Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 6, Issue, 07, pp.7608-7613, July, 2014

INTERNATIONAL JOURNAL
OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

# BUILDING A SYSTEM BASED ON NATURAL LANGUAGES PROCESSING TO AUTOMATIC QUESTION GENERATION FROM ARABIC TEXTS

MohamedElbasyouni, Elsaeed Abdelrazek and *AbeerSaad

Computer Teacher Preparation Department, Faculty of Specific Education, Damietta University, Egypt

ARTICLE INFO

*Article History:*
Received 07th April, 2014
Received in revised form
23rd May, 2014
Accepted 15th June, 2014
Published online 20th July, 2014

*Key words:*

Arabic natural language processing (ANLP), Arabic morphology, Question Generation (QG), Arabic Part-Of-Speech (POS), Named Entity Recognition (NER).

ABSTRACT

This paper aims to describe the called Arabic Questions Generation (AQG) system. It automates the process of generating questions fill in the blanks from Arabic texts, based on pre-generated corpus pattern. The system as well uses Stanford Arabic Natural Languages Processing (NLP) tools to generate a morphological tagged tree from the Arabic text, which then be matched with the patterns to form the question. The system provides the teacher with a Graphical user interface to facilitate the process of generating a test from the set of Questions, as it has many complex problems and issues in the Natural Languages Processing. The system useful for teachers for the generating questions automatically Instead of the manual method. It used to assessing learners' achievements of learning. Grades were calculated based only on generated questions from The System Out of all grading criteria, The System less in the syntactic correctness of generated questions. The System did the best in Relevance.

## INTRODUCTION

Natural language is a very important and ubiquitous part of human intelligence and society Natural Language Processing (NLP) (Abney 1991). NLPis an artificial intelligence branch which has the ultimate goal to invent theories, so NLP gives computers the ability to understand the way humans learn and use language is the most challenge inherent in natural language processing (Abd-el-Kader and Souilem Boumiza 2009). NLP is the automated approach to analyse text that is based on a set of theories and a set of technologies together (Ramesh and Marcus 1995). The NLP tools and techniques parse linguistic input (word, sentence, text,) according to the rules of the language (like lexicon, corpus, and dictionary) (Moath and Abd-el-Kader 2014). The Arabic alphabet counts twenty-eight letters (or 29 if we add the "hamza"). There is no difference between the handwritten letters and the printed letters; On the other hand, the letters have, a different shape depending on their position in words: isolated, in the beginning, in the middle or in the end (Miami and Etzioni 2003). Arabic language is considered to be a Semitic language with richness in morphology (Farghaly and Shaalan 2009). Over the last few years, Arabic natural language processing (ANLP) has gained increasing importance; these applications had to deal with several complex problems pertinent to the nature and structure of the Arabic language (Othman and Shaalan 2003). These issues in the ANLP in Classical Arabic (Verb, Subject, and Object) form is used, there are forms in dialects like SVO (Subject, Verb, Object this forms used in ANLP, and change the sentences structure and sometimes the meaning of the sentence (Chen and Gey 2003).

Arabic language morphology is much richer than English (Ibrahim 2003). Arabic language is considered as the most category complex natural language (Ditters 2001). Understanding Arabic requires the treatment of the language constituents at all levels morphology, syntax, and semantics (Lauert and Graessera 1992). Arabic morphology and syntax provide the ability to add a large number of affixes to each word which makes combinatorial increment of possible words (Graesser and Van Lehn 2001). Question Generation (QG) is important components in advanced learning technologies such as intelligent tutoring systems. QG is an essential element of learning environments, help systems, information seeking systems, and other applications (Chen and Aist 2009). Another benefit of QG is that it can be a good tool to help improve the quality of the Question Answering (QA) systems (Graesser and Chipman 2005). QG aims at generating questions from text and has become a vibrant line of research. Generating questions. It is useful for knowledge assessment-related tasks, by reducing the amount of time allocated for the creation of tests by teachers a time consuming and tedious task if done manually (Saidalavi 2010).

