추출 기반 한국어 개체 중의성 해결

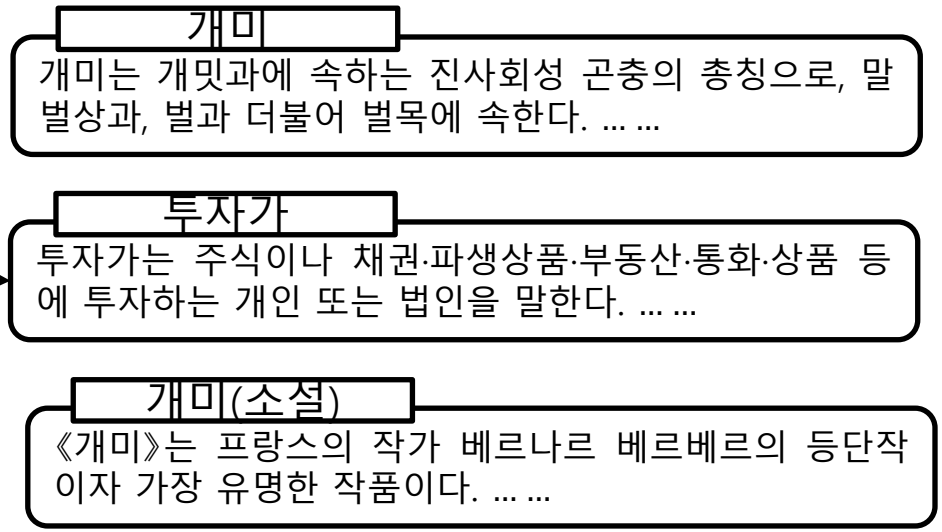
민진우, 나승훈
전북대학교

목차

- 개체 중의성 해결 & 개체 연결 소개
- 관련연구
- 한국어 추출 기반 한국어 개체 중의성 해결 모델
- 실험 세팅
- 실험 결과
- 결론 및 향후 연구

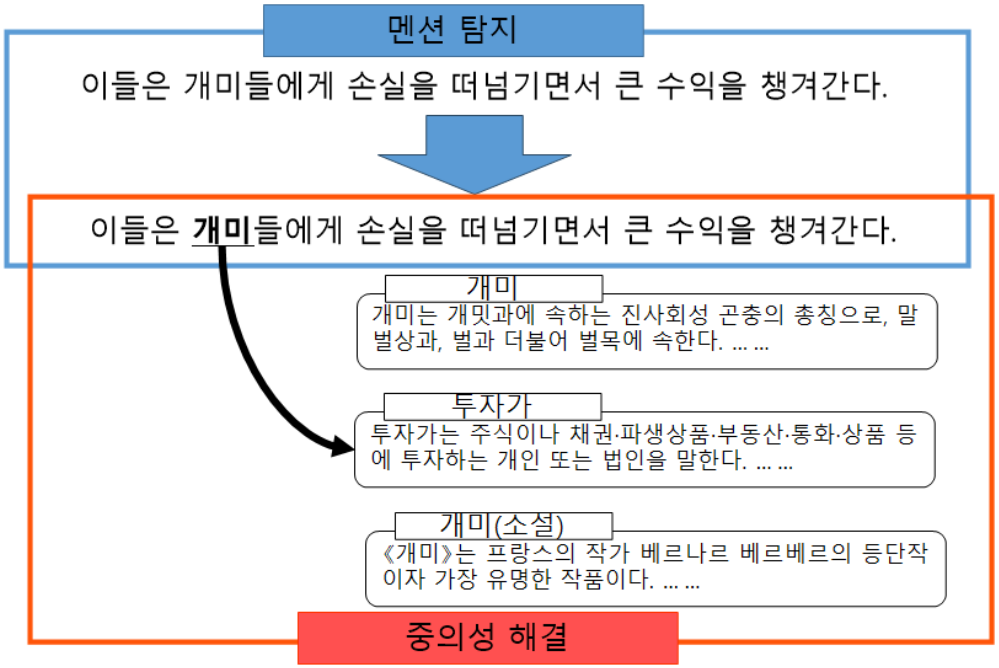
개체 중의성 해결

이들은 **개미**들에게 손실을 떠넘기면서 큰 수익을 챙겨간다.



