NIL-Aware KnowBERT를 이용한 개체 연결 모델의 성능 향상

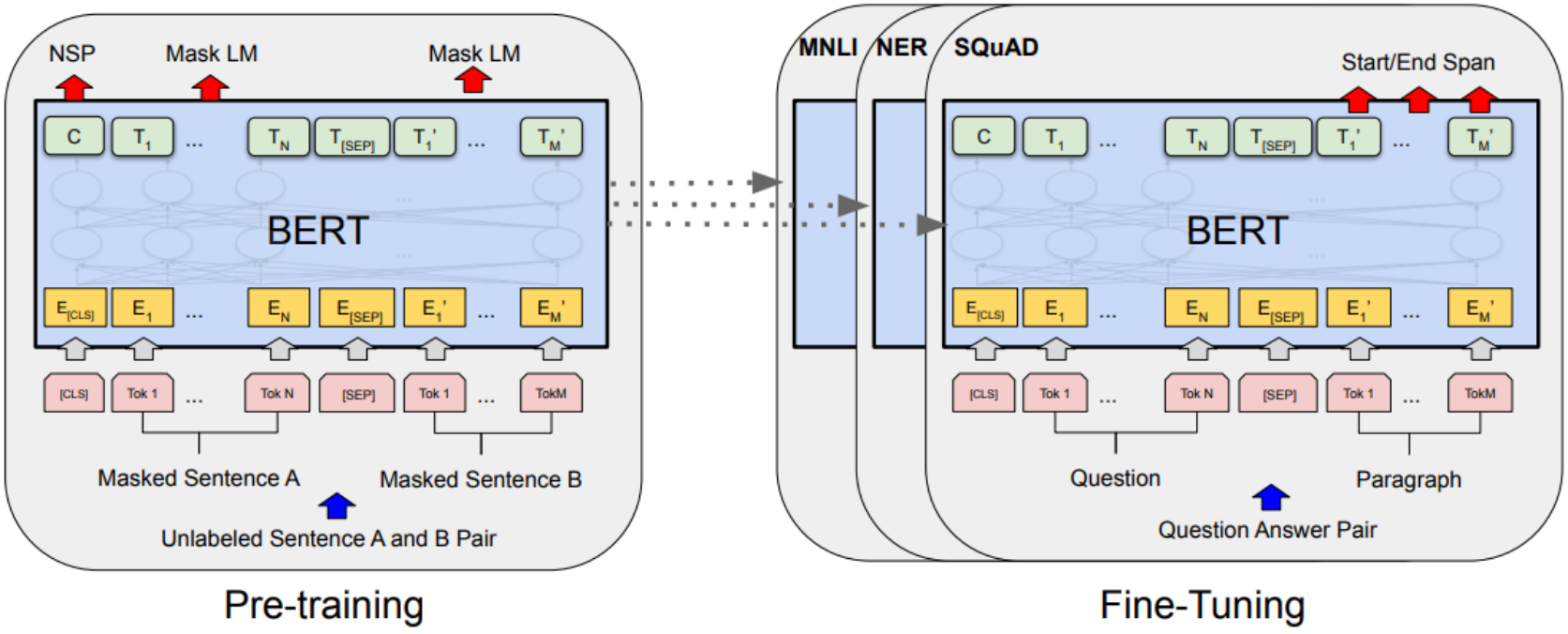
민진우[O1], 나승훈[1], 김현호[2], 김선훈[2], 강인호[2]

[1]전북대학교 인지컴퓨팅 연구실, [2]ETRI

목차

- 관련연구
- 한국어 nil_KnowBERT 모델
- 실험 세팅 & 실험 결과

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding(Yinhan Liu et al., '19)



- BERT : Bidirectional Encoder Representation from Transformer
  - 양방향의 Transformer를 이용하여 문장 내 임의의 단어를 예측하고 다음 문장을 예측하는 두 가지 Task로 언어 모델 학습
  - 응용 Task에 Fine-Tuning하는 방식으로 성능 향상

