KB-BERT: 금융 특화 한국어 사전학습 언어모델과 그 응용

Journal of Intelligence and Information Systems VOL 28. No. 2.
June 2022 191-206

2022254019 한병엽

01 연구 개발 개요

1. 연구개발 개요

1-1. 연구개발 배경

금융 특화 언어모델의 학습과 이를 응용한 금융 특화 자연어 처리 모델에서의 성능 향상을 목표

- 범용 사전학습 언어모델 -> 도메인 특화 능력 부재
  - OOV(Out of Vocabulary) 발생으로 언어 이해 부족
  - 특정 도메인에서만 사용되는 단어 및 유의어 관계에 대한 이해 부족
- 금융 전문 지식을 고려한 자연어 처리 모델의 학습이 필요
  - 금융 용어에 대한 이해
  - 상품명 등 고유명사의 이해
  - 문서내 여러 수치 및 값에 대한 이해(Numerical Reasoning)

1. 연구개발 개요

1-2. 선행연구

- BERT
  - Transformer 뉴럴 네트워크 구조를 활용한 첫 번째 언어모델
  - 텍스트의 일부분을 임의로 마스킹(Masking)하고 이를 다시 원본 텍스트로 복원하는 디코딩(Decoding)을 수행하는 Masked language modeling 방법
  - Labeled 데이터가 없어도 비지도 방식으로 언어 지식을 학습

- 도메인 특화 학습
  - 의료, 법률, 과학 등 목표 도메인 말뭉치를 수집 및 활용하여 From-scratch 방식으로 언어 모델 학습
  - 기존 학습된 범용 목적 언어모델을 기반으로 도메인 적응 기법 활용(DAPT, Adaptive Tokenizer)

02 연구 개발 프로세스

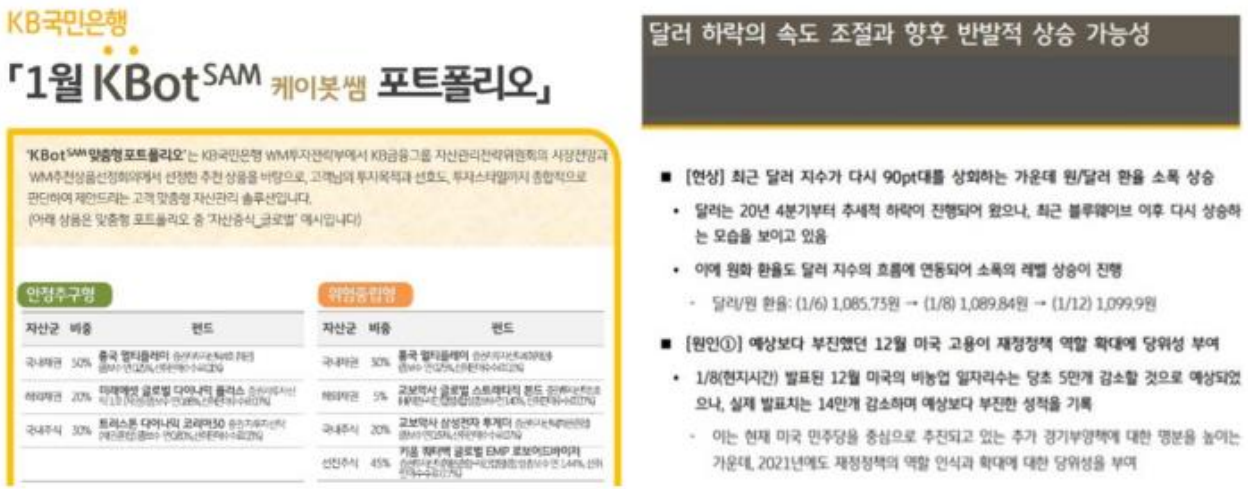
2. 연구 개발 프로세스

2-1. KB-BERT

- BERT와 동일한 Transformer 뉴럴 네트워크 구성
- Masked language modeling 방식을 통해 진행

〈표 1〉 모델 하이퍼파라미터

이름	Vocab	Word embedding	Layer	Hidden size	Self-attention heads
크기	35,000	786	12	786	12



2. 연구 개발 프로세스

2-2. 말뭉치 구성

- 위키, 뉴스, 웹문서 + 금융관련 문서
  - 금융상품 설명서
  - 투자 리포트

〈표 2〉 학습 말뭉치 크기

모델명	총 말뭉치 크기 (GB)	금융 말뭉치 크기(GB)
KoELECTRA-v3	34	-
KLUE-RoBERTa	62	-
KB-BERT	**90**	**40**

