

Knowledge Conflicts & LLM Improvements

EMNLP 2023

논문 세미나

이정섭

Contents

- **ALCUNA: Large Language Models Meet New Knowledge**
 - EMNLP 2023 long
- **Large Language Models Can Self-Improve**
 - EMNLP 2023 long
 - Citations: 152

ALCUNA: Large Language Models Meet New Knowledge

Xunjian Yin* and **Baizhou Huang*** and **Xiaojun Wan**

Wangxuan Institute of Computer Technology, Peking University

Center for Data Science, Peking University

The MOE Key Laboratory of Computational Linguistics, Peking University

{xjyin, hbz19, wanxiaojun}@pku.edu.cn

Introduction

Motivations

- LLM은 다양한 Tasks에서 뛰어난 성능을 보임
- 하지만, 세계의 지식은 빠르게 변화함. LLMs이 new knowledge를 맞닥뜨렸을 때, 이러한 지식을 처리하는 능력을 평가하는 벤치마크가 부족
- 본 논문에서는 이를 해결하기 위해, artificial entities로 new knowledge를 만들고, 이를 사용하여 새로운 Benchmark인 ALCUNA를 소개

Introduction

New Knowledge ?

- 모델이 실제로 new knowledge를 접하는 변화가 빠른 세계에 살고 있음.
 - LLM을 재학습 하는 것은 비용적으로 매우 비효율적
 - new knowledge를 LLM과 같이 사용하기 위해, 검색 모델과 같이 사용
- 학습 데이터와 테스트 데이터 간의 overlapping으로 LLM의 일반화 능력을 제대로 평가할 수 없음 (LLM이 이미 테스트 데이터를 학습했을 수도 있음)

Knowledge Generation

Artificial Entity

- ontology 설계를 바탕으로 artificial entity 생성
- Alpaca와 Vicuna는 Camels에 속하는 real entity이며, 이 둘을 이용해 Alcuna라는 artificial entity 생성
- 상속 (heredity): 부모 엔티티와 높은 유사성을 보이고, 클래스 공통성을 준수하는 엔티티 구축
- 변이 (variation): 형제 엔티티들과 다른 속성 값을 가지도록 변경하는 과정
- 확장 (extension): 부모 엔티티에는 없지만 형제 엔티티들이 가지고 있는 속성이나 관계를 추가
- Dropout: 특정 속성이나 관계를 제거

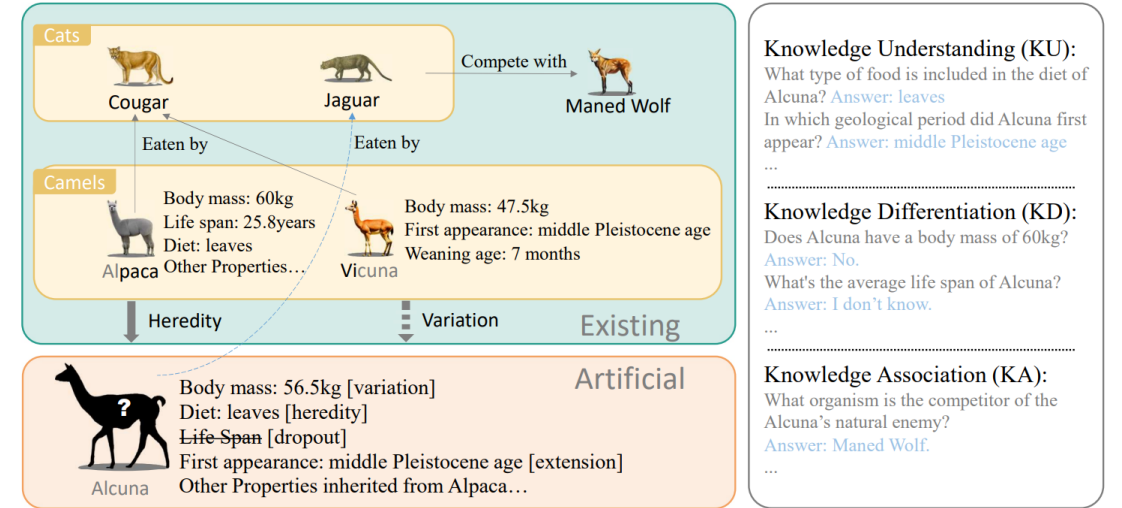


Figure 1: Demonstration of ALCUNA, including heredity, variation, extension and dropout operations in KnowGen, generated artificial entity named Alcuna and three types of questions related to it.

Knowledge Generation

Artificial Entity

- ontology 설계를 바탕으로 artificial entity 생성
- Alpaca와 Vicuna는 Camels에 속하는 real entity이며, 이 둘을 이용해 Alcuna라는 artificial entity 생성
- 상속 (heredity): 부모 엔티티와 높은 유사성을 보이고, 클래스 공통성을 준수하는 엔티티 구축
- 변이 (variation): 형제 엔티티들과 다른 속성 값을 가지도록 변경하는 과정
- 확장 (extension): 부모 엔티티에는 없지만 형제 엔티티들이 가지고 있는 속성이나 관계를 추가
- Dropout: 특정 속성이나 관계를 제거

