# 검색 모델 성능 향상을 위한 Hard Negative 추출 및 False Negative 문제 완화 방법

박성흠[0,1], 김홍진[1], 황금하[2], 권오욱[2], 김학수[3]
건국대학교 인공지능학과[1], 한국전자통신연구원[2], 건국대학교 컴퓨터공학부[3]
{tjdgma95,jin3430}@konkuk.ac.kr, {hgh,ohwoog}@etri.re.kr, nlpdrkim@konkuk.ac.kr

**Improving Dense Retrieval Performance by Extracting Hard Negative and Mitigating False Negative Problem**

Seong-Heum Park[0,1], Hongjin Kim[1], Jin-Xia Huang[2], Oh-Woog Kwon[2], Harksoo Kim[3]
Konkuk University, Department of Artificial Intelligence[1],
Electronics and Telecommunications Research Institute[2],
Konkuk University, Department of Computer Science and Engineering[3]

**요 약**

신경망 기반의 검색 모델이 활발히 연구됨에 따라 효과적인 대조학습을 위한 다양한 네거티브 샘플링 방법이 제안되고 있다. 대표적으로, ANN전략은 하드 네거티브 샘플링 방법으로 질문에 대해 검색된 후보 문서들 중에서 정답 문서를 제외한 상위 후보 문서를 네거티브로 사용하여 검색 모델의 성능을 효과적으로 개선시킨다. 하지만 질문에 부착된 정답 문서를 통해 후보 문서를 네거티브로 구분하기 때문에 실제로 정답을 유추할 수 있는 후보 문서임에도 불구하고 네거티브로 분류되어 대조학습을 진행할 수 있다는 문제점이 있다. 이러한 가짜 네거티브 문제(False Negative Problem)는 학습과정에서 검색 모델을 혼란스럽게 하며 성능을 감소시킨다. 본 논문에서는 False Negative Problem를 분석하고 이를 완화시키기 위해 가짜 네거티브 분류기(False Negative Classifier)를 소개한다. 실험은 오픈 도메인 질의 응답 데이터셋인 Natural Question에서 진행되었으며 실제 False Negative를 확인하고 이를 판별하여 기존 성능보다 더 높은 성능을 얻을 수 있음을 보여준다.

주제어: Dense Retrieval, Hard Negative, False Negative

## 1. 서론

정보 검색은 여러가지 자연어 처리 분야에 자주 활용되는 작업으로 대화 시스템[1], 질의 응답[2] 및 챗봇[3]등 다양한 곳에서 사용되고 있다. 최근 딥러닝(Deep Learning)이 발전함에 따라 정보 검색은 신경망 기반 검색 모델(Dense Retrieval)[4]에 대한 연구로 발전되었으며, 검색 성능을 향상시키기 위한 여러가지 연구가 활발히 진행되었다. 네거티브 샘플링은 검색 성능을 효과적으로 개선시키는 방법 중 하나로, 질문과 질문에 대한 정답 문서 및 정답이 아닌 부정문서들을 네거티브로 사용하여 하나의 인스턴스로 구축하고 학습에 사용된다. 대표적인 방법으로 In-Batch sampling[4]은, 배치 내의 다른 질문의 정답 문서를 현재 질문의 네거티브로 재사용하는 방식이다. 또 다른 방법으로는 BM25[5] 혹은 Dense Retrieval과 같은 검색기를 통해 질문에 대한 상위 k개의 후보 문서를 추출하고, 그 중 정답 문서로 부착된 문서를 제외한 나머지 후보 문서들을 네거티브로 사용한다. 이를 통해 정답 문서는 아니지만 정답과 유사한 정보를 가지고 있는 네거티브를 사전에 미리 구축하고 학습에 사용하는 하드 네거티브 샘플링 방법[6, 7, 8, 9, 10]이 있다.

