Current Evaluation Methods are a Bottleneck in Automatic Question Generation

Guher Gorgun GORGUN@UALBERTA.CA
Okan Bulut BULUT@UALBERTA.CA

Abstract

This study provides a comprehensive review of frequently used evaluation methods for assessing the quality of automatic question generation (AQG) systems based on computational linguistics techniques and large language models. As we present a comprehensive overview of the current state of evaluation methods, we discuss the advantages and limitations of each method. Furthermore, we elucidate the next steps for the full integration of automatic question generation systems in educational settings to achieve effective personalization and adaptation.

Keywords: automatic question generation, evaluation methods, machine translation, crowdsourcing, human evaluators, ablation studies

1. Introduction

Developing high-quality questions has been a major challenge for educators because this process requires not only content expertise (i.e., domain knowledge) but also assessment expertise including a solid understanding of how each element of a question (e.g., question stem, response options, or source materials) should be designed. Automatic question generation (AQG), also known as automatic item generation, has emerged as a viable solution to address the challenges around question development. AQG involves using computer algorithms to create test items, questions, or assessments automatically. AQG streamlines the question generation process, allowing educators and assessment developers to create a large number of items in a relatively short amount of time. This efficiency reduces the overall cost of test development, as it minimizes the need for manual labor and resources traditionally associated with creating assessments. AQG systems have been widely used in online learning platforms where students answer a large number of questions to evaluate the current state of their learning.

As educational assessments increasingly transition to digital platforms, AQG plays a crucial role in seamlessly integrating technology into the assessment process. However, current AQG systems fall short of consistently producing questions that are universally effective and readily deployable in digital learning and assessment environments. Each question generated by AQG necessitates thorough scrutiny for two primary purposes: (1) evaluating the overall quality and utility of the AQG system, and (2) discerning which questions are operationally suitable for educational settings. Efficiently generating numerous questions represents only the initial phase of a successful question-generation pipeline. This must be followed by robust evaluation methods to gauge the quality and usability of the questions. This study aims to comprehensively review the most frequently employed evaluation methods in AQG, elucidating the advantages and limitations of each method to diagnose the

evaluation needs of AQG for translating generated questions into real-world applications. The second objective is to bridge the gap between a psychometric approach and a computer science approach to question evaluation, providing a holistic overview.

2. Evaluation Methods in AQG

In this section, we discuss the most commonly used methods for evaluating AQG and their advantages and limitations. We present a summary of a comprehensive literature of recent peer-reviewed articles and publications focused on AQG (Amidei et al., 2018; Kurdi et al., 2020; Divate and Salgaonkar, 2017; Mulla and Gharpure, 2023; Soni et al., 2019).

2.1. Human Evaluators

Using manual coding practices and rating scales, human evaluators assess the quality of the generated questions. Human evaluators include experts, students, researchers, and teachers (Wang et al., 2022; Rodriguez-Torrealba et al., 2022; Attali et al., 2022; Gierl and Lai, 2016; Chung and Hsiao, 2022; Dugan et al., 2022; Maurya and Desarkar, 2020; Chughtai et al., 2022; Sarsa et al., 2022). Human evaluators may utilize a rating scale that may include quality criteria such as question difficulty (Rodriguez-Torrealba et al., 2022; Chung and Hsiao, 2022), distractor functioning (Maurya and Desarkar, 2020), domain relevance (Chughtai et al., 2022; Dugan et al., 2022; Elkins et al., 2023), fluency (Mostow et al., 2017), and grammatical accuracy (Chughtai et al., 2022).

Human evaluators have been one of the fundamental methods of question evaluation in automatic as well as traditional psychometric question evaluation approaches (Engelhard Jr et al., 1999; Osterlind and Osterlind, 1989). They may provide ground truth about the quality of the generated questions if a valid rating scale is used and a training process is employed to achieve standardization and inter-rater reliability among the raters. Yet, human evaluators are typically expensive. It takes a longer time to evaluate the generated questions because each question is evaluated individually and by at least two raters to assess the inter-rater reliability among the evaluators. Simply put, human evaluators are a violation of the basic AQG assumption—we can generate questions instantly in a cost-efficient manner. Unless questions are also evaluated in a similar cost-efficient way, human evaluations typically do not contribute toward the fundamental properties of AQG. This is perhaps the most serious limitation of human evaluators. Additionally, the lack of reporting and training practices is a threat to the validity of quality labels assigned to questions.

