Journal of Engineering and Applied Sciences 14 (22): 8289-8292, 2019
ISSN: 1816-949X

An Efficient Semantic Analysis Technique for the Question Answering Systems

[1]Ibrahim Mahmoud Ibrahim Alturani and [2]Mohd Pouzi Bin Hamzah
[1]Department of Applied Science, Ajloun College, Al-Balqa Applied University, Ajloun, Jordan
[2]School of Informatics and Applied Mathematics (PPIMG), Universiti Malaysia Terengganu, Terengganu, Malaysia

Abstract: Question Answering (QA) systems provide a natural way of requesting specific and concise information from a given data source. A crucial stage of such a system is the information retrieval stage which retrieves the possible passages based on their relevance to the question. Accordingly, this study introduces an approach of knowledge extraction of information retrieval from these corpus based on Conceptual Graph (CG). This study discusses how to enhance the accuracy of text-based QA system by modeling the knowledge automatically by using CGF and answers question semantically. The proposed approach showed efficient results of information retrieval measurement through compare recall and precision with another traditional method that has been applied in this same test collection. The result of experiments produced a score of 0.919 for precision and 0.853 for recall. We evaluate our model and show that for the answering task it performs better than standard QA Model.

Key words: Information retrieval, question answering systems, concepts extraction, keywords search, semantic search, conceptual graphs

INTRODUCTION

A QA system aims at taking a user's question in Natural Language (NL) as the input. Then, some analysis is done on the question, to find out what is being askedfor. The user of a question answering system is interested in a concise, understandable and correct answer which may refer to a word, sentence or paragraph. In contrast with classical Information Retrieval (IR) where complete documents are considered relevant to the information request (Kolomiyets and Moens, 2011).

The IRQA system (also, called text_based QA system) is organized as a set of documents from which it attempts to make a match between question and answer. Despite several text_based QA systems have been developed which can carry out the processing required to produce accepted answers. There are many problem that still make textbased QA systems, so, challenge. The main problem in textbased QA system is how to improve the answer accuracy. Hence, in this study, the new approach is proposed to model knowledge with employ conceptual graphs to design a retrieval model for an answer that uses both its structure and its content. Semantic network (called concept network) is a graphic notation for representing knowledge in forms of interconnected nodes and arcs. Semantic network at the level of ontology expresses vocabulary that is helpful primarily for human and usable for machine processing (Salem and Alfonse, 2008).

Question answering architecture: The general architecture of a QA system consists of three components as Fig. 1 (Allam and Haggag, 2012; Alturani and Hamzah, 2018) question processing, document processing and

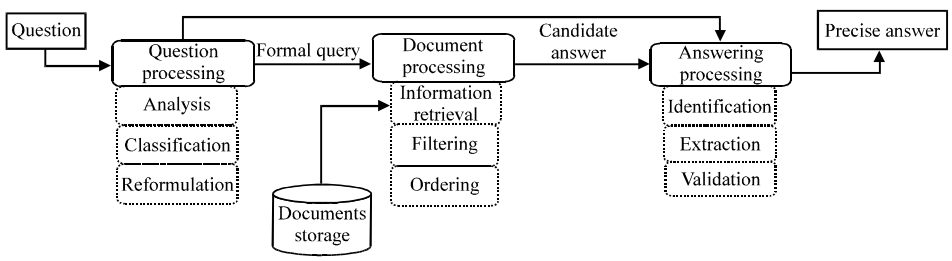

Fig. 1: General architecture of a question answering system

Corresponding Author: Ibrahim Mahmoud Ibrahim Alturani, Department of Applied Science, Ajloun College, Al-Balqa Applied University, Ajloun, Jordan

answer processing. The question processing component receives the question form user in the form of NL. Then, it runs several processes to determine the question type and to build a computational query. The query is passed to the document processing component where the relevant data is extracted and candidate answers are retrieved. The final step is to rank the answers and determine the precise answers based on the answer type.

