Smart Learning Environments

RESEARCH Open Access

Towards development of a system for automatic assessment of the quality of a question paper

Sujan Kumar Saha

Correspondence: sujan.kr.saha@gmail.com
Department of Computer Science and Engineering, Birla Institute of Technology Mesra, Ranchi 835215, India

Abstract

In this paper, we present a system for automatic evaluation of the quality of a question paper. Question paper plays a major role in educational assessment. The quality of a question paper is crucial to fulfilling the purpose of the assessment. In many education sectors, question papers are prepared manually. A prior analysis of a question paper might help in finding the errors in the question paper, and better achieving the goals of the assessment. In this experiment, we focus on higher education in the technical domain. First, we conducted a student survey to identify the key factors that affect the quality of a question paper. The top factors we identified are question relevance, question difficulty, and time requirement. We explored the strategies to handle these factors and implemented them. We employ various concepts and techniques for the implementation. The system finally assigns a numerical quality score against these factors. The system is evaluated using a set of question papers collected from various sources. The experimental results show that the proposed system is quite promising.

Keywords: Educational assessment, Question assessment, Question paper quality, Question difficulty, Question relevance

Introduction

The objective of this article is to design a framework for automatic estimation of the quality of a question paper. Question paper-based assessment is the most widely used method in various sectors of educational assessment. For this, a question paper is prepared to judge how well a student can demonstrate their acquired knowledge and understanding. A question paper contains a set of questions. A question is mainly composed of two parts: a short text through which the request for the information is made to the responder, and a maximum score which will be awarded to the responder based on the correctness of the answer. Various types of questions (e.g., objective or subjective, open-ended or cloze questions, recall or synthesis) are used in educational assessment depending on the level or purpose of assessment. Although a lot of research is being carried out on automatic question generation (Kurdi, Leo, Parsia, Sattler, & Al-Emari, 2019; Rao & Saha, 2020) and automatic answer grading (Burrows, Gurevych, & Stein, 2015; Mohler, Bunesc, & Mihalcea, 2011; Zhang, Huang, Yang, Yu,

& Zhuang, 2020), available systems are unable to generate/evaluate all types of questions correctly. Therefore, a fully automatic educational assessment is still not widely used. Many education sectors are reliant on manually prepared question papers.

'Quality' of the question paper is important to fulfill the purpose of the assessment. The essential set of qualities of a question paper varies from level to level. A few essential qualities of a question paper might be the relevance of the questions, coverage of the question paper concerning the syllabus, correctness and completeness of the questions, difficulty level, and answerability within the stipulated time. When the teacher himself is the paper-setter, then these quality metrics are mostly ensured. However, in many cases, the teacher himself is not the paper-setter. For instance, multiple colleges or schools might be affiliated with a university; all follow a particular syllabus, individual centers have their teacher, but a common examination is conducted using a single question paper. Sometimes one of those teachers prepares the question paper. Again, in certain cases to maintain uniformity, an external paper setter does the job. Prior assessment of the quality of the question paper is crucial there to avoid undesirable issues during or after the examination. Manual question paper assessment is a tedious task and not feasible in many cases. So, we investigate the possibility of designing a system that performs the automatic assessment of the quality of a question paper.

In this paper, we propose a novel system that assesses the quality of a human-generated question paper. There can be several metrics to define the quality of a question paper. We focus on an engineering study, and to understand the most crucial quality factors in the domain, we conduct a student survey. The identified factors are, (a) relevance of the questions to the stipulated syllabus, (b) difficulty level of the individual questions, and (c) time required to answer the questions. We studied these issues and identified appropriate techniques to handle them. For implementation, we used several concepts and techniques. These include the term or keyword extraction from the questions, Latent Dirichlet Analysis, question classification using multiple levels, Support Vector Machines, question-question similarity. For estimating the time required to answer the questions, we design an experimental study with a set of students. Finally, the system produces a score for all the considered difficulty factors given a question paper. The proposed system is tested on a dataset containing a set of real question papers. The system-generated scores are then compared with the scores given by human evaluators. In our experiments, we found that the system achieves 98.02% accuracy in predicting the relevance, 0.48 RMSE value (in 0–5 scale) in estimating the difficulty of the question papers, and the mean absolute error is 15.3 (in 0–180 scale) in predicting the response time. The background, methodology, implementation details, evaluation setup, and results are discussed in the subsequent sections.

