KorBERT와 Popularity 정보에 기반한 한국어 개체연결

허정○, 배경만, 임수종

한국전자통신연구원, 언어지능연구실
Jeonghur, kyoungman.bae, isj@etri.re.kr

Korean Entity Linking based on KorBERT and Popularity

Jeong Heo○, Kyung-Man Bae, Soo-Jong Lim
ETRI, Language Intelligence Research Section

요 약

본 논문에서는 KorBERT와 개체 인기정보(popularity)를 이용한 개체연결 기술을 소개한다. 멘션인식(mention detection)은 KorBERT를 이용한 토큰분류 문제로 학습하여 모델을 구성하였고, 개체 모호성해소(entity disambiguation)는 멘션 컨텍스트와 개체후보 컨텍스트 간의 의미적 연관성에 대한 KorBERT기반 이진분류 문제로 학습하여 모델을 구성하였다. 개체 인기정보는 위키피디아의 hyperlink, inlink, length 정보를 활용하였다. 멘션인식은 ETRI 개체명 인식기를 이용한 모델과 비교하였을 경우, ETRI 평가데이터에서는 F1 0.0312, 국립국어원 평가데이터에서는 F1 0.1106의 성능 개선이 있었다. 개체 모호성해소는 KorBERT 모델과 Popularity 모델을 혼용한 모델(hybrid)에서 가장 우수한 성능을 보였다. ETRI 평가데이터에서는 Hybrid 모델에서의 개체 모호성 해소의 성능이 Acc. 0.8911 이고, 국립국어원 평가데이터에서는 Acc. 0.793 이였다. 최종적으로 멘션인식 모델과 개체 모호성해소 모델을 통합한 개체연결 성능은 ETRI 평가데이터에서는 F1 0.7617 이고, 국립국어원 평가데이터에서는 F1 0.6784 였다.

주제어: KorBERT, 개체연결, 멘션인식, 개체모호성해소

1. 서론

개체연결(entity linking)은 문장 내에서 인식된 개체 멘션(entity mention)을 지식베이스의 개체(entity)와 연결하는 기술이다. 즉, 멘션이 있는 문장의 컨텍스트 정보에 기반하여 지식베이스 내의 다양한 개체후보들 중, 의미적으로 부합하는 개체를 찾아서 연결하는 개체 의미분석 기술로 볼 수 있다.

개체연결은 크게 세가지 기술로 구성된다. 첫째, 멘션인식 기술(mention detection)이다. 일반적으로 개체명 인식 기술을 활용한다. 개체명 인식은 문장 내에 표현되는 다양한 고유명사(인명, 지명, 기관명 등)를 인식하고, 의미적인 카테고리를 분류하는 기술이다. 둘째, 멘션이 연결될 수 있는 지식베이스 내의 개체후보들을 찾는 개체후보생성(candidate generation)이다. 셋째, 멘션의 문맥상 가장 의미적으로 부합하는 개체후보를 선정하는 개체 모호성해소 기술(entity disambiguation)이다[1,2].

개체연결을 위한 멘션인식과 개체 모호성해소 기술도 언어모델을 이용한 딥러닝 기술로 크게 개선되고 있다[1~3]. 그러나, 문서 도메인에 따른 특성과 인기(popularity) 정보를 활용하고 있지 않다. 본 논문에서는 멘션인식과 개체 모호성해소을 위해서 KorBERT 언어모델을 이용한 토큰분류(token classification) 모델과 이진분류(binary classification) 모델을 개발하였다[4]. 또한, 개체후보의 인기정보를 이용하여 성능개선을 하였다.

2. 관련 연구

CHOLAN 시스템[1]은 개체연결을 위한 두개의 트랜스포머기반 모델(멘션인식, 개체 모호성해소)을 순차적으로 연결한 개체연결 기술이다. End-to-end 방식의 딥러닝 모델을 사용하지 않은 이유는 개체후보 생성, 지식베이스와 사용되는 컨텍스트 정보의 유연성(flexibility)을 높이기 위한 것이다. 유연성을 높임으로써, 사용할 수 있는 컨텍스트 정보에 대한 운용성이 높이지는 장점이 있다.

