멀티 도메인 대화 데이터셋을 사용한 문서 검색 모델 성능 개선

성균관대학교 인공지능 학과 석사과정 배수영

2022 한국컴퓨터종합학술대회(KCC2022)

2022-06-30

목차

- 연구 필요성
- 관련 연구
- 접근 방법
- 실험
- 실험 결과
- 결론

연구 필요성

- 자연어 처리 분야인 Question Answering Task baseline model ->Dense Phrase Retriever 모델 : Single Domain Dataset 사용해서 학습.
- 실제로 인터넷 상에는 수많은 Domain이 있기 때문에 Domain 한 개 만으로 학습한 모델로는 사용자의 모든 질문에 대한 답변을 생성하기에 한계가 있다.
- 따라서 Question Answering 모델 학습을 Multi Domain Documents를 사용해서 학습을 진행한다면 Question Answering Task 성능 향상에 도달할 수 있을 것이라고 생각했다.
- 또한 모델이 질문에 대한 답을 할 때 이전 대화 내용을 반영하도록 했을 때 질문과 관련된 문서 검색 성능이 더 높아질 것이라고 생각했다.

관련 연구 – Open Domain Question Answering

- ODQA: 방대한 정보들을 포함하고 있는 문서 집합들을 참조해 질문에 대한 답변을 생성하는 Task
- 전통적인 접근 방법: Retriever-Reader Framework
  - Retriever: 질문과 관련 있을 법한 passage를 찾아옴 (ex: TF-IDF, BM25)
  - Reader: 주어진 문제에 대해 구체적인 답변을 찾아냄 (ex: Neural Network)
- Open domain question answering 에서는 후보 passages 를 고르는 Passage Retriever이 중요하다.
- Sparse vector model:
  - TF-IDF: 단어의 빈도와 역 문서 빈도를 사용해서 각 단어들마다 중요한 정도에 가중치를 주고
  - BM25: Bag of Words 개념을 사용해 query 에 있는 용어가 각 문서에 얼마나 자주 등장하는지 평가

관련 연구 – Dense Phrase Retriever (DPR)

- Retriever 부분을 기존에 사용하던 Sparse vector model(TF-IDF, BM25) 보다 dense vector representation을 사용해 질문과 관련된 document passage 후보들을 결정하는 모델
- Dual encoder architecture:
    1. question 과 passage 를 각각의 서로 다른 인코더에 통과시켜 d 차원으로 임배딩을 진행
    2. question 과 passage 벡터 간의 유사도를 Maximum Inner Product Search Algorithm (MIPS) 을 사용
- 추가적인 pretraining 없이 question, passage 쌍으로만 dense embedding model 학습 가능

Dense Passage Retrieval for Open-Domain Question Answering

Vladimir Karpukhin[*], **Barlas Oğuz**[*], **Sewon Min**[†], **Patrick Lewis**, **Ledell Wu**, **Sergey Edunov**, **Danqi Chen**[‡], **Wen-tau Yih**
Facebook AI [†]University of Washington [‡]Princeton University
{vladk, barlaso, plewis, ledell, edunov, scottyih}@fb.com
sewon@cs.washington.edu
danqic@cs.princeton.edu

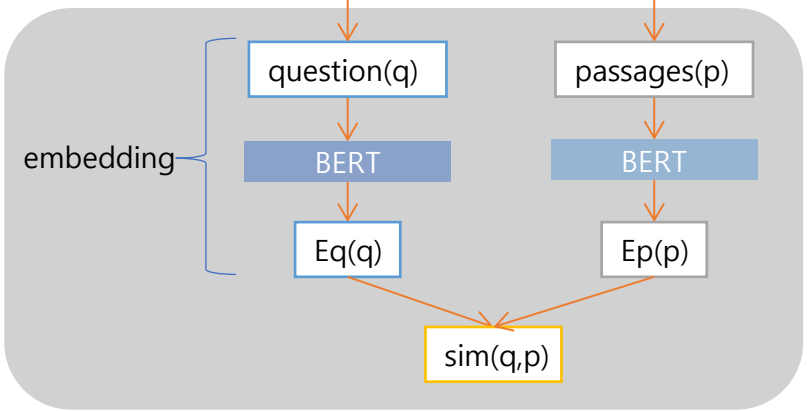
$$\text{sim}(q, p) = E_Q(q)^\intercal E_P(p)$$

관련 연구 – Dense Phrase Retriever (DPR)

- 학습(Training)

Dataset	Train	Dev	Test
Natural Questions	79,168	8,757	3,610
TriviaQA	78,785	60,413	8,837	11,313
WebQuestions	3,417	2,474	361	2,032
CuratedTREC	1,353	1,125	133	694
SQuAD	78,713	70,096	8,886	10,570

