# 해석 가능한 프롬프트 최적화에 관한 강화학습 연구

**최윤선°, 반성현, 김기웅**
**한국과학기술원**
**{cys9506, bansh123, kekim}@kaist.ac.kr**

# Explainable Prompt Optimization via Guided Exploration in Reinforcement Learning

**Yunseon Choi°, Seonghyun Ban, Kee-Eung Kim**
**KAIST**

**요 약**

거대 언어 모델은 다양한 자연어 태스크에 효과적으로 적용될 수 있는 잠재력을 지니고 있으나, 사전 훈련된 대규모의 매개변수를 특정 하위 작업에 맞게 업데이트하는 기존의 파인튜닝은 대량의 컴퓨팅 자원을 요구한다. 이에 모든 매개변수를 업데이트 하는 대신 기존 입력에 프롬프트를 추가하여 이를 학습하는 여러가지 프롬프팅 방법론들이 제시되어 왔으나, 학습된 프롬프트를 사람이 보았을 때 해석하는 것이 불가능하거나 친숙하지 않다는 공통적인 한계점이 있었다. 본 논문에서는 이러한 한계점을 해결하기 위해, 강화학습 분야의 행동 모사 학습을 기반으로 사람에게 친숙한 예제 프롬프트를 활용하는 해석 가능한 프롬프트 최적화 알고리즘을 제안한다.

## 1. 서론

최근 프롬프팅 (Prompting)은 사전 훈련된 거대 언어 모델 (LLMs) [1]을 사용하여 다양한 자연어 태스크를 해결하는데 유망한 접근법으로 부상하고 있다. 프롬프팅은 거대 언어 모델이 특정 하위 작업을 효과적으로 수행하도록 하기 위해, 프롬프트라 불리우는 특정 입력값을 기존 입력에 추가한 상태로 모델에 전달하는 기법이다. 예로, 소수 훈련 예제를 이용한 텍스트 분류 작업에서는 프롬프트와 분류해야 하는 텍스트를 합쳐 언어 모델에게 입력하였을 때, 주어진 텍스트에 해당하는 클래스의 예측 확률을 높이도록 만드는 프롬프트를 찾는 것이다. 이러한 프롬프팅은 프롬프트를 이루는 매개변수만 업데이트하기 때문에 모든 매개변수를 업데이트하는 전통적인 파인 튜닝에 비해 상대적으로 컴퓨팅 자원이 적게 드는 이점이 있다.

프롬프팅 방법의 하나인 소프트 프롬프트는 (Soft Prompts) [2] 언어 모델의 입력으로 들어가는 연속 임베딩 벡터를 경사 하강법으로 학습하는 것이다. 이렇게 학습된 소프트 프롬프트는 실제 어휘 형태로 존재하는 것이 아니라, 모델에 따라 크기가 상이한 임베딩 공간의 연속적 벡터 형태로 존재하기 때문에 사람이 해석하거나 다른 언어 모델에 적용하기 어려움이 있다. 어휘로 정의될 수 있는 이산적인 토큰의 조합으로 프롬프트를 구성한다면 이러한 문제를 개선할 수 있으나, 프롬프트에 직접적으로 경사 하강법을 적용하는 것이 어려워져 다른 형태의 학습방법이 요구된다. 강화학습은 이러한 이산 프롬프트 최적화에 적합하여 활용될 수 있음이 기존 연구 [3]에서 다뤄져 왔다.

수많은 어휘를 고려해야 하는 이산 최적화 문제는 규모 측면에서 해결하기 어려움이 있다. RLPROMPT [3]는 이를 우회하기 위해, 작은 규모의 MLP 계층으로 이루어지는 연속적 임베딩 공간 변환 함수를 학습한다. 그리고 기존의 언어 모델의 LM-HEAD를 이용하여 이산 토큰들로 구성되는 프롬프트를 결정한다. 하지만 이렇게 학습되어 나오는 이산 토큰들 또한 사람에게는 낯선 단어들로 구성되며, 보통 해석할 수 없는 어려움이 있다. 이러한 이유는 효율적으로 이산 최적화를 진행하였으나, 여전히 방대한 어휘 개수 때문에 사람에게 친숙한 어휘의 탐색이 부족하기 때문이다.

