Journal of The Institute of Electronics and Information Engineers Vol.60, NO.9, September 2023 https://doi.org/10.5573/ieie.2023.60.9.75
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2023-60-9-10

BERT 기반 금융 텍스트셋 구현을 위한 전이 학습 연구

(BERT-based Transfer Learning Research for Financial Dataset Implementation)

김 학 진*

(Kim Hackjin©)

요 약

한국어 자연어 처리 분야는 구글에서 공개한 BERT 언어모델을 중심으로 활발한 연구가 진행되고 있으나, 한국어의 특성상 많은 제약을 갖고 있다. 개체명 인식(NER)은 정형화되지 않은 방대한 텍스트에 숨어있는 개체명을 검출하고, 이를 미리 정의한 개체명 클래스에 따라 분류하는 자연어 처리(NLP) 태스크 중 하나이다. 개체명 정보는 텍스트 내에서 특정 도메인과 관련된 지식을 이해하는 데에 실마리를 제공하는 어휘 특징(Lexical Feature) 중 하나로 먼저 텍스트 문서로 부터 정보를 추출(Information Extraction)하여 토큰화(Tokenization)와 품사 태깅과 같은 전처리 작업이 필요하다. 본 연구에서는 한국어 처리에 앞서 금융에 특화한 영문 코퍼스를 제시하고자 한다. Financial_phrasebank 데이터셋와 IIRC 회원 약 83개 글로벌 기업의 재정보고서에서 약 8천개의 금융 텍스트를 추출, 통합하여 총 약 12.7천개 문장으로 구성된 데이터셋을 구축하였으며, 전이학습 과정에서 분류되는 개체명 클래스를 기존의 7개 클래스 분류에서 15개 클래스로 분류, 확장한 언어 모델을 제안한다. 레이블링을 거쳐 BERT_base 모델을 결합하여 모델 학습을 진행한 후 정확도, 재현율, F1 스코어 등의 검사를 통해 금융 분야에서의 최적의 데이터셋을 제안한다.

Abstract

In the field of Natural Language Processing (NLP) for Korean, research has been actively conducted centering on the BERT language model introduced by Google. However, its application to Korean still has some limitations due to the nature of this language. Named Entity Recognition (NER) is one of the NLP tasks that detects entity names represented in large amounts of unstructured text and classifies them according to predefined entity classes. As one of the lexical features, entity information provides a clue to understanding of domain-specific knowledge within a text. Extracting the entity information in text typically requires preprocessing, including tokenization and part-of-speech tagging. In this study, we present an English corpus specialized in finance for Korean language processing. We extracted and integrated about 8,000 financial texts from the Financial_phrasebank dataset and financial reports of about 83 global companies that are members of IIRC to build a dataset consisting of a total of about 12.7 thousand sentences. We propose a language model that extends the classification of object names from 7 classes to 15 classes in the process of transfer learning. After labeling and training the model by combining the BERT_base model, we propose an optimal dataset for the financial field through accuracy, recall, and F1 score.

Keywords : Natural language processing, Named entity recognition, BERT, Transfer learning, Financial dataset

Ⅰ. 서 론

자연어 텍스트의 이해는 텍스트를 분해한 후 재구성하는 역동적인 과정을 필요로 하는 것으로 연구되어 왔다. 이러한 과정은 두 단계로 구분할 수 있다. 먼저, 구문 구조의 분석을 통하여 텍스트의 의미를 파악하거나 보다 복잡한 의미를 이해하기 위해 상식을 적용하는 것이다[1]. 특히, 특정 분야에서 전문적인 맥락으로 사용되는 어휘 또는 개체명 등이 텍스트에 포함된 경우에 이 과정은 고도화 되어야 하는 필요성이 있다[2]. 이는 특정 분야에서의 텍스트의 이해를 위해 보다 전문적인 지식이 추가적으로 요구되기 때문이다. 그 예로, 금융 분야의 텍스트에서는 그 의미를 이해하는 데에 핵심이

*정회원, 서일대학교 컴퓨터전자공학과(Dept. of Computer Applications, Seoil University)
© Corresponding Author(E-mail : 20220010@seoil.ac.kr)
Received ; July 21, 2023 Revised ; July 28, 2023
Accepted ; August 1, 2023

되는 금융 분야에 한정된 개념과 용어가 빈번하게 발견된다. 또한 금융 관련 개체명은 종종 변형되거나 축약된 형태로 뉴스 기사 등에 사용되기도 한다. 이러한 변이된 형태의 개체명은 그 개체명이 언급된 문장 자체 내에서 의미를 파악하기보다는 텍스트 전체의 문맥이 고려되어야 정확한 이해가 가능하다[3]. 따라서, 독특한 개체명 및 언어의 사용은 텍스트의 이해를 보다 복잡하게 한다. 최근 개체명 인식 모델의 비약적인 발전에도 불구하고 금융 분야에서의 적용은 그 언어적 특성과 학습 데이터 부족으로 인해 여전히 어려움이 있는 것으로 파악된다. 이전 연구에서 Malo 등[4]은 금융 및 경제 영역에서 코퍼스(Corpus) 구축이 상대적으로 더딘 노력을 보이고 있음을 지적했다. 또한 개체명이 사전 레이블링된 코퍼스는 드물고, 경우에 따라 저작권 문제로 인해 모델 학습에 활용하기 어렵다[5]. 더욱이 금융 및 경제 분야에서 최근 현상을 반영하는 새로운 용어 및 개체명의 지속적인 탄생 역시 텍스트 이해의 어려움을 심화시킨다. 이러한 특성을 고려하여 금융 분야 개체명 인식은 일반 텍스트에 특화된 모델과 다르게 연구되어야 할 필요성이 있다.

