Automatic Question Generation using RNN-based and Pre-trained Transformer-based Models in Low Resource Indonesian Language

Karissa Vincentio and Derwin Suhartono
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, 11530, Indonesia
E-mail: felicia.vincentio@binus.ac.id, dsuhartono@binus.edu

Keywords: natural language processing, natural language generation, automatic question generation, recurrent neural network, long-short term memory, gated recurrent unit, transformer, fine-tuning

Received: June 14, 2022

Although Indonesian is the fourth most frequently used language on the internet, the development of NLP in Indonesian has not been studied intensively. One form of NLP application classified as an NLG task is the Automatic Question Generation task. Generally, the task has proven well, using rule-based and cloze tests, but these approaches depend heavily on the defined rules. While this approach is suitable for automated question generation systems on a small scale, it can become less efficient as the scale of the system grows. Many NLG model architectures have recently proven to have significantly improved performance compared to previous architectures, such as generative pre-trained transformers, text-to-text transfer transformers, bidirectional autoregressive transformers, and many more. Previous studies on AQG in Indonesian were built on RNN-based architecture such as GRU, LSTM, and Transformer. The performance of models in previous studies is compared with state-of-the-art models, such as multilingual models mBART and mT5, and monolingual models such as IndoBART and IndoGPT. As a result, the fine-tuned IndoBART performed significantly higher than either BiGRU and BiLSTM on the SQuAD dataset. Fine-tuned IndoBART on most of the metrics also performed better on the TyDiQA dataset only, which has fewer population than the SQuAD dataset.

Povzetek: Za indonezijščino, četrti najpogostejši spletni jezik, so za poučevanje razvili jezikovni pretvornik iz besedila v vprašanja.

1 Introduction

The current education system requires a process to efficiently evaluate students' understanding of lessons by reading a text's content [1]. Preparation of questions carried out by students can consume much time, while getting questions from external sources such as collections of questions makes it possible that they are irrelevant to the content studied by students [2]. In addition, questions designed to evaluate students' understanding of textual reading can also be influenced by their effectiveness, which can be seen from the development of various strategies in preparing questions [3]. From the emergence of these problems, many techniques have been investigated in the Question Generation process based on content, generally known as the Automatic Question Generation (AQG) system based on NLP in the NLG branch from various approaches from rule-based to attention-based models [4].

AQG is a job that automatically generates queries from various inputs such as original text, databases, or semantic representations [5]. From this understanding, the input type can take the form of various forms such as sentences, paragraphs, and poetry [6]. AQG has various applications such as healthcare systems, automated help systems, chatbot systems, and other AQG applications [7]. In this paper, AQG is the subject of research that requires text or related information to be processed using a sequence-to-sequence approach, namely Bidirectional Gated Recurrent Unit (BiGRU), Bidirectional Long Short Term Memory (BiLSTM), and Transformer architectures, as well as using the pre-trained fine-tuning approach of the mBART and mT5 architectural models.

1.1 AQG in English

Cohen first proposed AQG in 1929 to represent a question's content in a formula with one or more independent variables [8]. Since then, researchers have become interested in developing AQG in education for educational purposes, mainly because asking questions during teaching encourages students to understand what they are learning. One of the AQGs that Wolfe proposed supported learning in 1976 [9].

Recent AQG work showed that leveraging linguistic representation approaches such as Part Of Speech (POS) and Named Entity Recognition (NER) through deep neural networks based on Bidirectional Encoder Representations from Transformers (BERT) can achieve state-of-the-art results. The model architecture consists of a two-layer bidirectional Long Short-Term Memory (LSTM) encoder and a two-layer unidirectional LSTM decoder. The bidirectional LSTM encoder has been used for producing sequences of hidden states, and the unidirectional LSTM decoder has then used the representation to generate words [10].

Another recent work was fine-tuning a miniature version of a T5 transformer language model consisting of 220 million parameters using the SQuADv1.1 dataset, which contains 100,000 question-answer pairs. In order to generate questions, the model was trained by receiving the passage, and the 30% probability of the answer was replaced with the [MASK] token [11]. Some open English question-answer pair datasets can be leveraged for transfer learning approaches in creating AQG systems [12, 13, 14, 15, 16].

1.2 AQG in Indonesian

Various researches on AQG based on NLP have been conducted [17], but not many of them are observed in Indonesian. One study [18] that was conducted in Indonesian built a language model that utilizes a sequence-to-sequence approach and is trained on the SQuAD v2 [19] as well as TyDiQA [20] dataset, which has been translated into Indonesian using the Google Translate API v2 to the model with the Transformer architecture along with Recurrent Neural Network (RNN) such as BiLSTM & BiGRU [21]. This study found that the questions generated using BiLSTM and BiGRU were not significantly different. Meanwhile, the use of Transformers found difficulties in understanding the semantic context of the information provided [22].

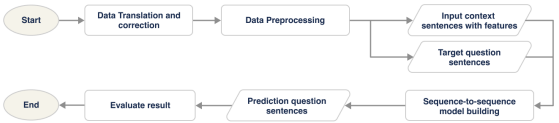

Figure 1: Related Research Modeling Process Diagram [18]

1.3 Models Benchmark in NLG Tasks

In this research, benchmark resources [23, 24] have not been involved in the results of the model studied for the Question Generation task in Indonesian [18]. Indonesian benchmark resources such as IndoNLU [25] & IndoNLG [4] can play a significant role in comparisons and literature reviews by other researchers so that they can be a reference in developing Automatic Question Generation that is more reliable and by the information provided in the form of textual or reading content.

