논문 2025-4-5 http://dx.doi.org/10.29056/jsav.2025.12.05

공유저작물 자동 정보 추출을 위한 미세조정 BERT-BiLSTM-CRF NER 모델의 성능 및 안정성 분석

황성훈*, 무사부부수구밀란두키스*, 박동주**†

Performance and Stability Analysis of a Fine-Tuned BERT-BiLSTM-CRF NER Model for Automated Information Extraction in Openly Licensed Works

SeongHun Hwang*, Milandu Keith Moussavou Boussougou*, Dong-Joo Park**†

요 약

스캔 기반 한국어 법률 문서는 복잡한 레이아웃과 OCR 오류로 인해 구조화 정보 추출이 어렵다. 본 연구는 Vision-Language 기반 OCR(Qwen-VL), Begin-Inside-Outside (B-I-O) 태깅 기반 학습 데이터셋, 그리고 BERT-BiLSTM-CRF 아키텍처를 결합한 한국어 법률 문서용 자동 개체명 인식(NER) 파이프라인을 제안한다. mBERT, KLUE-RoBERTa-Large, XLM-RoBERTa-Large를 Pure 구조와 BiLSTM-조건부 랜덤 필드(CRF) 구조로 미세조정하고, 약 30% OCR 노이즈를 포함한 데이터로 학습한 뒤 5-폴드 교차검증을 수행하였다. 실험 결과, CRF 결합 모델이 시퀀스 일관성과 엔티티 경계 인식에서 더 안정적인 성능을 보였으며, 특히 XLM-RoBERTa-Large-CRF는 평균 F1 0.9641을 기록했다. 본 연구는 OCR 노이즈 환경에서도 견고한 한국어 법률 NER 설계를 위한 실용적 방향을 제시한다.

Abstract

Korean legal documents pose challenges for information extraction due to complex layouts, Optical Character Recognition (OCR) noise, and agglutinative morphology. This paper proposes an automated Named-Entity Recognition(NER) pipeline that integrates Qwen-VL-based OCR, a Begin-Inside-Outside (B-I-O)-tagged training dataset, and fine-tuned BERT-family encoders with a BiLSTM-Conditional Random Field (CRF) decoder. We fine-tune mBERT, KLUE-RoBERTa-Large, and XLM-RoBERTa-Large under both Pure and BiLSTM-CRF settings, incorporating 30% OCR-style noise. A 5-폴드 cross-validation demonstrates that CRF-enhanced models achieve more stable and structurally consistent predictions, with XLM-RoBERTa-Large-CRF reaching an average F1-score of 0.998. The results highlight a practical design for robust NER in noisy OCR environments.

한글키워드 : 개체명 인식(NER), BERT, 조건부 랜덤 필드(CRF), 공유저작물

keywords : Named-Entity Recognition (NER), BERT, Conditional Random Field (CRF), Public Domain

* 숭실대학교 컴퓨터학과
** 숭실대학교 컴퓨터학부
† 교신저자: 박동주(email: djpark@ssu.ac.kr)

접수일자: 2025.12.02. 심사완료: 2025.12.13.
게재확정: 2025.12.20.

1. 서 론

국내 공공·공유저작물 문서는 스캔 이미지, 표, 다단 구조, 도장 영역 등이 혼합된 복잡한 레이아웃을 가지며, OCR 오류로 인해 텍스트가 누락되거나 노이즈가 포함되는 경우가 많다. 이러한 법률 문서의 권리정보 추출은 주로 수작업에 의존하고 있어 비용 증가와 오류 위험이 크며, 최근 문서량 증가로 자동화의 필요성이 더욱 높아지고 있다. 그러나 한국어 법률 문서에 특화된 NER 연구는 법률 문서의 정보 공개 제한이라는 구조적 제약으로 인해 매우 제한적으로 이루어져 왔다. 특히 계약서 및 동의서와 같은 핵심 법률 문서는 당사자 간 비밀유지서약(NDA) 및 개인정보 보호 문제로 인해 학습 데이터로의 활용이 사실상 불가능한 경우가 많았다. 이로 인해 기존 한국어 NER 연구의 대부분은 일반 문서 또는 개인정보 탐지 중심에 머물러 있으며, 법률 문서 도메인에 특화된 NER 모델을 OCR 환경까지 고려하여 체계적으로 분석한 연구는 극히 드물다. 또한 이러한 제약으로 인해 대규모 공개 데이터셋 구축 역시 현실적으로 어려운 상황이다. 본 연구는 이러한 연구 공백을 인식하고, 합성 데이터 기반 학습과 OCR 환경을 고려한 실험 설계를 통해 한국어 법률 문서 NER의 실용적 가능성을 탐색한다.

