ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse266942020.pdf
https://doi.org/10.30534/ijatcse/2020/266942020

Encoder-Decoder with Attention Mechanisms for Developing Question Generation Models in Education

Bambang Dwi Wijanarko[1, 2], Yaya Heryadi[1], Hapnes Toba[3], and Widodo Budiharto[1]
[1]Computer Science Department, BINUS Graduate Program-Doctor of Computer Science, Bina Nusantara University, Indonesia, bwijanarko@binus.edu
[2]Computer Science Department, BINUS Online Learning, Universitas Bina Nusantara, Indonesia
[3]Faculty of Information Technology, Maranatha Christian University, Indonesia

ABSTRACT

This research aims to create a system that generates questions that are used in higher education circles. This study uses a machine learning approach is to create questions that refer to the taxonomic bloom. This study also uses encoder-decoder and attention techniques that are used to look for context. The dataset used consisted of 93,000 pairs of questions. Data were trained repeatedly with a percentage of dataset 80:20 and classified using six levels in bloom's taxonomy and based on vector embedding. The results of this study prove that the encoder-decoder model with attention produces better and easier questions than other models. Evaluation of the question-producing system by comparing the two activation functions, namely hyperbolic tangential and sigmoid, shows that the sigmoid function can reduce the value of the loss function to 0.001, which means it can minimize the error factor.

Key words : Question generator, Attention mechanism, Encoder-Decoder, Education, Bloom's Taxonomy.

1. INTRODUCTION

The development of technology, especially in the field of education, has become a trend that researchers in recent years have focused on researchers in various fields. The phenomenon of the need for quality education is the reason for the importance of research in education. Develop a knowledge sharing system in the context of education [1], propose a recommendation system related to investment in education [2] build a contextual information-based recommendation system for higher education [3], Construct validation for Academic Application in Higher Education Institution [4] and this research, produce a system that generates questions with the same goal, namely to support a better learning system.

Question Generation (QG) that are relevant to the information in a sentence or paragraph. Various approaches have been made to the question generator related explicitly to the use of rules and human-made patterns to change descriptive sentences for various related questions [5] [6][7].

In the past decade, various studies have been carried out by several researchers in this field [8]–[12]. Based on the development of their previous research results, Du., et.all [13] introduced automatic question generation for sentences from text passages in reading comprehension, a mind-based sequential learning model to the task, and investigated the effect of encoding sentence-level information versus paragraph level. The automatic evaluation results show that our system significantly outperforms the state-of-the-art rule-based system. In individual evaluations, we also consider questions generated by their systems more natural (e.g., grammar, fluency) and more difficult to answer (in terms of syntactic and lexical differences from the original text and the reasons needed to answer).

The purpose of this study is to make questions with the use of taxonomy bloom in teaching in higher education. This research belongs to the category of Natural Language Processing (NLP), specifically text-generation. Various studies in the field of text-generation tend to produce regular expressions according to grammar (news sentences). The contribution of this research is to make questions following the context of source documents and taxonomy bloom. They use a machine learning approach with an encoder-decoder with attention technique (Attention used to look for context)

2. LITERATURE REVIEW

Rus, et.al [14] proposed a prominent definition of Question Generation as an automatic question-making machine, the questions generated have several types such as factual questions, Yes / No, Questions, etc. [15]. This machine input like text (specifically declarative sentences), raw data, and knowledge base. (Graesser, Ozuru, & Sullins,

in the media) stated that the success of other researchers in various fields had produced many criteria for good question models. Among them are taxonomic questions in the cognitive science literature, i.e., the Graesser-Person taxonomy, which classifies questions according to the nature of the information sought in the right answer to the question, Mosenthal's scale of the question depth, and Bloom's taxonomy of cognitive difficulty. It supported by various studies related to the question generator in several publications [14], [16], [17].

Jouault, et.al [18], generated automatically have the potential to strengthen students' deep historical understanding. They use an open data link (LOD), which clarified as a learning resource in generating questions related to learning history in students. They create an open learning space and make it easier for students to access information on topics learned in natural language. The results prove that the questions generated tend to be better with an average accuracy of above 80% after evaluation through an expert.