In addition, because the IndoNLU benchmark only covers NLU tasks in Indonesian, such as sentiment analysis [26, 27, 28, 29, 30], which is similar to the GLUE [31] benchmark for English Natural Language Understanding (NLU) tasks, while the Question Generation task is an NLG task, the GEM benchmark, which is a benchmark for various NLG tasks including Question Generation, should also be applied [32]. The resources of the GEM benchmark have selected and processed the most common dataset for the available NLG tasks. GEM also conducted baseline modeling using language models such as BART, T5 in one language, and the mBART model, mT5 for multilingual languages. Then GEM also provides a testbed in automated evaluations, including metrics according to the task. The GEM benchmark feature, which is regularly updated, makes it easier for researchers in other fields of NLG to compare the model he built with previously developed models [32].

Recently, many NLG model architectures have proven to have significantly improved performance compared to previous architectures, such as generative pre-trained transformers, text-to-text transfer transformers, bidirectional auto-regressive transformers, and many more. Meanwhile, the previous study that raised AQG in Indonesian was only carried out on the GRU, LSTM, and Transformer methods [18]. In this study, to develop research on AQG assignments in Indonesian, the performance of models in previous studies was compared with state-of-the-art models, such as

mBART and mT5.

Although Indonesian is the fourth most frequently used language on the internet, the development of NLP in Indonesian has not been studied intensively [25]. The automatic question generation process is one form of NLP application classified as an NLG task [18]. Generally, the task has proven well, using approaches such as rule-based and cloze tests, but these approaches depend heavily on the set of rules that have been created. So this approach is suitable for automated question generation systems on a small scale and can become less efficient as the scale of the system grows [33]. In this context, deep learning approaches, especially NLP, have better generalizations than rule-based approaches [34]. Although the deep learning approach is relatively highly complex, the system can construct its rules and evolve coherently to adapt its dataset if adequately trained and properly configured [35].

In the previous related research [18], the Stanford question-and-answer dataset (SQuAD v2 [36]), which consists of 536 articles with 161,550 collections of question-answer pairs in English, underwent translation and pre-processing into Indonesian and followed by improving some of the translations by using fuzzy string-matching to look for inconsistent translations, which then can be used for model training as well as model evaluation [18]. Language models based on RNN architecture such as GRU and LSTM in a bidirectional manner and language models based on transformer architecture with a sequence-to-sequence learning approach [18].

Several adaptations were made to the language model on the RNN-based (BiGRU and BiLSTM) and transformer from scratch, such as the use of several linguistic features (Ans, Case, POS, Named Entity (NE)), and the presence of a sentence embedding encoder. This research aims to measure how well the language model based on the RNN architecture and the state-of-the-art transformer-based models performs the question generation task in Indonesian [18]. Then, a validation process was followed by testing the model using the SQuAD dataset as it is the validation set to see how it performs on the same behavior dataset and followed by the evaluation on the TyDiQA dataset that is built naturally from Indonesian [20]. Overall flow can be seen in Figure 1.

2 Deep Learning Methods

2.1 RNN Based Models

RNN is a widely used neural network architecture for NLP, which has been proven to be relatively accurate and efficient for developing language models as well as in tasks of speech recognition. Essentially RNN uses what is known as feedback loops which allow the input sequence to be shared to different nodes as well as allowing RNN to have an internal memory that can help RNN generate predictions based on previous inputs. As much NLP research progresses from time to time, there are many novelty techniques, one of which showed that bidirectionally processing the input sequence can achieve a better understanding of the context [21]. Visualization for each model can be seen in Figure 2.

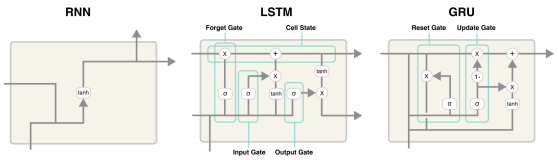

Figure 2: RNN vs. LSTM vs. GRU

In the first place, RNN was having a problem called the vanishing gradient problem, which occurs when using neural networks with gradient-based learning methods and backpropagation. GRU (Gated Recurrent Unit) was introduced to overcome this problem that utilizes an update gate and reset gate so the model can store information longer and remove irrelevant information for prediction. BiGRU is a model that uses two GRU in which one GRU will accept input by forwarding direction called forward GRU, and another will accept input by backward direction named backward GRU [21].

LSTM is an RNN enhancement that is capable of studying long-term dependencies [18]. This capability enables LSTMs to avoid long-term dependency problems. LSTM uses three gates to protect and control the cell gates: input gate, forget gate, and output gate. In this research, BiLSTM will be used as the representation of LSTM. The flow of each model in detail can be seen in Figure 3.

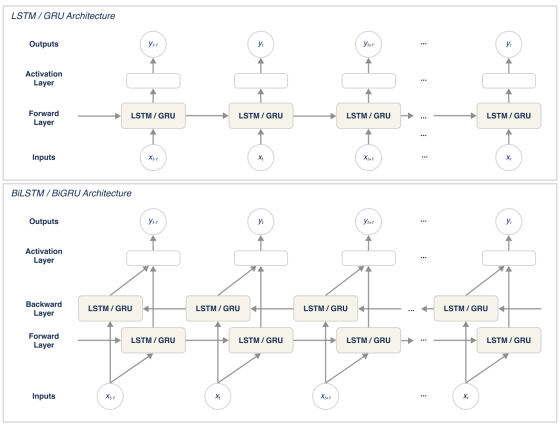

Figure 3: LSTM / GRU vs. BiLSTM / BiGRU Architecture

2.2 Transformer Based Models

BART (Bidirectional and Auto-Regressive Transformer) is a language model that is pre-trained by applying noise or corruption to the input sequence, and then the model is assigned to reconstruct the actual input sequence [33]. After that, the results of the model predictions will be calculated against the loss function generally in the form of cross-entropy and followed by the back-propagation gradients process and updating the model weights. A comparison between RNN with transformer architecture can be seen in Figure 4.

