ISSN: 2229-7359 Vol. 11 No. 25s,2025

https://theaspd.com/index.php

# Automated Generation Of Multiple-Choice Questions In Kazakh Using Transformer Architecture

Alibek Barlybayev 1\*, Assel Mukanova 2

- <sup>1</sup>School of Information Technology and Engineering, Astana International University, Astana, Kazakhstan. Email: frank-ab@mail.ru
- <sup>2</sup> School of Information Technology and Engineering, Astana International University, Astana, Kazakhstan. Email: asel\_ms@bk.ru

Abstract: This study presents a multi-stage pipeline for automated generation of multiple-choice questions (MCQs) in the Kazakh language, a low-resource, agglutinative language with limited NLP tooling and datasets. We cast question generation as a sequence-to-sequence problem and fine-tune a T5 model on a Kazakh adaptation of SQuAD and a geography-themed SQuAD-style set (85/15 train/validation; 50 epochs). Given a passage and an answer span, the generator produces a candidate question, after which a BERT-based semantic verifier filters incoherent or tautological pairs, improving validation accuracy from 44% (pre-trained) to 78% (fine-tuned). To construct MCQs, we integrate a SpaCy NER module that samples distractors from entities of the same type as the correct answer, increasing plausibility while preserving linguistic coherence. Automatic evaluation yields BLEU-1/2/3/4 of 42.57/25.78/18.46/13.42, METEOR of 17.81, and ROUGE-L of 41.09, indicating good lexical coverage with expected degradation at higher n-grams and adequate retention of key content. Qualitative analysis against GPT-40 suggests our system generally produces contextually relevant questions, with some tendency toward broader prompts in ambiguous contexts. The contribution is an end-to-end, replicable framework, combining transformer-based generation, semantic verification, and entity-aware distractor synthesis, and tailored to Kazakh and extensible to other low-resource educational settings.

Keywords: Kazakh language processing, question generation, multiple-choice assessment, transformers, educational NLP.

#### 1. INTRODUCTION

Manual creation of assessment questions is time-consuming and inconsistent, especially for educators in multilingual or under-resourced educational environments. Kazakh language lacks robust natural language processing (NLP) resources, tools, and annotated datasets, making automated educational content generation particularly difficult. Existing QA systems and multiple-choice questions generation tools are predominantly designed for English or other high-resource languages and are not directly applicable to Kazakh due to its agglutinative morphology, syntax, and semantic nuances. The emergence of transformer-based NLP models like BERT and T5 offers a promising avenue to overcome traditional limitations in rule-based or shallow-learning approaches, enabling more scalable, accurate, and semantically coherent question generation. The integration of such systems into educational technology can facilitate personalized learning, enable scalable assessments, and reduce instructors' workload, thereby contributing to broader digital transformation goals in Kazakh-language education.

In recent years, question generation has emerged as a significant area of research, particularly in the field of education, due to its broad applicability in enhancing learning outcomes. The primary objective of question generation is to automatically produce natural and contextually relevant questions from a given text, thereby facilitating improved comprehension and engagement among students [1]. Test questions play a crucial role in the educational process, serving as a means to evaluate learners' understanding and retention of material [2], [3]. However, the manual creation and assessment of such questions can be time-consuming and labor-intensive for educators [4]. As a result, the automation of question generation and answer evaluation has garnered considerable attention from both researchers and educational practitioners [5], [6].

In academic environments, assessments often include question formats such as multiple-choice questions (MCQs), true/false (T/F) statements, and fill-in-the-blank (FiB) tasks. These formats are widely adopted for their efficiency in measuring knowledge and for enabling objective scoring [7]. The development of automated systems capable of generating such questions can significantly reduce the workload of educators while maintaining assessment quality and consistency.

ISSN: 2229-7359 Vol. 11 No. 25s,2025

https://theaspd.com/index.php

Traditional question generation techniques predominantly rely on heuristic methods to transform descriptive text into corresponding questions. Existing rule-based approaches are generally classified into three main categories: template-based methods [8], syntax-based approaches [9]–[11], and semantic-based techniques [12]–[15]. Regardless of the method employed, question generation typically involves two fundamental stages: context selection and question formulation. These stages are executed by applying either syntactic or semantic parsers to the input text, allowing the system to identify salient topics that can be queried.

