INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A REVIEW ON AUTOMATIC QUESTION GENERATION SYSTEM FROM A GIVEN HINDI TEXT

Jaspreet Kaur[1], Ashok Kumar Bathla[2]
[1]*M.Tech. Research Scholar, jaspreet6dec@gmail.com*
[2] *Assistant Professor, CE Deptt., Ashokashok81@gmail.com*
Yadwindra College of Engineering, Talwandi Sabo, Punjab, India.

Abstract: - Automatic Question Generation is the task of generating reasonable questions from an input with the help of various NLP techniques. This paper presents a review of automatic question generation for Indian languages. Automatic question generation is the field of Natural Language Processing (NLP). In context of English a lot of work has been done in the field of question generation with the help of various tools like Semantic Role Labeller, POS Tagger, Annotated corpora tools. But these tools are not yet available for Indian languages such as Punjabi and Hindi. For Indian languages rules are formed to create the questions automatically from a given text using Rule based approach. So, A Named Entity Recognition (NER) tool is needed to generate the questions automatically.

Keywords: Automatic question generation, Named Entity Recognition, Natural Language processing, Rule based approach.

1. INTRODUCTION

A question is a linguistic expression used to make a request for information and information may be provided with an answer. So basically questions are asked to fulfill the informational needs. Questions are used in various fields for example teacher use questions to check the knowledge of the student, in interview questions are asked to candidate to check the skills of the candidate, for entrance exams and in many other areas. Automatic question generation is the task of generating questions automatically from a given text. Question generation is an interesting challenge of the NLP field. The various applications of Autoamtic Question Generation systems include Natural language summarization, closed domain Question Answering and intelligent tutoring systems. Also Automatic Question Generation can be helpful in:

- Knowledge evaluation system
- Helps in government exams generation
- Helps teachers to evaluate the students with the help of test
- Questions that human and computer tutors might ask to promote and assess deeper learning.
- In medicine by generating suggested questions for patients and caretakers.
- To generate questions for the security reasons.

For example consider the following paragraph:

सदियों की गुलामी के पश्चात भारत 15 अगस्त सन् 1947 के दिन आजाद हुआ। पहले हम अंग्रेजों के गुलाम थे।

From the above paragraph following questions can be generated:

1. भारत कब आजाद हुआ?
2. कौनसा देश 15 अगसत 1947 को आजाद हुआ?
3. हम किसके गुलाम थे?

Therefore from above paragraph it is clear that all the combination of the questions is generated.

Automatic question generation systems save the time in creating the questions because these systems can generate the questions faster than the human. These systems also generate all the questions from a given input that are possible.

Questions can be divided into shallow question generation and deep question generation based on the complexity. Shallow QG generates shallow questions that focus more on facts such as who, when, where, which, how many/much and yes/no questions. Whereas deep QG generates deep questions that involve more logical thinking such as why, why not, what-if, what-if-not and how questions. The task of question generation is a three step process: target selection, selection of the question type and question construction. The first step target selection is about deciding what the question should be about. In other words the first step is about deciding which type of information is conveyed. The second step is about deciding which type of question is created (who, where, when, why, what etc). For example, "गुरु नानक देव जी का जनम कब हुआ? (When Guru Nanak Dev Ji was born?)", "भारत के प्रधानमंत्री कौन है? (Who is the prime minister of India?)", "गुरू नानक देव जी का जनम कहाँ हुआ? (Where was Guru Nanak Dev Ji was born?)". In the last step the question is constructed to get the information. This can be shown in the following flowchart:

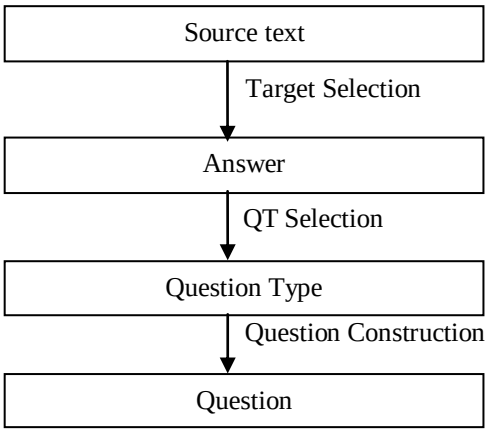

Figure 1: Flow chart of General process of question generation

2. NAMED ENTITY RECOGNITION (NER) SYSTEM

Named Entity Recognition is a tool that is used to classify the elements in text into classes such as names of person, location, organizations, date of wars etc. 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:

