엔티티 중심 멀티 턴 대화 증강을 이용한 Dense Retrieval

박준범◦ , 나승훈 [전북대학교 인지컴퓨팅 연구실]
홍범석, 최원석, 한영섭, 전병기 [LG 유플러스]

서론

[검색 모듈 성능 개선의 중요성]

1) 사용자가 입력한 발화에 답변을 하기 위해 해당 정보를 가지고 있는 문서를 찾는다.

2) 찾은 문서의 정보를 이용하여 적절한 답변을 생성한다.

-> 적절한 문서를 찾는 검색 모듈의 성능이 최종 답변 생성에 영향을 미친다.

[문서 기반의 대화 데이터 생성의 필요성]

- 검색 모듈로 Dense Passage Retrieval 방식을 사용할 경우
- 모델 학습에 필요한 문서 기반 대화 데이터 셋 양의 한계

-> 문서를 보고 진행된 두 턴 이상의 대화 데이터 증강 방식의 필요성 발생

[데이터 생성, 성능 개선 방식 제안]

1) 문서로부터 Named Entity Recognition 작업을 통한 엔티티 추출

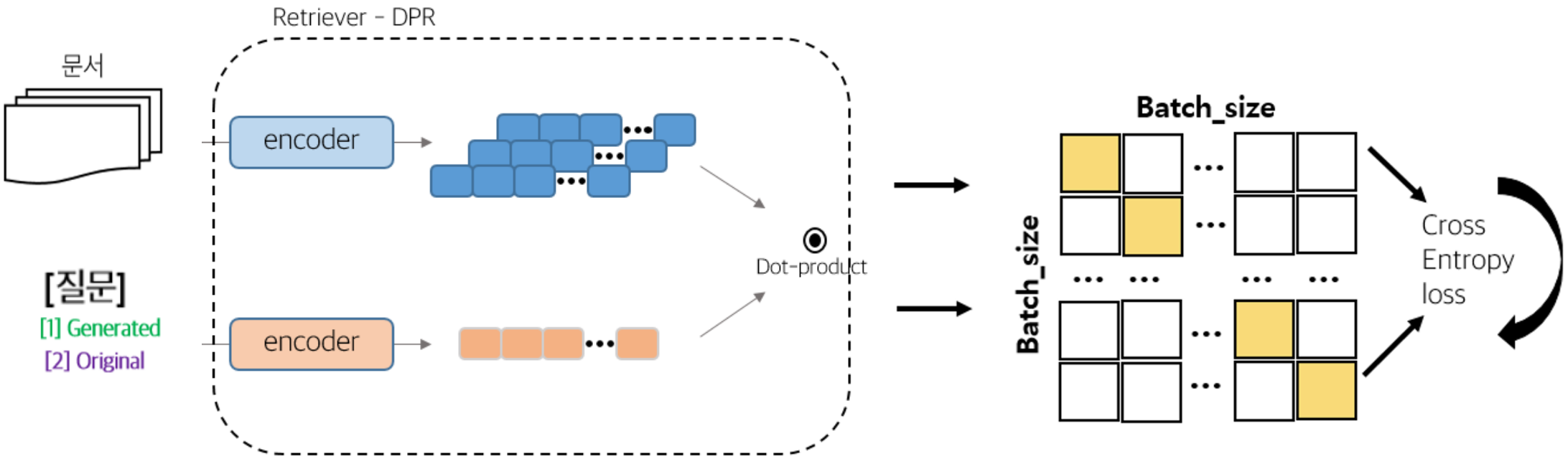
2) 추출한 엔티티를 답변으로 삼는 질문을 T5-base 모델로부터 생성

3) 생성된 질문들을 조합하여 두 턴의 대화 데이터 조합

4) 조합된 대화 데이터를 필터링 과정을 거쳐 DPR 사전학습에 적합한 학습 데이터로 증강

Dense Passage Retrieval

[학습 개요]



