한국어 개체명 인식 과제에서의 의미 모호성 연구*[1)]

김성현○, 송영숙, 송치성**, 한지윤***
스마일게이트 AI○, 경희대학교, 연세대학교***
seonghkim@smilegate.com○, youngsoksong@khu.ac.kr, daydrilling@gmail.com**,
clinamen35@yonsei.ac.kr***

A study on semantic ambiguity in the Korean Named Entity Recognition

Seonghyun Kim○, Youngsook Song, Chisung Song**, Jiyoon Han***
Smilegate.AI○, KyungHee University, Yonsei University***

요 약

본 논문에서는 맥락에 따라 개체명의 범주가 달라지는 어휘를 중심으로 교차 태깅된 개체명의 성능을 레이블과 스팬 정답률, 문장 성분과 문장 위치에 따른 정답률로 나누어 살펴 보았다. 레이블의 정확도는 KoGPT2, mBERT, KLUE-RoBERTa 순으로 정답률이 높아지는 양상을 보였다. 스팬 정답률에서는 mBERT가 KLUE-RoBERTa보다 근소하게 성능이 높았고 KoGPT2는 매우 낮은 정확도를 보였다. 다만, KoGPT2는 개체명이 문장의 끝에 위치할 때는 다른 모델과 비슷한 정도로 성능이 개선되는 결과를 보였다. 문장 종결 위치에서 인식기의 성능이 좋은 것은 실험에 사용된 말뭉치의 문장 성분이 서술어일 때 명사의 중첩이 적고 구문이 패턴화되어 있다는 특징과 KoGPT2가 decoder기반의 모델이기 때문으로 여겨지나 이에 대해서는 후속 연구가 필요하다.

주제어: 개체명, 의미 모호성, NER, BERT, KLUE

1. 서론

개체명 인식(NER, Named Entity Recognition)에서는 비정형의 자연어 텍스트에서 인명, 지명, 조직명 등 고유명을 주축으로 미리 정의된 특정한 언어 표현을 인식하고 분류한다[1]. 그렇기 때문에 개체명 인식에서는, 각각의 개체명에 범주적 특성을 명확히 하고 그에 따라 일관성 있게 태깅하는 것이 중요하다. 그런데 자연어 말뭉치에서는 동음이의어, 축약어 등의 결과로 다른 의미의 단어가 형태가 같아지기도 하고, 동일 개체명이라도 문맥에 따라 다른 의미로 사용되기 때문에 개체명 태깅의 범주가 달라지게 된다.

(1) 개체명 '한국'이 '장소(LC)'와 '조직(OG)'로 태깅된 예시

ㄱ. <한국:LC>에서 찍은 독립영화도 이보다는 탄탄할 수 있었을거라 생각하는데 왜 출연했는지 의문이 들었음 <NSMC>

ㄴ. 각 조 3위에 오른 6개국 중에서는 <한국:OG>이 승점 4, 골득실 0으로 가장 좋은 성적을 남겼고, 우즈베키스탄(승점 4·골득실 -1), 멕시코(승점 3·골득실 1), 가나(승점 3·골득실 0·5득점)가 뒤를 이어 16강에 진출했다. <Wikitree(2013. 6. 30.)>

예(ㄱ)에서 '한국'은 '한반도 지역'이라는 공간적 의미로 사용되었으므로 지명(LC)으로 태깅된다. 반면 (ㄴ)에 쓰인 '한국'은 "U20월드컵 대표팀, 콜롬비아와 16강 격돌 확정"이라는 헤드라인으로 실린 기사의 일부로 여기서 '한국'은 '지명'을 나타낸다기보다는 16강에 진출한 '한국 대표팀'으로 해석된다. 그에 따라 (ㄴ)의 '한국'은 기관 및 조직(OG , ORGANIZATION)으로 태깅하게 된다. 본 논문에서 다루려고 하는 것은 이와 같이 동일한 형태로 표현되지만 개체명의 범주가 달라지는 어휘들을 의미 모호성이 있는 것으로 판단하고 이렇게 의미 모호성이 있는 과제에서 사전 학습 모델(pretraining model)의 성능 차이를 비교하여 분석해 보고자 한다.

