The Role of Semantic Analysis in QA Systems

Dr. S. Jayalakshmi [#1] and Dr. V. Jayalakshmi[*2]

[#]*Department of Computer Applications, School of Computing Sciences,*
VISTAS, Chennai, India
[1]jayalakshmi.research@gmail.com
[2]jayasekar1996@yhaoo.co.in

Abstract—**The recent advancement of modern information retrieval method has created a high demand for Question Answering (QA) system to answer the question formulated by the user. The QA is an essential service that delivers the adequate sentences as answers to the specific natural language questions. Most of the QA systems often restrict the answer processing only to a syntactic pattern matching. To accurately provide the answers, the system must acquire fine-grained information regarding the question type and context. The Document processing is to generate and rank the patterns significantly for retrieving the lexical, syntactic, and semantic features based relevant sentences. Moreover, it intends to bring out the candidate answer sentences from the relevant documents, wherein the candidate answer sentence refers the most relevant sentences to the question. As the WH-operator answer processing methods are used to identify the precise answer from the candidate answer sentences.**

Keywords— **Question Processing, Answer Processing, Document Processing, Semantic, Lexical, Syntactic.**

I. INTRODUCTION

The dramatically pervading World Wide Web (WWW) is a globally interconnected system that comprises over 130 million domains and billion unique URLs with numerous internet users [1]. The exponential growth of the online information on WWW, the people, predominantly demands the sophisticated search tools to obtain significant Information retrieval (IR). The IR is the process of acquiring the information relevant to the user needs. The IR search tools allow the users to search full-text as well as short answers [2]. The IR tools are unable to explicitly deliver the information as the concise answers to the questions of the users. In contrast, the Question Answering (QA) [3,4] provides the specific and concise answers to the questions posed by the users in natural language. The ultimate goal of question answering is to interpret a set of documents in a concise and comprehensive form that is enough to provide the answer to the user query. Natural Language Processing (NLP) enables the IR systems to shift its input from natural language questions to the keywords. It also shrinks the answer from a set of related documents into concise answers. QA systems [5] have inherited the techniques from the several underlying technologies, namely machine learning, IR and NLP to automatically retrieve the precise answers based on the question key terms. The syntactic and semantic structure representation has been employed to represent the original representation of the question [6]. To satisfy the user requirements with ease of understanding, the QA system performs the several steps such as extracting the keywords from a natural language query, retrieving the query relevant information from sources, extracting and prioritizing the retrieved answers, and presenting the response as the answer in an efficient manner.

A. *Objectives And Information Retrieval Components of QA Systems*

The main objective of QA system is to determine the user Questions, relevant document, develop the algorithm, irrelevent content deduction suggest the accurate answer to given query. The IR plays a vital role in QA to retrieve the full-text and filter the irrelevant documents to speed-up the answer generation of the data source documents.

Full-text Retrieval: The IR retrieves the free-text documents based on the keyword of the queries submitted by the users. Normally, IR systems measure a numeric score for each retrieved data based on the input query and then, rank the retrieved documents to reveal the most relevant document to the user the determines whether the answer contains the appropriate answer of the question or not, NLP analyzes the retrieved passages in the source documents.

Filtering: Answer generation process often deals with the underlying problem of the incorrect answer generation due to the existence of mismatching between the NLP and IR component. NLP primitives such as lexical and semantic constraints of the question build the sophisticated representation of the information need and facilitate the system to generate the precise answer to the input question. IR component filters the irrelevant documents and considers the retrieved document as an answer which has precisely matched with the input question.

IR based QA system: To effectively provide the answer to the question, the QA system focuses on the NLP components and also significantly concentrates on the IR component. The IR component reduces the burden of the system by shrinking the search space around the relevant text information alone. The goal of an IR is to extract the documents based on the input question.

Answer extraction from relevant documents: The IR component initially retrieves the relevant documents from the collection of text data. It extracts and ranks the whole documents based on the coordinate matching, synonyms, stemming, proximity of words, and the order of words, significant words, and capitalization and quoted words in a question. Answer extraction and formulation is based on the heuristic pattern matching between the question and document collections.

II. THE ROLE OF SEMANTIC IN QA SYSTEM

In QA system, semantic analysis plays a crucial role that identifies the complex semantic structures of the questions. The QA has three cases such as recognizing the expected answer type based on the semantic structure, identifying the question class, and the topic of the information to be retrieved. Semantics [7] is the most complex and an essential factor for natural language. Hence, tagging the semantic elements along with the word text of a question is crucial in natural language processing based answering system. The semantic analysis is based on the different semantic elements such as ontology classes, WordNetsynsets, FrameNet frames, and semantic roles. The NLP task of the information extraction employs the Semantic Role Labeling (SRL) to understand the input question [8] precisely.

