멀티 도메인 대화 데이터 셋을 사용한 문서 검색 모델 성능 개선

배수영 ○, 김은총, 정윤경
성균관 대학교 인공지능학과
{sybae01, prokkec, aimecca}@g.skku.edu

Improving Document Retrieval Model Multi Domain Dialogue Dataset

Suyoung BaeO , Eunchong Kim, YunGung Cheong
SungKyunKwan University Artificial Intelligence Department

요 약

오픈 도메인 질의 응답 모델에서 DPR(Dense Phrase Retrieval)은 두 개의 인코더를 사용해서 질문과 문단 간의 정확도 높은 문단을 검색하는 모델이다. 기존 DPR은 위키피디아 도메인 한 개를 사용해 학습을 진행했지만 본 연구에서는 네 개의 도메인과 대화 형식의 질의 응답으로 이루어진 데이터 셋을 사용해서 DPR 학습을 진행하고 기존 모델과 검색 성능을 비교하는 실험을 진행했다. 그 결과 기존 모델보다 검색 성능이 정확도 성능 지표 면에서 약 10~20% 정도 향상되는 결과가 나왔고, 이전 대화 내용이 포함된 질문을 제시했을 때 모델이 답변과 관련된 문단을 잘 검색하는 결과가 나왔다.

1. 서 론

[1] 오픈 도메인 질의 응답(ODQA)은 방대한 정보들을 포함하고 있는 문서 집합들을 참조해 질문에 대한 답변을 만드는 문제다. 질의 응답 시스템은 주어진 질문에 대한 답변을 포함하고 있는 문단을 찾는 검색(Retriever) 과정과 검색된 문단에서 실제 답변을 찾는 읽기(Reader) 과정으로 이루어진 프레임 워크를 사용한다. 질의 응답 시스템의 성능을 높이기 위해서는 관련 문단을 검색하는 부분의 성능을 향상시키는 것이 중요한데 기존 연구들은 검색 모델에서 고차원 희소 벡터로 질문과 문단 표현하는 TF-IDF, BM25[1]를 사용해 질문 문장과 관련 있는 문단을 선택했다. 하지만 고차원 희소 벡터를 사용했을 때 동의어이지만 완전히 다른 토큰에 대해서 서로 가까운 벡터에 매핑하는 것이 어렵기 때문에 의미는 같지만 형태는 다를 경우 관련 문단으로 검색하는데 한계가 있다. 따라서 DPR[2] 논문에서는 밀집 벡터 표현을 사용해 의미가 문단들을 유사도가 높은 문단으로 판단할 수 있도록 개선했다.

DPR은 밀집 벡터 표현들 만으로 검색이 가능하게 검색(Retriever) 부분을 변형한 모델로 사전 학습된 BERT로 이루어진 두 개의 인코더를 사용해 질문과 문단을 밀집 벡터로 인코딩하고 질문 벡터와 문단 벡터 표현을 최대 내적 검색 알고리즘(MIPS: Maximum Inner Product Search Algorithm)을 사용해 점수가 가장 높은 N 개 문단을 검색한다. DPR 모델에서는 학습 단계에서는 위키피디아 문서를 100 단어 단위로 문단을 구성해 사용하고, 평가에서는 Natural Question, TriviaQA, WebQuestion, CuratedTREC, SquAD 로 총 5 개를 사용했다. DPR 모델을 사용한 오픈 도메인 질의 응답 시스템은 기존 검색 방식보다 검색 성능이 정확도가 약 13%(NQ 데이터셋으로 Top-100 개 문단을 검색한 결과 BM25 정확도는 73.7%, DPR 정확도는 86.0%)만큼 향상되었고, 최종 질의 응답 작업에 대해서도 더 좋은 성능을 보였다(NQ 데이터셋으로 실험한 결과 BM25 정확도는 26.5%, DPR 정확도는 41.5%로 약 15% 향상되었다).

하지만 DPR 모델은 두 가지 한계점이 존재한다. 첫 번째는 검색 모델 학습 시 오픈 도메인으로 위키피디아 문서 데이터 셋 하나 만을 사용했다는 점이다. 실제로 인터넷 상에 오픈 도메인은 위키피디아 말고도 수많은 도메인이 있기 때문에 도메인 하나 만으로 학습된 검색 모델을 사용하는 것으로는 완벽하게 답변을 생성하기 어렵다. 두 번째는 DPR 모델은 질문 한 개에 대한 대답 쌍으로 이루어진 데이터 형식으로 구성되어 있기 때문에 여러 개의 질의 응답 쌍으로 구성되어 있는 대화 형식에 적용한다면 문맥을 파악해 답변을 하는 학습이 부족하다. 실제로 질의 응답 작업이 적용되는

1. 이 논문은 2019년도 정부(교육과학기술부)의 재원으로 한국연구재단(No. 2019R1A2C1006316), 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(No.2019-0-00421, 인공지능대학원지원), 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업(IITP-2017-0-01642)의 지원을 받아 수행된 연구임.