RoBERTa: A Robustly Optimized BERT Pretraining Approach(Yinhan Liu et al., '19)

- BERT의 최적화 모델
  - BERT의 학습 과정에서 최적화 되지 않은 부분을 최적화
    - **Dynamic Mask LM** 기존의 BERT 모델과 달리 RoBERTa에서는 BERT와 달리 매 학습마다 마스킹하는 단어를 다르게 하는 Dynamic Masking 방식을 사용
    - **NSP 테스크 제외** 다음 문장을 예측하는 NSP 태스크는 학습하지 않고 최대 토큰 길이 512에 가깝도록 다른 문서의 문장으로 채워 넣어 문서 길이가 미리 설정한 하여 학습 효율을 높임
  - 기존 BERT에서의 성능 향상

Model	SQuAD 1.1 EM	SQuAD 1.1 F1	SQuAD 2.0 EM	SQuAD 2.0 F1
Single models on dev, w/o data augmentation				
$\text{BERT}_{\text{LARGE}}$	84.1	90.9	79.0	81.8
$\text{XLNet}_{\text{LARGE}}$	**89.0**	94.5	86.1	88.8
RoBERTa	88.9	**94.6**	**86.5**	**89.4**
Single models on test (as of July 25, 2019)				
$\text{XLNet}_{\text{LARGE}}$			86.3†	89.1†