Most researchers in the QA field were somehow heterogeneous with respect to their system architecture, approaches, scope, evaluation metrics, etc. But on the other hand researchers were mainly concerned with one or more of the three essential components of QA systems: question classification, document retrieval and answer extraction which merge IR with information extraction methods to identify a set of likely set of candidates and then to produce the final answers using some ranking scheme (Ko *et al*., 2010).

MATERIALS AND METHODS

Proposed system: The proposed model interested in finding the most relevant sentence that contains the answer to any given question, from a set of candidate sentences. So, it needs a mechanism that can measure how close a candidate answer is to the question. To achieve this, we look at the problem at two levels. First, need a representation of the sentences that captures useful information in order to accommodate the matching process. Second, need an efficient matching process that can utilize the chosen representation. In this sense, we propose a text-based QA system that obtains its knowledge from documents and answers question semantically. To achieve that this study highlights improvements in the design of a retrieval model in the text-based QA system. Figure 2 displays the architecture of our proposed model.

The proposed model is comprised of several stages, each one made up of one or several resources. In this approach, we use the concepts extraction documents to represent them as a semantic net and then in user question will extract the question terms, at last, the system able to compute the similarity between the terms of the question with semantic net concepts to retrieve the most relevant passages. It contains four main stages starts from preprocessing, extract the concepts, building the conceptual graph and keywords/semantic search each stage in the system described in the further section.

Preprocessing stage: Text pre-processing is a primary part of any NL processing applications like IR. The input of this stage is a text of the standard NL. It used to make a text more understandable and readable by a computer. Text pre-processing includes the following tasks (Indurkhya and Damerau, 2010).

Tokenization is the process of separating a plain text into sentences, words, symbols or other meaningful elements called tokens. The list of tokens becomes input for different processing such as parsing or text mining.

Normalization: Applying some linguistic models to tokens of text, it is the process of transforming text into a single canonical form.

Stops words removal: This process is to remove the meaningless words that recur very frequently. Stop words

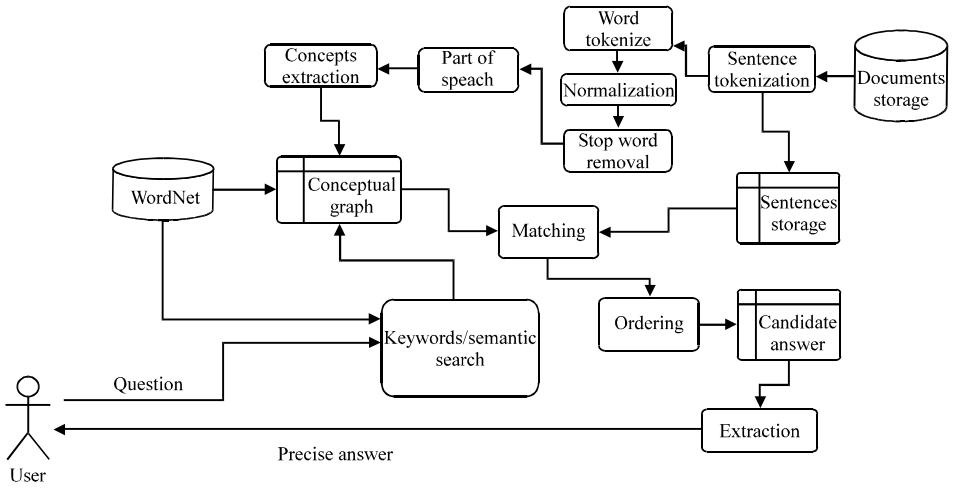

Fig. 2: Architecture of proposed approach

are used to join words together in a sentence. They are very frequently used common words like 'and', 'are', 'this', etc.

