Arabic Question Answering System Based On Data Mining

Waheeb Ahmed, Babu Anto P

Abstract: In this study, we describe An Arabic Question Answering(QA) system based on data mining approach. The system employs text mining techniques to determine the likely answers to factoid questions. It depends mainly on the use of lexical information and does not apply any complex language processing tools such as named entity recognizers, parsers and ontologies. The system achieved an accuracy of 61.5%.

Index Terms: Information Retrieval, Information Extraction, Natural Language Processing, Question Answering System, Text Mining

1 Introduction

Day by day, internet users number is increasing in percentage. The traditional search engines accept user's queries in the form of a set of keywords and they return a list of hyperlinks to documents. It is the sole responsibility of the user to search for the answer in the documents referred by these hyperlinks. In this regard, the motivation for systems that provide exact answers in response to users' questions given in natural language has risen. QA systems are extracting information provided by these hyperlinks. Question Answering provides perfect solution to return/retrieve correct and accurate answers to user. QA system enables the users to provide questions in natural language text and receive concise and relevant response as result. Many techniques from Artificial Intelligence(AI), Natural Language Processing(NLP), Information retrieval(IR), Machine Learning(ML) and Information extraction(IE) are brought together to propose a better QA system[1][2]. Recent developments in QA use several of linguistic tools/resources to help in understanding the natural language questions and the documents. However, these approaches have the drawback of development complexity and not optimal practically. In this study, we propose a QA system that can answer factoid questions like what, where, when, and who. This system is based on a text mining approach that requires a minimal information about the syntax and the lexicon of the specified language. It is built on the assumption that the questions and their answers are commonly expressed using the same set of terms. Therefore, it simply employs lexical information to identify the document containing relevant passages and to extract the candidate answers. The following sections present the details of the system. In particular, section 2 describes the method and architecture of the proposed QA system for factoid questions and section 3 discusses the results achieved by our system for Arabic language.

2 Proposed Question-answering System

It consists of three main modules: question classification identifies the type of the expected answer; passage retrieval in which passages with a high probability of containing the answer are retrieved from the document collection; and answer extraction, which selects candidate answers using a machine-learning approach, and the selection of final answer is done. The following sections describe in more detail each of these modules. Figure 1 shows the general process for answering factoid questions.

2.1 Question Classification

This module identifies the semantic class of the expected answer. The identified answer type will be used later by the answer extraction module to apply the proper technique for extracting the answer and hence reduce the searching space. The is to direct the answer extraction only to focus on those text blocks related to the expected type of answer. Our QA system performs this task using an approach based on regular expressions. It only deals with three main semantic classes of expected answers: names, dates, and quantities.

2.2 Document Retrieval Module

This module returns documents from the corpus that are possibly to contain answers to the user's question. It consists of text search engine. This module takes the user's question as input and creates a query containing a set of terms which are likely to occur in documents containing an answer. The generated query is passed to the text search engine, which employs it to return a set of documents[3].

2.3 Passage Retrieval Module

Passage extraction module take a document as input and attempts to identify passages from the document that contain an answer. the passage retrieval method identifies the passages with the relevant terms using a traditional information retrieval technique which is based on the vector space model. The passage retrieval module divides the document into passages, calculate a score for each passage and retrieves the passage having the highest score[4][5][6].

- *Waheeb Ahmed is currently pursuing Ph.D. program in Information Technology in Kannur University, India, E-mail: waheeb2003aden@yahoo.com*
- *Babu Anto P is currently working as associate professor at the department of Information Technology in Kannur University, India, E-mail: batop@gmail.com*

