This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2889100, IEEE Transactions on Learning Technologies

Automatic Multiple Choice Question Generation from Text : A Survey

Dhawaleswar Rao Ch and Sujan Kumar Saha

Abstract—Automatic Multiple Choice Question (MCQ) generation from a text is a popular research area. MCQs are widely accepted for large-scale assessment in various domains and applications. However, manual generation of MCQs is expensive and time-consuming. Therefore, researchers were attracted towards automatic MCQ generation since the late 90's. Since then, many systems have been developed for MCQ generation. We perform a systematic review of those systems. This paper presents our findings on the review. We outline a generic workflow for an automatic MCQ generation system. The workflow consists of six phases. For each of these phases, we find and discuss the list of techniques adopted in the literature. We also study the evaluation techniques for assessing the quality of the system generated MCQs. Finally, we identify the areas where the current research focus should be directed toward enriching the literature.

Index Terms—Automatic Question Generation, Multiple Choice Questions, Natural Language Processing, Text Analysis

1 Introduction

Question is an essential tool to assess the knowledge or understanding of a learner. Assessment is crucial in learning and question is essential for assessment. Multiple choice question (MCQ) is the most widespread form of a question for various levels of assessment. MCQs have many advantages including quick evaluation, less testing time, consistent scoring, and the possibility of an electronic evaluation. Many examinations use MCQ based question papers through a computerized environment. However, manual preparation of MCQs is time-consuming and costly. Therefore, the research community devoted substantial effort to find the techniques for automatic generation of MCQs. The research on automatic MCQ generation started at least 20 years ago. As an early attempt, we find the system developed by Coniam David in 1997 [1]. Since then, many MCQ generation systems have been developed in various languages and domains, and for various applications.

Since there exists strong literature on MCQ generation, a survey of the developed systems is necessary to leverage more research in this area. With this necessity, we perform a systematic review of the literature on MCQ generation. In this review, we cover the majority of the systems developed in the last two decades starting from 1997. The primary objective of this review is to summarize the techniques for MCQ generation. For that, we first tried to establish a generic workflow for the task. The workflow we established here consists of six phases. A number of techniques have been adopted by the developers for the execution of the individual phases. We categorized those techniques and exposed their relative strengths. Evaluation of a system is another important issue. Through this review, we also identified the approaches for the evaluation of the computer-generated MCQs. We found that there is no standardized evaluation technique, evaluation metrics, or benchmark data for MCQ system evaluation. Our final objective was to study the limitations of the existing systems and discover the next set of challenges that the field should focus on now.

The paper is organized as follows. A generic overview of MCQ type questions is summarized in §2. §3 presents the objective of the review. Then, §4 summarizes the methodology used in the systematic review process through the comprehensive data collection (§4.1), gathering of information (§4.2), and analysis of the research (§4.3). The subsequent sections contain the findings of the survey. §5 illustrates the major phases for the development of an automatic MCQ generation system. Summarization and grouping of the techniques applied to the individual phases are presented in six subsections. In §6, the system evaluation strategies are discussed. Finally, §7 concludes the paper.

2 Multiple Choice Questions

Multiple choice questions are a popular form of assessment in which the respondents select the best possible answer out of a set of choices. MCQ is often treated as a subcategory of objective type questions [2]. Primarily this type of questions has a specific scope, i.e., a question deals with the knowledge embedded in a very small sized text, often a single sentence. However, there exist a few systems that take more than one sentences into account to form a question ([3], [4], [5]). Some authors named MCQs as factual questions ([6], [7], [8]). Because these questions deal with the fact (or knowledge) embedded in the text. From this point of view, MCQ generation requires a non-fiction text that contains factual information rather than opinion. In education, factual questions are useful to assess the recall of specific information or the student's knowledge of the information embedded in the text.

• *Dhawaleswar Rao Ch and Sujan Kumar Saha are with the Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi 835215, India. E-mail: dhawaleswarrao@gmail.com, sujan.kr.saha@gmail.com.*

A set of choices or alternatives is the primary requirement for labeling a question as MCQ. The actual answer to the question must be included in those alternatives. Due to this property, the MCQs are also called as cloze questions in the literature ([9], [10], [11], [12], [6]). When we consider the nature of the question sentence of an MCQ, it holds multiple formats. Most popular among those are, fill-in-the-blank and wh-question.

A MCQ is composed of three core components. These are,

- stem
- key
- distractor.

The *stem* is also known as *item* or *question sentence*. Fundamentally this is the sentence from which the question is formed. So, a question without the alternatives can be treated as a stem. It can be in an assertive or interrogative form. The *key* (also named as *target word*) is the correct answer of the question. *Distractors* are the set of wrong answers or choices given along with the correct answer to befuddle the examinee.

For example, consider the following MCQ.
The largest planet in our solar system is _____.

1) Earth
2) Jupiter
3) Venus
4) Neptune.

The stem of this MCQ is "The largest planet in our solar system is". It could be presented in question form: "What is the largest planet of our solar system?". Four choices are given in the question. Among these, the correct answer is 'Jupiter'. So, this is the key. The rest three choices are the distractors.

Multiple choice questions have several advantages over other question categories. Most importantly,

- Quick evaluation. On the other hand, the evaluation of descriptive questions takes significant time.
- Scope of non-human evaluation. Electronic or automatic evaluation is possible for MCQs.
- Reliable and consistent scoring. On the other hand, scoring of opinion related questions can vary from evaluator to evaluator.
- Factors irrelevant to the assessed material do not come into play; for example, handwriting and clarity of presentation do not affect the marking.
- Test time is less.

Though, the MCQ type questions have a few weaknesses too. For example,

- MCQs can deal only the lower order knowledge levels.
- Guess can play a major role in answering the question.
- Partial knowledge has no credit. The student might know a partial amount of facts regarding the question, but due to a few unknown facts he is unable to answer it.

However, multiple choice tests are often chosen, not because of the type of knowledge being assessed, but because these are more affordable for testing a large number of students.

3 Research Motivation and Objectives

Multiple choice questions are popular for large-scale assessments. However, preparation of a set of good MCQs takes time and requires an in-depth knowledge of the subject and construction skill. Therefore, educational technology and natural language processing research community were attracted to the possibilities for automatic MCQ generation. The research on automatic MCQ generation started at least 20 years ago. A substantial amount of effort has been devoted since then. MCQ systems have been developed for a variety of applications, starting from language learning or vocabulary testing to assessment in active learning or e-learning framework ([13], [6], [14], [15], [16], [10], [11], [17], [18], [19], [20]). Also, MCQ systems have been developed in various languages and domains. In the literature, we found MCQ systems exist in many languages including English ([21], [22], [23], and many more), Portuguese ([24], [25]), Basque ([26], [27], [28]), French [29], Russian [30] and Chinese ([31], [7], [32]). Also automatic MCQ generation has been performed in several domains, including, language learning ([5], [33], [34]), grammar & vocabulary learning ([35], [36], [37]), science ([38], [39]), history ([40], [41]), general science ([42], [43]), biology & medical ([44], [45], [46], [47], [48], [9], [49]), technology ([50], [51], [52], [53], [54], [55]), generic domain ([56], [32], [57], [58]), e-learning & active learning ([59], [60], [61], [62], [11], [10], [63]), sports & entertainment ([64], [65], [66], [12], [67], [68], [69]) etc.

Automatic MCQ generation is still an active research area. Researchers have been aiming to develop MCQ systems for new and emerging applications or focusing on the unsolved challenges of this area. As a result, the literature on automatic MCQ generation is quite strong. Since there exist many systems for MCQ generation and different techniques have been devised, a survey of these systems is required. Knowledge of the existing approaches is essential to design a new approach. A survey article can provide the knowledge in a summarized form and it also helps to set a vision for work in this area.

With this need, we survey the existing literature on multiple-choice question generation. This survey primarily intends to accumulate the methods used in automatic MCQ generation, methods used for system evaluation, and a critical analysis of those methods. The specific research objectives are listed below.

1) Study the need and application of the MCQ generation systems.
2) Learn the generic workflow that takes an input text and generates a set of MCQs as output.
3) Find the techniques that have been applied for developing individual phases of the existing systems.
4) A critical analysis of those techniques.
5) Find the approaches that have been used for assessment of computer-generated MCQs.
6) Discover the research trends in this area and to point out the next challenges that the field should focus on now.

