Journal of KIIT. Vol. 20, No. 1, pp. 11-17, Jan. 31, 2022. pISSN 1598-8619, eISSN 2093-7571 11
http://dx.doi.org/10.14801/jkiit.2022.20.1.11

질의어 확장을 이용한 단어 중요도 기반 단락 검색

배용진*, 이공주**

Passage Retrieval based on Word Importance using Query Expansion

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

자연어 질의응답 시스템은 높은 검색 재현율과 정답을 추출하는 기계독해의 높은 정확률이 필요하다. 최근의 연구들은 대형 언어 모델을 기반으로 정확한 정답 추출에 초점을 맞추고 있으나, 1차적으로 정답단락을 검색하는 성능이 만족되어야 한다. 본 논문에서는 딥러닝의 문맥 정보를 고려한 전통 검색 방법을 제안한다. 문맥 정보를 고려하여 단어의 중요도를 추론하였으며, 단어의 중요도는 빈도 표현으로 변환하여 색인하였다. 단어 중요도 추론 모델은 정확한 단어 매칭 뿐 아니라 질문의 유의어 단어도 포함하여 기존의 제한적인 학습 대상을 확장하였다. 실험결과 일반상식 및 법률 도메인에서 선행연구 대비 MRR 평균 2.02%, 0.36% 향상되었으며, BR@Top1평가에서도 1.34%, 0.9% 향상되었다.

Abstract

Natural language question answer system requires the high search recall of the passage and the high precision of machine reading comprehension to extract the answer. Recent studies have focused on extracting the correct answer based on the large-scale language model, but the performance of searching for the correct answer passage should be satisfied primarily. In this paper, we propose a traditional retrieval system considering contextual information of deep learning. The importance of a word was inferred by considering the context of paragraph, and the word importance was indexed by converting it into a frequency. The word importance inference model extends the limited training targets of previous study including not only exact word matching but also synonym words of the question. As a result of the experiment, the average MRR improved by 2.02% and 0.36% compared to previous study in the common sense and legal domains, and also improved 1.34% and 0.9% via BR@Top1.

Keywords
information retrieval, deep learning, data augmentation, question answering system

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

· Received: Oct. 07, 2021, Revised: Jan. 07, 2022, Accepted: Jan. 10, 2022
· Corresponding Author: Kong-Joo Lee,
Dept. of Raid and Information Comm. Eng., ChungNam National University, Yuseong-gu, Daejeon, Republic of Korea,
Tel.: +82-42-821-6881, Email: kjoolee@cnu.ac.kr

Ⅰ. 서 론

최근의 텍스트 기반 질의응답 시스템은 일반적으로 단락 검색, 단락 재순위화, 기계독해 파이프라인을 걸쳐 질문에 응답한다. 딥러닝 알고리즘이 소개된 이후 단락 재순위화와 기계독해 모델의 정확도가 높아져, 질의응답 시스템의 성능이 향상되었다는 연구들이 소개되고 있다. 질의응답 시스템의 첫 번째 단계인 단락 검색 단계에서도 딥러닝 기술을 적용[1]하여 검색의 정확률과 재현율이 향상되고 있으나, 컬렉션 크기에 따라 연산량이 비례하여 증가하기 때문에 빠른 응답속도가 요구되는 환경에서는 아직 실용적으로 사용되기 어렵다.

본 논문에서는 전통적인 검색 모델을 사용하여 빠른 응답속도를 만족하면서, 딥러닝의 장점인 단락의 문맥 정보를 반영할 수 있는 검색 시스템을 제안한다. 문맥 정보를 기반으로 단어 중요도를 추론하였으며, 중요도 점수를 빈도 표현으로 변환하여 검색 모델에 반영하였다. 단어 중요도 추론 모델을 학습할 때 선행연구[2]는 질문-단락 간 정확히 매칭되는 단어를 중요 단어로 학습하였으나, 본 논문에서는 선행연구의 제한된 학습 환경을 개선하기 위해 질문 쿼리의 이형태 단어를 언어 모델을 기반으로 확장하여 단어 중요도를 학습하였다. 제안한 방법으로 일반상식 및 법률 도메인에서 검색 성능 평가 결과 선행연구[2] 대비 MRR 평균 2.02%, 0.36%이 향상되었으며, BR@Top1 에서는 1.34%, 0.9% 성능이 향상된 것을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 간략히 설명하고, 3장에서는 본 논문에서 제안하는 단어 중요도 학습 방법에 대해 설명한다. 4장에서는 실험을 통하여 제안한 방법의 우수성을 검증하고, 5장에서는 본 논문의 결론을 내린다.

