RoBERTa 기반 데이터 증강을 통한 국내 학술 논문 분야 분류 연구

김성식 [1], 양진환 [1], 최혁순 [1], 문남미 [2]
[1] 호서대학교 컴퓨터공학부 학부생
[2] 호서대학교 컴퓨터공학부 교수

sungsik001004@gmail.com, yjh970706@naver.com,
hyuksoon2001@gmail.com, nammee.moon@gmail.com

Classification of Domestic Academic Papers Through RoBERTa-based Data Augmentation

Sung-Sik Kim[1], Jin-Hwan Yang[1], Hyuk-Soon Choi[1], Nammee Moon[1]
[1]Dept. of Computer Science, Hoseo University

요 약

현재 대부분의 국내 학술 데이터 베이스는 개별 학술지 논문의 주제를 파악하는 표준화된 정보를 거의 제공하지 않고 있다. 본 연구에서는 논문의 제목만을 활용하여 학술 논문의 분야를 자동으로 분류하는 방법을 제안한다. 이를 위해 한국어로 사전 훈련된 KLUE-RoBERTa 모델을 사용하며, Back Translation 과 Chat-GPT 를 활용한 데이터 증강을 통해 모델의 성능을 향상한다. 연구 결과, Back Translation 과 Chat-GPT 를 사용하여 증강한 모델이 원본 데이터를 학습한 모델보다 약 11%의 성능 향상을 보였다.

1. 서론

최근 들어 텍스트를 비롯한 다양한 유형의 데이터가 기하 급수적으로 증가하고 있으며, 빅데이터 및 기계학습 기술의 발전으로 문서의 자동분류에 관한 연구도 활발히 진행되고 있다[1]. 그럼에도 불구하고 국외 학술 데이터베이스(WoS, SCOPUS, LISTA 등)와 달리, 대부분의 국내 학술 데이터베이스는 개별 학술지 논문의 주제를 파악하는 표준화된 정보를 거의 제공하지 않고 있다[2].

학술지 논문 분야 분류는 많은 비용이 요구되는 작업이다. 따라서 이러한 문제를 해결하고자, 기존의 수작업 분류 방법 대신 기계학습을 활용한 국내 학술지 논문 분야 분류를 적극적으로 모색할 필요가 있다 [3].

본 연구에서는 논문의 제목만을 활용하여 학술 논문의 분야를 자동으로 분류하는 방법을 제안한다. 이를 통해 대량의 학술지 논문을 효율적으로 분류하고, 작업의 효율성을 향상시키며, 연구 분야별로 정확한 분류를 수행하고자 한다.

2. 관련 연구

2.1 BERT 기반 모델

최근 들어 자연어처리분야에서 트랜스포머 구조를 활용해 더욱 성능을 향상시킨 모델들이 나타나고 있다. 그 중에서도 대표적인 모델이 구글의 BERT 모델이다[4]. 이전 연구에서는 BERT 모델을 활용하여 학술 문헌을 자동으로 분류하고 분류 성능을 분석하는 연구가 수행되었다[1].

RoBERTa 모델은 기존 BERT 모델의 훈련이 충분히 수행되지 않은 단점을 보완하며, 학습 결과의 성능을 높이기 위해 더 많은 데이터를 이용한다. 또한 큰 배치크기를 이용하여 학습하고 Masking 을 동적 할당하는 방식으로 BERT 를 학습하여 기존 BERT 모델의 성능을 향상시킨 모델이다[5]. 본 연구에서는 한국어로 사전 훈련된 KLUE-RoBERTa 모델을 사용하여 분류 작업을 수행한다.

3. 연구 방법

3.1 데이터수집 및 전처리

본 연구에서는『KISTI 인공지능 데이터 공유·활용 서비스』의 논문 연구분야 분류 데이터셋과 KCI 에서 수집한 데이터를 사용한다. 전체 데이터에 대해 정규식을 사용하여 제목이 없거나 영어 혹은 한문으로만 이루어진 제목을 제외한다. 전처리한 총 29,785 건의 문장에 대한 분류기준은 『국가 과학기술표준분류체계(2018 년 개정)』상의 분류 범주의 대분류명 30 가지 범주를 사용한다. 논문에 대한 30 개의 분류명은 모두 숫자로 바꿔 기입하였으며, 각 분야의 개수와 분류번호는 <표 1>과 같이 구성한다. 전체 데이터의 70%는 학습 데이터로, 15%는 검증 데이터, 15%는 평가데이터로 나눈다.