RoBERTa			86.8	89.8
XLNet + SG-Net Verifier			**87.0**†	**89.9**†

한국어 Contextual word representations 연구

- ETRI BERT
  - 한국어의 특성을 반영한 두 가지 단위의 BERT 언어 모델 제공
    - 형태소 분석 기반 모델 : 입력 문장을 형태소 분석기를 이용해 형태소 단위로 분리하여 입력 토큰으로 사용.

    - 어절 기반 언어 모델 : 문자의 어절에서 고빈도로 발생하는 문자(음절)들을 결합하여 토큰을 구성한 BPE(Byte Pair Encoding)방식.

  - 높은 성능 향상
    - BERT를 이용한 한국어 의존 구문 분석[박천음, KCC '2019]

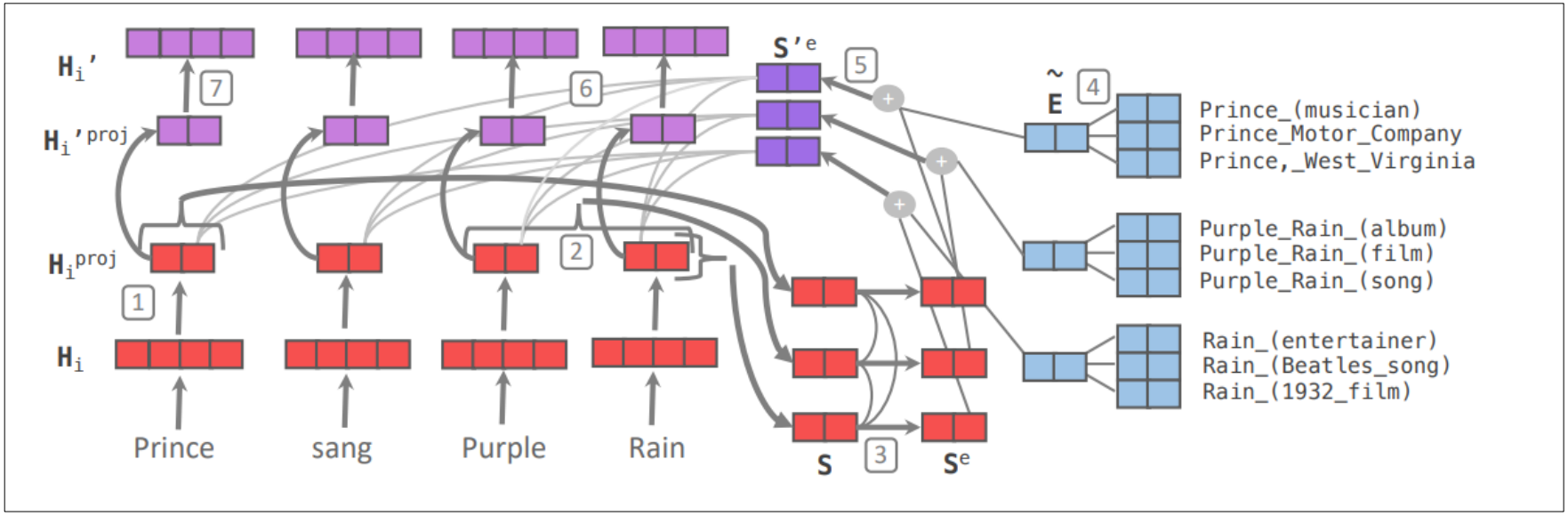
Dependency parsing	UAS	LAS
이창기[16] with MI	90.37	88.17
나승훈[6]: deep biaffine attention	91.78	89.76
박천음[5]: 포인터 네트워크	92.16	89.88
안휘진[10]: deep biaffine + 스택 포인터 네트워크	92.17	90.08
박성식[11]: ELMo + 멀티헤드 어텐션	92.85	90.65
BERT + LSTM deep bilinear	93.85	91.78
BERT + LSTM deep biaffine	**94.06**	**92.00**

