Automatic question-answer pair generation using deep learning

Alok Kumar **Aditi Kharadi** **Mala Kumari** **Deepika Singh**

Dept. of Computer Science & Engineering, Chhatrapati Shahu Ji Maharaj University, Kanpur

akumar.uiet@gmail.com | aditikharadi26@gmail.com | km.malakumari.12@gmail.com
ds.deepikasingh19@gmail.com

ABSTRACT

Automatic question-answer pair generation is gaining importance because of the time being saved when compared to manual creation of questions. It not only saves time but also reduces the effort involved in the manual question generation process. It is useful in many fields such as school assignments, law practicing, self-assessments and many more. Our main objective is to create wh-type questions from a paragraph and find its accurate answer as well. We introduce a Question Answer Generation(QAG) system using a deep learning approach by combining Answer Extraction(AE), Question Generation(QG) and Question Answering(QA) models. We have used a pre-trained language model - Bidirectional Encoder Representation from Transformers (BERT) which is then fine tuned as per our objective.

Keywords: Transformers, SQuAD dataset, torch, NLP (Natural language processing), NLTK (Natural Language Toolkit), T5Tokenizer.

1. INTRODUCTION

Critical advances in Question Answering (QA) have been accomplished by pretraining profound transformer language models on unlabeled dataset, e.g. , with BERT. The pretrained question-answering model is finetuned as per the need and then used to generate question-answer pairs from paragraphs.

To get the most relevant and possible questions, the system is divided into 3 subdivisions with AE, QG, QA modalities. The main idea behind using these three models is to increase accuracy of the system, which is achieved by comparing the answers generated by the first and third model. The subdivisions are outlined in Table 1.

For a given passage P, we generate a short answer A (Step (1) in Table 1). In Step (2), we produce a question Q using A and P, at that point (Step (3)) find the answer to the question generated in Step (2) using P and Q. If A and A' match, we add the triplet (P, Q, A) as another data to our preparing model (Step (4)). We train a different model for the above mentioned three stages, and afterward apply the models in sequence on a large number of unlabeled datasets.

• *P -> A*	*Answer Extraction*
• *P,A -> Q*	*Question Generation*
• *P,Q -> A'*	*Question Answering*
• *A = A'*	*Answer Validation*

Table 1

Training and evaluation is done on SQuAD. This paper consists of 5 sections. The proposed methodology is elaborated in Section III. The evaluation of the system is shown in the Section IV of this paper, while the results and conclusion-future work are shown in Section V and VI respectively.

2. Related work

Many researchers have proposed usage of linguistic features for the task of QAG. **Heilman et al.**, proposed a rule-based approach for transforming a sentence into its interrogative form. Another paper by **Lovenia et al.** suggests usage of named entities for gap selection in a sentence, and constituency parsing for Wh-question generation.

An alternative for the traditional linguistic approach would be usage of neural networks. **Du et al.** proposed automatic question generation from text with the help of a sequence to sequence model. **Lee et al.** propose usage of hierarchical conditional variational autoencoder(HCVAE) for the generation of QA pairs from an unstructured text. Their work also involves maximization of mutual information between QA pairs and Reverse Question Answer pair Evaluation (R-QAE).

A paper by **Zhou et al.** involves usage of the neural encoder-decoder model for QAG. **Han Xiao et al.** propose a novel two-way neural sequence transduction model, connecting the question, answer and context modalities. The model uses a hierarchical attention process for capturing the cross-model interaction. The dual task of QA and QG is solved at architecture level through mirroring the structure and sharing the components between them.

Another paper by **Lakhotia et al.** suggests the use of PyText to perform fast iterations over the models. **Anderson et al.** propose a way for deeper understanding by combining the bottom-up and top-down mechanism. An alternative way for deep multilingual abstractions

is designed by **Ruder et al.** by transferring monolingual models to new languages at their lexical level. **Batra et al.** question if the deep network models look at the regions the human way. This question suggests the formation of objective specific machine attention maps for more accurate results. **Lavie et al.** describe how unigrams can be used for matching translations produced by humans and machines. **Lewis et al.** propose the generative QA models using the joint distribution of question and answer. The model is not just trained to answer a question, but to explain it. An approach mentioned by **Zhilin et al.** presents a novel training framework -Generative domain-adaptive nets for QA. They employ reinforcement learning for decreasing the variance between model-generated data and human-generated data. While it is important to form questions from context and answer them, it is equally important to know if the question is answerable. **Rajpurkar et al.** presented a new dataset SQuADRUn, which combines traditional SQuAD with 50,000 unanswerable questions. Use of such a dataset can help in training a model to recognize questions that cannot be answered from the context.