그동안 개체명 인식에서 이와 같은 의미 모호성 해소를 위한 연구는 링킹 알고리즘에 의존하는 연구가 주도적이었다. 알고리즘의 적용은 위키피디아와 같은 백과사전의 정보를 개체와 연결시키는 방법[2, 3]과 트윗 개체 링킹과 같이 사용자 정보를 활용해 정확성을 높이는 방법[4] 등이 진행되었다. 이러한 방법의 공통점은 BERT[5], RoBERTa[6], GPT2[7]와 같이 문맥 단어 표현을 학습하는 딥러닝 모델들이 태깅한 개체 표현이 잘못 연결되었을 경우에 이를 보완하여 성능을 개선하는 방식이라는 것이다. 본 논문에서는 딥러닝 기반의 사전 학습 모델들이 맥락 정보를 해결하는 정도에 차이가 있는지 살펴보고, 그러한 차이가 생긴 원인을 파악하려는 것이 목적이다.

논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안하는 개체명의 의미 모호성의 원인을 세분화해서 살펴보고 그 특성을 파악한다. 3장에서는 트랜스포머 계열의 사전 학습 모델의 테스트 결과를 분석한다. 4장에서는 결론 및 발전 방향을 제시한다.

---

1) 이 논문은 Park, Sungjoon 외 (2021) KLUE:Korean Language Understanding Evaluation[8] 에서 구축된 KLUE-NER 말뭉치 (https://klue-benchmark.com/tasks/69/overview/description)와 개체명 인식을 위한 학습 및 평가 코드를 사용하였다. 연구자들 역시 KLUE 팀의 개체명 말뭉치 구축에 참여하면서 논의한 내용을 토대로 발표문을 작성하였다. 개체명 인식 연구가 가능하도록 토대를 만들어 준 KLUE 팀에 감사의 말씀을 드린다.

2. 개체명 태깅 범주에서 중의성이 발생하는 원인

분석에는 KLUE-NER[8]을 사용하였는데 개체명은 인물(PS), 장소(LC), 조직(OG), 날짜(DT), 시간(TI), 수량(QT)의 6개 범주로 구분되어 태깅되어 있다. 이들 개체명 간에 범주별 중복 유형이 발생하는 원인은 세 가지 정도로 정리할 수 있는데 먼저 동음이의어에 의해 형태가 같은 개체명이 생기는 경우이다. 가령 '하루'는 'DT'로 18번 쓰이고 'PS'로 한 번 쓰였는데 '하루네 방송'이 인명(PS)으로 쓰인 예이다.

두 번째로는 절단어 또는 축약어 유형이 있었는데 '하나'의 경우, 'QT'로 67번 태깅되었으나 '하나 은행'을 줄여서 '하나'로 표현한 기사에서는 'OG'로 태깅되었다.

(2) 개체명 '하나'가 '조직(OG)'으로 태깅된 예시

하나금융은 <하나:OG>·외환은행에 대한 합병을 외환노조와 합의했다고 13일 밝혔다.

<Wikitree(2015. 7. 13.)>

세 번째로 다음 예시 (3)은 서론의 예시 (1)과 같은 이유로 한국이 '장소'가 아닌 '조직'으로 태깅된 경우로 가장 자주 범주 교차적으로 태깅되는 주요한 예시라고 할 수 있다.

(3) 개체명 '한국'이 '조직(OG)'으로 태깅된 예시

유씨는 492억원의 횡령·배임 혐의를 받고 있어 <한국:OG>과 프랑스 양국 사이의 조약에 따라 일단 범죄인 인도 대상이다.

<Wikitree>

다음 3장에서는 이렇게 동일한 형태의 개체명이 다른 범주로 태깅될 때 개체명이 문장 내에서 어떠한 역할을 하고 있는지에 따라 구분하고 딥러닝 기반 모델들을 통해 성능을 비교해서 검증해 볼 것이다. 이를 통해 맥락 정보가 교차 범주의 개체명 성능에 어느 정도 영향을 주었는지 구체적으로 살펴볼 수 있을 것이다.