***Corresponding author: AbeerSaad**
*Computer Teacher Preparation Department, Faculty of Specific Education, Damietta University, Egypt.*

The automatic generation of questions is an important research area potentially useful in intelligent tutoring systemsdialogue systems, and educational technologies. (Alqrainy *et al.,* 2012) In Related work on question generation, Sneider (2002). used templates, Heilman and Smith (2009). Used general-purpose rules to transform sentences into question, and Saidalavi used Syntax and Keywords (2006). Parsing is defined as the process of identifying the structure of a specific sentence according to a given grammar. The term parser is used in cases where the sentences are made up of information units of any kind. Parsing Arabic sentences is a difficult task. The difficulty is due to the following reasons: first, the average length of an Arabic sentence is 20 to 30 words, and in some sentences, the number of words exceeds 100. Therefore, Arabic sentences, by nature, are long and complex. Second, the Arabic sentence is syntactically ambiguous and complex due to the frequent usage of grammatical relations, order of words and phrases, conjunctions, and other constructions such as diacritics (vowels), which is known in written Arabic as "altashkiil" (DeRose and Steven 1990). The syntactical analysis requires lexical (syntactical) information for each word in the sentence that needs to parse, such this information usually obtained from the output of a Part-Of- Speech (POS) tagger. POS tagging system is an important first step and an integral part for any parsing system. The aim of POS tagging system is to assign the lexical category (N: noun, V: verb: P: particle) to each word in the parsing sentences (Spector and Harrington 2012). Parts of speech tags of the Arabic language are Part of speech (POS) tags are widely used NLP tools and applications development, For Arabic, there is the Arabic Treebank tagset (Graesser and Chipman 2005).

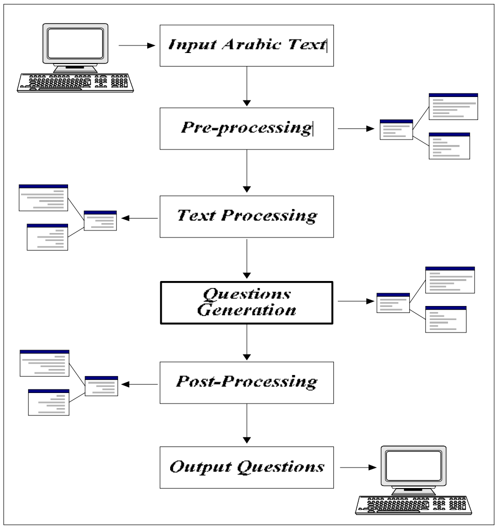
Arabic POS Tagging is the process of identifying lexical category of the Arabic word existing in a sentence based on its context (Abuleil 2004). The most used categories are noun, adverb, verb and adjective (Benajiba and Rosso 2008). There are three general approaches to deal with the tagging problem: Rule-based approach, Statistical approach, and Hybrid approach (Graesser and Lehn 2001). Named Entity Recognition (NER) is an information extraction subtask to classify proper names from unstructured texts into categories of names. NER task has been used to evolve many Natural Language Processing (NLP) subtasks, such as Information Retrieval and Question Answering (Saidalavi 2010).

## II. PREVIOUS WORKS

(Myller 2007) presented a way to automatically generate prediction questions during a program visualization automatically and a proof of concept implementation of it .also presented different types of questions that could be automatically generated with the same framework and ways to determine when those questions should be raised in order to support different ways of learning and testing, it seems that the activities performed by provide various example questions that could be automatically generated based on the data obtained during the visualization process. (Deeb 2012) Presented the system which aims to presentation an automatically generate questions given sentences, by using the dataset provided by the TREC 2007 Question Answering Track. dependent on filtered out important sentences from the dataset by following a target-driven method and simplified the process by extracting elementary sentences from the complex sentences using syntactic information and Part Of Speech Tagging (POS).After classifying the elementary sentences based on their subject, verb, object and preposition, generated the questions automatically from them using a predefined set of interaction rules. (26)Present Development of computer aided environment for drawing (to set) fills in the blanks that can generate for given paragraph. The System finds fill in the blanks, blanking key generates from the selected statement. Syntactic and lexical features are used in this process. The System is developed in Java using JDBC which is open source. The system select important sentence from the given paragraph, and generate fill in the blanks question on them. Syntactic features helped in quality of fill in the blanks generated. (27) Introduces a system for a Fill-in-the-blank questions or cloze items, with generating cloze items for prepositions, and poses problems for non-native speakers of English. The quality of a cloze item depends on the choice of distract- torsbased on collocations and on non-native English corpora to generate distractors for only on word frequency. (28) Develop a system to generate questions automatically from large text corpora User questions. a comparison of the retrieval performance using only automatically generated questions and manually-generated questions 15.7% of the system responses were relevant given automatically generated questions, while 84% of the system responses were deemed relevant with manually-generated questions.