4 Review Methodology

A systematic review is a summary of studies addressing a clear question, using systematic and explicit methods to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2889100, IEEE Transactions on Learning Technologies

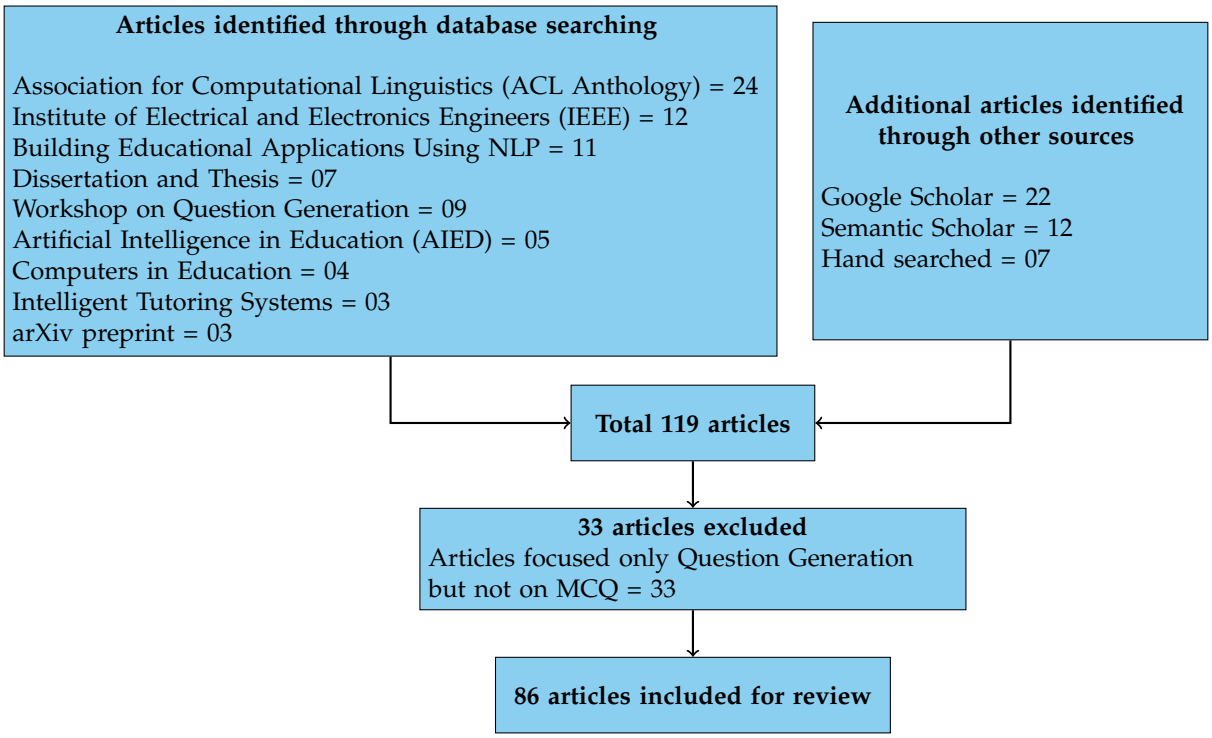

Fig. 1. Literature Search: a pictorial representation of the procedure

identify, select, and critically appraise relevant studies, and to collect and analyze data from them[1]. In order to ensure a systematic review process, this study followed the following steps for conducting a systematic review:

1) Formulate the research problem.
2) Literature Search.
3) Information Gathering.
4) Analyze and integrate the outcomes of research.

The first step in conducting a systematic review is to formulate the research problem, which has been specified in the previous section. Rest of the steps and outcome of the review are explained below.

4.1 Literature Search

The second step in a systematic review is to search the literature. This is an interdisciplinary research area. First, we started the exploration by searching journal articles but we found certain conferences and workshops are quite popular and cover MCQ related articles. Hence, we included both the journal and conference articles in this survey. We searched various databases to ensure comprehensive data collection. The starting point involved searching for a combination and variation of the keywords like automatic MCQ generation, question generation, MCQ generation, multiple choice questions, cloze questions, and fill-in-the-blank questions. Then we review the results against the following inclusion criteria.

- The primary focus of the paper is the automatic generation of multiple choice questions or fill-in-the-blank questions.
- The input to the system is unstructured text.
- The MCQ generation and the individual modules employ some automation. Studies that are entirely based on manual steps or random selection are not included.
- The paper should not focus on static or manual MCQ bank creation and its access related issues. Also, correction of MCQ answers related works are not included.
- The paper should not focus only on question formation (like the conversion of assertive sentences into interrogative form).
- The paper is written in English.
- The full-text of the article is publicly available or available through the institutional library subscriptions.

A total of 86 papers qualified these criteria and are included in this study. Fig.1 summarizes the literature search and paper selection policy.

4.2 Information Gathering

For each included article, the following data were extracted: Title, Journal or Conference Name, Year, Volume and Page no., Authors, Country, Question Type (MCQ, gap-fill, basic QG), Language, Domain, Application, Learning Outcome, Overall workflow (i.e., number of phases and what are those), Pre-processing, Sentence selection strategy, Key selection strategy, Distractor generation strategy, Post-processing, Metrics used for Evaluation, Who evaluated, Dataset, System accuracy. The basic analysis started using this extracted data.

4.3 Research Analysis

We analyzed the articles to satisfy the objectives of this survey. Our first objective is to establish a workflow for the task. So, we reviewed the articles to find a workflow.

1. http://getitglossary.org/term/systematic+review

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2889100, IEEE Transactions on Learning Technologies

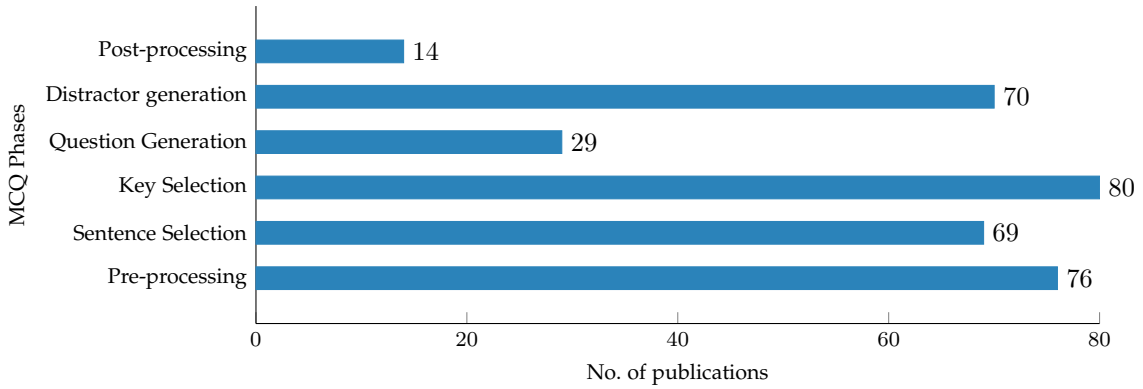

Fig. 2. Availability of the individual phases in 86 articles in consideration

Among these eighty-six research articles, 76 follow a workflow to generate the questions. Rest does not follow any specific strategy. Then we focused on the individual phases of the process. There we found, 76 articles employed one or more pre-processing steps. However, certain articles did not explicitly consider those as pre-processing. Similarly, 69 articles considered sentence selection, 80 articles presented key selection strategies, 29 systems performed question formation, 70 articles included discussion on distractor generation, and 14 systems applied post-processing steps. A summary of these findings is shown in Fig.2. These values indicate that four phases, namely, pre-processing, sentence selection, key selection and distractor generation are essential for automatic MCQ generation. Then we studied these articles to find the techniques used for implementation of the individual phases. Our findings are discussed in the subsequent sections.

5 Discussion on the Approaches for MCQ Generation

Now we discuss the approaches that have been devised for the development of the systems for automatic MCQ generation. We found that the researchers were primarily motivated by the methodology expected to follow for manual preparation of MCQs from a text. For the manual preparation of MCQs, the person first needs to acquire the information embedded in the input text. As one MCQ primarily demands one informative sentence, he also identifies the sentences that contain any questionable fact or information. The next task is to identify the word or phrase that acts as the key. Then he forms a suitable question from the sentence where the key becomes the answer. The final task is to pick a few distractors by analyzing the input text or a broader context.