Algorithm 1: Knowledge Generation

```
input : One Class  $C$ 
output : Property Set  $\mathcal{T}(\tilde{e})$  of  $\tilde{e}$ 

 $e^p \leftarrow \text{RandomSelect}(C)$ ;
 $E^{psb} \leftarrow \text{sib}(e^p)$ 
// Get the triplet set for heredity, variation, dropout
and extension
 $\mathcal{T}_R^h, \mathcal{T}_R^v, \mathcal{T}_R^d \leftarrow \text{RandomSplit}(\mathcal{T}_R(e^p))$ 
 $\mathcal{T}_A^h, \mathcal{T}_A^v, \mathcal{T}_A^d \leftarrow \text{RandomSplit}(\mathcal{T}_A(e^p))$ 
 $\mathcal{T}^e \leftarrow \text{RandomSample}(\mathcal{T}(E^{psb}))$ 

// Heredity and Dropout
 $\mathcal{T}_R(\tilde{e}) \leftarrow \mathcal{T}_R(C) \cup \mathcal{T}_R^h$ ;
 $\mathcal{T}_A(\tilde{e}) \leftarrow \mathcal{T}_A(C) \cup \mathcal{T}_A^h$ 

// Variation: replacing the object with an entity from
the same class
for  $(e^p, r, e')$  in  $\mathcal{T}_R^v$  do
     $E'^{sb} \leftarrow \text{sib}(e')$ 
     $e'^{sb} \leftarrow \text{RandomSelect}(E'^{sb})$ 
     $\mathcal{T}_R(\tilde{e}) \leftarrow \mathcal{T}_R(\tilde{e}) \cup \{(\tilde{e}, r, e'^{sb})\}$ 

// Variation: add gaussian noise to the value or copy
from  $E^{psb}$ 
for  $(e^p, a, v)$  in  $\mathcal{T}_A^v$  do
    if isnum( $v$ ) then
         $\tilde{v} \leftarrow v + \mathcal{N}(0, v/10)$ 
    else
         $e^{psb} \leftarrow \text{RandomSelect}(E^{psb})$ 
         $\tilde{v} \leftarrow \text{GetValue}(e^{psb}, a)$ 
     $\mathcal{T}_A(\tilde{e}) \leftarrow \mathcal{T}_A(\tilde{e}) \cup \{(\tilde{e}, a, \tilde{v})\}$ 

// Extension and get final property
 $\mathcal{T}(\tilde{e}) \leftarrow \mathcal{T}_A(\tilde{e}) \cup \mathcal{T}_R(\tilde{e}) \cup \mathcal{T}^e$ 
```

Question Answering as Evaluation Task

Question Answering

- attribute triplet (\tilde{e}, a, v) 을 사용하여, artificial entity에 대한 one-hop question $q(\tilde{e}, a, v)$ 생성
- chain of relation triplets $\mathcal{T}_C = (\tilde{e}, r, e_1) \rightarrow (e_1, r_1, e_2) \rightarrow \dots \rightarrow (e_{N-1}, r_{N-1}, e_N)$ 을 활용하여, artificial entity가 주어졌을 때, tail에 속하는 e_N 에 대해 질문하는 multi-hop question $q(\mathcal{T}_C)$ 구성

Question Answering as Evaluation Task

LLM이 New knowledge에 직면했을 때, 다음의 Question으로 세부화

- knowledge understanding (KU)
 - 새로운 지식을 기억하고 이해하는 능력 평가
 - 제공된 정보를 기반으로 정확한 답변이 가능한지 평가
- knowledge differentiation (KD)
 - 인공 엔티티와 부모 엔티티를 구별하는 능력 평가
 - Variation set과 Dropout set에서 attribute triplet을 샘플링하여 생성
- knowledge association (KA)
 - 인공 엔티티와 기존 지식 사이의 연관성을 바탕으로, multi-hop 추론이 가능한지 평가

The diagram illustrates two types of knowledge: 'Existing' and 'Artificial'. The 'Existing' section features a light blue rounded rectangle containing an image of a maned wolf with the text 'with' and an arrow pointing to it, and 'Maned Wolf' below. The 'Artificial' section features a light orange rounded rectangle containing the text 'kg', ': middle Pleistocene age', and 'months'. Below the 'Artificial' section is a label '[extension]'.

Knowledge Understanding (KU):
What type of food is included in the diet of Alcuna? **Answer: leaves**
In which geological period did Alcuna first appear? **Answer: middle Pleistocene age**
...

Knowledge Differentiation (KD):
Does Alcuna have a body mass of 60kg? **Answer: No.**
What's the average life span of Alcuna? **Answer: I don't know.**
...

Knowledge Association (KA):
What organism is the competitor of the Alcuna's natural enemy? **Answer: Maned Wolf.**
...

ity, variation, extension and dropout operations in KnowGen, es of questions related to it.

Question Answering as Evaluation Task

LLM이 New knowledge에 직면했을 때, 다음의 Question으로 세부화

- knowledge association (KA)
 - 인공 엔티티와 기존 지식 사이의 연관성을 바탕으로, multi-hop 추론이 가능한지 평가
 - 인공 엔티티 *Alcuna*가 *Jaguar*에게 먹히고, *Jaguar*가 *Maned Wolf*와 경쟁한다는 관계가 있다면, KA는

Knowledge Association (KA):

What organism is the competitor of the Alcuna's natural enemy?

Answer: Maned Wolf.

...

