Genuine Knowledge Prediction for Commonsense Knowledge Transfer

Jaewook Lee[1○], Jaehyung Seo[1], Dahyun Jung[1], Gyeongmin Kim[2*]
Department of Computer Science and Engineering, Korea University[1], Human-inspired AI Research[2]

Abstract

Compared to traditional pre-trained language models, Large Language Models (LLMs) trained on a large corpus of parameters seem to produce more natural sentences and even perform better on challenging tasks such as commonsense reasoning with a good base of human-instruction. In contrast, pre-trained language models have been shown to have difficulty acquiring enough implicit commonsense knowledge through self-supervision alone. In this work, we propose genuine commonsense prediction (GCP), a novel multi-choice dataset for language models to effectively learn commonsense knowledge. We utilize a large language model to generate non-commonsense knowledge and design a task to distinguish genuine knowledge from the presented choices so that the language model can be trained to distinguish between them. We validate our proposed dataset by performing the task on large langauge models, and asking for its interpretation.

Keywords: Commonsense, Dataest, Counterfactual

1. Introduction

Recent deep learning-based natural language generation research has been actively developing in the direction of improving the commonsense reasoning of language models. Representative natural language generation-based commonsense reasoning datasets include SWAG [1], GLUCOSE [2], and CommonGen [3]. However, most of the existing datasets deal with the generation of sentences that conform to commonsense, which limits their ability to accurately distinguish and interpret non-commonsense knowledge. To overcome this limitation, a study has emerged that deals with the discrimination and generation of non-commonsense knowledge by utilizing the triple of ConceptNet [4] instead of only dealing with commonsense content. We propose a new task, Genuine Commonsense Prediction (GCP), that goes beyond simply dealing with negative commonsense knowledge and utilizes the powerful inference capabilities of large language models and natural language generation to generate counterfactual knowledge and distinguish it from genuine commonsense knowledge.

The Genuine Commonsense Prediction (GCP) dataset proposed in this paper is generated in the following way. First, we select triples such that the *head node* is unique by *relation tag* in one of the representative commonsense knowledge graphs, **ATOMIC** [5]. Then, the Large Language Model is used to generate counterfactual knowledge corresponding to the selected triples. The language model is trained to distinguish between the generated counterfactual knowledge and the original triples so that it can efficiently learn commonsense knowledge.

2. Related Works

We utilize the commonsense knowledge graph (CKG) to build counterfactual knowledge. A commonsense knowledge graph is a structured representation of commonsense, the knowledge agreed upon by members of a society, extracted from natural language text. Representative research on CKGs has been centered on **ConceptNet**, which is built as a triple of (*head node*, *relation label*, *tail node*) to represent relationship information that conforms to general common sense based on semantics, and **ATOMIC**, which is built as a triple to represent commonsense knowledge as an If-Then relationship. Based on the triples of **ATOMIC**, **COMET** trained Knowledge Graph Completion on models such as BART and T5 to automatically generate *head node* and *tail node* given a relation tag. Based on this, $\mathbf{ATOMIC}^{20}_{20}$ was

(a) The *Head*, *Relation* and *Tail* inferred from it are given in the form below:
Head --(*Relation*)-> *Tail*
Your task is to generate counterfactual(Head) such that the above relationship between the "Tail" and the generated counterfactuals never make sense.

(b) PersonX is nice to PersonY--(xIntent)-> to get close PersonY
to keep PersonY away

Figure 1. Example of input prompt for counterfactual knowledge generation.

expanded to a graph that can represent a total of 23 relationship types by accepting various relationship representations of **ConceptNet**. Attempts to train negative knowledge into the language model showed that by training it into the language model, the discrimination ability for positive knowledge can be improved [6]. Based on this, a study was conducted to probe whether LLM actually understands knowledge through a task of generating sentences containing negative knowledge from constrained keywords and a boolean question answering task. As a study to infuse commonsense into PLM, commonsense text infilling, a task designed to learn triples of CKGs similar to the process of PLM acquiring knowledge, and commonsense prediction, a multi-choice QA task to select correctly matched triples, were proposed. We propose a task called genuine commonsense prediction (GCP), which can effectively learn commonsense knowledge by utilizing counterfactual knowledge, based on previous research where the task of commonsense prediction was effectively learned on a pre-trained language model.