The development of an automatic MCQ generation system might follow a similar strategy. The task is divided into multiple phases. Although the number of phases and the overall strategy slightly varies from system to system, most of the systems follow a generic workflow. Fig.3 shows the workflow of a system that contains six phases: (1) Pre-processing, (2) Sentence selection, (3) Key selection, (4) Question generation, (5) Distractor generation and (6) Post-processing. We discuss below the techniques used for the development of these individual phases in the literature.

5.1 Pre-processing of Input Text

Several pre-processing steps have been performed by the researchers for automatic generation of MCQs from a text. We summarize below the techniques.

- **Text Normalization:** Normalization refers to a conversion of the input text into the required format and removal of unnecessary content from the text [70]. The requirement of pre-processing is largely reliant on the domain and applications. Therefore, the technique to be employed here depends on the necessity of a particular task. Various types of text normalization and sentence normalization have been used in [71], [72], [73].
- **Structural Analysis:** This step identifies the chapters, sections, subsections, paragraphs, and other relevant tags in the text. Some authors take help of such structural information in MCQ generation. Additionally, as the MCQs deal with text portion only, removal of not text content becomes essential ([74], [6], [75], [16], [19], [76], [1], [77], [15], [78]).
- **Sentence Simplification:** One MCQ needs one sentence that contains single questionable fact. Long sentence might contain certain clue that help in identifying the correct answer. Sentence simplification involves converting complex and compound sentences into simple sentences. This is primarily done by using the external system or through parse structure analysis [79]. Sentence simplification has been performed in various systems; like, [80], [48], [75], [68], [28], [81], [82].
- **Lexical Analysis:** It involves splitting up the document text into a stream of words, symbols, and numbers. Additionally, inflections might affect various individual modules of the task. Therefore, stemming [83] that finds the root form of the word is performed in several systems. For example, [71], [72], [84].
- **Statistical Analysis:** Several systems used certain word level analysis including word frequency, n-gram frequency, Term frequency * inverse document frequency (Tf*Idf) [85], and co-occurrence statistics in various modules ([6], [16], [76], [49], [47], [69], [1], [77], [86], [87], [14], [88]). However, many of these articles did not explicitly mention these as a pre-processing step.
- **Syntactic Analysis:** Various levels of syntactic analysis including parts-of-speech tagging (assigning parts-of-speech of each words), named entity recognition

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2889100, IEEE Transactions on Learning Technologies

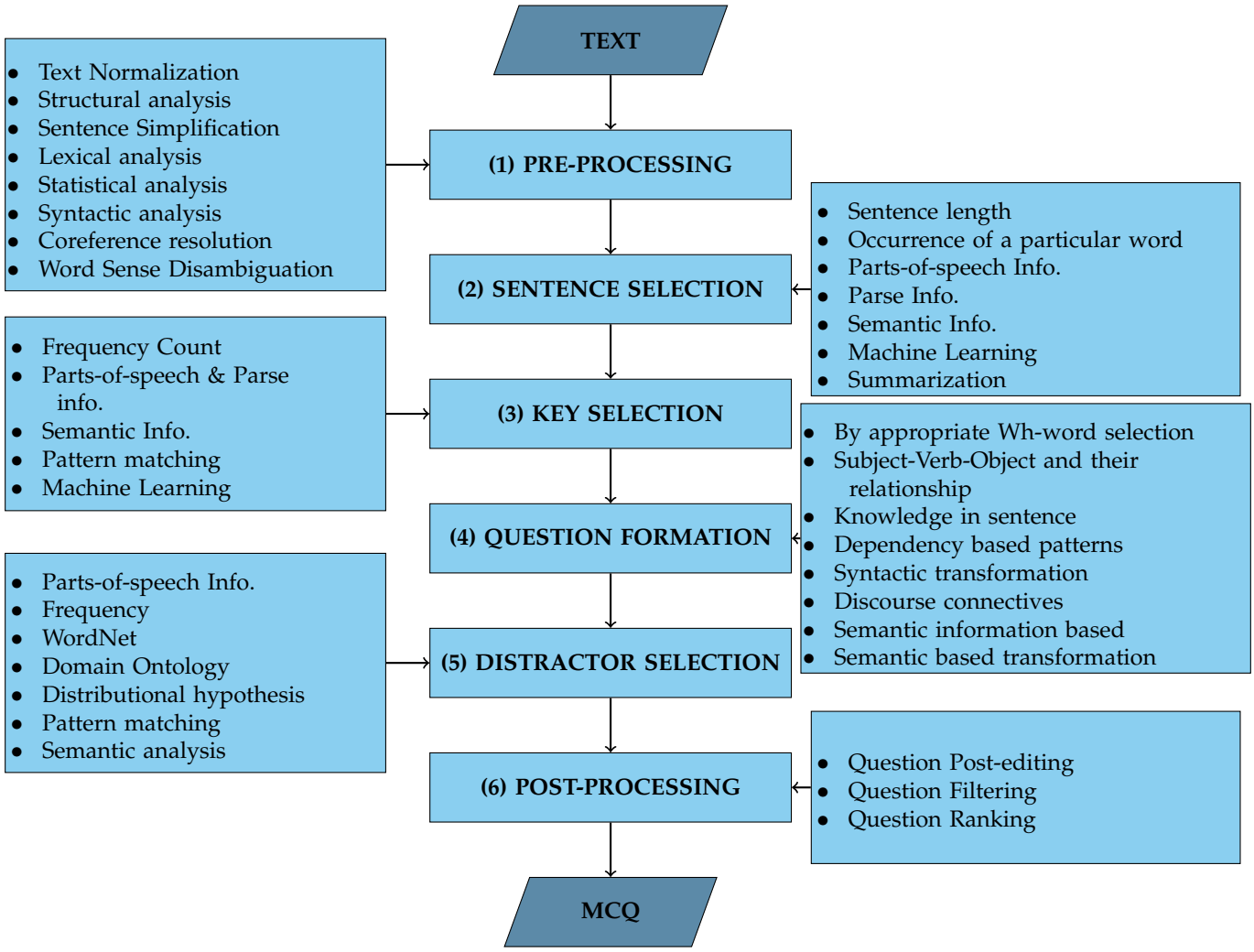

Fig. 3. The flowchart of the question generating system

(finding names of predefined categories from a text), syntactic parsing (generation of parse structure of the sentence) have been used in several systems. For example, [6], [75], [19], [76], [89], [17], [49], [67], [68], [1], [15], [87], [18].

- **Coreference resolution:** Generally, pronouns do not act as the subject of a question. Pronoun resolution is an important pre-processing step that aims to map the pronouns to their corresponding nouns. Coreference resolution has been used in [12], [90], [3] and some other systems.
- **Word Sense Disambiguation (WSD):** This is the task of identifying the exact sense of a word in a given sentence or text [91]. We find the use of this pre-processing step in several systems including [6], [76], [87], [14].

Our Observation on Pre-processing: Type of pre-processing required to develop an MCQ system depends on the nature of the input text and demand of the subsequent modules. For example, if the input text is collected through web crawling, then the document might contain some amount of noise and unnecessary content. Then text cleaning becomes an essential step. Again, we observed, the systems that use Wikipedia documents as input, often demand a sentence simplification step to handle the long and complex sentences. Some authors performed certain statistical analysis, applied basic natural language processing tools but these are not explicitly mentioned as a pre-processing step. However, we feel these are better to mention as pre-processing to avoid confusion to the readers.

The educational text also contains mathematical expressions, equations, figures, tables, and bullet points. The existing MCQ literature mostly avoided that information and focused only on the text portions. So, during the pre-processing phase, those systems often extracted the text portion by removing other contents. However, an MCQ generation system should have the capability to handle this information embedded in the text. Future research on MCQ generation should focus on the generation of MCQs from a hybrid educational text.

5.2 Sentence Selection

Each sentence of a text is not able to generate a valid question. Only the sentences that contain a questionable fact, can act as a candidate for generating MCQs. Therefore, sentence selection plays a major role in automatic MCQ generation task. Several approaches have been used in the literature

for selecting informative and questionable sentences from a text. We summarize below those approaches.

