# OPEN DOMAIN QUESTION ANSWERING SYSTEM USING SEMANTIC ROLE LABELING

**Sreelakshmi V[1], Sangeetha Jamal[2]**

[1]*Department of Computer Science, Rajagiri School of Engineering and Technology, Kochi, India*
[2]*Department of Computer Science, Rajagiri School of Engineering and Technology, Kochi, India*

## Abstract

*The World Wide Web is an attractive source of information retrieval and can be used for seeking simple factual answers for user questions. Although usual information retrieval systems like search engines helps in finding the relevant documents for a set of keywords, there are situations where we have more specific information need. Search engines only provide the actual facts as a ranked list of documents and links that contain the keywords. However, what a user really wants is often a precise answer to a question. A Question Answering system gives us a precise solution to this scenario. A Question Answering system retrieves a precise answer to any natural language question posed by the user. A typical QA system uses several language processing techniques. In this paper, an Open Domain Question Answering that answers simple Wh-questions using online search has been proposed.*

***Keywords:*** *Information Retrieval, Question Answering, Semantic Role Labeling*

----------***----------

## 1. INTRODUCTION

The Question Answering task is one of the major research area under Information Retrieval (IR) with applications from Natural Language Processing (NLP). The goal of a Question Answering system is to deliver an exact answer to the arbitrary questions formulated by users. In the process of normal Information Retrieval using search engines, the user is provided with a ranked list of documents and links, but not the exact answer. A Question Answering system addresses this problem. These systems allow a user to ask questions in everyday language and provide the user with a quick and precise answer.

Question Answering systems can be classified into two groups namely closed domain and open domain. Closed domain Question Answering Systems is restricted to a specific domain (like medical, music etc.). They also contain a domain specific ontology. Whereas, an open domain Question Answering system deals with questions about nearly everything and can on universal ontology and information such as World Wide Web.

A typical Question Answering system can be divided into 3 modules namely: Question Processing module, Document Processing or Information Retrieval module and Answer Processing module. Each module contains several sub modules and these modules use several Natural Language Processing Techniques inorder to extract the proper answer.

The usual Question Answering system is designed to answer simple wh-questions like "who", "what", "when", "where", etc. But the recent QA research focuses on extending the system to answer complex questions, summary questions, opinion questions etc. The paper proposes a Question Answering system that answers simple factoid, wh-questions by using a technique called Semantic Role Labeling.

The rest of the paper is organized as follows. The next section describes the general architecture of a Question Answering System. Section 3 discusses some of the related works in this area. The proposed system architecture is described in section 4. The paper concludes with the experimental setup and results.

## 2. QA SYSTEM ARCHITECTURE

A general Question Answering system is composed of three main modules. They are:

- Question Processing
- Information Retrieval or Document Processing
- Answer Processing module

In the Question Processing module, the user query is processed. In this phase, the user question is classified based on the stem word like 'who', 'what', 'when', 'where' etc. and hence the answer type information can be easily obtained from this. For example, the question "Where is the headquarters of WHO" looks for a location as answer. In this phase, the keywords from the questions are also extracted. This phase is crucial because the answer extraction strategy completely depends on the information collected from this module. An optional phase of query reformulation can also be included in this module.

In the Document Processing module, the question is question is passed for Information retrieval. This module uses a search engine for IR and the documents related to the user questions are saved. Then they are ranked based on certain heuristics.

Answer Processing module is the distinguishing feature between usual IR systems and QA systems. In this module, the exact answer for the user question is identified and extracted from the retrieved documents. This phase uses the query information collected from the first phase. Then the answer is validated for correctness and returned to the user.

## 3. RELATED WORKS

Before the emergence of web, Question Answering systems had a limited audience and were mostly domain restricted. The first QA system developed was BASEBALL (Green et al. 1961). It answered domain–specific questions about the baseball games played in American league. Another early work was LUNAR system (Woods et al. 1972) used to access, compare and evaluate the chemical analysis data on lunar rock and soil composition that was accumulating as a result of the Apollo moon mission.

When the web became popular, QA systems became open domain and millions were able to access these systems. START (Katz, 1997) Information Server built at the MIT Artificial intelligence Lab is one of the first web-based QA systems. A new method called Knowledge annotation approach was introduced in the system.

