패턴 추출 학습을 통한 한국어 주장 탐지 및 입장 분류

이우진°[1], 정석원[2], 김태일[3], 최성원[3], 김학수[1]
[1]건국대학교, [2]강원대학교, [3]네이버
[1]{shes100, nlpdrkim}@konkuk.ac.kr, [2]nlpsw@kangwon.ac.kr, [3]{eiji.kim, sungwon.choi}@navercorp.com

Claim Detection and Stance Classification through Pattern Extraction Learning in Korean

Woojin Lee°[1], Seokwon Jeong[2], Tae-il Kim[3], Sung-won Choi[3], Harksoo Kim[1]
[1]Konkuk University, [2]Kangwon University, [3]Naver

요 약

미세 조정은 대부분의 연구에서 사전학습 모델을 위한 표준 기법으로 활용되고 있으나, 최근 초거대 모델의 등장과 환경 오염 등의 문제로 인해 더 효율적인 사전학습 모델 활용 방법이 요구되고 있다. 패턴 추출 학습은 사전학습 모델을 효율적으로 활용하기 위해 제안된 방법으로, 본 논문에서는 한국어 주장 탐지 및 입장 분류를 위해 패턴 추출 학습을 활용하는 모델을 구현하였다. 우리는 기존 미세 조정 방식 모델과의 비교 실험을 통해 본 논문에서 구현한 한국어 주장 탐지 및 입장 분류 모델이 사전학습 단계에서 학습한 모델의 내부 지식을 효과적으로 활용할 수 있음을 보였다.

주제어: 주장 탐지(Claim Detection), 입장 분류(Stance Classification), 논증 마이닝(Argument Mining), 패턴 추출 학습(Pattern-Exploiting Training)

1. 서론

주장 탐지(Claim Detection)는 입력 텍스트가 주어진 주제를 지지하거나 반박하는지 즉, 주장에 해당하는지를 자동으로 식별하는 작업이며, 입장 분류(Claim Stance Classification)는 해당 주장이 찬성, 반대, 중립의 입장 후보 중 어떤 입장을 나타내는지 분류하는 작업이다. 주장 탐지와 입장 분류는 자동화된 토론 시스템[1], 가짜 뉴스 탐지[2]의 핵심 구성 요소이며, 소셜 미디어 포스트 및 정치적 온라인 댓글 등으로부터 대중의 반응을 분석하고 의견 다양성을 이해하기 위한 중요 도구로 활용되는[3, 4] 등 활용도가 높은 연구 분야이다.

BERT[5], RoBERTa[6]의 성공으로 인해 미세 조정(fine-tuning) 방식은 사전학습 모델을 활용하는 표준 기법으로 자리잡았으며, 기존의 주장 탐지 및 입장 분류 또한 미세 조정을 이용한 방법이 주로 연구되었다[7]. 그러나, 미세 조정만으로 온전히 활용하기 어려운 초거대 모델의 등장[8]과 이로 인한 환경 오염 문제[9]로 인해 사전학습 모델을 더 효율적으로 활용하기 위한 방법이 요구되었다. 패턴 추출 학습(PET; Pattern-Exploiting Training)[10]은 언급한 문제를 개선하기 위한 대안 중 하나로써, 감성 분석(Sentiment Analysis), 자연어 추론(Natural Language Inference) 등 여러 태스크에서 미세 조정 방식을 효과적으로 개선했다. 이후에 [11]에 의해 입장 분류 태스크에서도 패턴 추출 학습을 활용하는 방법이 제안되었다. 또한, 기존 주장 탐지와 입장 분류는 대부분 별개의 태스크로써 연구되었으나[12, 13, 14], 연관성이 높은 태스크를 함께 학습할 때 상호 보완적으로 성능을 향상시킨다는 연구 결과[15, 16]를 근거로 최근에는 두 태스크를 통합하는 방법이 제안되고 있다. 특히, [17]은 주장 탐지, 입장 분류, 근거 추출이라는 3가지 상호 보완적인 태스크에 대해 조사하고 이를 성공적으로 통합한 연구 결과를 보고했다.

