Journal of KIIT. Vol. 23, No. 9, pp. 65-74, Sep. 30, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.9.65

사전학습 언어모델 기반 한국어 개체명 인식 성능 비교 연구

김선겸*[1], 원지선*[2], 신재영*[3]

A Comparative Study on Korean Named Entity Recognition using Pretrained Language Models

Sun-Kyum Kim*[1], Jisun Won*[2], and Jaeyoung Shin*[3]

본 연구는 과학기술정보통신부 한국건설기술연구원 연구운영비지원(주요사업)사업으로 수행되었습니다
(20250211-001, 미래 건설산업 견인 및 신시장 창출을 위한 스마트 건설기술 연구)

요 약

본 연구는 KLUE-RoBERTa와 KoELECTRA 한국어 사전학습 언어모델을 활용하여, 2020~2022년 '모두의 말뭉치' 개체명 분석 데이터를 기반으로 연도별 NER 성능을 비교·분석하였다. 두 모델은 동일한 전처리 절차, 토크나이저, 학습 구조, 평가 지표를 적용하여 공정성을 확보하였다. 분석 결과, KLUE-RoBERTa는 평균 F1-score에서 KoELECTRA보다 10~15% 높은 성능을 보였으며, 인명, 기관명, 날짜, 수량 등 주요 라벨에서 Precision과 Recall이 균형 있게 나타났다. 반면, KoELECTRA는 형식이 일정한 수치·수량 라벨에서 높은 Recall을 기록했으나, 문맥 의존도가 높은 라벨에서는 Precision이 낮아 전체 F1-score가 떨어졌다. 또한, 연도별 데이터셋의 장르 분포와 라벨 다양성, 주석 불일치 사례가 성능 차이에 직접적인 영향을 주었음을 확인하였다.

Abstract

This study compares and analyzes the yearly NER performance of KLUE-RoBERTa and KoELECTRA Korean pre-trained language models using the named entity analysis data of the "Modu Corpus" from 2020 to 2022. To ensure fairness, both models were applied with identical preprocessing procedures, tokenizers, training architectures, and evaluation metrics. The analysis showed that KLUE-RoBERTa achieved 10–15 percentage points higher average F1-scores than KoELECTRA, with balanced Precision and Recall across major labels such as person names, institution names, dates, and quantities. In contrast, KoELECTRA recorded high Recall for numerals and quantity-related labels with fixed formats, but exhibited lower Precision for labels with high contextual dependency, leading to a lower overall F1-score. In addition, it was confirmed that the genre distribution, label diversity, and annotation inconsistencies of the yearly datasets had a direct impact on performance differences.

Keywords

Korean named entity recognition, pretrained language models, KLUE-RoBERTa, KoELECTRA, Modu corpus

* 한국건설기술연구원 연구원(*[1] 교신저자)
- ORCID[1]: http://orcid.org/0000-0002-4610-1961
- ORCID[2]: http://orcid.org/0000-0002-3690-8470
- ORCID[3]: http://orcid.org/0000-0002-5917-0472

· Received: Jul. 17, 2025, Revised: Aug. 18, 2025, Accepted: Aug. 21, 2025
· Corresponding Author: Sun-Kyum Kim
Dept. of Future & Smart Construction Research
Korea Institute of Civil Engineering and Building Technology
Tel.: +82-31-995-0962, Email: sunkyumkim@kict.re.kr

Ⅰ. 서 론

최근 자연어 처리(NLP, Natural Language Processing) 분야에서는 문서 요약, 질의응답, 기계번역 등 다양한 과제에서 개체명 인식(NER, Named Entity Recognition) 기술이 핵심 기반 기술로 주목받고 있다. 특히 한국어는 불규칙한 띄어쓰기, 조사 및 어미의 다양성 등으로 인해 개체의 경계를 판단하는 것이 매우 어려우며, 이에 따라 고성능의 NER 시스템을 구축하기 위한 연구가 활발히 이루어지고 있다.

기존 NER 시스템은 사전 기반 접근, 규칙 기반 기법, 통계적 학습 모델에 의존하였으나, 이러한 방법들은 문맥 정보의 제한성과 확장성 부족으로 인해 실제 응용에 제약이 있었다. Transformer 모델의 등장 이후, 사전학습 언어모델(Pretrained Language Model, PLM)을 활용한 NER 기법이 주류로 자리잡았으며[1], 특히 KLUE-RoBERTa와 KoELECTRA는 한국어에 특화된 대표적 PLM으로 활용되고 있다. KLUE-RoBERTa는 한국어 말뭉치를 기반으로 RoBERTa 구조를 따르는 정제된 문장 예측 학습을 수행한 모델이고[2], KoELECTRA는 ELECTRA 구조에 기반하여 Generator와 Discriminator의 대립적 학습을 통해 효율적인 학습과 경량화된 구조를 특징으로 한다[3]. 두 모델은 각각 KLUE benchmark와 AI-Hub 과제를 통해 공개되었으며, 실제 한국어 NER 실험에서 높은 성능을 보여주었다[4][5].

한편, 국립국어원에서는 한국어 언어자원의 공공성과 품질 향상을 위해 대규모 언어 자료인 ‘모두의 말뭉치(Modu Corpus)’를 지속적으로 구축하고 있으며, 그 중 개체명 분석 말뭉치는 한국어 고유 개체명 분류 체계를 반영한 고품질의 주석 데이터셋으로 평가받고 있다[6]. 그러나 이 말뭉치를 활용하여 동일 조건 하에서 다양한 PLM 기반 모델 성능을 비교한 연구는 아직 부족한 실정이다.