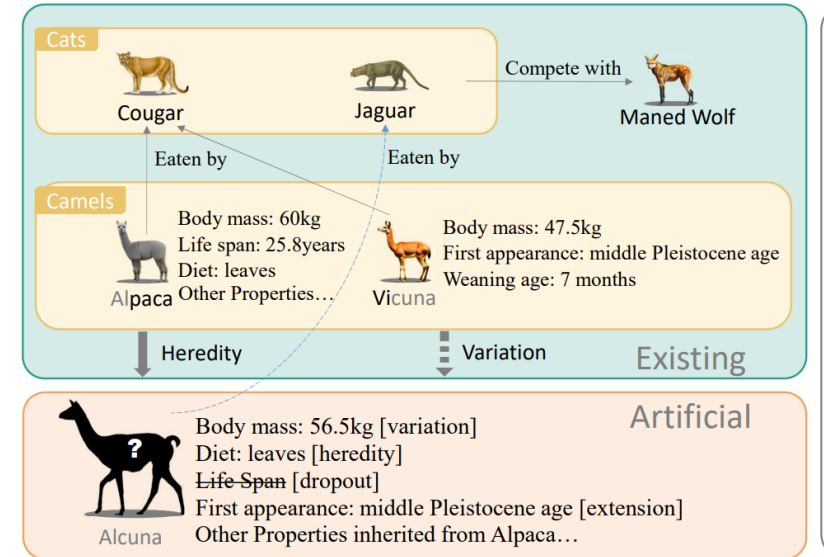


Figure 1: Demonstration of ALCUNA, including heredity, variation, extensic generated artificial entity named Alcuna and three types of questions relate

ALCUNA: Proposed Benchmark

EOL3 데이터베이스 사용

- 모든 생물학적 분류군을 tree로 구조화
- 669 property types / 2404790 entities / 13625612 properties

질문 생성은 ChatGPT 활용

| | multiple choice | Boolean | fill-in-the-blank |
|-------|-----------------|---------|-------------------|
| KU | 2459 | 4232 | 4625 |
| KD | 3487 | 34497 | 19698 |
| KA | 15353 | 0 | 0 |
| Total | 21299 | 38729 | 24323 |

Table 6: Number of different forms of KU, KD and KA questions.

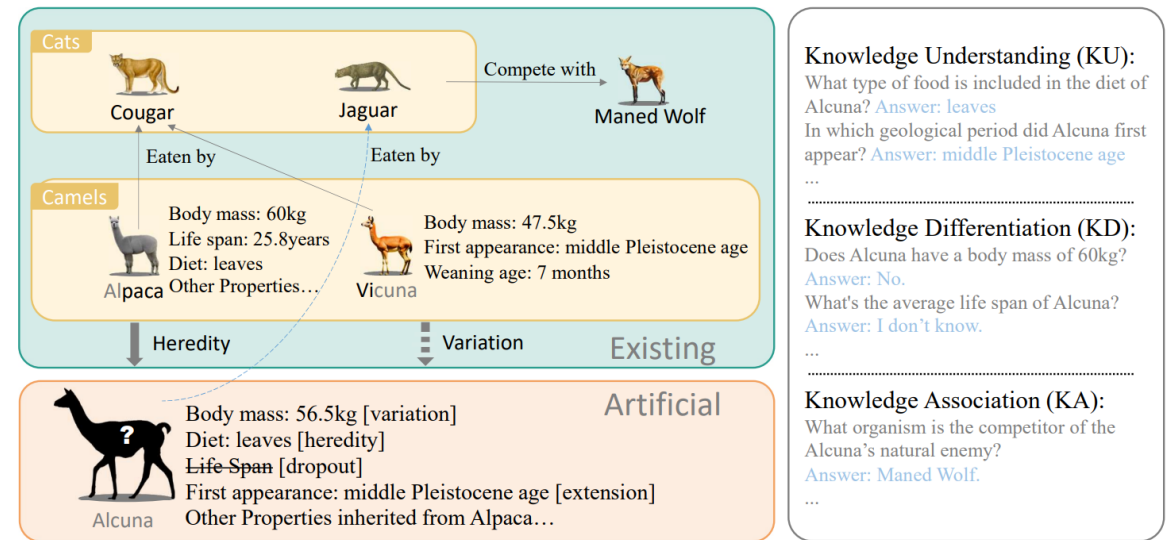


Figure 1: Demonstration of ALCUNA, including heredity, variation, extension and dropout operations in KnowGen, generated artificial entity named Alcuna and three types of questions related to it.

ALCUNA: Proposed Benchmark

System: You are a powerful multi-hop question generator with biological knowledge. Users will provide a chain of RDF triples, and you will help write a question to ask the tail entity from the head entity. You shouldn't include bridge entities in the generated question. The question should only include the head entity [T], which is placeholder. If you can't create a valid question, reply with \"[None]\" only. Don't reply with any explanation or other information.

/ exemplars */*

User: Given RDF triples ([T], compete with, x1), (x1, are eaten by, x2), (x2, are host of, x3), write a question to ask x3. Don't mention x1, x2, ... Write the possible question in natural English.

Agent: Multi-hop question: Which entities serve as the habitats for the creatures that feed on the organisms competing with [T]?

User: Given RDF triples ([T], parasitize, x1), (x1, prey on, x2), write a question to ask x2. Don't mention x1, x2, ... Write the possible question in natural English.

Agent: Multi-hop question: What is the prey of an organism that is parasitized by [T]?

User: Given RDF triples ([T], eat, x1), (x1, compete with, x2), (x2, are pathogens of, x3), write a question to ask x3. Don't mention x1, x2, ... Write the possible question in natural English.

Agent: Multi-hop question: Which species has pathogens that compete with species that are eaten by [T]?

/ query */*

User: Given RDF triples {\$Triple chain}, write a question to ask {\$Tail entity}. Don't mention x1, x2, ... Write the possible question in natural English.

Figure 4: Prompt for generating multi-hop question templates.

System: You are a powerful question generation model with biological knowledge. Given a biological taxon's property name and a set of possible values of the property. You need to generate a yes-no question template with placeholders [T] and [V] about the given property satisfying the following constraints:

1. Most important: the generated question template must contain [T] and [V].
2. The placeholder [T] can be replaced with any taxons name.
3. The placeholder [V] can be replaced with possible value of the given property. You can refer to the given value set for the format of possible property value. But don't use any value in the value set directly.
4. Don't do the replacement, just provide the question template with [T] and [V] placeholders.
5. Try to create the question even if the property is not a biological property.
6. Don't reply with any explanation or other information.

