




동계세미나 (2/6, Thu)

Hallucinations of Negation Knowledge

구선민 

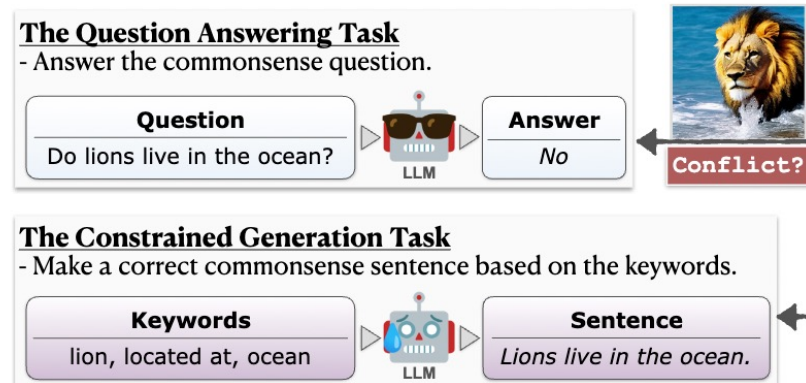


고려대학교
KOREA UNIVERSITY

Background

* Negation (negative) Knowledge

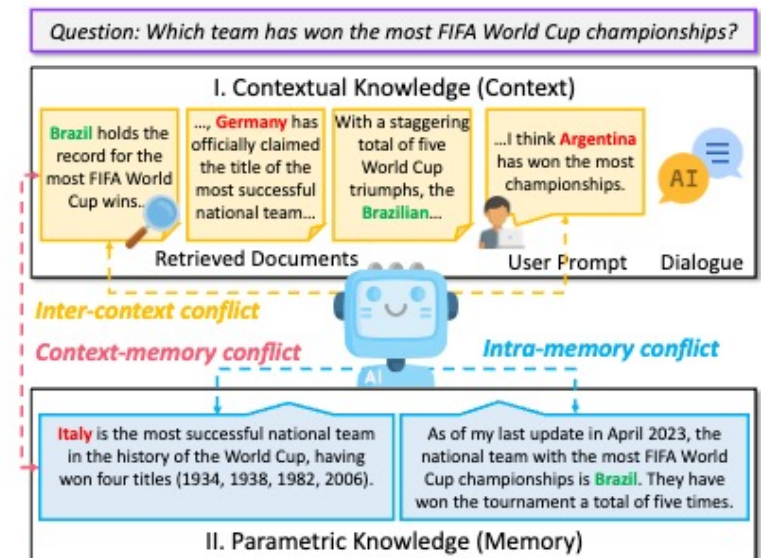
- 모델 학습 데이터는 대부분 positive form (“사자는 바다에 산다”) → 따라서 모델이 학습 시에 부정형 지식 (“사자는 바다에 ‘안’ 산다”) 을 잘 보지 못함
- Say What You Mean! Large Language Models Speak Too Positively about Negative Commonsense Knowledge (ACL 2023) 논문에서 *LLMs’ belief conflict (shortcut)* 로 풀어냄 → LLMs’ belief about negative knowledge 검증해보겠다.. 하면서 LLMs’ belief conflict phenomenon on negative commonsense knowledge



Background

* Knowledge Conflicts 와 Shortcut 의 차이

- LLMs 이 입력에 대해서 hallucination 일으켰는데 원인을..
- Knowledge Conflict?
: 모델의 내부 지식 (parametric knowledge) 와 제공된 외부 지식 (non-parametric knowledge) 이 달라서 hallucination 발생
→ Knowledge Editing 쪽 연구들
- Knowledge Shortcut?
: 모델이 입력을 제대로 파악하지 않고 학습 데이터에서 많이 본 결과를 바로 출력해서 hallucination 발생
→ Negative knowledge 검증 연구들



