# 후보 레이블 정보를 반영한 멀티 디코더 모델

박원재○, 최기현, 김학수
건국대학교
ten94216@konkuk.ac.kr, pluto32@konkuk.ac.kr, nlpdrkim@konkuk.ac.kr

# Multi-decoder Model Reflecting Candidate Label Information

Won-Jae Park○, Gi-Hyeon Choi, Hark-Soo Kim
Konkuk University

**요 약**

지도 학습을 하기 위해선 레이블이 부착된 데이터셋이 필요하다. 크라우드소싱 서비스를 통해 데이터셋을 구축하는데 다수의 주석자(Annotator)가 관여한다. 다수의 주석자가 레이블을 할당하고 과반수인 레이블을 최종 정답으로 결정한다. 이 과정에서 최종 정답과 다른 후보 레이블의 정보가 누락된다. 이를 완화하고 목표 작업에 대한 성능을 높이기 위해 후보 레이블에 대한 정보를 반영하는 멀티 디코더 모델을 제안한다. KLUE-TC, SNLI, MNLI 데이터셋으로 정량적 성능 평가를 수행하였으며 실험한 데이터셋 모두 일괄적인 성능 향상을 보였다.

주제어: 주석자, 후보 레이블, 멀티 디코더

## 1. 서론

감성 분석, 개체명 인식, 텍스트 분류와 같은 여러 자연어처리 작업에서 딥러닝(Deep learning) 기반 모델들이 높은 성능을 보이고 있다[1-3].

딥러닝 기반 모델의 학습 방식은 크게 지도 학습(Supervised learning) 방식과 비지도 학습(Unsupervised learning) 방식으로 나뉜다. 지도 학습 방식이 비지도 학습 방식에 비해 높은 성능을 보이지만[4] 지도 학습 방식을 적용하기 위해서는 레이블이 부착된 데이터셋이 있어야 한다. 이를 위해 지도 학습을 위한 데이터셋을 구축할 때, 주로 다수의 주석자(Annotator)들이 데이터를 보고 적절한 레이블을 할당한다. 구축 과정에서 명확한 작업지침(Guideline)이 주어지지 않거나, 데이터가 갖고 있는 정보가 제한적인 경우에 데이터에 할당되는 레이블은 주석자 개인의 주관적인 판단에 의존하게 된다[5]. 이로 인해 동일한 데이터에 대해 각 주석자들의 의견이 서로 일치되지 않는 경우가 존재한다. 아래의 표 1은 이에 대한 예시이다.

표 1 KLUE-TC 데이터셋 예시

| 예시 기사 제목 | 주석자 | | | | 합의된 범주 | 실제 범주 |
|---|---|---|---|---|---|---|
| | 순위 / 주석자 | 1순위 | 2순위 | 3순위 | | |
| 단말기 보조금만 떠들고 요금할인 침묵하면 과징금 | A | 경제 | 해당 없음 | 해당 없음 | 경제 | IT과학 |
| | B | 경제 | 해당 없음 | 해당 없음 | | |
| | C | IT 과학 | 사회 | 해당 없음 | | |
| 신화 속 도시 트로이의 건립연대는 기원전 3천500년 | A | 세계 | 사회 | 생활 문화 | 생활 문화 | 세계 |
| | B | 생활 문화 | 해당 없음 | 해당 없음 | | |
| | C | 생활 문화 | 해당 없음 | 해당 없음 | | |