Based on the extracted topic, intermediate linguistic representations are generated and subsequently converted into natural language questions. This transformation can be accomplished using either rule-based templates or more dynamic, transformation-based techniques. However, conventional AI-driven approaches often depend heavily on manual feature engineering, which is not only time-consuming but also requires substantial domain-specific expertise. Furthermore, these systems are typically composed of multiple interdependent modules, resulting in limited scalability and reusability, which ultimately hinders their robustness and adaptability across diverse domains.

## 2. MATERIALS AND METHODS

The primary aim of the study is to develop a robust and linguistically-informed automated system for generating high-quality multiple-choice questions in the Kazakh language using transformer-based neural architectures.

The objective of this study is to enhance MCQ systems for the Kazakh language, which poses unique challenges due to its low-resource status, agglutinative grammar structure, and limited availability of high-quality annotated datasets. This section outlines the models, datasets, and methodological framework employed in the research, providing a detailed step-by-step explanation of the fine-tuning procedures and evaluation processes applied to assess model performance on Kazakh-language QA tasks.

#### **Data Set Collection**

Given the scarcity of high-quality datasets in the Kazakh language, this study utilized both publicly available and custom-adapted versions of the Stanford Question Answering Dataset (SQuAD) for training and finetuning the proposed models. The SQuAD dataset is a widely recognized benchmark composed of questions generated by non-experts based on Wikipedia articles, with answers extracted directly from the corresponding text passages [16]. We adapted SQuAD-style QA material and a geography-focused Kazakh QA set into a unified MCQ/QG training corpus. Because SQuAD provides answer \*spans\* tied to character offsets, our pipeline preserved span fidelity during translation and post-editing. The translated dataset encompasses a diverse range of context-rich passages, which is essential for improving the model's ability to perform accurate question-answering tasks in Kazakh. Prior to training, all datasets underwent preprocessing steps to ensure consistent tokenization, normalization, and formatting, thereby facilitating effective model learning and evaluation. We also used an already developed dataset in the SQUAD notation on the topic of geography [17]. Each item (context, answer span, question, and MCQ options) passed through (i) two independent native-Kazakh translators, (ii) one senior linguist for reconciliation/adjudication, and (iii) a QA annotator for final acceptance. Before full production, a 100-item pilot was double-translated and adjudicated to align on: (a) preserving answer boundaries in agglutinative contexts, (b) consistent rendering of proper nouns/toponyms, and (c) register and style of interrogatives. A living style guide covered: script/orthography (Unicode NFC normalization), punctuation, numerals and units, transliteration for out-of-vocabulary entities, and question templates (who/what/where/when/why/how) with examples of acceptable paraphrase breadth.

#### **Training Model**

Neural question generation models can be broadly categorized into several architectural paradigms, including sequence-to-sequence (Seq2Seq) models, pre-trained language models, variational autoencoders, graph-based models, and adversarial networks. Among these, Transformer-based architectures have become the foundation of most state-of-the-art NLP systems due to their superior performance and scalability. In this study, we adopt the Bidirectional Encoder Representations from Transformers (BERT) model [18] as the core of our question generation framework. The proposed approach consists of two primary stages: pre-training and fine-tuning. During the pre-training phase, the BERT model is trained on large-scale unlabeled text corpora by solving self-supervised learning tasks, such as masked language modeling and next sentence

ISSN: 2229-7359 Vol. 11 No. 25s,2025

https://theaspd.com/index.php

prediction, to acquire general-purpose linguistic representations. For the fine-tuning stage, the model is initialized with the pre-trained weights and subsequently optimized on a task-specific, labeled Kazakh-language dataset. Although all tasks begin from the same pre-trained model parameters, each fine-tuning phase involves task-specific updates, resulting in specialized models tailored for the question generation objective. This process enables the model to adapt to the linguistic nuances of the Kazakh language while leveraging the generalization capabilities obtained during pre-training. Within the BERT framework, the modeling process consists of two main stages: pre-training and fine-tuning. The architecture remains largely unchanged between these stages, with the exception of the output layers, which are adapted for specific downstream tasks. The same pre-trained model parameters are used to initialize multiple task-specific models, ensuring knowledge transfer across applications. During fine-tuning, the entire network including all transformer layers is updated based on labeled task-specific data. Input sequences are constructed using BERT's predefined formatting scheme: each example is preceded by a special classification token [CLS], which is used for aggregating the final representation, and segment boundaries such as between questions and answers, are marked with the separator token [SEP]. This structure enables BERT to effectively model relationships between paired inputs and generate context-aware embeddings suited for question generation and answering tasks in the Kazakh language.

