International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 38, No. 4, August, 2024, pp. 1275-1284
Journal homepage: http://iieta.org/journals/ria

Enhancing Question Generation in Bahasa Using Pretrained Language Models

Renaldy Fredyan[1*], Ivan Satrio Wiyono[2], Derwin Suhartono[1], Muhammad Rizki Nur Majiid[3], Fredy Purnomo[1]

[1] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
[2] Mathematics Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
[3] Computer Science Department Semarang Campus, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Corresponding Author Email: renaldy.fredyan@binus.ac.id

https://doi.org/10.18280/ria.380421

Received: 8 December 2023
Revised: 20 March 2024
Accepted: 8 May 2024
Available online: 23 August 2024

Keywords:
context-based question generation, IndoBERT, IndoGPT, IndoBART, question-answering datasets

ABSTRACT

Automatic Question Generation (AQG) from text is difficult, especially in Indonesia, where research is scarce. Current research focuses on factual questions, leaving room for improvement. Previous studies used sequence-to-sequence models, which are effective for rule-based and cloze testing but rely on pre-existing rules. This article evaluates state-of-the-art pre-trained models such as IndoBERT, IndoGPT, and IndoBART as well as classical models such as BiGRU, BiLSTM, and Transformer to fill this gap. This paper tests model question generation using SQuAD-ID, IDK-MRC, and TyDi-QA, three popular question-and-answer datasets. This study uses BLEU and ROUGE-L to evaluate each model's ability to generate meaningful queries from the provided settings. This research aims to understand AQG in Indonesian and evaluate model performance. Discusses the background of AQG research, model limitations, as well as research topics and hypotheses. The paper also analyzes the expected contributions, such as the effectiveness of the trained model and the architectural effects on the AQG process. This research improves natural language processing and question generation systems, especially for Indonesian.

1. INTRODUCTION

The automated production and answering of questions are a critical feature that bears great importance in a variety of sectors of application, including but not limited to education [1, 2], personal assistance [3], and healthcare [4]. The process of generating questions manually demands significant investment of both effort and resources, resulting in high costs and time consumption. Conventional educational settings encompass scheduled evaluations such as tests, quizzes, and exams, in addition to spontaneous inquiries posed by the educator during or following each instructional session. This facilitates the learner's ability to assess their comprehension, while also allowing the instructor to evaluate the efficacy of their instructional methods. The process of generating and choosing questions can be a laborious undertaking. The process of formulating high-quality questions is a multifaceted task that necessitates specialized training and practical expertise.

In addition, a significant number of students have adopted a self-directed learning approach through Massive Open Online Courses (MOOCs) due to their unmatched convenience, either to acquire new knowledge or complement their conventional classroom education [5]. However, a considerable proportion of virtual learning environments exhibit inadequacies in terms of suitable and sufficient assessments for evaluating learners, primarily due to the temporal limitations encountered by the developers. Therefore, this arises as an urgent requirement. As a result, there has been a notable degree of attention devoted to the creation of a self-governing system capable of generating queries and their respective responses within the last ten years.

Interrogatives are an essential component of knowledge acquisition and a vital tool in the area of teaching. They may be used to elicit more information or to confirm the degree of engagement in the lesson. Posing questions to learners can provide several benefits, including feedback on their comprehension and potential misunderstandings, the opportunity to practice recalling information from memory, the identification of key learning material and subsequent concentration on it, the stimulation of motivation to participate in learning activities, and the repetition of fundamental concepts to bolster the process of knowledge acquisition. As per academic norms, investigations can be broadly categorized into two distinct classifications, specifically subjective inquiries and objective inquiries [6].

The responses are produced using data analysis from the specified context. The responses are organized into separate duties. In the case of objective inquiries, the response comprises quoting a word or phrase from the provided paragraph [7]. Conversely, the multiple-choice method necessitates the selection of the accurate answer from a set of possible options. The subjective question answering approach involves the extraction of a subsequence from the given context to serve as a response for short answer questions, while free answering techniques are employed for long answer questions.

Conventionally, there have been two predominant approaches in natural language processing: statistical approaches and rule-based [8]. The development of regulations and frameworks incurs significant costs, exhibits limited variability, and presents challenges in terms of applicability across various fields. The latest developments in deep neural networks have demonstrated superior performance in various aspects of natural language generation compared to previous methods. However, this comes at the expense of increased processing requirements and a significant amount of training. The adoption of neural techniques has been facilitated by the reduction in hardware expenses and the enhanced accessibility of cloud platforms. Recent advancements in deep neural networks field, including attention mechanisms [9], copy mechanisms [10], and memory networks [11], have demonstrated encouraging outcomes for the task of generating questions and answers. Nevertheless, the task of producing a category of question-answer pairs from the given context continues to pose a considerable difficulty. The present study endeavours to tackle this challenge by introducing a system that can generate a diverse set of question-answer pairs pertaining to a given article in an automated manner.

The aforementioned methodologies necessitated a step-by-step procedure, rendering the training process arduous and protracted. However, transformers [12] has enabled the parallelization of tasks through the ingestion of the entire sequence as input, as opposed to the token-by-token processing approach, which has resulted in its widespread adoption. The utilization of transformer-based networks has become a conventional practice for various natural language tasks, which is then succeeded by fine-tuning over a vast corpus to attain the most favorable outcomes [13]. Over the past few years, multiple iterations of transformers have been introduced. The GPT [14], BART [15] and BERT [13] transformers differ in their underlying architecture, with the former utilizing a left-to-right decoder and the latter employing a deep bidirectional long short-term memory encoder.

Its variants, IndoBART, IndoBERT, and IndoGPT, have become important frameworks in the field of pre-trained language models. These models have demonstrated remarkable efficacy across a range of natural language processing tasks. Through the process of fine-tuning these trained models to fit particular applications, one can effectively leverage the knowledge already present in them. Researchers and industry experts have used the capabilities of IndoBART, IndoBERT, and IndoGPT to tackle a variety of natural language processing issues, such as sentiment analysis, text classification, language generation, and machine translation. They are a priceless tool in the field of natural language understanding because of their effectiveness and versatility.

