

BERT와 Multi-level Co-Attention Fusion을 이용한 한국어 기계독해

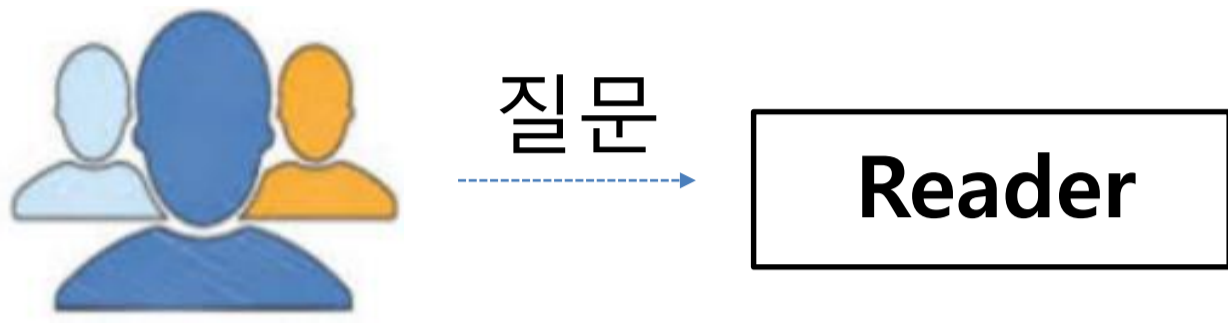
박광현¹, 나승훈¹, 최윤수², 장두성²
¹전북대학교, ²KT

khpark231@gmail.com, nash@jbnu.ac.kr, yunsu.choi@kt.com, dschang@kt.com

I. 기계독해

- 기계독해는 기계가 주어진 문맥을 이해하고 질문에 대한 답을 찾는 것을 목표로 함

Q: 대한민국의 수도는 어디야?



대한민국은 동아시아의 한반도 남부에 있는 공화국이다. 수도는 서울특별시이며, 국기는 태극기, 국가는 애국가(비공식), 공용어는 한국어와 한국 수어이다.

II. BERT

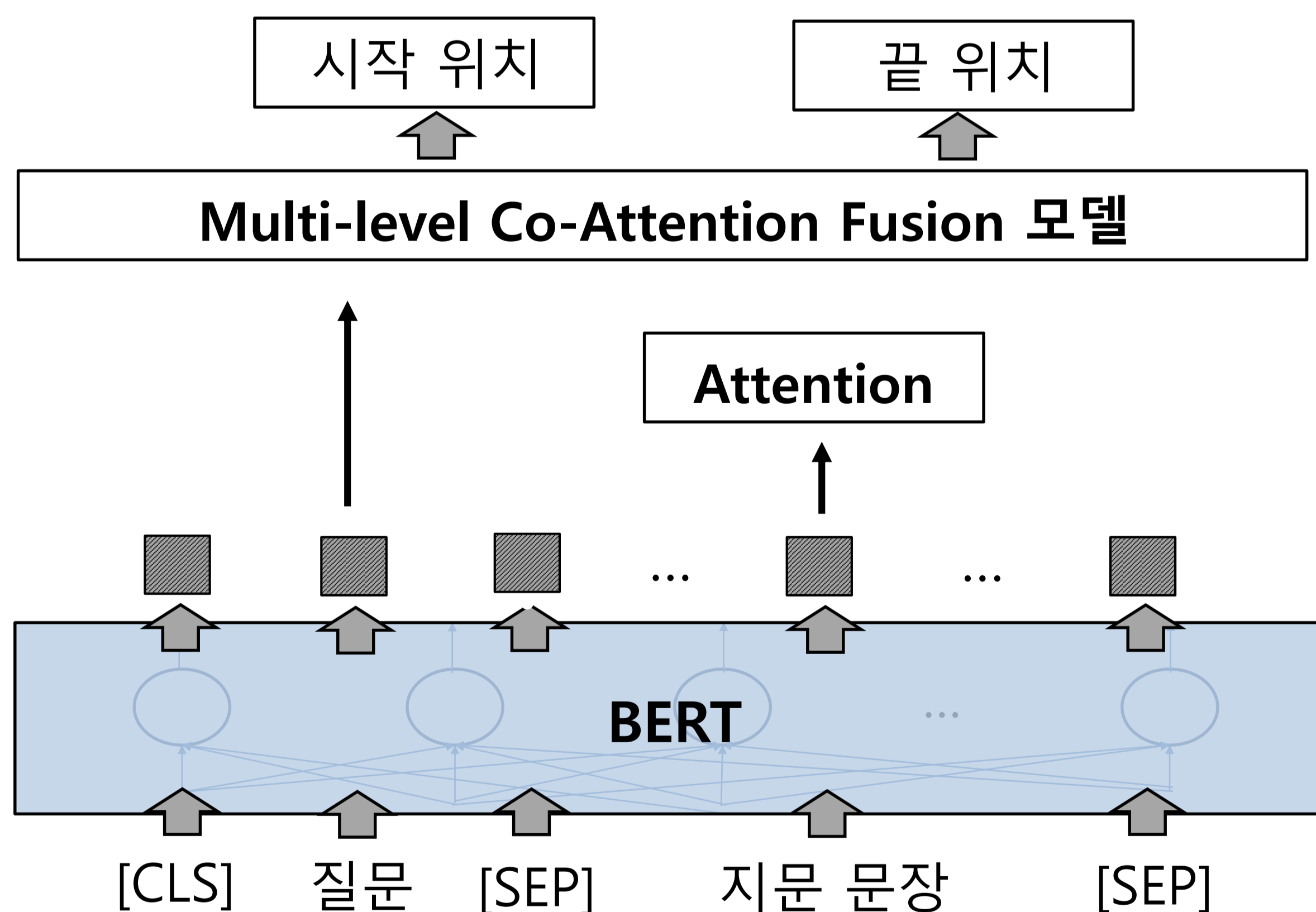
- Attention을 이용한 인코더인 Transformer 기반의 사전 학습 언어 모델
- Masking 단어 예측과 다음 문장이 적절한지 판단하는 2가지 문제로 사전 학습을 진행
- "[CLS] A문장 [SEP] B문장 [SEP]" 형태의 입력을 사용
- 시퀀스 길이가 512보다 긴 경우 처리를 할 수 없음
- 기계독해에서 긴 시퀀스를 한번에 처리하기 위해 문장을 여러 segment로 나누어 처리한 뒤 Attention을 적용해 전체 문장에 대한 정보를 활용