[2]에서는 단어와 개체 시퀀스의 쌍에 따라 개별 쿼리 파라미터를 두고 셀프 어텐션(self-attention)을 수행하는 LUKE 모델을 이용한 개체연결 기술을 소개하고 있다. 멘션과 개체후보 간의 점수는 LUKE 모델을 통해 인코딩(encoding)된 멘션표상과 개체후보의 위키피디아 문서 대상 RoBERTa 모델의 문맥표상인 [CLS]의 임베딩 정보를 입력으로 Biaffine Attention 함수를 이용하여 계산하였다.

본 논문은 CHOLAN 시스템과 유사하게 멘션인식과 개체 모호성해소를 독립적으로 구성하여, 순차적으로 수행되는 구성을 가지도록 개체연결 시스템을 개발하였다. 또한 CHOLAN 시스템과 [2]와 같이 언어모델을 이용한 멘션 컨텍스트와 개체후보 컨텍스트의 유사도 계산과 함께 개체후보의 인기정보를 추가하여 개체 모호성해소 성능을 개선하였다.

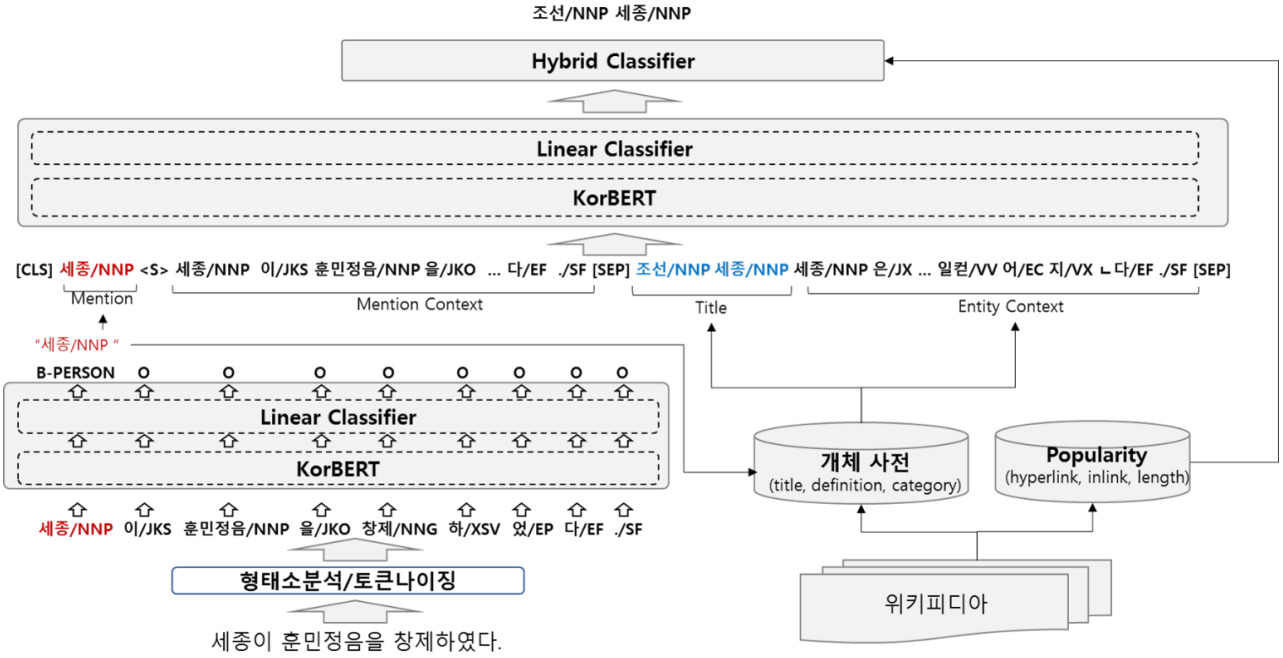

그림 1. KorBERT와 인기정보(popularity)를 이용한 개체연결 모델 구성

3. 개체연결 정의

입력 토큰열은 $W = \{w_1, w_2, \ldots, w_n\}$로 구성되고, 토큰열의 BIO 멘션 태그열은 $T = \{t_1, t_2, \ldots, t_n\}$로 구성된다. BIO 개체 태그셋은 $S = \{B_{MEN}, I_{MEN}, B_{PER}, I_{PER}, O\}$로 구성된다. B_{PER}과 I_{PER}는 인명에 대한 개체태그로 뉴스 도메인 데이터의 인물 상호참조(coreference)를 위해서 다른 멘션과 구분하였다. 따라서, $t_i \in S$의 관계를 가진다. 결국 멘션인식은 $W \rightarrow T$로 변환하는 작업이다. 입력 토큰열 내의 멘션은 $M = \{m_1, m_2, \ldots, m_k\}$로 구성되고, 멘션은 $m_x^{(i,j)} = (w_i, w_{i+1}, \ldots, w_j)$, $(0 < i, j \leq n)$로 구성된다. 멘션의 개체후보 집합은 $C(m_x) = \{e_1^x, e_2^x, \ldots, e_n^x\}$로 구성되고, 개별 개체후보 e_i^x는 지식베이스 K의 개체목록에 존재($e_i^x \in K$)한다. 최종적으로 개체연결은 아래의 수식 (1)과 같다.

$$e_{m_x} = arg \max_{e_i^x \in C(m_x)} p(e_i^x | m_x, W, D_{e_i^x}, P_{e_i^x}) \quad (1)$$