2.2. Crowdsourcing

Crowdsourcing is an alternative approach to human evaluators and may overcome several limitations of the human evaluators that we discussed above. Crowdsourcing is defined as mobilizing competence and expertise among a crowd (Zhao and Zhu, 2014). Thus, using crowdsourcing platforms (e.g., Amazon's Mechanical Turk; Litman and Robinson (2020)), the generated questions can be evaluated quickly and cost-effectively, and typically rating scales have been used to provide quality criteria to crowdsource workers (Becker et al., 2012; Lin et al., 2015; von Davier, 2018).

While crowdsourcing platforms offer a cost-effective means for evaluating a substantial number of generated questions, the reliability of evaluations through crowdsourcing has raised notable concerns (Fort et al., 2011). Issues related to ethical considerations (Zhao and Zhu, 2014) and the representativeness of crowdsource workers (Brunt and Meidell, 2017; Wang et al., 2023) further compound the apprehensions surrounding this evaluation approach. Beyond the concerns about the quality of crowdsourced evaluations, akin to concerns with human evaluators, the absence of standardized reporting and training processes emerges as another significant challenge. This lack of structure poses a serious threat to the validity and reliability of ratings obtained through crowdsourcing, highlighting the need for a more systematic and transparent approach to ensure the robustness of assessments in the AQG domain. Addressing these concerns is pivotal for establishing the credibility and integrity of crowdsourced evaluations in the context of AQG systems.

2.3. Ablation Studies

Ablation studies, the least frequently used evaluation method in AQG systems (Wang et al., 2021), involve the systematic removal of an essential component from the AQG system to assess the impact on the question generation process. This method allows researchers to dissect the contributions of individual components, shedding light on their relative importance and influence. For example, in an AQG system designed for language assessment, an ablation study could involve temporarily disabling the natural language processing module responsible for semantic analysis. By observing how the omission of this component affects the system's ability to generate contextually relevant questions, researchers can gain insights into the module's significance in ensuring the quality and appropriateness of the generated questions.

While ablation studies prove to be valuable in the evaluation of AQG systems, a significant limitation lies in the fact that these studies compare the system's performance against itself when specific components are removed. This approach, although insightful for understanding the overall impact of individual components on system performance, falls short in providing a nuanced assessment of the quality of each question generated by the system. The inherent drawback is that the evaluation result pertains to the system's collective performance rather than offering a granular analysis of the merit of each individual question produced. To overcome this limitation, researchers may need to complement ablation studies with additional evaluation methodologies that provide a more fine-grained analysis of the generated questions. This could involve employing metrics that assess the relevance, coherence, or difficulty of individual questions, thereby offering a more comprehensive understanding of the nuanced aspects of question quality within the AQG system.

2.4. Post-Hoc Analysis

This group of evaluation methods involves administering, typically, a subset of questions generated from a representative sample and then evaluating question quality using the learners' response data. Similar to human evaluators, this method has been one of the foundational approaches to question quality evaluation under traditional psychometric evaluation methods (Anastasi and Urbina, 1997; Clauser and Hambleton, 2012; Ebel and Frisbie, 1986). The AQG researchers may administer a subset of questions generated to a group of learners

to obtain item statistics such as difficulty, discrimination, or distractor functioning (Gierl and Lai, 2016; Attali et al., 2022). Alternatively, they can compare the effect of AQG systems on learner motivation and academic achievement (Yang et al., 2021; Van Campenhout et al., 2022).

Since post-hoc analysis provides statistical estimates regarding item quality, they are less susceptible to bias in human judgment (Seyler et al., 2017). Nonetheless, the assessment conditions should replicate the real assessment settings to minimize construct-irrelevant variance, which may contaminate the quality of statistical estimates obtained (Gorgun and Bulut, 2021, 2022). Another limitation is that only a subset of generated questions can be administered, and thus all of the generated items cannot be evaluated with this approach, limiting the appeal of employing post-hoc analysis for the evaluation of questions generated. Furthermore, this is a resource-intensive approach to question evaluation because a group of representative learners should be recruited to obtain item statistics. A final limitation is that question quality is obtained in a retrospective manner, meaning the question quality is typically unknown prior to assessment administration. This may lead to undesirable consequences for the learners, for example, catalyzing confusion and frustration in learners when a low-quality question is administered during the assessment (Ali and Ruit, 2015).

2.5. Machine Translation Metrics

AQG researchers adopted metrics developed for machine translation and text summarization tasks for evaluating generated questions automatically. These metrics typically include BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Lavie and Denkowski, 2009) and allow the AQG researchers to compare the similarity between generated questions and reference questions (Gao et al., 2018; Kumar et al., 2018; Rodriguez-Torrealba et al., 2022; Wang et al., 2021; Liang et al., 2023; Ghanem et al., 2022). Thus, this evaluation method required questions—typically human-authored ones—already available in the dataset.

