World Wide Web Based Question Answering System – A Relevance Feedback Framework for Automatic Answer Validation

Santosh Kumar Ray
Birla Institute of Technology, Muscat, Oman
sapin2@gmail.com

Shailendra Singh
Samsung India Software Centre, Noida, India
shailendra.s@samsung.com

B.P.Joshi
Birla Institute of Technology, Noida, India
bp_joshi@yahoo.com

Abstract

An open domain question answering system is one of the emerging information retrieval systems available on the World Wide Web that is becoming popular day by day to get succinct and relevant answers in response of users' questions. The validation of the correctness of the answer is an important issue in the field of question answering. In this paper, we are proposing a World Wide Web based solution for answer validation where answers returned by open domain Question Answering Systems can be validated using online resources such as Wikipedia and Google. We have applied several heuristics for answer validation task and tested them against some popular World Wide Web based open domain Question Answering Systems over a collection of 500 questions collected from standard sources such as TREC, the Worldbook, and the Worldfactbook. We found that the proposed method is yielding promising results for automatic answer validation task.

Keywords: Answer Validation, Question Answering System, Web validation

1. Introduction

Today, the World Wide Web has become the major source of information for everyone, from a general user to researchers for fulfilling their information needs. Virtually all kinds of information are available on the World Wide Web in several forms. A recently published article [10] says that number of web pages on the internet increased tremendously and crossed 1 trillion landmark in 2008 which was only 200 billion in 2006 as reported in [6]. Therefore, managing such a huge volume of data is not an easy task. Search engines like Google [8] Yahoo [23] etc. are helping users to retrieve information from the World Wide Web in form of links to the documents for the user query. Most often, web pages retrieved by search engines do not provide precise information and may contain irrelevant information in even top ranked results. This prompts researchers to look for an alternate information retrieval system that can provide answers of the user queries in succinct form. Question Answering System is one of such prominent information retrieval system that is getting popular among all type of users ranging from occasional surfers to specialist information seekers. Question Answering Systems, unlike other information retrieval systems, combine information retrieval and information extraction techniques to present precise answers to the user questions posed in a natural language.

As shown in figure 1, a typical pipeline Question Answering System consists of three distinct phases: Question Processing, Document Processing, and Answer Processing. The question processing phase classifies user questions (also termed as question classification), derives expected answer types, extracts keywords, and reformulates a question into semantically equivalent multiple questions. Reformulation of a query into similar meaning queries is also known as query expansion (Section 3) and it boosts up the recall of the information retrieval system. The document processing phase retrieves documents containing keywords in the original as well as expanded questions, applies ranking algorithms on the retrieved document set and returns the top ranked documents. In answer processing phase, system identifies the candidate answer sentences, performs validation of the correctness of the answers, ranks them and finally presents the answers to the user using information extraction techniques. This is very easy to realize that providing incorrect answer is worse than providing no answer.

978-1-4244-4457-1/09/$25.00 ©2009 IEEE

To emphasize the importance of answer validation, TREC [18] has introduced the notion of confidence since its 2002 edition. This paper focuses on answer validation task in answer processing phase to enhance the precision of a Question Answering System. The proposed method exploits the fact that correct answers have higher frequency of occurrence on the World Wide Web as compared to incorrect answers. We apply several heuristics on candidate answer sentences and finally validate the correctness of the answers based on the relevance feedback provided on top retrieved documents from the World Wide Web.