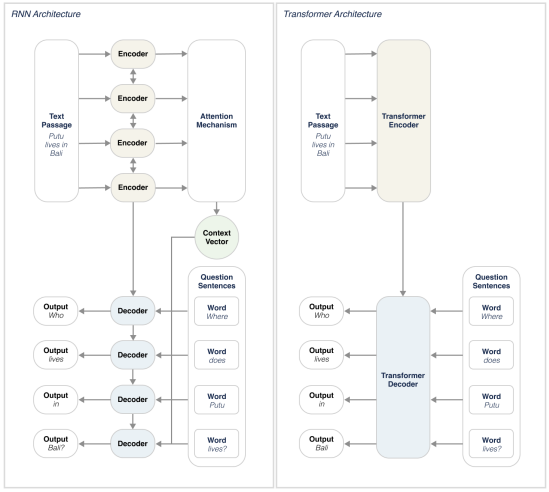

Figure 4: RNN vs. Transformer Architecture

The BART language model architecture utilizes an encoder (see Figure 5) on BERT (Bidirectional Encoder Representations from Transformers) [37] and

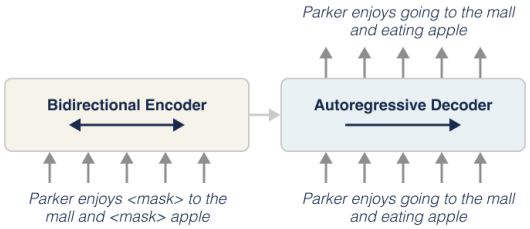

Figure 5: Encoder-decoder Illustration

a decoder on GPT (Generative Pre-Trained Transformer) [38] capable of performing NLP tasks in the form of NLU and NLG. mBART is a language model modified from the BART model, which utilizes autoencoder denoising and a sequence-to-sequence pre-trained model. mBART model was trained once using a dataset of multiple languages that could be customized in a fine-tuning process [39].

The second one, T5 (Text-to-Text Transfer Transformer), is a pre-trained model that utilizes a unified Text-to-Text format from NLP using text [40]. By using this model, when setting the configuration with hyperparameters, then it will be applied to another task.

The third one, GPT leverages what is known as masked self-attention, where it masks future tokens and only knows the present and the previous tokens. GPT works autoregressively by adding generated tokens to the input sequence, and that particular new sequence then will be used as the input to the model in its next step.

3 Materials and Methods

The approach that is going to be used in this research is transfer learning, which is the approach where a model is first pre-trained on a data-rich task before being fine-tuned on a downstream task [40]. The research is divided into four main steps; the first step (planning phase) is to identify the problem, followed by the dataset preprocessing phase, which was done to preprocess the dataset in a configured format so that the dataset can be forwarded to train the language model. The third step is to train the model, which leverages several preconfigured language models such as BiGRU and BiLSTM. As for the pre-trained language models like mBART and mT5 will require further fine-tuning.

3.1 Planning

In the first step, previous research regarding automatic question generators using BiGRU and BiLSTM in Indonesian was reviewed and evaluated. Both of those models are based on recurrent neural network architecture. This research explores multilingual language models based on the transformer (encoder-decoder) architecture such as mBART and mT5 and monolingual language models such as IndoNLG's IndoBART and IndoGPT [4].

3.2 Dataset Preprocessing

The dataset we are using is preprocessed SQuAD dataset that has been translated to Indonesian from the dataset itself for all the models, resulting in 102,657 training data, 11,407 validation data, and 10,567 testing data for SQuAD, and 550 testing data for TyDiQA. As for the sequence-to-sequence language model, the preprocessed SQuAD dataset was first added with special tokens.

Since some of the SQuAD question-answer pairs Indonesian dataset translation might be misleading from its true meaning because the translation process for the passage and the context were done separately, some corrections need to be made. The correction process was done by leveraging fuzzy string matching to search the translated question-answer pairs for inconsistent translations thoroughly. As long as the answer is found, it will update the start position of the answer, whereas if the answer is not found, then the start position of the answer will be set to negative one (-1) and removed.

In this step, we preprocessed the enhanced SQuAD dataset from previous research [36] by reusing some of the main dataset attributes (context/passage, question, and answer) that are going to be used by the model for training, excluding some of the linguistic features such as part of speech (POS) and named entity (NE) attributes that will be used only by BiGRU, BiLSTM, and Transformer model. For mBART, the input encoder structure will be formatted to `<context><sep><answer><eos><langid>`, and decoder `<langid><question><eos>`, whereas for mT5 the input encoder will be in `<context><sep><answer>` format, as for the decoder format is going to be `<bos><question>`. As for GPT since it is a decoder only transformer language model, then the input sequence will be formatted to `<context><sep><answer><bos><question><eos>`. The flow can be seen in Figure 6.

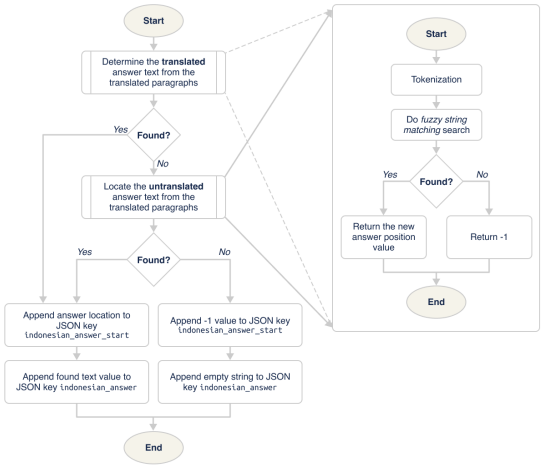

Figure 6: Process to Repair Answer Translation Result from Previous Research

3.3 Training Model

By utilizing the formatted dataset, these models were made by applying configuration from the Sequence-to-Sequence Learning method for the Indonesian Automatic Question Generator for several algorithms, BiGRU and BiLSTM. For the mBART and mT5, new fine-tuning models were made [41]. Alternately, the model training was conducted to ensure the computer uses the same resources.