한국어 NER은 교착어적 특성, 형태소 결합, 그리고 OCR 오류 전이 문제로 인해 더욱 복잡하다. 따라서 레이아웃을 보존하고 OCR 오류를 완화할 수 있는 전처리 기술과 도메인 특화 NER 모델이 함께 요구된다. BERT 기반 NER은 표준으로 자리 잡았지만, 한국어 법률 문서에서 BERT 계열 모델 간 체계적 비교 연구, 특히 Pure 구조 대비 BiLSTM-CRF[1] 구조의 안정성 분석은 거의 이루어지지 않았다.

한편, 국내에서는 문화체육관광부와 한국문화정보원을 중심으로 공공·공유저작물의 개방과 활용을 촉진하기 위한 정책이 추진되고 있으며, 공공누리(KOGL)[2] 라이선스를 통해 공유저작물의 이용 조건, 권리자 정보, 공개 범위 등이 체계적으로 관리되고 있다. 그러나 이러한 정책적 기반에도 불구하고, 실제 계약서·동의서와 같은 법률 문서에 포함된 권리 정보를 자동으로 추출·구조화하는 기술적 연구는 아직 충분히 축적되지 않은 상황이다. 이에 따라 공유저작물의 효율적인 관리와 활용을 위해서는, OCR 환경을 고려한 법률 문서 분석과 도메인 특화 NER 기술에 대한 실증적 연구가 요구된다.

이에 본 연구는 Vision-Language 모델 기반 OCR, 미세조정 BERT-BiLSTM-CRF NER, 그리고 BIO 태깅 기반 데이터 구성을 결합한 한국어 법률 문서 정보 추출 파이프라인을 구축하고 그 성능과 안정성을 종합적으로 분석한다.

2. 관련 연구

2.1 한국어 개체명 인식의 도전 과제

한국어 NER은 교착어적 형태론[3]의 특성 때문에 영어와는 다른 고유한 어려움을 가진다. 영어에서는 개체명이 보통 단일 단어로 분리되어 나타나는 반면, 한국어에서는 체언(명사)에 조사와 같은 문법 형태소가 결합되어 하나의 어절을 이룬다. 초기 연구에서는 형태소 분석을 선행 단계로 두는 형태소 기반 태깅 접근 방식을 사용하였으나, 형태소 분석기(MA)의 오류가 NER 단계로 그대로 전이되는 문제가 발생해 성능의 병목 현상을 야기하였다[4]. 최근에는 KLUE-BERT, KoELECTRA 등 사전학습 언어모델의 발전으로 인해 형태소 분석 의존도를 줄이기 위해 문자 또는 서브워드 기반 처리 방식이 주로 사용되고 있다[5].

2.2 법률 텍스트 처리와 OCR 오류 전이 문제

법률 문서는 높은 정확도가 필수적인 대표적 도메인이다. Au et al.[6]의 최근 연구는 일반 NER 모델을 법률 텍스트에 적용할 경우 성능이 크게 저하됨을 보여주었다. 일반 웹 텍스트와 달리, 법률 계약서는 고어적 어휘, 중첩된 절 구조를 가진 복잡한 문장, 그리고 한자 기반의 법률 전문 용어가 높은 밀도로 등장한다. 따라서 해당 도메인에 특화된 미세조정 NER 모델이 필수적이다.

2.3 문서 이해 및 OCR 모델 비교

최근 문서 기반 정보 추출 연구에서는 OCR 결과와 문서 레이아웃을 함께 활용하는 다양한 모델이 제안되어 왔다. 대표적으로 LayoutLM 계열 모델과 Donut은 문서 이해를 위한 선행 연구로 널리 사용된다. LayoutLM[7]은 OCR 텍스트와 좌표 정보를 결합하여 문서 레이아웃을 모델링하지만, 고품질 OCR 결과와 대규모 레이아웃 어노테이션이 필요해 기밀성이 강한 한국어 법률 문서 환경에서는 적용에 제약이 있다. Donut[8]은 OCR-free 방식으로 문서 이미지를 직접 구조화된 출력으로 변환하지만, 태스크별 파인튜닝 의존도가 높고 복잡한 법률 문서에 대한 일반화에는 한계가 보고되고 있다.