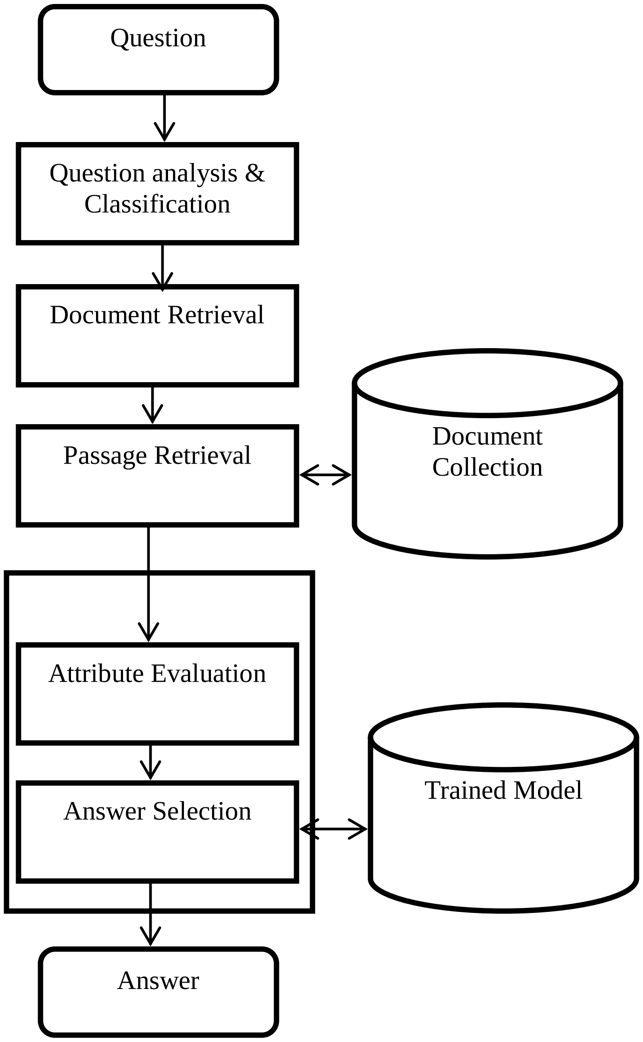

Fig. 1. *QA Architecture*

2.4 Answer Extraction

This module takes a passage from the passage retrieval module as input and tries to return an exact phrase as an answer . User needs a precise and very concise answers. It is constructed based on a supervised machine-learning approach. It is composed of two main modules namely: attribute extraction module and answer selection module. Attribute extraction . The retrieved passages are analyzed in order to determine all text blocks related to the expected type of answer. Each identified text block is considered as a candidate answer. For doing this analysis, A set of regular expressions are applied. Next, analysis of the lexical context of each candidate answer is performed with the intent of building its formal representation. That is, each candidate answer is represented by a set of 21 attributes, clustered in the following sets:

1. Attributes that describe the complexity of the question, for instance, number of words of the question (excluding stop-words).
2. Attributes that computes the similarity between the given question and the context of the candidate answer and. Some of them describe the common words between the context of the candidate answer and the question. Others computes the distribution of the question words in the context/text of the candidate answer.
3. Attributes that denote the relevance of the candidate answer in association with the set of recovered passages. Some of these attributes are: the repetition/redundancy of the candidate answer in the set of the retrieved passages, and the location of the passage containing the candidate answer.

Answer Selection. Based on the results of question analysis, answers are selected from the candidate answers. This module is based on a machine-learning approach. The main purpose of this module is to select the answer, from the set of candidate answers, the answer with the highest score of being the correct answer. The implementation of this module is done by Support Vector Machines(SVM) where the classifier is trained on a set of question-answer pairs.

3 Evaluation Results

Our QA system can only deal with factoid questions. From the 200 test questions. Table 1 shows the details our results on answering the 200 questions. The training data for the answer selection module consists of 500 question-answer pairs. The number of corrects answers was the highest and the system achieved 61.5% accuracy.

Table 1. *Evaluation of the system for factoid questions*

	# of Right	# of Wrong	# of Inexact	Accuracy
Evaluation	123	66	11	61.5%

4 Conclusions and Future Work

In this paper, we presented a question answering system that can answer factoid questions. The system is based on text mining approach. It is known that the questions and their answers are mostly expressed using the same set of words, and based on that it uses lexical information to detect the relevant passages and the candidate answers. The answer extraction module is based on a machine learning approach. Each candidate is expressed by a set of lexical attributes and a classifier identifies the most likely answer for the given question. The proposed method achieved nice results, however it requires a lot of training data. As future work we need to improve the final answer selection stage by applying an answer validation method.

References

[1] G. Nanda, M. Dua and K. Singla, “A Hindi Question Answering System using Machine Learning Approach”, In Proceedings of the International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 2016.

[2] Y. Liu, X. Yi, R. Chen, and Y. Song, “A Survey on Frameworks and Methods of Question Answering”, In Proceedings of the 3rd International Conference on Information Science and Control Engineering, IEEE, pp. 115-119, 2016.

[3] H. Hu, "A study on Question Answering System Using Integrated retrieval method", Phd Thesis Submitted to Graduate school of engineering at the University of Tokushima, February, 2006.

[4] S. Tellex, "Pauchok: A Modular Framework for question Answering", Master Thesis Submitted to the Department of Electrical Engineering and computer science, Maccachusetts institute of Technology, June 2003.

[5] H. Sundblad, "Question Classification in Question Answering systems", Phd Thesis Submitted to Department of Computer and information Science at Linkoping University, 2007

[6] Y. Niu, "Analysis of Semantic classes: Toward Non-Factoid question answering", Phd Thesis submitted to Department of Computer science, University of Toronto, 2007.