3. Methodology

We used GPT-4 [7] to generate counterfactual knowledge from ATOMIC, a CKG. GPT-4 is one of the most powerful large-scale language models in existence, with an excellent understanding of human-instruction and top-notch natural language generation capabilities, which allows it to effectively perform the given generation task. The model checkpoint we used is GPT-4 (gpt-4-0613), released on June 13, 2023. We first selected one triple per *head*, one per *relation tag* of $\textbf{ATOMIC}_{20}^{20}$. This helps reduce unnecessary iterations when generating counterfactual knowledge and reduces the cost of generating it. There are 97k unique heads in $\textbf{ATOMIC}_{20}^{20}$, which is about 9% of the whole $\textbf{ATOMIC}_{20}^{20}$. The unique triples of the selected $\textbf{ATOMIC}_{20}^{20}$ are inserted into the LLM by prompting it to generate counterfactual knowledge for each *relation tag* that is not related to that tag. In this case, the counterfactual knowledge can be generated for *head* and *tail* of a triple. Figure 1 shows an example of an input prompt. As an output of LLM, the generated

	Head	Relation	Tail
A.	PersonX loves dogs	xWant	to adopt one
B.	PersonX loves chocolate	xWant	to adopt one
C.	PersonX loves dogs	xWant	to work hard
D.	PersonX loves dogs	xIntent	to adopt one

output: A

Figure 2. Example of Genuine Commonsense Prediction

counterfactual knowledge consists of a genuine knowledge triple with multi choice QA. As depicted in figure 2, we are given the following four choices

- *genuine*: Original triples selected from ATOMIC
- *fake head*: Replace *head node* to counterfactual head
- *fake tail*: Replace *tail node* to counterfactual tail
- *fake relation*: Replace *relation tag* with any other tag

4. Experiment

To validate the GCP proposed in this paper, we randomly selected 100 samples from the GCP dataset and performed a zero-shot task in which GPT-3.5 [7] and GPT-4 solved multi-choice QA questions and interpreted whether each choice was correct or incorrect. The evaluation of the interpretations was expressed as a percentage of whether the human annotator could accept or reject GPT-4's interpretation of each choice. As shown in Table 1, both models are quite good at distinguishing between three types of counterfactuals and one type of genuine knowledge. However, GPT-4 does a good enough job of explaining why the counterfactual knowledge is wrong to be acceptable, whereas GPT-3.5's interpretation of each prophecy is less acceptable.

	Accuracy	Accept Rate
GPT-3.5	87/100	43%
GPT-4	98/100	96%

Table 1. Performance for Genuine Knowledge Prediction on GPT-3.5 and GPT-4

5. Conclusion

In this study, we proposed a task to generate counterfactual knowledge from CKG through LLM and distinguish it from the original triple as a new method for language models to learn commonsense knowledge.However, although the proposed task can be utilized as a method for language models to learn commonsense knowledge, it needs to be verified by applying it to language models in the future and checking its effectiveness through benchmarks.

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information Communications Technology Planning & Evaluation). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

Reference

[1] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference," *EMNLP*, 2018.

[2] N. Mostafazadeh, A. Kalyanpur, L. Moon, D. Buchanan, L. Berkowitz, O. Biran, and J. Chu-Carroll, "Glucose: Generalized and contextualized story explanations," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4569–4586, 2020.

[3] B. Y. Lin, W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, and X. Ren, "Commongen: A constrained text generation challenge for generative commonsense reasoning," *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1823–1840, 2020.

[4] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, Vol. 22, No. 4, pp. 211–226, 2004.

[5] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, No. 01, pp. 3027–3035, 2019.

[6] J. Chen, W. Shi, Z. Fu, S. Cheng, L. Li, and Y. Xiao, "Say what you mean! large language models speak too positively about negative commonsense knowledge," *arXiv preprint arXiv:2305.05976*, 2023.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.