이에 본 연구는 KLUE-RoBERTa와 KoELECTRA 두 가지 한국어 PLM을 활용하여, 2020~2022년 ‘모두의 말뭉치’ 개체명 분석 데이터를 기반으로 한 연도별 NER 성능 비교 실험을 수행하였다. 두 모델은 동일한 전처리, 토크나이저, 학습 구조 및 평가 기준을 적용함으로써, 모델 구조의 차이와 말뭉치의 특성이 NER 성능에 미치는 영향을 정량적으로 분석하였다. 이를 위해 한국어 NER 프레임워크인 pytorch-ko-ner[7]를 기반으로 하여, 개체명 분석 데이터에 맞춰 전처리, 라벨 처리, 평가 방식 등을 연구 목적에 맞게 수정·보완하여 실험을 수행하였다.

분석 결과, KLUE-RoBERTa는 KoELECTRA와 비교해 전반적으로 더 높은 F1-score를 나타냈으며, 인명, 기관명, 날짜, 수량, 수치 표현 등 다양한 라벨에서 안정적인 성능을 보였다. 반면, KoELECTRA는 형식이 일정한 라벨에서는 Recall이 높게 나타났으나, Precision이 낮아 전체 F1-score 에서는 낮은 성능을 보였다. 두 모델의 성능 차이는 라벨 특성과 구조적 차이를 반영하는 F1-score 분포 및 Precision - Recall 분석을 통해 확인할 수 있었다.

본 연구는 이러한 분석 결과를 바탕으로 향후 도메인 특화 NER 모델 개발 및 라벨 기반 데이터 개선 전략 수립에 유용한 기초 자료로 활용될 수 있으며, 동일한 실험 조건에서 모델 구조 차이와 연도별 말뭉치 특성이 성능에 미치는 영향을 실증적으로 분석한 최초의 시도로서 학술적 의의를 함께 지닌다.

본 논문은 다음과 같은 구성으로 이루어진다. 제2장에서는 관련 연구를 다루고, 제3장에서는 데이터 분석과 실험 설계를 한다. 제4장에서는 실험 결과 및 분석 내용을 제시하며, 마지막으로 제5장에서 본 연구의 결론과 향후 연구 방향을 논의한다.

Ⅱ. 관련 연구

2.1 통계 기반 및 기계학습 기반 NER

2000년대 이후, NER 기술에서는 HMM(Hidden Markov Model)이나 CRF(Conditional Random Fields)와 같은 통계적 기계학습 기법이 널리 활용되어 왔다. 이러한 기법들은 단어 주변의 n-gram 정보, 품사, 형태소 분석 결과 등을 특징으로하여 개체명과 일반 단어 사이의 경계를 학습하였다. 특히 CRF는 BIO 태그 기반 NER 작업에 효과적인 시퀀스 라벨링 기법으로 자주 활용되었으나, 긴 문맥을 반영하기에는 한계가 존재하였다[8].

2.2 PLM

2018년 BERT(Bidirectional Encoder Representations from Transformers)[9]의 등장 이후, NER을 포함한 대부분의 자연어 처리 과제에서는 PLM이 새로운 표준이 되었다[1]. Transformer 기반 구조는 문맥의 양방향 정보를 반영하며, 별도의 태스크 특화 학습 없이도 높은 성능을 발휘하였다. 이후 다양한 변형 모델들이 등장하였으며, 한국어에 특화된 PLM 모델 개발이 활발히 진행되었다[10]. 예를 들어 KLUE-RoBERTa는 대규모 한국어 위키, 뉴스, 웹 데이터를 기반으로 학습된 모델로 Masked Language Modeling(MLM)을 학습 목표로 하여 문장의 일부 토큰을 가린 후 이를 복원하는 방식으로 문맥 정보를 학습하며, RoBERTa 구조에 기반한 문장 예측에 강점이 있다[2]. 또한 KoELECTRA는 ELECTRA 프레임워크를 한국어에 적용한 모델로, Generator가 아닌 Discriminator가 판별하는 RTD(Replaced Token Detection) 방식을 채택하여 학습을 수행하며, 상대적으로 적은 파라미터 수로 빠른 수렴 속도가 특징이다[3].

2.3 한국어 개체명 말뭉치 활용 한계

국립국어원은 2020년부터 '모두의 말뭉치'라는 한국어 언어자원을 구축해 왔다. 이 중 개체명 분석 데이터는 매년 주석 범위를 확장해 다수의 라벨을 포함하는 구조로 발전해 오고 있다. 이외에도, 한국어 개체명 데이터는 Naver NER 데이터셋, AI Hub에서 제공하는 한국어 말뭉치 등이 있으며, 다양한 PLM의 성능 평가를 위해 활용되고 있다. 이러한 데이터를 바탕으로, Yang(2021)의 리뷰 연구[11]에서는 KLUE, KorBERT, KoELECTRA 등 다양한 PLM 모델의 성능을 비교 분석하였다. 그러나 기존 연구들은 대체로 단일 연도 데이터 또는 고정된 장르 기반의 성능 비교에만 제한되어 있는 경우가 많으며, 말뭉치 자체의 구성 차이, 라벨 불균형[12], 주석 일관성 변화[13] 등이 모델 성능에 미치는 영향을 분석한 연구는 드문 상황이다. 따라서 보다 신뢰할 수 있는 모델 성능 비교를 위하여 말뭉치 구조와 모델 간의 특성을 종합적으로 분석하는 접근이 필요하다.