3. RESEARCH METHODOLOGY

The question template data set in this study consisted of 93,602 pairs of questions, key phrases, and Bloom's taxonomy as training data, testing data with a ratio of 80:20, and classified in 6 levels of Bloom's taxonomy. The training data entity is a pair of question templates and key-phrases accompanied by question types. The number of types of questions used for model training was 20 variations, including duplicated data.

4. RESULT AND DISCUSSION

4.1 Preprocessing

We do vocabulary indexs to represent input and target words into word index. The selected vocabulary is only the words most often used uniquely, while we convert the other words into "unknown" tokens. We will use indexs that has got during the training model (for example, in table 1). The encoder takes the source key-phrase in a 1-of-K word vector format as input:

$$X = (x_1, \ldots, x_{T_x}), x_i \in \mathbb{R}^{K_x} \quad (1)$$

and generate question sentences from 1-of-K word vectors:

$$Y = (y_1, \ldots, y_{T_y}), y_i \in \mathbb{R}^{K_y} \quad (2)$$

Table 1: vocabulary indexs

X : Input; index to word mapping	**Y : Target; index to word mapping**
2 <start>	1 <start>
16 development	273 propose
15 process	54 ideas
1 ,	14 for
173 client	274 innovation
13 requirements	16 in
1 ,	368 client
8 creation	21 requirements
19 r	10 on
3 <end>	28 development
	30 process
	4 !
	2 <end>

4.2. QG Architecture Model.

We built the Question Generation model design through a machine learning approach using encoder-decoder techniques and attention mechanisms. We base the training process on vector operations that use the TensorFlow and NumPy libraries (see Fig 1).

