ICIC Express Letters
Volume **17**, Number **9**, September **2023**

ICIC International ©2023 ISSN 1881-803X
pp. **1067–1074**

LONG SHORT-TERM MEMORY-BASED ENCODER-DECODER WITH ATTENTION MECHANISM MODEL FOR QUESTION GENERATION

Dina Fitria Murad[1,*], Bambang Dwi Wijanarko[2], Riyan Leandros[1] and Silvia Ayunda Murad[3]

[1]Information Systems Department, BINUS Online Learning
[2]Computer Science Department, BINUS Online Learning
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
bwijanarko@binus.edu; riyan.leandros@binus.ac.id
*Corresponding author: dmurad@binus.edu

[3]Information Technology Department
Universitas Islam Syekh Yusuf (UNIS)
Jl. Maulana Yusuf No. 10, Babakan. Kec. Tangerang, Kota Tangerang, Banten 15118, Indonesia
silvia.ayunda@unis.ac.id

Received September 2022; accepted December 2022

Abstract. *This experimental study aims to produce a Question Generation model using an encoder-decoder technique with long-term memory-based attention mechanisms. This is necessary in relation to the dynamics of the education system which encourages changes to the online education system and the learning quality assessment system. The biggest challenge facing the online learning assessment system is how to prepare various question bank resources. This research uses the Natural Language Processing method, where Question Generation's ability to generate large and diverse, structured, and measurable question banks provides an opportunity to be a solution to the needs of online learning assessment. The results of this study prove that the performance of the Question Generation model is proven to be better, as shown by the results of measuring text creation, namely the BLEU score = 0.9210 and the Kappa Cohen coefficient = 0.6273. The algorithm developed to identify key phrases is a new thing from the results of research in this field because it is proven to be able to contribute to the formulation of problem templates according to Bloom's taxonomy. It turns out, the LSTM-based attention mechanism plays an important role in improving the performance of the Question Generation model in generating questions for learning assessment.*

Keywords: Attention mechanism, Bloom's taxonomy, Encoder-decoder, Key phrases, Question Generator

1. **Introduction.** Digital transformation in education has an impact on revolutionary trends in the development of learning methods and media. The dynamics of educational change that leads to improving the quality of teaching makes the online learning system a strategically important choice because online learning can bring education closer to students. Quality education is often based on product assurance and learning evaluation processes, including online learning. There are fundamental differences between the evaluation system in online learning and face-to-face learning, for example, the evaluation location, evaluation time, evaluation mechanism, and participants. This difference resulted in the need for evaluation materials in the form of questions used in the two learning methods. The online learning model applied at different times and places poses a challenge in providing various questions to evaluate quality learning. Multiple questions are

DOI: 10.24507/icicel.17.09.1067

needed in the learning process and in evaluating learning to develop a broader cognitive perspective. The main problem in this research is how to create a Question Generation model to support online learning.

Several Question Generation models have been carried out using a machine learning approach, where the model is trained with pairs of questions and answers extracted from a sentence. The questions generated from these methods can be categorized as Shallow questions that ask about facts, or circumstances, or questions that can answer briefly. When mapped in taxonomy Bloom's, these questions are only at the level of remembering and understanding. Meanwhile, a Question Generation model has not been found that generates questions that require answers based on actions. Bloom's taxonomy level is at the application, analysis, evaluate, and create groups [1]. Research opportunities in the field of Question Generation are still open, especially for mapping answers to questions, extracting factual statements, and resolving ambiguities from the document context [2]. This opportunity becomes motivation in taking research opportunities [3] to find automatic Question Generation, especially in online learning [4].

The primary purpose of this research is to produce a Question Generation model for learning evaluation purposes. Question Generation can generate a variety of learning questions that are classified into the six levels of Bloom's taxonomy. Question generation builds on the encoder-decoder using short-term, long-term memory LSTM cells. This research was conducted by adopting machine translation techniques in the natural programming language NLP. Learning evaluation is part of the learning cycle or universal education system. Question Generation contributes to supporting the availability of evaluation materials in the form of questions for the learning evaluation process. Question Generation could generate various questions that have a certain level of complexity.