While this approach provides swift and cost-effective quality indices for generated questions, it faces limitations in scenarios where reference questions are lacking. Moreover, the reliance on linguistic structure comparison with reference questions may lead to suboptimal scoring for questions that are well-crafted but differ linguistically (Kurdi et al., 2020). Therefore, the effectiveness of this method is contingent on the presence of a diverse and representative set of reference questions in the dataset.

3. Implications and Future Directions

While the fundamental assumption of AQG is that it generates questions instantly and efficiently, the evaluation of the questions generated is at the heart of determining the usability of the questions generated. In this study, we delved into the limitations inherent in commonly employed methods used for evaluating AQG systems and elucidated the reasons why these methods persist as a bottleneck in AQG research. This comprehensive overview underscores the pressing need for the development and adoption of automatic evaluation methods capable of efficiently assessing the entirety of questions generated by AQG systems.

To address these challenges inherent in current evaluation methods, researchers may benefit from sharing AQG publicly. That is a noteworthy opportunity for AQG researchers

to leverage publicly available datasets that encompass not only the questions automatically generated but also the corresponding evaluation metrics. The inclusion of such datasets can significantly enhance the robustness and applicability of AQG research outcomes. For instance, consider a scenario where a publicly accessible dataset contains a diverse array of automatically generated questions paired with comprehensive evaluation metrics, including BLEU, ROUGE-L, and METEOR scores. Researchers can utilize this dataset to not only train and validate their AQG models but also to benchmark and compare the performance of different models on a standardized set of questions. This approach fosters transparency and reproducibility in AQG research, as it allows others to independently assess and validate the efficacy of novel question-generation techniques.

Additionally, establishing standardized quality criteria for evaluating the generated questions is imperative. This not only ensures a benchmark for assessing the merit of generated questions but also facilitates the development of pre-trained models. These tools, in turn, can play a pivotal role in expediting the operational use of generated questions in real-world educational settings. For instance, envision a standardized set of criteria that not only evaluates the grammatical correctness and coherence of generated questions but also considers their alignment with specific educational objectives and the potential for promoting critical thinking skills. By incorporating such nuanced criteria, the evaluation process becomes more holistic and aligned with the multifaceted goals of education. Standardized criteria can also provide a framework for ongoing refinement and optimization of AQG models, encouraging researchers to continually enhance the quality of generated questions. This iterative process not only raises the overall standard of AQG but also ensures that the technology evolves to meet the dynamic needs of educators and learners.

References

Syed Haris Ali and Kenneth G Ruit. The impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspectives on medical education*, 4:244–251, 2015.

Jacopo Amidei, Paul Piwek, and Alistair Willis. Evaluation methodologies in automatic question generation 2013-2018. 2018.

Anne Anastasi and Susana Urbina. *Psychological testing*. Prentice Hall/Pearson Education, 1997.

Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A von Davier. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903077, 2022.

Lee Becker, Sumit Basu, and Lucy Vanderwende. Mind the gap: learning to choose gaps for question generation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 742–751, 2012.

Liam Brunt and Erik Meidell. When are crowdsourced data truthful, accurate, and representative. *The Journal of Business Inquiry*, 17(1):55–71, 2017.

Rudeema Chughtai, Farooque Azam, Muhammad Waseem Anwar, Wasi Haider But, and Muhammad Umar Farooq. A lecture centric automated distractor generation for post-graduate software engineering courses. In *2022 International Conference on Frontiers of Information Technology (FIT)*, pages 100–105. IEEE, 2022.

Cheng-Yu Chung and I-Han Hsiao. Programming question generation by a semantic network: A preliminary user study with experienced instructors. In *International Conference on Artificial Intelligence in Education*, pages 463–466. Springer, 2022.

Jerome C Clauser and Ronald K Hambleton. Item analysis procedures for classroom assessments in higher education. In *Handbook on measurement, assessment, and evaluation in higher education*, pages 316–329. Routledge, 2012.

Manisha Divate and Ambuja Salgaonkar. Automatic question generation approaches and evaluation techniques. *Current Science*, pages 1683–1691, 2017.

Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, Dayheon Choi, Chuning Yuan, and Chris Callison-Burch. A feasibility study of answer-agnostic question generation for education. *arXiv preprint arXiv:2203.08685*, 2022.