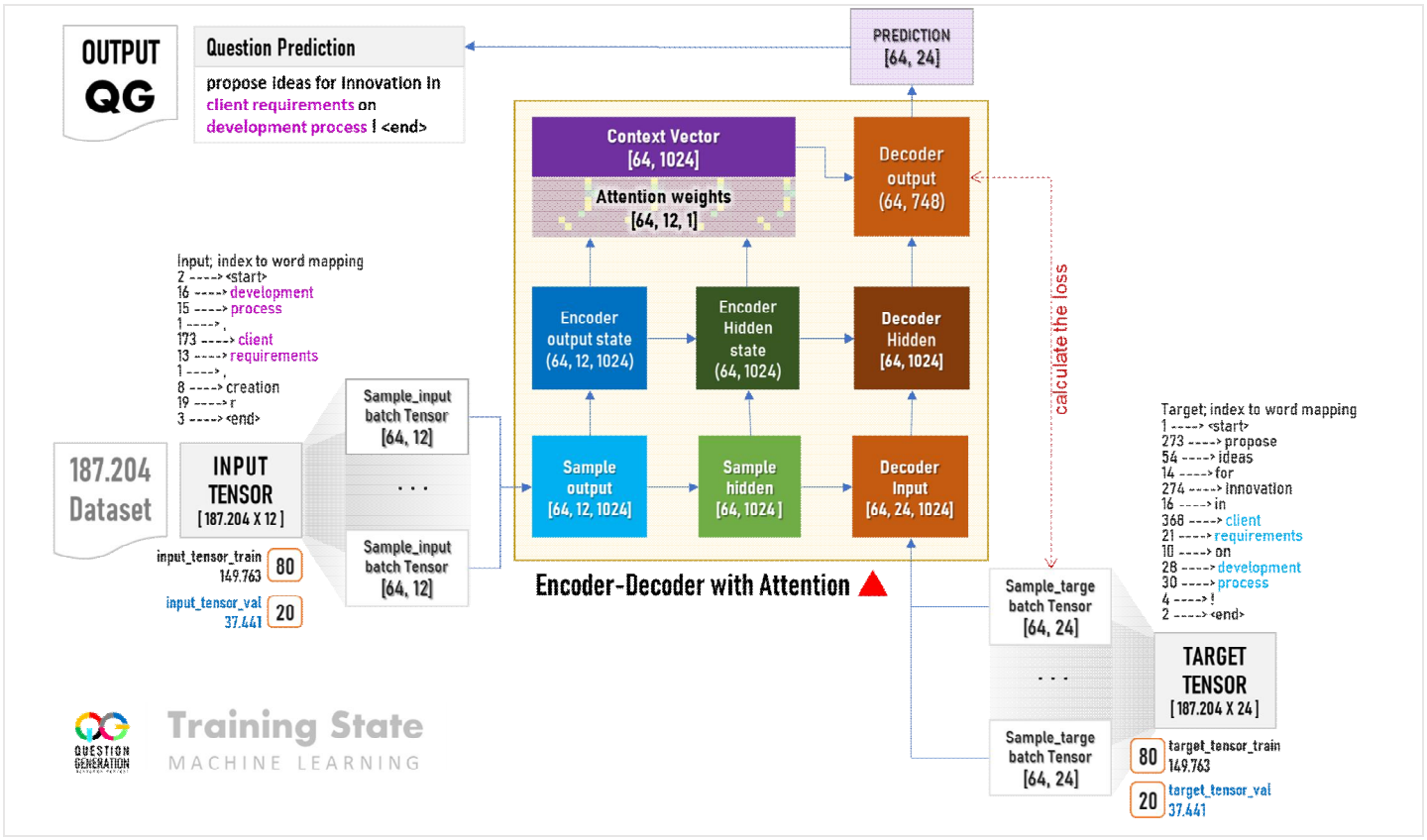

Figure 1: Question Generation architecture model

4.3. Tensor

Tensorflow in this study serves to prepare mathematical functions from computational functions or dataset calculations when the training process and predicting results. Data training in the form of input and target pairs. The target sample is in the form of a question template, while the input sample is the key phrase and level of Bloom's taxonomy. Tensorflow stores parts of words from input samples and target samples in numeric format after going through the tokenization process

The trial results show that the amount of training data determines the training results, meaning that the more training data, the better the results. It marked by the smaller the loss value. Before the training process, we separated the input sample and target sample into two parts, with a ratio of 80% for training data and 20% for testing data. In this study, we experimented several times. The ratio of training data and testing other than 80:20 was conduct, but the results of the training showed a higher loss value.

Tensor input uses dimensionless vector (187,204x12), this shows the amount of data as much as 187,604, and each has a maximum length of 12. Likewise, the target tensor uses dimensionless vector (187,204x24) because the target data is 187,204 and length 24 During the tensor input training process and Tensor targets do not allow to be processed simultaneously but must be done in stages, then divided into several batches, each batch measuring 64 data.

4.4. Encoder

The encoder in the Question Generation model is an input vector network that gets input from tensor_input and produces a feature vector to prepare data to be processed by Attention. The encoder works by mapping specific words connected with the order of other words..

The encoder dimensions are sized (64, 12, 1024) as a representation of the encoder model's configuration, i.e.

1. A batch size of 64 sentences used as training data,
2. Word embedding size 12 as a dimension size for input words, and
3. The encoder configured with 1 hidden layer measuring 1024 as a GRU (gated recurrent unit).

4.5. Attention

The introduction of knowledge base with the attention mechanism refers to [19] where knowledge can be detected dynamically with attention-based sequence-to-sequence techniques trained from the vocabulary. The attention

mechanism is a step to prepare data, especially to recognize important topics hidden in sentences..

The attention mechanism used adopts from the NMT (neural machine translation) system in [12], a detailed calculation of attention occurs at each step of the decoder process. The process of calculating attention follows the following stages.

1. Target_hidden_state is currently compared with all encoder_output_states to get attention_weight with the following formula:

$$\text{score}(\boldsymbol{h_t}, \bar{h}_s) = \begin{cases} h_t^T W \bar{h}_s & , Luong\ Attention \\ v_a^T sigmoid(W_1 \boldsymbol{h_t} + W_2 \bar{\boldsymbol{h}}_{\boldsymbol{s}}) & , Bahdanau\ Attention \end{cases}$$

$$\alpha_{ts} = \frac{exp(score(\boldsymbol{h_t}, \bar{h}_s))}{\sum_{S'=1}^{S} exp(score(\boldsymbol{h_t}, \bar{h}_s))} \ , Attention\ Weight$$

2. Based on attention_weight, context_vector is calculated as a weighted average of encoder_output_states.

$$c_t = \sum_s \alpha_{ts} \bar{\boldsymbol{h}}_{\boldsymbol{s}} \ , \qquad Context\ Vector$$

3. Combine context_vector with current Target_hidden_state to generate the last attention_vector.

$$a_t = f(c_t, \boldsymbol{h_t}) = \ sofmax(W_c[c_t; \boldsymbol{h_t}]) \ , \qquad Attention\ Vector$$

4. Attention_vector is entered as input to the next step (input feeding) in the decoder process.

4.6 Decoder

A decoder is a target vector network that takes feature vectors from the encoder through sample_hidden and encoder_hidden to produce decoder_hidden. The decoder generates a decoder_output based on calculations from decoder_hidden and attention_vector. The final step decoder calculates the closeness/match between sample_target with predicted_product. The best results from the proximity calculation if the smallest loss value is obtained based on mean-square-error (MSE), see figure 2.