3. 개체명 의미 모호성 판단 실험

본 연구에서는 KLUE-NER[8]의 데이터를 가공하여 실험을 진행하였다. 모델의 훈련과 검증에는 KLUE-NER의 학습 데이터를 사용하였으며, 모델 실험은 검증 데이터의 일부를 추출하여 진행하였다.

3.1. 실험 데이터 구성

[표 1] KLUE-NER 데이터 구성

Source data	Train	Dev/Validation	Total
Wikitree	11,435	2,534	13,969
NSMC	9,573	2,466	12,039
Total	21,008	5,000	26,008

[표 1]은 KLUE-NER 데이터의 구성을 보여준다. 전체 데이터는 소셜 미디어 뉴스로 분류되는 '위키트리(Wikitree)'와 웹 사용자의 발화 특성을 살펴볼 수 있는 '네이버 영화 리뷰(NSMC)'로 구성되어 있다. 각 장르의 비율은 데이터별로 1:1의 비율을 보이며, 학습 데이터는 21,008 문장, 검증용 데이터는 5,000 문장으로 구성되어 있다.

본 연구의 실험 데이터는 이 중 검증(dev) 데이터에서 두 개 이상의 레이블로 교차 태깅된 69 종의 개체명을 포함한 823 용례(중복 제외 648문장)을 추출하여 구축하였다. 개체명이 교차 태깅될 가능성이 높다는 것은 의미 모호성이 높다는 의미이다. 교차 태깅된 개체명 범주의 유형은 [표 3]에 제시된 바와 같다. 총 5개 유형으로, 'LC-OG'가 가장 높은 빈도로 전체 교차 태깅 용례의 85.54%를 차지한다. 그 다음은 OG-QT로 전체의 8.26%를 차지하며 DT-PS와 LC-OG-PS, QT-DT를 교차 태깅한 경우는 각각 전체 용례의 3% 이하를 차지한다.

실험 데이터는 개체명의 의미 모호성을 다루는 연구의 특성 상 연구진 간에 논의한 지침에 맞춰 KLUE-NER 데이터의 태깅 결과를 일부 변경하여 제작하였다. 각 범주에 해당하는 세부 개체명의 빈도와 종 수는 다음 표와 같고 이 실험 데이터에 대한 정답률은 3.2.2에서 분석한다.

[표 2] KLUE-NER 개체명 빈도

개체명	학습 데이터 (종 수)	평가용 데이터 (종 수)	합
PS	14,453 (5,428)	4,418 (2,706)	18,871 (8,134)
LC	6,663 (2,068)	1,649 (896)	8,312 (2,964)
OG	8,491 (3,008)	2,182 (1,291)	10,673 (4,299)
DT	8,029 (1,608)	2,312 (835)	10,341 (2,443)
TI	2,020 (573)	5,45 (268)	2,020 (841)
QT	11,717 (3,628)	3,151 (1,763)	14,868 (5,391)
Total	51,373 (16,313)	14,257 (7,759)	65,630 (24,072)

KLUE-NER 전체 데이터의 개체명 빈도를 통해 실험 데이터를 추출한 검증용 데이터는 7,759 종류의 개체명을 포함하고 있었으며, 이들의 총 빈도는 14,257임을 알 수 있다. 그 중에서 교차 태깅된 개체명의 범주별 빈도는 다음 표와 같다.

[표 3] 실험 데이터 내 교차 태깅 범주별 빈도
(단위: %, 괄호 안: 개체명 빈도수)

개체명 범주	빈도
LC-OG	85.54(704)
OG-QT	8.26(68)
DT-PS	3.52(29)
LC-OG-PS	2.07(17)
QT-DT	0.61(5)
합	100.00(823)

위의 [표3]을 통해 대체로 장소(LC)와 조직(OG) 범주에서 의미 모호성이 발생하는 예시가 많다는 것을 알 수 있다. 범주 교차가 빈번했던 어휘의 구체적 예시는 아래 [표4]에서 확인할 수 있다. 개체명이 속한 범주가 세 개 이상에 걸친 예도 있었지만 여기서는 고빈도 순으로 2개까지만 나타냈다.