The aim of this study is to aid or support to educators in the process of creating assessments for reading comprehension (RC) [16]. The focus of this work is on Automatic Question Generation with a Transformers-based language model. It is becoming more relevant and crucial to provide engaging experiences for in-person interactions in the classroom as teachers spend more time with students and less time on regular tasks. Consequently, this will lessen students' fear of the virtual learning environment, where they are isolated from both teachers and other pupils [17].

The primary investigation of this research inquiry is as follows: In what ways can the development of an Automated Question Generation (AQG) system intended for Reading Comprehension (RC) evaluation benefit from the progress made in natural language processing models such as IndoBERT and IndoGPT? This research can address the issue where educators encounter difficulty in formulating superior inquiries that effectively evaluate students' comprehension and higher-order cognitive capacities. To achieve this, our aim is to simplify the question generation process and enhance educational assessment through the implementation of an Automatic Question Generation (AQG) system based on Natural Language Processing (NLP). The challenges and benefits of implementing Automatic Question Generation (AQG) in multilingual education, with a concentration on Indonesian, will be the focus of our investigation. This study departs from prior investigations that primarily employed the English language. Our investigation seeks to comprehend the intricacies of the AQG model as it pertains to educational evaluation. We will conduct an analysis of the challenges, opportunities, and most effective strategies associated with this model in order to improve educational assessment methods in multilingual settings.

The following describes the organization of the subsequent sections of this study: The first section introduces the significance of conducting this research. Following this, a concise overview of the existing methodologies utilized for question generation and answer extraction is presented in Section 2. A comprehensive analysis of the structure and development of the query generation system's constituent components is provided in Section 3. Also, the algorithms used during each module's execution are examined. The results of the evaluation of the system are detailed in Section 4. Therefore, the concluding remarks and prospective directions for future research are provided in Section 5.

2. RELATED WORKS

Traditionally, rule-based approaches to question generation have depended on templates that were created manually [18]. However, this approach has its limitations since it restricts the range of questions that may be generated and their applicability to different areas. To address these constraints, researchers have created advanced techniques based on deep neural networks [19], such as recurrent neural networks (RNNs) [20], which include long short-term memory (LSTM) networks [21], gated recurrent units (GRUs) [22], and their variations. Nevertheless, these techniques frequently necessitate significant computational burden and lengthy instruction. Transformers [23] have transformed AQG by facilitating parallel processing, resulting in its extensive usage.

Consequently, a discernible transition has occurred in the field of natural language processing, moving away from rule-based and statistical approaches towards deep learning techniques, and more recently, transformers. This shift has been motivated by the superior outcomes achieved through the utilization of the latter. Historically, neural networks have exhibited superior performance in the majority of natural language generation tasks. However, this has been accompanied by a processing overhead and a substantial training requirement. The adoption of neural techniques has been facilitated by the reduction in hardware costs and increased availability, which has been further supported by the proliferation of cloud platforms. The utilization of memory networks [11], attention mechanisms [9], and copy mechanisms [10] has led to notable progress in the application of deep neural networks for the purpose of generating question-answer pairs.

The automated generation of questions through the use of a deep sequence-to-sequence neural model was first introduced by Du et al. in their study [20]. To generate answer-aware questions, the initial step involves extracting the positions of the answer span from the input sentence. Subsequently, questions that are specific to the answer are generated [24]. The majority of prior research endeavours have employed an encoder-decoder architecture that incorporates an attention mechanism [25]. Various techniques have been implemented by different models to extract answer information, including the approach of initially identifying the answer that is relevant to the question and subsequently formulating a question that is cognizant of the answer [26]. In addition, various techniques have been employed such as incorporating the relative distance between the answer and the context words [27], utilizing key-phrase extraction to identify potential answer candidates [28], implementing an indicator for the position of the answer [29], among others. In their study, Du and Cardie employed gated coreference knowledge to facilitate the generation of questions that are answer-aware at the paragraph level [25]. The authors Zhao et al. [30] introduced a sequence-to-sequence (seq2seq) neural network architecture that incorporates a gated self-attention encoder and a maxout pointer decoder for the purpose of generating single questions at the paragraph level. Nema et al. [31] proposed refined networks for the task of question generation. Ma et al. [32] employed deep neural networks to generate questions through the process of answer-separation. Vu et al. [33] employed a decoder-only transformer model to generate questions at the paragraph level. Liu [34] employed sequence-to-sequence (seq2seq) algorithms featuring copy and attention mechanisms in the context of question generation. Various methodologies have been employed for the task of generating questions and answers, however, they exhibit restricted capabilities when compared to the approach we have put forth.

Transformers were introduced as a result of the incorporation of attention [23]. Prior methodologies necessitated sequential processing, leading to a laborious and time-intensive training process. However, transformers have enabled the parallelization of tasks by accepting the complete sequence as input, rather than token by token, resulting in their widespread adoption. Over the recent years, a number of different iterations of transformers have been presented. The architecture of GPT is founded on a decoder that operates from left to right, while BERT is based on a deep encoder that utilizes bidirectional long short-term memory. The concept of transfer learning models centers on the retention of acquired knowledge from the training process of a particular problem and its subsequent application to a related yet distinct problem [35]. Dehghani and colleagues employed a transformer architecture in their study on open-domain question answering [36]. In their study, Lopez et al. [37] employed the GPT-2 transformer architecture to generate questions based on a provided paragraph.

The majority of extant question-answer generation models have been put forth with regard to a solitary sentence context and are capable of producing solely three categories of WH-question, namely what, who, and where. These systems rely on manually crafted rules for generating questions and answers, and struggle when presented with complex sentences [38]. Moreover, the current neural models produce a single question for each context [31]. Prior studies on question generation have also employed transformers [37], albeit in either an encoder or decoder capacity. Certain research endeavors employ an encoder-decoder methodology; however, they may exhibit a deficiency in the variety of the produced question-answer pairs [39].