멘션 m_x, 토큰열 W, 개체후보 e_i^x의 개체 컨텍스트 정보 $D_{e_i^x}$, 개체후보 e_i^x의 인기정보 $P_{e_i^x}$가 주어졌을 때, 멘션 m_x의 개체후보 $C(m_x)$에서 확률값이 가장 높은 개체후보 e_{m_x}를 선정하는 것이 개체연결이다.

4. KorBERT와 Popularity를 이용한 개체연결 모델

본 논문에서 소개하는 개체연결 모델은 [그림 1]과 같이 구성된다.

4.1. KorBERT 기반 멘션인식

문장 내의 멘션을 인식하기 위해서 KorBERT 언어모델을 이용한 토큰분류 모듈을 이용하였다. 자연어 문장이 입력되면, KorBERT 모델을 이용하기 위해 형태소 분석을 하고, KorBERT의 Vocabulary에 맞춰 문장을 토큰열(token sequence)로 변환한다. 토큰열이 KorBERT로 입력되고, 토큰 별로 BIO형태의 멘션 태그를 결정하기 위해 KorBERT 상단에 분류 레이어를 둔다.

$$m_i = KorBERT(t_i) \quad (2)$$

4.2. 개체후보 생성

본 논문에서 멘션을 연결할 지식베이스는 위키피디아 사전이다. 위키피디아 사전의 구조적 정보인 타이틀, 리다이렉트(redirect), 동음이의어 정보를 기반으로 멘션 사전을 구축하였다. 멘션이 인식되면 해당 멘션을 사전에서 검색하여 개체후보를 선정한다. 개체후보가 선정될 때, 해당 개체후보의 인기정보를 함께 가져온다. 인기정보는 개체후보 article 내의 hyperlink 개수, 해당 article을 참조하는 article의 개수인 inlink 개수, 해당 article을 구성하는 본문의 길이(length)로 구성된다. 개체후보의 인기도는 아래와 같이 계산된다.

$$p^{feature}(e_i^x) = \frac{count^{feature}(e_i^x)}{\sum_{i=1}^{n} count^{feature}(e_i^x)}, \quad e_i^x \in C(m_x) \quad (3)$$

$$p^{popularity}(e_i^x) = \frac{(p^{hyperlink}(e_i^x) + p^{inlink}(e_i^x) + p^{length}(e_i^x))}{3} \quad (4)$$

개체후보 e_i^x 의 인기도는 인기도 자질인 hyperlink($p^{hyperlink}(e_i^x)$), inlink($p^{inlink}(e_i^x)$), length($p^{length}(e_i^x)$)의 인기도 평균이다.