2. **Problem Statement and Preliminaries.**

Question Generation is a system that "automatically generates several questions from various inputs such as text, database, or semantic representation" [5]. Almost the exact definition is given by [6], which defines Question Generation as a system that functions to generate rational questions with input in the form of structured data such as databases or unstructured data such as text. From the two definitions above, it can be concluded that a Question Generation system is a system that can generate many questions related to an input sentence. The input for Question Generation can be extended to one or more sentences, paragraphs, or semantic maps.

According to [7], which examined several journal articles in the 2013-2018 range showed that the evaluation methodology in Automated Question Generation was classified into several categories: 1) according to input category, 2) based on dataset, and 3) algorithm used. Challenges in Question Generation research are still open, including 1) mapping answers to questions, 2) extracting factual statements, 3) resolving context ambiguity, 4) increasing the utility of question generation tools, 5) more profound questions and other genres, and 6) automatic evaluation metrics and use of languages other than English [8].

2.1. **Question Generation.**

Definition of Question Generation contained in [9] states that, if given the sentence x, to generate questions y related to the information in the sentence x, then y can be a sequence (sequence): $\left[y_1, \ldots, y_{|y|}\right]$. Let the length of the input sentence be M, and then x can be represented as a sequence of tokens $[x_1, \ldots, x_M]$. Question Generation is an effort to find y, such that

$$y = \arg\max_{y} P(y|x) \tag{1}$$

where $P(y|x)$ is a conditional probability function (log-likelihood) of the sequence of question predictions y, given the input x. arg max is the domain element of some processes

where the function values are maximal. In this case, arg max refers to the y part or argument where the function output $P(y|x)$ is the largest.

In the field of natural language processing, the design of the Question Generation system begins with creating questions derived from a sentence. Several references show that Question Generation has contributed to developing a QA question answering system and reading comprehension. The Question Generation research is inspired by neural machine translation, an exciting topic in natural language processing [10,11]. In general, its function is to transform answer sentences into relevant questions. For the question modeling process, the answer context and the question context are formulated mathematically to explain the technical question formation using the encoder-decoder architecture through the attention mechanism at a later stage.

2.2. **Encoder.** The encoder network is a Recurrent Neural Network (RNN) which maps input sequences into word vectors and then converts them into hidden states $h_1, \ldots, h_N$. The hidden states encoding is calculated as

$$h_t = LSTM(e(a_t), h_{t-1}) \tag{2}$$

where $e(a_t) \in \mathbb{R}^m$ represents the m-dimensional word from the word a_t and h_{t-1} is the previous hidden state. The encoder reads the input from left to right and only summarizes the information of the earlier words. They have no knowledge of the following words. In order to create a model based on the following words, the LSTM used is two-way, as suggested in [12], which consists of a forward and backward LSTM.

2.3. **Attention-based decoder.** The attention-based decoder used [13] allows the decoder to learn a specific range of input sequences during the creation task. This mechanism is similar to forming questions in mind, with humans paying attention to certain parts of a sentence and making questions about specific details. The decoder is an LSTM that takes the output of the encoder and creates a sequence of words as a query. The outcome of the encoder is called the context vector, which is used to initialize the hidden state of the decoder and encode the entire sentence. To lighten the load on the context vector, an attention mechanism is used.

