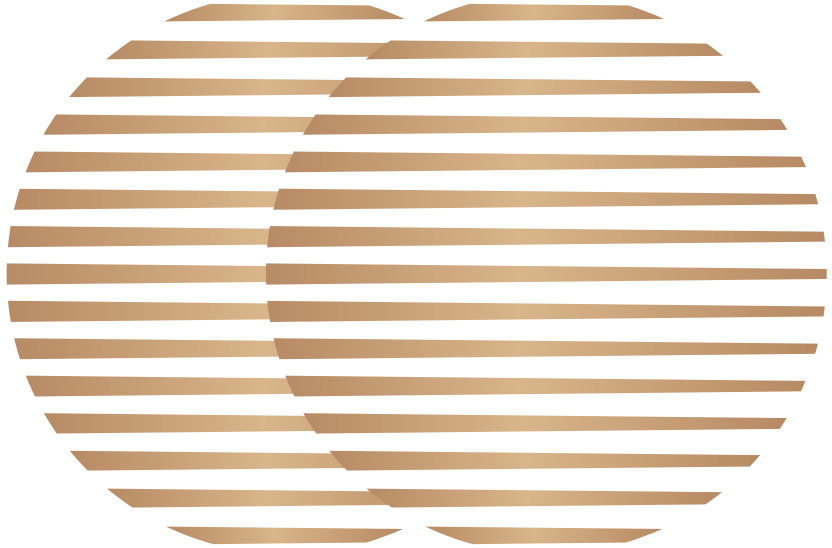
2024년 3월 28일 자료

Logical Fallacy Paper

Logical Fallacy Detection

정지원
성균관대학교 인공지능학과
석사과정
jwjw9603@g.skku.edu



성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)

Paper Follow – UP(Logical Fallacy)

ACL 2021

2021.08

Breaking Down the invisible wall of informal fallacies in online discussions

- Reddit 소셜미디어 데이터셋

EMNLP 2021

2022.02

Logical Fallacy Detection

- 13가지의 타입을 가지고 있는 논리 오류 데이터셋(LOGIC)
- 논리 오류의 구조적 형태를 학습하고자 함

Multitask Instruction-based Prompting for Fallacy Recognition

EMNLP 2022

Data	Ex	F	Genre	Domain
ARGOTARIO	880	5	Dialogue	General
PROPAGANDA	5.1k	15†	News	Politics
LOGIC	4.5k	13	Diverse	Education
COVID-19	621	9‡	SocMed/News	Covid-19
CLIMATE	477	9‡	News	Climate

2022.12

IJCAL 2023

2023.01

Case-based reasoning with Language Models for classification of Logical fallacies

- Case base reasoning기반으로 LOGIC 데이터셋 학습 및 평가
- 논리 오류를 발생시키는 문장에 counterarguments, goals, explanations, structure와 같은 추가적인 정보를 주고 CBR알고리즘 + attention 매커니즘 적용

Argument-based Detection and Classification of Fallacies in Political Debates

- 정치 연설문 데이터셋 생성
- Detection을 진행하기 위해 fallacy가 발생하는 시작, 끝 부분을 찾기 위해 Pos-tagging 방법 사용

2023.08

EMNLP 2023

NAACL 2024

2023.11

A closer Look at the Self-Verification abilities of Large Language Models in Logical Reasoning Paper

- LLM의 logical reasoning을 제대로 이해하는지 파악하기 위해 최신 LLM에 대해 성능 평가를 진행하였고, FALLACIES라는 데이터셋 생성
- Formal : 24개 타입, Informal : 208개 타입

Large Language Models are Few-Shot Training Example Generators : A case Study in Fallacy Recognition

- Multitask instruction-based prompting for fallacy recognition 논문의 저자가 낸 논문
- 클래스별 불균형을 줄이기 위해 prompting을 통해 data augmentation진행

2023.11

Preprint 2023

2023.12

Detecting argumentative fallacies in the Wild : Problems and Limitations of Large Language Models

- 기존의 논리 오류 관련 논문들은 각자 데이터셋을 만들고 LM을 평가하는 목적으로 나옴.
- 이러한 점은 일반적으로 실제 상황에서 적용 가능성과 유용성을 이해하는데 많은 한계점을 보임.
- 그래서, natural language argumentation schemes로 이루어진 검증 코퍼스를 만듦