QA has the advantage of the SRL improving the accuracy when exploiting SRL based NLP. The argument-predicate relationship based question and document analysis precisely improve the performance of the system. Semantic relational information [9] has been used to transform the input natural language questions into enhanced queries to determine the precise answers. The shallow semantic parser creates the semantic roles on the matching questions and answers. A FrameNet-style based semantic role information develops an answer extraction while considering the answer extraction as semantic relation based graph matching problem [10]

A. *Semantic Analysis of QA Processes*

The QA framework incorporates three major processes such as question processing, document processing, and answer processing. Question classification is the initial stage of the question processing [11].

In Question processing: The primary objective of question processing is to understand the questions of the users by analyzing the questions through semantic search [12]. It involves in determining the type of the question, deducing the expected answer type, recognizing the focus of the question, and formulating the query in the collection of documents based on the semantically represented keywords of the queries.

***In Document processing*:** It is the function of semantically-relevant document retrieval from the data sources. Even though it does not find the actual answers to the question, it determines the documents that contain the answer. This semantically-relevant document retrieval is the perfecting method for the answer generation due to the need of additional specific information than the traditional IR. Passage-based retrieval facilitates the system to identify the specific exact answer rather than exploiting the full document retrieval

In Answer processing: It depends on the results from the question and document processing steps. Answer extraction is based on the complexity of the question, expected answer type, retrieved semantically relevant documents, and context of the question. QA systems typically extract the answers based on the entities in the retrieved passages to match the entity with the expected answer type of the question to predict the exact answer.

B. *Types Of Questions And QA*

The types of user questions have an impact on the task of answer generation. Mostly, the errors happen in QA systems are due to misclassification of questions [13]. The QA systems classify the questions based on types of answers asked by users.

Factoid Type Questions

The factoid questions based QA system can be divided into two classes such as knowledge intensive and data intensive. The knowledge-intensive method highly relies on the meta-form of both the questions and answers. The factoid type of questions includes the questions starting with what, when, which, who, and how. Current research works can attain better performance in answering factoid type questions [14].

List and Hypothetical Type Questions

This kind of questions starts as 'what would happen' along with the 'if' condition.

Causal and Confirmation Questions

These kinds of questions are used for asking the explanations, reasons, and elaborations. The major challenge in the casual questions is to answer the term 'why.' Because the answers for why questions range from a sentence to a paragraph. Why Fu?, Why took lecture? andWhy in class k?

III. CHALLENGES IN QA SYSTEMS

The conventional QA systems address a variety of complex natural language question types. [5]. Moreover, the four fundamental challenges in QA systems are discussed as follows:

Indexing the Heterogeneous Sources

IR tools based indexing and retrieving semantic information decides the speed and accuracy of question answering systems. According to the type of data, several indexing techniques have been used in IR systems. The indexing techniques can be categorized into structured, semi-structured, and unstructured data. Most of the conventional QA systems deal with structured or semi-structured or unstructured data and make inverted indices.

Interoperability

It provides the answers with ease of searching the relevant information due to the accessing capability of the domain-specific knowledge source such as ontologies. The open domain QA system focuses on the questions about a specific domain as well as on the entire general topics using global knowledge and general ontologies

A. *Techniques Used in QA System*

The WWW data can be categorized into Data, Information, Knowledge, understanding, and wisdom [15]. The information refers a set of meaningful data and groups the data in a relational connection. Knowledge is a pattern which provides a high level of predictability. It needs to integrate the knowledge of various domains and analytical process ability and the true cognitive ability of human beings.

TABLE 1

QA SYSTEMS CLASSIFICATION BASED ON TECHNIQUES

ASPECT	QUESTION ANSWER SYSTEM BASED ON			
	A. Data Mining	*B.* Information Retrieval	*C.* Natural language	*D.* Knowledge Retrieval
E. Type of Question	*F.* Simple-find *G.*	*H.* Factoid questions-what, where, which, when	*I.* Definition questions	*J.* Hypothetical and confirmation Questions yes–no
K. Type of Answer	*L.* Short	*M.* Combined	*N.* Combined	*O.* Combined
P. Searching	*Q.* Factual Data	*R.* Querying for factual data	*S.* subjective,opinionated or fact	*T.* Searching for precise answers
U. Matching	*V.* Exact	*W.* Relevant	*X.* Relevant	*Y.* Exact
Z. Relevancy	*AA.* Objective	*BB.* Subjective	*CC.* Subjective	*DD.* Subjective
EE. Techniques	*FF.* Syntactic	*GG.* Syntactic and	*HH.* Pattern matching,	*II.* Discourse and pragmatic analysis

		Semantic analysis	syntactic, semantic analysis	
JJ. Knowledge Source	*KK.* Data base *LL.*	*MM.* Syntactic information	*NN.* Syntactic and pragmatic web	*OO.* Semantic and pragmatic web
PP. Models	*QQ.* BOW	*RR.* BOW	*SS.* Bag of Concepts	*TT.* Bag of Knowledge

B. *Evaluation Metrics*

In QA system, measuring the correctness of the answers corresponding to the question is crucial, at least in the perspective of QA campaign. The semi-automatic evaluation identifies the correctness of the answer based on the accurate answer patterns and document identifier of each question. This evaluation is only feasible for a known set of questions.

