Algorithm for answer extraction based on pattern learning

Muthukrishnan Ramprasath[1], Shanmugasundaram Hariharan[2]
Assistant professor, J.J. College of Engineering and Technology, Trichy-620009, Tamilnadu, India
Associate professor TRP Engineering College, Trichy, Tamil Nadu,
mramprasath@gmail.com, mailtos.hariharan @gmail.com

Abstract: *The rapid growth of information available on the internet has provoked the development of diverse tool for searching and browsing large document collections. Information retrieval (IR) system act as a vital tool for identifying relevant document for user queries posted to search engine. Some special kind of IR system, such as Google, yahoo and Bing which allow the system to retrieve the relevant information to user question form web. Question Answering System (QAS) play important role for identifying the correct answer to user question by relying on the many IR tools. In this paper, we propose a method for answer extraction based on pattern learning algorithm. Answer extraction component provide precise answer to user question. The proposed QA system uses the pattern learning algorithm which consists of following component such as question transformation, question and answer pattern generation, pattern learning, pattern based answer extraction and answer evaluation. The experiment has been conducted different type question on TREC data sets. Our system used different ranking metrics in the experimental part to find the correct answer to user question. The experimental results were investigated and compare with different type of questions*.

Keywords: *Question Answering System, pattern learning, question transformation, Answer Extraction, TREC data set*.

1. Introduction

In the past few years, question answering system problem has received considerable attention in the field of answer extraction. The answer extraction component is act as a core component of question answering system [9]. Initially the majority of the work has focused on the task of factoid question where the answer to question will we short segment, usually in the form of named entities. For instance, consider the user question "*when did X get selected as president*" (TREC 2001) but the current research is shifting toward more complex type questions such as definition (*what is operating system*) and list type questions (*List the names of boxers Floyd Patterson fought* (TREC 2004)) and WH-type questions. However, NIST conducting workshops since from 1999, such as Text REtrieval Conference (TREC) [2], annotated corpora of question and answer has become available for several languages. Subsequent success of TREC in CLEF and NTCIR workshops has started [8] multilingual and cross-lingual QA tracks starts respectively.

Since, the beginning of computing machine the QA problem has been started to address in research domain. The Natural Language Processing (NLP) communities were initially used structural methods to initiate work on question answering. Early experiment in QA system was operated in very restricted domain. IR system helps to process the large volume textual information on the internet: Nevertheless, IR system lacking with answering specific question formulated by the user. The IR system having problem with reviewing all retrieved document relevant to user question in order to find the correct answer. This limitation prompts the appearance of QAS. In recent times many traditional question answering system have started to change the original user queries to improve the possibility of retrieving the correct answer to user questions.

The proposed QA system aim at identifying exact answer to user question from given set of document collection. For instance, consider the user query formulated in natural language (*who developed the vaccination against polio?)*, our system find the text segment that having respond (*Jonas Salk*) instead of returning list of relevant document to user question. [6] QA system uses bootstrapping techniques to built semi-automatic hierarchy question types and it used to transfer the user question in to appropriate question classes. However, [5],[13] used synonym and hypernyms from WordNet database to extract additional relevant documents to user question. Nevertheless, the quality of retrieved document set given by these methods does not shown improvement in the results.

In addition to that, [1] Ask MSR uses manually hand crafted, question-to-query-translation methods to focus relevant answer to user question.
Submitting user question (*How many calories are there in a Big Mac?)* in original form to search engines (Google, yahoo) often does not work well. it gives similar documents likely contain the answer to given query. The retrieved document can be examined by human experts or directed to complicated answer extraction modules of question answering system [10].

Consequently, it is difficult to find the correct answer from set of initially retrieved documents. We use some formalized pattern to extract the answer form retrieved document sets.

Table 1. Question summary in TREC QA

QA at TREC Evaluation	Number of Questions
TREC 8(1999)	198
TREC 9 (2000)	692
TREC 10(2001)	491
TREC 11(2002)	499
TREC 12(2003)	413
TREC 13 (2004)	231

The table 1 shows the summary of the questions used in TREC QA conferences since from 1999 to 2004. These tracks consist of all type of question which includes (*factoid: TREC-8* Q.NO: 170 *who was President of Afghanistan in 1994? Definition TREC-13 Q.NO:1907 who is Alberto Tomba?, List type: TREC-13 Q.NO:56.4 who were the key players in negotiating the agreement?)* questions. The rest of the paper is organized as follow: section 2 discussed the related work, section 3 present the proposed QA architecture based on pattern learning. Section 4 discuss pattern based answer extraction, section 5 discuss answer evaluation using TREC data set section 6 discuss conclusion and future work.