III. 데이터 분석 및 실험 설계

3.1 데이터 분석

본 연구에서 활용한 '모두의 말뭉치' 개체명 분석 데이터는 문장 단위의 JSON 형식으로 구성되어 있으며, 각 문장에는 개체명과 해당 범주가 위치 정보(begin, end)와 함께 주석되어 있다. BIO 태그 형식은 포함되어 있지 않아, 추후 전처리 과정에서 별도로 수행해야 한다. 이 데이터는 표 1과 같이 2020년부터 2022년까지 구축된 세 개 연도의 말뭉치를 활용하였으며, 연도별로 문장 수, 장르 구성, 라벨 분포 등에서 차이를 가진다. 특히 2021년과 2022년 데이터에는 신문 기사, 공적 대화, 온라인 채팅 등 다양한 양식의 문서가 혼합되어 있어, 문체 및 표현 방식 측면에서도 상이한 특성을 가진다. 이러한 연도별 장르 구성 차이는 모델이 문맥을 해석하고 라벨을 분류하는 데 있어 직접적인 영향을 미칠 수 있다.

표 1. 연도별 데이터셋
Table 1 Dataset by year

Year	Number of sentences	Main content
2020	~665,000	SNS, blogs, etc.
2021	~161,000	Newspapers, lectures, broadcast scripts, etc.
2022	~315,000	Newspapers, daily conversations, chat messages, etc.

전체 데이터셋에는 PS(인명), OG(조직명), LC(지명) 외에도 DT(날짜), EV(이벤트), CV(속성), QT(수량), AM(단위 속성) 등 총 150여 개의 세분화된 라벨이 포함되어 있으며, 라벨별 분포에는 큰 불균형이 존재한다. CV(21.27%), QT(11.79%), PS(10.47%), DT(10.45%) 등 일부 고빈도 라벨이 전체의 과반 이상을 차지하는 반면, FD(0.51%), TMM(0.05%), PT(0.04%) 등은 차지하는 비중이 낮아, 학습 시 특정 라벨에 과도하게 편중되거나 저빈도 라벨의 예

측이 불안정할 가능성이 있다. 특히 TMM(교통 및 시설명)이나 PT(특허)처럼 전체 샘플 수가 수백 건에 불과한 라벨은 모델 학습 시 유의미한 특징 학습이 어려워 낮은 성능을 가질 수 있다.

또한 중복 문장은 전체 데이터의 약 1.6%에서 4.3%를 차지하였으며, 연도 간 중복 문장 수도 600건 미만으로 나타났다. 이러한 중복은 전체 성능에 미치는 영향은 크지 않을 수 있으나, 특정 표현의 과도한 반복 학습으로 인해 과적합을 유발하거나 오탐률을 높이는 요소로 작용할 가능성이 있다.

한편, 데이터셋 내부에는 동일한 표현에 대해 서로 다른 라벨이 부여된 사례도 일부 확인되었다. 예를 들어 숫자 표현인 '1.'이 QT_COUNT로, 다른 문맥에서는 QT_ORDER로 주석되어 일관되지 않은 경우를 확인할 수 있었다. 이러한 라벨 혼용과 태깅 기준의 불일치는 모델의 예측 불안정을 발생시키며, 특히 Discriminator 기반 구조를 사용하는 KoELECTRA에서는 Recall은 높게 유지되더라도 Precision은 낮아지는 원인이 될 수 있다. 따라서 본 연구에서는 데이터셋의 구조적 한계를 고려한 실험 설계와 결과 해석을 전제로 진행되었다.

3.2 실험 설계

본 연구에서는 국립국어원이 구축한 '모두의 말뭉치'의 2020년부터 2022년까지 총 3개 연도에 걸친 개체명 분석 데이터를 실험에 활용하였다. 각 연도의 데이터는 문장 수, 장르 분포, 개체명 라벨 구성 등에서 연도별 편차가 존재하며, 특히 2022년 데이터는 상대적으로 균형 잡힌 클래스 분포를 가진다.

전체 파이프라인은 표 2와 같이 JSON 처리, 전처리, 인코딩, 학습, 추론으로 구성된다. 데이터는 JSON 구조를 pandas 기반의 데이터프레임으로 변환한 뒤, 문장을 단어보다 작은 의미 단위의 조각으로 분해하고, 각 조각에 BIO 라벨을 대응시켰다. 이 과정에서 각 조각이 원래 문장의 어떤 단어에 해당하는지를 추적하는 정보를 활용하여 정렬하였으며, HuggingFace Tokenizer를 활용하여 WordPiece 단위로 토큰화하였다. 또한, 전체 데이터는 train/validation/test로 64:16:20 비율로 나누었으며, 라벨 불균형을 완화하기 위해 stratified sampling을 적용하였다. 모델 학습에는 대표적인 한국어 PLM인 KLUE-RoBERTa와 KoELECTRA를 사용하였다.

표 2. 단계별 기능
Table 2. Step-by-step functions

Step	Filename	Main Role
1	json_to_df	- Convert JSON → DataFrame - Integrate sentence IDs, raw text, and NER annotation info
2	preprocess	- Split the dataset into training and test sets
3	encoding	- Tokenize sentences with a tokenizer - Align BIO tags with tokens
4	hf_trainer	- Train KLUE-RoBERTa and KoELECTRA models - Compute evaluation metrics
5	inference	- Run inference on the test set - Save predictions and produce final metrics

실험 구성은 2020~2022년 3개 연도별 데이터셋 각각에 대해 KLUE-RoBERTa 및 KoELECTRA 모델을 적용하는 방식으로 이루어져, 총 6가지 실험을 수행하였으며, 이 결과를 바탕으로 모델 구조와 데이터 특성 간의 상호작용을 분석하였다. 또한, 보조 실험으로 3개년 데이터를 통합한 단일 데이터셋을 구성하고 동일한 두 모델을 적용하여, 데이터 출처가 혼합된 상황에서의 모델의 일반화된 성능도 함께 평가하였다. 이와 같은 실험 설계는 다양한 데이터 구성과 모델 구조 간의 상호작용을 실증적으로 비교할 수 있다.

