# Tryout Work for Creating A Better System For Questions Answering

**Hao Zou**
Department of Computer Science and Engineering
University of Minnesota, Twin Cites
Minneapolis, MN 55414
zou00080@umn.edu

## 1 Summerization

This is my second part of tryout project for DPR better retriever for Open-domain Question Answering. My first tryout report was written 18 days ago before my final weeks which you can see from the link: `https://github.com/zouhao0418/zouhao0418.github.io/blob/master/DPR_tryout/QA_tryout.pdf`. For last report, I mainly focused on DPR source code (`https://github.com/facebookresearch/DPR`) to reimplement the Retriever training/Inference, Reader training/Inference step by step for Natural Questions datasets based on the original code. I also tried Haystack toolkit to train the DPR Retriever for NQ datasets from scrach and got the performance report.

For this one, I aimed to apply DPR on other datasets apart from NQ such as SQuAD and Quiz Bowl. What I've done is as following:

- I implemented DPR retriever training for SQuAD dataset based on transformers DPR from Huggingface (`https://huggingface.co/transformers/model_doc/dpr.html`) and got specific dpr model successfully, I haven't finished the evaluatation process due to the time issue. The code can be checked here **Colab** or here **GitHub**.
- Tried to train DPR retriever on Quiz Bowl dataset by transformers DPR from Huggingface but failed. For Quiz Bowl dataset, I can connect to the corresponding wiki passages but I can only extract the question id from json files while fail to find the specific question sentences corresponded to the answers.
- Tried to fine-tuning DPR retriever on Quiz Bowl dataset by Haystack but failed. The code can be checked here **Colab** or here **GitHub**

## 2 Discussion

1. This time I totally avoid using DPR source code directly which might increase the complexity of the project. DPR source code requires high computing resource while training the retriever and reader, moreover, it requires specific type of json files for retriever training which contains positive_ctxs, negative_ctxs and hard_negative_ctxs. Hence SQuAD and Quiz Bowl datasets should be converted to that form before training and the process is not easy to perform.
2. Haystack toolkit has a lot of debugging issue while fine-tuning the DPR retriever on customized datasets.
3. Huggingface transformers DPR seems like the best way for future purpose. While fine-tuning DPR on Quiz Bowl dataset, I can only find the question id in json files but failed to extract the complete question sentences. Will figure out how to solve this problem soon.

Preprint. Under review.

## 3 Dense Passage Retriever (DPR)

1. A simple method to learn dense passage and question representations from only question-answer pairs
2. Without expensive pre-training
3. Given a set of question-passage pairs (passage=100-word text chunks), learning encoders for passages and questions. As for inference, 20-30M passages in Wikipedia have been encoded into sense vectors, any new question will be encoded by question encoder and maximum inner product search (MIPS) will be performed using FAISS. To conclude, there is a gap between training and testing. Training is that encoders are learned from a small set of question-passage pairs, and at the inference time, a lot of passages are indexed first and then the MIPS search is performed.
4. How does DPR work?
    - BERT encoders for question and passage $BERT\_Q$, $BERT\_P$, mapping questions and passages to d-dimensional vectors
    - Dot product as the similarity metric and Negative log-likelihood as the loss function for training.

## 4 Acknowledge

Thanks for **Danqi Chen**'s talk at **CLunch meeting** from which I got more sense about learning representations for dense retrieval.