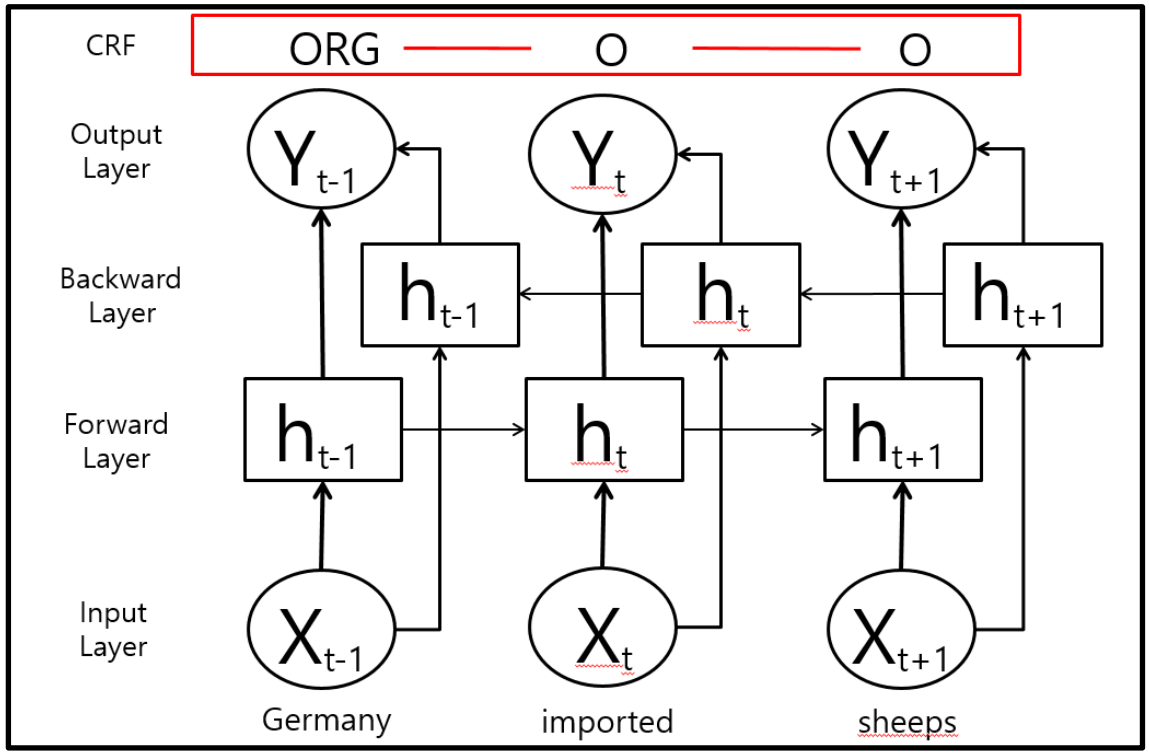
- 개체 연결의 subtask
  - 주어진 개체 멘션이 가질 수 있는 후보 엔티티 중 단 하나의 엔티티로 연결하는 과정
  - 딥러닝 기반의 중의성 해결 연구는 텍스트 상의 인코딩된 멘션 표상과 후보 엔티티 표상과의 내적으로 scoring하는 방식이 주를 이룸

