A Comparison of OpenNMT Sequence Model for Indonesian Automatic Question Generation

Yuniar Indrihapsari[1], Handaru Jati[2], Nurkhamid[3], Ratna Wardani[4,] , Pradana Setialana[5], Muhamad Izzudin Mahali[6], Danang Wijaya[7], Dhista Dwi Nur Ardiansyah[8] , Satya Adhiyaksa Ardy[9] , Maria Bernadetha Charlotta Wonda Tiala[10], Andi Hakim Al-khawarizmi[11], and Widya Ardiyanto[12]

[1,3,5,9,10,11,12] Information Technology, UNY, Yogyakarta, Indonesia
[2,4,7,8] Informatics Engineering Education, UNY, Yogyakarta, Indonesia
[1]Information Management, NTUST, Taipei, Taiwan
[6]Electronic and Computer Engineering, NTUST, Taipei, Taiwan
[*]E-mail: yuniar@uny.ac.id

ABSTRACT

Evaluation of learners is a crucial aspect of the educational system. However, creating evaluation instruments is a process that demands teachers' time and energy. The researcher developed the Indonesia Automatic Question Generator in this study using an architecture modified from past studies. The primary goals of this project are (1) to construct an AQG tool utilizing the OpenNMT series and (2) to analyze and compare the model's performance. As a data source, this study employs the SQuAD 2.0 dataset and numerous sequence techniques, including BiGRU, BiLSTM, and Transformer. The researcher trained the models using OpenNMT-py and Google Collaboratory. This approach generates questions that are relevant to the context of the source. This study found that the model was acceptable.

Keywords: openNMT, SQuAD 2.0, Indonesian automatic question generator, evaluation process

INTRODUCTION

Evaluation is an essential process in learning activities at school. Evaluation is the implementation of procedures to determine the extent to which students have achieved learning objectives. In addition to serving as a form of assessment, questions can influence student learning. As said by [1], the advantages of using the specific question include: (1) providing the possibility to practice retrieving stored information, (2) offering learners responses regarding with there incorrect assumptions, (3) concentrating students' learning attention on the vital educational substance, (4) supporting knowledge by replicating core concepts, and (5) encouraging students to participate in active learning (e.g., discussing and reading). Despite these benefits, manual question creation is a complicated task that requires expertise, experience, and funds.

The increase in the human population has led to an increase in the number of students in schools and other educational institutions. This resulted in a shortage of teachers and hence their busy schedules. In such a scenario, it becomes difficult for teachers to create questions for the student evaluation process and give them subjective feedback [2]. The other problem is that the teacher has difficulty developing instruments to make the questions for learning evaluation. Question evaluation creation is a complicated process that requires diverse knowledge, expertise, experience, and reference materials. Creating evaluation questions manually requires much time and effort from the teacher, who must condense a great deal of information from many sources [3]. The problem affects teachers to create and develop learning evaluation questions that are not optimal and often use questions used in previous evaluations [4]. Using past questions can cause the results of the evaluation process will not to match the real-time student's abilities. Researchers found that the automation of question development could significantly decrease the burdens placed on teachers after examining the process of reviewing student performance by creating questions.

Teachers can use Automatic Question Generation's (AQG) tool to solve this problem. Automatic question generation (AQG) plays a vital function in educational evaluation. Manual question writing is employment, time-consuming, and expensive. In the last two decades, academics have focused on creating an autonomous system for generating questions and evaluating the responses from students [5]. Several researchers have developed various methods to overcome this challenge. Some techniques were developed as a response to test developers' challenges while creating many excellent questions. The study field of automated question creation for academic purposes has garnered the interest of scholars from various disciplines. Question generation is defined as follows: Question generation is the autonomous generating of questions from a host of variables, such as unprocessed text, a database, or a semantic representation [6]. This definition indicates that the input for question generation could be in various forms, such as a paragraph, a sentence, or a semantic map. AQG focuses on developing algorithms for generating questions from organized (e.g., knowledge sources) or unorganized (e.g., text) sources of information. Recently, AQG has become more significant due to the development of MOOCs and other e-learning technologies [7][8][9].

In the last decades, the question-generating system QGSTEC used a dataset of 1,000 questions (generated by machines and humans). The algorithm produced questions for each category (who, what, which, where, when, and how many). To generate the questions' accuracy, use the variable relevance, type of question, grammatical accuracy, and ambiguity. Both the validity and syntactic accuracy measures scored low marks. Little to no consistency emerged here between the two human judges.

The 30MQA, SQuAD, RACE, NewsQA, MS MARCO, TriviaQA, and NarrativeQA datasets provide question-answer pairs primarily for enhancing mechanical reading comprehension in question-answer models. The design of the use of these data sources will produce questions that are not actual from text sources. Furthermore, the datasets could be more suitable for educational assessment due to their limiting variety of themes or a lack of data for developing questions and answering them.

According to most prior studies, AQG models constructed with a deep learning technique beat their rule-based equivalent significantly and offer greater flexibility and domain coverage. Since 1950 the strategy to assess student reading comprehension, the cloze test method, removes some words repeatedly (for example, one word every five words) and assigns the students to guess the missing word [4].