## III. THE PROPOSED SYSTEM

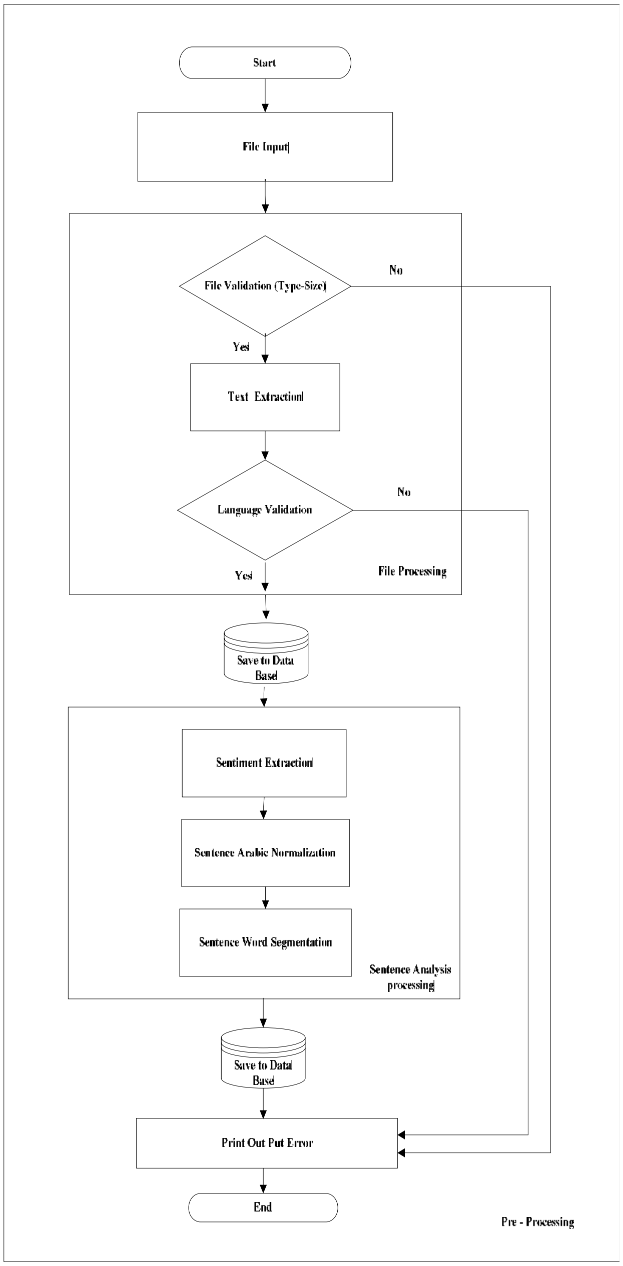
The proposed system was called Arabic Question Generation (AQG) system. It automates the process of generating questions from Arabic texts; it is useful for teachers for the generating questions fill in the blanks automatically Instead of the manual method. The system consists of four phases, and each phase contains some steps, the first phase is the pre-processing phase, second phase is the Text processing, the third phase is the questions Generation, and the fourth phase is the Post-Processing, and each phase contains some steps. The proposed system structure illustrated in Figure 1



**First phase: Pre-Processing**

This phase aims to validate input data and prepare it before processing text. The complex structure of Arabic language contains many issues. It has very complex morphology, such as a (Alhmza - Almadda - Diacritic- Tatweel). The end goal of this stage is to get pure text and remove diacritic and normalize Arabic characters and unify a free from of Alhmza. Also this stage contains a set of sub-stages which will be discussed in the following part.

