2024년 3월 15일 Study Meeting

실험 결과

Experiment Result

정지원
성균관대학교 인공지능학과
석사과정
jwjw9603@g.skku.edu

성균관대학교
1398
SUNG KYUN KWAN UNIVERSITY(SKKU)

진행 내용 Overview

- No Fallacy dataset
- Query(Question)을 만드는 과정에 대한 설명
- 실험

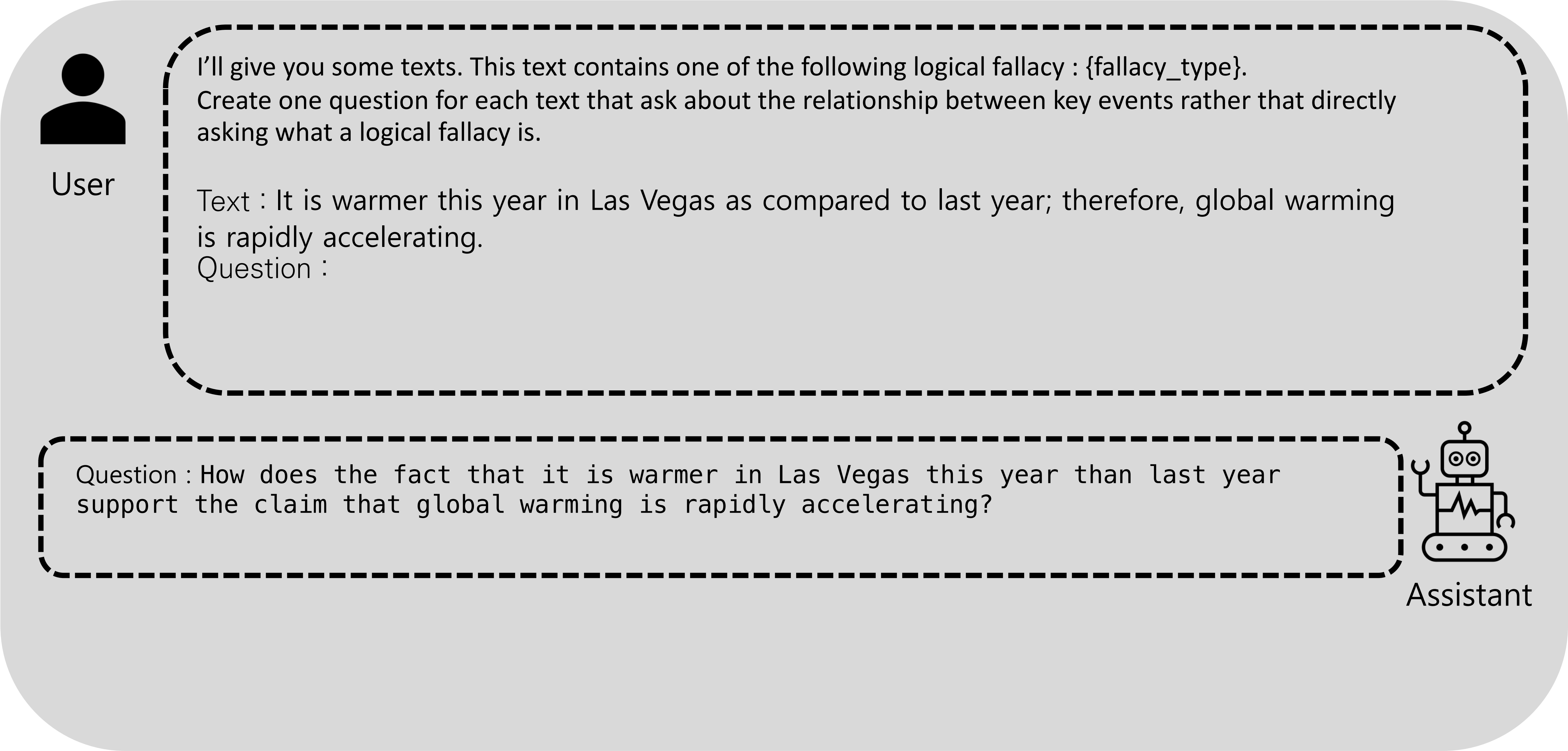
No Fallacy dataset

1. 우리는 Argotario, LOGIC, CLIMATE, COVID-19 4개의 데이터셋을 다룸
2. LOGIC 데이터셋을 제외하고 "No Fallacy" 클래스에 해당하는 데이터가 있음. Argotario(419개), CLIMATE(206개), COVID-19(61개)
3. Faulty Generalization, False Causality, Irrelevant Authority 클래스는 공통점이 있기 때문에 general한 질문을 만들 수 있었음.
   - Create one question for each text that ask about the relationship between key events within the text rather than directly asking what a logical fallacy is.
4. 하지만, "No Fallacy" 클래스는 어떻게??
   - 위 클래스의 prompt와 동일하게 하면 성능이 오히려 떨어짐
   - No fallacy는 관계에서 발생하는 문제가 아니기 때문에, 다르게 해야 할 듯
   - 차라리 No fallacy를 빼고 진행한다면?
5. No fallacy 데이터는 위 세 개의 클래스로 finetuned 된 LLM에 testing 용도로 쓰이면 어떨까? -> finetuning은 나중에,,,

Query를 만드는 과정

1. 우리는 Argotario, LOGIC, CLIMATE, COVID-19 4개의 데이터셋을 다룸
2. 각 데이터셋의 개수는 다음과 같다 :
   - Argotario : Faulty generalization(157), Irrelevant authority(156)
   - LOGIC : Faulty generalization(441), False Causality(216), Irrelevant authority(132)
   - CLIMATE : Faulty generalization(7), False Causality(19), Irrelevant authority(18)
   - COVID-19 : Faulty generalization(9), False Causality(60), Irrelevant authority(24)
3. 각 데이터셋은 학습, 검증, 테스트 데이터 모두 공개가 되어있으며, 정답(레이블)도 공개 되어있다.
   1) LOGIC데이터셋을 제외하고는, 학습, 검증, 테스트로 나누어져 있지 않음
   2) LLM을 학습하는 것이 아니며, 데이터셋을 테스트 하는 것이다. 그렇기 때문에 test 데이터만 가지고 진행한다면 너무 개수가 적다. 따라서 LLM(ChatGPT)에 테스팅을 해보기 위해 모든 데이터셋을 사용했다.
4. 주어진 문장(original text)과 이 문장의 레이블을 통해 Query를 만든다.
   1) 전처리 과정에서 진행하며, 주어진 문장(original text)과 이 문장의 레이블에 맞게 Query를 만든다.
   2) 각 데이터셋을 클래스별로 나눈 다음, 레이블{i}에 따라 {i}prompt를 적용시켜 Query를 만든다.(Class specific Query)
      - i ∈ $\{Faulty\ generalization, False\ causality, Irrelevant\ authority\}$
      - General Query일경우에는 클래스별로 나누지 않는다.
   3) 이렇게 진행되면 각 문장(original text)당 **한 개의 Query**가 만들어진다.