3.4 Evaluation

The results from each algorithm are evaluated by using BLEU and ROUGE metrics. From that, comparison and analysis results are conducted based on the results to choose the best from all of the implemented algorithms. The overall flow of benchmarking model can be seen in Figure 7.

4 Results & Discussion

Tokenization is a way to separate a piece of text into smaller units known as tokens, which can be words, characters, or subwords. In order to fine-tune the mBART pre-trained language model, the sequence that is going to be forwarded to the model will firstly be appended with some special tokens such as language id token for the multilingual model to identify

Table 1: Related Research Rerun Result on SQuAD Test Set

Model	Dataset	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE L	Epoch
BiGRU	Cased	33.87	17.01	8.42	3.88	37.98	20
	Cased-Copy	36.03	19.37	9.57	5.41	40.96	20
	Cased-Copy-Cov	36.16	18.79	10.99	6.52	40.49	20
	Uncased	36.96	17.23	8.46	5.11	40.08	20
	Uncased-Copy	39.62	22.26	12.02	5.88	43.38	20
	Uncased-Copy-Cov	**39.56**	**21.99**	**11.34**	**5.99**	**43.41**	20
BiLSTM	Cased	32.16	14.61	7.73	3.73	38.00	10
	Cased-Copy	36.67	19.28	10.85	5.29	40.67	10
	Cased-Copy-Cov	35.86	18.69	9.21	7.07	40.27	10
	Uncased	35.45	18.19	8.87	4.63	39.48	10
	Uncased-Copy	40.60	21.35	10.93	5.73	**43.79**	10
	Uncased-Copy-Cov	**39.90**	**22.23**	**12.49**	**5.98**	43.34	10
Transformer	Cased	30.72	12.63	4.44	2.46	34.25	300
	Cased-Copy	36.14	18.81	9.52	4.75	39.58	300
	Uncased	33.34	13.58	5.86	3.38	37.71	300
	Uncased-Copy	**39.09**	**21.21**	**10.83**	**5.39**	**43.69**	300

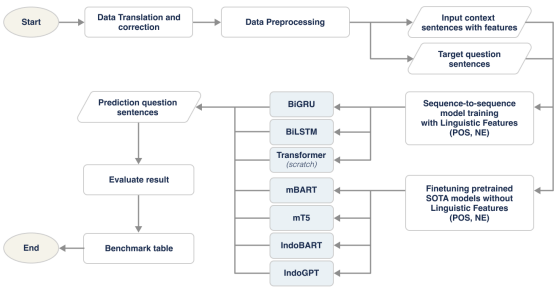

Figure 7: Models Benchmark creation process diagram.

the language, beginning of sentence token, end of sentence token, as well as separator token to be used as a separator between the answer and the question, which then can be used for the model to identify the label and the context. Then the tokenizer will tokenize the rest of the passage into token representation so that the model can understand the context. Unlike mBART, fine-tuning mT5 does not require a language id token to help the model identify the language that is supposed to be appended to the input sequence. Surprisingly when not considering the special tokens or skipping special tokens, the language model cannot parse the input sequence correctly; therefore, it performs poorly.

Table 3 and Table 4 are some samples of the generated questions in Indonesian from each of all the models on datasets SQuAD and TyDiQA. "Input Sentence & Answer" is the context or passage as the model input followed by the expected answer, and "Target Question" is the expected generated question.

RNN-based models evaluated on the TyDiQA dataset are performing lower than SQuAD dataset due to most of the text being translated on the SQuAD dataset, which consists of many faulty translations [18], while TyDiQA is in Indonesian by origin. It also applies to the transformer-based models, including those based on pre-trained multilingual and monolingual models. It can also be seen on the pre-trained models' row that the maximum score on ROUGE and BLEU on TyDiQA is up to 10 points higher than the SQuAD dataset. This evaluation on the TyDiQA test set is done to obtain a more reliable and comparable evaluation score since TyDiQA is a more natural Indonesian dataset [18].

On the RNN-based and transformer from scratch results, the TyDiQA and SQuAD do not show a significant difference in the scores, but they differ significantly on the pre-trained models, especially the monolingual models IndoBART and IndoGPT. With these monolingual models, the TyDiQA dataset that is already available in Indonesian while SQuAD is mostly

Table 2: Related Research Rerun Result on TyDiQA Test Set

Model	Dataset	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE L	Epoch
BiGRU	Cased	30.33	10.83	3.18	1.89	34.41	20
	Cased-Copy	34.05	15.01	7.47	3.12	38.15	20
	Cased-Copy-Cov	34.28	14.72	6.60	2.63	38.30	20
	Uncased	33.28	13.92	5.65	2.98	37.45	20
	Uncased-Copy	37.42	17.96	8.73	4.65	41.71	20
	Uncased-Copy-Cov	**37.78**	**18.68**	**9.15**	**5.46**	**41.92**	20
BiLSTM	Cased	30.74	12.00	4.09	1.48	34.74	10
	Cased-Copy	34.62	14.80	6.49	3.62	38.66	10
	Cased-Copy-Cov	34.13	14.57	6.40	2.84	38.17	10
	Uncased	32.92	13.02	4.81	2.28	36.98	10
	Uncased-Copy	37.63	18.38	8.73	**4.62**	42.12	10
	Uncased-Copy-Cov	**38.14**	**18.54**	**8.90**	4.55	**42.59**	10
Transformer	Cased	27.88	8.00	0.71	0.64	31.86	300
	Cased-Copy	31.95	12.43	4.63	2.27	36.37	300
	Uncased	29.39	8.62	1.12	0.52	33.27	300
	Uncased-Copy	**37.23**	**17.93**	**8.19**	**3.40**	**41.90**	300

translation proves that monolinguals can perform better on datasets within the same language.

