Q&A Chatbot in Arabic Language about Prophet's Biography

Somaya Yassin Taher[1], Mohammad Zubair Khan[2]

[1,2]Department of computer science, College of computer science and Engineering, Taibah University Madinah Saudi Arabia

Abstract
Chatbots have become very popular in our times and are used in several fields. The emergence of chatbots has created a new way of communicating between human and computer interaction. A Chatbot also called a "Chatter Robot," or conversational agent CA is a software application that mimics human conversations in its natural format, which contains textual material and oral communication with artificial intelligence AI techniques. Generally, there are two types of chatbots rule-based and smart machine-based. Over the years, several chatbots designed in many languages for serving various fields such as medicine, entertainment, and education. Unfortunately, in the Arabic chatbots area, little work has been done. In this paper, we developed a beneficial tool (chatBot) in the Arabic language which contributes to educating people about the Prophet's biography providing them with useful information by using Natural Language Processing.

Keywords:
Artificial intelligence; natural language processing; Arabic chatbot

1. Introduction

In the digital era, artificial intelligence (AI) has a huge impact on our daily activities through designing advanced applications and devices, called intelligent agents, which can perform many functions. The chatbot is an AI program that played an important role in the Human-Computer Interaction (HCI) field, allowing communication between humans and computers by using natural language [1].

1-2 Chatbot definition

According to the dictionary explained a chatbot as follows: A computer program which is developed to simulate conversation with humans, typically over the Internet [2].
In the scientific literature, Chatbot was considered a conversational agent (CA), which can talk with user about a specific subject by using natural language processing (NLP). Chatbots can be classified into different categories such as e-commerce (via chat), education, food, entertainment, games, medical (Heath), Human Resource (HR), news, travel etc.

1-3 Chatbot Background

In 1950, was published an article "Computing Machinery and Intelligence" written by Alan Turing; he was wondered if a computer program could talk to a group of people without perceiving that their interlocutor was artificial. A first-ever bot ELIZA was constructed in 1966 [3]. Figure 1 shows the most popular and famous ELIZA conversation, in which ELIZA simulated a psychotherapist's operation, giving very sophisticated therapeutic advice [4]. ELIZA uses the pattern matching technique to categorize the user input as <pattern> and then provide an appropriate replay kept in < template> [5]. Its ability to communicate and its knowledge were limited, but it was a source of inspiration to develop other chatbots. In 1995, " Artificial Linguistic Internet Computer Entity" A.L.I.C.E or Alicebot was a natural language processing bot, the first online chatbot inspired by ELIZA [6]. She engages in a human conversation by applying heuristical pattern matching rules to the human's input. A.L.I.C.E was developed with a new language Artificial Intelligence Markup Language (AIML) which was created for this purpose; therefore, this the main difference between A.L.I.C.E and ELIZA [7].

Manuscript received March 5, 2024
Manuscript revised March 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.3.25

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
::
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU?
My father
YOUR FATHER
You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don't want me to notice that.
WHAT MAKES YOU THINK I AM NOT AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Figure.1 Sample ELIZA dialogue from Weizenbaum (1966)

In 2010, Siri was developed by Apple as a voice-controlled personal assistant, and Siri has been around for several years. Siri is designed to offer you a seamless way of interacting with your iPhone, iPad and so on. Users can make conversations with Siri by using voice commands, including integration with audio, video, and image files [8]. Siri makes recommendations and answers user requests using a set of internet services. It adapts to users' language usages, searches, and desires with constant use. Although Siri is considered a sophisticated and advanced system, it has some weaknesses. It requires an internet connection. It is multilingual, but it does not support many languages and requires an internet connection [9].

As shown in Figure 2, according to Scopus [8], there was a quick interest growth in the chatbots area after 2016.

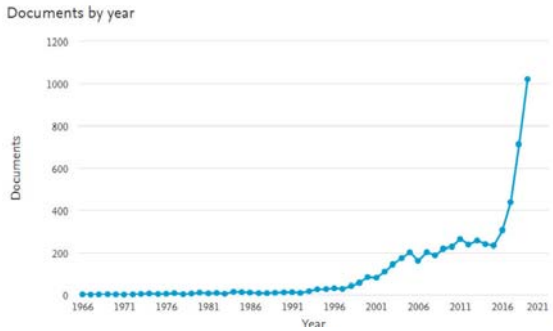


Figure.2 Search results in Scopus by year from 2000 to 2019 for "chatbot" or "conversation agent" as keywords.

1-4 Research Problem

When people want to know the answers to their questions about the Prophet's biographyالسيرة النبوية they must search in the books or website or asking other people. They consume lots of their time in the searching process. In the end, they may find their answer or not, or sometimes they got an incorrect answer.