표 3과 같이 실험에 사용한 두 모델 모두 HuggingFace Transformers 라이브러리를 기반으로 PyTorch 환경에서 구현되었으며, 학습 환경은 사용 가능한 GPU 메모리와 모델 안정성을 고려해 구성하였으며, 모델 구조 간 성능을 공정하게 비교하기 위해 동일한 하이퍼파라미터를 적용하였다. 각 하이퍼파라미터 값은 GPU 환경에서의 학습 안정성과 과적합 방지를 고려해 재현성과 효율성을 확보할 수 있는 최대 범위 내에서 설정하였다. 학습 환경은

사용 가능한 GPU 메모리와 모델 안정성을 고려해 구성되었으며, Batch size, Learning rate, Epoch 수 등은 실험의 재현성과 효율성을 함께 반영하여 설정되었다. 또한, 과적합을 방지하기 위해 Early Stopping 조건을 적용하였다. 성능 평가는 Precision (정밀도), Recall(재현율), F1-score(조화 평균)을 기준으로 진행되었으며, 전체 및 개별 라벨 기준으로 병렬 분석을 수행하였다.

표 3. 실험 환경
Table 3. Experimental environment

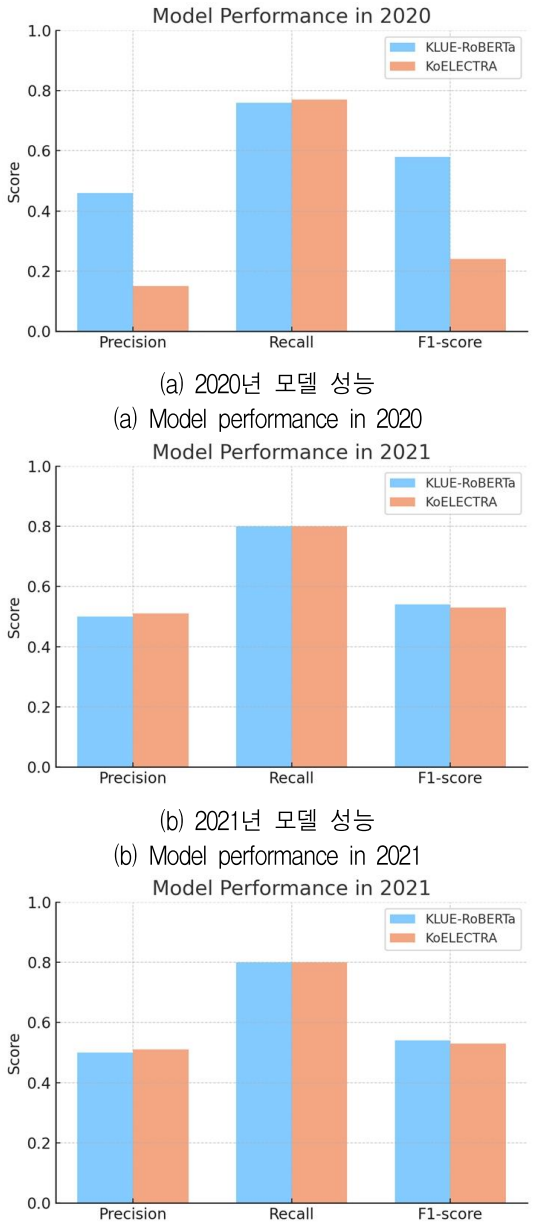
Item	Setting value
Optimizer	AdamW
Batch size	16
Epochs	10
Learning rate	5e-5
Tokenizer	token-label alignment
Metric	Precision, Recall, F1-score

Ⅳ. 실험 결과 및 분석

본 장에서는 KLUE-RoBERTa와 KoELECTRA 모델을 대상으로 2020~2022년 '모두의 말뭉치' NER 데이터를 활용한 실험 결과를 바탕으로 성능 비교 및 정량적 분석을 수행하였다. 분석은 연도별, 전체 라벨 기준, Precision - Recall 특성, 성능 분포 등을 중심으로 구성되며, 각 결과에 대해 데이터셋의 특성과 모델 구조를 종합적으로 고려하여 해석하였다.

4.1 연도별 모델 성능 비교

그림 1의 (a)~(c)는 2020년부터 2022년까지 연도별 실험 결과를 Precision, Recall, F1-score 지표 기준으로 비교한 결과이다. 각 연도에서 KLUE-RoBERTa는 KoELECTRA 대비 전반적으로 더 높은 성능을 보였다. 특히 Precision 측면에서 모든 연도에서 일관된 우위를 보였으며, F1-score 기준으로도 2020년과 2022년에서는 뚜렷한 차이가 확인되었다. Welch's t-검정 결과, 2020년과 2022년에서는 KLUE-RoBERTa가 KoELECTRA보다 통계적으로 유의미하게 높은 성능을 보였으나($p < 0.05$), 2021년에서는 두 모델 간 평균 차이가 통계적으로 유의하지 않았다($p > 0.05$).