गुरु नानक देव जी का जन्म लाहौर के निकट 'तलवंडी' नामक स्थान पर सन् 1469 में महिता कालू जी के घर में हुआ था[

From the above paragraph following are the named entities:

गुरु नानक देव जी

लाहौर

तलवंडी

1469

महिता कालू

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

3. RELATED WORK

Garg et al. [4] proposed a system for generating questions automatically from given Punjabi text. This system converts the declarative sentences into interrogative counterparts. It accepts sentences as input and produces questions automatically on the basis of entities found in the given input sentence. The system shows good results for some question types and shows low results for other question types. It is capable of generating only shallow questions starting with the words "what", "where", "when" "who" and is not capable of generating questions with "why" and "how" etc words. The overall precision of the system is 63.20%, the overall recall value of the system calculated over the test data is 46.42%.

Singh et al. [11] proposed a rule based question generation system from Punjabi text contain historical information. This system uses NER tool which is used to recognize the names from a given input text and generates the question according to the recognized entities. Punjabi corpus is also used to generate the questions automatically and this corpus contains different entities such as names of cities, countries, locations, person name etc. to generate the questions. The overall precision of the system is 85.50%, which is far more than the previous systems. With the help of this system various type of questions starting with "where", "whom", "what", "when", "how many/much", "why", "direction", "monetary expressions" words are generated. The results are evaluated with the Recall value, Precision, F-Measure parameters.

Gupta et al. [5] presents the Named entity recognition for Punjabi language text summarization. In the field of question generation for English a lot of work has been done with the help of various techniques like POS taggers, SRL etc but for Indian languages no such tools are available. This paper explains that for Indian languages NER system is used to generate questions automatically with the help of rule based system. Five rules have been developed prefix rule, suffix rule, proper name rule, middle name rule and last name rule for this system.

Stanescu et al. [12] presents the question generation for learning evaluation. This paper presents a software tool that can be used in the learning process in order to automatically generate questions from course materials, based on a series of tags defined by the professor. The Test Creator tool permits generation of questions based on electronic materials that students have. The solution implies teachers to have a series of tags and templates that they manage. These tags are used to generate questions automatically. For each tag teacher defines several questions for a specific category. This paper is basically helpful in evaluating the knowledge of the students.

Aldabe et al. [1] proposed an automatic question generator based on corpora and NLP techniques. This system can generate fill in blank, word formation, multiple choice and error correction type different types of questions. The quality of the questions obtained depends on the source corpus and NLP techniques used in the process of question generation. Well formed questions by this generation are more than 80%.

Ali et al. [2] proposed an efficient system for automation of question generation from sentences. This system will generate the elementary sentences from a given complex sentence using a syntactic parser. Depending upon the verb, subject, object, preposition each elementary sentence is classified and depending upon the classification we determine the type of the question to be generated.

Mannem et al. [6] proposed a system for Question generation from paragraphs. This system works in three steps: content selection, question formation and ranking. This system uses the predicate argument structures of sentences with semantic role labels to generate the questions. The generated questions are then ranked to pick the best six questions.

Thinh Le et al. [13] proposed an automatic question generation for supporting argumentation. In this paper three techniques are combined to generate the questions automatically: template based, syntax based and semantics based. In this paper authors addressed two questions to generate the questions automatically. First is how the system recognizes the main concepts in the text and second is how the system uses this information to generate the questions. This system analyzes all the grammatical structures and extracts main concepts from the given text. After this, the system generates the questions automatically.