- **Sentence Length:** It denotes the number of words in the sentence. A very short sentence may not contain sufficient information for question generation. Again, a very long sentence often contains multiple facts and relations that create difficulty in the task. Sentence length is used as sentence selection criteria in several articles including [75], [16], [49], [68], [88], [24], [57], [25].
- **Occurrence of a particular word:** Occurrence based filtering method, that checks availability of a particular word or n-gram in the sentence, is applied by several researchers; for example, [86], [14], [76], [92], [93], [25] etc.
- **Parts-of-speech Information:** Aldabe, Maritxalar, and Mitkov [88] used occurrences of a verb and verb form patterns along with the length of the sentence; Lin, Sung, and Chen [89] searched the occurrence of adjective-noun pairs for sentence selection.
- **Parse Information:** Mitkov and Ha [87], Mitkov, Ha, and Karamanis [6] used subject-verb-object (SVO) structure based filtering for sentence selection. Majumder and Saha (2014 & 2015) used a parse tree matching based algorithm for informative sentence selection. They prepared a set of reference parse tree structures from the existing MCQs, and if a sentence from the input text matches any of those structures, then it is selected.
- **Semantic Information:** Semantic information embedded in the text has also been used as a criteria for sentence selection. Fattoh [58] used feature extraction through semantic role labeling [94] and named entity information for sentence selection. Araki et al. [3] used various semantic processing including coreference resolution [95], paraphrase detection [96], and extraction of a relationship between the concepts.
- **Machine Learning:** Several machine learning algorithms have been used in the task. For example, Naive Bayes by Hoshino and Nakagawa [77], Support Vector Machine by Correia et al. [24], Ranking Voted Perceptron by Goto et al. [10] & [11], Neural Network by Kumar, Banchs, and D'Haro [46], Counter Propagation Network-based classification by Bednarik and Kovacs [71].
- **Summarization:** Kurtasov [30] adopted a summarization based approach for sentence selection. Shah [90], Narendra, Agarwal, and Shah [12] also used an extractive summarizer, named MEAD [97], for sentence selection.

Discussion: Many methods have been employed in the literature for sentence selection. However, determining the most appropriate technique for sentence selection in a particular MCQ task is quite difficult. Sometimes it depends on the application too. For example, in a vocabulary assessment task, parts-of-speech (POS) or frequency based simple methods are sufficient. Again, in certain domains where names are dominating, the occurrence of named entities can be a suitable method. However, in many applications, a single technique becomes inadequate to select informative sentences accurately. Hybridization of multiple techniques becomes essential there. In the literature we also find a few systems that use a hybrid approach for sentence selection; for example, [75], [49].

The existing approaches to MCQ generation primarily focused on generating questions from a single sentence. However, the text might contain certain facts that are expressed through multiple sentences. While studying the human-generated MCQs in different domains, we found that many MCQs deal with such multi-line facts. However, only a few attempts have been taken for the automatic generation of such questions. For instance, Araki et al. [3] proposed an approach for generating MCQs from multiple sentences. Therefore, we feel that the research on automatic MCQ generation should focus on MCQ generation from multi-line facts. The primary difficulty in MCQ generation from multi-line fact arises in establishing the inter-sentence relationship and knowledge flow in these sentences.

5.3 Key Selection

It is obvious that all the words in an informative sentence cannot serve as the key. Therefore, the key selection is an essential step that determines the word, n-gram [98] or phrase in the selected sentence that will be blanked out. We discuss below the approaches used in the literature.

- **Frequency Count:** Frequency count of the words has been used as a selection metric in a number of MCQ systems ([1], [86], [87], [6], [49]). However, only the frequency count does not provide sufficient clue; therefore, certain additional information is also used along with the frequency to make the final selection. Sometimes, Tf*Idf has been used instead of simple term frequency ([17], [89]).
- **Part-of-speech and Parse Information:** It is observed that a particular part-of-speech or parse category becomes dominating as a potential keyword in some specific domains or applications. For example, Sumita, Sugaya, and Yamamoto [16] selected the verbs as keywords; Lin, Sung, and Chen [89] nominated the sense association among the adjectives as the correct answer; Lee and Seneff [99] selected a preposition as the key. Aldabe et al. [18], Agarwal and Mannem [49], Mitkov, Ha, and Karamanis [6], Chen, Liou, and Chang [15], Gates et al. [100] also used part-of-speech or parse information for key selection.
- **Semantic Information:** Semantic information has been used in multiple ways. Sung, Lin, and Chen [101] used a semantic network structure to represent the relationship among the players, actions, and attributes; and used this relationship to select the key. Fattoh [58] used a predicate extraction based approach for key selection. Afzal and Mitkov [48] extracted semantic relations between key concepts in a text and generated questions from such semantic relations. Liu et al. [76] applied word sense disambiguation based method for selecting keywords; Papasalouros, Kanaris, and Kotis [41] generated the key from property instances in an ontology.
- **Pattern Matching:** Pattern matching is another technique for keyword selection. Chen, Liou, and Chang

[15] extracted the sentences having similar structural feature and identified the common patterns, primarily based on the verbs. Those patterns helped in identifying the keywords. Gates et al. [100] used syntactic patterns based on the parse structure of the sentences to find the keys.

- **Machine Learning:** We find a few works where machine learning is used for key selection. Hoshino and Nakagawa [77] generated verbs or part of idioms or adverbs as keywords by utilizing machine learning techniques. Goto et al. [10] & [11] used a conditional random field classifier for estimating the key.

Our Observation on Key Selection: Like in the sentence selection, the suitability of a technique for key selection also depends on the domain and application. Early systems primarily relied on basic statistical and syntactic information. However, the recent trend is to use semantic information or machine learning. For the generation of a question from a sentence, the information embedded in the sentence and semantic relationship among the entities play a big role. So, we feel semantic level analysis based approaches are quite generic and applicable to diverse domains.

5.4 Question Formation

After key selection, the next task becomes a transformation of the declarative sentence into the interrogative form. However, in the literature, we found that this step has been ignored in many MCQ systems. If the transformation is not done, then the sentence remains in its original form, and a blank replaces the key. As a result, it becomes a fill-in-the-blank type question with distractors. However, we found several works containing the transformation from the declarative to an interrogative sentence. Rule or pattern based methods are dominating here. The approach of defining the rule varies from system to system. These are summarized below.

- **By Appropriate Wh-word Selection:** Several works attempted to identify the appropriate wh-word and formed the question accordingly. For example, 'what' for a thing, 'who' for people, 'where' for location, 'when' for time, and so on. Majumder and Saha [68] analyzed the parse structure of the target sentence to identify the key and position. Then they identified the suitable wh-word and formed the question using rules.
- **Subject-Verb-Object and their Relationship:** Mitkov and Ha [87], Mitkov, Ha, and Karamanis [6] used the structure of the target sentence (SVO or SV), terms occurred in the sentence, their positions, and types to define the rules.
- **Knowledge in Sentence:** Pabitha et al. [84] also used a rule-based approach. Their rules were based on the type of knowledge embedded in the sentence. They used various type of knowledge labels. For example, concept, definition, example, calculation, procedure, and result. These knowledge labels escorted the rules. For example, if it is *definition* then the question is 'What is meant by *X*?'; if *procedure* then 'How do you perform X?'.
- **Dependency Based Patterns:** Afzal and Mitkov [48] used dependency based patterns for question formation. From the dependency tree of the sentence, they identified the main verb and the portion that will be asked as the question. Then they removed the unnecessary portions and selected appropriate wh-word to transform the sentence into a question.
- **Syntactic Transformation:** Heilman [72] identified the answer phrases and performed syntactic transformation to generate questions. Their major steps included: marking of unmovable phrases, generation of possible question phrases using a rule-based approach, decomposition of the main verb, subject-auxiliary inversion, removal of answers, and insertion of question phrases.
- **Discourse Connectives:** Shah [90] used discourse connectives to perform transformations on the content to get the final question. They used discourse relations (e.g., temporal, causal, elaboration, contrast, result) to determine the question type and question word.
- **Semantic Information Based:** Lindberg et al. [102] developed a template-based framework where they used semantic role labeling to identify patterns to generate questions. Mazidi and Nielsen [103] also used semantic role labeling based framework for question formation.
- **Semantic Based Transformation:** Yao and Zhang [104] proposed a semantics-based approach by transforming declarative sentences to interrogative forms.

Discussion on Question Generation: Question generation from a text is another related research area where a large amount of research effort has been devoted in past few years. The objective of that area is to generate a question from an input text. However, in MCQ, one informative sentence is picked first, then the key is identified, and finally, it is converted into a question form based on the key. So, here the question formation step is fundamentally a conversion of the sentence from assertive to interrogative form using the key as the center. When we focus on MCQs where the sentence and key are pre-decided, then rule-based or pattern based techniques are simple and effective.

