АҚПАРАТТЫҚ-КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР
ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION AND COMMUNICATION TECHNOLOGIES

ЖАСАНДЫ ИНТЕЛЛЕКТ ЖӘНЕ МАШИНАЛЫҚ ОҚЫТУ
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DOI 10.51885/1561-4212_2025_2_172
IRSTI 20.53.23

D.V. Lopatkin[1], N.P. Rokhas Kriulko[2], I.A. Kotlyarova[3], G.V. Popova[4], Y.A. Vais[5]
D. Serikbayev East-Kazakhstan Technical university, Ust-Kamenogorsk, Kazakhstan
[1]*E-mail: lopatkin.d@edu.ektu.kz*
[2]*E-mail: nrohas@edu.ektu.kz**
[3]*E-mail: IKotlyarova@edu.ektu.kz*
[4]*E-mail: gpopova@edu.ektu.kz*
[5]*E-mail: YuVais@edu.ektu.kz*

ADVANCES IN AUTOMATIC QUESTION GENERATION: A SURVEY OF AUTOMATIC QUESTION GENERATION TECHNIQUES, DATASETS, AND EVALUATION

СҰРАҚТАРДЫ АВТОМАТТЫ ТҮРДЕ ГЕНЕРАЦИЯЛАУ САЛАСЫНДАҒЫ ЖЕТІСТІКТЕР: СҰРАҚТАРДЫ АВТОМАТТЫ ТҮРДЕ ГЕНЕРАЦИЯЛАУ ӘДІСТЕРІНЕ, ДЕРЕКТЕР ЖИЫНТЫҒЫНА ЖӘНЕ БАҒАЛАУҒА ШОЛУ

ДОСТИЖЕНИЯ В ОБЛАСТИ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ВОПРОСОВ: ОБЗОР МЕТОДОВ, НАБОРОВ ДАННЫХ И ОЦЕНОК АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ВОПРОСОВ

Abstract. *Automatic Question Generation (AQG) is a rapidly growing area within artificial intelligence (AI) and natural language processing (NLP), focused on creating questions automatically from various sources like raw text, databases, and semantic representations. This review explores a wide range of AQG approaches, from traditional rule-based methods to advanced neural network models, including sequence-to-sequence, transformer-based, and graph-based architectures, as well as hybrid methods that combine linguistic rules with machine learning techniques. While rule-based systems offer clarity and control, they often struggle with complex language structures, whereas neural models, especially those using transformers like T5 and BART, have transformed AQG by enabling end-to-end learning and generating more contextually relevant questions. Hybrid models aim to balance the strengths of both approaches, enhancing flexibility and adaptability. The review also discusses evaluation methods, including automated metrics like BLEU, ROUGE, and METEOR, along with human assessments. Despite notable progress, challenges remain in achieving natural question fluency, semantic accuracy, and the generation of high-quality distractors for multiple-choice questions. Looking ahead, promising research directions include lifelong learning models, multimodal question generation that integrates text with images or code, and more robust evaluation frameworks. This review offers insights for researchers and practitioners, emphasizing AQG's potential to improve educational tools, conversational agents, and information retrieval systems.*
Keywords: *Automatic Question Generation, Natural Language Processing, Rule-based approaches, Neural Networks.*

Аңдатпа. *Сұрақтарды автоматты түрде жасау (AQG) – мәтіндік материалдар, дерекқорлар және семантикалық көріністер сияқты әртүрлі көздерден сұрақтарды автоматты түрде жасауға бағытталған жасанды интеллект (AI) және табиғи тілді өңдеудің (NLP) қарқынды дамып келе жатқан саласы. Бұл шолу AQG әдістерінің кең ауқымын қамтиды, дәстүрлі ережеге негізделген*

