Go to **ACL ARR 2024 February** homepage (/group?id=aclweb.org/ACL/ARR/2024/February)

Episodic Memory Retrieval from LLMs: A Neuromorphic Mechanism to Generate Commonsense Counterfactuals for Relation Extraction

(/pdf?id=rnWE63o2MN)

Anonymous

17 Feb 2024 ACL ARR 2024 February Blind Submission Readers: February, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers, Paper1140 Authors Show Revisions (/revisions?id=rnWE63o2MN)

Abstract: Large language models (LLMs) have achieved satisfactory performance in counterfactual generation. However, confined by the stochastic generation process of LLMs, there often are misalignments between LLMs and humans which hinder LLMs from handling complex tasks like relation extraction. As a result, LLMs may generate commonsense-violated counterfactuals like `eggs were produced by a box'. To bridge this gap, we propose to mimick the episodic memory retrieval, the working mechanism of human hippocampus, to align LLMs' generation process with that of humans. In this way, LLMs can derive experience from their extensive memory, which keeps in line with the way humans gain commonsense. We then implement two central functions in the hippocampus, i.e., pattern separation and pattern completion, to retrieve the episodic memory from LLMs and generate commonsense counterfactuals for relation extraction. Experimental results demonstrate the improvements of our framework over existing methods in terms of the quality of counterfactuals.

Paper Type: long

Research Area: Information Extraction

Contribution Types: NLP engineering experiment, Data analysis

Languages Studied: English

Revealed to Xin Miao, Yongqi Li, Shen Zhou, Tieyun Qian

15 Feb 2024 (modified: 16 Feb 2024) ACL ARR 2024 February Submission

Authors: *Xin Miao (/profile?id=~Xin_Miao4), Yongqi Li (/profile?id=~Yongqi_Li3), Shen Zhou (/profile?id=~Shen_Zhou2), Tieyun Qian (/profile?id=~Tieyun_Qian1)*

TL;DR: We propose a novel neuromorphic mechanism to guide LLMs to generate commonsense counterfactuals for relation extraciton.

Reassignment Request Action Editor: This is not a resubmission

Reassignment Request Reviewers: This is not a resubmission

Software: zip (/attachment?id=-QMOhgPzxv&name=software)

Data: zip (/attachment?id=-QMOhgPzxv&name=data)

Preprint: no

Preprint Status: We are considering releasing a non-anonymous preprint in the next two months (i.e., during the reviewing process).

Consent To Share Data: yes

Consent To Review: yes

Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.

A1: yes

A1 Elaboration For Yes Or No: 6

A2: no

A2 Elaboration For Yes Or No: There is no risk in RE counterfactual research.

A3: yes

A3 Elaboration For Yes Or No: 1

B: no

B1: n/a

B2: n/a

B3: n/a

B4: n/a

B5: n/a

B6: n/a

C: yes

C1: yes

C1 Elaboration For Yes Or No: 4

C2: yes

C2 Elaboration For Yes Or No: 4

C3: yes

C3 Elaboration For Yes Or No: 4

C4: yes

C4 Elaboration For Yes Or No: 4

D: yes

D1: n/a

D2: n/a

D3: n/a

D4: n/a

D5: n/a

E: yes

E1: yes

E1 Elaboration For Yes Or No: 4

Reply Type: all Author: everybody Visible To: all readers **16 Replies**

Hidden From: nobody

[−] **Meta Review of Paper1140 by Area Chair QMX5**

ACL ARR 2024 February Paper1140 Area Chair QMX5

08 Apr 2024, 14:58 ACL ARR 2024 February Paper1140 Meta Review Readers: Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Authors, Paper1140 Reviewers Submitted, Program Chairs Show Revisions (/revisions?id=h_vXvwOSGl)

Paper Summary:

This submission presents a framework of counterfactual generation for relation extraction. Inspired by the episodic memory mechanisms in neural science, the framework contains two modules: Pattern Separation to find/label the entities and Pattern Completion to generate the relational description between the entities. Based on GPT-3.5 the paper demonstrates slight improvement on three relation classification datasets (SemEval 2010 task-8, TACRED, and ACE2005) with simulated low-resource settings.

Summary Of Strengths:

- The studied topic (generating commonsense counterfactual) could be useful in low-resource scenarios for relation classification.
- This study contains human evaluation to show that the generation results are better than other baselines in terms of being more reasonable in commonsense.

Summary Of Weaknesses:

- The clarity of technical writing seems insufficient, making it significantly less readable in terms of important technical details. I have tried to read the paper by myself, but unfortunately failed to capture much more than just a high-level idea, probably because many ad-hoc terms have not been properly defined before being heavily used. Just take a few instances, "causal term" is not a standard terminology at all for relation extraction. Descriptions around Eq 6 and 7 on how PS & PC were implemented only contain vague verbs such as "decomposes" and "combines", without instantiated explanation. There's also no description on what exactly correspond to the percentages in the head of Table 2 & 4, although some experienced readers might have an educated guess. Probably this non-trivial issue is inherited from previous work that this paper closely follows (see e.g. review comments recorded here (https://openreview.net/forum?id=fi90p5364y)). I would suggest the authors thoroughly rewritten & proofread the technical descriptions from main methodology to experimental discussion.