이러한 하드 네거티브 샘플링은 질문에 대한 정답 문서와 헷갈릴만한 네거티브를 대조학습을 통해 효과적으로 구분할 수 있도록 Dense Retrieval을 학습하기 때문에 성능 개선에 큰 영향을 끼친다. 하지만 위와 같이 검색기를 통해 질문에 대한 후보 문서들을 추출하고 하드 네거티브(Hard Negative)를 구축

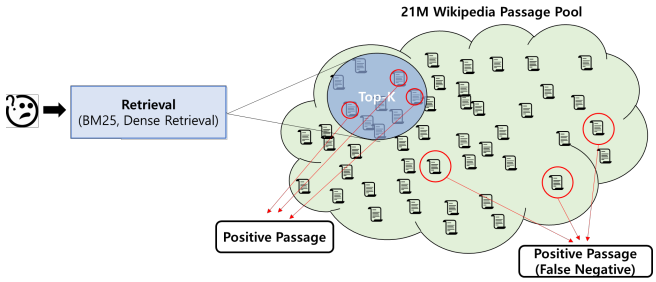


그림 1. 질문에 대한 정답 문서를 부착하는 검수 과정

하는 과정에서, 질문에 대한 정답 문서가 제대로 부착되어 있지 않으면 실제 정답을 유추할 수 있는 문서임에도 불구하고 네거티브로 간주되어 학습에 사용될 수 있다. 이는 Dense Retrieval이 대조학습(Contrastive Learning)을 통해 학습(Training)하는 과정에서 혼란을 야기시키며 따라서 성능을 완전히 이끌어내지 못하는 문제가 생긴다. 기존 연구에서는[11] 이러한 문제를 False Negative Problem으로 정의한다. False Negative Problem은 정보 검색을 위한 데이터셋을 구축하는 과정에서부터 시작된다. 그림 1을 보면, 질문에 대한 정답 문서를 부착하기 위해 임의의 구성된 문서풀(e.g. Wikipedia)에서 상위 k개의 후보 문서를 추출하고 검수자(Annotator)를 통해 질문과 관련된 문서들을 정답 문서로 라벨(Label)을 부착한다. 데이터셋 구축 과정에서 정답 문서가 더 존재함에도 상위 k개의 후보 문서에 추출되지 않았다면, 추후에 연구자들이 하드 네거티브 샘플링 전략을 활용하여 Dense Retrieval을 학습하는 과정에서 정답을

유추할 수 있는 후보 문서를 네거티브로 간주하여 학습에 사용될 수 있다. 특히 무수히 많은 문서 풀(pool)을 가지고 있는 오픈 도메인 질의 응답 데이터셋[12]에서 False Negative Problem이 더욱 두드러진다[13]. 이를 해결하기 위해 RocketQA[8]에서는 ERNIE-large[14]을 사용하여 질문에 대한 후보 문서들이 실제 질문과 관련있는 문서인지 아닌지를 분류한다.

본 논문에서는 오픈 도메인 질의 응답 데이터셋인 Natural Question(NQ)[12]에서 하드 네거티브 샘플링의 대표적인 방법 Approximate Nearest Neighbor(ANN)[6]을 기반으로 한 Dense Retrieval에 대한 성능을 보고하고 RocketQA[8]에서 사용된 판별기를 통해 False Negative Problem를 분석한다.

