A Survey of Existing Question Answering Techniques for Indian Languages

Poonam Gupta
University Institute of Engineering & Technology, Panjab University Chandigarh, India
Email: poonam_123z@yahoo.co.in

Vishal Gupta
University Institute of Engineering & Technology, Panjab University Chandigarh, India
Email: vishal@pu.ac.in

***Abstract*— With advancement in technology, the process of question answering is the main field of research of text mining. In this process user is assigned a particular answers in place of number of text documents or paragraphs. It provides a way for obtaining appropriate and useful answers for questions of user put in native natural language instead of formal query in that language. Hindi, Punjabi, Bengali, Kannada, Telugu and Marathi etc are main spoken languages in India. For these Indian languages, very less number of language resources is available and much of research is going on for developing basic language resources in these languages. This paper discusses a survey of different question answering techniques for Indian languages.**

***Index Terms*— question answering system, Indian question answering techniques, information retrieval, text mining**

I. Introduction

The process of question answering is a technique in information extraction and retrieval. Most of question answering techniques are applied to retrieve appropriate answers for queries typed in native natural language of that region. Usually this process of question answering has 03 components i) categorization of questions ii) retrieval of information and retrieval of appropriate answers [9] The main role of these three components of question answering. In categorization of questions we categorize questions on the basis of entity type. Second component of retrieval of information is applied for finding success using extracting appropriate answers of different types of questions put in system of question answering. Extraction of answers component is applied for scoring & validating the candidate answers for a particular question [10]. Question answering is novel field of research. In this we try to retrieve answers for different questions instead of systems of information extraction like search engines which usually retrieve appropriate text documents. In other sense we can say that process of question answering is the next new generation in case of search engines [11]. This paper discusses a survey of different question answering techniques for Indian languages.

II. Existing Question Answering Systems for Indian Languages

Kumar et al. (2005) [1] implemented the Hindi search engine for the retrieval of the relevant passages from the collection of the passages. The architecture for the question answering was proposed. The architecture involved the various modules. Automatic Entity Generator module which identified domain related entities to which the user wanted to ask questions. Then the question was classified based upon the categories in the question classification module. Question parsing module was searching for the domain entities from the question after removing stop words. Query formulation module which converted the question into a query which then input to the retrieval engine for the retrieval of the answers. Query expansion module expanded the query which increased the search process by including the terms whose meanings were same and thus retrieved the text in which query terms do not particularly appear. Answer was extracted by using the answer extraction module. Answers were selected among the candidate answers by using ranking in the answer selection module.

Sahu et al. (2012) [2] discussed an approach for finding out the answers for the questions in the Hindi language. The answers were extracted from Hindi text and the text was completely analyzed to understand the meaning of each sentence. In this paper the main focus was on four types of questions when, where, how many and what time. The architecture for Hindi question answering system has been given. The architecture consists of five stages. The first stage was used to classify the question based upon its type. The second stage used the query logic language (QLL) for the input question. In the next stage the answers were searched in the stored database. The answer produced by that stage was stored as a document. Then the answer was converted into Hindi and then it was presented to the user. In this paper, Query logic language (QLL) which was a subset of Prolog which was used to represent the questions. Hindi shallow parser was used for the identification of the verb, noun, and question word. These words were very helpful for the

doi:10.4304/jetwi.6.2.165-168

extraction of the answers. The answers were extracted using set of developed rules.

Reddy et al. (2006) [3] described the dialogue based question answering system in Telugu language for railway specific domain. The main part of this system was dialogue manager which was responsible for the handling of the dialogues between user and the system. The system architecture for railway domain had been given. This architecture was based on the keyword approach in which input query was analyzed by the query analyzer. The query analyzer was responsible for the generation of the tokens and keywords with the use of knowledge base. Based upon the keywords and tokens which were presented in the knowledge base, an appropriate frame was selected. The words that have some semantic information were needed to be present in the knowledge base. SQL statements were generated from the tokens. There were two main issues in the design of the railway information system, how to design railway database and knowledge base. The railway database had been constructed which contained the information about the arrival / departure time of each train, information regarding their fares. For these purposes, relational model had been used. Railway database contained the tables like schedule tables, route tables and temporal tables while knowledge base contained the tables like train name, station name and also contained alias tables for station name and train name. For each input statement, root words were identified by the query analyzer during query analysis. Based upon the detection of the keywords and tokens, a query frame was identified during query frame decision. The basic responsibility of the dialogue manager was to manage the flow of dialogues. Dialogue manager was also responsible for the coordination of the other components in the system. After the generation of the query frame, SQL query was generated. Then the answer was retrieved from the database using SQL query.