/ exemplars */*

User: Property name: skeleton contains Value set: ['chitin']

Agent: Question template: Does the skeleton of [T] contain [V]?

User: Property name: body shape Value set: ['cylinder', 'elongated', 'bilaterally symmetric', 'radially symmetric']

Agent: Question template: What's the body shape of [T]?

User: Property name: litters per year Value set: ['1.0 /year', '3.0 /year', '1.5 /year', '2.0 /year', '2.5 /year']

Agent: Question template: Does one [T] produce [V] litters per year?

/ query */*

User: Property name: {\$Property name} Value set: {\$Possible values of Property}

Figure 5: Prompt for generating Boolean question templates.

ALCUNA: Proposed Benchmark

System: You are a powerful of RDF triples, and you will bridge entities in the gener: placeholder. If you can't cre: other information.

/ exemplars */*

User: Given RDF triples ([T] x3. Don't mention x1, x2, ...)

Agent: Multi-hop question: 'competing with [T]?'

User: Given RDF triples ([T] Write the possible question

Agent: Multi-hop question:

User: Given RDF triples ([T] x3. Don't mention x1, x2, ...)

Agent: Multi-hop question:

/ query */*

User: Given RDF triples {\$T} possible question in natural

System: You are a powerful question generation model with biological knowledge. Given a biological taxon's property name. You need to generate a question template with a placeholder [T] about the given property so that the placeholder [T] can be replaced with any taxons name. Try to create the question even if the property is not a biological property. Don't reply with any explanation or other information.

/ exemplars */*

User: Property name: skeleton contains

Agent: Question template: Which organic compound is a component of a [T]'s skeleton?

User: Property name: body shape

Agent: Question template: What's the body shape of [T]

User: Property name: litters per year

Agent: Question template: How many litters can [T] have per year?

/ query */*

User: Property name: {\$Property name}

Given a biological taxon's yes-no question template with nts:

ou can refer to the given value et directly. aceholders.

ymmetric', 'radiallysymmetric']
, '2.0 /year', '2.5 /year']

plates.

Fig

Figure 6: Prompt for generating fill-in-the-blank question templates.

Evaluation of LLMs

Models

ChatGPT, Alpaca-7B, Vicuna-13B, ChatGLM-6B

Prompts

- Zero/Few-shot
- Standard/CoT Prompting

Metrics

- Accuracy

```
You are a powerful question-answering system with knowledge in the field of biology.
Users will provide some biological information along with a question.
Your task is to combine the information provided by the user with your biological knowledge to
answer the question.
If you are unable to answer the question, simply respond with "I don't know."
Here is the basic information about a taxon you can refer:
###
{
  "name": "Bainvillevillea spinosa",
  "property": {
    "cellularity": ["multicellular"],
    "conservation status": ["least concern"],
    "geographic distribution": ["Ecuador"],
    "habitat": ["terrestrial"],
    "leaf complexity": ["compound"],
    "leaf morphology": ["broad"],
    "leaf sheddability": ["evergreen"],
    "plant growth form": ["branched"],
    "produces": ["oxygen"],
    "woodiness": ["woody"]
  },
  "rank": "species"
}
###
Answer the following question a few words: What is the habitat of Bainvillevillea spinosa?
Desired format: Thought process: <Thought process>, Final answer: [Final answer].
Let's think step by step.
```

Table 12: Demonstration of the zero-shot prompt in the CoT form.

Evaluation of LLMs

You are a powerful question-answering system with knowledge in the field of biology.
Users will provide some biological information along with a question.

Model: ChatGPT
Prompt:
- Zero-shot
- Standard
Metrics:
- Accuracy

Question: What organism is parasitized by an organism that feeds on *Agabus ajax*?
Answer: *Rangifer tarandus*
Model's Output: "Thought process: *Agabus ajax* is a species of aquatic beetle that feeds on Acari, Culicidae, *Daphnia magna*, *Daphnia pulex*, and Ceratopogonidae. To answer the question, we need to identify an organism that is parasitized by an organism that feeds on *Agabus ajax*.
Final answer: *Thala gorii*"]

Question: Is *Sorghum versinatum* a type of hyperhydrite?
Answer: Yes
Model's Output: "ANSWER: Yes"

Question: What's the preferred water depth range for *Plana*?
Answer: 500.0 cm
Model's Output: ANSWER: 500 cm\n\nIt is important to note that the information

nowledge to

iosa?

Table 11: Example of question and answer from our models.

Table 12: Demonstration of the zero-shot prompt in the CoT form.

Evaluation Results

| | ChatGPT | Alpaca | Vicuna | ChatGLM | ChatGPT | Alpaca | Vicuna | ChatGLM |
|-------------|-------------------|--------|--------|---------|---------------|--------|--------|---------|
| | Zero-Shot-Vanilla | | | | Zero-Shot-CoT | | | |
| KU | 50.19 | 31.02 | 34.12 | 34.64 | 68.75 | 39.81 | 39.61 | 24.85 |
| KD | 58.70 | 15.35 | 38.65 | 32.29 | 61.78 | 14.29 | 38.53 | 22.84 |
| KA | 28.44 | 24.60 | 29.71 | 10.29 | 35.36 | 19.66 | 29.55 | 4.97 |
| Avg. | 52.85 | 25.12 | 35.98 | 31.26 | 63.34 | 29.44 | 38.00 | 22.04 |
| | Few-Shot-Vanilla | | | | Few-Shot-CoT | | | |
| KU | 75.44 | 33.80 | 41.22 | 47.97 | 82.18 | 40.77 | 43.67 | 40.91 |
| KD | 64.20 | 38.97 | 46.76 | 41.42 | 74.99 | 36.24 | 55.81 | 37.19 |
| KA | 41.52 | 27.63 | 30.10 | 27.47 | 37.88 | 25.73 | 25.07 | 26.93 |
| Avg. | 64.37 | 35.09 | 42.74 | 42.56 | 74.11 | 38.02 | 47.73 | 37.64 |

