

# A Chatbot as a Natural Web Interface to Arabic Web QA

[doi:10.3991/ijet.v6i1.1502](https://doi.org/10.3991/ijet.v6i1.1502)

Bayan Abu Shawar

Arab Open University, Amman, Jordan

**Abstract**—In this paper, we describe a way to access Arabic Web Question Answering (QA) corpus using a chatbot, without the need for sophisticated natural language processing or logical inference. Any Natural Language (NL) interface to Question Answer (QA) system is constrained to reply with the given answers, so there is no need for NL generation to recreate well-formed answers, or for deep analysis or logical inference to map user input questions onto this logical ontology; simple (but large) set of pattern-template matching rules will suffice. In previous research, this approach works properly with English and other European languages. In this paper, we try to see how the same chatbot will react in terms of Arabic Web QA corpus. Initial results shows that 93% of answers were correct, but because of a lot of characteristics related to Arabic language, changing Arabic questions into other forms may lead to no answers.

**Index Terms**—chatbot, QA, information retrieval, Question Answering System

## I. INTRODUCTION

Human computer interfaces are created to facilitate communication between human and computers in a user friendly way. For instances information retrieval systems such as Google, Yahoo, AskJeeves are used to remotely access and search a large information system based on keyword matching, and retrieving documents. However, with the tremendous amount of information available via web pages, what a user really needs is an answer to his/her request instead of documents or links to these documents. This is what a question answer system do. A question answering (QA) system accepts user's question in natural language, then retrieve an answer from its knowledge base rather than "full documents or even best-matching passages as most information retrieval systems currently do." [1]

QA systems are classified into two categories [2]: Open-domain QA; and closed-domain QA. Closed-domain question answering systems answers questions in specific domain such as medicine, education or weather forecasting. In contrast, open-domain question answering answers questions about everything and relies on general ontology and world knowledge. In recent years, "the combination of the Web growth and the explosive demand for better information access has motivated the interest in Web-based QA systems" [3].

Katz et al., [4] addressed three challenges face QA developers to provide right answers: "understanding questions, identifying where to find the information, and fetching the information itself". To understand questions and retrieve correct answers, QA systems use different NLP techniques such as: super vector machine to classify ques-

tions, and HMM based named entity recognizer to obtain right answer [5]; Others use a surface patterns to extract important terms from questions, constructs the terms' relations from sentences in the corpus, and then use these relations to filter appropriate answer candidates [6].

In contrast to English and other European languages, Arabic language suffers from shortage in NLP resources and tools. In this paper we will use an Arabic QA corpus to retrieve answers for questions without the need for using sophisticated NLP through using an interface which fools users into thinking that they speak/ask a real human; chatbot.

A chatbot is a conversational software agent, which interacts with users using natural language. The idea of chatbot systems originated in the Massachusetts Institute of Technology [7], where Weizenbaum implemented the Eliza chatbot to emulate a psychotherapist. After that, Colby developed PARRY [8] to simulate a paranoid patient. Colby [8] "regarded PARRY as a tool to study the nature of paranoia, and considered ELIZA as a potential clinical agent who could, within a time-sharing framework, autonomously handle several hundred patients an hour."

Nowadays, several chatbots are available online, and are used for different purposes such as: MIA<sup>1</sup> which is a German advisor on opening a bank account; Sanelma<sup>2</sup> a fictional female to talk with in a museum that provides information related to specific piece of art; Cybelle<sup>3</sup>, and AskJeeves<sup>4</sup>, a web-based search engine.

The remainder of this paper describes our ALICE/AIML architecture in section II. Arabic Information retrieval tools and the Arabic language characteristic are described in section III. The Arabic QA corpus that is used to retrain ALICE with the adapted program is described in sections IV and V consecutively. Results and conclusion are discussed in sections VI and VII consecutively.

## II. ALICE/AIML CHATBOT ARCHITECTURE

We have worked with the ALICE open-source chatbot initiative. ALICE5 is the Artificial Linguistic Internet Computer Entity, originated by Wallace in 1995. A sample of chatting with ALICE is shown in figure 1. ALICE stores knowledge about English conversation patterns in AIML files. AIML, or Artificial Intelligence Mark-up Language, is a derivative of Extensible Mark-up Lan-

<sup>1</sup> <http://www.aitools.org/livebots/>

<sup>2</sup> <http://www.mlab.uiah.fi/mummi/sanelma/>

<sup>3</sup> AgentLand.com

<sup>4</sup> <http://ask.co.uk/home>

<sup>5</sup> <http://www.Alicebot.org/>

guage (XML). It was developed by the Alicebot free software community during 1995-2000 to enable people to input dialogue pattern knowledge into chatbots based on the ALICE free software technology.