The present study employs a transformer model based on an encoder-decoder architecture to produce question-answer pairs from a given context [40]. The passage may consist of either a singular or multiple sentences within a given context. The system under consideration exhibits the potential to effectively accommodate varying perspectives and question-answer styles. The main aim of this research is to create a system that can assist educators and learners in producing educational materials for a wide range of subject areas. This system aims to serve multiple purposes, but the manual production of such content can be time-consuming. For a thorough evaluation, it is essential to incorporate inquiries that are both subjective and objective. The system under consideration exhibits the capability to generate aforementioned inquiries. One of the primary challenges encountered by educators involved in direct classroom instruction is the limited amount of time available to develop high-quality instructional materials. Generating question-answer pairs that are both diverse and grammatically accurate from a given text remains a substantial obstacle. The present study aims to tackle the aforementioned obstacles by introducing a novel system that can generate a diverse set of question-answer pairs pertaining to a specific paragraph in an automated manner.

3. METHOD

3.1 Dataset used

We outline the general architecture of our Question Generation (QG) system in this section. Every module is covered in detail. As previously mentioned, we take the task of question generation with an objective in mind. For this, we employ three datasets: IDK-MRC (Indonesian Machine Reading Comprehension), TyDi-QA (Typologically Diverse Question Answering), and SQuAD (Stanford Question Answering Dataset). These datasets are highly esteemed as significant assets for assessing machine reading comprehension and question-answering systems, especially in multilingual and cross-lingual scenarios.

The collection contains a number of distinct subjects. For every issue, a goal and a list of pertinent questions are provided. Along with the findings, the actual answers and the names of the files containing the answers were made available. We start by cleaning up the data to remove any extraneous information. To identify the sentences that are pertinent to either the target, the answer, or both, we use the actual answers along with the targets. The sentences may contain several clauses and a complex structure. Therefore, it would be challenging to formulate meaningful queries from the intricate language. Consequently, we streamline the procedure by employing syntactic information to separate simple statements from complicated sentences. To determine the potential kinds of inquiries that could be formed, we categorize the sentences according to their subject, verb, object, and preposition.

3.2 Implementation details

The cleaning and processing of raw data is the critical first step in any NLP operation. We remove any superfluous tags and text from each document and query in the dataset. The Oak system is used to tokenize the questions and sentences [41]. The dataset's text file contains the details of the actual answers to the questions. Parsing extracts the topic number, response, and file name containing the answers to the questions. The system opens the answer file and searches for relevant sentences that contain the answer, the aim, or both. We parse fewer sentences as a result. The Named Entity (NE) and Parts of Speech (POS) taggers receive the pertinent phrases after that. To create the POS tagged sentence, we employ the Oak system. We shall learn about the verbs and their tenses from the POS-tagged sentences. We use this information to extract every verb in a sentence. Once more, the NE labeled phrases are generated using the Oak system. Among the 150 Named Entity kinds that are possible, a sentence may contain the following Named Entity types: Person, Location, Organization, Facility, Date, Money, Percent, Time, etc.

This module accepts simple sentences as input. Based on the POS and NE tagged data, we obtain the subject, object, preposition, and verb for each simple sentence. Based on this information, we categorize the sentences. We utilize a two-layered taxonomy to reflect natural semantic classification for the sentences. It aims to provide a structured framework for categorizing sentences based on their content and purpose. This taxonomy typically consists of two main layers: sentence type and sentence function. At the first layer, sentences are classified according to their grammatical structure and form, including declarative, interrogative, imperative, and exclamatory types. For example, "The cat is sleeping" is a declarative sentence, while "Is the cat sleeping?" is interrogative, "Wake up, cat!" is imperative, and "What a beautiful sunset!" is exclamatory. The second layer of classification focuses on the intended function or purpose of the sentence in communication. This layer includes categories such as statement/assertion, question/interrogation, command/imperative, and exclamation/exclamatory. For instance, "The cat is sleeping" serves as a statement or assertion, "Where is the cat?" functions as a question or interrogation, "Wake up, cat!" gives a command or imperative, and "What a beautiful sunset!" expresses an exclamation or exclamatory sentiment. By employing this taxonomy, sentences can be systematically categorized, facilitating analysis of their structure, purpose, and function in various contexts.

Our sentence classifier module employs two fundamental classifiers in succession. The sentences are divided into fine (Fine Classifier) and coarse (Coarse Classifier) categories in the first. This strategy is comparable to that which is explained in the study [42]. Because the second classifier generates its candidate labels by condensing the first's set of retained fine classes into a set of coarse classes, it has an impact on the first classifier. The 150 types in the OAK System are tagged. They're part of a hierarchy. The candidate fine and coarse classes are created using this information. The five coarse classes are defined as follows:

1. ***Human***: *Any topic or thing with a person's name on it.*
2. ***Entity***: *Contains all objects, plants, animals, and mountains.*
3. ***Location***: *Terms denoting places, like nation, city, school, etc.*
4. ***Time***: *Words like year, Monday, 9 a.m., last week, etc. that denote a time, date, or period.*
5. ***Count***: *Gather all the items that have been tallied, including the weight and the nine men.*

To classify each sentence, we apply a top-down processing method. Let $C_0 = \{c_1, c_2, \ldots, c_n\}$, the set of all the coarse classes, be the confusion set of any sentence. The fine classifier first establishes the fine classes. The class hierarchy then narrows down the set of fine classes to a coarse class. In other words, the coarse class c_i is translated into the collection of fine classes $\{f_{i1}, f_{i2}, \ldots, f_{im}\}$. We take into account the relationship between the terms in the phrase based on the coarse classification. A statement with the structure "Human Verb Human," for instance, will be categorized as a "whom and who" query type.