2. 연구 개발 프로세스

2-3. 말뭉치 전처리

- 스팸 텍스트 분류 모델
  - 뉴스 및 웹 문서
  - 광고성 텍스트 빈번 -> 사실(Factual) 기반 예측 능력 저하, 윤리적 이슈

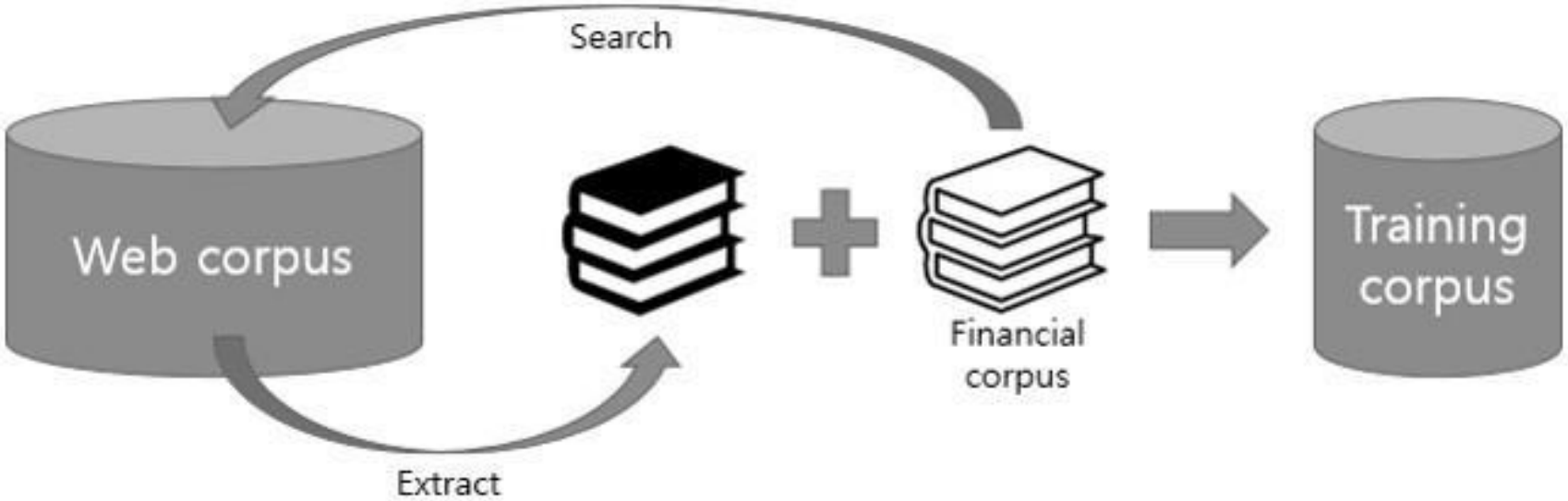
- 해시(MinHashLSH) 기반 문서 중복 제거
  - 중복 문서가 존재하기 쉬움 -> 성능 저하 및 불필요한 학습시간 증가

