사고 구조 시그니처 기반 인지왜곡 탐지 기술 연구

김민석 [1], 강미경 [1], S. Shyam Sundar[1, 2], 한진영 [1*]
[1] 성균관대학교, [2] 펜실베니아 주립대학교

seo9ky@g.skku.edu, gy77@g.skku.edu, sss12@psu.edu, jinyounghan@skku.edu*

Cognitive Distortion Detection based on Cognitive Structure Signatures

Minseok Kim[1], Migyeong Kang[1], S. Shyam Sundar[1, 2], Jinyoung Han[1*]

[1]Sungkyunkwan Univ., [2]Pennsylvania State Univ.

요 약

인지왜곡(Cognitive Distortion)은 개인이 외부 사건을 해석하는 과정에서 사고가 비합리적으로 왜곡되는 현상으로, 발화 속 인지왜곡을 탐지하는 것은 심리치료의 효과를 높이는 핵심 요소이다. 그러나 기존 인지왜곡 탐지 연구는 주로 발화의 표면적 내용 분석에 치중해, 인지왜곡 사고의 구조적 특성을 충분히 반영하지 못했다. 본 연구는 이러한 한계를 보완하기 위해 대규모언어모델(LLM)을 활용해 인지왜곡 사고 구조의 특성을 추론하는 Signature-of-Thought(SoT) 기법을 제안한다. SoT 는 LLM 이 다양한 인지왜곡 사고의 전형적 패턴과 단서를 통합해 사고 구조 시그니처(Signature)를 정의하고, 입력 발화와의 사고 구조 유사성을 비교하여 인지왜곡 여부를 판단하도록 설계된 방식이다. 인지왜곡 탐지 벤치마크 데이터셋을 활용한 실험 결과, 제안한 SoT 기법은 기존 접근법 대비 F1-Score 기준 최대 9% 이상 상대적 성능 향상을 달성함으로써 사고 구조 시그니처 기반 접근이 LLM 의 인지왜곡 탐지의 정확도와 일관성을 향상시키는 데 효과적임을 시사한다.

Ⅰ. 서론

인지왜곡(Cognitive Distortion)은 개인이 외부 사건을 해석하는 과정에서 비합리적인 사고 패턴을 보이는 심리적 현상이다. 이는 단순한 부정적 사고가 아니라, 논리적 비약, 과잉 일반화, 흑백논리적 결론 등 다양한 인지적 오류를 포함한다. Beck 은 인지왜곡을 "감정적 고통을 심화시키는 현실에 대한 왜곡된 사고"로 정의하며, 이를 우울, 불안, 자존감 저하 등 정신건강 문제의 근본 원인으로 지적하였다[1].

이러한 인지왜곡을 대규모언어모델(LLM)로 탐지하는 기술은 두 가지 측면에서 의의가 있다. 첫째, 상담자나 인공지능 상담 시스템이 개인의 비합리적 사고 신호를 조기에 탐지하여 정서 조절 및 인지 재구조화 치료를 효율적으로 수행할 수 있다. 둘째, 온라인 환경에서 생성되는 대규모 심리 데이터를 활용해 심리적 위험군을 조기 식별함으로써 자동화된 정신건강 지원 시스템 구축의 기반을 마련할 수 있다.

기존 인지왜곡 탐지 연구들은 주로 감정 어휘, 부정적 표현, 극단적 문장 패턴 등 언어적 표면 특성에 기반한 발화 분석에 집중해왔다[2]. 그러나 이러한 접근만으로는 인지왜곡이 형성되는 사고의 전개 과정과 구조적 특성을 충분히 반영하기 어렵다. LLM 의 논리적 추론 능력을 활용한 Diagnosis-of-Thought(DoT) 기법은 인지왜곡 탐지 성능을 향상시켰으나, 여전히 입력 발화 기반 추론에만 머물러 사고 구조 수준의 인지왜곡 특성은 충분히 포착하지 못했다[3].

