KorBERT 와 능동 학습 기법을 활용한 한국어 기반 유형 분류 모델에 관한 경험적 연구

박원주, 김연수*, 백명선, 이용태
한국전자통신연구원 국방·안전 ICT 연구단, *이화여자대학교 컴퓨터공학과
[wjpark, sabman, ytlee]@etri.re.kr, *wx12348@i.ewha.ac.kr

An Empirical Study on the Korean Text Classification based on Active Learning with KorBERT

Wonjoo PARK, Yeon-Soo Kim*, Myung-Sun Baek, Yong-Tae Lee
ETRI, *Ewha Womans University

요 약

능동 학습 기법은 지도 학습에서 데이터 레이블링을 위한 비용과 시간을 절감하기 위하여 널리 사용되고 있다. 최근 BERT 는 다양한 자연어 처리 태스크에서 높은 성능을 보여주고 있다. 본 논문은 ETRI KorBERT 모델과 능동 학습 기법을 활용하여 한국어 텍스트의 유형 분류 모델의 성능을 개선한 경험적 연구 결과를 제안한다. 특히 전문 지식이 필요한 치안 도메인에 다양한 능동 학습 기법을 적용하여 데이터 구축 비용 등 예산 절감 효과를 제시하고, 실제 데이터와 최신 딥러닝 모델을 결합한 실험 결과를 제공한다.

Ⅰ. 서 론

텍스트 유형 분류 태스크는 자연어 처리 분야에서 오랫동안 연구되고 있고 좋은 성능을 보여주고 있는 분야이다. 특히 구글의 BERT 모델과 코드가 공개된 이후 자국의 언어를 활용한 사전학습모델을 생성하고, 미세조정한 후, 다양한 과업에 적용하여 좋은 성능을 제시하고 있다. 한국어는 ETRI 에서 KorBERT 모델을 공개한 이후, SKT 의 KoBERT, 투블럭 AI 의 HanBERT 와 같은 한국어 사전학습모델이 공개되고 있다. 지도학습모델인 KorBERT 를 텍스트 유형 분류 태스크에 적용하기 위해 데이터 레이블링 작업이 선행되어야 한다. 그러나 데이터 레이블링은 많은 비용과 시간을 요구하고, 특히나 실제 환경에서는 도메인 전문 지식이 필요하다[1]. 이 때 능동 학습은 정보량이 풍부한 데이터, 즉 성능 향상에 효과적인 인스턴스들을 추출하여 데이터 레이블링을 수행하되, 적은 데이터로 최적의 성능을 달성하는 것을 목적으로 한다[2]. 본 논문에서는 치안도메인에서 다양한 능동 학습 기법을 적용하여 데이터 레이블링 비용과 시간을 절감한 효과를 제시함으로써, 실제 데이터와 KorBERT 모델을 결합한 실험 결과를 살펴본다.

Ⅱ. 능동 학습과 KorBERT 를 결합한 사건 유형 추론 모델 학습 및 검증

본 연구는 과거의 112 신고 내용을 학습하여 신규 신고 접수 시 사건 유형을 추론하는 기술 개발을 목적으로 하고 있다. 이 때, 사건 유형은 사회가 변화함에 따라 신규 유형이 추가되기도 하고, 경찰의 적절한 대응을 위하여 보다 세분화되기도 한다. 실제 현장에서는, 일반적으로 적절한 시스템 운영을 위하여 과거의 유형의 삭제는 최소화하고, 신규 유형을 추가하도록 설계되어야 한다. 예를 들어 '놀이터에서 청소년들이 크게 노래를 불러서 시끄럽다. 순찰 요망'이라는 신고가 접수되었을 때, 이 신고는 과거에는 '기타경범'이라는 유형으로 처리하였으나, 소음에 대한 신고량이 많아지고, 소음 발생 상황 등에 따라 적절한 대응이 필요하여 '소음'이라는 유형이 추가될 수 있다. 그러나 과거의 신고내용 중 소음 관련된 신고내용은 기타경범으로 이미 분류된 사례가 적지 않을 것이다. 이렇게 포괄적으로 운영되던 분류의 일부가 세분화되고, 이를 포함하여 전체 클래스를 지도 학습하여 유형을 분류하는 모델을 생성하기 위하여 본 연구는 포괄적 유형에 능동 학습을 적용한다. 112 의 대표적인 포괄적 유형은 '기타형사범'과 '기타경범'이다. '기타형사범'은 동물학대, 마약, 주거침입, 재물손괴 등 형사 사건 관련 유형과 신고 시 두개 이상의 범죄행위가 발생한 경우를 포함하고 '기타경범'은 소음, 교통불편, 보호조치, 노상방뇨, 무단투기 등 질서유지가 필요한 유형을 포함하고 있다. 본 시스템은 전체 112 신고사건 중 기타형사범(others1)과 기타경범(others2)으로 분류된 인스턴스만을 추출하여 능동 학습을 적용함으로써 포괄적 유형으로 분류된 인스턴스 중 정보량이 풍부한 데이터를 선별하여, 재레이블링을 수행하고, 이 인스턴스를 포함하여 전체 클래스를 분류하는 모델을 학습하여 사건 유형 분류 성능을 향상시키고자 한다. 그림 1 은 능동 학습을 사용하여 질의전략에 따른 데이터 샘플링과 레이블링 작업 프로세스를 보여준다.