(a) 2020년 모델 성능
(a) Model performance in 2020
(b) 2021년 모델 성능
(b) Model performance in 2021
(c) 2022년 모델 성능
(c) Model performance in 2022
그림 1. 연도별 모델 성능 비교
Fig. 1. Comparison of model performance by year

2020년은 SNS 및 블로그 기반의 비정형 문장이 포함되어 있어 모델의 예측 안정성에 영향을 미친 것으로 보인다. KoELECTRA는 Recall이 0.77로 높은 값을 보유했지만 Precision은 0.15로 과도한 오탐이 발생하는 것을 확인하였다. 이는 중복 문장과 태깅 불일치 등 데이터 품질 문제와 맞물려 KoELECTRA의 예측 혼란을 야기시킨 것으로 보인다. 반면에 KLUE-RoBERTa는 Precision 0.46, Recall 0.76으로 비

교적 균형 잡힌 성능을 보였다.

2021년은 신문과 공적 대화 등 문어체와 구어체가 혼합된 구성으로, 두 모델의 F1-score는 각각 0.53, 0.54 수준으로 유사하게 나타났다. 다만 Precision과 Recall 간 균형 면에서 KLUE-RoBERTa가 안정적인 성능을 유지하였다.

2022년은 2021년과 같이 신문, 온라인 채팅 및 일상대화 등 문어체와 구어체가 혼합된 데이터로 구성되어, KLUE-RoBERTa는 Precision 0.66, Recall 0.84로 2020, 2021년과 비교하여 가장 우수한 성능을 보였다. 반면 KoELECTRA는 Recall 0.8을 상회하였으나 Precision이 0.50에 머무르며 예측 정확도 측면에서 상대적 약점을 드러냈다.

4.2 전체 라벨별 F1-score 분석

모델별 F1-score 기준으로 전체 라벨의 성능을 분석한 결과, 그림 2와 같이 KLUE-RoBERTa는 대부분의 라벨에서 KoELECTRA 대비 전반적으로 높은 성능을 보였다. 특히 DT, LCP, QT, TMM, LCG, LC 등에서는 F1-score 0.5 이상을 기록하며 뚜렷한 우위를 나타냈다. 이러한 결과는 문맥 정보를 보다 정밀하게 반영하는 양방향 Transformer 구조의 특성으로 인해 라벨 전반에 걸쳐 예측 분산이 완만하고 평균 F1-score도 높은 경향을 보여 안정적인 성능을 유지할 수 있었다. 반면 KoELECTRA는 일부 라벨에서 유사한 수준의 F1-score를 보였고, 다수 라벨에서는 0.4 미만에 머물러 전반적으로 성능이 낮았다.

이러한 격차는 모델 구조의 차이뿐만 아니라, 앞서 확인된 데이터셋의 특성이 복합적으로 작용한 결과로 해석된다. CV, QT 등 상위 소수 라벨에 편중된 분포, TMM, PT 등 희소 라벨의 존재, 그리고 동일 표현에 상이한 라벨이 부여된 주석 불일치 현상은 KoELECTRA의 예측 일관성의 저해의 원인으로 작용했다. 예컨데 '1.'이 QT_COUNT 또는 QT_ORDER로 서로 다르게 주석된 사례는 KoELECTRA에서 오탐을 증가시켰고, 결과적으로 낮은 F1-score로 이어졌다.

결론적으로, 이러한 라벨별 성능 차이는 모델 구조의 특성에서 비롯된 것으로 해석된다. KLUE-RoBERTa는 문맥 전반을 활용하는 양방향 Transformer 구조 덕분에 복합 구조의 개체에서도 안정적인 성능을 보였으나, KoELECTRA는 Discriminator 중심의 단일 토큰 판단 방식에 치우쳐 Precision 손실이 상대적으로 크게 나타났다.

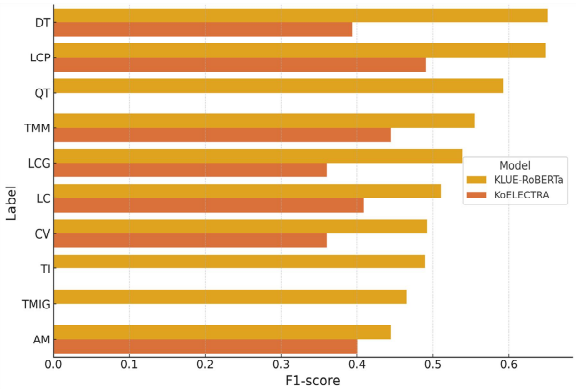


그림 2. 상위 10개 라벨 F1-score 결과
Fig. 2. Top-10 Label F1-score by model

4.3 라벨별 성능 분포 및 모델 안정성

전체 라벨에 대한 F1-score 분포를 분석한 결과, 그림 3과 같이 KLUE-RoBERTa는 대부분의 라벨이 F1-score 0.6 이상 구간에 집중되어 있었으며, 전반적으로 고르고 안정된 성능을 보였다. 반면 KoELECTRA는 성능 분포가 상대적으로 넓게 퍼져 있었고, 특히 0.1~0.4 구간에 집중된 라벨이 많아 저성능 라벨의 비중이 높은 양상을 나타냈다. 라벨별 F1-score를 대상으로 Welch's t-검정을 실시한 결과, $p = 0.104$로 통계적 유의수준($p < 0.05$)을 만족하지 않았다. 그러나 QT, DT, LC 등 일부 라벨에서는 수치적으로 뚜렷한 성능 차이가 확인되었다. 이는 모델 구조의 특성에 따라 특정 라벨에서 성능 민감도가 다르게 나타날 수 있음을 나타낸다.

