
Proceedings Paper:
Hamdelsayed, MA and Atwell, ES (2016) Using Arabic Numbers (Singular, Dual, and Plurals) Patterns To Enhance Question Answering System Results. In: IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies. IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies, 20-22 Dec 2016, Khartoum, Sudan.

Using Arabic numbers (singular, dual, and Plurals) Patterns To Enhance Question Answering System Results

Mohamed Adany Hamdelsayed[1,2,3a],Eric Atwell[4b]
[1]Computer Science and Information Technology Faculty,
Sudan University of Science and Technology-SUST, [2]Blue Nile University, [3] Gabrah Academic College
[4]School of Computing, Faculty of Engineering, UNIVERSITY OF LEEDS
Leeds LS2 9JT, England
a mohdn111@sustech.edu,b E.S.Atwell@leeds.ac.uk

ABSTRACT

In the field of information retrieval, it is very difficult to answer the question entered by the user, because the search engine retrieve a ranked documents that contain any key word or phrase inside the documents, this need another extra effort to search the answer inside the documents, and there may be no answer. The alternative of search engine is a question answering system, which it retrieves the exact answer of the question in the natural language if found. A question answering system accepts the question in the natural, then many processes were done to extract the exact answer. In general a question answering system is composed of three main components: question classification module, information retrieval module and answer extraction module. A question answering system is applied in holy Quran which written and cited in Arabic language, some characteristic of the Arabic language were used to enhance the answer extraction, one of these important characteristics is numbering, singular, dual and plural. A prototype build uses special pattern used to process the number in Arabic language, which enhance the answers by adding more words and meaning. A corpus of questions and its answers from holy Quran used to test and answers the question.

Keywords: question answering system, corpus, Quran, Arabic language, natural language processing (NLP).

1. Introduction:

A question answering system is the process of accepting the question in natural language and answer it after many processing. The main goal of question answering system is to enter the question in the natural language and understand the meaning to present answers from stored in a repository of information (B Hammo, Abuleil, Lytinen, & Evens, 2004),(Loni, 2011). Different related fields of research intersect with question answering systemsuch as: Information retrieval (IR), information extraction(IE), natural language processing(NLP) (Allam & Haggag, 2012) and Artificial Intelligence (AI).

A question answering system can classified to two main domains open, and closed domains (Mohamed, Allam, & Haggag, 2012), also a question answering system applied in many research area (Adany & Atwell, 2015).

In general question answering system consists of three main components: a question classification, an information retrieval, and an answer extraction (Heba Kurdi, Sara Alkhaider

& Department, 2014). A question classification have two stages : the first stage is question processing or reformulation the question by applying many processing such as: tokenizing or splitting the sentence to words, removing diacritics and stop words, replacing some characters with others, removing some characters that committed with sentences specially in Arabic language such as the character و, and query expansion to use the key words that produced to produce a new words using: ontology (Abouenour, Bouzoubaa, & Rosso, 2008),stemmer (Hammo, B., Sleit, A., El-Haj, 2007), language resources such as: thesaurus (Bassam Hammo & El-haj, 2008), lexicon (Shaalan, 2007), dictionary (Cimiano & Bielefeld, 2010), spell and grammar checker (Shaalan, 2010) etc. The second one is question classification in which we can classify the type of question by defining its category and the type of entity depending on taxonomy to identify the relevant answer. Second part is the information retrieval in which we can use the product of the first component (key word and the generated word from the key word) to find the relevant documents that contain one or more key words or the phrase of the query and many others processing to rank and present the documents. The third component is answer extraction in which the system can search inside the documents and find the paragraph or words that can match the criteria to find the suitable answer or answers, rank and display them.

2. Related work:

We can classify this part to two main parts depending the study, because we the question answering system applied in holy Quran, and the language uses is Arabic language. For this we scan Arabic language and holy Quran in more details.

2.1. Arabic language:

2.1.1. introduction :

Arabic language is Semitic language (Hammo, B., Sleit, A., El-Haj, 2007), it ranked as fifth spoken language in the world and it has 36 phonemes (Alotaibi & Selouani, 2009), it is the main language of Islamic religion, which is used by Muslims in their daily prayers(Gravano, 2010), it has three types: classical Arabic which is used by holy Quran, MSA which is used today in schools and news papers, and colloquial Arabic dialects(Kanaan, Hammouri, Al-Shalabi, & Swalha, 2009), it is written from right to left (Ishkewy, Harb, & Farahat, 2014), it is use diacritics to remove some Arabic ambiguity(Cavalli-sforza, Soudi, & Mitamura, 2000), the order sentence words is free, inflectional and clitic language, and drop pro language(Attia, 2008).

