MATHEMATICAL MODELING AND COMPUTING, Vol. 11, No. 3, pp. 763–772 (2024)

An Arabic question generation system based on a shared BERT-base encoder–decoder architecture

Lafkiar S., En Nahnahi N.

LISAC Laboratory, Faculty of Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University Fez, Morocco

(Received 21 January 2024; Accepted 16 August 2024)

A Question Generation System (QGS) is a sophisticated piece of AI technology designed to automatically generate questions from a given text, document, or context. Recently, this technology has gained significant attention in various fields, including education, and content creation. As AI continues to evolve, these systems are likely to become even more advanced and viewed as an inherent part of any modern e-learning or knowledge assessment system. In this research paper, we showcase the effectiveness of leveraging pre-trained checkpoints for Arabic questions generation. We propose a Transformer-based sequence-to-sequence model that seamlessly integrates with publicly accessible pre-trained AraBERT checkpoints. Our study focuses on evaluating the advantages of initializing our model, encompassing both the encoder and decoder, with these checkpoints. As resources for Arabic language are still limited and the publicly datasets for question generation systems in Arabic are not available, we collected our dataset for this task from various existing question answering, we used this latter to train and test our model. The experimental results show that our model yields performance was able to outperform existing Arabic question generation models in terms of the BLEU and METEOR scores,by achieving 20.29 as BLEU score and 30.73 for METEOR. Finally, we assessed the capability of our model to generate contextually relevant questions.

Keywords: *question generation system; natural language processing; reading comprehension; transformers; transfer learning.*

2010 MSC: 68T05 **DOI:** 10.23939/mmc2024.03.763

1. Introduction

Automatic question generation (AQG) refers to the process of using artificial intelligence (AI) and natural language processing (NLP) [1] techniques to generate questions from given texts, documents, or contexts without human intervention. This technology has found applications in various domains, including education, content creation and assessment. AQG systems aim to create relevant and coherent questions that can be used for quizzes, exams, content enrichment, enabling Chatbot to conduct a conversation [2], and more. They have the potential to save time, improve learning experiences, and enhance assessment processes. However, such systems are still under development and have the potential to revolutionize teaching and learning methods.

In the early stages of AQG research, studies focused on rule-based methods for generating questions. These approaches involved the manual formulation of grammatical rules and templates for question generation. While effective to some extent, they often struggled with handling the linguistic complexity [3–6]. Then, with the advent of machine learning and neural networks, researchers began to explore data-driven AQG methods. These approaches leveraged large corpora of text to train models capable of generating questions automatically. Some studies utilized sequence-to-sequence models, such as Recurrent Neural Networks (RNNs) [7, 8] and Transformer-based models [9, 10].

AQG is an important and ongoing area of research in natural language processing (NLP). It entails the generation of diverse question types, including interrogative, correctness/incorrectness, open-ended, fill-in-the-blank, and multiple-choice questions, based on input text, optionally accompanied by an answer. Notably, this task is more extensively explored in English compared to Arabic [11]. This can be explained by the lack of data and corpora in Arabic and the recent interest for tackling various natural language processing challenges related to this low-resource language.

In this research, we investigate an Arabic question generation system that operates on the principle of crafting questions with predefined answers, utilizing the capabilities of transfer learning from a finely tuned transformer model. Our study aims to demonstrate how a pre-trained transformer model, once fine-tuned, can proficiently produce questions that are both contextually fitting and logically coherent.

The remainder of this article follows this organization: Section 2 delves into the previous works of question generation, Section 3 outlines our novel question generation techniques, Section 4 provides a comprehensive breakdown of the implementations of these methods, Section 5 thoroughly examines and discusses the achieved results. Lastly, the concluding section serves as a comprehensive summary of our work and offers insights into future directions.

2. Related works

Question Generation System (QGS) has gained increasing attention in recent years, reflecting the broader interest in natural language processing and educational technology. Researchers and developers have explored various approaches and techniques to tackle the challenges specific to generating questions in different language. In this section, we review some of the notable works in the field of Arabic question generation.

The authors [12] implemented of a rule-based question generation system for Arabic text. This system employed a rule-based methodology to create questions, and the data source for this endeavor was the Arabic Language Book for the Fifth Standard in Yemen. The authors' approach involved several steps, including the removal of stop words and punctuation marks from the text. Furthermore, the paragraphs within the text were segmented using a two-step process, first at the sentence level and then at the word level. To ensure grammatical and linguistic accuracy, the authors developed Named Entity Recognition (NER) tagging and Part-of-Speech (POS) tagging components. In the process of question generation, special attention was given to nouns, particularly focusing on using them to construct wh-questions from the Arabic text.

In the study referenced as [13], the authors introduced an Arabic Question Generation (AQG) system designed to automatically create fill-in-the-blanks questions based on Arabic text. The system comprises four distinct stages: preprocessing, text processing, question generation, and post-processing. This system proves highly beneficial for educators as it offers a user-friendly graphical interface for generating a test using a set of questions.

The researchers in [5] have developed a new way to automatically generate questions in Arabic. This method can be used to create questions for the QUIZZITO platform, which is a platform for children's education. The method is flexible and can be used to generate a variety of different types of questions. It also takes into account the semantic role of the words in the question, which helps to ensure that the questions are meaningful and accurate. They created a set of rules that can be used to generate questions from text. These rules are based on the REGEX language [14], which is a language that is used to define patterns in text

The authors of paper [7] used two different Seq2Seq models, which are a type of neural recurrent network. They also explored several different types of attention mechanisms, which are a technique that can help neural networks to focus on the most important parts of an input sequence. They trained and tested their models on the ARCD dataset, which is a dataset of Arabic reading comprehension questions and answers. Their experimental results show that their models are able to generate relevant questions in the Arabic language.

