QUESTION ANSWERING SYSTEM FOR FACTOID BASED QUESTION

Prakash Ranjan
Department of Computer Science and Engineering
International Institute of Information Technology
Bhubaneswar, India 751003
Email: prakash.bitu92@gmail.com

Rakesh Chandra Balabantaray
Department of Computer Science and Engineering
International Institute of Information Technology
Bhubaneswar, India 751003
Email: rakeshbray@gmail.com

***Abstract*—Objective of question answering system (QA) is to generate concise answer of arbitrary question asked in natural language. This kind of information retrieval is required with growth of digital information. Analysis of natural language is complex task. Previously QAS were developed for specific domain and have limited efficiency. Present QAS Target on types of question commonly asked by users, characteristics of data source and correct answer generated. Our aim is to build web scale QA system. Early QA system are based on rewriting various kind of rules and pattern-generation methods for finding answer paragraph and for answer extraction. Most of QA system before answer extraction do question classification for predicting entity type of answer of question. In this paper we deals with open-domain factoid based question. Li and Roth (2002) classify questions into 6 coarse classes and 50 fine classes but it deals with limited classes of question. But we have classified question into 5 categories only and use the advance search engine technology and growth of web for QA system.**

I. INTRODUCTION

Question answering is different from traditional information retrieval systems. IR system presents the relevant document in response to the user query. Then user have to examine every document one by one for getting required answer. For a simple question user have to spend lot of time and effort with information retrial system (Ferret et al., 2001). Ideally question answering helps in providing exact answer rather than a ranked list of document that may contain the answer. Many number of QA have been developed since 1960s (Androutsopoulos et al., 1995; Kolomiyets, 2011). Current QAS attempt to answer question asked by user in natural languages after retrieving and processing information from data sources. The format of answer is also being changed from simple text to multimedia(Voorhees and weishedel, 2000).QA works can divided into three main different subtask, that are question analysis in which question classification plays main role, data retrieval , answer extraction. Almost all the QA follow these three step it may differ in how these task are implemented. QA systems can be classified on the basis of application domain, types of question asked by users, types of data sources and their characteristics, techniques used for retrieving answer.

In this paper, we deal with the problem of open- domain factoid based question answering, where the answers are in the form of word or phrase. For example

Question: Where did Bill Gates go to college?

Answer: Harvard University

These type of the question were focus on NIST TREC-QA tracks (Voorhees and Tice, 2000), which held from year from 1999 to 2007. It is very complex to find the answer of question Due to the formulations of the same question in variety of ways and presents of More than one correct answer in document. For building a good QA system it is need of deep natural language understanding. One of the top performer systems in TREC 2001, was AskMSR (Brill et al., 2001), which was built using leveraging the Web documents and generating textual patterns from the question and use it to find the answer from document For example, one of the patterns of the above question is Bill Gates went to college at. By making in consideration that answer would appear near the pattern in the document. Then system tries to extract answer from top retrieved document. Question pattern generation was major step for answer extraction in web-QA system like AskMSR (Yang et al., 2003; Zhao et al., 2007) but it use have become less now day due to popularity of QA sites like Yahoo! Answer, using of original question as query would give better relevant document instead of using derived question patterns and specially advance in modern search engine technologies by using of query reformulation and click logs analysis (Huang and Efthimiadis, 2009)

II. APPROACH

Our system consist of three steps. First step is question classification which is done to find the type of answer required by a given question. After that data collection in which we fire whole question as a query to the search and extract top 8 retrieved page. For this we have taken the help of Goggle search engine. Finally answer candidate extraction is done. Before describing each component we describe what factoid base question is about, what are its requirement, question structure.

