Automated Question Answering System Using Ontology and Semantic Role

S. Jayalakshmi
Research Scholar, Periyar University, Asst.Professor, Vels
University, India.
E-mail:jayalakshmi.research@gmail.com

Dr.Ananthi Sheshasaayee
Associate Prof.& Head, PG & Research Dept. of Computer Science, Quaid-e-Millath Government College for Women (Autonomous), Chennai, India.
E-mail: ananthi.research@gmail.com

Abstract- Semantic similarity is an essential part for question answering, it is used various fields such as Artificial Intelligence, Natural Language Processing, information retrieval, Document Retrieval and Automatic evaluations. This paper mainly focuses on similarity measure based on the posted query, and finding the appropriate meaning between the words. Accessing an accurate answer from the web document is challenging task. The proposed approach is used to analyze and measuring the similarity between the words. It presents the Web And semantic knowledge-Driven automatic question answering system (WAD). It encompasses three phases to enhance the performance of QA system using the web as well as the semantic knowledge. Initially, the WAD approach determines the user query, query expansion technique and entity linking method. The ontology based information is used in WAD to rank the answers and experimental results provide the result with high accuracy than the baseline method.

Key Words: **Semantic Relation; Ontology; Answer Ranking; Natural Language Processing.**

I. INTRODUCTION

Question Answering is the most challenging process when the QA system requires several components to achieve a standard quality. The components are Named Entity Recognition, Named Entity Disambiguation, and semantic analysis of the question, query expansion, query execution, and many more. The process of integrating these components into QA system is a difficult task to generating the candidate answers. In the web data, the most significant and interesting facts often repeatedly extend the processing time of the response generation. Hence, restricting the time consumption is primordial to avoid the web redundancy.

When exploiting a large information source, the concept relevance identification and irrelevant content deduction are considered as major constraints. Usually, a named entity has different forms such as full name, partial names, alternate spellings, aliases, and abbreviations. Due to this name variance and entity ambiguity, the determination of the specific entity becomes a complex task of the QA system. The Web contains a plenty of data which are highly ambiguous due to the frequent occurrences of named entities in which a single name represents the multiple named entities. In QA system, an answer type validation matches the type of the candidate answer to the question using named entity recognition. It provides only the raw type labels of the system. Hence, this answer type validation is limited to some QA systems only. The semantic relation of the query terms, consideration of the intention of the questions, and the precise answer type identification are necessary to provide an accurate answer with adequate information. Thus, the proposed QA system using ontology and web documents hierarchy entity linking to predict the answers accurately according to the questions of the users.

A. Natural Language Processing

NLP provides an automatic annotation of WebPages, community mining, extracting hyponyms, question answering, paragraph extraction and key word extraction for inter entity relation representation. The main goal of NLP based mechanism used to receives user query and provide an accurate answer for the posted query with the help of machine learning algorithms.

II. Question Answering Process

Question Answering (QA) system is a process of Information Retrieval (IR). The automatic QA system is the ability to provide the answer with most relevant technical information in human language in which questions are in either simple or complex form. The QA system [1] is targeted to furnish the correct answers for natural human language queries. Most of the QA systems use Natural language processing (NLP) with the IR component for searching the answer, as they depend on open-domain documents. The NLP performs two operations such as question analysis in front-end and answer generation in back-end of the QA system.

A. Web-based Question-Answering systems

The QA system [2] utilizes the rich web sources to answer the users' natural language questions. The web source is only the dynamic information source to provide the accurate result of the questions in a timely fashion. While missing entity relationship in the large-scale knowledge bases, the web-based searching is used to generate the appropriate answers [3]. Google search engine [4] identifies the complex mappings between the surface form of users' questions and entities through the Google API. The work in [5] presents the framework for a question-answering system that includes question analysis, search, hypothesis generation, and hypothesis scoring. An automatic QA system in [6] exploits the search engine results and n-gram co-occurrence statistics to answer the complex and non-factoid questions by implementing the dynamic programming algorithm. To predict the appropriate answer type of each question, the existing QA systems contemplate on the question analysis regarding question classification, and context based answer type validation [7,8].

B. Ontology-based Question-Answering systems

Ontology a common vocabulary, it extends service to expand the queries before submitting them to the web search engine. Conceptual semantic space [9] is used to expand the users' questions by exploiting the WordNet ontology-based semantic relations of the keywords appearing in the questions.