본 연구에서는 이러한 End-to-End 문서 이해 모델 대신, Vision-Language 모델 기반 OCR(Qwen3-VL-235B-a22b-instruct)과 NER를 분리한 파이프라인 구조를 채택하였다. Qwen3-VL은 문서 이미지로부터 레이아웃 순서를 반영한 텍스트 시퀀스를 생성할 수 있어, 기존 OCR 대비 구조 보존성이 높으며 후속 NER 모델과의 연계가 용이하다. 본 설계는 데이터 확보가 제한적인 한국어 법률 문서 환경에서 OCR 노이즈의 영향을 단계별로 분석하고, NER 모델의 안정성을 정량적으로 평가하기 위한 실험 목적에 부합하는 현실적인 선택이다.

2.4 NER을 위한 BERT와 CRF의 역할

기존 연구들은 BERT를 NER 태스크에 미세조정하면 최신(state-of-the-art) 성능을 달성할 수 있으며, 이는 사실상 표준 방식으로 자리 잡았음을 보여주었다. 이후 연구에서는 BERT에 BiLSTM과 CRF를 결합하면 라벨 간 의존성을 모델링할 수 있어 성능을 추가적으로 향상시킬 수 있음이 밝혀졌고, 이러한 구조는 널리 활용되고 있다[9].

여러 연구들에서 BERT 단독 모델과 BERT+CRF 모델을 비교한 결과, 많은 경우 CRF가 제공하는 성능 향상은 크지는 않지만 분명히 존재함이 보고되었다. 예를 들어 Hong과 Kim[10]은 한국어 NER을 위해 KoELECTRA-CRF 모델을 적용하였으며, ELECTRA 계열 모델에 CRF를 결합할 때 F1 점수가 개선된다는 것을 입증하였다.

요약하자면, BERT+CRF 모델은 한국어 법률 문서를 포함한 법률 NER에서 여전히 핵심적 접근 방식으로, 문맥 기반 이해 능력과 시퀀스 일관성을 동시에 확보할 수 있는 견고한 방법을 제공한다. BERT의 강력한 문맥 표현 덕분에 CRF의 기여도가 과거(pre-BERT) 모델에 비해 작아지긴 하였으나, 엔티티 수준 F1에서 CRF는 여전히 안정적인 소폭의 성능 향상을 제공한다.

3. 연구방법

본 연구에서는 비정형적인 법률 문서 이미지(raw, unstructured images)를 구조화된 메타데이터로 변환하기 위한 통합 파이프라인을 제안한다. 본 절에서는 제안된 파이프라인의 전체 시스템 구조, 데이터 구축 과정, 그리고 각 구성 요소의 구현 방법을 상세히 기술한다.