Stalin et al. (2012) [4] discussed the web based application for the extraction of answers for a question posed in Hindi language from Hindi text. If the answer was not present in the Hindi text then the answers were searched on Google. This paper proposed a question answering architecture that used words of sentence (question). This acts as a source. This architecture searched the answer for a question posed in Hindi language in Hindi text. Then the results were displayed to the user. The architecture of the system involved various modules. Query interface was used for the retrieval of the question from the user. Question classification was used for the recognition of the question type. Query formulation was used for the retrieval of the correct answer. Database was used for the retrieval of the document based upon the keyword present in the question. It required the knowledge of the pattern of the question. Database sent all the candidate answers to the next module which was responsible for the extraction of the answers from the retrieved documents. Then all the candidate answers were displayed on the screen. Working of the system involved the various steps:

a. In step 1, the user first had to click on the add story link to add stories.
b. After the addition of the stories, the user can give the title to each story then the user can go back to the HOME page after clicking on the HOME link for the searching of the answers.
c. If the user want to search for a particular question. Then firstly the user has to select the story related to the question he wants to ask.
d. Then the user searched the answer of a question. If the answer was not present in the story then user can also search the answer on Google.
e. User can also generate the graph of the story selected after go back to the HOME page.

This graph was a resultant graph. It displayed the number of answers displayed to the number of related answers.

Banerjee et al. (2012) [5] discussed how to classify a question. Since classification step was necessary towards the construction of the question answering system. In this paper discussion of the lexical, semantic and syntactic characteristics was given for the classification of the Bengali question. The proposed classification work was based upon the machine learning techniques. The question classification work for Bengali Language used the Bengali shallow parser. This parser was used to analyze the sentence in terms of chunking, POS tagging etc. This paper also discussed the various interrogatives present in the Bengali language. There were three types of Interrogatives classified in Bengali:
a. Simple Interrogatives or Unit Interrogatives: they were further of two types, Singular Simple Interrogatives and Plural Simple Interrogatives.
b. Dual Interrogatives: They were made up of using simple interrogatives twice.
c. Compound / Composite Interrogatives: They were made up of using multiple simple interrogatives.

The question types for Bengali had been discussed. For the classification of the questions, there were three types of classifiers which were used to train the classifiers. These are lexical features, semantic features and syntactic features. Lexical features of the question were extracted based upon the words were present in the question. These features include wh-word, wh-word position, wh-type, length of the question, end marker and the shape of the word. Syntactical features include parts of speech (POS) tags and head words. POS tags were like nouns, adjectives, noun phrases and verb phrases. Head words were those words that were used to identify the objects from the question. Semantic features can be extracted from the question based upon meaning of the words that were present in the question. These features include related words and named entities. For the retrieval of related words Bengali dictionary had been used and for named entities, Bengali NER system had been used. The classification module had three types of classifiers:
a. Naïve Bayes(NB) classifier which was based upon Bayes' Theorem. It assumed that the presence or the absence of a specific feature of a class was not related to the presence or the absence of another feature.

b. Kernel Naïve Bayes (KNB) classifier used the estimated kernel density.
Rule induction and Decision trees were also used for classification.

Sekine et al. (2003) [6] developed a question answering system for Hindi and English. The questions were created in Hindi Language and the answers were found in Hindi Language through Hindi newspapers and then these answers were converted into English Language back. With the help of the tagger, person names, location names, organization names were identified for the English paragraphs. First of all, the examiner examined the questions and searched their answers from Hindi newspapers. An English Hindi bilingual dictionary was used to find out the top 20 Hindi articles. These articles were used to find out the candidate answers. In the end, Hindi answers were returned back to the English language.

Pakray (2007) [7] presented a system in which the query is specified by the user by starting the dialogue with the system. Every question was received with the use of language specific shallow parser. The input question was semantically tagged with the help of domain ontology. The words that were tagged were divided into chunks. With the help of keywords that were presented in the chunks, the frame of the query was determined. Dialogue manager was used to obtain the missing information from the user query. SQL statements were generated corresponding to the query frame. With the help of SQL statements, answers were extracted from the database. The natural language answer was generated by the answer generator. For the retrieval of the answer templates, the answer generator consulted with the language specific Answer/ Response template. The Bengali and the Telugu Question Answering system were evaluated with and without dialogue management.

Gupta (2013) [8] discussed method of QA for documents in English & Punjabi. This approach accepts query in input typed by any user. It then eliminates stop words present in question. English and Punjabi stop words list have been developed in advance. Then key words are retrieved from rest of query string. This approach considers verbs, adjectives and nouns as keywords. It also retrieves synonyms for keywords by usage of dictionary of Punjabi & English by applying vector space technique. Then reformulation of question is done by applying keywords & synonyms retrieved. In the next step various required web pages are extracted by using matching of strings with question reformulation. In the last step this system gives various answers extracted from various online web documents retrieved by search tool & after this it calculates ranks for the candidate answers. Gupta et al. (2013) [12] proposed Punjabi QA algorithm which uses a new way for patterns & matching to recognize most relevant appropriate answers from multiple set of answers for a given question. The relevant answers are retrieved for different types of questions like: ਕਦੋਂ "when", ਕੀ "what", ਕੌਣ "who", ਕਿਉ "why" and ਕਿੱਥੇ "where". The accuracy of this proposed QA algorithm is 73% which is calculated over fifty documents in Punjabi.