1-5 Motivation

There is a large movement in (HCI) to streamline processes and improve methods of dealing between humans and computers. Although with the advancement of technology in our digital era, the chatbot has made a technological revolution in various fields. Accordingly, we searched about these questions:

1- What is the chatbot?
2- What is the background of the chatbot?
3- How could we Build a chatbot from scratch by using the Chabot design platform?

From this standpoint, we came up with the idea of building a chatbot that contributes to serving users by answering their questions the Prophet's biographyالسيرة النبوية easily and effectively.

1-6 Contribution

This research's main contribution is designing chatbot about the Prophet's biography in the Arabic language because this language has a great and significant position in the world. It is the language of the Holy Quran. It has deep characteristics and expressions that distinguish it from other languages, such as derivation, which mean derive multiple meaning of words from one origin such as كتاب – كتب- كاتب – كتابة – مكتوب . And also, create a dataset that contains questions and answers about our topic.

1-7 Related work

CAs have developed in several languages and fields. In this section, we will survey English and Arabic CAs. In 2020 in this paper [10], the authors proposed Artificial Intelligence Snapchat Visual Conversation Agent (AISVCA) they talked about a visual method that can be used in conversations. The method is highly driven to visual conversations such as images and graphics. AISVCA uses an artificial intelligence-driven visual conversation automation method to create received image caption and respond to any visual message in a visual conversation. This helps a lot to make the technical conversation much easier. These functionalities are achieved by using a combination of Convolutional

Neural Network (CNN) by processes the input image to extracts the image features and then encode it into a vectoral representation, Long Short-Term Memory Neural Network (LSTM) takes the vectorial representation as the initial input and sequentially predicts the next word by taking previous word into account and, Latent Semantic Indexing method (LSI). So, CNN and LSTM are used to create image captions. LSI is used to calculate and assess the semantic similarity between captions generated from personalized image dataset, and captions extracted from the received image content. To evaluate the system, they measured the proposed system's accuracy and conducted a user study to test communication quality. In the user study, they analyzed source credibility and interpersonal attraction of the AISVCA. The user study authors' results showed no significant differences in communication quality between a visual conversation with AISVCA and visual conversation with the human agent. According to the collected data from the accuracy evaluation experiment, they conclude that proposed method generates a reasonable visual response in 98.2% of cases.

In 2020 Naous, T., Hokayem, C., & Hajj, H [11], were proposed the first model for Arabic empathetic conversational bots and a dataset of conversations in Arabic. The proposed model is a sequence-to-sequence (Seq2Seq) model with LSTM units combined with Attention. The dataset is available in the Arabic language, they translated the EmpatheticDialogues dataset [12], which is the only available and include dataset in English for building empathetic chatbots. They used the Googletrans API [13] to perform the translations from English to Arabic. To assess the translation quality, they choose 100 random translated samples and compared them with the original English samples. The results indicated that only 6 of the 100 randomly chosen samples were found unreasonable while the rest were reasonable. These errors are rare in the generated conversation dataset. The translation system showed an accuracy of 96% on a sample of the data, which was deemed sufficient for model development and training conversational bots. The proposed model is evaluated using the Perplexity (PPL) automated metric and the BLEU score as an additional metric for evaluation. An embedding dimension of 500 reached state-of-the-art performance for Arabic with a PPL of 38.6 and a BLEU score of 0.5. human ratings are an important part of the overall evaluation. They collected ratings from 50 speakers of the Arabic language and rate them in terms of Empathy, Relevance, and Fluency, by answering some questions. Human evaluation of the generated responses also validated the success of the proposed model, shows performance for Arabic with average levels of Empathy and Fluency reaching of 3.7 and 3.92 respectively but the Relevance metric was at 3.16 the model did not always stay on topic while responses sometimes go off-topic.

In 2020 Al-Ghadhban, D., & Al-Twairesh, N [14] proposed Saudi dialect chatbot called Nabiha. This bot serves the student in the information technology IT department in King Saud University by conversing with bot in Saudi dialect. This chatbot's goal is to communicate with the students and answer their questions about the courses offered in the IT department or any query related to their academic progress. Nabiha was developed by using pattern matching and AIML.

2. Literature Review

In this section, the aim is to understand Arabic Language Morphology. First, we discuss the structure of the Arabic language, Arabic Morphology.

2-1 Arabic language

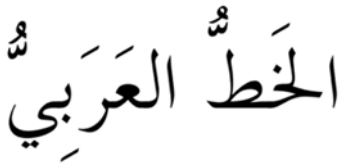
Recently the Arabic language has become the focus of many projects in NLP and computational linguistics CL [15,16]. In general, language is made up of a group of dialects and scripts.

However, there is a difference between standard Arabic and its dialects. The fact that standard Arabic is not any Arab’s native language. In fact, with the mixing of cultures and peoples from different countries, many dialects were formed, and each country or region had a special dialect such as:

- Egyptian Dialects Arabic covers the Nile valley: Egypt and Sudan.
- Levantine Dialects Arabic includes Lebanon, Syria, Jordan and Palestine
- Gulf Dialects Arabic includes the dialects of Kuwait, United Arab Emirates, Bahrain, Omani, Qatar and Saudi Arabia is typically included sub-dialects within it.