그림 1은 제안하는 한국어 법률 문서용 NER

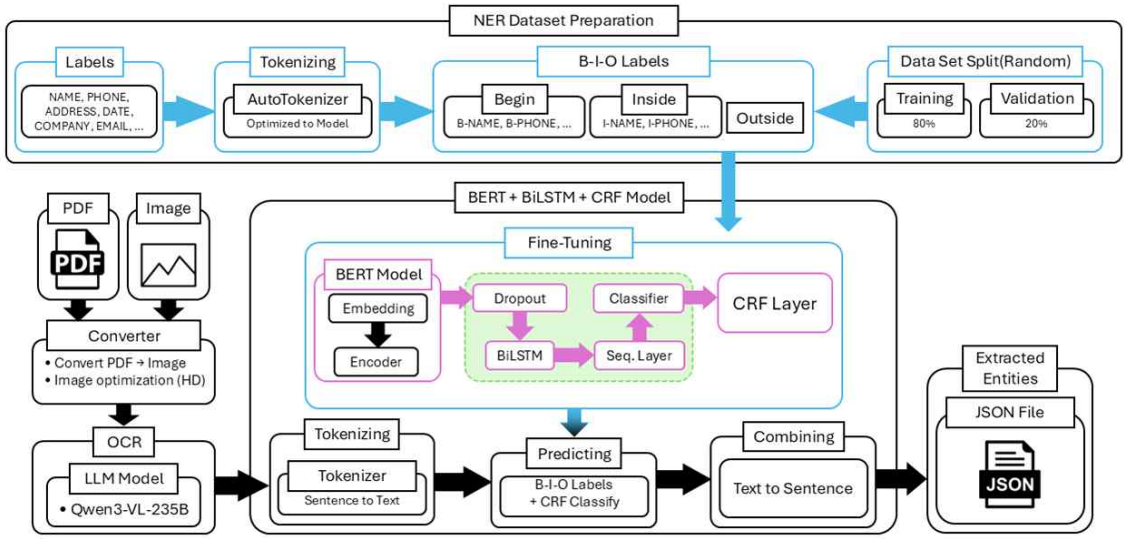

그림 1. 파이프라인 개요
Fig. 1. Pipeline Overview

시스템의 전체 흐름을 나타낸다. 시스템은 문서 입력 단계, OCR 단계, 그리고 미세조정된 모델을 활용한 개체명 인식 단계의 세 가지 과정으로 구성된다.

PDF와 이미지는 텍스트가 들어있는 문서를 찍은 사진이며, 따라서 문서 내부의 텍스트를 인식할 필요가 있다. 먼저 PDF는 이미지로 변환하고, 이미지는 해상도 업스케일링을 통해 텍스트가 선명히 보이도록 개선한다. 이렇게 작업한 결과물을 OCR을 통해 텍스트로 변환한다. 이는 OCR 결과물의 정확도를 높이기 위한 수단으로 작용한다.

사용할 모델 구조는 크게 BERT모델, BiLSTM, CRF 레이어로 나눌 수 있으며, 가장 중요한 역할을 담당하는 레이어는 CRF와 BERT 모델이다. BERT모델은 각 텍스트에 임베딩이 이루어진 후 인코딩 처리가 되어 텍스트 리스트에 담기게 된다. 이렇게 담긴 텍스트들 중 일부는 마스킹 처리하는 드롭아웃(Dropout)을 시행하여 BERT모델의 예측에서 과적합을 방지하고, 이를 다시 BiLSTM과 순차적 레이어(Sequential Layer)로 전체 텍스트를 정규화하는 과정을 거친다. 이후 학습에 없던 단어가 포함되어 예측이 되지 않은 경우를 대비하기 위해, CRF 레이어를 추가하여 어디 라벨에 가까울지를 가중치로 예측한다.

위와 같이 복잡한 과정을 거치는 이유는 학습할 데이터가 적어 과적합이 쉽게 일어날 가능성이 있고, 학습한 데이터가 아닌 데이터에 대해서 예측률이 떨어지는 경향을 보이는 BERT모델(이 부분은 REFERENCE 추가 필요, 추가 후 괄호 내부 문장 삭제) 특성상 결과물의 질이 하락할 우려가 존재하기 때문이다.

학습의 경우에는 BERT모델의 텍스트 토큰 방식을 고려하여 학습시킬 NER 데이터셋은 학습시킬 라벨과 데이터를 B-I-O로 구성하여 학습시켰다. 이는 예측 시에도 마찬가지로, 모든 텍스트는 반드시 글자당의 대응을 수행할 수 있게 토크나이징이 필요하다. 즉, OCR을 수행한 텍스트는 토크나이징을 통해 모델이 인식할 수 있는 형태로 재구성되어 각 글자마다 B-I-O로 예측이 수행되고, 이후 B-I-O 태깅을 제거한 후 엔티티-

라벨 관계로 다시 합치는 과정을 수행하여 JSON 파일로 저장된다.

3.1 비전-언어 모델을 활용한 OCR 처리

법률 문서는 “저작권 양도 계약서”와 같은 PDF 또는 이미지 형태로 입력되며, Qwen3-VL-235b-a22b-instruct 모델을 사용한 Vision-Language model(VLM) 기반 OCR을 통해 텍스트를 추출한다. Qwen3-VL[11], [12] 계열 모델은 문서 레이아웃을 정확하게 처리할 수 있으므로 한국어 법률 문서의 OCR 작업에 적합하다[13].

표준 OCR이 텍스트를 단순한 단어 집합 형태로 출력하는 것과 달리, Qwen-VL은 문서의 공간적 구조를 반영한 텍스트를 생성하여 레이아웃을 보존한다. 이러한 구조 인식 능력은 서명, 표 항목 등 위치가 의미에 직접 영향을 미치는 법률 서식 처리에서 특히 중요하다.

또한 Tesseract와 같은 기존 OCR 엔진은 표, 다단 구성 등 복잡한 법률 문서 레이아웃에서 성능 저하를 보이는 경우가 많다. 반면 VLM 기반 OCR(Qwen-VL)을 통합함으로써, 본 연구의 접근 방식은 모델의 의미적 이해를 활용하여 텍스트 생성 단계에서 시각적 모호성을 교정함으로써 OCR 오류 전이를 효과적으로 줄이는 데 초점을 맞춘다.