## 2. 관련 연구

기존 정보 검색을 위한 검색기는 단어 매칭을 활용한 통계 기반으로 BM25[5], TF-IDF등이 있다. 이는 질문과 정답 문서의 사이의 단어가 일치하는 경우에는 강력한 성능을 나타내지만, 단어가 불일치하는 경우에는 정답 문서가 존재함에도 찾지 못하는 단어 불일치 문제(Vocabulary Mismatching Problem)[4]가 발생한다. 최근 활발히 연구되고 있는 신경망 기반 검색 모델은 BERT,ERNIE[15, 14]와 같은 사전 학습 언어모델을 Encoder로 활용하여 Query encoder와 Passage encoder로 구성된 Dual-encoder를 기본 모델로 사용한다. 이는 질문과 문서에 대한 정보를 Encoder를 통해 각각 벡터로 표현하고 그 둘 사이의 내적을 통해 유사도를 구하여 문서를 검색하기 때문에 단어가 불일치하는 경우에도 의미적으로 관련된 정답 문서를 찾아낼 수 있도록 한다. 이는 단어 불일치 문제를 완화시킨다. 이러한 신경망 기반의 검색모델[4, 6]을 학습시키기 위해서는 네거티브 샘플링을 통해 질문과 질문에 대한 정답 문서, 그리고 네거티브들을 인스턴스로 구축해야 한다. 대표적인 네거티브 샘플링 방법은 In-Batch Sampling[4]으로, 배치 내에 다른 질문의 정답 문서를 현재 질문의 네거티브로 재활용하는 방법이다. 이 방법은 추가적인 사전준비 없이 질문과 그에 대한 정답 문서 쌍만 주어진다면 효율적으로 인스턴스를 구축할 수 있다. 또 다른 네거티브 샘플링 방법은 ANN(Approximate Nearest Neighbor)[6]으로, Dense Retrieval를 통해 질문에 대한 후보 문서들을 추출한 뒤 정답 문서를 제외한 나머지 문서들을 네거티브로 사용한다. 마찬가지로 BM25를 통해서도 네거티브들을 얻을 수 있다[4]. 이러한 네거티브들은 질문에 대한 정답 문서와 비슷하지만 답을 유추할 수 없는 문서들이기 때문에 하드 네거티브(Hard Negative)라고 언급되며, 이러한 하드 네거티브는 대조학습 과정을 더욱 효과적으로 만든다.

그러나 기존 하드 네거티브 메커니즘은 검색기를 통해 얻은 후보 문서가 정답을 유추할 수 있는 후보 문서임에도 질문에 대한 정답 문서로 부착되어 있지 않아 네거티브로 간주되는 False Negative Problem이 발생할 수 있다. 이는 Dense Retrieval의 성능을 완전히 이끌어내지 못하게 만든다[13]. 이러한 False Negative Problem을 해결하기 위해서 기존 연구[8]에서는 사전 학습 언어모델을 False Negative 분류기(Classifier)로 활용하고, 이를 통해 Hard Negative로 사용될 후보 문서들 중에서 실제로는 정답 문서로 쓰일 수 있는 문서인지 판별한다. 다른 연구[11]에서는 질문에 대한 각 후보문서가 얼마나 선택될 경향이 있는지 선택 가능성을 판단하는 선택 모델과 주어진 질문에 대한 각 문서들의 실제 관련성을 판단하는 관련성 모델을 학습하여 False Negative Problem을 완화하는 방법을 제시했다. 또 다른 연구[16]에서는 쉽지도, 어렵지도 않는 애매한 부정 문서(Ambiguous Negative)를 추출하는 방법을 제안하고 이를 Hard Negative로 사용하여 False Negative Problem으로 인한 Dense Retrieval의 성능 저하를 완화시키는 방법을 제안했다.

본 논문에서는 사전 학습 언어모델을 False Negative Classifier로 활용하여 ANN를 통해 학습되는 Dense Retrieval에 접목하는 방법을 제시한다. 실험 결과, False Negative Classifier가 검색 성능을 개선시키는데 도움을 주는 것을 확인했다.

## 3. 제안 방법

### 3.1 ANN 기반 Dense Retrieval

ANN은 대표적인 하드 네거티브 메커니즘으로 Dense Retrieval을 학습하는 과정에서 네거티브 문서를 주기적으로 갱신(Refresh)하는 전략이다. 이는 대조 학습에 도움이 되는 Hard Negative를 효과적으로 수집할 수 있게 만든다[6]. 그림 2의 왼쪽 모델은 Dual Encoder를 기반으로 하는 Dense Retrieval로, 질문(Query)과 그에 대한 Hard Negative를 학습 과정으로부터 일정한 Training Step마다 주기적으로 상위 K(Top-k)개의 후보 문서들을 통해 얻는 방법을 보인다. 정답 문서를 제외한 Top-k개의 후보 문서들은 Hard Negative와 False Negative가 섞여 다음 I번째 Training Step에 사용된다.