위의 표 1 예시는 한국어 주제 분류(Topic classification) 데이터셋인 KLUE-TC[6]에서 추출한 것이다. KLUE-TC는 저작권 문제로 인해 기사 본문 없이 뉴스 제목만을 보고 적합한 뉴스 범주(Category)를 할당한 데이터셋이다. 우선 3명의 주석자가 각각 관련성 순으로 최대 3개의 후보 범주를 할당하며 최종 정답으로 후보 범주 중 과반수인 범주가 할당된다. 위의 표 1의 두 예시를 보면 주석자가 할당한 여러 후보 범주 중 과반수가 아닌 후보 범주가 실제 뉴스 기사의 범주와 동일한 경우가 존재한다. 또, 첫 번째 예시의 기사 제목의 경우 중심 내용은 “IT 과학” 범주와 관련성이 높지만 “과징금”과 같은 일부 단어는 “경제” 범주와 관련성이 높다. 학습 단계에서 최종 정답으로 할당되는 과반수인 범주만 사용하게 되면 이러한 정보들이 모두 누락되는 문제점이 존재한다. 앞서 언급한 문제점을 완화하고 보다 정확한 예측을 위해서 후보 범주들에 대한 정보를 반영해 줄 필요가 있다. 따라서 본 논문에서는 지도 학습을 위한 데이터 구축 과정에서 생성된 후보 레이블들의 정보를 활용하여 목표 작업에 대한 성능을 향상시킬 수 있는 딥러닝 기반 모델을 제안한다. 제안 모델은 사전 학습 기반 언어 모델을 활용하여 입력 텍스트를 추상화하고 멀티 디코더를 통해 각 후보 레이블들에 대한 정보들을 학습한다. 제안 모델의 정량적 성능 평가를 위해 후보 레이블들이 주어진 KLUE-TC, SNLI[7], MNLI[8] 데이터셋을 사용한 실험에서 일괄적인 성능 향상을 보였다.

## 2. 관련 연구

지도 학습을 통한 예측 모델을 구축하기 위해서 레이블이 지정되지 않은 데이터셋에 범주를 할당하는 것은 중요한 작업이다. 기존에는 전문가들을 통해 데이터셋을 구축했기 때문에 많은 인적, 시간적 비용이 소요됐다

[9]. 크라우드소싱 서비스(Crowdsourcing services)의 등장으로 여러 주석자들이 모여 저렴하고, 짧은 시간 내에 레이블이 할당된 데이터셋을 얻는 것이 가능해졌다 [10]. 하지만 이 같은 방식은 주석의 품질을 관리하기가 어렵다는 문제점이 있다. 이러한 문제점을 완화하고 목표 작업에 대한 성능을 높이기 위해 여러 연구가 진행되었다. [11]은 주석자 편향 보정 기법과 범주형 데이터의 편향 보정 기법을 제안하여 다수의 비전문 주석자가 구축한 데이터셋의 품질을 전문가 구축한 데이터셋에 근접한 품질로 향상시켰다. [12]는 데이터 구축과정에서 주석자들이 할당한 레이블의 적합성을 평가할 수 있는 방법론을 제안하여 낮은 평가를 받는 주석자를 제외하는 방식으로 데이터의 품질을 향상시켰다. [13]은 각 주석자들을 기준으로 데이터를 분리한 후, 각 데이터들을 사용하여 모델을 독립적으로 학습시키는 앙상블(Ensemble)과 유사한 방법론을 제안하였다. 이를 통해 각 모델이 대응하는 주석자들의 관점을 학습할 수 있도록 유도하여 목표 작업에 대한 성능을 높였다.