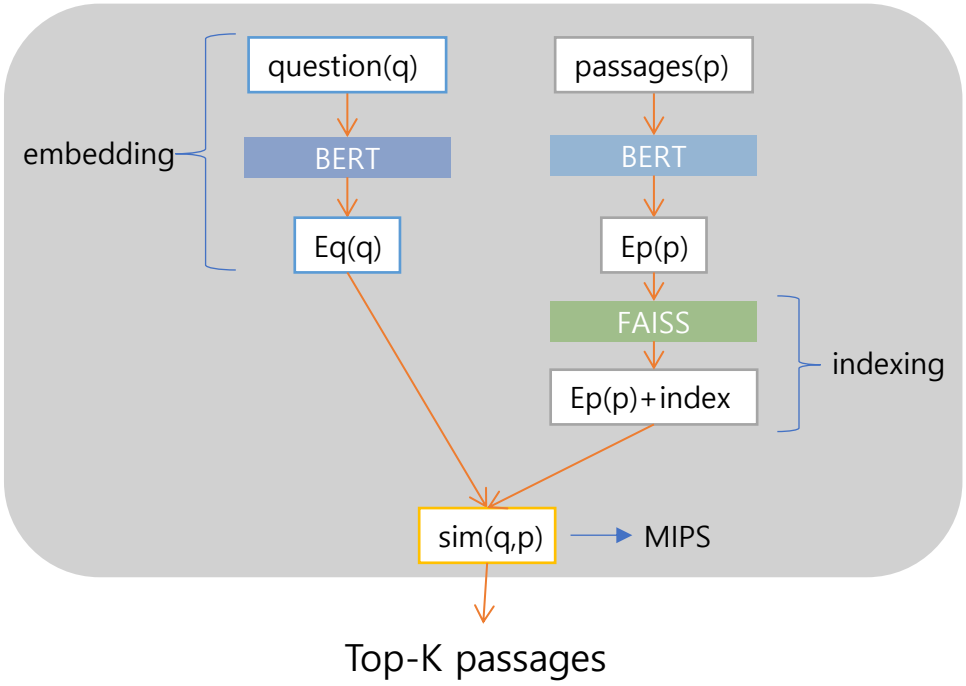
Positive 1개
Negative n개



$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)$$

$$= -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}}$$

- 추론(Inference)

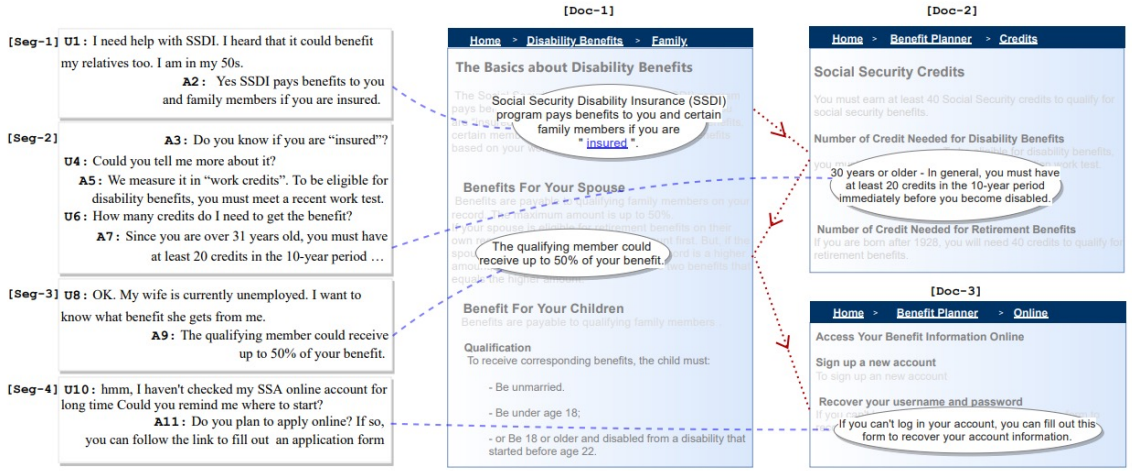


관련 연구 – MultiDoc2Dial Dataset

- MultiDoc2Dial : goal-oriented dialogues grounded in multiple documents
- Document: 4개 domain 에서 문서를 수집함 (ssa, va, dmv, student)
- Dialogue: Document-grounded Dialogue (Doc2Dial) 에서 가져온 dialogues.
- Dialogue 를 segment 로 나누어서 각 segment 마다 참조할 document 를 연결해준다. (파란 점선)
- Segment 가 바뀔 때마다 다른 document 를 참조하도록 dialogue 흐름을 재조합 한다. (빨간 점선)

domain	#doc	#dial	two-seg	>two-seg	single
ssa	109	1191	701	188	302
va	138	1337	648	491	198
dmv	149	1328	781	257	290
student	92	940	508	274	158
total	488	4796	2638	1210	948