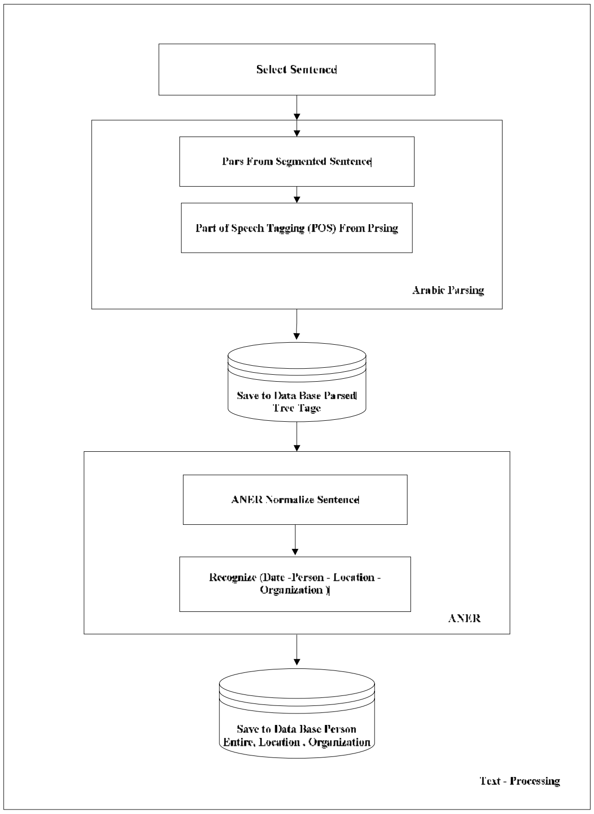
**2 The First phase is illustrated in Figure**



**Question From The Proposed system shown in table1**

| *No* | *Entered Arabic Sentence* | *Question Fill in the blank generate* |
|---|---|---|
| 1 | تم تأميم شركة قناة السويس عام ١٩٥٦ لتحرير اقتصاد مصر | ١ـ تم تاميم ...... عام ١٩٥٦ لتحرير اقتصاد مصر |
| | | ٢ـ تم تاميم شركه قناه السويس عام ...... لتحرير اقتصاد مصر |
| 2 | الوالي هو تاتب السلطان في مص ورئيس السلطة التنفيذية | ١ـ...... هو تاتب السلطان في مصر ورئيس السلطه التنفيذيه |
| | | ٢ـ الوالي هو تاتب السلطان في ...... ورئيس السلطه التنفيذيه |
| 3 | تم إلغاء الملكية وإعلان الجمهورية فى مصر سنة ١٩٥٣ م. | ١ـ تم الغاء الملكيه واعلان ...... في مصر سنه ١٩٥٣ م |
| | | ٢ـ تم الغاء الملكيه واعلان الجمهوريه في مصر سنه ..... م |
| 4 | تولى محمد على حكم مصر من عام ١٨٠٥ الى عام١٨٤٨ . | ١ـ تولي محمد علي حكم مصر من عام ...... الي عام...... |
| | | ٢ـ تولي ...... حكم مصر من عام ١٨٠٥ الي عام |

**The second phase is illustrated in Figure 3**


Select Sentence
Pars From Segmented Sentence
Part of Speech Tagging (POS) From Prsing
Arabic Parsing
Save to Data Base Parsed Tree Tage
ANER Normalize Sentence
Recognize (Date -Person - Location - Organization )
ANER
Save to Data Base Person Entire, Location , Organization
Text - Processing

## Second Phase :Text Processing

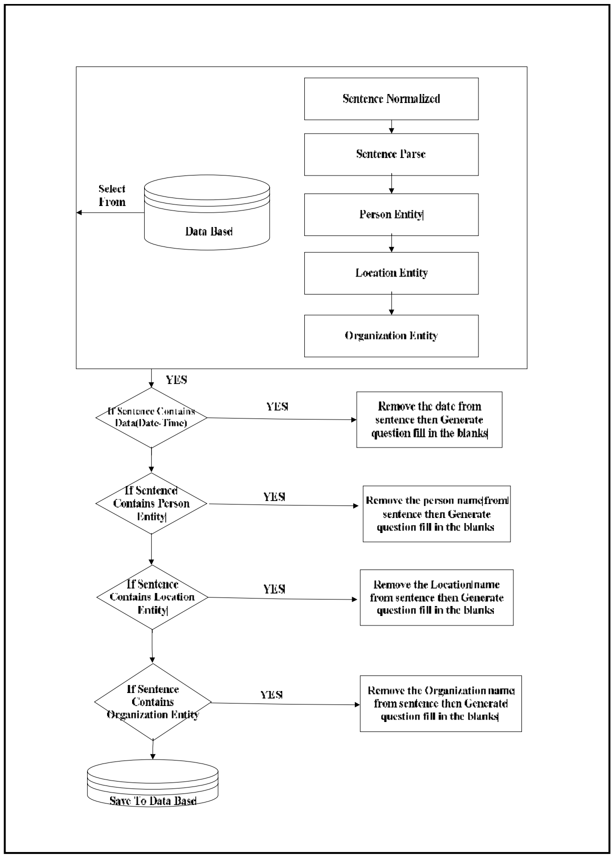
This phase aims to Text Processingfor question generation

This phase consists of two stages main the first stage: Arabic Parsing and the second stage ANER (Arabic name entity organization).

