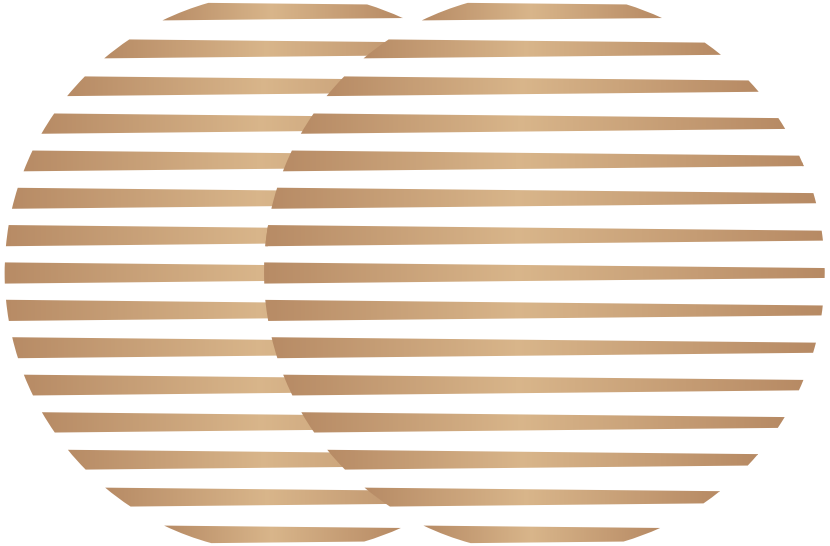
2024년 3월 20일 Study Meeting

논리 오류 감지

Logical Fallacy detection

정지원
성균관대학교 인공지능학과
석사과정
jwjw9603@g.skku.edu



성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)

진행 내용 Overview

- 실험결과
- 방법론

Result(LOGIC)

Type/Metric	Total Accuracy	Precision	Recall	F1	FG acc	FC acc	IA acc
No Question	0.64	0.52	0.54	0.50	0.68(±0.01)	0.71(±0.01)	0.91
General Question	0.73(±0.01)	0.56(±0.01)	0.57(±0.01)	0.55(±0.01)	0.77(±0.01)	0.78(±0.01)	0.92(±0.01)
Classwise Question - zero	0.78(±0.01)	0.60	0.63	0.60	0.79(±0.01)	0.83(±0.01)	0.95
Classwise Question - one	0.78	0.59	0.62	0.59	0.79	0.83	0.94
Classwise Question - two	0.76(±0.01)	0.58	0.60	0.57(±0.01)	0.78(±0.01)	0.81(±0.01)	0.93(±0.01)
Classwise Question - five	0.72(±0.01)	0.56	0.58(±0.01)	0.55	0.74(±0.01)	0.78(±0.01)	0.93

Result(COVID-19)

Type/Metric	Total Accuracy	Precision	Recall	F1	FG acc	FC acc	IA acc
No Question	0.61	0.41(±0.05)	0.40(±0.08)	0.37(±0.07)	0.83(±0.01)	0.64(±0.02)	0.77(±0.03)
General Question	0.68(±0.02)	0.70(±0.16)	0.57(±0.03)	0.56(±0.06)	0.86(±0.02)	0.70(±0.03)	0.80
Classwise Question - zero	0.87(±0.02)	0.84	0.75(±0.16)	0.72(±0.11)	0.89(±0.01)	0.91(±0.02)	0.95(±0.01)
Classwise Question - one	0.86(±0.01)	0.81(±0.14)	0.68(±0.08)	0.69(±0.09)	0.88(±0.01)	0.89(±0.03)	0.95(±0.01)
Classwise Question - two	0.84(±0.02)	0.83(±0.10)	0.66(±0.10)	0.66(±0.11)	0.88(±0.02)	0.88(±0.02)	0.94(±0.02)
Classwise Question - five	0.78(±0.05)	0.72(±0.24)	0.60(±0.13)	0.59(±0.15)	0.86(±0.02)	0.80(±0.06)	0.90(±0.04)

Result(CLIMATE)

Type/Metric	Total Accuracy	Precision	Recall	F1	FG acc	FC acc	IA acc
No Question	0.63(±0.03)	0.42(±0.05)	0.33(±0.03)	0.37(±0.04)	0.89(±0.01)	0.73(±0.03)	0.80(±0.02)
General Question	0.68(±0.01)	0.37(±0.01)	0.36(±0.01)	0.36(±0.01)	0.88(±0.01)	0.71(±0.03)	0.79(±0.02)
Classwise Question - zero	0.83(±0.01)	0.56(±0.06)	0.54(±0.07)	0.54(±0.07)	0.89(±0.01)	0.88(±0.02)	0.92(±0.02)
Classwise Question - one	0.78(±0.01)	0.53(±0.03)	0.49(±0.02)	0.50(±0.02)	0.81(±0.01)	0.81(±0.02)	0.88(±0.01)
Classwise Question - two	0.81(±0.01)	0.52(±0.03)	0.48(±0.01)	0.49(±0.01)	0.91(±0.01)	0.83(±0.02)	0.90(±0.02)
Classwise Question - five	0.71(±0.01)	0.47(±0.09)	0.46(±0.08)	0.46(±0.08)	0.87(±0.01)	0.73(±0.01)	0.82(±0.01)

