관세데이터를 활용한 개체명 인식

유경훈 [1]
[1] 고려대학교 정보통신대학원 석사과정

Hun0520@gmail.com

Named Entity Recognition Using Customs Data

KyoungHun yu[1]

[1] Department of Bigdata Convergence Korea University

요 약

본 연구는 관세 데이터를 BERT 기반 모델을 활용한 개체명 인식(NER)모델을 제안한다. 관세 분야 국내 첫 시도이며, 선행연구들과 달리 개체명 인식에 초점을 맞춘다. 관세 관련 텍스트에서 고유한 의미의 개체를 인식하는 것이 주요 목표이다. 이 연구는 관세 분야의 개체명 인식에 대한 이해도를 높이고 향후 HS 코드 검색 시스템 개발에 대한 기초 연구를 제공한다.

1. 서론

HS 코드는 국제 통일상품 분류 체계를 따르는 품목분류 코드로, 세계관세기구(WCO)에서 정한 원칙에 따라 관세, 무역통계, 운송, 보험 등 다양한 분야에서 사용된다. 국내 수출입 관련 업무에서 HS 코드는 중요한 역할을 하지만, 일반인이나 전문가에게도 분류가 어려운 경우가 많다. 이에 따라 본 연구에서는 관세분야 개체명 인식(NER)모델을 제시하고, HS 코드 검색 시스템을 만들기 위한 선행 연구로 진행된다.

본 연구에서는 최근 자연어 처리 분야에서 SOTA를 달성한 BERT 모델을 활용해 관세 분야 개체명 인식을 수행한다. 이는 관세분야에서 국내 첫 시도로 의의가 있다.

개체명 인식은 텍스트에서 고유한 의미의 개체(Entity)를 인식하는 것으로, 자연어 처리 전반에 걸쳐 중요한 역할을 한다. 제안하는 개체명 인식 모델을 이용하면 향후 HS 코드 검색 시스템 모델 연구에 성능 향상 방안으로 활용할 수 있다.

본 연구는 다음과 같이 구성된다. 제 2 장에서는 개체명 인식, BERT, HS 코드에 대한 이론, 제 3 장 에서는 선행연구를 소개한다. 제 4 장에서는 본 연구에서 제안하는 개체명 인식 모델을 소개하고, 실험결과, 결론 및 향후 연구 방향에 대해 논의한다.

관세 분야의 개체명 인식에 대한 이해도를 높이고, 향후 연구 및 시스템 개발에 기여할 것으로 기대된다.

2. 이론적 배경

2.1 개체명 인식

개체명 인식은 컴퓨터를 활용해 문서내에 있는 개체를 찾아 미리 정의된 라벨로 분류하는 다중 분류 모델이다. 이는 자연어 처리(NLP)에서 중요한 역할을 하며, 주요 정보를 추출해 다양한 방면에서 활용이 가능하다. 전통적인 개체명 인식 방법에는 규칙기반, 지도학습 기반, 지식기반 등이 있으나 최근에는 딥러닝 기반의 BiLSTM-CNN-CRF 모델과 사전 학습된 Transformer 계열의 BERT 모델을 활용한 연구가 주를 이룬다.

2.2 BERT

BERT 는 구글이 발표한 Transformer 기반 언어 학습 모델로 대용량 코퍼스와 많은 파라미터를 활용해 사전 학습을 수행한다. QA, NER, Classification 등 다양한 프로젝트에서 활용할 수 있다. BERT 는 사전 학습 시 두 가지 목적으로 학습하며, 이는 문장내 마스킹 된 단어를 예측하는 MLM 과 두 쌍의 문장이 연결되는 문장인지 예측하는 NSP 방식으로 학습한다.

1) MLM
   전체 토큰 중 15%를 임의로 마스킹 하여 모델이 학습을 통해 마스킹 한 토큰을 예측하도록 학습한다. 마스크 토큰은 사전 학습 단계 에서만 사용하기 때문에 Fine Tunning 단계에서 불일치가 발생할 수 있다. 이를 해결하기 위해 80%

만 마스킹, 10%는 랜덤 토큰 치환, 나머지 10%는 마스킹 하지 않는 방식으로 해결한다. [1]

2) NSP

두개 이상의 문장 간의 관계를 알아야만 해결할 수 있는 문제에서 좋은 성능을 내기 위해 BERT 는 50%의 문장은 A 와 B 가 실제로 이어진(Next Sentence)데이터를 선택하고, 나머지는 랜덤 한 데이터를 선택해 A, B 문장이 실제로 연결되는 문장인지 예측하는 방식으로 QA, NLI 같은 과제에서 높은 성능을 나타낸다.

