a single word to express the same thing. Moreover, the real life conversations are often grammatically incorrect and the conversation processing in the machine takes a long time [25][26]. Furthermore, the natural language used by humans is extremely rich in form and structure and very ambiguous.

These challenges increase when developing a CA for rich inflectional and derivational language such as Arabic. An efficient NLP based CA should interpret the utterance, determine and perform the actions that must be taken as response to the utterance [27]. For instance, if a user entered "I would like to buy it now". The CA should determine the meaning of the utterance that he/she wants to buy something maybe it mentioned in earlier conversation or it is out of the conversation context. Once the CA interpreted the utterance, it should know how to acts on this utterance such as continuing the purchase procedures. Finally, the CA should respond with appropriate actions that might include presenting information or accessing/writing to a database [27].

In the last few years, the Arabic based computational processing tools has gained a more concentration due to the large amount of the online Arabic available data [28]. These tools can be applied in different areas such as information retrieval, speech recognition, machine translation, localization, sentiment analysis, CAs and tutoring systems. The need of providing a user with high quality tools is important to keep up with the huge amount of Arabic data growth [29]. However, the Arabic language is still young in the NLP area in general [30].

An Arabic CA based upon NLP will face all of the above discussed challenges.The Arabic language has the flexibility of sentence structuring rather than the English. The sentence structure in Arabic has three forms [13], which are: (from right to left): [object][subject][verb] (أكل محمد التفاحة)"Mohammad eats an apple", [object][verb][subject] (محمد أكل التفاحة) "Mohammad eats an apple", and [verb][subject][object] (التفاحة أكلها محمد) "the apple has been eaten by Mohammad".In contrast, the sentence structure in English might be [subject][verb][object] ('Mohammad eats an apple'). Consequently, this flexibility of sentence structuring in Arabic will increase the complexity of building a CA based on NLP in terms of actual sentence understanding. In Arabic language, the research into computational semantics is much smaller than other areas in NLP, because of its higher complexity [13]. According to [31], "*There seems to be no agreement on an efficient way for having an adequate morphological analysis/generation*". Moreover, the NLP approach is expensive computationally. In contrast, in the English language, these challenges is much less in its complexity as the sentence's structure in English is simple as illustrated above.Another limitation that Arabic might faces is that NLP tools has been developed for western languages so it is not easy to adapt with Arabic language as it has different features [12]. In addition, the transfer of knowledge and technology as most recent publications in science and technology is from the English language.

Understanding the sentence is the most efficient way to handle conversations either in Arabic or in English. However, this computing based understanding of the sentence is still the dream for the researchers as it needs further research. Apparently, this dream is much longer for the Arabic language NLP researchers [31]. Even the Machine Translation (MT) from Arabic to English is a not easy task as it mainly based on understanding the Arabic sentence [32]. In Arabic, the sentence is long; its length is 20 to 30 words and sometimes exceeds the 100 words [32]. According to [32], "*The Arabic sentence is complex and syntactically ambiguous due to the frequent usage of grammatical relations, order of words and phrases, conjunctions, and other constructions. Consequently, most of the researches in Arabic MT mainly concentrated on the translation from English to Arabic*.". A few number of CAs based on NLP has been developed such as the GALAXY [33] which based on the English language and no similar related work has been reported for the Arabic.

Finally, when Arabic language computing based research progress to reach understanding the sentence, the NLP approach might be best way to build an Arabic CA.

2.2 Arabic Pattern Matching based Conversational Agent

Text-based Pattern Matching (PM) in computer science is the process of searching for a string or sequence of strings in a piece of text to find all occurrences of these strings inside that text. PM has been used in a large number of applications [34], such as computer parsers [35], spam filters [36], natural language processing [37], computational molecular biology [38], question answering systems [39] and CAs [3]. From a CA perspective, PM is a mechanism that uses an algorithm to handle user conversations by matching CA's patterns against a user's utterance [40]. A typical pattern consists of a collection of words, spaces, and wildcards. A wildcard is a symbol used to match a portion of the user's utterance.

PM algorithms have been successfully used in the development of a number of CAs such as InfoChat[6, 41] , ArabChat[9] and ALICE [3]. These CAs organise their scripted domain into contexts, each of which has a number of related rules. Each rule has structured patterns that represent a user's utterance. In addition, a rule has associated response to reply to the utterance. In the CA's engine that rely on the PM approach, it match the user's utterance with the scripted patterns. These patterns distributed in rules which themselves represents different topics. The matched rule will be fired and its response will send back to the user.

