Automatic Question Generation from Text: A Survey

Prof. Rajesh Lomte[1], Gaurav Borse[2], Manish Choudhary[3], Bhavit Dagade[4], Tanishq Dhussa[5]

[1]Pimpri Chinchwad College of Engineering

Abstract - *Students ask questions to satisfy their never-ending questions for getting knowledge. To perform competently assessment of students on their major concepts they learned from the study material. Preparing a set of questions for assessment can be time-consuming for teachers while getting questions from external sources like assessment books or question banks might not be relevant to content studied by students. An example, students ask questions to learn more from their teachers, teachers ask questions to help themselves evaluate the performance of the students, and our day-to-day lives involve asking questions in conversations. To perform competently assessment of students on their major concepts they learned from the study material. Automatic Question Answer Generation (AQAG) is the technique for generating a correct set of questions and answers from a paragraph, which can be text. Automatic question-answer generation (AQAG) is very important for many educational institutes but yet challenging problems in NLP. It is defined as the task of generating syntactically correct sound, semantically perfect and relevant questions and answers from several input formats like text, paragraphs. Question-Answer generation can be useful in many domains such as online assessment, e-tutoring sessions, chatbot systems, search engines and healthcare for analysing mental health. AQAG has got great consideration from researchers in a field of data science engineering.*

Keywords **- Natural Language Processing, Automatic Question Generation, NLTK, Pos-Tagger, NER**

1. INTRODUCTION

Generally, individuals ask the question to each other to asses or improve their knowledge Here we study recent progress in the generation of natural language questions by machine based on text passages provided as the input. Preparing a questions for assessment can be time-consuming for teachers. Getting questions from external sources like assessment books or question banks might not be relevant to content studied by students.

The machine understands the written language as human level. Questions are used to asses other individual knowledge or owns knowledge through information seeking behavior. Because most of the human knowledge is recorded as text, this would enable transformative applications. Here focus is more on developing the system that will generate the question from text rather than answering the system which answers the question based on questions asked. This study will focus on processes that will quality questions. The study comprises of various automatic question generation systems that generates various questions such as Gap fill questions i.e. Fill in the blanks type questions type questions, Generating Distractors for Multiple Type questions. Generally, Distractor Will Be Generated from WordNet of NLTK. The systems approach is focused generating quality questions by removing which are either not directly related to a topic or have no specific meaning.

Question-Answer generation can be useful in many domains such as online assessment, e-tutoring sessions, chatbot systems, search engines and healthcare for analyzing mental health. AQAG has got great consideration from researchers in a field of data science engineering. It is a very challenging task in spite of its functionality. It is difficult for teachers to perform so many tasks such as making questions and evaluating them for each student therefore such system will save much more time. Till now most of approaches were template-based approach where fixed set of patterns were predefined and from Gap fill questions were generated which had very less accuracy of generating meaningful questions

2. RESEARCH MOTIVATION

Motivation in building Automatic Question Generation System is that it will automate the current process of question generation by teachers which wastes a lot of time and sufficient time is not allotted for teaching or other activities that will help in overall development of student. Another thing is till now as per the survey no any highly accurate system is developed that can generate Semantically correct meaningful questions. So main motivation is to develop a system or a platform which can generate wide variety of questions such as Fill in the blanks, MCQ type questions and Wh-Type questions in one system. This system will be highly useful in educational institutions and in primary schools where easy questions are framed for students. Keeping in mind such vast uses of this system will highly beneficial for users who are using it.

3. SYSTEM COMPONENTS

(1) Input:

The input to our system is a paragraph which is taken from any website or any file. The paragraph contains different sentences which are useful for generating questions [5].

(2) Pre-Processing:

Every Sentence in a paragraph is not a question worthy. It is important to remove sentences containing URLS and Non-ASCII characters that are not important for generating questions. System will automatically find URLS and Non-ASCII characters and remove them. The paragraph is split into sentences depending upon terminating symbol or discourse-connective [5].

(3) Sentence Selection:

Identifying the most relevant sentences that can be useful for question generation is very important task. Sentences can be ranked depending upon their syntax and semantic correctness along with complexity level. It also important to Identify the types of sentences by extracting features from each of them. Depending on these features it selects all the important sentences on which questions can be generated. Following features are used to select the candidate sentences [6]:

First sentence: Usually gives information about a summary of paragraph.