한국어 Contextual word representations 연구

- 한국어 RoBERTa 모델(KCC ‘19)
  - RoBERTa를 한국어 코퍼스 상에서 학습
  - Hybrid Tokenizer 제안
    - 형태소 토큰 5만개와 BPE 토큰 2만개를 단어장으로 구성
    - 분석된 형태소를 형태소 단위를 우선적으로 단어장에서 매칭한 후 해당 형태소가 미등록어일 경우 형태소를 BPE 단위로 토크나이징 하는 하이브리드 방식을 사용

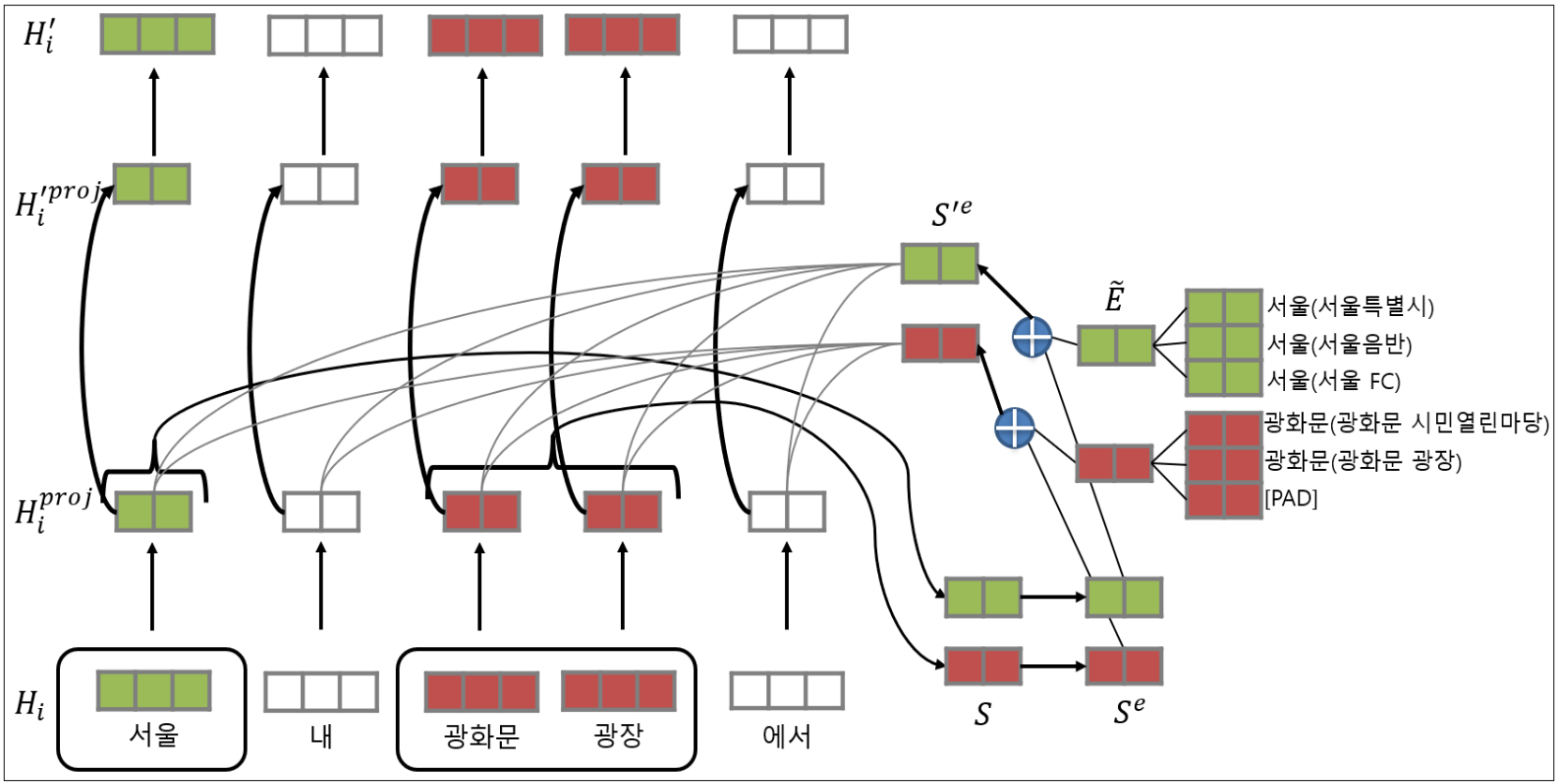
원문
고전주의와 바로크는 공통의 면이 있다.
형태소 분석 결과
고전주의/NNG, 와/JC, 바로크/NNG, 는/JX, 공통/NNG, 의/JKG, 면/NNG, 이/JKS, 있/VV, 다/EF, ./SF
토크나이징 결과
_고전, 주의, 와/JC, 바로크/NNG, 는/JX, 공통/NNG, 의/JKG, 면/NNG, 이/JKS, 있/VV, 다/EF, ./SF

Knowledge Enhanced Contextual Word Representations (Peters et al., '19)