In this study, researchers build a tool to help teachers create and develop the questions for learning evaluation. The AQG parsing approach based on syntax or semantics works by parsing the text depending on the paragraph's syntactic or semantic meaning. The following procedure continues with the process of creating questions [10]. Natural Language Processing is used with Deep Learning approach to generate questions from input paragraphs. For the database resource, the researcher used translated SQuAD 2.0 dataset in Indonesian as the resource to build the model. SQuAD 2.0 is a massive scale of the Stanford Question Answering Dataset, enabling researchers to design AI models for reading comprehension tasks under challenge. The researcher used OpenNMT-py and Google Colaboratory to train the models. The primary purposes of this study are (1) to build an AQG system using the OpenNMT series and (2) to evaluate and compare the model's performance. The researcher compares the result of the models trained with OpenNMT using Untrained Automatic Metrics such as BLEU, ROUGE, and METEOR [11]. In this AQG evaluation, researchers used metrics commonly used for machine translation tasks. This metric assesses scores based on predictions and target similarity (starts at 0-1, but the researcher multiplies by 100 to simplify reading and comparison).

METHODS

A. Research and Development Method

The researcher uses the Research and Development (R&D) method to produce the specific product and test the product. In this study, the researcher built a prototype of a deep learning model for autonomous question generation. The framework and tools used are PyTorch and OpenNMT with the python programming language and the SQuAD 2.0 dataset. Next step, the dataset is translated into Bahasa Indonesia. This research consists of 4 steps that shown in Figure 1.

(1) Dataset Translation: in this step, we translate the SQuAD dataset from English into Bahasa Indonesia. First, the researcher needs to identify the main resource in the English text. The SQuAD is an English language dataset, and the researcher only has Indonesian translation resources. (2) Dataset Preparation: after the

dataset is translated, the researcher creates feature extraction from the dataset, including POS Tagging, Named Entity, and answer location. (3) Model Training: researcher trains the model using the sequence model from OpenNMT with the RNN and LSTM-based model.

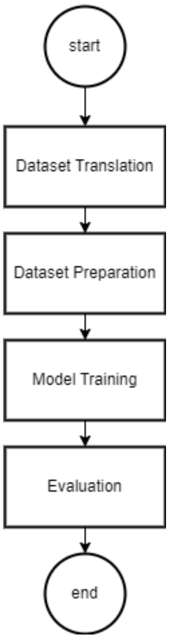


Figure 1. Research Flowchart

(4) Evaluation: after the researcher trained the model, the next step was to evaluate with ngl-eval tools and use several untrained automatic metrics such as BLEU, METEOR, and ROUGE.

B. Materials and Instrument of Study

The research used various software libraries and development kits. This study uses Google Colaboratory with the GPU that can run in a web browser using an internet connection for the development environment. The research also used some libraries such as PyTorch and OpenNMT-py for training the deep learning model and nlg-eval tools to evaluate each model that has been trained.

C. Data Collection Techniques and Instruments

This study using SQuAD Dataset as a resource. Dataset of Stanford Question Answering SQuAD is a reading comprehension dataset used for machine reading comprehension, which is the primary objective of natural language processing. SQuAD 2.0 is an improved version by adding 53.775 new questions that the same paragraphs could not answer in the last SQuAD version. The questions are arranged so that they are relevant to the statement paragraph, and the paragraph contains reasonable answers. SQuAD 2.0 has a higher quality than the previous version, with a state-of-the-art model getting a score of 66.3% during training and testing, while human accuracy is 89.5% [12].

D. Data Analysis Technique

Data analysis was carried out from the evaluation of each model configuration. That had gone through the training stage using untrained automatic metrics to compare the performance evaluation results from the model training process. This result was used to decide which model configuration had the best performance.

RESULT AND DISCUSSION

A. Dataset Translation

Our initial stage is translating the SQuAD dataset from English to Indonesian using the Google Translate API with the help of the google trans library. Before being translated, the dataset goes through several processes of splitting and selecting the data contained in the SQuAD dataset. The process flow is shown in Figure 2.

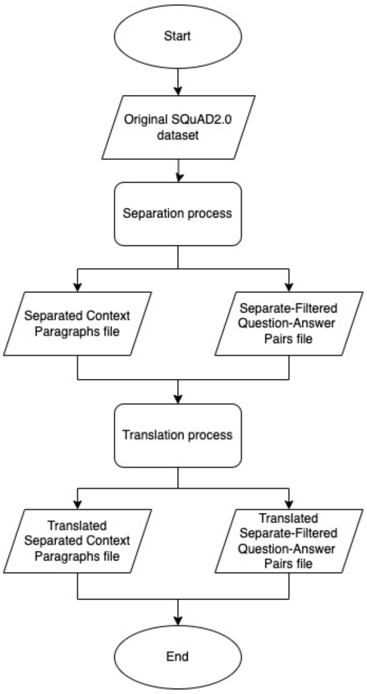

Start
Original SQuAD2.0 dataset
Separation process
Separated Context Paragraphs file
Separate-Filtered Question-Answer Pairs file
Translation process
Translated Separated Context Paragraphs file
Translated Separate-Filtered Question-Answer Pairs file
End