### 3.2 False Negative Classifier

False Negative Classifier은 단일 사전학습(Pre-Trained) 언어모델을 Cross Encoder로 사용하며, 학습할 데이터를 통해 미세조정(Fine-tuning)된다. Fine-tuning된 Cross Encoder는 질문과 Top-k개의 후보 문서들 간의 관계가 False Negative인지 아닌지를 구분하는 판별기로 사용된다. 그림 2의 오른쪽 모델은 False Negative Classifier로, 이전 단계에서 얻은 Top-k개의 후보 문서들이 주어지면 질문과 후보 문서들을 입력으로 받아 False Negative인지 아닌지를 판단한다. False Negative로 판단된 문서들은 Top-k개의 후보 문서들로부터 제외되며 나머지 N개의 후보 문서들이 Hard Negative로 학습에 사용된다. Cross Encoder의 입력구조는 {[CLS], Query, [SEP], Title,

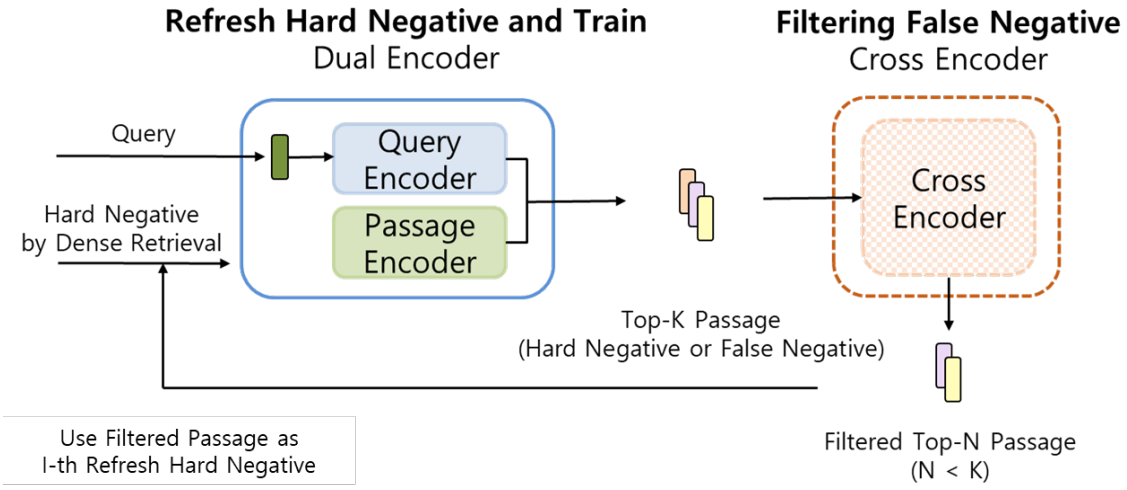


그림 2. ANN 기반 Dense Retrieval 및 False Negative Classifier 파이프라인 구조도

Candidate Passage, [SEP]}로, 질문과 후보문서의 유사도가 0.9보다 높으면 False Negative로 간주한다[8].

## 4. 실험 및 결과

### 4.1 실험 환경 및 데이터

본 논문에서 제안한 ANN에 기반한 Dense Retrieval과 False Negative Classifier의 파이프라인을 평가하고 False Negative Problem에 대한 분석을 시도하기 위해 오픈 도메인 질의 응답에서 대표적으로 쓰이는 Natural Question(NQ)[12] 데이터셋을 사용하였다. NQ 데이터셋은 구글 검색에서 쓰이는 실제 질의 응답에 기반해 구축되었으며, 그에 대한 문서 풀은 위키피디아(Wikipedia)를 통해 구성된 100 단어 크기의 21,014,324개의 문서를 다루고 있다. 구체적인 통계는 다음과 같다.

