Path Reranking 기반 전역적 End-to-End 엔터티 링킹

2021.6.24

홍승연[1] , 나승훈[1], 김현호[2], 김선훈[2], 강인호[2]

[1]전북대학교, [2]네이버

목차

- Background
- BASE Model
- Path Reranking
- Experiment Result
- Conclusion

개체명 연결 (Entity Linking)

주어진 문장에 출현한 단어의 중의성을 해결하여

지식 기반(Knowledge base)상의 하나의 특정 개체로 연결하는 작업


거미

위키백과, 우리 모두의 백과사전.
(거미 (동물)에서 넘어옴)

다른 뜻에 대해서는 거미 (동음이의) 문서를 참조하십시오.

거미 (가수)

위키백과, 우리 모두의 백과사전.

거미(한국 한자: 巨美, 본명: 박지연, 본명 한자: 朴志妍, 1981년 4월 8일 ~)는 대한민국의 여성 알앤비 가수이다. '거미'라는 예명은 클 거(巨), 아름다울 미(美)로 '크고 아름다워져라'라는 뜻이 있지만, '거미줄에 걸린 것처럼 헤어 나올 수 없는'이라는 뜻도 있다. 2018년 10월에 배우 조정석과 결혼하였다. 2020년 8월 6일 딸을 출산하였다.

목차 [숨기기]
1 생애
2 연애와 결혼
3 학력
4 음반 목록
4.1 정규 앨범
4.2 미니 앨범
4.3 라이브 앨범
4.4 디지털 싱글
4.5 OST
4.6 일본 음반
5 수상
5.1 시상식
5.2 가요 프로그램 1위
6 가수 외 활동
6.1 예능
7 가족 관계
8 각주
9 외부 링크

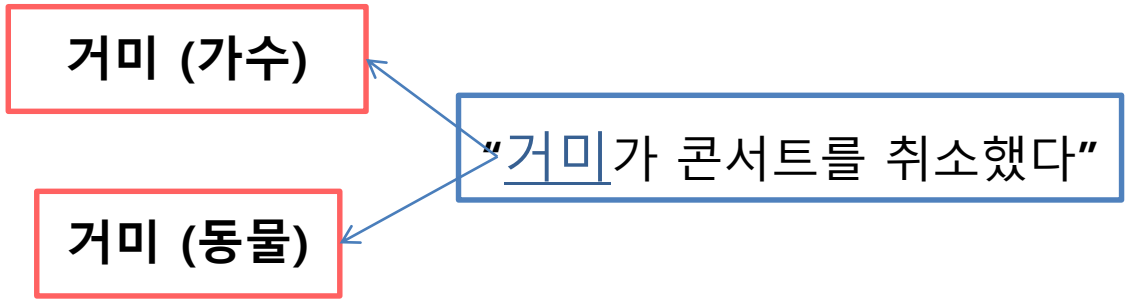
생애 [편집]

전라남도 완도군 금당면 울포리에서 태어났다. 2001년 YG 엔터테인먼트 대표 양현석을 만나게 되고, 2003년 정규 1집 《Like Them》으로 데뷔하였다. 이후 '그대 돌아오면', '친구라도 될 걸 그랬어' 등 히트곡을 내며 한창 인기를 누리는 것으로 보였으나 데뷔 두 달만에 성대에 이상이 생겨 활동을 중단하게 되었다. 약 1년 간의 재활 후 2004년

거미

기본 정보

본명	박지연
출생	1981년 4월 8일 (39세) 전라남도 완도군 금당면
직업	가수
장르	R&B, 발라드
활동 시기	2001년 ~
배우자	조정석
종교	개신교
레이블	카카오엠
소속사	씨제스엔터테인먼트
웹사이트	씨제스 엔터테인먼트 거미 거미 - 페이스북



개체명 연결 (Entity Linking)

1. Local Model(지역적)

- Input pair(Context, Mention)과 Candidate Entities의 유사도 비교

가수 이승철은 존박의 1집 앨범 '이너차일드(Inner Child)'에 대해 ...

