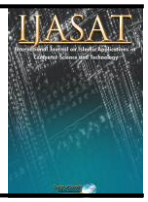

Quran Question Answering System Using Arabic Number Patterns (Singular, Dual, Plural)

Mohamed Adany Hamdels Ayed[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 (IR), it may be difficult to answer a question posed by the user, because the search engine retrieves a ranked list of documents that may contain the answer inside the documents, but this needs extra effort from the user to search for the answer inside the documents, and there may be no answer. The alternative to IR search engine is a question answering system, which retrieves the answer to the question in the natural language text if found. A question answering system accepts the question in natural language, then applies a series of processes to extract the answer. In general a question answering system is composed of three main components: question classification module, information retrieval module and answer extraction module. We developed a question answering system applied to the Holy Quran written in Classical Arabic. Some characteristics of the Arabic language were used to enhance the answer extraction: one of these important characteristics is number in nouns: singular, dual and plural. A version of the question-answering system was built which uses noun number patterns to process the number in Arabic questions and candidate answers, which enhances the result set of answers by adding more words and meaning. A corpus of questions and its answers about the Holy Quran was used to test and compare baseline and enhanced versions of our Quran Qurstion Answering system.

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

1. Introduction

A question answering system implements the process of accepting the question in natural language and answering it after processing. The main goal of a question answering system is to accept the question in natural language and understand the meaning so as to present answers drawn from a repository of information (Hammo et al, 2004), (Loni, 2011). Different related fields of research intersect with question answering systems such as: Information Retrieval (IR), Information Extraction (IE), Natural Language Processing (NLP) (Allam & Haggag, 2012) and Artificial Intelligence (AI).

A question answering system can be classified into two main types: open domain, and closed domain (Mohamed et al, 2012); also a question answering system can be applied in many research areas (Adany & Atwell, 2015).

In general question answering systems consist of three main components: question classification, information retrieval, and answer extraction (Kurdi et al, 2014). Question classification has two stages. The first stage is question processing or reformulation of the question by applying a series of processing such as tokenizing or splitting the sentence into words, removing diacritics and stop words, replacing some special characters with others, removing some problematic characters that come in the Arabic language such as the character و, and query expansion of the key words to produce a new words using ontology (Abouenour et al, 2008), stemmer (Hammo et al, 2007), language resources such as thesaurus (Hammo & El-haj, 2008), lexicon (Shaalan, 2007), dictionary (Lopez et al 2011), spelling and grammar checker (Shaalan, 2010) etc. The second stage is question classification in which we can classify the type of question by defining its category and the type of entity depending on a taxonomy to identify the relevant answer.

The second top-level component is the Information Retrieval in which we can use the product of the first component (key words and additional words generated from the key word query expansion) to find the relevant documents that contain one or more key words or the phrase of the query; this also requires 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 sentence or words that can match the criteria to find the suitable answer or answers, rank and display them.

2. Related work

There are two main areas of related work, given that the question answering system is applied to the Holy Quran, and the language this uses is Arabic. So, we review related work in Arabic language processing, and Quran computing.

2.1. Arabic language

2.1.1. Introduction

The Arabic language is a Semitic language (Hammo et al, 2007). Arabic is ranked as fifth most spoken language in the world. Arabic has 36 phonemes (Alotaibi & Selouani, 2009). It is the main language of the Islamic religion, and so is used by Muslims in their daily prayers (Gravano, 2009). There are three main types of Arabic: Classical Arabic which is used in the Holy Quran and other classical texts; Modern Standard Arabic (MSA) which is used today in formal settings including schools and newspapers; and colloquial Arabic dialects (Kanaan et al, 2009). Unlike Latin script used by many European languages, Arabic script is written from right to left (Ishkewy et al, 2014), and vowels are often not written but a writer can choose to use diacritics to remove ambiguity (Cavalli-sforza et al, 2000). The order of words in a sentence is free; Arabic is an inflectional and clitic language, and a pro-drop language (Attia, 2008).