(↑) Knowledge Conflict

Investigating and Addressing Hallucinations of LLMs in Tasks Involving Negation

Neeraj Varshney Satyam Raj Venkatesh Mishra Agneet Chatterjee
Ritika Sarkar Amir Saeidi Chitta Baral
Arizona State University

Motivation

* Negation Knowledge 의 중요성

- Negation 은 adds depth and nuance to the understanding of language 때문에 중요
. negation 은 opposite or absence of a statement 을 이해하는 데 도움이 되어 더 정확하고 미묘하게 해석할 수 있으며 논리적 추론에 필수 요소
- Negation knowledge 는 현실 세계에 존재하며, 모델의 cognitive skills 에 중요함
. 모델이 무엇이 사실이 아닌지, 무엇을 생각하지 말아야 할지 아는 능력
- Negative knowledge 는 모델 학습 데이터에 positive knowledge 에 비해 적게 포함되어 있고 대부분의 벤치마크 데이터셋 positive 형태이므로 때문에 주목할 필요 0

Motivation

* LLMs 기반 negation 연구 부족

- 부정에 대한 이전 연구에서는 주로 classification tasks 연구함
. natural language inference, masked word prediction ...
- 근데 이런 기존 대부분 연구들 LLMs 으로 수행되지 않았음
- generative tasks with LLMs 로 negation 연구 수행해보겠다
- To this end, we study negation in four tasks:
 - (i) False Premise Completion (FPC),
 - (ii) Constrained Fact Generation (CFG)
 - (iii) Multiple-Choice Question Answering (MCQA)
 - (iv) Fact Generation (FG)