Attention helps the decoder to focus on different parts of the encoder output. The output of Equation (3) is used to calculate the context vector c_t. To distinguish between the hidden state encoder and decoder, the remote state decoder is expressed as h_t the hidden state of the last layer of the encoder output of Equation (3) b_t. Thus, c_t, the sum of weight, b_t is calculated as follows:

$$c_t = \sum_{i-1}^{N} a_t(i) b_t \tag{3}$$

3. **Research Method.** The research framework in Figure 1 is structured to obtain research results and a method to answer research problems. According to the research objectives, the research question formulation consists of three critical parts to solve the main problem. The three research questions proposed for the preparation of the Question Generation design include the following processes: 1) Key phrase identification, 2) Question Generation based on Bloom's taxonomy, and 3) Question Generation modeling with a machine learning approach.

4. **Main Results.** An important goal when presenting results is to clearly indicate new (unpublished) results while appropriately citing previously published results. The discussion section aims to explain the results and show how to answer the research questions posed in the introduction. This discussion is generally carried out in stages: 1) summarizing the results, 2) discussing whether the results are expected or unexpected, 3) comparing the results with previous work, 4) interpreting and explaining the results by comparing

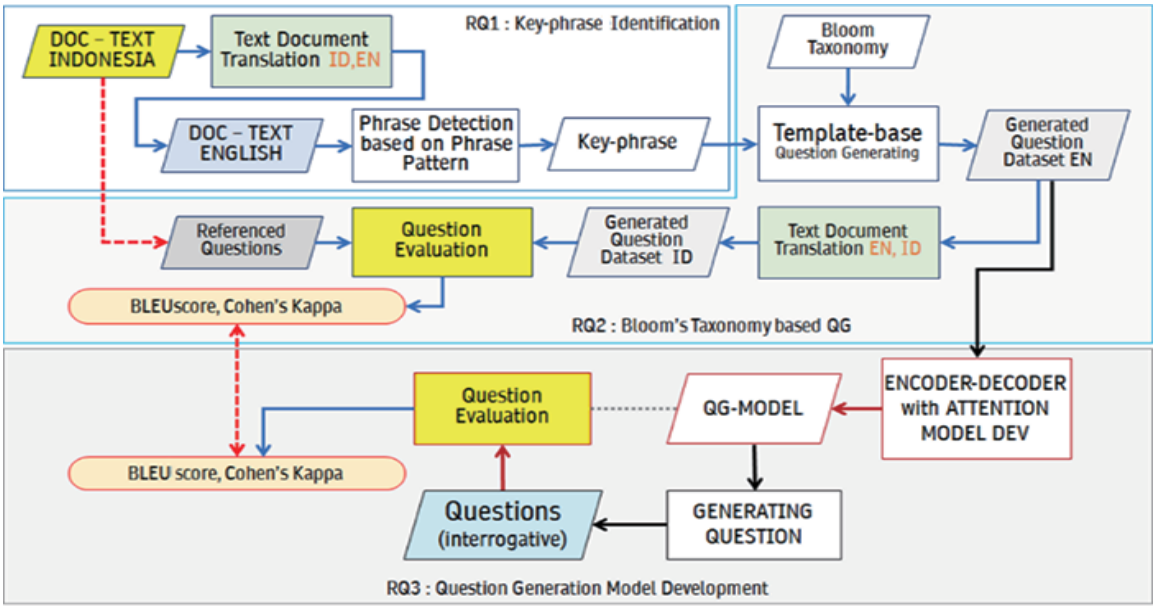

FIGURE 1. Research framework

them to theories or models, and hypotheses. The discussion section reverses the introduction format, moving from the specific (the results in this work) to the general (how these results demonstrate general principles that can be applied more broadly). Problems or shortcomings encountered during the work should also be discussed, especially if they may affect how the results are interpreted.

4.1. **Identifying key phrases.** Key phrase identification is carried out to find Context information hidden in semi-unstructured documents after going through the tokenization process and part-of-speech tagging (postag). The post tagging process for key phrases identifies whether the phrase is a noun phrase and an adjective in level one in Figure 1. The assumptions used are

- Noun-Phrase NN, in general, has a unique meaning or has a clear meaning,
- Adjective JJ cannot stand alone but as an explanation of a noun-phrase around it.