최근 국내에서도 도메인에 특화된 언어 모델의 연구가 활발히 시도되고 있으며 대표적인 언어 모델로는 의학 및 컴퓨터 과학 분야 언어 모델인 SciBERT[6], 금융 분야 특화모델인 FinBERT[7, 8]와 KB-BERT[9], 의생명 분야의 BioBERT[10], 특허 분야에 patentBERT[11]와 KorPatBERT[12], 과학기술 분야 KorSciBERT[13]와 뷰티 분야의 언어모델인 HwahaePLM[14] 등의 연구가 진행되고 있다. 본 연구에서는 금융 분야에 적용 가능한 언어 모델인 eFinBERT(English Financial BERT)를 구축하고자 한다. eFinBERT는 금융 분야의 전문 문서를 정확하게 인식하기 위한 언어 모델이다. 먼저, 금융 분야의 코퍼스를 구축하면서 언어 모델을 학습하기 위한 개체명 레이블링 스킴(Labelling Scheme)을 제안하고자 한다. 선행연구에서 사용된 다양한 스킴들을 참조하여 개체명의 분류와 개념을 통합 및 단일화하여 금융 분야에 특화한 레이블링 스킴을 제안하고, 사전학습 언어 모델에 필요한 코퍼스로는 Financial_Phrasebank 데이터셋[4]을 기본으로, 추가 확장은 국제통합보고위원회(International Integrated Reporting Council,이하 IIRC)에 신고하는 기업의 재정보고서에서 문장을 추출하고 수작업의 사전처리를 거친 후 두 코퍼스를 통합한 언어 모델을 구축함으로써 금융 분야에 특화한 개체명 인식의 성능 향상을 제고하고자 한다.

Ⅱ. 이론적 배경

1. 개체명 인식

개체명(Named Entity)이란 사람, 장소, 기관 등 고유한 명칭을 포함하는 구를 의미한다[15].

개체명 인식(Named Entity Recognition, 이하 NER)은 정형화되지 않은 텍스트에 나타난 개체명을 구별하고[16], 이를 미리 정의한 개체명 클래스에 따라 분류하는 자연어 처리(Natural Language Processing, 이하 NLP) 태스크 중 하나이다. 또한 개체명 정보는 텍스트 내에서 특정 도메인과 관련된 지식을 이해하는 데에 실마리를 제공하는 어휘 특성(Lexical Feature) 중 하나이다[17]. 이러한 개체명 정보를 인식하는 과정은 정보추출(Information Extraction), 질의응답 시스템(Questionon Answering System) 등에 활용되며, 다음과 같은 단계로 이루어진다. 첫 번째 단계는 텍스트 문서의 토큰화(Tokenization)와 품사 태깅과 같은 전처리 작업이다. 이후 적절한 분류 알고리즘을 적용하여 개체명을 식별하는데 대체로 지도학습이나 비지도학습 알고리즘을 적용한다.

지도학습기반의 NER은 대용량의 레이블링된 데이터셋을 사용하여 모델을 학습시키는 방식이다. 대표적인 지도학습 알고리즘으로는 Conditional Random Fields (CRF)[18]와 Hidden Markov Models(HMM)[19]이 있다. 이 알고리즘들은 주어진 문장 내에서 개체를 특정 레벨(label)에 할당하기 위해 토큰에 대한 특성을 추출하고, 이를 이용하여 모델을 학습시킨다.

비지도학습기반의 NER은 레이블링이 되지 않은 데이터셋에서 개체를 식별하는 방식이다. 이 방식은 대개 텍스트 규칙과 통계적 기법을 사용하는 방식으로, 개체의 특성을 추출하여 그룹핑하는 것으로, 기존 데이터셋에서 발견되지 않은 새로운 개체에 대해서도 적용될 수 있다는 장점을 갖고 있다.

2. BERT 언어 모델

BERT(Bidirectional Encoder Representations from Transformers) 언어 모델은 NLP 분야에서 매우 성능이 우수한 모델 중 하나이다. BERT는 Transformer 모델을 기반으로 하며, 양방향(bidirectional)으로 입력 시퀀스를 처리하는 방식을 통해 다양한 자연어처리 문제에 적용될 수 있다.