[표 4] 실험 데이터 내 교차 태깅 개체명 빈도

번호	개체명	범주1	빈도	범주2	빈도	합
1	한국	LC	109	OG	11	120
2	하나	QT	67	OG	1	68
3	일본	LC	55	OG	10	65
4	미국	LC	48	OG	10	58
5	중국	LC	51	OG	7	58
⋮	⋮	⋮	⋮	⋮	⋮	⋮
70	말레이시아	LC	1	OG	1	2

빈도에 따른 결과로 보면, '한국', '일본', '미국', '중국' 등 국가명과 같이 'LC-OG'로 교차 태깅될 수 있는 데이터가 고빈도를 보이며, '하나'와 같이 'QT'와 'OG'가 교차 태깅되는 경우도 눈에 띈다. 전반적으로 'OG'와 교차 태깅되는 경우 개체명이 'OG'로 태깅되는 비율보다, 'LC'로 태깅되는 비율이 높음을 알 수 있다. '한국'의 경우 용례 중 'LC'로 태깅되는 사례가 109개, 'OG'로 태깅된 사례가 11개이며 다른 국가명도 유사한 양상을 보이나 '북한'의 경우는 예외적으로 'OG'로 태깅되는 비율이 'LC'로 태깅되는 비율이 높다.

3.2. 실험 결과 및 분석

실험은 모델별 성능 평가를 진행한 뒤 교차 태그 유형별로 각 모델의 정답률을 살펴보았다. 모델별 성능 평가는 KLUE-NER의 dev 데이터를 test 데이터로 삼아 진행하였고, 교차 태그 유형별 각 모델의 정답률은 3.1에서 추출한 실험 데이터를 바탕으로 진행하였다.

3.2.1 모델별 성능 평가

사전 학습 모델에 따른 개체명 인식의 의미모호성 차이를 관찰하고자, 문맥 이해에 특화된 최신 언어모델인 BERT[5], RoBERTa[6], GPT2[7]를 활용했다. BERT 모델은 다국어 모델인 BERT-base-multilingual-cased (mBERT),RoBERTa와 GPT2는 각각 한국어로 학습된 KLUE-RoBERTa-base, 그리고 SKT에서 공개한 KoGPT2-base-v2를 사용하였다. mBERT는 한국어를 포함한 총 119,547개의 사전으로 구성된 WordPiece 토큰화 방법을 사용한다. KoGPT2는 Byte-pair encoding(BPE)[9] 을 통해 토큰화 과정을 거치며, 총 51,200개의 사전을 사용할 수 있다. KLUE-RoBERTa는 한국어에 특화된 형태소 기반 서브 워드 (Morpheme-based subword) 토큰화와 32,000개의 사전을 사용한다. 각 모델의 토큰화 방식은 [표 5]와 같다.

[표 5] 입력 문장 토큰화 예시

토큰화 방법	토큰 시퀀스
원시 텍스트	개체명 인식은 자연어 처리의 핵심과제이다.
형태소 기반의 Subword (KLUE-RoBERTa)	개체 / ##명 / 인식 / ##은 / 자연 / ##어 / 처리 / ##의 / 핵심 / ##과 / ##제이 / ##다 / .
WordPiece (mBERT)	개 / ##체 / ##명 / 인 / ##식은 / 자 / ##연 / ##어 / 처 / ##리의 / 핵 / ##심 / ##과 / ##제 / ##이다 / .
Character BPE (KoGPT2)	__개체 / 명 / __인식은 /__자연 / 어 / __처/ 리의 / __핵심 / 과제 / 이다.

개체명 분석을 위한 모델 학습은 각 토큰의 마지막 층에 범주를 분류하는 선형 계층(linear layer)를 부착하여 파인-튜닝하는 방식으로 설계하였다. 각 파인-튜닝의 학습률(learning rate)는 1e-4, 배치 크기는 8, 에포크는 4로 고정했으며, Huggingface의 Auto Model For Token Classification 클래스와 Colab pro의 Tesla P100-PCIE-16GB를 사용해 학습하였다. 학습과 평가 데이터는 KLUE-NER-v1.1의 train, dev 데이터를 활용하였다. 각 모델에 대한 평가 방법으로 KLUE-NER과 동일하게 음절 기준 F1 점수와 개체명 기준 F1 점수[9]를 사용하였으며, 결과는 [표 6]과 같다.

