한국어 개체명 인식 & 링킹 & 의존파싱 통합모델

Cognitive Computing Lab.

민진우[O1], 나승훈[2], 신종훈[3], 김영길[4]
[12] 전북대학교, [34] 한국전자통신연구원
jinwoomin4488@gmail.com, nash@jbnu.ac.kr, {jhshin82, kimyk}@etri.re.kr

Ⅰ. 서론

개체명 연결 주어진 문서에 나타난 인명, 지명, 등의 개체를 인식하고 위키피디아와 같은 지식 기반 상의 하나의 개체와 연결하여 특정 개체가 무엇인지 식별하여 중의성을 해결하는 작업

9일 오후 방송한 mbc 예능 프로그램
우리 결혼했어요 에서는 곽시양, 김소연이 화보 촬영하는 …

candidate entities
- **김소연_(1980년) : 0.65**
- **김소연_(2002년) : 0.05**
- **김소연_(1967년) : 0.01**
- …

의존 파싱 문장 내의 두 단어 사이에 지배소와 의존소를 결정해나가면서 문장을 분석하는 자연어처리 분야

Ⅱ. 제안 방법

Proposed System

의존파싱과 개체명 인식을 전이 기반 방식으로 통합한 후 수행하여 개체 연결 단계에서 개체명 정보 뿐 의존 파싱 정보까지 모두 반영하여 개체의 중의성을 해결하는 개체명 인식 & 링킹 & 의존 파싱 통합 모델을 제안하고 실험 결과를 보임

M_t	C_t	S_t	B_t	O_t	Action	M_{t+1}	C_t	S_{t+1}	B_{t+1}	O_t
M	u, C	S	B	O	$Shift$	u, M	C	S	B	O
$(u, \cdots, v), M$	C	S	B	O	$Reduce$	M	C	S	B	$g_e(u, v, r_y), O$
M	u, C	S	B	O	Out	M	C	S	B	$g_e(u, v, \emptyset), O$
M	C	S	u, B	O	P_Shift	M	C	u, S	B	O
M	C	u, S	B	O	P_Reduce	M	C	S	B	O
M	C	u, S	v, B	O	$Left(l)$	M	C	S	$g_l(v, u), B$	O
M	C	u, S	v, B	O	$Right(l)$	M	C	$g_l(u, v), S$	B	O

ⅰ.통합 전이 액션

개체명 인식은 형태소 단위로 수행되며 의존 파싱은 어절 단위로 수행되기 때문에 서로 구분되는 버퍼와 스택을 가지며 개체명 인식을 위한 버퍼와 스택은 C와 M, 의존파싱을 위한 스택과 버퍼는 B와 S로 표현하며 개체명 인식에서는 모든 생성된 결과를 저장하는 별도의 출력 스택 O를 가짐

▪ 개체명 인식 액션

∘ $Shift$: 버퍼 C의 형태소를 스택 M으로 이동하는 액션으로 현재 스택의 개체명의 구성 요소가 됨
∘ $Reduce$: 스택 M의 모든 형태소들을 Pop하여 Chunk를 생성하고 개체명 레이블을 부착하는 액션으로 생성된 개체명은 스택 O로 이동
∘ Out : 버퍼 C의 형태소를 바로 스택 O로 이동시키는 액션으로 이동된 형태소는 개체명에 해당하지 않는다.

▪ 의존 파싱 액션

∘ P_Shift : 버퍼 내의 어절을 스택으로 이동하는 액션
∘ P_Reduce: 스택의 Top에 위치하는 어절을 Pop하여 제거하는 액션
∘ $Left(l)$: 스택의 top과 버퍼의 Top 이 있을 때 v가 지배소 u가 의존소가 되도록 새로운 트리를 생성 후 스택 S 에 push하는 액션
∘ $Right(l)$: $Left$와 반대로 u 가 지배소, v 가 의존소가 되도록 의존 트리를 생성

▪ 전이 액션의 전환

① 개체명 인식→의존파싱: Shift 혹은 Out 액션을 통해 어절의 마지막 형태소가 버퍼에서 이동되었을 때
② 의존파싱→개체명 인식: P-Shift, RightArc 액션을 통해 어절 buffer의 item이 스택으로 이동되었을 때