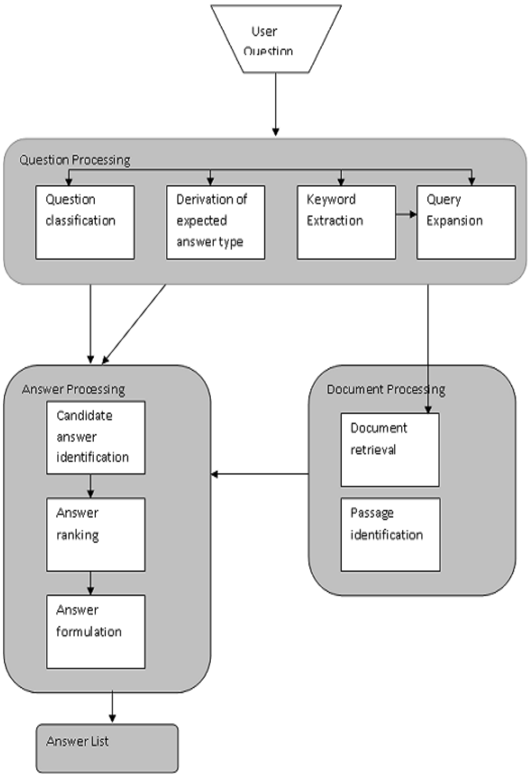

Fig. 1. Prototypical QAS architecture

In this paper, section 2 describes existing research work related to answer validation while section 3 briefly introduces the question classification and query expansion phases. Section 4 explains the heuristics applied for the proposed answer validation method. We have discussed our experiments and shown our results and observations in section 5. In the last section, we have stated our conclusion and future directions to automatically validate the answers retrieved by a Question Answering System.

2. Related Work

Question Answering System has an important role to deliver exact answers against questions asked by users. A number of Question Answering Systems such as START [16], AnswerBus [1], BrainBoost [3], PowerAnswer [14], Inferret [9], Yahoo Answering System [22], etc. are running on the World Wide Web to fulfill user needs. Researchers are involved in developing efficient approaches to decide criteria for judging the correctness of the answers returned by Question Answering Systems. Breck et al [7] describe the following list of criteria to evaluate retrieved answers:

- **Relevance:** Answer should be a response to the question.
- **Correctness:** Answer should be correct.
- **Conciseness:** Answer should not contain extraneous or irrelevant information.
- **Completeness:** Answer should be complete that is partial answer should not get credit.
- **Coherence:** Answer should be coherent, so that questionnaire can read it easily.
- **Justification:** Answer should be supplied with sufficient context to allow a reader to determine why this was chosen as an answer to the question.

There are two main approaches for judging the correctness of an answer: manual and automatic. In manual evaluation of an answer a team of assessors manually judge the correctness of the answer. One of the most widely known platforms using this method is TREC. TREC [12] adopted a method to judge the correctness of the answers which has been accepted widely by several Question Answering Systems. In this method, system's response to a natural language question is considered as a pair consisting of an answer string and a supporting document. All responses are manually judged by at least one human expert and answers are assigned one of the four labels: "correct", "unsupported", "inexact", or "incorrect". In order for a response unit to be judged "correct", the answer string must provide only the relevant information along with the supporting document appropriately justifying the answer string. However, the same answer string paired with a document that does not justify the answer would be judged as "unsupported". An answer string with extraneous words would be judged "inexact". Finally, the response would be judged as "incorrect" if the answer string does not provide the information requested in the question.

Automatic validation of the answer is a relatively new concept. Introduction of AVE (Automatic Validation Exercise) at QA@CLEF in 2006 [5] gave a major boost to research in this direction. Some of the best performing answer validation systems during initial AVE are discussed in [13][11]. [2] proposes a simple automatic validation system which searches for the presence or absence of the important terms of the questions in the answer snippets.

We are using the World Wide Web resources for validation of the answers returned by an open-domain Question Answering System. As per literature, World Wide Web has not been used for answer validation task.

3. Question Classification and Query Expansion

In this section, we are briefly describing the question classification and the query expansion phases of our Question Answering System. These phases are important as performance of the answer validation module is depending upon the correctness and exactness of the output of these phases.

3.1. Question Classification

The question classification module takes the user questions as input and classifies them into one of the eight question types namely, Functional word questions, How questions, What questions, When questions, Who questions, Where questions, Which questions, and Why questions. Each of these questions types, in turn, is independently described by several regular expressions. For the user questions matching with any of the regular expression, possible answer patterns are written. Our question classification module uses a procedure named *QC* for determination of the entity type expected to be present in the candidate answers for the given question. The procedure *QC* uses online resources such as Wikipedia [20] and WordNet [21] to compute the entity type. Consider the question for example, "*What well-known actor is the father of star Alan Alda?* ". After being parsed, this question matches with the pattern of "What questions". The Subject NP (Noun Phrase) "well-known actor" is passed by *QC* to the WordNet and the Wikipedia to determine the entity type of the NP which turns outs to be HUMAN-INDIVIDUAL. These answer patterns and entity types (such as LOCATION, HUMAN-INDVIDUAL, ORGANIZATION etc.) are used for similarity coefficient computation (Sec 4.1) and named entity matching (Sec 4.2) in the answer validation module.