---

User: Hello, my name is George and I am the human.  
 ALICE: Hi George.  
 User: Would you like to play a game?  
 ALICE: Sure.  
 User: Ok then. I will email it to you by use of the CD drive.  
 ALICE: What hobbies do you enjoy?  
 User: I like cinema.  
 ALICE: What types of films?  
 User: All.  
 ALICE: What movie did you see this year that you liked?  
 User: I don't have time for movies.  
 ALICE: I thought you liked cinema.  
 User: I did, but then you started to say stupid things.

---

Figure 1. A sample of conversation with ALICE (from Loebner prize contest 2004)

AIML consists of data objects called AIML objects, which are made up of units called topics and categories as shown in figure 2. The topic is an optional top-level element, it has a name attribute and a set of categories related to that topic. Categories are the basic unit of knowledge in AIML. Each category is a rule for matching an input and converting to an output, and consists of a pattern, which represents the user input, and a template, which implies the ALICE robot answer. The AIML pattern is simple, consisting only of words, spaces, and the wildcard symbols `_` and `*`. The words may consist of letters and numerals, but no other characters. Words are separated by a single space, and the wildcard characters function like words. The pattern language is case invariant. The idea of the pattern matching technique is based on finding the best, longest, pattern match.

---

```
<aiml version="1.0">
<topic name="the topic">
<category>
<pattern>USER INPUT</pattern>
<that>THAT</that>
<template>Chatbotanswer</template>
</category>
</topic> </aiml>
```

---

Figure 2. The AIML format

### A. TYPES OF ALICE/AIML CATEGORIES

There are three types of the AIML categories: atomic categories, default categories, and recursive categories.

*Atomic categories* are those with patterns that do not have wildcard symbols, `_` and `*`, e.g.:

```
<category>
<pattern>WHAT IS 2 AND 2</pattern>
<template> It is 4 </template>
</category>
```

In the above category, if the user inputs ‘What is 2 and 2’, then ALICE answers ‘it is 4’.

*Default categories* are those with patterns having wildcard symbols `*` or `_`. The wildcard symbols match any input but they differ in their alphabetical order. Assuming the previous input WHAT IS 2 AND 2, if the robot does not find the previous category with an atomic pattern, then it will try to find a category with a default pattern such as:

```
<category>
<pattern>WHAT IS 2 *</pattern>
<template>
<random>
<li>Two.</li>
<li>Four.</li>
<li>Six.</li>
</random>
</template>
</category>
```

So ALICE will pick a random answer from the list.

*Recursive categories* are those with templates having `<srai>` and `<sr>` tags, which refer to simply recursive artificial intelligence, and symbolic reduction. Recursive categories have many applications: symbolic reduction that reduces complex grammatical forms to simpler ones; divide and conquer that splits an input into two or more subparts, and combines the responses to each; and dealing with synonyms by mapping different ways of saying the same thing to the same reply as the following example:

```
<category>
<pattern>HALO</pattern>
<template>
<srai>Hello</srai>
</template>
</category>
```

The input is mapped to another form, which has the same meaning.

### B. ALICE/AIML PATTERN MATCHING

The AIML interpreter tries to match word by word to obtain the longest pattern match, as this is normally the best one. This behavior can be described in terms of the Graphmaster as shown in figure 3. Graphmaster is a set of files and directories, which has a set of nodes called nodemappers and branches representing the first words of all patterns and wildcard symbols. Assume the user input starts with word X and the root of this tree structure is a folder of the file system that contains all patterns and templates; the pattern matching algorithm uses depth first search techniques:

- If the folder has a subfolder starting with underscore then turn to, “`_`”, scan through it to match all words suffixed X, if no match then:
- Go back to folder, try to find a subfolder starts with word X, if so turn to “`X`”, scan for matching the tail of X, if no match then:
- Go back to the folder, try to find a subfolder start with star notation, if so, turn to “`*`”, try all remaining suffixes of input following “`X`” to see if one match. If no match was found, change directory back to the parent of this folder, and put “`X`” back on the head of the input. When a match is found, the process stops, and the template that belongs to that category is processed by the interpreter to construct the output.