개체 연결



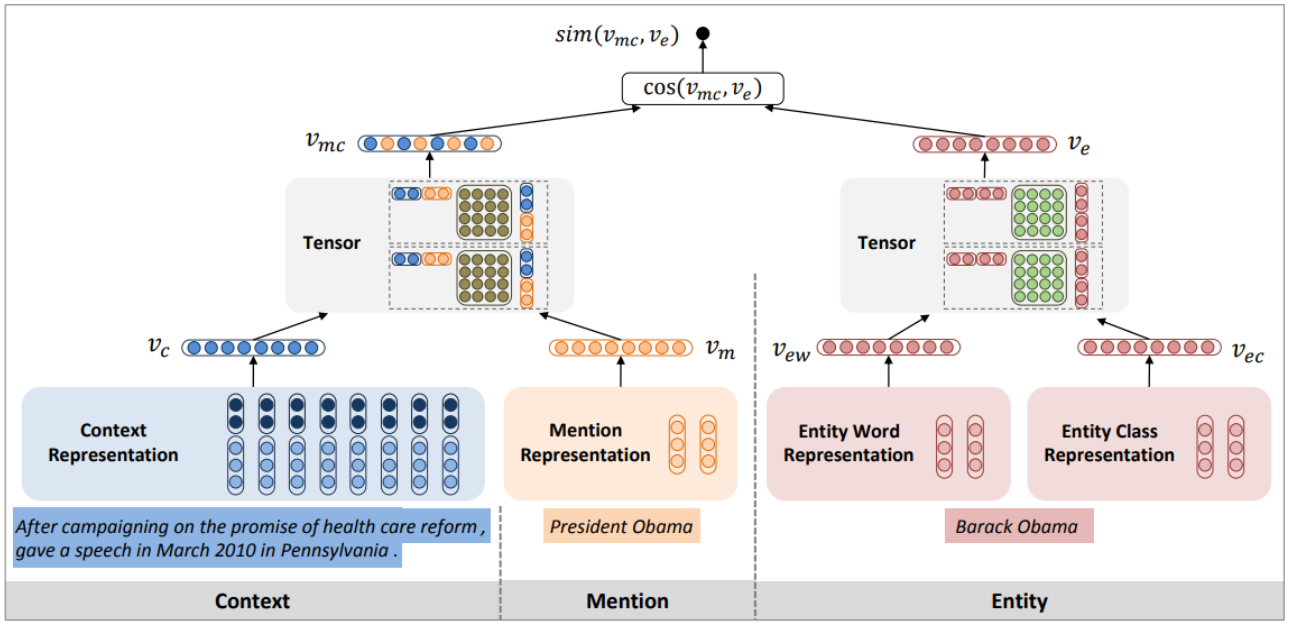
- 다음의 두가지 하위 태스크로 나뉨
  - 멘션 탐지 : 문서에서 등장하는 지명, 인명, 기관명 등을 나타내는 개체 표현인 개체 멘션을 찾는 과정
  - 중의성 해결 : 개체 멘션이 가질 수 있는 후보 엔티티 중 단 하나의 엔티티로 연결하는 과정
- End-to-End 개체 연결
  - 멘션 탐지와 중의성 해결을 일괄적으로 해결하는 모델

Bidirectional LSTM-CRF Models for Sequence Tagging (Zhiheng Huang et al, '16)