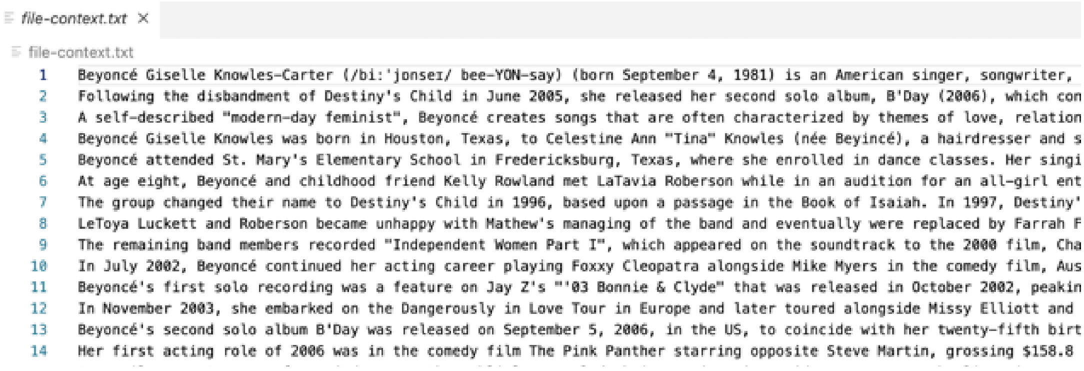
Figure 2. Dataset Translation Process

In the first step, researchers separate the context paragraphs from the dataset for further writing into a new file-the context paragraphs file shown in Figure 3.

The second step is separate the question-answer pairs from the dataset and writes them into a new file. Next, assign a number to determine the question-answer pairs referring to the context paragraphs previously separated. While doing the split, researchers also selected unanswered question-answer pairs not to be written to the file. The separated question-answer pairs file is shown in Figure 4 and Figure 5.

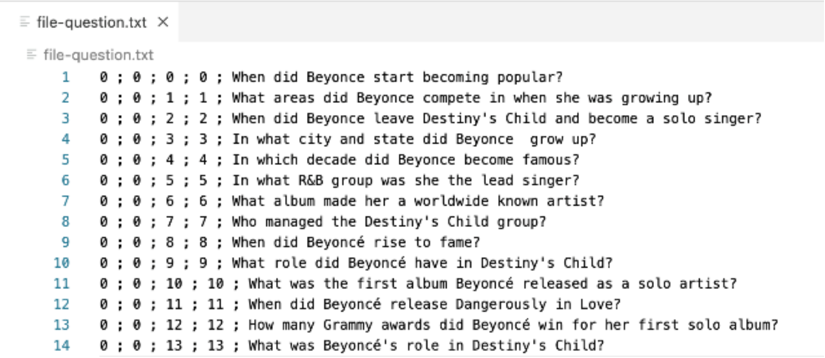
After all the data is separated, the following step is to translate all the context paragraphs and the question-answer pairs using the Google Translate API with the help of the Google trans library and write it into a new file. All translated dataset files are shown in Figure 6. The translated separated question-answer pairs file is shown in Figure 7 and Figure 8.

The process of splitting and selecting is carried out on both train-set and dev-set datasets. Next step, the researcher makes variations of cased and uncased from datasets that have been separated, selected, and translated into Indonesian.



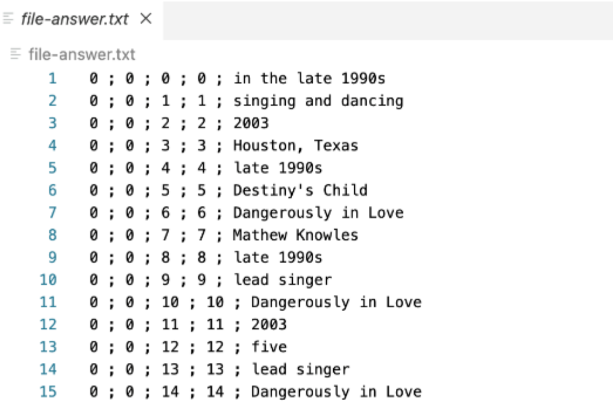
```
Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (born September 4, 1981) is an American singer, songwriter,
Following the disbandment of Destiny's Child in June 2005, she released her second solo album, B'Day (2006), which con
A self-described "modern-day feminist", Beyoncé creates songs that are often characterized by themes of love, relation
Beyoncé Giselle Knowles was born in Houston, Texas, to Celestine Ann "Tina" Knowles (née Beyincé), a hairdresser and s
Beyoncé attended St. Mary's Elementary School in Fredericksburg, Texas, where she enrolled in dance classes. Her singi
At age eight, Beyoncé and childhood friend Kelly Rowland met LaTavia Roberson while in an audition for an all-girl ent
The group changed their name to Destiny's Child in 1996, based upon a passage in the Book of Isaiah. In 1997, Destiny'
LeToya Luckett and Roberson became unhappy with Mathew's managing of the band and eventually were replaced by Farrah F
The remaining band members recorded "Independent Women Part I", which appeared on the soundtrack to the 2000 film, Cha
In July 2002, Beyoncé continued her acting career playing Foxxy Cleopatra alongside Mike Myers in the comedy film, Aus
Beyoncé's first solo recording was a feature on Jay Z's "'03 Bonnie & Clyde" that was released in October 2002, peakin
In November 2003, she embarked on the Dangerously in Love Tour in Europe and later toured alongside Missy Elliott and
Beyoncé's second solo album B'Day was released on September 5, 2006, in the US, to coincide with her twenty-fifth birt
Her first acting role of 2006 was in the comedy film The Pink Panther starring opposite Steve Martin, grossing $158.8
```