Ⅱ. 관련 연구

질의응답 시스템의 초기 검색 방법은 크게 두 가지 접근방법이 사용된다. 하나는 단어 빈도 중요도에 기반한 전통 검색 방법이고, 다른 접근 방법은 딥러닝을 기반으로 임베딩 및 유연 매칭 방법을 사용하는 뉴럴 검색이다.

단어 빈도 기반의 검색 모델은 과거부터 많은 연구가 진행됐으며 대표적으로 TF-IDF[3], BM25[4] 가 있다. 기존의 연구들은 문서 내 단어들의 출현 빈도 정보로 단어가 가지는 중요도를 측정하였다. 오래전부터 사용된 방법이지만 빠른 응답속도와 검색 결과의 직관성이 높아 현재도 산업현장에서 많이 사용되고 있다. 그러나 평면 텍스트로 부터 빈도 정보만 고려하기 때문에 문맥상 단어의 의미와 중요성을 판단하지 못하는 단점이 있다. 평면 텍스트상의 단어 출현의 정보 활용을 높이기 위해 그래프를 활용한 검색[5][6]도 연구되었다. 그래프 기반의 검색은 문서를 그래프로 인식하되 단어는 노드, 공기정보는 엣지로 사용하여 윈도우 N 크기 이내의 단어간에 관계를 고려하여 단어마다 점수를 부여하였고, 단어 점수를 사용하여 색인/검색에 사용하였다.

뉴럴 검색은 임베딩 정보를 사용하기 때문에 질문-단락 간 정확한 매칭이 되지 않아도 검색할 수 있는 장점이 있다. 딥러닝 소개 이후 초기에는 CNN, RNN 알고리즘을 사용하여 질문과 단락을 임베딩하여 두 벡터간 유사도를 계산하여 검색에 사용하였다. CNN 네트워크로 임베딩 된 질문과 단락의 풀링 층을 결합하여 두 텍스트 간 유사도를 측정[7][8] 하거나 Bi-LSTM 모델을 이용하여 질문-단락을 인코딩 하고, 상부에 CNN 층을 추가하여 질문 단락 간 유사도를 측정하는 선행연구[9]가 진행되었다. 최근에는 대용량의 말뭉치를 사용하여 사전학습한 BERT(Bidirectional Encoder Representations from Transformers)[10]를 이용하는 연구가 많이 진행되고 있다. BERT는 트랜스포머[11]의 인코더 부분을 활용한 모델로 다음 문장 예측과 입력 단어 중 일부 단어를 마스킹하고, 마스킹한 단어를 예측하면서 모델을 학습한다. BERT 기반으로 검색[1]도 질문-단락을 인코딩하여 유사도를 비교하는 것은 선행 뉴럴 검색 방법과 유사하지만 많은 양의 주요 문맥 정보를 활용한다는 측면에서 차이가 있다. 뉴럴 검색은 의미적 관계까지 고려한 검색이 가능하지만 실시간성을 고려해야 하는 질의응답 환경에서 활용하기 어렵다.

본 논문에서는 빠른 검색 속도를 유지하면서 단어의 문맥 정보를 고려한 색인/검색 시스템을 소개한다. 기존에도 유사한 연구[2]가 진행되었으나, 주

어진 질문-단락만 고려하기 때문에 학습시 제약사항이 있었다. 본 논문에서는 이런 제약상황을 극복하는 방법을 제시하였으며 실험을 통해 우수성을 증명하였다.

III. 제안 방법

본 장에서는 사전 학습된 언어모델을 기반으로 단어의 중요도를 학습하는 선행연구 방법과 선행연구가 가지는 문제를 해결하기 위해 본 논문에서 제안한 방법을 소개한다.