2.1.2. Related work

There has been growing attention to the computerization of the Arabic language in recent years, and much research has been done in Arabic language computation, for example:
Arabic light stemmer (ARS): (Al-Omari & Abuata, 2014) explain the two ways of reducing tokens in documents which are stemming and stop word removal. 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 can have many meanings, one root gives many words that have different meanings, and some letters have special functions in some words but are normal in other words. Three methods for stemming are: manually constructed dictionaries, statistical stemmer, morphological analysis. (Al-Omari & Abuata, 2014) built an algorithm which removes all affixes (antefixes, prefixes, suffixes and postfixes) from the Arabic word. They used a mathematical approach that divided the number of characters of the word by 2 to extract the middle character and took the two neighbouring characters which leads in general to the root. They looked up the candidate root in the dictionary; if it is found then stop search, else shift to right and check again if a root is found. If not found then return again to the left characters and start again. Also the algorithm has the option to remove or convert vowels. ARS was built based on 6,225 words and evaluated against two other algorithms. However the system makes a few errors from over-stemming, mis-stemming and under- stemming.

Al-Bayan (Abdelnasser, 2014) is a question answering system for the Holy Quran which uses Holy Quran and interpretation books (Tafseer books) to find matching answers for the question. The system has three main parts: the first used a semantic module to retrieve the related verses from holy Quran, secondly the system applies morphological analysis and disambiguition by using a Support Vector Machine (SVM) classifier to classify questions, extracting the top three ranked answers by using a Tafseer book. The system was evaluated by Quranic experts and the accuracy of the system is about 85%.

2.2. Holy Quran:

2.2.1. introduction :

The Holy Quran is the most sacred book for Muslims, it contains advice and legislation and instruction for their life (Adany & Atwell, 2015). Mulsims believe that Allah revealed the verse of the Quran to prophet Mohammed (peace be up on him) through the angel Gabriel. The Quran contains 114 chapters; each chapter has between 3 and 286 verses; the chapters are grouped into 30 parts. The Quran consists of 320015 characters, 77439 words (depending on tokenization metod used), 6236 verses, 114 chapters, or 30 parts.

2.2.2. Related work:

An ontology-based semantical approach was designed by (Yauri et al, 2013). For answering a user query, the ontology is used in alongside keyword matching to enhance the search results. The system includes three models: Quran ontology, semantic query reformulation, and the extraction model. Linguistic and semantic processing was done using Protégée Ontology Editor. A Quran ontology from Leeds University was used to evaluate the system. In principle, an ontology should enhance the system; however the Leeds ontology consists of only 300 concepts and 350 relationships which was not enough to handle all user queries.

The study of (Zeroual & Lakhouaja, 2016) re-used three corpora of the Holy Quran: the Quranic Arabic Corpus, the Boundary-Annotated Quran Corpus, and the Quran Corpus of Haifa. They found a lack of detailed grammatical information, so built an enriched corpus by using a semi-automatic technique "AlKhalil Morpho Sys" , then manual proofreading. Some further post-editing was done including: removing some symbols, and replacing some characters such as ٰin the word صلٰوت by the character "ا". Also, some words were added manually for three reasons: non-analyzed words, multiple analyses, and words that have wrong output analysis. The corpus has 1770 roots, voweled patterns for each stem and lemma, more than 100 POS tags used, and lemmas (1554 patterns). However, the corpus uses only one language, Classical Arabic, and has no linked translations to English or other languages.

3. Experiments

These experiments depend on theoretical linguistic features of the Arabic language specifically concerning nouns. In the Arabic language the noun class is divided into sub-categories as in the figure 1 below:

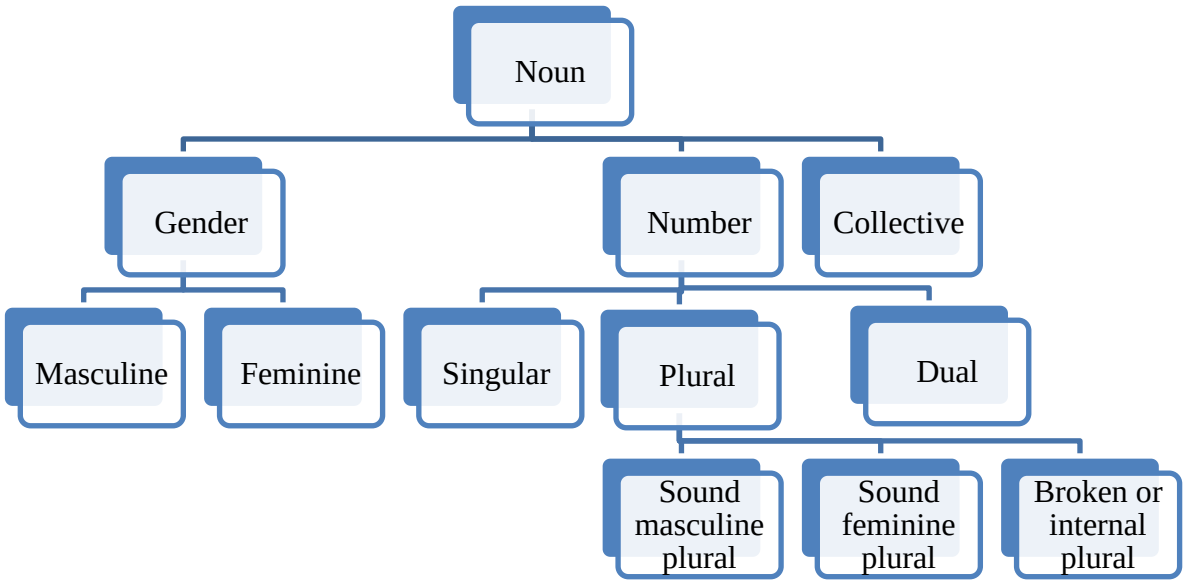

Figure 1: Noun category in Arabic language

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

- Singular is used to indicate 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 is used to indicate 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 is divided into sub-categories: regular plurals or sound plurals, irregular plurals or broken plurals, and the collective. Sound plural are also divided into two sub-categories: masculine and

feminine. The broken plural changes the singular weight in irregular changes without fixed rules, where fixed rules are used in sound masculine plural (adding: ون, ين at the end of the singular) and sound feminine plural (adding : ات at the end of the 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] gives some examples to clarify this.

Table 1: examples of noun number features in the Arabic language

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

3.1. The theoretical model

Depending on the above outline description of the Arabic language noun system, we use the characteristics of Arabic noun number to extend our Question-Answering system. The extension builds patterns from the words entered depending on three weights (4,6,8); these are general and used in the Holy Quran frequently. Then check its weight: if the weight is found then apply 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 practical implementation of the system:

A corpus of Quran questions and answers built by (Adany & Atwell, 2015) was used to answer the questions entered by the users to test the system. Only one question here is shown as an example. Finally we carried out a comparison between original and extended systems.

The system works as follows:

1. The system first accepts the query من هو المؤمن "Who is the believer?" from the user as in figure 2:

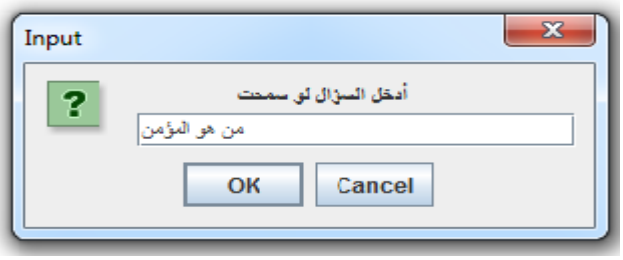

Figure 2: user query من هو المؤمن

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

سؤالك هو بعد تعديل الكلمات الأساسية في البحث :
المؤمن مؤمن مؤمنان مؤمنون مؤمنات مؤمنين المؤمنان المؤمنون المؤمنين المؤمنات مؤمنان المؤمنتان مؤمنتين مؤمنتان المؤمنتين

اسم السورة البقرة رقم الآية 91
وَإِذَا قِيلَ لَهُمْ آمِنُوا بِمَا أَنْزَلَ اللَّهُ قَالُوا نُؤْمِنُ بِمَا أُنْزِلَ عَلَيْنَا وَيَكْفُرُونَ بِمَا وَرَاءَهُ وَهُوَ الْحَقُّ مُصَدِّقًا لِمَا مَعَهُمْ قُلْ فَلِمَ تَقْتُلُونَ أَنْبِيَاءَ اللَّهِ مِنْ قَبْلُ إِنْ كُنْتُمْ مُؤْمِنِينَ
[When, it, is, said, to, them,, "Believe, in, what, Allah, Hath, sent, down,, "they, say,, "We, believe, in, what, was, sent, down, to, us:", yet, th