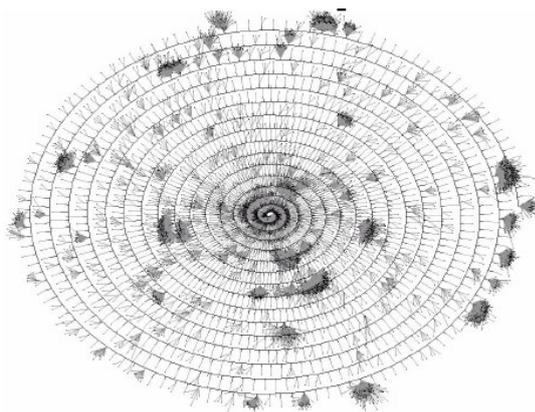


Figure 3. A Graphmaster that represents ALICE brain

III. STATE OF THE ART

A. Arabic Language Characteristic

Arabic language belongs to family of Semitic language that differs from Indo-European languages semantically, syntactically, and morphologically. Arabic is composed of 25 consonants and three long vowels that are written from right to left, and take different shapes according to its position in the word. In addition, Arabic has short vowels (diacritics) that appear in written text above or beneath alphabets and effect pronunciation and meaning of the word [9]. According to diacritics usage written Arabic text could be classified into Classical Arabic and Modern Standard Arabic (MSA) language. The Classical Arabic is vowelized, and is used in religious text that requires strict obedience to pronunciation rules such as Qur'an the holy book of Islam; some scripts of Al-Hadith (teachings of Prophet Mohammad (PBUH), classical poetries, in children literatures and in ordinary text when it is ambiguous to read [10, 11]. For example, Arabic word "علم" that is composed of three letters, can be ambiguous without vowels as shown in table 1.

MSA is the language of media, newspaper, books, magazines, and is used also as a communication media between different Arab nationalities. A third category of Arabic language is Colloquial Arabic dialects, which is the spoken language in different Arabic languages where each country has its own dialect. Colloquial Arabic dialects are used in informal settings, and between friends [10].

TABLE I.  
VOWEL DIACRITICS (FROM ELKHATEB ET AL. 2006)

Arabic	Transliteration	PoS	Meaning
علم	'alam	N	Flag
علم	'ilm	N	Science
علم	Ulima	Passive verb	Known
علم	'allama	verb	Teach
علم	'alam	N	Famous

Arabic is a derivative language where "most Arabic words are derived from a root, generally composed of three consonants; occasionally the root can be also formed of two, four or rarely five consonants" [12]. According to [11] Arabic words are classified into three categories:

- *Original Arabic words*: which include Arabic verbs and nouns that are formed according to Arabic derivation rules;
- *Fixed Arabic words*: which include words that do not belong to derivative rules, these words were modeled by Arabs in ancient times;
- *Arabized words*: which include nouns that are taken from foreign languages and become common within Arab people.

Even though, Arabic is an international language, rivaling English in number of mother-tongue speakers [13]; the progress in Arabic Natural Language Processing is slower than English and other European languages, and this is because of [14], 15]:

- Orthographic variations are prevalent in Arabic;
- Arabic has a very complex morphology;
- Arabic words are often ambiguous due to tri-literal root system;
- Synonyms are widespread, perhaps because variety in expression is appreciated as part of a good writing style;
- Broken plurals are common;
- The absence of diacritics in the written text creates ambiguity and therefore, complex morphological rules are required to identify the tokens and parse the text;
- The writing direction is from right-to-left and some of the characters change their shapes based on their location in the word;
- Capitalization is not used in Arabic, which makes it hard to identify proper names, acronyms, and abbreviations.

B. Arabic Information Retrieval

In 2002 Arabic-speaking internet users was about 4.4 million which is about 1.5% of the Arab population [16]. Within days, Arab Internet users increase, and the volume of Arabic online information also increases which necessitate the need for Arabic information retrieval systems (AIR) to facilitate accessing to this information. Abdelali et al., [12] classified Arabic IR into two categories:

- Full form based IR, which is used in commercial market as Yahoo, Google, and Ayna.
- Morphology-based IR. These systems use different NLP techniques to improve Arabic IR. These techniques involve: part of speech taggers [17, 18], using thesauri [14], using ontology [19], and using light stemmers [20, 21].