After all, the words in the sentence have a heading tagger; then this heading tag is used to identify phrase, key phrase, and Context. Context search, which is the center of the semantic meaning of a sentence or paragraph, is carried out based on one or more words in the sentence. A set of words that become the Context of a sentence is semantically interpreted as an object or type of noun (noun-phrase).

4.2. **Output of Bloom's taxonomy-based Question Generation.** The questions generated from the process of making questions based on Bloom's taxonomy with input from 778 phrase pairs and 240 templates are 93,602 questions (sample in Table 2) and refer to level two in Figure 1. Bloom's taxonomy level of each question can be identified from the verbs used although with different variations, including recognizing the presence of Context phrase in the question. An experiment using the colab google cloud to generate all questions took no less than 180 minutes.

In general, the resulting questions can be understood in Context, although there are some whose semantic meaning is still vague, so it needs to be redesigned with different techniques such as machine learning. The significance test study and evaluation of questions with Context and without Context are presented qualitatively and quantitatively using BLEU and Cohen's Kappa measurement metrics based on expert assessments in the field of software engineering.

Table 1. Input-output of QG

INPUT		Taksonomi	OUTPUT
Context	Phrase	Bloom	Generated Question based on Taksonomi Bloom
Question Generation without Context			
software engineering	software engineering	kno a	Explain what your assumes about the software engineering!
software engineering	software engineering	com a	Explain how to generalize software engineering!
software engineering	software engineering	apl a	Explain what is an example of software engineering!
software engineering	software engineering	anl a	What parts of software engineering?
software engineering	software engineering	eva a	What criteria do you use to assess software engineering?
software engineering	software engineering	cre a	What alternatives are suggested for software engineering?
scientific principles	scientific principles	kno b	Explain about definition of scientific principles!
scientific principles	scientific principles	com b	Explain how to manage scientific principles!
scientific principles	scientific principles	apl b	Explain how to present the scientific principles!
scientific principles	scientific principles	anl b	What do you infer about scientific principles?
scientific principles	scientific principles	eva b	What data used to evaluate scientific principles?
scientific principles	scientific principles	cre b	What must change to revise scientific principles?
Question Generation with Context			
software engineering	scientific principles	kno c	Explain how to identify scientific principles in software engineering!
software engineering	scientific principles	com c	Explain what can be concluded from scientific principles in software engineering!
software engineering	scientific principles	apl c	Explain the changing trends in scientific principles in software engineering!
software engineering	scientific principles	anl c	What ideas are for validating scientific principles in software engineering?
software engineering	scientific principles	eva c	Explain verification strategies for scientific principles in software engineering!
software engineering	scientific principles	cre c	How to generate the scientific principles in software engineering?

The resulting training model is then tested with unlabeled data: $x_j = (phrase_j^1, phrase_j^2, BT_j)$ to predict y_j. Each generated question has a different Bloom's taxonomy level from level: Knowledge – Understanding – Application – Analysis – Evaluation – Create. Question codes are given to make it easier to identify the classification of questions, for example, COM-S, which means the questions are at the level of understanding with variance s. The resulting question can be a question that requires a descriptive answer or a command so that the answer is an action that must be taken according to the Context of the question. This Question Generation can predict questions in two types, namely questions with Context and questions without Context; this depends on the input key phrases given in Table 2.

If the input given is a key-phrase in the form of phrase only then Question Generation will generate a question without Context, whereas if the key phrase is given consisting of Context and phrase, then the predicted question will have Context.

4.3. **Performance of Question Generation.** The question template dataset consisting of 93,602 pairs of questions and keyphrases – Bloom's taxonomy were used as training data and test data with a composition of 80%-20%. The training process works using epoch = 50 using the Tanh activation function and the Sigmoid function for five trials each, observing the Loss function. The illustration of the Loss function generated by SparseCategoricalCrossentropy from TensorFlow hard works better on attention with the Sigmoid activation function, which is an average value of 0.011, and in epoch 4 to epoch 50 shows a Loss value between 0.001 to 0.003, which means the difference in predictions based on training and testing is very low. The difference in the results of using the Tanh and Sigmoid activation functions also impacts the attention weight shown in the correlation between key phrases and words used in questions generated from the Question Generation engine and refer to level three in Figure 1.