әдістерден нейрондық желілердің кеңейтілген үлгілеріне, соның ішінде реттілікке, трансформаторға және графикаға негізделген архитектураларға, сондай-ақ лингвистикалық ережелерді машиналық оқыту әдістерімен біріктіретін гибридтік әдістерге дейін. Ережеге негізделген жүйелер анықтық пен бақылауды қамтамасыз еткенімен, олар күрделі тілдік құрылымдармен жиі күреседі, ал нейрондық модельдер, әсіресе T5 және BART сияқты трансформаторларды пайдаланатындар, соңына дейін оқуға мүмкіндік беру және контекстке қатысты сұрақтарды шығару арқылы AQG-ны өзгертті. Гибридті модельдер икемділік пен бейімделуді арттыра отырып, екі тәсілдің де күшті жақтарын теңестіруге бағытталған. Шолу сонымен қатар бағалау әдістерін, соның ішінде BLEU, ROUGE және METEOR сияқты автоматтандырылған көрсеткіштерді, сондай-ақ адам бағалауларын талқылайды. Елеулі прогреске қарамастан, сұрақтардың табиғи еркіндігіне, семантикалық дәлдігіне қол жеткізу және бірнеше таңдаулы сұрақтар үшін жоғары сапалы дистракторларды жасауда қиындықтар әлі де бар. Перспективалы зерттеу бағыттары өмір бойы оқыту үлгілерін, мәтінді суреттермен немесе кодпен біріктіретін мультимодальды сұрақтарды генерациялауды және сенімдірек бағалау жүйелерін қамтиды. Бұл шолу AQG-нің білім беру құралдарын, сөйлесу агенттерін және ақпарат іздеу жүйелерін жақсартудағы әлеуетін көрсете отырып, зерттеушілер мен практиктерге әсер етеді.

Түйін сөздер: *Сұрақтарды автоматты түрде құру, табиғи тілді өңдеу, ережеге негізделген тәсілдер, нейрондық желілер.*

Аннотация*. Автоматическая генерация вопросов (AQG) – стремительно развивающаяся область искусственного интеллекта (AI) и обработки естественного языка (NLP), ориентированная на автоматическое создание вопросов из различных источников, таких как текстовые материалы, базы данных и семантические представления. В данном обзоре рассматривается широкий спектр подходов к AQG, от традиционных методов, основанных на правилах, до продвинутых нейросетевых моделей, включая архитектуры на основе «последовательность-последовательность», трансформеров и графов, а также гибридные методы, сочетающие лингвистические правила с методами машинного обучения. Хотя системы, основанные на правилах, обеспечивают ясность и контроль, они часто не справляются со сложными языковыми структурами, в то время как нейронные модели, особенно использующие трансформеры, такие как T5 и BART, изменили AQG, обеспечив сквозное обучение и генерирование более контекстуально значимых вопросов. Гибридные модели призваны сбалансировать сильные стороны обоих подходов, повышая гибкость и адаптивность. В обзоре также рассматриваются методы оценки, включая автоматизированные метрики, такие как BLEU, ROUGE и METEOR, а также человеческие оценки. Несмотря на заметный прогресс, остаются проблемы, связанные с достижением естественной беглости вопросов, семантической точности и созданием высококачественных дистракторов для вопросов с несколькими вариантами ответов. Перспективные направления исследований включают модели обучения на протяжении всего существования, мультимодальную генерацию вопросов, объединяющую текст с изображениями или кодом, и более надежные системы оценки. В обзоре представлены наработки для исследователей и практиков, подчеркивающие потенциал AQG для улучшения образовательных инструментов, разговорных агентов и информационно-поисковых систем.*

Ключевые слова: *Автоматическая генерация вопросов, обработка естественного языка, подходы на основе правил, нейронные сети.*

Introduction. Automatic question generation (AQG) stands as a pivotal area within artificial intelligence (AI), focused on automatically crafting questions from diverse input sources, encompassing databases, semantic representations, and raw text (Tran, Nguyen, Tran, Vo, 2023). Previous surveys that have explored AQG have primarily concentrated on specific facets or methodologies, such as rule-based approaches or particular applications (Yuan et al., 2022). This limited focus has left a gap in terms of a comprehensive understanding of the broader field of AQG. This survey distinguishes itself by presenting a systematic and in-depth examination of AQG, encompassing both traditional and modern techniques, a wide array of question categories, and a range of evaluation methods.

Traditional rule-based methods for QG, grounded in predetermined linguistic rules and patterns, offer transparency and controllability (Tran, Nguyen, Tran, Vo, 2023). However, these methods often demand extensive manual effort and may encounter difficulties when confronted

with complex language structures. The emergence of neural QG models, particularly those built upon deep learning architectures such as sequence-to-sequence and transformer networks, has signaled a paradigm shift in the field (Dhole, Manning, 2022; Naeiji et al., 2023). These data-driven approaches present end-to-end trainable frameworks that can learn intricate patterns from expansive datasets and generate questions that are more contextually relevant and diverse. Notable advancements in neural QG include the incorporation of attention mechanisms, copy mechanisms, and pre-trained language models, contributing to enhancements in performance and fluency within the generated questions (Fei, Zhang, Zhou, 2021; Sun et al., 2022; Naeiji et al., 2023; Dhole, Manning, 2022).