- North Dialects African covers the dialects of Morocco, Algeria, Tunisia, Mauritania and Libyan.
- Iraqi Arabic has elements of both Levantine and Gulf.

2-2 Arabic Scrips



The Arabic script is the style and system used to write the Arabic language and several other languages of Asia and Africa, such as Persian, Kashmiri, Kurdish and Urdu.
The Arabic script is primarily used to write Modern Standard Arabic (MSA).
The first script was used to write texts in Arabic, most notably the Quran, the holy book of Islam.

2-2-1 Elements of the Arabic scripts

The Arabic script is an alphabet written from right to left in a cursive style; most of these letters are written in different forms according to whether they stand alone or join a following or preceding letter [16].

In the Arabic script for writing words, there are two types of symbols: letters and diacritics. Also, we discuss digits.

2-2-1-1 Letters

Arabic letters consist of two parts: letter form (رسم rasm) and letter mark (إعجام ejAm).
The letterform is a core component in every letter and includes 19 letterforms as shown in figure 3.
The letter marks, also called consonantal diacritics, it sub-classified into three types:

- First are dots, also called points, as shown in figure 4 of which one, two or three dots above the letterform or one or two under the letterform.
- Second is the short Kaf, which is used to symbol specific shapes of the letter Kaf.
- The third is Hamza (همزة). It can appear above or below specific letterforms. The term Hamza is used for both the letter form ءand mark which appears with other letterforms such as أÂ, ؤ w, and ئ y. The Madda letter mark (مدة mad ~ah) is called a Hamza variant.

Figure 3. Letter forms are the basic graphic backbones of Arabic letters.

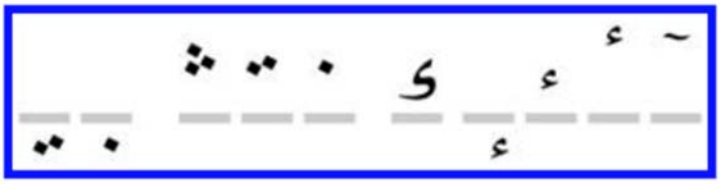


Figure 4. Letter marks are necessary to distinguish different letters.

Letter marks typically differentiate letters with different consonantal phonetic mappings but not always, as shown in figure 5.

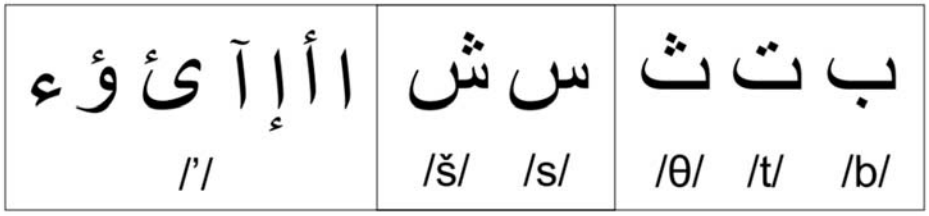


Figure 5. letter marks distinguish letters with different consonantal phonetic mappings

The Arabic letter form may be formed in different shapes according to its location globally; the figure below 6 illustrates the sample of letters with their different shapes.

`	w	r	d	A	l	k	h	T	S	s	q	f	m	γ	j	y	n	b	
ء	و	ر	د	ا	ل	ك	ه	ط	ص	س	ق	ف	م	غ	ج	ي	ن	ب	Isolated
					لـ	كـ	هـ	طـ	صـ	سـ	قـ	فـ	مـ	غـ	جـ	يـ	نـ	بـ	Initial
	ـو	ـر	ـد	ـا	ـل	ـك	ـه	ـط	ـص	ـس	ـق	ـف	ـم	ـغ	ـج	ـي	ـن	ـب	Medial
					ـل	ـك	ـه	ـط	ـص	ـس	ـق	ـف	ـم	ـغ	ـج	ـي	ـن	ـب	Final

Figure 6. A sample of letters with their different letter shapes.

2-2-1-1-1 Letter Shapes

In font and encoding architecture, distinct terminology is used: letters are characters, and symbols are shapes. Most of the letterforms are

written in a fully connected form, but a few are post-disconnected; they may connect to preceding letters but not to follow letters. All letter shapes following a post-disconnected letter form are either initial or isolated. One letter form(ء)is fully disconnected. As shown in Figure 2-4.

There are small white spaces between each disconnected letter in the word as isolated islands of connected letters. As shown in figure 7, this called word part illustrates two words and five-word parts. These spaces led to spelling errors that cause words split word parts or many words have attached without real space.

Figure 7. Arabic words are mostly connected but may contain small spaces from disconnective letters.

