Journal of The Korea Convergence Society
Vol. 13. No. 2, pp. 13-20, 2022
e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2022.13.02.013

사전학습 언어모델을 활용한 범죄수사 도메인 개체명 인식

김희두[1], 임희석[2*]
[1]고려대학교 빅데이터융합학과 석사과정, [2]고려대학교 컴퓨터학과 교수

A Named Entity Recognition Model in Criminal Investigation Domain using Pretrained Language Model

Hee-Dou Kim[1], Heuiseok Lim[2*]
[1]Master Student, Department of Bigdata Convergence, Korea University
[2]Professor, Department of Computer Science and Engineering, Korea University

요 약 본 연구는 딥러닝 기법을 활용하여 범죄 수사 도메인에 특화된 개체명 인식 모델을 개발하는 연구이다. 본 연구를 통해 비정형의 형사 판결문·수사 문서와 같은 텍스트 기반의 데이터에서 자동으로 범죄 수법과 범죄 관련 정보를 추출하고 유형화하여, 향후 데이터 분석기법을 활용한 범죄 예방 분석과 수사에 기여할 수 있는 시스템을 제안한다. 본 연구에서는 범죄 수사 도메인 텍스트를 수집하고 범죄 분석의 관점에서 필요한 개체명 분류를 새로 정의하였다. 또한 최근 자연어 처리에서 높은 성능을 보이고 있는 사전학습 언어모델인 KoELECTRA를 적용한 제안 모델은 본 연구에서 정의한 범죄 도메인 개체명 실험 데이터의 9종의 메인 카테고리 분류에서 micro average(이하 micro avg) F1-score 99%, macro average(이하 macro avg) F1-score 96%의 성능을 보이고, 56종의 서브 카테고리 분류에서 micro avg F1-score 98%, macro avg F1-score 62%의 성능을 보인다. 제안한 모델을 통해 향후 개선 가능성과 활용 가능성의 관점에서 분석한다.

주제어 : 범죄예방, 범죄수사, 사전학습 언어모델, 범죄 도메인 텍스트, 개체명 인식, KoELECTRA

Abstract This study is to develop a named entity recognition model specialized in criminal investigation domains using deep learning techniques. Through this study, we propose a system that can contribute to analysis of crime for prevention and investigation using data analysis techniques in the future by automatically extracting and categorizing crime-related information from text-based data such as criminal judgments and investigation documents. For this study, the criminal investigation domain text was collected and the required entity name was newly defined from the perspective of criminal analysis. In addition, the proposed model applying KoELECTRA, a pre-trained language model that has recently shown high performance in natural language processing, shows performance of micro average(referred to as micro avg) F1-score 98% and macro average(referred to as macro avg) F1-score 95% in 9 main categories of crime domain NER experiment data, and micro avg F1-score 98% and macro avg F1-score 62% in 56 sub categories. The proposed model is analyzed from the perspective of future improvement and utilization.

Key Words : Crime Prevention, Criminal Investigation, Pretrained Language Model, Crime Domain Text, Named Entity Recognition, KoELECTRA

*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received November 4, 2021
Revised December 7, 2021
Accepted February 20, 2022
Published February 28, 2022

1. 서론

1.1 서론

범죄 데이터 마이닝 기술은 미래에 발생할 수도 있는 범죄를 예방하거나 현재 수사 중인 사건을 해결하는데 있어 매우 유용한 도구이다. 개체 추출(Entity Extraction), 군집 분석(Clustering Analysis), 연관규칙 분석(Association Rule Analysis), 사회 연결망 분석(Social Network Analysis) 등은 범죄 데이터 마이닝의 기술로써 자주 사용되고 있으며[1], 이와 같은 분석을 통해 범죄의 특성을 분류한다면 발생한 사건의 범죄자·유사 사건 후보를 추천하거나, 범죄가 발생할 가능성이 높은 상황에 대해 미리 예측을 수행하는 등 다양한 목적으로 사용될 수 있다.

특히 최근 사이버 공간을 통한 신종 범죄가 증가하여 사법기관에 보고되지 않는 범죄 관련 정보가 급격하게 늘어남에 따라, 내부 DB에서 관리되는 정보 뿐만 아니라 외부 환경에서 존재하는 정보를 함께 연결하여 분석에 활용해야 할 필요성이 커지고 있는 상황이다. 이에 효율적인 범죄 정보의 추출과 통합의 중요성은 더욱 증대되고 있다.