표 1. Natural Question 데이터 통계

| Dataset | Train | Dev | Test | Passage |
|---|---|---|---|---|
| Natural Question | 58,880 | 6,515 | 3,610 | 21,015,324 |

Dense Retrieval을 Fine-Tuning하기 위해 Dense Passage Retrieval(DPR)[4]을 베이스 라인으로 사용한다. 학습을 위해 8 Batch Size, 40 Epoch, 1 Gradient Accumulate, 2e-5 Learning Rate, Adam, 0.1의 Warm-up과 Drop out을 포함한 Linear Scheduling을 학습에 사용하였다. In-Batch Sampling과 Hard Negative 1개를 추가하여 학습을 진행하였으며, 이때 Hard Negative는 BM25와 Dense Retrieval(DR)로 얻은 후보 문서들을 사용했다. 특히 DR로 얻은 Hard Negative는 실험과정에서 최대 2번까지 Refresh를 진행하며, 이때 후보 문서들은(Top-k)는 상위 후보 문서 100개이다. False Negative Classifier는 RocketQA[8][1]에서 사용된 False Negative Classifier를 사용하였으며, 재구현을 위해 ERNIE-large[14]로부터 Fine-tuning된 Cross-Encoder의 가중치로 초기화했다. 이는 256 Batch Size, 2 Epoch, 1e-5 Learning Rate, 0.1의 Warm-up과 Drop out을 포함한 Linear Scheduling을 통해 학습되었다.

### 4.2 실험 결과

#### 4.2.1 네거티브 샘플링에 따른 성능 결과

표 2. 네거티브 종류에 따른 Dense Retrieval 성능

| Negative type | Refresh | Natural Question | |
|---|---|---|---|
| | | Top-20 | Top-100 |
| In-Batch | - | 67.3 | 80.3 |
| BM25 | - | 71.9 | 82.3 |
| DR | 0 | **74.0** | **82.6** |

표 2는 네거티브 종류에 따른 Dense Retrieval에 대한 성능을 보여준다. 성능 평가를 위해 NQ의 시험 데이터(Test data)을 사용하였고, 상위 k 개의 검색된 후보 문서들 중 정답 문서의 포함여부(Top-k)를 지표로 사용하였다. 실험 결과, In-Batch Negative를 학습에 사용한 검색 모델이 가장 낮은 성능을 보이고 있다. 또한 BM25와 DR 네거티브(Negative)를 Hard Negative로 추가하여 학습한 결과, 이전 In-Batch Negative를 학습에 사용한 검색 성능보다 더 높은 성능을 보여주고 있다. 특히 DR Negative를 사용한 결과가 기존보다 Top-20에서 대략 10%, Top-100에서 3%가 향상되어 가장 높은 성능을 보여주고 있다. 이는 하드 네거티브가 Dense Retrieval의 학습과정에 도움이 되는 것을 보여준다.

[1]https://github.com/PaddlePaddle/RocketQA

| Datasets | Question and Answer | Positive passage | False Negative passage |
|---|---|---|---|
| Natural Question | (Question)<br>do veins carry blood to the heart or away<br><br>(Answer)<br>to | (Passage id: 406,356)<br>Vein **Veins are blood vessels that carry blood toward the heart**. Most veins carry deoxygenated blood from the ... | (Rank 1)<br>... They are roughly grouped as ₩"arterial₩" and ₩"venous₩", determined by whether the blood in it is flowing ₩"away from₩" (arterial) or **₩"toward₩" (venous) the heart**. The term ₩"arterial blood₩" is nevertheless ... |
| Natural Question | (Question)<br>big little lies season 2 how many episodes<br><br>(Answer)<br>seven | (Passage id: 18,768,923)<br>... Production on the second season began in March 2018 and is set to premiere in 2019. **All seven episodes are being written by Kelley** | (Rank 3)<br>... after the Critics' Choice Television Award and Golden Globe Award nomination voting periods were over, HBO officially renewed the series **for a seven-episode second** |