PM is considered one of the most successful methods for developing CAs that show or at least give the impression of some kind of intelligence [42]. The PM technique has many advantages compared with other techniques, including that it is easy to understand and it is language independent which it might be suitable for English and Arabic. In addition, PM based CAs do not require complex pre-processing stages such as "part of speech tagging" that require extra time to process and thus the PM is not expensive computationally. Consequently, CAs developed using PM can support conversations effectively for large numbers of users in a real-time environment like the Internet [42]. Moreover, it resolves many linguistic challenges that NLP technique faces, such as morphological changes occurring on a word through adding affixes to it. All of these morphological changes can be covered using pattern's wildcards. It is easy to resolve grammatical and spelling errors in a user's utterance using the pattern wildcards. Resolving spelling and grammatical errors is an important function that a CA needs to keep the conversation going, especially as research studies show that the amount of grammatical and spelling errors increases as conversations progresses [43].

The main disadvantage of a PM based CAs is that it requires a large number of patterns to implement a coherent domain. This large number of patterns comes from many issues. Firstly, there are many diverse ways that a user can construct his/her utterance (using different sentence forms or using different synonyms). Secondly, the natural language morphological richness forces a scripter to cover all expected keyword's changes to meet different cases of the scripted keyword such as singular/plural.

This disadvantage will be faced by both languages (English and Arabic). In Arabic, this challenge is bigger as it rich in flectional and derivational features as described in the previous section. The word transformation in Arabic is much large and complex than the English specially when adding infixes. Although, using the PM technique for Arabic language is extremely affected by the affixation features (which increase the pattern scripting effort), it is possible to alleviate the patterns scripting difficulties by using different techniques such as stemming [44]. Using the stemming technique as a pre-processing stage, a CA can eliminate a word from its affixes to result its root to be used in patterns scripting instead of covering many variations o f a word. However, this stemming technique might increase the computational cost but it can decrease the number of scripted patterns and thus decreasing the computational cost as well.

2.3 Sentence Similarity Measures based CA

Generally, similarity as a term is used to describe the similarity level among two objects. Similarity has been debated in number of fields such as philosophical, linguistic, and information theory [45]. Semantics is “the study of the meaning of linguistic expressions” [13]. In computing, generally, and in artificial intelligence, especially, similarity-based research has been conducted in many applications such as information retrieval, information extraction, machine translation, question answering, and conversational agents [46].In this paper, it will be focused on sentence similarity measures as an approach to build the CA.

Sentence similarity techniques try to measure the semantic similarity level between sentences. A sentence is a collection of words. Therefore, a semantic sentence similarity is related to those words semantic similarity [46].

Two main approaches have been used to measure the similarity between sentences are “Latent Semantic Analysis” (LSA) [47] and”sentence similarity based on semantic networks and corpus statistics” (SSSN) [48].

LSA analyses a large corpus of a text’s words using statistical computations and then generates a representation that captures the similarity of words and text passages [47]. A matrix of words is generated based on the number of times that a word appeared in a specific context without consideration for the order of words in a sentence [47]. Then, LSA uses the Singular Value Decomposition (SVD) technique to decompose the words matrix to reduce its size. SVD is an analysing technique that tries to reduce the dimensional representation of the matrix by keeping the entries that have the strongest relationship between words and their occurrences in sentences. However, LSA does not deal with syntactic relations (words order is not important) and with polysemy (words with multiple meanings). Consequently, this causes problems that might lead to the inability to correctly analyse the sentences. In addition, the SVD technique is expensive computationally.

The SSSN approach can overcome the LSA’s limitations [49]. SSSN is a technique based on computing the similarity through the use of lexical/semantic resources such as a WordNet [50]. WordNet is a machine-readable lexical database developed at Princeton University. WordNet has four word categories (nouns, verbs, adjectives, and adverbs). WordNet’s words are grouped into synonyms called synsets. Synsets are connected by means of semantic and lexical relations. However, Most of the published work in sentence similarity measures has been done for the English language [51]. The WordNet research work for Arabic is a young area as it has been started in 2006 with the collaboration of several companies and universities [52][31]. Where, the SSM research work for Arabic language has been started in 2012 according to [51] with creating of an Arabic benchmark dataset of only 70 word pairs. As a result, the sentence similarity area in Arabic is still young and needs further research to rely on it for CA building.

As discussed in the previous section that pattern matching suffering from the big number of needed patterns in order to build an robust CA. Sentence similarity approach came as an alternative way of building a CA to overcome PM technique challenge.

