مجلــة العلــوم
الهندسية والتقنية
مجلة علمية محكمة - تصدر عن كلية الهندسة وكلية الحاسبات والمعلوماتية – جامعة ذمار

A Literature Review on Arabic Automatic Question Generation

Abdulkhaleq Amin Abdullah[1,2], Khaled A. Al-Soufi[2]

[1]Information Technology, Faculty of Engineering and Information Technology
[2]Qalam University, Yemen
a7q2014@gmail.com, kalsoufi@gmail.com

Abstract:

This comprehensive literature review is dedicated to the field of Arabic Automatic Question Generation (AQG), which focuses on the development of computational models and algorithms for the automatic generation of questions. The review systematically covers key concepts in AQG, including question types and evaluation metrics. Additionally, it delves into the specific challenges associated with applying AQG techniques to the Arabic language, considering factors like complex morphology and dialectal variations. The review introduces a taxonomy of Arabic AQG approaches, classifying them into rule-based, template-based, and machine learning-based methods. It examines the pivotal role of datasets, resources, and evaluation methodologies in the training and assessment of AQG systems. Advancements in Arabic AQG are highlighted, and the review identifies emerging research directions, such as domain-specific question generation and integration into educational platforms.

In conclusion, the review provides valuable insights for researchers, developers, and educators interested in Arabic AQG. It addresses current advancements, challenges, and outlines potential future research directions, including the scarcity of labeled data and the necessity for domain-specific approaches. Overall, this review serves as a comprehensive resource in the realm of Arabic AQG.

Keywords: Question generation, Arabic language, Artificial intelligence, Contexts.

1. Introduction

Arabic Automatic Question Generation (AAQG) is an emerging field of research that focuses on developing systems and techniques for automatically generating questions in the Arabic language. This field is of great importance in various educational applications as it can be used to create language learning materials, assess comprehension skills, and enhance the overall learning experience for Arabic learners.

Automatic Question Generation is of great significance for the Arabic language for several reasons. Firstly, the Arabic language is one of the most widely spoken languages globally, with millions of Arabic speakers around the world. Secondly, the Arabic language has a rich literary and intellectual heritage, making it an important language for educational purposes. Furthermore, the increasing availability of Arabic educational content on the internet necessitates the development of efficient and effective methods for generating relevant and accurate questions to enhance learning outcomes for Arabic-speaking students [1].

Question answering (QA) can benefit from QG, where the training dataset of QA can be enriched using QG to improve the learning and performance of QA algorithms [2]. Additionally, in the future, QG can be greatly helpful to the education industry by generating teaching materials with less effort and motivating students' intention to read.

Arabic question-generation systems have been limited and are mainly focused on factoid questions and short-answer generation [3]. Current approaches rely heavily on rule-based methods or manual construction of question styles using specific text sources, which lack scalability and generalizability [4]. The complexity of the Arabic language presents challenges in morphological analysis, syntactic structures, and semantic nuances (Ahmed, 2025, Alharbi, 2022)[5][6]. Addressing these challenges requires the creation of large and diverse Arabic question-generation datasets, as well as the exploration of advanced Natural Language Processing techniques specifically tailored for the Arabic language [5]. Deep learning models have demonstrated their effectiveness in overcoming these obstacles by leveraging extensive Arabic datasets and educational resources to comprehend the intricacies of the language (Al-Ghobesi, 2025)[7].

To develop an effective Automatic Question Generation system for the Arabic language, it is imperative to adopt strategies that leverage recent advancements in deep learning techniques, particularly those that have been successful in the development of AQG systems for other languages [8]. Recent research has demonstrated the efficacy of Transformer-based models and recurrent neural networks in generating questions from textual sources [5][9][10]. These models have exhibited a high degree of accuracy when applied to languages with complex syntactic and semantic structures, making them viable candidates for developing Arabic question-generation systems [11].

The field of educational question generation has made significant advancements, particularly with the integration of deep learning techniques [12]. This attempts to create questions from a text paragraph, where certain sub-spans of the passage in question will answer the questions produced. Artificial intelligence (AI) techniques are now extensively used to minimize the manual effort required by teachers, and research progress has been made in Automatic Question Generation (AQG) from textual data based on syntax and semantics [13].

The integration of question-generation technology in education has the potential to greatly enhance the learning experience by automatically generating thought-provoking questions [14]. This advancement relies on deep learning techniques, which have shown promise in capturing complex patterns and semantic connections within Arabic QG systems used in educational contexts.

The review has made significant contributions, which are listed below:

-The current state of research on generating Arabic questions.

-What existing methods for creating Arabic questions could be adjusted to fulfill the specific demands of education.

-Strategies for future research to address deficiencies and fill gaps in educational Arabic question generation.