- Current experimental results could only support relatively weak conclusions. The shown performance improvement seems very marginal. Moreover, the current experiments are only conducted on simulated low-resource settings. In order to verify the true utility of the presented framework, additional experiments on more realistic low-resource relation classification (e.g., in specific domains, or in low-resource languages) are needed.
- The use of proprietary LLMs (GPT 3.5) makes this work difficult to reproduce. As a result, one reviewer has suggested conducting the experiments using open-weights LLMs, which makes sense to me.
- There are also a number of new details & new results appearing in the discussion. Some of them should be included in the next version of this draft.

Overall Assessment: 3 = There are major points that may be revised

Best Paper Ae: No

Needs Ethics Review: No

Information Regarding The New ACL Policy On Deanonymized Preprints: I confirm I have read the information above about changes to the anonymity policy.

[−]

Clarification Notice on the Different Expertise and Misunderstanding of Area Chair QMX5

ACL ARR 2024 February Paper1140 Authors Xin Miao (/profile?id=~Xin_Miao4) (privately revealed to you)

13 Apr 2024, 22:33 (modified: 19 Apr 2024, 17:18) ACL ARR 2024 February Paper1140 Author-Editors Confidential Comment Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Authors Show Revisions (/revisions?id=rHuAv0UUIL)

Comment:
Dear Area Chair,

Thanks for your time and effort in reading and meta reviewing our work. We would like to clarify your misunderstanding about our work.

Firstly, regarding your concerns on technical writing, we guess this might be caused by your different expertise and careless reading.

1) As Reviewer xkhh and Reviewer iyMB pointed out, '*the paper is well written and claims are supported by empirical findings*', '*the paper is clearly written and generally easy to follow*', respectively. Hence your mention that '*the clarity of technical writing seems insufficient, making it significantly less readable in terms of important technical details*' is NOT TRUE, which can be due to your different expertise.

2) Your mention that '*causal term is not a standard terminology at all for relation extraction*' and '*this non-trivial issue is inherited from previous work that this paper closely follows [1]*' is NOT TRUE, which can be due to your different expertise and careless reading.

As far as we know, previous studies [1][2] using the terminology 'causal term' have drawn attentions from prestige research groups including **Yoshua Bengio [R1]**, **Julian McAuley [R2]**, **Le Sun [R3]**, **Yulan He [R4]**, and **Shafiq Joty [R5]**. Please note that none of these references has ever posed the similar question like yours. It is quite common to use the term 'causal' to describe the determining words in relevant causal analysis works [3][4]. Additionally, we have explicitly defined the causal term as 'causal term c_i which determines the state between entities' in Line 161, and provided a detailed example in Figure 1.

3) Your mention that '*descriptions around Eq 6 and 7 on how PS & PC were implemented only contain vague verbs such as "decomposes" and "combines", without instantiated explanation*' is NOT TRUE, which can be due to your careless reading.

We have emphasized multiple times that we employ in-context learning to guide LLM to conduct PS & PC processes (Line 316, 342). This means that the example illustrated in Figure 3 (b) entirely serves as an in-context instantiated explanation for LLM. In other words, the entire example is used to guide LLM to mimic the execution of PS & PC processes. The complete prompt content is detailed in Table 9 of the Appendix. Therefore, we do not only use vague verbs to implement PS & PC without instantiated explanation.

4) Your mention that '*there's also no description on what exactly correspond to the percentages in the head of Table 2 & 4, although some experienced readers might have an educated guess*' is NOT TRUE, which can be due to your careless reading.

Due to the space limit, the detailed specifics of the dataset are in the Appendix A.1, and we have marked this in the main text (Line 413).

Secondly, your mention that '*the shown performance improvement seems very marginal*'. We have explained in our rebuttal to Reviewer BbhY that the effectiveness of counterfactually augmented data (CAD) needs to be indirectly reflected through the backbone (BERT or RoBERTa), and thus the impact of CAD is reduced. Existing CAD methods [1][2][5][6] also encounter similar situations. Our improvement is substantial compared with them. Therefore, it is reasonable that the improvement by our PSPC mechanism is not as dramatic as the methods directly guiding LLM to perform certain tasks. More importantly, we believe that this concern from the Reviewer BbhY has been addressed, because he/she has raised the overall and soundness scores.

Thirdly, regarding the experiments on open-weights LLMs, the corresponding results have been provided in the rebuttal to Reviewer BbhY. The results demonstrate that the trend is consistent with the results in the original submission, and conducting further experiments would only be incremental. Additionally, the version effectiveness analysis presented in the appendix (Line 1142) can also demonstrate that our PSPC mechanism remains effective across different models.

Finally, your mention that '*the current experiments are only conducted on simulated low-resource settings*'. We need to emphasize that we conducted both low-resouce (on SemEval, TACRED) and out-of-domain (on ACE2005) settings, which are constructed to accurately test the effectiveness of counterfactuals under out-of-distribution (OOD) scenarios. Many relevant studies [3][4][7][8] have demonstrated that the effects of counterfactuals cannot be properly validated in typical independent and identically distributed (i.i.d.) scenarios. Furthermore, such OOD scenarios are closer to reality than typical i.i.d. scenarios. We have explained the specific reasons for this in Appendix A.1, and we follow such OOD setups with previous works [2] [5].

In summary, the writing and contributions of our paper are clear, and the experiments are extensive. Therefore, the revisions needed are minor, which are consistent with the four reviewers' overall assessment.