### Third Phase :Questions Generation

This phase aims to generated fill in the blanks questionsare where one or more words are removed from Arabic sentence and replace it with space. System has the property of diversity it is Generates more than a question of the same type to the same sentence. The numbers are identified by using expressions with using Regular Expressions. It is a way to extract information from the tree of the Arabic text. The system as well uses Stanford Arabic Natural Languages Processing (NLP) tools to generate a morphological tagged tree from the Arabic text, and a tool to extract the Pattern and model question. We use the Gazetteer Arabic Named Entity Reorganization (ANER) to find the entity type. For generated Arabic question fill in the blanks from Arabic sentence.

**The third phase is illustrated in Figure 4**


Sentence Normalized
Sentence Parse
Person Entity
Location Entity
Organization Entity
Select From
Data Base
YES
If Sentence Contains Data(Date-Time)
YES
Remove the date from sentence then Generate question fill in the blanks
If Sentence Contains Person Entity
YES
Remove the person name from sentence then Generate question fill in the blanks
If Sentence Contains Location Entity
YES
Remove the Location name from sentence then Generate question fill in the blanks
If Sentence Contains Organization Entity
YES
Remove the Organization name from sentence then Generate question fill in the blanks
Save To Data Base

### Algorithm: Generation Questions fill in the blanks from Arabic Sentence

Algorithm// Question Formation algorithm Explains the procedures takes to form the question using the content. Templates generated.

**Illustrated in Figure 5**

```
Procedure ques Form (SentencesLst: text normalized sentences)
        QuestionsLst ← []
        For each sentence in SentencesLst do
                IF sentence contains digits”Date/Time” Then
                        Digits Value ← match (sentence,
        “\\s+\\d+\\s+”)
                        Remove date from sentence then
        generate question Fill in the blanks
                        questionsLst ← ADD (Question No,
        ques)
                End if
                IF sentence contains Person Entity Then
                        Person Entity ← match (sentence, Per
        Gazetteer)
                Remove that person from sentence then
generate question Fill in the blanks
                        QuestionsLst ← ADD (Question No,
        ques)
                End if
                IF sentence contains Location Entity Then
                        Location Entity ← match (sentence,
        LocGazetteer)
                        Remove that organization from
        sentence then generate question Fill in the blanks
                questionsLst ← ADD (Question No, ques)
                End if
```

### Fourth Phase: Post – Processing

These forth phase of the AQG System aims to Process the saved question from the previous phase and output the final question list. The system offer some additional features for the examiner to generate a complete exam, generate full questions and finally preview the result and save it to the hard disk.

## IV. EXPERIMENTAL RESULTS

Application experience, "a sample of teachers of history, psychology professors, Computer, curricula and teaching methods in Damietta." Graphical user interface of the proposed system is shown in Figure (6)

### SYSTEM PERFORMANCE EVOLUTION

Random sample was selected from the questions generated by the system. The system was tested by through the opinions of expert’s humans twenty (teachers of history, psychology professors, Computer, curricula and teaching methods). Statistical processing based on (suitable significantly, moderately suitable, appropriate degree weak, Inappropriate).

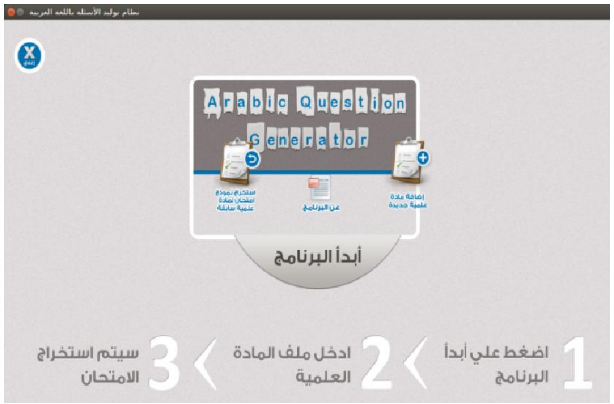
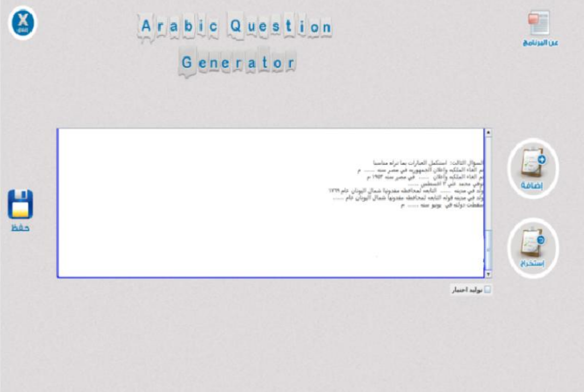