This paper is organized as follows: We first formally represent the problem of Arabic automatic question generation and discuss the various question categories, providing a technical overview of such a system. In section 2, an overview of previous works on Arabic question generation is provided. Section 3 presents the methodologies and techniques employed in Arabic question generation using deep learning, including Transformer-based models and recurrent neural networks. In section 4, we discuss the performance evaluation metrics used to assess the quality of generated questions and compare different approaches. Section 5 presented the challenges and limitations faced in Arabic question generation, such as the lack of large-scale annotated datasets and the complexity of Arabic language structures. Future works are discussed in section 6. Finally, the conclusion is presented in section 7.

1.1. ARABIC LANGUAGE Challenges

Arabic is a Semitic language known for its rich morphology, complex grammatical structures, and varied dialects. It is spoken by hundreds of millions of people worldwide and is the official language of over 20 countries [15]. Although Arabic is popular, work on the Arabic language is still limited, especially in QG [16]. To understand the challenges of the Arabic language, the basics of Arabic should be understood.

Given the complexity of the Arabic language, it presents unique challenges in the field of Natural Language

Processing. One of the notable challenges is the shortage of large datasets and limited tools for tasks such as sentiment analysis and question generation [17]. Although deep learning techniques have shown promising results in addressing these challenges for other languages, there is a lack of research specifically focusing on deep learning methods for the Arabic language. The intricate morphology and structural complexity of Arabic create obstacles that necessitate the adaptation and customization of existing deep learning models to meet the specific requirements of Arabic NLP [18].

1.2. Classification of question generation:

The question generation problem involves finding a model that approximates the generated question. The dataset is pre-processed to make the data available in the desired format, and an appropriate strategy is employed based on the question type for generating questions [19]. Depending on the type of system being used, either text or image datasets can be selected [20]. Categorization aids in outlining specific use cases, and the following is a classification of questions based on existing research in question generation.

1.2.1. Factual questions:

Factual questions aim to elicit specific information or facts. These questions typically start with words like "من who," "ماذا/ما what," "أين where," and "متى when." Examples of factual questions in an educational context could be: "ماهي عاصمة اليمن؟ What is the capital of Yemen?"

1.2.2. Inferential questions:

Inferential questions require the reader to analyze information and draw conclusions based on the given context. These questions often start with words like "لِمَاذا why," "كيف how," or phrases like "بماذا تستنتج؟ What can you infer from this?

1.2.3. Multiple sentences spanning questions:

Some questions may necessitate answers consisting of multiple sentences from a paragraph, as the relevant information is distributed across several sentences. These types of questions are typically W4H (What/Where/When/Why/How) and can be addressed using methods similar to those employed for answering factual questions.

1.2.4. Yes/no type questions:

Yes/no type questions aim to elicit a binary response of "yes" or "no." Examples of yes/no type questions in an educational context could be: "هل الشمس تدور حول الأرض ؟ Does the sun revolve around the Earth?".

2. Literature Review

The recent work on Arabic QA has focused on leveraging the advancements made in English QA, particularly the utilization of deep learning models such as BERT and recurrent neural networks [21]. These

models have shown promise in improving the performance of Arabic QG systems by capturing contextual information and semantic relationships [22]. These models have shown promising results in generating questions from Arabic healthcare texts based on word embeddings. However, there is still a lack of research on Arabic QG specifically for educational purposes.

In addition to applying advancements from English QA to Arabic QA, recent studies have also demonstrated the effectiveness of deep learning in the domain of Arabic text classification. Techniques such as word embeddings and deep learning models have shown promising classification accuracy, highlighting their potential for use in educational applications. Furthermore, the integration of deep learning in Arabic QG can significantly enhance the automatic generation of thought-provoking questions by capturing complex patterns and semantic connections within the educational context [23].

The literature on Arabic Question Generation using Deep Learning in educational contexts reveals a comprehensive landscape of research endeavors focused on advancing the integration of DL techniques for QG in Arabic language. Numerous studies have explored the utilization of recurrent neural networks, convolutional neural networks, and transformer models in the domain of Arabic QG, emphasizing the potential of these DL architectures in addressing the complexities inherent to Arabic language.

A study proposes a method to analyze the question and retrieve the passage answer in Arabic language, along with experimenting with the generation of a logical representation from the declarative form of each question [11]. This includes extensive evaluation of various Arabic language tasks such as Sentiment Analysis, Named Entity Recognition, and QA using different datasets. Additionally, it discusses advancements in word vector representations for Arabic compared to English models like Word2Vec and highlights initiatives such as AraVec which provides powerful word embedding models developed specifically for Arabic NLP through pre-trained word representations sourced from different domains like Wikipedia, World Wide Web pages, and tweets. Furthermore, it describes real-time embedding schemes that aim at improving performance in various NLP tasks without relying on tabulated pre-embedding or pre-trained transfer learning models [24].

In 2015, [26] introduced rule-based methods that utilized rule-based knowledge sources to assess the comprehension of crucial domain rules. Our interpretation aligns rule-based approaches with semantic-based methods due to their requirement for a deeper understanding beyond mere syntax. Regarding the fifth category, schemas are similar to templates but more abstract in nature. They group templates that represent variations of the same problem. However, we find the differentiation between template and schemas to be unclear.

