A New Approach For Open-Domain Question Answering System

Ibrahim Mahmoud Ibrahim Alturani[1] and Mohd Pouzi Bin Hamzah[2]

Universiti Malaysia Terengganu, Terengganu, Malaysia

Abstract
Many people have many queries related to the deferent topics. They are inquisitive to find these answers using search engines. The most problem in the search engine is that they provide you with the link to the web page contains passages include answer instead of the exact answer. The goal of a question-answering system is to retrieve answers to questions rather than full documents or passages.

QA systems can be running on a closed domain or open domain. In close domain they work on a specific field, like biology or medicine or other, they do not need to integrate different knowledge bases thus less from ambiguity, and they create faster and higher quality results. But closed-domain approaches, not flexibility and there are high costs when adopting such a system to a new domain or implementing a new system. Thus, the research focus has moved to domain-specific, adaptable extension or open-domain QA systems.

This paper proposed a new approach architecture for open domain question-answering system depends on the ontology and wordnet to improve answer accuracy. The result of experiments generated a score of 84.7% for WH questions and 82.6% for Yes/No questions.

Key words:

Question answering, Domain selection, Document processing, Answer processing, Ontology, WordNet, Named Entity Recognition.

1. Introduction

Nowadays the internet contains a vast number of documents, links and all other types of information such as digital libraries, newspapers collections, and others, stored in electronic format. To take advantage of this information through the combination of the web growth and the explosive demand for better information access has motivated the interest in QA systems [1][2].

QA considered a technology to find the target answer in large documents to the questions posed in natural language. QA is a research area combines research from different, but related, fields which are Information Retrieval (IR), Information Extraction (IE) and Natural Language Processing (NLP). Regardless of system architecture, or whether the system is operating over a closed text collection or the web, most QA systems are fed with the questions in natural language and output is either the proper answer recognized in a text or small text crumbs including the answer [3].

The QA System is a complex system that composed of many components which may interact with each other, and each component task has many challenges for building and for evaluating it. TREC (Text Retrieval Conference) and CLEF (Cross-Language Evaluation Forum) are main international competitions helping the researchers in this area to compare their systems [10], another workshop namely NTCIR also evaluates QA systems.

In a QA the user can ask a question immediately in NL to the system without any condition to have query syntax. The system allows answering the question in the form of extracting the exact answer from the documents. On the other hand, in IR, the input query is defined in the query language of the search engine such as Yahoo, Google or MSN. The output includes a ranked list of the documents supposedly containing the possible answers; the user then is responsible for reading the texts and find out relevant answers [4].

2. Process Model

The QA system architecture illustrated in Figure 1. From a general viewpoint, the system composed of the following components [5][9]: Question processing, Document processing, Answer processing and Ontology/Knowledge Base.

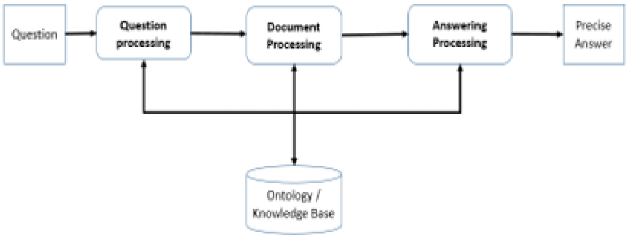


Fig. 1 QA system architecture

A) Question Processing

QA system first must understand what the question is asking about, the most critical step for a question to be answered correctly. Most QA systems depend on Natural Language Processing (NLP) tools that perform linguistic analysis to help in understanding the user’s query and matching (searching) sections in documents. That is an

Manuscript received June 5, 2018
Manuscript revised June 20, 2018

essential task of question and document processing. The most common NLP system contains: tokenization, part-of-speech tagging (POS), stemming, named entity recognition (NER), semantic relations, dictionaries, WordNet, etc. [6].
Question processing unit examines a question entirely and assigns labels to the question according to its likely answer type through a process. This process is called question classification. Also, new semantically equivalent keywords are added to the question to enhance the probability of retrieving relevant documents. That is called query expansion (Reformulation).

B) Document Processing

In document processing unit usually, compute the similarity between the expanded query and the documents to determine the relevancy of the documents and retrieve passages with the highest probability of containing the answer. IR system recall is significant for question answering because if no correct answers are present in a document, no further processing could be carried out to find an answer, Precision and ranking of candidate passages can also affect question answering performance in the IR phase.

C) Answer Processing

Answer processing module is an essential component in QAS that creates the correct answer from the passages of text. At first, it extracts and produces candidate answers from the paragraphs and then assigns them some ranks according to some functions (SVM, KNN, Base). It is a process to select an appropriate answer from the available collection of answers considering the constraints of the Question Analysis unit.