In [10], the authors introduced an Automatic Arabic Question Generation (AAQG) model founded on the Transformer architecture. This model possesses the capability to generate numerous interrogative questions from educational content contained within a single document, regardless of its length. The architecture of this model consists of two key components: the first component serves as a fundamental question generation model, forming the core of Arabic automatic question generation, and the second component is specifically designed to overcome the limitations of the basic model and facilitate the generation of multiple questions from lengthy texts. It achieves this by extracting pivotal sentences from the text, which are then employed as inputs to the foundational model.

The researchers in paper [15] introduced ARGEN, a large language model for Arabic that can perform seven important tasks, including question generation. The authors train three powerful T5-style models on Modern Standard Arabic (MSA) and a variety of Arabic dialects, and evaluate them on the ARGEN benchmark.

In contrast to the various related studies mentioned earlier, there have been relatively limited efforts dedicated to addressing this particular challenge within the context of the Arabic language. Furthermore, the approach introduced by [10] lacks an answer-guided mechanism, as it does not exert control over the type of question generated based on a provided passage. This paper aims to fill this gap by developing an Arabic Question Generation System (AQGS) that leverages a pre-trained Transformer sequence-to-sequence model, incorporating both a source passage and a target answer to enhance question generation.

3. Proposed model

In this work, we develop an approach for Arabic question generation, which outperforms previous methods in terms of both potency and adaptability. This approach does not require any human intervention, and it can generate questions from any type of text input, which could have a number of potential benefits for education and knowledge assessment.

The proposed model is based on the Transformer BERT architecture, a cutting edge machine learning model renowned for its prowess in handling natural language processing tasks. This model is able to grasp extensive textual relationships, which is essential for generating coherent and meaningful questions.

We present our Arabic question generation architecture, which consists of three main components: 1) A pre-processing module, 2) A question generation module, and 3) A post-processing module. The system's configuration is shown in Figure 1.

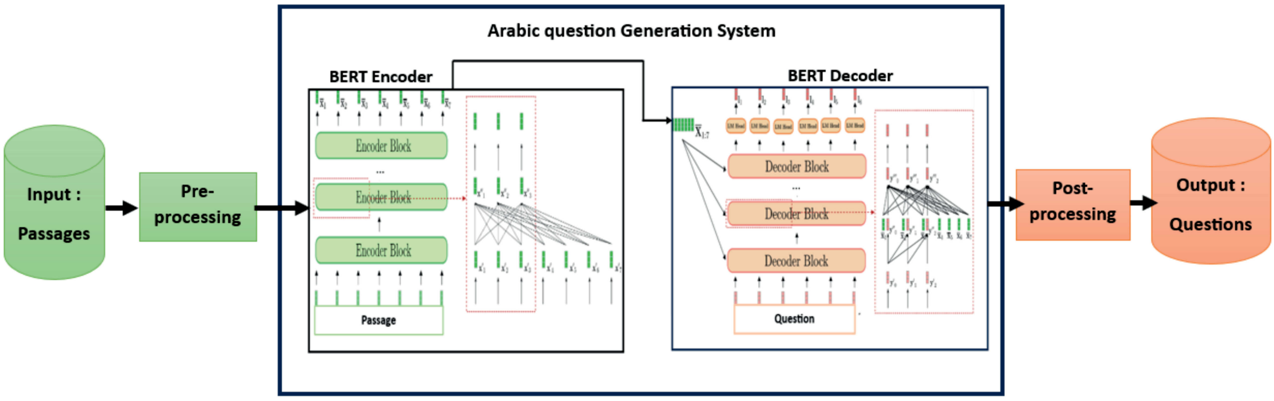


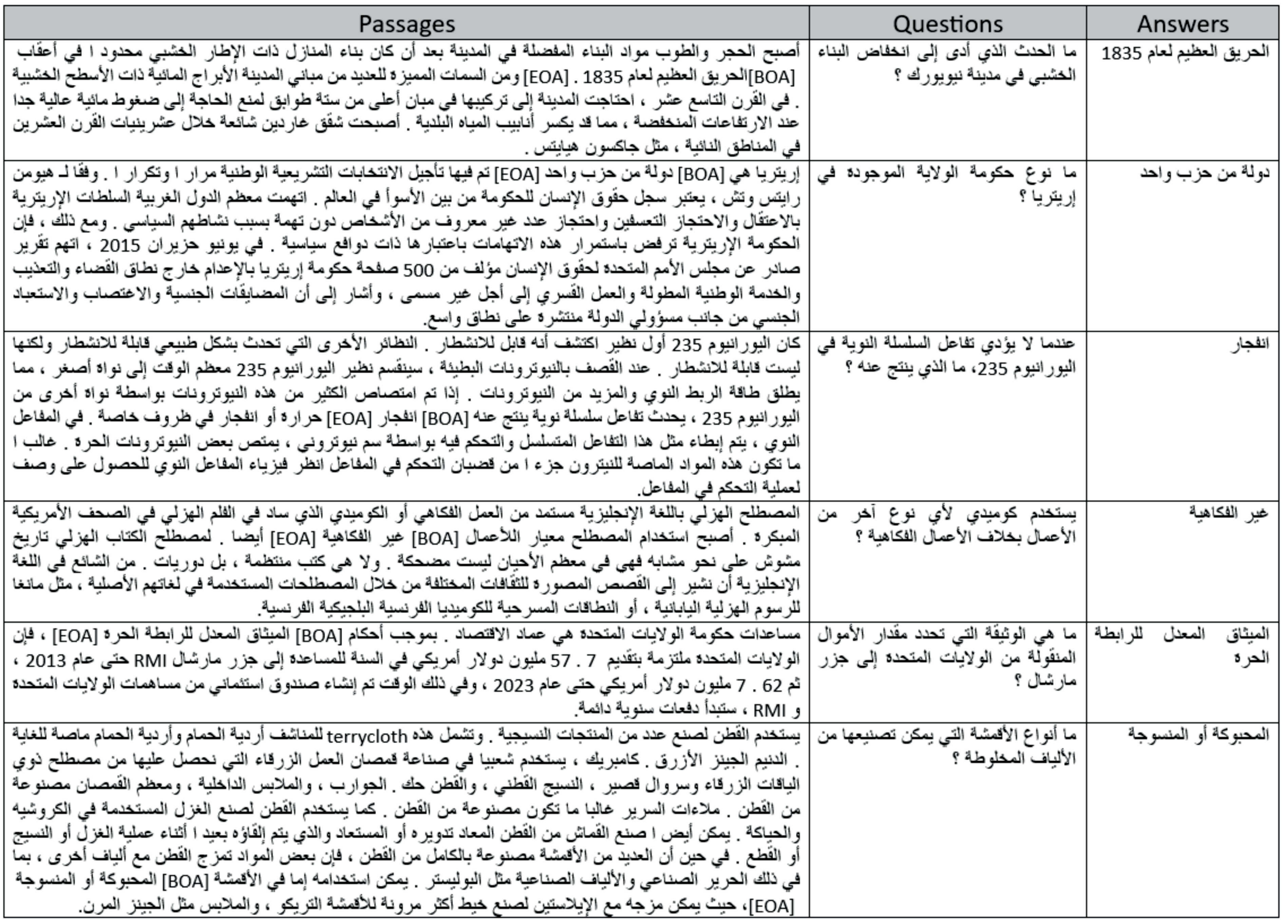
Fig. 1. The architecture of the proposed model by including checkpoints from BERT for Arabic questions generation.

