International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Special Issue No. 01, September 2015
www.ijates.com
ijates
ISSN 2348 - 7550

AUTOMATIC QUESTION GENERATION FROM HINDI TEXT USING HYBRID APPROACH

Jaspreet Kaur[1], Ashok Kumar Bathla[2]

[1,2]*Department of Computer Engineering, Yadavindra College of Engineering, Talwandi Sabo, Punjab, (India)*

ABSTRACT

Question generation is an interesting challenge of Natural Language Processing. Automatic question generation is the task of generating questions automatically from a given text. In English, a lot of work has been done in this field with the help of various tools like Semantic Role Labeller, POS Tagger, Annotated corpora tools but for Indian languages so such tools are available. In this paper, we are presenting a system that is used to generate the questions from Hindi text using hybrid approach and then a test is generated with the generated questions. Hybrid approach is the combination of rule based approach, dictionary lookup approach and example based approach. For this, a corpus is generated in Hindi language containing various different entities like names of persons, locations name, date formats, monetary expressions, etcetera. A named entity recognition tool is also created in this system to extract the entities to generate the questions automatically.

Keywords: Automatic Question Generation, Dictionary Lookup Approach, Example based Approach, Named Entity Recognition, Natural Language Processing, Rule based Approach.

I. INTRODUCTION

1.1 Question Generation

Questions are used to check the information from the existing contents or to extract information from the existing contents. So questions are the basic requirement in learning.

Automatic question generation is the process of generating reasonable questions from a given text. These systems save a lot of time as manually question generation is a very time consuming process. Question generation can be mainly categorized into two categories depending upon the target complexity, deep question generation and shallow question generation. Deep QG generates deep questions that involve more logical thinking (such as why, why not, what-if, what-if-not and how questions) whereas shallow QG generates shallow questions that focus more on facts (such as who, what, when, where, which, how many/much and yes/no questions).

The task of autoamtic question generation is basically divided into three parts. The first one is target selection, select content from input text from which a set of question can be generated. The second step is selection of question type; it is about deciding the most appropriate type of question. For example: who, why, where, how many, how much etc. The last one is question construction; it focuses on construction of actual question. Applications of automated question generation facilities are endless and far reaching. A few examples are listed below:

- These are used in question answering system in which question generated automatically.
- These systems are used by teachers in the task of setting tests.
- It is used in medicine by generating suggested question for patients and caretakers.
- It is helpful to learners by generating good question that are asked by learner while reading documents and other media.
- To human and computer tutors by generating questions that they might ask to promote and evaluate deeper learning.
- To generate suggested questions that might be asked in security contexts by interrogators or in legal contexts by petitioners.

1.2 Multiple Choice Question Generation

Multiple choice questions (MCQ) are very popular form of assessment in which respondents are asked to select the best possible answer out of a set of choices. Each test item is composed by a question and a group of answers, in which only one is correct. Incorrect answers are called distractors.

The traditional multiple-choice question generation is made up of three components: question sentence, keys and distractors. Fig. 1 shows the general structure of multiple choice question generation. Question sentence is the given sentence from which questions are generated. Then with the help of named entity recognition tool important entities are found which acts as keys. Distracters are the incorrect answers. Multiple choice question generation is an important method to evaluate students understanding and performance.

There are a lot of characteristics of multiple choice question generation. These are less influenced by guessing than true-false. Objective nature of this type of questions limits scoring bias. Items can be more efficiently and reliably scored than supply items. Difficulty can be manipulated by adjusting similarity of distractors. Items can be constructed to address various levels of cognitive complexity. Incorrect response patterns can be analyzed. Moreover, these are easy to evaluate.

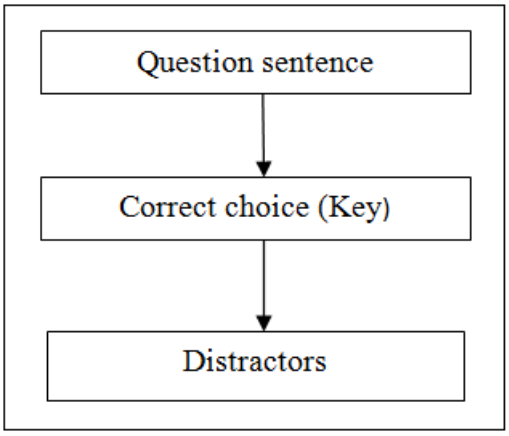

Figure: 1(Structure of Multiple Choice Question Generation System)