[표 6] KLUE-NER-dev를 기준으로 한 각 모델 성능 평가

모델	음절 기준 F1 Score	개체명 기준 F1 Score
KLUE-RoBERTa	90.73	83.09
mBERT	88.00	74.33
KoGPT2	83.15	70.93

다국어 모델인 mBERT와 비교하여 한국어로 학습한 KLUE-RoBERTa가 더 좋은 개체명 인식 결과를 보여주었으며, 동일한 Transformer encoder 모델인 mBERT가 Transformer decoder 모델인 KoGPT2보다 높은 성능을 나타냈다.

3.2.2 문장 성분에 따른 교차 태그의 모델별 성능

개체명의 의미 모호성을 해소하는데 가장 중요한 요소는 사전 학습 모델이 주어진 문장에 대한 문맥 정보를 올바르게 임베딩(embedding)할 수 있는 능력이다. 따라서 모델별 교차 태그 범주의 성능 비교를 위해 교차 태그 범주별로 정답률을 살펴 보았는데 결과는 [표 7]과 같다. [표 7]에서 모델의 정답률은 레이블(label)의 정답률과 스팬(span)의 정답률로 나누어 평가한 것이다. 레이블의 정답률은 개체명 스팬의 정확도와 상관없이, 개체명에 해당하는 문자열에 부여된 레이블과 개체명의 레이블이 일치하는지만을 평가의 기준으로 삼았다. 예를 들어 <한:OG>가 정답인 경우 모델이 <한일:OG>로 태깅한 경우 레이블은 맞춘 것으로 판별하였다. 다만 이 경우 스팬은 틀린 것으로 간주하였다.

[표 7] 교차 태그 범주별 모델 성능
(단위:%, 괄호 안:개체명 빈도수)

	레이블 정답률			스팬 정답률		
교차 태깅 범주	KLUE-RoBERTa	mBERT	KoGPT2	KLUE-RoBERTa	mBERT	KoGPT2
LC-OG	83.95 (591)	82.39 (580)	72.59 (511)	90.06 (634)	90.20 (635)	58.10 (409)
OG-QT	98.53 (67)	95.59 (65)	82.35 (56)	95.59 (65)	77.94 (53)	42.65 (29)
DT-PS	93.10 (27)	93.10 (27)	58.62 (17)	86.21 (25)	93.10 (27)	55.17 (16)
LC-OG-PS	88.24 (15)	94.12 (16)	64.71 (11)	29.41 (5)	100.00 (17)	0.00 (0)
DT-QT	100.00 (5)	80.00 (4)	40.00 (2)	100.00 (5)	100.00 (5)	40.00 (2)
총계	85.66 (705)	84.08 (692)	72.54 (597)	89.19 (734)	89.55 (737)	55.41 (456)

레이블 정답률의 경우 KLUE-RoBERTa는 85.66%, mBERT는 84.08%로 근소한 차이를 보였으나, KoGPT2의 경우 72.54%로 다른 모델에 비해 11% 이상 낮은 정답률을 보였다. KLUE-RoBERTa와 mBERT의 경우는 스팬 정답률이 레이블 정답률보다 높은 경향을 보이는데 mBERT가 89.55%, KLUE-RoBERTa가 89.19%로 mBERT의 성능이 근소하게 높았다. KoGPT2의 경우 스팬 정답율이 55.41%로 레이블 정답률보다 20% 가까이 낮아지는 경향을 보였다.