A word is formed when all of its letters are put together. Arabic is written from right to left when we match up the transliterations with the letters. Figure 8 letter Alif (A which is colored in green) is a disconnected letter and breaks the second word into two-word parts.

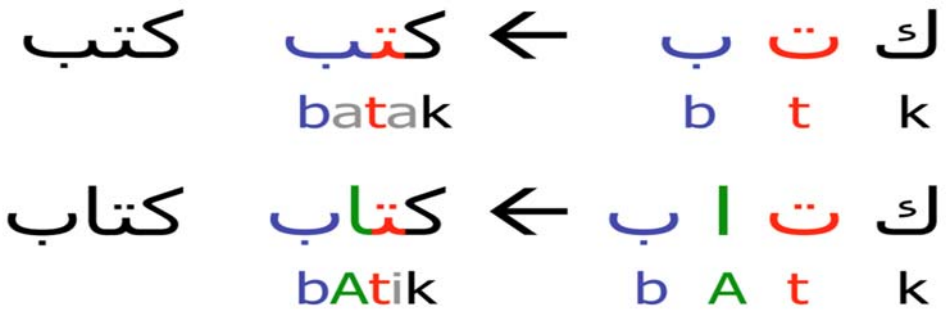


Figure 8. alif break a word into two-word parts

2-2-1-1-2 Ligatures

Arabic has a large set of common ligatures, different representations of two or even three letters. Ligatures involve vertical positioning of letters Figure 9 and vary by the font in Figure 15.

Figure 9. Example of two optional ligatures

2-2-1-1-3 Different Types of Letters

The 36 Arabic letters used in MSA can be classified into the following subsets:

- The basic Arabic's 28 letters with consonantal sounds. They are constructed using all letterforms except for the Hamza letter form. Each letter has a different shape, as shown in figure 10.

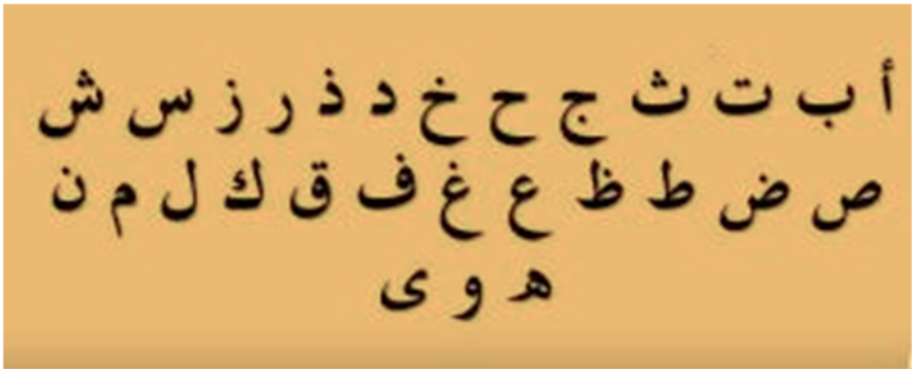


Figure 10. Arabic Letters

- The Hamza letters: There are six form of Hamza in Arabic: (ء، ئ، آ، ؤ، إ، أ). Each form has different position. One is the " Hamza-on-the-line. The rest use the Hamza and Madda letter marks with other letter forms. The Hamza letters all represent one consonant.

- A ة taa marbuta (تـاء مربوطـة) is a special version of the same letter ت. typically marking a feminine ending. It only appears in word final positions. For examples : كريمة– سعيدة.

2-2-1-2 Diacritics

The word diacritic means the marks that can appear above and below letters to alter their pronunciation. There are three types of diacritics in Arabic as shown in figure 11: Vowel, Nunation, and Shadda. It is optional to write letters with diacritics, whether fully diacritized, partially diacritized or undiacritized. Three short vowel diacritics (Fatha /a/, Damma/u/ and Kasra /i/) and the absence of any vowel (no vowel, Sukun). These three short vowels doubled, called tanween (Nunation diacritics) as shown in figure 12. tanween looks like a doubling of the short vowels and adds an "n" or ن sound to the end of the word. The Shadda which represents a doubled consonant. Figures 13 show example of fully diacritized words.