3.1 중요 단어 추론 모델 학습

단락 내의 단어 중요도 추론 모델을 학습하기 위해 단어 주위의 문맥과 전체 단락의 의미이해가 기본적으로 필요하다. 본 논문에서는 문맥 정보를 반영하여 단어 자질을 생성할 수 있는 BERT 모델을 사용하여 중요도를 학습하였다. 중요 단어 학습에 사용한 모델의 구조는 그림 2, 식은 아래 (1), (2)와 같다.

$$\hat{y}_{t,c} = \vec{w} T_{t,c} + b \tag{1}$$

$$loss_{MSE} = \sum_{c} \sum_{t} (y_{t,c} - \hat{y}_{t,c})^2 \tag{2}$$

$T_{t,c}$는 문맥 정보 c가 반영된 토큰 t의 자질 임베딩 정보이고 $\vec{w}$, b는 학습 파라미터와 바이어스 정보이다. 위의 수식과 손실 함수를 사용하여 모델을 학습한다. 학습데이터는 질문-단락 간 정확히 일치하는 단어를 1, 그렇지 않은 단어를 0 레이블을 부착하여 오차를 학습하였다. 선행연구[2]만으로 성능 향상이 가능하나, 질문-단락 간 매칭된 단어만 학습하기에는 제한적이고, 질문의 이형태 단어는 단락에서 인지하지 못하고 오답 데이터로 학습하기 때문에 이형태 단어로 구성된 단락을 추론하였을 때 동일한 효과를 얻기 어렵다.

본 논문에서는 기존의 학습 방법을 개선하기 위해 질문의 이형태 단어 정보까지 반영하여 학습하였으며, 질문의 단어 이형태 정보 확장은 언어 모델에 기반한 단어 단순화 방법[12]을 사용하였다. 단어 단순화 과정은 그림 1과 같이 언어 모델의 입력으로 동일한 질문 쌍을 입력으로 하되, 하나의 질문에는 이형태 단어를 생성할 곳에 [MASK]로 치환하여 입력하면, 출력으로 문맥상 가장 적합한 단어를 사용할 수 있다.

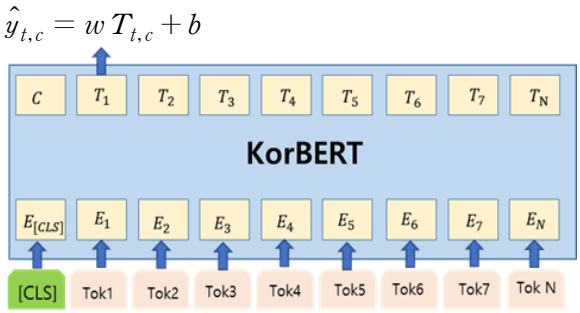

그림 2. 단어 중요도 학습을 위한 KorBERT 모델 구조
Fig. 2. KorBERT model structure for learning words importance

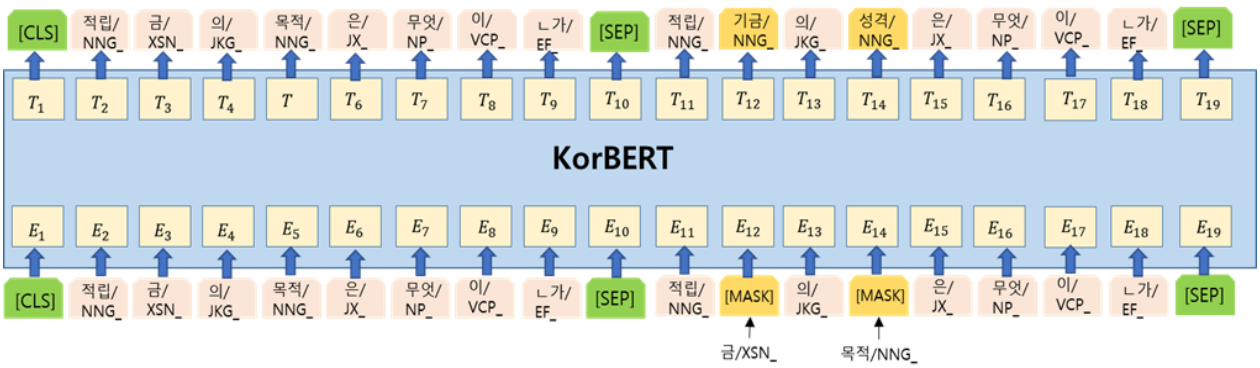

그림 1. 질문 단어 단순화를 위한 KorBERT 모델 구조
Fig. 1. KorBERT model structure for lexical simplification in question

