증강된 질문을 이용한 RoBERTa 기반 Dense Passage Retrieval

박준범, 홍범석, 최원석, 한영섭, 전병기, 나승훈
전북대학교, LGU+

발표 순서

- 서론
- 관련 연구
- 생성 모델을 이용한 질문 생성
- 증강 방법
- 실험 및 결과

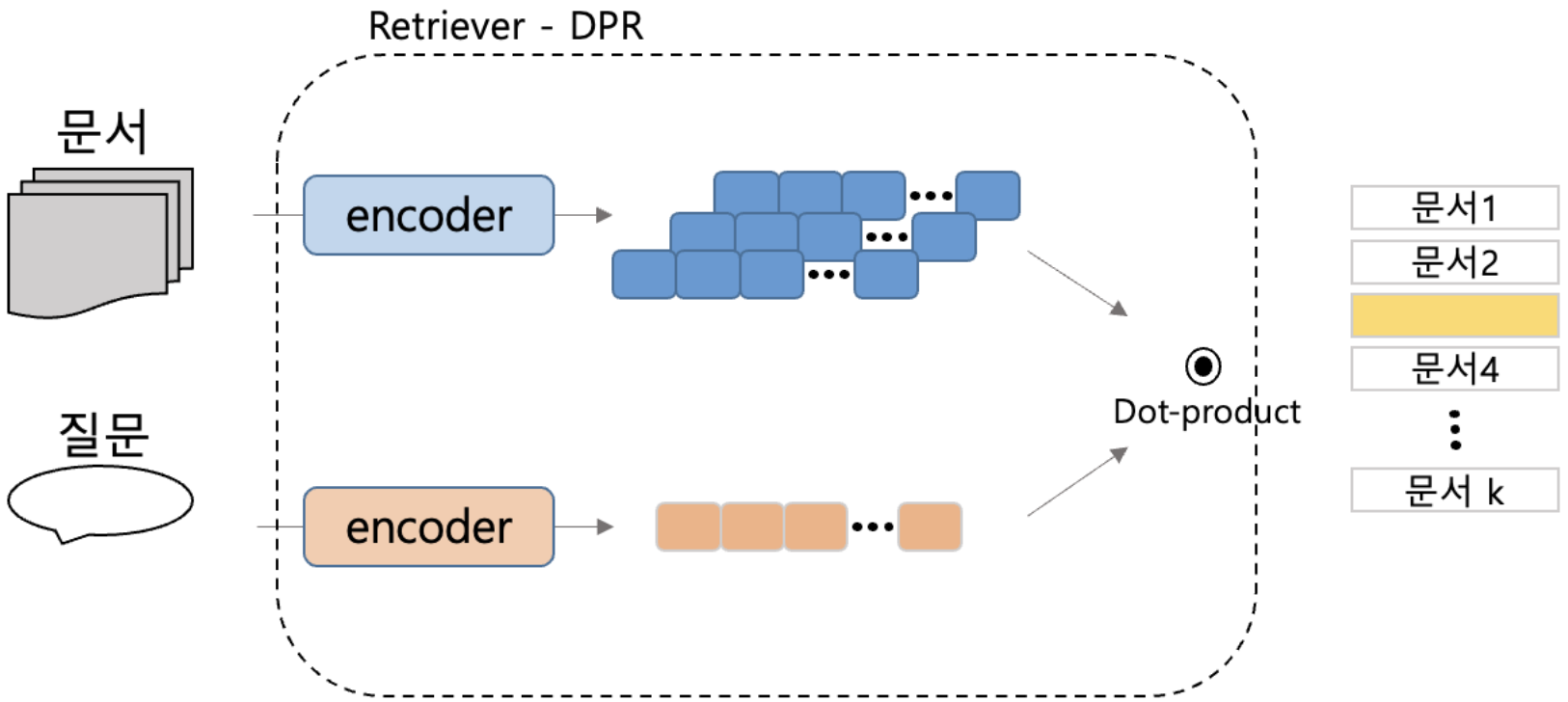
서론

(Dataset augmentation: 문서 기반 대화 모델의 사전학습 보강을 위한 데이터셋 증강)