KoELECTRA는 일부 라벨에서 0.5 이상의 F1-score를 기록하며 일정 수준의 성능을 보였지만, 전반적으로 성능 편차가 크고 불안정한 경향이 확인되었다. 이러한 결과는 KoELECTRA가 라벨 수가 적거나 주석이 일관되지 않은 경우에 성능이 쉽게 떨어진다는 것을 보여준다. 전체 라벨 중 CV, QT, PS 등 고빈도 라벨이 전체의 과반 이상을 차지하고, TMM, PT 등은 매우 희소한 분포를 보여 라벨 간 학습량 격차로 이어졌다.

한편, KLUE-RoBERTa는 라벨의 불균형이나 주석 상의 오류와 같은 데이터의 구조적 한계에도 비교적 영향을 덜 받았으며, 전반적으로 예측 결과의 신뢰도가 높고 분산도 낮은 안정적인 성능을 유지했다. 특히 2022년 데이터처럼 장르와 문체가 다양함에도 높은 일관성을 보일 수 있었으며, 이는 이 모델이 양방향으로 문맥 정보를 처리할 수 있기 때문에 다양한 라벨 유형과 문장 구성을 해석할 수 있었던 것으로 보인다.

특히 희소 라벨이나 의미가 모호한 개체의 경우, 문장 전체 흐름을 고려하는 RoBERTa 구조가 더 효과적이었다. 반면 KoELECTRA는 개별 토큰의 교체 여부를 중심으로 학습하기 때문에, 문맥 정보가 분산된 라벨에서는 오탐과 누락이 잦았다.

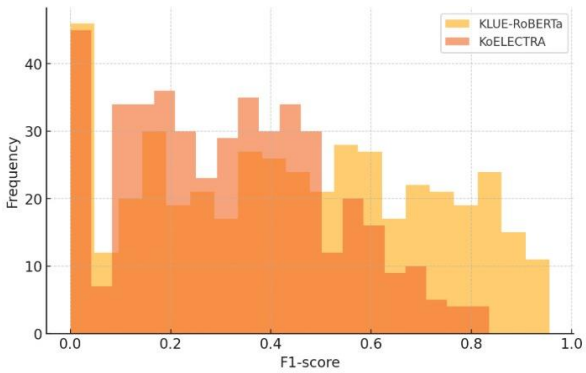


그림 3. 모델별 F1-score 분포
Fig. 3. Distribution of F1-scores by model

4.4 모델 간 라벨 그룹별 F1-score 격차 분석

KLUE-RoBERTa와 KoELECTRA 간 F1-score 차이가 크게 나타난 전체 라벨 상위 15개를 분석한 결과, 그림 4와 같이 대부분의 라벨에서 KLUE-RoBERTa가 KoELECTRA보다 우수한 성능을 보였다. 특히 OGG, AM, TM, TMI, TR 등의 라벨에서 두 모델 간 F1-score 격차는 0.15~0.30 수준으로 크게 나타났으며, 이는 RoBERTa 구조가 복합 명사 구조나 문맥 의존도가 높은 개체 표현에서 보다 안정적인 예측을 수행했기 때문으로 보인다.

KoELECTRA는 QT, DT, TI처럼 수치정보를 나타내는 비교적 형식이 고정된 라벨에서 KLUE-RoBERTa에 근접한 성능을 보이긴 했으나, 전체 라벨 중 어느 항목에서도 우위를 점하지는 못하였다. 특히 정형 표현에 반응해 높은 Recall을 기록하더라도, 라벨 경계를 정확하게 구분하지 못하거나 유사 표현 간 오류가 발생하면서 Precision은 상대적으로 낮게 나타났다.

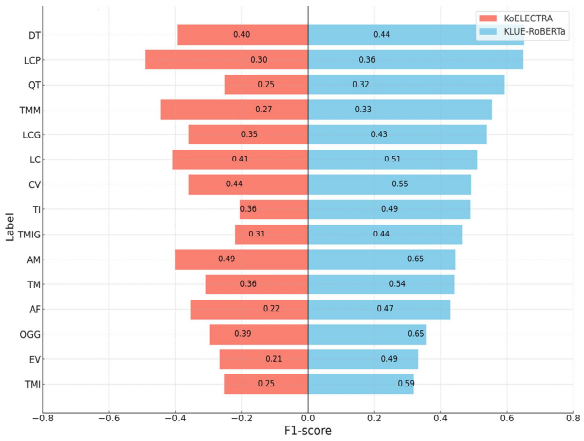


그림 4. 모델별 F1-score 기준 상위 15개 라벨 그룹
Fig. 4. Top-15 label by model F1-score

이러한 격차는 데이터셋의 태깅 불일치나 라벨 혼용 문제에 의해 더욱 심화되었다. 동일한 숫자 표현인 '1.'이 문맥에 따라 QT_COUNT 또는 QT_ORDER로 주석되는 경우처럼, 주석 기준의 일관성이 부족한 항목은 KoELECTRA의 예측 혼란을 야기해 성능 격차를 확대시키는 요인으로 작용하였다. 반면 KLUE-RoBERTa는 문맥 기반의 양방향 구조를 통해 이러한 라벨 간 의미 차이를 상대적으로 더 정확하게 판별하였다. 종합적으로, KLUE-RoBERTa는 전체 라벨에 걸쳐 보다 균형 잡힌 예측 성능을 유지한 반면, KoELECTRA는 일부 정형 표현에만 제한적으로 근접한 성능을 보이며 구조적 한계가 뚜렷하게 드러났다.

종합하면, KLUE-RoBERTa의 복합 표현 및 다의어 라벨에서의 강세는 문장 전체 맥락을 활용하는 양방향 학습 구조의 결과로 볼 수 있다. 반면 KoELECTRA는 형식이 고정된 단일 라벨에는 강점을 보였으나, 문맥 해석이 요구되는 복잡한 표현에서는 라벨 경계 인식이 흔들리며 예측 정확도가 낮았다.