3.1. Preprocessing module

The pre-processing data step is the process of preparing the input text for our model. This step helps the model to better understand the meaning of the input text and to generate more relevant and informative questions.

Although, we found specialized datasets like ARGENQG [15] but are not publicly available. Thus, due to the lack of datasets for Arabic Question generation systems, we utilized several datasets, including the Arabic Stanford Question Answering Dataset (Arabic-SQuAD), Arabic Reading Comprehension

Data (ARCD) [16], MultiLingual Question Answering (MLQA) [17], and Typologically Diverse Question Answering (TYDIQA) [18], as the basis for creating our custom dataset, referred to as ARabic Question Generation Data (ARQGData), Figure 2 shows an excerpt from our dataset. This dataset underwent a series of pre-processing steps, including text cleaning, normalization, stop word removal solely from passages, and the addition of special tokens $[BOA]$ and $[EOA]$ to delimit each answer within the passage.



Passages	Questions	Answers
أصبح الحجر والطوب مواد البناء المفضلة في المدينة بعد أن كان بناء المنازل ذات الإطار الخشبي محدود ا في أعقاب [BOA]الحريق العظيم لعام 1835 . [EOA] ومن السمات المميزه للعديد من مباني المدينة الأبراج المائية ذات الأسطح الخشبية . في القرن التاسع عشر ، احتاجت المدينة إلى تركيبها في مبان أعلى من ستة طوابق لمنع الحاجة إلى ضغوط مائية عالية جدا عند الارتفاعات المنخفضة ، مما قد يكسر أنابيب المياه البلدية . أصبحت شقق عاردين شائعة خلال عشرينيات القرن العشرين في المناطق النائية ، مثل جاكسون هيايتس .	ما الحدث الذي أدى إلى انخفاض البناء الخشبي في مدينة نيويورك ؟	الحريق العظيم لعام 1835
إريتريا هي [BOA] دولة من حزب واحد [EOA] تم فيها تأجيل الانتخابات التشريعية الوطنية مرار ا وتكرار ا . وفقا لـ هيومن رايتس وتش ، يعتبر سجل حقوق الإنسان للحكومة من بين الأسوأ في العالم . اتهمت معظم الدول الغربية السلطات الإريترية بالاعتقال والاحتجاز التعسفين واحتجاز عدد غير معروف من الأشخاص دون تهمة بسبب نشاطهم السياسي . ومع ذلك ، فإن الحكومة الإريترية ترفض باستمرار هذه الاتهامات باعتبارها ذات دوافع سياسية . في يونيو حزيران 2015 ، اتهم تقرير صادر عن مجلس الأمم المتحده لحقوق الإنسان مؤلف من 500 صفحة حكومة إريتريا بالإعدام خارج نطاق القضاء والتعذيب والخدمة الوطنية المطولة والعمل القسري إلى أجل غير مسمى ، وأشار إلى أن المضايقات الجنسية والاغتصاب والاستعباد الجنسي من جانب مسؤولي الدولة منتشره على نطاق واسع.	ما نوع حكومة الولاية الموجوده في إريتريا ؟	دولة من حزب واحد
كان اليورانيوم 235 أول نظير اكتشف أنه قابل للانشطار . النظائر الأخرى التي تحدث بشكل طبيعي قابلة للانشطار ولكنها ليست قابلة للانشطار . عند القصف بالنيوترونات البطيئة ، سينقسم نظير اليورانيوم 235 معظم الوقت إلى نواه أصغر ، مما يطلق طاقة الربط النوي والمزيد من النيوترونات . إذا تم امتصاص الكثير من هذه النيوترونات بواسطة نواه أخرى من اليورانيوم 235 ، يحدث تفاعل سلسلة نوية ينتج عنه [BOA] انفجار [EOA] حراره أو انفجار في ظروف خاصة . في المفاعل النوي ، يتم إبطاء مثل هذا التفاعل المتسلسل والتحكم فيه بواسطة سم نيوتروني ، يمتص بعض النيوترونات الحره . غالب ا ما تكون هذه المواد الماصة للنيترون جزء ا من قضبان التحكم في المفاعل انظر فيزياء المفاعل النوي للحصول على وصف لعملية التحكم في المفاعل.	عندما لا يؤدي تفاعل السلسلة النوية في اليورانيوم 235، ما الذي ينتج عنه ؟	انفجار
المصطلح الهزلي باللغة الإنجليزية مستمد من العمل الفكاهي أو الكوميدي الذي ساد في الفلم الهزلي في الصحف الأمريكية المبكره . أصبح استخدام المصطلح معيار اللأعمال [BOA] غير الفكاهية [EOA] أيضا . لمصطلح الكتاب الهزلي تاريخ مشوش على نحو مشابه فهي في معظم الأحيان ليست مضحكة . ولا هي كتب منتظمة ، بل دوريات . من الشائع في اللغة الإنجليزية أن نشير إلى القصص المصوره للثقافات المختلفة من خلال المصطلحات المستخدمة في لغاتهم الأصلية ، مثل مانغا للرسوم الهزلية اليابانية ، أو النطاقات المسرحية للكوميديا الفرنسية البلجيكية الفرنسية.	يستخدم كوميدي لأي نوع آخر من الأعمال بخلاف الأعمال الفكاهية ؟	غير الفكاهية
مساعدات حكومة الولايات المتحده هي عماد الاقتصاد . بموجب أحكام [BOA] الميثاق المعدل للرابطة الحره [EOA] ، فإن الولايات المتحده ملتزمة بتقديم 7 . 57 مليون دولار أمريكي في السنة للمساعده إلى جزر مارشال RMI حتى عام 2013 ، ثم 62 . 7 مليون دولار أمريكي حتى عام 2023 ، وفي ذلك الوقت تم إنشاء صندوق استئماني من مساهمات الولايات المتحده و RMI ، ستبدأ دفعات سنوية دائمة.	ما هي الوثيقة التي تحدد مقدار الأموال المنقولة من الولايات المتحده إلى جزر مارشال ؟	الميثاق المعدل للرابطة الحره
يستخدم القطن لصنع عدد من المنتجات النسيجية . وتشمل هذه terrycloth للمناشف أردية الحمام وأردية الحمام ماصة للغاية . الدنيم الجينز الأزرق . كامبريك ، يستخدم شعبيا في صناعة قمصان العمل الزرقاء التي نحصل عليها من مصطلح ذوي الياقات الزرقاء وسروال قصير ، النسيج القطني ، والقطن حك . الجوارب ، والملابس الداخلية ، ومعظم القمصان مصنوعة من القطن . ملاءات السرير غالبا ما تكون مصنوعة من القطن . كما يستخدم القطن لصنع الغزل المستخدمة في الكروشيه والحياكة . يمكن أيض ا صنع القماش من القطن المعاد تدويره أو المستعاد والذي يتم إلقاؤه بعيد ا أثناء عملية الغزل أو النسيج أو القطع . في حين أن العديد من الأقمشة مصنوعة بالكامل من القطن ، فإن بعض المواد تمزج القطن مع ألياف أخرى ، بما في ذلك الحرير الصناعي والألياف الصناعية مثل البوليستر . يمكن استخدامه إما في الأقمشة [BOA] المحبوكة أو المنسوجة [EOA]، حيث يمكن مزجه مع الإيلاستين لصنع خيط أكثر مرونة للأقمشة التريكو ، والملابس مثل الجينز المرن.	ما أنواع الأقمشة التي يمكن تصنيعها من الألياف المخلوطة ؟	المحبوكة أو المنسوجة