According to [53-56], applying the sentence similarity technique in CA building is more effective in terms of reducing the scripting effort to its minimum. The Sentence similarity technique replaces groups of scripted patterns into a few natural language sentences, which leads to the reduction of the scripting time and the effort of pattern maintenance [55]. For instance, table 2 [49] shows that the scripting differences in a rule handle the same topic between the PM and sentence similarity techniques. Where ‘P’ is a pattern expression, ‘S’ represents a sentence, and ‘R’ represents the rule’s response.

Table 2: shows an example of scripting differences in a rule handle the same topic between the PM and sentence similarity techniques (taken from [49]).

A Rule scripts using Pattern Matching	A rule scripts using sentence similarity
<Rule 1> P: * money P: * *cash* P: * funding . . R: I'm sorry to hear that	<Rule 1> S: I have no money R: I'm sorry to hear that

However, the measurement of similarity among sentences is an “uncertainty problem” because the real similarity process from the human judge might depend on information factors which are usually related to time and a situation.[46]. Although, capturing these hidden information factors, they are a big challenge as the current research cannot totally determine how the human brain works and how he/she capture such these information factors [46]. Moreover, no research in [49, 55, 56] have mentioned details as to if they tried to evaluate their CAs in an online environment like in terms of the elapsed time needed to handle an user's utterance.

In Arabic language, the research into computational semantics much smaller than other areas in NLP perhaps due to its higher complexity [13]. In addition, a useful database such as “WordNet” which might help the CA researcher to use it in order to build their CAs based on sentence similarity, is still a young area. Arabic WordNet (AWN) have been started in 2006 as mentioned above.

The Arabic CA based on a sentence similarity approach will face many challenges such as Arab people use three types of Arabic language (Classical, Modern and Colloquial) which might make their utterances mixed between them. Thus, increasing the need for a CA to deal with colloquial words as an example. Moreover, some Arabic words have different semantic meanings for instance, the word (عين) alone and without diacritics might have many meanings such as “eye”, “spy”, “assistance”, “a spring of water” or “ع” which is one of the Arabic letters. The word “بنك” could means “river side” or it means a “financial institution”. This variety of meaning for the same word could cause “WordNet” to misclassify such a word [31]. In addition, Arabic has neither capital letters nor acronyms in order to detect proper names such as names of people, names of months, and names of cities [31]. Moreover, some Arabic names increase the ambiguity of classifying them because they are sharing some animal, months and days names. For instance, the word “جمعه” might mean day of the week which is “Friday” and might be a male person name and the word “ليث” might means an animal name which is “Lion” or might means a male person name. The word “رمضان” might mean an Arabic month's name which is “Ramadan” or might means male person name.

3. CONCLUSION

In this paper, a brief description for Arabic language has been introduced with the comparing to the English language. This paper focused on the Arabic language complexities and its challenges in computing based work in general. In addition, this research work has determined its scope by discussing the Conversational Agent (linguistic text-based) and its building approaches. Three main approaches for CA building which are; NLP, SSM and PM. Each approach has advantages and disadvantages in building the CAs. This paper debated briefly the three approaches challenges and limitations in building the CA for the Arabic language comparing to the English language. It has been released that the Arabic language is more challengeable than the English in terms of a CA building for many general reasons. Firstly, the complexity of the Arabic language itself as it is a Semitic language with a high rich derivational and inflectional features.This complexity derived from different aspects as this paper described. Secondly, most of the research work in the NLP and SSM areas were for English language. This increased the challenge as it needs a further research for those areas before start building a CA.

ACKNOWLEDGMENT

The authors acknowledge the financial support that received from the Applied Science University in Jordan to present this work.

REFERENCES

[1] Turing, A., Computing machinery and intelligence. Mind, 1950: p. 433-460.

[2] Weizenbaum, J., ELIZA: A computer program for the study of natural language communication between man and machine. Communications of the ACM., 1966. Vol 10.: p. PP 36-45.

[3] Wallace, R.S. ALICE: Artificial Intelligence Foundation Inc. 2008 [cited; Available from: http://www.alicebot.org.

[4] ConvAgent. ADAM Conversational Agent Demo. 2001 [cited 23-02-2011]; Available from: http://www.convagent.com/convagent/adam3.aspx.

[5] Kopp, S., et al., A conversational agent as museum guide: design and evaluation of a real-world application, in Lecture Notes in Computer Science. 2005, Springer-Verlag. p. 329-343.

[6] Sammut, C. and D. Michie, InfochatTM Scripter's Manual, Convagent Ltd. 2001: Manchester.