   4) 이 과정은 테스팅 과정과 다른 모델(instruct)로 진행하며, 다 만들어지면 데이터셋이 준비가 된 것이다.
   5) 데이터셋에는 original text, label, query로 구성된다.

Prompt(General-Question)



I’ll give you some texts. This text contains one of the following logical fallacy : {fallacy_type}.
Create one question for each text that ask about the relationship between key events rather that directly asking what a logical fallacy is.

Text : It is warmer this year in Las Vegas as compared to last year; therefore, global warming is rapidly accelerating.
Question :

Question : How does the fact that it is warmer in Las Vegas this year than last year support the claim that global warming is rapidly accelerating?

Prompt(Classwise Question– Faulty Generalization)

User

A faulty generalization often follows the following format:

The proportion Q of the sample has attribute A. Therefore, the proportion Q of the population has attribute A.

Extracting Q and A from each text allows us to determine the scope of the claims made by Q and A.
When A > Q, it refers to cases where a minority of people make generalized claims based on their experiences, hearsay, or observations.
In such texts, it's appropriate to formulate questions that inquire about or challenge the claim A, or about the relationship between A and Q.
When A < Q, it refers to cases where questions are formulated to inquire about the content of Q.
I'll provide you with some texts, along with their Q and A, comparisons of Q and A, and the Q and A extracted from the texts.
These texts contain the logical fallacy of Faulty Generalization.
Considering the above, Please create one question for each text.

Here are some examples:

- text : "Annie must like Starbucks because all white girls like Starbucks." - Q : All white girls like starbucks. - A : Annie must like starbucks. - A < Q - question : Do all white girls like Starbucks?
- Text : It is warmer this year in Las Vegas as compared to last year; therefore, global warming is rapidly accelerating.