Related works

Preparation of a quality question paper needs expertise, experience, time, and care. The absence of these leads to an inappropriate question paper that results in disgraceful events. A few studies are there in the literature that identifies the features of good questions for academic assessment. For instance, Haladyna, Downing, and Rodriguez (2002) presented a taxonomy of 31 multiple-choice item writing guidelines. The taxonomy was intended for classroom assessment. The authors suggested that textbook writers should consider the taxonomy in future editions of their textbooks. The taxonomy is

also useful for developing test items for large-scale assessments. Later, Anderson and Rogan (2010) presented a list of criteria that instructors may use to evaluate the quality of their assessment instruments and overall program. They defined various criteria to be satisfied by the assessment instruments like questions. They defined eight 'before the assessment' criteria to be checked during the design of the assessment instrument. These criteria are well-defined and provide a clear guideline for the assessment of question papers. A dedicated teacher always tries to follow these to prepare a good question paper. However, there might be some exceptions in certain cases. These might result in poor question papers. The identification of such question papers is essential.

Automatic assessment of question papers is not a widely explored territory. In the literature, we find a few attempts that aim to assess the quality of computer-generated questions. Automatic question generation (AQG) is a related research area where a lot of works have been carried out to develop systems for the automatic generation of questions from the text. An overview of the literature on automatic question generation specifically for educational purposes may be found in Kurdi et al. (2019). Automatic multiple-choice question generation is a subproblem of AQG; a survey of the literature on automatic MCQ generation can be found in Rao and Saha (2020). Several approaches and metrics have been proposed by the researchers to assess the quality of system-generated MCQs. Chali and Hasan (2015) focused on evaluating the syntactic correctness of the questions. They proposed an approach that computes the syntactic similarity of each question with the associated content information. Araki et al. (2016) evaluated the questions based on grammatical correctness and distractor quality. Narendra, Agarwal, and Shah (2013) considered informativeness and relevance to assessing the quality of the system generated questions. Zhang and VanLehn (2016) used several parameters including relevance, fluency, ambiguity, pedagogy, and depth to evaluate the quality of the system-generated questions. Susanti, Tokunaga, Nishikawa, and Obari (2017) proposed a method named item analysis for the evaluation of the system-generated questions. Item analysis used two parameters difficulty index and discrimination index that helps to evaluate the standard of MCQs used in a test. Pandarova et al. (2019) aimed to estimate the difficulty scoring of grammar exercise items to be used in dynamic difficulty adaptation in an intelligent language tutoring system. Luger and Bowles (2013) and Luger (2016) proposed a few methods for assessing the quality and difficulty of MCQs. They proposed a method for automatically judging the difficulty and discriminating power of MCQs. Their approach for measuring question difficulty relies on models of how good pupils will perform and contrasts that with their lower-performing peers. As the questions to be assessed are generated by systems, most of these works employ human evaluators who assess the questions using the proposed metrics. However, the current article aims to the automatic evaluation of human-generated questions. Additionally, most of those methods focus on the evaluation of MCQs. These are not applicable in other types of questions including, subjective or open-ended questions.

In the literature, we find some works that aim to estimate the difficulty level of a question. Santos et al. (2012) conducted a study to examine the ability of teachers for categorizing questions by difficulty level and comparing it with the students' perception. They also developed an expert system for automatic classification of questions. Their results showed that the teachers are only slightly more accurate (closer to the

expert system), despite the general students' tendency to overestimate the difficulty level of less difficult questions. However, they failed to draw any general conclusions in that study. Liu, Wang, Lin, and Hon (2013) worked on estimating question difficulty in community question answering services. They proposed a competition-based model for estimating question difficulty by leveraging pairwise comparisons between questions and users. They compared their model with the PageRank-based approach. Huang et al. (2017) proposed a system for predicting question difficulty for reading problems in standard tests.

However, we did not find any system that assesses the quality of a human-generated whole question paper of university-level subjects.

Methodology

In Fig. 1, we have presented the overall workflow of the proposed system for automatic quality assessment of a human-generated question paper. The first step is to identify the most crucial factors that affect the quality. The opinion of the students is extremely important in deciding the quality factors. So, to understand the quality factors, we conducted a survey.

Student survey to identify quality factors

In this study, we focus on the issues of university-level question papers. So, in this survey, we engaged 112 graduate students at our university. The participants are Computer Science and Engineering students. During the survey, the participants were asked to point out three major causes that make a question paper poor. We did not provide any predefined list of possible issues during the survey. Also, pair-discussion was not allowed during the survey. So, the issues raised by the participants resulted from their individual thinking.