- 질문-문서의 Similarity

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

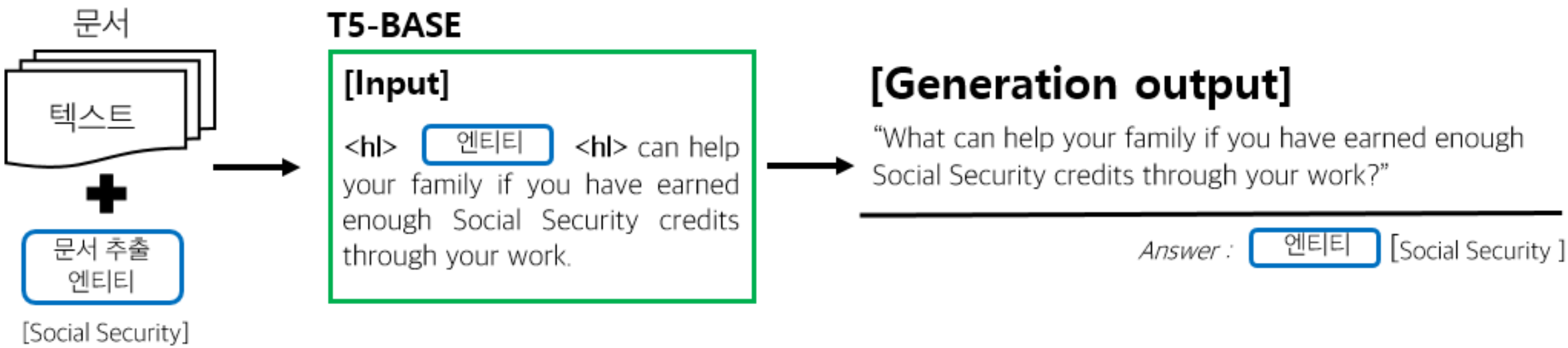
- In-Batch Cross Entropy loss

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

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

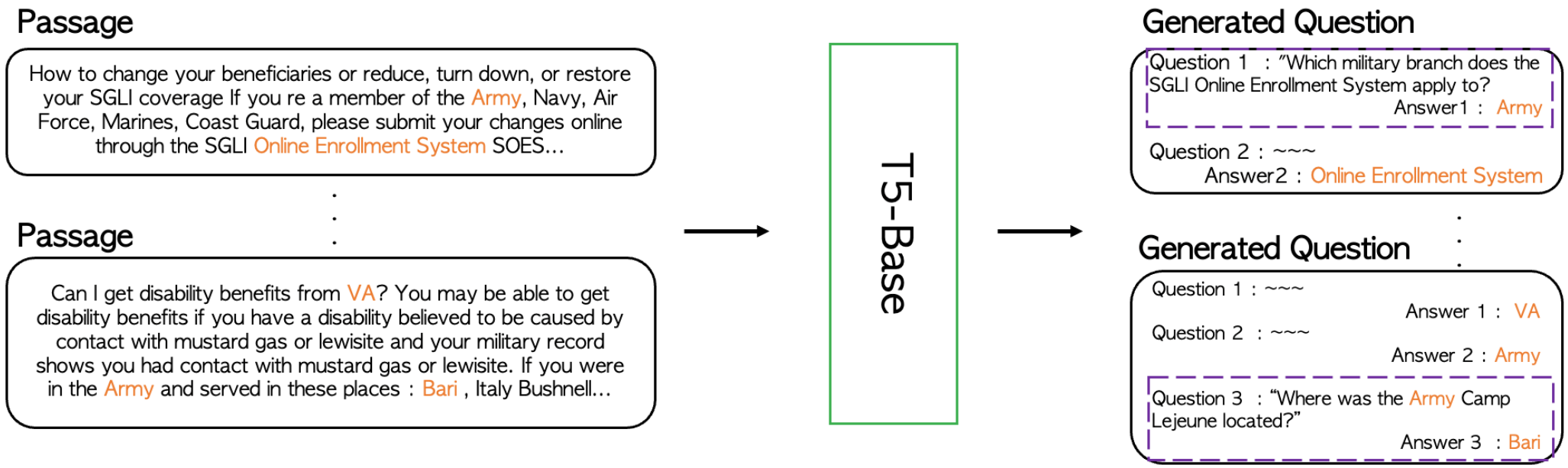
엔티티 기반 싱글 턴 대화 생성

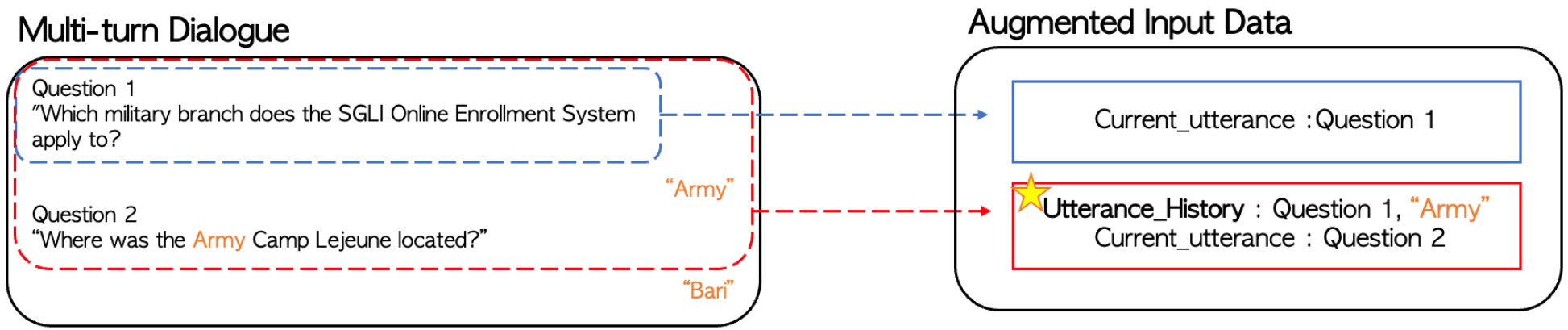
[엔티티 기반 싱글 턴 질문 생성 개요]