이에 본 연구는 LLM 이 인지왜곡 사고의 구조적 특성을 스스로 추론하도록 설계한 Signature-of-Thought (SoT) 기법을 제안한다. SoT 는 LLM 이 다양한 인지왜곡의 사고 패턴과 단서를 추론하여 구조화한 시그니처(Signature)를 정의하고, 입력 발화의 사고 구조와의 유사도를 기반으로 인지왜곡 여부를 판단하도록 설계된 방식이다. 이 접근법은 기존 발화 중심 분석에 인지왜곡 사고 구조 분석 계층을 추가함으로써 LLM 의 추론 과정에 보다 체계적인 논리 근거를 제공한다.

인지왜곡 탐지를 위한 공개 벤치마크 데이터셋을 바탕으로 한 실험 결과, SoT 기법은 기존 선행 접근법 대비 F1-Score 기준 최대 9% 이상 상대적으로 향상된 인지왜곡 탐지 성능을 보였다. 이는 LLM 이 사고 구조 시그니처 개념을 통해 인지왜곡의 복합적인 구조 특성을 효과적으로 추론할 수 있음을 보여준다.

Ⅱ. 방법론

본 연구의 Signature-of-Thought(SoT) 기법은 LLM 이 인지왜곡을 구성하는 사고의 구조적 특성을 추론하여 시그니처(Signature)라는 개념으로 정의하고, 입력 발화의 사고 구조와 이 시그니처 간의 구조적 유사성을 비교하여 인지왜곡 여부를 판단할 수 있도록 설계되었다.

SoT 에서 정의되는 인지왜곡 사고 구조 시그니처는 다음 네 가지 요소로 구성된다. 첫째, 인지왜곡의 근본 전제를 이루는 핵심 신념(Core Belief), 둘째, 사실에서 결론으로 이르는 비합리적 사고 전개 방식(Reasoning Pattern), 셋째, 과도한 일반화나 극단적 표현 등 언어적 왜곡 단서(Linguistic Cue), 넷째, 맥락적 예외나 비유적 표현 등 인지왜곡으로 오인하지 않아야 할 경계 조건(Boundary Condition)이다. 구체적으로, 인지왜곡 유형 중 하나인 개인화(Personalization)에 대해 정의된 시그니처 예시는 표 1 과 같다.

SoT 기법은 이 네 요소를 추론하여 LLM 이 다양한 인지왜곡 사고의 전형적 패턴과 단서를 구조화하고, 입력 발화가 해당 시그니처와 얼마나 유사한지를 판단하도록 유도한다. 이를 통해 LLM 이 단순한 문장 의미 분석을 넘어, 사고 구조 수준의 분석을 수행하도록 한다.

Key	Value
Core Belief	Everything is my fault
Reasoning Pattern	Relating external events to oneself
Linguistic Cue	Because of me
Boundary Condition	Expressing empathy without self-blame

표 1. 인지왜곡 사고 구조 시그니처 예시

Ⅲ. 실험

본 연구에서는 제안한 Signature-of-Thought(SoT) 기법 기반 LLM 의 인지왜곡 탐지 성능을 검증하기 위해 기존 접근법들과의 비교 실험을 수행하였다. 실험에는 Kaggle 에 공개된 Cognitive Distortion detection dataset [1] 을 사용하였다. 해당 데이터셋은 Therapist Q&A[2] 데이터셋을 기반으로 임상 전문가가 인지왜곡 여부를 주석한 자료이며, 총 2,530 개의 데이터로 구성되어 있다. 전체 데이터 중 약 63%는 인지왜곡 문장, 37%는 비왜곡 문장으로 라벨링되어 있다.

모든 실험은 GPT-4.1 기반의 LLM 환경에서 진행되었으며, 비교 대상은 네 가지 인지왜곡 탐지 기법으로 설정하였다. 첫째, 별도의 지침 없이 LLM 이 직접 추론하도록 하는 Zero-Shot(ZS) 기법, 둘째, 단계적 사고 전개를 유도하는 Chain-of-Thought(CoT) 기법[4], 셋째, 기존 연구에서 제안된 Diagnosis-of-Thought(DoT) 기법[3], 넷째, 본 연구에서 새롭게 제안하는 Signature-of-Thought(SoT) 기법이다.