그림 3. 분류된 False Negative 예시

#### 4.2.2 False Negative 필터링에 따른 성능 결과

표 3. DR Negative의 Refresh 및 False Negative 분류에 따른 성능

| Refresh | Negative Type | Filtered False Negative | Natural Question | |
|---|---|---|---|---|
| | | | Top-20 | Top-100 |
| 0 | DR | x | 73.99 | 82.58 |
| | | 0 | **75.73** | **84.13** |
| 1 | DR | x | 78.37 | 84.99 |
| | | 0 | **78.92** | **85.82** |
| 2 | DR | x | 78.70 | 85.57 |
| | | 0 | **79.89** | **86.15** |

표 3은 ANN 전략을 통해 DR Negative를 주기적으로 Refresh하여 검색 모델을 학습한 성능과 False Negative Classifier를 통해 False Negative 분류 작업 여부에 대한 성능을 보여준다. 실험 결과, 검색 모델 학습 과정에서 DR Negative를 여러 번 Refresh할수록 성능이 오르는 모습을 볼 수 있다. 추가적으로 False Negative Classifier를 통해 DR negative를 Refresh하는 과정에서 얻은 후보 문서들(Top-k)이 False Negative인지 분류하고 이를 하드 네거티브들(Top-n)로 학습한 결과, 분류되지 않는 채로 학습된 검색 모델의 성능보다 더 높은 성능을 확인할 수 있었다. 이는 False Negative Classifier를 통해 False Negative로 예측한 후보 문서들(k-n)이 검색 모델 학습 과정에 부정적인 영향을 미침을 보여준다.

### 4.3 필터링된 False Negative 분석

#### 4.3.1 False Negative

우리는 False Negative Problem을 분석하기 위해 False Negative Classifier로부터 분류된 후보 문서를 정성적으로 확인하였다. 그림 3은 실제 False Negative로 분류된 후보 문서들을 보여주며 노란색과 볼드체는 질문에 대한 정답을 유추할 수 있는 정보이다. 정답 문서(Positive passage)와의 비교를 통해, False Negative 또한 질문에 대한 정답을 유추할 수 있는 후보 문서임에도 정답 문서로 부착되지 않는 모습을 보여준다. 이러한 False Negative는 Dense Retrieval 학습과정에 혼란을 줄 수 있으며, 이로 인해 성능이 저하될 수 있음을 보여준다.

#### 4.3.2 False Negative Problem 비율

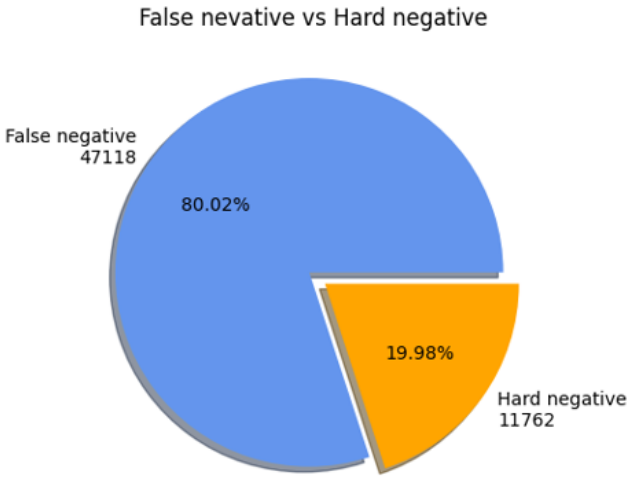


그림 4. NQ 데이터에서의 False Negative 비율

NQ데이터셋에서 False Negative Problem 발생 비율을 확인하기 위해, 58,888개의 학습 데이터(Train data)에서 False Negative의 존재 여부를 조사하였다(그림 4). 구체적으로, ANN 전략을 통해 각 질문에 대해 정답 문서를 제외한 후보 문서 100개를 추출하고 False Negative Classifier를 사용하여 그중 한 개 이상 False Negative로 분류된 후보 문서가 있다면 False Negative Problem이 있는 데이터로 판별하였다. 조사 결과, 총 58,888개중 11,762개를 제외한 47,118개의 학습 데이터에서 False Negative의 분류를 확인하였다. 이는 전체 학습 데이터의 80.02%의 수치이며 False Negative Problem이 빈번하게 발생하고 있음을 보여준다.