2.1.2. Related work:

There is a tendency for the attention of the computerization of the Arabic language in recent decades, many researches were done in Arabic language computation such as:

One of the important studies in language is Arabic light stemmer (ARS) (Al-Omari & Abuata, 2014), this study explains the two ways of reduces tokens in documents which are stemming and stop word, there are two types of stemming light(prefix and suffix) and heavy (prefix, infix, and suffix). Also the study explains the difficulties of stemming in Arabic language depending on many characteristics such as: one word have many meaning, one root gives many words that have different meaning, changing the letter form special vowel, and some letters have some functions but are original in other words. Four algorithms for stemming
which are: manually constructed dictionaries, statistical stemmer, morphological analysis. They built an algorithm removes all affixes (antefixes, prefixes, suffixes and postfixes) from the

Arabic word, they use a mathematical approach that divided the number of characters of the word by 2 to extract the middle character and take the two neighboring characters which leads in general to the root, by comparing the produced word and the dictionary if it is found then stop search, else shift to right until find or not find if not found then return again to the left characters and starts again. Also the algorithm have the ability of removing or converting vowels. ARS build based on 6,225 words and evaluated with two other algorithms. However the system has a few errors from over-stemming, mis-stemming and under- stemming and need more tests.

Al-Bayan (Abdelnasser, 2014) is a question answering system for holy Quran uses holy Quran and interpretation books (Tafseer books) to find exact answer for the question, the system has a three main parts: the first used asemantic modules to retrieve the related verses from holy Quran, secondly the system applies morphological analysis and disambiguity by using classifier(Vector space machine SVM) to classify questions, extracting the three ranked answers by using tafseer book. A named entity for Quran proposed and constructed Arabic question classifier. The system evaluated by Quranic experts and its accuracy of the system is about 85%.

2.2. Holy Quran:

2.2.1. introduction :

Holy Quran is sacred book of Muslims, it contains and legislation and instruction of their life(Adany & Atwell, 2015), Allah revealed it to prophet Mohammed (peace be up on him) by angle Gabriel, Quran contains 114 chapters, each chapter have verses (not less than 3 and more than 286), these chapters divided into 30 parts, these parts consist of 320015 characters, 77439 word, and 6236 verse.

2.2.2. Related work:

An ontology semantically approach designed by (Yauri, Kadir, Azman, & Murad, 2013) for answering user query, the ontology besides the keyword matching used to enhance the the search results. The system consists of three models: Quran ontology, semantic Query, reformulation, and the extraction models. Many processes linguistics and semantically were done using Protégée Ontology Editor. A Quran ontology from Leeds University used to evaluate the system. A knowledge enhanced by using the system, however the relationship of Leeds ontology consists of 300 concepts and 350 relationship which is not enough to handle users query.

The study of (Zeroual & Lakhouaja, 2016) explains that three corpora of holy Quran proposed (The Quranic Arabic Corpus, The Boundary-Annotated Quran Corpus, Quran Corpus of Haifa) which they have some problems and they have not grammatical information which is lead to build enrichment corpus by using a semi automatic technique by using "AlKhalil Morpho Sys" , then manual processes. Some important processes were done such as: removing some symbols, and replacing some characters such as ◌ٰ in the word صلوٰت by the character "ا". Also, some words added by manually for three reasons: non analyzed words, multiple output analysis, and Words that have one output analysis. The corpus has: 1770 roots, voweled patterns for each stem and lemma, more than 100 POS tags used, and true lemma(1554 patterns). However, the corpus uses only one language.