He et al. proposed usage of dual learning in machine translation(**Lample et al.**) across both directions has proved to be an efficient means to improve the quality of a model. Another way of enhancing a model like machine translation is back-translation(**Sennrich et al.**), i.e., using the generated data in training.

A paper by **Devlin et al.** introduced a new language representation model, BERT. BERT can be pre-trained on an unlabelled data. The pre-trained model can then be fine-tuned for obtaining state-of-art models. A similar yet novel language representation model ERNIE (**Sun et al.**), involves knowledge masking strategies for improved representation for the chinese language. A paper by **Alberti et al.** proposes the use of back translation for both supervised and unsupervised settings. It also suggests the use of round-trip consistency to improve the accuracy of the question-answer generation process. Another way is by analysing the attention mechanism(**Clark et al.**) of the pretrained BERT model. With a similar approach, our system uses a pretrained model, which is further fine tuned as per our objective. Several papers including **Kwiatkowski et al.**, **Noah et al.** have introduced benchmarks for evaluation of tasks like ours.

3. Proposed Methodology

Having a single model which can perform all the tasks including Answer Extraction, Question Generation and Question Answering, makes the model complex to generate. So the complete task is divided into three sub-tasks. Which are mainly,

1. extract answer from the sentence
2. generate question based on the answer and the sentence
3. generate answer for each question formed in *step2*

First answers(A) are extracted like spans from a set of sentences(paragraph) which is then used to generate the corresponding questions(Q) in the second step. Again, to improve the accuracy of the system, answers(A') are generated using the questions from the second step.

Both the answers (A and A') generated in steps (1) and (3) are matched to check the consistency and correctness of our system. For the Question-Answer pair generated to be perfect, the answers A and A' should match. Thus a Question-Answer pair is generated. The flow of the process is shown in Figure 2.

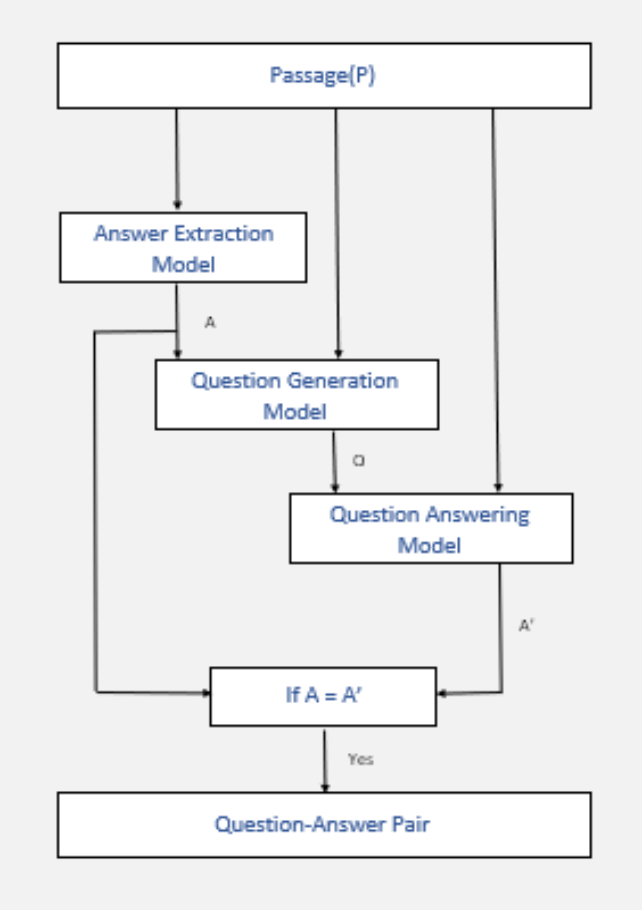

Figure 1: Process map

3.1 Data Pre-processing

For the System to support different data formats, pre-processed cached dataset is expected. Using the pre-processed dataset for training, further processes can be carried out in the way we want.

The dataset thus pre-processed contains a dictionary with values,

- source id
- target id
- Attention mask

Where, source id is the encoded source text, target id is the encoded target text and attention mask is the mask for the source id. To process the cached dataset, it uses the required tokenizer prepend or highlight formats and the respective tokens <sep> or <hl>.

3.2 Models

The task of QAG is performed with the help of following 3 models in our system:

1. Answer Extraction Model
2. Question Generation Model
3. Question Answering Model

These models are first trained on large QA data and then are applied in sequence of a large number of unlabeled text passages. We use BERT to model each of our sub-parts. A string is input to each of the distributions followed by a passage(P).