BERT의 전이학습은 미리 학습된 모델을 다른 자연어 처리 문제에 미세조정(Fine-tuning)하는 방식으로

동작한다. 미리 학습된 모델은 대규모의 텍스트 데이터를 학습한 후, 해당 모델 파라미터를 저장하고 이후 새로운 자연어 처리 문제가 발생하면 전이학습 모델은 새로운 데이터에 맞게 재학습을 거치는 것이 아니라, 이전에 학습한 모델 파라미터를 재사용하여 학습을 수행한다[20].

BERT 모델에서 분류 작업은 미세조정 단계에서 이루어진다. 분류 작업에서는 입력 문장이 특정 클래스에 속하는지를 예측하는 것으로 입력 문장의 첫 번째 토큰은 [CLS] (Classification)로 대체되며, 마지막 토큰은 문장의 끝을 나타내는 [SEP] (Separator)으로 대체된다. 새로운 데이터에 대한 전이학습은 다음과 같이 수행된다. 첫째, 미리 학습된 BERT 모델에 새로운 데이터를 입력한다. 둘째, 입력 데이터는 [CLS] 토큰과 [SEP] 토큰을 붙여서 입력 토큰으로 변환된다. 셋째, 입력 토큰은 BERT 모델의 pre-trained 파라미터를 활용하여 출력 벡터로 변환된다. 마지막으로, 출력 벡터는 추가적인 연산을 거쳐 각 클래스별 확률값을 계산하게 된다[20].

3. 전이 학습(Transfer Learning)

최근 기계 학습(Machine Learning)의 패러다임은 다양한 분야에서 보다 정확한 예측과 그에 따른 의사결정을 가능하게 하는 방향으로 발전해 왔다. 기계 학습 방법론은 이미 존재하는 데이터에서 구조적인 패턴을 찾아내고, 이를 논리적으로 검증 및 추론하는 과정에 방점을 두고 있다[21]. 이와 같은 맥락에서 양질의 학습 데이터, 특히 적용하려는 분야의 태스크와 동일한 분포 형태를 보이는 학습 데이터를 획득하는 것은 기계 학습에 있어 매우 중요하다. 하지만, 이러한 데이터를 수집하고 가공하여 연구에 적용하는 데에는 많은 자원을 필요로 하며, 이미 활용이 가능한 오픈 데이터 역시 많지 않다[22].

전이 학습(Transfer Learning)은 기계 학습의 하나의 방법론으로, 상이한 특성 공간과 분포를 지닌 데이터를 기반으로 이미 학습된 모델을 다른 태스크를 위해 학습 시키는 방식을 의미한다[23]. 이러한 방식은 학습 데이 터와 목표 태스크의 분포와는 상관없이 이미 모델이 학습한 결과를 다른 모델을 학습 시키는 데에 사용할 수 있다는 장점이 있다[24]. 특히, 활용 가능한 학습 데이터가 희박한 경우에 그 효용성이 더욱 부각된다. 이전 연구에서도 전이학습을 이용하면 적은 양의 레이 블링 된 학습 데이터로도 모델의 성능을 향상 시킬 수 있음을 증명하였다[24]. 전이학습이 학습 데이터를 획득 하고 가공하는 과정에서 많은 비용과 시간을 절감할 수 있는 효과적인 해결 방안이 될 수 있음을 시사하고 있다.

Ⅲ. 연구 방법

1. 코퍼스 구축

본 연구에서는 금융 분야에 특화된 코퍼스를 구축하기 위해 Malo 등이 수집한 영어 금융 뉴스 중심으로 약 4,846개 문장으로 구성된 Financial_Phrasebank 데이터셋[4]을 기본으로 채택하고, 여기에 국제통합보고위원회(IIRC)에 신고하는 글로벌 기업 중 83개 기업의 재정보고서에서 약 7,814천개의 문장을 추출하여 소문자 변환, 중복 및 불용어 제거 등 사전처리를 수작업으로 진행한 후 Financial_ Phrasebank 데이터셋과 통합하여 총 약 12.7천개 문장으로 구성된 데이터셋을 완성하여, eFinBERT 언어 모델 개발에 활용하였다. 아래 표 1은 구축된 데이터셋의 현황을 보여주고 있다.

표 1. 데이터셋 구축
Table 1. Building a dataset.

Classification	No. of Sentences	Lan.	No. of Tokens	Period
Corporate Financial reports	7,814	Eng.	213K	2021
Financial_ Phrasebank	4,846	Eng.	115K	2013
Total	12,660		328K	

2. 개체명 레이블링

모델 학습을 위한 데이터 구축을 위해 수집된 코퍼스에 토큰 단위의 개체명 레이블링을 진행하였다. Chiinchor와 Robinson이 1998년 MUC-7에서 소개한 개체명 어노테이션 가이드 라인[25]에서 아래 표 2와 같이 개체명 클래스를 7개로 제시하였다.

표 2. 7분류 개체 클래스[4]
Table 2. Class 7 entity class[4].

