IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Question Focus Recognition in Question Answering Systems

Waheeb Ahmed[1], Dr. Babu Anto P[2]

Research Scholar, Department of Information Technology, Kannur University, Kerala, India[1]

Associate Professor, Department of Information Technology, Kannur University, Kerala, India[2]

Abstract: Question Answering (QA) Systems are systems that attempts to answer questions posed by human in natural language. As a part of the QA system comes the question processing module. The question processing module serves several tasks including question classification and focus identification. Question classification and focus identification play crucial role in Question Answering systems. This paper describes and evaluates the techniques we developed for answer type detection based on question classification and focus identification in Arabic Question Answering systems. Question classification helps in providing the type of the expected answer and hence directing the answer extraction module to apply the proper technique for extracting the answer. While focus identification helps in ranking the candidate answers. Consequently, that has increased the accuracy of answers produced by the QA system. Question processing module involves analysing the questions in order to extract the important information for identifying what is being asked and how to approach answering it, and this is one of the most important components of a QA system. Therefore, we propose methods for solving two main problems in question analysis, namely question classification and focus extraction.

Keywords: Question classification, Question focus extraction, Question Answering Systems, Information Retrieval, Natural Language Processing.

I. INTRODUCTION

Traditional search engines like Google and Yahoo provides the user with a set of links based on the user query and it is the responsibility of the user to go through these links and try to get the answer that he/she is looking for. On the contrary, Question Answering (QA) systems tries to generated precise answers automatically for questions presented in natural languages. Developing a completely capable QA system, however, has challenges mostly due to several challenging sub-problems that require to be solved, such as question analysis (involving pre-processing, classification of questions and identification of focus), information retrieval and generation of answer (including extracting and formulating answer), along with some other lower level subtasks, such as reference resolution and paraphrasing.

In addition, the type of a QA system and the employed techniques usually depend on factors such as domain of question and language. Many researchers have opted for solving the individual problems involved in such systems separately. While some of these problems are considered to be solved, the majority of them are still open to further research [1, 2]. The main goal of question classification is to precisely assign labels to questions based on expected answer category [3].

II. METHODOLOGY

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it. Most QA systems- as part of question analysis- determine the answer type, i.e. the class of the entity, sought by the question [3]. For example, the question Q1:" من اكتشف التلفزيون؟" (Who invented the television?) is asking for the name of a Human entity (Person), whereas Q2:" ما هي الألياف الضوئية؟" (What are Optical fibres?) looks for a Definition type of discourse. Therefore, the answer types will be Human, and Definition respectively. Knowing the answer type correlated with a given question can assist during the answer extraction stage, where the system will use it to select the answer from a wide range of candidates. Moreover, the answer type will determine the technique used for extracting the correct answer. The answer type for question Q1(Human) means that the required answer is simply the name of a person, and can be identified using a named entity recognizer.

The question Q2 is of type definition, and it will involve techniques that identify paragraphs with definition structures concentrated on the question topic (optical fibres), or more complex techniques in which sentences on the question topic are automatically gathered from multiple documents into an answer paragraph that is presented in the form of a definition. The user use an explicit set of answer types emphasize the importance of summarizing the given natural language question into a single word which clarifies the type of the expected answer: according to [4] the question focus is defined as "a noun phrase composed of several words and in some

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

cases a simple noun, that is the property or entity being sought by the question". According to [4], the nouns country, city, population and colour are the focus nouns in the following questions:

Q 3 "في أي مدينة يقع متحف اللوفر؟"
("Louvre Museum is located in what city?")
Q 4 "كم سرعة الضوء؟"
("How much the speed of light?")
Q 5 " ما هو حرف العلة الثاني الأكثر استخداما في اللغة الإنجليزية؟"
(" What 's the second-most-used vowel in English? ")
Q 6 "ما هي أكبر شركة نشر في أمريكا؟"
("What company is the largest American publisher?")
Q 7 "ما هو أطول مبنى في العالم؟"
("What is the tallest building in the world?")