3. Experiments:

These experiments depend on theoretical parts in the Arabic language which is noun category, in Arabic language the noun divided into three main category and sub categories(for more details return to Arabic language grammar) as in the figure 1 below:

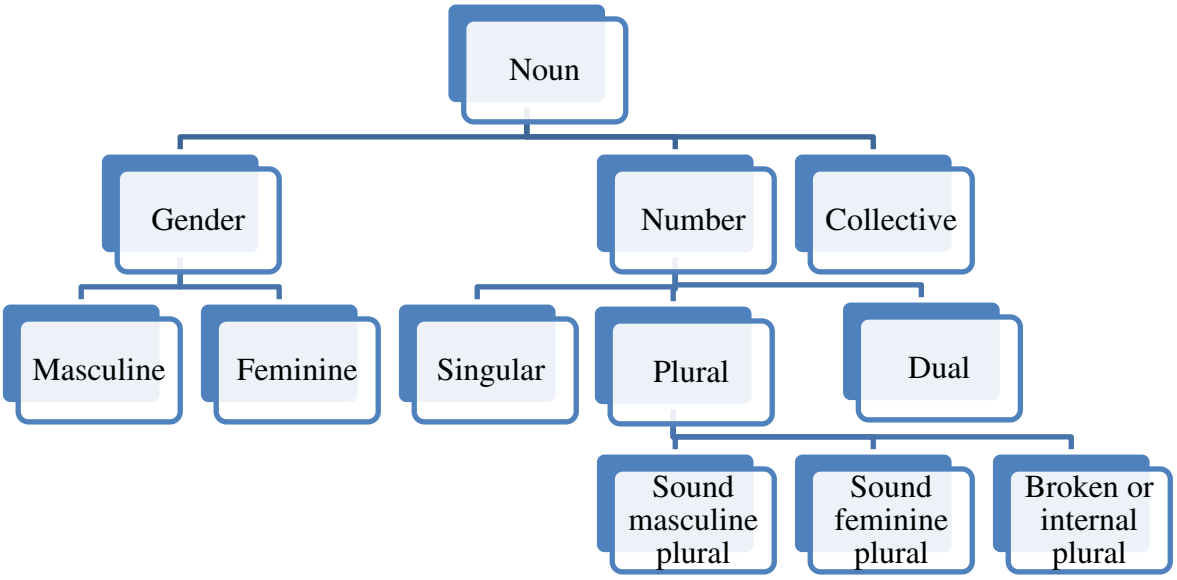

Figure 1: Noun category in Arabic language

Depending on this noun category, we built our prototype, which deals with the word, if the word in the singular then replace it by plural, or dual and vice verse. The following part explains the numbers in Arabic.

- Singular: it used to indicate that one person or thing, the Arabic word here must be abstracted from any mark for dual or plural, such as : ولد (boy), كرة (foot ball).
- Dual: it used to indicate that two persons or things, such as : ولدان (boy), كرتين (foot ball). The indicator here is : ان, ين depending on its position in the sentence.
- Plural: in the plural parts there is some complexity, because plural divided into two categories: regular plurals or sound plurals, irregular plurals (broken) and the collective. Sound plural also divided into two categories: masculine and feminine. The broken plural changes the singular weight in irregular changes without fixed rules, where fixed rules used in sound masculine plural (adding : ون , ين at the end of the singular) and sound feminine plural (adding : ات at the end of singular). The collective plural, which is called a noun plural is used as a singular to explain the plural by deleting the feminine mark (ة) or ي(called ya alnasab ياء النسب). Table [1] explains some examples to clarify this.

المفرد singular	نوعه	المثنى dual	علامته mark	الجمع	علامته mark	نوعه type	ملحوظات
معلم	مذكر	معلمان / معلمين	ان / ين	معلمون / معلمين	ون / ين	سالم	
معلمة	مؤنث	معلمتان	ان	معلمات	ات	سالم	
برتقالة	مؤنث	برتقالتان	ان	برتقالات	ات	سالم	
برتقالة	مؤنث	برتقالتان	ان	برتقال	حذف الـ ة	اسم جنس جمعي	
جندي	مذكر	جنديان	ان	جند	حذف الـ ي	اسم جنس جمعي	
ولد	مذكر	ولدان	ان	أولاد	لاعلامة	تكسير	إضافة حروف للمفرد
كتاب	مذكر	كتابان	ان	كتب	حذف الألف	تكسير	نقصان حرف من المفرد
دار	مذكر	داران	ان	دور	تغيير حرف	تكسير	تغيير في أصل الحرف
قاموس	مذكر	قاموسان	ان	قواميس	إضافة وتغيير	تكسير	إضافة وتغيير

