Journal of KIIT. Vol. 20, No. 7, pp. 11-17, Jul. 31, 2022. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2022.20.7.11

뉴럴 및 심볼릭 방법을 통합한 하이브리드 단락 검색

배용진*, 이공주**

Hybrid Passage Retrieval Combing Neural and Symbolic Methods

Yongjin Bae*, Kong-Joo Lee**

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2013-2-00131, (엑소브레인-총괄/1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)

요 약

뉴럴 검색은 심볼릭 검색의 미스매칭 문제를 해결하는 장점이 있으나, 여전히 특정 질문과 단락 환경에서는 심볼릭 검색의 효율이 높아 두 방법론이 병행되어 연구되고 있다. 본 논문에서는 한국어 텍스트 환경에서 두 검색 방법을 통합한, 뉴럴-심볼릭 하이브리드 검색을 제안한다. 하이브리드 검색시 사용되는 자질은 텀 빈도 기반의 심볼릭 자질 벡터와 언어모델 기반의 인코더로 생성한 뉴럴 자질 벡터이다. 유사도를 계산할 때는 생성된 두 자질 벡터를 결합하여 질문-단락 간 적합성 점수를 계산하였다. 실험 결과 단일 검색 방법 대비 위키피디아 및 오피스 문서 컬렉션에서 BR@Top1이 6.35%, 4.41% 각각 성능이 향상되었고, BR@Top50에서도 1.00%, 0.73% 성능이 향상되었다.

Abstract

Neural retrieval has the advantage of resolving the mismatch problem of symbolic retrieval, but the efficiency of symbolic retrieval is still high in a specific question and paragraph environment, so both methodologies are being studied in parallel. In this paper, we proposed a hybrid retrieval that integrates the two search methods in the Korean text. The features used in the hybrid search are a term frequency-based symbolic feature vector and a neural feature vector represented by language model based encoder. We calculated the relevance score of the question-passage by combining the two feature vectors. As a result of the experiment, the BR@Top1 improved by 6.35% and 4.41% compared to the single retrieval method in the wikipedia and office document collections, and also improved 1.00% and 0.73% via BR@Top50.

Keywords
information retrieval, deep learning, neural retrieval, question answering system

* 한국전자통신연구원 선임연구원 (교신저자)
- ORCID: http://orcid.org/0000-0002-0227-8933
** 충남대학교 전파정보통신공학과 교수
- ORCID: http://orcid.org/0000-0003-0025-4230

• Received: May 18, 2022, Revised: Jul. 07, 2022, Accepted: Jul. 10, 2022
• Corresponding Author: Yongjin Bae
Language Intelligence Research Section, ETRI, Yuseong-gu, Daejeon, Republic of Korea
Tel.: +82-42-860-6879, Email: yongjin@etri.re.kr

Ⅰ. 서 론

딥러닝과 사전학습 모델이 자연어처리 분야에 적용되면서 많은 성능 향상을 나타냈다. 검색에서도 질문-단락의 임베딩 정보와 소프트 매칭을 통한 뉴럴 검색 기법[1][2]이 소개되었으며, 기존 심볼릭 검색의 단점인 미스매칭 문제를 극복하는 방안으로 사용되고 있다. 뉴럴 검색이 심볼릭 검색의 단점을 보완하는 방법으로 대두되고 있지만, 심볼릭 검색의 필요성과 효율성에 관한 연구도 함께 병행되고 있다. 선행연구[3]는 심볼릭 검색에서 텀 빈도에 편향되게 검색되는 양상이 뉴럴 검색에서는 현저히 적어 빈도에 따른 견고함이 뛰어나다고 설명하였으나, 질문의 길이가 길어질수록 성능 하락의 폭이 심볼릭 검색보다 커 질문의 길이에 따른 성능의 견고성은 심볼릭 검색이 우세하다고 설명하였다.

또한, 질문의 유형과 관계없이 유사한 성능을 보이는 심볼릭 검색과 달리 뉴럴 검색은 일반적인 질문에 비해 엔티티 타입의 질문을 검색할 때는 성능 하락의 폭이 큰 경향이 있다고 설명하였다. 선행연구[4]에서는 질문-단락의 텀들이 정확히 일치하는 상황에서는 심볼릭 검색이 뉴럴 검색보다 성능이 높게 평가된다고 소개하였고, 질문과 단락의 적합성 정도에 따라서도 심볼릭 검색과 뉴럴 검색의 효율성 차이를 설명하였다.

