한국어 다중추론 질의응답을 위한 Dense Retrieval 사전학습

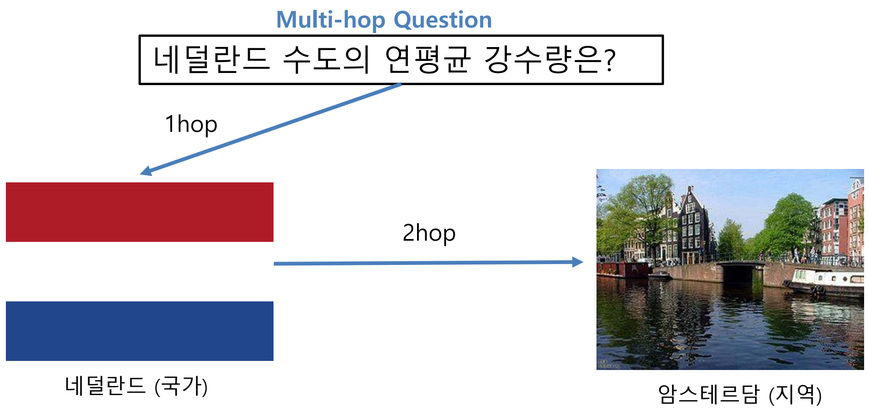
Cognitive Computing Lab.

강동찬, 나승훈, 김태형, 최윤수, 장두성
전북대학교, KT 융합기술원
kdc1430@gmail.com, nash@jbnu.ac.kr, {taehyeong_2019.kim, yunsu.choi, dschang}@kt.com

Ⅰ. 서론

다중추론 질의응답 태스크는 하나의 문서만 필요한 기존의 단일추론 질의응답(Single-hop QA)을 넘어서 복잡한 추론을 요구하는 질문에 응답하는 것이 목표이다. IRQA에서는 검색 모델의 역할이 중요한 반면, 주목받고 있는 Dense Retrieval 모델 기반의 다중추론 질의응답 검색 모델은 찾기 어렵다.

본 논문에서는 검색분야에서 좋은 성능 보이고 있는 Dense Retrieval 모델의 다중추론을 위한 사전학습 방법을 제안하고 관련 한국어 데이터 셋에서 이전 방법과의 성능을 비교 측정하여 학습 방법의 유효성을 검증하고 있다. 이를 통해 지식 베이스, 엔터티 링킹, 개체명 인식모듈을 비롯한 다른 서브모듈을 사용하지 않고도 다중추론 Dense Retrieval 모델을 학습시킬 수 있음을 보인다.

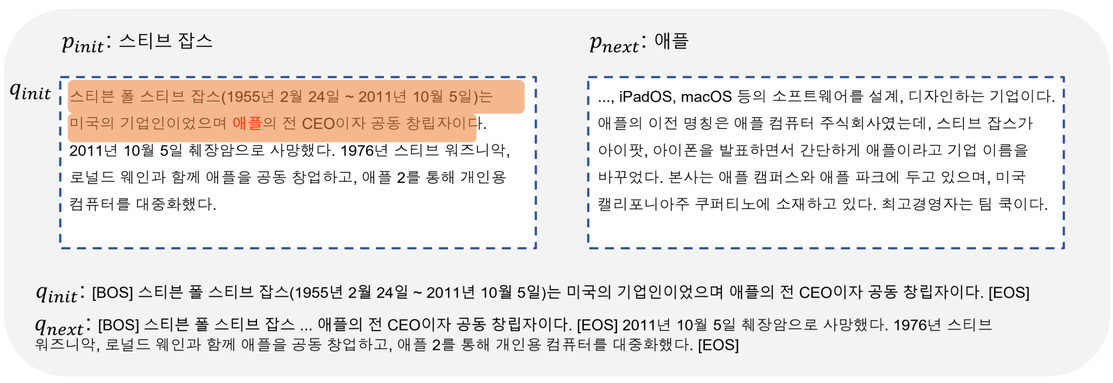


Ⅱ. 제안 방법

사전 학습

위키피디아 문서에서 시작 문장을 선택한 후 이것을 시작 쿼리로 하며, 시작 문장과 시작 문서가 이어진 입력을 다음 쿼리로 한다. 한편 시작 문서와 연결 된 문서를 다음 문서로 한다.