[7] Kadous, M. and C. Sammut, InCA: A Mobile Conversational Agent. Trends in Artificial Intelligence, 2004. Volume 3157: p. pp 644-653.

[8] Alobaidi, O., et al. Abdullah: An Intelligent Arabic Conversational Tutoring System for Modern Islamic Education. in WCE 2013. 2013. London, U.K.

[9] Hijjawi, M., et al. ArabChat: An Arabic Conversational Agent. in Computer Science and Information Technology (CSIT), 2014 6th International Conference on. 2014. Amman, Jordan.

[10] UN. United Nations. 2011 [cited; Available from: http://www.un.org.

[11] Ryding, K., A Reference Grammar of Modern Standard Arabic. 2005: Cambridge university press.

[12] Ali, F. and S. Khaled, Arabic Natural Language Processing: Challenges and Solutions. 2009. 8(4): p. 1-22.

[13] Habash, N., Introduction to Arabic Natural Language Processing, ed. U.o.T. Graeme Hirst. 2010: Morgan & Claypool.

[14] Al-Saidat, E.and I. Al-Momani, Future markers in modern standard arabic and jordanian arabic: A contrastive study. European Journal of Social Sciences, 2010. 12.

[15] Al Ameed, H., et al., Arabic light stemmer: A new enhanced approach, in The Second International Conference on Innovations in Information Technology (IIT'05). 2005.

[16] Abu-Chacra, F., Arabic: An Essential Grammar. 2007.

[17] Crystal, D., Dictionary of linguistics and phonetics. Sixth Edition ed. 2008: Blackwell.

[18] Khoury, R., F.Karray, and M. Kamel, Keyword extraction rules based on a part-of-speech hierarchy. International Journal of Advanced Media and Communications, 2008. V2: p. pp 138-153.

[19] Al-Fedaghi, S. and F. Al-Anzi. A New Algorithm to Extract Arabic Root-Pattern Forms. in Proceedings of the 11th National Computer Conference. 1989. Saudi Arabia.

[20] Ahmed, A.R. and F.S. Khaled, Lexical analysis of inflected Arabic words using exhaustive search of an augmented transition network. Softw. Pract. Exper., 1993. 23(6): p. 567-588.

[21] Al-Shammari, E., Improving Arabic text processing via stemming with application to text mining and Web retrieval, in Graduate Faculty. 2010, George Mason University. p. 201.

[22] Abuleil, S. and M. Evens, Extracting an Arabic Lexicon from Arabic Newspaper Text. Computers and the Humanities, 2002. 36(2): p. 191-221.

[23] Abuleil, S., K. Alsamara, and M. Evens, Acquisition system for Arabic noun morphology, in Proceedings of the ACL-02 workshop on Computational approaches to semitic languages. 2002, Association for Computational Linguistics: Philadelphia, Pennsylvania.

[24] Diab, M., K. Hacioglu, and D. Jurafsky, Automatic tagging of Arabic text: from raw text to base phrase chunks, in Proceedings of HLT-NAACL 2004: Short Papers. 2004, Association for Computational Linguistics: Boston, Massachusetts.

[25] Fernezelyi, M., Z. Szegedy, and L. Robert, Smalltalk: interactive installation, in Proceedings of the 14th annual ACM international conference on Multimedia. 2006, ACM: Santa Barbara, CA, USA.

[26] Michie, D., Return of the Imitation Game. Linköping Electronic Articles in Computer and Information Science, 2001(Vol. 6).

[27] Lester, J., K. Branting, and B. Mott, Conversational Agents, in The Practical Handbook of Internet Computing M. Singh, Editor. 2004, Chapman and Hall/CRC. p. 1144.

[28] Abdelali, A., J. Cowie, and S. Soliman. Building A Modern Standard Arabic Corpus. in Workshop on Computational Modeling of Lexical Acquisation. 2005. Croatia.

[29] ElHadj, Y., I. Al-Sughayeir, and A. Al-Ansari. Arabic part-of-speech tagging using the sentence structure. in Proc. Second International Conference on Arabic Language Resources and Tools 2009.

[30] Shaalan, K., Rule-based Approach in Arabic Natural Language Processing. 2010.

[31] Black, W., et al., Building aWordNet for Arabic, in In Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006). 2006: Italy.

[32] Shaalan, K., Machine Translation of Arabic Interrogative Sentence into English. National Agricultural Research Information Managment System, 2000.

[33] Seneff, S., et al. (1998) GALAXY: A reference architecture for conversational system development. Volume,

[34] Sedgewick, R. and K. Wayne, Algorithms. 4th edition ed. 2011.