Section 1 provides an overview of various approaches to Automatic Question Generation (AQG), including rule-based, sequence-based, transformation-based, graph-based models and hybrid approaches. Section 2 focuses on evaluation methods, covering both automatic metrics and human assessments. Finally, Section 3 wraps up with the conclusion.

Classification of automatic question generation approaches. Automatic question generation (AQG) is a task that involves generating questions from various input formats like text, data, and knowledge bases (Leite, Cardoso, 2023; Blstak, Rozinajova, 2022). AQG systems can be categorized into three main types: rule-based, neural network-based and hybrid approaches (Huang et al., 2021). The following section will delve into a detailed classification of these approaches, examining the diverse methodologies and techniques employed within each category.

Rule-based approaches. Rule-based question generation (QG) systems operate by applying predefined linguistic rules and patterns to transform declarative sentences into questions. These systems typically involve a multi-stage process (Leite, Cardoso, 2023; Zhang, Zhang, Wang, 2022; Naeiji et al., 2023): First, the input sentence is parsed using techniques like dependency parsing (Sewunetie, Kovacs, 2024; Huang et al., 2021; Zhang, Wang, 2022), part-of-speech tagging (Blstak, Rozinajova, 2022), named entity recognition (Sewunetie, Kovacs, 2024), and semantic role labeling (Zhang, Zhang, Wang, 2022; Sewunetie, Kovacs, 2024; Leite, Cardoso, 2023) to identify key linguistic features. Next, this information is matched against a set of predefined rules or templates that specify how different sentence structures should be transformed into questions (Huang et al., 2021; Zhang, Wang, 2022; Barlybayev, Matkarimov, 2024). Finally, the system uses these rules to generate a corresponding question, often targeting specific elements of the input sentence, like the subject, object, or location. While this approach may require considerable effort in crafting rules, rule-based QG systems offer transparency, controllability, and ensure grammatically well-formed questions (Dhole, Manning, 2022). They are particularly useful in scenarios with limited training data, such as low-resource languages, where neural network-based methods may not be feasible (Leite, Cardoso, 2023). Table 1 provides a comparative overview of rule-based approaches.

Neural Network-based approaches. Neural network-based approaches have achieved remarkable success in Automatic Question Generation (AQG), surpassing traditional rule-based methods by leveraging the power of deep learning to learn complex patterns from data and generate diverse and contextually relevant questions. These approaches can be broadly categorized into three main types: Seq2Seq, Transformer-based, and Graph-based approaches. The sequence-to-sequence (Seq2Seq) architecture is at the core of most neural automatic question generation (AQG) systems (Zhang, Zhang, Wang, 2022; Leite, Cardoso, 2023; Barlybayev, Matkarimov, 2024). This architecture is built around two key components: an encoder and a decoder. The encoder processes the input text, such as a sentence or paragraph, and converts it into a fixed-length vector, often called the context vector, which encapsulates the essential meaning of the input. The decoder then uses this context vector to generate a question, crafting it word by word. By leveraging patterns from training data and the context provided by the encoder, the decoder transforms the encoded information into coherent and relevant questions.