Last sentence: Usually gives information about conclusion of paragraph.

Common tokens: This feature counts the words, only nouns and adjectives that the sentence and the title or the subtitle of the paragraph have in common

Length: This is the number of tokens words in the sentence.

Number of Nouns: This is the count of the number of tokens that are tagged as noun (NN, NNS, NNP, NNPS) by the POS tagger.

Discourse Connective: This feature examines the presence of discourse connectives in the sentences. Discourse

4. LITERATURE SURVEY

An AOG system can be implemented using three different methodology classified as Rule based Approach, Corpus based Approach and Template-based Approach [1]. In rule-based approach interdependencies between columns are taken in consideration for question generation. While in Corpus based approach text knowledge of the language is acquired using text corpus is taken into account. Template based approach use a well-defined template with placeholder for question formation.

According to the literature survey we noticed that question can be formed from structured data and unstructured data. Area of question generation form structured data is unexplored [1].

Templated based methodology is used to question formation from Structured data (Data Table) [1]. While working on structured data there is need to select suitable table column data and row data. For these pre-processing is carried out using storing and trimming concept. Storing means to store only those data table which contains master row and master column. While in trimming redundant data from the master row and column is removed. Further entity recognizer is used for formation of metadata to annotate data table. For these categorization and annotation technique are used. After this Tuple generator is used to get a random tuples (data) from table. Tuples can be generated in varies combination such as (Row, column, Data), (Column) and (Column, row, row) to form question of different variety and Varying complexity. These generated tuples act an input for question Generator phase. Question generator phase process tuples with varies templates and forms relevant wh-questions. This system showed 63.15% matching questions with human question writers.

connectives make a vital role in making the text coherent and so wh-type of questions can be easily generated using them [6].

Table 1.Question type for various discourse connectives

Discourse Connective	Wh-Question
Since	When
When	When
Because	Why
As a result	Why

(4) Key Selection:

Key selection can be done in two ways manually, or by a system using Stanford CoreNLP tagger. In this appropriate Wh-word is selected by studying Subject-Verb-Object and their relationship. It can be also done depending on basis of dependency-based patterns, semantic based transformation, Syntactic transformation and Discourse connective [5].

Selected answer is encoded the source sentence using the BIO

(Begin, Inside, Outside) notation. Hence this selected key can be further used to generate different questions.

(5) Question Generation:

Different types of Question can be generated by our system depending upon the choice of the user. Questions like Fill in the blanks, MCQ's, Wh-questions can be generated depending upon the requirement. After questions are generated, we can filter them on the basis of confidence score. The Un answerable questions can be removed using BERTH question filtering Algorithm [5].

Computational Intelligence Framework can be used to generate question along with rank determining module based on semantic correctness and complexity level [2]. A rule-based approach is used in Framework. In following AQG system text is been analyzed using tree object and classified into various noun, pronoun, adverb categories. Tree object is created using the Stanford parser. Further classification is done using regular expression.

(1) Fill in the blank question

The Key point for fill in the blank question, is to identify the potential gap for which blanks need to be created [2].

The produce discussed for identifying the gap is as follow:

1. Train the model with predefined questions and gap.
2. Use NLP to identify the verbs, pronoun, noun (Stanford NLP parser).
3. Eliminate the stop words from being used as a gap.
4. Give Priority to numbers/places for gap creation.

```
Constituent tree:

(S (NP Firewall)
   (VP isolates
       (NP (NP organization 's)
           internal net)
       (PP from
           (NP larger internet
               (VP , allowing
                   (NP some traffic)
                   (S (VP to
                          (VP pass)
                          ,
                          (S (VP blocking
                                 (NP others)))))
                   .)))))
```

```
                                        Question  \
0  _____ is someone with access right to the system
1                                  Insider is _____
2  Insider is _____ with access right to the system
3                     Insider is someone with _____
4       Insider is someone with _____ to the system
5     Insider is someone with access right to _____

                                  Answer  Prediction
0                                Insider           2
1  someone with access right to the system           1
2                                someone           2
3                access right to the system          2
4                            access right           2
5                              the system           2
0: bad question; 1: okay question; 2: good question
```

But this system has Challenges certain such as a lexical, paraphrasing, generation of distractor and relevance of pronoun [2].