In 2010, [27] developed a method for creating definitional queries, which are inquiries that seek the meaning of a chosen term referred to as an "up-key". The identification of the "up-key" can be achieved through statistical approaches such as term frequency. Once the named entity of the "up-key" is identified, a question word can be selected. This "up-key" and question word are then inserted into the template "<Question word> is <up-key>?" to generate a definitional query.

In 2010, [28] proposed an approach to automatically generate questions to support students in writing literature reviews. The approach generates questions e.g. What is the research question framed by X? The approach automatically extracts and classifies citations from a student's review. The approach uses templates to generate questions based on the extracted information.

In 2018, [8] presented a Corpus-Based Approach that utilizes a large collection of texts, known as a corpus, to automatically generate questions. This approach involves analyzing the corpus to identify patterns and generate questions based on the content found in the texts. The generated questions are based on patterns and structures observed in the corpus. The corpus-based approach for automatic question generation relies on analyzing a large collection of texts to identify patterns and generate questions. These questions are generated based on patterns and structures observed in the analyzed corpus, allowing for a more data-driven approach to question generation.

In 2014, Google introduced a text-generation task defined as a sequence-by-sequence task. This approach facilitates end-to-end learning for tasks that involve input and output sequences of tokens. Encoder-decoder models are frequently applied to these types of problems and typically include an encoder to process the input sequence through layers of recurrent neural networks, along with a decoder designed to generate the output sequence.

The research by [12] focused on creating queries from a text paragraph for machine-reading comprehension using an attention-driven mechanism based on a sequence-to-sequence model and an RNN-based encoder-decoder system with two distinct encoders. These approaches demonstrate the use of templates and corpus-based analysis, as well as the application of sequence-to-sequence models with attention mechanisms for automatic question generation from text.

In 2019, [35] employed transformers to generate questions based on passages from the SQuAD dataset. They used word error rate as a metric to compare the generated questions with the target questions. The authors observed the syntactic correctness of the generated questions and their relevance to the passage. Furthermore, they found that WER was low for shorter questions but increased for longer ones.

In 2022, [5] improved the transformer model for generating questions on the mMARCO dataset and

optimized it using the AdamW optimizer model. They proposed an end-to-end Arabic automatic question generation model based on the Transformer architecture to generate N interrogative questions for educational content from a single unlimited-length document. This constitutes a transfer-learning process. The authors also incorporated TextRank and Sentence Extraction techniques in their model to improve the quality of the generated questions.

3. Categories of Arabic question generation Approaches

Based on the sources reviewed, the categories of Arabic question generation approaches can be classified into the following:

3.1. Rule-Based Approach

A method based on predefined rules is employed to generate questions from declarative sentences [25]. This method simplifies the sentence, applies a transformation technique for question generation, ranks the generated questions using logistic regression to assess their quality, and then labels them for approval. For example, the rule-based approach may involve transforming a declarative sentence like " أحمد ذهب إلى السوق" into a question like "إلى أين ذهب أحمد؟".

3.2. Template-Based Approach

A method based on templates entails the creation of pre-established question templates, which are then populated with pertinent information derived from the input text. These templates act as a framework for generating questions and ensure that the resulting questions adhere to proper grammar while remaining relevant to the content at hand (Ushio et al., 2023).

3.3. Corpus-Based Approach

The Corpus-Based Approach utilizes a framework for automatic question generation that involves analyzing a large corpus of text. This approach extracts patterns and linguistic features from the corpus to generate questions [51].

3.4. Sequence-to-Sequence Approach

One method for generating questions is the sequence-to-sequence approach, where a neural network model is trained to generate questions by treating it as a sequence-generation task. The model takes an input sequence (such as a sentence or passage) and produces an output sequence (the question). This output sequence is trained to match human-generated questions, allowing the model to learn how to generate questions based on the given input [51].

3.5. Transformer-based Approach

A transformer-based approach [29] involves utilizing the Transformer architecture, a popular deep-learning model, for automatic question generation. This approach relies on the use of transformer models, such as the Transformer architecture, to automatically generate questions by analyzing the syntactic and semantic structures of sentences [5].

These different categories of automatic question generation approaches provide a range of methods for generating questions from text, including rule-based approaches, template-based approaches, corpus-based approaches, sequence-to-sequence approaches, and transformer-based approaches.

Furthermore, the use of multilingual and cross-lingual models allows for the transfer of question-generation techniques and models from other languages to Arabic [49]. Overall, the existing research on Arabic question generation using deep learning techniques demonstrates various approaches and models, including template-based methods, sequence-to-sequence models with attention mechanisms, and transformer-based approaches different QG approaches used for Arabic question generation.

Table 1 provides a comparison summary of different approaches for Arabic question generation, including rule-based approaches, template-based approaches, corpus-based approaches, sequence-to-sequence approaches, and transformer-based approaches.