4.5 Precision vs Recall 산점도 분석

전체 라벨별 Precision과 Recall의 관계를 산점도로 분석한 결과, 그림 5와 같이 두 모델은 상이한 성능 특성을 보였다. KLUE-RoBERTa는 대부분의 라벨에서 Precision과 Recall이 균형 있게 분포되었으며, 특히 DT, LCP, QT 등 주요 라벨에서 Precision 0.5 이상, Recall 0.7 이상을 기록하였다. 이는 해당 모델이 다양한 유형의 개체에 대해 정밀성과 재현율을 고르게 확보하고 있음을 보여준다.

KoELECTRA는 일부 라벨에서 높은 Recall을 기록했지만, 전반적으로 Precision은 낮은 경향을 보였다. 예를 들어, DT, AM, LC, LCG 등에서는 Recall이 0.7 이상으로 우수했음에도 Precision은 0.3대를 넘지 못해 오탐이 빈번하게 발생하였다. 이는 모델이 개체를 과도하게 탐지하거나, 정확한 라벨 예측보다는 탐지 범위에 치중되었다는 것을 의미한다.

FD, AFA 및 AFW, TR 등에서는 두 모델 모두 Precision과 Recall이 낮은 값을 보였고, 이는 개체 경계가 불분명하거나 학습 샘플이 희소했던 라벨의 특성과 관련이 깊다. 앞서 분석한 바와 같이, 전체 데이터에서 TMM은 0.05%, PT는 0.04% 수준에 불과하며, 일부 표현은 중복되거나 서로 다른 라벨로 주석되어 라벨 구분의 혼란을 야기하였다. 또한 숫자 표현 '1.'이 QT_COUNT 또는 QT_ORDER로 혼용된 사례는 Precision 손실에 직접적인 영향을 미친다.

종합적으로, KLUE-RoBERTa는 정밀도와 재현율 모두에서 안정적인 성능을 보이며 다양한 도메인 라벨에 대해 일반화된 인식 능력을 확보한 반면, KoELECTRA는 Recall 중심의 탐지 특성을 보이되 라벨 분별 정확도 측면에서는 구조적 한계를 가진다. NER처럼 단어 경계 인식이 중요한 작업에서는, 문맥 정보를 세밀하게 반영할 수 있는 모델이 Precision과 Recall 모두에서 우수한 성능을 발휘한다.

본 연구에서도 Precision과 Recall의 불균형은 모델의 학습 방식의 차이에서 기인한 것으로 보인다. KoELECTRA는 RTD 방식의 특성으로 인해 탐지 범위를 넓히는 경향이 있어 Recall은 높게 유지되지만, Precision은 상대적으로 낮아지는 한계를 보였다. 반면 RoBERTa는 전체 문맥을 기반으로 판단하기 때문에 불필요한 탐지를 줄이고 Precision을 안정적으로 확보할 수 있었다.

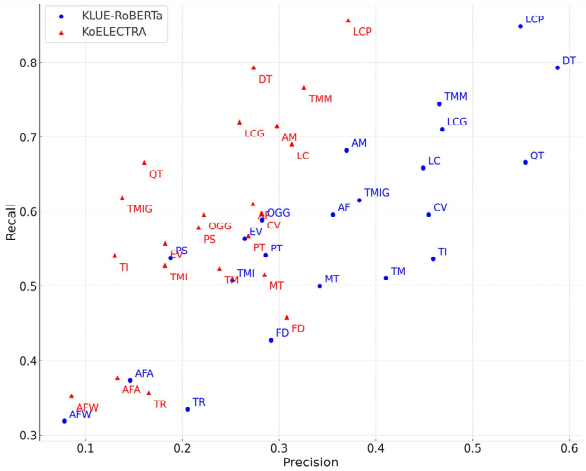


그림 5. 모델별 Precision, Recall 산점도
Fig. 5. Precision vs Recall by label group

V. 결론 및 향후 과제

본 연구는 한국어 NER 성능 비교를 위해, KLUE-RoBERTa와 KoELECTRA 두 PLM을 대상으로 국립국어원의 '모두의 말뭉치' 개체명 분석 데이터를 활용한 실험을 수행하였다. 실험은 2020~2022년까지 3개년의 연도별 데이터셋을 기반으로 구성되었으며, 각 모델의 예측 성능을 Precision, Recall, F1-score 지표를 통해 종합적으로 분석하였다.

KLUE-RoBERTa는 전반적으로 높은 Precision과 안정적인 성능 분포를 보였으며, 특히 문맥적 정보가 중요한 복합 개체명에서 우수한 성능을 나타냈다. 반면 KoELECTRA는 수량, 날짜 등과 같은 정형화된 표현에서 일정 수준의 Recall은 확보했으나, 전반적인 Precision은 낮아 예측 신뢰도 측면에서는 한계를 드러냈다. 이 같은 결과는 RoBERTa의 문맥 이해 방식과 ELECTRA의 판별 기반 구조 간의 차이에서 비롯된 것으로 보인다.

또한 본 연구는 단순한 모델 성능 비교를 넘어서, 데이터셋 자체의 구성 특성—연도별 장르 분포, 라벨 다양성, 주석 불일치가 모델 성능에 미치는 영향을 실증적으로 분석하였다. 예를 들어, 동일 문장에서 숫자 '1.'이나 '2'가 서로 다른 라벨로 주석된 경우에 모델의 예측 정확도 저하에 직접적인 영향을 미치는 요소로 작용하였다.

