A Proposed Online Approach of English and Punjabi Question Answering

Vishal Gupta

Assistant Professor, UIET, Panjab University
Chandigarh, India

***Abstract*— This paper discusses a proposed technique of question answering for online English and Punjabi text. Initially this system takes question as input text written by user. Then stop words are removed from input question. A list of stop words has been prepared in advance for English and Punjabi. After this key terms are extracted from remaining string of question. Nouns, adjectives and verbs are treated as key terms. Synonyms of these key terms are extracted using bilingual dictionary of English and Punjabi and using Vector Space Model. Query is then reformulated by usage of these key terms and synonyms. Next phase is to retrieve the necessary web pages by applying string matching with reformulation of query. At last our question answering system returns the answers from the web documents extracted by online search engine and then it gives scores to the answer candidates. Finally we can extract top scored twenty answers for our question.**

***Keywords*— Question answering system, information retrieval, text mining, natural language processing**

I. Introduction

Text Mining [1] is an approach for automatically extracting knowledge from text which is in unstructured format. In these days huge amount of information is available on internet in the form of online digital web documents and internet can fulfil our almost every need of information. But, without proper technique which assist the users for extracting the information required when they require it, all of these online documents are of no use. For solving it, various techniques of accessing the information are applied in the world. The best examples are: information extraction [8] (IE) and approach of question answering (QA). Information extraction solves the difficulties with extraction of documents from document collection for user query. The motive of any IE technique is to search online documents collection and gives in response the subset of online text documents in decreasing order of their relevance to input query. Popular IE systems in the world are different web search engines like Altavista, Yahoo and Google. The present IE techniques are used for extracting relevant web documents for need of user, but these not able to give the concise answer of any question [12]. Online question answering (QA) systems are used for this purpose. These approaches are sufficient for giving answers to the questions in natural language of the users. Latest improvements in question answering are concentrated on answering the factual-questions (which are simply having named entities in answer), and these are usually suitable to target language as English. This paper discusses the statistical question answering system which is able for extracting answers to online the factual questions in English and Punjabi language. This proposed approach is based on assumption that questions answers are usually using same set of key terms. So the answers can be obtained by simple lexical techniques of pattern matching. They are not using complicated linguistic analyses of both questions and online web documents. The other section of this research paper is structured as follows. Section 2 gives briefly the present techniques of question answering and Section 3 shows the architecture of our proposed system of question answering and shows the techniques for reformulation of questions and extraction of answers. Section 4 shows present development, implementation and plans of future, and at last section 5 finally describes the conclusions.

II. Literature Survey

The paradigm of question-answering i.e. technique of extracting to the point answers to questions in natural language [6], was proposed in 1960 and in the start of 1970 by applying natural language understanding. For particular domains, it was developed for solving problems. Discovery of world wide web has again created the need of GUI based question answering approaches which can minimize the overflow of information, and gives challenges for automatic question answering systems. Popular applications of question answering techniques are information retrieval from whole Web (i.e. “search engines which are intelligent”), databases which are online etc. Approaches of natural language processing are used in areas which can query to online databases, retrieve required information from text, extract necessary documents from online document collection, translate text into other language, create responses to text, or recognize the terms spoken and convert in form of text. Question answering systems based on natural language processing can use machine based learning techniques for improving the rules of their syntax, improving rules of semantic, improving lexicon rules. The information extraction approaches were used by first question answering systems[9][10][11] for extracting relevant sections of text basis on key terms of questions and text documents. Present techniques apply various linguistic resources for understanding of questions and pattern based matching parts of text. The very popular resources of linguistic involves: Named entity recognition, dictionaries with semantic relations, POS (part of speech tagging), Word-net and parsers [13][14][15]. Although there are good response of these techniques, there are 02 main in-convenients: (i) task of developing these resources of linguistics is very difficult and (ii) binding of these linguistic resources with a particular language. In present world, mixing of growth of web and the great need for good access to information has increased the demand of question answering techniques for the online web. Present techniques of question answering on world wide apply a different resources of linguistic for processing of online web documents and queries. But the web size has complicated its use. Due to this, novel approaches of probabilistic on basis of online web redundancy are increased. This research paper discusses statistical based question answering technique which is able for retrieving answers of English and Punjabi factual questions from online web. The main theme of this approach is that the answers and queries are usually represented by same terms. Probability of getting simple pattern based matching in them improves. So, for input query, this method creates various reformulations of question by changing the terms order in the query. After this each reformulation is sent to online search engine, and then gathers the summary of online web documents. Finally, n-grams (word sequences) with vary high frequency are extracted from these document summaries. These word sequences are treated as the possible answer for input query. The current extends the work of Brill [16]. This system applies application of this technique in questions answering for English and Punjabi online web documents. The reformulation of query phase is different. Brill applies lexicon for finding part of speech of question terms and morphological variants of this, we have developed reformulation of query by changing the order of words without having background information regarding these terms.