Bidirectional Encoder Representation from Transformers(BERT) is a state of the art language model or nlp that applies bidirectional training of transformers to language modelling. BERT shows that a model trained bidirectionally has more understanding of the language than the models trained in a single direction, either from left-to-right or right-to-left. It uses the concept of transfer learning- pre-training a neural network model, and then fine tuning the model to achieve the new purpose. The transformer used by BERT involves two mechanisms:

1. Encoder
2. Decoder

3.2.1 Answer Extraction Model

Answer Extraction Model takes the paragraph(P) as input and extracts answers(A) like spans (Figure 2). Answer extraction is done using methods like Named Entity Recognition or noun-phrase extraction. With the t5 model, answer extraction is done using text-to-format.

Basically, the answer in the sentence is separated by <sep> token and the sentence is highlighted with <hl> token.

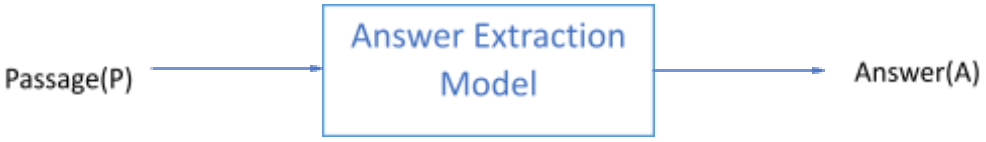

Figure 2: AE model

The above task is performed using the t5-base-qa-qg-hl model. It is a question answering and question generation model trained for both generating answer aware questions and question answering. For question generation, the answer span is highlighted with <hl> token and prefixed with *'generate-question:'*. For question answering, two prefixes are added: *'question;'* and *'context:'*. Furthermore, following are the steps involved to separate answer from the highlighted text:
1. split the text into sentences.
2. Highlight the sentence with <hl> tokens if it has an answer.
3. for the target text move the answers in that sentence to front and separate it with <sep> token.

The above processed spans are then input to our AE model.
Example,
Text : `In the past decade, machine learning has given us self-driving cars.`
Input text : `<hl>In the past decade, machine learning has given us self-driving cars. <hl>`
target text: `<hl> self-driving cars <sep> In the past decade, machine learning has given us. <hl>`

3.2.2 Question Generation Model

This model generates questions by taking the passage(P) and the answer extracted(A) by the answer-extraction model as inputs (as shown in Figure 3.). In the Question generation models, the input text can be processed in two way:

1. **prepend format**: The answer is moved to the front and separated with the context using sep token.

Example: `Machine learning [sep] Machine learning is the science of getting computers to act without being explicitly programmed.`

For the model, input is processed as shown below:

Answer: `Machine learning` *Context* : `Machine learning is the science of getting computers to act without being explicitly programmed.`

2. **Highlight format:** In this case, the answer is highlighted using <hl> highlight tokens.

Example: `<hl> Machine learning <hl> is the science of getting computers to act without being explicitly programmed.`

In this model a simple adaptation of publicly available BERT model is used which does not require any pre training but it can be directly fine-tuned and used as per our requirements by using subsets of datasets like: SQuAD2

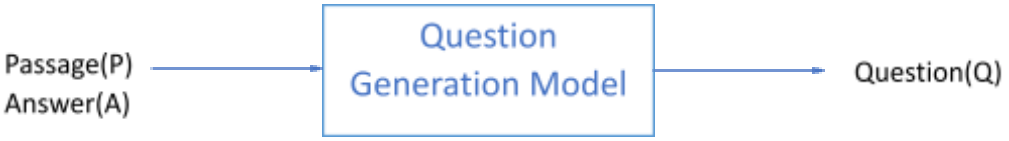

Figure 3: QG model

The model trained for this task is t5-base-qg-hl. It is the base question answering model trained for generating answer aware questions. Here the potential answer in the target text is highlighted with <hl> token. A sample input and output for the model is shown below.

Input text: `In the past decade, machine learning has given us self-driving cars.`
Output: `['What has machine learning given us in the past decade']`

3.2.3 Question Answering Model

This model generates answers using the passage(P) and the question generated(Q) by using the question-generation model(as shown in Figure 4).