نظام توليد الأسئله باللغه العربيه
إغلاق
Arabic Question Generator
استخراج نموذج امتحان لمادة علمية سابقة
عن البرنامج
إضافة مادة علمية جديدة
أبدأ البرنامج
1 اضغط علي أبدأ البرنامج
2 ادخل ملف المادة العلمية
3 سيتم استخراج الامتحان

**Fig.6. The Graphical user interface of the proposed system**

In the light of the following Evaluation Criteria of questions generated by the system (Relevance‹Question target‹Syntactic‹Variety‹clearness)

**A sample of the results in Table 2.**

| Evaluation Criteria | Relevance | Question Target | Syntactic | Clearness | Variety |
|---|---|---|---|---|---|
| Agreement % | 96.9% | 94.8% | 88.1% | 95.8% | 89.3% |



السؤال الثالث: استكمل العبارات بما تراه مناسبا
تم الغاء الملكيه واعلان الجمهوريه في مصر سنه ...... م
تم الغاء الملكيه واعلان ...... في مصر سنه ١٩٥٣ م
توفي محمد علي ٢ اغسطس ......
ولد في مدينه ...... التابعه لمحافظه مقدونيا شمال اليونان عام ١٧٦٩
ولد في مدينه قوله التابعه لمحافظه مقدونيا شمال اليونان عام ......
سقطت دولته في يونيو سنه ...... م

□ توليد اختبار

**Fig.7. Sample of questionsfrom  the proposed system**

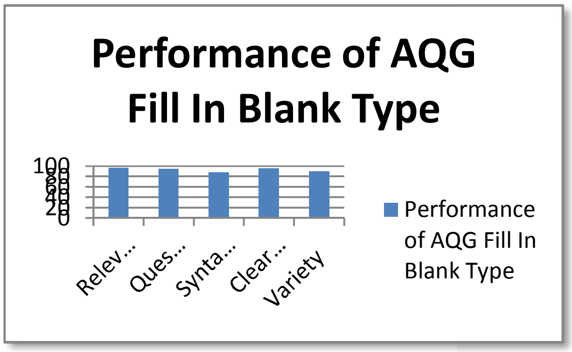


**Fig.8. Results for AQG from Arabic Sentences**

## Conclusion

In In this paper we presented a system to generated question fill in the blanks based on pre-generated corpus patterns. The system select Arabic sentence from the given text, and generated question fill in the blanks on them. It is Generates more than a question of the same type to the same sentence. The system as well uses Stanford Arabic Natural Languages Processing (NLP) tools to generate a morphological tagged tree from the Arabic text; we described the question generation method.

## REFERENCES

Abd-el-Kader A., D. Souilem Boumiza: “A categorization algorithm for the Arabic language”, International Conference on Communication, Computer and Power (ICCCP'09), Muscat, February 2009.

Abney, S. “Principle-based parsing: computation and psycholinguistics”, 1st Edn., Springer, Dordrecht, 1991. ISBN: 0792311736, pp. 408.

Abuleil S.: “Extracting Names from Arabic Text for Question-Answering Systems”, 7th International ConferenceOf Computer-Assisted Information Retrieval Applications, University of Avignon, RIAO 2004, France.

Alqrainy. S. *et al.* “Context-Free Grammar Analysis for Arabic Sentences “*International Journal of Computer Applications* (0975 - 8887) Volume 53 - No. 3, September 2012p8.

Benajiba Y. and P.Rosso:”Arabic Named Entity Recognition using Conditional Random Field”, 6th Int. Conf.On Language Resources and Evaluation”, LREC 2008.