Entity Type	Labels	Description
Organization	ORG	기관
Person	PER	인물
Location	LOC	장소
Date	DAT	날짜
Time	TIM	시간
Money	MON	화폐
Percent	PER	퍼센트

본 연구에서는 이 가이드라인을 기반으로 레이블링 분류를 적용하였다. 추가적으로 개념(Concept) 클래스를 도입하였다. 개념 클래스는 경제지표, 문서의 장르, 유형, 학문 주제와 같은 속성을 포괄하는 클래스로 특히 금융 분야에서는 이들을 분류하기 위한 클래스가 필요하다[26]. 이에 따라 기존의 7분류에 개념 클래스를 추가하여 금융 텍스트 NER에서 유용성을 확인해 보고자 한다. 또한 클래스가 늘어나면 모델이 복잡해지는 것을 방지하고자 금액 클래스와 퍼센트 클래스를 단일 클래스로 통합하였으며, 레이블링이 필요하다고 판단되나 7분류 클래스에 해당되지 않는 경우를 고려하여 기타 클래스를 추가하였다.

또한 개체명 어노테이션 가이드 라인으로 두 개 이상의 토큰이 결합되어 개체명을 이루는 경우 분할의 어려움이 있어, 이를 보완하기 위해 개체명을 인식 방법 중 가장 보편적인 방법으로 이전 연구에서 소개된 BIO 태깅 스킴[27]을 적용하였다. 두 개 이상의 토큰이 결합되어 하나의 개체명을 가리키는 경우에 개체명이 시작되는 토큰은 Begin을 의미하는 B로, 뒤에 따라 오는 토큰들은 Inside를 의미하는 I로 하였으며, 개체명에 포함되지 않는 토큰은 Outside를 의미하는 O로 태깅하는 방식으로 그 경계를 구분하고자 표 3과 같이 15개의 개체 분류 모델을 제안하였다.

주요 분류로는 인물(Person), 장소(Location), 날짜(Date), 기관(Organization), 가치(Value), 개념(Concept), 기타(Miscellaneous)등 7개 클래스와 BIO 태깅 스킴을 결합한 총 15개(클래스별 2개 태그 및 'O'태그) 클래스로 분류하여 학습 데이터 구축을 위한 개체명 레이블링에 사용하였다.

표 3. 제안한 15분류 개체 클래스(eFinBERT)
Table 3. Proposed 15 Classified Entity classes (eFinBERT).

Entity Type	Labels	Description
Organization	B-ORG I-ORG	기관
Person	B-PER I-PER	인물
Location	B-LOC I-LOC	장소
Date	B-DATE I-DATE	날짜
Value	B-VAL I-VAL	가치
Concept	B-CONC I-CONC	개념
Miscellaneous	B-MISC I-MISC	기타
Outside	OUT	미포함

Ⅳ. 실 험

본 연구에서는 정형화되지 않은 금융관련 영문 텍스트에서 기관, 인물, 장소, 날짜, 가치, 개념, 기타 등 7가지의 개체명을 인식하는데 필요한 금융 데이터셋을 구축하고, 언어 모델로 15분류 개체 클래스를 제안한다. 데이터셋의 검증을 위해 개체명 인식 실험을 통해 검증하였다.

1. 실험 배경

제안한 NER 모델 학습 및 데이터 전처리와 원시 코퍼스 정리 및 변환은 오픈소스 플랫폼인 Jupyter Notebook에서 실행하였다. 개체명 레이블링에는 INCEpTION[28] 프로그램을, 모델 학습은 Sterbak[29]이 개발한 코드를 활용하였다. 사전 학습된 모델은 규모와 특성을 고려하여 최첨단 BERT 모델 중 BERT_base 모델을 활용하였다. BERT_base 모델은 전이 학습 작업에 있어 BERT_large 보다 NER 작업에서 더 유용하다고 권

```
import os
import pandas as pd
import spacy

# DEFINE THE PATH OF CORPUS FILE
path_corpus = "fin_corpus.csv"
DIR_PATH_CORPUS = os.path.dirname(path_corpus)
FILE_PATH_CORPUS = os.path.join(DIR_PATH_CORPUS, "fin_corpus.csv")

# REMOVE NEW LINE CHARACTER & LINE SEPARATOR
df_corpus = pd.read_csv(FILE_PATH_CORPUS, sep=";", encoding="utf-8")
df_corpus = df_corpus.replace(u"\n", "", regex=True).replace(u"\u2028", " ",
regex=True)

# TRANSFORM CORPUS USING SPACY NLP PIPELINE
nlp = spacy.load("en_core_web_sm")
nlp.max_length = 2000000

df_corpus["text"] = df_corpus.loc[0:, "text"].apply(lambda x: [sent.text for
sent in nlp(str(x)).sents])
df_corpus = df_corpus.explode("text", ignore_index=True)
df_corpus.index.name = "sentence #"
df_corpus = pd.DataFrame(df_corpus)
```

그림 1. 전처리 관련 코드
Fig. 1. Preprocessing related code.