The lack of Arabic IR tools refers mainly to two reasons [10]:

- Complexity of Arabic language;
- Lack of adequate resources (corpora, morphological analyzers, lexicons, part-of- speech taggers, etc.)

Some researchers try to use what is known as Cross Language Information Retrieval (CLIR) which allows a user to insert query in his own natural language and obtain documents in one or several other languages [22]. Different approaches have been used to tackle Arabic CLIR problem as listed below [10]:

- *Machine translation and dictionary-based approach*, where input queries are translated using a dictionary into the language in which document may be found. This technique was adopted by Aljlal et al., [22], to build an Arabic-English IR. Diekema et al., [23] built an English-Arabic CLIR system that retrieves an English query and returns documents in Arabic language, at the same time those Arabic documents are automatically translated into English to facilitate reading for English analyst.
- *Transliteration/Transcription approach*, where input queries are converting the characters of an alphabetical or syllabic script to the characters of a conversion alphabet. This technique was used by AbdulJaleel and Larkey [24] to build Arabic-English IR.
- *Latent semantic indexing (LSI) approach*, where automatic statistical algorithms are used to improve the retrieval process Abdelali et al. [25]
- *Corpus-based approach*, Semmar and Fluhr [26] describes a new approach to align Arabic-French sentences retrieved from Parallel corpus based on CLIR system. In this approach a database of sentences of the target text is created, and each sentence of the source text is considered as a query to that database.

Generally speaking, IR systems return most relevant documents according to user request [27], this is insufficient in some how in this electronic age, sometimes what users really need is a specific answer instead of a set of relevant documents. The goal of a Question Answering (QA) is "to provide inexperienced users with a flexible access to the information allowing for writing a query in natural language and obtaining a precise answer" [28]. Most QA systems are developed for English as the target language, till now few QA systems has been developed in Arabic which also based on using sophisticated NLP, and machine translation such as: AQAS [29]; LMRA [30]; QARAB [15]; [28]; and [11].

In this paper we examine accessing an Arabic Web QA using a simple pattern matching technique simplified by ALICE chatbot without the need for using sophisticated NLP.

#### IV. USING WEB ARABIC QA TO RETRAIN ALICE

ALICE chatbot was originated for chatting and entertainment. In order to find other useful applications for ALICE, a Java program that converts a text corpus to the AIML chatbot language model format was developed.

The program generates two files: an atomic file and default files. The atomic file will hold the same questions and answers as appeared in the corpus where the pattern represents the questions and the template represents the answers. Since we can not guarantee that the user may enter the same questions as it is stored in the ALICE knowledge base, the default file was built using the idea of first word and the most significant word approaches.

The first word will act as question classifiers: so when differs than who, or what. The most significant word is the least frequent word in the question, which will have the highest information content? For example, when you ask "what is your name?" The least frequent word will be "name" and the answer will be generated according to this.

We modified the Java program we developed with the Qur'an, the holy book of Islam [31]. The generated chatbot accepts an Arabic question related to Islamic issues, and the answers are verses from Qur'an that match some keywords. However, because of the Qur'an nature as a monologue text, not as questions and its answers, evaluation for the Qur'an chatbot shows that most of responses were not related to the question. In this paper, we extend our FAQs chatbot systems [32] generated before in English, and Spanish to include Arabic QA.

In this term, we used different Web-pages to build a small corpus consist of 412 Arabic QA, and covers 5 domains:

- Mothers and pregnancy issues,
- Teeth care issues<sup>6</sup>,
- Fasting and related issues to health,
- Blood disease such as cholesterol, and diabetes,
- Blood charity issues<sup>7</sup>.

The questions and answers were extracted not from users' forums, but to guarantee its correctness, we gathered it from web pages like medical centers and hospitals.

Different problems raised up that is related to QA format and structural issues which necessitate some manual and automatic treatments as follows:

The questions in these sites were denoted using different symbols: stars, bullet points, numbers and sometimes with "س:" which mean "Q:". To facilitate programming issues, and unify these symbols, all questions were preceded with "Q:". Samples of those questions are presented in table 1.

Another problem was that some of these were in fact PDF files not as web pages, which required to convert it into text ones.