III. THE METHOD

This paper discusses a proposed technique of question answering for online English and Punjabi text. Initially this system takes question as input text written by user. Then stop words are removed from input question. A list of stop words has been prepared in advance for English and Punjabi. After this key terms are extracted from remaining string of question. Nouns, adjectives and verbs are treated as key terms. Synonyms of these key terms are extracted using bilingual dictionary of English and Punjabi and using Vector Space Model. Query is then reformulated by usage of these key terms and synonyms. Next phase is to retrieve the necessary web pages by applying string matching with reformulation of query. At last our question answering system returns the answers from the web documents extracted by online search engine and then it gives scores to the answer candidates. Finally we can extract top scored twenty answers for our question.

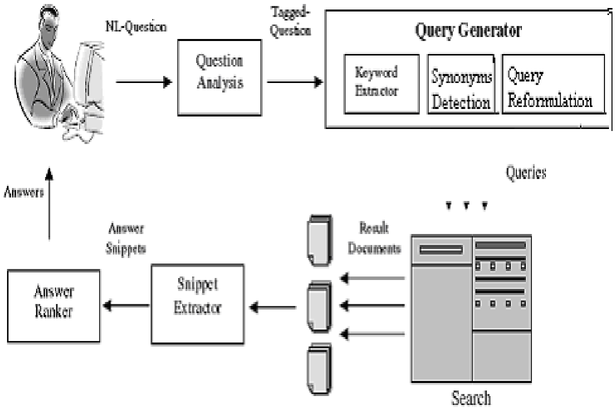

Fig. 1 Architecture of proposed system [4]

Fig. 1[4] represents the required architecture of our online question answering system. Different steps of this system are discussed below:

A. *Query Analysis*

In phase of analysis of query analysis[4][7] query string of user is analysed for extracting key terms. It accepts user queries in natural language. The query is then given to Part of Speech tagger. POS processes the query and finds part of speech of each term in query. Tagged query is then passes to generators of query. It creates various types of questions, and then is passed to a particular search engine.

B. *Query Generator Phase*

In this step reformulation of query is done. There is list of stop words for English and Punjabi . The motive of this step is to remove stop words in question string. It is having 03 sub steps.

1) Key Terms Retrieval: After removing stop words in the question string, next thing is key terms retrieval. Nons, verbs and adjectives are treated as key terms.

2) Identification of Key Terms Synonyms: In this sub step synonyms [1] [2][3] of key terms are extracted. Algorithm for identification of synonyms for Punjabi language is as:

Algorithm:
Step1: Bilingual dictionary of Punjabi and English is stored in database.
Step2: Punjabi Key terms are Input by user whose synonyms are to be determined.
Step3: Corresponding record of that term is fetched in record set. Example-
ਚੰਗਾ ---- nice, good, fine

Step4: All those records are fetched having any of the R.H.S. entries of previous record on R.H.S. For example

ਵਧੀਆ ---**Excellent,good,fine,fair**

Which means we will extract all those records in which R.H.S. field is having any of the entries among nice, good or fine.

Step5: These selected records are synonyms of the Punjabi language Key word.

We can use English Word-net for identifying synonyms of key terms in English language. Same approach can be applied on English Word-net.

3) Reformulation of Query: For input query, this sub step creates set of reformulations of query [5]. Reformulations are applied for writing expected answer of question. After removing stop words from query string, reformulations are made by key terms and synonyms of them. The below mentioned algorithm represents query as set of terms.

$Q = \{w_0, w_1, \ldots, w_{n-1}\}$.

Where w_0 represents wh-term, and n denotes the frequency of terms in question. R is notation for reformulation of query as string. It contains terms, quotation marks and spaces. It fulfils the notation of a typical question of any search engine.

R = wi wj represents the question wi AND wj.

For example: Who obtained the Nobel Physics Prize in 1999?

1st reformulation of this query as:

Obtained Nobel Physics Prize 1999

It is set of non stop-terms in the query.

2nd reformulation of this query is movement of verb:

We know that verbs are used with very high frequency after wh-term. For converting an interrogative line to declarative line it is essential to remove the verb or the other solution is to shift it to last position in any line. Reformulation of query is made by removing, or shifting at end of line, 1st & 2nd terms from query. Two examples:

i) the Nobel Physics Prize in 1999 obtained

ii) Nobel Physics prize in 1999

3rd reformulation: In 3rd reformulation there is split in components of input query. Component is type of any expression separated with preposition. So, query Q having m number of prepositions is denoted by component set

$C = (c_1, c_2, \ldots, c_{m+1})$.

Every component is subset of terms of original question string. For example:

i) “obtained the Nobel Prize” “of Physics” “in 1999”

ii) “in 1999 obtained the Nobel Physics Prize”

4th reformulation: In this main verb of query is removed and then reformulations by components is applied. Examples:

i) “in 1999 the Nobel Physics Prize”

ii) “the Nobel Prize” “of Physics” “in 1999”

C. *Online Search Tool*

Online search tool is very essential and relevant component of this proposed approach because knowledge base for this system is collection of online web documents. Answer quality is based on the assumption that there are rich quality of precise online web documents. Those online web pages are retrieved, which have necessary key terms in same lines. www. Google.com also have same technique for searching the web documents.

