Retrieval-based QA system in Russian

Imangali Zhumangali

Nazarbayev University

1 Introduction

The project that was implemented is Q/A system in Russian language using retrieval-based approach. The aim if this project was to create effective system that can read and comprehend some knowledge in Russian and answer to the question based on this knowledge similar to the models that were created based on SQuAD dataset so that developers could create chat bots for citizens of post-Soviet countries.

2 Data

For this project, the SQuAD-like dataset in Russian language called SberQuAD was used. It consists of 74,300 rows of data. 45,300 rows for training, 23,900 for testing and 5,040 rows for validation. The dataset was collected via crowdworkers that created different question and answers on the Wikipedia articles.

ID	Context	Question	Answers
62,310	В протерозойских отложениях органические остатки встречаются намного чаще, чем в архейских. Они представлены известковыми выделениями сине-зелёных водорослей, ходами червей, остатками кишечнополостных. Кроме известковых водорослей, к числу древнейших растительных остатков относятся скопления графито-углистого вещества, образовавшегося в результате разложения Corycium enigmaticum.	чем представлены органические остатки?	{ "text": известковыми выделениями сине-зелёных водорослей, "answer$_s tart$” : [109]}

Таблица 1. SberQuAD sample

3 Methods

The pretrained BERT multilingual base model was used for fine-tuning in this research. It was trained on the top 104 languages with the largest Wikipedia using a masked language modeling. The HuggingFace course about NLP and its article about Question Answering was also used as a information source. Also the following libraries were used during the project implementation: Transformers (4.38.2), Pytorch (2.2.1+cu121), Datasets (2.18.0), Tokenizers (0.15.2).

For the question answering evaluation F-Score was used. It is calculated as the harmonic mean of the precision P and recall R, however in question answering task, true positive answers are the tokens shared between the correct (or gold) tokens and the all predicted tokens, false positives are the predicted tokens absent in the correct gold) answer, and false negatives are the tokens from the correct (gold) answer absent in the predicted.

$$F = 2PR/(P + R) \tag{1}$$

$$P = shared/(shared + (predicted - shared)) \tag{2}$$

$$R = shared/(shared + (gold - shared)) \tag{3}$$

4 Experiments

The Google Collab Notebook and its A-100 GPU was used for the training of the model. The following hyperparameters were used during training:

learning$_{r}ate : 2e{-}05, train_{b}atch_{s}ize : 8, eval_{b}atch_{s}ize : 8, seed : 42, optimizer : Adamwithbetas = (0.9, 0.999)andepsilon = 1e{-}08, lr_{s}cheduler_{t}ype : linear, num_{e}pochs : 3, mixed_{p}recision_{t}raining : NativeAMP, max_{l}ength = 384ofcontext.$

5 Result and analysis

As a result after experimenting with different parameters, the final result of F-score was the following: precision : 0.7377772255699819, recall : 0.6429886302111533. Therefore, the f1 score : 0.6871293430150144.

Also the model answered questions like "Who invented the Transformer architecture?” better but fare poorly when given open-ended questions like “Why is the sky blue?”

6 Conclusion

As a conclusion, the multilingual Bert and retrieval-based model showed a good results and were able to answer to the most question correctly. In the future the generative models like GPT-2 could be experimented as well.