Candidate Entities

이승철_(가수)	0.65
이승철_(배우)	0.05
이승철_(기업인)	0.01
...	

2. Global Model(전역적)

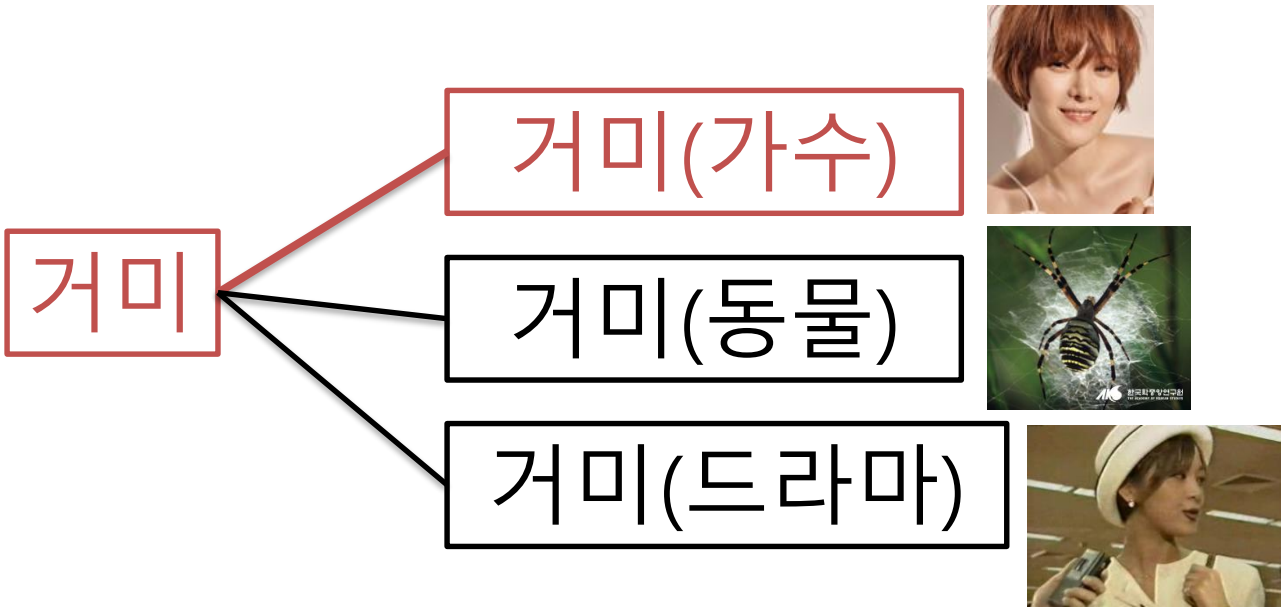
모든 Mention의 Entitys가 통일성(coherence) 있게 연결 되도록 하는 방법
주변 개체의 정보가 반영되어야함