4.3. KorBERT기반 개체 모호성해소

본 논문에서 개체 모호성해소는 이진분류문제로 처리한다. 멘션에 대한 컨텍스트 정보(mention context)와 개체후보의 컨텍스트 정보(entity context)가 의미적으로 연관성이 있는지 여부를 긍/부정으로 분류하는 것이다. 긍정과 부정에 대한 이진분류는 softmax에 의해서 긍정/부정에 대한 점수로 변환된다. 긍정 레이블에 대한 점수를 멘션 컨텍스트와 개체후보 컨텍스트의 의미적 연관성 점수로 사용한다. 멘션 컨텍스트 정보는 연결하고자 하는 멘션과 멘션이 포함된 문장을 연결한 것이다. 본 논문에서는 멘션이 있는 문장의 앞뒤 문장을 추가하여 3문장(± 1)으로 멘션 컨텍스트를 입력하였다. 개체 컨텍스트는 개체후보의 타이틀과 정의문을 입력하였다. 개체후보의 정의문은 위키피디아 본문의 첫번째 문장(단락)을 이용하였다. KorBERT의 입력은 "[CLS]멘션 컨텍스트[SEP]개체 컨텍스트[SEP]"로 구성된다. CLS 토큰을 이용하여 이진분류를 수행하였다.

$$p^{cls}(e_i^x) = KorBERT(m_x, W, D_{e_i^x}) \quad (5)$$

4.4. 개체 인기정보를 혼용한 개체 모호성해소

개체후보의 인기정보는 개체 모호성해소에 주요한 정보이다. 본 논문에서는 개체 인기정보를 KorBERT기반 이진분류 결과와 혼용하여 점수를 계산하고 개체 모호성을 해소한다. 멘션 m_x에 대한 개체후보 e_i^x의 개체 모호성해소 점수는 아래의 수식과 같이 계산된다.

$$p(e_i^x) = \left(\alpha * p^{popularity}(e_i^x)\right) + \left((1-\alpha) * p^{cls}(e_i^x)\right) \quad (6)$$

개체후보 인기정보의 반영 가중치 α는 실험을 통해서 결정된다. 최종적으로 개체연결은 아래의 수식과 같다.

$$e_{m_x} = arg \max_{e_i^x \in C(m_x)} p(e_i^x) \quad (7)$$

멘션 m_x의 개체후보 집합 $C(m_x)$에서 $p(e_i^x)$의 점수가 가장 높은 e_i^x를 연결하는 것이다. e_{m_x}의 점수가 실험을 통해 지정된 임계값을 넘지 못한 경우, NIL 연결로 처리한다. 임계값은 실험을 통해 0.5로 결정하였다.

5. 학습데이터 구축

KorBERT 기반 멘션인식과 개체 모호성해소 기술은 학습데이터가 필요하다. 멘션은 토큰분류에 기반한 시퀀스 레이블링 문제이고, 개체 모호성해소는 두 문장(문단)에 대한 의미적 연관성에 대한 긍/부정 이진분류 문제이다.

5.1. 멘션인식 학습데이터 구축

멘션인식 학습데이터는 국립국어원의 "모두의 말뭉치" 중, 개체명 분석 말뭉치(2019년, 2020년)를 이용하였다. 개체명 분석 말뭉치에서 개체연결의 대상이 되는 개체태그[1]가 부착된 멘션을 대상으로 학습데이터를 구축하였다. 개체명 분석 및 연결 말뭉치(2021년) 데이터는 평가데이터로 활용하였다. 개체명 분석 말뭉치(2019년, 2020년) 데이터는 9:1로 구분하여 학습데이터와 개발데이터로 활용하였다. 데이터의 구성 통계는 <표 1>과 같다.

표 1. 멘션인식 데이터 구성

구분	어절 개수	크기
학습데이터	17,801,894	180M
개발데이터	2,013,187	21M
평가데이터	6,439,884	64M

KorBERT 기반 멘션인식을 위한 BIO 토큰분류 모델의 평가결과는 <표 2>와 같다.