본 논문에서는 질문의 길이와 유형, 적합성과 일치율에 따른 두 검색 방법의 특징을 고려하여 심볼릭 검색과 뉴럴 검색을 통합한 하이브리드 검색을 소개하며, 성능 평가는 한국어 위키피디아 컬렉션과 오피스 문서 컬렉션을 사용하여 우수성을 검증하였다. 실험 결과 위키피디아 컬렉션에서는 하이브리드 검색 성능이 BR@Top1에서 6.35% 향상된 것을 확인하였고, 문서 컬렉션에서는 BR@Top1에서 4.41% 향상된 것을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 간략히 설명하고, 3장에서는 본 논문에서 소개하는 하이브리드 검색 방법에 관해 설명한다. 4장에서는 실험환경과 성능을 비교하여 우수성을 검증하고, 5장에서는 본 논문의 결론을 내린다.

Ⅱ. 관련 연구

정보 검색과 관련한 연구는 오래전부터 진행되어 왔다. 접근 방법은 딥러닝과 사전학습 모델 소개 전후로 나눌 수 있으며, 전통적 검색 방법인 팀 빈도를 이용한 심볼릭 검색과 사전학습 모델에 기반한 뉴럴 검색이 있다.

심볼릭 검색은 단락 내 텀 빈도와 컬렉션의 빈도를 고려하여 단락의 텀 중요도를 계산하며, 질문과 매칭되는 텀의 점수를 합하여 순위화한다. 대표적인 검색 모델은 TF-IDF[5], BM25[6]가 있다. 심볼릭 검색은 리소스의 효율성과 텀 매칭을 기반으로 하기 때문에 검색 결과에 대한 설명이 가능한 장점이 있어, 현재에도 다양한 분야에서 사용되고 있다. 그러나, 텀 빈도 기반의 검색은 빈도와 매칭 여부를 판단할 때 동음 이의어와 같은 의미적 구분과 어형은 다르지만, 어휘적 뜻이 유사한 유의어에 대한 구분이 어려워 동일한 의미를 가진 텀이 포함된 질문-단락 간에 미스매칭이 발생하는 문제를 가지고 있다. 이런 문제를 해결하기 위해 언어 자원이 필요하거나 딥러닝 기반의 소프트 매칭이 요구되고 있다.

뉴럴 검색은 질문-단락을 임베딩하여 거리 유사도를 측정하는 방법으로 검색한다. 딥러닝 소개 초기에는 CNN(Convolution Neural Network), RNN (Recurrent Neural Network) 알고리즘으로 텍스트 임베딩을 수행하였다. CNN은 대상 질문-단락을 각각 임베딩하고, 풀링층을 결합하여 질문-단락 간 유사도를 계산[7]하였으며, LSTM은 단방향 혹은 양방향 네트워크를 생성하여 질문-단락을 임베딩하였으며, 네트워크 최상단은 CNN층을 추가하여 검색[8]을 수행하는 연구가 진행되었다. 그리고, 최근에는 사전학습 모델인 트랜스포머[9]의 인코더 부분만 사용한 BERT[10]를 기반으로 질문-단락을 임베딩한 후 두 임베딩 간 유사도를 측정하여 검색을 수행하는 연구[1]가 진행되었다. 뉴럴 검색은 텀 빈도 기반의 검색에 비해 문맥적 정보를 고려하여 의미적 매칭이 가능하다는 장점이 있으나, 임베딩 정보이기 때문에 질문-단락 간 매칭된 이유에 대한 직관적 설명이 어려우며, 비교적 쉬운 매칭 난이도의 질문에 대해서도 많은 리소스가 필요한 단점이 있다.

본 논문에서는 기존의 두 검색 접근 방법의 단점을 보완하고자 하이브리드 검색을 소개하며, 특징이 다른 두 도메인에 적용하였을 때 기존의 단일 검색 대비 우수함을 증명하였다.