The work in QA system, for the past few years, has been flourished by the inclusion of QA task in the yearly Text Retrieval Conference (TREC), sponsored annually by the U.S. National Institute of Standards and Technology (NIST) and the Cross Language Evaluation Forum (CLEF). These competitions helped to develop and improve the research in this field.

A semantics based QA system accessible via the web was developed bt Hovy et al. The system was named Webclopedia. In the system, the parsed question is expanded and converted to multi unit words and related documents are retrieved. Answers are selected using machine learning based grammar parser.

Radev et al. Proposed a method called Probabilistic Phrase Reranking (PPR) in their web-based question answering system (NLQA). The method was fully implemented at the University of Michigan as a Web-accessible system, called NSIR which relies on an existing search engine to return documents that are likely to contain the answer to a user question.

A general-purpose, fully automated question-answering system available on the web was MULDER developed by Kwok et al. MULDERs architecture relies on multiple search engine queries, natural-language parsing, and a novel voting procedure to yield reliable answers coupled with high recall.

Harabagiu et al proposed a system called FALCON that was presented in TREC-9. The system used a method called boosting, which integrates different forms of syntactic, semantic and pragmatic knowledge for the goal of achieving better performance. The answer engine handles question reformulations, finds the expected answer type from a large hierarchy using WordNet semantics and extracts the answers after performing unifications on the semantic forms of the question and its candidate answers. FALCON also used Named Entity recognition technique and the answers were extracted from semantically rich passages that match the question type.

Dan Moldovan et al. proposed LASSO, A Tool for Surfing the Answer Net in TREC 8. In order to find answers in large collections of documents, novel language processing techniques were included in the system. The question was processed by combining syntactic information with semantic information that characterized the question in which eight heuristic rules were defined to extract the keywords used for identifying the answer.

Recent research in QA systems use the ontology based methods to improve the efficiency. Current works in QA are based on semantic web. Bollegala et al. proposed an automatic method to estimate the semantic similarity between words or entities using web.

Semantic similarity is computed by mapping terms to an ontology and by examining their relationships in that ontology. For that, ontologies like Wordnet are used. The similarity is measured by page counts and by analyzing the retrieved snippets. Coefficients like WebJaccard ,Web PMI, WebOverlap etc. can be used to find the semantic similarity between question answer pair, and hence improve the performance.

## 4. PROPOSED SYSTEM

A simple Question Answering system was implemented using the technique called Semantic Role Labeling. The system consisted of three phases called Query Processing, Document Processing, Answer Processing. In each of these phases, several language processing components were used. The system is web based and hence uses the search engine to extract information from the web.

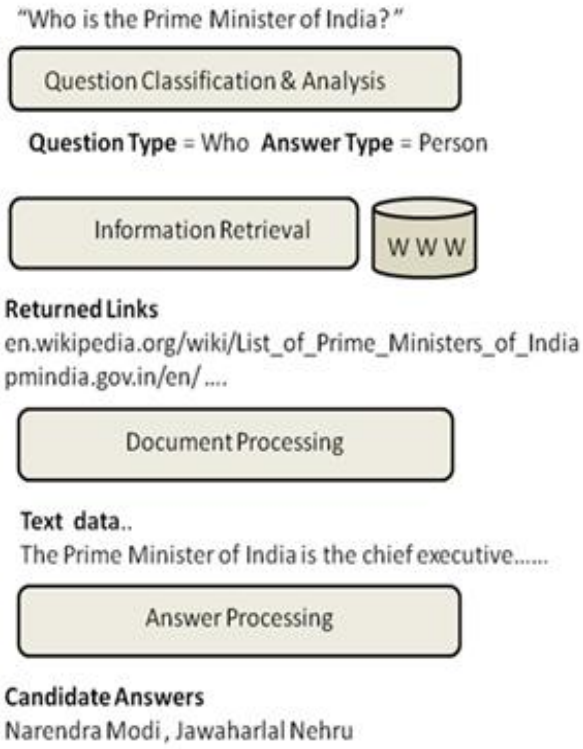


**Fig -1:** Proposed QA System Working Example

The first phase is to identify the keywords from the user question. In this phase, the user question is analyzed and classified based on the stem word. From the question stem word, The expected answer type information can be obtained.

The next phase is Information Retrieval using search engine. The user question is passed on to an online search engine and the documents related to query are retrieved. These documents undergoes some linguistic techniques like POS tagging, Parsing etc.