2. Related work

QA system is a hot issue of current research in the field of information retrieval. QA system is another form of information retrieval, where answer to the user question is directly identified based on the search engine we used. The QA system evaluation in TREC, each system given set of document collection, training question, test question and answer set. These text collections consist of newswire articles collected from many news agencies and also contain million of documents. The TREC data set contains larger number diverse type of questions (list, factoid, definition and other types) and all the question were closed class question types such as, *who won the noble price in 1991?, Where is Microsoft's corporate headquarters located? Name the first private citizen to fly in space*. In TREC question types are: Person, Location, Organization, Time and Date. It includes above type question and does not have pre-defined list of questions.

In common, growing information available in social media and internet make people difficult to receive correct information to their question posted to the search engine (Google, yahoo). QA system helps to discover relevant documents that satisfy user need from the large document collection. General QA system consists of question processing, information retrieval and answer extraction component. [3] Answer extraction is act as a core component of the QA system which is the tag of discrimination between information retrieval system and QA.

If the user question having ambiguous words than it is hard to get accurate answer to user question. In order to provide accurate answer to user question, it is needed to supply more information to narrow down the search area for question. [20] Discussed scenario based open domain QA system (HIITQA) which reports to satisfy given scenario template and this information obtain interactively. The current research in QA system has mainly promoted by TREC, CLEF, NICIR conferences. The final outcome has made known some attractive facts.

Several QA system participated in these conferences have shown their highly accurate response to certain kind of questions. For example, in the Portuguese QA track [4] properly responded to 89% of definition questions, where as in factual question it could respond only 35% of questions. Based on these facts, we focus the initial component of QA task, namely the retrieved documents that are likely contain the answer phrase to user question. Unsupervised method [14] is applied in answer extraction module to rank the candidate answer. In contrast, Supervised machine learning [15] method uses question and answer pair from the TREC data set and rank the candidate answer.

The several answer extraction modules were examined and these approaches described in general way. Initially they find the unseen information on the user question and answer sentence side and then they locate the answer suing some methods in QA system. Spending a few times to think about what kind of information is use full for user question to locate the answer. Near the beginning, QA system in TREC uses hand-crafted pattern or surface text pattern [16] & [17] to extract the answer to user questions. After analyzing diverse answer extraction method [17] & [22] we came to know the answer extraction methods mainly consider pattern matching grammar comparison between question and answers.[23] has presented (Textual Case-Based Reasoning) TCBR in intelligent fatawa (religious verdict) QA system which helps to answer religious inquires daily. A method of pattern based answer extraction method presented in this paper which process the passage given by the retrieval component and extract the accurate answer to the user question.

3. Proposed QA system architecture based on pattern learning

The information available internet can be used as a linguistic resource for learning question and answer pattern for diverse type of question. The following figure 1 shows the proposed architecture of the QA system based on pattern learning algorithm.

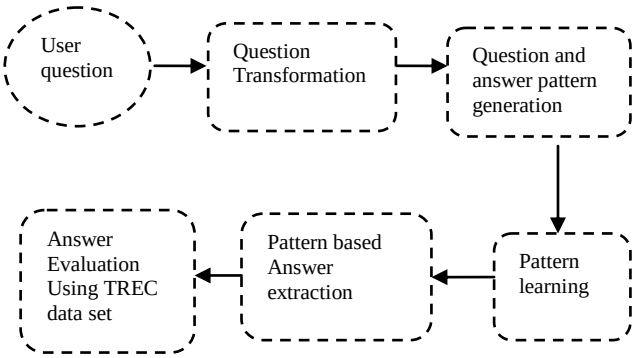

Fig 1. Pattern based answer extraction

Initially the user posts a query to the system then we transform the user question in to set of effective queries which contain term or phrases likely to appear in the retrieved documents that contain the answer to user question. Using these phrase or term we can generate the pattern for question and answer. After we construct the table of pattern for each question types using pattern learning algorithm. The constructed pattern for each question types will helps to extract the correct answer to user questions. Finally the extracted answer is evaluated using TREC data sets.

Algorithm (pattern based answer Extraction)

Input: Set of user question from TREC data set Q_u = $\{q_1, q_2, \ldots q_n\}$

Output: candidate answer based on pattern learning process.

Patten learning Process:

1) For each <question, answer> in TREC

2) Extract Q_{ph} query phrase using Q_{tri} question transformation

3) For each <question, answer> =

Generate (Q_{pt}, A_{pt}).