#### 3. RESULTS AND DISCUSSION

#### A. Model Training

Following the evaluation of various neural architectures, the T5 (Text-to-Text Transfer Transformer) model developed by Google [19] was selected as the most suitable for the task. The core principle of T5 is the unification of all natural language processing (NLP) tasks under a text-to-text framework, enabling diverse tasks such as summarization, classification, and question generation to be modeled as sequence transformation problems. For instance, summarization involves providing a source text as input and generating a concise summary as output, while sentiment analysis involves producing a sentiment label sequence from a given input. This flexibility makes T5 particularly well-suited for tasks such as question generation, even though it was not originally trained with this specific goal in mind.

In this study, question generation was approached by inputting both the context and the corresponding answer into the model, which then produced a relevant question as output. To implement and train the T5 model, the HuggingFace Transformers library [19] was utilized. This library provides convenient access to pre-trained transformer models and tokenizers, allowing for efficient model initialization and fine-tuning. The T5 base model was loaded along with its tokenizer, and the input data was encoded accordingly. During training, it was essential to ensure that padding tokens in the target sequences were replaced with a value of 100. This step prevents the model from incorporating padding tokens into the loss calculation, thereby improving training efficiency and avoiding artificially low loss values resulting from correct matches with padded positions.

The dataset was partitioned into 85% for training and 15% for validation. The model was fine-tuned for 50 epochs on the Kazakh-language question generation dataset. Upon evaluation, the generated outputs demonstrated syntactic correctness and coherent grammar, indicating that the model effectively learned to produce natural and relevant questions from the given inputs.

#### B. Evaluation of Generated Questions

To ensure the quality and relevance of the generated questions, it was essential to prevent the system from producing outputs that were unrelated to the given context, tautological in nature, or contained the answer within the question. To address this, a secondary evaluation model was introduced, designed to assess and filter the generated question–answer pairs. For this purpose, we employed a pre-trained BERT model [20], which is based on a transformer architecture trained using the masked language modeling (MLM) objective. This objective enables the model to predict masked tokens by leveraging bidirectional context, and capturing both preceding and succeeding textual information, which significantly enhances its capacity for deep language understanding. BERT has transformed traditional language modeling by facilitating the generation of context-aware embeddings, making it particularly suitable for downstream tasks such as question–answer evaluation. To train our evaluator, we utilized outputs from the question generation model, omitting the original context. The training dataset was constructed such that half of the samples included correctly paired

ISSN: 2229-7359 Vol. 11 No. 25s,2025

https://theaspd.com/index.php

questions and answers, while the other half contained deliberately corrupted pairs. Two manipulation strategies were employed to distort the answer: (1) substituting the correct answer with an unrelated answer from another sample within the dataset, and (2) extracting named entities from the question and inserting them into the answer to create semantically incongruent pairs.

Initially, the pre-trained BERT model achieved an accuracy of 44% on the validation set is only marginally above random guessing. However, after fine-tuning, the model's performance improved significantly, reaching an accuracy of 78%, thereby enabling effective filtering of low-quality or irrelevant question–answer pairs. The resulting two-stage system comprises: (1) a question generation model that synthesizes questions from provided answers and context; and (2) a question–answer evaluation model that verifies the semantic correctness of the generated pairs. The pipeline begins by segmenting the source text into individual sentences, which serve as candidate answers. These are then combined with the corresponding context and passed through the question generation model. The resulting question–answer pairs are subsequently evaluated by the BERT-based classifier, which assigns a confidence score indicating the pair's validity. These scores are used to rank the pairs, and the top-N highest-ranked question–answer sets are presented to the user as the final output.

