PRASHNOTTAR: A HINDI QUESTION ANSWERING SYSTEM

Shriya Sahu[1], Nandkishor Vasnik[2] and Devshri Roy[3]

[1]Department of Computer Science & Engineering ,MANIT, Bhopal, India
s.shriya88@gmail.com
[2]Department of Computer Science & Engineering ,MANIT, Bhopal, India
vasnik.nd@gmail.com
[3]Department of Computer Science & Engineering ,MANIT, Bhopal, India
droy.iit@gmail.com

ABSTRACT

This paper presents an approach to extract answers from Hindi text for a given question. It is based on understanding the meaning of the given question and expressing them in query logic language. The Hindi text is analyzed to understand the semantic of each sentence and relevant answer is extracted for the given question. The answers are extracted for the questions of type when, where, how many and what time. The experimental results are satisfactory.

KEYWORDS

Natural Language Processing, Question Answering, Parsing, Hindi Shallow Parser

1. INTRODUCTION

NLP focuses on interactions between computers and natural languages in terms of theoretical results and practical applications, and on information sharing now that information is exchanged as it never has been before and sharing information becomes the dominant theme in the domain of NLP systems. This trend leads to an explosion of activities like information retrieval, natural language understanding, etc. [13][14][15]. Information retrieval is the art and science of searching for information in documents, searching for documents themselves, searching for metadata which describes documents, or searching within databases, whether relational standalone databases or hypertext networked databases such as the Internet or intranets, for text, sound, images or data [7].

Question Answering (QA)[16][19] is the task of automatically answering a question posed in natural language. Hindi QA system research attempts to deal with a wide range of question types like *कब (when), कहाँ (where), किस समय (what time), कितने (how many).*

Current information retrieval systems allow us to locate documents that might contain the pertinent information, but most of them leave it to the user to extract the useful information from a ranked list. This leaves the user with a relatively large amount of text for getting the required information. There is a need for tools that would reduce the amount of text in order to obtain the desired information. People have questions and they need answers, not documents. Automatic question answering system will help for the above technology.

The rest of the paper is organized as follows. The review of Question Answering System in related fields is discussed in Section 2. In Section 3, the Architecture of the System and it's different phases are discussed. In Section 4, implementation and some related issues like

Question preprocessing , Question classification and Answer extraction algorithm are discussed. Section 5 is the discussion for experimental results with analysis. Finally, Section 6 includes conclusion and directions for future work.

1.1 Motivation: Hindi Question Answering (QA) system

The Internet today has to face the complexity of dealing with multilingualism. People speak different languages and the number of natural languages along with their dialects is estimated to be close to 4000. Of the top 100 languages in the world, Hindi occupies the fifth position with the number of speakers being close to 200 million [11]. The information need of this large section of humanity will place its unique demand on the web calling for knowledge processing of Hindi documents on the web.

All the work in Question-Answering system is done for different natural languages but as per our knowledge limited work is done in Hindi. The developed Question-Answering system in Hindi uses Hindi Shallow Parser which is developed by IIIT Hyderabad[8]. The shallow parser gives the analysis of a sentence in terms of morphological analysis, POS tagging, Chunking, etc. Apart from the final output, intermediate output of individual modules is also available. All outputs are in Shakti Standard Format (SSF)[8].

2. RELATED WORK

Semantic matching based QA system is the first generation of question answering systems. In 1973-1979, the first automatic question answering system SAM (Schank & Colby, 1973), Malaprop (Charniak, 1977), PAM (Wilensky, 1978) and POLITICS (Carbonell, 1979) are presented.

With the emergence of the World Wide Web, FAQ Finder (Burke et al. 1997)[12], AnswerBus (Zheng, 2002), as well as MULDER (Kwok, Etzioni & Weld, 2001) extend the answer extraction process from the local data source to the World Wide Web, which allows them to deal with large count of questions. In 1999, TREC[10][17] opens the first question answering task (Voorhees, 2004, Voorhees, 2001, Voorhees, 2000,Voorhees, 1999). In TREC-8, LASSO (Moldovan et al., 1999), makes use of syntax-based natural language understanding technique and question classification technique to win the question answering task.

