HR-평가 문장 Multi-classification 및 Unlabeled data 를 활용한 Post-training 효과 분석

최철 [1], 임희석 [1]
[1] 고려대학교 컴퓨터정보통신대학원
dusfkd@korea.ac.kr, limhseok@korea.ac.kr

HR-evaluation sentence multi-classification and Analysis post-training effect using unlabeled data

Cheol Choi[1], HeuiSeok Lim[1]
[1]Graduate School of Computer & Information Technology, Korea University

요 약

본연구는 도메인 특성이 강한 HR 평가문장을 BERT PLM 모델을통해 4 가지 class 로 구분하는 문제를 다룬다. 다양한 PLM 모델 적용과 training data 수에 따른 모델 성능 비교를 통해 특정 도메인에 언어모델을 적용하기 위해서 필요한 기준을 확인하였다. 또한 Unlabeled 된 HR 분야 corpus 를 활용하여 BERT 모델을 post-training 한 HR-BERT 가 PLM 분석모델 정확도 향상에 미치는 결과를 탐구한다. 위와 같은 연구를 통해 HR 이 가지고 있는 가장 큰 text data 에 대한 활용 기반을 마련하고, 특수한 도메인 분야에 PLM 을 적용하기 위한 가이드를 제시하고자 한다

1. 서론

최근 다양한 도메인 영역에서 ML/DL 의 활용도가 급격히 증가하고 있다. 마찬가지로 사람의 직관이 우선시 되던 HR 도메인 영역도 '사원 키워드 분석', '적합인재 추천 모델', '퇴직예측 분석' 등의 다양한 AI 모델이 만들어지고 있는 추세이다

본 연구에서는 HR 도메인의 다양한 AI 기술 활용의 한 측면에서, 매니저가 부하직원에 대해 평가한 Text 문서의 문장들을 PLM(pre-trained-Model) 기반의 모델을 적용하여 총 4 개의 class(평가긍정, 평가부정, 기대사항, 기타)에 대해 multi-label 예측을 수행한다

첫번째 단계로 HR 평가데이터 classification 문제를 다양한 BERT 모델을 통해 학습하고 그 성능을 살펴본다. 특히 모델 크기와 pre-trained 에 사용된 corpus 와의 유사성이 결과에 미치는 영향과, train data 개수에 따른 성능을 통해 도메인 적용에 적절한 모델과 라벨링된 train data set 의 크기를 제안한다

두번째로 쉽게 구할 수 있는 unlabeled data 문장을 바탕으로 기존의 BERT 모델을 post-training 한다. 해당 방법론은 회사 업무내용 및 HR 용어가 많이 등장하는 도메인 문장의 특징을 BERT 의 MLM(Masked Language Model) 방법을 통해 추가 학습한다면, 비슷한 문장 classification 성능의 긍정적인 영향을 줄 것이라는 점에 착안해 구상하였다. 이는 일반적으로 라벨링된 데이터의 개수가 부족한 상황속에서 unlabeled 데이터의 활용가능성을 탐색한다는데 큰 의의가 있다

2. 관련연구

BERT 모델은 "Attention is all you need" 논문에서 제안한 Encoder-Decoder 구조 중 Encoder 를 바탕으로 한 AE(Auto Encoding) 방식의 대표적인 모델이다.[1] pre-train 은 MLM(Masked Language Model) 방식을 통해 수행된다. bidirectional representation 을 학습하기 위해 input 토큰의 15%를 랜덤하게 마스킹하고 모델이 이를 맞추는 과정에서 학습이 이루어진다. [2]

이처럼 대량의 corpus 를 바탕으로 사전 학습된 모델은 classification 을 비롯한 다양한 downstream task 에 활용된다. 다만 pre-train 에 사용된 corpus 가 위키피디아 문서와 같이 범용적이기 때문에 도메인에 특화된 Task 적용에 한계를 가지는 것도 사실이다. [3]