Table 1: Performance of LLMs at our benchmark under different settings

knowledge association이 가장 어려움

→ New knowledge를 기존 지식과 연결하는 것에 큰 어려움을 겪음

Evaluation Results

Entity 유사성의 영향

- Attribute Similarity (Table 3)

엔티티간 Property가 겹치는 비율에 따라, KD Score 계산

- Property가 더 유사할 수록 모델이 혼란스러워 함 (ChatGPT 제외)
- New Knowledge와 기존 지식을 융합하지 못함 → 사람과 반대 !!!

- Name Similarity (Table 2)

인공 엔티티가 결국 이름으로 식별되기 때문에, 엔티티의 이름에 따라 성능이 달라질 것이라 예상

- 인공 엔티티에 무작위 할당 (random)
- 부모 엔티티의 한 문자를 무작위 할당 (similar)
- 인공 엔티티와 부모 엔티티의 이름이 유사할 수록 더 헛갈려 함

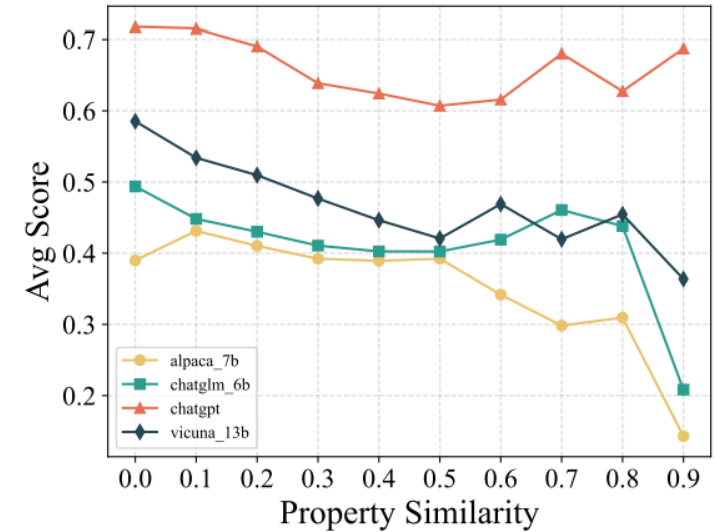


Figure 3: Relationship between model performance on KD questions and the property similarity between the artificial entity and its parental entity.

| | Alpaca | Vicuna | ChatGLM |
|---------|--------|--------|---------|
| similar | 38.74 | 47.23 | 41.61 |
| random | 39.52 | 47.64 | 43.20 |

Table 2: Results on KD questions of the entity with the same property and different name.

Conclusions

부모 엔티티와 유사할 수록 인공 엔티티를 더 헛갈려 함
→ 기존의 지식과 새로운 지식의 충돌 ?

생물학적 Entity에만 초점을 둔 것이 명확한 한계
→ New Knowledge에 대한 성능 평가를 완벽히 하지 못함

새로운 지식과 내부 지식 사이의 추론 관계
→ 기존 지식과 새로운 지식을 잘 엮을 수 있는 방법 필요

Large Language Models Can Self-Improve

**Jiaxin Huang^{1*} Shixiang Shane Gu² Le Hou^{2†} Yuexin Wu² Xuezhi Wang²
Hongkun Yu² Jiawei Han¹**

¹University of Illinois at Urbana-Champaign ²Google

¹{jiaxinh3, hanj}@illinois.edu ²{shanegu, lehou, crickwu,
xuezhiw, hongkunyu}@google.com

Introduction

Motivations

- LLM을 학습하기 위한 supervised dataset을 구축하는 것은 비용이 holy moly..
- 사람은 스스로 생각해서 추론 능력 향상할 수 있는데, LLM도 가능하지 않을까?

→ unlabeled data + CoT + self-consistency로 LLM을 fine-tuning하면 self-improve 가능!

Introduction

Table 1: Examples of 3 self-generated CoT reasoning paths given a question. Output 1 and 3 are the most consistent reasoning paths based on majority voting and kept as self-training data.

Question: Stefan goes to a restaurant with his family. They order an appetizer that costs \$10 and 4 entrees that are \$20 each. If they tip 20% of the total, what is the total amount of money that they spend?

Multiple Path Decoding:

Output 1: The appetizer costs \$10. The entrees cost \$20 each so in total $4 * \$20 = \80 . This means the total cost is $\$10 + \$80 = \$90$. They tip 20% of it, so the total amount they spend is $\$90 * 1.2 = \108 . The answer is 108.

Output 2: The appetizer costs \$10 and the entrees are \$20 each. There are 4 entrees so the sum is $\$20 * 4 = \80 . The waiter gets 20% of the total. 20% of \$80 is $\$80 * .2 = \16 . The answer is $\$80 + \$16 = \$96$. (*Incorrect reasoning path*)

Output 3: The appetizer costs \$10. The entrees cost $4 * \$20 = \80 . The tip is 20% of the total, so it is 20% of the \$90 they have spent. The tip is $0.2 * 90 = \$18$. The total they spent is $\$90 + \$18 = \$108$. The answer is 108.

