Question Classification for Arabic Question Answering Systems

Hani Maluf Al Chalabi
Information Technology Department
Al Khawarizmi International College
Al Ain, United Arab Emirates
hani_alchalabi@khawarizmi.com

Santosh Kumar Ray
Information Technology Department
Al Khwarizmi International College
Al Ain, United Arab Emirates
santosh.ray@khawarizmi.com

Khaled Shaalan
Faculty of Engineering and IT
British University in Dubai
Dubai, United Arab Emirates
khaled.shaalan@buid.ac.ae

***Abstract*— Due to very fast growth of information in the last few decades, getting precise information in real time is becoming increasingly difficult. Search engines such as Google and Yahoo are helping in finding the information but the information provided by them are in the form of documents which consumes a lot of time of the user. Question Answering Systems have emerged as a good alternative to search engines where they produce the desired information in a very precise way in the real time. This saves a lot of time for the user. There has been a lot of research in the field of English and some European language Question Answering Systems. However, Arabic Question Answering Systems could not match the pace due to some inherent difficulties with the language itself as well as due to lack of tools available to assist the researchers. Question classification is a very important module of Question Answering Systems. In this paper, we are presenting a method to accurately classify the Arabic questions in order to retrieve precise answers. The proposed method gives promising results.**

Keywords*— *Arabic Question Answering Systems; Question Classification; Interrogative Particles; Nooj

I. INTRODUCTION

Question Answering Systems (QAS) are emerging as important tools to retrieve precise information in real time. A typical QAS consists of three main modules namely, Question Processing, Document Processing, and Answering Processing. In order to answer natural language question, the question answering process analyzes the question first to generate some representation of the information required. The processing of the question is done by the 'question analysis' phase of the QAS. The question analysis phase of question processing consists of different sub-processes like question classification, keyword extraction, derivation of expected answer type, and query expansion. Document Processing and Answer Processing modules help to analyze, rank and retrieve exact answers. There has been a lot of research on QASs in English and Other Latin languages but the same could not be achieved for Arabic QAS s. This paper makes an attempt to reduce this gap by developing question analysis module of Arabic QASs. The rest of the paper is organized as follow: Section II of the paper underlines the significance of Question classification task. Section III describes the approaches used to classify the questions. Section IV introduces question taxonomies used in the proposed method. Section V describes, in details, the proposed method for question classifier. Section VI presents the results for the proposed method. Section VII provides conclusion and future scope.

II. WHY QUESTION CLASSIFICATION?

Question classification is a main part of QAS regardless of various types of architectures [1]. In addition, it has been observed that the pursuance of question classification has made an important influence on the pursuance of QASs[2] [3] [4]. In general, there are two motivations for question classification: predicting the answer entity, and developing answer pattern.

A. Predicting Answer Entity

To find the answer of a given question, the beforehand knowledge of the type of entities such as person, location, date etc. expected to be present in the answer is crucial. The question classification process helps in predicting the type of entities needed to be present in the candidate by classifying the question into various question classes. The answers of each of these question classes must contain specific type of entities. For instance, the question, "متى أصدرت لأول مرة نيويورك تايمز ؟" (When the New York Times was released for the first time?") has a class of type "سنة (year)". While retrieving the answer for this question, the QAS will assign higher ranks to the passages containing the information about year.

B. Developing Answer Pattern

The second motivation behind the question classification task is to develop the linguistic patterns for the candidate answers. These patterns are helpful in matching, parsing and identifying the candidate answer sentences. For instance, consider the question "من هو رائد الفضاء؟" (Who is an astronaut?). The question classification process predicts this question as "Definition" question, and creates the searching patterns for this question as "رائد الفضاء هو ...." (An astronaut is ….) or " ..... يسمى رائد فضاء" (…. is called an astronaut.).

III. QUESTION CLASSIFICATION APPROACHES

The researchers have proposed several methods for question classification. In general, these methods can be grouped into rule-based and learning-based approaches [5] [8]. Rule based approaches use manual approaches for matching the questions by applying hand-crafted rules [6] [7]. However, these approaches are agonized from the need to generate a large number of rules [8]. In addition, rule-based approaches may

perform very well on a specific dataset, but may face difficulties with updated or new datasets. On the other hand, Learning-based approaches perform the question classification task by taking out some lineaments from the questions, train a classifier, and using the trained classifier for predicting the class label. Unlike the rule based approaches, the learning approaches can be aware of recent changes in the existing dataset or can learn with a new dataset. Some research works are using combination of the both approaches to take the full advantage of the best features of the both approaches [9] [10] [11].