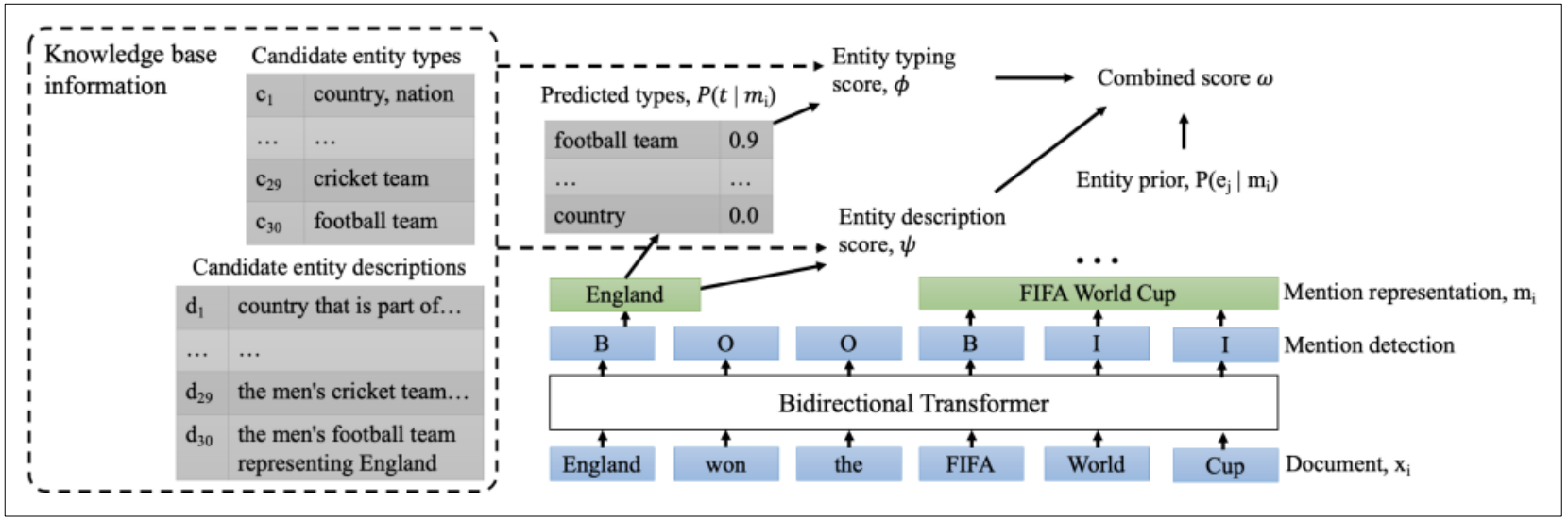
- Bi-LSTM-CRF 기반의 개체명 인식
  - 개체명 인식은 멘션 추출 과정의 또 다른 표현
  - 주어진 sequence에 대해서 label을 부여하는 sequence labeling(순차 태깅) 방식

Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation (Sun, IJCAI '15)



- 초기 딥러닝 기반 중의성 해결 모델
  - 신경망을 통해 mention, context, 후보 엔티티에 대한 vector를 구성하고 결합된 mention-context vector와 후보 entity vector 사이의 유사도를 구하고 유사도가 가장 높은 entity를 선택

ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking (Tom Ayoola et al, '21)



- 멘션 추출 및 중의성 해결 통합 모델
  - 문서 내의 모든 멘션에 대한 멘션 탐지(추출)과 엔티티 타이핑과 중의성 해결을 동시에 해결하는 end-to-end 방식의 개체 연결 모델
  - 엔티티 타입 정보와 description 정보를 이용하여 중의성 해결

ENTQA: ENTITY LINKING AS QUESTION ANSWERING

(Tom Ayoola et al, ‘21)