LLM을 사용할 때, 정답이 나올 때도 있고, 오답이 나올 때도 있음

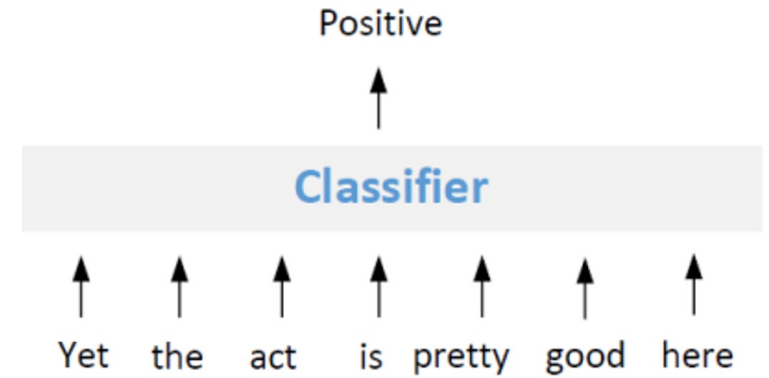
- 정답 CoT reasoning path를 잘 뽑아서 fine-tuning 하면, 모델의 확신도를 높일 수 있을 것

Backgrounds

Self-training

Self-training은 Semi-Supervised Learning 기법 중 대표적인 것 중 하나

- 1) 초기의 labeled dataset (original labeled data 혹은 gold data)으로 모델을 학습
- 2) 학습된 모델을 사용하여, unlabeled data의 label을 예측하여 추가적인 데이터 (pseudo-labeled data 혹은 silver data)를 얻음
- 3) 추가적으로 얻은 데이터를 이용하여 모델을 재학습
- 4) 위의 과정을 반복



(d) Self-training.



Self-training

Advantages:

1. unlabeled data를 이용할 수 있음
2. class imbalance 문제를 완화할 수 있음

Limitations:

1. unlabeled data가 필요함
2. unlabeled data의 도메인 등의 퀄리티가 보장되어야 함

Proposed Method

Language Model Self-Improved, LMSI

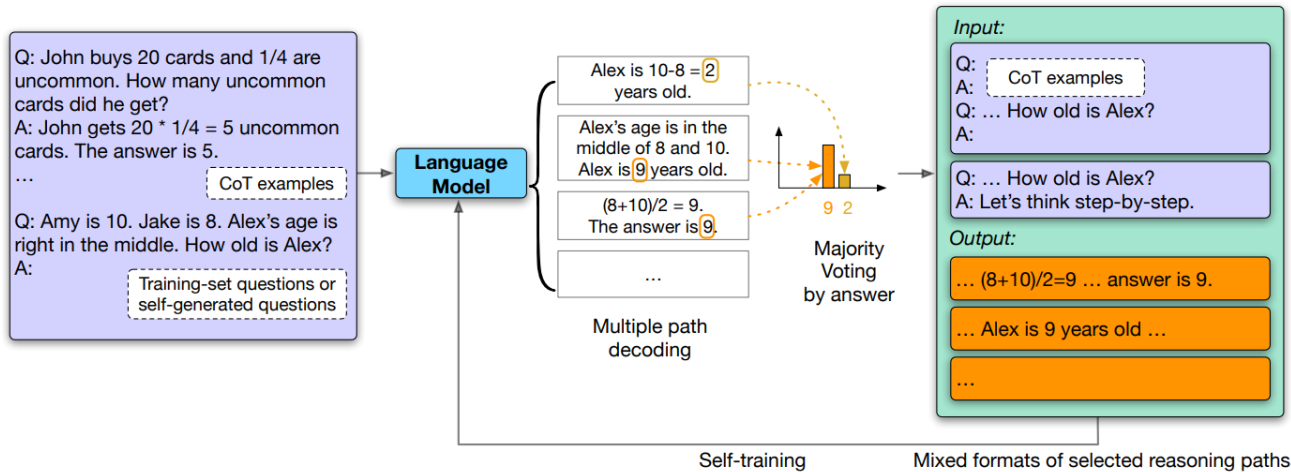


Figure 1: Overview of our method. With Chain-of-Thought (CoT) examples as demonstration (Wei et al., 2022c), the language model generates multiple CoT reasoning paths and answers (temperature $T > 0$) for each question. The most consistent answer is selected by majority voting (Wang et al., 2022c). The CoT reasoning paths that lead to the answer with the highest confidence are augmented by mixed formats, and are fed back to the model as the final training samples.

1. LLM을 이용하여 각 질문에 대해 여러 가지 추론 경로를 생성하고, 이러한 경로들에 대한 답변 예측
2. 다수결 방식(self-consistency)을 적용하여 여러 추론 경로 중에서 가장 일관된 답변을 선택. (답변이 정확한지는 고려 X)
3. 가장 일관된 답변으로 이어진 추론 경로만을 필터링하여 자체 생성된 훈련 데이터셋 구축
4. 특정 프롬프트나 답변 스타일에 대한 over-fitting을 방지하기 위해, 다양한 형식의 프롬프트와 답변을 혼합하여 자체 생성된 학습 데이터를 증강하고, 이 증강된 데이터로 LLM을 fine-tuning

Proposed Method

Language Model Self-Improved, LMSI

Table 2: An example of how a reasoning path is augmented into four formats of training data with different prompts (in input) and answer styles (in output). Specifically, the *CoT prompting examples* used for each tasks are listed in Appendix A.2. The *Standard prompting examples* are the same question-answer pairs with *CoT prompting examples*, except that reasoning is removed.