입력:

[CLS] 질문 [SEP] 지문 문장 1 [SEP]

...

[CLS] 질문 [SEP] 지문 문장 N [SEP]



BERT+MCAF 모델 구조

- N개의 지문 문장에 대해서 각각 인코딩을 한 후 질문 벡터는 N개의 지문 문장에 대해 동일하게 사용되기 때문에 N개 질문 벡터의 합의 평균을 사용하고 지문 벡터는 N개의 벡터를 시간축에 대해서 concat 하여 사용
- BERT의 L개 레이어의 정보를 모두 이용하기 위해 L개의 출력값의 가중평균을 사용

$$\bar{\mathbf{B}}^q = \frac{1}{N} \sum_i \bar{\mathbf{B}}_i^q \quad \bar{\mathbf{B}}^p = [\bar{\mathbf{B}}_1^p; \dots; \bar{\mathbf{B}}_N^p]$$

$$\mathbf{B}^q = \frac{1}{L} \sum_l \alpha_l^q \bar{\mathbf{B}}_l^q \quad \hat{\mathbf{B}}^p = \frac{1}{L} \sum_l \alpha_l^p \bar{\mathbf{B}}_l^p$$

$$\mathbf{B}^p = \text{MultiHeadAttention}(\hat{\mathbf{B}}^p)$$

III. Multi-level Co-Attention Fusion

- Multi-level Co-Attention Fusion에서는 3 단계로 질문과 지문에 대한 상호정보를 찾고 정답을 추출

- 인코딩 단계

- 앞에서 얻은 BERT의 출력값을 사용하여 지문(P)과 질문(Q)을 인코딩 하게 됨
- 모델이 깊어질수록 정보가 추상화 되기 때문에 여러 의미 수준의 정보를 고려하기 위해 여러 단계로 인코딩 하게 됨