분석 결과는 한국어 NER 성능을 높이기 위해 단순히 모델 구조만 개선하는 것에 그치지 않고, 데이

터 품질도 함께 관리되어야 함을 보여준다. 특히, 특정 시점이나 특정 장르에 최적화된 모델은 다양한 문맥이나 상황에서 일관된 결과를 내기 어렵기 때문에 보다 다양한 형식의 데이터를 활용한 다각적 평가가 필요하다. 또한, 실제 환경에서 NER 모델을 적용하려면 라벨 체계의 정비와 주석 기준의 표준화, 수치 표현과 반복 표현에 대한 명확한 지침 마련이 필요한 것으로 판단된다.

향후 연구에서는 이러한 분석 결과를 바탕으로 건설 도메인에 최적화된 NER 모델을 개발할 예정이다. 이를 위해 설계보고서, 법규, 시방서, 민원 공문 등 건설 문서의 데이터 특성과 포함 개체명을 분석하고, 공법명·자재명·시설명 등 도메인 특화 개체의 정의 및 주석 기준을 개발할 것이다. 이러한 전처리·라벨링 표준화 과정을 거친 학습 데이터를 활용하여, 두 모델의 장점을 결합한 고정밀 NER 시스템으로 확장함으로써 실제 산업 환경에서의 적용 가능성을 높일 계획이다.

References

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, Vol. 1, pp. 4171-4186, Jun. 2019. https://doi.org/10.18653/v1/N19-1423.

[2] S. Park, S. Kim, H. Kim, and H. Lee, "KLUE: Korean language understanding evaluation", arXiv preprint, arXiv:2105.09680, May 2021. https://arxiv.org/abs/2105.09680.

[3] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators", Proc. of International Conference on Learning Representations (ICLR), Virtual, pp. 1-14, Apr. 2020. https://openreview.net/forum?id=r1xMH1BtvB.

[4] National Institute of Korean Language, "Modu Corpus: Named Entity Annotated Data", Korean Corpus Portal, 2020-2022. https://corpus.korean.go.kr. [accessed: Jun. 01, 2025]

[5] National Information Society Agency, "Korean Named Entity Recognition Data", AI Hub, Apr. 2021. https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=122. [accessed: Jun. 01, 2025]

[6] National Institute of Korean Language, "Modu Corpus", Korean Corpus Portal, https://corpus.korean.go.kr. [accessed: Jun. 01, 2025]

[7] H. Lee, pytorch-ko-ner: Korean Named Entity Recognition using Transformers, https://github.com/ai2-ner-project/pytorch-ko-ner/. [accessed: Jun. 01, 2025]

[8] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", Proc. of the 18th International Conference on Machine Learning (ICML), San FranciscoCAUnited States, pp. 282-289, Jun. 2001. https://dl.acm.org/doi/10.5555/645530.655813.

[9] B. Y. Lin and W. Lu, "Neural Adaptation Layers for Cross-domain Named Entity Recognition", Proc. of Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 2012-2022, Oct.-Nov. 2018. https://doi.org/10.18653/v1/D18-1226.

[10] J. Jang, J. Min, and H. Noh, "KorPatELECTRA: A Pre-trained Language Model for Korean Patent Literature to Improve Performance", Journal of the Korea Society of Computer and Information, Vol. 27, No. 2, pp. 15-23, Feb. 2022. https://doi.org/10.9708/jksci.2022.27.02.015.

[11] K. Yang, "Transformer-based Korean pretrained language models: A survey on three years of progress", arXiv preprint, arXiv:2112.03014, Dec. 2021. https://doi.org/10.48550/arXiv.2112.03014.

[12] E. Merdjanovska, A. Aynetdinov, and A. Akbik, "NoiseBench: Benchmarking the Impact of Real Label Noise on Named Entity Recognition", arXiv preprint, arXiv:2405.07609, May 2024. https://doi.org/10.48550/arXiv.2405.07609.

[13] Y. Chen, K. Lim, and J. Park, "Korean Named Entity Recognition Based on Language-Specific Features", Natural language Engineering Vol. 30, pp. 625-649, Jun. 2023. https://doi.org/10.1017/S1351324923000311.

저자소개

김 선 겸 (Sun-Kyum Kim)

2010년 2월 : 세종대학교 컴퓨터공학과(공학사)
2016년 2월 : 연세대학교 컴퓨터과학과(공학박사)
2019년 3월 : 한국과학기술정보연구원 박사후연구원
2020년 7월 : 차세대융합기술연구원 선임연구원
2020년 8월 ~ 현재 : 한국건설기술연구원 미래스마트건설연구본부 수석연구원
관심분야 : 인공지능, 데이터분석, 블록체인

원 지 선 (Jisun Won)

2003년 2월 : 경희대학교 토목건축공학부(공학사)
2005년 2월 : 경희대학교 건축공학과(공학석사)
2024년 2월 : 경희대학교 건축학과 (박사수료)
2005년 12월 ~ 현재 : 한국건설기술연구원 미래스마트건설연구본부 수석연구원
관심분야 : 건설 데이터 표준, 인공지능, 자연어처리, BIM(Building Information Modeling)

신 재 영 (Jaeyoung Shin)

2015년 2월 : 한양대학교 실내건축디자인학과(이학사)
2017년 2월 : 한양대학교 실내건축디자인학과(이학석사)
2020년 9월 ~ 현재 : 연세대학교 실내건축학과 박사과정
2017년 3월 ~ 현재 : 한국건설기술연구원 미래스마트건설연구본부 전임연구원
관심분야 : 실내건축(설계, 실내동선), 인공지능, BIM(Building Information Modeling)