4) Answer Extraction using pattern learning Ans_{PL}

5) Evaluation of answer using TREC data sets

The algorithm takes set of question from TREC data set as an input and uses pattern learning process to produce exact answer to user questions. In each questions the algorithm extract query phrase using question transformation. Next, its generate question and answer patterns to extract correct answer based on the patterns. Finally, candidate answer evaluated using TREC data sets.

3.1 Question Transformation

Question transformation is the first stage of answer extraction process, which extracts phrases from user question that helps to identify different categories of question in the TREC data set.

Table 2. Question types and its phrases

Question types	Question phrase
who	"who is"
How	"how do i"
Where	"where is"
What	"what are"

The table 2 shows question phrases for each question types. For instance, "*what is mean by operating system?"* implies that the user looking for definition of operating system. The answer to the question can be inferred form question phrase *"what is mean"*. Set of user question can be given as input to the query transformation and output for this stage set of question phrases used categorize the question into individual question types.

3.2 **Pattern generation and learning**

In this second stage of learning algorithm, we need to generate pattern for question and answer pair. Question pattern is used to define the type of the question and answer pattern is searched form the retrieved document collection. [11] Presented query refinement approach based on pattern analysis where the system automatically learn text pattern from user question that can be apply to the retrieved document for extracting answer to user question. We used following procedure to construct pattern for individual question type in TREC data sets.

Step 1: select question form TREC collection for given question type. *(TREC Q.NO:1104 what year did the United States abolish the draft? ANS: 1973)*

Step 2: Extract the question and answer term and submit is as a queries to search engine. Thus, we given query +" year + United States abolish "1973" to search engine.

Step 3: download the top 25 document given by the search engine

Step 4: sentence breaker can be applied and retain only those sentence contain both question and answer term.

Step 5: use suffix tree constructor to find all substrings of all lengths along with their counts.

Step 6: suffix tree uses filter to keep only those phrases contain both question and answer terms. In our example, we extract only those phrases from suffix tree that contain the word "year "+ "united state abolish" + "1973".

Step 7: replace the question and answer term using tag "< Name>" and "<year>".

Table 3. Tags used for representing pattern

Tag	Meaning	Example
ADJ	Adjective	new, good, high, special, big, local
ADV	Adverb	really, already, still, early, now
CNJ	Conjunction	and, or, but, if, while, although
DET	Determiner	the, a, some, most, every, no
N	Noun	year, home, costs, time
NP	proper noun	Alison, Africa, April, Washington
NUM	Number	twenty-four, fourth, 1991, 14:24
PRO	Pronoun	he, their, her, its, my, I, us
WH	wh determiner	who, which, when, what, where

Table 3 list all tags used for generalization of question and answer pattern.

Table 4. Question and answer pattern for WH- type question

Wh- Question s	Question phrase	Answer Pattern
Who invented airplane?	"Who invented"	An/DT airplane/NN is/VBZ invented/VBD by/JJ
What is the color of horse?	"What is "	What/QW1 The/DT color/NN of horse/NP is/VBZ
Where is the Tajmahal?	"where is "	Where/QW2 The/DT Tajmahal/NN is/VBZ (in)…
When did the life on earth begin?	"when did"	When/QW4 The/DT life on the earth/NNP begins/VBD…
How the stock market works?	"how to"	How/QW5 The/DT stock market/NN works/VBD…
Why did David ask FBI for a word processor?	"Why did"	Why/QW6 David/NN ask/VBD FBI for a word processor/NNP….

The process is used for diverse example for same type of questions. The general pattern for INVENTOR question type uses the following output:
<ANSWER> invents <NAME>
<NAME> was invented by <ANSWER>.

Suffix tree used to record all substring partly overlapping string which allow as obtaining separate counts their occurrence of frequencies.

4. Pattern based answer extraction

In pattern based answer extraction, we used many sentences retrieved from the document collection that contain the answer and observe whether question and answer terms are present in that collection or not. For each question and answer pair extracted from TREC data set, we define question and answer terms which are likely contain the answer in the retrieved documents. Google search engine can used to post the query and examine the first 100 documents likely to contain the answer term. We used the procedure discussed in section 3.2 for retain the documents which only contain question and answer terms.

We have chosen different question types: LOCATION, DISCOVER, WHY- FAMOUS then we constructed pattern table for each question using algorithm discussed in section 3. Some of the patterns listed below along with question types.