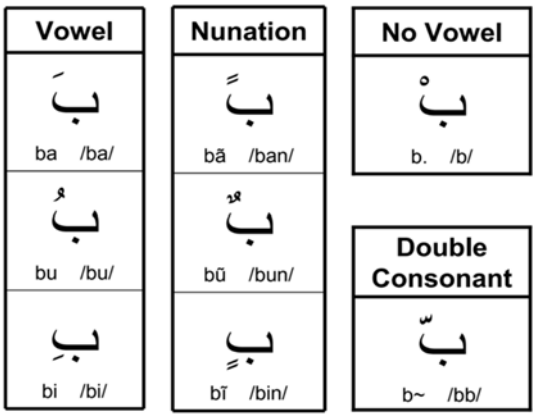


Figure 11. Types of Arabic diacritics

Nunation التَّنْوين		
[in] ◌ٍ	[an] ◌ً	[un] ◌ٌ
كِتَابٍ	كِتَابًا	كِتَابٌ a book
شَيْخٍ	شَيْخًا	شَيْخٌ an old man
شَيْءٍ	شَيْئًا	شَيْءٌ a thing
تُفَّاحَةٍ	تُفَّاحَةً	تُفَّاحَةٌ an apple
مَاءٍ	مَاءً	مَاءٌ water
خَيْرٍ	خَيْرًا	خَيْرٌ a good deed
بَيْتٍ	بَيْتًا	بَيْتٌ a house
جُزْءٍ	جُزْءًا	جُزْءٌ a section
طَبِيْبٍ	طَبِيْبًا	طَبِيْبٌ a physician

Figure 12. Example on Nunation

Figure 13. Example of fully diacritized words.

2-2-1-3 Digigts

Arabic digigts are written in a decimal system. As shown in figure 14 the three-digit sets are contrasted:

- First, the numbers used in some Arab countries such as (Morocco, Algeria, Tunisia and etc..) which called Western Arabic numerals.
- Second, the numbers used in Middle Eastern Arab countries (e.g., Egypt, Syria, Iraq, Saudi Arabia and etc..) is called Indo-Arabic numerals [5].
- Third, some non-Arab countries such as (Iran, Pakistan and etc.) have used the Eastern Indo-Arabic it similar to Indo-Arabic but with a difference only in numbers 4, 5 and 6.

	0	1	2	3	4	5	6	7	8	9
Western Arabic *Tunisia, Morocco, etc.*										
Indo-Arabic *Middle East*	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Eastern Indo-Arabic *Iran, Pakistan, etc.*	٠	١	٢	٣	۴	۵	۶	٧	٨	٩

Figure 14. Three sets of digits used in the Arabic script.

2-2-1-4 Arabic Typography

The Arabic script has a large number of fonts and styles as shown in Figure 15. some examples of Arabic script use. Most current operating systems, Windows, MacOS and Linux variants, support Arabic fonts.

Traditional Arabic	عربي	محمد	الجبر
Simplified Arabic	عربي	محمد	الجبر
Tahoma Arabic	عربي	محمد	الجبر
Andalus	عربي	محمد	الجبر
	ʕarabiy~ *Arabic*	muHam~ad *Muhammad*	Aljabr *Algebra*

Figure 15. Arabic script Fonts and styles.

2-3 Arabic Morphology

Arabic Morphology (علم الصـــرف العربي) is the sub-science of classical Arabic that studies the pattern of the word , extra letters within the template of a word and study the internal structure of the word . These meanings include tense, voice, and etc. In Figure 16 is the example of how one word gives us many meanings [17] It looks like a single word إسْـــتَنْصَـــرُوا, but it is actually a full sentence. The meaning of this word in English “They (group of males) sought help” three words communicate (1) they (2) sought (3) help. But Actually, this word has more than three, we explained that as several meanings come from this word as follow:

1) Meaning of "help" نصـــــر contains three consonants: ن [nun] – ص [saad] – ر [raa]. If we switch those three letters and put ط [Taa] – ع [‘ayn] – م [meem] which mean "food", then the meaning of the translation changes to "they sought food" اِسْتَطْعَمُوا, this called rhymes.

2) Notion of "seeking" it is coming from the س [seen] and ت [taa]. If we remove these two letters س [seen] and ت [taa] the word become نَصَرُوا "they helped", and the "seeking" meaning disappears.

3) Past tense meaning

4) Active Voice.

The active voice come from pure vowels cause what we have is اِسْتَنْصَـــرُوا (they sought help).

5) Masculine gender.

6) Plural.

Figure 16. explain the meanings of اِسْتَنْصَرُوا

7) Third Person.

Talking about them. This is masculine, plural third person. The three final meanings are coming from the "و" at the end.

This word in its different meanings showed us the beauty of the Arabic language. In English they needed three words, even then it wasn't precise. Because in English "they" is used for males and females. Whereas in Arabic اِسْتَنْصَـــــــرُوا example there is a distinction between males and females. This is what the prophet ﷺ meant when he said:

أُعْطِيتُ جَوَامِعَ الْكَلِمِ
وَ اخْتُصِرَ لِيَ الْكَلَامُ اِخْتِصارًا

I was given words of great comprehensive meaning, and speech was made concise for me

Figure 17. Hadith of the Prophet ﷺ

The majority of Arabic language meanings do not come from words. They come from vowels, patterns and grammatical structures. We distinguish two types of morphology approaches: form-based morphology and functional morphology as shown in Figure 18.