Result(Argotario)

Type/Metric	Total Accuracy	Precision	Recall	F1	FG acc	IA acc
No Question	0.65(±0.01)	0.59(±0.11)	0.58(±0.11)	0.57(±0.11)	0.65(±0.01)	0.65(±0.021)
General Question	0.72(±0.01)	0.74(±0.01)	0.72(±0.01)	0.71	0.72(±0.01)	0.72(±0.01)
Classwise Question - zero	0.76(±0.01)	0.83(±0.01)	0.76(±0.01)	0.75(±0.01)	0.76(±0.01)	0.76(±0.01)
Classwise Question - one	0.72(±0.01)	0.79(±0.01)	0.72(±0.01)	0.71(±0.01)	0.72(±0.01)	0.72(±0.01)
Classwise Question - two	0.80(±0.02)	0.84(±0.02)	0.80(±0.02)	0.79(±0.02)	0.80(±0.02)	0.80(±0.02)
Classwise Question - five	0.74(±0.01)	0.74(±0.13)	0.58(±0.13)	0.57(±0.13)	0.74(±0.01)	0.75(±0.01)

Result(Total)

Type/Metric	Total Accuracy	Precision	Recall	F1	FG acc	FC acc	IA acc
No Question	0.64(±0.01)	0.51(±0.01)	0.52	0.49	0.71	0.73(±0.01)	0.85
General Question	0.70	0.55	0.56	0.53	0.75	0.77	0.87
Classwise Question - zero	0.75	0.65(±0.09)	0.68(±0.09)	0.64(±0.09)	0.77	0.82(±0.01)	0.92(±0.01)
Classwise Question - one	0.74	0.50	0.60	0.56	0.76	0.82	0.91
Classwise Question - two	0.74(±0.01)	0.58	0.60	0.58(±0.02)	0.76(±0.01)	0.80	0.92
Classwise Question - five	0.69	0.55	0.67	0.53	0.72	0.76	0.90