Table 1. Summary of rule-based approaches used for AQG

Model	Automatic evaluation	Human evaluation		
Dependency parsing, part-of-speech tagging, chunking, and named entity recognition to identify sentence templates	-	412 questions evaluated in total		
			Simple sentences	Complex sentences
		Grammatical	76.95%	76.95%
		Make sense	52.31%	52.31%
		Vague	47.69%	47.69%
		Correct	329	77
		Spurious	300	105
		Missed	57	39
		Precision	0.52	0.42
		Recall	0.85	0.66
		F-Score	0.65	0.51
Tag-set sequence to capture syntactic and semantic information	-		Total	
		TSSP-DB entry pairs	122	
		QAPs generated	796	
		All correct	773	
		Syntactically acceptable	13	
		Semantically acceptable	8	
		Unacceptable	2	
Syntactic rules with universal dependencies, shallow semantic parsing, and back-translation	BLEU-1 BLEU-2 BLEU-3 BLEU-4	45.55 30.24 23.84 18.72		Grammatical correctness and relevance score. The inter-rater agreement (Krippendorff's coefficient) is 0.72
Syntactic information, semantic roles, dependency labels, discourse connectors, and relative pronouns/adverbs	BLEU-4 ROUGE-L BERTScore	32.33 53.14 85.49		Machine-generated questions are indistinguishable from human-authored questions in 51.9% of the cases
Dependency parsing, NER, and adverb/noun subtype analysis	Metrics	Type	Score	Overall score 3.67 out of 5.0
	BLUE	1-gram	0.86	
		2-gram	0.78	
		3-gram	0.77	
		4-gram	0.75	
	ROUGE-1	F1-Score	0.62	
		Precision	0.59	
	ROUGE-2	Recall	0.65	
		F1-Score	0.53	
	ROUGE-L	Precision	0.52	
		Recall	0.55	
		F1-Score	0.62	
		Precision	0.59	
		Recall	0.65	

Note: compiled by the author based on (Tran, Nguyen, Tran, Vo, 2023; Yuan et al., 2022; Dhole, Manning, 2022; Leite, Cardoso, 2023; Sewunetie, Kovacs, 2024)

Early sequence-to-sequence (Seq2Seq) models for automatic question generation (AQG) primarily relied on Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, for both the encoding and decoding stages (Barlybayev, Matkarimov, 2024; Leite, Cardoso, 2023; Huang et al., 2021). While RNNs are well-suited for processing sequential data, they often struggle to capture long-term dependencies in text (Blstak, Rozinajova, 2022; Fei, Zhang, Zhou, 2021). LSTMs address this challenge by introducing a memory cell that can selectively retain or discard information from earlier steps, allowing them to manage the complexities of language more effectively (Barlybayev, Matkarimov, 2024; Huang et al., 2021; Leite, Cardoso, 2023). Table 2 offers a comparative summary of various Seq2Seq models.

Table 2. Summary of Seq2seq-based models

<table>
<tr><th rowspan="2">Model</th><th rowspan="2">Dataset</th><th colspan="5">Automatic evaluation</th><th colspan="3">Human evaluation</th></tr>
<tr><th>BLEU-1</th><th>BLEU-2</th><th>BLEU-3</th><th>BLEU-4</th><th>ROUGE-L</th><th>Repetition</th><th>Incomplete</th><th>Accuracy</th></tr>
<tr><td>Fine-grained Question Type Classifier and BiLSTM Encoder</td><td>SQuAD
HotpotQA
TibetanQA</td><td>47.28
42.35
42.45</td><td>31.35
30.42
35.07</td><td>23.02
23.83
29.64</td><td>17.52
19.37
25.58</td><td>46.78
17.52
43.28</td><td>6.48%
5.63%
7.38%</td><td>19.18%
16.20%
39.34%</td><td>87.67%
65.50%
46.72%</td></tr>
<tr><td rowspan="3">Semantic Role Labeler (Seq2seq model and two semantic mappers)</td><td rowspan="3">SQuAD
Car Manuals
NewsQA</td><td colspan="5">Precision</td><td colspan="3" rowspan="3">Scale (1 – 5)
Clarity: 4.75
Relatedness: 4.61
Grammar: 4.93</td></tr>
<tr><td colspan="2">METEOR</td><td colspan="2">BLEU-4</td><td>ROUGE-L</td></tr>
<tr><td colspan="2">21.8
61.8</td><td colspan="2">20.02
85.1</td><td>46.9
93.7</td></tr>
<tr><td colspan="10">Note: compiled by the author based on (Sun et al., 2022; Naeiji et al., 2023)</td></tr>
</table>