- 답변 생성을 위한 다중 문서 기반 대화 시스템 속 검색 모듈의 중요성
- 문서에 기반한 응답 시스템 연구에 사용될 한국어 대화 데이터 셋이 부족함
- 생성 모델(T5-base)를 이용하여 문서와 관련된 질문을 생성하고 새로운 입력 쌍을 구성하여 검색 모듈의 강화 방법 모색
- RoBERTa-base를 이용한 Dense Passage Retrieval 검색 모듈에 데이터를 증강하여 강화시킨 PRAQ 제안
  (Pretrained RoBERTa with Augmented Query)

관련 연구

- Retrieval for response generation

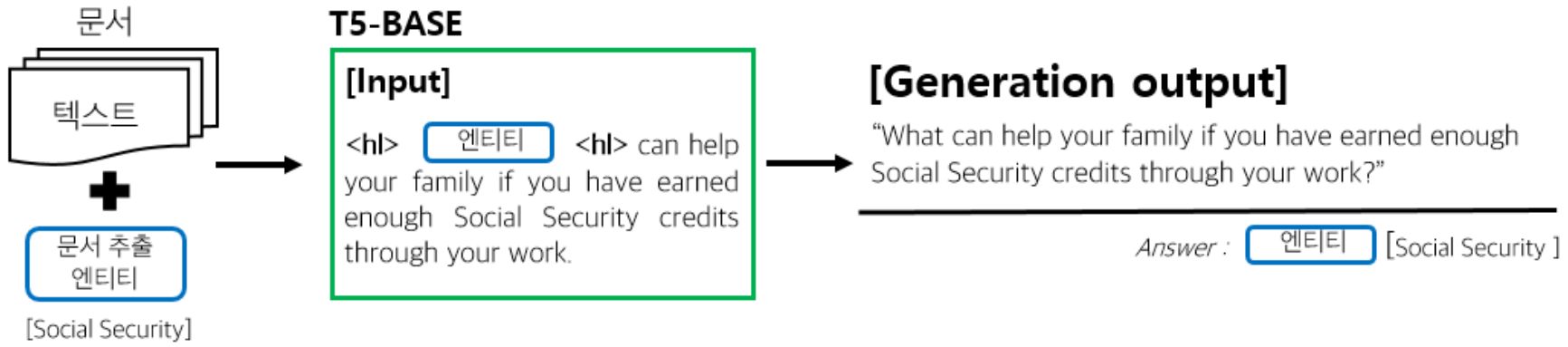


- Dense Passage Retrieval

  - 이중 인코더 프레임워크를 이용하여 얻어낸 질문과 문서의 dense representation간 유사도
  - 질문이 문서와 의미적으로 유사성을 갖는다면 질문과 문서를 관련이 있다고 매칭시킬 수 있음
  - Hard negative sampling을 활용한 성능 개선을 시도할 수 있음.

생성 모델을 이용한 질문 생성

- Generation Overview



- Generation Example

Entity	**Benefit Calculators**
Text	… check our **Benefit Calculators** for an estimate of the benefits your family could receive …
Generated Question	What can you use to estimate the amount of benefits your family could receive if you died right now?

Entity	**DHS** (Department of Homeland Security)
Text	…If **DHS** did not authorize you to work , we will change your name on our records but we cannot…
Generated Question	Who will change your name on our records but not issue you a corrected card?

문서-질문 세트

- Document Grounded Query-Passage set

Document Passages		
4,110		
Train	Valid	Test
21,451	4,201	4,094

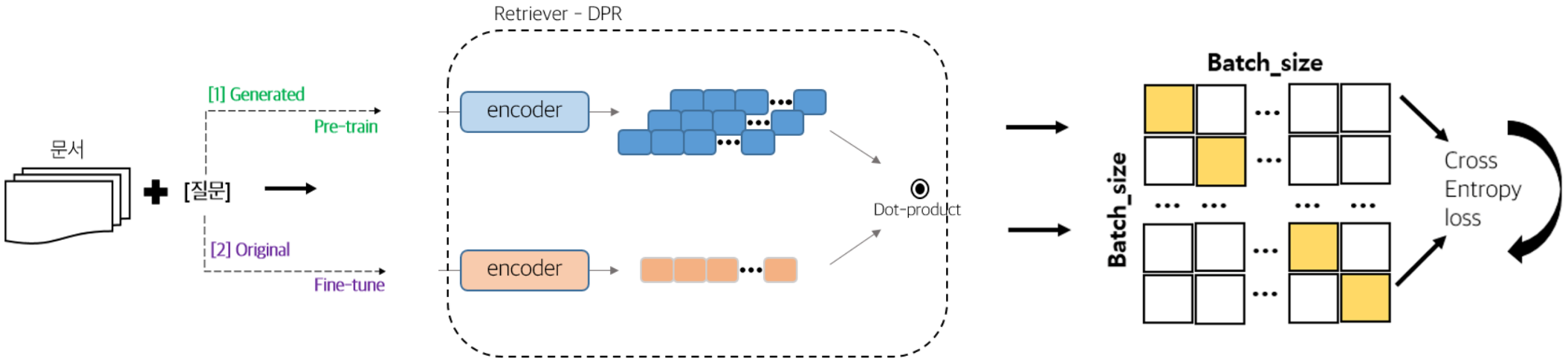
Document Passages	
4,110	
Entity	Generated Question
4,625	20,135