BERT 모델은 MLM(Masked Language Model) 태스크로 사전학습을 수행하기 때문에 주변 문맥을 고려하였을 때 가장 적합한 단어 출력에 효과적이다. 마스킹할 단어 대상은 질문 내에서 조사와 어미와 같은 기능어는 제외하고 내용어를 대상을 마스킹하여 질문의 이형태 단어를 생성한다. 본 논문에서는 형태소 분석[13]과 KorBERT[13] 언어 모델을 사용하여 단어 단순화 과정을 수행하였다. 그림 3은 원본 질문과 형태소 분석 결과의 질의어, 단락, 중요 단어 레이블링의 결과 예제이다.

원본질문: 구상 성단이 띄는 모양은?

	원본 데이터 학습(DeepCT)	단어 단순화 치환(제안방법)
질의어	구상/NNG 성단/NNG 이/JKS 띄/VV 는/ETM 모양/NNG 은/JX ?/SF	은하/NNG 별/NNG 을/JKO 띠/VV ㄴ/ETM 형상/NNG 이/JKS ./SF
중요 단어 레이블링	구상/NNG, 성단/NNG, 모양/NNG	구상/NNG, 성단/NNG, 모양/NNG, **별/NNG, 은하/NNG**
단락	**구상/NNG 성단/NNG** 은/NNG 구형/NNG 의/JKG 항성/NNG 의/JKG 모임/NNG 으로/JKB ,/SP **은하/NNG** 중심/NNG 의/JKG 주위/NNG 를/JKO 마치/MAG 위성/NNG 처럼/JKB 돌/VV ㄴ다/EF ./SF **구상/NNG 성단/NNG** 은/JX 중력/NNG 에/JKB 의하/VV 어/EC 단단히/MAG 유지/NNG 되/XSV 고/EC 있/VX 으며/EC ,/SP 이/NP 에/JKB 따르/VV 아/EC 구형/NNG 의/JKG **모양/NNG** 을/JKO 유지/NNG 하/XSV 고/EC ,/SP 중심/NNG 으로/JKB 들어가/VV ㄹ수록/EC **별/NNG** 의/JKG 개수/NNG 밀도/NNG 가/JKS 높아지/VV ㄴ다/EF ./SF	

그림 3. 단어 단순화를 이용한 질의어와 중요 단어 레이블링 예제

Fig. 3. Example of importance word labeling using lexical simplification

질의어를 비교하였을 때 '구상/NNG'은 '은하/NNG', '성단/NNG'은 '별/NNG'로 치환되어 질문이 재생성 되었으며, 단락과 매칭되어 두 단어 모두 중요한 단어로 레이블링하여 학습한다. '모양/NNG'은 '형상/NNG'으로 치환되었으나, 단락과 매칭되는 단어가 없어 레이블링 대상에서는 제외되었다. 다음 장에서는 두 데이터를 단어 중요도 학습에 각각 사용하여 검색성능에 미치는 영향을 비교 실험하였다.

3.2 색인/검색

단어 중요도 추론 모델 학습이 완료되면, 컬렉션 데이터의 단어 중요도를 추론한다. 단어의 중요도 점수가 텀 빈도 기반의 검색 시스템(BM25)에 반영되기 위해서는 단어 중요도 표현(0.0~1.0)을 단어 빈도 형식으로 변환이 필요하며, 아래 식 (3)을 사용하여 변환한다.