#### 4.3.3 질문 당 False Negative의 평균 개수

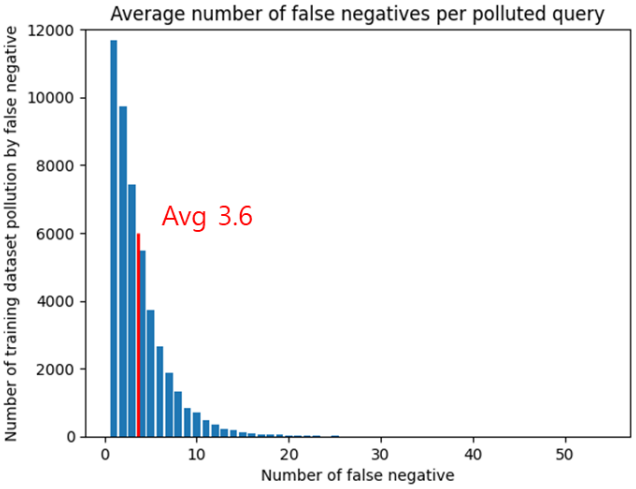


그림 5. 분류된 False Negative의 개수 분포

우리는 False Negative Problem을 가진 47,118개의 학습 데이터들 중에서 평균적으로 몇 개의 False Negative를 가지고 있는지 분석을 시도하였다(그림 5). 편의상 이러한 데이터를 오염된 질문(Polluted Query)으로 정의한다. 분석 결과, 47,118개의 오염된 질문 당 상위 후보 문서100개(Top-100) 중에서 평균적으로 3.6개의 후보 문서가 False Negative로 분류되었으며 최대 58개까지 있음을 보여준다. 이는 Dense Retrieval을 통해 학습에 도움이 되는 후보 문서를 추출하는 ANN 전략이 완전하지 않음을 보이며, False Negative Classifier를 통해 검색 성능을 저하시키는 False Negative를 필터링하여 더욱 높은 성능을 이끌어 낼 수 있음을 보인다.

