УДК 004.85

Linevych I. Yu., a 1st year Master's Student
of the Specialty «Software engineering»
Scientific Advisers: Korotun O. V., Candidate of Pedagogical Sciences, Associate Professor of Computer Science,
VakaliukTetiana, Ds.Sc. in Education, Professor,
Professor of the Software Engineering Department
Zhytomyr Polytechnic State University, Zhytomyr. Ukraine

TRAINING OF NATURAL LANGUAGE PROCESSING MODELS TO GENERATE QUESTIONS AND ANSWERS FROM THE TEXT

The automatic generation of text-based questions plays an important role in the education assessment process. Creating questions manually and checking them takes a lot of effort, time, and resources. All types of questions can be divided into two groups: objective and subjective. When answering objective questions, students are asked to choose the correct answer from the ready-made options. The most popular of this type are tasks with multiple-choice, matching, choosing true or false, filling in the blanks in a sentence, and others. At the same time, when answering subjective questions, students need to write the answer themselves. Tasks of this kind may require a short answer (from a few words to a few sentences) or a long one (for example, an essay). In this case, generating the question itself is a fairly easy task compared to checking it, because the boundaries of the correctness of the answer are quite blurred. Therefore, it is better to leave such a decision to the teacher.

The autogeneration of questions based on any text can be carried out using transformer models of natural language processing (NLP). The transformer is a deep learning model. It introduced an 'attention' mechanism that takes into account the relationship between all the words in the sentence. It creates differential weightings indicating which other elements in the sentence are most critical to the interpretation of a problem word. In this way, ambiguous elements can be resolved quickly and efficiently [1].

There are currently many ready-made natural language processing models, each of which has been pre-configured on large datasets to be able to perform specific NLP tasks. During this initial training, transformers learn different language constructions and basic functions, which is enough not to teach them from scratch.

Pre-trained models are most often used in NLP tasks because they are easier to configure, have high accuracy, and require less time for additional training compared to custom transformers. These models can be used to perform the following tasks: named entity recognition; sentiment analysis; machine translation; reduction of the text while preserving its basic meaning; sentence creation; speech recognition; generation of answers, etc. [2]. These features are quite useful and some of them can be easily used to automate the process of creating tests.

Although most ready-made transformers have all the above functionality, none of them from start can generate questions based on the analyzed text and create alternative incorrect answers that would complement the correct one. Therefore, it is important to choose a suitable NLP model and train it yourself. For that, you can use the Text-To-Text Transfer Transformer (T5) developed by Google. In T5 all «NLP tasks are transformed into a unified text-to-text format where the input and output are always text strings, in contrast to BERT-style (Bidirectional Encoder Representations from Transformers) models that can only output either a class label or a span of the input» [3].

First, to generate questions you need to find keywords in the text that will be considered potential answers. This can include names, dates, etc. Next, you need to pass the text and the found keywords to the input of the model, based on which the relevant questions will be created. For the model to be able to perform the generation correctly, it must be trained using data sets. It can include SQuAD (Stanford Question Answering Dataset), CoQA (Conversational Question Answering dataset), and MS MARCO (MAchine Reading COmprehension Dataset). These datasets contain a large number of ready-made questions and answers from various Internet resources, analyzing which, the model will learn to create their own. Also, don't forget to create fake answer options. To do this, there are many libraries that, based on the correct answer, will create synonyms, antonyms, or find other similar keywords from the text.

Thus, the training of transformer models of natural language processing is a rather laborious process, the results of which completely exceed all expectations.

References

1. S.p.A., E., 2021. Transformer NLP explained & natural language processing examples [online] Available at: https://www.eidosmedia.com/blog/technology/machine-learning-size-isn-t-everything.
2. NLP Pre-trained Models Explained with Examples [online] Available at: https://vitalflux.com/nlp-pre-trained-models-explained-with-examples.
3. Transformer, E., 2021. Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer. [online] Available at: https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html.