# DPR

Dense Passage Retrieval for Open-Domain Question Answering

# Contribution & brief

- ORQA(Lee et al. 2019)을 타겟으로 함
    - Inverse Cloze Task을 위한 additional pretraining 없이!
        - Context에서 특정 문장을 input으로 두고 다른 context와 함께 실제 포함하는 문장이 무엇인지 사전 학습하게 함.
        - 90%정도는 이렇게 하고 나머지는 그대로 -> 최소한의 지식은 제공해줌.
        - 이 pretraing이 자원이 많이 든다.
    - Passage encoder 학습이 여기서 끗. 그래서 fine tuning하는 도메인에 적용 없음.
- 그래서 이 논문은 특별한 pretraining 없이 버트 하나만으로! 특정 도메인에 fine tuning 해서! BM25을 넘는 SOTA을 만들어따!
- 그리고 retrieval의 validation metric 방법도 제시했음.
- In mini batch negative example을 사용해서 1개의 example만 네트워크에 통과시키지만 batch size의 negative case을 만드는 방법도 사용!
- BM25도 negative example 만드는데 사용!

# Training

## Training 자체는 OrQA와 유사

- Query와  가장 유사한 Context을 찾기 위해서
  - Query는 Query Encoder에 통과해서 임베딩
  - Context는 Context Encoder에 통과시켜서 임베딩
  - 두 벡터를 내적해서 코사인 유사도를 구한다.
  - 하나의 Query와 다른 모든 Context와의 유사도를 구해서 softmax로 확률화
  - Ground Truth와 비교해서 Cross Entropy Loss
- 이게 끗! 다른 pretraining이 없다! 그냥 encoder bert 두개만 학습!

# Inference

## OrQA와 유사

- Query와  가장 유사한 Context을 찾기 위해서
  - Query는 Query Encoder에 통과해서 임베딩
  - Context는 Context Encoder에 통과시켜서 임베딩
  - 두 벡터를 내적해서 코사인 유사도를 구한다.
  - 하나의 Query와 다른 모든 Context와의 유사도를 구해서 softmax로 확률화
  - argmax에 해당하는 context을 선택

# Negative Sampling

## 크게 3가지

- 1. 다른 Passage에서 random하게 뽑기
- 2. In batch negative: 현재 mini batch에서 다른 passage을 negative으로 하기
  - 계산을 재사용한다고 표현하는데, 매우 매우 효율적임.
- 3. 2에 덧붙여서 BM25에서 negative 1개 가져옴. -> 이게 성능이 제일 조아따고 한다.

# Validation Metric

- Retrieval의 Metric을 소개했음.
  - 현재 query에 대해서 top k으로 가지고 온 문서들에 실제 answer span에 해당하는 string이 있으면 True으로 해서 Accuracy을 구함.
  - BM25을 능가했다고 한다.

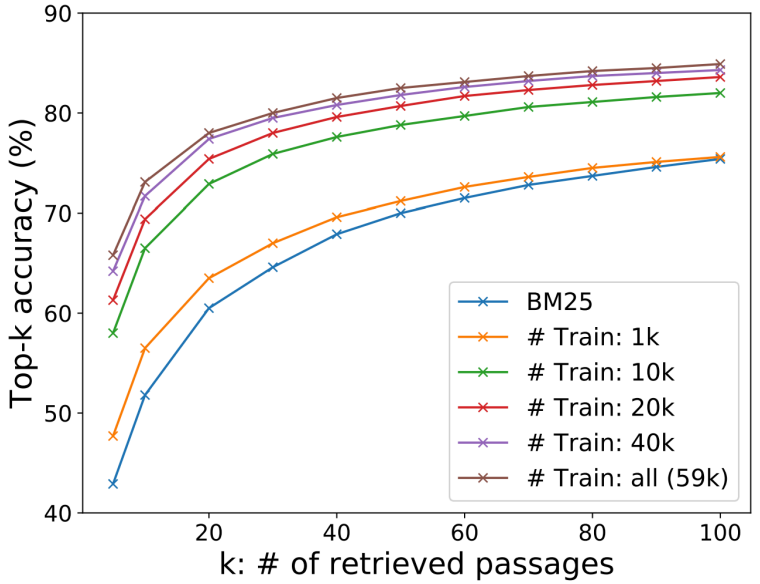


Figure 1: Retriever top-$k$ accuracy with different numbers of training examples used in our dense passage retriever vs BM25. The results are measured on the development set of Natural Questions. Our DPR trained using 1,000 examples already outperforms BM25.