3.2 미세조정된 NER 모델

OCR을 통해 추출된 텍스트에 대해, BiLSTM이 결합된 미세조정 BERT 계열 인코더와 B-I-O 태그 위에서 유효한 태그 시퀀스를 강제하는 CRF 디코더를 사용하여 NER을 수행한다. BERT-BiLSTM-CRF NER의 출력은 텍스트 시퀀스의 각 단어에 대한 예측 태그이다.

B-I-O 태깅 체계는 동일 유형의 인접한 개체를 구분하거나 여러 토큰으로 구성된 개체(예: “Ministry | of | Culture” → B-ORG, I-ORG, I-ORG)를 효과적으로 표현할 수 있다는 장점이 있다.

사전학습된 BERT 기반 모델을 한국어 법률 문서 NER에 적용하기 위해서는, 해당 언어의 개체를 정확하게 인식하도록 미세조정이 필요하다. 본 연구에서는 mBERT-Cased, KLUE-RoBERTa-Large, XLM-RoBERTa-Large 세 모델을 미세조정하여 성능을 비교하였다. 이러한 모델들은 한국어와 같이 자연어 처리 관련 데이터가 상대적으로 부족한 언어에서 우수한 성능을 보이기 때문에 선택하였다.

4. 실험 및 실험결과

4.1 실험 환경

모든 모델의 학습은 NVIDIA의 H200 GPU 환경에서 수행하였다. 학습 및 추론 코드의 경우 Python 3.x, PyTorch, HuggingFace의 Transformers 라이브러리를 기반으로 구현하였다. GPU 자원을 효율적으로 활용하기 위해 AdamW와 linear warmup scheduler를 사용하였고, 학습률 2e-5, epoch 20, 최대 시퀀스 길이 128, 배치 크기는 전 모델 128로 고정하였다. mixed precision(AMP)로 학습을 수행하였으며, gradient clipping을 적용해 학습이 안정적으로 이루어지도록 하였다. 데이터가 부족한 환경이기 때문에 전체 데이터셋을 훈련용 데이터셋과 테스트 데이터셋으로 나누고, 이후 검증용 데이터셋을 따로 복사하였다. 이후 훈련 중에는 검증용 데이터셋은 유지한 채로 훈련용 데이터셋과 테스트 데이터셋만 매번 새로 만들어가며 테스트를 진행하였다. 데이터가 부족한 환경의 특성과, 모델 간의 학습 결과가 절대성을 띨 수 있도록 Early-Stopping은 사용하지 않았다. 과적합이 발

생하더라도 이후 그래프에서 차이가 발생하므로, 과적합 진단이 가능하기 때문이다.

4.2 데이터셋 전처리

법률 NER에서 가장 큰 장애 요소는 라벨이 부착된 학습 데이터의 부족이다. 법률 문서는 개인정보 및 기밀 정보가 포함될 가능성이 높아 대규모 공개 데이터셋 구축이 어렵기 때문이다. 이를 해결하기 위해 본 연구에서는 법률 문서에서 빈번히 등장하는 문장 구조를 템플릿으로 만들고, 이름·전화번호·날짜 등 23개 엔티티 유형을 삽입하여 합성 데이터를 생성하였다. 특히 구성한 라벨 중 언어, 주소와 같은 라벨도 존재하는데, 이는 추후 공유저작물로 한국어에 한정하지 않고 공유저작물을 해외에도 적용되게 하거나, 해외의 저작물 또한 추출할 수 있도록 작용하는 것을 목표하고 있다. 또한 실제 문서처럼 엔티티가 전혀 없는 문장도 약 10% 포함해 모델이 불필요한 엔티티를 과대 검출하는 현상을 방지하였다.

최종적으로 약 30,000개의 문장으로 구성된 합성 데이터셋을 구축하였으며, 각 모델은 5-폴드 교차 검증을 통해 학습·평가의 편향을 최소화하였다. 즉, 전체 데이터셋을 5개의 부분집합으로 구성한 후 하나를 검증데이터로 사용, 나머지를 훈련 데이터로 사용하는 과정으로 검사를 진행하며, 이는 총 5번 반복된다. 이는 Fold 간 성능 편차로 인한 왜곡을 방지하고, 모델의 안정적 성능을 검증하기 위함이다.