분❶는 대부분 대화 형식으로 구성되어 있기 때문에 대화 형식에 대해서도 학습할 필요성이 있다. 따라서 본 논문에서는 DPR 모델에 멀티 도메인과 대화 형식으로 구성되어 있는 새로운 데이터 셋(MultiDoc2Dial)[3]을 적용해 추가 학습을 시도했다.

본 논문이 기여하는 바는 다음과 같다. 1) 여러 도메인을 학습함으로써 다양한 주제에 대한 질문이 들어왔을 때 모델이 관련된 문서를 이전보다 더 잘 검색할 수 있게 하였다. 2) 대화 형식 안에서 이전 대화 내용을 반영한 질문이 주어졌을 때 질문과 관련된 문서 검색 성능이 더 높아짐을 보였다.

1. 접근 방법

1.1 모델

검색 모델로 Dense Phrase Retrieval(DPR)를 사용했고, 문단, 질문 인코더 모델로 BERT를 사용해 인코딩을 진행했다. 문단 인코더의 입력으로는 멀티 도메인 데이터 셋을 100 단어 단위로 잘라서 사용했다.

1.2 데이터셋

데이터 셋은 DOC2Dial 데이터 셋 V1.0.1 [2]을 기반으로 구성된다. Doc2Dial 데이터 셋[4]은 도메인 데이터 셋과 대화 데이터 셋으로 이루어져 있다. 표 1에서 도메인 별 문서 수와 대화 수의 통계를 볼 수 있다. 도메인 문서 데이터 셋은 미국 공공 서비스 웹사이트인 ssa, va, dmv, studentaid 총 4 개에서 수집한 데이터들을 도메인 별로 제목, 문서 아이디, 문서 내용, HTML 마크 업 정보들을 담아서 재구성한 데이터 셋이다. 대화 데이터 셋은 각 도메인 별 대화 아이디와 함께 각 대화에서의 역할, 대화 내용, 참조하는 문서 아이디로 구성되어 있다. 본 논문에서는 MultiDoc2Dial 데이터 셋을 DPR 데이터 셋 형식으로 바꿔 검색 모델의 입력으로 사용했다.

도메인	*문서 수*	*대화 수*
Ssa	109	1191
Va	138	1337
Dmv	149	1328
Studentid	92	940
Total	488	4796

표 1: MultiDoc2Dial 데이터 셋 통계

2. 실험

2.1 데이터 셋 전처리

도메인 데이터 셋 구성을 위해 MultiDoc2Dial 도메인 데이터 셋의 도메인 별 텍스트에서 'Wn', 'Wt', 'Wr'를 제거하고 각 텍스트를 100 단어 단위로 잘라서 문단으로 나누어 주었다. 그 결과 총 4282 개 문단이 구성되었고, 각 문단이 포함된 문서 제목과 인덱스를 함께 저장했다. 대화 데이터 셋은 MultiDoc2Dial 대화 데이터 셋에서 질문 아이디와 질문, 답변을 저장하고, 참조하는 문단에 대해 BM25를 사용해 긍정 문단 (Positive Context)와 부정 문단(Negative Context)을 만들어 저장했다. 테스트 셋 구성을 위해서는 이미 만들어진 DPR 형식에서 질의 응답 쌍을 골랐고, 질문 부분에는 질문과 이전 대화 기록을 [SEP] 토큰으로 구분해 저장했다. 대화 기록 유무에 따른 실험 결과를 비교하기 위해 대화 기록을 제거한 질의 응답 쌍도 만들었다.

2.2 학습 방법

DPR 구조에서 문단 인코더와 질문 인코더 모델은 모두 사전 학습된 BERT를 사용했다. 입력으로는 문단 인코더에는 100 단어로 잘라 문단들을 구성한 멀티 도메인 데이터 셋을 사용했고, 질문 인코더에는 멀티 도메인 질의 응답 학습 데이터 셋을 사용해서 DPR 모델 학습을 진행했다. 학습 에폭 50, 배치 사이즈 16, warm up step 800으로 설정했다. 에폭 당 저장되는 모델 크기가 약 2~3GB 이기 때문에 메모리를 절약하기 위해 이전 에폭 저장 모델보다 검증 로스(Validation Loss) 값이 줄어드는 경우만 저장했고, 최종 실험 모델은 에폭 48 일 때 모델을 사용해서 실험을 진행했다.

2.3 성능 평가

실험은 총 두 개를 진행했다. 첫 번째는 기존 모델인 위키피디아 도메인 데이터 셋으로 학습한 DPR 검색 모델과 멀티 도메인 데이터 셋으로 학습한 검색 모델의 성능을 비교했다. 문서를 얼마나 정확히 검색하는지에 대한 성능을 평가하기 위해 정확도(Accuracy) 성능 지표를 사용한다. K 값은 1, 20, 60, 100으로 변경하면서 학습, 검증, 테스트 데이터 셋의 정확도를 비교했다.