The introduction of Transformer models marked a paradigm shift in neural AQG. Transformers, like T5 (Zhang, Zhang, Wang, 2022), BART (Barlybayev, Matkarimov, 2024; Su et al., 2020; Dijkstra et al., 2024), BERT (Barlybayev, Matkarimov, 2024), UniLM (Leite, Cardoso, 2023; Zhang, Zhang, Wang, 2022), GPT (Dijkstra et al., 2024; Murakhovska et al., 2022) and ERNIE-GEN (Leite, Cardoso, 2023; Murakhovska et al., 2022), have revolutionized the field by relying solely on self-attention mechanisms rather than the sequential processing of RNNs. This allows them to process the entire input sentence in parallel, enabling them to capture long-range dependencies more effectively and generate more fluent and coherent questions (Lopez et al., 2021; Dijkstra et al., 2024; Sun et al., 2022; Dhole, Manning, 2022; Yuan et al., 2022; Tran, Nguyen, Tran, Vo, 2023). The success of Transformer-based AQG is further amplified by pre-training on massive text datasets. These pre-trained models come equipped with rich language representations that can be fine-tuned for specific tasks like AQG. By leveraging these pre-trained representations, Transformer-based AQG models have achieved state-of-the-art results on various benchmarks (Su et al., 2020; Dijkstra et al., 2024; Murakhovska et al., 2022; Yuan et al., 2022). Automatic evaluation is often conducted on different datasets to assess their performance. A summary of Transformer-based models is

presented in Table 3.

Table 3. Summary of transformer-based models

Model	Dataset	Automatic evaluation							Human evaluation
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BERTScore	
Unified-QG employs a format-convert encoding to transform various QG formats into a unified representation	MCTest SQuAD RACE NarrativeQA Arc-easy Arc-hard OpenbookQA BoolQA	54.35 46.50 36.59 38.64 30.85 27.04 34.32 48.84	40.08 32.29 24.86 23.45 19.81 16.51 22.57 33.10	32.09 24.44 17.71 16.43 13.85 11.05 15.76 24.19	25.48 19.20 12.95 12.19 10.17 7.83 11.33 18.04	55.54 47.42 33.53 41.59 36.79 33.66 35.18 46.08	27.75 23.90 19.33 18.64 17.78 15.82 19.43 23.27	-	-
Uses a single pre-trained GPT-2 transformer for paragraph-level QG without additional features or answer metadata	SQuAD	55.28	30.81	16.55	8.21	44.27	21.11	-	-
EduQuiz is based on a GPT-3 model fine-tuned on text-quiz pairs for end-to-end quiz generation	EQG-RACE (processed RACE dataset with only examination questions)	-	-	-	11.61	36.11	25.42	-	Fluency, Relevancy, Answerability of: Question: 96.3% Answer: 89.4% Distractors: 85.8%
MixQG leverages a text-to-text framework with pre-trained language models like T5 and BART	SQuAD NQ QAConv Quoref DROP TweetQA	- - - - - -	- - - - - -	- - - - - -	23.46 31.25 22.74 27.36 28.53 18.66	50.10 57.84 44.40 44.42 51.12 45.94	44.15 55.90 39.93 42.06 47.83 46.60	0.5582 0.5351 0.4533 0.4137 0.5493 0.4645	No error: 68.4% Disfluent: 9.7% Off target: 5.8% Wrong Context: 16.2%
Leaf fine-tunes the T5 transformer on SQuAD 1.1 for multi-task question and answer generation	SQuAD RACE	34.47	-	-	-	-	-	-	Exact match: 41.51% F1-Score: 53.26%
TP3 fine-tunes pre-trained transformers on QAP datasets	SQuAD Gaokao-EN	-	-	-	22.62	50.98	48.98	55.82	Ratio of adequate QAPs over all QAPs being generated: 86.65
Develops a comprehensive system for MCQ generation, including question and answer generation and distractor formulation	SQuAD Quasar RACE CoQA MS MARCO	52.58	36.27	25.15	17.59	49.66	28.03	-	-

Note: compiled by the author based on (Yuan et al., 2022; Lopez et al., 2021; Dijkstra et al., 2024;

Murakhovska et al., 2022; Vachev et al., 2022; Zhang, Zhang, Wang, 2022; Barlybayev, Matkarimov, 2024)

Graph-based models, particularly Graph Convolutional Networks (GCNs), have been explored to incorporate syntactic structure information into neural AQG systems. GCNs excel at capturing dependency relationships between words in a sentence, providing a deeper understanding of sentence structure that aids in generating grammatically and semantically coherent questions (Huang et al., 2021; Fei, Zhang, Zhou, 2021; Su et al., 2020; Dhole, Manning, 2022). Table 4 gives comparative overview of graph-based models.