1.3 Named Entity Recognition

The process of identification of various elements from a given text and classification of these elements (known as named entities) into existing classes is known as named entity recognition like person name, location name, dates format, organization name, etcetera. Named entities can consist of any type of word: adverbs, prepositions,

adjectives, and even some verbs, but the majority of named entities are made up of nouns. It is a precursor for many natural languages processing tasks and now established as a key technology to understanding low- level texts. A corpus is required to generate the questions automatically which contain all the names related to person names, countries, and other entities. But the problem is that no such corpus is available for Indian languages. So to extract the various entities from the input a tool is required. These entities are used to generate the questions automatically. NER is a tool that is used to extract the entities from the input. NER systems can be used in various applications like machine translation systems, text summarization, question answering, text classification etc. A condition based approach is used to implement the NER systems for Indian languages. In condition based approach or rule based approach various rules are developed like prefix rules, suffix rules, proper names, middle names and last names to extract the entities from the text. The accuracy of the NER system depends on the rules created. More is the accuracy of NER systems; more is the accuracy of question generation.

For example consider the following paragraph:

महाराजा रणजीत सिंह का इलाका पंजाब के 30 किलोमीटर छत्तीसगड़ह के दुआले था।

From the above paragraph following are the named entities:

रणजीत सिंह

पंजाब

30 किलोमीटर

छत्तीसगड़ह

From the above entities different type of questions are generated such as based on person name format, based on number format, based on location format.

II. LITERATURE SURVEY

Bhatia et al. [1] presents an automatic generation of multiple choice questions using Wikipedia in sports domain. This system selects the sentences from the web with the help of pattern extracted from the existing questions. This system used a new technique for generating named entity distracters. This system generates the good quality of distracters by extracting extra attribute values on the key from the web. The generated questions and distracters are evaluated with the help of different parameters by a set of human evaluators.

Garg et al. [2] defines the system for generating questions automatically from given Punjabi text. This paper converts the declarative sentences into the interrogative. In this paper various Punjabi language dependant rules have been formed to generate the questions based on the named entity found in the given input sentence. The system shows the good results for some question types but for other question types system shows low results. The system is capable of generating shallow questions only with the wh- family like what, why, where, when.

Singh et al. [3] presents a rule based question generation from historical documents written in Punjabi languageA NER tool is used which recognizes the names from a sentence and generates the question accordingly. This system doesn't create all the questions starting with 'Wh-' family. This system is based on the rule based approach which requires various modifications to achieve high accuracy.

Aldabe et al. [4] defines an automatic question generator based on corpora and NLP techniques. This paper shows the results obtained in the development of a system, ArikIturri, an automatic question generator for Basque language test questions. This paper has proved the viability of this system when constructing, automatically, fill-in-the-blank, word formation and multiple choice question types. The results of this system represents that the automatic generator is good because well-formed questions are more than 80%.

Gupta et al. [5] represents the named entity recognition for Punjabi language text summarization. This paper shows that named entity recognition is used to locate and classify the elements such as names of persons, locations and different organizations, date and time formats etcetera. With the help of this approach various rules have been formed like propername rule, middlename rule, lastname rule, suffix rule and prefix rule in Punjabi language based on the various entities that are found in given text.

Mannem et al. [6] defines the question generation from the paragraphs. This system uses the predicate argument structures to select content for question generation and then series of transformations are applied to the content to generate a list of question. This system is divided into three sections: content selection, question formation and ranking.

Le et al. (2014) [7] defines an automatic question generation for educational applications. In this paper three research directions are proposed for educational purposes. First, question generation should be deployed in Intelligent Tutoring Systems in. The second research direction is deploying semantic information available on the Internet to generate semantics-based questions. The third research direction promotes applying automatic question generation.

Ali et al. (2010) [8] presents the automation of question generation from sentences. This system will generate all possible questions which the sentences contain. The given sentence may be a complex sentence; the system will generate elementary sentences, from the input complex sentences, using a syntactic parser. In this system questions are generated based on the subject, verb, object and preposition with the help of predefined interaction rules.

Zhao et al. (2011) [9] defines the automatically generating questions from queries. This method is based on search engine query logs. This system generates the question when a new query is submitted by using template instantiation. This paper explains that the search engine query logs are powerful data for the research of query to question generation, from which we get a large volume of question generation templates. This system is effective, which achieves good precision and outperforms a baseline method. The query-to-question generation technique can be used to improve the result.

Pabitha et al. (2014) [10] presents an automatic question generation system. This system uses the supervised learning approach and naïve bayes method to overcome several to overcome several problems. This system also extends the work to use summarization, noun filtering and question generation to generate the semantically correct questions. Summarization is used in case of larger file.