5.5 Distractor Generation

Distractors play an important role in multiple-choice question generation. The quality of an MCQ largely depends on the quality of the distractors. If the distractors are not capable of confusing the examinee sufficiently, then he picks the correct answer easily. As a result, overall quality and usability of the MCQ degrade. Several approaches have been used for distractor generation in the literature. These are summarized below.

- **Parts-of-speech Information:** Distractor is semantically close to the key. Therefore, both the key and distractors should belong to same parts-of-speech category. This observation has been used as a clue in [14], [1], [76], [25], [49], [10], [11], [87].
- **Frequency:** Frequency of the words is another hint. Both the distractors and key should have similar frequency count. Frequency is used in [14], [1], [76], [49], [19], [99]. Several systems used a combination of parts-of-speech, frequency, and other type of information.
- **WordNet:** WordNet [105] is a lexical database that groups the words into sets of synonyms (called

synsets), and records the relations among these synonym sets or their members. Semantically close concepts can be used as distractor. Therefore, the WordNet is used by many researchers for distractor generation ([87], [6], [89], [16], [10], [11]).

- **Domain Ontology:** Papasalouros, Kanaris, and Kotis [41] used the Web Ontology Language [106] for finding the distractors.
- **Distributional Hypothesis:** The distributional hypothesis states that similar words appear in similar context. Karamanis, Ha, and Mitkov [17], Mitkov et al. [107], Afzal and Mitkov [48], Smith, Sommers, and Kilgarriff [93] used distributional similarity based approach for distractor generation. Lee and Seneff [99], Agarwal and Mannem [49] used collocation based approach for distractor generation.
- **Pattern Matching:** Pattern matching is another approach. Patterns are defined using various information like parse information and the key. Those patterns are then used to extract the distractor. Hoshino and Nakagawa [19], Aldabe, Maritxalar, and Martinez [28], Chen, Liou, and Chang [15], Goto et al. [10] & [11] used pattern-based method for distractor generation.
- **Semantic Analysis:** As the distractors are the concepts that are semantically close to the keys, semantic analysis is another popular approach. Aldabe and Maritxalar [27] used Latent Semantic Analysis [108] to generate distractors. To generate distractors, Pino, Heilman, and Eskenazi [75] computed semantic similarity between two words using the Patwardhan and Pedersen's method [109]. Aldabe, Maritxalar, and Mitkov [88] computed semantic similarity between verbs according to the distributional data. Kumar, Banchs, and D'Haro [46] selected the distractors using semantic similarity computed through the word2vec tool [110].

Our Observation on Distractor Generation: From the study of the literature, we conclude that the technique to be adopted for distractor generation depends on the application and type of distractors. Frequency, parts-of-speech information and other basic syntactic or statistical information based distractors perform well in applications like second language study, vocabulary, and grammar assessment. However, MCQs of several domains including biomedical, history, and school level subjects require deep statistical or semantic closeness. Numerical distractors are essential for science and mathematics related MCQs. In some domains like sports, entertainment, and tourism named entities are dominant as the key or distractor. Similarity among the names is required to be computed to generate attractive distractors.

Most of the existing systems generate simple distractors. However, in real MCQs various types of distractors are possible. For example, named entity distractor, numerical values, distractor containing multiple words. Generation of such complex distractors is quite difficult and the techniques applied to simple distractor generation might fail there. A robust MCQ generation system should handle all such types of distractors. However, the literature does not include adequate techniques for generation of such variety of distractors. Deep semantic analysis of the text or use of neural embedding based methods might be a possible direction towards complex distractor generation.

5.6 Post-processing

Post-processing is the final phase that aims to improve the quality of the system generated MCQs. Various types of errors can be there in system generated MCQs. These include punctuation error, improper question word, too lengthy stem, availability of discourse words in question, error in number agreement, poor quality of distractors. The system should minimize these errors from the final output. The post-processing step tries to rectify these errors, otherwise, removes the incorrect questions. Primarily three types of post-processing have been used in the literature. These are, post-editing of ill-formed questions, question ranking, and filtering of unacceptable questions.

- **Question Post-editing:** Mitkov, Ha, and Karamanis [6] performed a set of manual post-editing steps. First, they performed a classification to determine the amount of post editing required: minor, fair, or major. Minor revision indicated limited post-editing of the item including correction of spelling and punctuation. The fair revision referred re-ordering, insertion or deletion of several words and replacement of one distractor at most. Major revision involved substantial rephrasing of the stem and replacement of at least two distractors. After the classification, they prepared a post-editing environment. Then a set of human experts were employed to edit the questions. Heilman [72] applied a few post-editing steps like, final periods are changed to question marks and removal of extra whitespace symbols.
- **Question Filtering:** Aldabe et al. [18], Aldabe, Maritxalar, and Martinez [28] designed a filter to reject the ill-formed questions. Primarily the rejecter judged the quality of the distractors (e.g., identical distractors - different words but the same sense, one is the inflection of another). If the required number of good distractors were not produced for a question then the question was rejected. Gates et al. [100] also used distractor filtering in their system. Their filter was based on length, phrase type, and semantic closeness. Sumita, Sugaya, and Yamamoto [16] proposed a filter method based on item information for reducing test size by selection of effective items. For measuring item information, they used a test taker parameter, proficiency of the test taker, discrimination, and difficulty. They selected the items whose contribution to test information is maximal. Heilman [72] included a filter as a post-processing step. The step filtered out questions with noun phrases consisting solely of determiners (e.g., those), noun phrases with unresolved pronouns, and the length is more than 30 tokens.
- **Question Ranking:** Ranking is another approach. Mannem, Prasad, and Joshi [111] employed a heuristic based ranking module. Example heuristics were the depth of the predicate from which the question was generated and the number of pronouns. Heilman and Smith [73] also employed a ranking of the generated questions as a post-processing step. A portion of the labeled data was separated out which was used to

TABLE 1
Evaluation of stem and key of a MCQ: few example systems

System	Type_of_evaluation	Evaluation_metrics	Accuracy
Liu et al. [76]	Semi-automatic evaluation	Quality of cloze items	System has generated 66.2%, 69.4%, 60.0% and 61.5% of correct sentences corresponding to the input request
Aldabe et al. [18]	Expert language teacher	Quality of questions	More than 80%
Pino, Heilman, and Eskenazi [75]	Five English teachers	sentence length, simplicity or difficulty level	66.53%
Agarwal and Mannem [49]	Two biology students	Useful for learning and answerable, or not	Evaluator-1: sentence selection 91.66%, key selection 94.16%, distractor selection 60.05% and Evaluator-2: sentence selection 79.16%, key selection 84.16%, distractor selection 67.72%
Bhatia, Kirti, and Saha [69]	Five evaluators having domain knowledge	difficulty, domain relevance, question formation, over-informative or under-informative	distractors average accuracy 88% and key accuracy 79.4%
Narendra, Agarwal, and shah [12]	Three evaluators and evaluation guidelines	informativeness and relevance	Average score of 3.18 (out of 4)
Kumar, Banchs, and D'Haro [46]	15 Human evaluators	sentence, gap and distractors are good	Question sentences 94%, gaps 87% and distractors 60%
Majumder and Saha [68]	Five human evaluators	Quality of questions	Informative sentences 93.21%, key selection 83.03% and distractor quality accuracy 91.07%
Shah, Shah, and Kurup [38]	Human tutors	Question acceptance	70.66%
Satria and Tokunaga [35]	Five English teachers	Quality of questions	65%
Santhanavijayan et al. [59]	Experimental results and discussions	Efficiency of the system	Informative sentences 72%, blank generation 77.6% and distractor generation accuracy 78.8%
Susanti et al. [5]	79 undergraduate students and eight English teachers	English proficiency and item analysis	Proficiency of 0.71 for TOEFL, 0.68 for TOEIC, 0.57 for CASEC and 74% of the items were acceptable

develop the discriminative question ranker. Questions were ranked by the predictions of a logistic regression model of question acceptability. Effenberger [57] incorporated an exercise grader module in their system. The grader took an exercise to assign grades for difficulty, relevance to the article, and grammatical correctness.

6 MCQ System Evaluation

In the literature we found, most of the systems adopted manual evaluation of the computer-generated MCQs. Since an MCQ contains multiple components, different metrics have been defined for assessing the quality of the individual components. We discuss below the approaches used for MCQ evaluation.