본 논문에서는 한국어 환경에서 주장 탐지 및 입장 분류를 수행하기 위해 앞서 언급한 [11]의 모델을 재구현한다. 또한, 재구현한 방법과 미세 조정을 활용한 표준 분류 모델을 비교함으로써 패턴 추출 학습을 통해 한국어 환경에서도 사전학습 모델을 더 효과적으로 활용할 수 있음을 보인다.

2. 관련 연구

2.1 주장 탐지 및 입장 분류

주장 탐지의 초기 연구에서 [18]은 주제를 직접적으로 지지하거나 반대하는 간결한 문장인 문맥 의존 주장(Context Dependent Claim)의 개념을 정의하였고, 이를 기반으로 [19, 20]은 각각 주장 탐지 및 입장 분류의 벤치마크 데이터셋 및 전통적 기계학습 활용 방법을 제시하였다. 이후에는 BERT[5]와 같은 사전학습 모델을 통해 주장 탐지와 입장 분류의 성능이 대폭 향상되었다[21, 22, 23]. 이러한 성능 향상을 바탕으로 주장 탐지는 주로 가짜 뉴스 탐지[12, 13], 입장 분류는 소셜 미디어나 제품 리뷰 분석[14]과 같은 실제적인 분야에 활용되고 있다.

2.2 패턴 추출 학습

GPT-3[8]은 효율적인 모델 활용을 위해 프롬프트를 통한 문맥 내 학습(In-Context Learning) 방식을 사용하였다. 작업 목표에 대한 설명과 입력 문장을 함께 프롬프트로 입력하면, 모델은 이를 통해 목표 작업에 대한 적절한 예측을 생성한다. 이러한 문맥 내 학습은 추가적인 매개변수 업데이트 없이도 모델이 다