## 3. 후보 레이블 정보를 반영한 멀티 디코더 모델

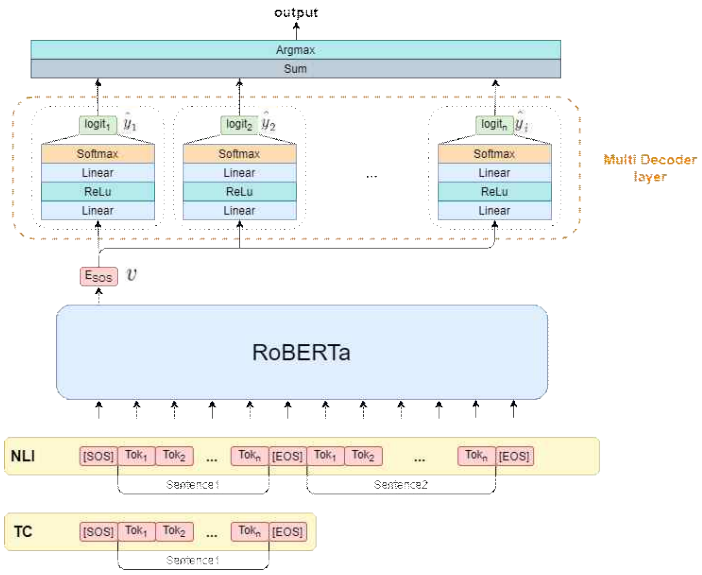


그림 1 제안 모델 전체 구조도

본 논문에서 제안하는 모델을 학습하려면 각 데이터의 후보 레이블에 대한 정보가 필수적으로 있어야 한다. 따라서 후보 레이블 정보가 포함된 주제 분류 데이터셋인 KLUE-TC와 자연어 추론(Natural language inference) 데이터셋인 SNLI, MNLI를 사용한다. 주제 분류는 입력 텍스트를 사전 정의된 주제에 맞춰 분류하는 작업을 의미하며 자연어 추론은 두 문장 사이의 의미적 관계를 연관(Entailment), 모순(Contradiction), 중립(Neutral)으로 정의하여 입력 문장쌍을 사전 정의된 관계에 맞춰 분류하는 작업을 의미한다.

제안 모델의 전체 구조는 그림 1과 같으며 크게 사전 학습 기반 언어 모델인 RoBERTa와 여러 개의 디코더로 구성된 멀티 디코더 계층으로 나뉘어진다. RoBERTa는 입력받은 문장 또는 문장쌍을 인코딩하여 입력 정보가 추상화된 벡터 $v$를 생성한다. $i$번째 디코더는 $v$를 입력받아 비선형 변환을 통해 $i$번째 디코더의 초점에 맞춘 벡터 $v_i$를 생성한다. 그리고 $v_i$를 활용하여 목표 작업에 대응하는 클래스 확률 분포 $\hat{y}_i$를 생성한다. 이에 대한 수식은 다음과 같다.

$$v_i = relu(W_i^1 \cdot v)$$
$$\hat{y}_i = softmax(W_i^2 \cdot v_i) \quad (1)$$

식 (1)에서 $W_i^1$, $W_i^2$는 가중치 행렬을 의미하며 랜덤 초기화(Random initialize)하여 사용하며 학습 과정 중에 미세 조정(Fine tuning) 된다. $i$번째 디코더는 $i$번째 후보 레이블 $y_i$ 사이의 크로스 엔트로피(Cross entropy)를 최소화하도록 학습되며 이에 대한 수식은 다음과 같다.

$$loss_i = -\sum_k y_i^k \log(\hat{y}_i^k) \quad (2)$$

식 (2)를 바탕으로 멀티 디코더 계층에 존재하는 각 디코더는 대응하는 후보 레이블에 맞춰 학습된다. 입력 데이터에 대한 최종 분류는 아래의 수식에 맞춰 예측된다.

$$\hat{y}_{total} = argmax(\sum_i \hat{y}_i) \quad (3)$$

## 4. 실험 및 결과

### 4.1 데이터셋

제안 모델의 정량적 성능 평가를 위해 실험 데이터로 KLUE-TC 데이터, SNLI, MNLI 데이터셋을 사용했다. KLUE-TC는 뉴스 본문 없이 제목을 보고 해당 기사의 뉴스 범주를 예측하는 데이터셋이다. 정치, 경제, 사회, 생활문화, 세계, IT과학, 스포츠로 총 7개의 뉴스 범주로 구성되어있다. 평가 데이터는 공개되어 있지 않으며 학습데이터 45,678개, 검증데이터 9,107개로 구성되어 있다. SNLI, MNLI 데이터셋은 자연어 추론을 위한 데이터셋으로 두 문장 사이의 의미적 관계를 분류하는 데이터셋이다. SNLI은 학습 데이터 550,152개, 검증 데이터 10,000개, 평가 데이터 10,000개로 구성되어 있다. MNLI는 학습 데이터 392,702개와 일치(Match) 검증 데이터 10,000개, 불일치(Mismatch) 검증 데이터 10,000개로 구성되어 있다. 일치 검증 데이터는 학습데이터와 동일한 출처에서 추출된 데이터를 의미하며, 불일치 검증 데이터는 학습데이터와 다른 출처에서 추출된 데이터를 의미한다.