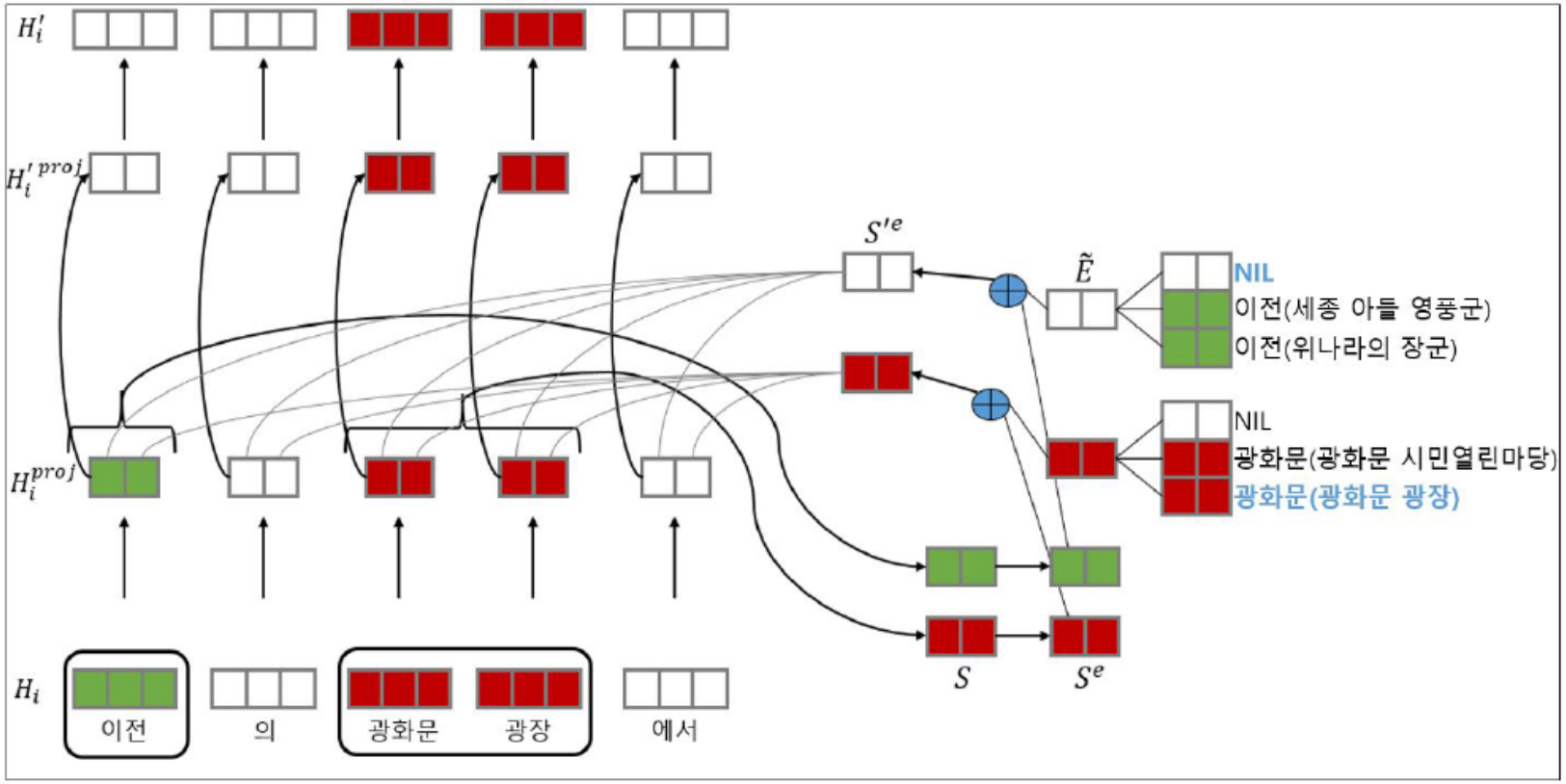
- KnowBERT - 지식 향상된 문맥 단어 표현
  - 다양한 지식 베이스로부터 엔티티 링커 모듈을 학습 후 링커 모듈을 통해 생성된 후보 엔티티 점수를 문맥 단어 표현에 반영하는 재문맥화 과정을 통해 지식 향상된 문맥 단어 표현을 얻음
  - relationship extraction, entity typing, and word sense disambiguation 등의 응용 테스크에서 기존 BERT 모델 대비 성능 향상.

KnowBERT를 이용한 지식 그라운딩 (Knowledge-grounded)된 한국어 자연언어처리, KCC '20)



- 한국어 KnowBERT
  - 한국어 wikipedia를 지식베이스로 하여 knowBERT 학습 후 개체명 타이핑, 링킹 등에서 개선된 성능을 보임

NIL-Aware KnowBERT 모델



- NIL-Aware KnowBERT
  - KnowBERT의 링킹 모듈에서 NIL을 탐지 한 후 NIL로 탐지된 멘션에 대해서는 점수를 반영하지 않고 재문맥화를 수행