Hybrid approaches. Hybrid approaches to question generation (QG) combine multiple techniques to leverage the strengths of each, often resulting in more resilient and adaptable systems (Blstak, Rozinajova, 2022; Alshboul, Baksa-Varga, 2024). These approaches typically incorporate rule-based methods that utilize linguistic knowledge with data-driven methods that depend on statistical models and machine learning (Blstak, Rozinajova, 2022; Alshboul, Baksa-Varga, 2024; Panchal, Thakkar, Pillai, Patil, 2021). Table 5 provides a summary of models using the hybrid approach.

Table 4. Summary of graph-based models

Model	Dataset	Automatic evaluation						Human evaluation
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	
MulQG (Multi-Hop Encoding Fusion Network for QG) extends the Seq2Seq QG framework with multi-hop context encoding. Employs GCN on dynamic entity graph to aggregate evidence related to questions	HotpotQA	40.08	26.58	19.61	15.11	35.35	20.24	Fluency: 0.43 Answerability: 0.61 Completeness: 0.33
IGND (Iterative Graph Network-based Decoder) models the structure information in the previous generation at each decode step using a GNN. Captures dependency relations in the passage	SQuAD MS MARCO	50.82 -	34.73 -	25.64 -	20.33 23.87	48.94 -	- -	Fluency: 4.24 Relevancy: 4.33 Answerability: 4.26
EGSS (Entity Guided Question Generation model) uses GCN and Bi-LSTM to capture structure and sequence information from context. Employs an entity-guided approach to obtain question type from the answer	SQuAD	50.86	34.42	25.44	19.45	47.29	23.54	Grammaticality: 4.06 Relevancy: 4.37 Answerability: 4.01

Note: compiled by the author based on (Su et al., 2020; Fei, Zhang, Zhou, 2021; Huang et al., 2021)

Evaluation techniques. Evaluating automatic question generation (AQG) systems is essential for ensuring that the generated questions are high-quality, relevant, and appropriately challenging.

Two main evaluation methods are used: automatic evaluation with computational metrics and human-based evaluation. Automatic evaluation compares generated questions to reference questions, typically human-authored, using metrics such as BLEU (Leite, Cardoso, 2023; Sewunetie, Kovacs, 2024; Blstak, Rozinajova, 2022), ROUGE (Zhang, Zhang, Wang, 2022; Leite, Cardoso, 2023), METEOR (Su et al., 2020; Barlybayev, Matkarimov, 2024; Sewunetie, Kovacs, 2024; Huang et al., 2021), and BERTScore (Murakhovska et al., 2022; Zhang, Zhang, Wang, 2022). These metrics focus on lexical and semantic similarity, providing quick and cost-effective assessments. However, they often fail to fully capture aspects like answerability, relevance, and cognitive demand, making them less reliable indicators of overall question quality.

Human-based evaluation is critical for a more nuanced analysis, as it accounts for qualities that automatic metrics might miss. Human evaluators focus on criteria like grammaticality, fluency, relevance, answerability, clarity, cognitive demand, and distractor quality in multiple-choice questions. Grammaticality ensures syntactic well-formedness, while relevance checks how well questions align with the content (Leite, Cardoso, 2023; Blstak, Rozinajova, 2022; Huang et al., 2021).

Table 5. Summary of hybrid approaches

<table>
<tr><th rowspan="2">Model</th><th rowspan="2">Dataset</th><th colspan="5">Automatic evaluation</th><th rowspan="2">Human evaluation</th></tr>
<tr><th>BLEU-1</th><th>BLEU-2</th><th>BLEU-3</th><th>BLEU-4</th><th>ROUGE-L</th></tr>
<tr><td>Sentence Structure Analysis, Rule Learning, and Evaluation Module</td><td>SQuAD</td><td>0.83</td><td>0.78</td><td>0.74</td><td>0.71</td><td>0.87</td><td>Syntax: 86.5%
Semantics: 76.8%</td></tr>
<tr><td>Semantic code conversion using ontology and an AI-based question generator - QuestGen AI</td><td>-</td><td colspan="5">-</td><td>Average validity of the generated questions is 4,2 out of 5</td></tr>
<tr><td rowspan="2">Machine learning classification-based Fill-in-the-Blank (FIB) generator and a rule-based approach to generate Wh-type questions</td><td rowspan="2">SQuAD</td><td>BLEU</td><td></td><td colspan="2">FIB questions (Total 100)</td><td>Wh-type questions (Total 100)</td><td rowspan="2">Total 200 questions:
Grammatical score: 0.64
Answerability score: 0.49
Difficulty score: 0.41
Context score: 0.77</td></tr>
<tr><td>0.73</td><td>Correct
Incorrect
Success rate</td><td colspan="2">59
41
59%</td><td>49
51
49%</td></tr>
<tr><td colspan="8">Note: compiled by the author based on (Blstak, Rozinajova, 2022; Alshboul, Baksa-Varga, 2024; Panchal, Thakkar, Pillai, Patil, 2021)</td></tr>
</table>