- 언어 판별 모델
  - 한국어 이외의 다양한 언어 존재 -> 외국어 문서 필터링 진행

2. 연구 개발 프로세스

2-4. 말뭉치 증강

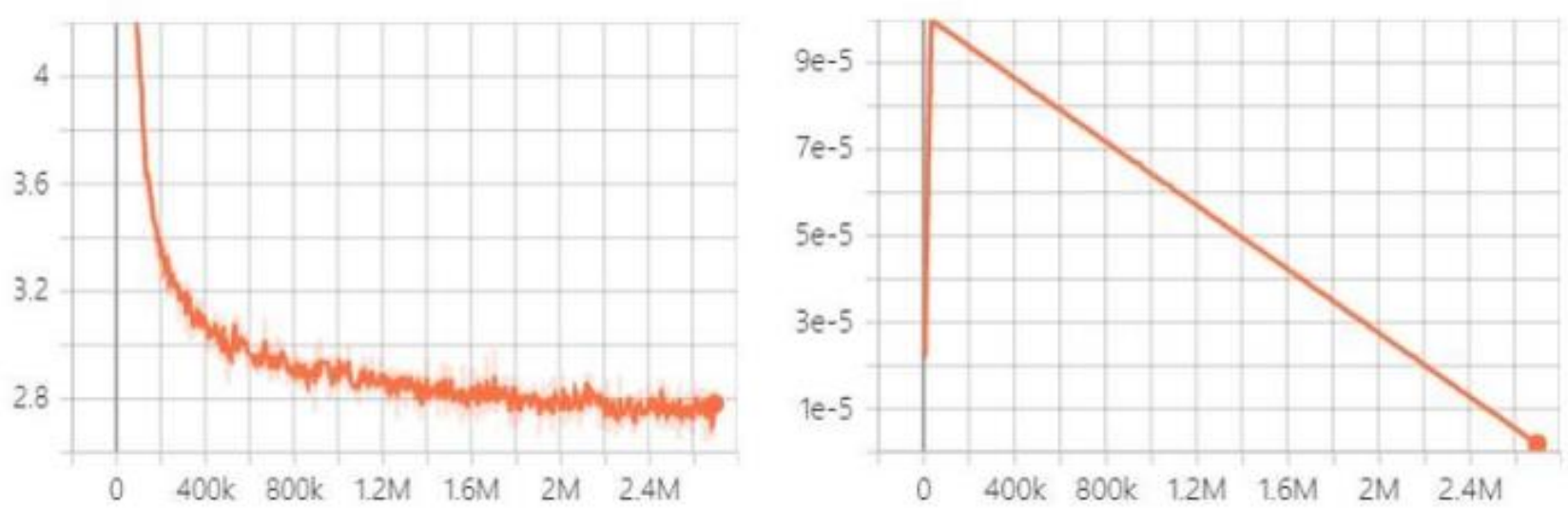
- 추가적인 금융 문서 말뭉치 획득
  - 금융 도메인 학습 말뭉치 수량이 충분치 않음



2. 연구 개발 프로세스

2-5. 학습 환경

- GPU 기반 20일간 학습
  - NVIDIA V100 32GB GPU * 8개



〈그림 3〉 학습 loss(좌) / Learning rate(우)

03 연구 결과

3. 연구 결과

3-1. 금융 특화 평가 데이터셋 – 토픽 분류

- 주어진 텍스트를 사전 정의된 토픽 클래스 중 하나로 분류

〈표 4〉 금융 특화 토픽 분류 데이터 예시

토픽	뉴스 텍스트
경제정책	은행 개인사업자 대출에 대한 예대율 규제 완화가 연말까지 연장된다. 금융위원회는 15 일 은행 개인사업자 대출 신규취급분에 적용하는 예대율 가중치를 기존 100%에서 85%로 인하하는 조치를 12 월 말까지 연장하는 내용의 은행업 감독규정 개정안을 규정변경예고했다.
수출/입	세계 수출 시장에서 점유율 1 위를 차지한 우리나라 제품이 70 개에 가까운 것으로 나타났습니다. 무역협회 국제무역통상연구원이 오늘(7 일) 내놓은 '세계 수출 시장 1 위 품목으로 본 우리 수출 경쟁력 현황' 보고서에 따르면 우리나라 세계 1 위 품목 수는 지난 2019 년 기준 69 개로 전년보다 7 개 …
투자	암호화폐 시장이 달아오르는 가운데 바이낸스코인(BNB)이 큰 주목을 받고 있다. 올해 들어 1600% 가까이 상승하면서 BNB 는 사람들을 열광시키고 있다. BNB 는 24 시간 기준으로 약 25% 상승한 후 12 일(이하 현지시간) 시가총액 950 억 달러를 돌파했다.
금융상품	삼성카드는 개인사업자에게 다양한 혜택을 제공하는 '삼성카드 BIZ LEADERS'를 출시했다고 15 일 밝혔다. 삼성카드 BIZ LEADERS 는 개인사업자들이 많이 사용하는 업종을 분석해 특화된 혜택을 제공하는 상품이다. 보험, 전기요금, 통신 업종에서 자동결제를 이용하면 결제금액의 10% 할인 …