$$TF_{DeepCT}(t,d) = round(\hat{y}_{t,d} * N) \quad (3)$$

$\hat{y}_{t,d}$은 단락d 내의 단어t의 중요도 추론 결과이며, N은 파라미터로 실험적으로 선택한다. 본 논문에서는 파라미터 N값을 10~30 범위를 적용하여 다양한 환경을 구성하여 실험을 수행하였다. 단어 중요도 $\hat{y}_{t,d}$값이 음수일 때는 제외하고, 양수의 결과값이 나왔을 때만 사용하였고, 단락의 길이가 언어모델의 제한된 입력 토큰 개수인 512보다 길 때는 중첩되는 문맥 정보 없이 슬라이딩하여 단어 중요도를 계산하였다. 단락d를 입력하였을 때 식 (3)의 결과로 출력되는 단어들을 위치 정보에 상관없이 평면 텍스트로 생성하고, bag-of-words 방식인 BM25를 사용하여 색인/검색을 수행하였다. 위와 같이 검색 시스템을 수행하면 빠른 검색 속도를 유지하면서 문맥을 고려한 단어 중요도를 기반으로 색인/검색이 가능하다.

Ⅳ. 평가 결과

본 논문에서 실험에 사용한 데이터는 두 가지 도메인을 사용하였다. 하나는 법률 분야의 질의응답 시스템을 위해 생성된 데이터이고, 다른 하나는 일반상식 분야 질의응답을 위해 생성된 데이터이다. 각 도메인에 따라 학습데이터를 사용하여 단어 중요도 추론 모델을 학습하고, 단어의 중요도를 추론하여 색인/검색 시스템에 반영하여 평가를 수행하였다. 또한, 본 논문에서 제안한 방법을 비교 평가하기 위해 원본 데이터로 학습한 모델과 단어 단순화 치환 과정을 거쳐 학습한 모델의 성능을 비교 실험하였다.

4.1 학습 데이터

본 논문에서는 제안한 방법의 성능 측정을 위해 두 도메인에서 실험을 수행하였다. 학습 데이터는 질문-정답-단락 쌍으로 구성된 법률과 일반상식 분야의 데이터이다. 학습수행에 사용된 데이터 수량과 통계는 아래 표와 같으며, 매칭된 단어들은 조사와

어미와 같이 기능어는 제외하였다.

전체 질문-정답-단락 쌍 데이터는 법률 도메인에서 76,787건이고, 일반상식 분야 데이터는 45,430건이다. 단어 중요도 모델 학습에 사용된 단어의 개수는 질문당 평균 약 6.67개, 3.37개이고, 질문의 단어를 단어 단순화 치환하였을 때는 0.4~0.7개가 늘어난 7.64개, 3.82건이다. 각 도메인별 학습데이터 2셋씩 중요 단어 추론 모델을 생성하여 검색 성능을 평가하였다.

표 1. 도메인 별 학습데이터
Table 1. Training data for each domain

Domain	# of training data	Lexical simplification	Average number of important words	Average length of questions
Law	76,787	X	6.67	27.62
		O	7.42	
Common sense	45,430	X	3.37	12.44
		O	3.82	

4.2 평가데이터 및 성능 측정 방법

이번 절에서는 평가데이터와 성능 평가 척도에 대해 설명한다. 베이스라인과 실험에 사용된 검색 모델은 BM25을 사용하였고, 베이스라인 성능은 원본 텍스트로부터 단어 빈도를 산출하여 검색하였다.

실험에 사용된 컬렉션 볼륨과 평가데이터 수량은 표 2와 같다. 일반상식 분야 도메인은 위키피디아 문서 약 51만 건이 대상이다. 위키피디아 문서를 소제목 및 문단 구분 단위로 구분하여 단락으로 사용하였고, 색인 단락 수는 약 820만건이다. 법률 도메인에서는 국내 법령과 국외의 헌법 등 1,457건의 법령이 사용하였고, 하나의 조항을 검색 단락 단위로 사용하였다. 실험에 사용된 전체 법 조항의 개수는 75,929건이다.

