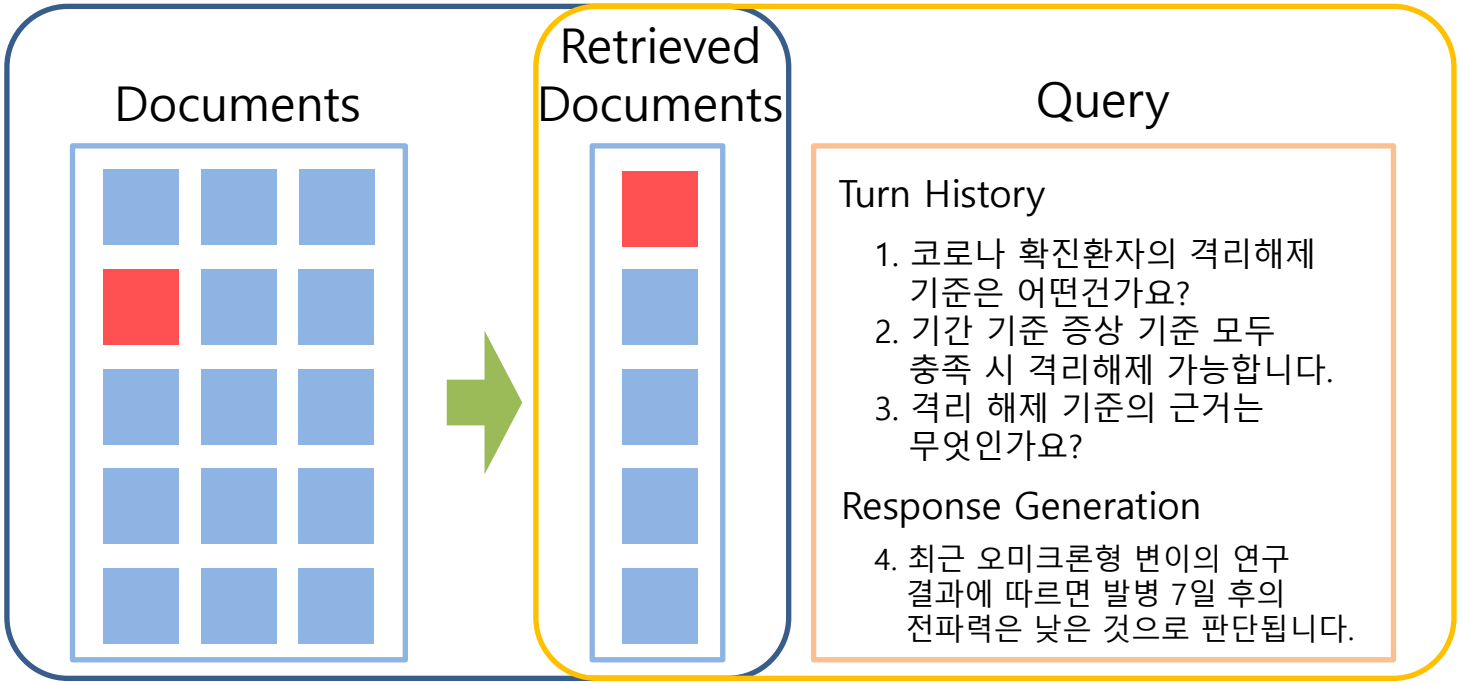
BART-SCA: Supervised Cross-Attention기반 BART를 이용한 Document-Grounded 응답생성

Heyon-Jun Choi, Seung-Hoon Na
Cognitive Computing Lab.
Jeonbuk National University

Document-grounded QA

- 주어진 질의에 대해 특정한 문서 내의 정보에 기반하여 답변을 생성
  - 소비자 상담이나 보험설계 같은 정확한 사실에 기반한 질의 답변 시스템에 해당
  - 답변 생성에 필요한 문서를 찾아내는 검색단계와, 검색된 문서에 기반하여 답변을 생성하는 생성단계 두 단계로 작업이 구성됨



MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents[Song Feng, et al., 2021]

- **사용자와 에이전트간 대화로 구성된 데이터셋**
  - **답변 생성에 필요한 문서는 별도로 주어지기 때문에, 검색 모델을 통해 이를 선택해야 함**
  - **하나의 대화 히스토리 내에서 답변 생성에 필요한 문서가 하나가 아니라 여러 개가 존재 할 수 있기 때문에, 각 답변에 맞는 문서를 각각 선택해야 함**
  - **각 문서는 문단으로 분리 될 수 있고, 각각의 문단에서 실제 답변 생성에 필요한 Reference가 포함됨**

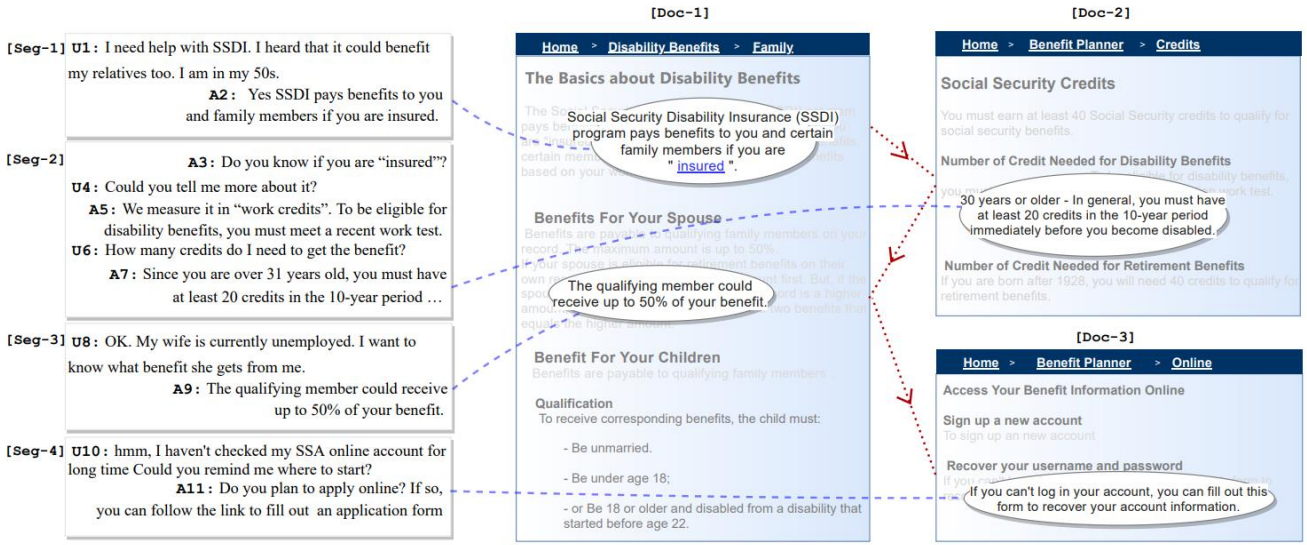


Figure 1: A sample goal-oriented dialogue (on the left) that is grounded in three relevant documents (on the right).

Distilling Knowledge from Reader to Retriever for Question Answering[Gautier Izacard, Edouard Grave, 2020]

- **QA에서 Response Generation 모델의 Cross-Attention의 가중치는 연관 있는 Text Segment일 수록 높게 계산 될 것이라는 가설을 세움**
  - **Fusion-in-Decoder 기반 Response Generation 모델을 통해 문서 검색이 가능하고, 문서 검색 모델보다 높은 성능을 낼 수 있음을 확인**
    - **FiD 모델은 여러 문서와 쿼리를 인코더를 통해 각각 인코딩 한 후, 디코더의 입력으로 하나로 Concatenate하여 입력되는 문서의 개수를 확장한 모델임**
  - **이 결과를 통해 생성 모델로부터 검색 모델로 지식 증류를 적용, 검색 모델의 성능을 더 끌어올릴 수 있는 방법을 제시**
- **이번 실험에서는 이 가정을 뒤집어서, 연관있는 Text Segment에 높은 가중치가 계산 되도록 할 경우, 생성 성능을 향상 시킬 수 있을 거라는 가설 하에 실험**