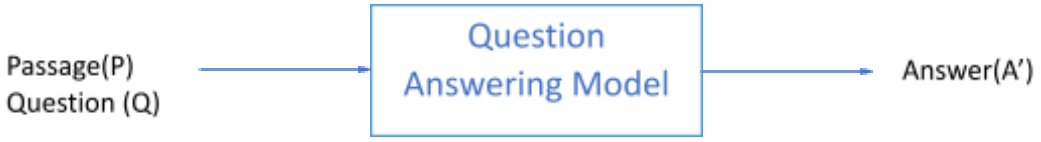

Figure 4: QA model

Example:

Input:

Context: `Machine learning [sep] Machine learning is the science of getting computers to act without being explicitly programmed.`

Question: `What is the Science of getting computers to act without being explicitly programmed?`

Output:

Answer: `Machine learning.`

4. Evaluation

For evaluating our system, we have used **S**tanford **Q**uestion **A**nswering **D**ataset (SQuAD), which is a question answering dataset, consisting of question-and-answer pairs generated from different paragraphs (like: Wikipedia articles). The benefit of using SQuAD is that it does not only answer questions generated from a paragraph, but also checks for the questions that are unanswerable.

To calculate the precision and recall values for our system's evaluation, we have tested our system against 600 paragraphs (Table 2). These paragraphs are taken from SQuAD and the output of our system is compared with question-and-answer pairs given in the dataset for the respective paragraphs. The formulae used for the calculations are listed in the equations 1,2,3.

$$Precision = \frac{No.\ of\ relevant\ question\ answer\ pairs\ generated}{Total\ no.\ of\ question\ answer\ pairs\ generated} \quad \text{.....(1)}$$

$$Recall = \frac{No.\ of\ relevant\ question\ answer\ pairs\ generated}{No.\ of\ question\ answer\ pairs\ generated\ manually} \quad \text{.....(2)}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{.....(3)}$$

The plot of mean of above metrics is shown in Figure 5.

	Precision %	Recall %	F1 Score
First Set (20 Paragraphs)	95.683	84.142	89.374
Second Set (20 Paragraphs)	93.825	74.414	82.629
Third Set (20 Paragraphs)	93.037	81.556	86.759
Average	**94.182**	**80.037**	**86.254**

Table 2: The Precision, Recall and F1-Score value after evaluation

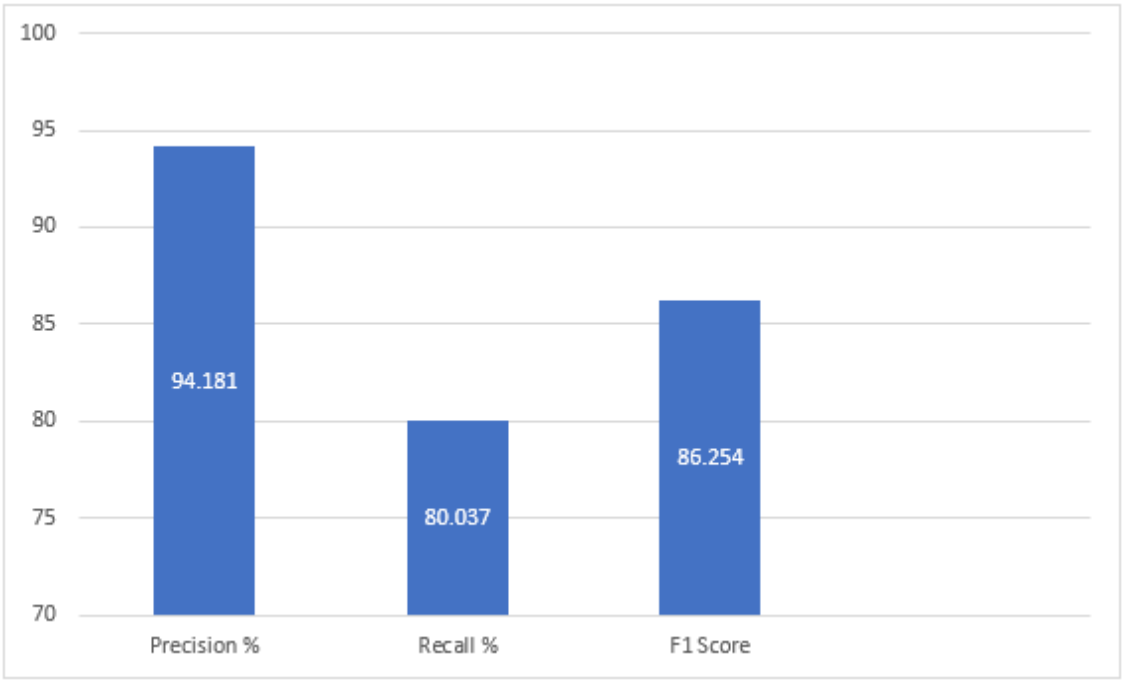

Figure 5: Distribution of average value for precision, recall and F1-score

5. Result

Our system successfully generates wh-type questions from a paragraph using the fine-tune transformers. A sample output is shown below.

```
Machine learning is the science of getting computers
 to act without being explicitly programmed. In the
past decade, machine learning has given us self-
driving cars, practical speech recognition, effectiv
e web search, and a vastly improved understanding of
 the human genome. Machine learning is so pervasive
today that you probably use it dozens of times a day
 without knowing it. Many researchers also think it
is the best way to make progress towards human-
level AI. In this class.
```