Fig. 2. Excerpt from our dataset ARQGData.

3.2. Question generation module

We describe our architecture for Arabic question generation which uses the Transformer model to convert one sequence (the passage) into another (the question). This transformation process encompasses the utilization of both an encoder and a decoder. At the core of this module lies the Transformer model, which takes as input the passage containing the Arabic answer delimiter $\mathbf{X}^p_{1:n}$ and produces an output question $\overline{Y}_{1:t}$ that is contextually linked to the answer within the passage. Where:

$$X^p_{1:n} = \left(x^p_1; \ldots ; x^{\langle BOA\rangle}; x^a_1; \ldots ; x^a_m; x^{\langle EOA\rangle}; \ldots ; x^p_n\right), \tag{1}$$

$$\overline{Y}_{1:t} = (y_1; \ldots ; y_t). \tag{2}$$

To generate an accurate Arabic question, the model tries to find the best question having height conditional likelihood.

Our architecture, based on Transformer encoder decoder models [19], both assembled using collections of residual attention units. The pivotal advancement of these models lies in the capability of these residual attention blocks to handle variable-length input sequences $X^p_{1:n}$ without relying on a recurrent structure. This unique trait eliminates the need for a repeating structure, making transformer based on encoder–decoder models greatly parallelized. When addressing a Seq2Seq problem, the primary goal is to establish a mapping between an input sequence $X^p_{1:n}$ and an output sequence $\overline{Y}_{1:t}$, where the output length might vary. Let us delve into how transformer built on encoder and decoder models are utilized to achieve this linkage. These models establish a conditional probability of target vectors $\overline{Y}_{1:t}$ based on an input sequence $X^p_{1:n}$:

$$p_{\theta_{\text{enc}},\theta_{\text{dec}}}\big(\overline{Y}_{1:t}|X^p_{1:n}\big). \tag{3}$$

We elaborate in the following section the functionalities of both the encoder and the decoder within this architecture.

* **Encoder:** The encoder of transformer-based processes the input sequence $X^p_{1:n}$, transforming it into a sequence of hidden states $\overline{X}^p_{1:n}$, thereby establishing a defined mapping [20]:

$$f_{\theta_{\text{enc}}}: X^p_{1:n} \to \overline{X}^p_{1:n}. \tag{4}$$

Upon closer examination of the architecture, the transformer-based encoder comprises a series of stacked residual encoder blocks. Every encoder block contains two sub-layers: the first being a multi-head self-attention mechanism, and the second, a straightforward position-wise fully connected feed-forward network. Within each sub-layer, we use a residual connection [21] followed by layer normalization [22]. This involves applying $\text{LayerNorm}(x+\text{Sublayer}(x))$ to the output of each sub-layer, where $\text{Sublayer}(x)$ represents the function executed by the sub-layer itself [19].