Supervised Cross-Attention

- Transformer 구조에서 인코더를 통해 인코딩된 모델의 입력X는 디코더의 Cross-Attention에서 디코더의 Self-Attention 출력 H와 합쳐짐

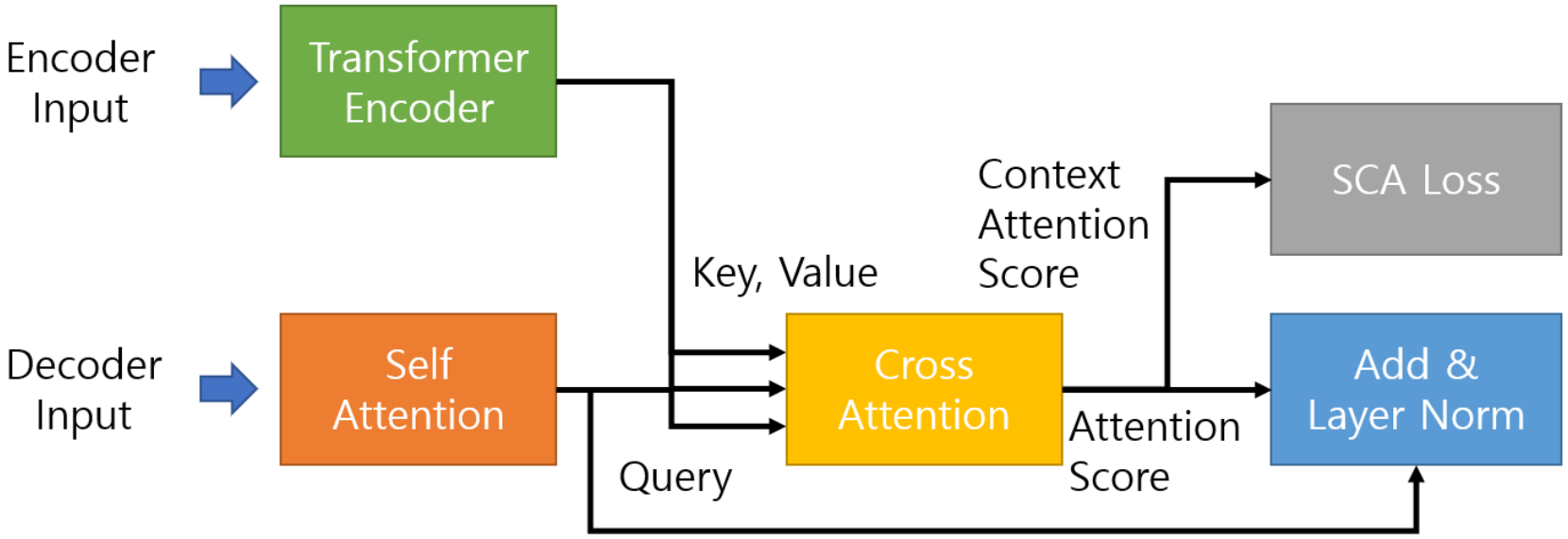
$$\mathbf{Q} = \mathbf{W}_q\mathbf{H}, \mathbf{K} = \mathbf{W}_k\mathbf{X}, \mathbf{V} = \mathbf{W}_v\mathbf{X}$$

$$\alpha_{i,j} = \mathbf{Q}_i^T\mathbf{K}_j, \tilde{\alpha}_{i,j} = \frac{\exp(\alpha_{i,j})}{\sum_m \exp(\alpha_{i,m})}$$

$$\mathbf{O}_i = \mathbf{W}_o \sum_j \tilde{\alpha}_{i,j}\mathbf{V}_{i,j}$$

- 따라서, α는 디코더가 인코더의 출력을 얼마나 참조 할 지에 대한 가중치에 해당
- Reference에 해당하는 토큰에 대해 높은 가중치가 계산되도록 Attention Grounding을 적용