표 2. 멘션인식 평가 결과 (BIO 인식 성능)

구분	Precision	Recall	F1 Score
MENTION	0.8372	0.8392	0.8382
PERSON	0.8657	0.8566	0.8612
Total	0.8476	0.8455	0.8466

5.2. 개체 모호성해소 학습데이터 구축

개체 모호성해소 학습데이터는 위키피디아의 hyperlink를 이용하여 자동으로 구축하였다. Hyperlink가 포함된 문장과 hyperlink로 연결되는 article의 첫번째 문장(단락)을 추출하여 개체 모호성해소 이진분류의 긍정 레이블 데이터로 활용하였다. 부정 레이블의 데이터는 hyperlink가 연결된 표현을 질의로 멘션 사전을 검색하여 개체연결 후보를 가져온다. 개체연결 후보 중, 긍정에 해당하는 개체후보를 제외한 후보들 중 랜덤으로 하나의 후보를 선정하고 해당 개체후보의 정의문(article의 첫 문장(단락))을 추출하여 부정 레이블 데이터로 활용하였다. 긍정 데이터와 부정 데이터는 1:1로 구성하였다. 구축된 데이터는 8:1:1로 학습데이터:개발데이터:평가데이터로

[1] 인명, 지명, 기관명, 작품명

구분하였다. 데이터의 구성 통계는 <표 3>과 같다. 긍/부정 이진분류에 대한 평가 결과는 F1 0.9588 이었다.

표 3. 개체 모호성해소를 위한 이진분류 데이터 구성

구분	문장 쌍 개수	크기
학습데이터	2,763,002	4.0G
개발데이터	346,058	511M
평가데이터	345,628	512M

6. 실험 및 평가

6.1. 평가데이터 구성

개체연결에 대한 평가데이터는 크게 두 종류로 구성하였다. 첫째, ETRI에서 구축한 평가데이터이고, 둘째, 국립국어원의 "모두의 말뭉치" 중 개체명 분석 및 연결 말뭉치(2021년)이다. 데이터의 구성 정보는 <표 4~표 5>와 같다.

표 4. ETRI 평가데이터

도메인	기사 개수	멘션 수
뉴스	98개	3,033개
위키피디아	100개	2,434개

표 5. 국립국어원 개체명 연결 말뭉치(2021년) 데이터

도메인	파일 개수	사이즈	멘션 수
뉴스	1,678개	10.7M	65,766개
대화(구어체)	270개	18.2M	48,591개

6.2. 개체연결 평가

개체연결 평가는 두가지 관점에서 진행하였다. 첫째, 멘션인식 방법론에 따른 평가로, 본 논문에서는 ETRI 개체명 인식기[5]를 이용한 모델(NER)과 KorBERT 기반 멘션인식 모델(KorBERT)을 비교하였다. 두번째, 개체 모호성해소에서 개체후보의 인기정보만을 이용한 모델(Popularity), KorBERT기반 이진분류에 기반한 모델(KorBERT)과 두 모델을 혼용한 모델(Hybrid)을 비교하였다. 실험에 사용된 인기정보 가중치 α는 실험을 통해 0.5로 결정하였다.

표 6. 멘션인식(MD) 모델 성능 비교

데이터	MD 모델	Precision	Recall	F1
ETRI	NER	0.8299	0.8175	0.8236
	KorBERT	0.8909	0.8215	0.8548
국립국어원	NER	0.6903	0.8087	0.7448
	KorBERT	0.8695	0.8418	0.8554

표 7. 개체 모호성해소(ED) 모델 성능 비교

데이터	ED 모델	ED	Entity Linking (MD+ED)		
		Acc.	P.	R.	F1
ETRI	KorBERT	0.8697	0.7748	0.7145	0.7434
	Popularity	0.8742	0.7788	0.7181	0.7472
	Hybrid	0.8911	0.7939	0.7320	0.7617
국립국어원	KorBERT	0.7738	0.6729	0.6514	0.6619
	Popularity	0.7720	0.6713	0.6499	0.6604
	Hybrid	0.7930	0.6896	0.6676	0.6784