Table 2 presents an analysis of papers that discussed Arabic QG from multiple perspectives.

Table 1 Summary of Arabic QG Approaches

Approach	Description	Advantages	Limitations
Rule-based approaches	Arabic Question generation methods have traditionally relied on pre-established rules and syntactic transformations.	These methods are simple to execute and can produce grammatically accurate questions.	These methodologies heavily depend on manually created rules and templates, which might not encompass all linguistic variations in Arabic.
Template-based approaches	These methods utilize predetermined templates to produce questions by substituting specific placeholders with pertinent details from the given text.	One approach that is relatively simple to implement and can generate questions with correct syntax is the use of templates.	The ability of these methods to handle complex or diverse sentence structures may be limited, and the quality of the produced questions relies heavily on the quality of the templates used.
Corpus-based approaches	These methods leverage a vast collection of pre-existing questions and their corresponding answers to generate novel questions.	Corpus-driven methods can capture a wide range of question patterns and enhance the diversity of generated questions.	Obtaining a sufficient and diverse question-answer corpus for Arabic may pose challenges for these approaches.

Approach	Description	Advantages	Limitations
Sequence-to-sequence approaches	These methods utilize neural networks to generate questions from the input text by considering the process of generation as a prediction task based on sequences.	Neural sequence-to-sequence methods have been found to produce a wider range of questions that are more relevant in context, compared to traditional rule-based or template-based approaches.	These methods might necessitate a substantial amount of training data and could be computationally demanding.
Transformer-based approaches	These methods make use of transformer models, like the Transformer architecture, to generate questions by capturing the connections between input and output sequences.	Recent advancements in natural language processing tasks have demonstrated the effectiveness of Transformer-based approaches, particularly for question generation.	Optimal performance of transformer-based methods may necessitate a substantial amount of computational resources and training data.

Table 2 Presents an analysis of papers that discussed Arabic QG:

Ref.	*Approach*	*Algorithm*	*Scientific additions*
[5]	Transformer-based approaches	The **TextRank** algorithm operates by extracting key sentences from paragraphs and organizing them into a list.	The system creates questions without answers and functions by segmenting the document into paragraphs while identifying key sentences.
[1]	Template-based approaches	It seamlessly incorporates models for semantic role labeling and question generation.	A novel method for generating questions from Arabic text involves the fusion of semantic role labeling (SRL) for capturing linguistic relationships and the flexibility inherent in question models.
[22]	Transformer-based approaches	ARAGPT2 is a stacked transformer-decoder model that has undergone training using the causal language modeling objective.	ARAGPT2-MEGA, with its impressive 1.46 billion parameters, stands as the most extensive Arabic language model currently accessible. This model, AraGPT2-mega, demonstrates remarkable proficiency in generating news articles that pose a challenge in distinguishing them from those crafted by human authors.

Ref.	*Approach*	*Algorithm*	*Scientific additions*
[30]	Transformer-based approaches	Fine-tuning AraT5 model on ARGENQG dataset.	The introduction involves three potent variants of the T5 model specifically designed for Arabic language generation. Furthermore, a groundbreaking proposition is made for a novel benchmark in Arabic natural language generation, denoted as ARGEN, comprising seven distinct tasks.
[9]	Transformer-based approaches	Fine-tuning GPT-2 architecture	The training of GPT-2 for the purpose of generating Arabic poems has resulted in superior performance, surpassing existing models in the realm of Arabic poetry generation.
[31]	Transformer-based approaches	The authors put forward a multilingual data-driven approach for the generation of reading comprehension questions, employing dependency trees in the process.	The assumed position for the question word in Arabic is the first processed word, essentially representing the end of the sentence in Arabic sentence structure.
[34]	Visual Arabic Question Answering (VAQA)	The envisioned system comprises five key modules: visual feature extraction, question preprocessing, textual feature extraction, feature fusion, and answer prediction.	The VQA task is conceptualized as a binary classification problem.

4. Datasets

The development and assessment of natural language processing models rely on the availability and quality of Arabic question generation datasets. However, obtaining such datasets for Arabic language poses challenges due to factors such as limited resources, the absence of standardized datasets, and the intricate linguistic nature of Arabic. This scarcity poses a notable challenge for researchers and developers aiming to enhance Arabic natural language processing models. Addressing this issue by creating extensive Arabic question-generation datasets is essential [1].

To address the scarcity of Arabic question generation datasets, recent translation techniques have been applied to crowd-sourced annotated datasets. This approach has produced reasonable results on training data for different languages, enabling researchers to overcome the lack of datasets by translating existing ones into Arabic [1]. For instance, the SQuAD dataset has been translated into Arabic, creating a valuable resource for Arabic question generation [38].

Although the availability of specific Arabic question generation datasets may still be limited compared to English, several datasets have been developed for this purpose [39]. These datasets, such as the translated SQuAD dataset and the Arabic Question Answering Web Corpus, provide valuable resources for training and evaluating Arabic question generation models, enabling the development of more accurate and robust systems for generating questions in Arabic.