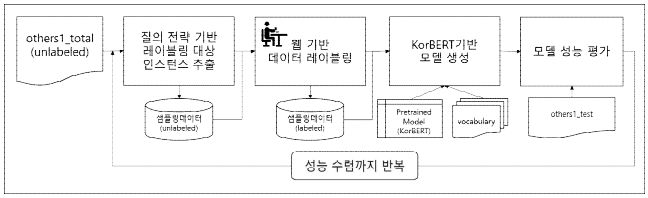


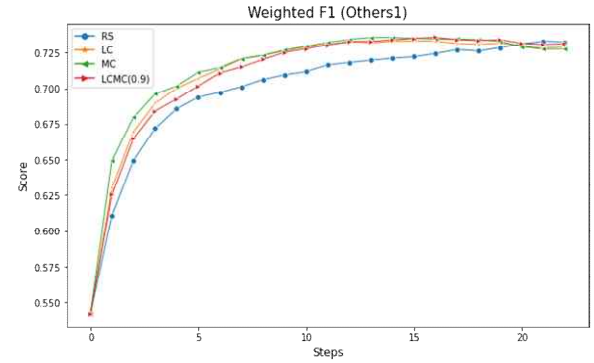
그림 1. 능동 학습 기반 데이터 구축 프로세스

능동 학습을 적용하기 위하여 대상 인스턴스를 샘플링하는 질의 전략은 다양하다. 본 연구는 베이스로 RS(Random Sampling)을 선정한 후, 능동 학습에서 많이 사용하는 LC(Least Confidence), MC (Margin Confidence)를 사용한다. LC 는 모델이 예측한

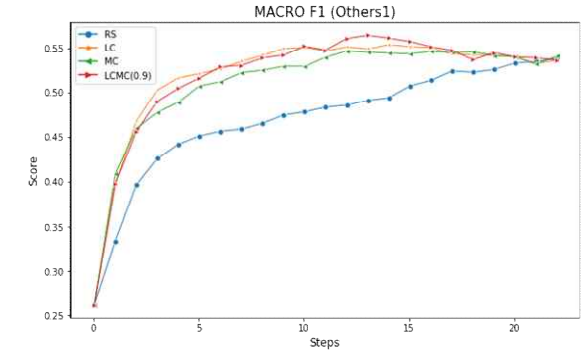
확률값이 가장 낮게 측정한 인스턴스를 우선 사용하는 방법으로, 확신도가 낮은 인스턴스를 우선 선택하는 질의 전략이다. MC 는 LC 전략이 모델에서 하나의 확률값만 사용하는 것을 보완하기 위해, 클래스 확률이 가장 큰 값과 두번째로 큰 값의 차이가 가장 작은 인스턴스를 우선 선택하는 전략이다[3]. 이것은 두 라벨의 차이가 적어 모호함을 많이 포함하는 인스턴스를 위주로 데이터 레이블링을 구축하려는 목적이다. 또한 본 논문은 LC 질의 전략이 확신도가 낮은 인스턴스를 우선 선택하여 모델을 학습시키는 경우, 일반성이 떨어지고 특성 인스턴스에 모델이 과적합될 수 있어, 확신도가 낮은 인스턴스를 샘플링하는 비율(α)과 확신도가 높은 인스턴스를 샘플링하는 비율(β)을 결합하여 질의하는 전략인 LCMC(Least Confidence & Most Confidence)를 제안한다. α 와 β 의 합은 1 이고, 사용자가 조정한다. 표 1 은 기타형사범(others1)과 기타경범(others2)에 대하여 능동 학습을 진행한 실험 환경이고, 그림 2 는 실험 결과이다.