This study was testing two models using the tanh and sigmoid activation functions. There is a difference in the training duration, namely: time per epoch sigmoid = 106.13 sec and time per epoch tanh = 130.72 sec. The illustration of the Loss function (4) generated by SparseCategoricalCrossentropy from hard TensorFlow. Attention can work better with the sigmoid activation function, the average (MSE) in Epoch 4 to Epoch 50, which shows a loss between 0.001 and 0.003. Based on this value, we can interpret that the prediction proportion based on training and testing is deficient.

Because of the complex nature of the language involving several vocabulary and grammar permutations, a practical model will require much data. It can be seen in the evaluation data (figure 2) where SIGMOID-D graphics that use more data can produce relatively small losses compared to others

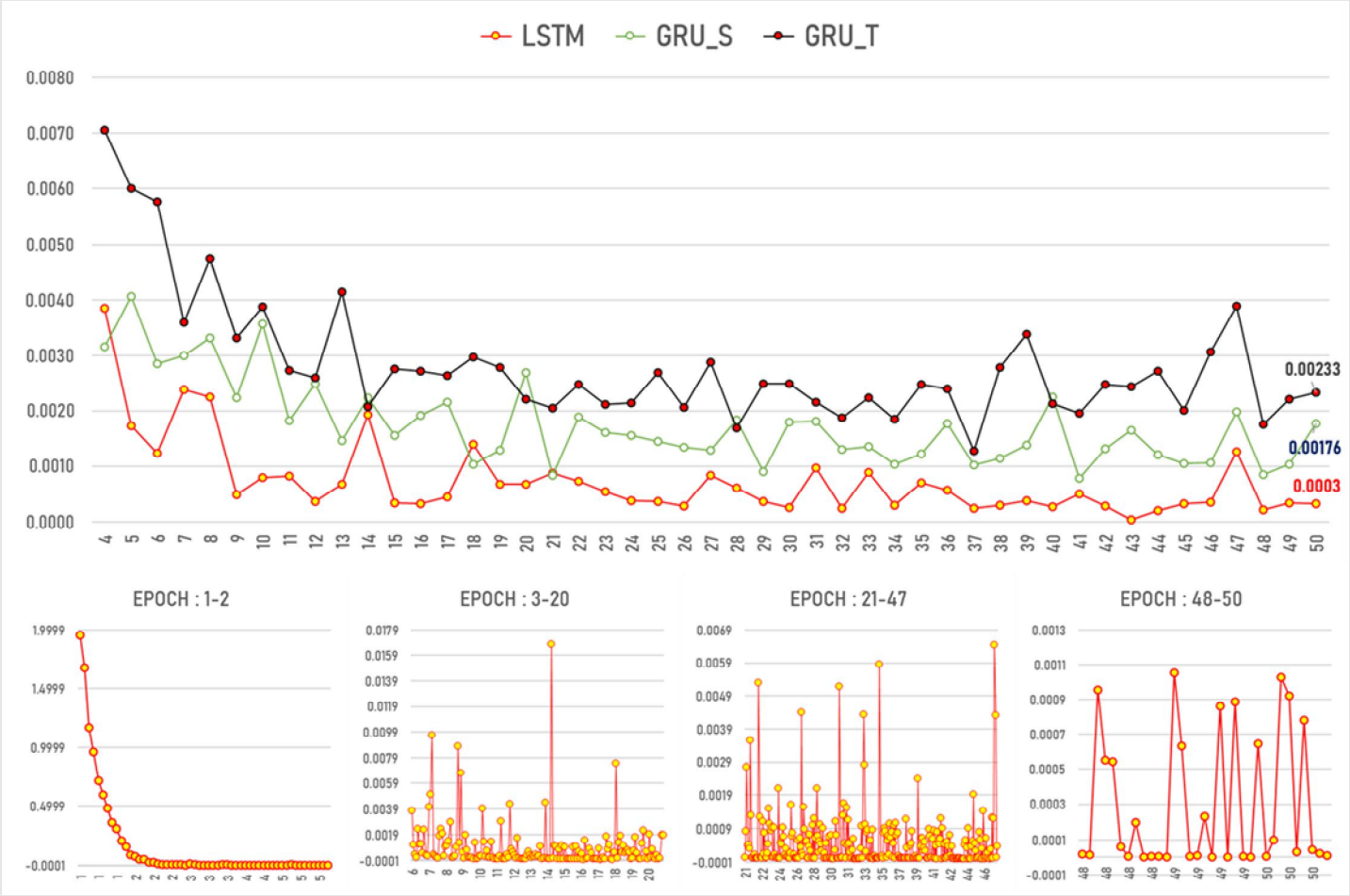

Figure 2: The Effect of Activation Function of QG Model to Training Loss (MSE)

4.7. Question Generation Results

Table 2 is 120 samples of 92,835 questions generated from the Question Generation model using an encoder-decoder and Attention mechanism. Each question has Bloom's taxonomic level that is different from the level: Remembering - Comprehension - Application - Analysis - Evaluate - Create. The question code is given to make it easier to recognize the classification of the question, for example, COM-S which means the question is at the level of comprehension with variant s. Questions can be questions that require descriptive answers or order so that the answer in the form of an action that must be done in accordance with the topic or context of the question.

Table 2: Question Generation Results from Machine learning

NUM	ID_QUESTION	QUESTION GENERATION
1	QG_83877 COM-S	What the great idea behind computational purpose on software engineering concepts ?
2	QG_55955 APP-M	What will happen if you change certain parts of scientific principles in software engineering concepts ?
3	QG_07993 COM-B	Explain how to manage reliable software product on software engineering !
4	QG_54563 COM-L	What the main idea of project management on software engineering process ?
5	QG_32620 CRE-H	How to solve software engineering process problems ?
6	QG_83814 EVA-S	Give your criticism about software engineering !
7	QG_91801 CRE-T	Give the recommended method for reviewing reliable software product on software engineering !
8	QG_55876 COM-M	What do you think of software engineering ?