THE DETAILS OF THE DATASETS USED FOR QUESTION GENERATION ARE SUMMARIZED IN TABLE 3.

Table 3: *Summary of Arabic QG datasets*

DATASET NAME	**Source**	**Content**
ARABIC SQUAD: Arabic Stanford Question Answering Dataset [38].	Translated from English	Translated into Arabic, providing a valuable resource for Arabic question generation.
ARCD: Arabic Reading Comprehension Dataset [38].	Custom	Crowdsourced Arabic questions based on the CNN and Daily Mail datasets in English.
MLQA: Multilingual Question Answering [40].	Multilingual	Parallel sentences in 7 languages, including Arabic, for machine translation evaluation.
XQUAD: Cross-Lingual Question Answering Dataset [41].	Multilingual	Multilingual dataset for question-answering tasks, including Arabic.
TYDI QA: Typologically Diverse Question Answering [41].	Multilingual	The data set is a question-answering benchmark based on Wikipedia articles for 11 typologically diverse languages. Eight of these languages have available UD treebanks and trained dependency parsers in the Stanza package
LAREQA: Low-Resource Cross-Language Question Answering [42].	Custom	A dataset for cross-lingual question answering in low-resource languages, including Arabic.
DAWQAS: Dataset For Web-Based Question Answering In Arabic Script [43].	Custom	A dataset for Arabic web-based question answering.
EXAMS [44].	Custom	A dataset for evaluating extractive question-answering models, including Arabic passages.
MS MARCO [5].	Custom	The MMARCO dataset is a machine-translated multilingual version of the MS MARCO passage ranking dataset covering 13 typologically diverse languages [45].
QUIZITO'S Platform [1].	Custom	The collection comprises 40,435 questions gathered from kids' books and summaries manually curated by Quizito.
Image-Question-Answer (IQA) Triplet [43].	VAQA	The VAQA dataset contains 5000 images. The first Visual Arabic Question Answering (VAQA) dataset is generated. first dataset and system for VQA in Arabic.
ARGENQG [30]	Custom	The researcher built ARGENQG by extracting 96K (passage, answer, and question) triplets from the ARCD dataset, and three multi-lingual QA datasets: MLQA, XQuAD, and TyDi QA.

4. Evaluation Metrics

To evaluate Arabic question generation, several metrics can be used to assess the quality and effectiveness of the generated questions. Some commonly used metrics for automatic evaluation include BLEU-4, METEOR, and ROUGE-L. These metrics compare the generated questions with reference questions to measure the similarity and overlap in terms of n-grams and recall scores. Apart from these automatic metrics, human evaluation can also be conducted to gather subjective feedback on the quality and comprehensibility of the generated questions [3]. The commonly used similarity metrics for evaluating Arabic question generation are as follows:

4.1. BLEU Score:

This metric measures the similarity between the generated questions and reference questions based on n-gram overlap. BLEU-4 is a commonly used automatic evaluation metric in question generation tasks. It measures the n-gram similarity between the generated questions and the reference questions. The BLEU-4 metric calculates the precision of n-grams up to four words in length, comparing the generated questions with the reference questions [46].

4.2. ROUGE Score:

This metric evaluates the quality of the generated questions by comparing them with reference questions based on n-gram overlap and word order. ROUGE-L is a metric specifically designed for evaluating the quality of summaries or generated questions. It calculates the recall score by comparing the generated questions with the reference questions based on the longest common subsequence between them. This metric is particularly useful in assessing the quality of generated questions when they may have different complexities than the input questions. ROUGE-L is a metric that measures the recall-oriented evaluation of generated questions by comparing them with reference questions [47].

4.3. METEOR Score:

This metric takes into account various linguistic aspects such as synonyms, stemming, and word order to measure the quality of the generated questions. METEOR is another automatic evaluation metric commonly used in question generation. METEOR is a metric that combines precision and recall to measure the quality of generated questions compared to the reference questions. It takes into account not only the overlap in n-grams but also considers synonyms and paraphrases to evaluate the semantic similarity between the generated questions and the reference questions [48].

4.4. CIDEr

CIDEr, also known as Consensus-based Image Description Evaluation, is an automated metric that was

developed to assess the quality of image descriptions. This metric compares a sentence generated by a model with a set of human-written sentences considered as ground truth. It measures how similar the model-generated sentence is to the consensus among the ground truth sentences for that particular image. The concept of consensus in CIDEr refers to how often most of the sentences used to describe an image are similar. Additionally, CIDEr claims that its metric inherently captures aspects such as grammar, importance, accuracy, and saliency [50].

4.5. Human Evaluation:

In human evaluation studies, human evaluators assess the quality of the generated questions based on criteria such as relevance, syntactic correctness, ambiguity, and difficulty. They evaluate if the generated questions are relevant to the given information, if they are grammatically correct if there is any ambiguity in the wording, and if the difficulty level of the questions aligns with the intended target audience. Additionally, the evaluation criteria used in human evaluation studies provide guidelines for the evaluators and help assess the performance of each template. The human evaluators rate the difficulty level of the questions to ensure that there is a syntactic divergence between the input sentence and the generated question [8].