[35] Alessandro, W. and P. Ian, OMeta: an object-oriented language for pattern matching, in Proceedings of the 2007 symposium on Dynamic languages. 2007, ACM: Montreal, Quebec, Canada.

[36] Freschi, V., A. Seraghiti, and A. Bogliolo. Filtering Obfuscated Email Spam by means of Phonetic String Matching. in In Proceedings of ECIR. 2006.

[37] David, D.L., S. Karen, and J. rck, Natural language processing for information retrieval. Commun. ACM, 1996. 39(1): p. 92-101.

[38] Rouchka, E., Pattern Matching Techniques and Their Applications to Computational Molecular Biology - A Review, in WUCS. 1999.

[39] Hang, C., K. Min-Yen, and C. Tat-Seng, Soft pattern matching models for definitional question answering. ACM Trans. Inf. Syst., 2007. 25(2): p. 8.

[40] Pudner, K., K. Crockett, and Z. Bandar, An Intelligent Conversational Agent Approach to Extracting Queries from Natural Language, in World Congress on Engineering, International Conference of Data Mining and Knowledge Engineering. 2007. p. pp 305-310.

[41] ConvAgent. ConvAgent Foundation- ADAM Conversational Agent. 2011 [cited 20-02-2011]; Available from: www.ConvAgent.com.

[42] Timothy, B. and G. Toni, Health dialog systems for patients and consumers. J. of Biomedical Informatics, 2006. 39(5): p. 556-571.

[43] Maragoudakisa, M., et al. Natural language in dialogue systems. A case study on a medical application. in Proceedings of Panhellenic Conference with International Participation in Human–Computer Interaction. 2001. Patras, Greece.

[44] Hijjawi, M., et al., An Application of Pattern Matching Stemmer in Arabic Dialogue System, in 5th International KES Conference on Agents and Multi-agent Systems – Technologies and Applications. 2011: Manchester.

[45] Hatzivassiloglou, V., J. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. in In Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 1999.

[46] Feng, J., Y. Zhou, and T. Martin, Sentence Similarity based on Relevance, in Information Processing and Managment of Uncertainty in Knowledge-Based Systems. 2008: Spain.

[47] Landauer, T., P. Foltz, and D. Laham, Introduction to Latent Semantic Analysis. Discourrse Processes, 1998. V 25: p. pp 259-284.

[48] Li, Y., et al., Sentence similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering, 2006. V 18(8): p. 1138-1149.

[49] O'Shea, K., Z. Bandar, and K. Crockett, A Conversational Agent Framework using Semantic Analysis. International Journal of Intelligent Computing Research, 2010. V 1(1).

[50] George, A.M., WordNet: a lexical database for English. Commun. ACM, 1995. 38(11): p. 39-41.

[51] Almarsoomi, F., et al., ArabicWord Semantic Similarity. World Academy of Science, Engineering and Technology2012. Vol:6.

[52] Black, W., et al., Introducing the Arabic WordNet Project, in In Proceedings of the Third International WordNet Conference. 2006: Korea.

[53] O'Shea, J., et al., A comparative study of two short text semantic similarity measures, in Proceedings of the 2nd KES International conference on Agent and multi-agent systems: technologies and applications. 2008, Springer-Verlag: Incheon, Korea.

[54] Li, Y., et al., A Method for Measuring Sentence Similarity and its Application to Conversational Agents, in FLAIRS. 2004.

[55] Karen O'Shea, Z.B., and Keeley Crockett, A Novel Approach for Constructing Conversational Agents using Sentence Similarity Measures. 2008.

[56] O'Shea, K., Z. Bandar, and K. Crockett, Towards a New Generation of Conversational Agents Based on Sentence Similarity, in Advances in Electrical Engineering and Computational Science. 2009, Springer Netherlands. p. 505-514.

Authors

Mohammad Hijjawi is currently an assistant professor at the Applied Science University in Amman, Jordan. Dr.Hijjawi holds a Ph.D. degree in Computer Science from Manchester Metropolitan University, UK. His research interests include: Conversational Agents, machine learning, Data Mining and Arabic language processing.

Yousef Elsheikh is an assistant professor of Information Technology at the Applied Science University. He holds PhD in Informatics from University of Bradford, UK and MSc in Information Technology from University of the West of England, UK. He is currently working as a head of the Computer Information Systems department at the Applied Science University. His research interests includes conceptual modeling, ebusiness applications, information systems engineering, knowledge based representations, Ontologies and issues in software engineering.