Concepts extraction: At each text, there are number keywords which provide important information about the content of that text (Coursey *et al*., 2009). And the correct choosing of keywords for a text plays a vital role in the right representing of that text. The proposed model uses NL processing techniques to analyze the contents of a document works as follow:

Part-of-Speech tagging (POS) is the process of marking up a word in a text to a corresponding part of a speech tag, based on its context and definition. It used in building Named Entity Recognitions (NERs) (most named entities are nouns) and extracting relation with adjacent and related words in the sentence. e.g., if the input sentence is 'Elephants are mammals' the output tags is ('Elephants', 'NNS'), ('are', 'VBP'), ('mammals', 'NNS'). Convert the words found in the sentence into concepts using WordNet ontology lexical database of concepts with semantic relations between them which maps each word with one or more concepts. To find out which concept is meant, the system disambiguates these words depending on the context of the sentence.

Construct conceptual graph: Conceptual Graphs (CGs) are a kind of semantic networks proposed by Sowa (1984) by Peirce's existential graphs. A CG is a finite direct graph, each node in the graph is either a concept node or relation node. Concept nodes represent entities, attributes, states and events and relation nodes show how the concepts are interconnected. The proposed model constructs CG based on the following:

Syntactic information (POS and dependency tree) a dependency tree (Biasotti *et al*., 2008) is a directed tree that is different from a phrase structure in representing a sentence. Each node in a dependency tree can be represented by a word from a sentence and each edge in a dependency tree may link two words by a syntactic relation. In contrast, each node in a sentence structure is either a concept or a word from a sentence and each edge in a phrase structure may link two nodes without a syntactic relation between them.

Chunking is the frame that matches the sentence pattern, it implements shallow processing techniques to group together words to larger syntactic and meaning holding components (Cimiano, 2006), for example:

[NP: ('Elephants', 'NNS')≥Subject | VP: ('are', 'VBP')≥Verb | NP: ('mammals', 'NNS')≥Object]

Semantic information (WordNet ontology).

Keyword/semantic search process in QA systems: In QA systems two types of search are available namely shallow keywords-based search and semantic search. Standard search engines are working under the keyword-based searching concept. Use shallow keyword-based expansion techniques to locate interesting sentences from the retrieved documents, based on the presence of words that refer to entities of the same type of the expected answer type. The ranking is based on syntactic features such as word order or similarity to the query. But sometimes, there is a problem of taking the wrong answer for a different meaning of the same word. So, semantic search is used to solve this problem. Semantic search is used to enhance the accuracy of search by understanding the aim of the user and the meaning of the words in the searching sentence. Semantic search uses semantics to produce extremely appropriate searching results. Also, this semantic search technique can be used to retrieve the knowledge from the data source (Kalaivani and Duraiswamy, 2012; Pasca, 2003; Zhang, 2006). In the proposed model when a user asks the system a question, the system creates its CG's as explained previously and then matches each question's CG to each sentence's CG in the same domains of the question to extract the exact answer from the part of the sentence's CG that has been projected under the question target's concept.

Evaluation metrics: The evaluation of QA systems is provided depending on the criteria for judging an answer. The following list takes the most criteria for answer evaluation (Pease *et al*., 2002), relevance, correctness, conciseness (the answer should not contain irrelevant information) and completeness. Based on these criteria, there are three different judgments for an answer extracted from a document, correct, inexact or unsupported. Different metrics have applied over the years but the following measures are the most commonly used measures that are typically utilized for evaluation:

$$\text{Precision} = \frac{\text{No. of correct answers}}{\text{No. of questions answered}}$$

$$\text{Recall} = \frac{\text{No. of correct answers}}{\text{No. of questions to be answered}}$$

RESULTS AND DISCUSSION

The proposed model tests and validates on 110 questions that selected from a sample of 5,000 that was obtained from the three datasets; Yahoo non-factoid question dataset, TREC 2007 question answering data and a Wikipedia dataset that was generated by Smith *et al*.