In question Q6, the word company emphasizes the type of the answer. In question Q7 and the noun phrase "the largest city in Germany" specifies the focus of the question, while at the same time its head noun city specifies a type/class of the answer. The following table gives examples of natural language questions along with answer type/question class and question focus.

TABLE I QUESTIONS WITH THEIR ANSWER TYPE & FOCUS

Question	Answer Type	Focus
ما هو الكمبيوتر؟ What is Computer?	Description: definition	كمبيوتر Computer
ما هو لون الشمس؟ What is the colour of the sun?	Entity: colour	لون الشمس Colour of the sun
من هو أول مدير لناسا؟ Who is the first director of NASA?	Human: individual	أول مدير لناسا First director of NASA
أين يوجد أطول نهر في العالم؟ Where does the longest river in the world exist?	Location: river	أطول نهر في العالم؟ Longest river in the world

The following sections explains the process of identifying the answer type and question focus;

A. Answer Type Detection/Question Classification

The question processing module tries to provide each question an answer type, which is a label that specifies what kind of answer the question is looking for. For question classification we used the same taxonomy proposed by Li & Roth [5].

In our previous work [6] we used Support Vector Machine (SVM) classifier which is trained using a training data consisting of 300 questions derived from the Arabic Wikipedia and tested over 200 questions translated from Text Retrieval Conference (TREC 10) [7].

For example:
Question 8: "من اكتشف لقاح البنسلين؟"
("Who invented the Penicillin vaccine ?")
Question Class = PERSON
Furthermore, question analysis tries to get a more general type. It tries to find a noun or a noun phrase. For example,
Question 9: " ما هي السيارة التي لديها أعلى قوة حركية في العالم؟"
("What car has the highest horsepower in the world?")
General Type = "سيارة" (car)
Question 10: "ما هو اسم أول شركة تأمين في نيو يورك؟"
("What is the name of the first insurance company in New York?")
Named Entity List = ORGANIZATION
General Type = company
The following architecture shows the overall process of answer type detection and question focus identification:

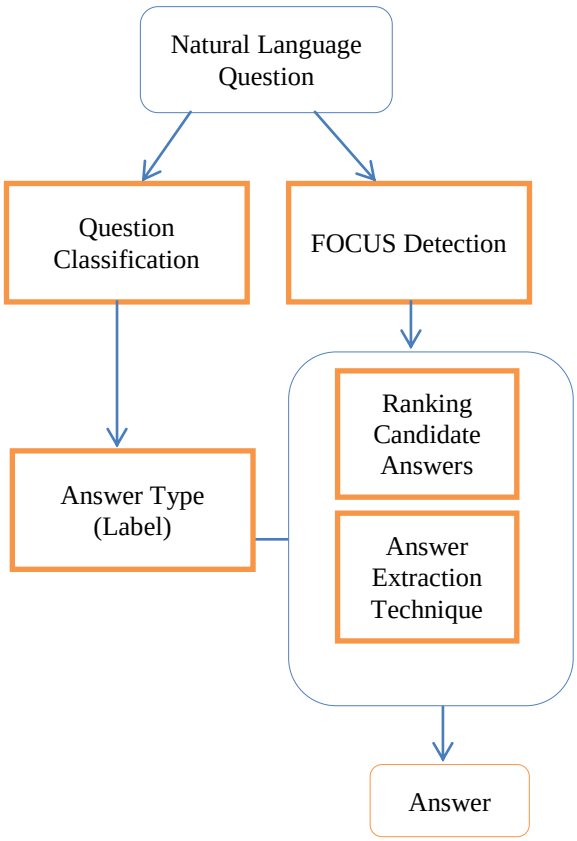

Fig. 1 Architecture for answer type detection and question focus identification.

B. Defining Question Focus

To the best of our knowledge, all previous literature on answer typing considers only the question class for identification of answer type. The question focus is the set of all the noun phrases available in the question that clarify the type of the answer. In this regard, question processing tries to find the question focus, which represents a noun or a noun phrase that is likely to be present in the answer. For each question, we will determine a focus, a focus head (the main noun) and the "modifiers" of the focus head (adjective, complement...).