9	QG_23350 EVA-F	What changes to software engineering concepts as your recommendation ?
10	QG_37318 CRE-I	How the best way to resolve problems of software engineering concepts ?

5. CONCLUSION

This study uses encoder-decoder and attention techniques with a machine learning approach until it used 93,602 pairs of questions. Evaluation results using 2 (two) test models using the tanh and sigmoid activation functions, there are differences in the training's duration, namely: time per epoch sigmoid = 106.13 seconds and time per epoch tanh = 130.72 seconds.

The lost - function illustrated generated by SparseCategoricalCrossentropy from TensorFlow. Attention can work better with the sigmoid activation function, which is the average (MSE) in Epoch 4 to Epoch 50, which shows losses between 0.001 and 0.003. Based on this value, we can interpret that the proportion of predictions based on training and testing is deficient. Because of the complex nature of language that involves several vocabulary and grammar vocabularies, a practical model will require much data.

This study shows that the Attention model with encoder-decoder to extract critical phrases in question-making scenarios trained using templates from Bloom's taxonomy can achieve competitive performance as shown by creating metric texts and evaluations by humans. Reviewers have the same assessment of performing the Question Generation model.

In future work, we plan to investigate the method of extracting answers from corpus texts based on questions generated by the machine. We also plan to investigate the performance of the question model with variations in the encoder-decoder for various fields or topics of study.

ACKNOWLEDGEMENT

This research was supported partially by the Directorate General of Research and Development, Ministry of Research, Technology, and Indonesian Higher Education, as part of the Research Doctoral Dissertation Research Grant to Binus University entitled "Generating Questions Automatically in Online Learning Using Encoders-Decoders Using Attention" with Contract number: 049 / LL3 / PG / 2020.

REFERENCES

[1] K. Salmen, O. Ali, and E. Al-moataz, “Management Science Letters,” vol. 10, pp. 3477–3484, 2020 doi: 10.5267/j.msl.2020.7.001.

[2] V. H. Pham, T. Yen, and M. Hoa, “The effects of different factors influencing the results of public investment in educational institu- tions in Vietnam,” vol. 10, pp. 3545–3552, 2020 doi: 10.5267/j.msl.2020.6.040.