데이터셋 출처와 LLM을 대상으로 fallacy dataset 평가

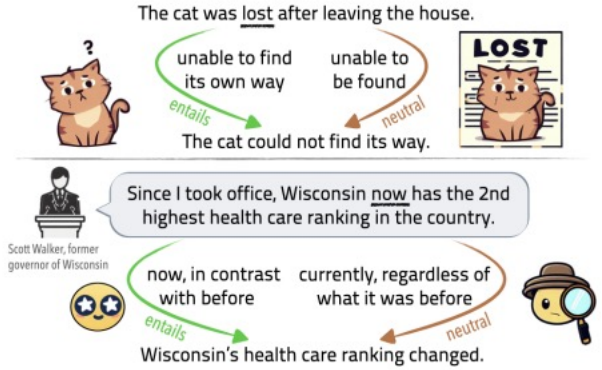
Paper Follow – UP(LLM Evaluation)

EMNLP 2023

2023.04

We' re Afraid Langauge Models aren't Modeling Ambiguity

- 중의성(애매모호함)을 가지는 텍스트들을 모아 데이터셋을 만듦
- 전제와 가설, 두 애매모호한 문장을 보여주고 이 문장 간의 관계를 확인한다.



2023.08

EMNLP 2023

This is not a Dataset : A Large Negation Benchmark to Challenge Large Language Models

- 부정어(negation)을 가지는 텍스트들을 모아 데이터셋을 만듦.
- 데이터는 wornet에서 11가지의 relation을 정하고, 그 relation이 들어가는 triple을 추출함.
- 추출한 triple을 기반으로 template(prompt)를 만들고 template에 맞춰 데이터를 생성함.
  - Triple : 〈part, bill, bird〉
  - Template : 〈noun1+(e)s〉[are commonly | may be] part of 〈noun2 +(e)s〉.
- 만들어진 데이터셋은 두 명의 native speakers들이 데이터셋으로부터 220개의 문장을 랜덤 샘플링을 진행해서 평가함.

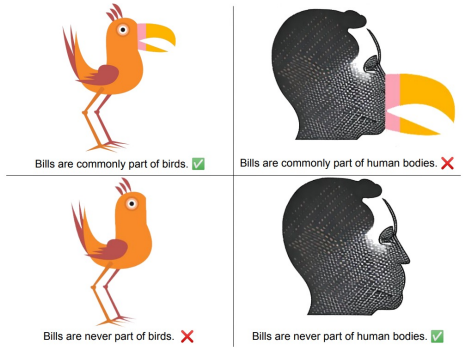


Figure 1: Affirmative and negative sentences in the dataset.

Paper Follow – UP(LLM with KGs)

Unifying Large Language Models and Knowledge Graphs : A Roadmap

- LLM과 KGs의 통합을 위한 전망적인 로드맵 제시

MindMap : Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models

- KG를 사용해서LLM을 최신 지식과 연결하고 LLMs의 추론 경로를 유도하기 위한 방법을 탐구함
- Evidence graph mining -> Evidence graph aggregation -> LLM reasoning on the mindmap
- Entity linking 방법은 얻어갈만한 정보

Knowledge-Driven CoT : Exploring faithful reasoning in LLMs for knowledge-intensive Question Answering

- CoT Collection을 미리 만들고 이를 기반으로 Retrieve-reader-verifier 모듈을 거침

ICLR 2024 — 2023.06

Boosting Language Models Reasoning with Chain-of Knowledge Prompting

- Evidence triple, explanation hints 사용
- F2-verificiation

2023.06 — IEEE 2023

EMNLP 2023 — 2023.08

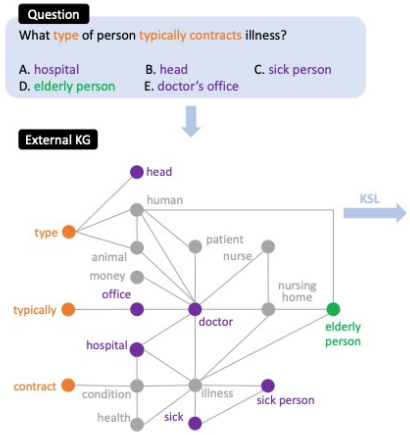
Reasoning on Graphs : Faithful and interpretable large language model reasoning

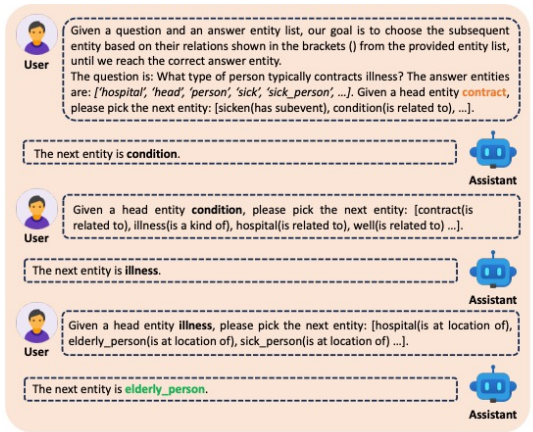
- Plan-and-solve를 차용한 planning-retrieval-reasoning framework 제시
- ELBO를 사용하여 수식적으로 설명