Paper1140 Authors

[−]

Rerefrence Supplement

ACL ARR 2024 February Paper1140 Authors Xin Miao (/profile?id=~Xin_Miao4) (privately revealed to you)

13 Apr 2024, 22:37 ACL ARR 2024 February Paper1140 Author-Editors

Confidential Comment Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Authors Show Revisions (/revisions?id=MCUajglZFa3)

Comment:

Rerefrence:

[1] Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, Tieyun Qian: Large Language Models as Counterfactual Generator: Strengths and Weaknesses. CoRR abs/2305.14791 (2023)

[2] Xin Miao, Yongqi Li, Tieyun Qian: Generating Commonsense Counterfactuals for Stable Relation Extraction. EMNLP 2023: 5654-5668

References to [1][2]

[R1] Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, Yoshua Bengio. Efficient Causal Graph Discovery Using Large Language Models. arXiv:2402.01207.

[R2] Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, Furong Huang. Large language models and causal inference in collaboration: A comprehensive survey. arXiv:2403.09606

[R3] Xinlin Peng, Ying Zhou, Ben He, Le Sun, Yingfei Sun. Hidding the Ghostwriters: An Adversarial Evaluation of AI-Generated Student Essay Detection. arXiv:2402.00412

[R4] Zhang Y, Li P, Lai Y, et al. Large, Small or Both: A Novel Data Augmentation Framework Based on Language Models for Debiasing Opinion Summarization. arXiv preprint arXiv:2403.07693, 2024.

[R5] Ding B, Qin C, Zhao R, et al. Data augmentation using llms: Data perspectives, learning paradigms and challenges. arXiv preprint arXiv:2403.02990, 2024.

[3] Kaushik D, Hovy E, Lipton Z C. Learning the difference that makes a difference with counterfactually-augmented data. arXiv preprint arXiv:1909.12434, 2019.

[4] Wang Z, Culotta A. Robustness to spurious correlations in text classification via automatically generated counterfactuals. AAAI, 2021, 35(16): 14024-14031.

[5] Zhang M, Qian T, Zhang T, et al. Towards Model Robustness: Generating Contextual Counterfactuals for Entities in Relation Extraction. The ACM Web Conference 2023, 1832-1842.

[6] Wen J, Zhu Y, Zhang J, et al. AutoCAD: Automatically Generate Counterfactuals for Mitigating Shortcut Learning. Findings of EMNLP 2022: 2302-2317.

[7] Wu S, He Y. Enriching pre-trained language model with entity information for relation classification. CIKM

2019: 2361-2364.
[8] Geva M, Wolfson T, Berant J. Break, perturb, build: Automatic perturbation of reasoning paths through question decomposition. TACL, 2022, 10: 111-126.

[−]
Official Review of Paper1140 by Reviewer FJB2

ACL ARR 2024 February Paper1140 Reviewer FJB2

22 Mar 2024, 21:09 ACL ARR 2024 February Paper1140 Official Review Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=st4mYMtqRv0)

Recommended Process Of Reviewing: I have read the instructions above

Paper Summary:

The paper proposes a counterfactual-based data augmentation method for relation extraction tasks. Based on the LLMs' In-context Learning and Chain-of-Thought ability, they divide the counterfactual data generation into two parts: get the properties of entities and pair the attributes of two entities to get new expressions with new relations.

Summary Of Strengths:

1. Motivated by aligning the LLMs' relation discovery process with humans, they propose a commonsense counterfactual data generation method to augment the training data for relation extraction.
2. They design a task-specific workflow (PCPS) for counterfactual data generation which utilizes the properties of entities and pairs the properties.

Summary Of Weaknesses:

1. There is a lack of analysis of the impact of different relationship relation types. For example, the relations in TACRED are far from SemEval. The former is more about the relation between person-place or person-organization, while the latter is common sense in life. Whether it behave very differently in different relations?
2. The introduction of counterfactual data is interesting, however, the experiments only demonstrate the effectiveness of data augmentation. I think it might be more novel if we could explain that counterfactual data can solve some problems in RE, such as relation bias for entities. Previous research has shown that good performance can also be obtained with two entities as input (which ignores the context).
3. Performance improvements are relatively limited. Especially on ReTACRED, the overall baseline performance in low resources is among 15-40 F1, and the method in this paper only improves by less than 2 points.

Comments, Suggestions And Typos:

Have you tried adapting your method for LLM-RE (directly predict the relation via LLM)? For example, let LLM generate the attributes of the entity, then summarize the causal terms of the input sentence, and finally output the relation type.

Soundness: 3.5

Overall Assessment: 3.5

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Best Paper: No

Ethical Concerns:

None

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: FJB2

[−]
Response to Reviewer FJB2

ACL ARR 2024 February Paper1140 Authors Xin Miao (/profile?id=~Xin_Miao4) (privately revealed to you)

30 Mar 2024, 03:02 (modified: 30 Mar 2024, 12:14) ACL ARR 2024 February Paper1140 Official Comment Readers: Program Chairs, Paper1140 Senior Area

Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors   Show Revisions (/revisions?id=n0LKEW9SpV)

Comment:
Dear Reviewer,

Thank you very much for your recognition and valuable comments. For your questions and suggestions, we provide the following responses.

Question: There is a lack of analysis of the impact of different relationship relation types. For example, the relations in TACRED are far from SemEval. The former is more about the relation between person-place or person-organization, while the latter is common sense in life. Whether it behave very differently in different relations?
Response: Indeed, our PSPC behaves consistently in terms of guiding LLM to avoid commonsense issues. Firstly, the results in Table 2 of our paper demonstrate the effectiveness of PSPC across different types of datasets. Secondly, we also provide an example from TACRED for illustration, as shown in Table 1. Through the counterfactual of CF-CoT, we can observe that commonsense issues also arise in this example from TACRED i.e., an organization cannot be a marriage partner, but can be as a job title. We believe the reason is that LLM is misled by the word ‘People’ within the organization’s name. Meanwhile, the decomposition of entity attributes in the PSPC mechanism can effectively alleviate this issue.