Table 1: Experimental results

Types of model	Measures	
	Precision	Recall
Proposed model	0.919	0.853
Standard QA Model (using shallow keywords based search)	0.846	0.801

(2008). Also, at the same time, we have applied the shallow keywords-based search in the same sample of questions. Table 1 shown contains the experimental results that retrieved from the proposed model using semantic net and standard QA approaches.

CONCLUSION

QA system still faces many challenges which rely on determining the appropriate answer source for answering the question posed. So, to solve these challenges should enhance the knowledge representation. One of these suggested solutions to improve the accuracy of QA system is a semantic net. A semantic net helps QA systems to access the target answer domain.

This study showed the semantic net to improve the QA system from different resources. The experimental results of 110 questions showed the degree of the performance of the proposed model. The proposed model showed an efficient result of information retrieval measurement through compare recall and precision with standard QA Model that has been applied in this same test collection.

REFERENCES

Allam, A.M.N. and M.H. Haggag, 2012. The question answering systems: A survey. Intl. J. Res. Rev. Inf. Sci., 2: 1-12.

Alturani, I.M.I. and M.P.B. Hamzah, 2018. A new approach for open-domain question answering system. Intl. J. Comput. Sci. Network Secur., 18: 100-103.

Biasotti, S., L. De Floriani, B. Falcidieno, P. Frosini and D. Giorgi *et al*., 2008. Describing shapes by geometrical-topological properties of real functions. ACM. Comput. Surv., 40: 1-87.

Cimiano, P., 2006. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Vol. 27, Springer, Berlin, Germany, ISBN-13: 978-0-387-39252-3,.

Coursey, K.H., R. Mihalcea and W.E. Moen, 2009. Automatic keyword extraction for learning object repositories. Proc. Am. Soc. Inf. Sci. Technol., 45: 1-10.

Indurkhya, N. and F.J. Damerau, 2010. Handbook of Natural Language Processing. 2nd Edn., CRC Press, Boca Raton, Florida, ISBN-13:978-1-4200-5893-8, Pages: 679.

Kalaivani, S. and K. Duraiswamy, 2012. Comparison of question answering systems based on ontology and semantic web in different environment. J. Comput. Sci., 8: 1407-1413.

Ko, J., L. Si and E. Nyberg, 2010. Combining evidence with a probabilistic framework for answer ranking and answer merging in question answering. Inf. Process. Manage., 46: 541-554.

Kolomiyets, O. and M.F. Moens, 2011. A survey on question answering technology from an information retrieval perspective. Inf. Sci., 181: 5412-5434.

Pasca, M., 2003. Open-Domain Question Answering from Large Text Collections. 1st Edn., Center for the Study of Language and Information, Street Chicago, Illinois, USA., ISBN-13:978-1575864273, Pages: 157.

Pease, A., I. Niles and J. Li, 2002. The suggested upper merged ontology: A large ontology for the semantic web and its applications. Proceedings of the AAAI-2002 International Workshop on Ontologies and the Semantic Web Vol. 28, July 28-29, 2002, Edmonton Convention Centre, Edmonton, Canada, pp: 7-10..

Salem, A.B.M. and M. Alfonse, 2008. Ontology versus semantic networks for medical knowledge representation. Proceedings of the 12th WSEAS International Conference on Computers, July 23-25, 2008, World Scientific and Engineering Academy and Society (WSEAS), Heraklion, Greece, ISBN:978-960-6766-85-5, pp: 769-774.

Smith, N.A., M. Heilman and R. Hwa, 2008. Question generation as a competitive undergraduate course project. Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge, September 25-26, 2008, Naval Support Facility Arlington, Arlington, Virginia, pp: 4-6.

Sowa, J.F., 1984. Conceptual Structures: Information Processing in Mind and Machine. Addison Wesley, UK.

Zhang, A., 2006. Research and implementation of ontology-based intelligent question answer system. Comput. Appl. Software, Vol. 5,