The answers for some questions were long and found in many lines which requires a concatenation procedure to merge these lines together.

TABLE II.  
TABLE 1: SAMPLES OF QUESTIONS OF ARABIC QUESTIONS

English translation	Arabic question
Q: Why does the wisdom tooth have this name?	س: لماذا سمي ضرس العقل بهذا الاسم؟
1) What does blood mean?	1) ماهو الدم؟
* What cloths should a pregnant wear?	* ماهي الثياب التي يفضل أن ترتديها الحامل؟

#### V. PROCESSING THE ARABIC QA

The Java program that was developed and used before to convert a readable text to the AIML format is adapted to handle the Arabic QA corpus. The program is composed of three modules as follows:

- Module 1: Generating the atomic file by reading questions and answers.

<sup>6</sup> D:\ArabicQA\_corpora\ مركز عالم الابتسامة التخصصي لطب الأسنان وزراعتها - أسئلة متكررة.htm

<sup>7</sup> D:\ArabicQA\_corpora\ اسئلة المتكررة من المتبرعين - منتدى شباب الخير.htm

- Module 2: Constructing the frequency list, and a file of all questions.
- Module 3: Generating default files.

A. Module 1: Generating Atomic file

The first program is generating the atomic file; during this program the following steps are applied:

1. Reading the questions which are denoted by "س:" ("Q:")
2. Normalizing the question by: removing punctuations, and un-necessary symbols
3. Adding the question as a pattern.
4. Reading the answer which is coming in a separate line after question mark.
5. Concatenating answer lines till the next question mark found.
6. Adding the answer as a template.

For example: if the Q/A is

What is blood? ما هو الدم؟

مادة بديعة التركيب تحتوي على خلايا بأنواع مختلفة ، فهناك الكريات البيضاء التي لها أشكال عديدة ، وهناك الكريات الحمراء التي تمنح الدم لونه ، كما توجد عناصر ضئيلة الحجم تدعى الصفائح ، وهناك عوامل عديدة تؤدي لحدوث التخثر وعوامل أخرى تعاكس الأولى .....

The AIML category will be:

```
<category>
<pattern>ما هو الدم</pattern>
<template>- مادة بديعة التركيب تحتوي على خلايا بأنواع مختلفة ، فهناك الكريات الحمراء التي تمنح الدم لونه ، كما توجد عناصر ضئيلة الحجم تدعى الصفائح ، وهناك عوامل عديدة تؤدي لحدوث التخثر وعوامل أخرى تعاكس الأولى .....
```

B. Module 2: Generating the frequency list

The frequency list created using the questions only, since the most significant words will be used within the questions. All questions denoted by <pattern> are read form the atomic file. A file of these questions is generated. After that a tokenization process is applied to have lexical and found its frequencies. As a result a frequency list is created.

C. Module 3: Generating the default file

1. Reading the questions and extracting the two most significant words (content words only) which are the least frequent words.
2. Different categories are added to extend the chance of finding answers as shown below:
  - o Build four categories using the most significant word (least 1) in four positions as patterns and the set of links it has as templates.
  - o Repeat the same process using the second-most significant word (least 2)
  - o Build four categories using the first word and the most significant words (least 1) where the most significant word is handled in four positions.

- o Build two categories using most significant 1 and most significant 2, keeping the order of position as in the original question.
- o Build a category using the first word, most significant word 1, and most significant word 2 where the template is a direct answer.

At the end of this stage, two files were generated: an atomic file and a default one. One of the default categories for the above atomic category is:

```
<category>
<pattern>* الدم</pattern>
<template>- مادة بديعة التركيب تحتوي على خلايا بأنواع مختلفة ، فهناك الكريات البيضاء التي لها أشكال عديدة ، وهناك الكريات الحمراء التي تمنح الدم لونه ، كما توجد عناصر ضئيلة الحجم تدعى الصفائح ، وهناك عوامل عديدة .....
```

VI. RESULTS AND EVALUATIONS

Before training ALICE with the generated AIML files, these files were converted into "UTF-8" code to recognize the Arabic letters. For this purpose two steps are taken:

1. All Arabic AIML file are started with: <?xml version="1.0" encoding="UTF-8"?>
2. An online tool was used (Foxe2318) to convert encoding into UTF-8.

As a result five versions of ALICE were generated to cover the five domains as shown in table 3.