<표 1> 분류명별 데이터

분야	논문 수	분야	논문 수
보건의료(11)	1,932	환경(29)	854
문학(6)	1,723	생활(13)	844
교육(2)	1,712	화학(28)	818
철학/종교(27)	1,604	에너지/자원(17)	694
정보/통신(23)	1,570	원자력(20)	678
문화/예술/체육(7)	1,535	정치/행정(24)	665
여성학(18)	1,528	미디어(9)	663
농림수산식품(4)	1,435	지구과학(25)	620
경제/경영(1)	1,390	지리/지역/관광(26)	545
역사/고고학(19)	1,179	물리학(8)	525
기계(3)	1,170	뇌 과학(5)	477
생명과학(12)	1,140	수학(14)	477
건설/교통(0)	1,034	법(10)	374
전기/전자(22)	975	언어(16)	367
재료(21)	889	심리(15)	304
합계		29,785	

3.2 데이터 증강

데이터 증강은 모델의 성능 향상을 위한 핵심적인 전략 중 하나로, 다양한 기술과 방법을 활용하여 데이터의 다양성과 양을 높이는 과정이다. 이를 위해 본 연구에서는 두 가지 방법을 사용해서 데이터를 증강한다.

가. Back Translation

한국어로 이루어진 논문 제목을 파파고 번역 서비스를 이용해 영어로 번역하고 번역된 영어 논문 제목을 다시 원본 언어인 한국어로 역 번역한다. 이 과정에서 언어 간 변환 과정을 거치게 되어 데이터의 다양성을 증가시키고 언어 특성을 보존한다.

학습데이터에 대해 증강을 진행하는데 증강 비율은 해당 분야 원본 개수의 50%로 설정한다. 증강된 데이터 중에서 원본 데이터와 동일한 데이터 및 결측 데이터를 제외한 총 11,501 개의 데이터를 증강한다.

나. Chat-GPT

Chat-GPT 는 OpenAI 에서 GPT 모델을 기반으로 제작된 챗봇 서비스이다. 분야에 해당하는 질문을 Chat-GPT 에 prompt 로 제공하여 새로운 데이터를 생성하는 방법으로 증강한다.

데이터 개수가 하위 15 개인 분야의 학습데이터에 증강을 진행한다. 이때 증강 개수는 분야당 100 개씩 증강을 진행하고 원본 데이터와 동일한 데이터를 제외한 총 1,475 개의 데이터를 증강한다.

4. 실험

4.1 실험 환경

본 연구에서는 Hugging Face 에 공개되어 있는 한국어로 사전 훈련된 KLUE-RoBERTa-large 를 활용하여 논문 제목의 분야 분류 실험을 진행한다. 실험에 사용된 하이퍼파라미터는 학습 횟수 10, 배치크기 16, 학습률 1e-5 로 설정하며, 최적화 알고리즘은 AdaBelief 를 사용한다.

4.2 실험 결과

학습 결과 평가 데이터의 정확도와 F1-Score 는 <표 2>와 같다.

<표 2> 학습의 정확도, F1-Score

Model	증강방법	정확도	F1-Score
RoBERTa	X	63.14	61.83
	GPT	64.35	62.14
	BT	73.81	72.19
	BT+GPT	74.13	72.78

<표 2>를 보면 모든 데이터 증강을 적용한 RoBERTa 모델의 성능이 정확도 74.13%, F1-Score 72.78%로 원본 데이터 학습 모델보다 약 11% 높은 정확도를 기록하였다.

5. 결론

본 연구에서는 RoBERTa 모델을 활용하여 논문 제목을 통해 30 개의 분야로 분류하는 작업을 수행하였다. 연구 결과 데이터 증강을 통해 모델의 성능을 향상시킬 수 있음을 확인하였으며, Back Translation 과 Chat-GPT 를 사용하여 증강한 모델이 원본 데이터를 학습한 모델보다 약 11%의 성능 향상을 보였다. 이러한 결과는 추가적인 데이터 확보와 클래스 균형을 맞추게 된다면 성능을 더욱 향상시킬 것으로 기대된다.

본 연구는 논문 제목만을 활용하여 분류를 시도함으로써 보다 적은 노력과 예산을 투입하여 실질적인 자동 분류를 수행할 수 있는 효율적인 방안을 제시하였다. 더불어, 향후 연구에서는 요약 및 키워드와 같은 부가 정보를 활용하여 더욱 정확하고 효과적인 분류 시스템을 연구할 계획이다.

사사문구

이 논문은 2023 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF- 2021R1A2C2011966).

참고문헌

[1] 김인후, 김성희, 딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류, 정보관리학회지, 39 권, 3 호, 293-310, 2022

[2] 김판준, 자질 선정을 통한 국내 학술지 논문의 자동분류에 관한 연구, 정보관리학회지, 39 권, 1 호, 69-90, 2022

[3] 김판준, 기계학습에 기초한 자동분류의 성능 요소에 관한 연구, 정보관리학회지, 35 권, 2 호 37-62, 2018

[4] 이치훈, 이연지, 이동희, 사전 학습된 한국어 BERT 의 전이학습을 통한 한국어 기계독해 성능 개선에 관한 연구, 한국 IT 서비스학회지, 19 권, 5 호, 83-91, 2020

[5] 허주은, 박도제, 이선아, 깃허브 이슈 보고서 관리를 위해 RoBERTa Fine-Tuning 을 이용한 다중 레이블 분류 기법과 성능 비교, 한국정보과학회 2022 한국 컴퓨터종합학술 대회, 한국 정보과학회, 2022, 258-260