Figure 3. Context paragraphs



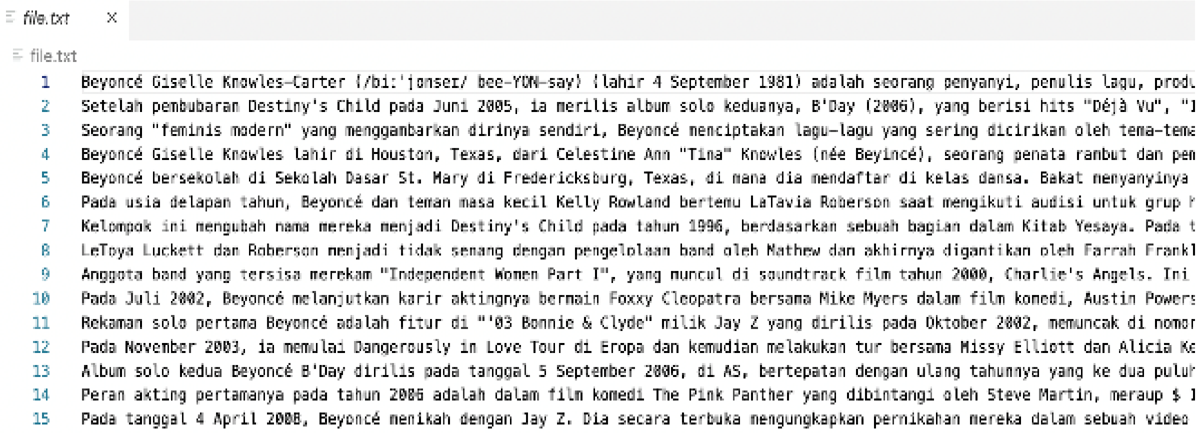
```
0 ; 0 ; 0 ; 0 ; When did Beyonce start becoming popular?
0 ; 0 ; 1 ; 1 ; What areas did Beyonce compete in when she was growing up?
0 ; 0 ; 2 ; 2 ; When did Beyonce leave Destiny's Child and become a solo singer?
0 ; 0 ; 3 ; 3 ; In what city and state did Beyonce  grow up?
0 ; 0 ; 4 ; 4 ; In which decade did Beyonce become famous?
0 ; 0 ; 5 ; 5 ; In what R&B group was she the lead singer?
0 ; 0 ; 6 ; 6 ; What album made her a worldwide known artist?
0 ; 0 ; 7 ; 7 ; Who managed the Destiny's Child group?
0 ; 0 ; 8 ; 8 ; When did Beyoncé rise to fame?
0 ; 0 ; 9 ; 9 ; What role did Beyoncé have in Destiny's Child?
0 ; 0 ; 10 ; 10 ; What was the first album Beyoncé released as a solo artist?
0 ; 0 ; 11 ; 11 ; When did Beyoncé release Dangerously in Love?
0 ; 0 ; 12 ; 12 ; How many Grammy awards did Beyoncé win for her first solo album?
0 ; 0 ; 13 ; 13 ; What was Beyoncé's role in Destiny's Child?
```

Figure 4. The separated questions from the dataset



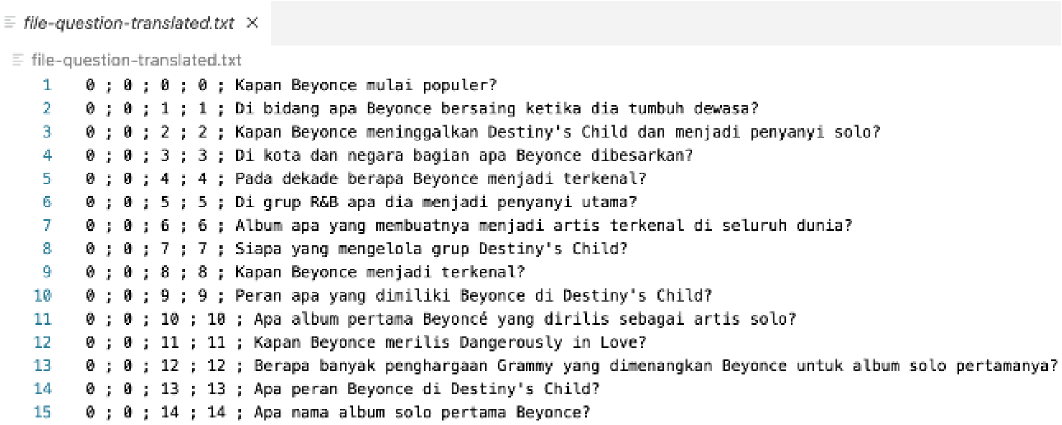
```
0 ; 0 ; 0 ; 0 ; in the late 1990s
0 ; 0 ; 1 ; 1 ; singing and dancing
0 ; 0 ; 2 ; 2 ; 2003
0 ; 0 ; 3 ; 3 ; Houston, Texas
0 ; 0 ; 4 ; 4 ; late 1990s
0 ; 0 ; 5 ; 5 ; Destiny's Child
0 ; 0 ; 6 ; 6 ; Dangerously in Love
0 ; 0 ; 7 ; 7 ; Mathew Knowles
0 ; 0 ; 8 ; 8 ; late 1990s
0 ; 0 ; 9 ; 9 ; lead singer
0 ; 0 ; 10 ; 10 ; Dangerously in Love
0 ; 0 ; 11 ; 11 ; 2003
0 ; 0 ; 12 ; 12 ; five
0 ; 0 ; 13 ; 13 ; lead singer
0 ; 0 ; 14 ; 14 ; Dangerously in Love
```