To prove the difference between the average Loss function using GRU cells and LSTM cells with Sigmoid activation function shown in Figure 2. Experiments with the LSTM model take longer (about 6 hours for one experiment with 50 epochs) but are able to

Table 2. Question Generation results

Context	Key phrase	Bloom	Question Generation by GRU – Model#2
software engineering	software engineering	rem q	Where can you find software engineering?
software engineering	software engineering	com q	Explain how to show the software engineering!
software engineering	software engineering	apl q	How if you changes software engineering?
software engineering	software engineering	anl q	How to explain the truth of software engineering?
software engineering	software engineering	eva q	How to modify software engineering?
software engineering	software engineering	cre q	How would you improve software engineering?
software engineering	scientific principles	rem r	Write down your observational findings in scientific principles of software engineering!
software engineering	scientific principles	com r	Gather the evidence that supports for scientific principles in software engineering!
software engineering	scientific principles	apl r	What are the experimental results of scientific principles in software engineering?
software engineering	scientific principles	anl r	How do you deconstruct scientific principles to software engineering?
software engineering	scientific principles	eva r	How would you modify scientific principles to software engineering?
software engineering	scientific principles	cre r	Propose ideas for innovation of scientific principles in software engineering!
Context	Key phrase	Bloom	Question Generation by LSTM – Model#2
software engineering	software engineering	rem q	What is your favourite reference about software engineering?
software engineering	software engineering	com q	Explain the software engineering in your own words!
software engineering	software engineering	apl q	Interview your classmate about software engineering and show the result!
software engineering	software engineering	anl q	How to overcome the weaknesses of software engineering?
software engineering	software engineering	eva q	Evaluation of persuasive cases with software engineering!
software engineering	software engineering	cre q	How would you modify software engineering?
software engineering	scientific principles	rem r	Write down your observational findings in scientific principles of software engineering!
software engineering	scientific principles	com r	Gather the evidence that supports for scientific principles in software engineering!
software engineering	scientific principles	apl r	What are the experimental results of scientific principles in software engineering?
software engineering	scientific principles	anl r	How do you appraise of scientific principles to software engineering?
software engineering	scientific principles	eva r	How would you argue scientific principles to software engineering?
software engineering	scientific principles	cre r	Propose ideas for innovation of scientific principles in software engineering!

provide improvements to very good results, namely 1) there are no error questions, 2) able to generate new questions using sub-key phrases. The model training process used pairs of question sentences according to Bloom's taxonomy and keyphrases consisting of context, and question variation codes. If the model is tested with a key phrase that has not been trained before, it produces decent results.

4.4. **Bloom's taxonomy-based question performance.** The Question Generation algorithm based on Bloom's taxonomy is designed to process input in the form of a text document of learning materials totaling 683 declarative sentences into 93,602 questions belonging to 6 levels of Bloom's taxonomy. The resulting questions consist of 41,196 questions using Context and 52,406 without Context. The questions are built using 777 key phrases in the form of unique context phrases obtained from the extraction of the same input document.