While there is other system in which pre-processing is carried out along with ARKref tool for pronoun resolution [3]. Sentence selection is done using NER information and Predicate-Argument-Based Selection. While Key or Blank word identification is implemented through NER output or predicate argument. This system shows prominent result with accuracy of 84.73% but this system can be for topic-based fill in the blank question generation.

One of the major issues addressed in fill in the question is questions with ambiguity in answers.

For example, the question "_____ is someone with access rights to the system." May have various correct answers, which are not restricted to "Insider" only. Due to this one can follow an approach for MCQ formation using Concept of Distractor generator.

(2) Multiple Choice Questions

The produce discussed for MCQ formation is as follow:

1. Find potential keywords which is identified as the answer key answer key.
2. Provide keyword (answer key) input to distractor.
3. Use the output of distractor as possible choices of MCQ.

Distractor:

It processes given input keywords to generate possible synonyms and antonyms using help of NLTK dictionary and WordNet. Antonyms can directly be classified as the potential distractors. While synonyms generated are passed through the checker module, which further internally calls the usage checker using online resources to see similar sentences, which is employed in classifying the synonyms into relevant and irrelevant. Irrelevant synonyms are considered as potential distractors. While relevant synonyms are passed through semantic checking

Sentence selection methods of MCQ:

1]Text rank

The PageRank algorithm by Google Search uses to rank the selected sentences in the paragraph. The outline of Text **Rank**is as follows [11].

1. Each sentence in passage *P* is added as a vertex in the graph.

2. Calculate the similarity between every pair of sentences and use it as an edge in the graph.

3. Normalize the similarity values.

4. Run the PageRank algorithm until convergence.

5. Ranking the sentences based on their score.

6. Select top- *N* sentences as the summarization.

2] Multi-word phrase extraction (MWPE)

This is the second method for the selection of the sentences, the system used TextRank to extract key phrases instead of the highest-ranked sentences [11]. Then, the repetition of words that formed the key phrases was counted in each sentence of the passage. The important sentences were the ones that had the highest number of repetitions. The following is the outline of multi**-word phrase extraction**.

1. Each tokenized word is annotated with part-of-speech (POS) tags.

2. Filter the words (only leave those that are nouns or adjectives).

3. Adding the words to the graphs as vertices.

4. Adding an edge between words that co-occur within window size *W* words of each other.

5. Run the PageRank algorithm until convergence.

6. Ranking words based on their scores.

7. Picking top- *K* words as keywords.

8. Pairing the keywords together as a key phrase, if they are adjacent in the graph.

9. Calculating the occurrence of words that form the key phrases in each sentence.

10. Picking top- *N* sentences with the highest number of occurrences as the summarization.

3] Latent semantic analysis:

LSA deciding the meaning of a sentence using the word it has, and the meaning of a word using the sentences that contain the word [11]. Then, interlinkage between words and sentences came across with singular value decomposition (SVD). The outline of latent **semantic analysis**is as follows.

1. Extract terms from sentences

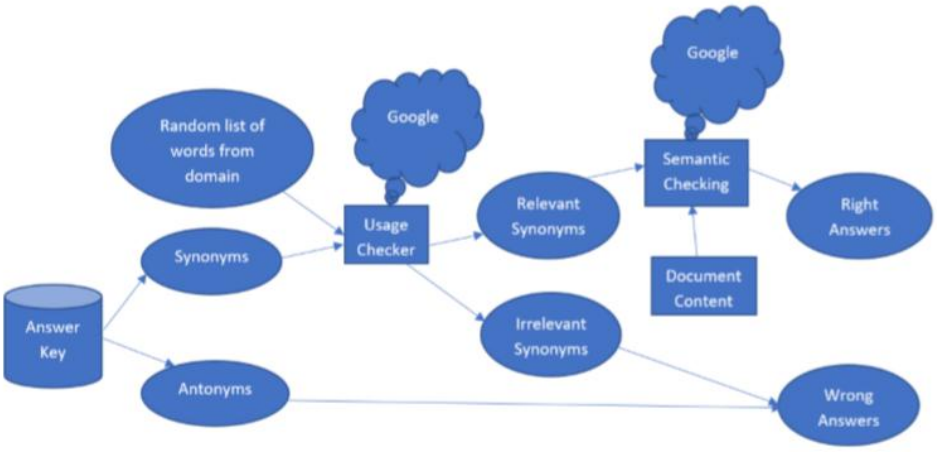
2. Building an input matrix with an approach on the terms as a binary representation