Figure 5. The separated answers from dataset



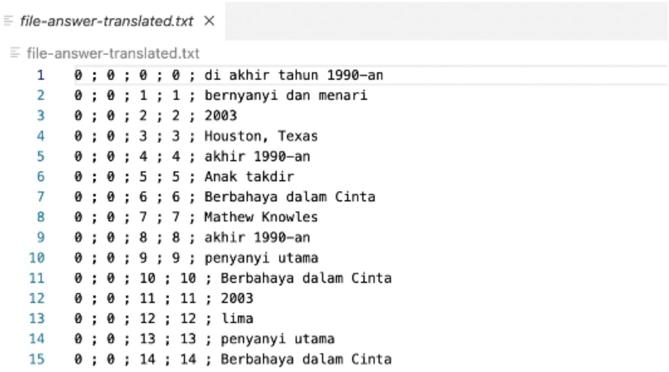
```
Beyoncé Giselle Knowles-Carter (/biː'jɒnseɪ/ bee-YON-say) (lahir 4 September 1981) adalah seorang penyanyi, penulis lagu, produ
Setelah pembubaran Destiny's Child pada Juni 2005, ia merilis album solo keduanya, B'Day (2006), yang berisi hits "Déjà Vu", "I
Seorang "feminis modern" yang menggambarkan dirinya sendiri, Beyoncé menciptakan lagu-lagu yang sering dicirikan oleh tema-tema
Beyoncé Giselle Knowles lahir di Houston, Texas, dari Celestine Ann "Tina" Knowles (née Beyincé), seorang penata rambut dan pem
Beyoncé bersekolah di Sekolah Dasar St. Mary di Fredericksburg, Texas, di mana dia mendaftar di kelas dansa. Bakat menyanyinya
Pada usia delapan tahun, Beyoncé dan teman masa kecil Kelly Rowland bertemu LaTavia Roberson saat mengikuti audisi untuk grup H
Kelompok ini mengubah nama mereka menjadi Destiny's Child pada tahun 1996, berdasarkan sebuah bagian dalam Kitab Yesaya. Pada t
LeToya Luckett dan Roberson menjadi tidak senang dengan pengelolaan band oleh Mathew dan akhirnya digantikan oleh Farrah Frankl
Anggota band yang tersisa merekam "Independent Women Part I", yang muncul di soundtrack film tahun 2000, Charlie's Angels. Ini
Pada Juli 2002, Beyoncé melanjutkan karir aktingnya bermain Foxxy Cleopatra bersama Mike Myers dalam film komedi, Austin Powers
Rekaman solo pertama Beyoncé adalah fitur di "'03 Bonnie & Clyde" milik Jay Z yang dirilis pada Oktober 2002, memuncak di nomor
Pada November 2003, ia memulai Dangerously in Love Tour di Eropa dan kemudian melakukan tur bersama Missy Elliott dan Alicia Ke
Album solo kedua Beyoncé B'Day dirilis pada tanggal 5 September 2006, di AS, bertepatan dengan ulang tahunnya yang ke dua puluh
Peran akting pertamanya pada tahun 2006 adalah dalam film komedi The Pink Panther yang dibintangi oleh Steve Martin, meraup $ 1
Pada tanggal 4 April 2008, Beyoncé menikah dengan Jay Z. Dia secara terbuka mengungkapkan pernikahan mereka dalam sebuah video
```