6.1 Evaluation of the Stem and Key

We found that the majority of the systems have been assessed by human evaluators. There exists no standard public dataset for evaluation of system generated MCQs. So, the developers often created private test data through which they evaluated the system with the help of human evaluators. In Table 1 we present a summary of the evaluation process of a few MCQ systems. This is only a small part of the actual table we created during the information gathering phase (Section 4.2). From the table, it can be observed that also there is no standard performance metric for the task. Various metrics and parameters have been employed by the researchers. For example, well-formed, sentence length, sentence simplicity, difficulty level, useful for learning and answerable, sufficient context, difficulty of the item, relevant to the domain, over-informative, under-informative, grammatical correctness and correctness of the sentences.

In the table, we have also incorporated the evaluation results of a few systems. Majority of the systems used their own test data for the evaluation. As the test dataset is not common, it is not worthy to compare the systems based on these accuracy values. Still, we show the accuracy values to provide a synopsis of the performance of the existing systems. These values also indicate that ample research opportunity is there to develop a more accurate system.

6.2 Evaluation of the Distractors

Similarly, there is no standard dataset or evaluation metrics for evaluation of distractors. Hard comparison between gold-standard distractors and system generated distractors cannot be used as the accuracy. In many domains and applications, an MCQ can have a large set of distractors. A gold standard dataset cannot accommodate all of those. So, there is a possibility that due to the inadequacy of the gold standard dataset, a system generated correct distractors is marked as wrong. Therefore, human experts often performed the distractor evaluation task.

Various metrics have been used in the literature depending on the domain, application, and type of the distractors. For example, Mitkov et al. [107] used difficulty, discriminating power and usefulness for distractor evaluation. Aldabe and Maritxalar [27] followed a similar strategy for distractor evaluation. Additionally, they identified the distractors that were never chosen by the examinees. To judge the quality

of the distractors Pino, Heilman, and Eskenazi [75] replaced the keyword by a distractor and measured the grammaticality and collocation criteria of the sentence from the syntactic and semantic point of view. Agarwal and Mannem [49] used a similar approach through the readability metric. They asked the evaluators to substitute the distractor in the gap and check the readability and semantic meaning of the sentence. For distractor evaluation Bhatia, Kirti, and Saha [69] used the closeness value. The distractor set is 'good' if at least one of the distractors are close to the key. Araki et al. [3] measured the distractor quality using a three-point scale. In that scaling, 1 (worst) specifies that the distractor is confusing because it overlaps the correct answer partially or completely; value 2 means that the distractor can be easily identified as an incorrect answer; 3 (best) indicates that the distractor can be viable.

7 Conclusion

Evaluation is essential in the teaching-learning process and MCQs are popular for educational assessment. In this paper, we reviewed the works presented in the literature of automatic MCQ generation from a text. We discussed the existing approaches for MCQ generation. We established a generic workflow consisting of six broadly classified dependent phases, namely, pre-processing, sentence selection, key selection, question formation, distractor generation, and post-processing. Various techniques have been employed for implementation of the individual phases. We presented a comparative discussion of these techniques. We also discussed the methods for assessment of system generated MCQs.

During the study, we observed that the literature on automatic MCQ generation experiences a few challenges that need to be solved. The researchers should focus on these areas where attention is required to make the field stronger. We have identified a few extents where future research on MCQ generation should focus on. These are, MCQ from multi-line facts, ability to handle complex knowledge text, complex distractor generation, standard evaluation techniques, and gold-standard test data. So the area still contains plenty of scopes to continue future research.

Acknowledgments

This work is supported by the project grant (file no.: YSS/2015/001948) provided by the Science and Engineering Research Board (SERB), Govt. of India.

References

[1] D. Coniam, "A preliminary inquiry into using corpus word frequency data in the automatic generation of english language cloze tests," *Calico Journal*, vol. 14, no. 2-4, pp. 15–33, 1997.

[2] R. D. Nielsen, J. Buckingham, G. Knoll, B. Marsh, and L. Palen, "A taxonomy of questions for question generation," in *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, 2008.

[3] J. Araki, D. Rajagopal, S. Sankaranarayanan, S. Holm, Y. Yamakawa, and T. Mitamura, "Generating questions and multiple-choice answers using semantic analysis of texts." in *COLING*, 2016, pp. 1125–1136.

[4] S. Subramanian, T. Wang, X. Yuan, S. Zhang, Y. Bengio, and A. Trischler, "Neural models for key phrase detection and question generation," *arXiv preprint arXiv:1706.04560*, 2017.

[5] Y. Susanti, T. Tokunaga, H. Nishikawa, and H. Obari, "Evaluation of automatically generated english vocabulary questions," *Research and Practice in Technology Enhanced Learning*, vol. 12, no. 1, p. 11, 2017.

[6] R. Mitkov, L. An Ha, and N. Karamanis, "A computer-aided environment for generating multiple-choice test items," *Nat. Lang. Eng.*, vol. 12, no. 2, pp. 177–194, Jun. 2006.

[7] M. Liu, V. Rus, and L. Liu, "Automatic chinese factual question generation," *IEEE Transactions on Learning Technologies*, vol. 10, no. 2, pp. 194–204, 2017.

[8] D. Metzler and W. B. Croft, "Analysis of statistical question classification for fact-based questions," *Inf. Retr.*, vol. 8, no. 3, pp. 481–504, May 2005.

[9] M. Agarwal, "Cloze and open cloze question generation systems and their evaluation guidelines," *International Institute of Information Technology, Hyderabad*, 2012.

[10] T. Goto, T. Kojiri, T. Watanabe, T. Iwata, and T. Yamada, "An automatic generation of multiple-choice cloze questions based on statistical learning," in *Proceedings of the 17th International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education, 2009, pp. 415–422.

[11] ——, "Automatic generation system of multiple-choice cloze questions and its evaluation," *Knowledge Management & E-Learning: An International Journal (KM&EL)*, vol. 2, no. 3, pp. 210–224, 2010.

[12] A. Narendra, M. Agarwal, and R. shah, "Automatic cloze-questions generation," in *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. INCOMA Ltd. Shoumen, BULGARIA, 2013, pp. 511–515.

[13] J. Mostow and W. Chen, "Generating instruction automatically for the reading strategy of self-questioning." in *AIED 2009: 14th International Conference on Artificial Intelligence in Education*, 2009, pp. 465–472.

[14] J. C. Brown, G. A. Frishkoff, and M. Eskenazi, "Automatic question generation for vocabulary assessment," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 819–826.

[15] C.-Y. Chen, H.-C. Liou, and J. S. Chang, "Fast: An automatic generation system for grammar tests," in *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, ser. COLING-ACL '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 1–4.

[16] E. Sumita, F. Sugaya, and S. Yamamoto, "Measuring non-native speakers' proficiency of english by using a test with automatically-generated fill-in-the-blank questions," in *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, ser. EdAppsNLP 05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 61–68.

[17] N. Karamanis, L. A. Ha, and R. Mitkov, "Generating multiple-choice test items from medical text: A pilot study," in *Proceedings of the Fourth International Natural Language Generation Conference*, ser. INLG '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 111–113.

[18] I. Aldabe, M. L. de Lacalle, M. Maritxalar, E. Martinez, and L. Uria, "Arikiturri: An automatic question generator based on corpora and nlp techniques," in *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, ser. ITS'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 584–594.

[19] A. Hoshino and H. Nakagawa, "Assisting cloze test making with a web application," *TECHNOLOGY AND TEACHER EDUCATION ANNUAL*, vol. 18, no. 5, pp. 2807–2814, 2007.

[20] C. M. Feeney and M. Heilman, "Automatically generating and validating reading-check questions," in *International Conference on Intelligent Tutoring Systems*. Springer, 2008, pp. 659–661.

[21] A. Y. Satria and T. Tokunaga, "Automatic generation of english reference question by utilising nonrestrictive relative clause," in *Proceedings of the 9th International Conference on Computer Supported Education*, 2017, pp. 379–386.

[22] Y. Susanti, H. Nishikawa, T. Tokunaga, and O. Hiroyuki, "Item difficulty analysis of english vocabulary questions." in *CSEDU (1)*, 2016, pp. 267–274.

[23] A. Malinova and O. Rahneva, "Automatic generation of english language test questions using mathematica," in *Conference: CBU International Conference on Innovations in Science and Education (CBUIC)*, 2016, pp. 906–909.