Question: Amy is 10 years old. Jake is 8 years old. Alex’s age is right in the middle. How old is Alex?
Selected Chain-of-Thought: Amy is 10 years old. Jake is 8 years old. Alex’s age is in the middle of Amy and Jake, so Alex is $(8 + 10) / 2 = 9$ years old. The answer is 9.

Mixed-formats of training data:

Format 1: Input: *[CoT prompting examples]* + ‘\n’ + *[Question]* + ‘\n’ + ‘A:’

Output: Amy is 10 years old. Jake is 8 years old. Alex’s age is in the middle of Amy and Jake, so Alex is $(8 + 10) / 2 = 9$ years old. The answer is 9.

Format 2: Input: *[Standard prompting examples]* + ‘\n’ + *[Question]* + ‘\n’ + ‘A:’

Output: The answer is 9.

Format 3: Input: *[Question]* + ‘\n’ + ‘A: Let’s think step by step.’

Output: Amy is 10 years old. Jake is 8 years old. Alex’s age is in the middle of Amy and Jake, so Alex is $(8 + 10) / 2 = 9$ years old. The answer is 9.

Format 4: Input: *[Question]* + ‘\n’ + ‘A:’

Output: The answer is 9.

4. 특정 프롬프트나 답변 스타일에 대한 over-fitting을 방지하기 위해, 다양한 형식의 프롬프트와 답변을 혼합하여 자체 생성된 학습 데이터를 증강하고, 이 증강된 데이터로 LLM을 fine-tuning

5. 추론 경로 생성, 필터링, 증강, fine-tuning의 과정을 반복하여, 지도 학습 데이터에 의존하지 않고도 모델의 추론 능력을 점진적으로 향상

Proposed Method

Language Model Self-Improved, LMSI

Question Generation

few-shot example의 Question을 통해 (label 사용 X), Question을 연속으로 이어서 시퀀스를 만든 다음에, 모델이 연속적으로 Question을 생성할 수 있도록 함

Prompt Generation

Zero-shot 설정에서, “Let’s think step by step.”으로 답변을 시작하도록 프롬프팅 하여, 모델이 여러 추론 경로를 생성하게 한 후에 이를 few-shot examples로 사용

Experimental Setup

Tasks and Datasets

Arithmetic reasoning

- **GSM8K** (Cobbe et al., 2021)
- **DROP** (Dua et al., 2019)

Commonsense reasoning

- OpenBookQA (Mihaylov et al., 2018)
- AI2 Reasoning Challenge (ARC) (Clark et al., 2018)

Natural Language Inference

- Adversarial NLI (ANLI) (Mihaylov et al., 2018)

Experimental Setup

Models, Prompt, Reasoning paths

Model

- PaLM 540B (Chowdhery et al., 2022)
- DROP (Dua et al., 2019)

Reasoning paths

- 각 question 별로 32개 생성
- 초기 LLM의 Temperature : 0.7
- self-improved LLM의 temperature : 1.2

Natural Language Inference

- Adversarial NLI (ANLI) (Mihaylov et al., 2018)ççç

Prompt

Table 10: Few-shot CoT prompts for OpenBookQA, from (Wang et al., 2022b).

Q: Poison causes harm to which of the following? (a) a Tree (b) a robot (c) a house (d) a car

A: Poison will harm living things, only a tree is a living thing. The answer is (a).

Q: As you look deeper into a Marbel you can see (a) the future (b) minut defects (c) colors (d) the other side

A: Marbel is not transparent, so you can not see the other side. Marbel does not necessarily have multiple colors. You will see minut defects. The answer is (b).

Q: When food is reduced in the stomach (a) the mind needs time to digest (b) take a second to digest what I said (c) nutrients are being deconstructed (d) reader's digest is a body of works

A: The food is being deconstructed in the stomach during digestion. The answer is (c).

Q: The sun is responsible for (a) puppies learning new tricks (b) children growing up and getting old (c) flowers wilting in a vase (d) plants sprouting, blooming and wilting

A: The sun can affect the growing of living things, like plants. The answer is (d).



Let's think step-by-step.

Evaluation Results

In-domain Benchmarks

Table 3: Accuracy results on six reasoning benchmarks with or without **LMSI** using different prompting method.

| Prompting Method | w. or w/o LMSI | GSM8K | DROP | ARC-c | OpenBookQA | ANLI-A2 | ANLI-A3 |
|--------------------|----------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| Standard-Prompting | w/o LMSI | 17.9 | 60.0 | 87.1 | 84.4 | 55.8 | 55.8 |
| | w. LMSI | 32.2 (+14.3) | 71.7 (+11.7) | 87.2 (+0.1) | 92.0 (+7.6) | 64.8 (+9.0) | 66.9 (+11.1) |
| CoT-Prompting | w/o LMSI | 56.5 | 70.6 | 85.2 | 86.4 | 58.9 | 60.6 |
| | w. LMSI | 73.5 (+17.0) | 76.2 (+5.6) | 88.3 (+3.1) | 93.0 (+6.6) | 65.3 (+6.4) | 67.3 (+6.7) |
| Self-Consistency | w/o LMSI | 74.4 | 78.2 | 88.7 | 90.0 | 64.5 | 63.4 |
| | w. LMSI | 82.1 (+7.7) | 83.0 (+4.8) | 89.8 (+1.1) | 94.4 (+4.4) | 66.5 (+2.0) | 67.9 (+4.5) |

Table 5: Ablation study: **LMSI** with different combinations of training format on GSM8K dataset.