Passage

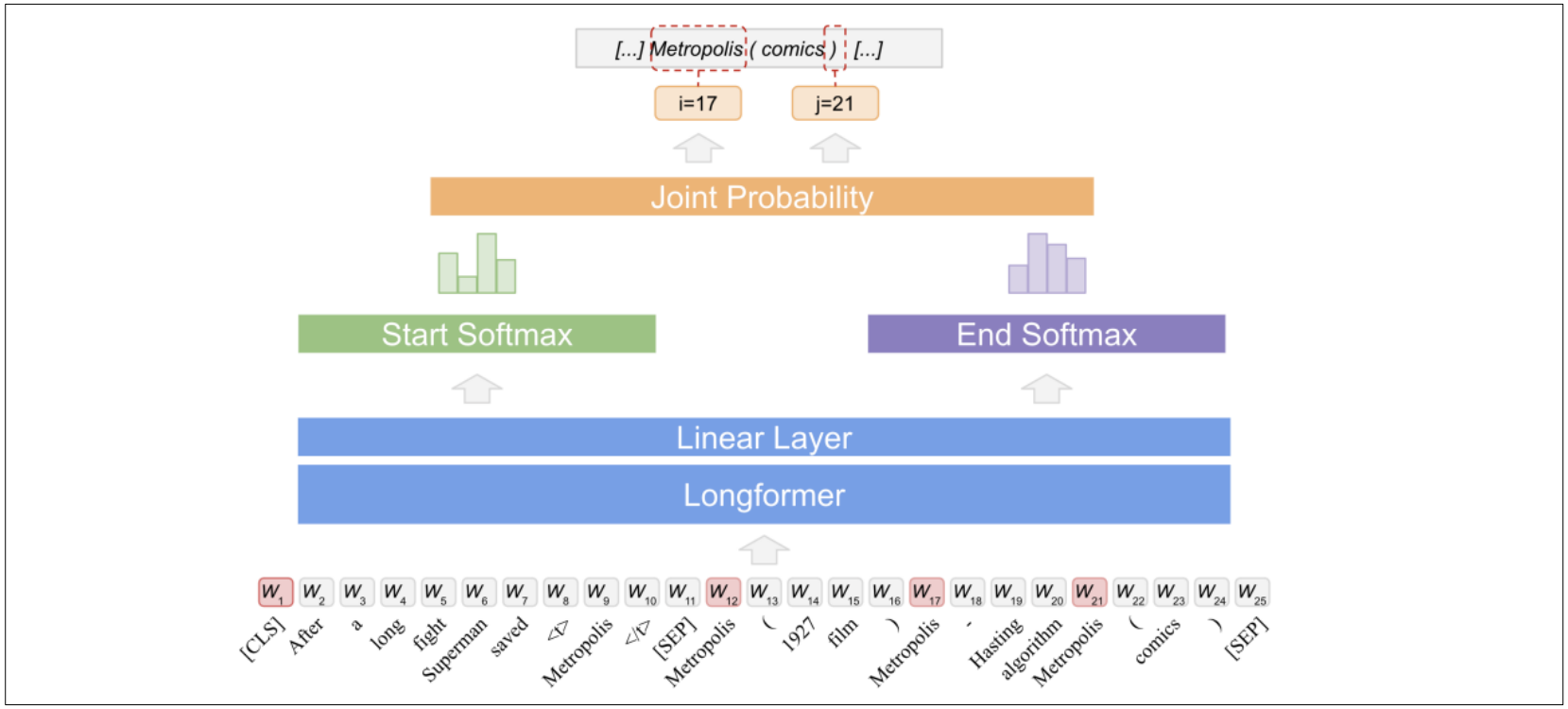
After bowling [Somerset]$_3$ out for 83 on the opening morning at [Grace Road]$_2$, [Leicestershire]$_1$ extended their first innings by 94 runs before being bowled out for 296 with [England]$_{11}$

Top-K candidate entities

1. Leicestershire County Cricket Club
2. Grace Road
3. Somerset County Cricket Club
✗ 4. Durham County Cricket Club
✗ 5. Nottinghamshire County Cricket Club
✗ 6. Derbyshire County Cricket Club
✗ 7. Warwickshire County Cricket Club
✗ 8. Leicestershire
✗ 9. Worcestershire County Cricket Club
✗ 10. Yorkshire County Cricket Club
11. England cricket team
✗ 12. Marylebone Cricket Club
✗ 13. Sussex County Cricket Club
✗ 14. Kent County Cricket Club
✗ 15. Leicester
✗ 16. Aylestone Road
✗ 17. County Cricket Ground, Derby
⋮

- 기존 방법의 단점
  - 기존 Entity Linking 모델은 [Mention Detection] => [Entity Disambiguation]. 이러한 방식은 멘션에 대한 후보 엔티티 정보 없이 멘션을 추출
- 과정의 전환(QA 방식으로 해결)
  - 먼저, Retriever에서 문서가 가질 수 있는 Entity 집합을 추출 후 Reader에서 해당 Entity에 대한 멘션의 Start, end position 포인팅하는 QA 방식

ExtEnD: Extractive Entity Disambiguation (Edoardo Barba, ACL '22)