이러한 한계를 극복하기 위한 방법 중 하나로 post-training 이 활용될 수 있다. post-training 이란 범용적인 문장으로 pre-train 된 PLM 모델에 도메인 특화 corpus 를 추가 학습시켜 downstream task 를 적용하는 방법을 의미한다.[4] 이를 통해 도메인 특화 언어 지식을 모델에 학습시키는 것이 가능하며, 일반적인 Corpus 만으로 학습된 PLM 모델 보다 특정 도메인 문제 해결에 대해 더 좋은 성능을 발휘하는 것이 가능하다

Class_1 (Evaluation_Positive)	장애없이 안정적인 청구 서비스 운영을 수행하는데 크게 기여하였고, 고객과의 협업으로 고객 만족도 목표 달성함
	새로운 시스템에 대한 개선 노력 및 프로젝트 팀원에 대한 지원까지 다양하고 중요한 역할을 수행합니다.
	적극적인 업무 참여와 본인의 분야가 아닌 업무에서도 적극적으로 개선 사항을 도출하였습니다.
	하이프라자 고객에 대한 응대 및 조직원 리딩을 매우 잘 하고 있어, 특히 고객만족도 측면에서 매우 우수함
	가점 소형 및 자동차 FCS까지 관리하며 소규모프로젝트 PM으로 많은 역할을 담당하였음
Class_2 (Evaluation_Negative)	팀 뿐만 아니라, 담당 전체적으로 상반기에는 TEC 산출물 등록 활동이 미흡했습니다.
	지역적인 핸디캡으로 학습활동이 아쉬움이 있습니다.
	본 프로젝트 PM의 평가 또한 평균 대비 다소 저조합니다.
	상반기 vmCube 웹서버 장애시 장애 감지 및 보고 체계에 대한 혼선이 있었습니다.
	AD 파트 활동이 없었음.
Class_3 (Expectation)	시스템 중설전에 로드 증가를 줄일방안은 없을지 찾아서 제안하면 더 좋을 것 같습니다.
	사업 영역 확대를 위해 많은 제안을 부탁드립니다.
	향후 이행단계에서 그간의 경험을 바탕으로 주도적인 역할을 기대합니다.
	CERT가 팀내 기술역량 강화에 핵심 역할을 수행해야 합니다.
	보다 많은 사람이 참여하고 능동적으로 진행할 수 있도록 적극 참여 바랍니다.
Class_4 (ETC)	어려운 프로젝트를 맡아 상반기 내내 많은 고생을 했던 점 감사합니다.
	한해동안 팀내 최고참으로서 과제수행하시느라 고생하셨습니다.
	많은 배움과 성장의 계기가 되셨을 것으로 생각합니다.
	특이한 사항이 없으며, 정도경영을 잘 준수하였습니다.
	프로젝트 PM 평가 반영합니다.
	76.9(%) - 목표대비 95.4% 달성 (목표 80.6%) - 제출 아이디어(3/19) : 미제출 - 기타 직간접 의견제시 및 활동 : 없음
	상반기 목표 수준 50점 기준 1. 지식등록 : 0건 2. 자산등록 : 1건 3. 분석 LL 공유 : 0건 4. 조직내외 발표 및 강의활동 : 1건

표 1. HR 평가 Text 문장 예시

3. 연구방법

1) 데이터 수집

데이터 원본은 국내 L 그룹 계열사 두 곳의 3 년치 성과평가 / 역량평가 / 동료평가를 바탕으로 구성되었다. 해당 데이터는 문단으로 구성되어 있어, NLTK 모듈을 통해 문장단위로 split 하여 labeling 되지 않은 문장 총 30 만 건을 생성하였다.

2) 데이터 전처리

전체 텍스트데이터에 대해 정규식을 사용하여 한글, 영어 문자만 남기고 숫자 및 특수문자는 모두 제거하였다. 또한 문장 split 이 잘못되거나 문장자체가 너무 단순한 '수고하셨습니다', '노력 바랍니다'와 같이 문장전체의 토큰 개수가 10 개를 넘지 않는 문장들을 데이터셋에서 제외하였다. tokenizing 이후 문장 별 token 개수의 분포를 확인하고, 이를 바탕으로 max token 값은 128 로 설정하고, 128 보다 작은 부분에 대해서 padding 처리를 수행하였다.