III. 제안 방법

3.1 심볼릭 검색

심볼릭 검색은 검색 속도와 리소스 효율을 고려하여 역색인 기법으로 검색을 수행한다. BM25 랭킹 모델이 대표적인 방법이고, 질문-단락 간 유사도 계산 방법은 식 (1)과 같다.

$$BM25(Q,P) = \sum_{i=1}^{n} \frac{IDF(q_i) * freq(q_i,P) * (k+1)}{freq(q_i,P) + k*(1-b+b*\frac{m}{m_{avg}})} \quad (1)$$

여기서 q_i는 단락 P에 출현한 텀이고, $freq(q_i,P)$는 단락 P내의 q_i의 빈도를 의미한다. k, b는 단락의 길이와 빈도 정보에 부여하는 가중치이고, m, m_{avg}은 단락 P의 길이와 컬렉션의 평균 길이를 의미한다.

본 논문에서는 질문-단락 간 유사도 계산시 하이브리드 검색을 고려하여 텀의 중요도 정보를 벡터로 변환[11]하여 내적 연산을 하였다. 질문에 대한 벡터 변환은 $q^{bm25} \in [0,1]^{|V|}$로 표현한다. V는 색인시 사용되는 텀 사전이고, 질의어 내 텀이 사전 포함 여부에 따라 1, 0으로 벡터를 생성한다. 단락의 벡터 변환은 $p^{bm25} \in \mathbb{R}^{|V|}$로 변환하고, 단락의 텀 $p^{bm25}[i]$는 식 (2)와 같이 계산한다.

$$p_i = \frac{IDF(p_i) * freq(p_i,P) * (k+1)}{freq(p_i,P) + k*(1-b+b*\frac{m}{m_{avg}})} \quad (2)$$

3.2 뉴럴 검색

뉴럴 검색은 질문과 단락의 텍스트를 각각 임베딩하여 두 벡터의 거리 계산을 통해 유사도를 측정한다. 본 논문에서 뉴럴 검색은 선행연구[1] DPR에서 제안한 방법으로 수행하였다.

$$sim(q,p) = E_Q(q)^T E_P(p) \quad (3)$$

질문-단락 간 유사도는 식 (3)과 같이 계산하며, E_Q는 질문 인코더, E_P는 단락 인코더이다. 인코더를 통해 생성된 질문과 단락의 임베딩 결과는 q^{dpr}, p^{dpr}로 정의한다. 인코더로 사용한 사전학습 모델은 한국어 언어모델인 KorBERT[12]를 사용하였고, 질문과 단락을 인코더에 입력하였을 때 출력된 토큰 중 [CLS]를 임베딩 결과로 사용하였다. DPR 모델의 사후 학습시 사용한 손실 함수는 음의 로그 우도를 사용하였다.

$$L(q_i,p_i^+,p_{i,1}^-,p_{i,2}^-,\cdots,p_{i,n}^-) = -\log \frac{e^{sim(q_i,p_i^+)}}{e^{sim(q_i,p_i^+)} + \sum_{j=1}^{n} e^{sim(q_i,p_{i,j}^-)}} \quad (4)$$

3.3 하이브리드 통합 검색

하이브리드 검색은 한 번의 연산으로 심볼릭 검색과 뉴럴 검색을 통합한 스코어를 계산한다. 하이브리드 검색 스코어를 계산하기 위해 3.1장과 3.2장을 통해 질문과 단락을 두 검색 방법에 사용되는 자질 벡터로 변환하였으며, 변환된 벡터들은 식 (5)를 사용하여 유사도를 계산한다. 하이브리드 검색 전체 구성도는 그림 1과 같다.

본 실험은 한국어 텍스트를 기반으로 진행하기 때문에 효율적인 한국어 정보를 사용하기 위해 형태소 분석[12] 결과를 사용하였다. 심볼릭 자질 벡터 생성 단계에서는 질문과 단락이 입력되고 형태소 분석이 완료되면 색인어를 추출한다. 색인어의 대상은 동사, 명사와 같은 내용어를 중심으로 추출하고, 기능어는 제외하였다. 명사에 접두사나 접미사가 있으면, 하나의 색인어로 간주하여 추출하였다. 추출된 색인어는 미리 구축된 사전을 사용하여 자질 벡터를 생성하는데, 질문은 색인어의 포함 여부로 벡터를 생성하고, 단락은 식 (2)의 p_i값으로 자질 벡터를 생성한다.