IV. QUESTION TYPE TAXONOMIES

The categories of question classes are usually known as question taxonomy. Researchers have offered various question taxonomies. The most widely used question taxonomy was proposed by Li and Roth (2002) which is based on two layer taxonomy containing six classes of coarse-grained and fifty classes of fine-grained entities. This dataset, known as UIUC dataset also, consists of 5500 questions used for training set and the 500 questions used as test set.

TABLE **I**: The question classes of coarse and fine

Coarse	Fine
ABBR	abbreviation, expansion
DESC	definition, description, manner, reason
ENTY	animal, body, color, creation, currency, disease, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUM	description, group, individual, title
LOC	city, country, mountain, other, state
NUM	code, count, date, distance, money, order, other, percent, percent, period, speed, temperature, size, weight

Metzler and Croft (2005) developed the taxonomy dataset of UIUC by adding two classes named list and yes-no-explain. Also, they created 250 questions dataset taken from the MadSci archive (http://www.madsci.org/). There are some more question taxonomies, most notable among them is by Hermjakob et al. (2002) that contains one hundred and eighty classes and represents the widest question taxonomies proposed till now.

V. QUESTION CLASSIFICATION IN ARABIC LANGUAGE

Questions in Arabic language can be classified according to the question words known as *Interrogative Particles* (IP) أدوات الأستفهام. IPs help to identify the suitable answer for a given question. As shown in Fig.1, there are seven IPs in Arabic; some of them have two or more subclasses. IP represents the Arabic words like; kem-كم , min-مِن, ma-ما, ayn-أَينَ mata-مَتى, ay-أَيّ and kaif-كيف. To present the questions in a machine processable format, we have used regular expressions for each IP class. We have used Stanford Parser (http://nlp.stanford.edu:8080/parser/index.jsp) to parse the questions. In this paper, we have designed the patterns to identify wh-questions. Hence, non-wh questions like " أشرح تشكيل الأرض" (Explain the formation of the earth.) will not follow these patterns Also, there can be some exceptions for wh-questions. The logical representation for each IP class is described as follows:

• **kem- كَم** : This IP class holds different meanings like "How much", "How many", "How long", and "How far". The classified answer type of this class is a number. Furthermore, this class can be classified into two subclasses; 1) depends on noun phrase (NP), for instance; كم مدينة زرتها؟ (How many cities did you visit?) and 2) the other one depends on verb phrase (VP), for instance; كم كتابا قرأت اليوم؟ (How many books did you read today?), both subclasses comes after "**kem- كَم** " the IP word (IPW). In all the subsequent patterns, WF (Word Form) represents set of words which do not affect the classification of the questions.

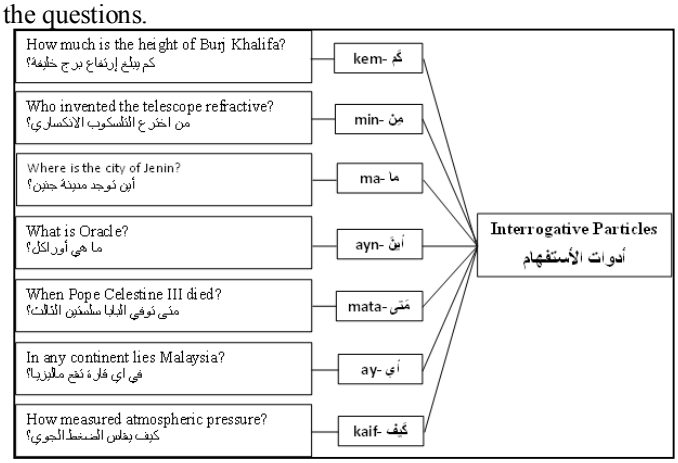

Fig. 1. Interrogative Particles (IP) Classes

The logical representation for all of the questions in this class can take one of the following patterns:

Pattern I:

IPW NP VP WF

Pattern II:

IPW VP WF

An NLP tool, Nooj (http://www.nooj4nlp.net/pages/nooj.html), has been used to graphically represent, train, and test the questions and logical patterns described in this paper. The pattern used for the logical expression grammar of "**kem- كَم**" subclasses are shown in Fig. 2 for Pattern I, and Fig. 3 for Pattern II.