2023.08 — Preprint 2023

Preprint 2023 — 2023.08

Knowledge Solver : Teaching LLMs to search for domain Knowledge from Knowledge graphs





Preprint 2023 — 2023.08

-> LLM이 entity를 Step by Step방법으로 생각하도록 구성함

아이디어 및 방법론

Reference

1. SAHAI, Saumya; BALALAU, Oana; HORINCAR, Roxana. Breaking down the invisible wall of informal fallacies in online discussions. 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)*. 2021. p. 644-657.
2. ALHINDI, Tariq, et al. Multitask Instruction-based Prompting for Fallacy Recognition. *arXiv preprint arXiv:2301.09992*, 2023.
3. JIN, Zhijing, et al. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*, 2022.
4. GOFFREDO, Pierpaolo, et al. Argument-based Detection and Classification of Fallacies in Political Debates. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. p. 11101–11112.
5. SOURATI, Zhivar, et al. Case-based reasoning with language models for classification of logical fallacies. *arXiv preprint arXiv:2301.11879*, 2023.
6. HONG, Ruixin, et al. A Closer Look at the Self-Verification Abilities of Large Language Models in Logical Reasoning. *arXiv preprint arXiv:2311.07954*, 2023.
7. PAN, Shirui, et al. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
8. WANG, Jianing, et al. Boosting Language Models Reasoning with Chain-of-Knowledge Prompting. *arXiv preprint arXiv:2306.06427*, 2023.

Reference

9. ALHINDI, Tariq; MURESAN, Smaranda; NAKOV, Preslav. Large Language Models are Few-Shot Training Example Generators: A Case Study in Fallacy Recognition. *arXiv preprint arXiv:2311.09552*, 2023.

10. LIU, Alisa, et al. We're Afraid Language Models Aren't Modeling Ambiguity. *arXiv preprint arXiv:2304.14399*, 2023.

11. GARCÍA-FERRERO, Iker, et al. This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models. *arXiv preprint arXiv:2310.15941*, 2023.

12. LUO, Linhao, et al. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*, 2023.

13. WEN, Yilin; WANG, Zifeng; SUN, Jimeng. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*, 2023.

14. FENG, Chao; ZHANG, Xinyu; FEI, Zichu. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*, 2023.

15. WANG, Keheng, et al. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*, 2023.

Reference

16. CRESWELL, Antonia; SHANAHAN, Murray; HIGGINS, Irina. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.

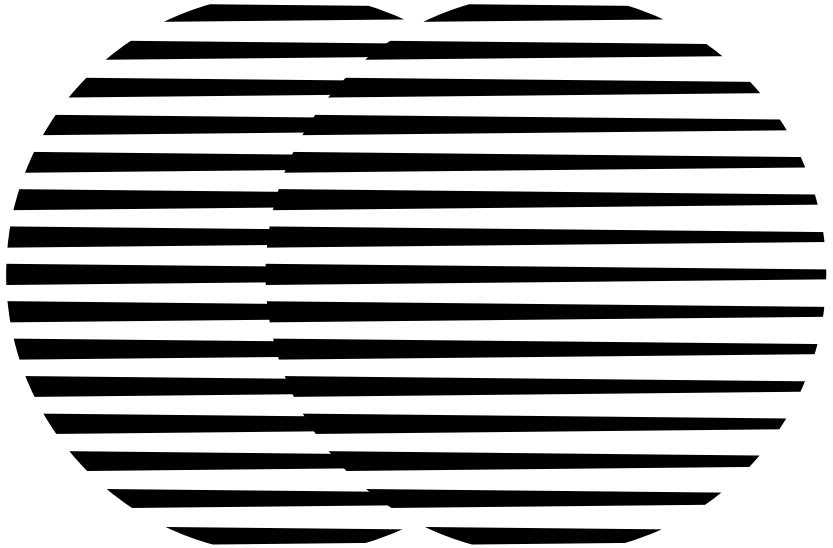
17. KRISHNA, Satyapriya, et al. Post hoc explanations of language models can improve language models. *Advances in Neural Information Processing Systems*, 2024, 36.

18. ZHAO, Ruochen, et al. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *arXiv preprint arXiv:2305.03268*, 2023.

19. XU, Xiaohan, et al. Re-reading improves reasoning in language models. *arXiv preprint arXiv:2309.06275*, 2023.

감사합니다

발표 경청해 주셔서 감사합니다



정지원 성균관대학교 인공지능학과 석사 과정
jwjw9603@g.skku.edu

성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)