3) 데이터 라벨링

전처리한 총 1 만건의 문장에 대해서 현업 HR 담당자 2 명이 cross-check 를 하면서 4 가지 class(multi-label)로 labeling 을 수행하였다. 클래스별 예시 문장은 표 1 과 같다. 결론적으로 총 30 만 건의 문장에서 1 만 건을 labeling 하여 Bert classification 모델에 학습 데이터로 사용하였고, labeling 하지 않은 29 만 건의 문장은 post-training 에 사용하였다.

4) PLM 모델 사용

연구에 사용된 BERT 모델은 Huggingface 에 공개되어 있는 한국어 Model 인 KoBERT-base, KcBERT-base, KcElectra, KoElectra-dialog 이다. hyperparameter 의 경우 batch size=16, optimizer=AdamW, learning rate=1e-5 로 설정하였다. 해당 PLM 모델을 기반으로 parameter 크기 및 pre-train 에 사용된 corpus 가 최종 결과에 어떠한 영향을 주는지, train data 개수에 따른 성능 차이를 비교 연구한다.

5) Post-training

labeling 하지 않은 29 만개의 문장으로 BERT 모델을 MLM 방식으로 추가 학습하여 HR 도메인 문장에 특화된 HR-BERT 를 만든다. 사용된 PLM 모델은 학습효율성을 고려하여 base 크기의 모델인 KcBERT-base, KoBERT-base 2 가지로 선정하였다. Electra 계열의 경우 pre-train 학습방법 자체가 MLM 이 아닌 generator-discriminator 방식이기 때문에 post-training 에서 제외하였다. [6]

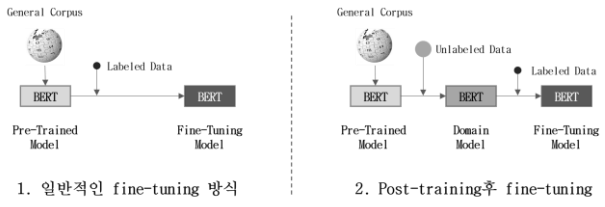

그림 1. Post-training 개념

한국어 PLM 종류		train data 개수에 따른 f1-score(micro)						
이름	Size	100개	200개	500개	1,000개	2,000개	4,000개	8,000개
KcBert-base	417M	0.813	0.853	0.847	0.878	0.887	0.886	0.900
KoBERT-base	368M	0.776	0.831	0.859	0.855	0.877	0.890	0.904
KcELECTRA-base	475M	0.850	0.870	0.884	0.881	0.895	0.893	0.896
Dialog-KoELECTRA-Small	58M	0.789	0.845	0.867	0.880	0.889	0.893	0.899

표 2. Train_data 개수에 따른 F1-score(micro)

post-training		training data개수에 따른 모델 f1-score(micro)						
Model	Epoch	100개	200개	500개	1,000개	2,000개	4,000개	8,000개
KoBert-base	0 epoch	0.776	0.831	0.859	0.855	0.877	0.890	0.904
	1 epoch	0.832	0.842	0.873	0.875	0.891	0.897	0.902
	2 epoch	0.855	0.863	0.872	0.885	0.901	0.899	0.902
	3 epoch	0.858	0.867	0.882	0.884	0.891	0.893	0.898
KcBert-base	0 epoch	0.813	0.853	0.847	0.878	0.887	0.886	0.900
	1 epoch	0.854	0.867	0.876	0.878	0.889	0.893	0.899
	2 epoch	0.811	0.854	0.879	0.881	0.888	0.896	0.897
	3 epoch	0.832	0.857	0.874	0.883	0.894	0.894	0.898

표 3. Post-training에 따른 F1-score(micro)

4. 연구결과

1) BERT Model 비교

사용된 PLM 모델별 train data 개수에 따른 f1-score 의 결과는 표.2 와 같다. 4 개 모델 모두 train data 개수가 500 개 이하인 구간에는 train data 투입에 따른 성능 증가폭이 크지만, 1,000 개 이후부터는 데이터의 개수가 2 배씩 증가함에도 불구하고 marginal 한 성능 향상만을 보여주었다. labeling 의 비용이 data 개수에 선형적인 것을 감안하면 500 개 이후부터는 labeling 작업의 ROI(Return of Investment)가 급격히 감소한다는 것을 확인할 수 있다.