Evaluation of the questions qualitatively is carried out by human reviewers who have competence in the field of study according to the discussion material used as the object of the question. The output of this qualitative evaluation process results in the tabulation of data about the suitability of the semantic relationship between words in the question candidate sentences and the number of reference sentences as their pairs. The tabulation data from the reviewer is in the form of a sign with the notation 0: for sentences that are not easily understood and 1: for interrogative sentences that can be understood easily and clearly. They were evaluated quantitatively using the BLEU metric and Cohen's Kappa

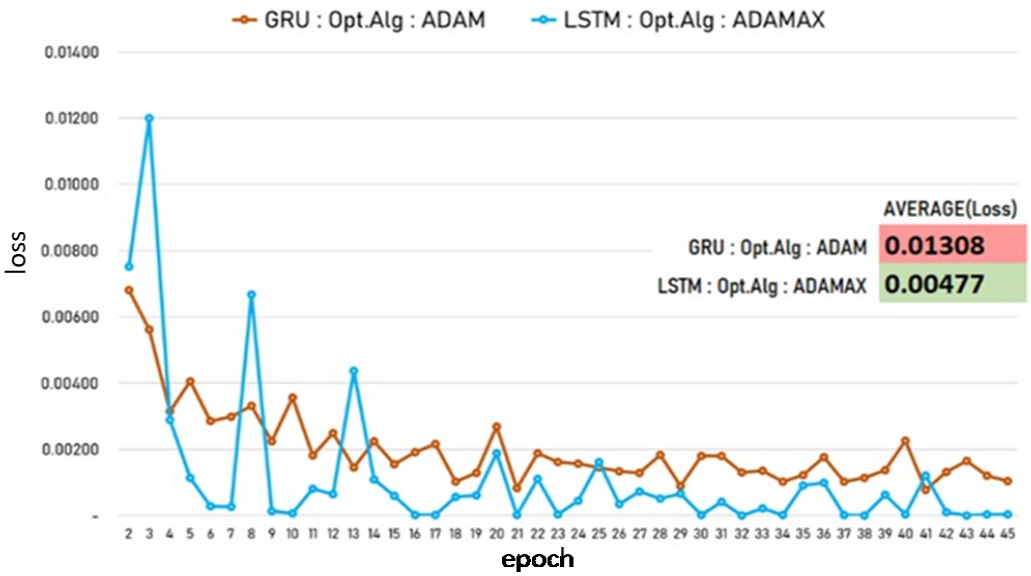

FIGURE 2. Activation function loss (MSE) model QG

TABLE 3. BLEU score and Cohen's Kappa calculation

(+) Context	N-Gram	BLEU 1	BLEU 2	BLEU 3	BLEU 4	BLEU 5	Average
	1	0.9646	0.9543	0.9510	0.9551	0.9882	0.9626
	2	0.9475	0.9318	0.9275	0.9345	0.9830	0.9449
	3	0.9346	0.9155	0.9089	0.9190	0.9786	0.9313
	4	0.9238	0.9022	0.8950	0.9072	0.9755	0.9208
	5	0.9145	0.8907	0.8832	0.8972	0.9732	0.9118
	6	0.9057	0.8801	0.8727	0.8883	0.9716	0.9037
	7	0.8984	0.8705	0.8637	0.8812	0.9715	0.8970
	8	0.8965	0.8664	0.8607	0.8800	0.9771	0.8961
	Average	0.9232	0.9014	0.8953	0.9078	0.9773	0.9210
Cohen's Kappa with Context: 0.6273							

score calculation based on tabulated data and candidate and reference sentence pairs. The selection of these two metrics is based on the state-of-the-art tools used by researchers to measure text generation performance. Table 3 shows the results of measuring BLEU and Cohen's Kappa scores obtained from 5 human reviewers on 120 candidate sentences.

The average BLEU score obtained from questions using Context 0.921 looks higher than questions without Context 0.861. This shows that the Question Generation algorithm is able to generate questions using words that are similar to the questions compiled by the reviewer when making similar questions. Cohen's Kappa scores on questions using Context of 0.6273 appear lower than questions without Context of 0.760. However, these two scores are in a good category, which means that the reviewers have a good agreement in capturing the Context of the question.