* **Decoder:** The decoder component models the conditional probability distribution of the target vector sequence $\overline{Y}_{1:t}$, considering the sequence of encoded hidden states $\overline{X}^p_{1:n}$,

$$p_{\theta_{\text{dec}}}\big(\overline{Y}_{1:n}|\overline{X}^p_{1:n}\big). \tag{5}$$

Applying Bayes' rule allows decomposed this distribution into a product of conditional probability distributions. To be specific, it encompasses the conditional probability distribution of the target vector y_i, given the encoded hidden states X_p, and all previous target vectors $Y_{0:i-1}$,

$$p_{\theta_{\text{dec}}}\big(\overline{Y}_{1:n}|\overline{X}^p_{1:n}\big) = \prod_{i=1}^{n} p_{\theta_{\text{dec}}}\big(y_i|\overline{Y}_{0:i-1}, \overline{X}^p_{1:n}\big). \tag{6}$$

The transformer-based decoder transforms the sequence of encoded hidden states $\overline{X}^p_{1:n}$ along with all previous target vectors $\overline{Y}_{0:i-1}$ into the logit vector I_i. Subsequently, this logit vector I_i undergoes a Softmax function to establish the conditional distribution $p_{\theta_{\text{dec}}}\big(y_i|\overline{Y}_{0:i-1}, \overline{X}^p_{1:n}\big)$. At this stage, it becomes feasible to auto-regressively generate the output, thereby defining a mapping from an input sequence $X^p_{1:n}$ to an output sequence $\overline{Y}_{1:t}$ during the inference process.

The decoder is likewise constructed with an equivalent number of blocks as the encoder. In the decoder, apart from the two sub-layers within each layer, there is an additional third sub-layer. This specific sub-layer conducts multi-head attention over the encoder stack's output. Similar to the encoder's setup, we incorporate residual connections surrounding each sub-layer, followed by layer normalization. Additionally, we make alterations in the uni-directional self-attention sub-layer within the decoder stack to prevent positions from attending to subsequent ones. This masking, along with the offset of output embeddings by one position, guarantees that predictions for position i solely rely on the known outputs at positions less than i [19].

Models like BERT, which are autoencoding models, share an identical architecture with the transformer models based on encoder. Models based on autoencoding [20] utilize this structure for extensive self-supervised pre-training on open-domain textual data, enabling them to convert any word sequence into a comprehensive bidirectional representation. The study conducted by [23] showcased that a pre-trained BERT model, augmented with a single task-specific classification layer, could attain cutting-edge performance across eleven natural language processing (NLP) tasks. We undertook the task of re-implementing from scratch the shared BERT-base encoder–decoder architecture as described in [24]. This re-implementation serves as our foundational model, depicted in Figure 1. This model is an encoder–decoder model built upon the architecture of BERT. Whereby, the parameters between the encoder and decoder are shared, significantly reducing the model's memory requirements, with only 136 million parameters compared to the original 221 million parameters.

3.3. Post-processing module

To evaluate and analyze the results of our experiment, we implemented a post-processing module aimed at enhancing the quality of the generated questions through:

- Eliminating redundant spaces, including spaces between words, spaces preceding and following punctuation marks, and consecutive multiple spaces.
- Eliminating words that appear after a question mark in the generated text.
- Appending a question mark to generated questions that do not conclude with one.

3.4. Metrics and evaluation

The most popular metrics used to evaluate the quality of natural language generation (NLG) systems are BLEU [25], METEOR [26], and ROUGE [27]. These metrics work by calculating the n-gram overlap between the reference sentence and the generated sentence. An n-gram is a sequence of n consecutive words in a sentence. In other words, these metrics compare the generated sentence to the reference sentence by counting the number of n-grams that they have in common. The higher the n-gram overlap, the higher the score, and the better the quality of the generated sentence is assumed to be. However, there are some limitations to these metrics. For example, they do not take into account the meaning of the words in the sentence, or the grammatical correctness of the sentence. This means that it is possible for a generated sentence to receive a high score even if it is not very meaningful or grammatically correct. Despite their limitations, BLEU, METEOR, and ROUGE are still widely used to evaluate the quality of NLG systems. This is because they are relatively easy to compute and interpret, and sound to correlate well with human judgments of NLG quality in some cases.

4. Experiment

To evaluate the performance of our question generation model, we conducted our experiment on the ARQGData dataset. We first describe the used corpus then we provide details on the model implementation and the experiment settings.

4.1. Dataset

We collected our dataset from four existing datasets as mentioned in Section 3. From this dataset, we extracted questions and their corresponding answers for each passage. Ultimately, this gathering produced a dataset containing 71 315 triplets, encompassing passages, questions, and corresponding answers. We trained and evaluated our model using consistent data split. We randomly partitioned our dataset into three distinct parts: we used 80% of dataset as training set to train our question generation model, 10% of dataset as validation set to evaluate the model during training and to tune the hyper-parameters of the model, and 10% of dataset as test set to evaluate the model after it has been trained.

4.2. Details of the implementation

We implemented our question generation model using the Encoder Decoder Model from the Transformers Python library. This library provides a variety of pre-trained models for natural language processing tasks. The Encoder Decoder Model is specifically designed for text-to-text generation tasks, such as translation, summarization and question generation. We used the same tokenization process as the pre-trained BERT models. The BERT models use a specific tokenization process that is designed to preserve the meaning of the text.

For our experiment the encoder and decoder were adjusted and initialized using publicly available AraBERT checkpoints to support Arabic language, leveraging the optimal weights of a pre-trained model that have been trained on large corpora of text. This pre-training helps the encoder and decoder to learn the general properties of language. It learns different aspects of language at different layers. It performs better than older methods and can potentially improve performance and reduce training costs by sharing the weights between layers.