D. *Extraction of Document summaries*

This sub phase retrieves document summaries (i.e. snippets) from online web documents given by online search tool. Same technique for online web pages has been applied as of google. Document summaries are extracted having sentences that contain all query terms and one sentence before and one sentence after that sentence. This condition is forced for retrieving document summaries. This approach gives very high accuracy than that of approach allowing sentences not having all key terms or allowing sentences having key terms spread over many sentences.

E. *Ranking of Answers*

Web documents extracted online are automatically scored and properly ranked [17] using search engine regarding suitability with query. We know that possible suitable answers can be determined from starting few extracted online web documents. So this proposed system takes care of only starting twenty online web pages out of thousands of documents retrieved. The lines, having maximum number of key terms from question string are retrieved and scored according to frequency of key terms of input query string.

IV. CURRENT IMPLEMENTATION AND FUTURE RESEARCH

Presently, half of this proposed system has been implemented. Implementation of key terms retrieval and synonyms identification is over. After testing, the accuracy of synonyms identification sub step is around 70%. Thirty percent errors are because of lack of consistency and errors due to syntax in dictionary of Punjabi. The performance can be increased by eliminating these errors. Implementation of remaining phases will be taken care of in future for this proposed system. Some parameters for increasing performace of this system are: applying large number of possible reformulations, Implementation of stemmer of Punjabi and English, Applying technique of binary search for identification of synonyms of Punjabi and applying other good techniques for scoring answers.

V. CONCLUSIONS

Nouns adjectives, verbs and adverbs etc. are treated as Key terms for this system. Punjabi language synonyms are detected by fetching all those records containing any of the R.H.S. entries of previous record on R.H.S. All different patterns of question are obtained by applying reformulation of query. Web pages containing lines with all key terms in same sentence are preferred and retrieved than other web pages. This proposed system can only analyse starting twenty online web pages with high scores than thousands of extracted web pages.

REFERENCES

[1] M. W. Berry, "Survey of Text Mining: Clustering, Classification and Retrieval," Springer Verlag, New York, pp. 24-43, 2004.

[2] G. Singh, M. S. Gill and S.S. Joshi, "Punjabi to English Bilingual Dictionary," Punjabi University, Patiala, 1999.

[3] V. Gupta and G.S. Lehal, "Creation of thesaurus from bilingual Punjabi dictionary using text Mining," International Conference of Challenges of E- commerce and Networks, APIIT SD panipat, India,2005.

[4] J. Parikh and M. N. Murty, "Adapting Question Answering Techniques to the Web," Proceedings of the Language Engineering Conference IEEE, 2002.

[5] A. Del-Castillo-Escobedo , M. Montes-y-Gómez and L. Villaseñor-Pineda, "QA on the Web: A Preliminary Study for Spanish Language," Proceedings of the Fifth Mexican International Conference in Computer Science, IEEE, 2004.

[6] A. Andrenucci, and E. Sneiders, "Automated Question Answering: Review of the Main Approaches," Proceedings of the Third International Conference on Information Technology and Applications (ICITA) IEEE, 2005.

[7] O. Mason, "QTAG-A portable probabilistic tagger," Corpus Research, the University of Birmingham, U.K, 1997.

[8] R. Baeza and B. Ribeiro, "Modern information retrieval," ACM Press, New York, Addison-Wesley, 1999.

[9] J. Allan, M. Connel, W. Croft, F. Feng, D. Fisher and X. Li. "INQUERY and TREC-9," TREC-10, 2000.

[10] G. Cormack, A. Clarke, C. Palmer and D. Kisman, "Fast Automatic Pasaje Ranking (MultiText Experiments for TREC-8)," In TREC-8, 1999.

[11] M. Fuller, M. Kaszkiel, S. Kimberly, J. Sobel, R. Wilson and M. Wu,"The RMIT/CSIRO Ad Hoc, Q&A, Web, Interactive, and Speech Experiments at TREC-8," In TREC-8, 1999.

[12] L. Hirshman and R. Gaizauskas, "Natural Language Question Answering: The View from Here," Natural Language Engineering, vol. 7, 2001.

[13] J. Chen, A. Diekema, M. Taffet, N. McCracken, N. Ozgencil, O. Yilmazel and E. Liddyl, "Question answering: CNLP at the TREC-10 question answering track," In TREC 2001, 2001.

[14] E. Hovy, L. Gerber, U. Hermajakob, M. Junk and C. Lin, "Question answering in Webclopedia," In TREC-9, 2000.

[15] E. Hovy, U. Hermajakob and C. Lin, "The use of external knowledge in factoid QA," In TREC'01, 2001.

[16] E. Brill, J. Lin, M. Banko, S. Dumais and A. Ng, "Data-intensive question answering," In TREC '01, 2001.

[17] C A. MONTERO and K. ARAKI, "Information-Demanding Question Answering System," Intematiorial Symposium on Coinmumcations and Information Tcchnologes ISClT , Japan, 2004.