The third and final phase is the Answer Processing phase, which distingushes question answering systems from the usual text retrieval systems. The proper answer for the user question is identified and extracted from the retrieved documents. The query information collected from the first phase can be used for validating the answer. For identifying the exact answer, a technique called semantic role labeling is been used.

### 4.1 Semantic Role Labeling

The technique called SRL or Semantic Role Labeling is used to label the semantic roles in a sentence. Also known as Shallow Semantic Parsing, Case Role Analysis, Thematic Analysis etc., this technique has applications in the research areas of Question Answering systems, Text summarization, Machine Translation etc. The goal is to identify all the constituents that fill a semantic role, i.e. to determine the roles like Agent, Patient, Instrument, Location, Time etc. in a sentence.

SRL can be used for identifying the exact answer for a user question. For example, 'Who' question uses Agents or Actors i.e. Persons as the answer. Similarly, 'Where' question requires a Location as answer and 'When' question uses Time or Temporal tags as answer. By combining these information, a simple factoid Question Answering system can be implemented.

## 5. RESULT

A simple QA system that answers simple factoid questions was implemented in Python. For the purpose of Semantic Role Labelling, SENNA [20] tool was used. The information retrieval was done using Google search engine.

The proposed system performance was compared with a QA system that used simple pattern matching technique for answer extraction. Results shows that the proposed system that uses SRL has got more accuracy than a normal pattern matching QA system. The results can be seen in the following table.

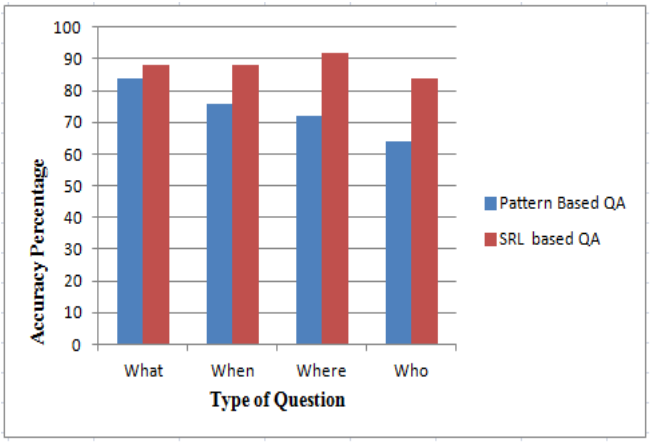


**Chart -1:** Result Comparison

The accuracy was calculated as:

Accuracy = (Correct Answer / Total Questions)

The system was tested for 4 types of questions namely 'who', 'what', 'when' and 'where'. The result was classified into two types. If the system provides the necessary information, then it was classified as a correct answer. And if the retrieved result does not contain the necessary information, or if the answer was not produced, then it is classified as a wrong answer. The answer size varied from a word , a few words, to a sentence based on the question type. It can be observed that the QA system accuracy has improved by using the semantic role labeling technique.

## 6. CONCLUSION

The wealth of information on the web makes it an attractive source for seeking information. Retrieval of accurate information like quick answers to simple factual questions from web is done with the help of a QA system. A Question Answering system provides the user a precise answer for the wh-question. The question Answering systems helps the

user to get a quick and precise answer for simple wh-questions. For this purpose, QA systems heavily use language processing components like Parser, Tagger etc. In this paper, a simple QA system was implemented using a technique called Semantic Role Labeling and the performance of the system is compared with a QA system that uses pattern matching. As a part of future enhancement, the system can be extended for answering complex questions.