[24] R. Correia, J. Baptista, M. Eskenazi, and N. Mamede, "Automatic generation of cloze question stems," in *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*, ser. PROPOR'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 168–178.

[25] R. Correia, J. Baptista, N. Mamede, I. Trancoso, and M. Eskenazi, "Automatic generation of cloze question distractors," in *Proceedings of the Interspeech 2010 Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology, Waseda University, Tokyo, Japan*, 2010.

[26] I. Aldabe and M. Maritxalar, "Semantic similarity measures for the generation of science tests in basque," *IEEE transactions on Learning Technologies*, vol. 7, no. 4, pp. 375–387, 2014.

[27] ——, "Automatic distractor generation for domain specific texts." in *IceTAL*. Springer, 2010, pp. 27–38.

[28] I. Aldabe, M. Maritxalar, and E. Martinez, "Evaluating and improving the distractor-generating heuristics," *corpus*, vol. 1060, 2007.

[29] L. Perez-Beltrachini, C. Gardent, and G. Kruszewski, "Generating grammar exercises," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, 2012, pp. 147–156.

[30] A. Kurtasov, "A system for generating cloze test items from russian-language text." in *RANLP*, 2013, pp. 107–112.

[31] M. Liu, V. Rus, and L. Liu, "Automatic chinese multiple choice question generation using mixed similarity strategy," *IEEE Transactions on Learning Technologies*, vol. 11, pp. 193–202.

[32] M.-H. Chu, W.-Y. Chen, and S.-D. Lin, "A learning-based framework to utilize e-hownet ontology and wikipedia sources to generate multiple-choice factual questions," in *Technologies and Applications of Artificial Intelligence (TAAI), 2012 Conference on*. IEEE, 2012, pp. 125–130.

[33] J. Hill and R. Simha, "Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and google n-grams," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 2016, pp. 23–30.

[34] Y. Susanti, R. Iida, and T. Tokunaga, "Automatic generation of english vocabulary tests." in *CSEDU (1)*, 2015, pp. 77–87.

[35] A. Y. Satria and T. Tokunaga, "Evaluation of automatically generated pronoun reference questions," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 2017, pp. 76–85.

[36] A. Hoshino and H. Nakagawa, "Predicting the difficulty of multiple-choice close questions for computer-adaptive testing," *Special issue: Natural Language Processing and its Applications*, pp. 279–292, 2010.

[37] ——, "Webexperimenter for multiple-choice question generation," in *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Association for Computational Linguistics, 2005, pp. 18–19.

[38] R. Shah, D. Shah, and L. Kurup, "Automatic question generation for intelligent tutoring systems," in *Communication Systems, Computing and IT Applications (CSCITA), 2017 2nd International Conference on*. IEEE, 2017, pp. 127–132.

[39] J. Mostow and H. Jang, "Generating diagnostic multiple choice comprehension cloze questions," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, 2012, pp. 136–146.

[40] M. Al-Yahya, "Ontoque: a question generation engine for educational assesment based on domain ontologies," in *Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on*. IEEE, 2011, pp. 393–395.

[41] A. Papasalouros, K. Kanaris, and K. Kotis, "Automatic generation of multiple choice questions from domain ontologies." in *e-Learning*, 2008, pp. 427–434.

[42] B. Sun, Y. Zhu, Y. Xiao, R. Xiao, and Y. G. Wei, "Automatic question tagging with deep neural networks," *IEEE Transactions on Learning Technologies*, 2018.

[43] R. Conejo, E. Guzmán, and M. Trella, "The siette automatic assessment environment," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 1, pp. 270–292, 2016.

[44] K. Stasaski and M. A. Hearst, "Multiple choice question generation utilizing an ontology," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 2017, pp. 303–312.

[45] D. Pugh, A. De Champlain, M. Gierl, H. Lai, and C. Touchie, "Using cognitive models to develop quality multiple-choice questions," *Medical teacher*, vol. 38, no. 8, pp. 838–843, 2016.

[46] G. Kumar, R. E. Banchs, and L. F. D'Haro, "Automatic fill-the-blank question generator for student self-assessment," in *Frontiers in Education Conference (FIE), 2015 IEEE*. IEEE, 2015, pp. 1–3.

[47] A. E. Awad and M. Y. Dahab, "Automatic generation of question bank based on pre-defined templates," *International Journal of Innovations Advancement in Computer Science*, vol. 3, no. 1, pp. 80–87, 2014.

[48] N. Afzal and R. Mitkov, "Automatic generation of multiple choice questions using dependency-based semantic relations," *Soft Comput.*, vol. 18, no. 7, pp. 1269–1281, Jul. 2014.

[49] M. Agarwal and P. Mannem, "Automatic gap-fill question generation from text books," in *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, ser. IUNLPBEA '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 56–64.

[50] L. Zavala and B. Mendoza, "On the use of semantic-based aig to automatically generate programming exercises," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. ACM, 2018, pp. 14–19.

[51] N. Stancheva, A. Stoyanova-Doycheva, S. Stoyanov, I. Popchev, and V. Ivanova, "A model for generation of test questions," *Compt. Rend. Acad. bulg. Sci*, vol. 70, no. 5, pp. 619–630, 2017.

[52] T. Alsubait, B. Parsia, and U. Sattler, "Ontology-based multiple choice question generation," *KI-Künstliche Intelligenz*, vol. 30, no. 2, pp. 183–188, 2016.

[53] F. de Assis Zampirolli, V. R. Batista, and J. A. Quilici-Gonzalez, "An automatic generator and corrector of multiple choice tests with random answer keys," in *Frontiers in Education Conference (FIE), 2016 IEEE*. IEEE, 2016, pp. 1–8.

[54] N. L. Bhale, S. G. Patil, M. R. Pawar, T. G. Katkade, and N. D. Jadhav, "Automatic question generation using wikipedia," *Imperial Journal of Interdisciplinary Research*, vol. 2, no. 4, 2016.

[55] B. Liu, "Sarac: a framework for automatic item generation," in *Advanced Learning Technologies, 2009. ICALT 2009. Ninth IEEE International Conference on*. IEEE, 2009, pp. 556–558.

[56] C. Kwankajornkiet, A. Suchato, and P. Punyabukkana, "Automatic multiple-choice question generation from thai text," in *Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on*. IEEE, 2016, pp. 1–6.

[57] T. Effenberger, "Automatic question generation and adaptive practice," Master's thesis, Masaryk University Faculty of Informatics, 2015.

[58] I. E. Fattoh, "Automatic multiple choice question generation system for semantic attributes using string similarity measures," *Computer Engineering and Intelligent Systems*, vol. 5, no. 8, pp. 66–73, 2014.

[59] A. Santhanavijayan, S. Balasundaram, S. H. Narayanan, S. V. Kumar, and V. V. Prasad, "Automatic generation of multiple choice questions for e-assessment," *International Journal of Signal and Imaging Systems Engineering*, vol. 10, no. 1-2, pp. 54–62, 2017.

[60] S. Adithya and P. K. Singh, "Web authoriser tool to build assessments using wikipedia articles," in *Region 10 Conference, TENCON 2017-2017 IEEE*. IEEE, 2017, pp. 467–470.

[61] N. S. Stancheva, I. Popchev, A. Stoyanova-Doycheva, and S. Stoyanov, "Automatic generation of test questions by software agents using ontologies," in *Intelligent Systems (IS), 2016 IEEE 8th International Conference on*. IEEE, 2016, pp. 741–746.

[62] T. KOJIRI, T. GOTO, T. WATANABE, T. IWATA, and T. YAMADA, "Statistical learning-based approach for automatic generation system of multiple-choice cloze questions," in *Workshop Proceedings of the 18th International Conference on Computers in Education: ICCE2010*, 2010, p. 84.

[63] L.-C. Sung, Y.-C. Lin, and M. C. Chen, "The design of automatic quiz generation for ubiquitous english e-learning system," in *Technology Enhanced Learning Conference (TELearn 2007), Jhongli, Taiwan*, 2007, pp. 161–168.

[64] E. Vinu, T. Alsubait, and P. S. Kumar, "Modeling of item-difficulty for ontology-based multiple-choice questions," *arXiv preprint arXiv:1607.00869*, 2016.

[65] D. Seyler, M. Yahya, and K. Berberich, "Knowledge questions from knowledge graphs," *arXiv preprint arXiv:1610.09935*, 2016.

