International Journal of Pure and Applied Mathematics
Volume 119 No. 10 2108, 131-137
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Effective Information Retrieval through Question –Answering using NLP Techniques.

*S.Jayalakshmi **
Research Scholar, Periyar University
Asst.Professor, VISTAS
jayalakshmi.research@gmail.com

**Dr. Ananthi Sheshaayee
Associate Professor
Quaid-e-Millath Government College for Women (Autonomous), Chennai,
Anathi.research@gmail.com

ABSTRACT

The 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. Several natural language questions which are not properly specified with 'W' and 'H'(WH-operator),tends to mislead the question processing syntactically. Therefore, it is crucial to focus the missing WH-operator questions that ensurethe relevant answerto a given question. Eventhough this acquires the answer related information, only focusing on the question leads to inaccurate answer generation. In essence, the conventional QA systems lack to furnish the accurate answer when there is an increasing number of ambiguitiesin the candidate answers. Conceptually, it is paramount to measure how the questionable arguments are semantically close to the answer sentence.

Keywords*: Question answering, Semantic, Document and Question processing, natural language processing, web search engine.*

I. Introduction

The question answering is the task of automatically generating answers to the natural language questions posed by the users [1]. A wide range of research in the IR field improves the ease of accessing and quality of the information. Also, QA system facilitates users with the understanding capability of information along with the benefits of IR field. In web searching, the users need a rapid response as well as the most relevant information from the web search engine.

The users randomly submit the natural language questions on web search engine. Even though the submitted questions are lexically incorrect, the QA system provides the accurate result as the answer at the top of the web page results rather than retrieving a set of documents as the answer. Conventionally, BASEBALL and LUNAR have been the most popular QA systems that provide answers about the United States (US) baseball league and geological analysis of

the Apollo moon missions. Further, closed as well as open-domain QA systems have been developed in IR systems.

QA is a typical NLP application, which understands the natural language question and then, develops the natural language interface between a human and a system in such a manner to provide an answer that is concise and comprehensive.

The question answering is the task of automatically generating answers to the natural language questions posed by the users. A wide range of research in the IR field improves the ease of accessing and quality of the information. Also, QA system facilitates users with the understanding capability of information along with the benefits of IR field. In web searching, the users need a rapid response as well as the most relevant information from the web search engine. The users randomly submit the natural language questions on web search engine. Even though the submitted questions are lexically incorrect, the QA system provides the accurate result as the answer at the top of the web page results rather than retrieving a set of documents as the answer. Conventionally, BASEBALL and LUNAR have been the most popular QA systems that provide answers about the United States (US) baseball league and geological analysis of the Apollo moon missions. Further, closed as well as open-domain QA systems have been developed in IR systems. QA is a typical NLP application, which understands the natural language question and then, develops the natural language interface between a human and a system in such a manner to provide an answer that is concise and comprehensive.

1.1 Techniques used in QA system

The WWW data can be categorized into Data, Information, Knowledge, understanding, and wisdom [2]. 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. There is a large number of users upload the generated data on the web such as sites, blogs, reviews, and so on. On the social network, the online users interact and share their experiences. However, such resources comprise fake opinions or false reviews.

The question answering system has been utilized to retrieve the factual information on the web through the search tool.To do the question answering task, the search tools incorporate various components of knowledge discovery, survey, and selection in making decisions related to answer generation. The QAs techniques are classified based on technology used as shown in Table 1.

	Question Answer System Based on

Aspect				
	Data Mining	**Information Retrieval**	**Natural language**	**Knowledge Retrieval**
Type of Question	Simple-find	Factoid questions-what, where, which, when	Definition questions	Hypothetical and confirmation Questions yes–no
Type of Answer	Short	Combined	Combined	Combined
Searching	Factual Data	Querying for factual data	subjective, opinionated or fact	Searching for precise answers
Matching	Exact	Relevant	Relevant	Exact
Relevancy	Objective	Subjective	Subjective	Subjective
Techniques	Syntactic	Syntactic and Semantic analysis	Pattern matching, syntactic, semantic analysis	Discourse and pragmatic analysis
Knowledge Source	Data base	Syntactic information	Syntactic and pragmatic web	Semantic and pragmatic web
Models	BOW	BOW	Bag of Concepts	Bag of Knowledge

Table 1. Question Answering Process

2. SSM-ANS Methodology

The SSM-ANS (Semantic Similarity Measure-Answer) approach is targeted to present the concise answer to the natural language factoid questions by exploiting syntactic as well as semantic structures. It intends to generate the answer for both proper and improper natural language questions, wherein improper questions represent the question without WH-operator. To support the improper questions, it determines the missing WH-operator of the input question by constructing the training corpus and semantically extracts the answers by lexical, syntactic, and semantic relation.

2.1. Question processing

The input questions are often in the form of NLP. To obtain an accurate answer to the question, the SSM-ANS approach precisely identifies the most relevant documents using Web Search Engine (WSE).

To achieve this objective, it is essential to construct the training corpus to classify the questions in SSM-ANSdue to the question type is beneficial to identify the specific named entity of the answer..

2.2. Document processing

The set of documents retrieved in question processing phase is used to extract a set of potentiallysuitable candidate answer sentences. The SSM-ANS approach initially matches the question arguments with the title of the documents and its snippets to discard the documents that are irrelevant for answering the original question. Finally, it assigns the rank to the answer patterns after filtering the relevant candidate answer sentences from the retrieved documents.