A. Factoid based question [what, when, who, which, how]:

These question are easy and based on fact which require answers in single words or phrase e.g. who is producer of movie XYZ. It generally start with wh-word Pros of factoid based question are:

978-1-5090-5256-1/16/$31.00 ©2016 IEEE

- Generally named entities are expected answer types for many factoid based question which could be easily find in document by using named entity tagging software (Kolomiyets, 2011; Vanessa, 2011). They are based on wh-category of questions.
- There are large number of wh-factoid type question asked in QAS. Wikipedia or news text can be used as source to find the answer for such QAS

Cons of factoid based question:

- Recognition of factoid based question and its sub-classifications is a research issue in QAS.
- Fuzzy type question: it is types of question which cannot represent information needed by users correctly. These question have fuzzy terms and evaluative adjective
  e.g.: find sat off all tall guys in town?
- Identifying question which contain relationship among named entities for example: ABC is working in WXY. Here abc is employee and wxy is company. Information extraction is finding out semantic information from the text because it covers co-reference resolution, named entity recognition, relationship extraction etc.
- Syntactically incorrect question which are very hard for system to identify the users requirement in answer.
- Ambiguous question: these type of badly worded, misspelled which are very difficult to get processed for generating correct answers. Such as what make him happyyy?

B. Question classification:

Classification of question into one of several predefined categories is very important for finding answer candidates accurately (hovy et al., 2001; Li and Roth, 2002). Before question classification we have done question preprocessing. We process the question to check weather asked question is ambiguous question or misspelled such as what make him cooool. Here we see that words cool is written as cooool , so if we does not remove these types of ambiguous then it will difficult for us to classify it, for these types of word we take help of WorldNet:

Algorithm for removing ambiguous word:

Step1: check weather given words have repeated characteristics.

If (repeated character)
Goto step 2:
Else return words:

Step2: if (repeated character)
Then first match that word with words in WorldNet
If (match):
Return word from WorldNet
Else
Remove one by one repeated character of words
And match it with word in WorldNet
Return word that match with WorldNet words:

Step3: Repeat Step1 until all words are checked:

Misspelled words should also be corrected before classification. After this we classify the question. Most of the error occurs happened due to miss-classification of question (Moldovan et al.2003). We have classified the question into 5 categories. These categories are based on the named entity types, which help in finding answer candidate. The full list of 5 categories and their rules are listed in table 1.

TABLE I
5 QUESTION TYPES AND THE CORRESPONDING STRUCTURE RULES.

Types	Rules
Location	Starts with where, what/which country/city/state/airport
person	start with who what date/day/year, Contains birthplace
Date	Start with when or in what year
Name	Contains name
Others	All the other question

C. Data collection

After question classification we need to collect data from which we will extract required information. For this we have taken help of Google. We fire complete user question to Google and Google give us the link we take top 8 link from that. After that we extract data from these top 8 link. Now we have got data regarding the given question. We have taken 8 link into consideration because some question does not have answer in top link so we collect data from every top 8 link. We do same process for every users question. As it can limit our QAS because we can give answer to those question only which data Google have but at present Google have one of the best database so we have taken help from Google. There are many constant while taking help from Google as Google gives answer based on location, from browser history etc. but we have utilize this constant such that if any user ask about question who is prime minister? This question ask about person name but here it is not specified that user is asking about prime minister of which country. Google respond on location based also so it will take the location of user for this and most cases these may give write answer

D. Answer extraction

We have done classification and collected data now next step is to extract answer from top 8 retrieved page. Before that we do some analysis on question:

- We find all the possible synonyms and root words of the words other than stop words that are present in the question because there is possibilities that not exact words are find in retrieved document their synonyms or root words may be present.
- We do NER tagging of question.
- After that we do POS tagging of question

If question is not other type we apply NER on the document retrieved and find the answer candidates. If the question is

of type other then we follow the following steps to find their answer

- we have consider that answer candidates are present only on those sentence which contains the words that are present in question other than stop words their synonyms, root words or phrase from sentences. So we have taken only those sentence into consideration for finding answer.
- We consider all unigram, bigram, and trigrams at sentence which we have taken in consideration for answer extraction
- Problem while using N-gram is that it may return many number of none meaningful phrase to improve the quality we remove that N-gram that satisfy any of the following (1) containing a verb (2) starting or ending with stop words (3) if the phrase that are consisting of words that are present in question (4) punctuation as a single token. On doing this it increase the possibility of correct candidate answer.
- As we have done POS tagging of question we give the preference to words that are verb after noun in the question. It is higher preference of occurring answer near that.
- NER tagging of verb words in the question may be possible answer type.
- Phrases or words that occurs between two phrase or words that are present in the question can be answer candidate.
- There may be duplicate information among all candidates even when each answer candidate is a meaningful phrase for example Stanford and Stanford university should no less count than Stanford university. Due to nature of N-gram because both of them be correct answer. To resolve this we constructed longer N-grams from sequence of overlapping shorter N-grams. For example: table 3 is result after performing this on table 2.after this only longer answer candidate are kept(table 3).