Method	Sentence	Relation
Original	<e1> Enkhbold </e1> decided to step down for the chairman of the ruling <e2> Mongolian People 's Revolutionary Party </e2>	per:employee_of
CF-CoT	<e1> Enkhbold </e1> is married to <e2> Mongolian People 's Revolutionary Party </e2>	per:spouse
PSPC	<e1> Enkhbold </e1> holds the position of the ruling <e2> Mongolian People 's Revolutionary Party </e2>	per:title

Table 1: An instance from TACRED. CF-CoT represents the standard counterfacutal generation pipeline. The only difference between CF-COT and ours is the absence of the PSPC mechanism.

Question: Performance improvements are relatively limited. Especially on ReTACRED, the overall baseline performance in low resources is among 15-40 F1, and the method in this paper only improves by less than 2 points.
Response: The effectiveness of counterfactual data augmentation (CAD) needs to be indirectly reflected through the base models, and thus the impact of CAD is reduced. Existing CAD methods [1-3] also encounter similar situations. Therefore, it is reasonable that the improvement by our PSPC mechanism is not as dramatic as the methods directly guiding large models to perform certain tasks.

Suggestion: The introduction of counterfactual data is interesting, however, the experiments only demonstrate the effectiveness of data augmentation. I think it might be more novel if we could explain that counterfactual data can solve some problems in RE, such as relation bias for entities. Previous research has shown that good performance can also be obtained with two entities as input (which ignores the context).
Response: To verify whether counterfactual data augmentation can mitigate entity bias issues in the base model, we conduct the following manual validation. We first pick out instances where the base models (BERT and RoBERTa) initially make mistakes but perform correctly after counterfactual data augmentation. Then, we randomly select 100 instances from these cases for manual annotation. We define instances where entities are semantically related to false predicted relations as entity bias instances, as shown in Table 2. In the first example, due to the semantic correlation between the entities ‘bottle’ and ‘container’ with the relation ‘content-container’, the model assigns the relation as ‘content-container’ without considering the context. According to our statistics, among all error corrections, the correction of entity bias issues accounts for 43% in BERT and 42% in RoBERTa. This demonstrates that entity bias is a major challenge faced by relation classification models, and our generated counterfactuals are effective in mitigating this issue.

Sentence	False Prediction	True Relation
PET <e1> bottle </e1> was put inside of a metal <e2> container </e2>	content-container	entity-destination
The <e1> project </e1> uses <e2> art </e2> as an instrument	message-topic	agency-instrument
She pulled out an <e1> atlas </e1> from the <e2> cabinet </e2>	content-container	entity-origin

Table 2: Entity bias instances where our counterfactuals have successfully conducted correction.

Suggestion: Have you tried adapting your method for LLM-RE (directly predict the relation via LLM)? For example, let LLM generate the attributes of the entity, then summarize the causal terms of the input sentence, and finally output the relation type.

Response: We wish to give feedback as soon as possible and thus do not provide results due to the time constraint, we will do our best to supplement relevant experiments before the deadline of rebuttal.

[−] **Experiemnt Supplement**

ACL ARR 2024 February Paper1140 Authors Xin Miao (/profile?id=~Xin_Miao4) (privately revealed to you)

01 Apr 2024, 14:50 ACL ARR 2024 February Paper1140 Official Comment Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=RnLVIxsFziE)

Comment:

Dear Reviewer,

Thanks for your patience. We finally got the results for LLM-RE per your suggestion. Below is the response.

Suggestion: Have you tried adapting your method for LLM-RE (directly predict the relation via LLM)? For example, let LLM generate the attributes of the entity, then summarize the causal terms of the input sentence, and finally output the relation type.
Response: We conduct corresponding validation experiments based on your suggestion. The example for in-context learning is as follows, where it first guides LLM to generate the attributes of the entities and summarizes the state between them, then discovers relevant causal terms in the sentence, and finally outputs the relation.
S1**: **gpt-3.5-turbo-0613 baseline with entity decomposition
***Input**: <e1> eggs </e1> are moved into a <e2> box </e2>*
***Entities**: eggs-box*
***Extract entity relation**: Based on the commonsense, we know that "eggs" are small and fragile items, while "box" is an item with storage space. Based on the attributes of the entities and the semantics of the sentence, the state of the entities can be summarized as: small fragile items are placed inside storage space. The causal context words in the sentence that best reflect this state are "moved into". Based on the attributes of the entities and the causal context words, the "eggs" is "entity", the "box" is "destination", hence the most appropriate relation is entity-destination.*
***Output**: entity-destination*

The contrast example is as follows, which directly requires LLM to find the relevant causal term in the sentence, and then outputs the relation.
S2**: **gpt-3.5-turbo-0613 baseline without entity decomposition
***Input**: <e1> eggs </e1> are moved into a <e2> box </e2>*
***Entities**: eggs-box*
***Extract entity relation**: Based on the semantics of entities and causal context words "moved into", "eggs" is "entity", "box" is "destination", hence the most appropriate relation is entity-destination.*
***Output**: entity-destination*

The experiments are conducted on the test set of SemEval, which comprises 2717 instances with 19 types of relations. The LLM used for testing is *gpt-3.5-turbo-0613*. Guided by our approach, the accuracy of ***S1*** (*gpt-3.5-turbo-0613 baseline with entity decomposition*) reaches 36.22%, showing a 4.71% improvement over ***S2*** (*gpt-3.5-turbo-0613 baseline without entity decomposition*). This clearly demonstrates that our entity decomposition strategy has a positive impact on LLM-RE. However, the results of S1 are still significantly worse than that (66.91%) of our PSPC in the 3% training data. We believe the reason can be due to the lack of fine-tuning of the backbone method and the counterfactual data augmentation.