Here we would like to provide a background to understand the scenario. Multiple extension centers are attached to the university, and a particular course often runs in

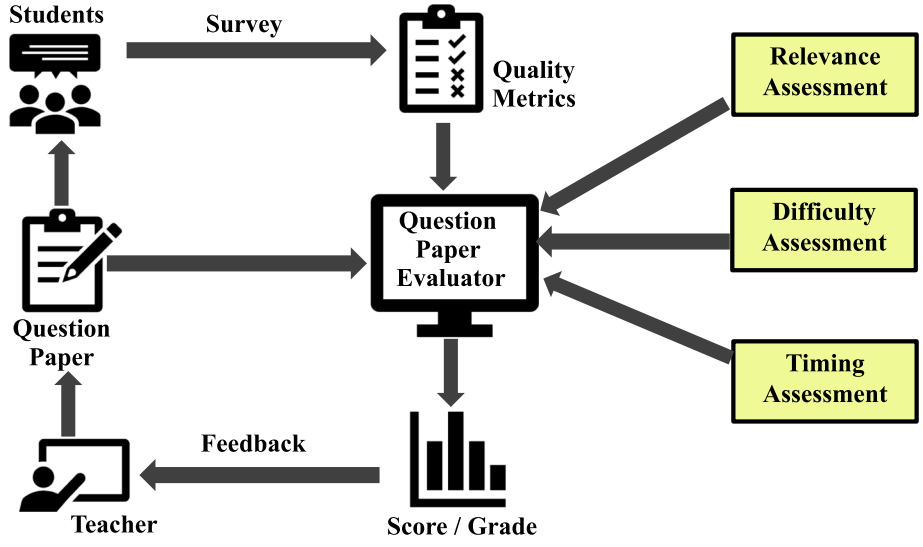

Fig. 1 Workflow of the proposed system to evaluate the quality of a question paper

multiple centers simultaneously. Therefore, the course is simultaneously taught by multiple faculty members. However, the examination of all the registered students is conducted through a single question paper. To maintain uniformity in the course content, each course follows a syllabus provided by the university. The syllabus also specifies the chapters or modules, a list of topics and sub-topics from each chapter, and a class-wise lecture plan. Additionally, the syllabus recommends one textbook and one or more reference books. The extent of discussion is primarily guided by the textbook. A paper-setter is appointed by the university examination office. The paper-setter may be one of the current teachers or an external expert. The paper-setter is instructed to follow the specified syllabus equitably.

Outcome of the survey

Students had written the issues by their wordings. During analysis, these statements were manually normalized. We found that majority of the participants pointed out three issues. These are listed in Table 1. Other mentioned issues are, module-wise coverage is not uniform, unbalanced question paper, the question is too generic to answer, the scope of the numerical question is there but the theory is asked, learning taxonomy is not followed, etc. However, those are not commonly raised by the majority of the students. So, during implementation, we considered the top 3 issues.

Key factors for question paper quality

As per the outcome of the survey, we considered three factors for implementation of the system. These factors are discussed below.

A. *Relevance Assessment:* This is the most important quality factor that a question paper should fulfil. The questions should follow the stipulated syllabus, out of syllabus question is not expected. This is an individual question-based factor. Therefore, each question of a question paper should be checked whether it confirms the syllabus.
B. *Difficulty Assessment:* Assigning a difficulty level to a question paper is not an easy task. From a teacher's point of view, a difficult question might have several motives and consequences. However, we found in our survey, the students primarily treat a question as difficult if it demands a deeper concept or the teacher did not discuss the concept properly. It is not feasible to implement all these. In this system, to determine the difficulty, we consider the factors: cognitive domain taxonomy, depth of the terms in the corresponding chapter, and multi-concept questions.
C. *Timing Assessment:* Question paper length or response time is another important factor in time-bound examinations. The length of the answers to the whole question paper should be answerable within the stipulated time. If one question is

Table 1 The top 3 issues identified by the students

Sl. No.	Question paper issue	No. of participants
1	Out of syllabus question, topic/sub-topic not mentioned in the specified syllabus	104
2	Questions are difficult	95
3	Lengthy question paper, unable to finish within stipulated time	92

lengthy, then another question should balance it so that the overall question paper remains manageable.