TABLE II

Answer candidate	count
Stanford	12
College	10
university	8
Stanford university	7
Stanford college	5

TABLE III

Stanford college	12
Stanford university	8

E. Answer candidate ranking:

Now we have obtain the answer candidate there may be possibility that more than one answer candidate can be generated but we have to return only one appropriate answer for this we have to do answer ranking. For this we have taken help of WorldNet similarity to measure semantic similarity between a question and an answer candidate and also taken help of Wikipedia introduction matching.

Wikipedia introduction matching: As answer candidate only consists of few words, we use introduction section in the corresponding Wikipedia page for textual information. We use Wikipedia page where the title exactly matches the answer candidate which means no entity disambiguation is performed. If answer string is ambiguous, it will no match as title in Wikipedia have some additional description to differentiate an ambiguous concept (e.g.: insurgent (novel) VS insurgent (film)). Each question and answer candidate is represent by bag-of-word vector. The i th component of an vector indicates the term frequency of the i-word in the vocabulary. We take the inner product between the vectors of a question and an answer as the score, which indicate how relevant the answer candidate is to question. For a pair of question and answer candidate, we add the features like words in the question and the answer candidate, named entity types in the question and the answer candidate, total number of occurrences of the answer candidate in the retrieved snippets, and the initial ranking (by the number of occurrence in snippets) .

III. EVALUATION

we have created new dataset by revising answer from five year of TREC shared task; the TEXT REtrieval conference (TREC) held a question answering track (Voorchees and Tice, 2000) each year from 1999 to 2007. the date which were used in TREC are available publicly and is popular benchmark for evaluating various QA system. so we have taken factoid QA from 1999 to 2003(TREC 8-12) for testing our system. some of the question were outdated so we have removed it. we use 1651 questions from 1999 to 2001 to test our system.the question which falls into 4 categories that are Location, person, Date ,Name give higher accuracy but the question which falls into ”other” categories have lower accuracy about 57.33% while using wikipedia introduction matching accuracy increase upto 62.11%

IV. CONCLUSION AND FUTURE WORK

In this paper we have taken help of Google for data collection and retrieve answer from it. We presented a QA which answer more appropriate to factoid based .we have classified question only into 5 categories and worked on the question which fall in other categories by defining certain rule and regulation we have taken help from Wikipedia for ranking candidate answer . In future we can give answer by not classifying question answer into one of categories as types of question asked by user are increasing day by day so it is not possible to add as many number of categories as required. So by deriving or extracting feature from question can be future work without classification.

REFERENCES

[1] Xin Li and Dan Roth. 2002. Learning question classifiers. In Proceedings of the 19th international conference on Computational linguistics-Volume 1, pages 17. Association for Computational Linguistics.

[2] Eric Brill, Jimmy J Lin, Michele Banko, Susan T Dumais, and Andrew Y Ng. 2001. Data-intensive question answering. In TREC.
[3] Brill, Eric, Susan Dumais, and Michele Banko. "An analysis of the AskMSR question-answering system." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
[4] Hui Yang, Tat-Seng Chua, ShuguangWang, and Chun- Keat Koh. 2003. Structured use of external knowledge for event-based open domain question answering. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 3340. ACM
[5] Silviu Cucerzan and Eugene Agichtein. 2005. Factoid question answering over unstructured and structured web content. In TREC, volume 72, page 90
[6] Mishra, Amit, and Sanjay Kumar Jain. "A survey on question answering systems with classification." Journal of King Saud University-Computer and Information Sciences (2015).
[7] Hartawan, Andrei, and Derwin Suhartono. "Using Vector Space Model in Question Answering System." Procedia Computer Science 59 (2015): 305-311.