Remembering that the finetuned pre-trained models such as mBART, IndoBART, IndoGPT, and mT5 do not use the linguistic features such as POS and NE provided in the preprocessed dataset, these models outperform the RNN-based models and the scratch models. Without extra context in the form of POS and NE, the transformer-based pre-trained models have proven that transfer learning helps the models have a better understanding than the models that do not have any base knowledge.

Generally, monolingual language models have a smaller number of parameters than multilingual language models, resulting in faster model training and smaller model size, whereas, in this research, monolingual language models were pre-trained on a large monolingual corpus (Indonesian). On the other hand, multilingual language models were pre-trained on a large multilingual corpus, hence the term multilingual. As for the performance of both languages, they only perform slightly differently [4].

Furthermore, numerous incorrect and unnatural translations, especially one of the SQuAD datasets on the input sentences and target questions, impact our model predictions. Nonetheless, semantically, those sentences were still understandable. The same projected questions were agreed upon by all models, resulting in highly identical questions. There were some variances in the verbs in the created questions, but they are all synonyms and have similar meanings.

mT5 model seems to have the lowest automatic evaluation score among other pre-trained models. However, reading directly from the generated predictions, the mT5 prediction seems to have the most fluent prediction. mT5 encoders that could affect this are based on BERT language models, known for their novelty through a bidirectional approach that was able to capture the context deeper of the input sequence [37]. These encoders also take account of the relation between words, which helps capture its meaning [42], consisting of a self-attention layer and a feedforward network to process the sequence to the decoders. As for the decoders, it is similar to the initially proposed transformer language model [43]. The decoders were leveraging the auto-regressive approach, which will be used to produce the output sequence. mT5 were pre-trained on a large multilingual corpus that covers over 100 languages [41].

The hyperparameters used for each model are configured in Table 5. As for the maximum sequence, the length hyperparameter was set to 512 for every fine-tuned language model.

Table 5 shows the configuration used in each model

Table 3: All Models AQG Task Sample Predictions 1 SQuAD

Sample Prediction 1 - SQuAD	
Input Sentence & Answer	*Samudra Pasifik atau Lautan Teduh (dari bahasa spanyol Pacifico, artinya tenang) adalah kawasan kumpulan air terbesar di dunia, serta mencakup kira-kira sepertiga permukaan Bumi, dengan luas sebesar 179,7 juta km2 (69,4 juta mi2). Panjangnya sekitar 15.500 km (9.600mi) dari Laut Bering di Arktik hingga batasan es di Laut Ross di Antartika di selatan. Samudra Pasifik mencapai lebar timur-barat terbesarnya pada sekitar 5 derajat U garis lintang, di mana ia terbentang sekitar 19.800 km (12.300mi) dari Indonesia hingga pesisir Kolombia. Batas sebelah barat samudra ini biasanya diletakkan di Selat Malaka. Titik terendah permukaan Bumi—Palung Mariana—berada di Samudra Pasifik. Samudra ini terletak di antara Asia dan Australia di sebelah barat, Amerika di sebelah timur, Antartika di sebelah selatan dan Samudra Arktik di sebelah utara.*
Answer	*179,7 juta km2*
Target Question	*Berapa luas Samudera Pasifik?*
BiGRU Uncased-Cop-Cov	*berapa luas samudra pasifik ?*
BiLSTM Uncased-Copy-Cov	*berapa luas bumi pasifik ?*
Transformer Uncased-Copy	*berapa luas air terbesar di dunia ?*
mBART-Large	*berapakah luas samudra pasifik?*
IndoBART	*berapakah luas samudra pasifik?*
IndoGPT	*berapa luas total wilayah lautan pasifik ?*
mT5-Small	*Berapa luas samudra pasifik?*

to generate the sentences and the training time needed for the SQuAD and TyDiQA datasets. The training step and valid for the mBART-L, IndoBART, IndoGPT, and mT5-Base pre-trained models are not listed because they are not explicitly defined in this modeling.

Fine-tuned mBART performed the best with the average BLEU 31.71 and ROUGE-L score of 46.27 on the SQuAD dataset (Table 6) for the Indonesian question generation task. Fine-tuned IndoBART also performed the best with an average score of BLEU 17.26 and ROUGE L score is 33.73 on the TyDiQA dataset (Table 6) for the Indonesian question generation task.

On the other hand, RNN-based and transformer from scratch results on TyDiQA and SQuAD datasets do not show a significant difference in the scores, but they differ significantly from the pre-trained models. With these monolingual models, the TyDiQA, whose origin is in Bahasa while SQuAD is mostly translation, proves that monolinguals can perform better on datasets within the same language.

5 Conclusions

Based on the results achieved in this research, language models based on transformer architecture that leverage self-attention mechanisms were able to achieve state-of-the-art results in generating questions compared to language models based on bidirectional recurrent neural network architecture such as BiLSTM and BiGRU.