In 2001, the question answering system of INSIGHT (Soubbotin et al., 2001), which uses some surface patterns, wins the question answering task in TREC-10. AQUA (Vargas-Vera, Motta & Domingue, 2003),which is presented in 2003, is another more sophisticated automatic question answering system, which combines natural language understanding technique, ontological knowledge, logical reasoning abilities and advanced knowledge extraction techniques[1].

In order to improve the speed of the answer extraction, Multitext (Clarke et al., 2000), IBM (Ittycheriah, Franz & Roukos, 2001), as well as SiteQ (Lee et al., 2001) use the density-based extraction method to retrieve related passages first and then extract the exact answers in them, which can greatly improve the extraction speed.

BuyAns (2005) proposes a user-interactive question answering system, which attempt to use knowledge deal to promote the enthusiasm of collaborative user. In addition, Feng et al (2006) introduce answer clustering methods n BuyAns to divide the answers into several clusters so that the users can browse the answers easily. Chen et al (2006) use the answer evaluation techniques to estimate the credibility of the answers which can greatly help users find the correct answers.

Nowadays, when users need some knowledge, they will probably relay on question answering systems, such as START (Katz et al., 2005), Baidu Zhidao[4], and so on. These systems play a more and more important role in daily life. However, there still exist some shortcomings in these QA systems.

But in Hindi there is no such Question-Answering system and this motivates for developing Hindi question-answering system in which user will pose a question in Hindi and also get answer in Hindi.

3. ARCHITECTURE

The user writes a question in Hindi using the user query interface. Then this query is used to extract all the possible answers for the input question. The architecture of Hindi Question-Answering system is as shown in Figure 1.

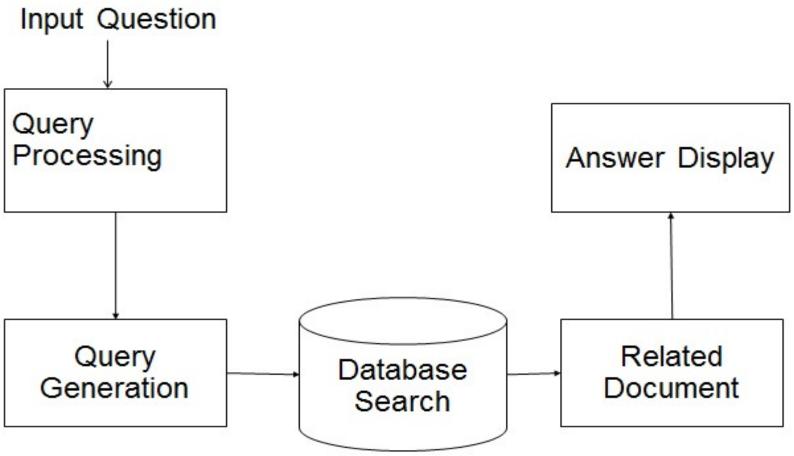


Figure 1. Architecture of Question-Answering System

The architecture given in Figure 1 works in 5 stages. The function of each stage is as follows:

3.1 Query Preprocessing

Given a natural language question as input, the overall function of the question preprocessing module is to process and analyze the input question. This leads to the classification of question as belonging to any of the types (types of question are defined in Table 1.) supported by the system.

3.2 Query Generation

In query generation we will use Query Logic Language (QLL)[1] which is used to express the input question.

3.3 Database Search

Here the search of the possible results is done in the stored database, the relevant results that satisfy the given query with selected keyword and rules are sent to the next stage.

3.4 Related Document

The result generated by the previous stage is stored as a document.

3.5 Answer Display

The result stored in the document is in wx format and the result is converted into Hindi text and displayed to the user.