Figure 3: Input Paragraph

```
{'answer': 'Machine learning',
 'question': 'What is the science of getting computers to act without being
explicitly programmed?'},
{'answer': 'self-driving cars',
 'question': 'What has machine learning given us in the past decade?'},
{'answer': 'dozens',
'question': 'How many times a day do you use machine learning without knowing
it?'},
{'answer': 'human-level AI',
 'question': 'What do many researchers think machine learning is the best way to
make progress towards?'},
{'answer': 'the most effective machine learning techniques',
 'question': 'What will you learn about in this class?'},
{'answer': 'Machine learning',
 'question': 'What is the science of getting computers to act without being
explicitly programmed?'},
{'answer': 'self-driving cars',
 'question': 'What has machine learning given us in the past decade?'},
{'answer': 'dozens',
 'question': 'How many times a day do you use machine learning without knowing
it?'},
{'answer': 'human-level AI',
 'question': 'What do many researchers think machine learning is the best way to
make progress towards?'},
```


Figure 4: Question-Answer pairs generated by our system

6. Conclusion and Future Work

Our system successfully generates QA pairs based on the input context. System's performance improves significantly upon fine-tuning the pre-trained transformer models. Back translation is another factor crucial for better learning within the encoder-decoder architecture. Further work in the light of this approach could be generation of subjective question-answer pairs and long-answer questions.

Fusion of an automated assessment system with our QAG system would be another step in making the complete examination process easier. This will make a complete automated system and reduce the expenses associated with the educational activities.

REFERENCES

1. Ahmed Aly, Kushal Lakhotia, Shicong Zhao, Mrinal Mohit, Barlas Oguz, Abhinav Arora, Sonal Gupta, Christopher Dewan, Stef Nelson-Lindall, and Rushin Shah. 2018. Pytext: A seamless path from nlp research to production. arXiv preprint arXiv:1812.08729.
2. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 6077–6086.
3. Mikel Artetxe , Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. CoRR, abs/1910.11856.
4. Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pages 65–72.
5. Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. J. Artif. Intell. Res., 49:1–47.
6. Kevin Clark, Urvashi Khandelwal, Omer Levy and Christopher D Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. arXiv preprint arXiv:1906.04341.
7. Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding, 163:90–100.
8. Chris Alberti, Kenton Lee, and Michael Collins. 2019. A bert baseline for the natural questions. arXiv preprint arXiv:1901.08634.
9. Maria-Florina, Balcan and Avrim Blum. 2005. A pacstyle model for learning from labeled and unlabeled data. In Proceedings of the 18th Annual Conference on Learning Theory, COLT'05, pages 111– 126, Berlin, Heidelberg. Springer-Verlag.
10. Tom Kwiatkowski, Jennimaria Palomaki, Olivia Rhinehart, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. Transactions of the Association of Computational Linguistics.
11. Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5039–5049.

12. Mike Lewis and Angela Fan. 2019. Generative Question answering: Learning to answer the whole question. International Conference on Learning Representations (ICLR).
13. Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you dont know: Unanswerable questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 784–789.
14. Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. CoRR, abs/1904.09223.
15. Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1040–1050.
16. Dong Bok Lee, Annie Lee, Won Tae Jeong, Dong Hwan Kim and Sung Ju Hwang. 2020. Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs. The 58th Annual Meeting of the Association for Computational Linguistics, Pages 208-224.
17. Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. ACL.
18. Michael Heilman. 2011.Automatic factual question generation from text.International Conference on Innovation in Intelligent Systems and Application.
19. Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao and Ming Zhou. 2017. Neural Question Generation from Text: A Preliminary Study. arXiv:1704.01792 .
20. Han Xiao, Feng Wang, Jianfeng Yan and Jingyao Zheng. 2018. Dual Ask-Answer Network for Machine Reading Comprehension. arXiv:1809.01997v2 .
21. Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. NIPS. arXiv:1611.00179 .
22. Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Pages 86–96.
23. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
24. Heilman, Michael & Smith, Noah. 2010. Good Question! Statistical Ranking for Question Generation.. 609-617.
25. Lovenia, H., Limanta, F., Gunawan, A. 2018. Automatic Question-Answer Pairs Generation from Text. Technical report, Bandung Institute of Technology.