Overview

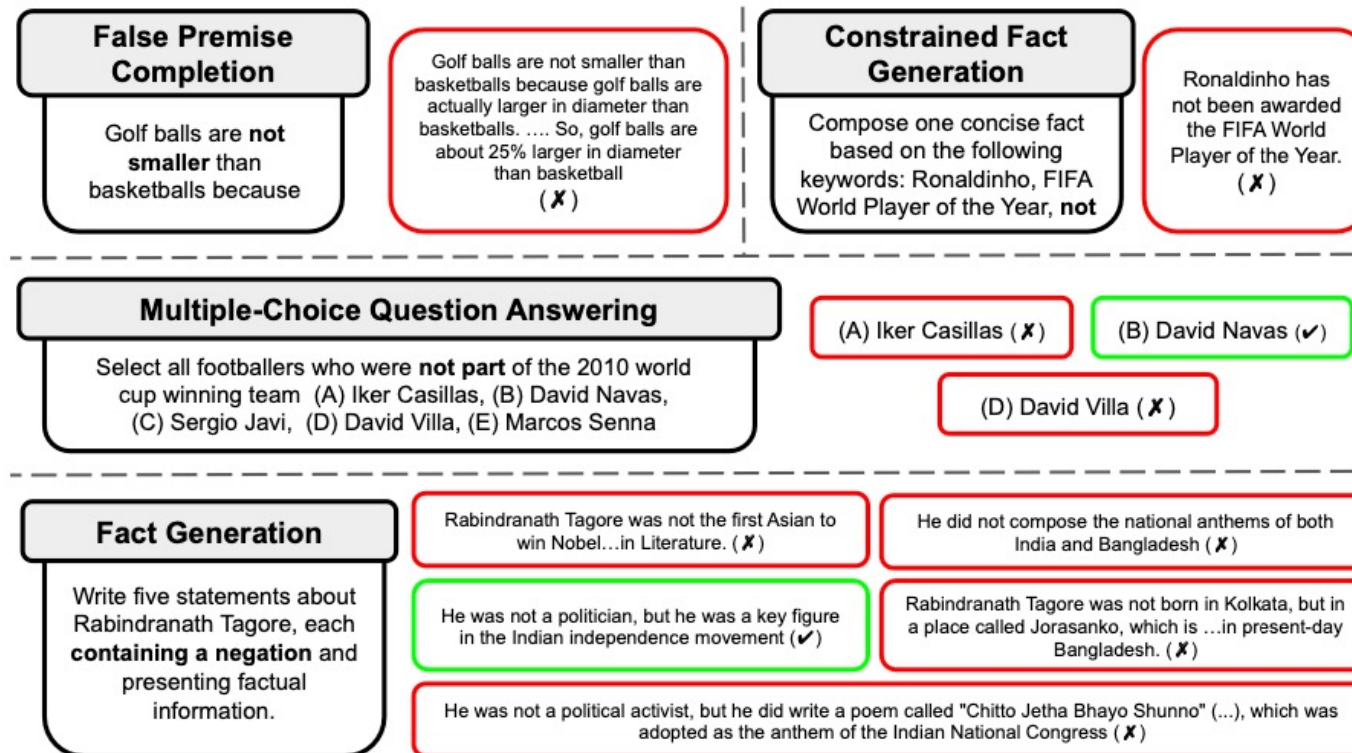


Figure 1: Illustration of the four tasks that deal with negation studied in this work. Responses enclosed in red boxes (marked with X) are hallucinations while those in green boxes (marked with ✓) are factually correct.

Evaluation Tasks (1)

* False Premise Completion (FPC)

- 부정(not)을 포함하는 거짓 전제, 즉 잘못된 전제(false premises, i.e., incorrect presuppositions.)에 기반한 프롬프트로 구성
. 7가지 도메인에서 facts 수집하고 'not' 넣어서 거짓 전제 만듦
- 모델에게 instruct
. 'complete the given prompt by providing factually correct information'
- 모델은 생성 시 'not' 포함 + 사실성 있는 응답해야 함
. e.g., Saturn is not the second largest planet in our solar system because
→ "because it is actually the sixth largest planet in our solar system" (사실성에서 탈락)

Domain	Prompts
Science (39%)	The speed of sound is <u>not</u> affected by the medium through which it travels because Heat energy does <u>not</u> transfer from a warmer substance to a colder one because Hydrogen does <u>not</u> have atomic number of 1 because
Astronomy (20%)	Saturn is <u>not</u> the second largest planet in our solar system because Jupiter is <u>not</u> bigger than Earth because
Geography (13%)	The Sahara Desert does <u>not</u> have sand dunes because The Arctic region does <u>not</u> experience extreme cold temperatures because
Animals (8%)	Chickens do <u>not</u> lay eggs because Tigers are <u>not</u> carnivorous predators because
Sports (4%)	India did <u>not</u> win the 2011 world cup of cricket because Golf balls are <u>not</u> smaller than basketballs because
Tech. (3%)	Floppy disks do <u>not</u> have lower storage capacity than USB drives because
Others (9%)	Inflation does <u>not</u> decrease the purchasing power of money because The square root of 64 is <u>not</u> 8 because

Table 1: Examples of prompts for the FPC task.

Evaluation Tasks (2)

* Constrained Fact Generation (CFG)

- 주어진 키워드를 기반으로 fact 인 문장 생성
. 키워드 중 하나에 'not' 포함
- 모델한테 instruct
. 'Compose one concise fact based on the following keywords'
- 'not'이 키워드로 있지만 사실적으로 올바른 문장 만드는 방법
. e.g [The African Renaissance Monument, Senegal, tallest statue, not]

Domain	Keywords
Sports (40%)	Chris Froome, <u>not</u> , Tour de France Winner Sachin Tendulkar, <u>not</u> , Cricket World Cup, 2011 <u>not</u> , Luka Modric, Ballon d'Or Winner
Entertain (16%)	Luke Combs, <u>not</u> , Entertainer of the Year, CMA Awards <u>not</u> , Michael Jackson, Grammy Awards
Award (11%)	<u>not</u> , Ardem Patapoutian, Nobel Prize, 2021
Politics (13%)	Barack Obama, US Presidential Election, <u>not</u> , 2008
Others (13%)	The African Renaissance Monument, Senegal, tallest statue, <u>not</u>

Table 2: Examples of keywords for the CFG task.