Figure 6. Context paragraphs (translate in Indonesia)



```
0 ; 0 ; 0 ; 0 ; Kapan Beyonce mulai populer?
0 ; 0 ; 1 ; 1 ; Di bidang apa Beyonce bersaing ketika dia tumbuh dewasa?
0 ; 0 ; 2 ; 2 ; Kapan Beyonce meninggalkan Destiny's Child dan menjadi penyanyi solo?
0 ; 0 ; 3 ; 3 ; Di kota dan negara bagian apa Beyonce dibesarkan?
0 ; 0 ; 4 ; 4 ; Pada dekade berapa Beyonce menjadi terkenal?
0 ; 0 ; 5 ; 5 ; Di grup R&B apa dia menjadi penyanyi utama?
0 ; 0 ; 6 ; 6 ; Album apa yang membuatnya menjadi artis terkenal di seluruh dunia?
0 ; 0 ; 7 ; 7 ; Siapa yang mengelola grup Destiny's Child?
0 ; 0 ; 8 ; 8 ; Kapan Beyonce menjadi terkenal?
0 ; 0 ; 9 ; 9 ; Peran apa yang dimiliki Beyonce di Destiny's Child?
0 ; 0 ; 10 ; 10 ; Apa album pertama Beyoncé yang dirilis sebagai artis solo?
0 ; 0 ; 11 ; 11 ; Kapan Beyonce merilis Dangerously in Love?
0 ; 0 ; 12 ; 12 ; Berapa banyak penghargaan Grammy yang dimenangkan Beyonce untuk album solo pertamanya?
0 ; 0 ; 13 ; 13 ; Apa peran Beyonce di Destiny's Child?
0 ; 0 ; 14 ; 14 ; Apa nama album solo pertama Beyonce?
```

Figure 7. The separated questions from the dataset (translate in Indonesia)



```
0 ; 0 ; 0 ; 0 ; di akhir tahun 1990-an
0 ; 0 ; 1 ; 1 ; bernyanyi dan menari
0 ; 0 ; 2 ; 2 ; 2003
0 ; 0 ; 3 ; 3 ; Houston, Texas
0 ; 0 ; 4 ; 4 ; akhir 1990-an
0 ; 0 ; 5 ; 5 ; Anak takdir
0 ; 0 ; 6 ; 6 ; Berbahaya dalam Cinta
0 ; 0 ; 7 ; 7 ; Mathew Knowles
0 ; 0 ; 8 ; 8 ; akhir 1990-an
0 ; 0 ; 9 ; 9 ; penyanyi utama
0 ; 0 ; 10 ; 10 ; Berbahaya dalam Cinta
0 ; 0 ; 11 ; 11 ; 2003
0 ; 0 ; 12 ; 12 ; lima
0 ; 0 ; 13 ; 13 ; penyanyi utama
0 ; 0 ; 14 ; 14 ; Berbahaya dalam Cinta
```

Figure 8. The separated answers from dataset (translate in Indonesia)

B. Dataset Preparation

The researcher carried out several dataset preparation processes after the dataset was translated into Indonesian. The first step is determining the answer position in the context paragraph, based on the word's position. The second step is stemming and tokenizing to perform the process of extracting linguistic features with POS Tagging and Named Entity. The next step is to perform linguistic feature extraction with the NLTK library. Furthermore, the final step is dividing the dataset that would be used as a training dataset, validation, and testing. This dividing is performed on every dataset variation, which is uncased and cased. In every training, validation, and testing data, two types of data are saved in a different file, the source file containing paragraph data, which has been given linguistic features and split by space, and the target file containing questions targeted by the source file.

C. Model Training

Model training process carried out in Google Collaboratory with GPU runtime type and using several toolkits, libraries, and frameworks such as PyTorch and OpenNMT-py. Researchers use RNNs such as GRU and LSTM with bidirectional counterparts [13], and transformers architectures use linguistic features in the input context. RNN and transformers-based models use the same dataset with cased and uncased variations. This study showed a better result, except for the transformer's training time, which was 1,07 times slower. The number of steps has the same value. However, the accuracy rate is lower but still acceptable. The author realizes that the accuracy value could be more optimal because the SQuAD database translation process produces unnatural Indonesian. It takes much human effort to fix different patterns for syntactical and semantic translation [14]. Table 1 is a result of the training process for each model.

Table 1. *Model Training Result*

Model Name	Validation Accuracy	Step	Time (s)
BiGRU *Uncased*	50.15	32100	6136
BiGRU *Cased*	49.63	32100	3105
BiLSTM *Uncased*	50.91	16050	1888
BiLSTM *Cased*	50.50	16050	1861
Transformer *Uncased*	50.89	120600	11998
Transformer *Cased*	50.39	120600	12854

D. Evaluation

The model evaluation step is carried out using the Automatic Measures Method. The model evaluation on the prototype was built using UAM (untrained automatic metric) such as BLEU, ROUGE, and METEOR metrics. Before evaluating each model, the inference is made, making questions using each model built from the test dataset. The inference is made by using test data from each dataset variation. Table 2 shows the question results of the inference from each model.

Table 2 displays a subset of SQuAD test questions produced by our systems. "Input sentence" is the model input, "Answer" is the predicted response, and "Target" is the predicted created question. In this research, specific model names have been abbreviated. From the result of case 1, case 2, case 3, and case 4, the researcher got that tool generates questions that are very close to the context of the source. Those models were proven to predict the same questions, resulting in nearly identical questions. There were also some variations in the verb used in the generated questions. They are all alternatives that have the exact meaning. There were unnatural translations of input phrases and target questions, which negatively impacted the accuracy of our model's predictions. Nevertheless, these statements might still be semantically comprehended.