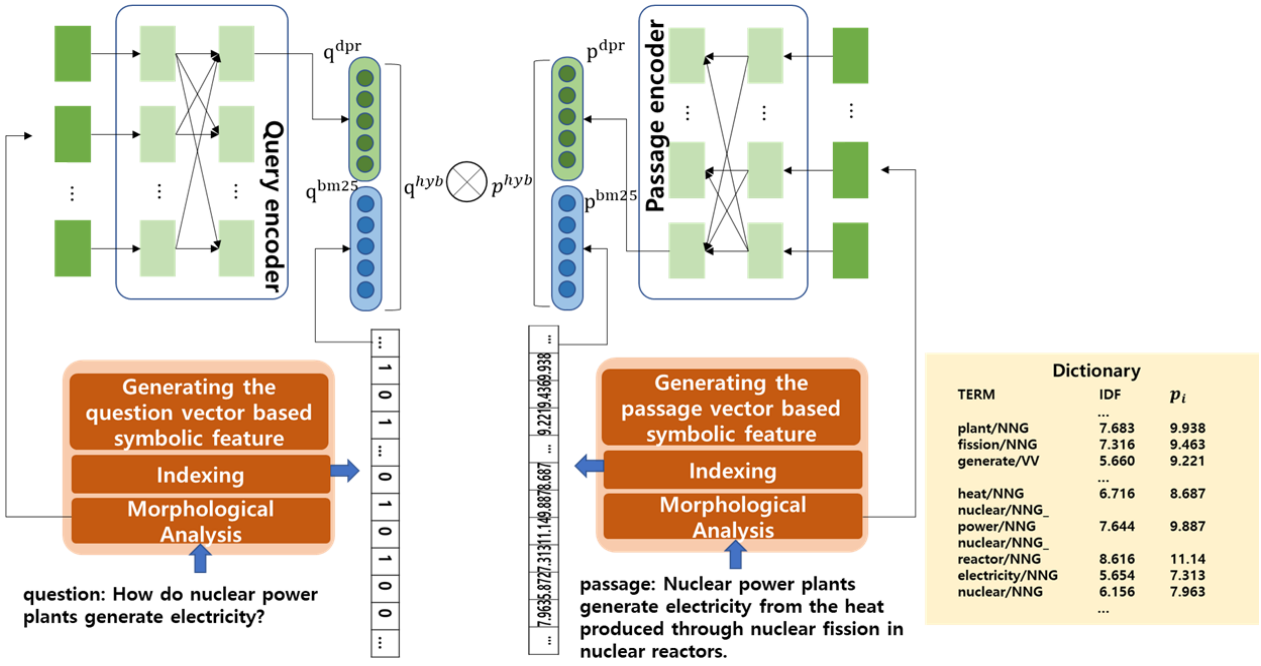

그림 1. 뉴럴-심볼릭 자질 기반 하이브리드 단락 검색의 구성도
Fig. 1. Structure of the hybrid passage retrieval based on neural and symbolic features

그림 1의 예제에서 단락 자질 벡터에 사용된 파라미터는 k는 1.1.2, b는 0.75이고, 단락의 길이는 추출된 색인어를 고려하여 10, 단락 평균 길이 22.458을 사용하여 계산하였다. 뉴럴 자질 벡터는 3.2장의 듀얼 인코더를 기반으로 생성한다. 심볼릭 자질과 동일하게 형태소 분석 결과를 입력으로 사용하며, 출력값 중 [CLS] 토큰의 결과 벡터를 사용하였다. 최종적으로 생성된 두 심볼릭 및 뉴럴 자질 벡터를 사용하여, 질문-단락 간 유사도를 계산한다. 하이브리드 검색은 전체 단락에 대해 미리 심볼릭 및 뉴럴 자질 벡터로 변환하여 저장이 가능하기 때문에 검색시 중요한 요소인 실시간성도 고려할 수 있다.

$$\begin{aligned} & sim\left(q^{hyb}, p^{hyb}\right) \\ & = \langle q^{hyb}, p^{hyb} \rangle \\ & = \langle [q^{bm25}, q^{dpr}], [p^{bm25}, p^{dpr}] \rangle \\ & = \alpha \langle q^{bm25}, p^{bm25} \rangle + \beta \langle q^{dpr}, p^{dpr} \rangle \end{aligned} \quad (5)$$

여기에서 <,>는 내적 연산을 의미하고, [,] 벡터 간 결합을 의미한다. q^{hyb}, p^{hyb}는 q^{bm25}와 q^{dpr}, p^{bm25}와 p^{dpr}의 벡터 결합한 결과이고, α, β는 두 검색기의 중요도를 반영하기 위해 사용한 가중치 파라미터이다.