After analyzing the examples, we find that through decomposing and summarizing entity attributes, LLM can more accurately determine causal terms and entity relations. In the example below, the relation between entities is mainly determined by the meanings of the entities. However, the contrast method ***S2*** encounters the illusion problem by generating non-existent causal words and ultimately producing the incorrect relation.
***Input**: Roni Music is a <e1> software </e1><e2> company </e2> with an emphasis on powerful music software.*
***Entities**: software-company*
***Extract entity relation**: Based on the semantics of entities and causal context words "working for", "software" is "component", "company" is "whole", hence the most appropriate relation is component-whole.*
***Output**: component-whole*

In contrast, with summarization (***S1***), LLM knows that there are no explicit causal words in the sentence, and it generates the correct entity relation based on entity attributes.
***Input**: Roni Music is a <e1> software </e1><e2> company </e2> with an emphasis on powerful music software.*
***Entities**: software-company*
***Extract entity relation**: Based on our commonsense, we know that "software " is a program item, while "company"

is a manufacturer item. Based on the attributes of the entities and the semantics of the sentence, the state of the entities can be summarized as: a manufacturer that specializes in program related to music. There is no explicit causal context words in the sentence that indicate a specific relation between the entities. Based on the attributes of the entities, the "software" is "product", the "company" is "producer", and hence the most appropriate relation is product-producer.
Output*: product-producer*

[−]
Rerefrence Supplement

ACL ARR 2024 February Paper1140 Authors Xin Miao (/profile?id=~Xin_Miao4) (privately revealed to you)

30 Mar 2024, 03:09 ACL ARR 2024 February Paper1140 Official Comment Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=BkE4e9UGDMB)

Comment:
Rerefrence:
[1] Wen J, Zhu Y, Zhang J, et al. AutoCAD: Automatically Generate Counterfactuals for Mitigating Shortcut Learning[C]//Findings of the Association for Computational Linguistics: EMNLP 2022. 2022: 2302-2317.
[2] Zhang M, Qian T, Zhang T, et al. Towards Model Robustness: Generating Contextual Counterfactuals for Entities in Relation Extraction[C]//Proceedings of the ACM Web Conference 2023. 2023: 1832-1842.
[3] Li Y, Xu M, Miao X, et al. Large language models as counterfactual generator: Strengths and weaknesses[J]. arXiv preprint arXiv:2305.14791, 2023.

[−]
Official Review of Paper1140 by Reviewer xkhh

ACL ARR 2024 February Paper1140 Reviewer xkhh

21 Mar 2024, 05:47 ACL ARR 2024 February Paper1140 Official Review Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=Ew5YJJJKWP4)

Paper Summary:
Authors propose a novel mechanism to align LLMs with humans during the experience gaining process in counterfactual generation for relation extraction. Again, authors claim that their methods (pattern separation and pattern completion) enables retrieval of entities’ scenarios from the model’s extensive memory providing factual basis for the relation between counterfactuals and entities, and support their claims by conducting experiments on 3 different ER datasets.

Summary Of Strengths:
Validation of the proposed framework with both data augmentation and human evaluation on 3 RE datasets. Consistent improvements are obtained across datasets.

Summary Of Weaknesses:
None, paper is well written and claims are supported by empirical findings.

Comments, Suggestions And Typos:
I am not sure if this is possible at all but it would help if you can improve readability of Table 2,3 and 4 (especially 2). It was a little bit hard for me to read it.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.

Overall Assessment: 4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Best Paper: No

Ethical Concerns:
None

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 1 = No usable datasets submitted.

Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: xkhh

[−] Response to Reviewer xkhh

ACL ARR 2024 February Paper1140 Authors *Xin Miao (/profile?id=~Xin_Miao4) (privately revealed to you)*

30 Mar 2024, 02:00 (modified: 30 Mar 2024, 02:04) ACL ARR 2024 February Paper1140 Official Comment Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=Ym9i2afOb7m)

Comment:

Dear Reviewer,

So many thanks for your recognition! Here is our response to your comment.

Suggestion: I am not sure if this is possible at all but it would help if you can improve readability of Table 2,3 and 4 (especially 2). It was a little bit hard for me to read it.
Response: We really appreciate and thank you for your intensive suggestion, we will find ways to improve the readability of these tables.

[−] Official Review of Paper1140 by Reviewer iyMB

ACL ARR 2024 February Paper1140 Reviewer iyMB

18 Mar 2024, 00:30 ACL ARR 2024 February Paper1140 Official Review Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=-R7dIKkR3E0)

Recommended Process Of Reviewing: I have read the instructions above

Paper Summary:
The paper describes a neuromorphic mechanism to generate commonsense-compliant counterfactuals for relation extraction. It uses a pattern separation function and a pattern completion function to mimic the process of recalling episodic memory within human hippocampus, both implemented through prompt-engineering with LLMs as in-context examples. The pattern separation step identifies properties of the involved entities, and the pattern completion step combines entity-properties to raise commonsense-compliant statements (scenarios) about the entities, where counterfactual relations are extracted from the generated statements. Experiments on various Relation Extraction datasets show the effectiveness of the proposed method.