#### C. Distractor Generation

To enhance the functionality of the system, MCQ generation was integrated, enabling rapid assessment creation and simplifying the evaluation process for learners by allowing selection from predefined answer options. However, naive selection of alternative answer choices can lead to trivial or poorly constructed questions that lack alignment with the intended inquiry. Such oversimplified distractors may limit the depth of student engagement and reduce the effectiveness of the learning experience. To address this limitation, a more comprehensive strategy was implemented, leveraging Named Entity Recognition (NER) to generate contextually relevant and semantically plausible distractors.

The NER module, provided by the SpaCy library [21], was utilized to extract named entities from the text passages. These entities served as the basis for constructing both correct answers and distractors. For each identified entity type (e.g., person, organization, location), alternative responses were selected from similar entity classes within the same dataset, ensuring greater contextual relevance and cognitive challenge in the MCQs.

The complete pipeline for question generation, evaluation, and distractor synthesis was divided into four stages:

- Step 1 Dataset Collection: This phase involved compiling a training dataset comprising text passages, corresponding questions, and accurate answers. These pre-generated examples formed the foundation for supervised model training.
- Step 2 Generation of QA Pairs: A T5-based sequence-to-sequence model was fine-tuned on the collected dataset to automatically generate question-answer pairs from input text and answer spans.
- Step 3 Evaluation of QA Pairs: The quality and semantic correctness of the generated question-answer pairs were assessed using a BERT-based classifier. The model was trained to distinguish between valid and invalid pairs, as described in previous sections.
- Step 4 Distractor Generation: SpaCy's NER model was employed to extract named entities from the source text, which were then used to generate distractors. For each valid question-answer pair, distractors were selected from existing entities of the same type, ensuring both linguistic coherence and increased complexity. This multi-step architecture enables the generation of high-quality multiple-choice questions that are contextually accurate, cognitively engaging, and suitable for automated assessment in the Kazakh language.

## D. Analysis of Results

To evaluate the performance of our model, we employed the evaluation package developed by Chen et al. [22], originally designed for the assessment of image captioning systems. This framework incorporates several established metrics, including BLEU-1 through BLEU-4 [23], METEOR [24], and ROUGE-L [25]. The BLEU (Bilingual Evaluation Understudy) metric assesses the average n-gram precision between candidate and reference sentences, while incorporating a brevity penalty to account for excessively short outputs. The BLEU-n variant specifically refers to the use of n-gram sequences (e.g., unigram to 4-gram) to compute co-occurrence statistics and enhance the sensitivity of the evaluation. METEOR (Metric for Evaluation of Translation with Explicit ORdering) evaluates the similarity between generated and reference texts by considering linguistic

ISSN: 2229-7359 Vol. 11 No. 25s,2025

https://theaspd.com/index.php

variations such as synonymy, stemming, and paraphrasing. This enables a more nuanced assessment compared to strict n-gram matching. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence) is employed to measure the recall of generated summaries relative to reference texts. It captures the longest common subsequence (LCS) as an indicator of structural similarity. In this study, we report the results based on ROUGE-L scores, as they provide insight into the coverage of essential information in the generated outputs. The results showed the following values: BLEU-1=42.57; BLEU-2=25.78; BLEU-3=18.46; BLEU-4=13.42; METEOR=17.81; ROUGEL=41.09.

The evaluation of the proposed model using standard automatic metrics demonstrates a promising level of performance in generating contextually appropriate questions and answers. The BLEU scores progressively decrease from BLEU-1 (42.57) to BLEU-4 (13.42), which is consistent with the expected reduction in n-gram precision as the n-gram length increases. The relatively high BLEU-1 and BLEU-2 values indicate that the model effectively captures local lexical patterns and key content words. However, the lower BLEU-3 and BLEU-4 values suggest some limitations in maintaining longer contextual dependencies and complex phrase structures, which is a common challenge in natural language generation tasks. The METEOR score (17.81) further highlights this trend by reflecting a moderate ability to account for semantic and lexical variations, such as synonymy and paraphrasing. This suggests that while the model is capable of generating semantically relevant content, there is still room for improvement in capturing diverse linguistic expressions and maintaining semantic fidelity across various contexts. The ROUGE-L score (41.09) demonstrates that the model achieves adequate coverage of the essential information present in the reference sentences. This metric, which emphasizes the longest common subsequence, implies that the generated outputs often align structurally with reference questions and answers, thereby ensuring that key information is preserved.