## REFERENCES

[1]. M Ramprasath, S Hariharan Improved Question Answering System by semantic reformulation, IEEE- Fourth International Conference on Advanced Computing, 2012.
[2]. Ali Mohamed Nabil Allam, and Mohamed Hassan Haggag, The Question Answering Systems: A Survey, International Journal of Research and Reviews in Information Sciences (IJRRIS), September 2012 Science Academy Publisher, United Kingdom.
[3]. B.F. Green , A.K. Wolf, C. Chomsky, and K. Laughery, Baseball: An automatic question answerer, In Proceedings Western Computing Conference, 1961.
[4]. Woods W., Progress in natural language understanding an application to lunar geology, American Federation of Information Processing Societies (AFIPS) Conference Proceedings 1973.
[5]. Hovy E., Gerber L., Hermjakob U., Junk M., and Lin C.Y., Question answering in webclopedia ; The Ninth Text REtrieval Conference (TREC 9), 2002.
[6]. Moldovan D., Lasso: A Tool for Surfing the Answer Net, In Proceedings of the Eighth Text Retrieval Conference (TREC-8).
[7]. Text Retrieval Conference (TREC), http://trec.nist.gov
[8]. Cross Language Evaluation Forum (CLEF) (http://www.clefcampaign.org/).
[9]. Deepak Ravichandran, Eduard Hovy, Learning Surface Text Patterns for a Question Answering System , In Proceedings of the ACL Conference, 2002.
[10]. Soubbotin M. M., Patterns of potential answer expressions as clues to the right answers, In NIST Special Publication, The Tenth Text Retrieval Conference TREC-10, Gaithersburg 2002.
[11]. L. Hirschman, R. Gaizauskas, Natural language question answering: the view from here, Natural Language Engineering 7, 2001 Cambridge University Press.
[12]. Marco De Boni, Suresh Manandhar, Implementing clarification dialogues in open domain question answering, Natural Language Engineering, Pages 343-361, 01/2005.
[13]. Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka , A Web Search Engine-Based Approach to Measure Semantic Similarity between Word , IEEE Transactions On Knowledge And Data Engineering, July 2011.
[14]. Abney S., Collins M., and Singhal A., Answer extraction , In the Proceedings of ANLP 2000.
[15]. A Harabagiu, Dan Moldovan, Marius Paaca, Rada Mihalcea, Mihai Surdeanu, Rzvan Bunescu, Roxana Girju, Vasile Rus, Paul Morarescu, FALCON: Boosting Knowledge for Answer Engines , In NIST Special Publication 500-249:The Ninth Text REtrieval Conference (TREC 9)
[16]. Kwok C., Etzioni O., and Weld D. S., "Scaling question answering to the web", In the Proceedings of the 10th World Wide Web Conference (WWW 2001), Hong Kong.
[17]. Kangavari M., Ghandchi S. & Golpour M, Information Retrieval: Improving Question Answering Systems by Query Reformulation and Answer Validation, World Academy of Science, Engineering and Technology, 2008.
[18]. Boris Katz,From Sentence Processing to Information Access on the World Wide Web , AAAI Technical Report, 1997
[19]. Sreelakshmi V, Sangeetha Jamal, Web Based Question Answering System using Pattern Matching , in International Conference on Information Science - ICIS, July 2014, College of Engineering Cherthala.
[20]. Ronan Collobert, Weston , L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. , Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research (JMLR), 2011.
[21]. R. Collobert, J. R. Collobert. Deep Learning for Efficient Discriminative Parsing, in International Conference on Artificial Intelligence and Statistics (AISTATS), 2011.
[22]. D. Gildea, D. Jurafsky, Automatic labeling of semantic roles , Computational Linguistics.
[23]. S. Narayanan, Harabagiu, Question Answering on Semantic Structures, In Proceedings of the 19th COLING.
[24]. Sreelakshmi V, Sangeetha Jamal, Survey Paper : Question Answering Systems, in National Conference on Computing and Communication - (NCCC ), March 2014, GEC Idukki.
[25]. Poonam Gupta, Vishal Gupta, "A Survey of Text Question Answering Techniques", International Journal of Computer Applications, Volume 53, September 2012.
[26]. SENNA SRL Tool[Online], Available : http://ml.nec-labs.com/senna/
[27]. Li X. , Roth D. , "Learning question classifiers" , In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), p.556-562.
[28]. Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, Amardeep Grewal, "Probabilistic Question Answering on the Web", Proceedings of the 11th international conference on World Wide Web, ACM 2002

## BIOGRAPHIES

Sreelakshmi V is a postgraduate engineering student in Computer Science and Engineering-Information Systems at Rajagiri School of Engineering and Technology, Kochi, India. She completed her graduation in Computer Science and Engineering from the University of Calicut.

Ms. Sangeetha Jamal is currently working as Assistant Professor in the Department of Computer Science and Engg., RSET, Cochin, Kerala. She has completed her M.Tech Degree in Software Engineering from Cochin University of Science & Technology, and her B.Tech Degree in Computer Science & Engg from the University of Calicut. She is currently doing her Ph.D. in Natural Language Processing.