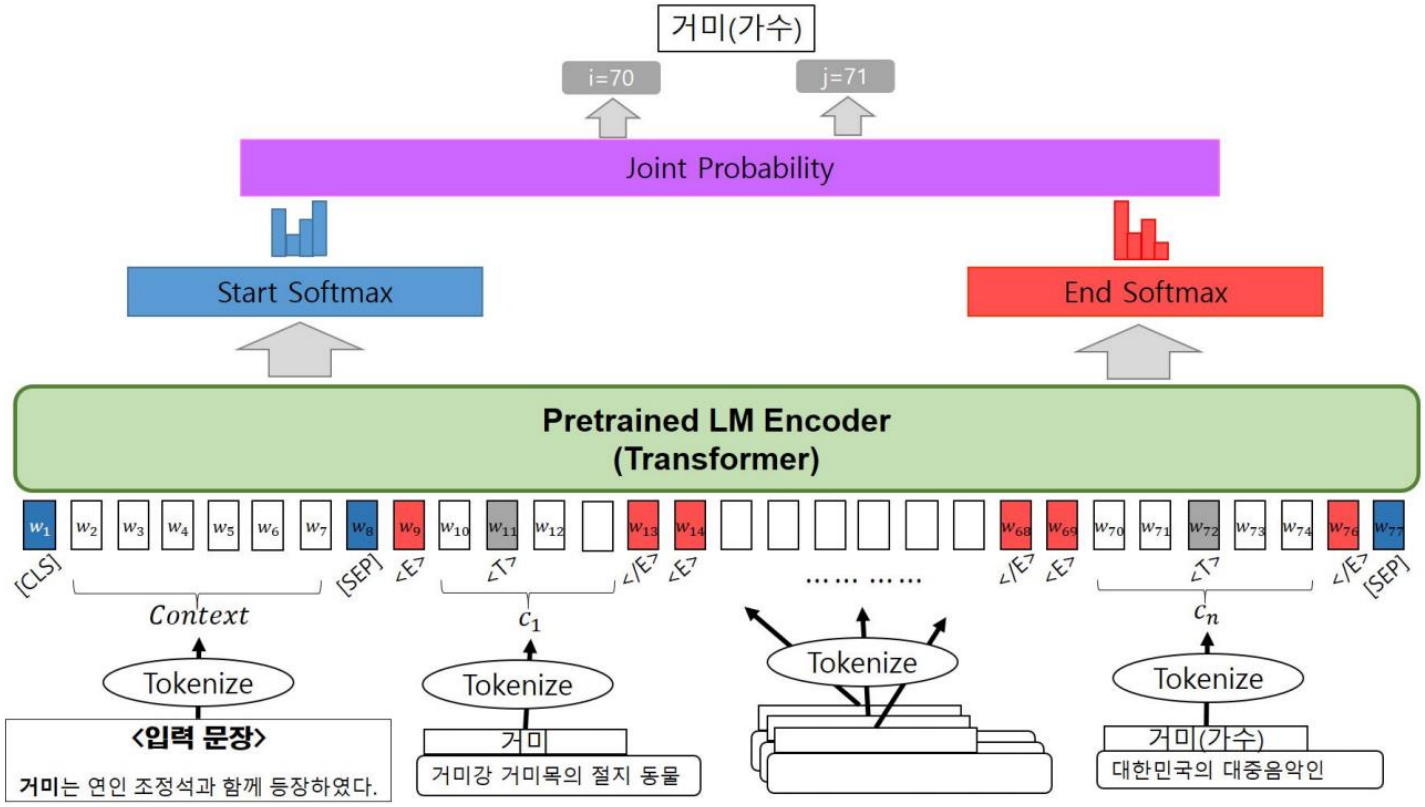
- Extractive 기반 모델
  - 입력 시퀀스와 후보 엔티티 타이틀 정보만을 활용하는 cross-encoder 방식으로 인코딩 한 후 정답 엔티티의 시작, 끝 포지션의 확률을 최대화 하는 방식으로 학습