4. IMPLEMENTATION

4.1 Question Preprocessing

QLL is used to express input questions. QLL is a subset of Prolog. The translation between a query written in Hindi and a logical form is performed using developed rules. The form of the logical predicates introduced by each syntax category is described as follows:

4.1.1 Predicates for when *(कब):*

i) *महात्मा गाँधी का जन्म कब हुआ?* (When was Mahatma Gandhi born?) The predicate for this interrogative sentence is:

जन्म(गाँधी, X)

ii) *ईस्ट इंडिया कंपनी ने भारत का दौरा कब किया था?* (When was East India Company visited India?) The predicate for this interrogative sentence is:

दौरा(ईस्ट_इंडिया_कंपनी, भारत, X)

4.1.2 Predicates for where *(कहाँ):*

i) *ताजमहल कहाँ हैं?* (Where is the Taj Mahal?) The predicate for this interrogative sentence is:

स्थान(ताजमहल, X)

ii*) महात्मा गाँधी का जन्म कहाँ हुआ?* (Where was Mahatma Gandhi born?) The predicate for this interrogative sentence is:

जन्म(गाँधी, X))

4.2 Question Classification

This step involves processing the question to identify the category of answer the user is seeking[18]. Further parsing the question using Hindi Shallow parser is done. Table I shows the category of the question.

The Question Processing results are a list of parts of speech(POS) plus the information for asking point. For example, the question:

ताजमहल कहाँ है? (Where is TajMahal ?)

After parsing the question by "Hindi Shallow Parser", all the parts of the sentence like verb, noun, adjective, question word(WQ) etc. are identified. For example in above sentence WQ is "kahAz"*(कहाँ)*. In this way it can be inferred that the sentence is of interrogative category. On the basis of the WQ tag the category of the question is determined. The parts of speech leads to identification of keywords (e.g. ताजमहल). The value of WQ and keywords present in the Question are further used for answer extraction.

Table 1. Question classification

Question Type	**Example Question**	**Answer Type**
कब(when)	महात्मा गाँधी का जन्म कब हुआ?(When was Mahatma Gandhi born?)	समय/तारीख (time/date)
कहाँ(where)	ताजमहल कहां है?(Where is TajMahal?)	स्थान (location)
कितने(how many)	दुनिया में कितने महाद्वीप हैं?(How many continents are in the world?)	संख्या (number)
किस समय (what time)	स्टीव जॉब्स की मृत्यु किस समय हुई?(At what time did Steve Jobs die?)	समय (time)

4.3 Answer Extraction

Answer extraction is a difficult process. It depends on the following:

- complexity of the question
- actual data where the answer is searched
- search method
- question focus and context

In most of the cases non-relevant results are often retrieved. Some of the examples are discussed below:

Example 1:

Question: भारत के प्रथम राष्ट्रपति कौन थे? (Who was the first president of India?)

The system may give the answer

Answer: ए पी जे अब्दुल कलाम भारत के ११वे राष्ट्रपति थे । (A. P. J. Abdul Kalam was 11th president of India.)

The main reason is that, traditional methods take words as independent words during matching and just check the existence of the query keywords in the stored data. Hence, they ignore the constraint relations between words in a phrase or neighbourhood. However, some results that contain most of keywords may still be non-relevant. Since the above answer contains most keywords of the question, it is still not a correct answer to the question. This is because the important immediate modifier "प्रथम (first)" of "राष्ट्रपति (president)" is ignored when we match the question to the stored data. Taking "*राष्ट्रपति*(president)" and its immediate modifier "*प्रथम*(first)" together for matching can avoid this problem to some extent.

The reason why we obtain the above non-relevant answer using these methods is that they use keyword vector to represent the question and the stored data which ignores many information, such as term position, term sequence, synonyms, and so on.

But sometimes the use of synonyms may change the actual meaning of question which is shown in example 2.

Example 2:

*"क्षीर सागर में कौन निवास करता है ?" (*Who lives in the kshir saagar *?)*

In the above sentence if we replace the word "*क्षीर*" by "*दूध*" (milk which is a synonym of क्षीर), it changes the meaning of the above question and we will not be able to extract the possible answers.