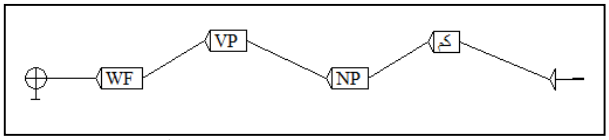

Fig. 2. "**kem- كَم**" logical expression grammar (Pattern I)

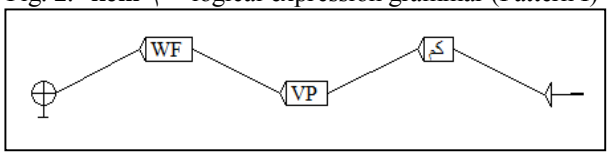

Fig. 3. "**kem- كَم**" logical expression grammar (Pattern II)

• **min-مِن** : This IP classes is equivalent to English question word "Who" that indicates the persons, organizations , etc. For instance, consider the question; من هو رئيس الهند؟ (Who is the president of India?). The classified answer type of this class is a person. The syntax of using "min-مِن" in Arabic language is

based on noun phrase that always comes after "**min-مِن**". The logical representation for all of the questions in this class can take the following pattern:

IPW NP WF

The pattern used for the logical expression grammar of "**min-مِن**" class shown in Fig. 4.

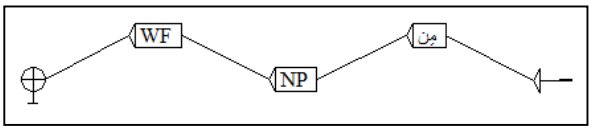

Fig. 4. "**min-مِن**" logical expression grammar

• **ma-ما**: This class of the IP is equivalent to "What" that indicated things, for instance; "ما هي وحدة قياس الزلازل؟" (What is the unit of earthquake measurement?). The answer type of this class can be device, geographical location, sports, organization, art, person, etc. Furthermore, "**ma-ما**" can be classified into three subclasses; 1) depends on NP that usually starts with the word "hoa-هو", for instance; "ما هو السي ++؟" "What is C++?", 2) depends on NP that usually starts with the word "hea-هي", for instance; "ما هي اليونسكو؟ " "What is UNESCO?", and 3) depends on NP that usually starts with verb, for instance; "ما معنى بي أو سي؟ " (What is the meaning of BOC?). The logical representation for all of the questions in this class can take one of the following patterns:

Pattern I:

IPW HOA NP WF

Pattern II:

IPW HEA NP WF

Pattern II:

IPW NP WF

The pattern used for the logical expression grammar of "**ma-ما**" subclasses are shown in Fig. 5 for Pattern I, Fig. 6 for Pattern II, and Fig. 7 for Pattern III.

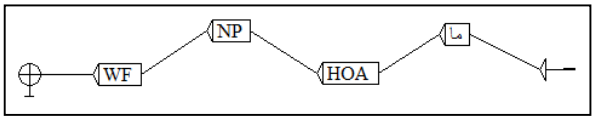

Fig. 5. "**ma-ما**" logical expression grammar(Pattern I)

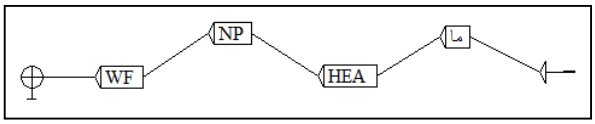

Fig.6. "**ma-ما**" logical expression grammar(Pattern II)

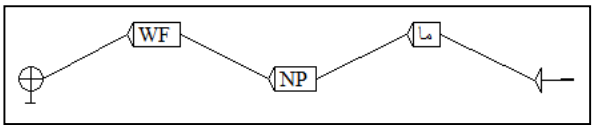

Fig.7. "**ma-ما**" logical expression grammar (Pattern III)

• **ayn-أَينَ**: It is the fourth class of IP classes used in Arabic question answering which holds a meaning of "Where" that indicates the place. For instance consider the question; أين ولد ابن بطوطة؟ (Where did Ibn Battuta born?). The classified answer type of this class is the geographical location. The syntax of using "ayn-أَينَ" is based on verb phrase that always comes after "**ayn-أَينَ**". The logical representation for questions in this class can take the following pattern;

IPW VP WF

The pattern used for the logical expression grammar of "**ayn-أَينَ**" class shown in Fig. 8.