3.2. Query Expansion

Query expansion phase is generally used to solve the fundamental problem of term mismatch between the authors' vocabularies for description of documents and users' search keywords. Query expansion process typically involves taking the keywords extracted from the user query and looking them up in to thesaurus or other resources to add similar search terms in order to fetch as many relevant documents as possible. Use of multiple domain ontologies for query expansions is recent and popular approach and we are using query expansion method described in our earlier work [15]. This method uses a semantic web search engine named Swoogle [17] for retrieval of domain ontologies and combines them with WordNet for expanding the user query. For example, consider the question "*Explain the reason for sky's blue color*". This question is expanded as "Explain the (Reason OR Cause) for (Sky OR Rainbow OR Cloud OR Lightning) (Blue OR Sky-blue) Color". These expanded queries are then used for similarity computation and web validation heuristics in the answer validation module.

4. Answer Validation

Answer validation is not a trivial task. It requires inputs from all the previous phases of Question Answering Process (Section 3) so that the correctness of the candidate answers can be judged appropriately. In this section, we are describing three heuristics applied in our answer validation module: Similarity computation, Named entity matching, and World Wide Web based Answer validation.

4.1. Similarity Computation

The first heuristic for answer validation is to find the similarity between the expected answer patterns and candidate answers extracted from the World Wide Web. For a given question, a set of possible answer sentences are produced by question classification and query expansion phases of the Question Answering System. These answer patterns are then represented as syntactic tree and the candidate answer sentences are parsed in bottom up fashion to match with any of these syntax trees. Similarity score is then assigned on the basis of what fraction of the candidate answer sentence has been successfully parsed. If a candidate answer sentence gets score higher than the threshold similarity score with any of the syntactic trees, it is sent for the second heuristic. Consider the question "*In what year did Arundhati Roy receive a Booker Prize?*" The expanded query "*In what year did ("Arundhati Roy" OR Arundhati) Receive OR Get) Booker (Prize OR Award)?*" was passed to the Google and top 10 answer sentences from top ranked documents were fetched. These 10 answer sentences were then parsed and matched with the expected answer patterns for the expanded query. Five answer sentences out of these 10 answer sentences got required similarity score and hence, they were sent to the next heuristic.

4.2. Entity Type

The second heuristic for answer validation searches for the presence of some specific type of entities in candidate answer sentences. The expected answer type for a given question is determined using the function *QC*. This phase uses named entity tagger to assign tags to the important keywords in candidate answers. The expected entity type of the question is then matched with the expected entity types of the important keywords of the candidate answers. If both type of entities match then candidate answer sentence is sent to the next stage of processing. The candidate answers failing to these criteria are discarded. If the procedure *QC* fails to find suitable expected answer type for a question, all candidate answer sentences are sent for further processing. Let us consider the question "*In what year did Arundhati Roy receive a Booker Prize?*" Five candidate answer sentences were received by the second stage for answer validation. The question classification module computes "date" as expected entity type for this question. We are considering date to be a number (optionally with month name or word "year"). Four candidate answer sentences (out of the five answer sentences) contained some number. Therefore, these four sentences were sent to the next stage for further processing.

4.3. World Wide Web Based Answer Validation

The World Wide Web answer validation heuristic is based on the fact that frequency of correct answers on the World Wide Web is higher than incorrect answers. Hence, we are using the World Wide Web as a source of answer validation. The original query is sent to a search engine such as Google or Yahoo search for retrieval of relevant documents. The candidate answers are then matched with the answers retrieved in top *"n"*documents. The answers validated by at least 20 % percent of the retrieved top documents are judged as correct answers. Let us consider the question as "*In what year did Arundhati Roy receive a Booker Prize?*" Four candidates passed the first two heuristics. Three candidate answer sentences contained "1997" as answer and the fourth returned "£ 20,000". Only the first answer (1997) was validated by the top most documents returned by Google. Hence, the three candidate answer sentences containing this answer were validated as correct answers.

5. Experiments and Results

5.1. Data Collection

To test the performance of the proposed answer validation method, we have chosen a set of 300 questions which is compiled using TREC [18], WorldBook [19], Worldfactbook [4], and other standard resources. For the given question collection set, the Question Answering System extracted 2604 candidate answer sentences. Out of these candidate answer sentences, 684 sentences are YES answers (containing correct answers) and remaining 1920 sentences are NO answers.

5.2. Results

We carried out our experiments on the set of collected questions and their corresponding answer sentences. We are presenting our results and observations in this section.