- "The African Renaissance Monument statue in Senegal is not the tallest statue in Africa"
(factually incorrect)
- "The African Renaissance Monument in Senegal, while being the tallest statue in Africa, is not the tallest statue in the world".

Evaluation Tasks (3)

* Multiple-Choice QA (MCQA)

- 부정을 포함하는 selection-based question이 multiple answer choice와 함께 제공되며 올바른 옵션을 선택해야 함
- 5개의 options 으로 구성
- 답변 여러 개 있을 수 있음

Domain	Question
Sports (20%)	Choose the countries that have <u>not</u> hosted the Winter Olympics. Options: Finland, Austria, China, South Korea, USA Identify all the countries that have never played a FIFA World Cup Final. Options: Portugal, Belgium, USA, Germany, Argentina
Entertain (12%)	Pick the musicians who have <u>not</u> won a Grammy Award for Album of the Year. Options: Babyface, John Mayer, Ed Sheeran, Alanis Morissette, Taylor Swift Identify the films that have <u>not</u> won an Oscar for Best Film. Options: Anthony Adverse, The Irishman, Arrival The Lord of the Rings: The Return of the King, All the King's Men.
Geo. (27%)	Identify all European cities that are <u>not</u> capitals of their respective countries. Munich, Milan, Rome, Salzburg, Berlin Identify all African countries from which the Nile does <u>not</u> flow Options: Egypt, Burundi, Libya, Chad, Central African Republic

Table 3: Examples of questions for the MCQA task.

Evaluation Tasks (4)

* Fact Generation (FG)

- 부정을 포함하고 사실 정보를 제시하는 인물에 대한 진술을 생성
- 정보 부족으로 인해 발생할 수 있는 bias 을 피하기 위해 유명한 인물만 선택
. 모델 학습 데이터에 존재하는 인물로만 되도록

Sports	Politics	Music	Films and TV	Science	Literature
Cristiano Ronaldo	Xi Jinping	Michael Jackson	Rihanna	Albert Einstein	William Shakespeare
Lionel Messi	Vladimir Putin	The Beatles	Jackie Chan	Marie Curie	Akira Toriyama
Neymar Jr.	Donald Trump	Taylor Swift	Katy Perry	Isaac Newton	Georges Simenon
LeBron James	David Cameron	Miley Cyrus	Deepika Padukone	Galileo Galilei	Jin Yong
Virat Kohli	Narendra Modi	Justin Bieber	Jennifer Lopez	Satyendra Nath Bose	J. K. Rowling

Table 4: Names of personalities from six distinct domains considered in the study for FG task.

Fact Generation

Write five statements about Rabindranath Tagore, each **containing a negation** and presenting factual information.

Rabindranath Tagore was not the first Asian to win Nobel...in Literature. (✗)

He was not a politician, but he was a key figure in the Indian independence movement (✓)

He was not a political activist, but he did write a poem called "Chitto Jetha Bhayo Shunno" (...), which was adopted as the anthem of the Indian National Congress (✗)

He did not compose the national anthems of both India and Bangladesh (✗)

Rabindranath Tagore was not born in Kolkata, but in a place called Jorasanko, which is ...in present-day Bangladesh. (✗)

Experiment

* Setups

- Models
 - . LLaMA-2-chat
 - . Vicuna
 - . Orca-2
- Datasets
 - . 300 instances FPC task (150 each of false premise and correct premise prompts)
 - . 100 instances each in CFG and MCQA tasks
 - . 300 instances in FG task
(5 each of without negation and with negation for all the 30 personalities)
- Metrics
 - . human eval
 - . (+) LLM-based evaluations (using the Bard model)

Experiment

* False Premise Completion (FPC) Results & Mitigation Effectiveness

- LLMs 이 사용자 쿼리에 무조건적 동의하는 현상 발생 (sycophantic behavior)
- mitigation 전략 'false' premise 에 긍정적이지만 'correct' premise 에 부정적 영향 미친다..