This research introduces a more extensive comparison between RNN-based and transformer-based models, including the state-of-the-art variation on the Indonesian AQG system. In the previous research, it has already been proven that the Indonesian AQG system can be built using an as-is machine-translated question answering dataset (SQuAD v2.0) with acceptable results, and this research is shown that better performance can be achieved with different varieties of

Table 4: All Models AQG Task Sample Predictions 2 TyDiQA

Sample Prediction 2 - TydiQA	
Input Sentence & Answer	*Kadipaten Normandia , yang mereka bentuk dengan perjanjian dengan mahkota Prancis , adalah tanah yang indah bagi Prancis abad pertengahan , dan di bawah Richard I dari Normandia ditempa menjadi sebuah pemerintahan yang kohesif dan tangguh dalam masa jabatan feodal.*
Answer	*Kadipaten Normandia*
Target Question	*Siapa yang memerintah kadipaten Normandia*
BiGRU Uncased-Cop-Cov	*siapa yang memerintah pemerintahan normandia ?*
BiLSTM Uncased-Copy-Cov	*siapa yang mendirikan kadipaten normandia ?*
Transformer Uncased-Copy	*siapa yang memerintah normandia di normandia ?*
mBART-Large	*siapakah kadipaten normandia di bawah raja normandia ?*
IndoBART	*siapa yang memimpin normandia ?*
IndoGPT	*dengan siapa prancis membentuk kadipaten normandia ?*
mT5-Small	*Siapa yang memerintahkan kadipaten normandia?*

Table 5: Model Configuration

Model	**Dataset**	**Learning Rate**	**Training Step**	**Valid**	**Epoch**	**Batch Size**	**Training Time**
BiGRU	Uncased-Cop-Cov	1.00E-03	32.100	3.210	20	64	55m
BiLSTM	Uncased-Cop-Cov	1.00E-03	17.655	3.210	10	64	1h13m
Transformer	Uncased-Cop	1.00E+00	120.600	4.020	300	256	5h40m
mBART-L	Uncased-Large	1.00E-03	-	-	20	8	40h42m
IndoBART	Uncased-v2	1.00E-03	-	-	20	64	7h26m
IndoGPT	Uncased	1.00E-03	-	-	20	32	10h36m
mT5-Base	Uncased-Small	3.00E-05	-	-	3	4	8h4m

transformer-based models such as mBART and mT5, as well as the monolingual models built on Indonesian dataset; IndoBART.

5.1 RNN-based vs Transformer-based

Transformer-based models outperformed all the RNN-based models. As seen in Table 6 & Table 7, Transformer-based models perform better in generating natural Indonesian questions on the TyDiQA dataset, which contains 550 pairs of question-answering Indonesian.

5.1.1 Monolingual vs. Multilingual

Since monolingual language models were pre-trained using a monolingual dataset, the model resulted in a lower number of parameters, hence faster training than multilingual language models. In terms of performance, it does not differ very much from multilingual language models and monolingual language models.

5.2 Future Improvements

The system of building an Indonesian AQG can achieve better results with a more natural labeled Indonesian QA or AQ dataset. It should be followed with more robust preprocessing data to avoid syntactically incorrect data and biases. Experiments on more

Table 6: Model Evaluation Metric Performance Comparison on SQuAD Test Set

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	Average BLEU	ROUGE L	Epoch
BiGRU	39.56	21.99	11.34	**5.99**	19.72	43.41	20
BiLSTM	**39.90**	**22.23**	**12.49**	5.98	**20.08**	43.34	10
Transformer	39.09	21.21	10.83	5.39	19.13	**43.69**	300
mBART-L	53.58	32.41	**23.25**	**17.59**	**31.71**	44.70	20
IndoBART	**55.03**	**31.88**	22.27	16.42	31.40	**46.27**	20
IndoGPT	54.07	30.56	21.21	15.72	30.39	44.31	20
mT5-Base	41.13	14.92	7.16	3.86	16.77	40.51	3

Table 7: Model Evaluation Metric Performance Comparison on TyDiQA Test Set

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	Average BLEU	ROUGE L	Epoch
BiGRU	37.78	**18.68**	**9.15**	**5.46**	**17.77**	41.92	20
BiLSTM	**38.14**	18.54	8.90	4.55	17.53	**42.59**	10
Transformer	37.23	17.93	8.19	3.40	16.69	41.90	300
mBART-L	36.85	15.96	**9.56**	**6.05**	17.10	32.64	20
IndoBART	**38.65**	**16.01**	8.95	5.43	**17.26**	33.73	20
IndoGPT	35.77	12.55	6.55	3.78	14.66	28.93	20
mT5-Base	32.23	7.98	2.39	0.92	10.88	**36.10**	3

precise hyperparameters can also help improve getting the best-performing models.

Future work concerns a deeper analysis of particular mechanisms and proposals to explore different techniques. Many other language models varying in parameter count can be explored for automatic question generation tasks. Various hyperparameter configurations can be optimized for the best language model, fine-tuning results through hyperparameter tuning. Leveraging different evaluation metrics can result in much more comprehensive results to see the model's capabilities within the bigger picture. It is also worth mentioning that the enhanced SQuAD dataset from previous research still has much room for improvement.

no

References

[1] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, 11 2019. [Online]. Available: https://doi.org/10.1007/s40593-019-00186-y

[2] N.-T. Le, T. Kojiri, and N. Pinkwart, "Automatic question generation for educational applications – the state of art," *Advances in Intelligent Systems and Computing*, vol. 282, pp. 325–338, 01 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-06569-4_24

[3] J. Jamiluddin and V. Ramadayanti, "Developing Students' Reading Comprehension Through Question Generation Strategy," *e-Journal of ELTS (English Language Teaching Society)*, vol. 8, no. 1, Apr. 2020.

[4] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, "IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation," pp. 8875–8898, Nov. 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.699

[5] C. A. Nwafor and I. E. Onyenwe, "An automated multiple-choice question generation using natural language processing techniques," *International Journal on Natural Language Computing*, vol. 10, no. 02, p. 1–10, Apr 2021. [Online]. Available: http://dx.doi.org/10.5121/ijnlc.2021.10201

[6] A. D. Lelkes, V. Q. Tran, and C. Yu, "Quiz-style question generation for news stories," New York, NY, USA, p. 2501–2511, 2021. [Online]. Available: https://doi.org/10.1145/3442381.3449892

[7] A. Graesser, V. Rus, S. D'Mello, and G. Jackson, "Autotutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner," *Current Perspectives on Cognition, Learning and Instruction: Recent Innovations in Educational Technology that Facilitate Student Learning*, pp. 95–125, 01 2008.