표 2. 도메인별 컬렉션 볼륨과 평가데이터
Table 2. Collection volume and evaluation data for each domain

Domain	Collection volume		Evaluation data
	# of document	# of passage	
Law	1,457	75,929	1,000
Common sense	519,508	8,238,116	1,500

평가 방법은 MRR(Mean Reciprocal Rank)과 BinaryRecall@TopN을 사용하여 평가하였다.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank} \tag{4}$$

$$\begin{aligned} &BinaryRecall@TopN \\ &\cdot BinaryRecall = \\ &\quad \frac{\#of\,questions\ including\ correct\ passage}{\#\,of\,total\ questions} \\ &\cdot TopN = Ranked\ top\ N\ passages \end{aligned} \tag{5}$$

MRR 평가는 검색 결과에서 정답 단락의 랭킹 분포를 파악하는데 유용한 평가 방법이고, 정답 단락을 분모로 하여 계산한다. BR(BinaryRecall)@TopN은 랭킹 분포 별 다각적 평가를 위해 사용하는 방법으로 상위 TopN개 이내에 정답단락의 포함 여부를 고려하여 성능을 측정하는 방법이다. 본 논문에서는 실험에 사용된 검색 개수는 상위 20개를 대상으로 MRR과 BR@TopN을 평가하였다.

4.3 평가 결과

도메인별 검색 성능의 MRR 결과는 아래 표 3과 4를 통해 확인할 수 있다. 표 3과 4에서 N은 수식 3 계산시 사용한 파라미터이고, 성능 평가에서 일관성을 확인하기 위해 다양한 파라미터를 사용하였다. 일반상식과 법률 도메인에서 베이스라인 검색 성능이 54.92%, 77.82%인 것을 감안하였을 때 문맥 정보를 고려한 단어 중요도 기반의 검색 방법이 더 높은 성능을 나타낸다는 것을 알 수 있다.

표 3. 일반상식 도메인에서 MRR 검색 성능 평가
Table 3. Performance of search via MRR measure in common sense domain

N	10	15	20	30
DeepCT[2]	67.86%	71.20%	71.91%	73.46%
Proposed method	70.74%	73.29%	74.06%	74.42%

표 4. 법률 도메인에서 MRR 검색 성능 평가
Table 4. Performance of search via MRR measure in law domain

N	10	15	20	30
DeepCT[2]	78.42%	78.99%	78.89%	78.52%
Proposed method	78.49%	79.19%	79.30%	79.26%

특히 일반상식 분야에서는 최대 약 18~19% 정도 효과가 있는 것으로 나타났다. 법률에서는 성능 향상의 정도가 약 1~1.4% 라 일반상식 분야만큼 크진 않지만, 검색 성능이 향상되는 것을 볼 수 있다. 표 3과 4에서 선행연구와 제안한 방법에서의 성능 차도 확인할 수 있는데, 파라미터 N의 수치에 따라 성능 향상의 차이는 있으나, 평균적으로 일반상식 도메인에서는 2.02%, 법률 도메인에서는 0.36% 성능이 향상된다는 것을 확인할 수 있다. 실험 결과를 통해 단어 중요도를 학습할 때 주어진 질문-단락 쌍에서 정확히 매칭되는 단어뿐 아니라 질문을 확장하여 단락 내의 유사 의미의 단어까지 중요 단어로 학습하는 것이 효과적이라는 것을 알 수 있다.

또한 평가 결과에서 N의 수치가 30, 법률 도메인일 때는 N의 수치가 20일 때 성능이 가장 높았는데, 도메인에 따라 수치는 다르지만, N의 수치가 커질수록 성능 향상 기여가 크다는 것과 수식 3에서 N의 값은 단어 중요도를 빈도로 바꿀 때 부여하는 가중치이기 때문에, 단어 중요도 학습이 검색에 많은 영향을 미친다는 것을 확인할 수 있었다.

표 5와 6은 도메인 별 BR@TopN의 성능 결과이며 각 도메인은 MRR 평가에서 높은 성능을 낸 파라미터 환경에서 실험한 결과이다. BR@TopN으로 평가하였을 때도 두 도메인 모두 MRR 평가와 동일하게 베이스라인 대비 높은 성능 향상이 된 것을 알 수 있다.