<표 6>은 멘션인식(MD)의 두 모델에 대한 실험 결과이다. 두 데이터 모두에서 KorBERT기반 모델이 ETRI 개체명 인식기를 이용한 모델보다 멘션인식 성능이 우수하였다. 특히, 정확률이 크게 개선되었다. ETRI 데이터의 경우, ETRI 개체명 인식기의 결과를 기반으로 데이터를 구축하였기 때문에 국립국어원 데이터와 비교하여 두 모델간 성능차이가 상대적으로 크지 않은 것으로 분석된다.

<표 7>은 개체 모호성해소 모델 별 성능(ED)과 멘션인식과 개체 모호성 해소 모델이 통합된 개체연결(Entity Linking) 성능을 보여주고 있다. 개체후보의 인기정보만을 이용한 모델(Popularity)이 KorBERT기반 이진분류를 이용한 모델(KorBERT)과 성능이 비슷하였다. 그리고, 두 모델을 혼용한 모델(Hybrid)이 가장 우수하였다.

7. 결론

본 논문에서는 KorBERT와 개체 인기정보를 이용한 개체연결 기술을 소개하였다. 멘션인식은 KorBERT를 이용한 토큰분류 문제로 학습하여 모델을 구성하였고, 개체 모호성해소는 멘션 컨텍스트와 개체후보 컨텍스트 간의 의미적 연관성에 대한 KorBERT기반 이진분류 문제로 학습하여 모델을 구성하였다. 개체 인기정보는 위키피디아의 hyperlink, inlink, length 정보를 활용하였다.

멘션인식은 ETRI 개체명 인식기를 이용한 모델과 비교하였을 경우, ETRI 데이터에서는 F1 0.0312, 국립국어원 데이터에서는 F1 0.1106의 성능 개선이 있었다.

개체 모호성해소는 KorBERT 모델과 Popularity 모델을 혼용한 모델(Hybrid)에서 가장 우수한 성능을 보였다. ETRI 데이터에서는 Hybrid 모델에서의 개체 모호성 해소의 성능이 Acc. 0.8911이고, 국립국어원 데이터에서는 Acc. 0.7930이였다. 최종적으로 멘션인식 모델과 개체 모호성해소 모델을 통합한 개체연결의 성능은 ETRI 데이터에서는 F1 0.7617 이고, 국립국어원 데이터에서는 F1 0.6784 였다. 본 논문의 실험에서는 개체후보의 인기정보만을 이용하여도 개체 모호성해소의 성능이 우수함을 알 수 있었다.

향후연구는 개체후보의 컨텍스트 정보로 위키피디아 문서의 전체내용과 다양한 구조적 정보를 이용하는 방법과 문서내에 존재하는 다양한 상호참조 문제 및 개체후

보 검색을 위한 방법 개선에 대해서 연구를 진행할 계획이다.

* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2013-2-00131, [엑소브레인-총괄/1세부] 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)과 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. RS-2022-00187238, 효율적 사전학습이 가능한 한국어 대형 언어모델 사전학습 기술 개발)

참고문헌

[1] Manoj Prabhakar Kannan Ravi, et al., CHOLAN: A Modular Approach for Neural Entity Linking on Wikipedia and Wikidata, EACL 2021.
[2] 민진우 외 4명, “LUKE를 이용한 한국어 자연어 처리: 개체명 인식, 개체 연결”, KIISE Transactions on Computing Practices, Vol. 28, No. 3, pp. 175-183, 2022.
[3] Wei Shen, et al., Entity Linking Meets Deep Learning: Techniques and Solutions, IEEE TKDE, 2021.
[4] Jacob Devlin, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAALC, 2019.
[5] C.K. Lee et al., “Fine-Grained Named Entity Recognition Using Conditional Random Fields for Question Answering,” Proc. Asia Conf. Inform. Retrieval Technol., Singapore, Oct. 16-18, 2006, pp. 581-587.