Coarse-grained 엔터티 타이핑에 기반한 멘션 탐지

2020.12.20

전북대학교 홍승연

목차

- Background
- Related Works
- Model
- 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

거미 (가수)

거미 (동물)

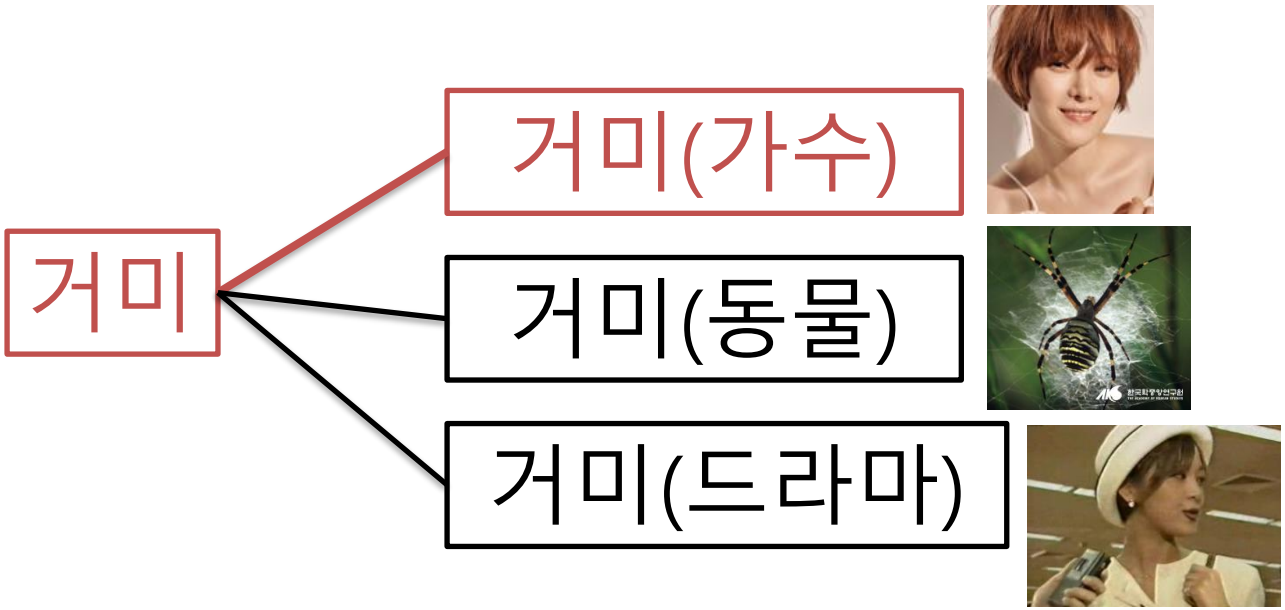
"거미가 콘서트를 취소했다"