Summary of question answering systems for different Indian Languages is given in TABLE I.

TABLE I..
SUMMARY OF QUESTION ANSWERING SYSTEMS FOR INDIAN LANGUAGES

Sr. No.	Question Answering Techniques for Indian Languages	Accuracy
1.	A Hindi Question Answering system for E-Learning Documents by Praveen Kumar et al. in 2005. [1]	The system directly answered 75% of the questions.
2.	Prashnottar: A Hindi Question Answering System by Shriya Sahu et al. in 2012. [1]	Overall accuracy of the system is 68%.
3.	Dialogue based Question Answering System in Telugu by Rami Reddy et al. in 2006. [2]	Dialogue Success Rate is 83.96% and Precision is 96. 34%.
4.	Web Based Application for Hindi Question Answering System by Shalini Stalin et al. [3]	Inconclusive results are shown.
5.	Bengali Question Classification: Towards Developing QA System by Somnath Banerjee et al. in 2012. [4]	The baseline system based on Naïve Bayes classifier (using only lexical features) = 80.65% accuracy. Lexical, semantic and syntactic features accuracy= 87.63%.
6.	Hindi-English cross-lingual question-answering system by S. Sekine et al. in 2003. [5]	MRR= 0.25
7.	Multilingual Restricted Domain QA System with Dialogue Management (Bengali and Telugu) by P. Pakray in 2007. [6]	Without Dialogue Management: Bengali Query System: Precision= 85.70% Recall= 80.00%. Telugu Query System: Precision= 97.63% Recall= 93.93% With dialogue Management: Bengali Query System: Dialogue Success Rate= 72.91% and Precision= 83.67%. Telugu query System: Dialogue Success Rate= 89.06% Precision= 96.49%.
8.	Algorithm of Question Answering for Punjabi by Gupta et. al. (2013) [11]	Accuracy=73% Which is evaluated over fifty documents in Punjabi

III. Conclusions

We can conclude that although very less number of language resources are existing for Indian languages but still lot of research and development is going on for them. Regarding question answering for Indian languages, only nine papers were found for question answering systems in Hindi, Punjabi, Telugu and Bengali. On the other hand much of research is going on for developing question answering systems for other languages in the world.

References

[1] P. Kumar, S. Kashyap, A. Mittal and S. Gupta, "A Hindi Question Answering System for E-learning documents," *In Proceedings of IEEE International Conference on Intelligent sensing and Information processing*, Bangalore, India, 2005, pp. 80-85.

[2] S. Sahu, N. Vashnik and D. Roy, " Prashnottar: A Hindi Question Answering System", *International Journal of Computer Science and Information Technology (IJCSIT)*, vol.4, no.2, 2012, pp.149-158.

[3] R. Reddy, N. Reddy and S. Bandyopadhyay, "Dialogue based Question Answering System in Teulgu," *In Proceedings of EACL Workshop on Multilingual Question Answering*, 2006, pp. 53-60.

[4] S. Stalin, R. Pandey and R. Barskar, "Web based Application for Hindi Question Answering System," *International Journal of Electronics and Computer Science Engineering*, vol. 2, 2012, pp. 72-78.

[5] S. Banerjee and S. Bandyopadhyay, "Bengali Question Classification: Towards Developing QA System," *In Proceedings of the 3rd Workshop on South and Sotheast Asian Language Processing (SANLP), COLING*, 2012, pp. 25-40.

[6] S. Sekine and R. Grishman, "Hindi-English Cross-Lingual Question Answering System," *ACM Transaction on Asian Language Information Processing*, vol. 2, 2003, pp.181-192.

[7] P. Parkray, "Multilingual Restricted Domain QA System with Dialogue Management (Bengali and Telugu)," *Master's Thesis Report*, Jadavpur University, Kolkata, 2007.

[8] V. Gupta, "A Proposed Online Approach of English and Punjabi Question Answering," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 6, 2013, pp. 292-295.

[9] M. Ramprasath, and S. Hariharan, "A Survey on Question Answering System," *International Journal of Research and Reviews in Information Sciences*, vol. 2, 2012.

[10] M. D. Smucker, J. Allan, and B. Dachev, "Human Question Answering Performance Using an Interactive Document Retrieval System," *In Proceedings of ACM lliX*, Netherlands, 2012.

[11] T. Gunawardena, M. Lokuhetti, N. Pathirana, R. Ragel, and S. Deegalla, "An Automatic Answering System with template matching for Natural Language Questions,", *In Proceedings of IEEE ICIAFS*, 2010.

[12] P. Gupta and V. Gupta, "Algorithm for Punjabi Question Answering System," *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*, vol.3, 2013, pp. 902-909.