장하고 있다[20]. 모델 학습의 batch 크기는 32로 설정하고 미세 조정을 하면서 3회에 걸쳐 수행하였으며, 학습률은 $3e$-5로 기본 최적화 도구인 Adam을 사용하여 진행하였다[20]. 아래 그림 1은 제안한 금융 코퍼스의 전처리에 필요한 코드를 보여주고 있으며, 그림 2는 금융 코퍼스의 개체명 레이블링 처리에 필요한 코드를 보여주고 있다.

```
# TRANSFORM THE DATA EXPORTED FROM INCEPTION
def clean_inception_data(data, mainUser, otherUserList):

    data = data.drop(["Type", "Collection", "Document", "Layer", "Feature"],
axis=1)
    data = data.drop(otherUserList, axis=1)
    data["Sentence #"] = None
    data["POS"] = None
    data = data.rename(columns = {"Position": "Word"})
    data = data.rename(columns = {mainUser: "Tag"})
    data = data[["Sentence #", "Word", "POS", "Tag"]]

    # ASSIGN SENTENCE NUMBERS
    n = 1

    for i in range(len(data)):
        if ("[.]" or "[?]" or "[!]") in data["Word"][i]:
            data["Sentence #"][i] = "Sentence: " + str(n)
            n += 1
            i += 1
        else:
            i += 1

    data = data.fillna(method="bfill")

    for i in range(len(data)):
        if data["Sentence #"][i] == None:
            data["Sentence #"][i] = "Sentence: " + str(n)
        else:
            i += 1

    # CLEAN INCEPTION DATA FORMATTING
    for i in range(len(data)):
        data["Word"][i] = data["Word"][i][data["Word"][i].find("[")+1 :-1]
        i += 1
```

그림 2. 개체명 레이블링 관련 코드
Fig. 2. Code related to named entity labeling.

2. 실험 결과

1) 금융 코퍼스 구축

IIRC의 기업 재정보고서와 Malo 등의 Financial-Phrasebank [4] 데이터셋을 결합한 후 개체명 레이블링을 완료한 코퍼스는 약 32천개의 개체명을 갖으며, 약 328천개의 토큰으로 구성되었다. 아래 표 4는 제안한 언어모델의 코퍼스 구성 현황을 보여주고 있다.

그림 3은 INCEpTION 프로그램에서 진행한 금융 코퍼스의 레이블링 실행 화면을 보여주고 있다.

2) 개체명 인식 검증

제안한 모델의 성능은 기존 연구에서 활용했던 정량적 분석의 하나인 정확도(Precision), 재현률(Recall) 및 F1스코어를 활용하였다. 정확도는 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율[30]이며, 재현율은 True인 것 중에서 모델이 True라고 예측한 것의 비율이다[30]. F1 스코어는 데이터 레이블이 불균형일 때, 모델의 성능을 정확하게 평가할 수 있다[31].

표 4. 제안한 금융 코퍼스 구성 현황
Table 4. Proposed financial corpus construction status.

Entity Type	Labels	No. of Corpus	No. of Tokens
Organization	B-ORG	7,208	7,208
	I-ORG		5,468
Person	B-PER	993	993
	I-PER		851
Location	B-LOC	4,155	4,155
	I-LOC		1,292
Date	B-DATE	5,244	5,244
	I-DATE		6,840
Value	B-VAL	6,682	6,682
	I-VAL		8,425
Concept	B-CONC	5,850	5,850
	I-CONC		4,810
Miscellaneous	B-MISC	628	628
	I-MISC		384
Outside	OUT	1,482	269,715
Total		32,242	328,545

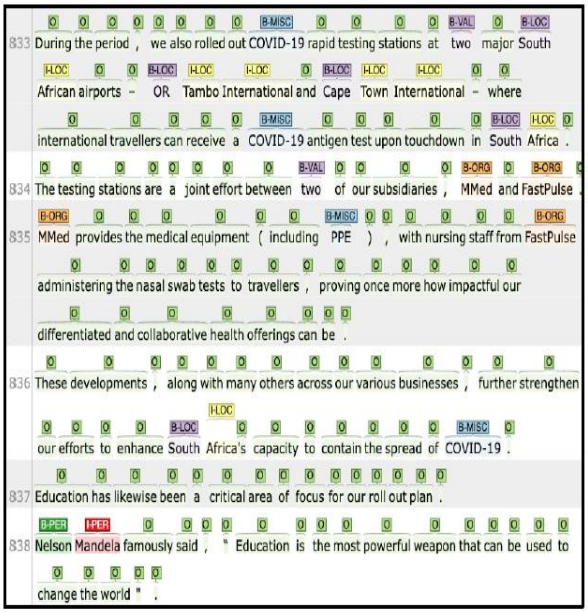

그림 3. 개체명 레이블링
Fig. 3. Named entity labeling.

$$Precision = True\ Positive\ Accuracy = \frac{TP}{TP+FP}$$

$$Recall = True\ Positive\ Rate = \frac{TP}{TP+FN}$$

$$F1 = \frac{2PR}{P+R}$$

* TP(True Positive), FP(False Positive), FN(False Negative)