Regarding the encoder, it retains the same architecture as BERT base model's [23]. The encoder takes a passage as input and converts each token within the passage into an embedding to generate a feature vector. Similarly, the decoder handles an input question by transforming each token into embeddings for further processing during the training phases. However, notable modifications have been made to the decoder in order to let it compatible with proposed hybrid architecture. Within each BERT block, a cross-attention layer is introduced with random initialization, inserted between the self-attention layer and the feed-forward layer. This adaptation allows the decoder to attend to the output from the encoder while generating text. To facilitate auto-regressive generation, the bidirectional self-attention layers in the decoder are transformed into unidirectional self-attention layers. As a result, the decoder's attention mechanism is constrained to consider only tokens that follow in the sequence, without the ability to account for tokens preceding it. Furthermore, the final decoder block incorporates an LM (Language Model) Head layer. This layer is responsible for predicting the subsequent token in the sequence. Typically, the LM Head layer is initialized with the same weights as the word embedding layer.

We fine-tuned model uses the same hyper-parameters as the BERT model. This included using 12 encoder and decoder blocks, 768 hidden units (d_{model}), a filter size of 3072, and 12 attention heads per mechanism. The key-value matrices were set to 64, following the guidance from [28]. The regularization parameters were also kept the same, with an epsilon value of 1e-12 using the AdamW optimizer and a dropout rate of 0.1. The checkpoints have a vocabulary size of around 64k word-pieces. The model was trained on a vocabulary of all the tokens that appear in the passages in the training dataset. This vocabulary includes words, numbers, and punctuation marks. The longest sentence in the training dataset is 509 tokens long, and the longest question in the training dataset is 42 tokens long. During training, the model was updated using batches of 8.

5. Results and discussion

The existing literature on machine learning techniques for generating Arabic questions is notably sparse and several factors contribute to this scarcity, including the lack of extensive and well-annotated datasets containing Arabic text and corresponding questions. The complexity of the Arabic language adds an additional challenge, making the development of high-quality question generation models a formidable task. Limited resources also hinder research efforts in Arabic question generation. Despite these obstacles, recent years have seen some progress in developing machine learning models for this purpose. Nevertheless, there is a substantial amount of work yet to be undertaken in this domain. Further research is essential to enhance machine learning models for Arabic question generation, and creating larger meticulously annotated datasets of Arabic text and questions should be a priority for future endeavors.

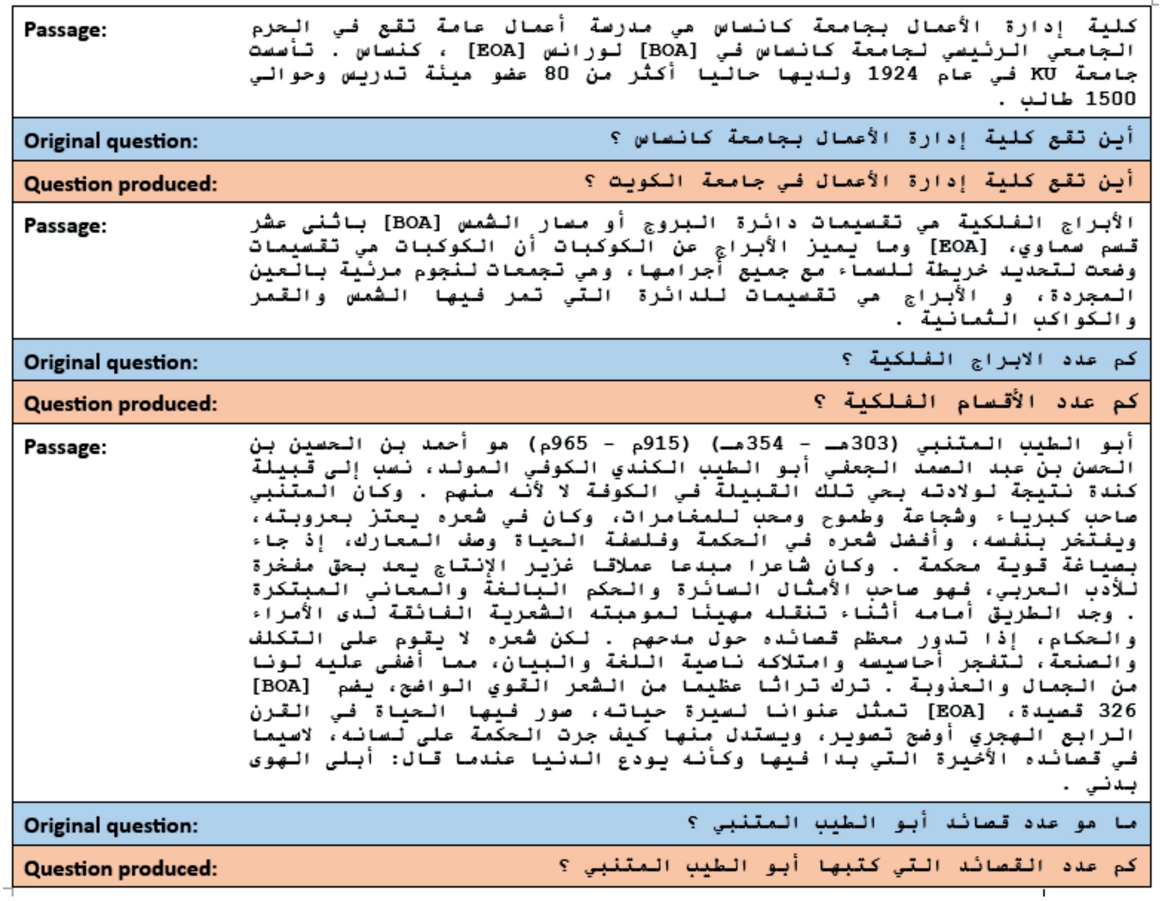
For these reasons, we conducted this experiment and juxtaposed their outcomes with those from previous Arabic language studies and that is for being able to assess the performance of our model and to identify areas where further improvement is needed. The recorded values of the $BLEU_4$, $ROUGE_L$, and $METEOR$ metrics in referenced studies are presented in Table 1.