To evaluate the quality of the generated output, we performed a qualitative analysis by examining sample questions produced by our system and comparing them with those generated by GPT-40. The assessment focused on the semantic alignment of the generated questions with the original input text. Table 1 presents representative examples of questions generated by both models.

Table 1. Comparison of questions and answers generated by our model and GPT-40 model.

Table I. Comparison of Questions and Answers Generated by Proposed Model and GPT-40 Model

Sentence	Gpt-40	Our model
Türkïyanıñ Ewropadağı Şığıs Frakïya nemese Rwmesley dep atalatın böligi bïiktigi 1000 m-dey üstirtti töbeli bolıp keledi. Azïya böligin (Kişi azïya nemese Anadolı) tügeldey Kişi azïya tawlı qıratı alıp jatır.	Question 1: Türkïyanıñ Ewropadağı böliginiñ jer bederi qanday sïpatta? Answer 1: Türkïyanıñ Şığıs böligi bïiktigi şamamen 1000 metr bolatın üstirtti töbeli jer bederimen sïpattaladı.	Question 1: Türkïyanıñ Şığıs Frakïya nemese Rwmelïya dep atalatın Ewropadağı böligi qanday jer bederimen erekşelenedi? Answer 1: Bïiktigi şamamen 1000 metr bolatın üstirtti töbeli jer
Ewropanıñ köpşilik jeriniñ klïmatı qoñırjay. Atmosferalıq jalpı aynalımınıñ qalıptaswına Atlant muxïtınan keletin Golfstrïm jılı ağısınıñ Soltüstik Atlant jılı tarmağı äser etedi.	Question 2: Ewropa qurlığınıñ köpşilik böliginde qanday klimat türi basım? Answer 2: Ewropanıñ köpşilik jerinde qoñırjay klimat basım boladı.	Question 2: Atmosferalıq jalpı aynalımnıñ qalıptaswına qanday tabiği faktor äser etedi? Answer 2: Golfstrim jılı ağısınıñ Soltüstik Atlant jılı tarmağı äser etedi.
Azïyanıñ oñtüstik, oñtüstik-şığıs şetki aymaqtarı öte ılğaldı keledi. Gïmalaydıñ oñtüstik etegindegi Çerrapwndjï öñirinde jıldıq jawın-şaşınnıñ ortaşa mölşeri 12665 mm (Jer şarındağı eñ ılğaldı awdan).	Question 3: Jer şarındağı eñ ılğaldı awdan qay jerde ornalasqan jäne onıñ jıldıq jawın-şaşın mölşeri qanşa? Answer 3: Jer şarındağı eñ ılğaldı awdan — Gïmalaydıñ oñtüstik etegindegi Çerrapwndji öñiri, munda jıldıq jawın-şaşınnıñ ortaşa mölşeri 12665 mm quraydı.	Question 3: Azïyanıñ qay aymaqtarı öte ılğaldı bolıp keledi? Answer 3: Oñtüstik jäne oñtüstik-şığıs şetki aymaqtarı.

ISSN: 2229-7359 Vol. 11 No. 25s,2025

https://theaspd.com/index.php

Sentence	Gpt-40	Our model
Qola däwirinde Sibirdiñ, Jayıq öñiriniñ, Qazaqstan men Orta Azïyanıñ ulan-ğayır dalaların tegi jäne tarïxï tağdırlarınıñ ortaqtığı jağınan twis taypalar mekendedi.	Question 4: Qola däwirinde twis taypalar qanday aymaqtardı qonistandı? Answer 4: Olar Sibirdiñ, Jayiq öñiriniñ, Qazaqstan men Orta Aziyanıñ ulan-ğayır dalaların qonistandı.	Question 4: Qola däwirinde qanday aymaqtardı twis taypalar mekendegen? Answer 4: Qola däwirinde Sibirdiñ, Jayıq öñiriniñ, Qazaqstan men Orta Azïyanıñ ulan-ğayır dalaların twis taypalar mekendegen.