Evaluation is carried out on the inference results from each model in each dataset variant by comparing the inference results with the target test data using the nlg-eval tool that contains untrained automatic untrained metrics such as BLEU, METEOR, and ROUGE-L. These metrics evaluate scores based on projection and objective similarity (ranging from 0 to 1 but multiplied by 100 to relieve readability and comparability). These metrics are suitable for the supervised assessment of generated questions. BLEU is the precision-based metric that counts the total n-grams ratio shown in (1).

$$BLEU - N = bp.exp(\textstyle\sum_{n=1}^{N} W_n logprecision_n) \tag{11}$$

ROUGE-L is the F-measure based metric where the recall and precision are calculated using the length of the LCS (Longest Common Subsequence) between pairs of sentences, shown in (3).

$$P_{lcs} = \frac{|LCS(p,r)|}{\#words\_in\_hypothesis}, R_{lcs} = \frac{|LCS(p,r)|}{\#words\_in\_refrerence} \tag{12}$$

$$ROUGE - L = F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \tag{13}$$

METEOR use F-score and relaxed matching criteria. METEOR metric is the upgrade from BLEU limitation to the unigram from reference or synonym. The F-score from METEOR is calculated by shown in (5).

$$P(Precision) = \frac{\#mapped\_unigrams}{\#unigrams\_in\_candidate}, \quad R(Recall) = \frac{\#mapped\_unigrams}{\#unigrams\_in\_reference} \tag{14}$$

$$\boldsymbol{F}_{score} = \frac{\boldsymbol{10PR}}{\boldsymbol{R+9P}} \tag{15}$$

Table 2. Question results from each model

1	
Input	*Teori kompleksitas komputasi adalah cabang dari teori komputasi dalam ilmu komputer teoretis yang berfokus pada pengelompokan masalah komputasi sesuai dengan kesulitan yang melekat, dan menghubungkan kelas-kelas tersebut satu sama lain.*
Answer	*Cabang dari teori komputasi dalam ilmu komputer teoretis yang berfokus pada pengelompokan masalah komputasi sesuai dengan kesulitan yang melekat, dan menghubungkan kelas-kelas tersebut satu sama lain*
Target	*Apa yang dimaksud dengan teori kompleksitas komputasi ?*
BiGRU *Uncased*	*apa yang dimaksud dengan teori kompleksitas komputasi ?*
BiGRU *Cased*	*Apa yang dimaksud dengan teori kompleksitas komputasi ?*
BiLSTM *Uncased*	*teori kompleksitas komputasi berfokus pada apa ?*
BiLSTM *Cased*	*Teori kompleksitas komputasi adalah cabang dari teori apa ?*
Transformer *Uncased*	*teori kompleksitas komputasi sesuai dengan apa ?*
Transformer *Cased*	*Teori kompleksitas komputasi sesuai dengan apa ?*
2	
Input	*Mesin uap adalah mesin pembakaran eksternal, di mana fluida kerja terpisah dari produk pembakaran.*
Answer	*Mesin pembakaran eksternal*
Target	*Apa itu mesin uap?*
BiGRU *Uncased*	*apa itu mesin uap ?*
BiGRU *Cased*	*Mesin uap apa yang digunakan untuk membuat motor uap ?*
BiLSTM *Uncased*	*apa itu mesin uap ?*
BiLSTM *Cased*	*Apa itu mesin uap ?*
Transformer *Uncased*	*apa mesin uap ?*
Transformer *Cased*	*Apa mesin uap ?*
3	
Input	*Siklus termodinamika ideal yang digunakan untuk menganalisis proses ini disebut siklus Rankine.*
Answer	*Siklus Rankine*
Target	*Apa nama siklus termodinamika dalam proses mesin uap?*
BiGRU *Uncased*	*apa siklus termodinamika ideal yang digunakan untuk menganalisis proses ini ?*
BiGRU *Cased*	*Apa nama lain untuk siklus termodinamika ideal ?*
BiLSTM *Uncased*	*apa siklus termodinamika ideal yang digunakan untuk menganalisis proses ini ?*
BiLSTM *Cased*	*Siklus termodinamika ideal disebut siklus apa ?*
Transformer *Uncased*	*apa siklus termodinamika ideal yang digunakan untuk menganalisis proses termodinamika ?*
Transformer *Cased*	*Apa siklus termodinamika yang digunakan untuk menganalisis proses ini ?*
4	
Input	*Oksigen adalah unsur kimia dengan simbol O dan nomor atom 8.*
Answer	*8*
Target	*Nomor atom oksigen adalah?*
BiGRU *Uncased*	*berapa nomor atom yang oksigen ?*
BiGRU *Cased*	*Berapa angka atom ?*
BiLSTM *Uncased*	*berapa nomor atom oksigen ?*
BiLSTM *Cased*	*Berapa nomor atom untuk Oksigen ?*
Transformer *Uncased*	*berapa nomor atom yang dimiliki oksigen ?*
Transformer *Cased*	*Berapa jumlah nomor atom yang digunakan untuk hidrogen ?*