Most of the conventional web-based QA systems present the candidate answers by exploring named entity recognition and N-grams measurement. Hence, this work focuses on linking the entities of search engine results with the ontology which induces the accurate candidate answer prediction. The WAD approach provides the appropriate answer including the adequate, and significant information on the given question by exploiting both the web corpus and ontology sources. This approach determines the answer type of each query based on the probability of the query and candidate answer sentences or context in the ontology structure [10]. Then, it validates the answer type with the list of answer sentences to predict the precise answer to a given question.

III. WAD METHODOLOGY

Query Generation: Receiving user query and identify the query type-WH

- **Preprocessing:** The task of question processing is to retrieve the relevant information from a large collection of documents against a query using search engine. In order to filter out the documents that do not contain the answer, the named entity recognition needs to map the entity types of query.

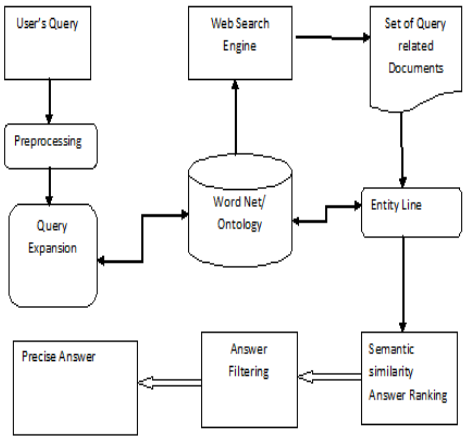

Fig 1. Process of WAD Methodology

Query Expansion: The proposed query expansion motivates the system to provide the answer with the named entity, as well as including additional information related to the named entity. In QA system, the numerous web users submit more and more variety of requests as questions on the web search engine.

Word Net: is a lexical database for English it contain Noun, verbs, adjectives and adverb are grouping information, set of cognitive synonyms. It is very useful database for semantic matching process.

Ontology: is domain knowledge used for representing by means of ontology based semantic matching question answering system.

Web search for relevant document: for the applied query it retrieve relevant documents for the related documents it is used to fetch accurate answers.

Answer Filtering: from the relevant documents the NLP WAD approach ranks the documents according to the similarity of the posted queries.

Entity link: It selects the set of snippets, including query terms with novel terms based on the threshold value. It contemplates the terms as the novel terms when the occurrence of the new term within the seven words along with the original query terms generated by the users.

Answer Ranking: The WAD approach ranks the candidate sentences based on the semantic relationship between the expanded query terms and context-aware candidate sentences.

Precise Answer: The WAD approach applies the conditional probability model on the semantic query terms with the answer type models in ontology and then, validates the answer type with the ranked list of candidate answer sentences.

A. WAD Algorithms

Input: User's question (Q_i)

Output: Corresponding answer for a given question (A_j)

//Generating user query and candidate answer

For each Q_i do

Step1.Preprocessing- - > Convert user query into NL Question format for input.

Step 2. Query Expansion - - > Snippets Selection{identify the novel term}Semantic Relation

{Synonyms, Hypernyms, Hyponyms}<- - (Search Dog Results)

Step 3. Relevant Snippets –> Discard the irrelevant contents

Step 4. Conditional Probability - -> to avoid ambiguity information retrieval.

//Ranking candidate answer Sentences

For each S_i based sentences do

Step 5. Ranking - - > use Head Word, NP, New Word

Step 6. Ranking factors - -> Path link distance and threshold limit

//**Predicting precise answer**

for each scored candidate sentence do

Step 7. Answer Type identification- -> use head word and WH

Step 8. Answer Validation - -> Determine Question Type(Q_t) and Information Selection(A_i)

Algorithm 1: The WAD Algorithm

IV. PERFORMANCE ANALYSIS OF QUESTION ANSWERING

Accuracy: Accuracy as the ratio of no of correctly classified questions to the total no of question in the testing collections.

$$\text{Accuracy} = \frac{\text{No of correctly classified Questions}}{\text{Total no. of testing samples}}$$

A. *Evaluation Metrics*

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

$$\text{Precision} = \frac{\text{No .of Correct Answers}}{\text{Total No. of Related Answers}}$$

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

$$\text{Recall} = \frac{\text{No. of Related Record Retrieved}}{\text{Total No. of Questions}}$$