Cognitive demand assesses the intellectual level required, and distractor quality impacts the effectiveness of multiple-choice questions (Leite, Cardoso, 2023; Blstak, Rozinajova, 2022). Various evaluation methods, such as Likert scales (Leite, Cardoso, 2023), binary judgments (Dijkstra et al., 2024), or comparative assessments (Sun et al., 2022), are used to capture these

qualities. Researchers often combine both automatic and human evaluations for a more comprehensive assessment, as each offers unique strengths (Leite, Cardoso, 2023). By tailoring evaluation strategies to the specific goals and characteristics of AQG systems, a more accurate and meaningful measure of question quality can be achieved (Blstak, Rozinajova, 2022).

Conclusion. This survey provides an overview of recent work in the area of Automatic Question Generation (AQG), focusing on the methodologies, techniques, and evaluation techniques. AQG systems are categorized into rule-based, neural network-based, and hybrid approaches, each offering distinct strengths and weaknesses. Rule-based systems are transparent and controllable, but require extensive manual effort. In contrast, neural network-based methods, particularly sequence-to-sequence, transformer-based, and graph-based models, have made significant strides in generating diverse, contextually relevant questions, driven by deep learning techniques. The survey also underscores the importance of robust evaluation techniques, emphasizing the need to balance automatic metrics like BLEU and ROUGE with human assessments to capture nuanced aspects of question quality, such as relevance, answerability, and cognitive demand. Despite the progress made, challenges remain in ensuring the generated questions meet human standards for fluency, naturalness, and semantic accuracy. The limitations of current evaluation metrics further highlight the need for more comprehensive measures of question quality. Looking ahead, AQG research is poised for exciting developments, particularly in multimodal question generation, which will integrate textual data with visual and auditory inputs. Hybrid models, which combine rule-based and neural network techniques, may offer improved performance and adaptability. The continued evolution of AQG holds great potential for transforming applications in education, information retrieval, and human-computer interaction, fostering a deeper and more engaging learning experience.

Conflict of interest. The authors declare that there is no conflict of interest.

Acknowledgements. This research is funded by the authors' own funds.

"Notification of the use of generative AI and technologies using it in the process of writing the manuscript". The authors did not use tools of artificial intelligence services in the preparation of this paper.

References

Alshboul J. and Baksa-Varga E. (2024). A Hybrid Approach for Automatic Question Generation from Program Codes, https://doi.org/10.14569/ijacsa.2024.0150102.

Barlybayev A. and Matkarimov B. (2024). Development of system for generating questions, answers, distractors using transformers. International Journal of Electrical and Computer Engineering, vol. 14, no. 2, pp. 1851–1863, https://doi.org/10.11591/ijece.v14i2.pp1851-1863.

Blstak M. and Rozinajova V. (2022). Automatic question generation based on sentence structure analysis using machine learning approach. Nat Lang Eng, vol. 28, no. 4, pp. 487–517, https://doi.org/10.1017/S1351324921000139.

Dijkstra R., Genc Z., Kayal S., and Kamps J. (2022). Reading Comprehension Quiz Generation using Generative Pre-trained Transformers. Accessed: Dec. 30, 2024. [Online]. Available: https://e.humanities.uva.nl/publications/2022/dijk_read22.pdf

Dhole K. D. and Manning C. D. (2022). Syn-QG: Syntactic and Shallow Semantic Rules for Question Generation, https://doi.org/10.48550/arXiv.2004.08694.

Fei Z., Zhang Q., and Zhou Y. (2021). Iterative GNN-based Decoder for Question Generation, https://doi.org/10.18653/v1/2021.emnlp-main.201.