Q.NO: 246 (TREC 9) what did Vascoda Gama discover?
A ans : sea route to India
DISCOVER
<A_{ans}> discover by <$Entity_{Name}$>
<A_{ans}> discover <$Entity_{Name}$>
<$Entity_{Name}$> discover <A_{ans}> in

Q.NO: 55 (TREC 8) Where is Microsoft's corporate headquarters located?
A ans : Redmond, Wash
LOCATION
<A_{ans}>'s <$Entity_{Name}$>
in <A_{ans}> 's <$Entity_{Name}$> in
at the <A_{ans}> 's <$Entity_{Name}$> in
Q.NO:146 In what year did Ireland elect its first woman president?
A ans: 1990
YEAR
<$Entity_{Name}$> elected < A_{ans}>
<$Entity_{Name}$> was elected in < A_{ans}>
<$Entity_{Name}$> was elected < A_{ans}>
DEFINITION
<$Entity_{Name}$> and related < A_{ans}>
Form of <$Entity_{Name}$> < A_{ans}>
As <$Entity_{Name}$> < A_{ans}> and
We used the question from TREC 8, 9 for each question types. These questions were given as input to the algorithm presented in section 3.

5. Answer evaluation using TREC data sets.

5.1. Answer assessment in TREC

We used TREC-8, TREC-9 and TREC-10 judgments set and guidelines for the candidate answer validation. It required that document returned with answer string actually support the answer contained in the string. If the answer string did not contained the correct answer the response was judged "*incorrect*". If the string hold correct answer but the document did not support the user answer, the response was judged "*unsupported*" and otherwise the response was judged correct. In TREC-10, sometimes system returns NIL as a one of the response to user question. This will affect the overall performance of the system. The proposed system uses World Wide Web data base to extract relevant passage to user question. For evaluating candidate answer [5] used data pre processing relevance scoring metric. The modification of okapi formula [21] used to score the passages retrieved from the search engine. Extracting accurate answer based on the pattern learning algorithm will check the

presence of question and answer term in the retrieved passages. Here we present some of the complicated metric used in evaluation section such as precision, recall and F-measure.

5.2 Precision Recall and F-measure

Several metrics have been used for evaluating the result of the question answering system. Precision and recall metrics are used to measure the performance of the system. In the field of QAS precision is the fraction of retrieved documents that are relevant to the user question. Recall is the probability that a relevant document is retrieved in the search. Accuracy is used as a one of the major evaluation metrics, for which the answers are judged to be a globally correct.

$$Preciion = \frac{|R_d| \cap |R_r|}{|R_d|} \qquad (1)$$

$$Recall = \frac{|R_d| \cap R_r}{|R_d|} \qquad (2)$$

Where$|R_d|$, $|R_r|$ denotes relevant document and retrieved documents related to user queries. In pattern based answer extraction measure F-Measure can be used to test the accuracy of the system. It considers both precision and recall of the test to compute the score.

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

While evaluating the QAS recall, precision and F-measure metrics are used to evaluate the performance of the system.

We presented our proposed system based on the algorithm discussed in previous section. Here we collection all WH-Type (*what*, *where*, *who*, *when*, *how*, *why*) question from TREC data set and used for evaluation based on patterning learning algorithm. Google search engine can used as a knowledge base for providing answer to all type of question posted by user. We retrieved the answer form search engine based on the key term present in the user question. [12] has presented semantic based reformulation to improve the performance of QA system. We used Perl scripting language used to implement our system. We used 493 question and answer pair from [20] TREC data set. Table 5 present Our proposed system result is compared with both in pattern based answer extraction and keyword based extraction. Thus result is reported in table 5. We used precision metric for comparing the results and number of question with at least one candidate answer. The result in the table shows the slight improvement in precision.

Table 5. The comparison of result of the each question type based on keyword search and pattern learning.

Type of question	Number of question	Candidate answer extraction based on keyword search		Candidate answer extraction using patterning learning		Precision	
		Question with at least one candidate answer	Top 5corect answer for given query	Question with at least one candidate answer	Top 5corect answer for given query	Keyword search	Pattern learning
Who	52	48	43	43	38	0.89	0.88
What	266	83	55	54	42	0.66	0.77
Where	39	38	27	28	22	0.71	0.78
When	71	49	32	35	27	0.65	0.77
How	53	32	18	20	12	0.56	0.6
Which	12	12	8	5	3	0.66	0.6
Total	**493**	**263**	**183**	**185**	**145**	**0.688**	**0.733**

6. Conclusion and Future work

We presented the method for answer extraction based on pattern learning algorithm. The experimental result show that using pattern leaning algorithm for answer extraction will helps to improve the performance of the QA system. Our system mainly focusing the question and answer key term match with the retrieved passage results. We use manually generated pattern in our experimental to retrieve the answer to user question. The work could be easily extended in future if we try

to use automatic generation pattern for accurate answer extraction.