ExtEnD: Extractive Entity Disambiguation (Edoardo Barba, ACL '22)

	Model	In-domain	Out-of-domain				Avgs	
		AIDA	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	Avg	Avg_{OOD}
Wikipedia + AIDA	Ganea and Hofmann (2017)	92.2	93.7	88.5	88.5	77.9	77.5	86.4	85.2
	Guo and Barbosa (2018)	89.0	92.0	87.0	88.0	77.0	84.5	86.2	85.7
	Yang et al. (2018)	**95.9**	92.6	89.9	88.5	**81.8**	79.2	88.0	86.4
	Shahbazi et al. (2019)	93.5	92.3	90.1	88.7	78.4	79.8	87.1	85.9
	Yang et al. (2019)	93.7	93.8	88.2	90.1	75.6	78.8	86.7	85.3
	Le and Titov (2019)	89.6	92.2	90.7	88.1	78.2	81.7	86.8	86.2
	Fang et al. (2019)	94.3	92.8	87.5	91.2	78.5	82.8	87.9	86.6
	De Cao et al. (2021b)	93.3	94.3	89.9	90.1	77.3	87.4	88.8	87.8
	ExtEnD$_{Large}$ + BLINK	92.6	**94.7**	**91.6**	**91.8**	77.7	**88.8**	**89.5**	**88.9**
AIDA	De Cao et al. (2021b)	88.6	88.1	77.1	82.3	71.9	71.7	79.5	78.2
	Tedeschi et al. (2021)	**92.5**	89.2	69.5	**91.3**	68.5	64.0	79.2	76.5
	ExtEnD$_{Base}$	87.9	92.6	84.5	89.8	74.8	74.9	84.1	83.3
	ExtEnD$_{Large}$	90.0	**94.5**	**87.9**	88.9	**76.6**	**76.7**	**85.8**	**84.9**

- 실험 결과
  - 타이틀 정보만을 활용하여 considerable한 결과 도출
  - 저자는 엔티티 type 정보나 definition 정보를 활용하면 성능이 향상될 것을 기술함

추출 기반 한국어 개체 중의성 해결



- ExtEnD의 확장 & 한국어 적용
  - 타이틀 정보 뿐 아니라 개체의 description 정보까지 활용
  - 이를 위해 한국어 wikipedia title 및 해당 title의 wikidata의 description 정보 추출
  - 지정된 한국어 개체 중의성 해결 데이터 셋에 적용 개선된 결과 도출

- 컨텍스트 및 후보 엔티티 시퀀스 구성
  - 모델의 입력 : 입력 문장(문서)와 모든 후보 엔티티 정보를 하나의 시퀀스로 결합
  - 컨텍스트 시퀀스
    - 문서를 토큰화 한 후 멘션이 나타난 위치를 기준으로 설정한 최대 길이 이내의 멘션의 앞,뒤 시퀀스를 추출
    - 멘션의 경계는 <M >, </M>특수 토큰을 추가하여 식별
    - 컨텍스트의 앞에는 [CLS], 뒤에는 [SEP] 특수토큰을 추가

- 컨텍스트 및 후보 엔티티 시퀀스 구성
  - 후보 엔티티 시퀀스
    - 후보 엔티티 시퀀스의 템플릿과 그 예시

Template

<E> Wikipedia Title <T> Wikidata Definition </E>

Example

<E> 거미 (가수) <T> 대한민국의 대중음악인 </E>

    - 타이틀과 scription을 분리하기 위한 특수 토큰으로 <T> 사용. 해당 후보 타이틀에 대한 위키데이터 description이 존재하지 않으면 단순히 <Empty >특수 토큰으로 대체
    - 각 후보 엔티티를 구분하기 위해 후보 엔티티 시퀀스의 앞, 뒤에 각각 <E>, </E> 특수 토큰이 삽입