MultiDoc2Dial 데이터 통계



MultiDoc2Dial 흐름

접근 방법

- Dense Phrase Retrieval 한계점

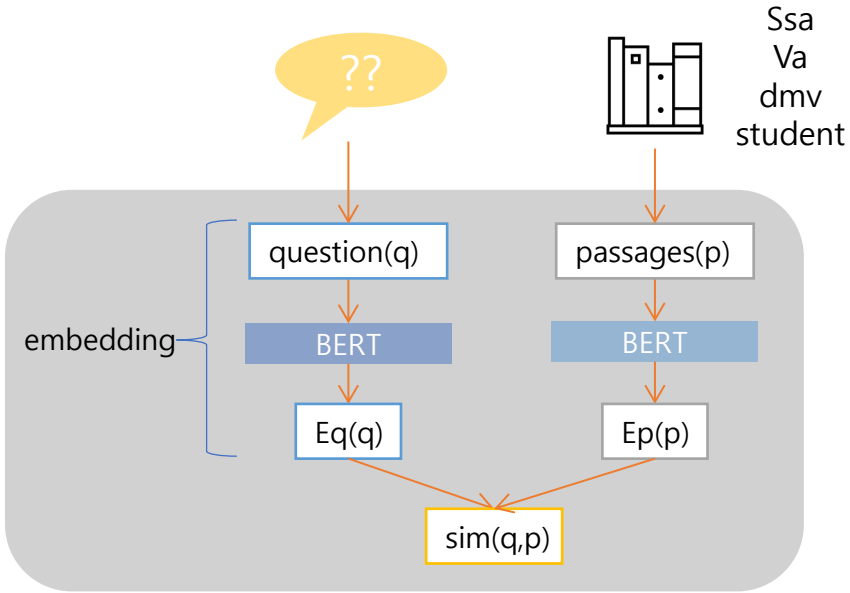
1. Retrieval 학습 시 위키피디아 문서 도메인 하나만을 사용해서 학습 -> 실제 인터넷 상에 오픈 도메인은 수많은 도메인 존재
   -> 여러 도메인을 반영해 학습을 해보자!
2. 한 개 질문에 대한 답변을 생성하는 것을 목적으로 하는 모델 -> 대화 형식에서 문맥을 파악해 답변을 생성하는 능력 학습 부족
   -> 대화 형식 데이터셋을 사용해보자!

- Model : Dense Phrase Retrieval Dual encoder 구조, BERT 사전 학습 모델을 인코더 모델로 사용
- Dataset : Multidoc2dial Dataset -> DPR question, passages 입력 형식으로 전처리 진행 후 입력으로 넣음

실험 – 데이터셋 전처리

1. Domain Dataset
   - '₩n', '₩t', '₩r' 등 불필요한 부분 제거
   - 각 텍스트를 100 words 로 잘라서 passage 로 설정
   - 각 passage 가 포함된 문서 제목, 도메인 정보, 인덱스 함께 저장
   - 4282개 passages 구성
2. Dialogue Dataset
   - MultiDoc2Dial 질문과 답변을 저장
   - 질문에 대해 참조하는 passage 에 대해 긍정 문단, 부정 문단을 만들어 저장함 (question, passage 쌍 생성)
   - 대화 기록 유무에 따른 성능 비교 평가를 위해 질문에 이전 대화 기록을 [sep] 토큰으로 구분해 함께 저장한 테스트셋도 별도로 만들어 저장

실험 – 학습



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

- Epoch : 50
- Batch size : 16
- Warmup step: 800
- OS: Linux-x86_64
- GPU: NVIDIA GeForce RTX 3090

- 학습 결과 loss 최소 모델인 epoch 48 사용.

실험 – 성능 평가

1. 기존 DPR 모델과 멀티 도메인 데이터셋으로 학습한 검색 모델의 성능 비교
   - 정확도(Accuracy)
   - Top-k 에서 k 값 1, 20, 60, 100 으로 변경하면서 비교
2. 멀티 도메인 데이터셋으로 학습한 모델을 사용해 질문에 대한 이전 대화 기록을 포함했는지 여부에 따라 정확도가 얼마나 차이가 나는지 비교
   - 정확도(Accuracy)
   - Top-k 에서 k 값 1, 20, 60, 100 으로 변경하면서 비교

실험 결과

1. 멀티 도메인 데이터셋 사용한 검색 모델 평가 결과

- Single

K	1	20	60	100
Test	52	81	-	87

- Multi

K	1	20	60	100
Test	0	81	94	96

실험 결과

2. 입력에 대화 기록 포함 유무에 따른 성능 평가 결과

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

결론

- Retriever 모델 학습 시 2개 이상 도메인을 사용해 학습을 했을 때 검색 성능이 더 높아졌다는 것을 검증했다.
- 질문 형식에 이전 대화 기록을 포함했을 때 질문과 관련된 문서 검색을 더 잘 한다는 결과가 나왔다.

Q & A