| | Results on GSM8K | |
|------------------------|------------------|---------------|
| | Std. Prompting | CoT Prompting |
| w/o LMSI | 17.9 | 56.5 |
| LMSI w/o CoT formats | 23.6 (+5.7) | 61.6 (+5.1) |
| LMSI only few-shot CoT | 29.2 (+11.3) | 69.4 (+12.9) |
| LMSI w/ CoT formats | 32.2 (+14.3) | 73.5 (+17.0) |

Main Results

- Few-shot examples을 Standard, CoT, Self-Consistency (Majority vote)의 성능 비교
- GSM8K에서 +7.7%, DROP에서 +4.8%, OpenBookQA에서 +4.4%, ANLI-A3에서 +4.5%의 Gain

Ablation study

- 4가지 프롬프트를 모두 사용하는 Mixed-formats이 프롬프트에 over-fitting 되는 것을 막아 높은 성능을 가져옴

Evaluation Results

In-domain Benchmarks

Table 2: An example of how a reasoning path is augmented into four formats of training data with different prompts (in input) and answer styles (in output). Specifically, the *CoT prompting examples* used for each tasks are listed in Appendix A.2. The *Standard prompting examples* are the same question-answer pairs with *CoT prompting examples*, except that reasoning is removed.

Question: Amy is 10 years old. Jake is 8 years old. Alex’s age is right in the middle. How old is Alex?
Selected Chain-of-Thought: Amy is 10 years old. Jake is 8 years old. Alex’s age is in the middle of Amy and Jake, so Alex is $(8 + 10) / 2 = 9$ years old. The answer is 9.

Mixed-formats of training data:

Format 1: Input: *[CoT prompting examples]* + ‘\n’ + *[Question]* + ‘\n’ + ‘A:’

Output: Amy is 10 years old. Jake is 8 years old. Alex’s age is in the middle of Amy and Jake, so Alex is $(8 + 10) / 2 = 9$ years old. The answer is 9.

Format 2: Input: *[Standard prompting examples]* + ‘\n’ + *[Question]* + ‘\n’ + ‘A:’

Output: The answer is 9.

Format 3: Input: *[Question]* + ‘\n’ + ‘A: Let’s think step by step.’

Output: Amy is 10 years old. Jake is 8 years old. Alex’s age is in the middle of Amy and Jake, so Alex is $(8 + 10) / 2 = 9$ years old. The answer is 9.

Format 4: Input: *[Question]* + ‘\n’ + ‘A:’

Output: The answer is 9.

Table 5: Ablation study: **LMSI** with different combinations of training format on GSM8K dataset.

| | Results on GSM8K | |
|-------------------------------|-------------------------|---------------|
| | Std. Prompting | CoT Prompting |
| w/o LMSI | 17.9 | 56.5 |
| LMSI w/o CoT formats | 23.6 (+5.7) | 61.6 (+5.1) |
| LMSI only few-shot CoT | 29.2 (+11.3) | 69.4 (+12.9) |
| LMSI w/ CoT formats | 32.2 (+14.3) | 73.5 (+17.0) |

Ablation study

- 4가지 프롬프트를 모두 사용하는 Mixed-formats이 프롬프트에 over-fitting 되는 것을 막아 높은 성능을 가져옴

Evaluation Results

Out-domain Benchmarks

Table 4: Comparison of CoT-prompting accuracy results on six Out-Of-Domain benchmarks with or without training on six In-Domain (GSM8K, DROP, ARC-c, OpenBookQA, ANLI-A2, ANLI-A3) training-set questions.

| | Self-training data | AQUA | SVAMP | StrategyQA | ANLI-A1 | RTE | MNLI-M/MM |
|----------|--------------------|-------------|-------------|-------------|--------------|-------------|-----------------------|
| w/o LMSI | - | 35.8 | 79.0 | 75.3 | 68.8 | 79.1 | 72.0/74.0 |
| w. LMSI | GSM8K + DROP + ... | 39.0 (+3.2) | 82.8 (+3.8) | 77.8 (+2.5) | 79.2 (+10.4) | 80.1 (+1.0) | 81.8/82.2 (+9.8/+8.2) |

OOD Results

- GSM8K, DROP, ARC-c, OpenBookQA, ANLI-A2, ANLI-A3 데이터를 혼합해서 학습 후, 학습하지 않은 도메인 추론
- 전반적으로 성능 향상을 얻음

Evaluation Results

Self-Generating Questions VS Questions in the Training set

Table 6: Accuracy on GSM8K test set after self-training on different question sets. Results are shown for both CoT-Prompting (CoT) and Self-Consistency (SC).

| | Questions used for Self-Training | GSM8K | |
|----------|-------------------------------------|--------------|-------------|
| | | CoT | SC |
| w/o LMSI | - | 56.5 | 74.4 |
| w. LMSI | Generated | 66.2 (+9.7) | 78.1 (+3.7) |
| w. LMSI | Training-set | 73.5 (+17.0) | 82.1 (+7.7) |

Further Study

현재 Knowledge Conflict을 중심으로, LLMs의 평가와 개선에 대한 연구 중

"LLMs이 기존에 가지고 있는 지식에 어긋나는 새로운 지식이 주어졌을 때, 이를 잘 반영하여 생성을 잘 할 수 있는가. 그렇지 않다면 해결 방법은?"

1번째 논문

→ 기존에 가지고 있는 지식에 어긋나면, 생성을 잘 할 수 없을 것으로 보임

2번째 논문

기존 Data Augmentation 중, Semi-supervised Learning 방법을 사용

→ 여러 Task에 맞는 개선 방법이 추후 등장, LLM도 마찬가지?

→ 검색기반 LLMs의 개선에 특화 할수 있을 가능성?

(현재 Task별 Learning 방법이 가장 현명한 방법으로 생각됨)



HAPPY NEW YEAR!



HAPPY NEW YEAR