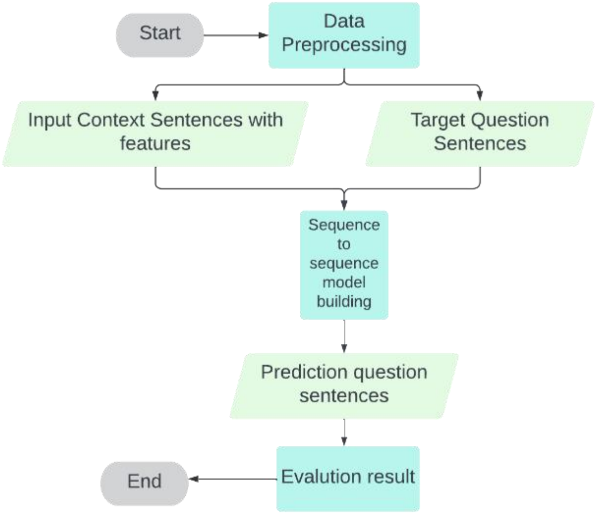
Start
Data Preprocessing
Input Context Sentences with features
Target Question Sentences
Sequence to sequence model building
Prediction question sentences
Evalution result
End

Figure 1. Diagram of the related research modelling process [43]

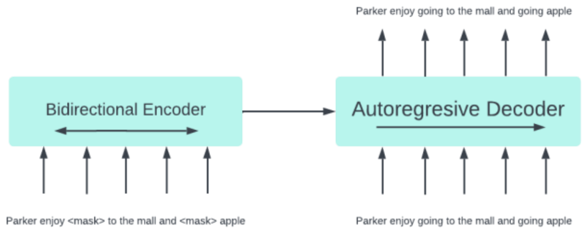
Parker enjoy going to the mall and going apple
Bidirectional Encoder
Autoregresive Decoder
Parker enjoy <mask> to the mall and <mask> apple
Parker enjoy going to the mall and going apple

Figure 2. Illustration of the Transformer based encoder-decoder model [44]

Our model is based on a sequence-to-sequence approach language model developed by one study in Indonesian, as shown in Figure 1, and trained using the SQuAD and TyDi QA datasets. The datasets were translated into Indonesian using the Google Translate API v2, and the model was created by combining the Transformer architecture with Recurrent Neural Network (RNN) models such as BiLSTM and BiGRU [43]. A language model known as IndoBART (Indonesia Bidirectional and Auto-Regressive Transform) is pre-trained by adding noise or corruption to the input sequence. The model is then tasked with reconstructing the original input sequence. The cross-entropy result of the model predictions will then be computed against the loss function, followed by the back-propagation gradients process and model weight updates. In the IndoBART language model architecture, an encoder on IndoBERT (Indonesia Bidirectional Encoder Representations from Transformers) and a decoder on IndoGPT (Indonesia Generative Pre-Trained Transformer) that can execute NLP tasks in the form of NLU and NLG are employed. In order to generate text based on this approach, we intend to perform a comparative analysis of IndoBART, IndoBERT, and IndoGPT, as can be seen in Figure 2. Furthermore, to validate earlier research, we rebuilt RNN-based methods employing LSTM and GRU models, as well as Transformer.

3.3 Training model

Six different models have been carefully created and optimized with the sole intention of generation questions. Each of these meticulously designed and refined versions has distinct qualities and abilities. A list of all these models is provided below:

- BiLSTM: Recurrent neural network architectures like Long Short-Term Memory (LSTM) are well known for their capacity to handle sequential data and preserve context over extended periods of time.
- BiGRU: Gated Recurrent Unit is a gating mechanism that is integrated into recurrent neural networks. It is another important actor in the field of autonomous question production.
- Transformer: A revolutionary neural network architecture, the Transformer has transformed several tasks, including natural language processing. Because it uses self-attention processes to gather contextual information, it is incredibly good at figuring out the dependencies and correlations between input sequences. The Transformer has played a significant role in the field of automatic question generation by generating context-aware inquiries in a variety of languages and domains.
- IndoBERT: Specifically developed to perform very well in Indonesian, IndoBERT is a language model. It is built around the BERT (Bidirectional Encoder Representations from Transformers) architecture and is specifically designed to understand and generate questions in the context of the Indonesian language.
- IndoGPT: Carefully trained and tuned for the Indonesian language, IndoGPT is a language model. By utilizing the GPT (Generative Pre-trained Transformer) architecture.
- IndoBART: An enhanced model based on the BART (Bidirectional and Auto-Regressive Transformers) architecture, designed to consider the subtleties of the Bahasa Indonesian.

3.4 Evaluation

The generated data set serves as a valuable instrument for conducting a comprehensive evaluation of the efficacy of the generation question system. These metrics are essential for assessing the system's ability to generate queries and responses accurately, as they provide insight into the system's overall performance and quality.

- The ROUGE-L metric is employed to assess the quality of text produced by a text summarization or machine translation model that emulates the behavior of natural language processing. Its principal objective is to assess the recall of the generated text, with a particular emphasis on the degree of similarity between the reference text and the anticipated text. The "L" in ROUGE-L represents "Longest," and the recall is computed using a most extended common subsequence-based approach. As the longest possible word sequence, it is referred to as the "longest" common subsequence. A subsequence that occurs frequently in both the reference and predicted texts consists of a collection of terms [45].
- The effectiveness of the model in recalling or replicating the content from the reference text is assessed by ROUGE-L, which quantifies the percentage of words in the reference text that can be matched by the predicted text. ROUGE-L fundamentally facilitates our understanding of the degree to which the generated text reproduces the vocabulary and concepts found in the source text. As the generated text more closely resembles the reference in terms of vocabulary and substance, the ROUGE-L score increases. This score indicates a greater level of recall and similarity. This metric holds significant importance in assessments pertaining to tasks such as summarization or machine translation, where the accuracy and integrity of the source text must be maintained.
- The BLEU (Bilingual Evaluation Understudy) metric is utilized to assess the quality of machine-generated text, encompassing translations and summaries. Its primary purpose is to evaluate the generated text's accuracy relative to the reference text. A comparison is made between terms present in the candidate (generated) text and those present in the reference (human-authored) text by BLEU. BLEU has the capability to be configured to consider a wide range of "n-grams," which are textual sequences consisting of n words [46]. The designations "BLEU-1," "BLEU-2," "BLEU-3," and "BLEU-4" represent the different n-grams that were considered in the evaluation process:

BLEU-1: This metric evaluates the degree of similarity between the candidate and reference texts with respect to unigrams, or individual words. It assesses the correspondence between the individual terms in the candidate text and those in the reference text.