Question : Does the temperature change in Las Vegas necessarily indicate a global trend?(zero-shot)

Assistant

Prompt(Classwise Question– False Causality)

User

A false causality format often follows the following format:
A occurred
B occurred
Therefore, A caused B
I'll give you some Texts. These texts contain the logical fallacy of False Causality.
Create one question for each text that asks whether the relationship between A and B is correct.
It would be good to refer to the text format.

Here are some examples:

- text : "Every time I go to sleep the sun goes down. Therefore, my sleeping causes the sun to set."
- - question : Does the fact that the sun goes down every time you go to sleep necessarily mean that your sleeping causes the sun to set?

– Text : Children who play violent video games act more violently than those who don't.
– Question :

Question : Does playing violent video games directly cause children to act more violently?
(zero-shot)

Assistant

Prompt(Classwise Question– Irrelevant Authority)

User

A irrelevant authority format often follows the following format:

Person A claims that X is true.
Person A is an expert in the field concerning X.
Therefore, X should be believed

I'll give you some texts. This text contains the logical fallacy of irrelevant authority.
Create one question for each text to verify the credibility of Person A claiming X.
It would be good to refer to the text format.

Here are some examples :

- text : You ask your mother if you can go to the mall with your friends. She says "no". You ask why? She says, "because I'm the mom and I say so".
- Question : How does the fact that someone is a mother justify their decision to not allow their child to go to the mall with friends?
- Text : Lebron James, one of the most decorated basketball players of all time, says you need to eat breakfast so you need to eat breakfast

Question : Can we trust Lebron James' claim that breakfast is necessary based solely on his expertise in basketball?(zero-shot)

Assistant

Prompt(Method-Question)

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 Question.

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?
Label :

Faulty Generalization ○

Assistant

Prompt(Method-No Question)

User

Your task is to detect a fallacy in the Text. The label can be “Faulty Generalization” and 'False Causality' and 'Irrelevant Authority'.
Text : It is warmer this year in Las Vegas as compared to last year; therefore, global warming is rapidly accelerating.
Label :

The Other Fallacy ✗

Assistant

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)

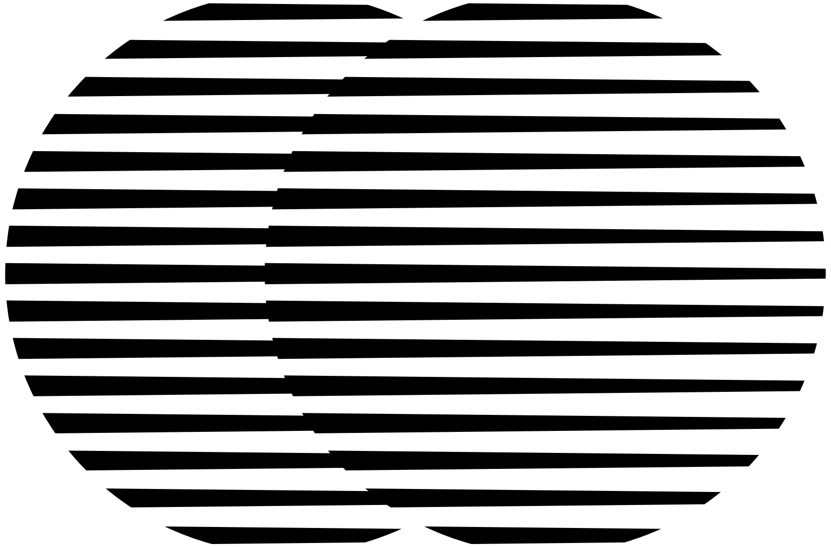
Type/Metric	Total Accuracy	Precision	Recall	F1	FG acc	FC acc	IA acc
No Question	0.64(±0.01)	0.51(±0.01)	0.52	0.49	0.71	0.73(±0.01)	0.85
General Question	0.70	0.55	0.56	0.53	0.75	0.77	0.87
Classwise Question - zero	0.75	0.65(±0.09)	0.68(±0.09)	0.64(±0.09)	0.77	0.82(±0.01)	0.92(±0.01)
Classwise Question - 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

지식 그래프

1. Query는 텍스트당 한 개의 Query가 나온다.
   1) Query로부터 생성되는 Reasoning path가 어떤 fallacy에 적합한지?
   2) Query로부터 생성되는 Reasoning path로 fallacy detection
2. 더 자세한 건 다음 미팅 전, 공유 필요.. (아이디어 확인)
3. 지식 그래프를 사용하는 아이디어와 진행상황은 추후에 공유하도록 하겠습니다.

감사합니다

발표 경청해 주셔서 감사합니다



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

1398
성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)