[66] S. Knoop and S. Wilske, "Wordgap-automatic generation of gap-filling vocabulary exercises for mobile learning," in *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013; May 22-24; Oslo; Norway. NEALT Proceedings*

Series 17, no. 086. Linköping University Electronic Press, 2013, pp. 39–47.

[67] M. Majumder and S. K. Saha, "Automatic selection of informative sentences: The sentences that can generate multiple choice questions," *Knowledge Management & E-Learning: An International Journal (KM&EL)*, vol. 6, no. 4, pp. 377–391, 2014.

[68] ——, "A system for generating multiple choice questions: With a novel approach for sentence selection," *ACL-IJCNLP Workshop on NLPTEA 2015*, pp. 64–72, 2015.

[69] A. S. Bhatia, M. Kirti, and S. K. Saha, "Automatic generation of multiple choice questions using wikipedia," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2013, pp. 733–738.

[70] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," *Comput. Speech Lang.*, vol. 15, no. 3, pp. 287–333, Jul. 2001.

[71] L. Bednarik and L. Kovacs, "Implementation and assessment of the automatic question generation module," in *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*. IEEE, 2012, pp. 687–690.

[72] M. Heilman, "Automatic factual question generation from text," Ph.D. dissertation, Language Technologies Institute, School of Computer Science, Pittsburgh, PA, USA, 2011, aAI3528179.

[73] M. Heilman and N. A. Smith, "Good question! statistical ranking for question generation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 609–617.

[74] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: Two sides of the same coin?" *Commun. ACM*, vol. 35, no. 12, pp. 29–38, Dec. 1992.

[75] J. Pino, M. Heilman, and M. Eskenazi, "A selection strategy to improve cloze question quality," in *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada*, 2008, pp. 22–32.

[76] C.-L. Liu, C.-H. Wang, Z.-M. Gao, and S.-M. Huang, "Applications of lexical information for algorithmically composing multiple-choice cloze items," in *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, ser. EdAppsNLP 05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 1–8.

[77] A. Hoshino and H. Nakagawa, "A real-time multiple-choice question generation for language testing: A preliminary study," in *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, ser. EdAppsNLP 05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 17–20.

[78] S. d. S. L. Curto, "Automatic generation of multiple-choice tests geraçao automatica de testes de escolha m ultipla," Master's thesis, Instituto Superior Tecnico, Information Systems and Computer Engineering, 2010.

[79] D. Vickrey and D. Koller, "Sentence simplification for semantic role labeling." in *ACL*, 2008, pp. 344–352.

[80] B. Das, M. Majumder, and S. Phadikar, "A novel system for generating simple sentences from complex and compound sentences," *I.J. Modern Education and Computer Science*, pp. 57–64, 2018.

[81] M. Heilman and M. Eskenazi, "Application of automatic thesaurus extraction for computer generation of vocabulary questions," in *Workshop on Speech and Language Technology in Education*, 2007.

[82] B. Das and M. Majumder, "Factual open cloze question generation for assessment of learner's knowledge," *International Journal of Educational Technology in Higher Education*, vol. 14, no. 1, p. 24, 2017.

[83] J. Allan and G. Kumaran, "Stemming in the language modeling framework," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, ser. SIGIR '03. New York, NY, USA: ACM, 2003, pp. 455–456.

[84] P. Pabitha, M. Mohana, S. Suganthi, and B. Sivanandhini, "Automatic question generation system," in *Recent Trends in Information Technology (ICRTIT), 2014 International Conference on*. IEEE, 2014, pp. 1–5.

[85] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Trans. Inf. Syst.*, vol. 26, no. 3, pp. 13:1–13:37, Jun. 2008.

[86] C.-C. Shei, "Followyou!: An automatic language lesson generation system," *Computer Assisted Language Learning*, vol. 14, no. 2, pp. 129–144, 2001.

[87] R. Mitkov and L. A. Ha, "Computer-aided generation of multiple-choice tests," in *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing - Volume 2*, ser. HLT-NAACL-EDUC '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 17–22.

[88] I. Aldabe, M. Maritxalar, and R. Mitkov, "A study on the automatic selection of candidate sentences distractors." in *AIED 2009: 14th International Conference on Artificial Intelligence in Education Workshops Proceedings*, 2009, pp. 656–658.

[89] Y.-C. Lin, L.-C. Sung, and M. C. Chen, "An automatic multiple-choice question generation scheme for english adjective understanding," in *Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007)*, 2007, pp. 137–142.

[90] R. Shah, "Automatic question generation using discourse cues and distractor selection for cloze questions," Master's thesis, Language Technology and Research Center (LTRC), International Institute of Information Technology, Hyderabad, 2012.

[91] S. Bhingardive, "Introduction to word sense disambiguation." [Online]. Available: http://www.cfilt.iitb.ac.in/viva_workshop/Day5-WSD-Sudha.Bhingardive.pdf

[92] S. Smith, P. Avinesh, and A. Kilgarriff, "Gap-fill tests for language learners: Corpus-driven item generation," in *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, 2010, pp. 1–6.

[93] S. Smith, S. Sommers, and A. Kilgarriff, "Learning words right with the sketch engine and webbootcat: Automatic cloze generation from corpora and the web," in *Proceedings of the 25th International Conference of English Teaching and Learning & 2008 International Conference on English Instruction and Assessment, Lisbon, Portugal*, 2008, pp. 1–8.

[94] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Comput. Linguist.*, vol. 28, no. 3, pp. 245–288, Sep. 2002.

[95] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, "Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, ser. CONLL Shared Task '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 28–34.

[96] S. Srivastava and S. Govilkar, "A survey on paraphrase detection techniques for indian regional languages," *International Journal of Computer Applications*, vol. 163, no. 9, pp. 42–47, 2017.

[97] D. R. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu *et al.*, "Mead-a platform for multidocument multilingual text summarization." in *LREC*, 2004.

[98] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. Noida, India: Pearson Education, 2014.

[99] J. Lee and S. Seneff, "Automatic generation of cloze items for prepositions," in *Eighth Annual Conference of the International Speech Communication Association*, 2007, pp. 2173–2176.

[100] D. Gates, G. Aist, J. Mostow, M. McKeown, and J. Bey, "How to generate cloze questions from definitions: A syntactic approach," in *2011 AAAI Fall Symposium Series*, 2011, pp. 19–22.

[101] L.-C. Sung, Y.-C. Lin, and M. C. Chen, "An automatic quiz generation system for english text," in *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 196–197.

[102] D. Lindberg, F. Popowich, J. Nesbit, and P. Winne, "Generating natural language questions to support learning on-line," *Proceedings of the 14th European Workshop on Natural Language Generation*, p. 105–114, 2013.

[103] K. Mazidi and R. D. Nielsen, "Linguistic considerations in automatic question generation." in *ACL proceedings*, 2014, pp. 321–326.

[104] X. Yao and Y. Zhang, "Question generation with minimal recursion semantics," in *Proceedings of QG2010: The Third Workshop on Question Generation*, 2010, pp. 68–75.

[105] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[106] G. Antoniou and F. Van Harmelen, "Web ontology language: Owl," in *Handbook on ontologies*. Springer, 2004, pp. 67–92.
[107] R. Mitkov, L. A. Ha, A. Varga, and L. Rello, "Semantic similarity of distractors in multiple-choice tests: Extrinsic evaluation," in *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, ser. GEMS '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 49–56.
[108] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.
[109] S. Patwardhan and T. Pedersen, "Using wordnet-based context vectors to estimate the semantic relatedness of concepts," in *Proceedings of the eacl 2006 workshop making sense of sense-bringing computational linguistics and psycholinguistics together*, vol. 1501. Trento, 2006, pp. 1–8.
[110] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
[111] P. Mannem, R. Prasad, and A. Joshi, "Question generation from paragraphs at upenn: Qgstec system description," in *Proceedings of QG2010: The Third Workshop on Question Generation*, 2010, pp. 84–91.

Dhawaleswar Rao Ch is currently working toward the Ph.D. degree at Birla Institute of Technology Mesra, India. His research interests include Natural Language Processing, Machine Learning and Internet algorithms.

Dr.Sujan Kumar Saha is currently working as Assistant Professor at Birla Institute of Technology Mesra, India. He received his Ph.D. degree from IIT Kharagpur in 2010 and M.tech degree from IIT Delhi in 2005. His research interests include Educational NLP, NLP in new languages and Machine Learning.