2.3. Result Discussion

The SSM-ANS approach is implemented by employing Java platform and an expert system of Java Expert System Shell (JESS) rule engine. The questions of the user are in the form of NLP. Hence, this approach exploits WordNet ontology to provide a semantic relation of each word.

It exploits SearchDocs search engine as the IR engine. The SearchDocs search engine [3] provides the question relevant information including snippets. The proposed implementation depends on the Java API for WordNet Searching (JAWS) that provides the interface for retrieving the information from the WordNet database.

Table 1 illustrates the specification of software tools used in the QA system implementation.

3. WH-operator classification

The SSM-ANS approach separates the WH-questions into three types such as what, who, which, when, where, whom, and whose are named as the first level question type. Similarly, in the case of second level question type, why and how question words are added. In third or complex level question type, how-many, how-much, how-old, how-long, and so on. Also, if more than one WH words are present in a question that is also added in third level question type.

The WH-covert questions are named as the fourth or most complex level question type. In preprocessing stage, the SSM-ANS approach employs Porter stemmer and Stanford Parser for extracting the question arguments.

4. Dataset

The SSM-ANS approach collects the questions together from TREC 8, TREC 9, and TREC 10 data sets that comprise 5952 questions[4][5].

The TREC-QA question set comprises factoid questions. Hence, it is answered by the short noun phrases. Other than the factoid type of questions are answered by the explanation in which question types

are in the format of why or how questions.

Among the 5952 questions, the SSM-ANS approach has taken into account of 5452 questions as a training set and 500 questions as a test set. The details of the dataset involved in the implementation are shownin Table 4.3.

Sample trained questions:

Trained questions	Answers
Q1: How many calories are there in a Big Mac?	257 calories
Q2: Who was the first American in space?	Alan B. Shepard
Q3: How many Grand Slam titles did Bjorn Born win?	11

Table 2: Sample Trained questions of the TREC QA dataset

5. Evaluation metrics

Precision: Precision is the ratio between the number of correct answers and the number of questions to answer returned.

Recall: Recall is the ratio between the number of correct answers and the total number of questions.

5.1. Result and Discussions:

- The implementation of the SSM-ANS and IQAS systems reveal its performance variation using four evaluation metrics such as
- Precision vs. the number of trained questions, recall vs. the number of trained questions,
- F-measure vs. complex question factor.
- Accuracy vs. the number of trained questions.

Screenshots:

Figure 1: Input screenshot of the SSM-ANS approach

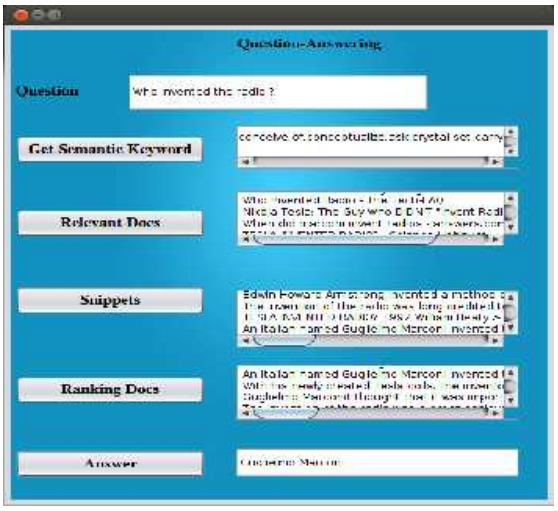

Figure 2: Output screenshot of the SSM-ANS approach

5.2. Precision Vs Training Questions

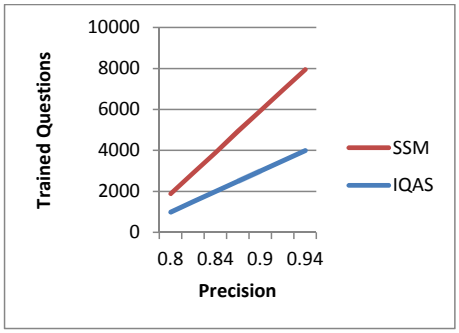

Fig. 3. Precision Vs Training Questions

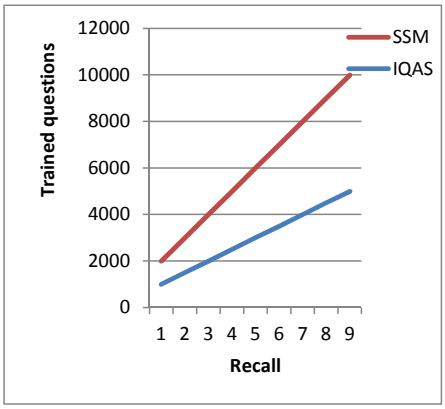

Fig 4: Recall Vs Training Questions

6. Conclusion

The concise answer generation methodology that enriched question, document, and answers processing modules of the QA system. The significant performance improvement is accomplished by contemplating the improper questions, and syntactic and semantic relation. This section has presented the introduction, brief overview about the SUPREME-ANS, and the overall methodology of the SSM-ANSapproach. Finally, it demonstrated the evaluation results of both the SSM-ANSand IQAS approach when testing on the TREC evaluation dataset.

References:

1. Abdelghani Bouzianea, Djelloul Bouchihaa, Noureddine Doumib, and Mimoun Malkic, ”Question Answering Systems: Survey and Trends”, International Conference on Advanced Wireless Information and Communication Technologies, Vol.73, pp.366–375, 2015
2. 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
3. http://searchdocs.net/
4. http://trec.nist.gov/data/qa.html
5. http://www.cs.cmu.edu/~ark/QA-data.