Table 3. Performance comparison of each model

Model Name	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
BiGRU Uncased	37.58	20.21	10.19	5.47	42.80	17.97
BiGRU Cased	33.92	17.47	8.64	4.72	39.63	17.54
BiLSTM Uncased	**38.10**	**20.69**	**10.58**	**5.78**	**42.98**	**18.31**
BiLSTM Cased	34.37	17.70	8.93	4.91	39.62	17.75
Transformers Uncased	35.90	18.78	8.96	4.65	41.70	17.10
Transformers Cased	32.71	16.53	7.98	4.08	38.67	16.79

Researchers discovered that all uncased variants outperformed cased versions in terms of metrics. Overall, the most performing model was BiLSTM Uncased, BiGRU Uncased, and Transformers Uncased. The models performed consistently across all three tasks. In all three experiments, the non-cased models outperformed the cased versions. These results suggest that uncasing may benefit neural language models since it allows the embeddings to be trained on more context-independent representations and thus achieve better overall performance. This study demonstrated that the OpenNMT implementation is effective. BiLSTM Uncased performed better in all trials than BiGRU Uncased and Transformer Uncased. Researchers discovered that uncasing the input sentences boosted model performance.

CONCLUSION

This research has demonstrated that develop an Indonesian Automatic Question Generator system using a machine-translated question-answering dataset (SQuAD v2.0) is feasible with satisfactory results. However, this causes the model to learn from partial and unnatural data, occasionally affecting the generation of questions. Further, researchers observed that implementing OpenNMT enables us to create the AQG system more effectively and efficiently than when we implemented it ourselves. Moreover, the model varieties Uncased/Cased, model architectures produce improve the model performance. From the perspective of a native Indonesian, researchers believe the queries generated by our best models regarding their best scenarios to be acceptable and reasonably valuable. For further research, the researcher will use other databases, such as TyDiQA, and compare the results. Apart from that, the researcher can use the Indobert model that recently appeared, fortunately, to avoid mistakes in the translation process because Indobert already uses Indonesian language.

REFERENCES

[1] W. Thalheimer, “The learning benefits of questions,” 2003. http://www.worklearning.com/ma/PP_WP003.asp (accessed Aug. 31, 2007).

[2] P. K. M. Debajit Datta, Rishav Agarwal, Ishita Tuteja, V Bhavyashri Vedula, “Optimization of an Automated Examination Generation System Using Hybrid Recurrent Neural Network,” *International Journal of Interdisciplinary Global Studies*, vol. 14, no. 04, 2020.

[3] H. R. S. Flores Veronica Ambassador, Jasa Lie, “Pembangkit Pertanyaan Otomatis pada Materi Pelajaran IPA Berbahasa Indonesia di Tingkat SD Berdasarkan Revisi Taksonomi Bloom,” *Majalah Ilmiah Teknologi Elektro*, vol. 20, no. 2.

[4] S. Rakangor and Y. R. Ghodasara, “Literature Review of Automatic Question Generation Systems,” *International Journal of Scientific and Research Publications*, vol. 5, no. 1, pp. 2250–3153, 2015.

[5] M. Divate and A. Salgaonkar, “Automatic question generation approaches and evaluation techniques,” *Current Science*, vol. 113, no. 9, pp. 1683–1691, 2017, doi: 10.18520/cs/v113/i09/1683-1691.

[6] V. Rus, B. Wyse, P. Pivek, M. Lintean, S. Stoyanchev, and C. Moldovan, “The first question generation shared task evaluation challenge,” *INLG 2010 - Proceedings of the 6th International Natural Language Generation Conference*, pp. 251–254, 2010.

[7] S. Panda and S. Garg, *Open and Distance Education in Asia, Middle East Africa and the National Perspectives in a Digital Age*. 2019. doi: 10.1007/978-981-13-5787-9_4.

[8] M. Divate *et al.*, *E-Learning in European higher education institutions: Results of a mapping survey conducted in october-december 2013*, vol. 113, no. December. 2014. doi: 10.1007/978-981-13-5787-9_4.

[9] I. R. Goldbach and F. G. Hamza-Lup, “Survey on e-Learning Implementation in Eastern-Europe Spotlight on Romania,” *International Conference on Mobile, Hybrid, and On-line Learning*, vol. 9, no. February, pp. 5–12, 2017.

[10] X. Yuan *et al.*, “Machine Comprehension by Text-to-Text Neural Question

Generation." arXiv, May 15, 2017. Accessed: Jun. 05, 2023. [Online]. Available: http://arxiv.org/abs/1705.02012

[11] S. Sharma, L. E. Asri, H. Schulz, and J. Zumer, "Relevance of Unsupervised Metrics in Task-Oriented Dialogue for Evaluating Natural Language Generation," *arXiv:1706.09799 [cs]*, Jun. 2017, Accessed: Mar. 13, 2022. [Online]. Available: http://arxiv.org/abs/1706.09799

[12] P. Rajpurkar, R. Jia, and P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD," *arXiv:1806.03822 [cs]*, Jun. 2018, Accessed: Feb. 22, 2022. [Online]. Available: http://arxiv.org/abs/1806.03822

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

[14] F. J. Muis and A. Purwarianti, "Sequence-to-Sequence Learning for Indonesian Automatic Question Generator," in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, Tokoname, Japan: IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/ICAICTA49861.2020.9429032.