또한 대부분 구조화되어 있지 않은 범죄 데이터 특성상 다양한 목적에 의해 파생되는 범죄 예측·분류의 범위와 정확도는 원천 데이터에서 추출 가능한 범죄 정보의 범위와 정확도에 의존할 수 밖에 없게 된다. 그런데 만약 범죄 정보가 정확하게 추출이 되어 있지 않거나, 사용가능한 특성의 범위가 적다면 범죄 예측 및 분류 결과의 활용과 신뢰도에 있어 많은 문제점을 야기할 수 있다.

먼저, 범죄 분석관·수사관들은 수사 중인 사건과 미해결 범죄와의 연관성을 분석하거나, 동일한 수법의 범죄자를 찾아 수사의 단서를 취득하기 위해 다양한 검색 쿼리를 활용하고, 범죄 특성별 통계적 유사도로 측정된 추천 알고리즘을 이용하게 된다. 이때 추출된 특성의 범위가 너무 적은 문제로 광범위한 추천 결과를 받아들인다면, 신속한 사건 처리에 어려움이 생길 수 있다. 둘째로, 범죄 특성이 정확히 추출되어 있지 않음에도 불구하고 범죄 발생 가능성이 높은 상황과 지역을 섣불리 예측하여 치안 인력을 배치하거나 범죄 예방 관련 정책을 세운다면, 오류로 측정된 범죄 예측 결과로 인해 막대한 예산 낭비와 민간의 피해를 발생시킬 수 있다.

따라서 데이터 분류와 예측을 통한 범죄수사 및 치안활동의 한계점을 극복하기 위해서는 우선 정확도가 높고 많은 데이터를 연결할 수 있는 공통된 범죄 정보 추출에 대한 연구가 먼저 선행되어야 한다.

한편, 개체명 인식(Named Entity Recognition)이란 사람, 장소, 기관 등 자연어에서 명명된 개체를 자동으로 추출할 수 있는 자연어 처리의 하위 분야이다. 최근 사전학습된 언어모델(Pretrained Language Model)을 이용하여 미세조정(fine-tuning) 학습을 수행하는 BERT[2] 방식의 연구가 높은 개체명 인식 성능을 보임에 따라 대화형 챗봇, 의생물학 분야 개체명 인식 등 다양한 분야에서 활용되고 있다[3]. 이처럼 정확도가 높아진 딥러닝 기반의 개체명 인식 기술은 효율적인 범죄 정보 추출의 도구로서도 기능할 수 있으며, 다양한 범죄 분석을 위해 활용될 수 있다. 그러나 국내에서는 아직 관련된 개발이나 연구 사례가 부족하여 적절히 활용되고 있지 못하고 있는 상황이다.

따라서 본 연구에서는 사전학습 언어 모델을 활용한 개체명 인식 기술을 범죄 수사라는 도메인에 적용하여, 범죄 예측과 분류에 있어 활용도를 높일 수 있는 딥러닝 기반의 범죄 도메인 개체명 인식 모델을 제안하고자 한다. 본 연구에서 제안하는 모델은 사람, 장소와 같은 일반적인 개체 이외에도 범행도구, 가명, 차량번호, 범행에 사용된 전화번호, 계좌번호 등 범죄 문서 상에서만 추출할 수 있는 특수한 개체명을 미리 정의함으로써 분석가능한 범죄 특성의 범위를 확장하였고, 사전학습 언어모델을 활용해 학습을 하였기 때문에 이전보다 정확도 높은 범죄 분석이 가능한 기반 기술이 될 것이다. 본 논문의 기여는 다음과 같다. 첫째, 한국어 범죄 수사 분야의 개체명 인식을 위한 개체명 분류 체계를 제시하고, 학습데이터를 구축하였다. 두 번째, 구축된 학습 데이터를 이용한 범죄 수사 분야 최초의 한국어 개체명 인식 기술을 제안하였다.

본 논문의 구성은 다음과 같다. 2장에서는 범죄정보 추출과 분류에 대한 해외 연구 사례들을 살펴볼 것이다. 3장에서는 본 논문에서 제안하고자 하는 딥러닝 기반 범죄 수사 도메인 개체명 인식 모델의 구조를 소개한다. 4장에서는 제안한 모델을 이용한 실험을 통해 범죄 도메인 텍스트에서의 개체명 인식 성능을 비교 분석한다. 5장에서는 본 연구의 결론과 향후 연구를 소개하고자 한다.