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

For example:
Question:"من هو أول رئيس لجامعة هارفارد؟"
("Who was the first rector of Harvard university?")
FOCUS = "أول رئيس لجامعة هارفارد"
(first rector of Harvard university)
FOCUS-HEAD = "رئيس" (rector)
MODIFIERS-FOCUS-HEAD = ADJ "أول" (first), COMP "جامعة هارفارد" (Harvard university)

C. Question Focus recognition

The focus of a question is considered as follows: (a) the head of the focus, (b) a list of modifiers. The question processing module tries to identify this focus in the sentences of the retrieved documents. It first finds the head of the focus, and then finds the noun phrase in which the head is enclosed. To determine the borders of this noun phrase, we define grammar rules for the NP in Arabic. This grammar relies on the output of Stanford Part-Of-Speech Tagger(POS) for Arabic language. For example, for the following question:

"من هو مؤلف هاري بوتر؟"
("Who is the author of Harry Potter?"),
the focus is " مؤلف هاري بوتر" (the author of Harry Potter), with the head : "مؤلف" (author).
we found the following NP: "مؤلف هاري بوتر" (the author of Harry Potter)
which fits the following expression:
Noun + Proper Noun + Proper Noun
Other grammar rules that we used for extracting various noun phrases are listed below:
Noun + Adjective
Plural Noun + Adjective
Noun + Proper Noun+ Proper Noun
Noun + Adjective + Adjective
Adjective + Noun + Noun

When the identification of the focus in the question failed, Question processing module looks for the proper nouns in the question, and it attempts to recognize NPs that contain these proper nouns. The scoring algorithm always considers the number of words available in the question phrase and in the document noun phrase.

For example the score given to the NP:
"في الأخير، المستضيف والمرشحة للبطولة أوروغواي هزمت الأرجنتين 4-2 أمام حشد من 93000 شخص وأصبحت أول دولة تفوز بكأس العالم".("In the final, hosts and pre-tournament favourites Uruguay defeated Argentina 4–2 in front of a crowd of 93,000 people, and became the first nation to win the World Cup.") Obtained for the question:

"ما هو أول بلد يفوز بكأس العالم؟"
("Which country won the first world cup for football?"), has been assigned a higher score because it has been obtained from the focus "بلد" (country), and from the noun phrase "أول كأس عالم" (first world cup), because of the availability of all the words of this noun phrase: "أول كأس عالم" (first world cup).

For each sentence of the selected document, Question processing module assigns all the relevant NPs according to the preceding algorithm, with the associated scores. It only retains the NPs which got the best scores, which in turn provides an evaluation of the relevance of the sentence, which will be used in the sentence selection module.

D. Answer Extraction and Ranking

The answer extraction module selects of a set of sentences that may contain the answer to a question is based: it compares each sentence from the selected documents for a question to this question and constantly stores them in a list of sentences that are the most similar to the question (The total number of selected sentences are 5). The comparison between the question and each sentence depends on a set of features extracted both from the questions and the sentences of the selected documents:
Terms, focus and named entities and distribution of terms in the sentence.
A similarity score is derived individually for each of these features. The last feature enables the module to decide between two sentences having the same score for the first three features. In this paper we tried several weighting schemes for terms[8]. The one we choose here was to sum the weights of the terms of the question that are in the document sentence. The term score is combined with the focus score and the resulting score constitutes the first condition for evaluating two document sentences S1 and S2: if S1 has a combined score much higher than S2, S1 is ranked higher than S2. If not, the named entity score is used in the same way.