사전학습은 시작 쿼리로 시작 문서를 맞추는 태스크와 다음 쿼리로 다음 문서를 맞추는 태스크로 구성된다.



사전학습 예제의 생성

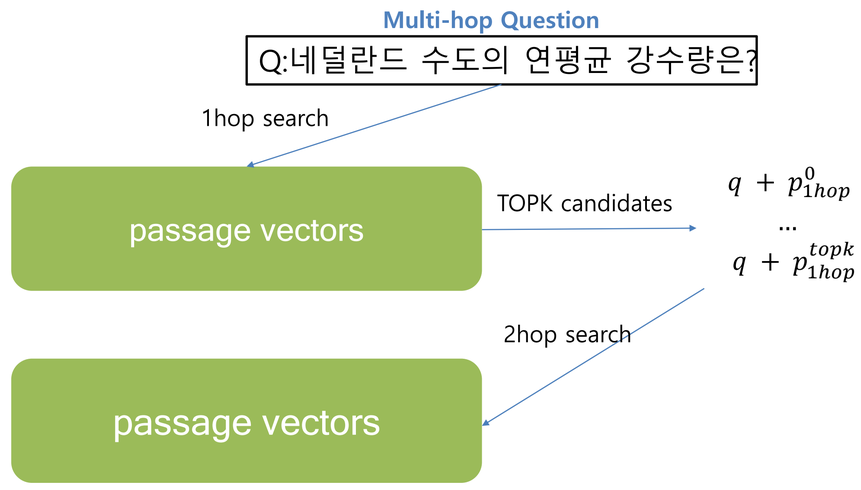
파인 튜닝

파인튜닝 시에는 이전 연구에서 구축한 한국어 다중추론 데이터 셋을 사용했다. 본 연구에서는 제안하는 다중추론 사전학습을 통해 문서의 임베딩을 충분히 얻을 수 있는지에 집중하기 위해 파인튜닝 시에는 문서 인코더를 고정시키고, 쿼리 인코더만 학습했다.

Train	Dev	Test
50708	500	1000

파인튜닝 데이터 통계

추론



추론 시에는 n-hop마다 반복적으로 검색하는 MDR 모델과 동일하게 하이퍼 링크를 사용하지 않고, 위키피디아 전체 문서에서 반복적으로 검색한다.

Ⅲ. 실험

실험 결과

Model	@5	@10	@20	@100
HopRetriever	74.6	80.6	84.7	-
MDR	39.2	46.9	55.1	70.5
Model	39.8	49.9	60.1	79.7
Model + Link	38.6	49.3	60.0	80.1

본 논문은 사전학습이 끝난 인코더를 기반으로 파인튜닝을 진행하고 성능을 측정했다. 문서쌍 열의 길이 2 설정을 사용하고 있으며 실험 시 1hop Topk의 경우 이전 연구와 동일하게 250을 사용했다. 평가 방법 또한 이전연구와 동일하게 랭킹된 문서쌍의 나열에 정답이 있는지 없는지로 측정 (Hits@k).

본 연구에서는 링크 제약의 사용은 성능에 큰 영향이 없었다. 제안한 모델은 전체 탐색공간의 크기가 870000 * 870000라는 점을 감안하면 다중추론이 가능한 것으로 보인다. 또한 제안한 모델은 파인튜닝 데이터만을 이용해 문서, 질문 인코더를 얻는 MDR 비해 문서 임베딩을 고정해 놓은 상태에서도 우수한 성능을 보였다. 그러나 HopRetriever보다는 낮은 성능을 보여 성능을 끌어올리기 위해서 학습이 더 정교해질 필요가 있다.

Ⅳ. 결론

본 논문에서는 한국어 상에서 오픈 도메인 다중추론을 위한 Dense Retrieval 모델의 사전학습 방법을 제안하고 한국어 데이터 셋에서 이전 방법과의 성능을 비교 측정하여 학습 방법의 유효성을 검증한다. 이를 통해 지식 베이스, 엔터티 링킹, 개체명 인식모듈을 비롯한 다른 서브모듈을 사용하지 않고도 다중추론 Dense Retrieval 모델을 학습시킬 수 있음을 보였다. 차후 사전학습, 파인튜닝 방법을 더 정교하게 만들어 검색 성능을 높일 계획이다.