2.3 HS 코드

HS 코드는 무역으로 거래되는 모든 물품에 번호를 부여해 구분할 수 있도록 고안한 것이며, 세계적으로 통일된 6 자리(소호)를 사용하고 있으며 한국은 10 자리를 사용한다. 관세율, 세관장 대상 여부 판단, 관세 감면, 보험요율 등 다양한 무역관련 업무를 HS 코드 기반으로 진행된다. [2]

HS 코드로 분류할 수 있는 품목은 약 100 만 종류에 해당하며, 이러한 품목을 올바르게 분류하는 것이 관세사의 전문적인 업무 중 하나이다. 특히, 기술 발전으로 인해 관세율표에 존재하지 않는 새로운 기능의 제품이나 다기능 복합물품이 생산되어 여러 개의 HS 코드가 경합되는 일관된 분류의 어려움이 존재한다. 이러한 문제를 해결하기 위해 품목분류 사례를 참고해 HS 코드를 정하기도 한다.

2.4 GPT

GPT 는 'Generative Pre-Trained Transformer'의 약어로 OpenAI 에서 개발한 자연어 처리 모델이다. GPT 모델은 대규모 코퍼스를 기반으로 사전 학습된 언어 모델이며 다음 단어 예측, 문장생성, 기계 번역, 질의 응답 등 다양한 자연어 처리에 활용 할 수 있다. GPT 모델중 'text-davinci-002'은 2020 년에 발표된 모델이다. 175 억개의 파라미터를 가지며 대화형 AI, 문장 생성, 질의 응답에서 높은 성능을 나타낸다.

3. 관세 분야 선행 연구

관세 데이터를 활용한 선행 연구는 주로 HS 코드 자동 추천을 위한 연구가 많이 진행되었다. 우효창(2019)은 단어 임베딩 기법인 Word2Vec 를 활용해 코사인 유사도 기반 HS 코드를 추천하는 연구를 수행했다. 이동주(2020)은 CNN 기반의 모델 중 VGGNET, ResNET50, Inception-V3 모델을 활용해 5 개 HS 코드에 대한 추천 모델을 연구했다. 마지막으로 이종권(2021)은 LSTM 모델을 활용한 자연어 처리 관점에서 HS 코드 추천 모델을 개발했다.

선행 연구를 통해 관세 분야에서 HS 코드 자동 추천을 위한 다양한 접근 방식이 존재함을 확인하였으나, 관세 데이터를 활용한 개체명 인식에 대한 연구는 아직 부족한 상태이다. 이에 따라 본 연구의 필요성이 대두되며 이를 통해 품목, 규격, 원산지 등과 같은 중요 정보를 자동으로 추출하여 HS 코드 분류의 정확성과 효율설을 개선할 수 있는 기회를 제공할 것이다.

4. 실험 구성

4.1 데이터

본 연구에서는 관세 분야의 데이터 수집을 위해 관세정보법령포털의 품목분류 국내사례(2013 년 1 월 1 일부터 2022 년 12 월 31 일까지)에 대한 POST 방식의 동적 크롤링을 수행하였다. 품목분류 국내사례는 위원회 결정사항, 협의회 결정사항, 품목분류사례 등으로 구성되어 있으며, 총 65,483 건의 데이터를 Python 을 사용하여 수집하였다. 데이터 수집 대상이 방대한 양이므로, 멀티프로세싱 기법을 도입하여 작업 노드를 10 개로 설정함으로써 데이터 수집의 효율성을 높였다. 수집된 데이터의 메타데이터에는 참조번호, 시행일자, 결정 세번(HS 코드), 물품 설명, 결정 사유, 이미지 등이 포함되어 있으나, 본 연구에서는 결정 세번과 물품 설명에 포함된 텍스트 데이터를 주로 활용하였다.

4.2 데이터 전처리

수집된 데이터는 특수문자 제거, 공백 조정, 문장 앞뒤의 공백 제거 등 일반적인 자연어 처리 데이터 전처리 과정을 거쳤다. 또한, 데이터의 균형을 고려하여 중복 문장을 제거하였다. 전처리가 완료된 최종 데이터셋의 HS 코드 분포는 그림 1 과 같이 나타난다.

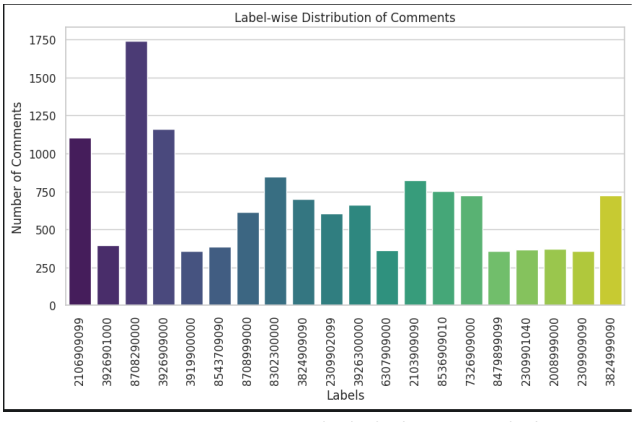

그림 1. HS 코드별데이터 분포 현황

HS 코드별 데이터 분포는 다양한 범위를 나타내며, 이를 통해 전처리 과정을 거친 데이터셋의 구성이 어

떠한 형태로 되어 있는지 파악할 수 있다. 이러한 전처리된 데이터를 활용하여 모델을 학습시키면, 높은 정확도의 결과를 얻을 수 있을 것으로 예상된다.