교차 태깅 범주별로 레이블 정답률을 살펴보면 KLUE-RoBERTa와 mBERT는 LC-OG에서 다른 범주들보다 낮은 성능을 보이는 공통적인 경향을 보였다. KLUE-RoBERTa의 평균 레이블 정답률은 85.66%인데, LG-OG의 정답률은 83.95%이고, mBERT의 평균 정답률은 84.05%인데 LC-OG의 정답률은 82.39%였다. 이에 반해 KoGPT2의 경우 평균 정답률은 72.54%이고 LG-OG의 정답률이 72.59%로 정답률에 큰 차이가 없었다. 또한 KLUE-RoBERTa의 경우 LC-OG를 제외하면 다른 모든 범주가 해당 모델의 평균적인 정답률을 상회하였고, mBERT의 경우는 DT-QT의 정답률이 80%로 평균 정답률보다 낮았으나 이는 DT-QT의 전체 용례 수가 5개뿐이기 때문에 유의미한 차이로 보기는 어렵다. KoGPT2의 경우 OG-QT의 정답률이 82.35%로 가장 높고, DT-PS는 58.62%, LC-OG-PS는 64.71%로 해당 모델의 평균 정답률보다 낮았다. 결론적으로, 전체 출현 빈도와 양쪽 범주에서 각각 등장한 빈도가 높았던 'LC-OG'의 경우 정확도가 평균 정답률보다 낮거나 비슷하다고 볼 수 있다. 다른 범주의 경우에는 현재의 결과만으로 성급히 결론을 내기에는 무리한 측면이 있지만 BERT 계열의 모델과 GPT 모델 사이에는 어느 정도 성능 차이를 보였다.

[표 8]은 개별 레이블별 모델의 정답률을 보여주는 것으로, 각 레이블에 대한 태깅 성능이 교차 태깅 범주별 성능에 영향을 미쳤음을 알 수 있다. KLUE-RoBERTa와 mBERT의 경우 OG의 정답률이 각각 72.61%, 66.09%로 다른 레이블에 비해 현저히 낮은 경향을 보여주고, KoGPT2의 경우에도 OG의 정답률이 55.65%로 낮은 동일한 경향을 보여주나, PS의 정답률이 9.09%로 유독 PS에 대하여 낮은 성능을 보였다.

[표 8] 레이블 별 모델 성능
(단위:%, 괄호 안:개체명 빈도 수)

레이블	KLUE-RoBERTa	mBERT	KoGPT2
LC	89.38 (429)	90.42 (434)	81.88 (393)
OG	72.61 (167)	66.09 (152)	55.65 (128)
QT	98.59 (70)	95.77 (68)	80.28 (57)
PS	90.91 (20)	90.91 (20)	9.09 (2)
DT	95.00 (19)	90.00 (18)	85.00 (17)
총계	85.66 (705)	84.08 (692)	72.54 (597)

[표 9]는 가장 높은 비율을 차지하고 있는 LC와 OG로 교차 태깅되는 개체명를 포함한 용례를 분석하여, 해당 개체명의 문장 성분을 살펴본 예시이다. 문장 성분은 각각의 개체명이 문장 내에서 어떠한 역할을 하고 있는지를 살펴 보기 위한 것으로 용례가 완전한 문장을 구성하지 못하여 어떤 성분으로 사용되었는지 판별할 수 없는 경우는 '기타 범주'로 처리하였다.

[표 9] 교차 태그 범주가 'LC-OG'인 유형의 문장 성분별 정답률(단위:%, 괄호 안:개체명 빈도수)

문장성분	KLUE-RoBERTa	mBERT	KoGPT2
부사어	88.10 (237)	88.48 (238)	84.39 (227)
주어	81.97 (150)	79.78 (146)	60.11 (110)
관형어	78.88 (127)	75.78 (122)	66.46 107)
목적어	81.94 (59)	79.17 (57)	69.44 (50)
서술어	100.00 (12)	91.67 (11)	91.67 (11)
기타 범주	85.71 (6)	85.71 (6)	85.71 (6)
총계	83.95 (591)	82.39 (580)	72.59 (511)

KLUE-RoBERTa와 mBERT의 경우 주어와 목적어에서의 정답률이 평균보다 낮지만, 비슷한 경향을 보이고 관형어일 때가 모든 범주 중 가장 낮은 정답률을 보였다. KoGPT2의 경우는 주어에서의 정답률이 가장 낮고 관형어와 목적어에서도 평균보다 낮은 성능을 보였다. 반면 세 모델 모두 서술어에서 가장 좋은 성능을 보였다. 세 모델이 관형어에서 비교적 낮은 성능을 보이는 것은 개체명이 명사 나열 구성일 때 모델이 이를 판별하는데 어려움을 겪기 때문일 수 있다. 개체명이 관형어로 사용되는 경우는 서술어에서 사용되는 경우에 비해, '<미국:OG><소련:OG>의 관계를'에서의 '미국 소련'처럼 명사가 연이어 오는 구성이 있었는데 이러한 구문의 특성이 영향을 주었을 것으로 생각된다. 반면에 서술어에서 사용된 경우는 그 예가 많지 않은 한계가 있기는 하지만 다음 예시와 같이 개체명 뒤에 '이다'가 연이어 나오는 패턴화된 구성들이 영향을 주었을 것으로 여겨진다.