TABLE III. TABLE 2. ONLINE CHATBOTS VERSIONS

Chatbot version	Link
Mother&kids issues	http://www.pandorabots.com/pandora/talk?botid=a9b5adeeae36a18a
Blood disease_issues	http://www.pandorabots.com/pandora/talk?botid=e5bab5f0ee36a350
Fasting&health_issues	http://www.pandorabots.com/pandora/talk?botid=d3a868e02e36a18d
Teeth_issues	http://www.pandorabots.com/pandora/talk?botid=ff2870b17e36a14a
Blood_charity_issues	http://www.pandorabots.com/pandora/talk?botid=a555a37e7e36a14d

Table 4 shows the number of categories generated from each WWW FAQs. In total 5,665 categories were generated.

Fifteen questions were submitted to the generated versions, 93% of answers were correct. Sample of chatting is shown in figure 4.

The same questions were submitted to Google and Ask-Jeeves, the recall was 87% for both. However, because Google and AskJeeves return documents that hold the answers, we measure how much it is easy to find the answers inside the documents, based on if correct document is the first one in the returned list, and if the answer is found at beginning. In both search engines AskJeeves and Google 50% of the answers were found in files, where users need to search again in these file to find their requests.

<sup>8</sup> http://www.firstobject.com/dn\_editor.htm

TABLE IV.  
AIML CATEGORIES GENERATED FROM ARABIC WWWFAQS

WWW FAQs	Atomic categories	Default categories
Mother&kids issues	164	2,057
Blood_disease	97	1,292
Blood_charity_issues	69	960
Health&fasting_issues	25	318
Teeth_care_issues	42	614

**Human:** ماهي أسباب السكري  
(What are the reasons of diabetes?)

**Arabic\_Q/A\_blood\_disease:**  
يحدث الداء السكري نتيجة فشل البنكرياس في إفراز ما يلزم من الأنسولين  
لخفض مستوى سكر الدم الذي يرتفع عادة بعد تناول الطعام وإجراء  
التمارين والتعرض للشمس

**Human:** ما سبب رائحة الفم؟  
(What is the reason of mouth smell?)

**Arabic\_Q/A\_Teeth\_issues:**  
سبب هضمي 10% - سبب فموي 80% - عوز بعض الفيتامينات 10%.

**Human:** ماهي الثياب التي تلبسها الحامل؟  
(What cloths should a pregnant wear?)

**Arabic\_Q/A\_mother&kids\_issues:**  
يجب أن تكون الثياب في أثناء الحمل عملية وبسيطة وفضفاضة ، لذلك  
تستبعد المشدات لعوقها الدوران الوريدي ومساعدتها في حدوث الدوالي ،  
وتجنب استعمال الأحذية ذات الكعب العالي التي تسبب إلى التوازن وتسبب  
آلام الظهر.

Figure 4. A sample of chatting with Arabic QChatbots

Unfortunately, changing the form of the Arabic question as presented in figure 5 leads to one of two cases: either no answer is generated or related but not correct answer is obtained.

In contrast, AskJeeves and Google give right answers or related ones even in case the Arabic form of question is changed. There are many reasons which may cause this as listed below:

- Arabic nouns and verbs are heavily prefixed. Nouns are usually preceded with the definite article al, and many conjunctions and prepositions are also attached as prefixes to nouns and verbs [12].
- Arabic word formation is a complex procedure that is entirely based on root-and-pattern system. A large number of words can be retrieved from one root [12].
- Information retrieval is language dependent operation, so retrieving Arabic documents implies retrieving all the variants of search terms using stemmer, morphological analysis, etc, and this is what AskJeeves and Google do.
- The Arabic chatbots does not apply any NLP techniques; all what a chatbot does is matching with the keywords which were found in the original QAcor-

pora without any modification. This was to aim to see how it works without any sophisticate NLP.

Another important reason is that the size of our corpora was small, as a result not a lot of lexical words variants are generated in the frequency list; we believe that if we increase the size of Arabic QA corpora, the possibility of having answers will increase even if the Arabic question form is changed without the need to any NLP techniques.