D) Ontology/Knowledge Base

In QA systems two types of search are available namely keywords-based search and semantic search. Typical search engines are working under the keyword-based searching concept. But sometimes, there is a problem of getting the wrong answer for a different meaning of the same word. So, semantic search is used to resolve the keywords-based search problem. Semantic search is used to advance the accuracy of search by explaining the intention of the user and the meaning of the words in the searching sentence. Semantic Search uses semantics to construct highly relevant searching results. Any semantic search technique can be used to retrieve the knowledge from the data source like ontology [7].
QA system works to find answers among knowledge-based data sources, using ontologies for query expansion, utilizing ontologies to analyze the question and create queries to find answers in a knowledge-based data sources.

3. Open domain QA System Architecture

Figure 2 displays the architecture of our system. Each unit in the system described in further section.

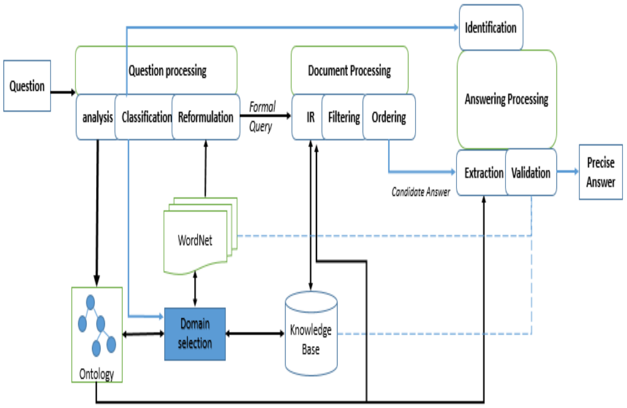


Fig. 2 Open domain QA System architecture

Part 1: Question processing

1. Question Analysis

Question analysis involves identifying and analyzing the structure of words, analysis of words in the sentence for grammar and ordering words in a manner that shows the relationship between the words. The results of this unit are:

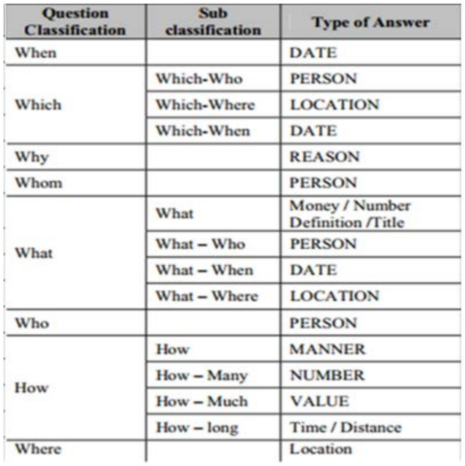
1. Extraction of the question keywords
2. Extraction of the question named entities
3. Extraction of the relation between entities used part of speech(POS) keywords

NER is one of the important operations in Question analysis unit that seeks to identify and classify named entities in text retrieved (questions, passages or other) into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. [6].

2. Question Classification

used to find the type of question and answers, because knowing the type of a question can provide constraints on what constitutes relevant data (the answer), which helps other modules to locate and verify an answer correctly. The question classified according to the entity type like who for answer type person for more cases show Table1 [15]. Questions can also be classified based on algorithms such as Support Vector Machine [16]. The type of a question may provide constraints on domain selection.

Table 1: Question classification



Question Classification	Sub classification	Type of Answer
When		DATE
Which	Which-Who	PERSON
	Which-Where	LOCATION
	Which-When	DATE
Why		REASON
Whom		PERSON
What	What	Money / Number Definition /Title
	What – Who	PERSON
	What – When	DATE
	What – Where	LOCATION
Who		PERSON
How	How	MANNER
	How – Many	NUMBER
	How – Much	VALUE
	How – long	Time / Distance
Where		Location

3. Question reformulation

The input question will be reformulated to another form using the set of rules and list of keywords to fit of the internal query (formal query of structural Database) which acts as input to a document retrieval engine.
WordNet (lexical database containing all the words that are in related domains and used to search the type of words and their synonyms [12][17]) is an important part because of support the re-formulation process and to ensure the validity extract the answer in the last stage.

Part 2: Document Processing

IR searches for the documents (stored in the knowledge base) based on the set of keywords in the reformulated query and ontology to retrieve accurate results in response to a query submitted by the user, and to rank these results according to their relevancy.