Table 1. Results obtained from the fine tuning Transformer model for generating questions in Arabic.

$Model$	$BLEU_4$	$ROUGE_L$	$METEOR$	$Dataset$
AraT5 [15]	16:99	—	—	$ARGEN_{QG}$
Seq2Seq [7]	17.45	31.00	—	ARCD
AraBert2AraBert [10]	19.12	51.99	23.00	$mMARCO$
Our Model	**20.29**	38.54	**30.73**	**ARQGData**

The results presented in Table 1 yield several noteworthy findings. It is evident that our experiment has showcased superior performance when compared to the model proposed by [15], which relied on the AraT5 pre-trained model. Furthermore, in contrast to the approach taken by [10], which utilized the AraBERT pre-trained model but did not incorporate answers, our experiment has demonstrated better performance, as indicated by BLEU and METEOR metrics. It is worth highlighting that our model has also outperformed the results of our previous work [7] across all the metrics assessed. Our model is able to generate high-quality questions in Arabic. However, the score calculated based on the tokens lonely may not be an accurate reflection of the performance of our model, as some generated questions have the same meaning as the original questions, even though they use different tokens. Our model also has some limitations, such as some generated questions can be not well formed as the reference ones. Additionally, the model's performance can be impacted by limiting passage length to 512 tokens. Also, the generated questions can not be as complex as expected for some passages. There is still room for improvement in the development of our models for Arabic question generation.

Figure 3 presents examples of questions generated by our model, accompanied by the original question and the passage containing the answer span used in question generation.



Passage:	كلية إدارة الأعمال بجامعة كانساس هي مدرسة أعمال عامة تقع في الحرم الجامعي الرئيسي لجامعة كانساس في [BOA] لورانس [EOA] ، كنساس . تأسست جامعة KU في عام 1924 ولديها حاليا أكثر من 80 عضو هيئة تدريس وحوالي 1500 طالب .
Original question:	أين تقع كلية إدارة الأعمال بجامعة كانساس ؟
Question produced:	أين تقع كلية إدارة الأعمال في جامعة الكويت ؟
Passage:	الأبراج الفلكية هي تقسيمات دائرة البروج أو مسار الشمس [BOA] باثنى عشر قسم سماوي، [EOA] وما يميز الأبراج عن الكوكبات أن الكوكبات هي تقسيمات وضعت لتحديد خريطة للسماء مع جميع أجرامها، وهي تجمعات لنجوم مرئية بالعين المجردة، و الأبراج هي تقسيمات للدائرة التي تمر فيها الشمس والقمر والكواكب الثمانية .
Original question:	كم عدد الابراج الفلكية ؟
Question produced:	كم عدد الأقسام الفلكية ؟
Passage:	أبو الطيب المتنبي (303هـ - 354هـ) (915م - 965م) هو أحمد بن الحسين بن الحسن بن عبد الصمد الجعفي أبو الطيب الكندي الكوفي المولد، نسب إلى قبيلة كندة نتيجة لولادته بحي تلك القبيلة في الكوفة لا لأنه منهم . وكان المتنبي صاحب كبرياء وشجاعة وطموح ومحب للمغامرات، وكان في شعره يعتز بعروبته، ويفتخر بنفسه، وأفضل شعره في الحكمة وفلسفة الحياة وصف المعارك، إذ جاء بصياغة قوية محكمة . وكان شاعرا مبدعا عملاقا غزير الإنتاج يعد بحق مفخرة للأدب العربي، فهو صاحب الأمثال السائرة والحكم البالغة والمعاني المبتكرة . وجد الطريق أمامه أثناء تنقله مهيئا لموهبته الشعرية الفائقة لدى الأمراء والحكام، إذا تدور معظم قصائده حول مدحهم . لكن شعره لا يقوم على التكلف والصنعة، لتفجر أحاسيسه وامتلاكه ناصية اللغة والبيان، مما أضفى عليه لونا من الجمال والعذوبة . ترك تراثا عظيما من الشعر القوي الواضح، يضم [BOA] 326 قصيدة، [EOA] تمثل عنوانا لسيرة حياته، صور فيها الحياة في القرن الرابع الهجري أوضح تصوير، ويستدل منها كيف جرت الحكمة على لسانه، لاسيما في قصائده الأخيرة التي بدا فيها وكأنه يودع الدنيا عندما قال: أبلى الهوى بدني .
Original question:	ما هو عدد قصائد أبو الطيب المتنبي ؟
Question produced:	كم عدد القصائد التي كتبها أبو الطيب المتنبي ؟

Fig. 3. Examples of questions generated by our model.

6. Conclusion and perspectives

In this paper, we developed an Arabic question generation model that uses transfer learning to adapt the BERT encoder-decoder model to the Arabic language. We fine-tuned the model for the task of question generation and found that it was able to generate high-quality Arabic questions that were similar to reference questions. The model also performed well on a variety of metrics. However, the performance of the model was affected by the length of the text and the complexity of the questions. Our research aims to provide valuable insights and contributions that can assist fellow researchers interested in the development of question generation models for the Arabic language. Through creating a more interactive educational experience that meets the specific needs of Arabic-speaking learners and applied this approach in different educational contexts, including schools, universities, online courses, and self-learning platforms. To improve our model's performance, additional experiments should be conducted to search for optimal parameters. In our forthcoming research endeavors, we plan to expand our dataset by gathering a more extensive and diverse collection of text and question pairs. This enlarged dataset will be instrumental for training and evaluating Arabic question generation models. Furthermore, we intend to introduce novel evaluation metrics that measure the quality of generated questions based primarily on contextual comprehension rather than mere word similarity.