Summary Of Strengths:

1. The method proposed in this paper is theoretically motivated and supported by experiments
2. The paper is clearly written and generally easy to follow

Summary Of Weaknesses:
The margins of improvements with the proposed method tend to decrease with rising percentage of augmented data (as shown in Table 2, Table 4), it would be good to discuss this diminishing return and its implications in the application of these counterfactuals.

Comments, Suggestions And Typos:
line 189-190: by "the causal term is proper", do you mean the textual expressions should make sense? What exactly is the difference between that and "the flipped label is reasonable for the entity pair"?

Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.

Overall Assessment: 4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it.

Confidence: 2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.

Best Paper: No

Ethical Concerns:
None

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 3 = Potentially useful: Someone might find the new datasets useful for their work.

Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: iyMB

[−]

Response to Reviewer iyMB

ACL ARR 2024 February Paper1140 Authors Xin Miao (/profile?id=~Xin_Miao4) (privately revealed to you)

30 Mar 2024, 02:16 ACL ARR 2024 February Paper1140 Official Comment Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=rTIoPup5iOt)

Comment:
Dear Reviewer,

We really appreciate and thank you for your recognition! Here are our responses to your comments.

Suggestion: The margins of improvements with the proposed method tend to decrease with rising percentage of augmented data (as shown in Table 2, Table 4), it would be good to discuss this diminishing return and its implications in the application of these counterfactuals.
Response: Actually, we have provided an explanation for this phenomenon in Appendix section A.1 (Line 772). We apologize for not detailing it in the main text due to space constraints. In brief, due to the datasets (SemEval and TACRED) satisfying the independent and identically distributed (i.i.d.) assumption, the spurious correlations in the training set and test set follow the same distribution. As the proportion of the training set increases, the overlap of spurious correlations also increases. In this situation, the spurious correlations can assist the model in finding shortcuts and improving accuracy [1]. Unfortunately, when the counterfactuals block spurious correlations, they may not help the model in terms of accuracy and could even have a counterproductive effect under such an i.i.d. situation [2-5]. We will supplement explanations for this in the main text.

Question: line 189-190: by "the causal term is proper", do you mean the textual expressions should make sense? What exactly is the difference between that and "the flipped label is reasonable for the entity pair"?
Response: The sentences generated by LLM typically make sense. The meaning of “the causal term is proper” is that the causal term should semantically match the relation. While “the flipped label is reasonable for the entity pair” means that the relation should semantically match the entities. The former corresponds to causal word replacement in the pipeline (Figure 2), while the latter corresponds to potential relation discovery. For a more intuitive explanation, we illustrate with the examples from Appendix A.2. Say the original sentence is “cheese is flowed into food banks”, with an “entity-destination” relation. (1) The causal term is improper: “cheese is donated to food banks,” with the new relation “content-container”. The replacement words “donated to” not semantically match the new relation “content-container”. (2) The flipped label is unreasonable: “cheese is contained in food banks”, with the new relation “content-container”. The new relation does not semantically match the entity pair. The issues of causal term mismatch typically arise from unreasonable potential relations, because LLM tends to generate common scenarios. In summary, the first issue typically compounds the occurrence of the second issue, introducing more noise.

Rerefrence:
[1] Sen I, Samory M, Flöck F, et al. How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs?[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 325-344.
[2] Kaushik D, Hovy E, Lipton Z. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data[C]//International Conference on Learning Representations. 2019.
[3] Sen I, Samory M, Flöck F, et al. How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs?[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 325-344.

[4] Wang Z, Culotta A. Robustness to spurious correlations in text classification via automatically generated counterfactuals[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(16): 14024-14031.
[5] Geva M, Wolfson T, Berant J. Break, perturb, build: Automatic perturbation of reasoning paths through question decomposition[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 111-126.

[−]

Official Review of Paper1140 by Reviewer BBhY

ACL ARR 2024 February Paper1140 Reviewer BBhY

11 Mar 2024, 17:53 (modified: 03 Apr 2024, 00:32) ACL ARR 2024 February Paper1140

Official Review Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=0uj76egahQ)

Recommended Process Of Reviewing: I have read the instructions above

Paper Summary:
The work provides a new way of performing entity extraction by generating common sense counterfactuals using in-context learning with LLM. They use two modules, Pattern Separation to find/label the entities and Pattern Completion to generate the relationship between the entities. Using GPT-3.5 they show slight improvement on three datasets

Summary Of Strengths:

- A new pipeline for extracting relationships by generating commonsense counterfactual that beats other baselines on three different datasets.
- Human evaluation shows that their suggested pipeline generates better common sense counterfactuals.