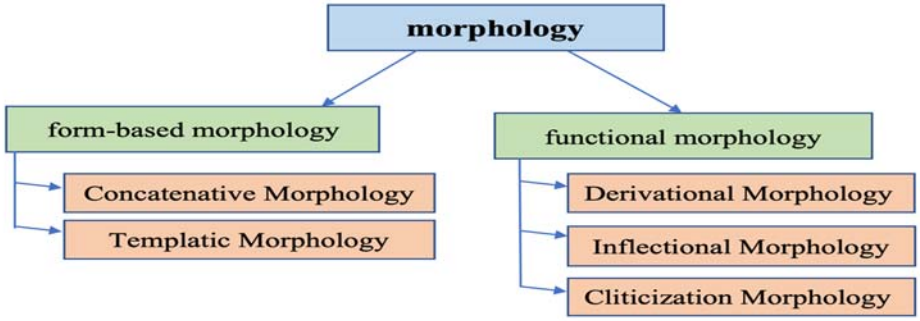


Figure 18. A chart of some related morphological terms.

2-3-1 Form-based morphology

Form-based morphology is about units that make up a word, their interactions with each other and how they relate to the word's overall form. Morpheme, the smallest meaningful unit in a language [20]. The nature of templatic morphemes in addition to concatenative morphemes is a defining characteristic in Arabic morphology. Concatenative morphemes are used in the creation of the word by a sequential method of concatenation, whereas the template morpheme is interleaved.

2-3-1-1 Concatenative Morphology

There are three types of concatenative morphemes: stems, affixes and clitics.

- Stem is a core of concatenative morphology and it is essential for every word.

- Affixes attach to the stem. There are three types of them:
  - Prefixes attach before the stem, are attached to the beginning of a root word. For example: +ن+ ‘first person plural of imperfective verbs ‘.
  - Suffixes attach after the stem, are attached to the end of a root word. For exampleون :+ +wn ‘nominative definite masculine sound plural’.
  - circumfixes surround the stem. can be considered a coordinated prefix-suffix pair.

Clitics attach to the stem after affixes. A clitic is a morpheme with a word's syntactic features but displays signs of being connected to another word phonologically. A clitic is markedly distinct from an affix in this regard and is phonologically and syntactic. Proclitics are clitics, such as the conjunction +و+w 'and' or the definite article +ال Al+ 'the', which precede the word (like a prefix).

Enclitics are clitics (like a suffix) that follow the word such as the object pronoun هم+ +hm ‘them’. In one word, multiple affixes and clitics can occur.For example, the word وسيكتبونها wasayaktubuwnahA in Figure 19 indicates that there are two proclitics, one circumfix and one enclitic.

wasayaktubuwnahA

wa+	*sa+*	*y+*	*aktub*	*+uwna*	*+hA*
and	will	3person	write	masculine-plural	it

‘and they will write it’

Figure 19. Multiple affixes and clitics can appear in a word

2-3-1-2 Templatic Morphology

Each word in the language, whether noun or verb, is a combination of base letters and a given template [21].

Templatic morphemes come in two types that are equally needed to create a word templatic stem: roots, patterns as Figure 20 An examples on templatic morphology.

عالَم, مِفْتاح, مِلْعَقَة

are instantiations of the patterns

فاعَل, مِفْعال, مِفْعَلَة

using the base letters

(ع، ل، م), (ف، ت، ح), (ع، ل، ق)

respectively. And all of the templates belong to a set called اسم الآلة (nouns of utilization).

Final Meaning	Template Set	Template	Base Meaning	Base Letters	Word
spoon (that with which you lick)	اسم الآلة	مِفْعَلَة	to lick	ل، ع، ق	مِلْعَقَة
key (that with which you open)	noun of usage	مِفْعال	to open	ف، ت، ح	مِفْتاح
world (that through which you learn (about God))		فاعَل	to know	ع، ل، م	عالَم

Figure 20. An example of templatic morphology

2-3-2 Functional Morphology

In functional morphology, as opposed to the shape of the morphemes they are formed from, we research words in terms of their morpho-syntactic and morpho-semantic behavior. Three functional operations are distinguished: derivation, inflection and cliticization. The Arabic differentiation between these three operations is close to that in other languages. The next three parts address the morphology of derivation, inflection and cliticization.

2-3-2-1 Derivational Morphology

Derivational morphology is concerned with producing new words from other words, a phase in which the word's core meaning is altered. It also teaches us how to make the different patterns of derivatives. Such as the words shown in Table 1 which are related to one root كتب [18].