2. 관련 연구

2.1 범죄 정보 추출(Crime Information Extraction)

범죄 정보 추출과 관련된 연구가 부족한 국내와 달리 해외에서는 비정형의 텍스트에서 범죄 정보를 추출하기 위한 연구가 많이 진행되어 왔다. 범죄 정보 추출 관련 연구는 크게 활용 데이터와 추출 방법이라는 두 가지 측면에 따라 나눠서 살펴볼 수 있다. 먼저 활용 데이터의 측면에서 살펴보면 초기에는 주로 경찰서 등 사법기관에서 제공한 적은 양의 수사 보고서, 진술조서, 이메일 피해 접수내역 등에서 자동으로 범죄 정보를 추출하는 알고리즘에 대한 연구가 진행되었다[4-7,13,14]. 이후 온라인 상에서 빠른 속도로 생산되는 대량의 웹 문서에 대한 빅데이터 분석이 필요해지면서, 뉴스·블로그·SNS 등의 텍스트에서 범죄 정보를 추출하여 마약 및 무기거래·성범죄·사이버범죄 등의 정보를 추적하는 연구가 진행되었다[8-12]. 최근에는 온라인 뉴스 및 SNS 상의 텍스트와 그에 포함된 이미지 등에서 추출된 범죄 정보를 동일한 엔티티를 기준으로 결합하여 지식베이스를 구축하기 위한 시도의 연구가 진행되었다[16].

다음으로, 추출 방법에 따른 연구를 살펴보면 사람, 장소, 조직, 마약명 등의 개체에 대해 주로 특정 범죄를 대상으로 구축된 어휘집(Lexicon) 또는 지명사전(Gazetteer)을 이용한 규칙 기반의 방식(Rule-based approach)이 많이 사용되었으나[6, 8, 9, 11, 13, 14], 자연어 처리 기술이 발달하면서 CRF 등 통계적 머신러닝 기법을 사용한 방식[12], 그리고 CNN(Convolutional Neural Network) 및 LSTM(Long Short-Term Memory)을 사용한 딥러닝 기반의 신경망 알고리즘을 사용한 방식도 연구되었다[15].

3. 범죄수사 도메인 개체명 인식 모델

3.1 범죄 도메인 개체명 정의

국내에서 범죄 수사 도메인에 특화된 개체명 인식 모델은 아직 제안된 바가 없다. 따라서 본 연구에서는 범죄 분석용 텍스트 내에서 분류가 가능한 개체명 체계를 새롭게 정립하고, 이와 같은 체계에 자연어 처리 기술을 적용하였을 때 범죄 정보 추출의 효용성과 활용 가능성을 증명하기 위해 딥러닝 기반의 모델을 설계하여 성능을 비교해본다.

우선 한국정보통신기술협회에서 표준으로 지정한 한국어 개체명 분류*에 따르면, 일반적인 문서에서 추출할 수 있도록 사전에 정의된 개체는 인물, 장소, 조직, 시간, 전화번호, 법률명칭 등이 있다. 반면 범죄 수사 도메인 텍스트인 수사보고서, 피의자 신문조서, 참고인 조서, 형사 판결문 등에서 추가적으로 활용할 수 있는 개체 중 기존 개체명 표준에서 정의되지 않은 개체는 Table 1과 같다.

Table 1. Newly defined entities in criminal investigation

Entity	Definition
PS_COURT	재판·수사 과정상 신분 및 법적인 지위
PS_NAME_ACCUSED	혐의가 확인되었거나 의심받는 사람
PS_NAME_VICTIM	피해가 확인되었거나 주장하는 사람
PS_NAME_WITNESS	목격하였거나 아는 바를 진술한 사람
PS_NAME_JUSTICE	기소 또는 수사 담당하는 검사, 경찰
PS_NAME_JUDGE	재판을 진행하는 판사
PS_NAME_ATTORNEY	변호사 또는 법무법인
PS_NAME_APPRAISER	감정,진단 등을 수행한 의사, 응급처치사, 구급대원
PS_NICKNAME	범죄 실행 과정에서 사용된 가짜 이름
DT_LAW	유기징역 등 형벌 선고에 의한 형기
OG_ENTERTAINMENT	노래방, 포장마차 등 유흥주점
OG_CONVINIENCE	편의점
OG_FAKE	사칭한 기관명
CV_SEX	성별
CV_POSITION_FAKE	사칭한 직급명
QT_ACCOUNT	계좌번호
QT_CASE	사건번호
QT_VEHICLE_PLATE	차량번호
QT_IDENTITY	주민번호
QT_LAW	법률·규칙의 조항
QT_URL	인터넷 url주소
QT_IP	ip주소
TMC_TOOL	범행에 사용된 도구
TMC_DAMAGED_PRODUCT	범죄로 인해 훔치거나 빼앗긴 물건