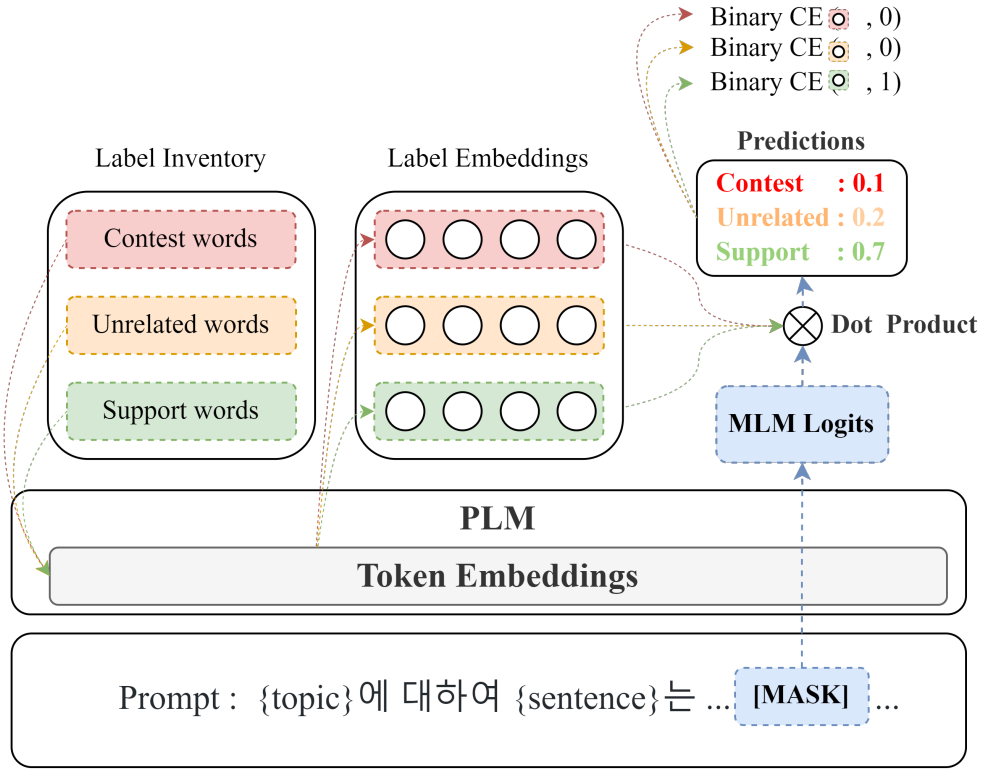

그림 1. 패턴 추출 학습 모델 구조 예시

양한 추론 작업을 수행할 수 있게 만들었으며, 이 방식을 통해 GPT-3는 여러 자연어 처리 작업에서 최고의 성능을 달성했다. 하지만 GPT-3는 약 1,750억 개의 매개변수로 이루어져 있는 초거대 모델로 저자원 환경에서의 활용이 어렵다. 이러한 문제를 극복하기 위해 [10, 24]는 BERT, RoBERTa와 같은 (상대적으로) 소규모 사전학습 모델에서 내부 지식을 최대한 활용할 수 있는 프롬프트 조정(prompt-tuning) 방법론인 패턴 추출 학습을 제안하였다. 패턴 추출 학습은 입력 문장 X에 템플릿 함수 T를 적용하여 T(X)="X It was [MASK]." 형태의 템플릿 문장을 생성한다. 그리고 이 템플릿 문장을 모델에 입력하여 [MASK] 토큰을 verbalizer 함수에 해당하는 토큰(예시: *bad*, *great*)으로 예측하도록 모델을 학습시킨다. [11]은 기존의 패턴 추출 학습을 변형하여 다국어 입장 분류에 적합한 모델을 제안하였다. 이 모델은 verbalizer 함수를 사용하는 기존의 패턴 추출 학습과 달리 레이블 인코더를 도입하여 단일 토큰 레이블과 다중 토큰 레이블에 대한 일관된 처리 방식을 제공하며, [10, 24]에서 주로 사용한 구두점 기반 프롬프트 대신 작업의 목적을 더 명확하게 반영하는 프롬프트 형식을 채택하였다.

3. 제안 모델

본 연구에서는 패턴 추출 학습을 활용하는 [11]의 모델을 구현하였다. 그림 1은 본 논문에서 구현한 모델의 구조이다. 모델은 사전 학습된 언어 모델에 특정 프롬프트를 입력하고 [MASK] 토큰의 임베딩과 레이블 임베딩의 내적(Dot Product)을 통해 레이블을 예측한다. 레이블 임베딩을 비롯한 프롬프트와 관련된 세부 구조는 각 절에서 자세히 설명한다.

3.1 프롬프트 설정

제안 모델은 주제, 입력 텍스트 등을 사전에 정의된 프롬프트 형태로 입력한다. 프롬프트 구성에 대한 예시는 그림 1에서 확인할 수 있다. 예시에서 {topic}은 주제, {sentence}는 입력 문장을 나타낸다. 제안 모델은 주장 탐지를 위해 "{topic}에 대하여 {sentence}는 주장으로 판단할 수 [MASK]다."라는 프롬프트를 사용하고, 입장 분류를 위해 "{topic}에 대하여 {sentence}는 [MASK]한 입장을 보인다."라는 프롬프트를 사용한다. 두 태스크를 통합하는 경우에도 동일한 프롬프트 형식을 사용한다.

3.2 레이블 임베딩

레이블 임베딩은 각 레이블 별로 대표 단어들의 표현을 평균한 값으로 초기화한다. 각 레이블의 대표 단어를 추출하기 위해 별도의 사전학습 모델에 프롬프트를 입력하고 [MASK] 위치에 예측되는 단어 중 레이블의 의미에 적합하다고 생각되는 단어를 임의로 선택하여 각 레이블 인벤토리에 포함한다. 예를 들어, 입장 분류의 경우 "반박(contest)", "지지(support)"라는 입장을 분류해야 한다. 이 때, 프롬프트를 입력하여 [MASK] 위치에

	train	dev	test
# sentences as claim candidates	55,544	7,057	7,065
# claims	3,871	492	527
# support claims	2,098	259	256
# contest claims	1,773	233	271

표 1. 데이터셋 통계

Task	Model	Unrelated	Contest	Support	Macro F1
Claim Extraction	RoBERTa	95.59	50.86		73.22
	RoBERTa + Label Embedding	95.61	50.52		73.06
	Ours	**95.66**	**51.26**		**73.46**
Stance Classification	RoBERTa	-	77.26	76.02	76.64
	RoBERTa + Label Embedding	-	80.35	76.41	78.38
	Ours	-	**80.42**	**77.21**	**79.87**
CESC	RoBERTa	95.53	42.01	37.93	58.49
	RoBERTa + Label Embedding	**95.66**	40.07	35.39	57.04
	Ours	95.06	**43.08**	**40.76**	**59.63**