4.2.1 B-I-O 태깅

본 논문에서 사용한 엔티티 타입은 총 23개이다. 따라서 B-, I- 접두사를 붙여서 총 46개 라벨이 만들어졌고, 이외 엔티티가 아닌 O 타입까지 포함해 총 47개의 B-I-O 라벨을 사용하였다. BERT 토크나이저의 특성에 따라 서브워드(Sub-Word) 단위의 토큰(Token)으로 분해할 때, 모든 어절의 라벨들은 서브워드에 복제하고, BERT 모델에서 단어 사이의 관계를 정의하게 되는 CLS, SEP, PAD 토큰의 경우 모두 -100으로 마스킹 처리한 후, 손실 계산 및 평가에서 제외하였다. 다음 표 1는 모델 미세조정 후 생성된 B-I-O 태깅 예시를 보여준다.

학습 데이터의 현실성을 높이기 위해, 위처럼 B-I-O 태깅으로 파일을 생성할 때 어절 단위 토큰 중 약 30%의 비율로 OCR에서 발생할 수 있는 노이즈(공백, 문자오류 등)를 발생시켰다.

표 1. B-I-O 태깅의 예시
Table 1. Example of B-I-O tagging

Token	Tag	Token	Tag
가격은	O	1431만원입니다.	B-MONEY
남기의	I-COMPANY	주소는	O
10월	I-DATE	9일에	I-DATE
허오의	B-NAME	지급한다.	O
서울시	B-ADDRESS	왕라구	I-ADDRESS
주민번호:	O	210419-1******	B-ID_NUM

4.3 BERT-BiLSTM-CRF 모델의 미세조정

본 연구에서는 세가지 사전학습 BERT모델에서 두가지 구조를 조합하여 나오는 6개의 모델을 실험하였다. 실험에 사용한 모델은 BERT-Base-Multilingual-Cased(mBERT), Klue-RoBERTa-Large(Klue-Ro), XLM-RoBERTa-Large(XLM)이다. 비교하는 구조가 각 모델의 그 자체 구조(Pure), 그리고 BERT에 BiLSTM과 CRF 레이어를 새로 추가한 구조이다.

BERT-BiLSTM-CRF 구조의 경우 BiLSTM은 BERT가 제공하는 문맥 정보 위로 추가 분류를 수행하며, 엔티티 경계 주위의 패턴(예를 들어 성명-홍길동, 계약일-2025년)을 더 민감하게 찾도록 돕는다. BiLSTM에서 경계 패턴을 찾고

나서, 이어지는 CRF 레이어는 예측한 문맥 정보들 중 B-I-O 태깅 문법을 만족하는 이들만을 선택한다. Pure BERT 모델과 달리, 각 토큰의 예측은 독립적이지 않고 시퀀스의 전체 점수를 최대화하도록 설계되어 있고, B-태그 다음에 B-태그가 다시 등장하는 등의 비정상적인 라벨링을 개선하는 경향이 있다. 이것이 기존 Pure 모델과 차이가 나는 부분이며, 이를 통해 전체적인 성능은 유지하면서 안정성을 끌어올릴 수 있다.

4.4 실험결과

모델의 평가에는 F1점수(F1-Score)를 주 지표로 해서, Precision, Recall를 위주로 산출하였다. 표 2는 학습 이후 세가지 미세조정 NER모델의 최종 평가 결과를 보여준다.

BERT-BiLSTM-CRF 모델의 최종적 목표는 성능 향상이 아닌, 성능 유지와 대비한 노이즈 분석의 안정화에 있다. 따라서 일부 설정에서는 전반적으로 Pure BERT모델보다 더 높은 F1 점수를 볼 수 있다.

표 2. 학습 후의 평가 결과
Table 2. Post-learning evaluation results

Model	Prec	Rec	F1
mBERT-Pure	0.8851	0.9518	0.9153
Klue-Ro-Pure	0.9085	0.9531	0.9293
XLM-Pure	0.9449	0.9795	0.9617
mBERT+BiLSTM+CRF	0.8925	0.9518	0.9239
Klue-Ro+BiLSTM+CRF	0.9105	0.9568	0.9354
XLM+BiLSTM+CRF	**0.9474**	**0.9804**	**0.9631**