[1] Rakangor S., Ghodasara Y. Literature review of automatic question generation systems. International Journal of Scientific and Research Publications. **5** (1), 1–5 (2015).

[2] Achtaich K., Achtaich N., Fagroud F. Z., Toumi H. ALMA: Machine learning breastfeeding chatbot. Mathematical Modeling and Computing. **10** (2), 487–497 (2023).

[3] Chali Y., Hasan S. A. Towards Topic-to-Question Generation. Computational Linguistics. **41** (1), 1–20 (2015).

[4] Yao X., Bouma G., Zhang Y. Semantics-based Question Generation and Implementation. Dialogue Discourse. **3** (2), 11–42 (2012).

[5] Bousmaha K. Z., Chergui N. H., Mbarek M. S. A., Belguith L. H. AQG: Arabic Question Generator. Revue d'Intelligence Artificielle. **34** (6), 721–729 (2020).

[6] Banou Z., Elfilali S., Benlahmar H. Towards a polynomial approximation of support vector machine accuracy applied to Arabic tweet sentiment analysis. Mathematical Modeling and Computing. **10** (2), 511–517 (2023).

[7] Lafkiar S., Hamza A., Zouitni M., Burmani N., Badir H., En Nahnahi N. Attentive Neural Seq2Seq for Arabic Question Generation. International Conference on Advanced Intelligent Systems for Sustainable Development. 802–816 (2022).

[8] Du X., Shao J., Cardie C. Learning to Ask: Neural Question Generation for Reading Comprehension. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 1342–1352 (2017).

[9] El Moatez B. N., Elmadany A., Abdul-Mageed M. AraT5: Text-to-text transformers for Arabic language generation. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 628–647 (2022).

[10] Alhashedi S., Suaib N. M., Bakri A. Arabic Automatic Question Generation Using Transformer Model. EasyChair Preprint no. 8588 (2022).

[11] Kurdi G., Leo J., Parsia B., Sattler U., Al-Emari S. A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education. **30**, 121–204 (2020).

[12] Alazani S. A., Mahender C. N. Rule based question generation for Arabic text: question answering system. Proceedings of the International Conference on Data Science, Machine Learning and Artificial Intelligence. 7–12 (2021).

[13] Re O. C. Building a system based on natural question generatio. Mohamed Elbasyouni (2014).

[14] Freydenberger D. D. Extended regular expressions: Succinctness and decidability. Theory of Computing Systems. **53**, 159–193 (2013).

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

[16] Mozannar H., Maamary E., El Hajal K., Hajj H. Neural Arabic Question Answering. Proceedings of the Fourth Arabic Natural Language Processing Workshop. 108–118 (2019).

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

[18] Clark J. H., Choi E., Collins M., Garrette D., Kwiatkowski T., Nikolaev V., Palomaki J. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. Transactions of the Association for Computational Linguistics. **8**, 454–470 (2020).

[19] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need. Advances in Neural Information Processing Systems. **30**, 1–11 (2017).

[20] Bas E. A robust optimization approach to diet problem with overall glycemic load as objective function. Applied Mathematical Modelling. **38** (19–20), 4926–4940 (2014).

[21] He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 770–778 (2016).

[22] Ba J. L., Kiros J. R., Hinton G. E. Layer normalization. Preprint arXiv:1607.06450 (2016).

[23] Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Preprint arXiv:1810.04805 (2018).

[24] Rothe S., Narayan S., Severyn A. Leveraging pre-trained checkpoints for sequence generation tasks. Transactions of the Association for Computational Linguistics. **8**, 264–280 (2020).

[25] Papineni K., Roukos S., Ward T., Zhu W.-J. Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 311–318 (2002).

[26] Lavie A., Denkowski M. J. The METEOR metric for automatic evaluation of machine translation. Machine translation. **23** (2), 105–115 (2009).

[27] Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of the Workshop on Text Summarization Branches Out. 74–81 (2004).

[28] Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research. **21** (1), 5485–5551 (2020).

Арабська система генерування запитань на основі спільної базової архітектури кодера–декодера BERT

Лафкіар С., Ен Нахнахі Н.

Лабораторія LISAC, Факультет наук Дхар Ель Мараз,
Університет Сіді Мохамед Бен Абделла Фес, Марокко

Система генерації запитань (QGS) — це складна частина технології штучного інтелекту, призначена для автоматичного генерування запитань із певного тексту, документа чи контексту. Останнім часом ця технологія привернула значну увагу в різних сферах, включаючи освіту та створення контенту. Оскільки штучний інтелект продовжує розвиватися, ці системи, швидше за все, стануть ще більш досконалими та розглядатимуться як невід'ємна частина будь-якої сучасної системи електронного навчання чи оцінки знань. У цій дослідницькій роботі демонструється ефективність використання попередньо підготовлених контрольних точок для створення питань на арабській мові. Пропонується засновану на Transformer модель послідовностей, яка легко інтегрується із загальнодоступними попередньо навченими контрольними точками AraBERT. Наше дослідження зосереджено на оцінці переваг ініціалізації нашої моделі, що охоплює як кодер, так і декодер, за допомогою цих контрольних точок. Оскільки ресурси для арабської мови все ще обмежені, а загальнодоступні набори даних для систем генерування питань арабською мовою недоступні, зібрано набір даних для цього завдання з різних існуючих відповідей на запитання, які використані для навчання та тестування запропонованої моделі. Експериментальні результати показують, що продуктивність запропонованої моделі змогла перевершити існуючі моделі генерації запитань арабською мовою за оцінками BLEU та METEOR, досягнувши 20.29 бали для оцінки BLEU та 30.73 бали для METEOR. Нарешті, оцінено здатність нашої моделі генерувати контекстуально відповідні запитання.

Ключові слова: *система формування питань; обробка природної мови; розуміння прочитаного; трансформери; трансферне навчання.*