MultiDoc2Dial : 488개의 문서로부터 기반하여 평균적으로 약 14번의 발화가 오가는 4,796개의 대화 데이터

- 문서에서 추출해낸 엔티티 하나 당 약 네 개 이상의 질문 생성
- 단순한 1회성의 발화로 대화가 진행되면서 쌓이는 기록이 포함된 대화의 형태를 띄고 있지 않음
- 생성 질문을 이용한 파인 튜닝시 기존 대화를 이용했을 때 만큼의 성과는 기대하기 어려움

증강 방법

- Training Process



[1] : 생성된 질문 - 문서 쌍을 DPR 인코더에 사전 학습
-> PRAQ

[2] : 기존 질문 - 문서 쌍을 이용한 파인 튜닝

Similarity of query-passage

$$\mathrm{sim}(\mathbf{q},\mathbf{p}) = \mathbf{E}_Q(\mathbf{q})^T \mathbf{E}_P(\mathbf{p})$$

In-Batch Cross Entropy loss

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

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

실험

- 실험 셋팅

- In-Batch 학습법(60) : 다른 질문-문서 쌍들과 배치를 형성하여 질문-정답 문서, 질문-오답 문서의 유사도 차이를 효율적으로 계산할 수 있음
- Dataset : MultiDoc2Dial
- Epoch : 30
- Learning Rate : 2e-05
- Metric : MRR(Mean Reciprocal Rank) , R@K (Recall at top-k)

1) Pretrained RoBERTa with Augmented Query

2) + Hard negative (TF-IDF)

- TF-IDF 유사도를 기준으로 각 질문마다 상위 100개의 문서들을 추림.
- 파인 튜닝시, 매 epoch마다 무작위로 선정한 하나의 오답 문서를 hard negative 로 활용하여 학습
- 1 개의 negative -> batch size : batch + 1

결과

- 실험 결과

	MRR	R@1	R@5	R@10	R@50
RoBERTa-base	57.16	43.62	73.61	82.95	94.77
PRAQ(ours)	**58.12**	**44.82**	**74.76**	**83.53**	**95.02**
RoBERTa-base + TF-IDF	57.41	44.74	72.64	81.72	93.94
PRAQ + TF-IDF (ours)	**58.97**	**46.31**	**74.76**	**83.14**	**94.11**

표 3. Multidoc2dial 셋에서의 RoBERTA-base, PRAQ의 DPR 성능 비교 실험 결과

추가 연구 방향

- Generating more sophisticated Question/ Template - based

Entity	**Benefit Calculators**
Text	… check our Benefit Calculators for an estimate of the benefits your family could receive …
Generated Question	What can you use to estimate the amount of benefits your family could receive if you died right now?

Entity	**Benefit Calculators**
Text	… check our Benefit Calculators for an estimate of the benefits your family could receive …
Retrieved Text	…use our **Benefit Calculators** to determine how much you would get if you became disabled right…
Generated Question	What can you use to determine how much you would get if you became disabled right now?

- 동일한 엔티티를 갖는 다른 텍스트를 검색하여 생성될 질문의 기본 틀로 삼음
- 템플릿 기반 질문 생성, Ex) [Wh, 검색된 텍스트, 수식 문구]
- **기존 맥락과 일치하지 않을 수 있는 검색된 문장이 단순한 엔티티 일치보다 더 복잡한 관계를 학습시킬 수 있음**

Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering [Fabbri et al' 2020,ACL]

감사합니다