두 번째는 멀티 도메인 데이터 셋을 통해서 학습한 모델을 사용해 질문에 대한 이전 대화 기록을 포함했는지 여부에 따라 정확도가 얼마나 차이가 나는지 실험을 진행했다. 평가 지표는 첫 번째 실험과 동일하게 K 값을 1, 20, 60, 100으로 변경하면서 학습, 검증, 테스트 데이터 셋의 정확도를 비교했다. 여기서 Top-K 정확도 값이란 각 질문에 대한 검색 문단들 K 개 중 답변이 포함되어 있는 문단에 해당될 비율을 말한다. 본 논문에서는 각 질문에 대한 정확도 평균을 계산해 평가 지표로 사용했다.

3. 실험 결과

3.1 멀티 도메인 데이터 셋 검색 모델 평가 결과

표 2, 표 3 은 위키피디아 도메인 데이터 셋으로 학습한 DPR 검색 모델(Single)과 멀티 도메인 데이터 셋으로 학습한 DPR 검색 모델(Multi)의 Top-K 정확도 값을 정리한 표이다. Single 검색 모델은 위키피디아 데이터셋 크기가 너무 커서 공개된 테스트 정확도 값 외에는 추가적으로 실험을 하지 못했다. 또한 Single, Multi 모델의 테스트셋이 달라 비교

[2] http://doc2dial.github.io/multidoc2dial/

분석은 수행하지 않았다. 정확도가 높은 100 개의 문단을 검색한 결과 Single 모델에서 테스트 데이터 셋 정확도가 87%가 나왔고, Multi 모델의 정확도는 96%가 나왔다. Multi 모델에서 학습 데이터보다 정확도가 높은 이유는 Multi 테스트 셋이 5 개 밖에 안되기 때문에 정확도가 90 이 넘는 결과가 나왔다고 본다. 따라서 다중 도메인 데이터 셋을 사용해서 학습한 모델에서 100 개의 문서를 검색했을 때 검색 능력이 좋다고 볼 수 있다.

k	1			20			60			100		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Single	.	.	52	.	.	81	.	.	.	.	.	87

표 2: Single 모델의 정확도

k	1			20			60			100		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Multi	72	33	0	82	60	81	83	69	94	83	73	**96**

표 3: Multi 모델의 정확도

3.2 대화 기록 포함 유무에 따른 성능 평가 결과

표 4 은 멀티 도메인 데이터 셋을 사용해서 DPR을 학습할 때 질문 데이터 셋에 질문 이전 대화 정보를 포함하는지 여부(O, X)에 따라 Top-K 정확도를 계산한 결과이다. Top-1, 20, 60, 100 모두 이전 대화 정보를 포함한 질문을 입력으로 넣었을 때 정확도가 높은 결과가 나왔다. Top 100 테스트 데이터 셋의 경우, 정보가 있는 경우 96, 없는 경우 59 로 큰 차이가 있는 것으로 보인다. 이는 대화 이전 기록을 포함하고 학습했을 때 더 DPR 검색 성능이 좋다고 할 수 있다.

k	1			20			60			100		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
O	72	33	0	82	60	81	83	69	94	83	73	**96**
X	40	20	0	57	39	56	62	48	59	64	52	59

표 4: MultiDoc2Dial 데이터 셋 통계

4. 결론 및 향후 연구 계획

이 논문에서는 질문 답변 시스템을 다양한 도메인에 적용할 수 있는 모델을 만들고자 시도했고, 실험 결과를 통해 멀티 도메인 데이터 셋을 사용하는 것이 효과적임을 검증했다. 검색 모델로 사용하는 DPR 모델을 멀티 도메인 데이터 셋을 사용해서 학습을 진행하고 Top-100 개를 검색했을 때 정확도가 96% 가 나온 것을 확인했다. 또한 질문 데이터 셋에 이전 대화 내용을 반영한 데이터 셋을 사용했을 때 검색 정확도가 높아지는 결과를 보였다. 실험 시간 부족이 100 단어로 나누어 문단 데이터셋을 만들 때 문장이 끊기는 문제를 해결하지 못했지만, 문장 구조를 기준으로 문단 데이터를 나누어 구성한다면 더욱 높은 결과를 도출해낼 수 있을 것으로 기대한다.

참고문헌

[1] Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4):333-389.

[2] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering.

[3] Song Feng, Siva Sankalp Patel, Wan Hui and Sachindra Joshi MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents.

[4] Song Feng, Hui Wan, Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, Luis A. Lastras doc2dial: A Goal- Oriented Document-Grounded Dialogue Dataset