[8] N.-T. Le, T. Kojiri, and N. Pinkwart, "Automatic question generation for educational applications – the state of art," *Advances in Intelligent Systems and Computing*, vol. 282, pp. 325–338, 01 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-06569-4_24

[9] J. H. Wolfe, "Automatic question generation from text - an aid to independent study," in *SIGCSE '76*, 1976. [Online]. Available: https://doi.org/10.1145/952989.803459

[10] W. Yuan, T. He, and X. Dai, "Improving Neural Question Generation using Deep Linguistic Representation," in *Proceedings of the Web Conference 2021*. Ljubljana Slovenia: ACM, Apr. 2021, pp. 3489–3500. [Online]. Available: https://doi.org/10.1145/3442381.3449975

[11] K. Vachev, M. Hardalov, G. Karadzhov, G. Georgiev, I. Koychev, and P. Nakov, "Leaf: Multiple-choice question generation," *CoRR*, vol. abs/2201.09012, 2022. [Online]. Available: https://doi.org/10.1007/978-3-030-99739-7_41

[12] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. H. Hovy, "RACE: large-scale reading comprehension dataset from examinations," *CoRR*, vol. abs/1704.04683, 2017. [Online]. Available: https://doi.org/10.18653/v1/d17-1082

[13] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? A new dataset for open book question answering," *CoRR*, vol. abs/1809.02789, 2018. [Online]. Available: https://doi.org/10.18653/v1/d18-1260

[14] P. Clark, O. Etzioni, D. Khashabi, T. Khot, B. D. Mishra, K. Richardson, A. Sabharwal, C. Schoenick, O. Tafjord, N. Tandon, S. Bhakthavatsalam, D. Groeneveld, M. Guerquin, and M. Schmitz, "From 'f' to 'a' on the N.Y. regents science exams: An overview of the aristo project," *CoRR*, vol. abs/1909.01958, 2019. [Online]. Available: https://doi.org/10.1609/aimag.v41i4.5304

[15] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the AI2 reasoning challenge," *CoRR*, vol. abs/1803.05457, 2018. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33017063

[16] O. Tafjord, P. Clark, M. Gardner, W. Yih, and A. Sabharwal, "Quarel: A dataset and models for answering questions about qualitative relationships," *CoRR*, vol. abs/1811.08048, 2018.

[17] F. C. Akyon, D. Cavusoglu, C. Cengiz, S. O. Altinuc, and A. Temizel, "Automated question generation and question answering from Turkish texts using text-to-text transformers," *arXiv:2111.06476 [cs]*, Nov. 2021, arXiv: 2111.06476. [Online]. Available: https://doi.org/10.55730/1300-0632.3914

[18] F. J. Muis and A. Purwarianti, "Sequence-to-sequence learning for indonesian automatic question generator," *2020 7th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2020*, 9 2020. [Online]. Available: https://doi.org/10.1109/ICAICTA49861.2020.9429032

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://doi.org/10.18653/v1/D16-1264

[20] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki, "TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 454–470, 2020. [Online]. Available: https://doi.org/10.1162/tacl_a_00317

[21] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997. [Online]. Available: https://doi.org/10.1109/78.650093

[22] K. Kriangchaivech and A. Wangperawong, "Question Generation by Transformers," *arXiv:1909.05017 [cs]*, Sep. 2019, arXiv: 1909.05017.

[23] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization," *CoRR*, vol. abs/2003.11080, 2020.

[24] P. Colombo, N. Noiry, E. Irurozki, and S. Clémençon, "What are the best systems? new perspectives on NLP benchmarking," *CoRR*, vol. abs/2202.03799, 2022.

[25] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 843–857. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.699

[26] W. Etaiwi, D. Suleiman, and A. Awajan, "Deep Learning Based Techniques for Sentiment Analysis: A Survey," *Informatica*, vol. 45, no. 7, Dec. 2021. [Online]. Available: https://doi.org/10.31449/inf.v45i7.3674

[27] A. C. Mazari and A. Djeffal, "Sentiment Analysis of Algerian Dialect Using Machine Learning and Deep Learning with Word2vec," *Informatica*, vol. 46, no. 6, Jul. 2022. [Online]. Available: https://doi.org/10.31449/inf.v46i6.3340

[28] S. T. Al-Otaibi and A. A. Al-Rasheed, "A Review and Comparative Analysis of Sentiment Analysis Techniques," *Informatica*, vol. 46, no. 6, Jul. 2022. [Online]. Available: https://doi.org/10.31449/inf.v46i6.3991

[29] D. Suleiman, A. Odeh, and R. Al-Sayyed, "Arabic Sentiment Analysis Using Naïve Bayes and CNN-LSTM," *Informatica*, vol. 46, no. 6, Jul. 2022. [Online]. Available: https://doi.org/10.31449/inf.v46i6.4199

[30] A. A. Al-Rasheed, "Finding Influential Users in Social Networking using Sentiment Analysis," *Informatica*, vol. 46, no. 5, Mar. 2022. [Online]. Available: https://doi.org/10.31449/inf.v46i5.3829

[31] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. [Online]. Available: https://doi.org/10.18653/v1/W18-5446