TABLE V.  
ANALYSIS OF ANSWERS GENERATED BY DIFFERENT ARABIC FORMAT OF THE SAME QUESTION

Question attempts	Question Form	Answer is found Y/N	Reason
Original	ماهي الثياب التي يفضل أن ترتديها الحامل؟ What cloths should a pregnant wear?	Y	It is the same question as found in corpus, so an atomic match occurs
Form1	ماهي الثياب التي تلبسها الحامل؟ What cloths should a pregnant wear?	Y	An Arabic synonym of lexical "wear" is replaced: "ترتديها" is replaced with "تلبسها". A right answer is returned because the match is generated according to the words "cloths" word "pregnant" and (الحامل, الثياب)
Form2	ماهي ثياب الحامل؟ What are the pregnant cloths?	N	This generates no answer, because the noun "cloths" is found in the corpus_ with an article "the" ("ال") so no match with the least word was found.

## VII. CONCLUSION

We managed to demonstrate that simple ALICE-style chatbot engine could be used as a tool to access the ArabicWeb QA. We did not need sophisticated natural language analysis or logical inference; a simple (but large) set of pattern-template matching rules is sufficient.

## REFERENCES

- [1] Dumais, S., Banko, M., Brill, E., Lin, J., and Ng, A. 2002. Web question answering: is more always better?. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (SIGIR 2002). Tampere, Finland, pp. 291-298.
- [2] Kangavari M., Ghandchi S., and Golpour M. A new model for question answering systems. 2008. In proceeding of world academy and science, engineering and technology volume. 2008. Pp. 536-543
- [3] Rosso P., Lyhyaoui A., Penarrubia J., Gomez M., Benajiba Y., and Raissouni N. 2005. Arabic-English question answering. In: Proceeding of Information Communication Technologies Int. Symposium, ICTIS-2005, Tetuan, Morocco, June 3-6, Pp. 36-41.
- [4] Katz B., Felshin S., Yuret D., Ibrahim A., Lin J., Marton G. 2002. Omnibase: Uniform Access to Heterogeneous Data for Question Answering. In Proceedings of the Seventh International Workshop on Applications of Natural Language to Information Systems (NLDB 2002), Pp. 1-5.