표 5. 일반상식 도메인에서 BR@TopN 검색 성능 평가
Table 5. Performance of search via BR@TopN measure in common sense domain

BR@TopN	Baseline	DeepCT[2]	Proposed method
Top1	45.07%	65.33%	66.67%
Top5	67.00%	83.40%	84.47%
Top10	74.13%	86.47%	87.00%
Top20	79.47%	88.69%	88.80%%

표 6. 법률 도메인에서 BR@TopN 검색 성능 평가
Table 6. Performance of search via BR@TopN measure in law domain

BR@TopN	Baseline	DeepCT[2]	Proposed method
Top1	69.90%	70.20%	71.10%
Top5	88.30%	89.80%	89.80%
Top10	93.00%	94.10%	94.30%
Top20	94.50%	96.80%	96.90%

표 5의 일반상식 분야에서는 정답 단락을 Top1으로 검색하는 성능은 약 20%이상 성능 향상되었으며, Top20에서의 성능 향상에 약 9%인 것을 감안하면 제안한 검색 방법이 재현율뿐 아니라 정확률까지 향상시킨다는 것을 알 수 있다. 또한, 선행연구의 Top1 성능과 비교하였을 때도 1.34% 향상되어 제안한 방법이 우수하다는 것을 확인할 수 있었다.

표 6의 법률 도메인에서는 성능 향상의 폭이 일반상식 도메인에 비해 적지만 Top1 기준으로 보았을 때 베이스라인 대비 최대 1.2% 성능이 향상되었고, 선행연구 대비 제안 방법이 BR@Top1 0.9%향상된 것을 확인할 수 있었다.

V. 결론 및 향후 과제

본 논문에서는 전통 검색의 빠른 검색과 딥러닝을 기반으로 문맥 정보를 반영하는 학습 방법을 제안하였으며, 기존 연구가 가진 제한적 학습을 극복하기 위해 질문을 치환하여 질문-단락 간 중요 단어를 학습하는 방법을 제안하였다. 실험 결과 두 도메인에서 사전연구 대비 MRR 평균 2.02%, 0.36%이 향상되었으며, BR@Top1 에서는 1.34%, 0.9% 성능이 향상되어 제안 방법이 우수하다는 것을 확인하였다. 그러나, 실험에 사용한 두 도메인은 동일한 방법을 사용하였음에도 성능 향상의 폭은 크게 나타났고, 도메인 별로 사용되는 파라미터에서도 차이있었다. 향후 연구에서는 단어의 중요도 학습을 위해 다양한 도메인에서도 견고하게 동작할 수 있게 일반화시킨 모델을 생성하는 방법을 연구할 계획이다.

References

[1] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih, "Dense Passage Retrieval for Open-Domain Question Answering", Proc. of the EMNLP 2020, Punta Cana, Dominican Republic, pp. 6769-6781. Nov. 2020. http://dx.doi.org/10.18653/v1/2020.emnlp-main.550.

[2] Zhuyun Dai and Jamie Callan, "Context-Aware

Term weighting for First Stage Passage Retrieval", Proc. of the SIGIR 2020, Xi'an, China, pp. 1533-1536. Jul. 2020. http://dx.doi.org/10.1145/3397271.3401204.

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

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

[5] R. Blanco and C. Lioma, "Graph-based term weighting for information retrieval", Information Retrieval, Vol. 15, No. 1, pp. 54-92, Feb. 2012. http://dx.doi.org/10.1007/s10791-011-9172-x.

[6] Rada Mihalcea and Paul Tarau, "TextRank: Bringing Order into Text", Proc. of the EMNLP 2004, Barcelona, Spain, pp. 404-411, Jul. 2004.

[7] A. Severyn and A. Moschitti, "Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks", Proc. of the SIGIR 2015, Shanghai, China, pp. 378-382, Aug. 2015. http://dx.doi.org/10.1145/2766462.2767738.

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

[9] 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://dx.doi.org/10.1109/DSC.2016.37.

[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] 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. http://dx.doi.org/10.1109/ICASSP39728.2021.9413901.

[12] Jipeng Qiang, Yun Li, Zhu Yi, Yunhao Yuan, and Xindong Wu, "Lexical simplification with pretrained encoders", Proc. of the AAAI 2020, New York, USA, pp. 8649-8656, Feb. 2020. https://doi.org/10.1609/aaai.v34i05.6389.

[13] ETRI AIOpen, "Provide the learning model and training data", http://aiopen.etri.re.kr/service_dataset.php. [accessed: Sep. 23, 2021]

저자소개

배 용 진 (Yongjin Bae)

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

이 공 주 (Kong-Joo Lee)

1992년 : 서강대학교 전자계산학과(학사)
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