Table 1: examples of numbers in Arabic language

3.1. The system theoretical part of the system:

Depending on the above part of the Arabic language, we use the characteristics of Arabic numbers to build our prototype. The prototype builds patterns from the words entered depending on three weights (4,6,8), they are general and used in the holy Quran frequently, then check its weight, if the weight found then applies the rule of this weigh as in the following algorithms:

1. Check the weight of the word :
   if (6 or 8) then apply the following :
   originalword = acceptedword
   if the first 2 characters of the acceptedword are ال then remove it

for I = 1 to 8
begin
finalword= the original word
newword = the original word+ ان Or newword = ال +the original word+ان;
finalword= Finalword+ newword
newword = the original word+ ات Or newword = ال +the original word+ات;
finalword= Finalword+ newword
newword = the original word+ ون Or newword = ال +the original word+ون;
finalword= Finalword+ newword
newword = the original word+ ين Or newword = ال +the original word+ين;
end;

3.2. The system practical part of the system:

A corpus build by (Adany & Atwell, 2015), used to answer the questions entered by the users. Only one question used as an example, then a comparison between two systems to judge the result of the system. The system work as follows:

1. The system first accepts the query from users as in figure 2:

Figure 2: user query من هو المؤمن

The system displays the following results as in figure[3].

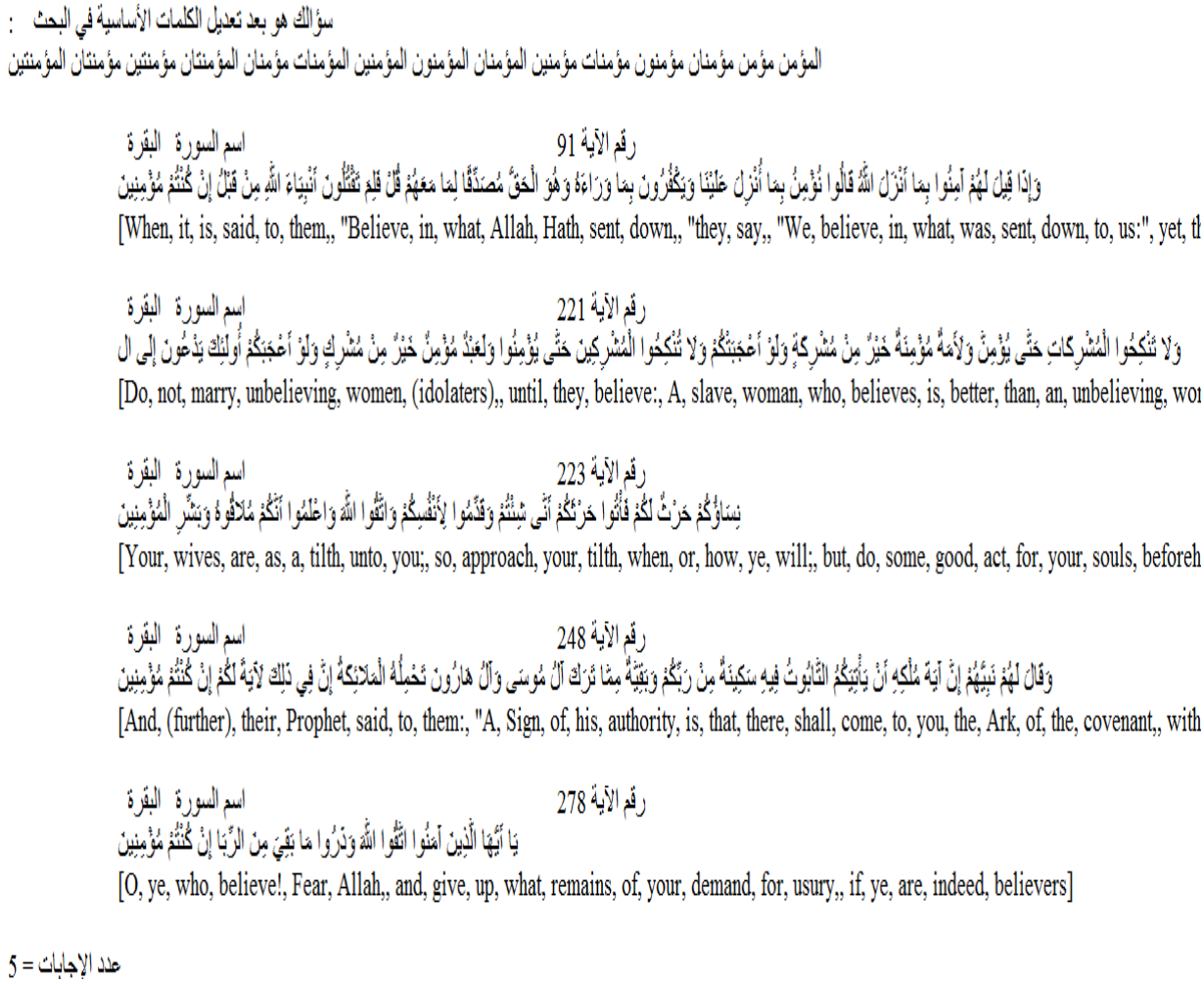

Figure 3: answer to the user query من هو المؤمن

3.3. Result discussions:

1. The system search for the query used by the user: many question were entered

(من هو مؤمن، المؤمن، المؤمنان، المؤمنون، المؤمنات)

The question asks about the word : مؤمن، المؤمن، المؤمنان، المؤمنون، المؤمنات , which is one word uses many patterns.

2. First system displays the new pattern for only one word the المؤمن generate the following patterns (new 14 words) for search as in table [2]:

ملاحظات	نوعها من حيث التذكير و التأنيث		نوعها من حيث العدد			الكلمة The word