표 1. 능동학습 실험 환경

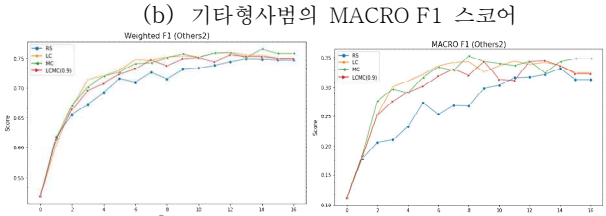
	기타형사범	기타경범
전체 인스턴스의 수	399,459	33,093
스텝당 인스턴스의 수	3,000	400
배치크기(batch size)	32	8
에포크(epoch)	3	3
학습률(learning rate)	2e-5	2e-5
테스트 인스턴스의 수	19,000	1,600



(a) 기타형사범의 Weighted F1 스코어



(b) 기타형사범의 MACRO F1 스코어



(c) 기타경범의 Weighted F1 (d) 기타경범의 MACRO F1

그림 2. 각 질의 전략에 따른 테스트 실험 결과

실험 결과는 유형별 데이터 인스턴스의 불균형을 고려하여 Weighted F1, MACRO F1 두 지표로 평가하였다. 결과를 분석하면, 레이블링을 위한 인스턴스 선택 시 무작위로 선택하는 RS 에 비해 LC, MC, LCMC(α :0.9) 등 능동 학습에 따른 인스턴스를 선택하는 기법이 성능 수렴에 빠르게 도달한다는 것을 알 수 있다. 기타형사범의 경우, RS 전략으로 60,00 인스스턴스를 학습해야 Weighted F1 스코어가 73%에 도달하는 데에 비해 LC, MC, LCMC 질의 전략으로 33,000~36,000 인스턴스를 사용하여 유사한 성능을 달성할 수 있다. 특히 MACRO F1 스코어의 경우 LCMC 질의 전략을 적용한 경우 가장 높은 결과를 보임에 따라, LC 질의 전략이 가질 수 있는 과적합을 보완할 수 있음을 확인하였다. 112 사건 유형 중 포괄적 유형으로 분류되었던 기타형사범과 기타경범을 능동 학습에 따라 재분류함으로써, 전체 신고데이터에 대한 사건 유형 분류 성능은 표 2 와 같이 MACRO F1 스코어가 약 11.6% 향상되었다.

표 2. 112 사건 유형 분류 성능

	AL 적용 전	AL 적용 후
Weighted F1	80.5%	85.4%
MACRO F1	67.0%	75.8%

Ⅲ. 결 론

본 논문은 과거의 신고내용 텍스트를 학습하여 신규 사건에 대한 유형을 추론하는 모델을 학습할 때, 기존 신고 사건에 대한 유형 레이블링의 품질을 보완하기 위한 능동 학습 기법을 개발하고 검증하였다. 특히 기존의 능동 학습 기법의 단점을 보완하기 위하여, 확신도가 낮은 인스턴스뿐만 아니라 확신도가 높은 인스턴스를 샘플링하여 모델을 학습시킴으로써, 과적합에 건강한 질의 전략을 새롭게 제안하고, 실험 결과로 검증하였다. 또한 능동 학습 기법으로 선택된 인스턴스를 사건 유형 추론 모델 학습에 활용하여 사건 유형 추론 성능을 향상시켰다. 향후 이 모델이 치안현장에서 112 신고 접수 시 경찰관이 사건 유형을 결정할 때, 의사지원 솔루션으로 활용될 것을 기대한다.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2018-0-00440, 위험 상황 초기 인지를 위한 ICT 기반의 범죄 위험도 예측 및 대응 기술 개발).

참 고 문 헌

[1] 정동호, 허민강, 김형철, 박상원, “능동 학습 기법을 활용한 한국어 금융 도메인 개체명 인식 데이터 구축”, 한글 및 한국어 정보처리학술대회 논문집(2020), pp.82-85

[2] Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim, ”Active Learning for BERT: An Empirical Study”, EMNLP2020, .pp.7949-7962, Nov. 2020

[3] Burr Setteles, “Active Learning Literature Survey”, University of Wisconsin-Madison. 2009