멘션 탐지와 개체 중의성 해결

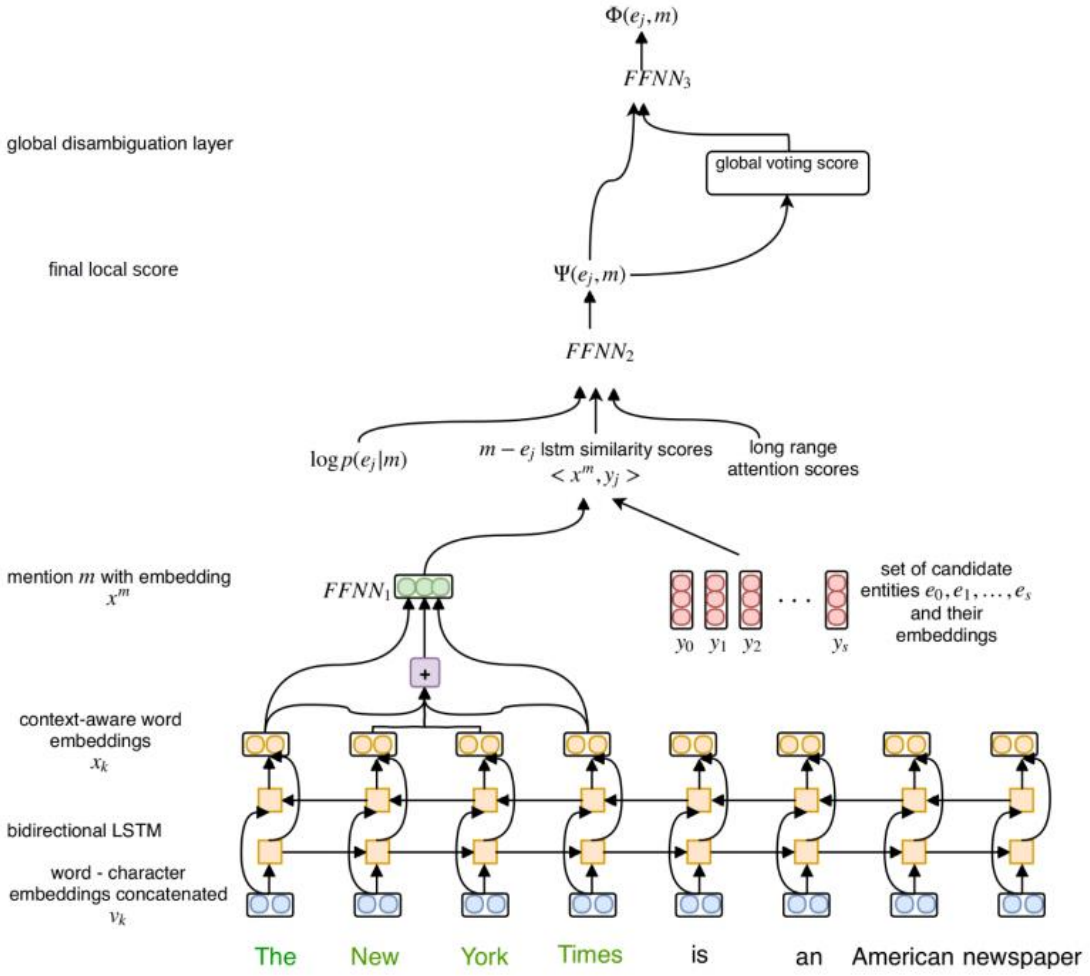
- Mention Detection
  - 문장에서 멘션 탐지

2018년에 거미는 결혼을 했다

- Entity Disambiguation
  - 얻어진 멘션에서 개체 결정

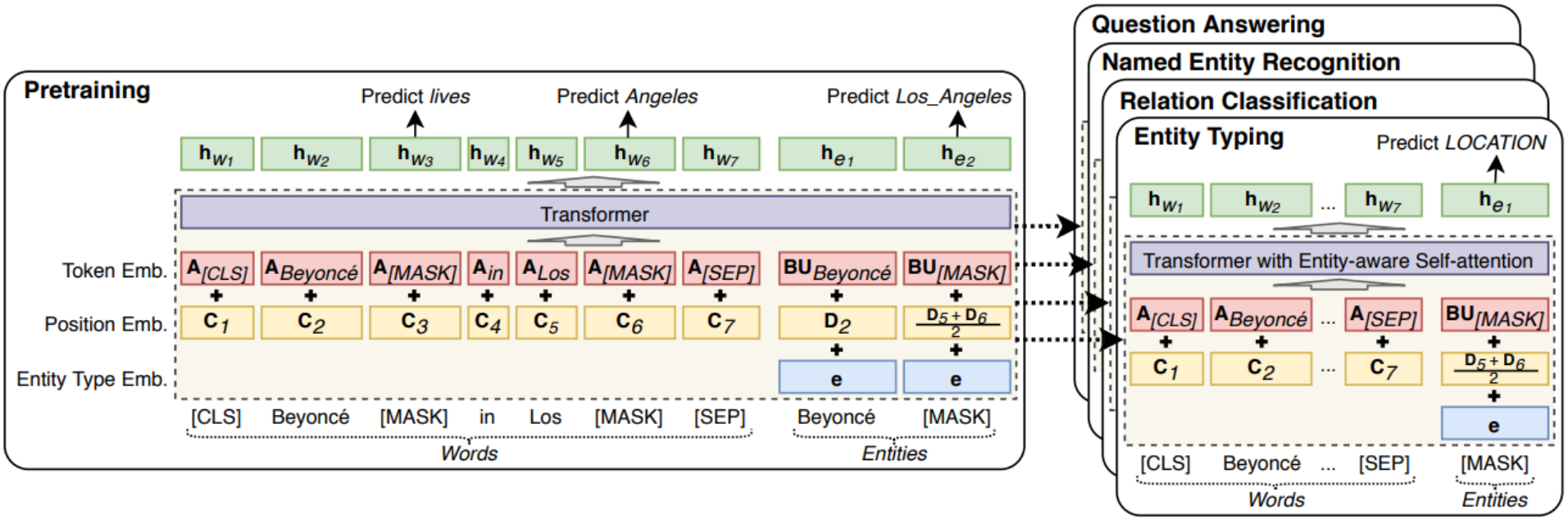