معرف بال / نكره Definite Noun with AL / indefinite	مؤنث feminine	مذكر masculine	جمع plural	مثنى dual	مفرد singular	
نكره indefinite		√			√	مؤمن
نكره indefinite	√				√	مؤمنات
نكره indefinite	√			√		مؤمنان
نكره indefinite	√			√		مؤمنين
نكره indefinite		√	√			مؤمنون
معرف بال /Definite		√			√	المؤمن
معرف بال /Definite	√		√			المؤمنات
معرف بال /Definite		√		√		المؤمنين
معرف بال /Definite		√	√			المؤمنون
معرف بال /Definite		√		√		المؤمنان
معرف بال /Definite	√			√		المؤمنتان
نكره indefinite	√			√		مؤمنتين
نكره indefinite	√			√		مؤمنتان
معرف بال /Definite	√			√		المؤمنتين

Table 2: the result word (patterns) of the query and some notices

3. The second process is removing stop word, some symbols, and diacritics which affect the search results.
4. The system uses the generated patterns to make matching between the patterns and the corpus, if any matching then store the result to the results found before, if there is no any new answers,then display the results.

4. Comparison between two prototypes:

The following part discuss a prototype build by (Adany & Atwell, 2015), and the new prototype. First we apply the same question من هو المؤمن، المؤمنان، مؤمن، : the answer is :

Figure 4: the answer of the question من هو المؤمن

When the question is : من هم المؤمنين, the answer is من هم المؤمنين, only one answer appeared as appeared in figure [5]:

Your Question is :
المؤمنين

رقم السورة رقم الآية اسم السورة
المؤمنين

2 البقرة 223
نِسَاؤُكُمْ حَرْثٌ لَكُمْ فَأْتُوا حَرْثَكُمْ أَنَّى شِئْتُمْ وَقَدِّمُوا لِأَنْفُسِكُمْ وَاتَّقُوا اللَّهَ وَاعْلَمُوا أَنَّكُمْ مُلَاقُوهُ وَبَشِّرِ الْمُؤْمِنِينَ:الآية
عدد الإجابات = 1
BUILD SUCCESSFUL (total time: 9 seconds)

Figure 5: the answer of the question من هو المؤمنين

We notice there is only one answers appeared because it depends on the key word matching techniques, instead of 5 answers as it appeared in the proposed prototype.

5. Experiments and results:

In this part we discuss our experiments, which it judged By Scholars form Jabrah college. We use 30 question from corpus designed by (Adany & Atwell, 2015) applied in the three prototypes. The following tables and figure, explain the results.