BLEU-2: This increases the evaluation of bigrams, which are word sequences composed of two consecutive words. The correspondence between the word pairs of the candidate text and the reference text is assessed utilizing BLEU-2.

BLEU-3: Analogous to BLEU-2, but it incorporates trigrams, which consist of word sequences composed of three adjacent words. It determines how closely the clusters of three adjacent words in the candidate text correspond to those in the reference text.

BLEU-4: This metric evaluates the precision of four-grams, which are sequences of four adjacent words. The primary focus of BLEU-4 is the correspondence between word sequences in the candidate text and those in the reference text. The numerous BLEU versions provide diverse perspectives on the quality of the generated text. A greater BLEU score indicates a more refined creation process as it signifies that the produced text exhibits a greater resemblance to the reference text with respect to the chosen n-grams.

4. RESULT AND DISCUSSION

Several datasets, including SQuAD-ID, IDK-MRC, and TyDi-QA, were employed in this work to test the performance of various models in autonomous question generating. Among the performance measures are ROUGE_L, the average BLEU score, and BLEU scores at various n-gram levels (BLEU_1, BLEU_2, BLEU_3, and BLEU_4). Table 1 displays all results for the SQuAD-ID dataset. Table 2 describes the model using the IDK-MRC dataset, while Table 3 uses the TyDi-QA dataset.

Among all n-gram levels, the Bi-GRU model exhibited superior performance compared to all other models in the SQuAD-ID dataset, as evidenced by its highest BLEU scores (44.44 for BLEU_1 and 23.57 for BLEU_2). The Transformer and Bi-LSTM models both exhibited commendable performance, with the Bi-LSTM model attaining an exceptionally high score of 34.72 on the ROUGE_L scale. In all n-gram levels, the IndoBERT model achieved the highest BLEU scores (29.77 for BLEU_1 and 9.45 for BLEU_2) on the IDK-MRC dataset, surpassing the performance of the other models. In addition to its strong performance on this dataset, the Transformer model demonstrated its adaptability to a variety of question-generation tasks.

The Bi-LSTM model achieved the highest BLEU_1 score (32.75 on the TyDi-QA dataset), demonstrating its ability to generate inquiries. In contrast, the Bi-GRU model and IndoBERT exhibited competitive performance with regard to this dataset. Upon comparison across all datasets, the IndoBERT model demonstrated consistent excellent performance and emerged as a dependable solution for autonomous question generation.

The aforementioned results furnish crucial insights into the advantages and disadvantages of different approaches to automated question generation, enabling professionals and scholars to discern the most effective methodology for their particular undertaking. On the contrary, the consistent low BLEU scores of IndoBART suggest that it may be advantageous to utilize this model for tasks involving the generation of Bahasa Indonesia questions. The results of this study offer valuable insights into the comparative merits and demerits of each model across a range of linguistic contexts.

It is noteworthy that the BiLSTM, BiGRU, and Transformer models achieve zero scores in both BLEU_3 and BLEU_4 across all datasets. In BLEU_2, BLEU_3, and BLEU_4, we additionally observed that the model constructed using the IDK-MRC and TyDi-QA datasets received a score of zero. The presented evidence implies that in comparison to alternative models, its ability to capture context dependencies and higher-order linguistic nuances might be diminished.

The lack of fundamental model contribution in BLEU_2, BLEU_3, and BLEU_4 necessitates further inquiry regarding question generation. The performance of the model in generating reference query-corresponding bigrams, trigrams, and four-grams is evaluated using these BLEU metrics.

The inability of the base model to achieve nonzero scores in these criteria indicates that it will be time-consuming to develop queries with intricate language structures and contextual dependencies. In contrast to its exceptional performance in BLEU_1 and ROUGE-L metrics, the base model exhibits deficiencies in higher-order BLEU metrics, suggesting its inability to encompass the complete spectrum of reference queries. The IndoBART, IndoBERT, and IndoGPT models, on the other hand, exhibit superior performance with nonzero scores for every BLEU criterion. The pre-trained models exhibit enhanced comprehension of the diverse levels of linguistic intricacy that are intrinsic to Bahasa Indonesian.

Table 1. Performance metrics for automatic question generation on SQuAD-ID dataset

Model	BLEU_1	BLEU_2	BLEU_3	BLEU_4	ROUGE_L	BLEU_Avg
Bi-LSTM	27.27	16.51	0	0	34.72	10.94
Bi-GRU	44.44	23.57	0	0	51.15	17
Transformer	30	18.26	0	0	36.45	12.06
IndoBART	19.91	3.78	0.35	0.01	22.31	6.01
IndoBERT	21.72	4.95	0.84	0.19	25.51	6.92
IndoGPT	21.3	4.88	0.84	0.16	24.62	6.79

Table 2. Evaluation scores of models on IDK-MRC dataset for question generation

Model	BLEU_1	BLEU_2	BLEU_3	BLEU_4	ROUGE_L	BLEU_Avg
Bi-LSTM	16.37	0	0	0	17.89	4.09
Bi-GRU	16.67	0	0	0	16.67	8.18
Transformer	30.33	0	0	0	38.61	7.58
IndoBART	22.88	3.76	2.3	0.59	25.56	7.38
IndoBERT	29.77	9.45	3.53	1.38	33	11.03
IndoGPT	18.61	4.76	0.94	0.38	21.97	6.17