RL Ebel and DA Frisbie. Using test and item analysis to evaluate and improve test quality. *Essentials of educational measurement (*, pages 223–242, 1986.

Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie CK Cheung. How useful are educational questions generated by large language models? In *International Conference on Artificial Intelligence in Education*, pages 536–542. Springer, 2023.

George Engelhard Jr, Melodee Davis, and Linda Hansche. Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education*, 12(2):199–210, 1999.

Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, pages 413–420, 2011.

Yifan Gao, Lidong Bing, Wang Chen, Michael R Lyu, and Irwin King. Difficulty controllable generation of reading comprehension questions. *arXiv preprint arXiv:1807.03586*, 2018.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer McIntosh von der Ohe, and Alona Fyshe. Question generation for reading comprehension assessment by modeling how and what to ask, 2022.

Mark J Gierl and Hollis Lai. A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice*, 35(4):6–20, 2016.

Guher Gorgun and Okan Bulut. A polytomous scoring approach to handle not-reached items in low-stakes assessments. *Educational and Psychological Measurement*, 81(5):847–871, 2021.

Guher Gorgun and Okan Bulut. Considering disengaged responses in Bayesian and deep knowledge tracing. In *International Conference on Artificial Intelligence in Education*, pages 591–594. Springer, 2022.

Vishwajeet Kumar, Kireeti Boorla, Yogesh Meena, Ganesh Ramakrishnan, and Yuan-Fang Li. Automating reading comprehension by generating question and answer pairs. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*, pages 335–348. Springer, 2018.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204, 2020.

Alon Lavie and Michael J Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23:105–115, 2009.

Yuanyuan Liang, Jianing Wang, Hanlun Zhu, Lei Wang, Weining Qian, and Yunshi Lan. Prompting large language models with chain-of-thought for few-shot knowledge base question generation. *arXiv preprint arXiv:2310.08395*, 2023.

Chenghua Lin, Dong Liu, Wei Pang, and Edward Apeh. Automatically predicting quiz difficulty level using similarity measures. In *Proceedings of the 8th International Conference on Knowledge Capture*, pages 1–8, 2015.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Leib Litman and Jonathan Robinson. *Conducting online research on Amazon Mechanical Turk and beyond*. Sage Publications, 2020.

Kaushal Kumar Maurya and Maunendra Sankar Desarkar. Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1115–1124, 2020.

Jack Mostow, Yi-Ting Huang, Hyeju Jang, Anders Weinstein, Joe Valeri, and Donna Gates. Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children's comprehension while reading. *Natural Language Engineering*, 23(2):245–294, 2017.

Nikahat Mulla and Prachi Gharpure. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32, 2023.

Steven J Osterlind and Steven J Osterlind. Judging the quality of test items: Item analysis. *Constructing test items*, pages 259–310, 1989.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Ricardo Rodriguez-Torrealba, Eva Garcia-Lopez, and Antonio Garcia-Cabot. End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, 208:118258, 2022.

Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43, 2022.

Dominic Seyler, Mohamed Yahya, and Klaus Berberich. Knowledge questions from knowledge graphs. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 11–18, 2017.

Sonam Soni, Praveen Kumar, and Amal Saha. Automatic question generation: A systematic review. In *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India*, 2019.

Rachel Van Campenhout, Martha Hubertz, and Benny G Johnson. Evaluating ai-generated questions: A mixed-methods analysis using question data and student perceptions. In *International Conference on Artificial Intelligence in Education*, pages 344–353. Springer, 2022.

Matthias von Davier. Automated item generation with recurrent neural networks. *psychometrika*, 83(4):847–857, 2018.

Ding Wang, Mark Díaz, Alicia Parrish, Lora Aroyo, Chris Homan, Greg Serapio-García, Vinodkumar Prabhakaran, and Alex Taylor. All that agrees is not gold: Evaluating ground truth labels and dialogue content for safety, 2023.

Zichao Wang, Andrew S Lan, and Richard G Baraniuk. Math word problem generation with mathematical consistency and problem context constraints. *arXiv preprint arXiv:2109.04546*, 2021.

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G Baraniuk. Towards human-like educational question generation with large language models. In *International conference on artificial intelligence in education*, pages 153–166. Springer, 2022.

Albert CM Yang, Irene YL Chen, Brendan Flanagan, and Hiroaki Ogata. Automatic generation of cloze items for repeated testing to improve reading comprehension. *Educational Technology & Society*, 24(3):147–158, 2021.

Yuxiang Zhao and Qinghua Zhu. Evaluation on crowdsourcing research: Current status and future direction. *Information systems frontiers*, 16:417–434, 2014.