The factors to be considered for question paper assessment depends on various aspects, including the level of study (pre-school or primary school question papers do not have some of these components but need other factors to have), type of assessment (interim or weekly assessment vs final or annual assessment), or the purpose of the assessment (college examination question paper vs interview question paper). In this study, we have considered three factors identified from the survey. However, these factors may not be enough or implementable to assess numerical questions, questions dealing with graphs, algorithms, code segment, flow diagrams, and figures.

System implementation

The system is composed of multiple components. Each component has its own strategy to follow. We discuss below the strategies we adopt to develop the individual components of the system. Essential resources for these components are the stipulated syllabus and softcopy of a textbook in a readable format.

Relevance assessment

The purpose of this phase is to test whether a question is confirming the specified syllabus. The core idea is based on keyword matching and is summarized in Fig. 2. We identify the keywords from the question and match these with the keywords extracted from the textbook. If the matching is higher than a threshold, then it indicates that the question is according to the syllabus and relevant.

To identify the keywords from the question and textbook, we use different approaches: structural pattern-based, Tf*Idf based, and Latent Dirichlet Analysis (LDA). The syllabus itself contains a list of topics, that are regarded as keywords. Additionally, the textbook contains a set of subtopics under a particular topic. Those are often not mentioned in the syllabus. To extract the list of subtopics, we defined a set of patterns that uses the structural information, including heading, subheading, section, subsection,

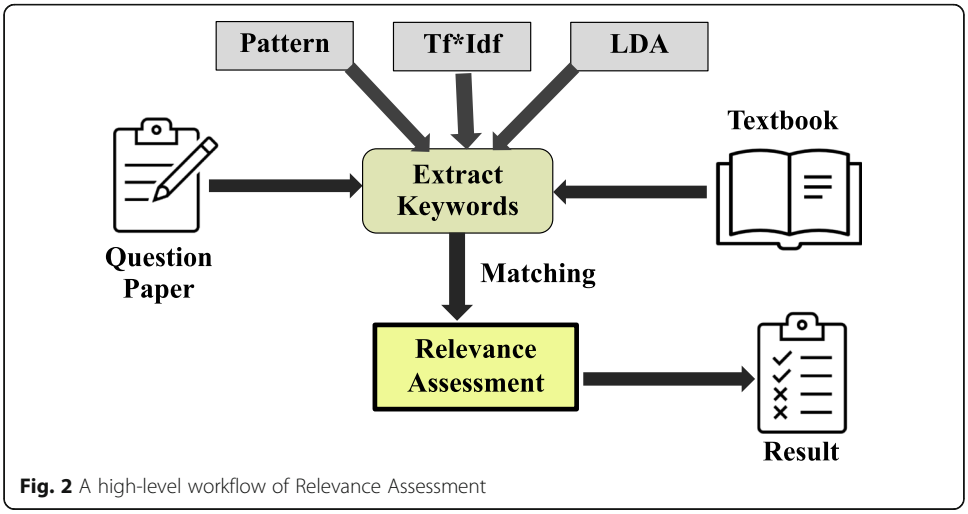

Fig. 2 A high-level workflow of Relevance Assessment

section number, and caption. These patterns extract a larger set of keywords. However, this is also not enough. The text sentences contain additional keywords that are used in factual or recall-type questions. So, we need to extract all the keywords. Tf*Idf and LDA-based approach is applied to the whole text for extraction of the keywords.

Tf-Idf (tf: term frequency, idf: inverse document frequency) is used to find a set of keywords. Subject-specific keywords have a higher occurrence in the textbook, but those rarely occur in other domains (e.g., general or newswire). To implement this fact, we used the Tf-Idf measure that identifies a set of domain-specific keywords. LDA (Blei, Ng, Jordan, & Lafferty, 2003) is a probabilistic topic model of documents where each document is assigned a set of topics. The distribution of documents is over a fixed number of topics, while each topic is represented as a distribution over words. LDA helps find useful structure in an unstructured collection as it learns distributions over words. This method generates a list of topics from an input document. As LDA is a probabilistic method and starts with a set of random seeds, it might generate different outputs in different runs. So, to guarantee the extraction of a list of good topics, we run it for three iterations and take the intersection to get the final set of topics or keywords.

We then extract the domain-specific terms or keywords from the question. If all the terms of a question belong to the list of terms, then the question is treated as relevant.