Question Number	Prototype (1)		Prototype (2)	
	Right answer	Matching	Right answer	Matching
2	1	1	1	1
12	0	3	7	7
13	0	0	2	2
16	1	1	2	2
22	1	1	2	4
31	1	8	8	8
32	1	29	13	30
76	1	4	6	8
91	0	0	1	6
110	1	7	4	7
111	1	31	1	32
119	1	4	1	7

121	1	8	1	10
132	1	6	1	6
137	1	2	1	2
165	1	1	1	5
168	0	7	0	10
179	1	16	0	6
187	1	1	2	2
188	0	1	0	2
199	0	4	0	9
201	0	0	1	2
214	0	0	1	73
226	1	5	1	36
238	1	9	3	8
239	1	6	1	6
241	0	4	0	1
251	1	13	1	16
275	1	13	1	4
281	1	8	3	23
Summation	21	193	66	355

Table 3: General table results

From table [3] depending on the final summation we can generate the following table :

Prototype	Prototype1	Prototype2	Differences
No of questions	**30**	**30**	**0**
No of right answers	**21**	**66**	**45**
No of wrong answers	**0**	**0**	**0**
% wrong	**42.8**	**7.6**	**35.2**
% right	**47.2**	**92.4**	**45.2**
% wrong (matching)	**5.5**	**1.4**	**4.1**
% right (matching)	**10.8**	**18.5**	**7.7**

Table 4: comparative table of results

Also from table 4 we can generate the following chart as in figure []:

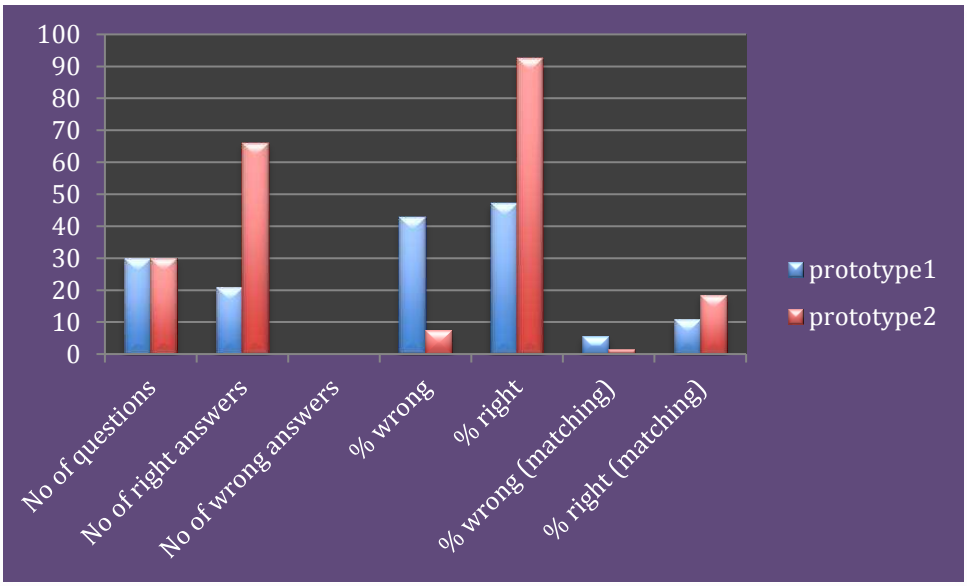

Figure 6: the chart of percentages

From all the above experiments we notice that:

1. The increase of right answer in prototype 2.
2. The decrease of wrong answers in prototype 2.

6. Conclusion:

Computation in the holy Quran and Arabic language now a day is going ahead, but need more attention from all. More efforts to fill the gap in Arabic language computation. Using patterns and learning the rules of the Arabic language in general can help more in the field of information retrieval and question answering systems.

7. Recommendation:

Applying these patterns in information retrieval, and designing more pattern can help Arabic language users. Designing more patterns also can add more values to these systems. Also, we need to make comparisons between patterns and mathematical methods to find the best

8. References:

Abdelnasser, H. (2014). Al-Bayan : An Arabic Question Answering System for the Holy Quran. *Proceedings of the 9th International Workshop on Semantic Evaluation*, 57–64.

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2008). Improving QA Using Arabic WordNet. In *the 2008 International Arab Conference on Information Technology*.

Adany, M. A. H., & Atwell, E. (2015). Islamic Applications of Automatic Question-Answering. *SUST Journal of Engineering and Computer Science (JECS)*, *1*(2), 51–57.