3.2 Bi-LSTM

LSTM[17]은 문장에서 분절된 토큰이 순차적으로 신경망(Neural Network)의 입력으로 들어가면서 직전 입력까지의 정보(memory)가 축적된 은닉상태(hidden state)에서의 출력을 함께 입력으로 받아 결과를 출력해주는 RNN 구조를 기본으로 한다. 이때 긴 문장을 정확하게 이해하기 위해서는 한참 전의 입력토큰 정보까지 기억해서 출력을 결정해야 하는 장기 의존성 문제(Long term dependency)가 존재하는데, RNN의 경우 긴 문장에 대한 학습 시 기억 손실이 일어나는 경우가 많다. LSTM은 이와 같은 RNN 구조의 한계를 보완하기 위해 은닉 상태 안에 cell state를 추가하고 입력

* https://committee.tta.or.kr/

게이트(Input gate), 망각게이트(Forget gate), 출력게이트(Output gate)로 설정된 계산 방식을 도입한다. 이때 망각게이트는 이전 은닉상태의 값을 얼마나 잊어버릴지를 결정하고, 입력게이트는 반대로 얼마나 반영할지를 결정한다.

Bi-LSTM은 기존 순방향의 LSTM 출력과 더불어 역방향의 입력을 통한 출력을 함께 사용하며, Fig. 1에서 볼 수 있듯 본 연구 모델은 위 Bi-LSTM 모델을 사용하여 각 은닉상태(Hidden state)에서 최종으로 출력되는 값에 선형 결합 층(Dense layer)를 추가하여 활성화 함수인 softmax 함수를 적용한 후, 가장 확률이 높은 분류군을 도출하는 시퀀스 레이블링(Sequence Labeling)을 통해 각 입력 토큰에 대한 정답 개체명을 출력하는 구조이다.

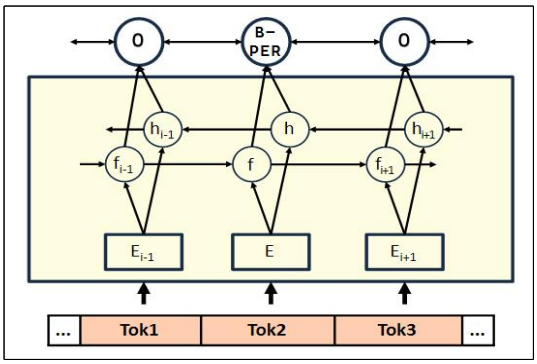

Fig. 1. Bi-LSTM model for NER in crime domain

3.3 사전학습 언어모델

BERT는 Transformer[18] 인코더 구조 기반의 사전학습 언어모델로, 두 개의 문장을 입력 받아 임의로 마스킹된 단어를 예측하는 문제(Masked Language Model)와 이전문장과 다음 문장이 이어지는 문장인지 예측하는 문제(Next Sentence Prediction)를 해결함으로써 대규모 말뭉치에 있는 양방향의 문맥 표현을 학습한다. BERT 모델의 등장 이후 사전학습된 언어모델의 파라미터를 이용해 감정분석, 기계독해, 개체명 인식, 의미역 인식 등 다양한 자연어 처리의 다운스트림 태스크(downstream task)에 미세조정 학습(fine-tuning)을 하는 방식이 기존 task-specific한 아키텍처보다 뛰어난 성능을 보임으로써 많은 분야에서 활용되고 있다. 한국어 역시 BERT의 구조와 동일한 KoBERT, KorBERT 등이 공개되면서 많은 한국어 자연어 처리에서 뛰어난 성능을 보여주고 있다.

ELECTRA[19]는 BERT와 마찬가지로 Transformer의 인코더 구조를 기반으로 하였으나, RTD(Replaced Token Detenction)라는 새로운 사전 학습 태스크 방식을 사용하였다. RTD는 Generator를 이용해 입력 토큰 중 일부를 가짜 토큰으로 바꾸고, Discriminator가 기존에 있던 토큰인지 생성해낸 토큰인지를 구별하는 이진 분류 문제를 학습함으로써, BERT 모델의 MLM 방식보다 학습의 효율성을 높인 사전학습 모델이다.