### 4.2 5-fold 교차검증(Cross validation) 성능 측정

각각의 데이터셋을 활용하여 5-fold 교차 검증을 수행

하였다. KLUE-TC 데이터셋은 공개된 학습, 검증 데이터들을 모두 합친 후 전체 데이터를 4:1 비율로 나누어 실험하였다. 실험 결과 0.15%p ~ 0.50%p의 성능 향상을 보였으며, 그림 2는 이에 대한 실험 결과를 나타낸다. SNLI, MNLI 데이터셋은 5개의 후보 레이블이 모두 존재하는 데이터만 추출하여 실험에 사용하였으며 SNLI 데이터셋에서 추출한 데이터는 55,949개, MNLI 데이터셋에서 추출한 데이터는 19,647개이다. 추출한 데이터를 4:1 비율로 나누어 실험하였다. SNLI 데이터셋을 사용한 실험에서 0.23%p ~ 0.79%p의 성능 향상을 보였으며 그림 3은 이에 대한 실험 결과를 나타낸다. MNLI 데이터셋을 사용한 실험에서는 0.43%p ~ 0.87%p의 성능 향상을 보였으며 그림 4는 이에 대한 실험결과를 나타낸다.

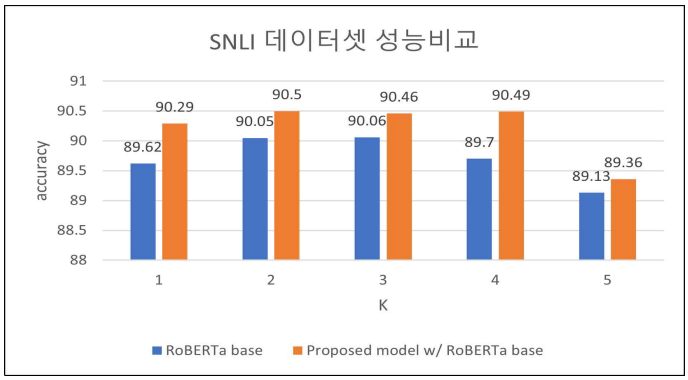


그림 2 SNLI 데이터셋 성능비교

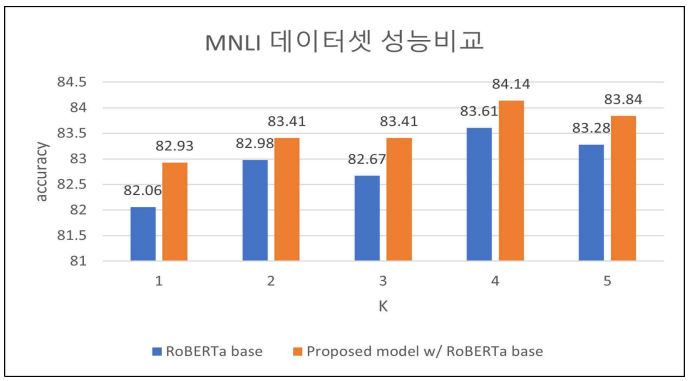


그림 3 MNLI 데이터셋 성능비교

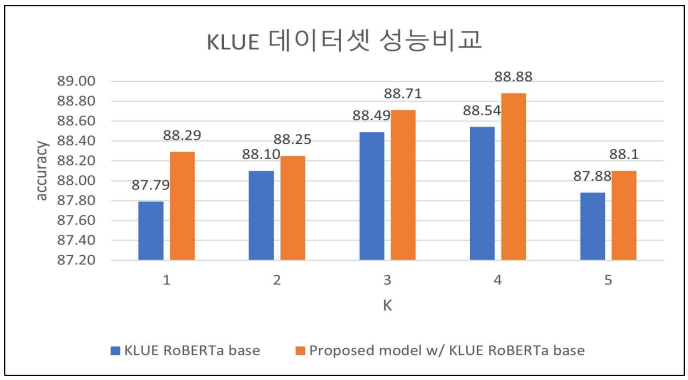


그림 4 KLUE 데이터셋 성능비교

### 4.3 언어모델 크기에 따른 성능 비교

표 2 언어모델 크기에 따른 성능 비교 실험

| 모델 | accuracy |
|---|---|
| KLUE RoBERTa small (our implements) | 86.2 |
| KLUE RoBERTa base (our implements) | 86.64 |
| KLUE RoBERTa large (our implements) | 87.16 |
| Proposed model w/ KLUE RoBERTa small | 86.9 |
| Proposed model w/ KLUE RoBERTa base | 87.15 |
| Proposed model w/ KLUE RoBERTa large | 87.77 |