Ⅳ. 평가 결과

4.1 컬렉션 정보

검색 성능 평가를 위해 컬렉션으로 한국어 위키피디아와 오피스 문서를 사용하였다. 각 컬렉션의 볼륨은 표 1과 같다.

표 1. 도메인 별 컬렉션 볼륨
Table 1. Collection volume for each domain

Domain	# of documents	# of passages
Wikipedia	519,509	8,238,117
Office document	209	16,214

위키피디아 컬렉션은 약 51만 건의 문서에서 약 823만 건의 단락을 추출하여 컬렉션을 구성하였다. 단락은 위키 페이지 내에서 줄 바꿈 구분자와 소제목과 같은 구문적 패턴을 이용하여 분할하였다.

오피스 문서는 209건의 규정/메뉴얼 문서에서 약 1.6만 건의 단락을 추출하였다. 오피스 문서에서 단락은 위키피디아 단락과 유사하게 규정 문서에서는 조항 단위로 구분하거나 메뉴얼 문서에서는 소제목과 장 번호와 같은 힌트 구분자를 사용하여 단락을

분할하였다. 오피스 문서 컬렉션은 위키피디아 컬렉션 대비 문서 건수는 적지만 일반적으로 사용하는 용어가 아닌 전문 용어로 작성된 문서이기 때문에 검색 난이도가 높다.

4.2 학습 데이터 및 평가 데이터

3.2장의 뉴럴 검색에서 인코더 학습에 사용한 데이터는 KorQuad1.0[13] 데이터로 전체 질문/정답/단락 쌍 66,301건이고, 학습 셋은 60,517건, 개발 셋은 5,784건으로 구성되어 있다. KorQuad1.0은 위키피디아 데이터를 기반으로 생성한 기계독해 데이터이기 때문에 기본적으로 정답 단락만 포함되어있다. 3.2장의 뉴럴 검색을 학습하기 위해서는 오답 단락이 추가로 필요하며 KorQuad1.0에서 제공하는 질문을 3.1장에서 구축한 심볼릭 검색기를 기반으로 검색을 수행하여 오답 단락을 쌍을 생성하였다. 오답 단락은 검색 결과 Top100에서 정답 텍스트가 포함된 단락은 제외하고 랜덤으로 30건의 단락을 추출하여 사용하여 학습하였다. 오피스 문서 임베딩을 위한 학습은 추가로 하지 않고, 위키피디아 학습 데이터로 사후학습 된 인코더를 사용하여 색인/검색을 수행하였다. 제안한 하이브리드 검색의 성능 검증을 위한 평가 데이터는 위키피디아 도메인 관련 질문 2,000건과 오피스 문서 검색을 위해 구축한 질문 3,309건을 사용하였다.

4.3 실험 환경 및 평가 측정 방법

본 논문에서 실험에 사용한 파라미터와 실험환경을 설명한다. 3.1장의 심볼릭 검색에서 사용한 파라미터 k, b는 일반적으로 사용되는 1.2, 0.75를 각각 부여하였다. 3.2장의 뉴럴 검색을 위한 인코더 학습 시 사용한 주요 파라미터는 학습 배치 크기 16, 학습율 2e-05, epoch 40으로 학습하였고, 마지막 epoch 모델을 인코더로 사용하였다. 학습시 질문/정답 단락/오답 단락은 1:1:1 비율로 학습하였고, 오답 단락은 매 epoch 마다 오답 단락 집합으로부터 랜덤 선택하여 학습하였다. 뉴럴 검색과 하이브리드 통합 검색은 python 언어로 래핑 된 Faiss 1.7.2 라이브러리를 사용하여 실험하였으며, 임베딩 된 벡터의 정보 무손실 유사도 계산을 위해 IndexFlatL2 함수를 사용하였다. 뉴럴 검색의 인코더를 위한 학습은 torch1.4, transformers3.1 환경에서 수행하였다. 3.3장의 하이브리드 검색을 위한 검색기의 중요도는 위키피디아 컬렉션에서는 α, β를 2.0, 1.0으로 사용하였고, 오피스 문서 컬렉션에서는 0.9, 0.1로 사용하였다. α, β에 대한 중요도는 실험적으로 평가하였다.