The three heuristics for answer validation were applied on the candidate answer sentences in a sequential manner. We applied Similarity computation heuristic on all 2604 extracted answer sentences and computed the similarity scores for them. There were 1560 questions which got higher score in comparison of threshold similarity score (0.5) and were passed as input to the second heuristic. In the second stage of validation, entity types of the candidate answer sentences were matched with the expected entity type of the questions. During the processing in the second stage, 557 sentences were rejected because their entity types were not matching with expected entity types. Finally, World Wide Web validation heuristic was applied on the remaining 1003 sentences and found 702 sentences as validated correct answers. In 450 cases, answers validated by the proposed method matched correctly with the answers judged as correct by the human assessors.

Precision and recall have been widely used parameters for the performance evaluation of an information retrieval system. Higher value of recall indicates that the system is able to retrieve higher number of relevant documents (or sentences here) from the knowledge base. On the other hand, higher value precision indicates the better accuracy of informational retrieval process. Both of these parameters are important for the overall performance of the information system. F-measure was introduced in TREC to encourage researchers to improve recall and precision simultaneously. We computed the precision, recall, and F-measure for the answer validation module using equation 1, 2 and 3 respectively.

$$precision = \frac{Number\ of\ Answers\ \Pr edicted\ As\ YES\ Correctly}{Number\ of\ Answers\ \Pr edicted\ As\ YES} \quad (1)$$

$$recall = \frac{Number\ of\ Answers\ \Pr edicted\ As\ YES\ Correctly}{Number\ of\ YES\ Answers} \quad (2)$$

$$F - measure = \frac{2 * precision * recall}{precision + recall} \quad (3)$$

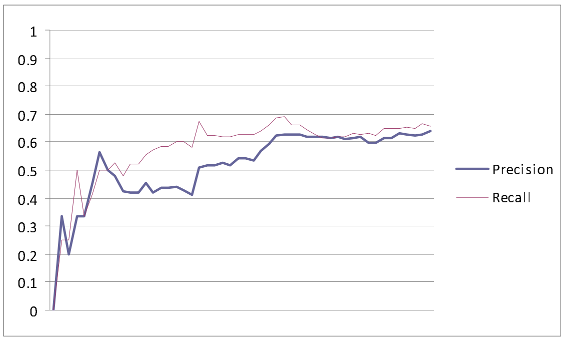

Figure 2. Precision and Recall for answer sentences validated as YES answers

Results of the experiments for our question collection are shown in figure 2 where x axis indicates question numbers and Y axis indicates precision (or Recall) value. It is obvious from the figure that we obtained a competitive result for both precision and recall. The system achieved a precision up to 0.7 and then stabilized at about 0.65. Recall also follows the similar trajectory and stabilizes for values ranging from 0.62 to 0.66. The overall performance of the systems has been summarized in table 1.

Table 1. Performance analysis of the answer validation module (with Google as back-end search engine)

Performance Parameters	YES Answers	NO Answers
Precision	0.6410	0.8675
Recall	0.6579	0.8593
F-measure	0.6493	0.8633

To verify the performance of the answer validation module we experimented with the same set of questions using Yahoo as back-end search engine. The results, as shown in table 2, are almost similar.

Table 2. Performance analysis of the answer validation module (with Yahoo as back-end search engine)

Performance Parameters	YES Answers	NO Answers
Precision	0.7058	0.9096
Recall	0.6315	0.9475
F-measure	0.6665	0.9281

Table 3. Results for answer validation as reported in [2]

Performance Parameters	YES Answers	NO Answers
Precision	0.54	0.73
Recall	0.58	0.87
F-measure	0.56	0.79

In literature, none of research work is reporting automated answer validation for World Wide Web based Question Answering Systems. Therefore, we performed comparison with Google, Yahoo, and with [2] as shown in table 1, table 2, and table 3 respectively. The comparison reflects that we have achieved a significant improvement in automated answer validation process.

Though the proposed method reports a competitive performance, there were certain issues in performance evaluation. Many of the questions in the question collection were the list questions and the explanation questions, where complete answer is either a list of entities or a complete paragraph. Answer validation module searched for some specific entity type (derived by Question classification module) in candidate sentences.. Let us consider the question, for example, “Give me the countries that border India”. Question classification module returns “location” as entity type. Answer validation module judges the answer as correct answer if it contains name of any country and question keywords (such as border) in same sentence or neighboring sentences. However, the correct answer of this question is a list of seven neighboring countries of India. This creates a mismatch between the judgments by answer validation module and that by human assessors though the candidate sentence contains partially correct answer. This reflects in reporting of the lower performance of answer validation process.