III. PROPOSED SYSTEM

The proposed system generates the questions from a given text written in the Hindi language with the help of a hybrid approach. Hybrid approach is the combination of rule based approach, dictionary lookup approach and example based approach. The proposed system embodies four stages: First, accept the input written in Hindi

text and tokenize the given sentence. In the second stage of the proposed system, any one approach is used from the rule based approach, dictionary lookup approach and example based approach to generate the questions according to the requirement of given sentence. In the wake of completion of this stage, four different options are generated for each question out of which one is correct and other three are incorrect. In the eventual stage of the proposed system answers are evaluated. Proposed system is also integrated with the NER tool to extract and classify the named entities to generate the questions automatically written in Hindi language. Fig. 2 displays the proposed flowchart of the system.

3.1 Rule Based Approach

In rule based approach handcrafted rules are developed to generate the questions using the key entities from a given Hindi text such as "location names", "person names", "date formats", "numbers" etcetera. Ten different rules are generated in the proposed system: Name rule, Location rule, Monetary rule, Measurement rule, Day rule, Numeral rule, City rule, Year rule, Direction rule.

Some general guidelines how these rules are used in the proposed system:

Rule 1: If any name is found in the sentence then replace it with "किस" word.

Rule 2: If any city name, state name, country name or any location name is found in the given sentence then replace it with the "कहाँ" word.

Rule 3: If any date format or any year is found then replace it with "कब" word.

Rule 4: If any integer or number is found replace it with "कितने" word.

Rule 5: If any direction is found then replace it with "किस" word.

Rule 6: If any abbreviation is found then replace it with "क्या" word.

Rule 7: If any word "इसीतर्हा" is found in the sentence then replace it with "किसतर्हा" word.

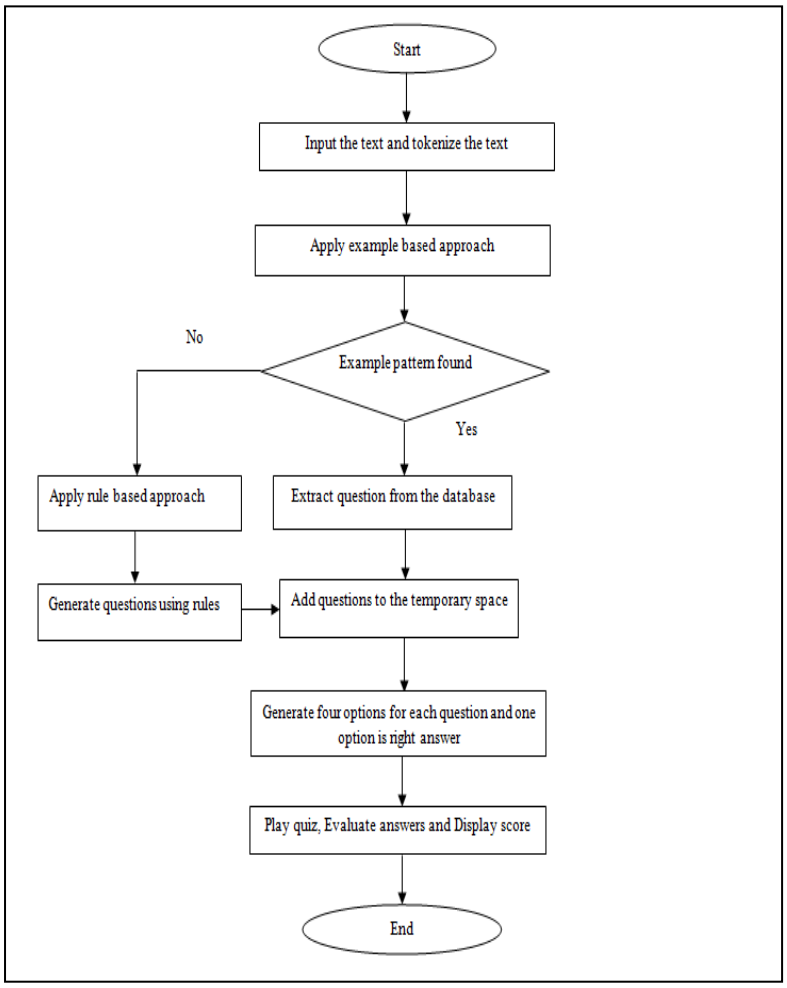

Figure: 2(Flowchart of Proposed Algorithm)

3.2 Dictionary Look-Up Approach: This approach is used with the rule based approach and example based approach. To use this approach a corpus of the named entities for Hindi language is formed. In this corpus different types of location names, person names, organization names etcetera are stored. The input text from which entities are to be extracted is compared using dictionary lookup approach with the corpus stored in the system. The extracted entities are used to generate different questions according to the various rules formed in the rule based approach.

3.3 Example Based Approach: This approach is used in the proposed system to improve the performance of the system. To use this approach some general patterns are stored. So, whenever any input is given to the system similar to the patterns stored then questions are generated by replacing required entities. For example if a pattern is stored: who is the father of Ram and input text is Gurjeet Singh is the father of Manpreet Singh. The output of this system is generated from the stored pattern by simply replacing Ram by Manpreet Singh.