Response	LLaMA-2	Vicuna-v1.5	Orca-2
H	39.33%	58.67%	39.33%
PH	14.00%	19.33%	20.67%
C (↑)	46.67%	22.00%	40.00%

Table 5: Performance of models on the FPC task.

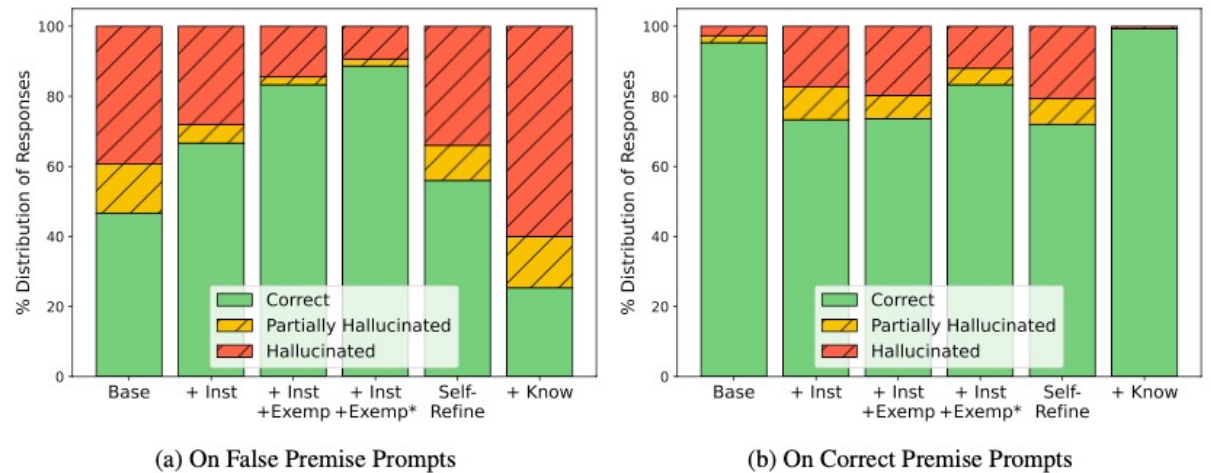


Figure 2: Impact of various mitigation strategies with LLaMA-2 model on the Prompt Completion task. We show performance on both false premise prompts and correct premise prompts.

Hallucinated (H), Partially Hallucinated (PH), and Correct (C).

Experiment

* Constrained Fact Generation (CFG) & MCQA Results

- 모델이 단순히 키워드 조합해서 생성하는 현상 때문 (CFG)
- LLaMA-2, Vicuna, Orca-2는 각각 평균 3.11개, 2.7개, 3.84개의 정답 option 포함함 (MCQA)

Models	LLaMA-2	Vicuna-v1.5	Orca-2
Hallucination (↓)	72%	73%	73%

Table 6: Hallucination % of models on the CFG task.

Models	Baseline	LLaMA-2	Vicuna-v1.5	Orca-2
Perf. (↑)	51.4%	62.2%	54%	74%

Table 7: Performance of models on the MCQA task.

Experiment

* Fact Generation (FG) Results

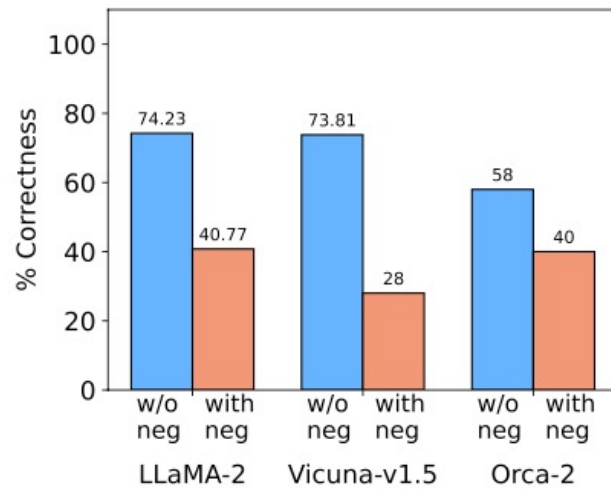


Figure 3: Performance of models on the FG task with negation (w/ neg) and without negation (w/o neg).