Precision and Recall: Precision and recall are the evaluation metrics of IR, which are related to the user query and relevant documents. Precision reveals the accuracy and Recall measures the exhaustively. The weighted harmonic mean of both the precision and recall is the F-measure.

Mean Reciprocal Rank: MRR measures the ability of QA system to answer a set of factoid questions.

Confidence Weighted Score: CWS is also defined as the average. It measures the confidence score of each answer to a question.

Non-Interpolated Average Precision: This metric is used for measuring the accuracy of a list of questions or answers at the same time.

IV. CONCLUSION

QA system presents the role and its current challenges in the web searching field. It acknowledges the issues which are often handled in the recent QA works and it contain the processes of the QA systems, including with its question answering techniques, tasks or components and semantic techniques. Also it included the details about QA in which way the natural language questions are processed about the evaluation campaigns and metrics. It deals with various measure are used to evaluate the accuracy of the retrieved text. Huge amount of information is available on the web but retrieving the relevant and most accurate answer is a challenging one, but the semantic similarity and NLP techniques are widely used to resolve these kinds of problems.

REFERENCES

[1] O. Etzioni, "Search needs a shake-up", Nature, Vol.476, No.7358, pp.25–26, 2011

[2] Lu, Wenpeng, Jinyong Cheng, and Qingbo Yang, "Question answering system based on web", IEEE Fifth International Conference on Intelligent Computation Technology and Automation (ICICTA), pp.573-576, 2012

[3] Sanjay K Dwivedia, and VaishaliSinghb, "Research and reviews in question answering system", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), Vol.10, pp.417-424, 2013

[4] Bouziane, A., Bouchiha, D., Doumi, N., and Malki, M. "Question Answering Systems: Survey and Trends", Procedia Computer Science, Vol.73, pp.366-375, 2015

[5] AbdelghaniBouzianea, DjelloulBouchihaa, NoureddineDoumib, and MimounMalkic, "Question Answering Systems: Survey and Trends", International Conference on Advanced Wireless Information and Communication Technologies, Vol.73, pp.366–375, 2015

[6] Silvia Quarteroni, Alessandro Moschitti, Suresh Manandhar, and Roberto Basili, "Advanced Structural Representations forQuestion Classification and Answer Re-ranking", Springer, pp.234–245, 2007

[7] Wang Wei, Payam M. Barnaghi, and AndrzejBargiela, "Search with Meanings: An Overview of Semantic Search Systems", pp.76-82, 2008

[8] Palmer, Martha, Daniel Gildea, and NianwenXue, "Semantic role labeling", Synthesis Lectures on Human Language Technologies, Vol.3, No.1, pp.1-103, 2010

[9] Kaisser, Michael, and Bonnie Webber, "Question answering based on semantic roles", ACM Proceedings of the Workshop on Deep Linguistic Processing, pp.41-48, 2007

[10] Shen Dan, and MirellaLapata, "Using semantic roles to improve question answering", Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.12–21, 2007

[11] BabakLoni, "A Survey of State-of-the-Art Methods on Question Classification", Literature Survey Published on TU Delft Repository, 2011

[12] A. Lally, J. M. Prager, M. C. McCord, B. K. Boguraev, S. Patwardhan, J. Fan, and P. Fodor, "Question Analysis: How Watson Reads a Clue", IEEE IBM Journal of Research and Development, Vol.56, No.3.4, pp.2.1-2.14, 2012

[13] Bu, Fan, Xingwei Zhu, Yu Hao, and Xiaoyan Zhu, "Function-based question classification for general QA", In Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp.1119-1128, 2010

[14] Lopez, Vanessa, Victoria Uren, Marta Sabou, and Enrico Motta, “Is question answering fit for the semantic web?: a survey”, Semantic Web, Vol.2, No.2, pp.125-155, 2011
[15] Cambria, E., White, B., “Jumping NLP curves: a review of natural language processing research”, IEEE Computational Intelligence Magazine, Vol.9, No.2, pp.48–57, 2014