Summary Of Weaknesses:

- No error analysis or insights into what works and what doesn't work. Where does the pipeline fails? Table 5 provides almost no insight.
- No ablation study to see which module (Pattern Seperation/ Pattern Completion) helps the most? Or are both necessary for this task.
- No results with open-source models. It would be interesting to see if the same pipelines provides improvement for other models and not just closed-source GPT model.
- Some of the examples are not clear: L211 -- why can’t cheese be donated to a food bank?
- I understand that this might be related to the way the benchmark is created, but I would like to see some experiments without explicitly providing the entities in the prompt. The model should be able to figure out the entities from the natural language itself.
- Only slight improvement over other baselines. (minor)

Comments, Suggestions And Typos:
Please refer to the summary of weakness. Additionally,

- L127: "relies on external knowledge to keep consistency with commonsense" -- do you think this is not a good way of keeping consistent with the commonsense. What would happen if the LLMs information is outdated w.r.t. some commonsense fact? This could be a good direction to explore.
- grammar: "eggs were taken out from the box" fig 1
- It would be good to have a footnote in the abstract regarding hippocampus
- I think its more natural to write the figure part and then the description. Eg: "Fig 1 a) this is on the left and b) this is on the right " instead of Fig 1 this is on the left a) and this is on the right b).

Soundness: 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.

Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Best Paper: No

Ethical Concerns:
None

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 2 = Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)

Software: 3 = Potentially useful: Someone might find the new software useful for their work.

Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

[−]
Response to Reviewer BBhY

ACL ARR 2024 February Paper1140 Authors Xin Miao (/profile?id=~Xin_Miao4) (privately revealed to you)

01 Apr 2024, 17:28 (modified: 01 Apr 2024, 17:40) ACL ARR 2024 February Paper1140 Official Comment Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=9QRm3CX-bZB)

Comment:
Dear Reviewer,

We apologize for bothering you. Could you please check our responses? We wonder if we have addressed your concerns. If they have, we sincerely hope you could consider revising your rating. If you have any further questions, please don't hesitate to tell us, so we can make a response again before the rebuttal deadline.

So many thanks for your time and patience. We are looking forward to communicating with you further!

Best regards!

[−]
Response to Reviewer BBhY (2/2)

ACL ARR 2024 February Paper1140 Authors Tieyun Qian (/profile?id=~Tieyun_Qian1) (privately revealed to you)

30 Mar 2024, 03:58 (modified: 30 Mar 2024, 13:01) ACL ARR 2024 February Paper1140 Official Comment Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=Mcw7B05CmQy)

Comment:
Suggestion: No results with open-source models. It would be interesting to see if the same pipelines provides improvement for other models and not just closed-source GPT model.
Response: Due to time constraints, we conduct validation experiments based on open-source models Llame-2-7b and Llame-2-13b under the 3% setting, and the experimental results are shown in Table 1 below. The experimental results demonstrate that our PSPC mechanism is consistently effective for open-source large models. In the future version, we will supplement more results from open-source large models.

Method	R-BERT	R-RoBERTa
Original	59.31 (±1.46)	64.27 (±3.20)
	Llama-2-7b	
CF-CoT	61.59 (±1.20)	64.58 (±2.51)
PSPC	**62.75 (±1.38)**	**66.23 (±1.30)**
	Llama-2-13b	
CF-CoT	61.97 (±1.09)	65.69 (±2.69)
PSPC	**63.61 (±1.84)**	**66.79 (±1.67)**

Table 1: Results on open-source models for data augmentation evaluation under 3% low-resource setting on SemEval.

Suggestion: I understand that this might be related to the way the benchmark is created, but I would like to see some experiments without explicitly providing the entities in the prompt. The model should be able to figure out the entities from the natural language itself.
Response: The definition of relation extraction is: extracting the relation between given entities. Therefore, the given entities are the default setting for relation extraction and the related counterfactual data augmentation [1-3].

The task you mentioned refers to entity and relation extraction [6], which first extracts entities and then extracts the relation between them. Although this is also a possible direction, the focus of our paper, as well as counterfactual data augmentation for RE, is on counterfactual generation for relation extraction.

Suggestion: L127: "relies on external knowledge to keep consistency with commonsense" -- do you think this is not a good way of keeping consistent with the commonsense. What would happen if the LLMs information is outdated w.r.t. some commonsense fact? This could be a good direction to explore.
Response: The external knowledge is also a good way. However, what we want to emphasize in our paper is the possibility of exploring the solution to this problem through the inherent capabilities of LLM. Thank you for your suggestion, and we believe the outdated information in LLM is a direction worth exploring.

Suggestion: (1) grammar: "eggs were taken out from the box" fig 1. (2) It would be good to have a footnote in the abstract regarding hippocampus. (3) I think its more natural to write the figure part and then the description. Eg: "Fig 1 a) this is on the left and b) this is on the right " instead of Fig 1 this is on the left a) and this is on the right b).
Response: Thank you for pointing these out. We will make adjustments per your suggestions.

Rerefrence:
[1] Miao X, Li Y, Qian T. Generating Commonsense Counterfactuals for Stable Relation Extraction[C]//The 2023 Conference on Empirical Methods in Natural Language Processing. 2023.
[2] Li Y, Xu M, Miao X, et al. Large language models as counterfactual generator: Strengths and weaknesses[J]. arXiv preprint arXiv:2305.14791, 2023.
[3] Zhang M, Qian T, Zhang T, et al. Towards Model Robustness: Generating Contextual Counterfactuals for Entities in Relation Extraction[C]//Proceedings of the ACM Web Conference 2023. 2023: 1832-1842.
[4] Wen J, Zhu Y, Zhang J, et al. AutoCAD: Automatically Generate Counterfactuals for Mitigating Shortcut Learning[C]//Findings of the Association for Computational Linguistics: EMNLP 2022. 2022: 2302-2317.
[5] Kumaran D, Hassabis D, McClelland J L. What learning systems do intelligent agents need? Complementary learning systems theory updated[J]. Trends in cognitive sciences, 2016, 20(7): 512-534.
[6] Zhong Z, Chen D. A Frustratingly Easy Approach for Entity and Relation Extraction[C]//2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021. Association for Computational Linguistics (ACL), 2021: 50-61.