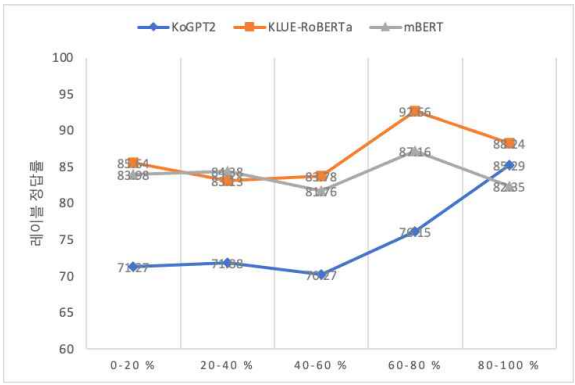
(4) 개체명 '하나'가 서술어 위치인 예시

12월 25일이라고 하셨는데 그 날은 저희 매장이 일년 중 가장 바쁜 날 중에 <하나:QT>입니다.

<Wikitree(204. 1. 20.)>

3.2.3 문장 위치에 따른 교차 태그의 모델별 성능

사전 학습 모델의 구조가 교차 태그 인식에 미치는 영향을 관찰하기 위해, 문장 내에서 등장하는 개체명 토큰의 상대 위치에 따른 모델별 레이블 정답률을 [그림 1]에서 비교해 보았다.



[그림 1] 교차 태그 포함 개체명의 문장 내 상대위치에 따른 레이블 정답률

mBERT와 KLUE-RoBERTa는 교차 태그 개체명의 상대 위치에 비교적 영향을 덜 받는 경향을 보인다. Transformer encoder로 구성된 두 사전 학습 모델의 경우, 양방향 인코딩을 통해 각 토큰의 임베딩 벡터를 만들기 때문에 어떤 위치에 있는 개체명이라도 문맥의 의미가 고려될 가능성이 높다.

반면에, KoGPT2는 개체명의 위치가 문장 내에서 끝에 위치할 때 가장 좋은 성능을 나타냈다. Transformer decoder 기반의 KoGPT2는 문맥 정보가 단방향으로 점차 누적되는 형태를 통해 토큰을 임베딩한다. 이런 형태의 임베딩은 전체 문장의 의미를 함유한 문맥 벡터(context vector)를 구하는 데는 특화되어 있으나, 특정 위치에 존재하여 문장의 일부분만을 활용할 수 밖에 없는 개체명 인식에서는 취약할 수 있다. 이 부분에 대해서는 decoder와 encoder의 임베딩 과정의 차이가 인식기에 미치는 영향에 대한 보다 정교한 후속 연구를 통해 밝혀질 수 있을 것이다.