표 2. 실험 결과

예측되는 단어 중 "부정", "반대", "상반"은 "반박" 레이블과 의미적으로 일치하므로, "반박" 레이블의 인벤토리에 해당 단어들을 추가한다. 각 단어 L_t의 임베딩을 평균하여 레이블 표현 LE_L을 생성한다.

$$LE_L = \frac{1}{N}\sum_{t=0}^{N} TokEmb(L_t); \quad \forall L \in \{\text{Labels}\} \qquad (1)$$

3.3 손실 함수

모델은 각 레이블 표현 LE_L과 [MASK] 토큰 위치에 출력된 MLM logit 값의 내적 간에 이진 교차 엔트로피 손실 함수를 적용하여 각 레이블의 오차를 역전파한다. 수식 표현은 다음과 같다.

$$\mathcal{L}_{LE} = \sum_{y' \in y^p} \text{BCE}(p(y'|x), 1) + \sum_{y'' \in y^n} \text{BCE}(p(y''|x), 0)) \quad (2)$$

4. 실험

4.1 데이터셋

본 논문에서는 한국어 환경에서의 주장 탐지 및 입장 분류를 위해 IAM 데이터셋[17]을 한국어로 번역하여 사용하였다. IAM 데이터셋은 123개 주제와 관련된 1천 개 이상의 기사에서 수집되었으며 주장 탐지, 입장 분류, 증거 추출을 위해 구축된 고품질 데이터셋이다. 데이터셋의 자세한 통계는 표 1에서 확인할 수 있다.

4.2 비교 모델 및 평가 지표

본 논문의 실험에서는 두 가지의 비교 모델을 활용했다. 먼저, 미세 조정과 패턴 추출 학습의 차이를 확인하기 위해 제안 모델을 미세 조정 방식의 표준 분류 모델(RoBERTa)과 비교했다. 또한, 제안 모델의 성능 향상이 구조적 차이로 인한 것이 아니라 프롬프트를 통한 패턴 학습으로 인한 것임을 확인하기 위해 프롬프트를 제외한 나머지 구조를 동일하게 설정한 모델(RoBERTa + Label Embedding)과 성능을 비교했다. 성능 평가를 위한 지표로는 레이블별 F1 점수를 측정했다. 또한, 레이블 불균형을 고려하고 빈도가 적은 레이블의 성능이 무시되는 것을 방지하기 위해 Macro-F1 점수를 측정했다.

4.3 실험 환경

본 연구에서는 PyTorch-Lightning[1] 프레임워크와 Huggingface[2] 라이브러리의 klue/roberta-base[25] 모델을 사용했다. 실험을 위한 최적화 알고리즘은 AdamW를 사용하였고, 배치 크기는 16, 학습률은 $5e{-}5$, 시드 값은 42로 설정하여 총 10 에포크(epoch)를 학습했다.

4.4 실험 결과

실험 결과는 표 2와 같다. 미세 조정을 활용한 기존의 표준 분류 모델에 비해 패턴 추출 학습을 활용한 제안 모델은 모든

[1]https://lightning.ai/docs/pytorch/stable/
[2]https://huggingface.co/

태스크에서 더 좋은 성능을 달성했다. 특히, 표준 분류 모델에 레이블 임베딩을 추가하여 모델 구조적인 차이를 최소화했음에도 불구하고 제안 모델이 더 높은 성능을 보이는 것을 확인할 수 있다. 이는 프롬프트를 활용하는 방식이 기존 방식보다 모델의 내재적 지식을 효과적으로 활용하여 목표 작업을 성공적으로 수행한다는 것을 의미한다. 또한, 본 논문에서 설정한 레이블 인벤토리 구성 및 프롬프트 설정에서 패턴 추출 학습 모델이 효과적으로 동작한다는 것을 확인할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 한국어 주장 탐지와 입장 분류에 대해 사전학습 모델을 효율적으로 활용하는 패턴 추출 학습 방식의 모델을 구현했다. 비교 실험 결과, 패턴 추출 학습은 미세 조정 방식에 비해 주장 탐지 및 입장 분류에서 더 향상된 성능을 보이는 것을 확인했다. 이러한 결과는 프롬프트를 활용한 패턴이 모델에 내재된 지식을 더 효과적으로 활용하게 함으로써 목표 작업의 성능을 향상시키는 역할을 하는 것으로 해석할 수 있다.