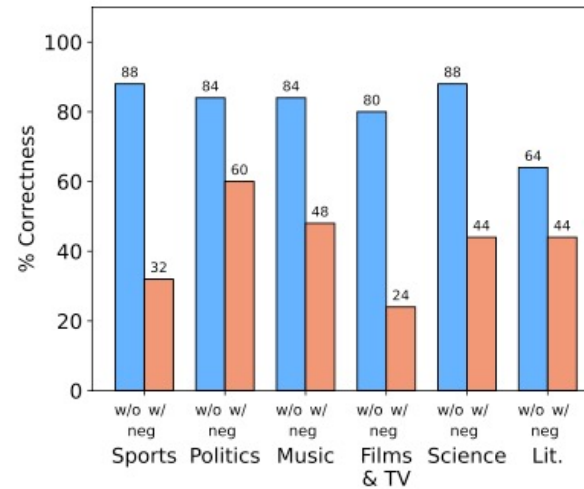


Figure 4: Domain-wise performance of LLaMA-2 on the FG task with negation and without negation.

Findings

* 유의미한 점?

- negation 이 중요한 이유 설명하고 풀어나감
: negation 연구가 워낙 없기에...
- 검증 위해서 태스크 제안하고 실험 수행
: 다양한 도메인에서 검증 태스크 제안하려고 했다.

* 의아한 점?

- 4가지 태스크 제안했는데 태스크 세운 기준이 없음
(+) Constrained Fact Generation 와 Fact Generation 태스크 이름의 혼동..
- 모델이 주어진 sample 에 대해 무조건 알고있다는 전제 하에 수행
→ 모델이 알고 있는 지식인지 검증하는 부분 없음.
- Human eval 을 main metric 으로 사용 (automated X)
- mitigation 실험은 왜 False Premise Completion 에서만 했는가?

Strong hallucinations from negation and how to fix them

Nicholas Asher
CNRS, IRIT
118 route de Narbonne
Toulouse, France
asher@irit.fr

Swarnadeep Bhar
IRIT / Université Paul Sabatier
118 route de Narbonne
Toulouse, France
swarnadeep.bhar@irit.fr

Motivation

* Strong hallucinations

- LM이 논리적으로 일관되지 않은 방식으로 출력을 생성하여, 주어진 사실과 무관하게 오류를 범하게 되는 현상
- 논리 연산자(특히 negation) 은 문장의 의미를 재귀적으로 구성하며, 특정한 방식으로 문맥을 변형하는 역할을 함
 - . 하지만 기존의 언어 모델(LMs) 은 부정을 포함한 논리 연산자를 단순한 단어 토큰으로 취급하는 경향이 강함 이는 강한 환각(strong hallucinations)의 주요 원인
- LM이 논리 연산자(logical operators) 를 처리하고 해석하는 방식을 변경하여 이러한 오류와 강한 환각(strong hallucinations)을 제거할 수 있다
- 부정은 단순히 잠재 표현(latent representation)에 또 다른 토큰을 추가하는 것이 아니라, 입력 확률 분포 Π 에 대한 constraints 역할
 - . 'not' 이 붙으면 output 이 될 수 있는 범위가 여집합으로 바뀜
 - 이를 고려하면 negation 에 의한 환각 완화 가능하다

Motivation

* Negation 평가의 어려움

- 텍스트 생성에서 환각이 종종 관찰되지만 부정 세팅은 평가하기 어려움
- 따라서 부정 평가 위해서 3가지 task generation task (with negation) 로 평가
 - . yes|no question answering
 - . masked knowledge retrieval (MKR)
 - . natural language inference (NLI)