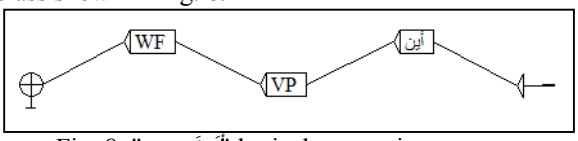

Fig. 8. "**ayn-أَينَ**" logical expression grammar

• **mata-مَتى**: This class of IP holds a meaning of "When" that indicated the date, for instance; متى اصدرت لأول مرة نيويورك تايمز؟ (When did the first time New York Times was issued?). The syntax of using "**mata-مَتى**" in Arabic language is based on verb phrase that always comes after "**mata-مَتى**". The logical representation for all of the questions in this class can take the following pattern:

IPW VP WF

The pattern used for the logical expression grammar of "**mata-مَتى**" class shown in Fig**.** 9.

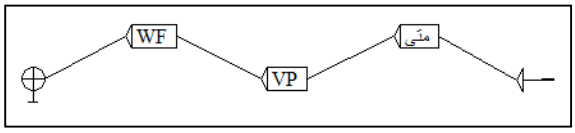

Fig.9. "**mata-مَتى**" logical expression grammar

• **ay-أَي**: This class of the IP means "Which" that indicated sapiens and non-sapiens, for instance; "أي دولة حكمها الملك فيصل؟" (Which country ruled King Faisal?). The answer type of this class can include number, geographical location, history, and sports. Furthermore, this class can be classified into five subclasses; 1) "**ay-أَي**" comes after a preposition word "in-في" and followed by NP, for instance; " في أي يوم أنتخب جورج واشنطن لرئاسة الولايات المتحدة الامريكية ؟" "In which day George Washington elected to the presidency of the United States?", 2) "**ay-أَي**" comes after a preposition "to-الى" and followed by NP, for instance; "الى أي مدى يصل ارتفاع برج إيفل؟" "To what extent rising up the Eiffel Tower?", 3) "**ay-أَي**" comes after a preposition "from-مِن" and followed by noun phrase NP, for instance; "من أي أسم أشتق أسم مدينة قورنيا؟ " "From which name the name of the city of Cyrene has been derived?", 4) "**ay-أَي**" comes after a preposition "on-عَن" and followed by noun phrase NP, for instance; "عن اي دولة استقلت تيمور الشرقية؟" "On which country became independent East Timor?", and 5) finally, "**ay-أَي**" comes at the beginning of the question followed by NP, for instance; "أي سنة تم بناء برج خليفة؟" "Which year was building Burj Khalifa?".

The logical representation for all of the questions in this class can take one of the following patterns;

Pattern I: For the first five types of subclasses mentioned above;

PP IPW NP WF

Pattern II: For the sixth type;

IPW NP WF

The patterns used for the logical expression grammar for Pattern I and Pattern II are shown in Fig. 10 and 11.

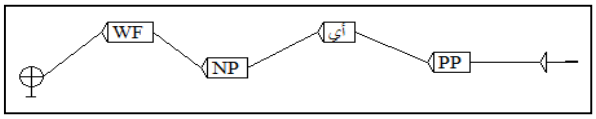

Fig.10. "**ay-أَي**" logical expression grammar(Pattern I)

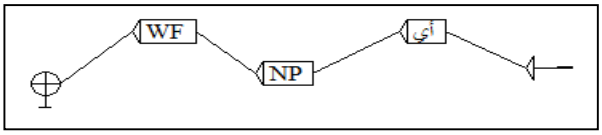

Fig.11. "**ay-أَي**" logical expression grammar (Pattern II)

• **kaif-كَيف**: The last class of IP holds a meaning of "How" that indicated sapiens and non- sapiens, for instance; كيف يعمل جهاز الكمبيوتر؟ (How does the computer work?). The classified answer type of this class is science. The syntax of using "**kaif-كَيف**" in Arabic language is based on verb phrase that always comes after "**kaif-كَيف**". The logical representation for questions in this class can take the following pattern;

IPW VP WF

The pattern used for the logical expression grammar of "**kaif-كَيف**" class shown in (**Fig. 3.12**).