Derivative	كاتِبٌ	مَكْتُوبٌ	مَكْتَبٌ	مَكْتَبَةٌ	كُتِبَ
Meaning	Writer	Written	Desk - office	Library-stationary	Was wrote
Type of derivative	اسم فاعل	اسم مفعول	اسم مكان	اسم مكان مؤنث	فعل مبني للمجهول
Meaning	Active participle	Passive participle	Noun of place	Feminine noun of place	Passive voice
Morphologic pattern	فاعلٌ	مَفْعُولٌ	مَفْعَلٌ	مَفْعَلَةٌ	فُعِلَ

Table 1. derivation words from one root

2-3-2-2 Inflectional Morphology

Inflectional affixes are those that are added to words to display grammatical function. Inflectional morphology, the word's central sense and part-of-speech POS remain unchanged and the extensions are still predictable and restricted to a collection of possible features. Inflectional operations leave the base's syntactic group unchanged, but they incorporate additional elements too [22]. There are eight inflectionary characteristics in Arabic. For verbs only, aspect, mood, person and voice apply, while case and state apply only to nouns/adjectives. Gender and number apply to both verbs and nouns/adjectives.

2-3-2-3 Cliticization Morphology

Cliticization is tightly linked to inflectional morphology. Cliticisation, analogous to inflection, does not modify the word's central meaning. However, inflectional morphology is expressed using both templatic and concatenative morphology, while cliticization is only expressed using concatenative morphology.

3 Methodology

In this section we design and implementation of the proposed Chabot.

3-1 System Architectural Design

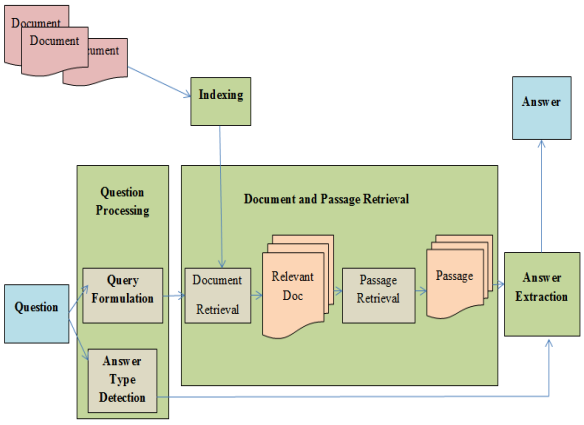


Figure 21. Question Answering (QA) Pipeline

3-1-1 Dataset Collection and Preparation

We created our own dataset by forming 65 questions and answers about the Prophet's biography by searching in the book (اللؤلؤ المكنون في سـيرة النبي المأمون)[23] which is related to the topic in order to obtain correct and accurate information. Then we arranged and placed the data in a text file.

3-1-2 Question Processing

The main goal of the question-processing phase is to extract the query, the keywords passed to the IR system to match potential documents.
we summarize the two most commonly used tasks, query formulation and answer type detection as follow:

- **Query Formulation**

Query formulation is the task of creating a query—a list of tokens— to send to an information retrieval system (chatbot) to retrieve document that might contain answer strings.

- **Answer Type Detection**

Chatbot will check the existence of the most sentence of the text which is similar to the user queries. Queries can be categorized as the following types that are illustrated in Table 2 and each type is paired to a distinct type of an answer.

Answer Type	Query (Question) Type
معلومات تخص غير العاقل	ما
معلومات تخص العاقل	من
الطريقة	كيف
الزمان	متى – كم
المكان	أين

Table 2. Q & A Type

3-1-3 Document and Passage Retrieval

Chatbot will find the most sentence of the text, which Natural Language Processing already processed, is similar to the user queries. Then,

Chatbot will get this sentence in order to return it as a response back to the user.

3-1-4 Answer Extraction

The final stage of question answering is to extract a specific answer from the passage.

3-2 System Design and Implementation

We have implemented the chatbot code using Colaboratory, or in short "Colab", is a product from Google Research. It allows anyone to write and execute python code through the browser. More technically, Colab is a free-hosted Jupyter notebook environment that runs completely in the cloud. The most important aspect of this tool does not require a setup to use. In addition, it supports writing and executing code in Python.

The codes start preparation by the following steps. Each step is implemented and written in Appendixes.

1. Installing the package NLTK. See Appendix A.
   - NLTK stands for Natural Language Tool Kit which has platform to write python programs in order to work with some data of human natural language.
2. Importing all the required libraries. See Appendix B.
3. Load the dataset and convert every text into lowercase. See Appendix C.
4. Data pre-processing. See Appendix D.
   - Data cleaning and preprocessing by tokenizing the text into a list of sentences and word.
5. Creating dictionary to remove punctuations. Creating a function to return a list of lemmatized lower-case words from the tokenized text after the punctuations are removed. See Appendix E.
6. Creating a greeting for user's input and Chatbot response as well as returning the greeting response randomly. See Appendix F.
7. Generating the response. See Appendix G.