도메인 적응 훈련 결과는 표 5에서 보여주고 있다 제안한 언어모델(eFinBERT)에 대한 정확도, 재현율, F1 스코어를 측정하였다. 정확도, 재현율 및 F1 스코어의 전체 평균은 0.85, 0.88 및 0.86를 기록하였다. 정확도에서는 가치, 장소, 날짜 등에서 높게 나타났으며, 재현율과 F1 스코어에서는 가치, 장소, 날짜, 기관 등에서 높은 성능을 보이고 있다. 정확도와 F1 스코어에서 가장 낮은 성능을 보이고 있는 것은 개념 클래스로 개념 클래스에는 일부 감정, 감성, 문서 명, 전문 용어 등을 묘사하는 도메인으로 이들을 표현하는 용어의 경우의 수가 다른 도메인보다 많이 있었으며, 특히 개체명이 변이 형태로 사용되는 경우가 상대적으로 빈번하게 발견되었는데 이 같은 현상이 정확도에 일부 영향을 미친 것을 확인하였다. 이에 대한 성능을 향상시킬 수 있는 별도의 모델이 필요하다.

표 6는 IIRC 데이터셋에 대한 성능 검증 결과로 정확도, 재현율, F1 스코어 등 모든 부분에서 제안한 언어모델보다 약 0.2~0.3정도 떨어지는 결과를 확인하였다. IIRC 데이터셋의 사전 처리과정에 수작업으로 많이 처리하면서 텍스트 관리에 보다 세심한 주의가 필요함을 확인할 수 있었다.

그림 4는 INCEpTION[28] 프로그램을 활용하여 12.7천개의 문장에 대한 레이블링을 수행한 후 개체명의 인식 검증을 수행한 결과 값을 Jupyter Notebook에서 직접 확인할 수 있었다.

표 5. 제안한 언어모델(eFinBERT)
Table 5. Proposed Language Model (eFinBERT).

Entity Type	Precision	Recall	F1	Total
Concept	0.62	0.73	0.67	720
Date	0.86	0.93	0.89	576
Location	0.86	0.88	0.87	512
Miscellaneous	0.89	0.77	0.82	205
Organization	0.85	0.88	0.87	1,691
Person	0.85	0.83	0.84	149
Value	0.95	0.94	0.95	1,263
Average	0.85	0.88	0.86	5,116

표 6. IIRC 데이터셋 검증
Table 6. IIRC Dataset Verification.

Entity Type	Precision	Recall	F1	Total
Concept	0.54	0.63	0.58	523
Date	0.84	0.92	0.88	380
Location	0.83	0.86	0.85	185
Miscellaneous	0.98	0.92	0.95	200
Organization	0.85	0.88	0.87	646
Person	0.87	0.92	0.89	71
Value	0.87	0.89	0.88	153
Average	0.82	0.86	0.84	2.158

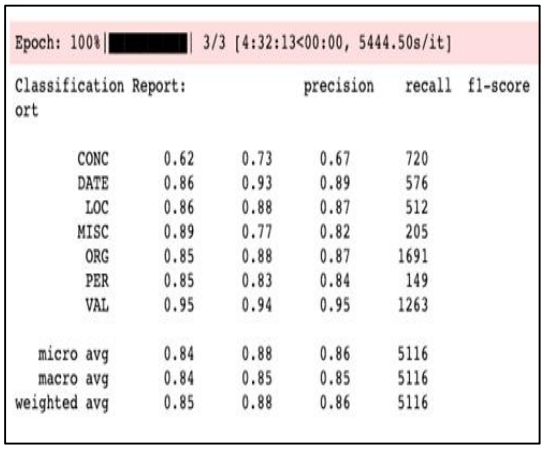

그림 4. Jupyter Notebook에서의 모델학습 결과화면
Fig. 4. Model training result screen in Jupyter Notebook.

3) 성능 평가

본 절에서는 제안한 eFinBERT의 성능을 평가하기 위하여 BERT 영문 모델로 NER을 위한 금융 감성 분석 모델인 FinBERT[7], 한국어 모델로는 군사 분야 언어 모델인 MIL-BERT[32], 금융 분야 언어 모델인 KB-BERT[9] 등과 정확도 및 f1 스코어를 비교하였다.

성능 평가 결과는 표 7에서 확인할 수 있다. 정확도에서 제안 모델은 영문모델인 FinBERT 모델보다는 낮은(0.02) 수준을 보였으나, 한국어 모델인 KB-BERT보다는 일부 우수(0.01)한 성능을 보였으며, f1 스코어에서

표 7. 성능 평가 결과
Table 7. Performance evaluation results.

Model	Precision	F1	Domain
BERT_base	0.84	0.83	General
FinBERT	0.87	0.85	Emotions
MIL-BERT	0.94	0.94	Military
KB-BERT	0.84	0.92	Financial
Proposed Model	0.85	0.86	Financial

는 KB-BERT(0.92), 제안 모델(0.86), FinBERT(0.85) 순으로 성능 결과를 확인하였다. 군사 분야 MIL-BERT는 정확도, f1 스코어에서 모두 좋은 성능을 보이고 있으나 자연어 처리 방법이 아닌 키워드 검출의 방법으로 언어 모델을 구현하고 실험한 결과로 본 연구에서 직접 비교 검증은 하지 않았다.