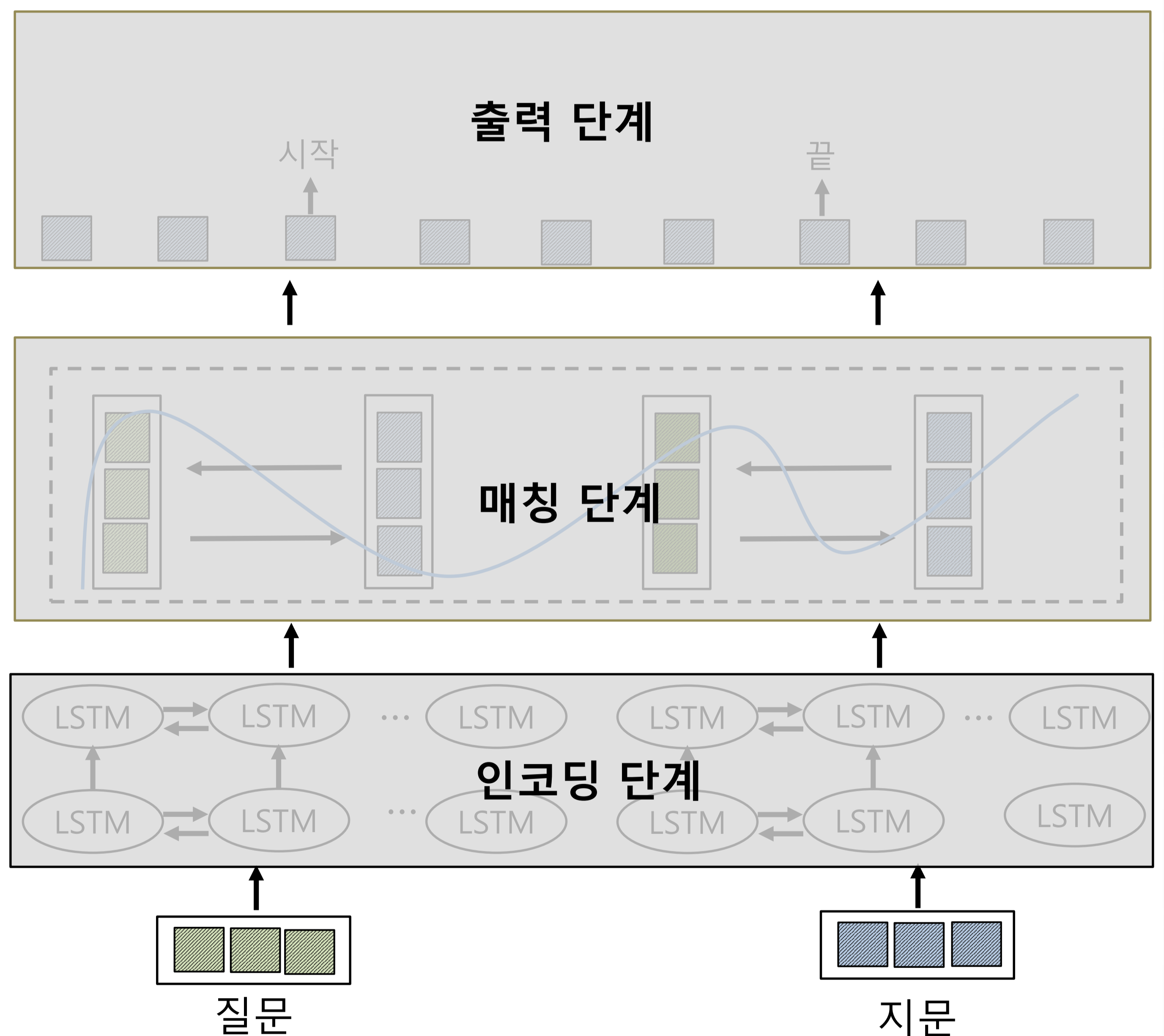
전주 → 전라북도 → 한국의 도시

- 매칭 단계

- 지문과 질문에 대한 상호정보를 얻는 단계
- 인코딩 단계에서 여러 단계로 얻어진 벡터들에 대해서 각각 어텐션을 수행한 후에 LSTM을 통해 정보를 하나로 결합

- 출력 단계

- 매칭 단계에서 얻은 지문과 질문의 상호정보를 이용해 지문에서 정답의 시작과 끝 위치를 찾는 단계



Multi-level Co-Attention Fusion 모델 구조

IV. 실험

- WordPiece 단위인 BERT(Multilingual)모델을 사용
- KorQuAD 데이터를 이용하여 실험을 진행

| 모델 | EM | F1 |
|-------------------|--------|--------|
| BERT+MCAF | 83.01% | 91.43% |
| BERT(Global)+MCAF | 84.04% | 92.43% |

KorQuAD Dev 데이터 셋 성능

- MCAF는 Multi-level Co-Attention Fusion을 나타냄
- BERT(Global)+MCAF는 본 논문에서 제안하는 BERT의 전체 문장에 대한 정보를 활용하였을 때의 성능임