Table 3. Comparative analysis of question generation models using TyDi-QA dataset

Model	BLEU_1	BLEU_2	BLEU_3	BLEU_4	ROUGE_L	BLEU_Avg
Bi-LSTM	32.75	0	0	0	35.78	8.18
Bi-GRU	28.57	0	0	0	31.2	7.14
Transformer	16.67	0	0	0	20.75	4.16
IndoBART	19.51	1.85	0.43	0.04	22.27	5.46
IndoBERT	24.43	4.84	1.09	0.34	27.31	7.68
IndoGPT	23.11	6	0.98	0.41	26.62	7.63

Table 4. Task sample predictions and model performance in automatic question generation - SQuAD-ID dataset

	Sample Prediction – SQuAD-ID	English Translation
Paragraph	Masalah fungsi adalah masalah komputasi di mana output.	A function problem is a computational problem in which the output.
Target Question	Apa yang dimaksud dengan solusi fungsi?	What is meant by function solution?
Answer	Masalah komputasi.	Computational problem.
Bi-LSTM	Apa yang mungkin disebabkan oleh pemilik atau penolakan efektif untuk meningkatkan harga dan catatan?	What may be due to the owner or effective refusal to increase the prices and records?
Bi-GRU	Apa yang bisa dilakukan dengan akses ke NPO?	What can be done with access to NPO?
Transformer	Apa yang dilakukan oleh sistem BeiDou?	What does the BeiDou system do?
IndoBART	Apa yang tidak pernah digunakan untuk membuat katalog Kobylanska?	What has never been used to catalog Kobylanska?
IndoBERT	Apa yang tidak bisa dilakukan oleh beberapa cara yang lebih tinggi?	What can't some higher means do?
IndoGPT	Apa yang tidak pernah digunakan untuk menentukan sistem?	What has never been used to define a system?

Table 5. Task sample predictions and model performance in automatic question generation – IDK-MRC dataset

	Sample Prediction – IDK-MRC	English Translation
Paragraph	Luas wilayah Dinasti Qing pada masa puncaknya pernah mencapai 12 juta kilometer persegi.	The area of the Qing Dynasty at its peak reached 12 million square kilometers.
Target Question	Berapakah luas kekuasaan Dinasti Qing?	How extensive was the Qing Dynasty?
Answer	12 juta kilometer persegi	12 million square kilometers
Bi-LSTM	Kapan Dinasti Tang berdiri?	When was the Tang Dynasty founded?
Bi-GRU	Apa yang dimaksud dengan frekuensi?	What is the meaning of Frequency?
Transformer	Berapa luas kekaisaran?	How big was the empire?
IndoBART	Berapa luas kekuasaan?	How extensive is the power?
IndoBERT	Berapa luas kekaisaran Tang?	How large was the Tang empire?
IndoGPT	Berapa luas kekuasaan Tang?	How extensive was Tang power?

Table 6. Task sample predictions and model performance in automatic question generation – TyDi-QA dataset

	Sample Prediction – TyDi-QA	English Translation
Paragraph	Ekologi adalah ilmu yang mempelajari interaksi antara organisme dengan lingkungannya dan yang lainnya.	Ecology is a science that studies the interactions between organisms and their environment and others.
Target Question	Apa yang dimaksud dengan ekologi?	What is meaning of ecology?
Answer	Ilmu yang mempelajari interaksi antara organisme dengan lingkungannya dan yang lainnya.	Science that studies interactions between organisms and their environment and others.
Bi-LSTM	Apa yang dimaksud dengan ilmu ekonomi ?	What is meaning of economics?
Bi-GRU	Apa itu ilmu kimia?	What is chemistry?
Transformer	Kapan istilah Windows pertama kali ditemukan?	When was the term Windows first coined?
IndoBART	Apa yang dimaksud dengan atom?	What is the meaning of atom?
IndoBERT	Apa yang dimaksud dengan ilmu?	What is the meaning of knowledge?
IndoGPT	Apa yang dimaksud dengan bahasa?	What is the meaning of language?

The automatic question generating algorithms (Bi-LSTM, Bi-GRU, Transformer, IndoBART, IndoBERT, and IndoGPT) produce questions based on sample predictions in the SQuAD-ID dataset with varying degrees of success (Table 4). By often generating queries regarding possible causes or explanations, the Bi-LSTM model exhibits its capacity to probe into underlying characteristics. Bi-GRU, on the other hand, exhibits its versatility by being particularly adept at producing inquiries about possibilities or actions that necessitate unique access. The Transformer model's understanding of operational difficulties is demonstrated by its emphasis on querying system functionality. IndoBART exhibits its interest in historical cataloging by challenging the use or lack thereof of specific methodologies. The concerns posed by IndoBERT revolve around the limitations or impossibility of specific higher-level techniques, suggesting a deep knowledge of restrictions. IndoGPT generates questions about unused methods in defining systems, showcasing its exploration of unique perspectives.

The models' question generating performance is assessed using example predictions regarding the Qing Dynasty from the IDK-MRC dataset (Table 5). By inquiring about historical foundation dates, Bi-LSTM successfully exhibits its historical context proficiency. The questions about word meanings in Bi-GRU illustrate the emphasis on semantic comprehension. Transformer creates queries about empire size, demonstrating that it is aware of spatial issues. Because IndoBART and IndoBERT stress geopolitical matters, they frequently express reservations about the scope of authority. IndoGPT puts its historical knowledge to the test by asking about the size of the Tang empire.

The TyDi-QA dataset is used to examine the models' capacity to create questions about ecology and other scientific ideas, as shown in Table 6. When asked to characterize economic science, Bi-LSTM reveals that it comprises a wide range of scientific disciplines. Bi-GRU emphasizes its focus on specific scientific principles by providing a question about the meaning of chemistry. The Transformer model analyzes the significance of the name "Windows," exhibiting its command of antiquated jargon. IndoBART invites a question about defining the term "atom," emphasizing its emphasis on fundamental scientific notions. IndoBERT offers concerns about the notion of general science, reflecting its broad scope. Finally, IndoGPT demonstrates its exploration of linguistic concepts by asking questions regarding the meaning of language.