3. 연구 결과

3-2. 금융 특화 평가 데이터셋 – 감성 분석

- 자연어 텍스트에 내포된 사람의 감성 상태를 분석 및 예측

<표 5> 금융 특화 감성분류 데이터 예시

감정	텍스트
낙관	과거수익률 종목명 5 년 기준 연평균 수익률 당월 5 년 기준 연평균 수익률 전월 DHS, PEY, SPHD 모두 장기투자 했을때는 연 환산 수익률이 7 이상을 기록하고 있어서 매달 배당을 받는다는 점을 가만했을 때 캐시카우용 종목이라고 생각합니다.
	외인이 던지는 건 미국 헷지펀드 등에서 고객 환매 요청을 대비해서 어쩔 수 없이 매도 하는거죠. 시장이 안정화 될 무렵 외인은 무조건 다시 삼전을 살겁니다. 그때 저렴하게 매수하기 위해 훈련 안된 개미들 공포에 손절매하게 할거고 가격 내려서 줍줍. 쓸 돈으로 투자한 개미들은 쫄아서 팔거고 …
비난/반대	시장경제에 그냥 맡기면 될걸 억지로 규제하니 풍선효과로 이난리지. 불과 4 년전 미분양 나서 난리났던 수도권이 서울 묶이자 지금은 투기과열지구까지 됐자나. 다 규제 풀어버리면 더 내려간다니까
	리딩증권사로서 주식시장 전체에 대하여 그리고 주식투자자전체에 대하여 심각한 심리적 물질적 영향을 끼쳤다 따라서 계속 영업하려면 주식투자자 전체에 보상하던가 아니면 자진상폐해라

3. 연구 결과

3-3. 금융 특화 평가 데이터셋 – 질의 응답

- 주어진 문서와 사용자 질문을 바탕으로 문서내 정답을 찾아 제공
  - 검색엔진을 위한 정답 제공
  - 챗봇 등 대화모델

〈표 6〉 금융 특화 질의 응답 데이터 예시

본문
KB 국민은행이 새롭게 단장한 스타뱅크 출시를 기념해 모바일뱅킹 전용 서비스를 시행한다고 28 일 밝혔다. 이번 환전 서비스는 미국 달러, 유로 등을 포함해 총 17 개 통화로 하루에 최대 3000 달러(미화 기준)까지 바꿀 수 있다. 특히 미국달러, 유로, 일본 엔화의 경우 3000 달러(미화 기준)까지 조건 없이 90% 우대 환율을 제공한다. 이는 전 금융권 모바일 환전 서비스 중 최대 우대 한도다. 스타뱅크 앱에서 환전을 신청하고 20 영업일 내에 기업은행 지점을 통해 외화를 찾아야 하며, 미국 달러, 유로, 일본 엔화, 중국 위안은 전국 모든 지점에서, 그 외 통화는 고객이 지정한 지점에서 수령 가능하다. 미화 1 만 달러까지는 여러 번 환전하고 한 번에 은행에서 찾을 수 있다. 외화 수령기간 내에는 스타뱅크 앱을 통해 외화예금에 입금하거나 원화로 재환전도 가능하다.
질문: 스타뱅크 3000 달러까지 환금시 우대 환율은? 답변: 90%
질문: 스타뱅크 4000 달러까지 환금시 우대 환율은? 답변: 답변불가

3. 연구 결과

3-4. 성능 평가

- 2종의 한국어 학습모델과 KB-BERT 성능비교
  - KoELECTRA-v3, KLUE-RoBERTa, KB-BERT
  - 감성분석, 토픽분류, 질의응답 3가지 항목 측정

〈표 7〉 범용 데이터셋 성능평가

모델명	NSMC (ACC)	KLUE-YNAT (F1)	KorQuAD v1 (F1)
KoELECTRA-v3	90.52	83.40	93.09
KLUE-RoBERTa	**90.75**	84.28	94.45
KB-BERT	90.72	**84.52**	**94.66**

〈표 8〉 금융 특화 데이터셋 성능평가

모델명	F-sentiment (F1)	F-news (F1)	F-QA (F1)
KoELECTRA-v3	43.96	58.30	71.72
KLUE-RoBERTa	46.19	61.71	71.08
KB-BERT	**47.86**	**64.10**	**72.94**

3. 연구 결과

3-5. 응용

- 자연어 분석 파이프라인 구축
  - 형태소 분석, 문서 분류, 감성 분석, 키워드 추출 -> 금융 문서 분석 및 활용
- 중요한 이벤트 추출 및 파악 업무
  - 경제 뉴스, SNS, 투자 및 경제 리포트 기반
- 딥러닝 문서 검색 및 질의응답 시스템 구축
  - 직원 및 고객의 정보 탐색

과거 수작업 규칙 또는 ML 기반 시스템보다 획기적인 성능 향상

04 한계점 및 발전 방향

4. 한계점 및 발전 방향

4-1. 한계점

- NLU(Natural language understanding) 기법만 사용
  - NLG(Natural language generation) 기반 Autoregressive 방식도 연구가 필요함

4-2. 향후 연구 발전 방향

- 도메인 특화 말뭉치의 추가적인 수집 및 개선을 통한 성능 향상
- 토픽분류, 감성분석, 질의응답 외의 태스크 분야 연구
- Autoregressive 기반 NLG 모델 연구