4.3.1. Algorithm for answer extraction

If the given question is: महात्मा गाँधी का जन्म कब हुआ? (When was Mahatma Gandhi born?) then extraction of the POSs "गाँधी" (Gandhi), and "जन्म" (born) is done. The Split(S) performs partitioning of the sentence S into individual words and returns the number of words. The is_a_digit() checks whether a word is a numeric or not. The algorithm for when(कब) is given in Figure2 in which rules are implemented.

```
Algorithm for "when( कब )":
Algorithm Answer_Extraction( Data_File F , Set_of_Pos POS)
{
  found ← FALSE;
  while not at end of file do
  {
      read one sentence S from F;
      i ← index of POS_1 in S;
      j ← index of POS_2 in S;
      n<- Split(S);
      for p ← 0 to n-1 do
      {
          if( is_a_digit ($k_p$) ) then
           k ← p;
      }/*end of for loop*/
  /*Rules*/
    if( i<j AND j<k ) then select S;  /* Rule_1 */
    found ← TRUE;
    else if( k<i AND i<j ) then select S; /* Rule_2 */
    found ← TRUE;
    else if( k<j AND j<i) then select S;  /* Rule_3 */
    found ← TRUE;
    else if(i<k AND k<j) then select S;  /* Rule_4 */
    found ←TRUE;
 }  /* end of while loop */
 if( found ==FALSE) then write(“Answer not found”);
 }   /*end of algorithm */
```

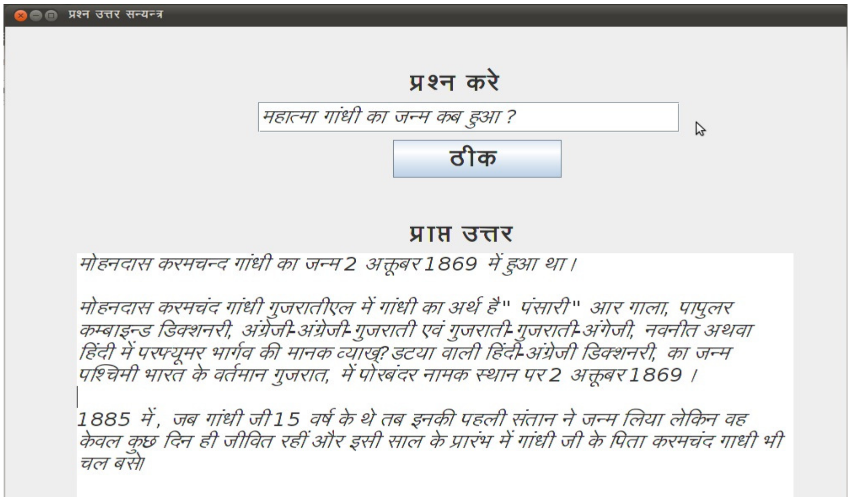
Figure 2. Algorithm for when

The algorithm given in Figure 2 has got four rules out of which one of the rules must satisfy to generate a result. If Rule_1 satisfies the sentence then the result generated contain pattern similar to the sentence “*गाँधीजी का जन्म 2 अक्टुबर, 1869 को पोरबंदर में हुआ*”. Similarly if Rule_2 satisfies the sentence the result will have pattern like “*2 अक्टुबर को गाँधीजी का जन्म हुआ*”. Further if Rule_3 satisfies the sentence we have resulting pattern like “*2 अक्टूबर को जिस महापुरुष का जन्म हुआ वो महात्मा गाँधी थे*”. Lastly if Rule_4 satisfies the sentence the result will have pattern like “*महात्मा गाँधी ने २ अक्टूबर को जन्म लिया*”. Similarly, the above implementation for ‘When’ (*कब*) can be implemented for ‘*Where*’ (*कहाँ*), ‘*What time’ (किस समय), and ‘How many’ (कितने).*

5. RESULTS AND ANALYSIS

5.1 Results

The screen shot of Hindi question answering system is given in Figure 3. Since there is no benchmark test set for the analysis, and the technology of answer extraction in Hindi is also not very mature. The experiment is performed on stored Hindi text data. The Hindi text data is collected from web. There are 60 questions of types *‘When’ (कब), ‘Where*’ (*कहाँ*), ‘*What time’ (किस समय), and ‘How many’ (कितने).* Each type has 15 questions. The accuracy for each type of question is given in table 2 and overall accuracy of the system is approximately 68.00 %.