다른 모델과 비교했을 때, K-폴드로 평가된 표에서 XLM-RoBERTa-Large가 F1-점수 96.31%로 가장 우수한 성능을 보였다. RoBERTa가 BERT를 개선한 모델인 점을 고려할 때, RoBERTa-BiLSTM에 CRF 레이어까지 추가함으로써 전통적 NER모델 대비 노이즈가 많은 텍스트를 시험할 때, 설정에 따라 폴드 간 분산(또는 최악 폴드 성능) 측면에서 변동이 완화되는 경향이 관측되었다. 이는 CRF가 라벨 전이 제약을 반영함으로써 예측 시퀀스의 구조적 안정성을 개선한다는 설계 의도와 부합한다. K-폴드로 평가하였기 때문에 F1점수, Recall, Precision 등에

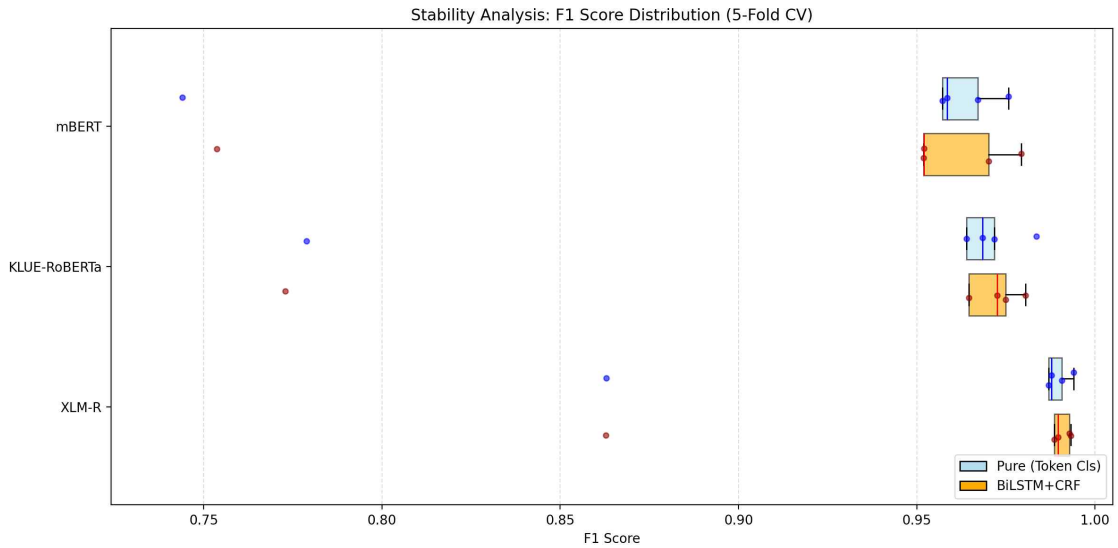

그림 2. 안정성 분석: F1 점수 분포 (5-겹 교차 검증)
Fig. 2. Stability Analysis: F1 Score Distribution (5-폴드 CV)

서 근소한 차이여도 실제로의 안정성이 Pure 대비 CRF 구조가 더 좋을 수 있다고 평가할 수 있다. 그림 2는 안정성 분석 F1 점수 분포(5-겹 교차 검증)를 보여준다.

실제로 전체적인 모델에서 평균 폴드 성능이 개선되었다는 점을 확인할 수 있었고, 각 모델이 가지는 특성에 따라 박스 플롯에서 꼬리(outlier) 부분이 완화되는 경향을 보이는 경우도 있고, 바닥(worst)점이 덜 무너지는 경향을 보이는 경우도 있었다.

5. 결 론

본 연구는 개인정보 및 라벨 부족으로 공개 학습이 어려운 한국어 법률 문서 환경을 대상으로, Qwen-VL 기반 OCR과 BERT 계열 NER 모델을 결합한 정보 추출 파이프라인을 제안하였다. 특히 기존 토큰 분류(Pure) 구조와 비교하여, BIO 구조를 반영한 BiLSTM-CRF 레이어가 예측 시퀀스의 구조적 일관성을 강화함을 확인하였다.

5-폴드 교차검증 결과, 제안 모델은 기존 모델과 유사한 F1 성능을 유지하면서 일부 조건에서 폴드 간 성능 변동이 감소하였다. 다만 본 연구는 템플릿 기반 합성 데이터를 사용했기 때문에 실제 문서·OCR 환경에서는 동일한 성능이 보장되지 않는 한계가 있다. 향후 실제 문서 기반 평가, BIO 오류 지표 추가, BiLSTM-CRF 레이어의 기여도 정량화 등이 필요하다.