Difficulty assessment

We aim to assign a numeric difficulty value to the questions. For representing the difficulty, we assign a 5-level difficulty value. Difficulty value 1 indicates the lowest difficulty, and 5 indicates the highest level of difficulty. For difficulty assessment, a hybrid technique is used that combines three different approaches. That is summarized in Fig. 3, and individual modules are discussed below.

Approach 1

The first approach is based on question classification. We perform question classification for difficulty assessment motivated by Mishra, Kumar, and Saha (2015). For question classification, we define a taxonomy based on different categories of questions that normally appear in the question papers. The taxonomy contains seven classes. These

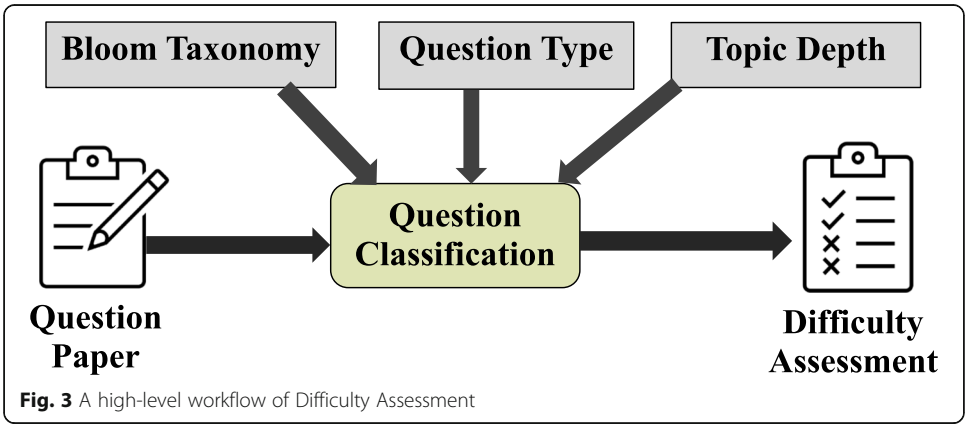

Fig. 3 A high-level workflow of Difficulty Assessment

are factual, define, example, discuss or note, comparative analysis, numerical, advantages, and cause.

To estimate the difficulty of the input question, first, its category level is determined, and a numerical difficulty value is assigned as per the category level. For implementation of the classifier, we prepare a dataset containing questions collected from various sources. Then we train a Support Vector Machine classifier (Cortes & Vapnik, 1995); particularly we used the LibSVM toolkit.[1] We used various features including unigram, bigram, wh-words, other question words, terms, parse-structure similarity. To compute the parse-structure similarity, we use the tree kernel as defined in Moschitti (2006). Then to assign a numerical difficulty, we use the mapping presented in Table 2.

Approach 2

The second approach is also based on classification. Here the basis is Bloom's Taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). Bloom's taxonomy is a hierarchical model used to classify educational learning objectives into levels of complexity and specificity. As per the revised Bloom's taxonomy defined in 2011, the levels are, Remember, Understand, Apply, Analyze, Evaluate, Create. These levels are considered as a measure of difficulty. Answers to the remember and understand level questions are directly available in the textbook. Answering these questions requires reading the topics and memorizing those. To answer apply level questions, the students have to understand the related concepts and apply those to get the result. Such questions are generally tougher than the previous category. Similarly, analyze and evaluate level questions are more difficult. So, a higher level implies a higher level of difficulty (Remember – 1, Create – 6). The mapping is presented in Table 3. The taxonomy level is determined by another classifier. The working of this classifier is the same as the previous one. Given a question, the classifier maps it into Bloom's Taxonomy-based difficulty level.

Approach 3

The third approach is based on the depth of the topic in the corresponding chapter. Casual students often tend to avoid the deeper portions of the syllabus. They do not go through the entire chapter and usually miss the topics written at a greater depth in the chapter hierarchy. On the other hand, the questions that are at lesser depth in the chapter hierarchy are easier for them. To identify the depth of a term, we represent the terms of the chapter as a directed graph. The nodes of the graph represent the terms, and an edge between two nodes indicates that one node calls the other. The term occurred in the chapter heading is denoted as a special node of the graph; all search operation in the graph should start from that node. Now the length of the path from the special node to a particular node denotes the depth of the corresponding term. The depth of the terms presented in the question is computed. These depth values are normalized and mapped to a difficulty score.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Table 2 Question type and corresponding numerical difficulty value

Question Type	Example Question	Numerical Difficulty
Define / Factual question	Define internal fragmentation.	1
Example question	Name three preemptive scheduling algorithm.	2
Discuss question	Discuss the role of time quantum in Round Robin scheduling.	3
Advantage / Cause type	What are the benefits of parallel sorting?	4
Comparative analysis and Numerical questions	Prove that 2-way marge sort is better than 3-way merge sort.	5

Response time estimation

Answer length is yet another metric for measuring the quality of the question paper given the fact that the questions need to be answered in a limited time frame in examinations. One question might take a longer time to answer, which can be adjusted by another quickly answerable question. To assess the time requirement, we use a question-question similarity and model-answer-based approach.