In summary, the evaluation of Arabic question generation can be done using a combination of automatic metrics and human evaluation studies. These evaluation methods provide a comprehensive analysis of the generated questions, taking into account both quantitative metrics for similarity and quality, as well as qualitative assessments by human evaluators, as well as human evaluation studies to assess the relevance, syntactic correctness, ambiguity, and difficulty of the generated questions.

In Table 4, we list the different evaluation metrics used for question generation.

Table 4 Summery of Evaluation metrics and results

Ref.	*BLEU-4*	*ROUGE*	METEOR	CIDEr	f-measure
[5]	19.12	23.00	51.99	-	-
[1]	-	-	-	-	86%
[22]	-	-	-	-	98.7%
[30]	16.99	-	-	-	-
[9]	0.187	-	-	-	-
[31]	2.90	13.12	24.69	22.70	-
[34]	-	-	-	-	84.94%

5.Arabic Question Generation Challenges:

When it comes to Arabic question generation, several unique challenges need to be considered. These challenges include the complexity of Arabic language, which has different grammatical rules and structures

compared to other languages. Additionally, Arabic Question Generation faces several unique challenges that need to be addressed for effective implementation [1]:

5.1. Lack of Arabic-specific datasets:

One of the major challenges in Arabic question generation is the scarcity of high-quality datasets specific to the Arabic language. The availability of comprehensive datasets is crucial for training and evaluating question generation models effectively [22].

5.2. Limited linguistic resources:

Arabic question generation requires robust linguistic resources, including morphological analyzers, part-of-speech taggers, and syntactic parsers. These resources are essential for accurately understanding the semantics and syntax of Arabic text and generating grammatically correct and meaningful questions [1].

5.3. Inconsistencies in discretization:

Arabic words can have different meanings depending on the placement of diacritical marks. This creates challenges in accurately representing the intended meaning of words and generating questions that capture the desired information [36].

5.4. Complex sentence structures:

Arabic sentences often follow different syntactic structures compared to other languages. These complex sentence structures pose a challenge for question-generation models, as they need to understand and parse the sentences correctly to generate relevant and accurate questions [3].

5.5. Lack of domain-specific knowledge:

Arabic question generation for educational purposes requires a deep understanding of the subject matter and the ability to generate contextually relevant questions that align with the educational content [4].

6. Future work

To address these challenges and improve Arabic question generation, future research can focus on the following areas:

6.1. Development of Arabic-specific datasets:

Developing high-quality datasets dedicated for generating Arabic questions is crucial for the effective training and evaluation of models. These datasets should cover a wide range of diverse topics and domains, including educational content, to ensure the generation of contextually relevant questions [38].

6.2. Integration of advanced linguistic resources:

Research efforts can focus on improving the availability and quality of linguistic resources for Arabic

question generation. This can include developing more accurate morphological analyzers, part-of-speech taggers, and syntactic parsers specifically tailored for Arabic language processing [3].

6.3. Application of transfer learning:

Investigating transfer learning techniques provides an avenue to utilize pre-existing question-generation models trained in different languages and modify them for application in Arabic. This strategy has the potential to address the constraints posed by the scarcity of Arabic linguistic resources, thereby enhancing the effectiveness of models dedicated to generating questions in Arabic [4].

6.4. Designing hybrid models:

The development of hybrid models, integrating rule-based methodologies with deep learning techniques, offers a promising solution to tackle the complexities associated with Arabic question generation. These hybrid models can harness the advantages of rule-based approaches, adept at managing intricate sentence structures and incorporating domain-specific knowledge, while concurrently leveraging the capabilities of deep learning models to generate questions that are diverse and contextually relevant [4].

In conclusion, future research in Arabic question generation should focus on the development of Arabic-specific datasets, integration of advanced linguistic resources, application of transfer learning, and design of hybrid models to improve the accuracy and contextual relevance of generated questions.

7. Conclusion

This literature review discussed the current state of research on Arabic question generation using deep learning techniques for educational purposes. It highlighted the challenges faced in Arabic question generation, such as the lack of Arabic-specific datasets and linguistic resources, and suggests potential solutions, including the development of Arabic-specific datasets, integration of advanced linguistic resources, application of transfer learning, fine-tuning pre-trained multilingual models and the design of hybrid models. Overall, the literature review provided insights into the current state of research on Arabic question generation using deep learning techniques for educational purposes. Future research in Arabic question generation should focus on the development of Arabic-specific datasets, integration of advanced linguistic resources, application of transfer learning, and the design of hybrid models to improve the accuracy and contextual relevance of generated questions.

References

[1] Bousmaha, K. Z., Chergui, N. H., Mbarek, M. S. A., & Belguith, L. H. (2020). AQG: Arabic Question Generator. Rev. d'Intelligence Artif., 34(6), 721-729.