References

[1] Abney, S., Collins, M., and Singhal, A. "Answer extraction". In Proceedings of the Applied Natural Language Processing" Conference ANLP 296–30, 2000.

[2] Attardi, Giuseppe, and Harman D.K. "Selectively using relations to improve precision in question answering". In Proceedings of the 10th Text REtrieval Conference (TREC-10), 2001 Gaithersburgh, MD, USA,November.

[3] Brill E. "A simple rule-based part of speech tagger," In Proceedings of the Applied NaturalLanguage Processing Conference ANLP,152–155,1992.

[4] Forner, P., Peñas, A., Agirre, E., Alegria, I.,Forascu, C., and Moreau, N., et al. "Overview of the clef multilingual question answering track," In Working notes for the CLEF 2008 workshop.

[5] Franz .M, J.S.McCarley, and Roukes.S "*Ad-hoc and multilingual information retrieval at ibm*" E.M.Voorhees and D.K.Harman, editor proceeding of seventh Text Retrieval Conference (TREC-7) NIST Special publication 500-242,1999.

[6] Harabagiu, s. M., pasca, m. A., and maiorano,s.j. "Experiments with open-domain textual question answering. In Proceedings of the International Conference on Computational Linguistics (COLING-2000). 292–298.

[7] Hovy, E., Gerber, L., Hermjakob, U., junk, M., and LIN, C.-Y. "Question answering in Webclopedia". In Proceedings of the TREC-9 Question Answering Track. 655–672, 2000.

[8] John 0' Connor." Retrieval of answer sentences and answer-figures from papers by text searching," Information Processing & management, 11(5/7): 155- 164, 1975.

[9] Li Peng, Teng Wen-Da, Zheng Wei "Formalized Answer Extraction Technology Based on Pattern Learning" in IFOST 2010 Proceedings 978-1-4244-9036-3/10/$26.00 ©2010 IEEE.

[10] Magnini, Bernardo, and Richard Sutcliffe. "Overview of the clef 2006 multilingual question answering track". In Proceedings of the Cross-Language EvaluationForum workshop (CLEF), Alicante, Spain, September, 2006.

[11] MuthukrishnanRamprasath,Shanmugasundaram Hariharan, "Query refinement based question answering system using pattern analysis" Adv. in Nat. Appl. Sci., 8(17): 8-15, 2014.

[12] MuthukrishnanRamprasath, Shanmugasundaram Hariharan, "improved question answering system by semantic reformulation," Advance computing (ICoAC) Dec 13-15, 2012..

[13] MILLER, G. A. Wordnet: A lexical database for English. Comm. ACM 39–41, 1995.

[14] Robertson S.R, S.Walker, S.Jones, M.Hanncock Beaulieu, M.Gatford and okapi, at TREC3. In Proceeding of third Text retrieval conference (TREC-3).NIST special publication.500-225.

[15] Ravichandran, Deepak and Eduard Hovy. "Learning surface text patterns for a question answering system," In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL)*, Philadelphia, PA, USA, July 2002.

[16] Sasaki, Y. "Question answering as question-biased term extraction'" A new approach toward multilingual QA. In Proceedings of ACL pp. 215–222, 2005.

[17] Small.S, Strzalkowski.T, Kelly.D, Rittman R, et al., Hitiqa:Scenario based question answering, in: Proceedings of HLT, 2004.

[18] Shirai.k, H. Tokue, Fundamental studies on generation of questions to users in an interactive question answering system, in: IPSJ SIG 2005-NL-165, pp. 53–56, 2005 (in Japanese).

[19] Soubbotin, Martin M. and Sergei M. Soubbotin. Patterns for potential answer expressions as clues to the right answers. In Proceedings of the 10th Text REtrieval Conference(TREC-10),Gaithersburgh,MD, USA, November, 2001.

[20] E.M. Voorhees, D.K. Harman (Eds.), Proceedings of the 11th TextREtrieval Conference (TREC 2002), Gaithersburg, Maryland, NIST,2002

[21] Yang, H., Chua, T. QUALIFIER: Question answering by lexical fabric and external resources. In Proceedings of EACL pp. 363–370, (2003).

[22] Zhang Z-Z, Zhou Y-Q, Huang X-J, Wu L-D, Answering Definition Questions Using Web Knowledge Bases[C], Proceedings of Second International Joint Conference (IJCNLP 2005), Jeju Island, Korea: 498-506. 2005.

[23] Islam Elhalwany, Ammar Mohammed, Khaled Wassif and Hesham Hefny "Using Textual Case-based Reasoning in Intelligent Fatawa QA System" The International Arab Journal of Information Technology, Vol. 12, No. 5, September 2015.