The core of this approach is a knowledge base that contains the length of the answer to the previous set of questions. Initially, the system requires questions and corresponding student-written answers of the previous years. From that, the system identifies the answers that obtained full marks for each question. The average length of these answers is computed. The facts embedded in these answers are extracted, to generate a model answer with human supervision. Now, the system first searches the availability of the target question in the knowledge base. If the question is available, then the corresponding answer helps in estimating the time requirement. Otherwise, the model answer is collected from the paper-setter. The length of the answer is the basis of response time estimation. Additionally, to estimate the time required to write the answer, we conduct a study which is discussed below.

Study design

The objective of the study is to estimate the amount of time required per token during the examination. Six students participated in the study. Among them, two students are high-scoring students having an excellent academic record, two are low-scoring students who struggle to get pass marks, and the rest two are average students. One chapter from the Computer Science subject Operating System was taught to the students for this study. For their preparation for the examination, a study material was provided that also contains a few sample questions and their model answers. A test was

Table 3 Bloom taxonomy and corresponding numerical difficulty value

Bloom Taxonomy Level	Numerical Difficulty
Remember	1
Understand	2
Apply	3
Analyze	4
Evaluate	5
Create	6

conducted after a couple of days. The question paper contained three questions; all are long or descriptive questions. One question was directly from the question-answer pairs provided in the study material. The concept of the second question was discussed in the study material, but the compact answer was not available. The answer of the third question was not directly available in the study material, however, the concept was discussed in class. The time taken to write the individual answers was recorded. The average of the time values and answer lengths are computed, and their ratio is taken as a gold-standard value for response time estimation in our experiment.

System evaluation

This section discusses the evaluation setup, performance of the individual modules and the performance of the overall system.

Evaluation setup

The system is composed of three different modules for assessing three factors, namely, relevance, difficulty, and response time. To evaluate the performance, we collected/compiled a set of question papers from different sources. As these modules need certain subject-specific information and tuning, we choose two subjects for this evaluation: Computer Science subjects 'Operating Systems' and 'Natural Language Processing'. For each subject, 50 question papers were taken in the test set. Each question paper contains 10–15 long questions. Many of these question papers do not contain errors. To test the accuracy of the system, we randomly incorporated certain irrelevant and erroneous questions in these question papers. The system computes various scores for the individual questions as well as the whole question papers. The system-generated scores are then compared with the scores given by human evaluators.

Evaluation of relevance

The relevance module aims to find whether the individual questions of the question-paper are relevant. This module aims to identify the out-of-syllabus questions so that they can be revised beforehand. It is a question-specific metric, and the system provides a binary decision: relevant or irrelevant. For the evaluation, the test set question-papers were given as input from which the system identified the irrelevant questions. A similar task was done by human evaluators. Then the identified irrelevant questions were compared as a measure of system efficiency. We use two metrics namely precision and recall. Precision is defined as the fraction of system retrieved questions that are actually irrelevant. The recall is the fraction of actual irrelevant questions that are successfully identified by the system.

The system marked a total of 65 questions as irrelevant. However, human evaluators felt that a total of 78 questions are irrelevant. When we compared these two sets, we found a match is there in 59 cases. So, among 76 actual irrelevant questions, 32 are identified by the system. So, the recall is 59/76 = 75.64%. The precision of the system is 59/65 = 90.77%. When we analyzed the overall accuracy of the system considering both the irrelevant and relevant questions, then the accuracy of the system became 98.02%. These values indicate that the system is effective in identifying irrelevant questions.