본 연구의 모델은 Fig. 2와 같이 사전 학습된 모델을 사용한 미세조정 학습이라는 동일한 구조를 가지고 있다. 사전학습 모델로는 구글에서 공개한 다국어 기반 BERT-base 모델과, 한국어 기반의 위키, 뉴스 데이터로 사전학습하여 공개된 KoBERT 모델 및 KoELECTRA 모델을 사용하며, Bi-LSTM 모델과 마찬가지로 각 입력 토큰의 최종 은닉 상태(hidden state) 값들에 선형 결합 층을 추가하여 softmax 함수를 적용한 후 가장 확률이 높은 분류군의 후보를 정답 개체명으로 출력하는 구조를 가진다.

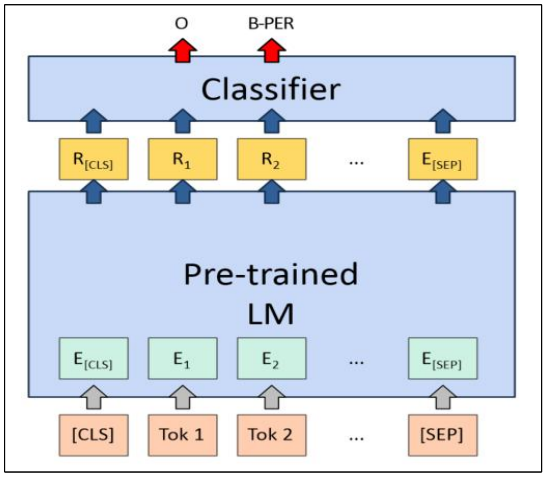

Fig. 2. Finetuning the pre-trained model for NER in crime domain

4. 실험 및 분석

4.1 데이터

본 연구는 범죄 수사 도메인 텍스트에서 새롭게 추가한 개체명의 분류 정확도를 살펴보는 실험을 진행하기 위해 자체 데이터셋을 구축하였다. 또한 개체명 분류 태스크를 9개의 메인 카테고리에 대한 분류 실험과

56개의 서브 카테고리에 대한 분류 실험으로 나누었으며, Fig 3과 같이 실험을 위해서 훈련(Train), 검증(Dev), 시험(Test) 데이터를 7:1:2 비율로 분리하였다. 훈련 데이터는 실제 파라미터 업데이트를 위해 사용되는 데이터이며, 검증 데이터는 학습 epoch 당 훈련의 성과를 검증하여 과적합을 방지하기 위해 사용하고, 시험 데이터는 학습이 완료된 모델에 대해 최종 정확도를 측정하기 위해 1번만 사용한다.

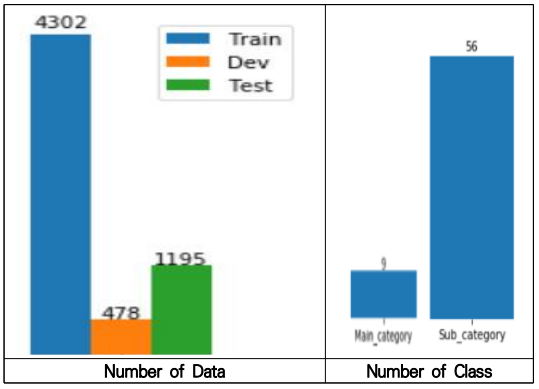

Fig. 3. Data Information

4.2 개체명 태깅

각 데이터는 형사판결문, 가상의 수사보고서에서 수집한 말뭉치 중 살인, 절도, 사기, 아동폭행, 강간 등의 죄종에 해당하는 범죄사실을 분리한 후, 새로 정의된 개체명 분류 범위 내에서 추출 가능한 개체들을 Table 2의 예시와 같이 태깅한 후 BIO 방식으로 토큰화하였다. 그 결과 본 실험 에서 범죄 도메인 개체명 분류 범위 내에서 실제 문장에서 각 개체명 별로 태깅된 개체의 수는 Fig. 4와 같다.

Table 2. BIO tagged texts

Task	Example
Main category	〈피해자 : PS〉를 〈살해 : CV〉하기로 마음먹고, 〈피해자 : PS〉가 식사를 하기 위하여 식당에 가려고 밖으로 나오자 미리 소지하고 있던 〈칼 : TMC〉(칼날길이 24.5㎝)로 〈피해자 : PS〉의 복부를 1회 찔러..
Sub category	〈피고인 : PS_COURT〉은 〈2008.7.29. : DT_OTHERS〉 〈11:30경 : TI_OTHERS〉 〈서울특별시 : LC_CITY〉 〈종로구 : LC_COUNTY〉 〈창신동 : LC_COUNTY〉 561-80에 있는 〈렛츠홈 홈피스텔 : AF_BUILDING〉 ***호에 사는 〈피해자 : PS_COURT〉 〈강** : PS_NAME_VICTIM〉 (〈47세 : QT_AGE〉)이 자신을 죽인다고 여기어