train data 개수가 2,000 개 이상인 경우에는 4 개 모델이 모두 좋은 성능을 보여주었고, Dialog-KoElectra 는 Small 모델을 사용하여 parameter 의 개수가 58M 이었음에도 300M 의 Base 모델과 거의 유사한 성능을 내는 것을 확인할 수 있다.

train data 개수가 500 개 미만인 경우에는 KoBert, KoElectra 보다 KcBert, KcElectra 가 더 높은 성능을 보여주었다. 이는 Ko 계열 pretrain 에 사용된 위키피디아, 뉴스기사와 같은 정제된 corpus 보다, Kc 계열에서 학습한 네이버리뷰와 같은 구어체 문장들이 HR 평가데이터와 더 유사했다는 것을 의미한다고 볼 수 있다. 반대로 Ko 계열의 모델 또한 1,000 개 이상의 데이터부터는 HR 도메인의 특성을 대부분 학습했다는 결론을 내릴 수 있다.

2) post-training 적용결과

일반적인 corpus 를 통해 pretrained 된 BERT 보다 해당 도메인의 특화된 문장을 바탕으로 post-train 한 HRBERT 는 도메인에 연관된 contextualized representation 을 학습한다. 그리고 이러한 학습을 바탕으로 HR 과 관련된 downstream task 에 훨씬 높은 결과를 보여줄 수 있다.

표.3 에서는 KcBert 와 KoBert 를 29 만 문장의 corpus 로 Epoch 1,2,3 회 학습하였을 때의 성능을 비교하고 있다. data 개수가 2,000 개 보다 많은 경우에는 post-training 을 통해서 비교적 marginal 한 향상만을 얻을 수 있었다. 이는 데이터의 개수가 풍부한 상황에서는 train data 만을 통해서도 post-training 을 통해 학습하는 도메인 지식을 BERT 가 학습할 수 있다는 것을 의미한다.

데이터의 개수가 적은 경우에는 post-training 을 통해 매우 큰 성능향상 결과를 보였다. 특히 train data 가 100 개인 경우 KoBERT 의 f1-score 가 10%(0.776 → 0.858)이상 향상되었다. 이는 post-training 을 통해 적은 양의 train data 에는 들어있지 않은 정보를 추가 학습할 수 있었다는 것을 의미하며, 향후 few-shot learning 과 같이 train data 가 적은 상황에서 post-training 을 적극 활용하여 성능을 개선할 수 있음을 시사한다.

5. 결론

본 논문은 BERT 모델을 통해 평가문장이라는 HR 도메인 특성이 강한 문장을 4 개의 class 로 분류하는 문제를 다루었다. 이 과정에서 BERT 모델별 train data 개수에 따른 성능차이를 비교하였으며, post-training 을 통해 성능을 높일 수 있다는 의미 있는 결과를 도출하였다. 이는 데이터가 적은 상황에서 unlabeled 되어 있는 도메인 corpus 를 활용할 수 있는 방법론을 제안하였다는데 의의가 있다. 본 연구를 발전시켜 문장 classification 문제와 더불어

text generation task 에도 post-training 을 통한 도메인 지식 학습도 가능할 것으로 기대되며 이는 추후 연구과제로 남긴다.

참고문헌

[1] Vaswani et al, "Attention is all you need", Advances in neural information processing systems, p1-2, 2017

[2] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, p7, 2018

[3] Whang et al, "An effective domain adaptive post-training method for bert in response selection", arXiv preprint arXiv:1908.0482, p3-4, 2019

[4] Xu, Hu and Liu, Bing and Shu, Lei and Yu, Philip S, "BERT post-training for review reading comprehension and aspect-based sentiment analysis", arXiv preprint arXiv:1904.02232, p2, 2019

[5] Loshchilov, Ilya and Hutter, Frank, "Decoupled weight decay regularization", arXiv preprint arXiv:1711.05101, p4, 2017

[6] Clark et al, "Electra: Pre-training text encoders as discriminators rather than generators", arXiv preprint arXiv:2003.10555, p2, 2020