[32] S. Gehrmann, T. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Aremu, A. Bosselut, K. R. Chandu, M.-A. Clinciu, D. Das, K. Dhole, W. Du, E. Durmus, O. Dušek, C. C. Emezue, V. Gangal, C. Garbacea, T. Hashimoto, Y. Hou, Y. Jernite, H. Jhamtani, Y. Ji, S. Jolly, M. Kale, D. Kumar, F. Ladhak, A. Madaan, M. Maddela, K. Mahajan, S. Mahamood, B. P. Majumder, P. H. Martins, A. McMillan-Major, S. Mille, E. van Miltenburg, M. Nadeem, S. Narayan, V. Nikolaev, A. Niyongabo Rubungo, S. Osei, A. Parikh, L. Perez-Beltrachini, N. R. Rao, V. Raunak, J. D. Rodriguez, S. Santhanam, J. Sedoc, T. Sellam, S. Shaikh, A. Shimorina, M. A. Sobrevilla Cabezudo, H. Strobelt, N. Subramani, W. Xu, D. Yang, A. Yerukola, and J. Zhou, "The GEM benchmark: Natural language generation, its evaluation and metrics," in *Proceedings of the 1st Workshop on Natural Language Genera-*

tion, Evaluation, and Metrics (GEM 2021). Online: Association for Computational Linguistics, Aug. 2021, pp. 96–120. [Online]. Available: https://doi.org/10.18653/v1/2021.gem-1.10

[33] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.703

[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-demos.6

[35] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 11 2020. [Online]. Available: https://doi.org/10.1162/tacl_a_00343

[36] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 784–789. [Online]. Available: https://doi.org/10.18653/v1/P18-2124

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/N19-1423

[38] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., Jul. 2020, pp. 1877–1901, arXiv: 2005.14165.

[39] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation from denoising pre-training," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3450–3466. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-acl.304

[40] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[41] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.41

[42] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettle-

moyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://doi.org/10.18653/v1/N18-1202

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

Appendix A

Table 8: Translated texts from Indonesian to English for all Indonesian texts mentioned above.

#	Indonesian	English
1	*Samudra Pasifik atau Lautan Teduh (dari bahasa spanyol Pacifico, artinya tenang) adalah kawasan kumpulan air terbesar di dunia, serta mencakup kira-kira sepertiga permukaan Bumi, dengan luas sebesar 179,7 juta km2 (69,4 juta mi2). Panjangnya sekitar 15.500 km (9.600mi) dari Laut Bering di Arktik hingga batasan es di Laut Ross di Antartika di selatan. Samudra Pasifik mencapai lebar timur-barat terbesarnya pada sekitar 5 derajat U garis lintang, di mana ia terbentang sekitar 19.800 km (12.300mi) dari Indonesia hingga pesisir Kolombia. Batas sebelah barat samudra ini biasanya diletakkan di Selat Malaka. Titik terendah permukaan Bumi—Palung Mariana—berada di Samudra Pasifik. Samudra ini terletak di antara Asia dan Australia di sebelah barat, Amerika di sebelah timur, Antartika di sebelah selatan dan Samudra Arktik di sebelah utara.*	*The Pacific Ocean or Ocean of Shades (from the Spanish Pacifico, meaning calm) is the largest area of water body in the world, and covers about a third of the Earth's surface, with an area of 179.7 million km2 (69.4 million mi2). It extends about 15,500 km (9,600mi) from the Bering Sea in the Arctic to the ice cap of the Ross Sea in Antarctica in the south. The Pacific Ocean reaches its greatest east-west width at about 5 degrees N latitude, where it extends about 19,800 km (12,300mi) from Indonesia to the coast of Colombia. The western boundary of this ocean is usually placed in the Malacca Strait. The lowest point on Earth's surface—the Mariana Trench—is in the Pacific Ocean. This ocean is located between Asia and Australia to the west, America to the east, Antarctica to the south and the Arctic Ocean to the north.*
2	*179,7 juta km2*	*179.7 million km2*
3	*Berapa luas Samudera Pasifik?*	*How wide is the Pacific Ocean?*
4	*berapa luas samudra pasifik ?*	*how wide is the pacific ocean?*
5	*berapa luas bumi pasifik ?*	*how big is the pacific earth?*
6	*berapa luas air terbesar di dunia ?*	*what is the largest area of water in the world?*
7	*berapakah luas samudra pasifik?*	*how wide is the pacific ocean?*
8	*berapakah luas samudra pasifik?*	*how wide is the pacific ocean?*
9	*berapa luas total wilayah lautan pasifik ?*	*What is the total area of the Pacific Ocean?*
10	*Berapa luas samudra pasifik?*	*How wide is the Pacific Ocean?*
11	*Kadipaten Normandia , yang mereka bentuk dengan perjanjian dengan mahkota Prancis, adalah tanah yang indah bagi Prancis abad pertengahan , dan di bawah Richard I dari Normandia ditempa menjadi sebuah pemerintahan yang kohesif dan tangguh dalam masa jabatan feodal.*	*The Duchy of Normandy, which they formed by treaty with the French crown, was a beautiful land for medieval France, and under Richard I of Normandy was forged into a cohesive and formidable government in feudal tenure.*
12	*Kadipaten Normandia*	*Duchy of Normandy*
13	*Siapa yang memerintah kadipaten Normandia*	*Who ruled the duchy of Normandy*
14	*siapa yang memerintah pemerintahan normandia ?*	*who governs the normandy government?*
15	*siapa yang mendirikan kadipaten normandia ?*	*who founded the duchy of normandy?*

Appendix B

Table 9: Translated texts from Indonesian to English for all Indonesian texts mentioned above.

#	Indonesian	English
16	*siapa yang memerintah normandia di normandia ?*	*who rules normandy in normandy ?*
17	*siapakah kadipaten normandia di bawah raja normandia ?*	*who is the duchy of normandy under the king of normandy ?*
18	*siapa yang memimpin normandia ?*	*who is in charge of normandy?*
19	*dengan siapa prancis membentuk kadipaten normandia ?*	*With whom did France form the Duchy of Normandy?*
20	*Siapa yang memerintahkan kadipaten normandia?*	*Who ruled the duchy of normandy?*