End-to-End Neural Entity Linking[Kolitsas' 18]



- Mention dectection도 동시에 진행하는 entity linking 방식에 모델로 현재까지도 좋은 성능을 보이고 있는 모델

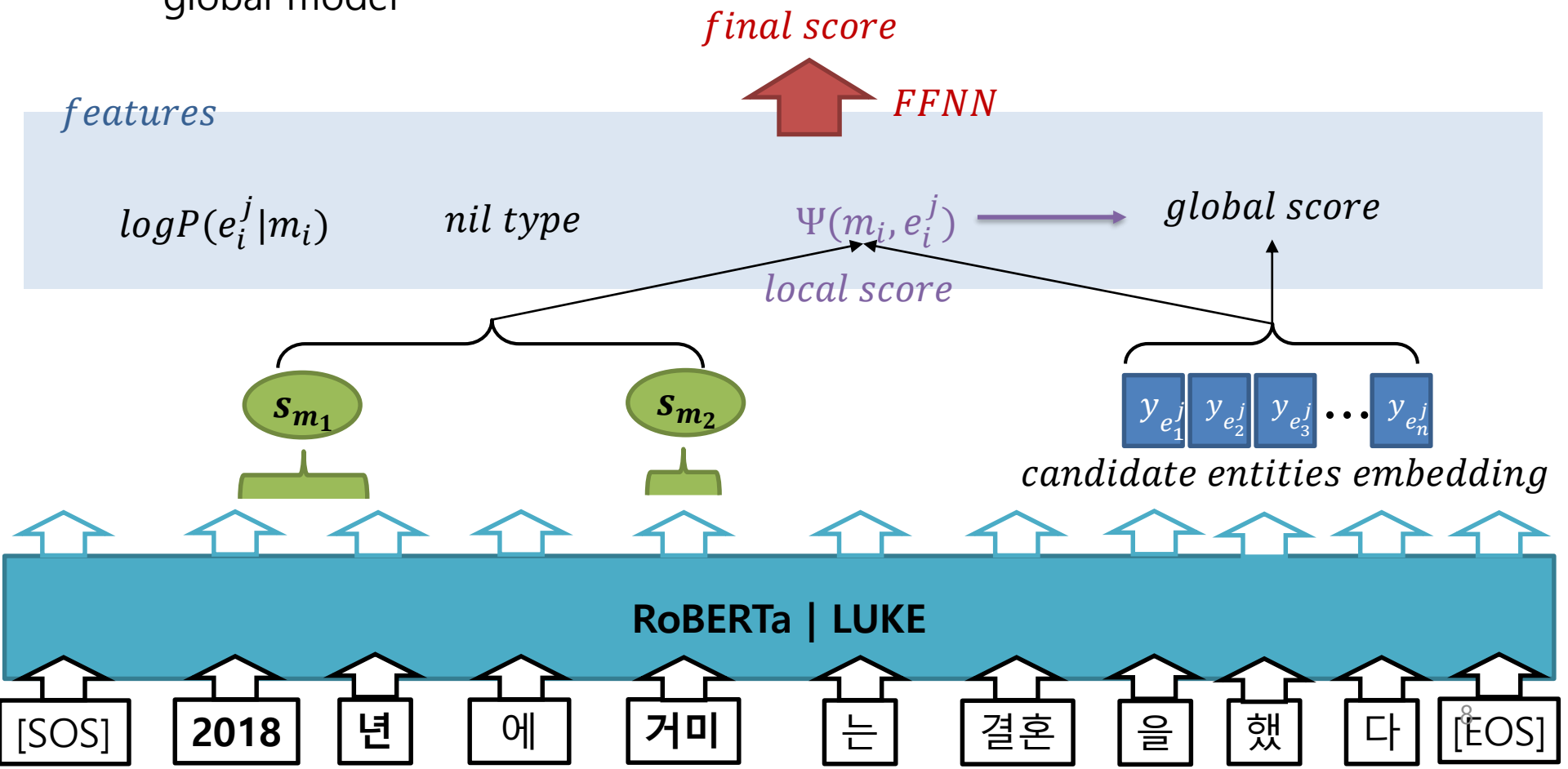
LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention [Yamada' 20]