Part 3: Reasoner section

1. Ontology Construction:

Ontology is a formal representation of knowledge (questions and answers) by a set of concepts within a domain and the relation between this concepts. Ontology represented via Classes, Relations, and Instances. Because all the results of the previous step are available then can construct ontology using the set of rules. Ontology created for

1. Semantic analysis of the questions and answers
2. Domain selection
3. Information retrieval

2. Domain selection

The primary task of domain selection is enriching the query through separating the initial request into several more precise queries and adding information the query to restrict the searching space. [11]
Depending on the previous sections (I and II), it had proved that the main factors support to select the domain are:

1. Ontology Construction for input question
2. Question Classification
3. WordNet maps various semantically related questions [13]

Part 4: Answer Processing

The answer type which determined during question classification depends on question type. And check the status of passage answer related to question type or not. If not, it is necessary to rely on a parser to recognize named entities. Also, using a part-of-speech tagger can help to enable recognition of answer candidates within identified paragraphs. The extraction of the answer and its validation based on a set of heuristics [14].
Extract the information satisfying the reformulated question from the retrieval passage answer depending on Ontology. Answer ranking based on keywords distance and rate of keywords in the answer. Finally, the confidence in the validity of an answer can increase through WordNet and specific knowledge base.

4. Experimental result

The proposed model tests and validates on 204 questions that selected randomly from Yahoo Non-Factoid Question Dataset, TREC 2007 Question Answering Data, and a Wikipedia dataset that was collected by Carnegie Mellon University cooperate with the University of Pittsburgh between 2008 and 2010. [15]
Different metrics have used over the years, but the current measurement is merely the percentage of questions correctly answered. Table 2 shown below contains the Experimental results that retrieved from the close domain using semantic search, suggested model without using reasoner section and proposed model using reasoner section.

Table 2: Experimental results

	percentage of the correct answer in WH Question	percentage of the correct answer in Y/N Question
close-domain using semantic search	86.4%	88.2%
open-domain without using reasoner	78.6%	79.9%
open-domain using reasoner	84.7%	82.6%

Although open domain QA gave less accurate results than close domain QA, at the same time these results showed a definite and robust improvement on the performance of the system after the use of reasoner section.

5. Conclusion

In recent years, there are many studies on question answering systems that interact with the user. Therefore, this study, we concentrate on the fact that can make domain selection for open domain QA system after taking the input question. We proposed an open domain QA system architecture depends on the essential elements for any QA system and using WordNet, ontology maps and question classification for domain selection. Also, we have seen the most prominent units that need these components to improve system performance such as ontology using in IR part.

Reference

[1] Baeza, R., and Ribeiro, B., Modern Information Retrieval. ACM Press, New York, Addison-Wesley, 1999.

[2] Burger, J. et alii. Issues, tasks, and program structures to roadmap research in question & answering (q&a), in NIST, 2002.

[3] Hirschman, L. and R. Gaizauskas. 2001. Natural language question answering: The view from here. Journal of Natural Language Engineering, Special Issue on Question Answering, Fall–Winter.

[4] Al Chalabi, Hani Maluf. "Question Processing for Arabic Question Answering System." (2015).

[5] M. R. Kangavari, S. Ghandchi, and M. Golpour, "Information retrieval: improving question answering systems by query reformulation and answer validation," World Academy of Science, Engineering and Technology, vol. 48, pp. 303-310, 2008.

[6] M. Shaheen and A. M. Ezzeldin, "Arabic Question Answering: Systems, Resources, Tools, and Future Trends," Arabian Journal for Science and Engineering, pp. 1-24, 2014.

[7] S. Kalaivani and K. Duraiswamy, "Comparison of Question Answering Systems Based on Ontology and Semantic Web in Different Environment," Journal of Computer Science, vol. 8, 2012.

[8] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," ed: Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, 2001.

[9] Asiaee, Amir Hosein. "A framework for ontology-based question answering with application to parasite data." (2013).

[10] Y. Benajiba and P. Rosso, "Arabic question answering," Diploma of advanced studies.
Technical University of Valencia, Spain, 2007.

[11] González-Bernal, J. A., et al. "Natural language dialogue system for information retrieval." Proceedings of the international workshop on research and development of human communication technologies for conversational interaction and learning, Puebla, Mexico. 2002.

[12] B. Magnini and G. Cavagli_a. 2000. Integrating subject field codes into WordNet. In Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, June 2000, pp. 1413- 1418.

[13] Moldovan, Dan, Marius Pasca, and Mihai Surdeanu. "Some Advanced Features Of Cc's Poweranswer." Advances in Open Domain Question Answering. Springer Netherlands, 2008. 3-34.

[14] Moldovan, D., 1999. Lasso: A Tool for Surfing the Answer Net. In Proceedings of the Eighth Text Retrieval Conference (TREC-8).

[15] Smith, N.A., Heilman, M., Hwa, R.: Question generation as a competitive undergraduate course project. In: Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge (2008)

[16] D. Zhang, W. S. Lee, “Question classification using support vector machines”, Proceedings of SIGIR, 2003.

[17] Kolte, S. G., and S. G. Bhirud. "WordNet: a knowledge source for word sense disambiguation." International Journal of Recent Trends in Engineering 2.4 (2009).