اسم السورة البقرة رقم الآية 221
وَلا تَنْكِحُوا الْمُشْرِكَاتِ حَتَّى يُؤْمِنَّ وَلَأَمَةٌ مُؤْمِنَةٌ خَيْرٌ مِنْ مُشْرِكَةٍ وَلَوْ أَعْجَبَتْكُمْ وَلا تُنْكِحُوا الْمُشْرِكِينَ حَتَّى يُؤْمِنُوا وَلَعَبْدٌ مُؤْمِنٌ خَيْرٌ مِنْ مُشْرِكٍ وَلَوْ أَعْجَبَكُمْ أُولَئِكَ يَدْعُونَ إِلَى ال
[Do, not, marry, unbelieving, women, (idolaters),, until, they, believe:, A, slave, woman, who, believes, is, better, than, an, unbelieving, wor

اسم السورة البقرة رقم الآية 223
نِسَاؤُكُمْ حَرْثٌ لَكُمْ فَأْتُوا حَرْثَكُمْ أَنَّى شِئْتُمْ وَقَدِّمُوا لِأَنْفُسِكُمْ وَاتَّقُوا اللَّهَ وَاعْلَمُوا أَنَّكُمْ مُلَاقُوهُ وَبَشِّرِ الْمُؤْمِنِينَ
[Your, wives, are, as, a, tilth, unto, you;, so, approach, your, tilth, when, or, how, ye, will;, but, do, some, good, act, for, your, souls, beforeh

اسم السورة البقرة رقم الآية 248
وَقَالَ لَهُمْ نَبِيُّهُمْ إِنَّ آيَةَ مُلْكِهِ أَنْ يَأْتِيَكُمُ التَّابُوتُ فِيهِ سَكِينَةٌ مِنْ رَبِّكُمْ وَبَقِيَّةٌ مِمَّا تَرَكَ آلُ مُوسَى وَآلُ هَارُونَ تَحْمِلُهُ الْمَلَائِكَةُ إِنَّ فِي ذَلِكَ لَآيَةً لَكُمْ إِنْ كُنْتُمْ مُؤْمِنِينَ
[And, (further), their, Prophet, said, to, them:, "A, Sign, of, his, authority, is, that, there, shall, come, to, you, the, Ark, of, the, covenant,, with

اسم السورة البقرة رقم الآية 278
يَا أَيُّهَا الَّذِينَ آمَنُوا اتَّقُوا اللَّهَ وَذَرُوا مَا بَقِيَ مِنَ الرِّبَا إِنْ كُنْتُمْ مُؤْمِنِينَ
[O, ye, who, believe!, Fear, Allah,, and, give, up, what, remains, of, your, demand, for, usury,, if, ye, are, indeed, believers]

عدد الإجابات = 5

Figure 3: answer to the user query من هو المؤمن

3.3. Example question and answers

1. The system searches for the query used by the user
2. The query expansion generates the expanded query: مؤمن، المؤمن، المؤمنان، المؤمنون، المؤمنات , “Who is the believer, the believer, the believers, the believers, the believers” which is one word that uses many patterns.
3. First the system displays the new pattern for only one word the المؤمن generate the following patterns (new 14 words) for search as in table [2]:

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

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

4. The second process is removing stop words, some symbols, and diacritics which affect the search results.
5. The system uses the generated patterns to make a match between the patterns and the corpus. If any match is found then store the results; if there is not any new answers then display the results.

4. Comparison between baseline and extended systems:

The following section compares a baseline Quran QA system built by (Adany & Atwell, 2015), and the new extension. First we apply the same question; 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]:

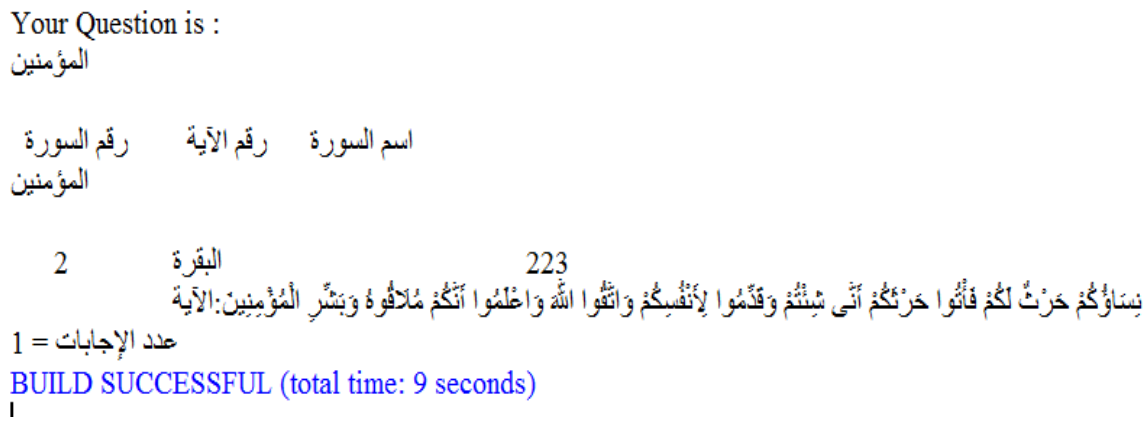

Figure 5: the answer of the question من هو المؤمنين

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

5. Experiments and results:

In this part we discuss our experiments and results, which were judged by Islamic scholars from Jabrah college. We used 30 questions from our Quran question answering corpus designed by (Adany & Atwell, 2015) applied in the systems. The following tables and figure explain the results.

Table 3: General table results

Question Number	QA (1)		QA (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

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

Table 4: comparative table of results

QA	QA1	QA2	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**

Also from table 4 we can generate the following chart as in figure [6]:

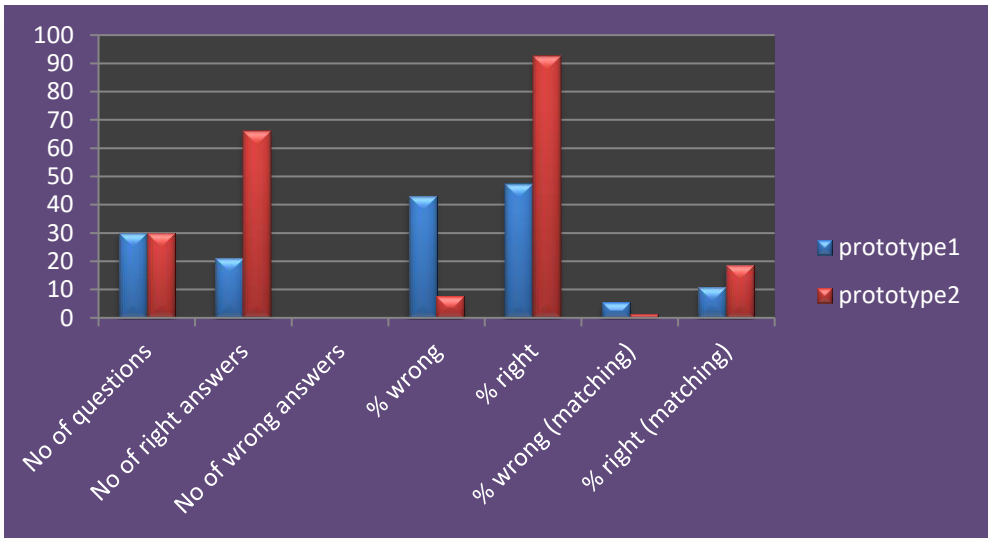

Figure 6: the chart of percentages

From all the above experiments we notice that:

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

6. Conclusion

Computation on the Holy Quran and Arabic language processing is making progress, but needs more attention from researchers. 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. Applying these patterns in information retrieval, and designing more patterns can help Arabic language users. Designing more patterns can add more value to these systems. Also, we need to make comparisons between patterns and mathematical methods to find the best combination of methods.

7. 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. PhD Thesis, University of Manchester. 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.

Gravano, A. (2009). Turn-taking and affirmative cue words in task-oriented dialogue. PhD Thesis, Columbia University. http://www.cs.columbia.edu/speech/ThesisFiles/agustin_gravano.pdf

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. 19th Midwest Artificial Intelligence And Cognitive Science Conference.

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

Lopez, V. et al. (2011). Is Question Answering fit for the Semantic Web ?: a Survey. Semantic Web Journal, 2:2 pp.125-155.

Mohamed, A., Allam, N., & Haggag, M. H. (2012). The Question Answering Systems : A Survey. International Journal of Research and Reviews in Information Sciences (IJRRIS) Vol. 2, No. 3,

Shaalan, K. (2007). Person Name Entity Recognition for Arabic, Proceedings of 2007 Workshop On Computational Approaches To Semitic Languages pp. 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