멘션 탐지와 개체 중의성 해결

- Mention Detection
  - 문장에서 멘션 탐지

2018년에 거미는 결혼을 했다

- Entity Disambiguation
  - 얻어진 멘션에서 개체 결정



엔터티 타이핑

- Entity Typing
  - 문서의 나타난 개체의 Type(eq. 지역, 기관)을 결정

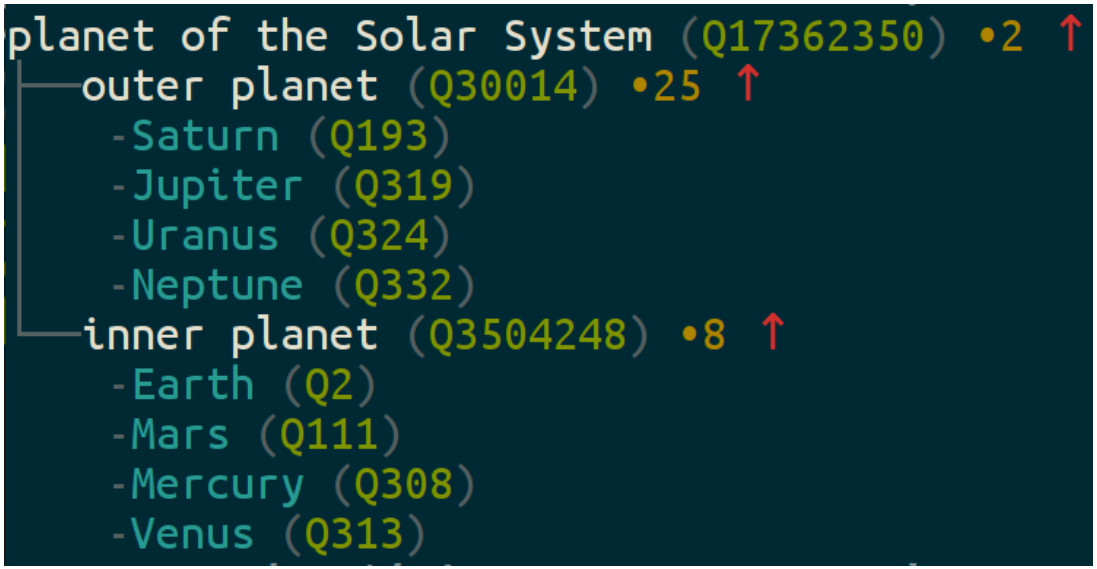
Entity	Type
거미 (가수)	사람
서울	수도 대한민구의 특별시 대도시
…	…