NIL-Aware KnowBERT 모델

- 재문맥화 모듈
  - RoBERTa 내의 특정 층 i에서의 문맥 표현 H_i을 변환, 은닉 표현 H_i projection

$$H_i^{proj} = H_i W_1^{proj} + b_1^{proj}$$

  - 문서 내의 m 번째 멘션에 대한 엔티티-Span 표상 s_m^e은 projection된 문맥 단어 표현에서 멘션의 시작, 끝 위치의 표상의 평균

  - 멘션 m의 후보 엔티티 점수 및 엔티티 링킹 손실 계산

$$\varphi_{mk} = MLP(p_{mk}, \mathbf{s}_m^e \cdot \mathbf{e}_{mk})$$

$$L_{EL} = \sum_m \left(\frac{exp(\varphi_{mg})}{\sum_k exp(\varphi_{mk})} \right)$$

  - 향상된 엔티티-Span 표상 $s_m'^e$
    - 엔티티 점수와 엔티티 임베딩의 weight sum 결과를 기존 엔티티-Span 표상 s_m^e에 반영하여 계산

$$\tilde{\mathbf{e}}_m = \sum_k \varphi_{mk} \mathbf{e}_{mk}$$

$$\mathbf{s}_m'^e = \mathbf{s}_m^e + \tilde{\mathbf{e}}_m$$

NIL-Aware KnowBERT 모델

- 재문맥화 모듈
  - 재문맥화 과정
    - 향상된 엔티티-Span 표상 $s_m'^e$을 키(key),값(value)으로 하는 멀티 헤드 어텐션을 수행한 후 projection을 한번 더 수행

$$H_i'^{proj} = MLP(MultiHeadAttn(H_i^{proj}, s'^e, s'^e))$$

$$H_i' = H_i'^{proj} W_2^{proj} + b_2 + H_i$$

    - 지식 향상된 문맥 단어 표상 H_i'에 대해 Masked LM을 목적함수로 언어 모델 학습

- NIL-Aware KnowBERT
  - 재문맥화 과정
    - 내부 엔티티 링커에서는 NIL을 추가한 총 $K+1$개의 후보 엔티티 중에 후보 멘션의 NIL을 포함한 손실 함수를 계산
    - 주어진 멘션에 대해 예측한 엔티티가 NIL인 경우 다음 수식과 같이 재문맥화 과정 속에 해당 멘션의 엔티티 정보를 반영하지 않도록 다음과 수식과 같이 제한하도록 한 후 재문맥화 과정을 수행

$$\tilde{\mathbf{e}}_m = 0 \quad (pred_m == NIL)$$

데이터 셋

- 사전 학습
  - 동일한 한국어 위키피디아 문서를 사용하였으며 위키피디아 문서 내에서 엔티티는 아래 그림과 같이 링크(Hyper-Link)되어 있음
  - 엔티티사전을 이용하여 매칭된 멘션 중 하이퍼링크가 없는 텍스트에 대해서는 NIL 엔티티로 간주하여 학습을 수행

... ... 이전의 광화문 광장에서 옮겨진 스케이트장에서 특별 행사가 치뤄진다는 뉴스가 보도되었다. ... ...

<사전 매칭이 완료된 위키피디아 문서>

- 응용 태스크 [엔티티 링킹]
  - 위키피디아 문장 6만 문장 중에서 학습 데이터 3만문장, 개발 데이터 1만 문장, 평가 데이터는 2만 문장으로 구성

실험 결과

- 실험 세팅
  - 개체-링킹 모델 : 멘션 임베딩을 이용한 nil 멘션 탐지와 개체 연결의 통합 모델 [홍승연 HCLT'20]을 이용. RoBERTa 언어 모델을 제안한 NIL-Aware KnowBERT 로 변경하여 실험결과를 비교

- 실험 결과

	F1
(Roberta) 개체명 연결	85.57%
(NIL-aware Knowbert) 개체명 연결	86.15%

  - NIL-Aware KnowBERT모델이 베이스라인 RoBERTa 기반 개체 연결 모델 대비 F1 기준 0.28% 성능 향상

Q&A

감사합니다.