Supervised Cross Attention



Supervised Cross Attention

- **Attention Grounding을 위해 k_0을 추가**
  - **k_0 는 학습 가능한 파라미터로 Reference의 경계선이 되도록 학습되며 Training 시에만 사용됨**

$$\mathbf{Q} = \mathbf{W}_q\mathbf{H}, \mathbf{K} = \mathbf{W}_k\mathbf{X}, \mathbf{K}' = [k_0, \mathbf{K}]$$

$$\alpha_{i,j} = \mathbf{Q}_i^T\mathbf{K}'_j$$

  - **r 에 해당하는 Reference 토큰에 대해 k_0의 어텐션 가중치 $\alpha_{i,0}$보다 높은 가중치를, 그 이외에는 낮은 가중치를 가지도록 학습**

$$\mathcal{L}_{sca} = \begin{cases} -\log\dfrac{\exp(\alpha_{i,j})}{\exp(\alpha_{i,0}) + \exp(\alpha_{i,j})} & j \in r \\ -\log\dfrac{\exp(\alpha_{i,j})}{\exp(\alpha_{i,0}) + \exp(\alpha_{i,j})} & else \end{cases}$$

실험 구성

- 문서 검색 모델, 리랭킹 모델, 답변 생성 모델로 구성된 프레임워크를 통해 실험
  - 문서 검색 모델은 RoBERTa base 기반 DPR 구조를 사용
    - 인코더는 Query와 Document가 하나의 인코더를 공유하는 Shared-encoder를 적용
    - 인코딩된 Query와 Documet를 벡터 내적을 통해 유사도 계산

$$sim(q, d) = E_Q(q)^T E_D(d)$$

  - 리랭킹 모델은 RoBERTa base 기반 Cross-Encoder 구조를 사용
    - Query와 Document를 결합, 인코더 출력 중 <cls>토큰의 값을 통해 유사도 계산

Model	R@1	R@5	R@10	MRR@5
Retriever	50.44%	78.05%	85.34%	61.09%
Reranker	62.46%	82.74%	89.07%	70.46%

실험 구성

- 문서 검색 모델, 리랭킹 모델, 답변 생성 모델로 구성된 프레임워크를 통해 실험
  - 답변 생성 모델은 BART base 모델을 사용
  - 256토큰의 대화, 768 토큰의 문서를 인코더 입력으로 하여 입력된 질의에 대한 다음 답변을 생성
  - Baseline 모델과 SCA를 추가한 모델 사이의 차이는 Attention Grounding 이외에는 없음
- 성능 측정은 Multidoc2dial 데이터셋의 Seen data를 통해 측정
  - Multidoc2dial은 4800여개의 대화와 490여개의 문서, 3800여개의 문단으로 구성됨
  - Seen Data는 Train set과 Dev set이 동일한 도메인에 존재함을 의미

실험 결과

Model	F1	Sacrebleu	Meteor	Rouge
Baseline	44.83%	28.89%	45.20%	42.63%
+ SCA	46.18%	30.48%	46.27%	43.98%

결론 및 향후계획

- 디코더의 Cross-Attention에 지도학습 과정을 추가하여 F1 지표에서 1.35%p의 성능 향상을 확인함
- Cross-Attention을 활용하여 추가적인 활용이 가능 할 것으로 보임
  - Self-Attention Grounding
  - Cross-Attention 기반 Retriever 지식 증류

감사합니다