महात्मा गांधी का जन्म कब हुआ ?

मोहनदास करमचन्द गांधी का जन्म2 अक्तूबर1869 में हुआ था।

मोहनदास करमचंद गांधी गुजरातीएल में गांधी का अर्थ है" पंसारी" आर गाला, पापुलर कम्बाइन्ड डिक्शनरी, अंग्रेजी-अंग्रेजी-गुजराती एवं गुजराती-गुजराती-अंगेजी, नवनीत अथवा हिंदी में परफ्यूमर भार्गव की मानक व्याख्? डटया वाली हिंदी-अंग्रेजी डिक्शनरी, का जन्म पश्चिमी भारत के वर्तमान गुजरात, में पोरबंदर नामक स्थान पर2 अक्तूबर1869 ।

1885 में, जब गांधी जी15 वर्ष के थे तब इनकी पहली संतान ने जन्म लिया लेकिन वह केवल कुछ दिन ही जीवित रहीं और इसी साल के प्रारंभ में गांधी जी के पिता करमचंद गाधी भी चल बसे

Figure 3. The screen shot of system

5.2 Analysis

For the questions of category ‘When’, ‘Where’, ‘What time’, and ‘How many’, the accuracy of the result is quite satisfactory. The accuracy of question type ‘*Where*’ is low because the answer type of this question is location. Location is proper noun and it is very difficult to identify the correct proper noun according to the question.

Table 2. Answer extraction

Question Type	Number of Question	Number of Error	Accuracy
When (कब)	15	5	66.66%
Where(कहाँ)	15	7	53.00%
How many(कितने)	15	4	73.33%
What time (किस समय)	15	3	80.00%
Total	60	19	68.00%

The accuracy of the question type ‘When’, ‘What time’, and ‘How many’ is relatively high because the identification of date and time is easy. For the questions that cannot get an answer, there is no further processing, and it is a factor which causes low accuracy. Some question has weak intellect, and it is difficult for people to answer. For example, the question “*हाल ही में शिमला यात्रा के लिए मुझे क्या तैयारी करने की जरुरत है*”(What need I prepare to do to travel to Shimla recently)” belongs to such category of low intellect.

6. CONCLUSION

In this paper an implementation for the question answering system in Hindi language has been done. There are wide range of rules that are employed to extract all possible set of answers from Hindi text for the input question. The focus of the system has been basically on four kind of questions of type *What, Where, How many, and what time.* On analysis of the system the overall efficiency of the system was found to be considerable. With a futuristic approach the efficiency of the algorithm can be improved through application of semantic approach and introducing a probability distribution scenario for optimal results. Further, in place of using static data set this algorithm can be extended for dynamic data set present over the internet.

REFERENCES

[1] Maria Vargas-Vera, Enrico Motta & John Domingue, (2003) “ AQUA: An Ontology-Driven Question Answering System”, 2003 American Association for Artificial Intelligence .

[2] Mehdi Rohaninezhad & Nazlia Omar, (2011) “Towards a Question Answering System Based on Precisiated Natural Language” ,2011 International Conference on Semantic Technology and Information Retrieval 28-29 June 2011, Putrajaya, Malaysia .

[3] Zhang Yi, (2004) “ ANSWER EXTRACTION ALGORITHMS IN MULTI- LINGUAL QUESTION ANSWERING SYSTEM(in Chinese)”, Master Degree Thesis of Shanghai Jiaotong University, 2004.

[4] YU ZhengTao, FAN XiaoZhong, GUO JianYi & GENG ZengMin, (2006) “Answer Extracting for Chinese Question—Answering System Based on Latent Semantic Analysis(in Chinese)”, CHINESE JOURNAL OF COMPUTERS, V01.29 No.10 Oct.2006.

[5] Zhou Zhibin , Shi Shuicai , Li Yuqin & Lv Xueqiang , (2010) “An Answer Extraction Method of Simple Question Based on Web Knowledge Library” , 2010 Second International Workshop on Education Technology and Computer Science.

[6] Jos6 L. Vicedo , (2001) “Using Semantics for Paragraph Selection in Question Answering Systems” , 0-7695-1192-910, 2001 IEEE .