감사의 글

본 연구는 NAVER(주)에 의해 지원된 과제(뉴스 기사의 심층성 및 다양성 측정 모델 개발)로 수행되었음. 또한 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2023-00216011, 사람처럼 개념적으로 이해/추론이 가능한 복합인공지능 원천기술 연구)

참고문헌

[1] N. Slonim, Y. Bilu, C. Alzate, R. Bar-Haim, B. Bogin, F. Bonin, L. Choshen, E. Cohen-Karlik, L. Dankin, L. Edelstein *et al.*, "An autonomous debating system," *Nature*, Vol. 591, No. 7850, pp. 379–384, 2021.

[2] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, Vol. 57, No. 2, p. 102025, 2020.

[3] A. AlDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, Vol. 58, No. 4, p. 102597, 2021.

[4] Y. Li, T. Sosea, A. Sawant, A. J. Nair, D. Inkpen, and C. Caragea, "P-stance: A large dataset for stance detection in political domain," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2355–2365, 2021.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[7] S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh, "Stance detection in web and social media: a comparative study," *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pp. 75–87, 2019.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[9] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland, "Risks and benefits of large language models for the environment," *Environmental Science & Technology*, Vol. 57, No. 9, pp. 3464–3466, 2023.

[10] T. Schick and H. Schütze, "Exploiting cloze questions for few shot text classification and natural language inference," *arXiv preprint arXiv:2001.07676*, 2020.

[11] M. Hardalov, A. Arora, P. Nakov, and I. Augenstein, "Few-shot cross-lingual stance detection with sentiment-based pre-training," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 10, pp. 10 729–10 737, 2022.

[12] J. Beltrán, R. Míguez, and I. Larraz, "Claimhunter: An unattended tool for automated claim detection on twitter." *KnOD@ WWW*, 2021.

[13] J. Ding, Y. Hu, and H. Chang, "Bert-based mental model, a better fake news detector," *Proceedings of the 2020 6th international conference on computing and artificial intelligence*, pp. 396–400, 2020.

[14] H. Karande, R. Walambe, V. Benjamin, K. Kotecha, and T. Raghu, "Stance detection with bert embeddings for credibility analysis of information on social media," *PeerJ Computer Science*, Vol. 7, p. e467, 2021.

[15] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[16] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, Vol. 5, No. 1, pp. 30–43, 2018.

[17] L. Cheng, L. Bing, R. He, Q. Yu, Y. Zhang, and L. Si, "Iam: A comprehensive and large-scale dataset for integrated argument mining tasks," *arXiv preprint arXiv:2203.12257*, 2022.

[18] E. Aharoni, A. Polnarov, T. Lavee, D. Hershcovich, R. Levy, R. Rinott, D. Gutfreund, and N. Slonim, "A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics," *Proceedings of the first workshop on argumentation mining*, pp. 64–68, 2014.

[19] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim, "Context dependent claim detection," *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1489–1500, 2014.

[20] R. Bar-Haim, I. Bhattacharya, F. Dinuzzo, A. Saha, and N. Slonim, "Stance classification of context-dependent claims," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 251–261, 2017.

[21] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "Stancy: Stance classification based on consistency cues," *arXiv preprint arXiv:1910.06048*, 2019.

[22] E. Allaway and K. McKeown, "Zero-shot stance detection: A dataset and model using generalized topic representations," *arXiv preprint arXiv:2010.03640*, 2020.

[23] B. Liang, Q. Zhu, X. Li, M. Yang, L. Gui, Y. He, and R. Xu, "Jointcl: a joint contrastive learning framework for zero-shot stance detection," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 81–91, 2022.

[24] T. Schick and H. Schütze, "It's not just size that matters: Small language models are also few-shot learners," *arXiv preprint arXiv:2009.07118*, 2020.

[25] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh *et al.*, "Klue: Korean language understanding evaluation," *arXiv preprint arXiv:2105.09680*, 2021.