위의 표 2는 언어모델의 크기에 따른 성능 비교 실험을 나타낸다. 실험결과 제안 모델이 언어모델의 크기에 상관없이 모두 분류 성능을 향상시키는 것을 확인할 수 있었다. 이러한 실험결과는 후보 레이블들에 대한 정보 반영이 목표 작업에 도움을 줄 수 있음을 의미한다.

## 5. 결론

본 논문에서는 데이터셋 구축과정에서 생성된 주석자들의 후보 레이블 정보를 활용하여 목표 작업에 대한 성능을 높인 멀티 디코더 모델을 제안했다. 실험을 통해 본 논문에서 제안하는 방법이 데이터셋이나 언어모델의 크기 변화에 상관없이 일관되게 성능을 향상시킬 수 있음을 보였다. 향후 연구로 각 주석자의 정보를 통해 모델이 각각의 주석자를 선별할 수 있고 보다 정답률이 높은 주석자의 정보를 활용할 수 있는 방안을 연구할 예정이다.

## 감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2020-0-00368, 뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발)

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2021-2016-0-00465)

## 참고문헌

[1] 최기현, 장영진, 김학수, 김관우, “독소 조항 분류를 위한 딥러닝 기반 텍스트 분류 모델”, 정보과학회논문지, 제47권, 제11호, pp. 1054-1060, 2020.

[2] 김홍진, 김담린, 김보은, 오신혁, 김학수, “감정

단어 등장 순서를 고려한 영화 리뷰 감성 분석”, 제 32회 한글 및 한국어 정보처리 학술대회 논문집, pp. 313-316, 2020.
[3] 김홍진, 오신혁, 김학수, “ELECTRA와 Label Attention Network를 이용한 한국어 개체명 인식”, 제 32회 한글 및 한국어 정보처리 학술대회 논문집, pp. 333-336, 2020.
[4] Bradley C.Love , "Comparing supervised and unsupervised category learning", Psychonomic Bulletin & Review 9, pp. 829-835, 2002.
[5] Igor Mozetic, Miha Grcar, Jasmina Smailovic, "Multilingual Twitter Sentiment Classification: The Role of Human Annotators", PLoS one, vol.11, 2016.
[6] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha and Kyunghyun Cho, KLUE: Korean Language Understanding Evaluation, https://arxiv.org/abs/2105.09680, 2021.
[7] Samuel R. Bowman, Gabor Angeli, Chirstopher Potts, Christopher D.Manning, “A large annotated corpus for learning natural language inference”, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632-642, 2015.
[8] Adina Williams, Nikita Nangia, Samuel Bowman, “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol .1, pp. 1112-1122, 2018
[9] Adam R Aron, Mask A Gluck, Russell A Poldrack , Long-term test-retest reliability of functional MRI in a classification learning task, NeuroImage, vol. 29, 3, pp. 1000-1006, 2005.
[10] Kevin Crowston, "Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars", Shaping the Future of ICT Research. Methods and Approaches, vol. 389, 2012.
[11] Rion Snow, Brendan O’Connor, Daniel Jurafsky, Andrew Ng , Cheap and fast but is it good? evaluating non-expert annotations for natural language tasks, Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 254-263, 2008.
[12] V. C. Raykar and S. Yu, "An Entropic Score to Rank Annotators for Crowdsourced Labeling Tasks," 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 29-32, 2011
[13] Sohail Akhtar, Valerio Basile, Viviana Patti, “Modeling Annotator Prespective and Polarized Opinions to Improve Hate Speech Detection”, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol.8, no. 1, pp. 151-154, 2020