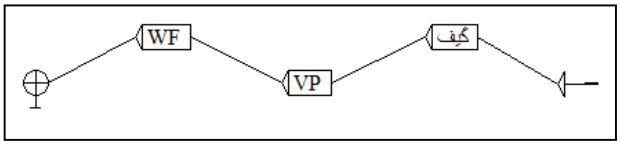

Fig.12. "**kaif-كَيف**" logical expression grammar

VI. Performance Evaluation of Question Classifier

Generally, the performance of a question classifier can be measured by computing the precision and recall of the system. These are defined as

Precision(c) = No. of samples correctly classified as c / No. of samples classified as c

Recall = No. of samples correctly classified as c / No. of samples in class c

We used *Regular Expression* module of *Nooj* tool to write regular expression. We first tested these regular expressions over a set of 200 Arabic questions called *training questions* developed by Y.Benajiba (http://users.dsic.upv.es/~ybenajiba/) and did the necessary modification in the regular expressions. In the second phase, we wrote context free grammar for these regular expressions using *Grammar* module of *Nooj*. The question classification module was then evaluated over another set of 200 Arabic questions called *test questions* developed for TREC and CLEF. 186 questions out of 200 questions satisfied the context free grammars written by us (recall 0.93). All of these questions were classified correctly which is equivalent to precision of 1. This result is promising as compared to some recent English Question Answering Systems with recall 0.63 and precision 0.7 [14] and recall 0.73 and precision 0.73 [15] . Thus, our results shows the effectiveness of regular expressions and context free grammars in classifying the questions.

VII. Conclusion and Future Scope

In this paper, we proposed question classification methods for Arabic questions using regular expressions and context free grammars. We used Nooj for designing the logical expressions. The results are very promising. We are planning to design methods for other modules of Arabic QASs.

References

[1] Ellen M. Voorhees. Overview of the trec 2001 question answering track. In In Proceedings of the Tenth Text REtrieval Conference (TREC, pages 42–51, 2001.

[2] A. Ittycheriah, M. Franz, W. J. Zhu, A. Ratnaparkhi, and R. J. Mammone. IBM's statistical question answering system. In Proceedings of the 9th Text Retrieval Conference,NIST, 2001.

[3] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin yew Lin, and Deepak Ravichandran.Toward semantics-based answer pinpointing, 2001.

[4] Dan Moldovan, Marius Pa¸sca, Sanda Harabagiu, and Mihai Surdeanu. Performance issues and error analysis in an open-domain question answering system. ACM Trans. Inf. Syst., 21:133–154, April 2003.

[5] Xin Li and Dan Roth. Learning question classifiers. In Proceedings of the 19th interna-

[6] tional conference on Computational linguistics, COLING '02, pages 1–7. Association for Computational Linguistics, 2002.

[7] John Prager, Dragomir Radev, Eric Brown, and Anni Coden. The use of predictive annotation for question answering in trec8. In In NIST Special Publication 500-246:The Eighth Text REtrieval Conference (TREC 8, pages 399–411). NIST, 1999.

[8] Xin Li and Dan Roth. Learning question classifiers: The role of semantic information. In In Proc. International Conference on Computational Linguistics (COLING, pages 556–562, 2004.

[9] Zhiheng Huang, Marcus Thint, and Zengchang Qin. Question classification using head words and their hypernyms. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP '08), pages 927–936, 2008.

[10] Santosh Kumar Ray, Shailendra Singh, and B. P. Joshi. A semantic approach for question classification using wordnet and wikipedia. Pattern Recogn. Lett., 31:1935–1943, October 2010.

[11] João Silva, Luísa Coheur, Ana Mendes, and Andreas Wichert. From symbolic to subsymbolic information in question classification. Arti_cial Intelligence Review, 35(2): 137–154, February 2011.

[12] Donald Metzler and W. Bruce Croft. Analysis of statistical question classification for fact-based questions. Inf. Retr., 8:481–504, May 2005.

[13] Ulf Hermjakob, Eduard Hovy, and Chin yew Lin. Automated question answering in webclopedia - a demonstration. In Proceedings of ACL-02, 2002.

[14] Borhan Samei, Haiying Li, Fazel Keshtkar, Vasile Rus, and Arthur C. Graesser. Context-based speech act classification in intelligent tutoring systems. In Intelligent Tutoring Systems (pp. 236-241). Springer International Publishing, January 2014.

[15] Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngonga, Elena Cabrio, Philipp Cirniano, Sebastian Walter. Question Ansering over Linked Data (QALD-4). In Working Notes for CLEF 2014 Conference, 2014.