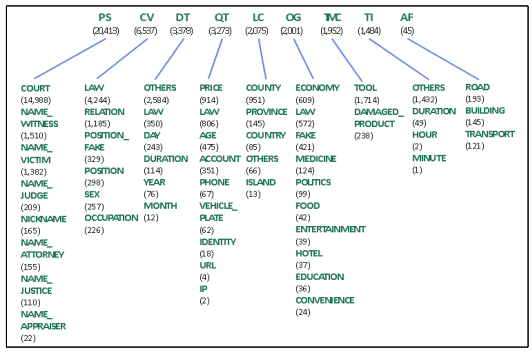

Fig. 4. Entity Statistics

4.3 평가지표

각 모델에 대한 실험 결과는 각 개체별 분류 결과에 대한 정확도(accuracy), 정밀도(precision), 재현율(recall), 그리고 정밀도와 재현율의 조화평균인 F1-score를 측정하여 평가한다. 또한 모든 개체 분류군별 샘플수를 고려한 분류 성능을 종합적으로 평가하기 위해 micro 평균과, macro 평균이라는 기준을 도입한다.

본 연구에서 설계한 각 모델은 토큰화된 문장을 입력으로 받아 토큰별로 가장 확률이 높은 label을 출력하게 된다. 이때 출력 결과는 같은 개체이더라도 BIO 태그에 의해 분리되어 있으므로, 이를 개체 단위(Entity-level)로 통합한 뒤 평가를 진행한다.

또한 개체 단위별 정확도, 정밀도, 재현율을 계산하기 위해서 Table 3과 같이 오차행렬(confusion matrix)이라는 개념을 활용한다. 오차행렬을 기준으로 각 개체에 대한 모델의 예측값과 실제값을 비교하여 실제값이 참(True)인지 거짓(False)인지, 예측값이 긍정(Positive)인지 거짓(Negative)인지에 따라 네 가지 경우 나누어 구분해 카운트하며 이 값의 통계치를 이용해 정확도, 정밀도, 재현율을 계산한다.

Table 3. Confusion matrix

		Actual Label	
		True	False
Predicted Label	Pos	TP	FP
	Neg	TN	FN

개체명 인식의 경우 각 개체 단위에 부여된 토큰들의 label을 모두 정확하게 예측하였을 경우 TP, 모든 토큰들을 잘못 예측하였을 경우 FN으로 카운트하며, 만약

한 개체의 토큰 중 일부만을 정확하게 예측하였을 경우에는 FN과 FP에 모두 해당하는 것으로 카운트한다. 오차행렬 구분법에 따른 결과로 정밀도, 재현율, 정확도, 그리고 F1-score의 계산 방법은 아래 수식과 같다.

$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{recall} = \frac{TP}{TP+FN}$$

$$\text{accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

$$\text{F1-score} = 2 \cdot \frac{preision \cdot recall}{preision+recall}$$

4.4 실험

본 연구는 범죄 수사 도메인 텍스트로 태깅된 데이터 셋에 대한 개체명 인식 성능 분석을 위해 9개의 메인 카테고리로 통합된 개체명에 대한 인식과 56개의 서브 카테고리로 구분된 개체명을 인식하는 두 가지 실험을 진행한다. 모든 실험은 구글 코랩의 TPU 사용환경에서 수행하였으며, BiLSTM모델의 경우 mecab 분석기를 통해 분리된 형태소 단위의 토큰을 입력으로 사용하였고, BERT-base, KoBERT, KoELECTRA 모델의 경우 해당 모델의 사전학습 시 사용된 자체 토크나이저를 동일하게 사용하였다. learning rate는 0.000002로 설정하였고, 학습 에포크는 10, 배치사이즈는 32로 설정하였다.

4.5 실험결과

Table 4와 Table 5에서 볼 수 있듯 메인 카테고리에 대한 실험과 서브 카테고리에 대한 실험에서 모두 가장 좋은 성능을 보인 KoELECTRA 모델은 메인 카테고리 실험에서 micro 평균 기준 precision 99%, recall 99%, F1 99%의 성능을 보이며, 서브 카테고리 실험에서는 micro 평균 기준 precision 98%, recall 98%, F1-score 98%의 성능을 보인다. 이는 사전학습 언어모델 기반이 아닌 Bi-LSTM 모델에 비해 메인 카테고리 실험에서는 precision 23%, recall 26%, F1-score 25%나 높은 성능을, 서브 카테고리 실험에서는 precision 21%, recall 24%, F1-score 23%나 높은 성능을 보인 것이다.