- 기존의 사전 학습 모델과 달리 추가로 entity도 입력을 받아 처리하여 entity 관련 태스크에서 좋은 성능을 보이고 있는 사전 학습 언어모델

Base Model

- 제안 Base 모델(global model)
  - Local score로부터 멘션의 엔터티를 예측 후 예측된 엔터티를 통해 global score를 얻고 여러 features를 FFNN을 통해 결합하여 최종 점수를 얻는 global model



Base Model

- Global score
  - 각 멘션의 local score를 통해 엔터티를 결정한 후 결정된 엔터티 정보를 통해 global score를 얻음
  - Local score $\Psi(m, e^j)$를 통해 각 mention의 예측된 엔터티를 얻음(V)
  - V^m은 현재 mention m의 예측된 엔터티를 제외한 다른 mention의 예측된 엔터티 집합
  - y_e^{gl}은 mention m의 예측된 엔터티를 제외한 주변 mention의 예측 entity embedding의 합으로 global 정보가 반영된 표상.
  - i번째 멘션에 $y_{e_i}^{gl}$와 후보 엔터티 e_i^j와의 Biaffine 함수를 통해 global 점수를 얻음

$$V = \{(m, e) | m \in M, e = top(\Psi(m, e^j))\}$$

$$V^m = \{e | (m', e) \in V \wedge m' \neq m\}$$

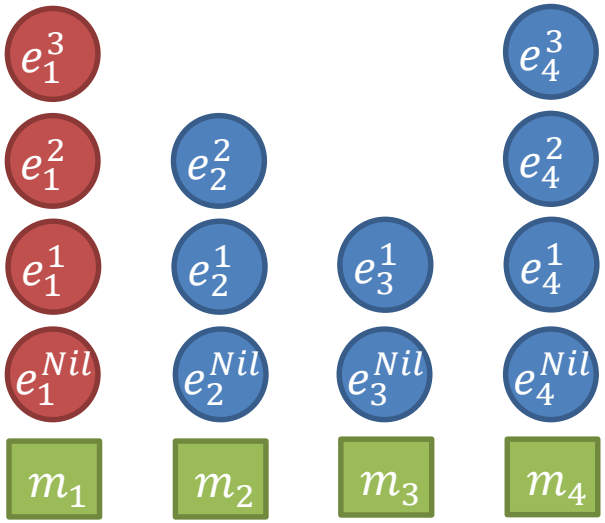
$$y_e^{gl} = \sum_{e \in V^m} y_e$$

e^j : m candidate entities

y_e: $entity\ embedding$

$$\Phi(m, e^j) = Biaffine(y_e^{gl}, e^j)$$

Base Model



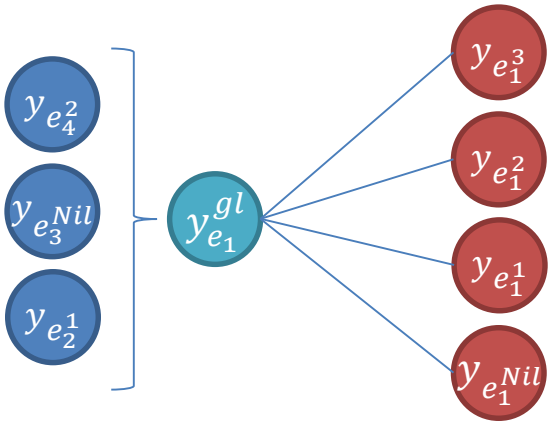
$\Psi(m_i, e_i^j)$

j \ i	1	2	3	4
Nil	0.7	0.1	0.9	0.2
1	0.15	0.8	0.1	0.2
2	0.1	0.1		0.5
3	0.05			0.1