검색 성능 측정 방법은 BR(BinaryRecall)@TopN을 사용하였다. BR@TopN은 상위 검색 결과 N개 이내에 정답 단락의 포함 여부를 판단하여 검색 성능을 평가하였으며 식 (6)과 같다.

$$
\begin{aligned}
&BinaryRecall@TopN \\
&\cdot BinaryRecall = \\
&\quad \frac{\#of\ \ Q\ \ including\ \ correct\ \ passage}{\#of\ \ total\ \ \ Q} \\
&\cdot TopN = Ranked\ top\ N\ passages
\end{aligned}
\tag{6}
$$

4.4 평가 결과

위키피디아 컬렉션에서 단일 검색의 성능과 하이브리드 검색 성능은 표 2와 같다.

표 2. 위키피디아 컬렉션 BR@TopN 검색 성능 평가
Table 2. Performance of retrieval via BR@TopN measure in wikipedia collection

BR@TopN	Symbolic retrieval	Neural retrieval	Hybrid retrieval
Top1	64.85%	56.00%	71.20%
Top5	84.75%	81.50%	89.35%
Top10	90.00%	88.10%	92.65%
Top30	93.40%	93.45%	95.75%
Top50	94.85%	95.60%	96.60%

단일 검색의 BR@Top1에서는 심볼릭 검색이 64.85%로 뉴럴 검색의 56.00%보다 8.85%높은 성능을 나타내었다. 이와는 반대로 BR@Top50에서는 뉴럴 검색이 심볼릭 검색보다 0.75%높은 성능을 나타내었다. 심볼릭 검색이 랭킹 정확도에서는 성능이 높으나, 전체 재현율은 뉴럴 검색 높게 평가되었다. 이 두 검색 방법을 통합한 하이브리드 검색은 BR@Top1에서 71.20%로 심볼릭 검색보다 6.35%가

높은 성능을 나타내었고, BR@Top50에서는 뉴럴 검색보다 1.00% 높은 성능을 나타내었다. 두 검색의 장점이 결합되어 검색의 정확률도 상승하면서 미스매칭 문제도 일부 해결되어 전체 재현율까지 동반 상승한 것을 실험 결과로 알 수 있다.

오피스 문서 컬렉션에서 실험한 결과는 표3과 같으며, 검색 성능의 경향성은 위키피디아 컬렉션에서의 결과와 유사하게 나타났다.

표 3. 오피스문서 컬렉션 BR@TopN 성능 평가
Table 3. Performance of retrieval via BR@TopN measure in office document collection

BR@TopN	Symbolic retrieval	Neural retrieval	Hybrid retrieval
Top1	67.48%	63.07%	71.53%
Top5	80.42%	87.22%	88.73%
Top10	86.40%	93.35%	93.90%
Top30	92.78%	96.55%	98.16%
Top50	95.71%	97.73%	98.46%

BR@Top1에서 심볼릭 검색이 67.48%로 뉴럴 검색보다 4.41%높은 성능을 나타내었고, BR@Top50에서는 뉴럴 검색이 심볼릭 검색보다 2.02%높은 성능을 나타내었다. 하이브리드 검색은 단일 검색기의 최대 성능과 비교했을 때 BR@Top1에서는 4.05%가 높은 성능을 나타냈고, BR@Top50에서는 뉴럴 검색보다 0.73%높은 성능을 나타내었다.

두 컬렉션을 통해 실험한 결과 단일 검색에서는 심볼릭 검색이 뉴럴 검색 대비 검색의 정확도 측면에서 높은 성능을 나타냈고, 뉴럴 검색은 전체 재현율 측면에서 심볼릭 검색 대비 높은 성능을 나타냈다. 그리고, 두 검색기를 통합한 하이브리드 검색은 각 컬렉션의 BR@TopN의 모든 구간에서 단일 검색보다 높은 성능을 나타내, 하이브리드 검색이 단일 검색보다 성능상에서나 견고성에서 성능 측면이나 견고성 측면에서 우수한 것을 알 수 있었다.