- Types 갯수

Ultra Fine-grained types > Fine-grained types > Coarse-grained types

Coarse-grained 엔터티 타이핑

- Wikidata taxonomy
  - 상위로 갈수록 더 넓은 의미를 포함하는 계층 구조를 가진 wikidata 집합



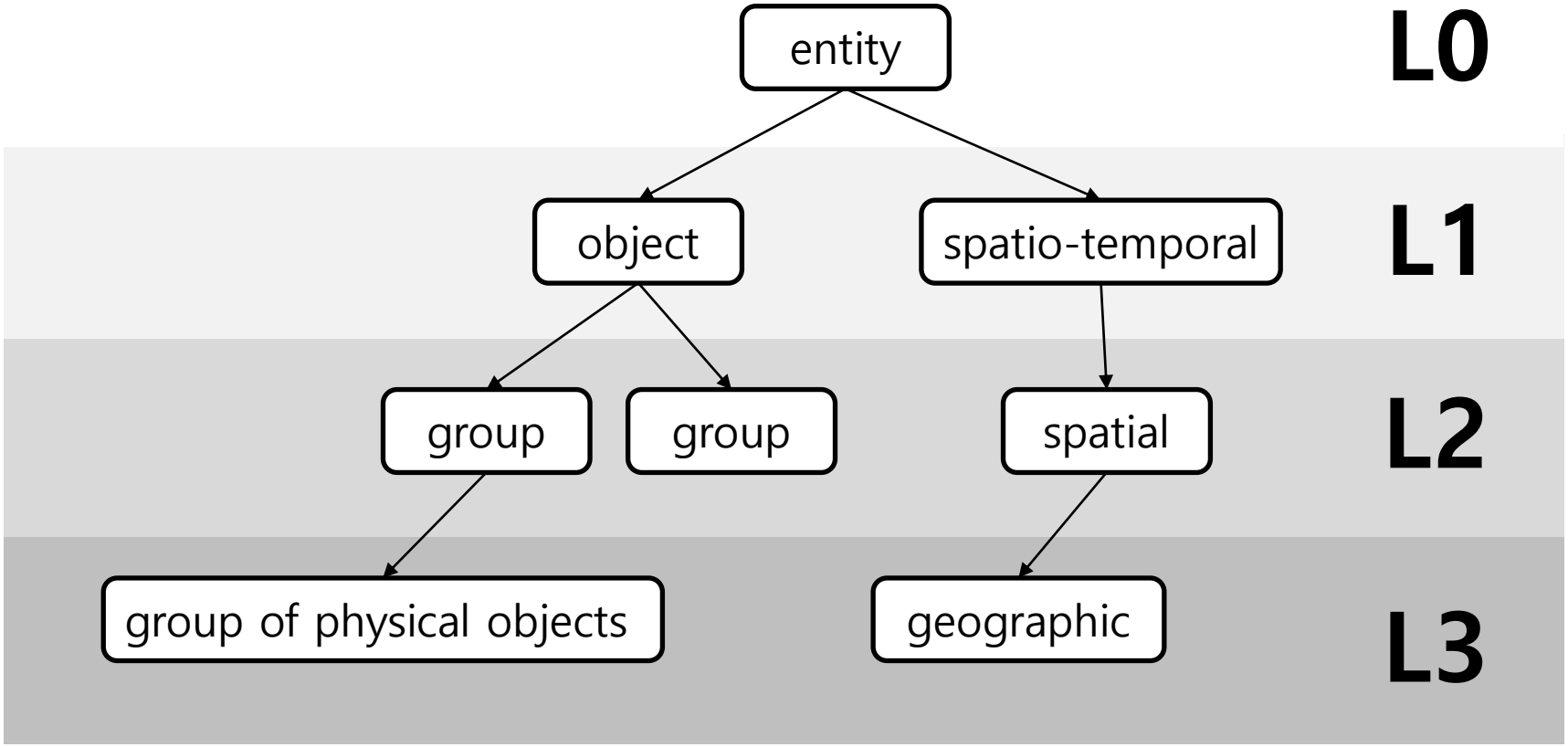
Coarse-grained 엔터티 타이핑

- Wikidata taxonomy
  - Wikidata의 taxonomy 정보를 이용하여 개체 타입 결정
  - Leaf node들의 frequency로 부터 entropy 계산하여 상위 노드 (50, 100)개의 타입 결정
  - parent node(root)로부터 각각 child node를 확장하여 얻은 leaf node들의 entropy 계산
  - frequency는 학습 데이터에서 entity의 types의 빈도수를 측정하여 얻음

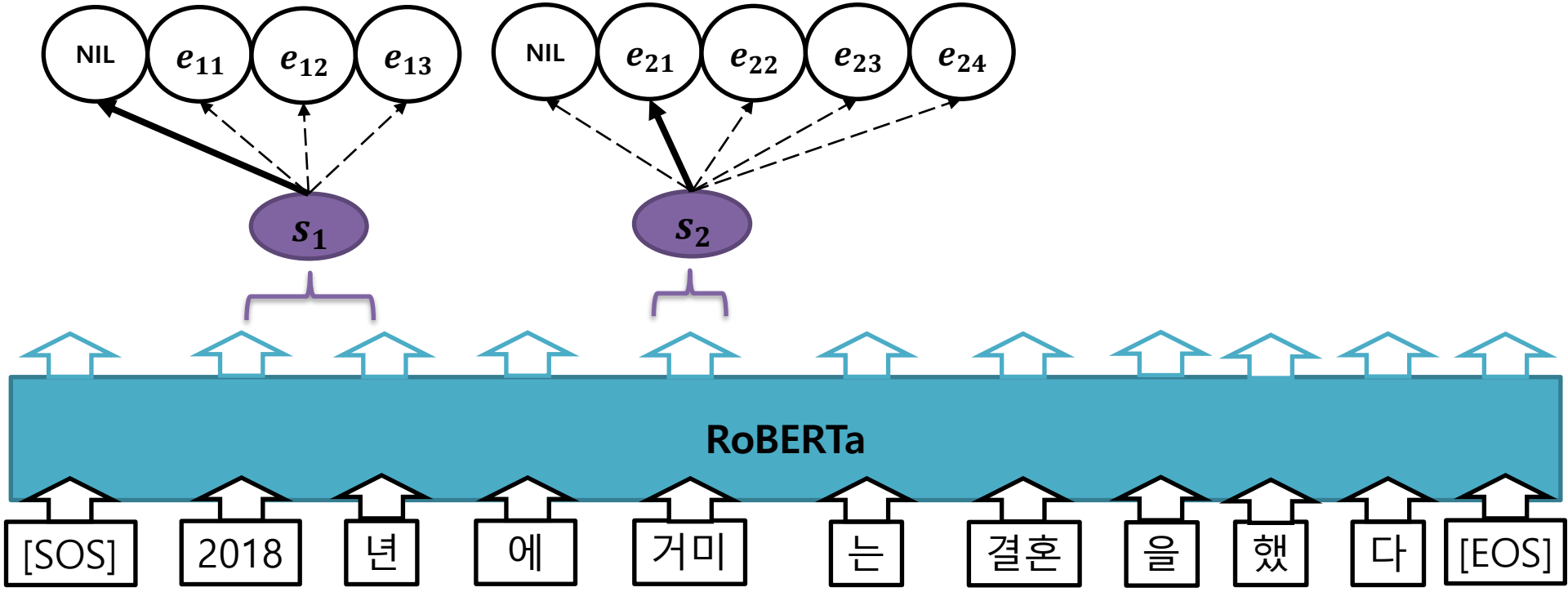
$$G = (V, E), \qquad V_0 = \{root\}$$
$$v' = argmax_V \Big(Entropy\big(Expand(G, v)\big)\Big), G = Expand(G, v')$$