Table 4. Performance for main category prediction task

	acc	precision		recall		F1	
		micro avg	macro avg	micro avg	macro avg	micro avg	macro avg
BiLSTM	0.96	0.76	0.69	0.73	0.68	0.74	0.69
BERT-base	0.99	0.98	0.94	0.98	0.96	0.98	0.95
KoBERT	0.99	0.94	0.76	0.92	0.70	0.93	0.71
KoELECTRA	**0.99**	**0.99**	**0.95**	**0.99**	**0.97**	**0.99**	**0.96**

Table 5. Performance for sub category prediction task

	acc	precision		recall		F1	
		micro avg	macro avg	micro avg	macro avg	micro avg	macro avg
BiLSTM	0.96	0.77	0.51	0.74	0.46	0.75	0.47
BERT-base	0.99	0.97	0.65	0.97	0.60	0.97	0.61
KoBERT	0.99	0.92	0.33	0.89	0.28	0.90	0.29
KoELECTRA	**0.99**	**0.98**	**0.67**	**0.98**	**0.63**	**0.98**	**0.62**

4.6 토의

실험 결과는 범죄 분석관의 참여를 통해 정의된 다운스트림 태스크들에 대해 사전학습 된 딥러닝 언어 모델로 미세조정을 하면, 향후 전문 인력의 개입 없이도 비정형 데이터에서 다양하게 원하는 범죄 정보를 높은 정확도로 추출할 수 있다는 가능성을 보여준다.

다만 몇 가지 지표를 자세히 살펴보면 모델과 데이터의 개선 가능성을 분석할 수 있다. 우선 micro 평균 기준의 평가지표가 90% 이상의 점수로 높게 나왔던 데 비해, macro 기준의 평가지표는 가장 좋은 성능이 precision 67%에 불과했던 이유에 대해 살펴볼 필요가 있다. 범죄 사실은 문장의 특성 상 '〈피고인 : ps_court〉은 〈2011. 1. 1. : dt_year〉경 〈서울시 : lc_city〉에서'와 같이 시간·장소 등의 전형적인 사실 표현 문장이 많고, 범행도구나 특정 음식점명과 같이 비전형적인 표현이 있는 문장은 상대적으로 빈도가 적다. 따라서 개체명 분류군별로 정확도에 큰 편차가 있기 때문에 macro 평균 지표에서는 낮은 성능이 나온 것으로 보인다. 둘째로, 한국어 데이터셋 만을 가지고 학습한 KoBERT에 비해 오히려 BERT-base 모델의 성능이 좋게 나왔는

데, 이는 KoBERT 모델의 경우 어휘 집합의 수가 상대적으로 적었고, 이에 따라 토큰화 과정에서 언노운 토큰이 많이 발생했기 때문이다.

이 점을 볼 때, 범죄 도메인 텍스트에서 높은 성능을 보이는 모델을 개발하여 실제 활용에 이르기 위해서는 도메인에 맞는 말뭉치를 이용한 사전학습이 필요하고, 비전형적인 표현의 문장이 충분한 다운스트림 태스크용 데이터가 확보되어야 함을 시사해준다.

5. 결론 및 시사점

본 연구에서는 범죄 분석에서 다양한 목적으로 활용될 수 있는 범죄 정보에 대한 자동 추출을 위해 딥러닝 기반 자연어 처리 기술을 응용한 최초의 범죄 수사 분야 개체명 인식 시스템을 개발하였다. 특히 사전학습된 언어 모델인 KoELECTRA 모델을 이용하여 미세조정 학습을 수행한 결과, 9종의 메인 카테고리 분류에서 micro avg F1-score 99%, macro avg F1-score 96%의 성능을, 56종의 서브 카테고리 분류에서 micro avg F1-score 98%, macro avg F1-score 62%라는 가장 높은 성능을 얻을 수 있었다. 또한 모델별 정확도의 차이와 분류 결과에 대한 macro 평균 지표를 분석하여 향후 개선된 모델을 통해 시스템의 활용도를 높여야 할 필요성을 확인하였다.

본 연구는 최신 자연어 처리 기술을 적절히 응용한다면 범죄 예방과 수사의 효율성을 높임으로써 국내 치안에 이바지 할 수 있음을 시사한다. 따라서 후속 연구를 통해 범죄 수사 도메인에 특화된 충분한 데이터 확보를 바탕으로 더욱 높은 활용성을 가진 모델을 개발해 나갈 것을 기대한다.