4. EXISTING WORK

In the existing systems to generate the questions automatically rule based approach is used with dictionary lookup approach. Existing systems are able to generate only shallow questions and the systems generate the questions starting with the words "what", "where", "when", "how many", "who" etc., so a lot of work is left in this field. Also these systems show good results for some questions and shows bad results for some questions. Hence there is need to improve the existing system to improve the performance of the system.

4.1 Rule based approach

Rule based approach is the first strategy ever developed in the field of machine translation. It is extensible and maintainable. Rule based approach plays a vital role in the process of creating question generation system because handcrafted rules are created according to the grammatical rules of the language used in the system. Rule based approach is basically a condition based approach that is used with the NER tool. In this approach rules can be created to extract the person names, location names, date/time formats to generate the questions from the given text. Various rules have been developed like prefix rule, suffix rule, proper name rule, middle name rule and last name rule etc. For Punjabi and Hindi languages there is no corpus available for generating the questions automatically. For this reason we use rule based approach to generate the questions automatically. Some of the rules are:

Rule 1: If a person name is found in the input sentence then question is created with "कौन (who)" word.

Rule 2: If location name is found in the input sentence then question is created with "कहाँ (where)" word.

Rule 3: If we found date or time in the input sentence then question is created with "कब (when)" word.

Rule 4: If we found any abbreviations in the input sentence then question is created with "क्या (what)" word.

Rule 5: If we found any integer in the sentence then question is created with "कितने (how many)" word. This can be shown in the following flowchart:

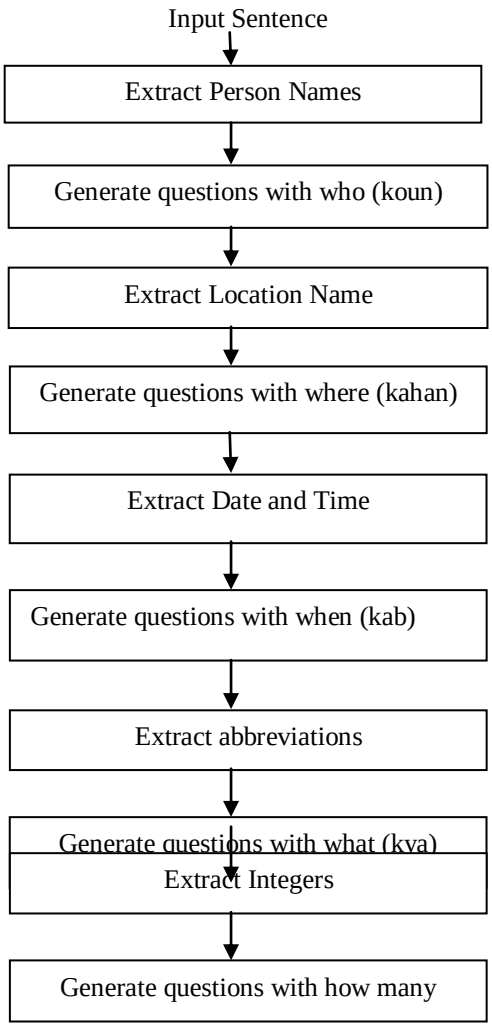

Figure 2: Flowchart of rules

4.2 Dictionary lookup approach

In dictionary lookup approach we have a corpus and this corpus defines various mappings. Dictionary lookup is the most natural way to store mappings. A dictionary is a data structure that directly is useful in word analysis. This data structure can be optimized for lookup. Lookup operations are usually simple and quick than the other approaches. Dictionary lookup approach is also used with the rule based approach. By using these two approaches many systems are designed to generate questions. These systems are evaluated with the help of three parameters Recall, Precision, F-Measure. Recall value is a value to total number of questions generated using system to the total number of questions generated with the help of human being. Precision is the value of total number of accurate questions generated by the system. F-Measure is the average of recall value to the precision. Table 1 shows the results of two different systems.