[3] D. F. Murad, Y. Heryadi, S. M. Isa, and W. Budiharto, “Personalization of Study Material based on Predicted Final Grades using Multi-criteria User-collaborative Filtering Recommender System,” *Educ. Inf. Technol.*, vol. May, no. III, 2020 doi: 10.1007/s10639-020-10238-9.

[4] N. R. M. Suradi, S. Kahar, A. A. Aziz, and N. A. A. Jamaludin, “Construct validation for academic application in higher education institution (HEI) with rasch model,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 160–166, 2020 doi: 10.30534/ijatcse/2020/24912020.

[5] R. Mitkov and L. A. Ha, “Computer-aided generation of multiple-choice tests,” in *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing* -, 2003, vol. 2, pp. 17–22 doi: 10.3115/1118894.1118897.

[6] C. Yllias, H. Sadid, Y. Chali, S. A. Hasan, C. Yllias, and H. Sadid, “Towards Topic-to-Question Generation,” *Comput. Linguist.*, vol. 70, no. 8, pp. 1–20, Mar. 2015, doi: 10.1162/COLI_a_00206.

[7] O. Zahour, E. H. Benlahmar, A. Eddaoui, and O. Hourrane, “Towards a system for predicting the category of educational and vocational guidance questions using bidirectional encoder representations of transformers (BERT),” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 505–511, 2020, doi: 10.30534/ijatcse/2020/69912020.

[8] W. Hu, B. Liu, R. Yan, D. Zhao, and J. Ma, “Topic - Based Question Generation,” *Iclr 2018*, pp. 1–10, 2018, [Online]. Available: https://github.com/GajjarMihir/Topic-Based-Question-Generation.

[9] D. Adamson, D. Bhartiya, B. Gujral, R. Kedia, A. Singh, and C. P. Rosé, “Automatically Generating Discussion Questions,” in *AIED*, 2013, pp. 81–90, doi: 10.1007/978-3-642-39112-5_9.

[10] M. Heilman and N. A. Smith, “Good question! Statistical ranking for question generation,” *NAACL HLT 2010 - Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist. Proc. Main Conf.*, pp. 609–617, 2010.

[11] A. Willis, G. Davis, S. Ruan, L. Manoharan, J. Landay, and E. Brunskill, “Key phrase extraction for generating educational question-answer pairs,” 2019, doi: 10.1145/3330430.3333636.

[12] M.-T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” 2015.

[13] X. Du, J. Shao, and C. Cardie, “Learning to ask: Neural question generation for reading comprehension,” *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 1342–1352, 2017, doi: 10.18653/v1/P17-1123.

[14] V. Rus, Z. Cai, and A. C. Graesser, “Evaluation in Natural Language Generation: The Question Generation Task,” *Work. Shar. Tasks Comp. Eval. Nat. Lang. Gener. Position Pap.*, pp. 20–21, 2007.

[15] B. D. Wijanarko, D. F. Murad, Y. Heryadi, Lukas, H. Toba, and W. Budiharto, "Questions Classification in Online Discussion Towards Smart Learning Management System," in *2018 International Conference on Information Management and Technology (ICIMTech)*, 2018, no. September, pp. 251–255, doi: 10.1109/ICIMTech.2018.8528131.

[16] V. Rus, A. Graesser, and Z. Cai, "Question Generation: Example of A Multi-year Evaluation Campaign," *Work. Quest. Gener. Shar. Task Eval. Chall.*, no. January, 2008.

[17] A. C. Graesser, V. Rus, Z. Cai, and X. Hu, "Question Answering and Generation," *Appl. Nat. Lang. Process. Identification, Investig. Resolut.*, no. 1, pp. 1–16, 2009, doi: 10.4018/978-1-60960-741-8.ch001.

[18] C. Jouault, K. Seta, and Y. Hayashi, "Content-Dependent Question Generation Using LOD for History Learning in Open Learning Space," *New Gener. Comput.*, vol. 34, no. 4, pp. 367–394, Oct. 2016, doi: 10.1007/s00354-016-0404-x.

[19] M. Yu, W. Yin, K. S. Hasan, C. dos Santos, B. Xiang, and B. Zhou, "Improved Neural Relation Detection for Knowledge Base Question Answering," Apr. 2017.