The qualitative analysis, as illustrated in Table 1, provides additional insight into the model's strengths and limitations. Compared to GPT-40, the proposed system tends to generate questions that are more general in scope and sometimes less specific to the input context. For instance, in Sentence 2, the GPT-40 output focuses on identifying the dominant climate type in Europe, whereas the proposed model generates a broader question regarding the factors influencing atmospheric circulation. Similarly, in Sentence 3, the GPT-40 output directly identifies the world's wettest region and its rainfall statistics, while the proposed model's question is less detailed, targeting the general identification of humid regions. These patterns suggest that the proposed model is adept at capturing the overall thematic content of the input but may struggle to consistently generate fine-grained or highly specific questions.

The analysis also reveals several instances where the proposed model performs comparably to GPT-40 in terms of semantic alignment and relevance. For example, in Sentences 1 and 4, the generated questions and answers are closely aligned with the input text and retain a high level of contextual relevance. This indicates that the model is capable of producing accurate and contextually appropriate outputs when the source text is more descriptive or less ambiguous.

This work demonstrates that automatically generated Kazakh MCQ/QA items, that produced via a T5 generator, filtered with a semantic verifier, and augmented with NER-consistent distractors, and can be integrated into existing educational infrastructure with minimal friction. By targeting open standards (QTI/GIFT for exchange; LTI for tool embedding) and exposing a small, stable API surface, the system aligns with national and university platforms (e.g., Moodle, Platonus, SmartENU) without bespoke engineering. In practical terms, this lowers the cost of item creation, broadens topical coverage, and enables rapid iteration: teachers can transform textbook passages into validated items during lesson planning or even in class, while assessment teams can route machine-drafted items through familiar review queues.

#### 4. CONCLUSIONS

This study presents a comprehensive comparison of the proposed Pre-trained T5 + BERT + NER system against various baseline and conventional models for automated question generation. The experimental results demonstrate that the integrated model, which incorporates sentence-level contextual information and entity-based distractor generation, consistently outperforms alternative approaches across multiple evaluation metrics. While traditional Seq2Seq and M2S+cp models showed competitive performance in terms of sentence-question alignment, they were limited in their ability to capture deeper contextual dependencies and semantic nuances. In contrast, the proposed model, which leverages pre-trained embeddings and a multistage architecture, exhibited superior performance in generating relevant, context-aware, and grammatically coherent questions. Although the inclusion of paragraph-level encoding introduced a slight decline in certain metrics, it contributed to a reduction in irrelevant information and helped preserve essential contextual cues. Notably, the Pre-trained T5 + BERT + NER model demonstrated its effectiveness by producing more precise and detail-rich questions, confirming the value of integrating transformer-based generation, contextual evaluation, and named entity recognition for improving question quality in low-resource languages such as Kazakh.

The scientific novelty of this study lies in the development of a comprehensive, multi-stage system for automated MCQ generation in the Kazakh language, which is considered a low-resource language. Unlike most prior research that primarily targets high-resource languages such as English, this work:

ISSN: 2229-7359 Vol. 11 No. 25s,2025

https://theaspd.com/index.php

Adapts state-of-the-art transformer architectures (T5 and BERT) for the generation and evaluation of QA pairs specifically in Kazakh.

- Integrates NER into the MCQ generation pipeline for context-aware distractor synthesis, thereby improving the cognitive complexity and plausibility of the answer options.
- Implements a two-stage validation process that includes a BERT-based semantic evaluator to filter irrelevant or incoherent question-answer pairs, an approach rarely applied in Kazakh NLP systems.
- Builds and utilizes a translated and domain-adapted version of the SQuAD dataset for Kazakh, addressing the critical challenge of dataset scarcity in this language.
- Presents a complete end-to-end architecture that includes data preparation, QA generation, quality evaluation, and distractor generation, is offering a replicable framework for low-resource educational NLP tasks.

Future work should therefore (i) run controlled studies on learning outcomes and time-to-author reductions, (ii) expand genre coverage (procedural texts, graphs, multimedia), (iii) couple MCQs with short-answer and rubric-scored items for higher-order reasoning, and (iv) refine adaptive sequencing using calibrated item banks rather than heuristic bands. National repositories could host vetted "gold" sets and reference style guides, while community review days and quarterly refresh cycles keep banks current and reduce drift.

## Acknowledgement:

This research is funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP19577922).