[7] Wen Zhang, Taketoshi Yoshida & Xijin Tang, (2008) “TFIDF, LSI and Multi-word in Information Retrieval and Text Categorization”, International Conference on Systems, Man and Cybernetics (SMC 2008), 1-4244-2384-2/08, 2008 IEEE.

[8] Himanshu Gahlot, Awaghad Ashish Krishnarao & D. S. Kushwaha , (2009) “Shallow Parsing for Hindi - An extensive analysis of sequential learning algorithms using a large annotated corpus” , 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009 .

[9] HU Dawei, (2010) “ Research and Implementation on Answer Acquisition for Question Answering Systems” ,Submitted to Department of Computer Science in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in CITY UNIVERSITY OF HONG KONG ,May 2010 .

[10] Edward Whittaker, Sadaoki Furui & Dietrich Klakow , (2005) “A Statistical Classification Approach to Question Answering using Web Data”, Proceedings of the 2005 International Conference on Cyberworlds (CW’05), 0-7695-2378-1/05, 2005 IEEE.

[11] Shachi Dave, Pushpak Bhattachary & Dietrich Klakowya, (2001), “Knowledge Extraction from Hindi Text”, Journal of Institution of Electronic and telecommunication engineers, 18(4). http://www.cse.iitb.ac.in/~pb/papers/iete.pdf

[12] K. Hammond, R. Burke, C . Martin & S. Lytinen, (1995) “FAQ Finder: A Case-Based Approach to Knowledge Navigation”, 1043-0989B5, 1995 IEEE.

[13] DU Jia-li & YU Ping-fang, (2010) “Towards natural language processing: A well-formed substring table approach to understanding garden path sentence”, 978-1-4244-6977-2/10, 2010 IEEE.

[14]O. S. Suárez, F. J. C. Riudavets, Z. H. Figueroa & A. C. G. Cabrera, (2007) "Integration of an XML electronic dictionary with linguistic tools for natural language processing," Information Processing & Management, vol. 43, July 2007, pp. 946-957.

[15]E. Métais, (2002) "Enhancing information systems management with natural language processing techniques," Data & Knowledge Engineering, vol. 41, June 2002, pp. 247-272..

[16] Caixia YUAN & Cong WANG, (2005) "Parsing Model for Answer Extraction in Chinese Question Answering System", 0-7803-9361-9105, 2005 IEEE.

[17] Text REtrieval Conference (TREC) Data, TREC 2003,http:H/trec.nist.gov/data/qamain.html.

[18] Kepei Zhang & Jieyu Zhao,(2010) "A Chinese Question-Answering System with Question Classification and Answer Clustering", 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010).

[19] Manuel E. Sucunuta & Guido E. Riofrio, (2010) "Architecture of a Question-Answering System for a Specific Repository of Documents", 2010 2nd International Conference on Software Technology and Engineering (ICSTE).

[20] Wenpeng Lu, Jinyong Cheng & Qingbo Yang, (2012), "2012 Fifth International Conference on Intelligent Computation Technology and Automation".

Authors-

Nandkishor Vasnik received his B.E. degree in Computer Science & Engineering from Rajeev Gandhi Technical University, Bhopal, India, in 2010. Now he is an MTech student at Computer Science and Engineering Department in Maulana Azad National Institute of Technology, Bhopal, India. His interests involve Natural Language Processing (NLP) and Ontology.

Shriya Sahu received her B.E. degree in Computer Science & Engineering from Chhattisgarh Swami Vivekanand Technical University, Bhilai, India, in 2009. Now she is an MTech student at Computer Science and Engineering Department in Maulana Azad National Institute of Technology, Bhopal, India. Her interests involve Natural Language Processing (NLP) and Ontology.

Dr. Devshri Roy is a University Distinguished Scholar Professor of Computer Science and Engineering at Maulana Azad National Institute of Technology, Bhopal, India . She has done her PhD from Indian Institute of Technology, Kharagpur, India. She is specialized in Application of Computer and Communication Technologies in E-learning , Personalized Information Retrieval,and Natural Language Processing. She published many research papers including writing of books.