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2025년도 글로벌 저작권 현안 신속 대응(R&D) 사업으로 수행되었음 (과제명: 공유저작물의 글로벌 확산을 위한 콘텐츠 분석 및 유형정보 판단 기술 개발, 과제번호: RS-2025-02305397, Rate: 100%)

참 고 문 헌

[1] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition”, in Proc. NAACL-HLT, San Diego, CA, USA, Jun. 2016, pp. 260 - 270, DOI: doi.org/10.18653/v1/N16-1030.

[2] Korea Culture Information Service Agency (KCISA), Korea Open Government License (KOGL) License Guidelines, Korea Culture Information Service Agency, 2021. [Online]. Available: https://www.kogl.or.kr

[3] Y. Chen, K. Lim, and J. Park, “Korean named entity recognition based on language-specific features”, Natural Language Engineering, vol. 30, no. 3, pp. 625 - 649, 2024. DOI: doi.org/10.1017/S1351324923000311

[4] H. Kim and H. Kim, “Fine-Grained Named Entity Recognition Using a Multi-Stacked Feature Fusion and Dual-Stacked Output in Korean”, Applied Sciences, vol. 11, no. 22, p. 10795, 2021. DOI: doi.org/10.3390/app112210795

[5] S. Jang, Y. Cho, H. Seong, T. Kim, and H. Woo, “The Development of a Named Entity Recognizer for Detecting Personal Information Using a Korean Pretrained Language Model”, Applied Sciences, vol. 14, no. 13, p. 5682, 2024. DOI: doi.org/10.3390/app14135682

[6] T. W. T. Au, V. Lampos, and I. Cox, “E-NER --- An Annotated Named Entity Recognition Corpus of Legal Text”, in Proceedings of the Natural Legal Language Processing Workshop 2022, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 246 - 255. DOI: doi.org/10.18653/v1/2022.nllp-1.22

[7] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, “LayoutLM: Pre-training of text and layout for document image

understanding", in Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (KDD), New York, NY, USA, 2020, pp. 1192 - 1200,
DOI: doi.org/10.1145/3394486.3403172.

[8] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, "OCR-free document understanding transformer",in Proc. ECCV, Tel Aviv, Israel, Oct. 2022, pp. 498 - 517, DOI: doi.org/10.1007/978-3-031-19815-1_29

[9] D. Premasiri, T. Ranasinghe, R. Mitkov, et al., "Survey on legal information extraction: current status and open challenges", Knowl. Inf. Syst., vol. 67, pp. 11287 - 11358, 2025.
DOI: doi.org/10.1007/s10115-025-02600-5.

[10] J. Hong and H. J. Kim, "Korean named entity recognition based on ELECTRA-CRFs", in Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology (HCLT), 2020, pp. 473 - 476. [Online]. Available: https://www.koreascience.kr/article/CFKO202030060831856.pdf

[11] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, et al., "Qwen3 Technical Report", arXiv e-print arXiv:2505.09388, May 2025.
DOI: doi.org/10.48550/arXiv.2505.09388

[12] S. Bai et al., "Qwen2.5-VL Technical Report", arXiv preprint arXiv:2502.13923, Feb. 19, 2025.
DOI: doi.org/10.48550/arXiv.2502.13923

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", in Proc. NAACL-HLT, Minneapolis, MN, USA, Jun. 2019, pp. 4171 - 4186,
DOI: doi.org/10.18653/v1/N19-1423.

저 자 소 개

황성훈(SeongHun Hwang)

2025.2 숭실대학교 미래교육원 컴퓨터학과 학사
2025.3 - 현재 : 숭실대학교 컴퓨터공학과 석사
<주관심분야> 인공지능, 자연어처리, LLM

무사부부수구밀란두키스(Milandu Keith Moussavou Boussougou)

2014.7 가봉 Institut Supérieur de Technologie 컴퓨터공학과 학사
2021.2 숭실대학교 컴퓨터공학과 석사
2021.3 - 현재 : 숭실대학교 컴퓨터공학과 박사
<주관심분야> 인공지능, 자연어처리, LLM, 보안 및 사이버보안 응용

박동주(Dong-Joo Park)

1995.2 서울대학교 컴퓨터공학과 학사
1997.2 서울대학교 컴퓨터공학과 석사
2001.8 서울대학교 컴퓨터공학부 박사
2004.3 - 현재 : 숭실대학교 컴퓨터학부 교수
<주관심분야> 데이터베이스, 멀티미디어 데이터베이스, 임베디드 소프트웨어