F-measure: F-measure is defined as the combination of precision and recall value i.e. harmonic mean.

$$\text{F-measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

B. *Result Analysis:*

The below table discuss about performance analysis of WAD and RP-SQALD approaches

Table 1: Performance analysis Table

No. of Questions	Approach	AA	Precision
200 – 400	RP-SQALD	**10%**	**High**
	WAD	**10%**	**High**
400 -800	RP-SQALD	**9%**	**Medium**
	WAD	**10%**	**High**
800 – 1000 and above	RP-SQALD	**8%**	**Low**
	WAD	**10%**	**High**

C. *Dataset*

The WAD and RP-SQALD systems are evaluated on the Question-Answer dataset. This dataset contains the question-answer pairs from the year of 2008, 2009, and 2010. It provides factoid questions and answers from the ontology source. It roughly comprises 1000 factoid questions and the corresponding answers [11][12].

D. *Result: Precision vs. Number of questions*

Both WAD and RP-SQALD approach on QA dataset varying the number of input questions

and answers. The precision value of RP-SQALD is low after reached 800 questions while compared with WAD.

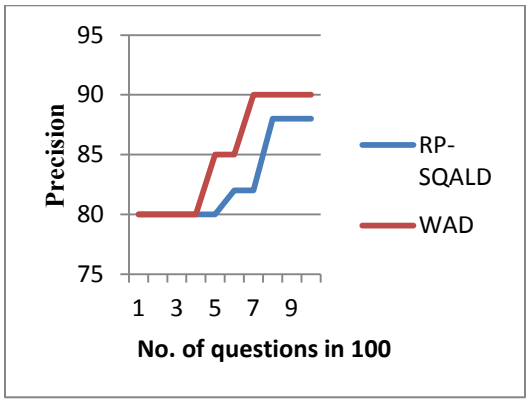

Fig 2: Performance analysis Result

V. CONCLUSION

This study has presented a WAD approach for managing the ambiguity of answers and answer selection complexity in QA system. The main goal of WAD approach is to increase the accuracy of the document matching. The user answer is ranked according to the document matching with semantic contents. The WAD approach has further improved the QA system, including relevent answer type based on the conditional probability and ontology structure. The improved system accurately validates the recognized answer type with the corresponding list of user answers and identifies the precise answer with the significant information for the users' question. Compared with the baseline QA system, the WAD approach has attained a high recall value.

REFERENCES

[1].L.Hierschman and R.Gaizauskas,"Natural Language question Answering: the view from here: ACM proceddings of the Natural Language Engineering, vol7, No.4,pp 275-300,2001.

[2]. F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge", ACM proceedings on WWW, pp.697-706, 2007

[3]. E. Brill, S. Dumais, and M. Banko, "An analysis of the AskMSR question-answering system", In EMNLP, pp.257-264, 2002

[4] West R, Gabrilovich E, Murphy K, Sun S, Gupta R, and Lin D, "Knowledge base completion via search-based question answering", ACM Proceedings of the 23rd international conference on World wide web, pp.515-526, 2014

[5] S. Monahan, J. Lehmann, T. Nyberg, J. Plymale, and A. Jung, "Cross-lingual cross-document coreference with entity linking", in TAC 2011 Workshop, 2011

[6] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, and J. Prager, "Building watson: An overview of the deepqa project", AI magazine, Vol.31, No.3, pp.59-79, 2010

[7] Soricut R, and Brill E, "Automatic question answering using the web: Beyond the factoid", In Journal of Information Retrieval-Special Issue on Web Information Retrieval, Vol.9, No.2, pp.191-206, 2006

[8] A. Lally, J. M. Prager, M. C. McCord, B. Boguraev, S. Patwardhan, J. Fan, P. Fodor, and J. Chu-Carroll, "Question analysis: How watson reads a clue", IBM Journal of Research and Development, Vol.56, No.3.4, pp.2-1, 2012

[10]. Lopez V, Uren V, Motta E, and Pasin M, "AquaLog: An ontology-driven question answering system for organizational semantic intranets", Web Semantics: Science, Services and Agents on the World Wide Web, Vol.5, No.2, pp.72-105, 2007

[11]. Web site:www.searchdocs.net.

[12]. Web site:www.cs.cmu.edu/QA-data.