ⅲ.의존 파싱 & 개체명 인식 & 연결 통합 모델

▪ 전이 액션의 결정 먼저, 개체명 인식과 의존 파싱은 버퍼 C와 스택 M, 의존파싱 버퍼 B와 스택 S, 그리고 출력 스택 O의 Top 노드로부터의 추출된 자질을 각각 $\boldsymbol{C}_t$, $\boldsymbol{M}_t$, $\boldsymbol{B}_t$, $\boldsymbol{S}_t$, $\boldsymbol{O}_t$라 하자. 이를 연결하여 비선형 변환을 통해 최종 상태 표상 $\boldsymbol{F}_t$를 만들고 이를 MLP 출력층으로 연결하여 다음 에 수행될 전이 액션을 결정.

$$\boldsymbol{F}_t = Relu(\boldsymbol{W} \cdot [\boldsymbol{C}_t; \boldsymbol{M}_t; \boldsymbol{B}_t; \boldsymbol{S}_t; \boldsymbol{O}_t])$$

▪ 개체명 연결 개체명 연결이 수행되는 시점은 앞에서 결정된 액션이 개체명의 경계를 결정하고 타입을 부여하는 $Reduce$이 수행될 때이며 이 때 개체 맨션에 대한 20여개의 후보 엔티티들의 정보를 가져오고 각 후보 엔티티 c_i에 표상은 다음과 같이 구성

$$c_i = [c_{ei}; \boldsymbol{m}; \boldsymbol{T}_t;]$$

c_{ei}는 i번째 후보 엔티티에 대한 임베딩이며 $\boldsymbol{m}$ 은 Mention Span에 해당하는 은닉 표상의 평균을 취해 얻은 고정크기의 벡터이며 $\boldsymbol{F}$는 $Reduce$액션이 수행된 시점의 모든 버퍼와 스택의 상태 표상으로 이들 정보를 받음으로 개체명 인식 정보 뿐 아니라 의존 정보 역시 반영하도록 다음 수식과 같이 가장 확률이 높은 후보 엔티티를 선택함으로 개체명 연결을 수행

$$\boldsymbol{l}_t = affine(\tanh(affine(c_i, \cdots, c_n)))$$
$$\boldsymbol{r}_t = softmax(\boldsymbol{l}_t)$$
$$\hat{y}_{t_{EL}} = argmax(\boldsymbol{r}_t(c))$$

Ⅲ. 실험 결과

▪ 실험 집합 약 20여만의 문장의 개체명 링킹 데이터 셋으로부터 3천 문장을 추출 후 각각 2500 문장, 250문장, 250문장으로 학습셋, 개발셋, 평가셋을 구성 파싱 정보를 부착하기 위해서 세종 구문 분석 데이터 셋에서 학습한 KorBERT 기반 한국어 의존 파서[11]를 활용하여 자동 태깅

▪ 개체명 인식 & 링킹 실험 결과

모델	F1	Link F1
Baseline Pipeline 모델	55.72%	40.32%
[NER + EL] Joint 모델	54.51%	40.05%
[NER + EL + Paring] Joint 모델	55.27%	39.70%

▪ 의존 파싱 실험 결과

모델	UAS	LAS
Baseline Arc Eager 파서 모델	89.02%	86.17%
[NER + EL + Paring] Joint 모델	88.78%	85.91%

Ⅳ. 결론 및 향후 연구

본 연구에서는 개체명 인식 & 링킹과 의존파싱을 동시에 수행하는 통합 모델의 실험 결과를 보였으나 기존의 성능보다 소폭 하락한 결과를 보임

제안 모델은 예측된 엔티티의 정보의 활용 즉, knowledge-grounded 기반의 형태가 아니며 연결된 엔티티의 범주, 설명 등의 정보를 다시 다음 전이 액션에 활용하는 방법에 대해 연구할 예정

개체 중의성 문제

- **Comparing the similarities between an**

  input pair(context , mention) and candidate entities

context mention ⟺ **entities**

Compare Similarities

Entity disambiguation → Ranking Task

9일 오후 방송한 mbc 예능 프로그램
우리 결혼했어요 에서는 곽시양, 김소연이 화보 촬영하는 …

candidate entities

- **김소연_(1980년) : 0.65**
- **김소연_(2002년) : 0.05**
- **김소연_(1967년) : 0.01**
- **…**

- 개체 중의성 문제
  - 개체 표현이 2개 이상의 개체와 연결될 때 개체 중의성 문제가 발생
  - 개체 중의성 문제는 순위를 매기는 문제로 연결