[2] Rahmani, A. M., Yousefpoor, E., Yousefpoor, M. S., Mehmood, Z., Haider, A., Hosseinzadeh, M., & Ali Naqvi, R. (2021). Machine learning (ML) in medicine: Review, applications, and challenges. Mathematics, 9(22), 2970.

[3] Biltawi, M. M., Tedmori, S., & Awajan, A. (2021). Arabic question answering systems: gap analysis. IEEE Access, 9, 63876-63904.

[4] Alwaneen, T H., Azmi, A M., Aboalsamh, H., Wang, Z., & Hussain, A. (2021, July 12). Arabic question answering system: a survey. https://doi.org/10.1007/s10462-021-10031-1.

[5] Alhashedi, S., Suaib, N. M., & Bakri, A. (2022). Arabic Automatic Question Generation Using Transformer Model (No. 8588). EasyChair.

[6] Al-Janabi, A., Al-Zubaidi, E. A., Al-Sagheer, R. H. A., & Hussein, R. (2020). Encapsulation of semantic description with syntactic components for the Arabic language. Indonesian Journal of Electrical Engineering and Computer Science, 22(2), 961-967.

[7] Fuad, A., & Al-Yahya, M. (2022). Recent developments in Arabic conversational AI: a literature review. IEEE Access, 10, 23842-23859.

[8] Keklik, O. (2018). Automatic question generation using natural language processing techniques (Doctoral dissertation, Izmir Institute of Technology (Turkey)).

[9] Beheitt, M. E. G., & Hmida, M. B. H. (2022). Automatic Arabic Poem Generation with GPT-2. In ICAART (2) (pp. 366-374).

[10] Ushio, A., Alva-Manchego, F., & Camacho-Collados, J. (2023). A Practical Toolkit for Multilingual Question and Answer Generation. arXiv preprint arXiv:2305.17416.

[11] Alsubhi, K., Jamal, A., & Alhothali, A. (2022). Deep learning-based approach for Arabic open domain question answering. PeerJ Computer Science, 8, e952.

[12] Jayarajan, A K. (2020, March 30). Automatic Question Generation u sing Sequence to Sequence RNN Model. https://scite.ai/reports/10.35940/ijitee.e2675.039520.

[13] Das, B., Majumder, M., Phadikar, S., & Sekh, A. A. (2021). Automatic question generation and answer assessment: a survey. Research and Practice in Technology Enhanced Learning, 16(1), 1-15.

[14] Khan, S., & Emara, S. A. (2018). Effect of Technology Use in Education. International Journal of Pedagogical Innovation, 6(2).

[15] Alotaibi, F., Abdullah, M T., Abdullah, R., Rahmat, R W O K., & Murad, M A A. (2019, November 22). A Method for Arabic Handwritten Diacritics Characters. https://scite.ai/reports/10.35940/ijeat.f1034.0986s319

[16] Ali, M. N., Tan, G., & Hussain, A. (2018). Bidirectional recurrent neural network approach for Arabic named entity recognition. Future Internet, 10(12), 123.

[17] Abdelali, A., Mubarak, H., Chowdhury, S. A., Hasanain, M., Mousi, B., Boughorbel, S., ... & Alam, F. (2023). Benchmarking Arabic AI with Large Language Models. arXiv preprint arXiv:2305.14982.

[18] Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP), 8(4), 1-22.

[19] Pranav, D S., & R, B P V. (2023, May 28). Histobot: Question Generation System Using Deep Learning Techniques. https://doi.org/10.47392/irjash.2023.s071

[20] Krishna, R., Bernstein, M S., & Li, F. (2019, March 26). Information Maximizing Visual Question Generation. http://arxiv.org/abs/1903.11207

[21] Abuali, B., & Kurdy, M. B. (2022). Full Diacritization of the Arabic Text to Improve Screen Readers for the Visually Impaired. Advances in Human-Computer Interaction.

[22] Antoun, W., Baly, F., & Hajj, H. (2020). AraGPT2: Pre-trained transformer for Arabic language generation. arXiv preprint arXiv:2012.15520.

[23] Alami, H., El Mahdaouy, A., Benlahbib, A., En-Nahnahi, N., Berrada, I., & Ouatik, S. E. A. (2023). DAQAS: Deep Arabic Question Answering System based on duplicate question detection and machine reading comprehension. Journal of King Saud University-Computer and Information Sciences, 35(8), 101709.

[24] Alkhurayyif, Y., & Sait, A. R. W. (2023). Developing an Open Domain Arabic Question Answering System Using a Deep Learning Technique. IEEE Access.

[25] Wyse, B., & Piwek, P. (2009). Generating questions from openlearn study units.

[26] Alsubait, T., Parsia, B., & Sattler, U. (2012). Automatic generation of analogy questions for student assessment: an Ontology-based approach. Research in Learning Technology, 20.