#### REFERENCES

- M. Heilman and N. A. Smith, "Good question! Statistical ranking for question generation," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 609-617, 2010.
- G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," International Journal of Artificial Intelligence in Education, vol. 30, pp. 121-204, 2020.
- R. Zhang, J. Guo, L. Chen, Y. Fan, and X. Cheng, "A review on question generation from natural language text," ACM Transactions on Information Systems (TOIS), vol. 40, no. 1, pp. 1-43, 2021.
- R. Ch and S. K. Saha, "Automatic multiple-choice question generation from text: A survey," IEEE Transactions on Learning Technologies, vol. 13, no. 1, pp. 14-25, 2018.
- O. Rodríguez Rocha and C. Faron Zucker, "Automatic generation of quizzes from DBpedia according to educational standards," in Companion Proceedings of the The Web Conference 2018, pp. 1035-1041, 2018.
- M. Divate and A. Salgaonkar, "Automatic question generation approaches and evaluation techniques," Current Science, pp. 1683-1691, 2017.
- B. Das, M. Majumder, S. Phadikar, and A. A. Sekh, "Automatic generation of fill-in-the-blank question with corpus-based distractors for e-assessment to enhance learning," Computer Applications in Engineering Education, vol. 27, no. 6, pp. 1485-1495, 2019.
- J. Mostow and W. Chen, "Generating Instruction Automatically for the Reading Strategy of Self-Questioning," in AIED, pp. 465-472, 2009.
- Varga and L. A. Ha, "WLV: a question generation system for the QGSTEC 2010 Task B," in Proceedings of QG2010: The third workshop on question generation, pp. 80-83, 2010.
- S. Kalady, A. Elikkottil, and R. Das, "Natural language question generation using syntax and keywords," in Proceedings of QG2010: The Third Workshop on Question Generation, vol. 2, pp. 5-14, 2010.
- H. Ali, Y. Chali, and S. A. Hasan, "Automatic question generation from sentences," in Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts, pp. 213-218, 2010.
- P. Mannem, R. Prasad, and A. Joshi, "Question generation from paragraphs at UPenn: QGSTEC system description," in Proceedings of QG2010: The Third Workshop on Question Generation, pp. 84-91, 2010.
- Y. Huang and L. He, "Automatic generation of short answer questions for reading comprehension assessment," Natural Language Engineering, vol. 22, no. 3, pp. 457-489, 2016.
- X. Yao and Y. Zhang, "Question generation with minimal recursion semantics," in Proceedings of QG2010: The Third Workshop
  on Question Generation, pp. 68-75, 2010.
- Copestake, D. Flickinger, C. Pollard, and I. A. Sag, "Minimal recursion semantics: An introduction," Research on language and computation, vol. 3, pp. 281-332, 2005.
- T. Parshakova, F. Rameau, A. Serdega, I. S. Kweon, and D. S. Kim, "Latent question interpretation through variational adaptation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1713-1724, 2019.
- Mukanova, A. Barlybayev, A. Nazyrova, L. Kussepova, B. Matkarimov, and G. Abdikalyk, "Development of a Geographical Question-Answering System in the Kazakh Language," IEEE Access, vol. 12, pp. 105460-105469, 2024.

ISSN: 2229-7359 Vol. 11 No. 25s,2025

https://theaspd.com/index.php

 D. M. Gates, "How to generate cloze questions from definitions: A syntactic approach," in 2011 AAAI Fall Symposium Series, 2011

- L. Qin, A. Gupta, S. Upadhyay, L. He, Y. Choi, and M. Faruqui, "TIMEDIAL: Temporal commonsense reasoning in dialog," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, vol. 1, pp. 7066-7076, 2021.
- K. Lo, "YiSi-a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources," in Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 507-513, 2019.
- Kleinberg, M. Mozes, A. Arntz, and B. Verschuere, "Using named entities for computer-automated verbal deception detection," Journal of forensic sciences, vol. 63, no. 3, pp. 714-723, 2018.
- X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," arxiv.org/abs/ 1504.00325v2, Apr. 2015.
- K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 2002-July, pp. 311-318, 2002.
- M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in Proceedings
  of the Annual Meeting of the Association for Computational Linguistics, pp. 376-380, 2014.
- C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Text summarization branches out, pp. 74-81, 2004.