$$V = \{(m_1, e_1^{Nil}), (m_2, e_2^1), (m_3, e_3^{Nil}), (m_4, e_4^2)\}$$

$$V^{m_1} = \{e_2^1, e_3^{Nil}, e_4^2\}$$
$$V^{m_2} = \{e_1^{Nil}, e_3^{Nil}, e_4^2\}$$
$$V^{m_3} = \{e_1^{Nil}, e_2^1, e_4^2\}$$
$$V^{m_4} = \{e_1^{Nil}, e_2^1, e_3^{Nil}\}$$

$$y_{e_1}^{gl} = y_{e_2^1} + y_{e_3^{Nil}} + y_{e_4^2}$$
$$y_{e_2}^{gl} = y_{e_1^{Nil}} + y_{e_3^{Nil}} + y_{e_4^2}$$
$$y_{e_3}^{gl} = y_{e_1^{Nil}} + y_{e_2^1} + y_{e_4^2}$$
$$y_{e_4}^{gl} = y_{e_1^{Nil}} + y_{e_2^1} + y_{e_3^{Nil}}$$



$$global\ score = Biaffine(y_{e_i}^{gl}, e_i^j)$$

Base Model

- Final score
  - Feature들을 FFNN을 통해 Final score 얻음
  - FFNN는 2개의 fully connected layer(100 hidden dim)과 gelu로 구성된 뉴럴 네트워크
  - local score, prior, nil type, global score 정보를 concatenate한 후 FFNN이 적용
  - 최종 loss function은 local 모델과 final 모델을 같이 학습

$$X(m, e^j) = FFNN([\Psi(m, e^j); logP(e^j|m); nil\ type; \Phi(m, e^j)])$$

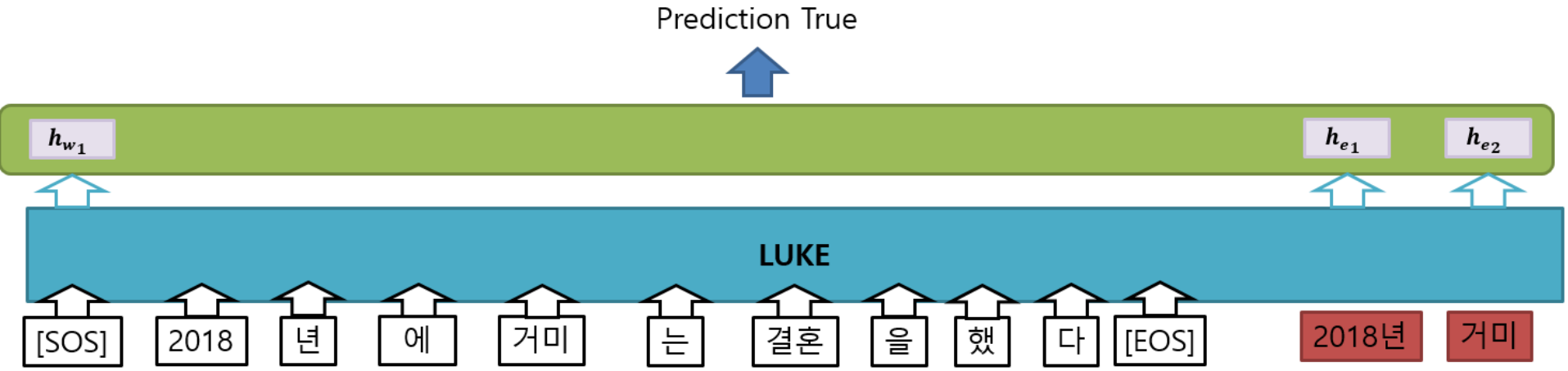
$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{d \in D} \sum_{m \in M} \sum_{e \in Candidate(m)} \Psi(m, e) + X(m, e)$$