Huang Q. et al. (2021). Entity Guided Question Generation with Contextual Structure and Sequence Information Capturing, https://doi.org/10.1609/aaai.v35i14.17544.

Leite B. and Cardoso H. L. (2023). Do Rules Still Rule? Comprehensive Evaluation of a Rule-Based Question Generation System. International Conference on Computer Supported Education, CSEDU - Proceedings, Science and Technology Publications, Lda, 2023, pp. 27–38, https://doi.org/

10.5220/0011852100003470.
Lopez L. E., Cruz D. K., Cruz J. C. B., and Cheng C. (2021). Simplifying Paragraph-level Question Generation via Transformer Language Models, https://doi.org/10.48550/arXiv.2005.01107.
Murakhovska L., Wu C.-S., Laban P., Niu T., Liu W., and Xiong C. (2022). MixQG: Neural Question Generation with Mixed Answer Types, https://doi.org/10.48550/arXiv.2110.08175.
Naeiji A., An A., Davoudi H., Delpisheh M., and Alzghool M. (2023). Question Generation Using Sequence-to-Sequence Model with Semantic Role Labels, doi: 10.18653/v1/2023.eacl-main.207.
Panchal P., Thakkar J., Pillai V., and Patil S. (2021). Automatic Question Generation and Evaluation. Journal of University of Shanghai for Science and Technology, vol. 23, no. 05, pp. 751–761, https://doi.org/10.51201/JUSST/21/05203
Sun Y., Liu S., Dan Z., and Zhao X. (2022). Question Generation Based on Grammar Knowledge and Fine-grained Classification. Accessed: Dec. 30, 2024. [Online]. Available: https://aclanthology.org/2022.coling-1.562
Su D., Xu Y., Dai W., Ji Z., Yu T., and Fung P. (2020). Multi-hop Question Generation with Graph Convolutional Network, https://doi.org/10.18653/v1/2020.findings-emnlp.416.
Sewunetie W.T. and Kovacs L. (2024). Automatic question generation using extended dependency parsing. Indonesian Journal of Electrical Engineering and Computer Science, vol. 33, no. 2, pp. 1108–1115, https://doi.org/10.11591/ijeecs.v33.i2.pp1108-1115.
Tran P., Nguyen D. K., T. Tran, and Vo B. (2023). Using Syntax and Shallow Semantic Analysis for Vietnamese Question Generation. KSII Transactions on Internet and Information Systems, vol. 17, no. 10, pp. 2718–2731, https://doi.org/10.3837/tiis.2023.10.007.
Vachev K., Hardalov M., Karadzhov G., Georgiev G., Koychev I., and Nakov P. (2022). Leaf: Multiple-Choice Question Generation, https://doi.org/10.48550/arXiv.2201.09012.
Yuan W., Yin H., He T., Chen T., Wang Q., and Cui L. (2022). Unified Question Generation with Continual Lifelong Learning. Proceedings of the ACM Web Conference 2022, Association for Computing Machinery, Inc, pp. 871–881, https://doi.org/10.1145/3485447.3511930.
Zhang C. and Wang J. (2022). Tag-Set-Sequence Learning for Generating Question-Answer Pairs, https://doi.org/10.48550/arXiv.2210.11608.
Zhang C., Zhang H., and Wang J. (2022). Downstream Transformer Generation of Question-Answer Pairs with Preprocessing and Postprocessing Pipelines, https://doi.org/10.48550/arXiv.2205.07387.

Information about authors

Lopatkin Dmitriy Vitalievich – D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan, lopatkin.d@edu.ektu.kz , ORCID: 0009-0003-8735-5997, +77775264039

Rokhas Kriulko Natalia Pedrovna – D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan, nrohas@edu.ektu.kz , ORCID:0009-0006-6841-0593, +77772462861

Kotlyarova Irina Aleksandrovna – Master of Engineering and Technology, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan, IKotlyarova@edu.ektu.kz, ORCID:0009-0006-5372-4207, +77772549051

Popova Galina Vladimirovna – Candidate of Physical and Mathematical Sciences, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan, gpopova@edu.ektu.kz, ORCID: 0000-0002-6935-1066, +77473371938

Vais Yuriy Andreevich – Candidate of Technical Sciences, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan, YuVais@edu.ektu.kz, ORCID: 0000-0002-2964-8260, +77052502872