#### 4.3.4 False Negative의 랭킹 순위

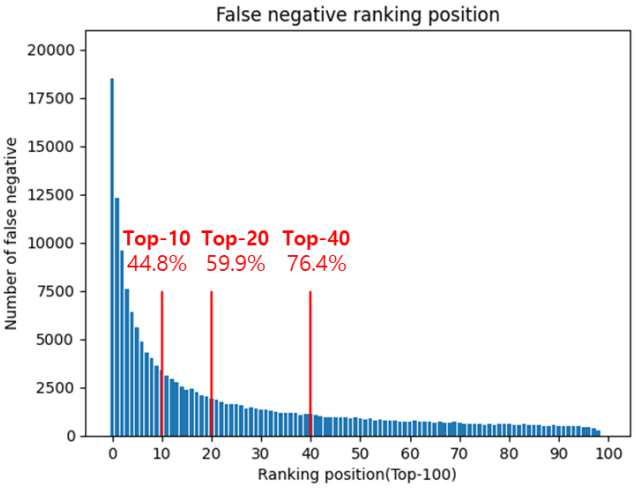


그림 6. 분류된 False Negative의 랭킹 순위 분포

마지막으로 우리는 분류된 False Negative들이 상위 후보 문서 순위(Top-100)에서 어느 랭킹 순위에 분포되어 있는지 분석을 시도하였다(그림 6). 47,118개의 오염된 질문으로부터 분석을 시도한 결과, False Negative가 Top-10 이내에는 44.8%, Top-20과 Top-40까지 후보를 늘리면 각각 59.9%와 76.4%로 대부분의 False Negative가 포함됨을 확인할 수 있었다. 특히 Top-1에서 가장 많이 위치해 있음을 보이는데, 이는 False Negative로 분류된 후보 문서가 실제로 상위 랭킹 순위에 분포되어 있음을 보여준다. 따라서 False Negative는 검색기를 통해 상위 후보 문서를 추출하고 학습에 도움이 되는 후보 문서를 얻는 하드 네거티브 메커니즘의 효과를 방해할 수 있음을 보여준다.

## 5. 결론

본 논문은 하드 네거티브 메커니즘인 ANN전략과 그로 인해 발생되는 False Negative Problem을 분석하고, Cross Encoder에 기반한 False Negative Classifier을 Dense Retrieval과 결합한 파이프라인을 제안하였다. 실험 결과, False Negative Classifier을 통해 필터링된 Hard Negative를 학습에 사용하는 것이 더 높은 성능을 얻을 수 있음을 보였다. 또한 분석 결과, False Negative로 분류된 후보 문서들이 실제로 정답을 유추할 수 있는 문서임을 보이며, 이로 인해 Dense Retrieval의 성능을 저하시킬 수 있음을 보였다.

## 감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

## 참고문헌

[1] J. Lim, M. Kang, Y. Hur, S. W. Jeong, J. Kim, Y. Jang, D. Lee, H. Ji, D. Shin, S. Kim *et al.*, "You truly understand what i need: Intellectual and friendly dialog agents grounding persona and knowledge," *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1053–1066, 2022.

[2] X. Chen, K. Lakhotia, B. Oguz, A. Gupta, P. Lewis, S. Peshterliev, Y. Mehdad, S. Gupta, and W.-t. Yih, "Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one?" *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 250–262, 2022.

[3] S. Feng, S. S. Patel, H. Wan, and S. Joshi, "Multidoc2dial: Modeling dialogues grounded in multiple documents," *Proceedings of the 2021 Conference on Empir-*

*ical Methods in Natural Language Processing*, pp. 6162–6176, 2021.

[4] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. T. Yih, "Dense passage retrieval for open-domain question answering," *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pp. 6769–6781, 2020.

[5] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp. 333–389, 2009.

[6] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. N. Bennett, J. Ahmed, and A. Overwijk, "Approximate nearest neighbor negative contrastive learning for dense text retrieval," *International Conference on Learning Representations*, 2020.

[7] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, and S. Ma, "Optimizing dense retrieval model training with hard negatives," *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1503–1512, 2021.

[8] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang, "Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5835–5847, 2021.

[9] J. Lu, G. H. Abrego, J. Ma, J. Ni, and Y. Yang, "Multi-stage training with improved negative contrast for neural passage retrieval," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6091–6103, 2021.

[10] S.-H. Park, K. Hongjin, J.-X. Huang, O.-W. Kwon, and H. Kim, "Efficient contrastive learning method through the effective hard negative sampling from dpr," *제34회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 348–353, 2022.

[11] Y. Cai, J. Guo, Y. Fan, Q. Ai, R. Zhang, and X. Cheng, "Hard negatives or false negatives: Correcting pooling bias in training neural ranking models," *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 118–127, 2022.

[12] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee *et al.*, "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 453–466, 2019.

[13] S. Wang, S. Zhuang, and G. Zuccon, "Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval," *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval*, pp. 317–324, 2021.

[14] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, No. 05, pp. 8968–8975, 2020.

[15] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[16] K. Zhou, Y. Gong, X. Liu, W. X. Zhao, Y. Shen, A. Dong, J. Lu, R. Majumder, J.-R. Wen, and N. Duan, "Simans: Simple ambiguous negatives sampling for dense text retrieval," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 548–559, 2022.