V. 결 론

본 연구에서는 금융 분야에 특화된 코퍼스를 구축하기 위하여 기존연구에서 발표된 코퍼스를 미세 조정하여 새로운 금융 언어 모델을 만들고 성능을 검증하였다. Financial_Phrasebank 데이터셋[4]과 국제통합보고위원회(IIRC)에 신고하는 기업의 재정보고서에서 문장을 추출하여 소문자 변환, 중복 및 불용어 제거 등 사전 처리를 수작업으로 진행한 후 두 개의 코퍼스를 통합함으로써 문장 수 12,660개, 토큰 수 328K개의 eFinBERT 언어 모델을 구축하였다. 성능평가 결과 정확도 0.85, f1 스코어 0.86이라는 성능을 확인하였다.

자연어 텍스트에는 문맥 또는 도메인에 따라 고유한 용어 사용이 두드러지며, 방대한 양의 블로그, 기사 등 다양한 텍스트들이 웹 사이트, 웹 포털 및 소셜 미디어에 끊임없이 생산 및 게시되고 있다. 따라서, 자연어 처리 모델 또한 필요에 따라 조정될 필요성이 있다. NER은 다양한 텍스트에 포함된 사람, 조직, 장소 등의 정보를 찾고, 특히 긴 콘텐츠에서 주제적 정보를 제공할 수 있는 개체명을 인식 및 분류하는 데에 적합한 도구이다. 또한, NER은 텍스트 요약, 정보 검색, 질문 답변 시스템, 의미 체계 구문 분석 등 NLP 작업에도 사용될 수 있다

본 연구에서는 자연어 처리가 가장 어렵다는 한국어 자연어 처리를 좀 더 용이하게 접근해 보고자 먼저 영문 코퍼스를 통해 검증하였다. 한국어 인식의 어려움은 토큰화에서 시작된다. 영어는 띄어쓰기를 기준으로 토큰화를 수행해도 유용하게 활용할 수 있으나, 한국어는 띄어쓰기만으로는 토큰화에 어려움이 있다. 한국어는 띄어쓰기 단위가 되는 단위를 “어절”이라고 하는데 먼저, 어절 토큰화의 개념에 대한 이해가 필요하며, 또 다른 형태로 영어의 단어 토큰화와 유사한 형태인 형태소 토큰화를 수행해야 한다. 형태소(morpheme) 중 자립형태소를 토큰화해야 영어의 단어 토큰화와 유사한 결과를 얻를 수 있다. 또한 품사 태깅(Part-of-speech tag ginging)의 어려움이다. 동음이의어, 다의어 등을 토큰화 과정에서 어떤 품사로 활용할 것인지 구분하는 연구 등이 함께 필요하다.

향후 연구로는 한국어 자연어 처리 연구의 일환으로 한국어 금융 NER 데이터셋 모델의 구축을 위한 연구를 진행하고자 한다. 현재 국내 NER 데이터셋은 다소 부족한 상태로 현재 국립국어원 NER 데이터셋, 한국해양대학교 자연어처리 연구실 NER 데이터셋, NAVER NLP Challenge 2018 NER 데이터셋 등이 연구에 활용할 수 있는 범주에 있다.

REFERENCES.

[1] Basu, K. et al. (2021) “Knowledge-driven Natuural Language Understanding of English Text and Its Applications”, *Proceedings of the AAAI Coference on Artificial Intelligence, 35(14), pp. 12554-12563. https://doi.org/10.1609/aaai.v35i14.17488.*

[2] Heuss, T. et al. (2014) “A comparison of NER tools w.r.t. a domain-specific vocabulary”, *in Proceedings of the 10th International Conference on Semantic Systems-SEM’14. the 10th International 75 Conference, Leipzig, Germany: ACM Press, pp. 100-107. https://doi.org/10.1145/2660517.2660520*

[3] Wang, S. et al. “Financial named entity recognition based on conditional random fields and information entropy”, *in 2014 Interational Conference on Machine Learning and Cybernetics (ICMLC), Lanzhou, China: IEEE 2014, pp. 838-843.*

[4] Malo, P. et al. “Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts”, *arXiv.1307.5336 (2013)*

[5] Remy, P. and Ding, X. “Financial News Dataset from Bloomberg and Reuters”. *https://github.com/Philipperemy/financial-news-dataset*

[6] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A Pretrained Language Model for Scientific Text,” *arXiv preprint arXiv:1903.10676, 2019.*

[7] Dogu Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models”, *submitted in partial fulfillment for the degree of master of science“, university of amsterdam 2019-06-25*

[8] Y. Yang, M. C. S. Uy, and A. Huang, “Finbert: A Pretrained Language Model for Financial Comm.” *arXiv preprint arXiv:2006.08097, 2020.*

[9] D. G. Kim, D. W. Lee, J. W. Park, S. W. Oh, S. J. Kwon, I. Y. Lee, and D. W. Choi, "KB-BERT : Training and Application of Korean Pre-trained Language Model in Financial Domain," *Korea Intelligent Information Systems Society, Vol. 28, No. 2, pp. 191-206, 2022.*

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics, Vol. 36, No. 4, pp. 1234-1240, 2020.*