3. Compute singular value decomposition (SVD) on the input matrix

4. Compute ranking based on sigma and V_T from SVD

5. Picking top- *N* sentences as the summarization.

Method	**Averaged cosine similarity**	
	No stop words	**With stop words**
LSA (tf-idf)	0.48 1	0.53 6
LSA (binary representation)	0.71 9	0.72 4
MWPE (w=5,k=7)	0.78 3	0.74
TextRank (Jaccard)	0.69 4	0.69 8
TextRank (cosine)	0.66 9	0.59 3



(3) Wh-Question

The Wh-question can be generated in two ways. First is from the selected pivotal answer and second is from types of sentence identified in sentence selection section [5].

For generating wh-question depending upon pivotal answer, the pivotal answer is encoded in the source sentence using the BIO notation, and train a sequence-to-sequence model augmented with dynamic dictionary, copy mechanism and global sparse-max attention. Our question generation module consists of a paragraph encoder and a question decoder. The encoder represents the paragraph input as a single fixed-length continuous vector. This vector representation of paragraph is passed to the decoder with reusable copy mechanism and sparse max attention to generate different questions [5].

For generating wh-question depending upon sentence identified, now we will identify different types of classes using NER (Name Entity Recognizer) like [6]

Human: This includes the name of a person.

Entity: This includes animals, plant, mountains and any object.

Time: This will be any time, date or period such as year, Monday, 11 am, last week, etc.

Location: This will be the words that represent locations, such as country, city, school, etc.

Count: This class will hold all the counted elements, such as 22 men, 8 workers, measurements like weight and size, etc.

Organization: Organizations which include companies, institutes, government, market, etc.

Once the sentence words have been classified to coarse classes, we consider the relationship [6] between the words in the sentence.

Table 1. Sample rules for generating questions based on the classes

Subject	Object	Preposition	Question type
H	H		who,whom, what?
H	H	L	who, whom, what, where?
L	H		where, when?
C	C		How many?

Here H=human/person, L=Location, O=Organization, C=Count, T=Time, E=Entity.

For example: Sachin plays cricket at 5 am

Sachin is a subject of coarse class Human

Cricket is an object of type Entity

At 5 am is a preposition of type Time

Sample generated questions based on the rule "Human Entity Time" will be:

Who plays cricket?

Who plays cricket at 5 am?

What does sachin play?

When does sachin play cricket ?