4. 결론

본 논문에서는 웹 기반 뉴스 말뭉치인 위키트리와 네이버 영화 댓글 말뭉치인 NSMC를 토대로 형태는 같지만 태깅된 개체명의 범주는 달라지는 어휘들을 분석하고 그에 따라 개체명 인식기에 성능 차이가 있는지 살펴 보았다. 개체명 인식기는 Transformer encoder 기반의 다국어 모델인 BERT-base-multilingual-cased (mBERT), 한국어로 학습된 KLUE-RoBERTa-base 모델, 그리고 Transformer decoder 기반의 KoGPT2-base-v2를 사용하였다. 모델별 성능 평가에서는 다국어 모델인 mBERT와 비교하여 한국어로 학습한 KLUE-RoBERTa가 더 좋은 개체명 인식 결과를 보여주었으며, 동일한 Transformer encoder 모델인 mBERT가 Transformer decoder 모델인 KoGPT2보다 높은 성능을 나타냈다. 교차 태깅된 개체명의 성능을 레이블과 스팬 정답률, 문장 성분에 따른 정답률, 문장 위치에 따른 정답률로 나누어 보면 먼저 레이블 정답률의 경우 KLUE-RoBERTa는 85.66%, mBERT는 84.08%로 근소한 차이를 보였으나, KoGPT2의 경우 72.54%로 다른 모델에 비해 11% 이상 낮은 정답률을 보였다. KLUE-RoBERTa와 mBERT의 경우는 스팬 정답률이 레이블 정답률보다 높은 경향을 보이는데 mBERT가 89.55%, KLUE-RoBERTa가 89.19%로 mBERT의 성능이 근소하게 높았다. KoGPT2 인식기는 스팬 정답율이 55.41%로 레이블 정답률보다 20% 가까이 낮아지는 경향을 보였다. 두 번째로 문장 성분에 따른 교차 태그 범주별 성능에서는 세 모델 모두 관형어에서 다소 낮은 성능을 보였고 서술어에서 가장 좋은 성능을 보였다. 세 번째로 문장 위치에 따른 정답률에서 mBERT와 KLUE-RoBERTa는 교차 태그 개체명의 상대 위치에 비교적 영향을 덜 받았지만 KoGPT2는 개체명의 위치가 문장 내에서 끝에 위치할 때 가장 좋은 성능을 나타냈다. 문장 종결 위치에서 성능이 좋은 것은 실험에 사용된 말뭉치가 서술어에서 명사의 중첩이 적고 구문이 패턴화되어 있다는 특징과 KoGPT2가 decoder기반의 모델이기 때문으로 여겨진다.

이후에 의미 모호성에 대한 대규모 주석 말뭉치나 법률이나 의료 분야 말뭉치와 같은 특수 분야 말뭉치에서 어느 정도 의미 모호성을 지니는 어휘들이 등장하는 지 등을 정밀히 살펴본 후 본 논문에서 살펴본 인식기의 특성을 중심으로 개체명 인식기의 성능 개선을 이루어나가는 후속 연구가 이어질 필요가 있을 것이다.

참고문헌

[1] 유현조, 정유남, 송영숙, 김민수, 윤기현 (2021). 딥러닝 기반 한국어 개체명 인식의 평가와 오류 분석 연구, 한국언어학회 46-3 (발간 예정).

[2] 이호경, 안재현, 윤정민, 배경만, 고영중. "위키피디아 기반의 효과적인 개체 링킹을 위한 NIL 개체 인식과 개체 연결 중의성 해소 방법." 정보과학회논문지.

[3] 민진우, 나승훈, 김현호, 김선훈, 강인호 (2020). NIL-Aware KnowBERT를 이용한 개체 연결 모델의 성능 향상. 한국정보과학회 학술발표논문집, 284-286.

[4] 김서현, 서영덕, 백두권 (2016). 개체 중의성 해소를 위한 사용자 유사도 기반의 트윗 개체 링킹 기법, 정보과학회논문지, 43(9), 1043-1051.

[5] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv:1907.11692, 2019.

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[8] Park, Sungjoon and Moon, Jihyung and Kim, Sungdong and Cho, Won Ik and Han, Jiyoon and Park, Jangwon and Song, Chisung and Kim, Junseong and Song, Yongsook and Oh, Taehwan and Lee, Joohong and Oh, Juhyun and Lyu, Sungwon and Jeong, Younghoon and Lee, Inkwon and, Seo, Sangwoo and Lee, Dongjun and Kim, Hyunwoo and Lee, Myeonghwa and Jang, Seongbo and Do, Seungwon and Kim,

Sunkyoung and Lim, Kyungtae and Lee, Jongwon and Park, Kyumin and Shin, Jamin and Kim, Seonghyun and Park, Lucy and Oh, Alice and Ha, Jung-Woo and Cho, Kyunghyun. (2021). KLUE: Korean Language Understanding Evaluation. arXiv:2105.09680v3.

[9] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subwordunits. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers), pages 1715-1725, Berlin, Germany, August 2016. Association for Computational Linguistics.