[11] J. S. Lee and J. Hsiang, "Patentbert: Patent Classification with Fine-tuning A Pre-trained Bert Model", *arXiv preprint arXiv:1906.02124, 2019.*

[12] J. W. Park, W. Chul. Sim, S. Hun. Lee, B. S. Ko, and H. Sung. Noh, "A Study on Automatic CPC Classification based on Korean Patent Sentence-\A Deep Learning Approach using Artificial Intelligence Language Model KorPatBERT-\," *Korea Institute of Intellectual Property, Vol. 17, No. 3, pp. 209-256, 2022.*

[13] KISTI (2021). KorSciBERT [Online]. *Available: https://doi.org/10.23057/46.*

[14] Hwahae (2021). Hwahae PLM [Online]. Available: http://blog.hwahae.co.kr/all/tech/tech-tech/5876/.

[15] Tjong Kim Sang, E.F. and De Meulder, F. "Introduction to the CoNLL2003 Shared Task : Languguage Independent Named Entity Recognition", *in Proceedings of the in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142-147.*

[16] Pasca, M. "Acquisition of categorized named entities for web search", *in Proceedings of the Thirteenth ACM conference on Information and Knowledge management-CIKM'04. the Thirteenth ACM conference, Washington, D.C.USA : ACM Press, 2014, p. 137.*

[17] Chen, Y. et al. (2015) "Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks", *in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers). ACL-IJCNLP 2015, Beijing , China : Association for Computational Linguistics, pp. 167-176. https://doi.org/10.3115/v1/P15-1017.*

[18] Sutton, C. and McCallum, A. (2010) "An Introduction to Conditional Random Fields" *.arXiv.100.arXiv.1011.4088. https://doi.org/10.48550/arXi*

[19] Rabiner, L. and Juang, B. (1986) "An introduction to hidden Markov models", *IEEE, ASSP Magazine, 3(1), pp. 4-16. https://doi.org/10.1109/MASSP.1986.1165342.*

[20] Devlin, J. et al. (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *arXiv. pp. 2-3. http://arxiv.org/abs/1810.04805.*

[21] Witten, I.H., Frank, E. and Hall, M.A. (2011) "Data mining: practical machine learning tools and techniques". *3rd ed. Burlington, MA: Morgan Kaufmann (Morgan Kaufmann series in data management systems).*

[22] Pan, S.J. and Yang, Q. "A Survey on Transfer Learning", *IEEE Transactions on Knowledge and Data Engineering, 2010, pp. 1345-1359.*

[22] Weiss, K., Khoshgoftaar, T.M. and Wang, D. (2016) "A survey of transfer learning", *Journal of Big Data, 3(1), p. 9.*

[24] Arnold, A., Nallapati, R. and Cohen, W.W. (2007) "A Comparative Study of Methods for Transductive Transfer Learning", *in Seventh IEEE International Conference on Data Mining Workshops(ICDMW 2007). Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), pp. 77-82.*

[25] Chinchor, N. and Robinson, P. (1998) "Appendix E: MUC-7 Named Entity Task Definition (ver 3.5)", *in Seventh Message Understanding Conference (MUC-7) : Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. MUC 1998.*

[26] Mendes, P., Jakob, M. and Bizer, C. (2012) "DBpedia: A Multilingual Cross-domain Knowledge Base", *in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). LREC 2012, Istanbul, Turkey: European Language Resources Association (ELRA), pp. 1813-817.*

[27] Ramshaw, L. and Marcus, M.(1995) "Text Chunking using Transformation Based Learning", *in Third Workshop on Very Large Corpora.*

[28] Klie, J.-C. et al. (2018) "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation", *in Proceedings of the 27th International*

Conference on Computational Linguistics: System Demonstrations. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 5-9. https://aclanthology.org/C18-2002.

[29] Sterbak, T. (2018) "Named entity recognition with Bert", *Depends on the definition. https://www.depends-on-the-definition.com/named-entity-recognignition-with-bert/*

[30] Buckland, M. and Gey, F. (1994) "The relationship between Recall and Precision", Journal of the American Society for Information Science, 45(1), pp. 12-19. https://doi.org/10.1002/(SICI)*1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L*

[31] Sasaki, Y. (2007) "The truth of the F-measure", p. 5

[32] H. S. Heo1, C. M. Yoon, Y. H. Ryu1, S. H. Yong, D. Y. Kim. *"MIL-BERT: Military Domain Specialized Korean Pre-trained Language Model" Korea Society for Naval Science & Technology, Vol. 6, No. 2, 2023, pp. 201-206.*

저 자 소 개

김 학 진(정회원)
1986년 건국대학교 전산계산학과 학사 졸업
1986~2019년 신용보증기금 부장
1997년 연세대학교 전산계산학과 석사 졸업
2004년 광운대학교 컴퓨터공학과 박사 졸업.
2001년~2020년 명지전문대학 강사, 겸임교수
2022~현재 서일대학교 컴퓨터전자공학과 조교수
<주관심분야 : 신호처리, 음성인식, 자연어 처리>