본 논문에서는 사람에게 친숙한 예제 프롬프트를 활용하는 행동 모사 학습을 기반으로, 모델의 탐색 범위를 친숙한 어휘들을 중심으로 설정함으로써 상대적으로 해석하기에 용이하며 서로 다른 언어 모델 사이에서 상호 호환할 수 있는 프롬프트를 찾는 강화학습 알고리즘을 제안한다.[1]

## 2. 연구 배경

주어진 태스크를 수행하기 위해 어휘 공간 내에서 효과적인 프롬프트를 식별하기 위한 목적식은 다음과 같다.

$$\max_{\vec{\boldsymbol{w}} \in \mathcal{V}^T} \mathbb{E}_{x \sim \text{LM}(\vec{\boldsymbol{w}})}[R(x)] \tag{1}$$

여기서 LM은 언어 모델을 나타내며, 어휘의 조합인 프롬프트 $\vec{\boldsymbol{w}}$을 이용하여 언어 모델이 텍스트 $x$를 생성할 때 기대 보상 $R(x)$을 최대화하는 프롬프트 $\vec{\boldsymbol{w}}$을 찾는 것이다. $\mathcal{V}$는 이산 어휘 공간을 뜻하며, $T$는 프롬프트의 최대 토큰 길이이다.

**효율적 매개 변수화 정책** 목적식 (1)을 최적화하는 어휘를 찾기 위해 단순히 모든 어휘에 대한 기대 가치값 (Q-value)을 매개변수화하여 Q-러닝을 진행하면, 방대한 어휘 공간으로 인한 학습의

1) 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2022-0-00311, 일상적 물건들의 다접촉 로봇 조작을 위한 목적지향 강화학습 기술 개발, No.2020-0-00940, 안전한 강화학습 원천 기술 개발 및 자연어 처리에의 응용)

어려움이 있다. RLPROMPT [3]에서는 이를 완화하기 위해 다음과 같은 구조로 Q-값을 효율적으로 추정하고자 하였다.

$$\begin{aligned}\hat{z}_{<t} &= \text{LM}(w_{<t}) \\ z_t &:= \pi_\theta(\hat{z}_{<t}) \\ p_\theta^{\text{LM}}(w_t|w_{<t}) &:= \text{softmax}(\hat{W}^{\text{LM}}\pi_\theta(\hat{z}_{<t}))\end{aligned} \tag{2}$$

여기서 $\mathcal{Z} : \mathbb{R}^{\dim(\mathcal{Z})}$는 토큰의 임베딩 공간을 의미하며, 매개변수화된 정책은 $\pi_\theta : \mathcal{Z} \to \mathcal{Z}$ 임베딩 공간의 변환 함수이다. 이전 시간까지의 어휘 $w_{<t}$의 입력으로 나오는 언어 모델의 출력값 $\hat{z}_{<t}$을 현재 상태로 정의하며, 학습을 진행해야 하는 정책 $\pi_\theta$은 $\hat{z}_{<t}$을 바탕으로 새로운 임베딩 벡터 $z_t$를 결정한다. 기존의 언어 모델이 가지고 있는 LM-HEAD, $\hat{W}^{\text{LM}} \in \mathbb{R}^{|\mathcal{V}|\times\dim(\mathcal{Z})}$와 $z_t$의 행렬곱셈으로 결정되는 로짓에 소프트맥스 활성화 함수가 적용되어, 프롬프트의 t번째 토큰에 해당하는 어휘의 확률이 결정된다.

**소프트 Q-러닝** 식 (2)로 정의된 구조를 사용한다면, 현재 상태 $\hat{z}_{<t}$에서 단어 $w_i$에 대한 Q-값 $Q(w_i, \hat{z}_{<t})$은 다음과 같이 표현된다.

$$\begin{pmatrix} Q(w_1, \hat{z}_{<t}) \\ \cdots \\ Q(w_i, \hat{z}_{<t}) \\ \cdots \\ Q(w_{|\mathcal{V}|}, \hat{z}_{<t}) \end{pmatrix} = \begin{pmatrix} -\hat{w}_1^T- \\ \cdots \\ -\hat{w}_i^T- \\ \cdots \\ -\hat{w}_{|\mathcal{V}|}^T- \end{pmatrix} \pi_\theta(\hat{z}_{<t}) = \hat{W}^{\text{LM}}\pi_\theta(\hat{z}_{<t}) \tag{3}$$