Coarse-grained 엔터티 타이핑

- Wikidata taxonomy
  - 추출한 entropy Top 50의 taxonomy의 일부



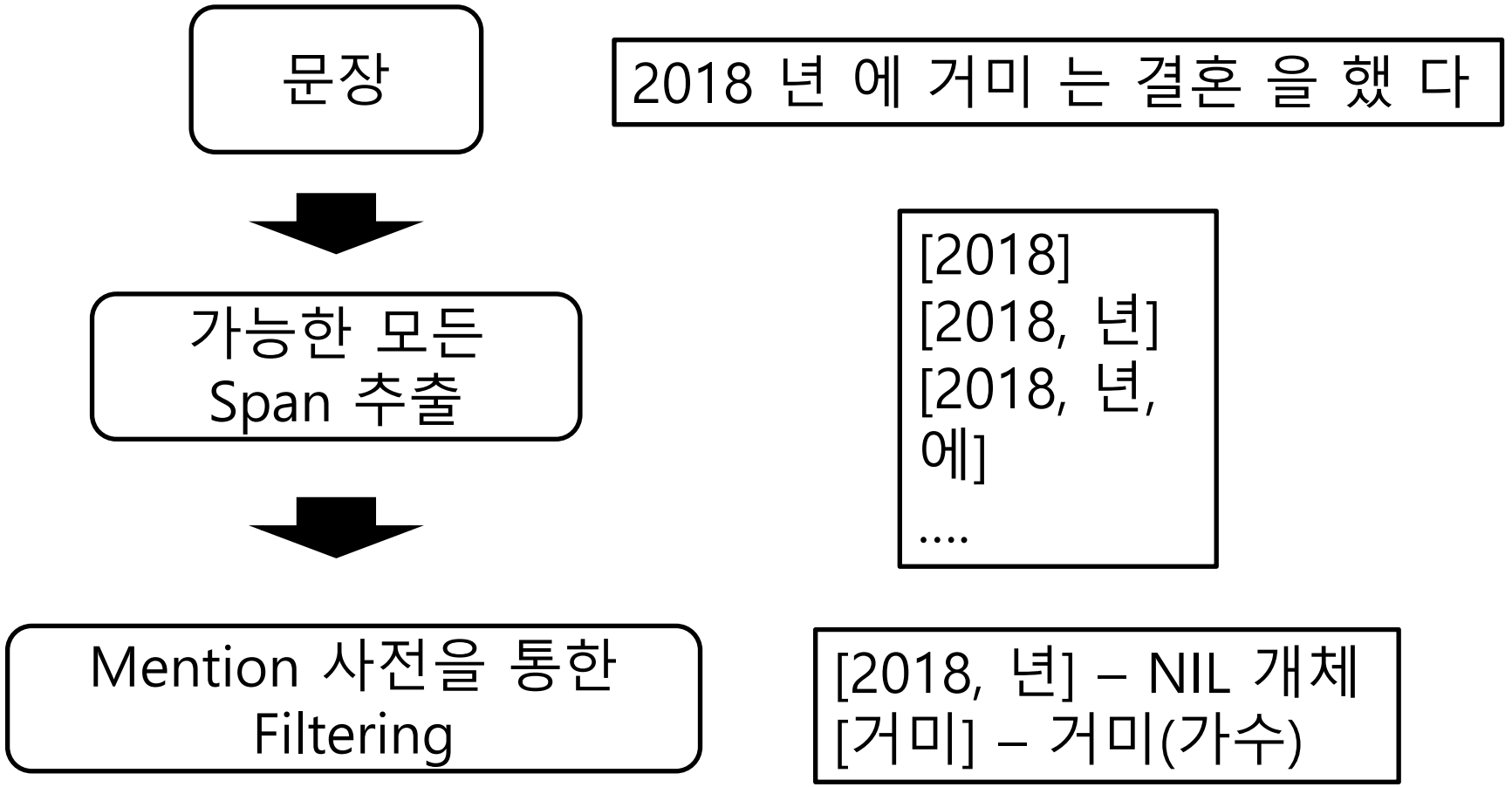
멘션 임베딩을 이용한 NIL 멘션 탐지와 개체 연결의 통합 모델[홍승연'20]