Path Reranking

- Motivation
  1. 각 멘션의 score 정보는 가지고 있고 이를 여러 조합으로 재구성하면 성능 향상을 보일 것
     - Beam search를 통해 여러 path를 추출하고 oracle 성능 측정시 기존보다 7-8%로 향상된 성능을 보임을 확인
  2. LUKE를 사용시 Entity 관련 모델에서 개선된 성능을 보임
     - 이를 이용하여 문장과 개체 정보만을 가지고 점수화하면 잘작동할 것이라 판단

Path Reranking

- Scoring model
  - 여러 Path를 얻기 위해 사전 학습한 개체 연결모델의 final score를 가지고 beam search를 사용하여 여러 path를 뽑음
  - Scoring model은 LUKE를 사용하여 입력으로 문장과 개체들을 입력으로 하여 해당 개체 path가 정답인지 판별하는 모델



Path Reranking

- Scoring model
  - 모델은 해당 path를 판별하기 위해 LUKE를 통해 얻은 문장 표상과 개체 표상을 결합하여 최종 표상을 얻고 분류
  - 학습을 위해 golden path를 positive data로 사용하였고 추출된 path중에 golden path가 아닌 path를 negative data로 구성하여 학습을 진행
  - 학습된 scoring model을 통해 여러 path의 점수를 결정하고 reranking을 진행하여 가장 높은 점수를 가지는 path를 통해 엔터티 링킹 결과 도출

실험 세팅

- 데이터셋
  - 위키피디아 말뭉치를 이용하여 데이터 구성

데이터 셋	학습 셋	개발 셋	평가 셋
#Context	30000	10000	20000

… SM 아카데미 대표인 이솔림 씨의 추천으로 참가해 **거미**&휘성의 곡인 <Do It>을 부르고 ….

실험 결과

- Base Model 실험 결과
  - LUKE vs RoBERTa
  - Global vs Local(Base Model에서 global score만 제외)

Base Models	링킹 F1
Local Model-RoBERTa	86.61%
Global Model-RoBERTa	**87.22%**
Local Model-LUKE	87.66%
Global Model-LUKE	**87.81%**

실험 결과

- Path reranking 실험 결과
  - LUKE를 사용한 모델을 통해 Path Reranking을 진행한 성능

Base Models	링킹 F1
Global Model-LUKE	87.81%
Global Model-LUKE(Path Reranking)	**87.93%**

결론 및 향후 연구

- 결론
  - LUKE를 적용하여 기존보다 개선된 성능을 얻음
  - Path Reranking을 통한 엔터티 링킹 모델을 제안하였고 실험 결과 기존보다 개선된 성능을 보임
  - 여러 path 추출할때 전이 확률을 고려하지 못하고 독립적으로 구함
  - Path scoring 모델이 간단한 판별 모델로 구성하여 좀 더 정교한 모델 필요

- 향후 연구
  - LUKE 외의 최근에 많이 연구되고 있는 지식이 부여된 여러 사전 학습 모델을 통해 실험을 진행하고자함

감사합니다.
Q&A