$\hat{w}_i^T$는 $\hat{W}^{\text{LM}}$을 이루는 $i$-번째의 행벡터를 의미한다. 소프트 Q-러닝 [4]을 통해서 Q-값을 추정할 수 있으며, 이때 타겟값 $y$은

$$y = \begin{cases} V(\hat{z}_{\leq t}) & t < T \\ V(\hat{z}_{\leq t}) + R(x) & t = T \end{cases} \tag{4}$$

이다. 여기서 상태 함수 $V(\hat{z}_{\leq t})$는 템퍼러처 상수 $\beta \in \mathbb{R}$와 함께 $V(\hat{z}_{\leq t}) = \beta \log \sum_w \exp(Q_\theta(w, \hat{z}_{\leq t})/\beta)$ 로 정의한다. 최종적으로 Q-값을 학습하기 위한 목적 함수는 식 (5)이며, $\bar{y}$는 타겟값 $y$ 자체를 구할 때 얻어지는 $\theta$에 대한 그레디언트는 반영하지 않음을 나타낸다.

$$\min_\theta \mathbb{E}_{w_t \sim p_\theta^{\text{LM}}}[(Q_\theta(w_t, \hat{z}_{<t}) - \bar{y})^2] \tag{5}$$

## 3. 본론

효율적인 매개 변수화 정책을 소프트-Q 러닝을 통해 성공적으로 학습할 수 있으나, 학습된 프롬프트를 살펴보면 사람한테 낯선 단어들의 조합으로 보통 나타난다. 즉, 방대한 어휘 공간에서 사전 지식 없이 해석할 수 있는 프롬프트를 학습하는 것은 어려움이 존재하는 것이다. 따라서 본 논문은 사전 예제를 이용하여, 해석할 수 있는 프롬프트를 학습하는 것을 목표로 한다.

**사전 행동 모사 학습** 프롬프트로 사용할 수 있는 사전 예제 $\mathcal{D}_{\text{demo}}$를 초기에 정책한데 주어, 행동 모사 학습을 수행한다. 그 후, 학습되어진 값을 바탕으로 소프트맥스 활성화 함수 이전의 로짓, Q-함수의 초기값으로 사용한다. 행동 모사 목적식 (6) 이며, 사전 예제 $\mathcal{D}_{\text{demo}}$에 존재하는 단어의 Q-값은 증가시키는 부분과 모든 단어에 대한 Q-값의 로그-합-지수 값은 발산하지 않도록 하는 정규식으로 표현된다.

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{w}\sim\mathcal{D}_{\text{demo}}}[\log p_\theta^{\text{LM}}(\boldsymbol{w})] \\ &\quad= \mathbb{E}_{w_t\sim\mathcal{D}_{\text{demo}}}[\log p_\theta^{\text{LM}}(w_t|w_{<t})] \\ &\quad= \mathbb{E}[Q_\theta(w_t, z_{<t}) - \log \sum_w \exp Q_\theta(w, z_{<t})]\end{aligned} \tag{6}$$

이로 초기화된 Q-값으로부터 기존의 소프트 Q-러닝 알고리즘을 통해 학습을 진행하면 사전 예제 $\mathcal{D}_{\text{demo}}$에서 나타난 어휘와 그 어휘와 토큰 임베딩 공간에서 가까운 곳에 위치한 어휘들을 중심으로 초기 탐색이 시작된다. 사전 예제에서 나온 어휘들의 조합이 사람한테 익숙하고 해석할 수 있는 문장이었다면, 이러한 어휘들의 토큰 임베딩 지점을 시작점으로 탐색하여 기대 보상을 높이는 어휘들 찾아간다면, 해석할 수 있으며 더 좋은 성능을 가지는 프롬프트를 찾을 가능성이 있다.

## 4. 실험

**소수의 예시를 이용한 텍스트 분류 (Few-shot text classification)** 적은 수의 레이블이 부착된 예제만 이용하여 텍스트를 분류해야 하는 태스크이다. BERT와 같은 MLM을 위한 토큰 채우기나 GPT와 같은 왼쪽에서 오른쪽으로 읽는 LM을 위한 다음 토큰 예측을 통해 분류 태스크를 수행 할 수 있다. 분류는 미리 결정된 클래스 레이블 집합에 해당하는 토큰을 선택하는 것으로, 예로, “great”은 긍정적인 감정을 위해, “terrible”은 부정적인 감정을 위해 표현하는 verbalizer가 있다. MLM을 사용하여 입력 문장의 감정을 분류하려면 먼저 프롬프트 Prompt와 분류해야 하는 입력 문장 Sentence을 템플릿 “[Sentence] [Prompt] [MASK]”에 채워넣고, 그 다음 [MASK] 위치에 채워넣을 때 가장 높은 확률을 가진 verbalizer 토큰을 선택한다. “great”와 “terrible” 과 같은 감정을 분류하기 위해, Manual Prompt로는 “It was”를 사용하였다.