- [5] Zhang D., Lee W. A web-based question answering system. [Online]: <http://dspace.mit.edu/handle/1721.1/3693>
- [6] Cheng-Lung Sung, Cheng-Wei Lee, Hsu-Chun Yen, Wen-Lian Hsu. 2008. An alignment-based surface pattern for a question answering system. In proceeding of Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on. Las Vegas, 2008, Pp. 172-177
- [7] Weizenbaum, J. 1966. ELIZA-A computer program for the study of natural language communication between man and machine, *Communications of the ACM*, Vol. 10, No. 8, pp36-45. [doi:10.1145/365153.365168](https://doi.org/10.1145/365153.365168)
- [8] Colby, K. 1999. Human-computer conversation in a cognitive therapy program. In Wilks, Y. (eds.) *Machine conversations*. Kluwer, Boston/Dordrecht/London. Pp. 9-19.
- [9] Elkateb, S., Black, W., and Farwell, D. 2006. Arabic WordNet and the challenges of Arabic. In *Proceedings of Arabic NLP/MT London-UK 2006*.
- [10] Hammo, B. 2008. Towards enhancing retrieval effectiveness of search engines for diacritized Arabic documents. *Journal of Information Retrieval*. LLC 2008 Springer
- [11] Kanaan, G., Hammouri, A., Al-Shalabi, R., and Swalha, M. 2009. A new question answering system for the Arabic language. *American Journal of Applied sciences* 6(4): 797-805, 2009. [doi:10.3844/ajassp.2009.797.805](https://doi.org/10.3844/ajassp.2009.797.805)
- [12] Abdelalali, A., Cowie, J., and Soliman, H. 2004. Arabic Information Retrieval Perspectives. In *proceedings of: JEP-TALN 2004, Arabic Language Processing*, pp. 19-22, 2004.
- [13] Graddol, D. 1997. *The Future of English*. London: British Council.
- [14] Xu, J., Fraser, A., and Weischedel, R. 2002. Empirical studies in strategies for Arabic Retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.86-91.
- [15] Hammo B., Abu-Salem H., Lytinen S. 2002. QARAB: a question answering system to support the Arabic language. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*. 2002. Pp. 1-11
- [16] Levini, J. 2002. The Internet Minute: Languages on the Net September. [Online]: <http://sonomabusiness.com/archives/2002-09-column-levini.html> Retrieved January 5, 2004
- [17] Khoja, S. 2001. APT: Arabic part-of-speech taggers. In *Proceedings of the student Workshop at NAACL 2001*, pp. 20-25
- [18] Diab, M. 2007. Improved Arabic base phrase chunking with a new enriched POS tag set. In *Proceedings of the 5th Workshop on Important Unresolved Matters*, pp.89-96.
- [19] Abdelalali, A., Cowi, J., Farwell, D., Ogden, W, AND Helmreich, S. 2003. Cross-language information retrieval using ontology. In *proceedings of TALN 2003, Batz-sur-France*.
- [20] Larkey, L., Ballesteros, L., and Connell, M. 2002. Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In *proceedings of the 25th Annual International ACM SIGIR Conference on Research & Development in IR*, pp. 275-282.
- [21] Semmar, N., Laib, M., and Fluhr, Ch. 2006. Using stemming in morphological analyzer to improve Arabic information retrieval. In *proceedings of TALN 2006*, pp.317-327.
- [22] Aljlayli, M., Frieder, O., and Grossman, D. 2002. On Arabic-English cross-language information retrieval: A machine translation approach. In *Proceedings of the Third International Conference on Information Technology*, 2-7.
- [23] Diekema, A.R., Hannouche, J., Ingersoll, G., Oddy, R.N., and Liddy, E.D. 2005. Analyst-Focused Arabic Information Retrieval (poster). In *Proceedings of the 2005 International Conference on Intelligence Analysis*, McLean, VA, May 2-6, 2005.
- [24] AbdulJaleel, N., and Larkey, L. 2003. Statistical transliteration for English-Arabic cross language information retrieval. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, 139-146.
- [25] Abdelali, A., Cowi, J., and Soliman, H. 2006. Improving query expansion precision using latent semantic analysis: Application on Arabic retrieval. *Journies d'Etudes sur le Traitement Automatique de la Langue Arabe (JETALA)*, Rabat, Morocco.
- [26] Semmar, N., AND Fluhr, C. 2007. Arabic to French sentence alignment: Exploration of a cross-language information retrieval approach. In *Proceedings of the 5<sup>th</sup> Workshop on Important Unresolved Matters*, pp. 73-80.
- [27] Baeza, R., and Ribeiro, B. 1999. *Modern Information Retrieval*. ACM Press, NewYork, Addison-Wesley.
- [28] Rosso, P., Lyhyaoui, A., Penarrubia, J., Gomez, M., Benajiba, Y., and Raissouni, N. 2005. Arabic-English question answering. In *Proceedings of Information Communication Technologies Int. Symposium (ICTIS)*, Tetuan, Morocco, June 2005.
- [29] Mohammed, F.A., Nasser, K., and Harb, H.M. 1993. A knowledge-based Arabic Question Answering System (AQAS). In: *ACM SIGART Bulletin*, pp. 21-33
- [30] Al-Johar, B., and McGregor, J. 1997. A Logical Meaning Representation for Arabic (LMRA). In *Proceedings of the 15th National Computer Conference*, Riyadh, Saudi Arabia, 31-40.
- [31] Abu Shawar, B., Atwell, E. 2004. An Arabic chatbot giving answers from the Qur'an. In: Bel, B & Marlien, I (editors) *Proceedings of TALN04*. Vol 2, pp. 197-202 ATALA.
- [32] Abu Shawar, B., Atwell, E. and Roberts A. 2005. FAQchat as an information retrieval system. In: Zygmunt V. (ed.), *Human Language Technologies as a Challenge for Computer Science and Linguistics: Proceedings of the 2nd Language and Technology Conference*, Wydawnictwo Poznanskie, Poznan, pp. 274-278.
- [33] H. Moukdad, Lost in cyberspace: how do search engines handle Arabic queries? 2004. In: *Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science*, Winnipeg, 2004 Available at: [www.cais-acsi.ca/proceedings/2004/moukdad\\_2004.pdf](http://www.cais-acsi.ca/proceedings/2004/moukdad_2004.pdf).

AUTHOR

**Bayan Abu Shawar** is with Information Technology and Computing Department, Arab Open University, Amman, 1339, Jordan, ([b\\_shawar@aou.edu.jo](mailto:b_shawar@aou.edu.jo)).

Submitted November 1<sup>st</sup>, 2010. Published as resubmitted by the author March 1<sup>st</sup>, 2011.