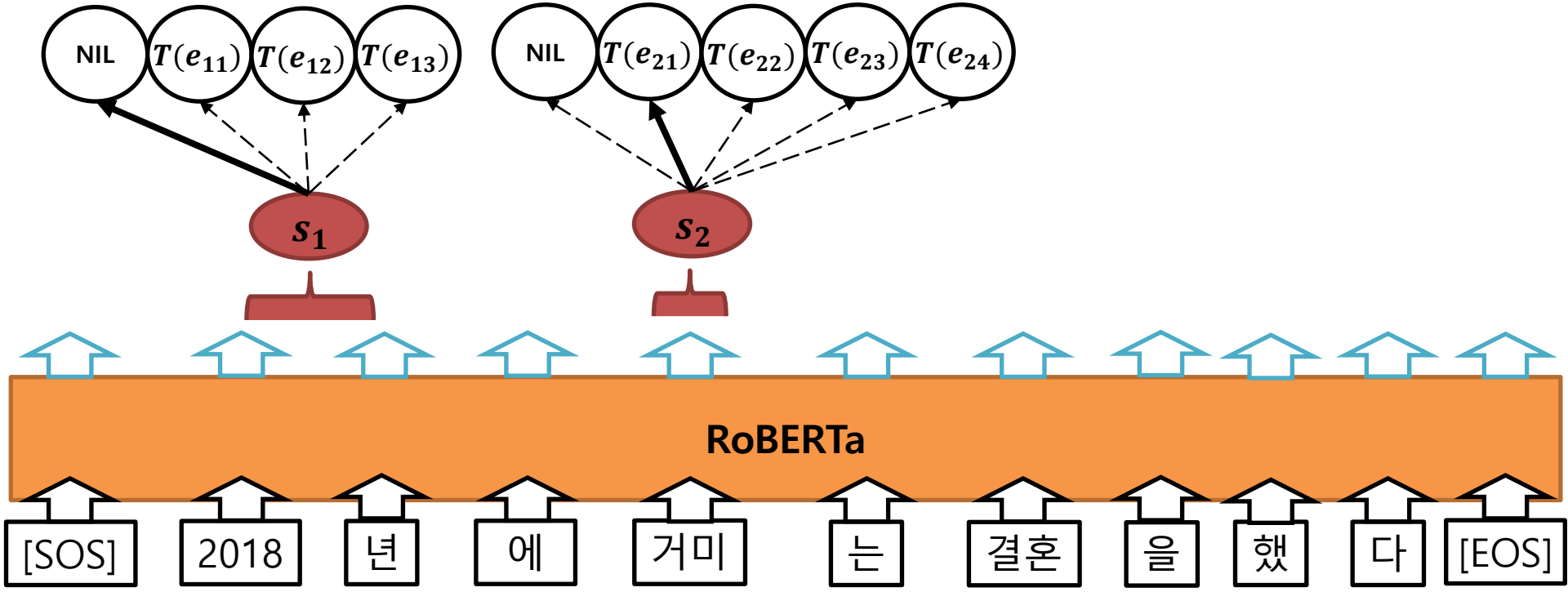
- 멘션 사전을 이용하여 가능한 모든 멘션을 탐지하고 중의성을 해결하는 모델
- 멘션 span 표상과 후보 개체 간의 유사도를 계산하여 개체 결정
- Base model