Allam, A., & Haggag, M. (2012). The Question Answering Systems: A Survey. *International Journal of Research and Reviews in Information Sciences, IJRRIS*, *2*(3). Retrieved from http://aliallam.com/QA Survey Paper (IJRRIS).pdf

Al-Omari, A., & Abuata, B. (2014). Arabic light stemmer (ARS). *Journal of Engineering Science and Technology*, *9*(6), 702–717.

Alotaibi, Y. A., & Selouani, S.-A. (2009). Evaluating the MSA West Point Speech Corpus. *International Journal of Computer Processing of Languages*, *22*(4), 285–304. http://doi.org/10.1142/S1793840609002111

Attia, M. (2008). Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. *Linguistics*. Retrieved from http://attiaspace.com/Publications/Attia-PhD-Thesis.pdf

Cavalli-sforza, V., Soudi, A., & Mitamura, T. (2000). Arabic Morphology Generation Using a Concatenative Strategy. *In Proceedings of NAACL*, 86–93.

Cimiano, P., & Bielefeld, U. (2010). Is Question Answering fit for the Semantic Web ?: a Survey .

Gravano, A. (2010). Turn-taking and affirmative cue words in task-oriented dialogue. *Dissertation Abstracts International, B: Sciences and Engineering*, *70*(8), 4943. http://doi.org/10.1162/COLI

Hammo, B., Sleit, A., El-Haj, M. (2007). Effectiveness of Query Expansion in Searching the Holy Quran. *Colloque Internationale Traitement Automatique de La Langue Arabe:, CITALA*, *7*, 18–19. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Effectiveness+of+Query+Expansion+in+Searching+the+Holy+Quran#0

Hammo, B., Abuleil, S., Lytinen, S., & Evens, M. (2004). Experimenting with a question answering system for the Arabic language. *Computers and the Humanities*, *38*(4), 397–415. http://doi.org/10.1007/s10579-004-1917-3

Hammo, B., & El-haj, M. (2008). Enhancing Retrieval Effectiveness of Diacritisized Arabic Passages Using Stemmer and Thesaurus. *Artificial Intelligence*, (APRIL).

Heba Kurdi, Sara Alkhaider, N. A., & Department. (2014). DEVELOPMENT AND EVALUATION OF A WEB BASED QUESTION ANSWERING SYSTEM FOR ARABIC LANGUAGE. In *Logic-based approach for improving Arabic question answering. Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on. IEEE, 2014.* (pp. 187–202).

Ishkewy, H., Harb, H., & Farahat, H. (2014). Azhary: An Arabic Lexical Ontology. *International Journal of Web & Semantic Technology*, *5*(4), 71–82. http://doi.org/10.5121/ijwest.2014.5405

Kanaan, G., Hammouri, A., Al-Shalabi, R., & Swalha, M. (2009). A New Question Answering System for the Arabic Language. *American Journal of Applied Sciences*, *6*(4), 797–805. http://doi.org/10.3844/ajas.2009.797.805

Loni, B. (2011). A survey of state-of-the-art methods on question classification. *Literature Survey, Published on TU Delft Repository*, *1999*(October). Retrieved from http://repository.tudelft.nl/assets/uuid:8e57caa8-04fc-4fe2-b668-20767ab3db92/A_Survey_of_State-of-the-Art_Methods_on_Question_Classification.pdf

Mohamed, A., Allam, N., & Haggag, M. H. (2012). The Question Answering Systems : A Survey ., *2*(3).

Shaalan, K. (2007). Person Name Entity Recognition for Arabic, (June), 17–24.

Shaalan, K. (2010). Rule-based Approach in Arabic Natural Language Processing. *International Journal on Information and Communication Technologies*, *3*(3), 11–19.

Yauri, A. R., Kadir, R. A., Azman, A., & Murad, M. A. A. (2013). Ontology semantic approach to extraction of knowledge from holy quran. *2013 5th International Conference on Computer Science and Information Technology*, 1–5. http://doi.org/10.1109/CSIT.2013.6588804

Zeroual, I., & Lakhouaja, A. (2016). A new Quranic Corpus rich in morphosyntactical information. *International Journal of Speech Technology*, pp. 1–8. http://doi.org/10.1007/s10772-016-9335-7