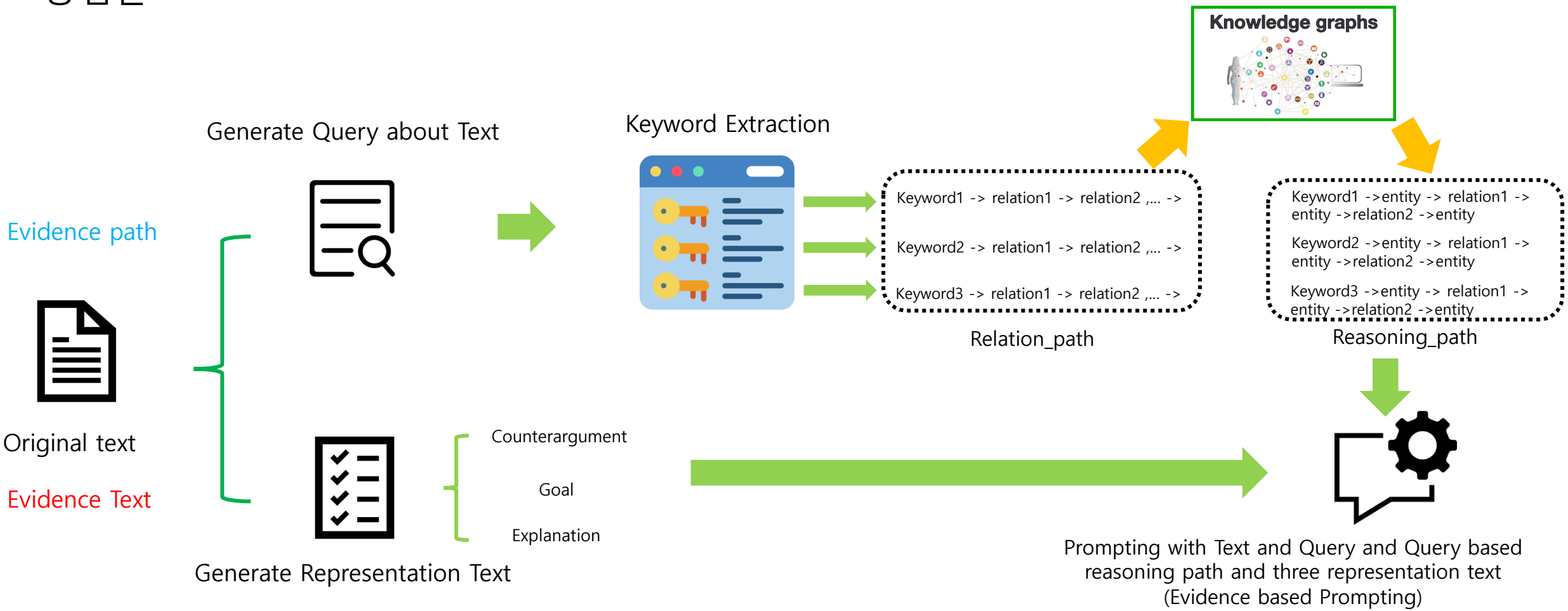
방법론

1. 논리 오류를 가지고 있는 **text로부터 Query를 생성한다**.
    1) LLM을 사용해서 **Query로부터 Keyword를 추출**한다.(Evidence Path or Triple) -〉 구현 완료
    2) RoG 방법론 -〉 **구현 완료**
        1) LLM에게 keyword를 헤드 엔티티로 하는 relation_path를 생성하라 요청한다.
        2) 생성된 relation_path가 KG에 있는지 확인한다.
        3) 있는 것들만 선택한 후 reasoning path로 변환한다.
2. Text로부터 Counterargument, goal, explanation text를 생성한다(Evidence Text).
3. 1-1과 1-2내용의 정보를 주고 논리 오류를 감지 및 분류를 진행한다.
    1) 감지 및 분류를 진행할 때, Prompt에 어떤 것이 유용했는지를 물어보는 질문을 함께한다(Explainability)

방법론

1. LLM을 사용해서 Query로부터 Keyword를 추출해서 진행하는 Rog 방법론 외에 또 다른 방법(Evidence Text)을 진행했지?
    1) 원래 RoG 방법론만을 사용하려 했으나, RoG를 통해 나오는 최종적인 reasoning path가 짧다.
        ① 왜냐하면 conceptnet의 relation(17개)은 다양하지 않기 때문에 LLM을 통해 생성한 것이 실제로 conceptnet에 있기 드물다.
        ② 최종적으로 Query로부터 생성되는 reasoning_path가 없는 Query가 있다.
    2) Query는 논리 오류를 가진 질문에 되묻는, 관계를 묻는 형태의 질문이다.
        ① 묻는 것에 넘어서 추론 과정의 path를 보여주고자 RoG를 사용한다.
        ② 하지만 Reasoning path는 Query의 키워드로부터 나온 내용이다. 이것만으로는 텍스트의 속뜻을 알기에 부족하다.
        ③ 최종적으로 질문을 할 때, Text, Query, representation text(Evidence Meaning), Reasoning_path 를 주고 분류를 진행하고자 한다.
2. 1-1번과 1-2번의 정보를 기반으로 논리 오류를 예측하는데, 이 정보들을 선별해서 추론하는 과정이 필요하지 않아?
    1) RoG 방법론에서 KG를 통해 필터링을 거친다.
    2) Evidence Meaning은 필터링이 필요 없다. -> 왜냐하면 이것은 LLM을 사용해서 text로부터 얻는 정보이므로, LLM에서 생성하는 것 자체가 내부적으로 필터링을 진행한 것이지 않을까?
    3) 이미 자체적으로 필터링을 거친 정보이므로 다 필요하지 않을까?

방법론



Zero-shot

Prompt(Method)

User

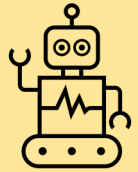
Your task is to detect a fallacy in the Text.
The label can be 'Faulty Generalization' and 'False Causality' and 'Irrelevant Authority'.
Please detect a fallacy in the text based on the Query.

Text : It is warmer this year in Las Vegas as compared to last year; therefore, global warming is rapidly accelerating.
Question : Does the fact that Las Vegas is warmer this year compared to last year necessarily imply that global warming is rapidly accelerating?

Evidence Paths :
Evidence Texts :

Label :

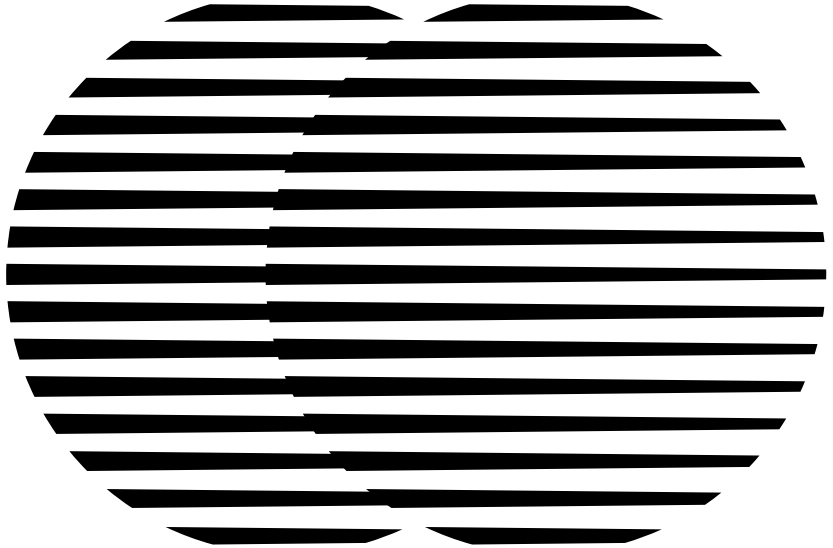
Faulty Generalization



Assistant

감사합니다

발표 경청해 주셔서 감사합니다



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

성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)