엔티티 기반 멀티 턴 대화 생성

[엔티티 기반 멀티 턴 대화 생성 개요]





대화 데이터 증강

[선행 질문에 대한 후속 질문 필터링 규칙]

- 정답 엔티티가 문장에 포함되어 있는 경우 제외.
- 문장 발생 빈도가 높은 경우의 엔티티 제외. Ex) ['Social Security', 245346], ['one', 107371], ['New York', 83873]
- 선행 질문과 후속 질문의 정답을 포함하는 passage가 서로 같은 경우 제외.

  -> 다른 passage에서 후속 질문을 선택함으로써 선행 질문의 passage로부터 알 수 없는 정보를 학습에 제공.

  -> 검색 기능 학습에 변별력 부여 기대.

[생성된 질문 데이터 셋]

- 기존 데이터 구성 : 1) passage – 3,820 2) 엔티티 – 4,640 // 학습 데이터 셋 : **21,451**
- 싱글 턴 데이터 생성 : 18,261
- 멀티 턴 데이터 조합 : 약 3,400,000
- 필터링 이후 증강된 데이터 셋 : **약 340,000**

Dense Retrieval 증강 실험

[실험 세팅]

- 배치 사이즈-128, 학습률- 2e-05, Optimizer-Adam, In Batch negative
- Epoch - baseline 및 파인 튜닝 : 30 epoch, 사전학습 : 10 epoch
- Dataset - MultiDoc2Dial
- Metric - MRR(Mean Reciprocal Rank) , R@K (Recall at top-k)

[증강 데이터 사전학습을 추가한 DPR]

- 생성한 질문들 중 무작위로 선정한 20,000개의 멀티 턴 대화 사전학습 데이터를 구성.

  각각의 멀티 턴 대화 마다 총 두개의 입력 데이터 1){선행 질문} // 2){대화 기록[선행 질문, 정답 엔티티], 후속 질문}

- 40,000개의 학습 데이터를 이용한 사전학습은 RoBERTa-base를 인코더로 삼아 10 Epoch 진행, 기존 데이터로 미세 조정하여 baseline 성능과 비교.

[증강 데이터 수에 따른 DPR 성능]

- 증강 데이터의 수를 실험마다 10,000개, 20,000개, 30,000개로 다르게 하여 증강 데이터 수에 따른 성능의 변화 확인.

Dense Retrieval 증강 실험 결과

[증강 데이터 사전학습을 추가한 DPR 성능]

	MRR	Recall@1
RoBERTa-base	58.97	46.11
Pretrain with Data Augmentation	**60.35**	**47.70**

- 증강 데이터로 사전학습 후 기존 데이터로 미세조정한 결과, baseline보다 높은 성능 확인

[증강 데이터 수에 따른 DPR 성능 변화]

Number of Augmented Data	MRR	Recall@1
10,000	60.11	47.14
20,000	60.19	47.19
30,000	60.22	47.38
40,000	**60.35**	**47.70**

- MRR 지표의 경우 최소 1.1%p, Recall 지표의 경우 최소 1%p의 상승
- 증강 데이터 수 증가에 따른 성능의 상승치도 확인 가능

결론

- 문서 기반 대화 시스템을 위한 검색 모듈 (DPR) 학습에 필요한 대화 데이터 셋의 부족
- 엔티티를 중심으로 한 멀티 턴 대화 생성 방식 및 학습 데이터로 증강 방법 제시
- 생성된 멀티 턴 대화 데이터의 검색 모듈 학습에 대한 기여를 확인하기 위해 DPR의 인코더 모델에 사전학습 실험 진행.
- 단발성 질문들을 조합하여 만들어진 멀티 턴의 대화도 사전학습을 통해 검색 모듈에 긍정적 영향 확인.