Table 1: Table of evaluation

Authors	**Garg et al. []**			**Singh et al. []**		
Question Type	**Recall (%)**	**Precision (%)**	**F-Measure (%)**	**Recall (%)**	**Precision (%)**	**F-Measure (%)**
ik~Qy (Kithe)	56.6	74.5	64.32	92.89	91.56	91.25
iks (Kis)	50.0	62.1	55.39	62.86	74.52	65.67
kI (Ki)	16.21	41.6	23.32	85.50	70.97	77.56
kdoN (Kado)	48.23	73.17	58.13	89.79	94.79	91.75
ikMny (Kinne)	52.33	51.7	52.01	91.24	98.30	94.64
ikauN (Kyon)	Not Done	Not Done	Not Done	82.01	95.09	88.06
ikvyN (kive)	Not Done	Not Done	Not Done	70.81	84.79	77.17
idSw (Direction based)	Not Done	Not Done	Not Done	90.68	88.22	89.43
Monetary expressions	Not Done	Not Done	Not Done	89.22	99.27	93.97

5. CONCLUSION

This paper presents the review to generate the questions automatically for Indian languages automatically. As discussed only rule based approach is used to generate the questions automatically for Indian languages from a given text. The system can generate the questions only starting with “who”, “what”, “where”, “when” etc words. So a lot of work has been left. The system can be improved by generating the deep questions also starting with

the "how", "why", "what-if", "what-if not" etc words. A hybrid approach is required to generate the questions automatically from a given text. The rule based approach can be combined with dictionary lookup approach and example based approach to obtain the hybrid approach. Also we can modify the existing systems by generating the multiple choice questions giving four options out of which only one is correct and other three are incorrect.

REFERENCES

[1] Itziar Aldabe, Maddalen Lacalle de Lacalle, Montse Maritxalar, Edurne Martinz, Larraitz Uria, 2006, ArikIturri: An automatic question generator based on corpora and NLP techniques, Springer-Verlag Berlin Heidelberg, pp. 584-594.

[2] Husam Ali, Yllias Chali, Sadid A. Hasan, 2010, Automation of question generation from sentences, Proceedings of the third workshop on question generation of University of Lethbridge, pp. 58-67.

[3] Yllias Chali, Sadid A. Hasan, 2012, Towards automatic topical question generation, Proceedings of Coling technical papers, pp. 475-492.

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

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

[6] Prashanth Mannem, Rashmi Prasad, Aravind Joshi, 2010, Question generation from paragraphs at UPenn: QGSTEC system description, International institute of information technology, pp. 1-8.

[7] Dr. P.Pabitha, M.Mohana, S.Suganthi, B.Sivanandhini, 2014, Automatic question generation system, IEEE International conference on recent trends in information technology.

[8] Sheetal Rakangor, Dr. Y. R. Ghodasara, 2015, Literature review of automatic question generation systems, International journal of scientific and research publications, pp. 1-5.

[9] Shriya Sahu, Nandkishor Vasnik, Devshri Roy, 2012, Prashnottar: a Hindi question Answering system, International journal of computer science & information technology, pp. 149-158.

[10] Parshan Singh, Rajbhupinder Kaur, 2014, A review on question generation system from Punjabi text contain historical information, International journal of computer science and mobile computing, pp. 185-189.

[11] Parshan sSingh, Rajbhupinder Kaur, 2014, Rule based question generation system from Punjabi text contain historical information, International journal of computer science and mobile computing, pp. 86-91.

[12] Liana Stanescu, Cosmin Stoica Spahiu, Acna Ion, Andrei Spahiu, 2008, Question generation for learning evaluation, IEEE proceedings of the international multiconference on computer science and information technology, pp. 509-513.

[13] Nguyen-Thinh Le, Nhu-Phuong Nguyen, Kazuhisa Seta, Niels Pinkwart, 2014, Automatic question generation for supporting argumentation, Springer, pp. 117-127.

[14] Xuchen Yao, Gosse Bouma, Yi Zhang , 2012, Semantics based question generation and implementation, International journal of dialogue and discourse, pp. 11-42.

[15] Shiqi Zhao, Haifeng Wang, Chao Li, Ting Liu, Yi Guan, 2011, Automatically generating questions from queries for community-based question answering, Proceedings of the 5th international joint conference on natural language processing, pp. 929-937.