멘션 임베딩을 이용한 NIL 멘션 탐지와 개체 연결의 통합 모델[홍승연'20]

- Mention 추출



Coarse-grained 엔터티 타이핑에 기반한 멘션 탐지



- NIL 멘션 탐지와 개체 연결의 통합 모델에서 엔터티 정보를 사용하지 않고 엔터티 타이핑 정보만을 사용하여 멘션 탐지 진행하는 모델

Coarse-grained 엔터티 타이핑에 기반한 멘션 탐지

- Entity Typing
  - 엔터티 타이핑은 여러 개의 Type이 존재하는 경우 빈도수에 근거하여 대표 Type을 정하고 Type이 없는 경우 UNK Type으로 설정
  - 엔터티 타이핑 정보는 wikidata에서 추출
  - Coarse-grained 엔터티 타이핑은 wikidata-taxonomy를 통해 얻은 대표 타입 사용

Entity	Type	Type Entity
거미 (가수)	사람	사람
지포스 20	X	UNK
서울	수도 대한민구의 특별시 대도시	수도
…	…	…

Coarse-grained 엔터티 타이핑에 기반한 멘션 탐지

- Mention Detection
  - 탐지된 멘션의 시작 위치 표상과 끝 위치 표상을 결합하여 Span 표상을 얻음
  - 얻어진 Span 표상 s_i과 후보 개체 타입 표상 $\mathrm{T}(e_{ij})$와 Biaffine 연산을 통해 점수를 얻고 멘션 여부 결정
  - 개체 타입 표상 $\mathrm{T}(e_{ij})$은 해당 개체 설명 문서에 RoBERTa를 적용하여 얻음(j는 j번째 후보 개체)

$$C_t = [c_1, \ldots, c_n]$$

$$h_t = RoBERTa(C_t)$$

$$s_i = \left[h_{start(i)}; h_{end(i)}\right]$$

$$score_{ij} = Biaffine(s_i, \mathrm{T}(e_{ij}))$$

Coarse-grained 엔터티 타이핑에 기반한 멘션 탐지

- 데이터 구축
  - 위키피디아 문장 6만개를 사용하여 학습셋 3만, 개발셋 1만, 평가셋 2만 문장으로 구성하여 학습
  - 위키피디아 문서의 링크 정보를 이용하여 멘션-개체 사전 구축

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

Coarse-grained 엔터티 타이핑에 기반한 멘션 탐지

- 실험 결과
  - Base Model은 멘션 임베딩을 이용한 NIL 멘션 탐지와 개체 연결의 통합 모델[홍승연'20] 결과
  - 성능은 기존 모델에 비해 감소 되었지만 총 개체 임베딩 수가 감소하여 메모리 부분에서 이점

모델	F1	개체 임베딩 수
Base Model	89.05%	54489
Typing Model	87.20%	3482
Coarse-grained Typing Model +Top 50 Taxonomy	87.36%	50
Coarse-grained Typing Model +Top 100 Taxonomy	87.27%	100

Coarse-grained 엔터티 타이핑에 기반한 멘션 탐지

- 결론
  - 엔터티 타이핑을 통해 멘션 탐지를 진행하여 기존보다 향상된 성능을 얻지 못하였지만 개체 임베딩 수를 줄여 메모리 부분에서 이점을 얻음
  - 엔터티 타이핑을 통해 좀 더 일반화된 엔터티를 통해 문제를 해결하였기 때문에 다양한 데이터에서 실험을 진행하여 기존 모델과 비교 실험 필요(기존 모델은 특정 도메인에서 데이터만 잘처리)

감사합니다.
Q&A