III.RESULTS AND PERFORMANCE EVALUATION

To evaluate our proposed method we used a set of 100 question translated from TREC 10. We calculated the precision ,recall and F- measure of the question processing module while classifying the questions and identifying the question focus of the 50 questions supplied to the module. For evaluating the performance of the question focus identification subtask we used the following formulas:

$$\text{Precision} = \frac{\text{No. of QFOCUS Correctly Detected}}{\text{No. of Detected QFOCUS}}$$

$$\text{Recall} = \frac{\text{No. of QFOCUS Correctly Detected}}{\text{Actual No.of QFOCUS available in the question set}}$$

$$\text{F1-Measure} = 2\frac{(\text{Precision})(\text{Rcall})}{\text{Precision} + \text{Recall}} \quad [9]$$

For evaluating the performance of the question classification/answer type detection we used the following formulas:

$$\text{Precision} = \frac{\text{No.of samples correctly classified as c}}{\text{Total No.of samples classified as c}}$$

$$\text{Recall} = \frac{\text{No.of samples correctly classified as c}}{\text{Actual No.of samples in class c}}$$

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

TABLE II Performance Evaluation of Question Focus (QFOCUS) Identification subtask Performance Evaluation

Criteria	Precision	Recall	F1-measure
Question Focus Identification	79%	73%	75%

TABLE III Question Classification (QC) Subtask Performance Evaluation

Criteria	Precision	Recall	F1-measure
Question Classification	96.3%	91%	93%

TABLE IV Impact on Question Answering System

Criteria	Question-Answering accuracy (%)
Using Question class (QC) Only	63.4%
Using Both Question Class and Question Focus (QFocus)	71%

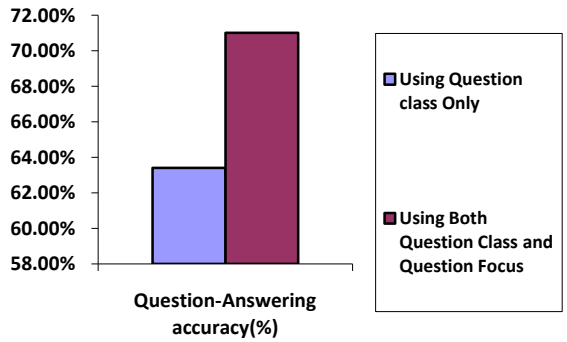

Fig. 2 Accuracy distribution of QA system performance evaluation using two features (QC & QFocus)

From Figure 2, it is obvious that the accuracy of the question answering system increased when the question focus identification is employed. This result proves that using question classification and question focus recognition tasks in the question processing module can highly improve the accuracy of the Question Answering System.

IV. CONCLUSION

We have developed techniques for Question classification and Question focus recognition for the purpose of increasing the accuracy of the Arabic QA Systems. The result we got is promising and these techniques can be used effectively in developing better QA Systems. As a future work, additional features will be incorporated besides the techniques we provided to get a more efficient QA system.

REFERENCES

[1] P. Gupta, V. Gupta: "A survey of text question answering techniques." International Journal of Computer Applications Vol. 53, No. 4, pp. 1–8, September 2012.
[2] A. Allam, M. Haggag: "The question answering systems: A survey." International Journal of Research and Reviews in Information Sciences (IJRRIS), Vol. 2, No. 3, pp. 211-221, September 2012.
[3] X. LIU and L. LIU," Question Classification Based on Focus", International Conference on Communication Systems and Network Technologies, pp. 512-516, May 2012.
[4] J. Prager, "Open-Domain Question Answering". Foundations and Trends in Information Retrieval, Vol. 1, No. 2, pp. 91–231, 2006.
[5] X. Li and D. Roth, "Learning Question Classifiers: The Role of Semantic Information", Journal of Natural Language Engineering, Vol. 12, pp. 229 - 249, September 2006.
[6] A. Waheeb and A. Babu,"Classification of Arabic Questions Using Multinomial Naïve Bayes And Support Vector Machines", International Journal of Latest Trends In Engineering And Technology, pp. 82-86, SACAIM, November 2016.
[7] E. M. Voorhees," Overview of the TREC 2001 question answering track," Proceedings of the 10th Text Retrieval Conference, pp. 42–52, 2001
[8] O. Ferret et al,"QALC – The Question Answering System of LIMSI-CNR". In Proceedings of the 9th Text Retrieval Conference (TREC-9), June 2000.
[9] A. Lally et aL," Question analysis: How Watson reads a clue.", IBM Journal of RESEARCH & DEVELOPMENT, VOL. 56, No. ¾, PAPER 2, JULY 2012.