5. COMPARATIVE STUDY

Sr no.	Paper	Procedure	Question Types
1	Automated Question Generation Tool for Structured Data	It uses the Template based approach. It takes input as the Data table and generates random tuples and passes to question generator.	Wh-question for structured data
2	Computational Intelligence Framework for Automatic Quiz Question Generation	It uses Rule-based approach. It makes use of production rules, LSTM neural network model and	Fill in the blank questions MCQ Wh- question
3	Automatic Generation of Fill-in-the-Blank Questions from History Books for School-Level Evaluation	It deals with question formation for topic-based content. It is based on the corpus-based approach.	Fill in the blank questions
4	Para QG: A System for Generating Questions and Answers from Paragraphs	This paper mainly focuses on Wh-question generation depending upon pivotal answer selected by user or system. BERTH –based question filtering to remove un-answerable questions. Grouping of questions having similar answer and selecting one out of them.	Generate Wh -questions for the paragraph passed paragraph and group similar type of questions.
5	Automatic Question Generation from Paragraph	This focuses on Wh-question generation depending upon identified class of sentence using stanfordNER is process over subject and object to obtain class. This class is further used to generate question by studying Subject-Verb-Object relation.	Generate Wh -questions for selected pivotal answer.
6	Machine Learning Approach to the process of Question Generation	This follows the data driven approach, which is realized by extracting features from text and choosing most suitable class of questions (on basis of the features). Each class of sentences has assigned set of possible questions and questions are chosen based on similarity of features.	Generates Factoid question, generates Gap fill, Wh-type Questions
7	A system for Generating Multiple Choice Question: with novel approach for sentence Selection	This paper mainly focuses on MCQ Generation for sports specifically for cricket where they have specific pattern of MCQ defined and based on that they generate the new question. This not a general and more specific towards question generation for specific domain	Generates MCQ for Cricket Domain. Gap-fill Question, Multiple Choice Question
8	Gap-fill Question Generation	The system automatically extracts the informational sentences from the comprehension or paragraph, and generates gap from it, by using first blanking keys from the sentences and then finding the similar words for the distractors for these keys.	Gap-fill Question, Multiple Choice Question
9	Automatic Question Generation for Intelligent Tutoring Systems	The system consists of two primary sub-modules: Knowledge Base Generation and the generation of MCQs using the Knowledge Base. Inverse Document Frequency (IDF) score is used to choose the word that will serve as a blank (a missing word) for a given MCQ. Distractors are generated by using context similarity derived with Paradigmatic relation discovery on the self-made corpus and dictionary.	Generates Multiple Choice Questions
10	Automatic Question-Answer Pairs Generation from Text	The system presented a rule-based automatic question generation system for reading comprehension. System uses multiple methods to select noteworthy sentences in a paragraph, then use named entity recognition and constituent parsing to generate possible question-answer pairs. The sentence is then transformed to an interrogative form based on a set of rules and possible answers.	Generates Interrogative Questions

6. CONCLUSION

The paper present the review of various approach of automatically questions and answers generation from text paragraph. Most of the algorithms mainly based on natural language processing. The work has basically been donated for the English language for the generation of questions which can be categories as Multiple Choice, Factoid andGap-fill questions. this motivational field of Automatic Questions Answers Generation (AQAN) has always been the section needs of continuous improvement like we can consider FAQ, TRUE/FALSE questions and answers generated in the future.

REFERENCES

[1]A. Shirude, S. Totaia, S. Nikhar (Author), Dr. V. Attar (CoAuthor) and J. Ramanand (CoAuthor) - "Automated Question Generation Tool for Structured Data" - 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCl)

[2]Sumeet Pannu, Aishwarya Krishna, Shiwani Kumari, Rakesh Patra and Sujan Kumar Saha – "Automatic Generation of Fill in the Blank Questions From History Books for School-Level Evaluation" - Springer Nature Singapore 2018

[3]Akhil Killawala, Igor Khokhlov, Leon Reznik – "Computational Intelligence Framework for Automatic Quiz Question Generation" - 2018 IEEE

[4] Rakesh Patra, Sujan Kumar Saha – "A hybrid approach for automatic generation of named entity distractors for multiple choice questions" - Springer Nature 2018

[5] Vishwajeet Kumar1,3,4, Shivanand Muneeswaran2, Ganesh Ramakrishnan3, and Yuan-Fang Li4 – "Para QG :A System for Generating Questions and Answers from Paragraphs".

[6] Dhaval Swali1, Jay Palan2, Ishita Shah3– "Automatic Question Generation from Paragraph"- International Journal of Advanced Engineering and Research Development.

[7] Blšták M., Rozinajová V. (2017) Machine Learning Approach to the Process of Question Generation. In: Ekštein K., Matoušek V. (eds) Text, Speech, and Dialogue. TSD 2017.

[8] A System for Generating Multiple Choice Questions: With a Novel Approach for Sentence Selection10.18653/v1/W15-4410

[9] Agarwal, M., Mannem, P.: Automatic Gap-fill Question Generation from Text Books. In: Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 56–64 (2011)

[10] Riken Shah, Deesha Shah, Prof. Lakshmi Kurup – "Automatic Question Generation for Intelligent Tutoring Systems" - 2017 IEEE

[11] Holy Lovenia, Felix Limanta, Agus Gunawan – "Automatic Question-Answer Pairs Generation from Text"- researchgate.net/publication/328916588(2018)