This study also looks at how to determine the difficulty of an inquiry by using factors like grammatical complexity, contextual comprehension requirements, and notion abstraction level. Several measures, including sentence length, word diversity, and grammatical complexity, help to quantify the difficulty of the questions. Furthermore, the feasibility of producing variations of related questions with continuous difficulty levels is studied using strategic approaches such as phrase structure changes, synonym usage, word order changes, and contextual emphasis. This ensures a diverse selection of questions while maintaining a consistent level of difficulty. When assessing the alignment between generated questions and expected standards, metrics such as BLEU and ROUGE are employed to evaluate similarity.

5. CONCLUSIONS

To avoid biases and mistakes, a naturally labeled Indonesian QA or AQ dataset and extensive data pretreatment are required for developing an efficient Indonesian Automatic Question Generation (AQG) system. Improving model performance requires optimizing settings, experimenting with new methodologies, and fine-tuning hyperparameters. Assessing the system using many metrics provides a more complete picture. While there is a risk of bias and unnatural data, using the machine-translated SQuAD into SQuAD-ID dataset can yield acceptable results. In addition, we use TyDi QA and IDK-MRC to compare our experiment to another dataset. The OpenNMT implementation proves to be more efficient, and model improvements significantly enhance speed. The best models generate questions that native Indonesians find suitable and valuable.

To automate tasks connected to question generation, a variety of language models with varying numbers of parameters might be examined. The performance of the language model can be enhanced by optimizing various hyperparameter configurations, and better results can be obtained by fine-tuning the hyperparameter tuning technique. Furthermore, taking into account a range of assessment measures may provide a more comprehensive understanding of the model's capabilities in a larger context. It's worth noting that the revised SQuAD dataset from previous studies still has a lot of room for improvement.

Further study might conduct a full evaluation using a combination of automatic and expert assessment, with criteria such as relevance, coherence, and conformity with the defined level of difficulty taken into account. When these strategies are coupled, they contribute to a comprehensive understanding of the acceptability of the generated questions in relation to the expected criteria.

ACKNOWLEDGMENT

Our research is supported by the 2023 "Penelitian Internasional Binus" grant (029/VRRTT/III/2023) from Bina Nusantara University, and we are grateful to them. The School of Computer Science at Bina Nusantara University and UNSW Sydney in Australia provided support for conducting all experiments in this research.

REFERENCES

[1] Kumar, G., Banchs, R.E., D'Haro, L.F. (2015). Revup: Automatic gap-fill question generation from educational texts. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 154-161. https://doi.org/10.3115/v1/W15-0618

[2] Tarek, A.I.T., El Hajji, M., Youssef, E.S., Fadili, H. (2022). Towards highly adaptive edu-chatbot. Procedia Computer Science, 198: 397-403. https://doi.org/10.1016/j.procs.2021.12.260

[3] Adio, A., Simeon, A. (2023). A linguistic variable of product-related question answering review system. Systems and Soft Computing, 5: 200047. https://doi.org/10.1016/j.sasc.2022.200047

[4] Shen, S., Li, Y., Du, N., Wu, X., Xie, Y., Ge, S., Yang, T., Wang, K., Liang, X.Z., Fan, W. (2020). On the generation of medical question-answer pairs. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(5): 8822-8829. https://doi.org/10.1609/aaai.v34i05.6410

[5] Al-Rahmi, W., Aldraiweesh, A., Yahaya, N., Kamin, Y. B., Zeki, A.M. (2019). Massive Open Online Courses (MOOCs): Data on higher education. Data in Brief, 22: 118-125. https://doi.org/10.1016/j.dib.2018.11.139

[6] Liu, S., Zhang, X., Zhang, S., Wang, H., Zhang, W. (2019). Neural machine reading comprehension: Methods and trends. Applied Sciences, 9(18): 3698. https://doi.org/10.3390/app9183698

[7] 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): 5. https://doi.org/10.1186/s41039-021-00151-1

[8] Song, L., Wang, Z., Hamza, W., Zhang, Y., Gildea, D. (2018). Leveraging context information for natural question generation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, pp. 569-574. https://doi.org/10.18653/v1/N18-2090

[9] Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., Zhou, M. (2018). Neural question generation from text: A preliminary study. In Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6, pp. 662-671. https://doi.org/10.1007/978-3-319-73618-1_56

[10] Kumar, V., Ramakrishnan, G., Li, Y.F. (2018). Putting the horse before the cart: A generator-evaluator framework for question generation from text. arXiv preprint arXiv:1808.04961. https://doi.org/10.48550/arXiv.1808.04961

[11] Wen, J., Jiang, D., Tu, G., Liu, C., Cambria, E. (2023). Dynamic interactive multiview memory network for emotion recognition in conversation. Information Fusion, 91: 123-133. https://doi.org/10.1016/j.inffus.2022.10.009

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.1109/2943.974352

[13] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[14] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pre-training. Work in Progress, 1-12.