From the last experiment on questions generated from the Question Generation model using LSTM cells, the results of the BLEU and Cohen's Kappa evaluations were obtained, as shown in Table 3. BLEU's average score increased to 0.9210 and Cohen's Kappa to 0.6273. The interpretation of Cohen's Kappa value = 0.6273 shown in Table 3 is included in the substantial category according to Landis and Koch [14] or good and almost excellent according to Fleiss et al. [15]. These results indicate that the use of LSTM cells can improve the performance of the Question Generation model.

5. **Conclusions.** The Question Generation generated in this study can accommodate learning evaluation needs that rely on six levels of Bloom's taxonomy. The performance of the proposed Question Generation was tested by measuring the BLUE score and Cohen's Kappa, which involved human experts in validating the output. Thus, this Question Generation research becomes strategic value to support the achievement of specific learning objectives or the success of education in general. What is new in this research is a mechanism to find a key phrase in a sentence.

The results of this study indicate that the encoder-decoder technique that is processed with a good mechanism can find the Context of sentences in making questions that are trained using Bloom's taxonomy based questions. And, this proves that the results can achieve competitive performance. In this study, we have not used a recommendation system, so the questions and answers also have not adapted to the personalized needs of students. This is our limitation for now, but will be our next concern.

In future work, we plan to investigate methods of extracting answers from corpus text based on machine-generated questions. We also plan to investigate the performance of the question generator model with the encoder-decoder variation to extract answers from questions.

REFERENCES

[1] B. S. Bloom, *Taxonomy of Educational Objectives: The Classification of Educational Goals*, Longman Group, 1956.

[2] M. Heilman, *Automatic Factual Question Generation from Text*, Ph.D. Thesis, Carnegie Mellon University, 2011.

[3] Z. Wang, A. S. Lan, W. Nie, A. E. Waters, P. J. Grimaldi and R. G. Baraniuk, QG-Net: A data-driven question generation model for educational content, *Proc. of the 5th Annu. ACM Conf. Learn. Scale*, doi: 10.1145/3231644.3231654, 2018.

[4] B. D. Wijanarko, Encoder-decoder with attention mechanisms for developing Question Generation models in education, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol.9, no.4, pp.5994-6000, doi: 10.30534/ijatcse/2020/266942020, 2020.

[5] V. Rus, A. Graesser and Z. Cai, Question Generation: Example of a multi-year evaluation campaign, *Work. Quest. Gener. Shar. Task Eval. Chall.*, 2008.

[6] X. Yao, G. Bouma, Y. Zhang, P. Piwek and K. E. Boyer, Semantic-based Question Generation and implementation, *Dialogue & Discourse*, vol.3, no.2, pp.11-42, doi: 0.5087/dad.2012.202, 2012.

[7] J. Amidei, P. Piwek and A. Willis, Evaluation methodologies in automatic Question Generation 2013-2018, *Proc. of the 11th International Conference on Natural Language Generation*, Tilburg University, The Netherlands, pp.307-317, 2018.

[8] M. Heilman and N. A. Smith, Question Generation via overgenerating transformations and ranking, *Framework*, 2009.

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

[10] G. Neubig, Neural machine translation and sequence-to-sequence models: A tutorial, *arXiv: 1703.01619*, pp.1-65, https://doi.org/10.48550/arXiv.1703.01619, 2017.

[11] T. Baghaee, *Automatic Neural Question Generation Using Community-Based Question Answering Systems*, Master Thesis, University of Lethbridge, Canada, 2018.

[12] K. Cho, B. Van Merriënboer, D. Bahdanau and Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, *arXiv Preprint arXiv: 1409.1259*, 2014.

[13] M.-T. Luong, H. Pham and C. D. Manning, Effective approaches to attention-based neural machine translation, *arXiv: 1508.04025*, https://doi.org/10.48550/arXiv.1508.04025, 2015.

[14] J. R. Landis and G. G. Koch, An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers, *Biometrics*, pp.363-374, 1977.

[15] J. L. Fleiss, B. Levin and M. C. Paik, The measurement of interrater agreement, in *Statistical Methods for Rates and Proportions*, 1981.