In this step, machine learning aspect of Chatbot is taking a place as the following:

1. It takes user response (query).
2. It appends user response to the end of the sentence list of the tokenized text.
3. It does the product of term frequency (TF) and inverse document frequency (IDF).
   - Term frequency measures how frequently a term occurs in a document.
   - Inverse document frequency measures how rare a term is.
4. It converts the tokenized text to a matrix of TF-IDF.
5. It returns the index of the most similar sentence to the user response.
6. It prints the most similar sentence.
7. If there is no similar sentence to the user response, it prints an apologize message.
8. Lastly, creating a loop for the conversation between Chatbot and the user. See Appendix H.

4-1 Experimental Results

In this section, we show the result of our Chatbot program, which is called EduBot.

It produces four observable functionalities. These functionalities are mentioned as the following:

1. Response back to user's greeting.
2. Apologize to user if any misconception has occurred.
3. End the conversation with the user by displaying ending response after the user types his/her ending message like: (thank you شكرا or bye إلى اللقاء).
4. Return a precise answer to the user about his/her query.

Figure 22 below shows the result of the first function which is "Response back to user's greeting".

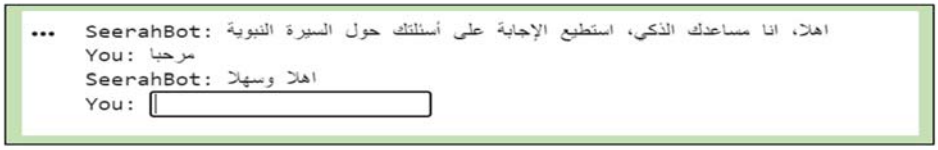


Figure 22

Figure 23 below shows the result of the second function which is "Apologize to user if any misconception has occurred".

Figure 23

Figure 24

Figure 24 below shows the result of the third function which is “End the conversation with the user by displaying ending response after the user types his/her ending message”.

Figure 25

Figure 25 below shows the result of the fourth function which is “Return a precise answer whether short or long answer to the user about his/her query”.

4-2 Conclusion and Future Work

In this paper, we have presented a Chabot, which is called (SeerahBot). We implemented the proposed system by using NLP and python as a programming language. As a result of the system, SeerahBot has an incredible ability to answer users’ queries which are written as input text. These queries are about the Prophet’s biography السيرة النبوية in the Arabic language. For Future work, we recommend that SeerahBot speech recognition and converse with users by their speech to communicate with the user by different languages and increase a dataset size.

References

1- Shawar, B.A., Atwell, E.: Different measurements metrics to evaluate a chatbot system. In: Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies, Association for Computational Linguistics, pp. 89–96 (2007)
2- Chatbot | definition of chatbot in english by Lexico Dictionaries, . (2019). Lexico Dictionaries | English website: https://www.lexico.com/en/definition/chatbot. (Retrieved16 July 2019).
3- Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM 9(1), 36–45 (1966)
4- Daniel Jurafsky, James H.(2019). Martin.Speech and Language Processing. Third Edition
5- Brandtzaeg, P. B., & Følstad, A. (2017, November). Why people use chatbots. In International Conference on Internet Science (pp. 377-392). Springer, Cham.
6- Wallace, R.S.: The anatomy of ALICE. In: Parsing the Turing Test, pp. 181–210. Springer, Dordrecht (2009)
7- https://en.wikipedia.org/wiki/Artificial_Linguistic_Internet_Computer_Entity
8- Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. Machine Learning with Applications, 2, 100006.
9- https://en.wikipedia.org/w/index.php?title=Siri&oldid=942524254
10- Arsovski, S., Cheok, A. D., Govindarajoo, K., Salehuddin, N., & Vedadi, S. (2020). Artificial intelligence snapchat: Visual conversation agent. *Applied Intelligence*, 1-10.
11- Naous, T., Hokayem, C., & Hajj, H. (2020, December). Empathy-driven Arabic Conversational Chatbot. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop* (pp. 58-68).
12- Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. L. (2018). Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*
13- https://pypi.org/project/googletrans/
14- Al-Ghadhban, D., & Al-Twairesh, N. (2020). Nabiha: An Arabic dialect chatbot. *Int. J. Adv. Comput. Sci. Appl*, *11*(3), 1-8.
15- Nizar Y. Habash, "Introduction to Arabic Natural Language Processing", Morgan & Claypool, 2010.
16- https://en.wikipedia.org/wiki/Arabic_script
17- https://www.shariahprogram.ca/comprehensiveness/
18- https://www.madinaharabic.com/arabic-language-course/lessons/L090_001.html
19- Olive, J., Christianson, C., & McCary, J. (Eds.). (2011). *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation.* Springer Science & Business Media.
20- https://en.wikipedia.org/wiki/Morpheme
21- https://www.learnarabiconline.com/arabic-morphology-introduction/
22- Shamsan, M. A. H. A., & Attayib, A. M. (2015). Inflectional morphology in Arabic and English: a contrastive study. *International Journal of English Linguistics*, *5*(2), 139.
23- العازمي، موسى. اللؤلؤ المكنون في سيرة النبي المأمون