LLM 의 인지왜곡 탐지 성능 평가는 Accuracy, Macro F1, Weighted F1 세 가지 지표를 활용하여 이루어졌다. Accuracy 는 전체 예측의 일치율을 의미하고, Macro F1 은 클래스 간 불균형을 보정한 평균 F1-score 를 나타내며, Weighted F1 은 실제 데이터 분포를 반영한 가중 평균으로 LLM 의 전반적인 성능을 현실적으로 평가하는 지표이다.

Method	Accuracy	Macro F1	Weighted F1
ZS	0.7209	0.6342	0.6810
CoT	0.7273	0.6484	0.6921
DoT	0.7182	0.7031	0.7206
SoT	**0.7470**	**0.7211**	**0.7434**

표 2. 인지왜곡 탐지 성능 실험 결과

실험 결과, SoT 기법은 모든 평가 지표에서 가장 우수한 성능을 보였으며, 실험 결과 내용은 표 2 와 같다. Weighted F1 기준으로 ZS 기법은 0.6810, CoT 기법은 0.6921, DoT 기법은 0.7206 을 기록한 반면 SoT 기법은 0.7434 로 가장 높은 성능을 달성하였다. 이는 ZS 기법 대비 약 9.2%, CoT 기법 대비 7.4%, DoT 기법 대비 3.2%의 상대적 성능 향상에 해당하며, Accuracy 와 Macro F1 지표에서도 SoT 기법이 각각 0.7470, 0.7211 로 모두 최고 수치를 기록하였다. 이러한 결과는 SoT 방식이 인지왜곡 사고의 구조적 패턴과 단서를 추론하여 통합한 시그니처 개념을 기반으로 입력 발화를 해석함으로써, 결과적으로 LLM 의 인지왜곡 탐지 능력을 보다 정밀하고 일관되게 향상시킬 수 있음을 입증한다.

Ⅳ. 결론

본 연구는 LLM 기반 인지왜곡 탐지를 사고 구조 분석의 관점에서 접근한 SoT 기법을 제안하였다. 이 방법론은 인지왜곡 사고의 전형적 패턴과 단서를 시그니처 개념으로 구조화하고, 입력 발화의 사고 구조와의 유사성을 기반으로 인지왜곡 여부를 판단함으로써, 실험 결과 기존 선행 접근법 대비 F1-score 기준 최대 9% 이상 상대적 성능 향상을 달성하였다. 이는 단순한 입력 발화 분석을 넘어 인지왜곡 사고의 구조적 특성을 반영하는 접근이 인지왜곡 탐지 정확도를 높일 수 있음을 보여준다. 즉, 개인의 인지왜곡이 어떤 패턴으로 전개되며 어떤 단서들이 나타나는지를 LLM 이 구조적으로 추론할 때 인지왜곡의 복합적 특성을 더 정확히 식별할 수 있음을 시사한다. 향후 연구에서는 실제 임상 상담 데이터를 활용하여 SoT 기법의 일반화 성능과 실질적 적용 가능성을 검증할 예정이다. 또한 세부 인지왜곡 유형별 분류로 확장하여, 사고 시그니처 기반 인지왜곡 탐지 기술의 응용 범위를 넓히는 것을 목표로 한다.

ACKNOWLEDGMENT

이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행되었으며 (No.RS-2025-02217960, 인간의 성장 발달과정을 모사한 인간지향적 AI 모델 개발), 과학기술정보통신부 및 정보통신기획평가원의 디지털분야해외석학유치지원 연구결과로 수행되었음 (RS-2024-00459638).

참 고 문 헌

[1] Beck, Aaron T. Cognitive therapy and the emotional disorders. Penguin, 1979.

[2] Shickel, Benjamin, et al. "Automatic detection and classification of cognitive distortions in mental health text." 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2020.

[3] Chen, Zhiyu, Yujie Lu, and William Yang Wang. "Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting." arXiv preprint arXiv:2310.07146 (2023).

[4] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in neural information processing systems 35 (2022): 24824-24837.

1 https://www.kaggle.com/datasets/sagarikashreevastava/cognitive-distortion-detetction-dataset

2 https://www.kaggle.com/datasets/arnmaud/therapist-qa