4.3 데이터 라벨링

BERT 기반 모델에서 도메인 특화 사전학습을 수행하기 위해서는 학습 데이터에 라벨링이 되어 있어야 한다. 그러나 현재 공개된 한국어 개체명 데이터 중 관세와 관련된 데이터는 존재하지 않는다. 따라서 수집한 데이터에 대한 라벨링 작업이 필수적으로 요구된다.

전이학습이나 액티브 러닝과 같이 라벨이 부여된 데이터를 활용하여 새로운 데이터에 대한 라벨링을 수행하는 자동화 라벨링 방법도 있지만, 본 연구에는 적합하지 않았다. 따라서 본 연구에서는 최근 사회에서 주목받고 있는 Chat GPT 를 활용한 데이터 라벨링을 수행하였다.

HS 코드분류시 활용할 수 있는 6 개의 개체를 정의하고 Chat GPT API 와 프롬프트 엔지니어링을 통해 데이터 라벨링을 수행했다.

개체명	설명
품명(IN)	이름, 명칭
규격(ST)	구체적인 크기, 형상, 성능 등의 기준
재질(MT)	물질이나 소재
기능(FC)	수행하는 구체적인 작업, 동작을 의미
성별(AG)	남성용 혹은 여성용
용도(PP)	사용하는 목적 또는 이유

표 1 라벨링 기준

4.4 실험 결과

각 모델에 대한 실험 결과는 정밀도(precision), 재현율(recall)의 조화평균인 F1-Score 를 활용해 평가하며, 라벨의 크기와 분포에 따라서 종합적인 모델 평가를 위해 Micro 와 Macro 를 활용한다.

1) Micro
분류 기준의 TP(True Positive), FN(Fales Negative), FP(False Positive)를 각각 합산하여 하나의 Confusion Matrix 를 통해 F1-Score 를 계산하며 분류 기준 별 크기의 차이가 큰 경우에 적합하다.

2) Macro
분류 기준 별 F1-Score 를 모두 합산한후, 기준 개수로 나누어 평균을 계산한다. 모든 기준에 대한 평균적인 성능을 계산할 수 있기 때문에 분포가 균일할 때 적합하다.

<표 1> 사전학습 모델 별 성능

모델	F1-Score (Micro)	F1-Score (Macro)
BERT Base	66%	45%
BERT-Multi	68%	48%

Chat GPT 활용해 생성된 데이터(6,000 건)를 9:1 로 Train/Test 데이터 셋으로 나눈 후, 사전 학습된 BERT 모델에 학습하여 검증데이터를 통해 검증한 결과 BERT Base 는 F1-Score 기준 66%, Bert Multilingual 는 68% 로 개체명 분석 성능을 나타냈다.

5. 결론 및 제언

본 연구는 관세 분야에서 국내 첫 시도로 의의가 있으며, 선행연구들과 달리 개체명 인식에 초점을 맞추었다. 하지만 모델의 성능을 높이기 위해서는 학습을 위한 학습용 데이터셋이 충분히 필요하지만, OpenAPI 를 사용하며 발생하는 비용에 의해 데이터셋을 충분히 확보하지 못했다.

다양한 데이터 셋을 사용해 모델의 학습 데이터를 늘림으로써 모델의 성능을 높일 수 있다. 더 많은 데이터를 활용하면 개체명 인식 모델이 더욱 정확하게 개체를 인식할 수 있을 것으로 기대 된다.

추가적인 성능 개선 방향으로 사전학습 모델로 사용한 BERT 계열의 모델은 한국어에 특화된 모델이 아니다. 따라서 한국어에 특화한 KoBERT, KcBERT 를 활용하거나, DAPT 방법론을 활용해 관세 도메인에 특화한 사전언어 학습 모델을 활용하면 개체명 인식율을 높일 수 있을것이다.

본 연구를 통해 개체명 인식 기반의 관세 데이터 처리 및 HS 코드 검색 시스템의 발전이 기대되며, 이를 바탕으로 더욱 효율적인 관세 관리 및 무역 환경 개선에 기여할 것으로 예상된다.

참고문헌

[1] Ashish Vaswan “Attention is all you need”
[2] 남대정 “무역 물류 실무”
[3] Suchin Gururangan "Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”
[4] Jing Li, Aixin Sun “A Survey on Deep Learning for Named Entity Recognition”
[5] Qi Zhang “Adaptive Co-Attention Network for Named Entity Recognition in Tweets”
[6] 이동주 “HS 코드 분류를 위한 CNN 기반의 추천 모델 개발”
[7] Suchin Gururangan “Don’t Stop Pretraining: Adapt Language Models and Tasks”
[8] 최신아, “빅데이터 기반 HS CODE 자동 제안 시스템 설계”