[15] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv Preprint arXiv:1910.13461. https://doi.org/10.48550/arXiv.1910.13461

[16] Rogers, A., Gardner, M., Augenstein, I. (2023). QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. ACM Computing Surveys, 55(10): 1-45. https://doi.org/10.1145/3560260

[17] Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R., Sindhi, S. (2018). Online education: Worldwide status, challenges, trends, and implications. Journal of Global Information Technology Management, 21(4): 233-241. https://doi.org/10.1080/1097198X.2018.1542262

[18] Chali, Y., Hasan, S.A. (2015). Towards topic-to-question generation. Computational Linguistics, 41(1): 1-20. https://doi.org/10.1162/COLI_a_00206

[19] Ahmadnia, B., Dorr, B. (2019). Enhancing phrase-based statistical machine translation by learning phrase representations using long short-term memory network. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 25-32. https://doi.org/10.26615/978-954-452-056-4_004

[20] Du, X., Shao, J., Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. arXiv preprint arXiv:1705.00106. https://doi.org/10.48550/arXiv.1705.00106

[21] Upadhya, B.A., Udupa, S., Kamath, S.S. (2019). Deep neural network models for question classification in community question-answering forums. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-6. https://doi.org/10.1109/ICCCNT45670.2019.8944861

[22] Li, S., Gong, B. (2021). Word embedding and text classification based on deep learning methods. In MATEC Web of Conferences, 336: 06022. https://doi.org/10.1051/matecconf/202133606022

[23] Wang, Z., Hamza, W., Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. arXiv preprint arXiv:1702.03814. https://doi.org/10.48550/arXiv.1702.03814

[24] Serban, I.V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., Bengio, Y. (2016). Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. arXiv preprint arXiv:1603.06807. https://doi.org/10.48550/arXiv.1603.06807

[25] Du, X., Cardie, C. (2018). Harvesting paragraph-level question-answer pairs from Wikipedia. arXiv preprint arXiv:1805.05942. https://doi.org/10.48550/arXiv.1805.05942

[26] Ghanem, B., Coleman, L.L., Dexter, J.R., von der Ohe, S.M., Fyshe, A. (2022). Question generation for reading comprehension assessment by modeling how and what to ask. arXiv preprint arXiv:2204.02908. https://doi.org/10.48550/arXiv.2204.02908

[27] Sun, X., Liu, J., Lyu, Y., He, W., Ma, Y., Wang, S. (2018). Answer-focused and position-aware neural question generation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 3930-3939. https://doi.org/10.18653/v1/D18-1427

[28] Willis, A., Davis, G., Ruan, S., Manoharan, L., Landay, J., Brunskill, E. (2019). Key phrase extraction for generating educational question-answer pairs. In Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale, pp. 1-10. https://doi.org/10.1145/3330430.3333636

[29] Liu, B., Zhao, M., Niu, D., Lai, K., He, Y., Wei, H., Xu, Y. (2019). Learning to generate questions by learning what not to generate. In the World Wide Web Conference, San Francisco, CA, USA, pp. 1106-1118. https://doi.org/10.1145/3308558.3313737

[30] Zhao, Y., Ni, X., Ding, Y., Ke, Q. (2018). Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 3901-3910. https://doi.org/10.18653/v1/D18-1424

[31] Nema, P., Mohankumar, A.K., Khapra, M.M., Srinivasan, B.V., Ravindran, B. (2019). Let's ask again: Refine network for automatic question generation. arXiv preprint arXiv:1909.05355. https://doi.org/10.48550/arXiv.1909.05355

[32] Ma, X., Zhu, Q., Zhou, Y., Li, X. (2020). Improving question generation with sentence-level semantic matching and answer position inferring. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(5): 8464-8471. https://doi.org/10.1609/aaai.v34i05.6366

[33] Vu, N., Van Nguyen, K. (2022). Enhancing vietnamese question generation with reinforcement learning. In Asian Conference on Intelligent Information and Database Systems, pp. 559-570. https://doi.org/10.1007/978-3-031-21743-2_45

[34] Liu, B. (2020). Neural question generation based on Seq2Seq. In Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence, pp. 119-123. https://doi.org/10.1145/3395260.3395275

[35] Bashath, S., Perera, N., Tripathi, S., Manjang, K., Dehmer, M., Streib, F.E. (2022). A data-centric review of deep transfer learning with applications to text data. Information Sciences, 585: 498-528. https://doi.org/10.1016/j.ins.2021.11.061

[36] Dehghani, M., Azarbonyad, H., Kamps, J., de Rijke, M. (2019). Learning to transform, combine, and reason in open-domain question answering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 681-689. https://doi.org/10.1145/3289600.3291012

[37] Lopez, L.E., Cruz, D.K., Cruz, J.C.B., Cheng, C. (2021).

Simplifying paragraph-level question generation via transformer language models. In PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18, pp. 323-334. https://doi.org/10.1007/978-3-030-89363-7_25

[38] Kumar, A., Kharadi, A., Singh, D., Kumari, M. (2021). Automatic question-answer pair generation using deep learning. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 794-799. https://doi.org/10.1109/ICIRCA51532.2021.9544654

[39] Akyon, F.C., Cavusoglu, D., Cengiz, C., Altinuc, S.O., Temizel, A. (2021). Automated question generation and question answering from Turkish texts. arXiv preprint arXiv:2111.06476. https://doi.org/10.48550/arXiv.2111.06476

[40] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140): 1-67.

[41] Ngo, H.Q., Nguyen, H.D., Le-Khac, N.A. (2023). Building legal knowledge map repository with NLP toolkits. In Conference on Information Technology and its Applications, pp. 25-36. https://doi.org/10.1007/978-3-031-36886-8_3

[42] Sayed, M.A., Braşoveanu, A.M., Nixon, L.J., Scharl, A. (2023). Unsupervised topic modeling with BERTopic for coarse and fine-grained news classification. In International Work-Conference on Artificial Neural Networks, pp. 162-174. https://doi.org/10.1007/978-3-031-43085-5_13

[43] Muis, F.J., Purwarianti, A. (2020). Sequence-to-sequence learning for Indonesian automatic question generator. In 2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA), pp. 1-6. https://doi.org/10.1109/ICAICTA49861.2020.9429032

[44] Vincentio, K., Suhartono, D. (2022). Automatic question generation monolingual multilingual pre-trained models using RNN and transformer in low resource Indonesian language. Informatica, 46(7): 103-118. https://doi.org/10.31449/inf.v46i7.4236

[45] Lin, C.Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out, Barcelona, Spain, pp. 74-81. https://aclanthology.org/W04-1013

[46] Papineni, K., Roukos, S., Ward, T., Zhu, W.J. (2002). 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. https://doi.org/10.3115/1073083.1073135