Evaluation of difficulty

The system assigns a numerical difficulty value on a 0 to 5 scale to the questions. Primarily it is a question-specific metric. However, the difficulty value of the whole question paper is also essential. An ideal question paper contains a good distribution of easy, medium, and tough questions. So, the system finally generates an overall difficulty score for the question paper. The difficulty score for the question paper is the weighted average of the question specific difficulty scores and is computed as,

$$\text{Question\_Paper\_Difficulty} = \sum(\text{Question\_Difficulty} * \text{Question\_Weightage})/\text{Full\_Marks}.$$

To assess the performance of the difficulty module, we compare the system-generated difficulty scores with human evaluator scores for the test-set question papers. The final score is represented through root-mean-square error (RMSE). RMSE is calculated as the squared root of the average of square differences between the system scores and human evaluator scores. To compute the RMSE value, we considered all the question papers in the test dataset. The RMSE value obtained here is 0.48 on a 0–5 scale. This good RMSE value indicates that the system is fairly accurate in predicting the difficulty value of the question papers.

We also analyze the question-specific scores. In Table 4, we show the system generated and human evaluator assigned difficulty scores of a few test-set papers. In the table, the "Average Human Score" represents the average of the scores given by the teachers and students. These values also indicate that the system is good at predicting the difficulty value of the question papers. However, the teacher and student scores vary in many cases. That implies the notion of difficulty is not the same for teachers and students. Here we would also like to mention that, both the subjects we considered in this evaluation are from the same domain, and the concerned students were taught by the same teacher. But, there may be cases that there are differences in the way the students are taught and they perceive question difficulty differently. Even, subject-dependent difficulty issues might be there. In this study, we did not consider such cases. The evaluation result may vary in such a scenario.

Evaluation of timing estimation

The system estimates the time-requirement of the test set question papers. The teachers also predict the time-requirement of the question papers. The difference

Table 4 Assessment of question paper difficulty

Question Paper No	System Score	Teacher Score	Student Score	Average Human Score
OS-1	3.42	3.24	3.72	3.48
OS-3	2.64	2.59	3.02	2.81
OS-9	3.6	3.66	4.14	3.9
OS-16	1.95	1.66	1.61	1.64
OS-19	2.1	1.48	1.94	1.71
NLP-2	3.15	2.75	3.22	2.99
NLP-7	4.06	3.88	4.47	4.18
NLP-10	3.58	3.75	3.66	3.71
NLP-11	2.86	2.6	2.74	2.67
NLP-15	2.48	1.78	2.62	2.2

between these two sets of values is taken as the accuracy of the system. The difference is represented through the mean absolute error, and the value is 15.3 min.

Prediction of time-requirement by a teacher might differ from the actual time required by the students. To understand the gap, we perform another study. There we use the actual time required by the students while answering the question paper as the basis of comparison. We selected three question papers for each subject for this study. We conduct a test using these question papers. In each test, we considered three students who answered all the questions and finished the examination within the stipulated time. The time taken to complete their test is recorded. The average of the time is considered as the gold-standard time-requirement for the corresponding question paper. Table 5 presents the values we obtained in this experiment. In the table, all the time-values are in minutes and float values (if any, after taking the average) are mapped to the nearest integer.

In this experiment, we found that there is a gap between the system's predicted value and actual value. However, we also found that a substantial difference is there between the teacher-predicted value and actual response time in some cases. In reality, estimating the time requirement is not straightforward. Several aspects are associated with the actual time taken by a student to answer a question paper. It is very difficult to consider all these factors during system implement. So, the gap is expected. We adopted a feasible-to-implement approach and achieved a moderate accuracy. The necessity of this module during the real application scenario is to detect the question papers that might take time longer than the stipulated time of the examination. The proposed system can be tuned accordingly to solve that purpose.

Conclusion

This paper presented a novel technique for assessing the quality of a question paper using a computer platform. The key factors that affect the quality of a question paper were identified through a student survey. Individual modules for automatic assessment of those factors were implemented using various techniques. The implemented modules were tested using a set of real question papers. The evaluation results indicated that the proposed technique is effective in estimating the question relevance, question paper difficulty, and response time estimation.

The system opens several directions to work in the future. In this first attempt, we have achieved a reasonable accuracy; however, it requires further improvement for real applications. Also, the proposed workflow is complicated and makes use of a certain amount of subject-specific information. It might be difficult to adopt this technique for all possible subjects. Sophisticated techniques may be explored for the individual

Table 5 Assessment of question paper time requirement (in minute)

Question Paper No	System Estimated Time	Teacher Predicted Time	Actual Time Taken by Students
OS-2	172	165	180
OS-4	125	138	144
OS-12	98	120	126
NLP-2	128	140	158
NLP-4	148	162	145
NLP-12	152	124	130

modules. Additionally, to use the system in real applications, a large amount of resources is required. Manual creation of a large resource repository is costly and time-consuming, semi-automatic approaches can be explored there. Additionally, the workflow assumes that the course has a stipulated pre-defined syllabus, and a specific textbook (one or more) is followed. However, in many autonomous settings, a predefined syllabus is not followed; the topics taught in the class form the syllabus. The system might require certain changes with additional inputs to adopt in such a scenario. We feel, the idea is generic and expected to work well in other levels of educations, like school education. However, it requires experimental verification. For timing estimation, we used student examination-based data. Finding some alternative approach to estimate the time can be another direction for future work.