- 컨텍스트 및 후보 엔티티 시퀀스 구성
  - 시퀀스에 대한 길이 설정
    - 최대 시퀀스 토큰 길이는 1024로 설정
    - 문서 컨텍스트의 최대 길이는 128, 각 후보 엔티티의 타이틀과 description의 최대 길이는 각각 10, 15로 설정
    - 각 멘션의 최대 후보 엔티티 개수는 30개로 제한

- 학습 및 디코딩
  - 모델의 출력
    - 통합 시퀀스를 토큰화 한 후 언어 모델을 통해 입력 시퀀스 인코딩
    - 인코딩된 표상을 얻은 후 두 개의 독립적인 classification 헤드를 통해 각 단어가 정답 개체 타이틀의 시작, 끝인지에 대한 확률 분포를 얻음
  - 학습
    - 교차 엔트로피 함수를 통해 각각 정답 개체 타이틀의 시작, 끝 위치의 확률을 최대화하도록 학습
  - 디코딩
    - 두 개의 헤드를 통해 얻어진 확률의 결합 확률이 가장 높은 후보 엔티티를 선택
    - 즉, 모든 후보 엔티티 타이틀의 시작, 끝 위치의 토큰의 확률을 결합을 구하고 결합 확률이 가장 높은 후보 엔티티를 선택함으로써 중의성을 해결

실험 세팅

- 데이터 셋 (민진우, ksc ‘22와 동일)
  - 모두의 말뭉치 개체 연결 데이터 셋으로부터 일부 문서를 추출하여 구축된 실험 집합 사용
  - 구성
    - 총 1678문서. 학습셋 1378 문서, 개발셋 100 문서, 평가셋 200문서로 나뉨
  - 멘션-후보 엔티티 사전
    - 모두의 말뭉치 개체 연결 데이터 셋의 전체 문서 내의 모든 멘션에 대해 연결된 위키피디아 타이틀 빈도수를 구하고 내림차순으로 정렬

- 언어 모델
  - Extend 논문에서는 긴 시퀀스를 인코딩하기 위한 언어모델인 Longformer를 사용. 공개되어 있는 한국어 Longformer 모델이 존재하지 않아 1026 토큰 길이까지 지원하는 한국어 BART 모델 사용(skt 개발)

https://huggingface.co/gogamza/kobart-base-v2

실험 결과

- 베이스라인
  - 기존 방법
    - [민진우, ksc '22]의 Encoder, Cross-Encoder 기반의 중의성 해결 모델을 제시. 두 모델 모두 위키피디아 타이틀과 해당 텍스트를 함께 사용.
  - 2가지 세팅
    - Extractive(title) : title 정보만을 사용.
    - Extractive(title+description) : title, description을 모두 활용한 세팅

- 실험 결과

	정확도(acc)
Bi-Encoder	87.60%
Cross-Encoder	92.90%
Extractive (title)	89.12%
Extractive (title + description)	**93.77%**

결론 및 향후 연구

- 결론
  - 결과 분석
    - Extractive (title): 타이틀만을 사용하여 Bi-Encoder 모델 보다 1.52% 높은 성능을 보임. 이는 타이틀 정보만으로도 모델이 높은 확률로 정답 엔티티를 유추 가능함.
    - Extractive (title + description): Bi-Encoder, Cross-Encoder 모델 대비 높은 성능 달성. 컨텍스트와 후보 엔티티 정보 사이의 정보 교환 뿐 아니라 각 후보 엔티티 간의 문맥을 반영하는 것이 도움이 될 수 있음.
  - Extend 모델에서 지식 베이스 정보인 위키데이터 상의 description을 활용하도록 확장 후 한국어 데이터 셋에서 베이스라인 모델 대비 성능 향상 달성

- 향후 연구
  - 위키 데이터의 지식 트리플(fact) 활용
    - prompt 방법 혹은 Graph Neural Networks(GNNs) 등으로 인코딩
  - Autoregressive 방법론 적용(GPT-4등 초거대 언어모델)
    - 출력 층에서 해당 멘션에 대한 위키피디아 상의 타이틀 생성

Q&A

감사합니다.