IV. EVALUATION

The analysis of questions generated by the proposed system has been done manually. The proposed system has been tested on 300 different sentences that have been collected from various Hindi websites and Hindi text

books. To evaluate the performance of question generation system for Hindi, we have used three standard metrics namely Precision, Recall and F-measure. The following table 1 will compare the existing system with the proposed system on the basis of features.

TABLE: 1(Comparison of Existing System and Proposed System)

Features	Existing System	Proposed System
Rule based approach	Yes	Yes
Dictionary look up approach	Yes	Yes
Example based approach	No	Yes
Mutiple choice questions	No	Yes
Evaluation	No	Yes

Parameters used in the proposed system: Three parameters are used to evaluate this system. Table 2 displays the results of the proposed system.

Recall: The ratio of the the total number of questions that are generated by the system to the total number of questions that are generated manually by humans is known as recall value.

Precision: The ratio of the accurate questions generated by the system to the total number of questions generated by the system is known as precision.

F-Measure: The ratio of the product of the recall and precision valuse to the sum of the recall and precision valuse is known as f-measur. In other words, F-Measure is the harmonic mean of recall and precision values.

TABLE: 2 (Recall, Precision and F-Measure Values Obtained for Different Types of Questions)

Question Type	Recall	Precision	F-Measure
कहाँ (Kahan)	95.50%	92.35%	93.89%
किस (Kis)	95.79%	90.50%	93.06%
कब (Kab)	92.33%	96.85%	94.53%
कितने (Kitne)	93.50%	98.44%	95.90%
कया (Kya)	87.48%	71.88%	78.91%
कौन (Koun)	85.23%	87.88%	86.53%
दिशा (Direction Based)	93.21%	93.50%	93.35%
Monetary Based	91.33%	99%	95.01%

V. CONCLUSION AND FUTURE WORK

In this paper we present a system to generate the questions automatically from Hindi text by using hybrid approach (rule based approach, dictionary lookup approach, example based approach). In this system example based approach is used to improve the performance of the existing systems. We also proposed a corpus in Hindi

language to generate the questions automatically using rule based approach. In this system multiple choice questions are generated with four options, out of which three are incorrect and one is correct. The answers given by the user are evaluated by comparing the answer with the stored correct answer and at the last system give the total score.

The proposed method may be extended in several directions. The system can be improved by improving the rule based approach used in the proposed system by generating more rules. Since human generated questions tend to have words with the different words with different meanings, the system can be improved by removing word disambiguation. The options generated in the proposed system for multiple choice questions are not confusing so system can also be improved by more reliable distracter generation procedure

REFERENCES

[1] Bhatia A., Kirti M., Saha S., "Automatic question generation of multiple choice questions using wikipedia", Springer: Proceedings of 5th International conference (Kolkata, India), pp.733-738, 2013.

[2] Garg S., Goyal V., "System for generating questions automatically from given Punjabi text", International journal of computer Science and mobile computing, pp. 324-327, 2013.

[3] Singh P., Kaur R., "Rule based question generation system from Punjabi text contain historical information", International journal of computer science and mobile computing, pp. 86-91, 2014.

[4] Aldabe I., Lacalle M., Maritxalar M., Martinz E., Uria L., "ArikIturri: An automatic question generator based on corpora and NLP techniques", Springer-Verlag Berlin Heidelberg (Donostia, Spain), pp. 584-594, 2006.

[5] Gupta V., Lehal G., "Named entity recognition for Punjabi language text summarization", International journal of computer applications, pp. 28-32, 2011.

[6] Mannem P., Prasad R., Joshi A., "Question generation from paragraphs at UPenn: QGSTEC system description", Proceedings of the third workshop on question generation (Hyderabad, India), pp. 84-91, 2010.

[7] Le N., Kojiri T., Pinkwart N., "Automatic question generation for educational applications – The state of art", Springer: Proceedings of the 2nd international conference on computer science, applied mathematics and applications (Switzerland), pp. 325-338, 2014.

[8] Ali H., Chali Y., Hasan S., "Automation of question generation from sentences", Proceedings of the third workshop on question generation, (Pittsburgh, USA),pp. 58-67, 2010.

[9] Zhao S., Wang H., Li C., Liu T., Guan Y., "Automatically generating questions from queries for community-based question answering", Proceedings of the 5th international joint conference on natural language processing (Chiang Mai, Thailand), pp. 929-937, 2011.

[10] Pabitha P., Mohana M., Suganthi S., Sivanandhini B., "Automatic question generation system", IEEE: 2014 international conference on recent trends in information technology (Chennai, India), pp. 1-5, 2014