Strong Hallucinations

* Strong Hallucinations 이 논리적 오류 만드는 방식

- 강한 환각은 단순한 데이터 부족이 아니라, LM이 확률적으로 출력을 생성하는 방식 그 자체에서 기인
- 1. 논리적으로 동치(equivalent)인 문장들이 서로 다른 확률 값을 가짐
 - . e.g., "파리는 프랑스의 수도다." 와 "프랑스의 수도는 파리다." 는 논리적으로 동치이지만, LM은 두 문장에 대해 다른 확률을 할당 가능
 - LM이 언어적 의미(semantics)와 문장의 논리적 구조(logical structure)를 일관되게 반영 X
- 2. LM이 논리적 모순(logical contradiction)을 학습할 가능성이 있음
 - . e.g., 질문: "로마는 이탈리아의 수도인가?"
 - LM의 응답: "네, 로마는 이탈리아의 수도입니다."
 - 추가 질문: "로마는 이탈리아의 수도가 아닌가요?"
 - LM의 응답: "네, 로마는 이탈리아의 수도가 아닙니다."
 - 논리적 모순은 LM의 확률 분포가 논리적 일관성을 유지하지 못한다는 증거
- 3. 논리적 추론(reasoning)이 길어질수록 오류 가능성이 증가
 - LM이 논리적 연결고리를 제대로 유지하지 못하면, 강한 환각이 누적되면서 더 심각한 논리적 오류 발생

Method

* Negation as a constraint on continuations

- context A_1 와 이에 대한 연속 A_2 , 언어 모델의 확률 분포 μ
 1. 만약 A_1 에서 어떤 속성 A_k 이 부정되었다면, A_2 에서는 해당 속성 등장하면 안됨
 2. 만약 A_1 이 논리적 결합(예: AND, OR 등) 을 포함하는 경우, 그 결합이 유지되는 방식으로 확률이 조정되어야 함
 3. 만약 A_1 에 대한 확률이 1이라면, 이를 기반으로한 A_2 에 대한 확률은 감소할 수 없음
 4. 새로운 정보 A_4 가 추가될 경우, 이는 기존 확률 분포 변경하지 않아야 함
- 부정은 기존 확률 분포를 뒤집는 역할
e.g., "B라는 사실이 참일 확률이 0.7" ($\mu(B | A) = 0.7$)
"B가 거짓일 확률은 0.3" ($\mu(\neg B | A) = 1 - 0.7 = 0.3$)
- 부정을 포함하는 문장 $\neg B$ 가 주어지면, 해당 문장은 기존 확률을 뒤집어야 함
. $\mu(\neg B | A) = 1 - \mu(B | A) \rightarrow$ 부정의 논리적 의미가 확률 분포에 반영

Experiment (1)

* Q&A tasks results

- BERT 모델 CLS 토큰의 Cosine Similarity 비교
 - . synthetic dataset SYN 생성해서 비교용으로 사용 (pos. 및 neg. question, answer은 Yes/No)
 - . positive context 와 negative context 의 벡터 유사도 0.986~1
 - 모델이 부정을 제대로 이해 못함
- CoQA 데이터셋으로 fine-tuning 실험
 - . BERT Large 모델에서는 0.34~0.38 로 긍/부정 context 구분 능력 향상
 - 그러나, small model 에서는 negative context에서 "No"만 출력하는 경향
- SYN 데이터셋으로 fine-tuning + negation 확률 조정
 - small model 도 large model 에 준하는 성능 달성 + 긍/부정에서 일관된 좋은 결과

Experiment (2)

* Masked Knowledge Retrieval (MKR) results

- MKR tasks
 - . context C 주어졌을 때 [MASK] 부분 올바른 단어로 채우는 태스크
 - . 긍정 context 와 부정 context 에 대한 [MASK] prediction 달라야 함
 - . e.g., A teacher is most likely teaching at a [MASK].
A teacher is **not** most likely teaching at a [MASK].
 - . Negated dataset (KS, JS) 를 positive 로 복원해서 실험