Chen A.  and F.Gey:” Building an Arabic Stemmer for Information Retrieval”, Defense Advanced Research Projects Agency (D ARP A), grant number N66001-00-1-8911 (Mar 2000-Feb 2003).

Chen W.and G. Aist:” Generating questions automatically from informationalText”, In the 2nd Workshop on Question Generation, 2009.

Deeb H.:” Automatic Question Generation: A Syntactical Approach to the Sentence-To-Question Generation Case”, Thesis (Master),  Department of Mathematics and Computer Science, University of Lethbridge, Canada, 2012.

DeRose and Steven:"Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages." PhD. Dissertation. Providence, RI: Brown University Department of Cognitive and Linguistic Sciences, 1990.

Ditters E.:” A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic”, In the proceeding of Arabic NLP Workshop at ACL/EACL, 2001.

Farghaly A.  and K. Shaalan:”Arabic Natural Language Processing: Challenges and Solutions”, ACM Transactions on Asian LanguageInformation Processing (TALIP), the Association for Computing Machinery (ACM). TALIP Vol 8, Issue 4, December 2009.

Forbus U. K. and C.Riesbeck,” Question generation for learning by reading”, Defense Technical Information Center, 2005.

Graesser A. C. and K. Van Lehn. P. W:” Intelligent Tutoring Systems with Conversational Dialogue and Harter”, Almagazine, 22 (4): 39-52,2001.

Graesser A.C. and, P. Chipman: “Auto tutor: an intelligent tutoring system with mixed-initiative dialogue”, IEEE Transactions on Education, 48(4), 612–618, 2005.

Heilman M. and Smith “Question generation via over generating transformations and ranking”, Technical Report CMU-LTI-09-013, Carnegie Mellon University, 2009.

Ibrahim S.:” Arabic language and the importance of Morphology”, 2003. Available: http://ccisdb.ksu.edu.sa /files/rep1120000.doc 1.

Kumar N. and A. Dalal:” Hindi part of speech tagging and chunking”, NLPAI machine learning contest, 2006.

Lafferty J. and A. Callum:” Conditional random fields: Probabilistic models for segmenting and labelling sequencedata”, C. N. Conditional, Pereira, In ICML, 2001.

Lauert W. and P. Graessera:”Questions and Information Systems”, 1992.

Lee J. and S. Seneff:”Automatic Generating of Cloze Items for Prepositions”, 8th Annual Conference of the International Speech Communication Association, Antwerp.

Maria A. and O. Etzioni: “Towards a theory of Natural Language Interfaces to Databases”, IUI ‘03 Miami, Florida USA, ACM 1-58113-586-6/03/0001.

Moath M. and A. Abd-el-Kader:” Arabic Natural Language Processing Laboratory serving Islamic Sciences”, (IJACSA) *International Journal of Advanced Computer Science and Applications*, Vol. 5, No. 3, 2014, p114.

Moldovan D. and A. Hickl:” Experiments with interactive question-answering”, In Proceedings of the 43rd annual meeting on Association for Computational Linguistics, pages 205–214,

Monroeand W. *et al.*” Word Segmentation of Informal Arabic with Domain Adaptation”, Computer Science Department, Stanford University.

Myller N.:”Automatic Generation of Prediction Questions during Program Visualization”, Electronic Notes in Theoretical Computer Science, 178 (2007) p43–49, Available on line at Science direct .com.

Othman E. and K. Shaalan:” A Chart Parser for Analyzing Modern Standard Arabic Sentence, In proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches, New Orleans, Louisiana, U.S.A, 2003.

Ramesh L.A., and M.P. Marcus, “Text chunking using transformation-based learning”, Proceeding of the 3rd ACL Workshop on Very Large Corpora, May 23-23, Computation and Language, Lance Ramshaw, 1995, pp. 82-94.

Saidalavi: K.” Natural Language Question Generation Using Syntax and Keywords”, Proceedings of the Third Workshop on Question Generation p1.

Sneiders E. ”Automated question answering using question templates that cover the conceptual model of the database”, In Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems pp. 235-239, 2002.

Spector L. and K. Harrington:” Tag based modularity in tree-based genetic programming”, In GECCO ’12: Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference, pages 815–822, Philadelphia, Pennsylvania, USA, 7-11 July, IJNLC) Vol. 2, No.6, December 2013+p9/ , 2012.

*******