6. Conclusion and Future Scope

The World Wide Web has become the chief source of information as it adds millions of new data regularly. To retrieve information from such a vast source, we need to have efficient information retrieval systems. Therefore, Question Answering Systems are playing an important role to retrieve relevant answers for specific users’ questions. There are highly sophisticated Question Answering Systems available on the World Wide Web but still there is a scope of improvement in their performance.

The performance of an automated Question Answering System, however, depends on the accurate judgment of the correctness of the answers to a great extent. In this paper, we have proposed a World Wide Web based method for automatic answer validation. The proposed method combines the vastness of the World Wide Web with Natural Language Processing (NLP) techniques. This provides us two-folded advantage. First, the use of NLP helps us to exactly pinpoint the type of entity expected by the user in a specific question and formulate the expected answer sentences accordingly. Secondly, the World Wide Web is offering us numerous semantic resources such as WordNet, Swoogle, and Wikipedia etc. These resources can be efficiently exploited to extract answers which are syntactically and semantically closer to the formulated answers for a specific user query. We have demonstrated it with our experiments where the system has achieved an accuracy of approximately 65% for answer validation results.

We are planning to extend the proposed method over the question collections in the social networking domain where information is relatively rough and unstructured and poses higher challenges to any kind of information retrieval system.

7. References

[1] AnswerBus Question Answering System, Website: http:// www.answerbus.com

[2] A-L. Ligozat, B.Grau, A. Vilnat, I. Robba, and A. Grappy, "Lexical Validation of Answers in Question Answering", In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, Washington DC, 2-5 Nov, 2007, pp.330-333.

[3] Brainboost Question Answering System, Website: http://www.answers.com/bb/

[4] CIA the world Factbook, Website: https://www.cia.gov/library/publications/the-world-factbook/

[5] Cross Language Evaluation Forum, 2006, Website: http://clef-qa.itc.it/CLEF-2006.html

[6] D. Wirken, "The Google Goal of Indexing 100 Billion Web Pages", Website: www.sitepronews.com/archives/2006/sep/20.html

[7] E. Breck, J. BuRger, L.Ferro, L.Hirschman, D.House, M.Light, and I.Mani 2000, "How to Evaluate your Question Answering System Every Day … and still Get Real Work Done", In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece, 31 May - 2 June 2000, pp. 1495-1500.

[8] Google Search engine, Website: www. google.com

[9] Inferret Question Answering System, Website: http://asked.jp.

[10] J. Alpert, and N. Hajaj, "We Knew the Web was Big...", 7/25/2008, Website: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

[11] J. Herrera, A. Rodrigo, A. Penas, and F. Verdejo, " UNED Submission to AVE 2006", In *Workshop CLEF 2006*, Alicante, Spain, 2006.

[12] J. Lin and B. Katz, "Building a Reusable Test Collection for Question Answering", Journal of the American Society for Information Science and Technology, Vol. 57(7), John Wiley & Sons, Inc. , New York, May 2006, pp. 851-861.

[13] M. Tatu, B. Iles, and D. Moldovan, "Automatic Answer Validation using COGEX", In Working Notes, CLEF Workshop, Alicante, Spain, September 20-22, *2006*, pp. 494-501.

[14] PowerAnswer Question Answering System, Website: http://www.languagecomputer.com

[15] S. K. Ray, S. Singh, and B. P. Joshi, "Question Answering Systems Performance Evaluation – To Construct an Effective Conceptual Query Based on Ontologies and WordNet", In Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives , Rome, Italy, December 15-17, 2008, CEUR Workshop Proceedings, ISSN 1613-0073.

[16] START Question Answering System, Website: http://start.csail.mit.edu.

[17] Swoogle Ontology Search engine, Website: http://swoogle.umbc.edu

[18] Text Retrieval Conference, Website: http://trec.nist.gov

[19] The World Book, Website: www.worldbook.com/

[20] Wikipedia, Website: www.en.wikipedia.org.

[21] WordNet, Website: http://wordnet.princeton.edu.

[22] Yahoo Question Answering System, Website: http://answers.yahoo.com/ .

[23] Yahoo Search Engine, Website: search.yahoo.com