**실험 환경** 소수 예시를 이용한 텍스트 분류 태스크를 위해 감정 분류의 대표적인 데이터셋 SST-2[5], MR[6]와 CR[7]에서 실험을 진행하였다. distilRoBERTa-base를 기본 모델로 사용하였으며, Manual Prompt 와 RLPrompt와 성능을 비교하였다. 여기서 Manual Prompt는 우리 알고리즘이 사전에 모사하는 프롬프트와 동일하다.

**결과** 생성한 문장의 보상 함수 $R$은 [3]의 정의를 따랐다. 결과는 표 1와 같으며, 학습된 프롬프트는 표 2에 나타나있다. RLPrompt와 제안한 알고리즘의 하이퍼파라미터 $\beta$는 동일하게 1로 하였다. 결과표를 보면, 제안한 알고리즘이 비교 알고리즘에 비해 높은 정확도를 가자며 학습된 프롬프트를 살펴봐도 사람이 이해할 수 있다는 것을 알 수 있다.

| Dataset | SST-2 | MR | CR |
|---|---|---|---|
| Manual Prompt | 82.2 | 77.7 | 84.4 |
| RLPrompt | 79.5 ±0.8 | 76.2 ±0.3 | 84.7 ±0.9 |
| Ours | 84.1 ±0.1 | 79.7 ±0.2 | 86.1 ±0.1 |

표 1: 각 데이터 셋의 분류 정확도. 16개의 예제만 이용하여 프롬프트를 찾았으며, 3번 반복 실험 결과의 평균값과 표준 오차를 나타냄.

| Dataset | SST-2 |
|---|---|
| Manual Prompt | It was [MASK] |
| RLPrompt | 409265 highly [MASK] |
| Ours | It is particularly [MASK] |
| Ours | is especially equally [MASK] |
| Dataset | MR |
| Manual Prompt | It was [MASK] |
| RLPrompt | Ramos613 equally [MASK] |
| Ours | is very Quite [MASK] |
| Ours | It is absolute [MASK] |
| Dataset | CR |
| Manual Prompt | It was [MASK] |
| RLPrompt | 308608 feels [MASK] |
| Ours | It is Truly [MASK] |
| Ours | incredibly feeling especially [MASK] |

표 2: 각 데이터 셋에서 이용 또는 학습한 프롬프트 결과. Manual Prompt를 제외하고, 프롬프트의 토큰 길이는 3으로 함.

## 5. 결론

본 논문에서는 기존 프롬프팅 기법의 공통적인 한계인 프롬프트의 해석 가능성을 다루었다. 우리의 방법론은 사전 예제를 기반으로하는 행동 모사 학습을 도입함으로써 친숙한 어휘를 중심으로 초기 탐색을 시작하며, 이를 통해 사람이 해석하기에 용이한 프롬프트를 찾아내는 것을 목표로 한다. 다양한 데이터셋을 사용한 텍스트 분류 태스크에서의 검증을 통해, 우리의 방법론이 실제로 사람에게 친숙한 어휘로 구성되는 프롬프트를 생성해내며, 더 나아가 기존 방법론 보다 좋은 성능을 달성함을 보였다.

## 참고 문헌


[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[2] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 4582–4597, Association for Computational Linguistics, Aug. 2021.

[3] M. Deng, J. Wang, C.-P. Hsieh, Y. Wang, H. Guo, T. Shu, M. Song, E. Xing, and Z. Hu, "RLPrompt: Optimizing discrete text prompts with reinforcement learning," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, (Abu Dhabi, United Arab Emirates), pp. 3369–3391, Association for Computational Linguistics, Dec. 2022.

[4] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361, PMLR, 06–11 Aug 2017.

[5] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (Seattle, Washington, USA), pp. 1631–1642, Association for Computational Linguistics, Oct. 2013.

[6] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, (Ann Arbor, Michigan), pp. 115–124, Association for Computational Linguistics, June 2005.

[7] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, (New York, NY, USA), p. 168–177, Association for Computing Machinery, 2004.