V. 결론 및 향후 과제

본 논문에서는 심볼릭 검색과 뉴럴 검색의 통합한 하이브리드 검색을 소개하였다. 실험 결과 위키피디아 컬렉션에서는 BR@Top1 6.35% 오피스 문서 컬렉션에서는 BR@Top1 4.41%으로 단일 검색 성능 대비 향상된 것을 확인할 수 있었고, 하이브리드 검색 방법이 전체 재현율 측면에서도 효과적인 것을 실험으로 확인하였다. 후속 연구로 각 검색 방법과 다양한 검색 난이도 상황을 고려하여 특징을 상세 분석할 계획이다.

References

[1] V. Karpukhin, B. Oguz, S. Min, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense Passage Retrieval for Open-Domain Question Answering", Proc. of the EMNLP 2020, pp. 6769-6781. Nov. 2020. https://doi.org/10.48550/arXiv.2004.04906.

[2] R. Nogueira and K. Cho, "PASSAGE RE-RANKING WITH BERT", https://arxiv.org/pdf/1901.04085. Apr. 2020. https://doi.org/10.48550/arXiv.1901.04085.

[3] H. Padigela, H. Zamani, and W. Bruce Croft, "Investigating the Successes and Failures of BERT for Passage Re-Ranking", May. 2019. https://doi.org/10.48550/arXiv.1905.01758.

[4] D. Rau and J. Kamp, "How Different are Pre-trained Transformers for Text Ranking?", Apr. 2022. https://doi.org/10.48550/arXiv.2204.07233.

[5] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries", Proc. of the ICML 2003, Washington D.C., USA, pp. 133-142, Aug. 2003.

[6] B. Aklouche, I. Bounhas, and Y. Slimani, "BM25 beyond query-document similarity", Proc. of the SPIRE 2019, Segovia, Spain, pp. 65-79, Oct. 2019. http://doi.org/10.1007/978-3-030-32686-9_5.

[7] M. Nicosia and A. Moschitti, "Semantic Linking in Convolutional Neural Networks for Answer Sentence Selection", Proc. of the EMNLP 2018, Brussels, Belgium, pp. 1070-1076, Nov. 2018. http://doi.org/10.18653/v1/D18-1133.

[8] Z. Li, J. Huang, Z. Zhou, H. Zhang, S. Chang, and Z. Huang, "LSTM-based Deep Learning

Models for Answer Ranking", Proc. of the DSC, Changsha, China, pp. 90-97, Jun. 2016. http://doi.org/10.1109/DSC.2016.37.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", Proc. of the NIPS 2017, CA, USA, pp. 6000-6010, Dec. 2017.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proc. of the NAACL, Minneapolis, USA, pp. 4171-4186, Jun. 2019. http://dx.doi.org/10.18653/v1/N19-1423.

[11] J. Ma, I. Korotkov, K. B. Hall, and R. T. McDonald, "Hybrid First-stage Retrieval Models for Biomedical Literature", CLEF 2020, Thessaloniki, Greece, Sep. 2020.

[12] ETRI AIOpen, http://aiopen.etri.re.kr/service_dataset.php. [accessed: Sep. 23, 2021]

[13] KorQuad1.0, https://korquad.github.io/KorQuad%201.0. [accessed: Apr. 30, 2022]

[14] Faiss, https://github.com/facebookresearch/faiss. [accessed: Apr. 30, 2022]

저자소개

배 용 진 (Yongjin Bae)

2012년 : 목원대학교 컴퓨터교육과(공학사)
2014년 : 과학기술연합대학원(UST) 컴퓨터소프트웨어 및 공학(공학석사)
2014년 ~ 현재 : 한국전자통신연구원 선임연구원
관심분야 : 정보검색, 질의응답, 딥러닝

이 공 주 (Kong-Joo Lee)

1992년 : 서강대학교 전자계산학과 (공학사)
1994년 : 한국과학기술원 전산학과(공학석사)
1998년 : 한국과학기술원 전산학과 (공학박사)
1998년 ~ 2003년 : 한국마이크로소프트(유) 연구원
2003년 : 이화여자대학교 컴퓨터학과 대우전임강사
2004년 : 경인여자대학 전산정보과 전임강사
2005년 ~ 현재 : 충남대학교 전파정보통신공학과 교수
관심분야 : 자연어처리, 기계번역, 정보검색, 정보추출