Acknowledgements
I thank all the undergraduate, postgraduate students and research scholars who had taken part in the student survey, study and evaluation of the system.

Author's contributions
The author has designed the study, implemented the system, and performed system evaluation. He has also written the paper. No other person has contributed substantially to meet the criteria of authorship. The author(s) read and approved the final manuscript.

Funding
Not Applicable.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The authors declare that:
- This article does not contain any studies with human participants or animals performed by any of the authors.
- For this type of study formal consent is not required.

Competing interests
The authors declare that they have no competing interests.

Received: 29 December 2020 Accepted: 4 March 2021
Published online: 19 March 2021

References

Anderson, T. R., & Rogan, J. M. (2010). Bridging the educational research-teaching practice gap: Tools for evaluating the quality of assessment instruments. *Biochemistry and Molecular Biology Education*, *38*(1), 51–57.
Araki, J., Rajagopal, D., Sankaranarayanan, S., Holm, S., Yamakawa, Y., & Mitamura, T. (2016). Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING 2016*, (pp. 1125–1136).
Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.
Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. David McKay Company.
Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*, 60–117.
Chali, Y., & Hasan, S. A. (2015). Towards topic-to-question generation. *Computational Linguistics*, *41*(1), 1–20.
Cortes, C., & Vapnik, V. N. (1995). "Support-vector networks" (PDF). *Machine Learning.*, *20*(3), 273–297.
Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333.
Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., … Hu, G. (2017). Question difficulty prediction for READING problems in standard tests. In *Proceedings of the thirty-first AAAI conference on artificial intelligence (AAAI-17)*, (pp. 1352–1359).
Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2019). A systematic review of automatic question generation for educational purposes. *Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-019-00186-y.
Liu, J., Wang, Q., Lin, C.-Y., & Hon, H.-W. (2013). Question difficulty estimation in community question answering services. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, (pp. 85–90).
Luger Sarah. (2016) Algorithms for assessing the quality and difficulty of multiple choice exam questions. Ph.D. Thesis, School of Informatics, University of Edinburgh.
Luger, S., & Bowles, J. (2013). Two methods for measuring question difficulty and discrimination in incomplete Crowdsourced data. In *Proceedings of the first AAAI conference on human computation and crowdsourcing (HCOMP-13)*.

Mishra, S. K., Kumar, P., & Saha, S. K. (2015). A support vector machine based system for technical question classification. In *Proceedings of mining intelligence and knowledge exploration (MIKE), volume 9468 of the series lecture notes in computer science*, (pp. 640–649).
Mohler, M., Bunesc, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of ACL-2011*, (pp. 752–762).
Moschitti, A. (2006). Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of the 17th European Conference on Machine Learning*.
Narendra, A., Agarwal, M., & Shah, R. (2013). Automatic cloze-questions generation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, (pp. 511–515).
Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R. D., & Brefeld, U. (2019). Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*, *29*(3), 342–367.
Rao, D. C. H., & Saha, S. K. (2020). Automatic multiple choice question generation from text: A survey. *IEEE Transaction on Learning Technologies.*, *13*(1), 14–25.
Santos, L. M. R., Fernandez, d. C., Pablo, J., Martin, R. G., Perez, E. V., & Perez, M. J. V. (2012). Automatic classification of question difficulty level: Teachers' estimation vs. students' perception. In *Proceedings of the 2012 IEEE Frontiers in Education Conference*, (pp. 1–5).
Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2017). Evaluation of automatically generated english vocabulary questions. *Research & Practice in Technology Enhanced Learning*, *12*(1), Art. no. 11.
Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2020). An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2019.1648300.
Zhang, L., & VanLehn, K. (2016). How do machine-generated questions compare to human-generated questions? *RPTEL*, *11*(7). https://doi.org/10.1186/s41039-016-0031-7.

Publisher's Note