[27] Kalady, S., Elikkottil, A., & Das, R. (2010, June). Natural language question generation using syntax and keywords. In Proceedings of QG2010: The Third Workshop on Question Generation (Vol. 11, pp. 1-10). questiongeneration. org.

[28] Liu, M., Calvo, R. A., & Rus, V. (2010). Automatic question generation for literature review writing support. In Intelligent Tutoring Systems: 10th International Conference, ITS 2010, Pittsburgh, PA, USA, June 14-18, 2010, Proceedings, Part I 10 (pp. 45-54). Springer Berlin Heidelberg.

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[30] Nagoudi, E. M. B., Elmadany, A., & Abdul-Mageed, M. (2021). AraT5: Text-to-text transformers for Arabic language generation. arXiv preprint arXiv:2109.12068.

[31] Kalpakchi, D., & Boye, J. (2023). Quinductor: a multilingual data-driven method for generating reading-comprehension questions using universal dependencies. Natural Language Engineering, 1-39.

[32] El-Khair, I. A. (2016). 1.5 billion words Arabic corpus. arXiv preprint arXiv:1611.04033.

[33] Zeroual, I., Goldhahn, D., Eckart, T., & Lakhouaja, A. (2019, August). OSIAN: Open source international Arabic news corpus-preparation and integration into the CLARIN-infrastructure. In Proceedings of the fourth arabic natural language processing workshop (pp. 175-182).

[34] kamel, S. M., Hassan, S. I., & Elrefaei, L. (2023). VAQA: Visual Arabic Question Answering. Arabian Journal for Science and Engineering, 1-21.

[35] Kriangchaivech, K., & Wangperawong, A. (2019). Question generation by transformers. arXiv preprint arXiv:1909.05017.

[36] Shaalan, K., Siddiqui, S., Alkhatib, M., & Abdel Monem, A. (2019). Challenges in Arabic natural language processing. In Computational linguistics, speech and image processing for Arabic language (pp. 59-83).

[37] Suárez, P. J. O., Romary, L., & Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. arXiv preprint arXiv:2006.06202.

[38] Mozannar, H., Hajal, K. E., Maamary, E., & Hajj, H. (2019). Neural Arabic question answering. arXiv preprint arXiv:1906.05394.

[39] Al-Bataineh, H., Farhan, W., Mustafa, A., Seelawi, H., & Al-Natsheh, H. T. (2019, November). Deep contextualized pairwise semantic similarity for Arabic language questions. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1586-1591). IEEE.

[40] Lewis, P., Oğuz, B., Rinott, R., Riedel, S., & Schwenk, H. (2019). MLQA: Evaluating cross-lingual extractive question answering. arXiv preprint arXiv:1910.07475.

[41] Artetxe, M., Ruder, S., & Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. arXiv preprint arXiv:1910.11856.

[42] Roy, U., Constant, N., Al-Rfou, R., Barua, A., Phillips, A., & Yang, Y. (2020). LAReQA: Language-agnostic answer retrieval from a multilingual pool. arXiv preprint arXiv:2004.05484.

[43] Ismail, W. S., & Homsi, M. N. (2018). Dawqas: A dataset for Arabic why question answering system. Procedia computer science, 142, 123-131.

[44] Hardalov, M., Mihaylov, T., Zlatkova, D., Dinkov, Y., Koychev, I., & Nakov, P. (2020). EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. arXiv preprint arXiv:2011.03080.

[45] Bonifacio, L., Jeronymo, V., Abonizio, H. Q., Campiotti, I., Fadaee, M., Lotufo, R., & Nogueira, R. (2021). mmarco: A multilingual version of the ms marco passage ranking dataset. arXiv preprint arXiv:2108.13897.

[46] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

[47] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

[48] Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72).

[49] Riabi, A., Scialom, T., Keraron, R., Sagot, B., Seddah, D., & Staiano, J. (2021, January 1). Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering.https://scite.ai/reports/10.18653/v1/2021.emnlp-main.562

[50] Vedantam, R., Zitnick, C.L., Parikh, D. (2015). CIDEr: Consensusbased image description evaluation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 07–12-June, pp. 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087

[51] Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, *12*(1), 1-32.

[52] Ahmed, M. R. A. (2025). Accreditation and Quality Assurance: Exploring Impact and Assessing Institutional Change in the US and Saudi Arabian Higher Education Institutions. *Arts for Linguistic & Literary Studies*, *7*(1), 626–639. https://doi.org/10.53286/arts.v7i1.2419

[53] Al-Ghobesi, A. A. H. (2025). Risks of Relying on Artificial Intelligence in Learning Arabic Language Sciences Through the Meta Application. *Arts for Linguistic & Literary Studies*, *7*(1), 396–419. https://doi.org/10.53286/arts.v7i1.2420.

[54] Alharbi, K. N. . (2022). Female Saudi ESL Learners' Attitudes Toward Communication in Mixed Gender Classes in the USA. *Arts for Linguistic & Literary Studies*, *1*(13), 7–30. https://doi.org/10.53286/arts.v1i13.838.