REFERENCES

[1] H. Hassani, X. Huang & E. S. Silva. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining, 9(3),* 139-154.
DOI : 10.1002/sam.11312

[2] J. Devlin, M. W. Chang, K. lee & K. Toutanova. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *In proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1,* 4171-4186,
DOI : 10.18653/v1/N19-1423

[3] J. H. Lee, W. J. Yoon, S. D. Kim, D. H. Kim, S. K Kim, C. H. So & J. W. Kang. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36(4),* 1234-1240.
DOI : 10.1093/bioinformatics/btz682

[4] M. Chau, J. J. Xu & H. Chen. (2002). Extracting meaningful entities from police narrative reports. *In Proceedings of the 2002 Annual National Conference on Digital Government Research, Los Angeles*

[5] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, & M. Chau. (2004). Crime data mining: a general framework and some examples. *Computer, 37(4),* 50-56.

[6] C. H. Ku, IA. riberri & G. Leroy. (2008). Natural language processing and e-government: crime information extraction from heterogeneous data sources. *Proceedings of the 9th Annual International Digital Government Research Conference, Canada.* 162-170.

[7] R. Bache, F. Crestani, D. Canter & D. Youngs. (2007). Application of Language Models to Suspect Prioritisation and Suspect Likelihood in Serial Crimes. *Third International Symposium on Information Assurance and Security,* 399-404.
DOI : 10.1109/IAS.2007.58

[8] K. R. Rahem & N. Omar. (2014). Drug-related crime information extraction and analysis. *Proceedings of the 6th International Conference on Information Technology and Multimedia,* pp. 250-254.
DOI : 10.1109/ICIMU.2014.7066639

[9] A. Alkaff & M. Mohd. (2013). Extraction of naitonality from crime news. *Journal of Theoretical and Applied Information Technology, 54,* 304-312.

[10] S. Sathyadevan, M. S. Devan & S. S. Gangadharan (2014). Crime analysis and prediction using data mining. *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*, 406-412.
DOI : 10.1109/CNSC.2014.6906719.

[11] M. Asharef, N. Omar & M. Albared. (2012). Arabic named entity recognition in crime documents. *Journal of Theoretical and Applied Information Technology, 44(1),* 1-6.

[12] Arulanandam, R., Savarimuthu, B. T. R. & Purvis.

M. A. (2014). Extracting crime information from online newspaper articles. *Proceedings of the Second Australasian Web Conference, Auckland, New Zealand,* 31-38.

[13] J. Johnson, A. Miller, L. Khan, B. Thuraisingham, & M. Kantarcioglu. (2011). Extraction of expanded entity phrases. *Proceedings of the IEEE International Conference on Intelligence and Security Informatics, Beijing, China,* 107-112. DOI : 10.1109/ISI.2011.5984059

[14] K-S. Yang, C-C. Chen, Y-H. Tseng & Z-P. Ho. (2012). Name entity extraction based on POS tagging for criminal information analysis and relation visualization. *Proceedings of the 6th International Conference on New Trends in Information Science and Service Science and Data Mining (ISSDM), October, Taipei.* 785-789.

[15] P Gohel. (2016) *Crime information extraction from news articles.* M Tech Dissertations. Dhirubhai Ambani Institute of Information and Communication Technology. Gandhinagar.

[16] K. Srinivasa & P. S. Thilagam (2019). Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers. *Information Processing & Management, 56.* DOI : org/10.1016/j.ipm.2019.102059

[17] S. Hochreiter & J. Schmidhuber. (1997). Long short-ter memory. *Neural computation, 9(8),* 1735-1780.
DOI : 10.1162/neco.1997.9.8.1735

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez & I. Polosukhin. (2017). Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems,* 6000-6010.

[19] K. Clark, M. T. Luong, Q. V. Le & C. D. Manning. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.*

김 희 두(Heedou Kim) [학생회원]

· 2014년 3월 : 경찰대학 법학과 (법학사)
· 2020년 3월 ~ 현재 : 고려대학교 빅데이터융합학과 석사과정
· 관심분야 : 범죄예측, 머신러닝
· E-Mail : heedou123@police.go.kr

임 희 석(Heuiseok Lim) [종신회원]

· 1992년 2월 : 고려대학교 컴퓨터학과(학사)
· 1994년 2월 : 고려대학교 컴퓨터학과(석사)
· 1997년 2월 : 고려대학교 컴퓨터학과(박사)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 기계학습, 인공지능
· E-Mail : limhseok@korea.ac.kr