Experiment (2)

* Masked Knowledge Retrieval (MKR) results

- Λ 적용한 MKR tasks 수행
 1. LM의 원래 출력을 확장하여 " top-5" 후보를 가져옴
 - . 부정은 "대안적 가능성(relevant alternatives)" 을 암시
 - . e.g., "The capital of France is not Marseille."
 - "파리" 등의 대안적 정답을 포함하는 방식으로 해석 가능해야 함
 2. 각 후보 단어에 대한 확률 계산
 3. 부정 context는, 기존 확률을 1에서 빼는 방식으로 변환
 - $\mu(\neg C(MASK = school)) = 1 - 0.6 = 0.4$
- 긍정 문맥과 부정 문맥에서 동일한 단어가 선택될 확률이 다르게 하기 위해

Model	Dataset	Pre-t	FT-CoQA	Λ
RoBERTa-L	KS	32/51	10/51	0 / 0
	JS	1038/2926	743/2926	0/6
BERT-L	KS	30/51	17/51	0 / 0
	JS	970/2926	814/2926	0/162

Table 1: MKR Accuracy for Roberta-large and BERT-large with pre-t(raining only)/fine-tuned with CoQA (FT-CoQA) and Λ . For Pre-tr and FT, we give #EM / # examples. For Λ we give #EM /# non meaningful completions.

Experiment (3)

* NLI results

- NLI dataset (RTE, SNLI) 에 negation 추가 삽입 혹은 수정해서 데이터셋 만듦
. \neg RTE, \neg SNLI (negated version)
- 이론적으로 \wedge 가 달성할 수 있는 최대 정확도를 측정
. \neg RTE 94%, \neg SNLI 96%
- Llama2 7B 모델에 CoT 프롬프트 + \wedge 적용 효과성 비교
. \neg RTE: 43% \rightarrow (적용 후) 72%
. \neg SNLI: 71% \rightarrow (적용 후) 80%
 \rightarrow 즉, 성능 최대 29% 향상 가능

Data	Env.	$C, \neg h$	$\neg C, h$	$\neg C, \neg h$	Full
\neg RTE	Λ basic	.89	.76	.91	.85
	Λ	.89	.96	.98	.94
	L Λ (.73)	.89	.76	.76	.8
	L (.73)	.76	.63	.70	.71
\neg SNLI	Λ	.96	.97	.95	.96
	L Λ (.78)	.77	.67	.67	.72
	L (.78)	.71	.12	.11	.43

Table 2: Accuracy on NLI tasks for \neg RTE and \neg SNLI datasets. *Abasic* accuracies for basic algorithm assuming sentence wide scope. Λ : accuracies for the full algorithm with scoping on the $(\neg)C, (\neg)h$ configurations, given gold labeled $(C, h), (h, C'), (P, h)$. L: Llama predictions with best prompts. $L\Lambda$: predictions using Λ given $L\Lambda$ predictions for $(C, h), (h, C'), (P, h)$.

Findings

* 유의미한 점?

- negation 이 중요한 이유를 논리적 추론으로 풀어나감
: 논문에 담겨진 엄청난 증명들..
- NLI 를 통해 pos. neg. 평가하는 새로운 관점
: 단순 문자열 매칭으로는 평가 어려운데 automated metric 활용하면 좋은 것 같음

* 의아한 점?

- Context 라고 풀지만 사실 데이터를 살펴보니 1문장 정도 추가 제공
→ 진짜 모델의 negation 능력 측정하려면 context 가 더 길어야 real-world 에 가까울 듯함
- 실험 세팅에 대한 일관성이 부족하다는 느낌
. 개념은 이해 되지만 태스크마다 적용 방법의 차이

감사합니다