[−] Response to Reviewer BBhY (1/2)

ACL ARR 2024 February Paper1140 Authors Xin Miao (/profile?id=~Xin_Miao4) (privately revealed to you)

30 Mar 2024, 03:49 (modified: 01 Apr 2024, 14:52) ACL ARR 2024 February Paper1140 Official Comment Readers: Program Chairs, Paper1140 Senior Area Chairs, Paper1140 Area Chairs, Paper1140 Reviewers Submitted, Paper1140 Authors Show Revisions (/revisions?id=xoKFkeZUSvj)

Comment:
Dear Reviewer,

Thank you for your valuable suggestions. Firstly, based on your summary, we are afraid that you may have some misunderstandings about our work. Therefore, please allow us to provide a brief clarification. We primarily focus on the commonsense constraint issues in counterfactual data augmentation for relation extraction [1]. To alleviate this issue in LLM [2], we introduce PSPC, which simulates the memory retrieval mechanism in the human hippocampus for the first time. PSPC retrieves scenarios relevant to entities within LLM, enabling LLM to discover commonsense potential relations and generate reasonable counterfactuals. The generated counterfactuals serve as augmented data to assist in training downstream models (e.g. BERT or RoBERTa), rather than directly guiding LLM to conduct extraction tasks. Secondly, based on your advice and questions, we have made the following clarifications in an effort to address your concerns.

Question: No error analysis or insights into what works and what doesn't work. Where does the pipeline fails? Table 5 provides almost no insight.
Response: In Appendix Table 11, we present the detailed pipeline of an example from Table 5, with a detailed explanation provided in Appendix Section A.4 (Line 1001). In the pipeline, due to the lack of specific processing mechanisms, traditional approaches and LLM primarily make mistakes in potential relation discovery, as indicated by previous work [1, 2], which is the motivation behind our paper. For example, given instance ‘CSU Stanislaus students take <e1> complaints </e1> are placed in President's <e2> door </e2> .’ with ‘entity-destination’ relation. LLM incorrectly identifies ‘content-container’ as the potential relation, leading to the un-commonsense counterfactual ‘CSU Stanislaus students take <e1> complaints </e1> are placed in President's <e2> door </e2>’. Obviously, the ‘door’ cannot serve as a ‘container’, the entities do not match the new relation. We apologize for not explicitly stating this in the main text, and we will make the necessary additions.

Question: No ablation study to see which module (Pattern Seperation/ Pattern Completion) helps the most? Or are both necessary for this task.
Response: In fact, pattern separation and pattern completion are two complementary computational processes [5], hence they cannot be separated. However, the comparison with CF-CoT can be regarded as a comprehensive ablation study for our PSPC mechanism, since CF-CoT shares the same settings as our method except for the absence of the PSPC mechanism.

Question: Some of the examples are not clear: L211 -- why can't cheese be donated to a food bank?
Response: In this example, there is nothing wrong with the semantics of the sentence 'cheese is donated to food banks', but rather a mismatch between the entities and their relation 'content-container'. We typically do not consider 'food banks' as 'container'. In the definition of relation extraction, a complete instance consists of a sentence containing given entities and the relation between them. Therefore, we must ensure the correctness of relation labels. The commonsense constraint issues we emphasize mainly refer to the rationality between entities and their relations. More explanations about this are provided in Section 3.1.

Question: Only slight improvement over other baselines. (minor)
Response: The effectiveness of counterfactually augmented data (CAD) needs to be indirectly reflected through the base models, and thus the impact of CAD is reduced. Existing CAD methods [1-4] also encounter similar situations. Therefore, it is reasonable that the improvement by our PSPC mechanism is not as dramatic as the methods directly guiding LLM to perform certain tasks. More importantly, the comparison with CF-CoT demonstrates the significant effectiveness of our PSPC mechanism and underscores the immense potential of LLM in counterfactual-related tasks.

[−]
Supplementary Materials by Program Chairs

ACL ARR 2024 February Program Chairs

17 Feb 2024, 05:45 ACL ARR 2024 February Paper1140 Supplementary Materials Readers: Program Chairs, Paper1140 Reviewers, Paper1140 Authors, Paper1140 Area Chairs, Paper1140 Senior Area Chairs Show Revisions (/revisions?id=0VEBrIdAYD)

Software: zip (/attachment?id=0VEBrIdAYD&name=software)

Data: zip (/attachment?id=0VEBrIdAYD&name=data)

Reassignment Request Action Editor: This is not a resubmission

Reassignment Request Reviewers: This is not a resubmission

A1: yes

A1 Elaboration For Yes Or No: 6

A2: no

A2 Elaboration For Yes Or No: There is no risk in RE counterfactual research.

A3: yes

A3 Elaboration For Yes Or No: 1

B: no

B1: n/a

B2: n/a

B3: n/a

B4: n/a

B5: n/a

B6: n/a

C: yes

C1: yes

C1 Elaboration For Yes Or No: 4

C2: yes

C2 Elaboration For Yes Or No: 4

C3: yes

C3 Elaboration For Yes Or No: 4

C4: yes

C4 Elaboration For Yes Or No: 4

D: yes

D1: n/a

D2: n/a

D3: n/a

D4: n/a

D5: n/a

E: yes

E1: yes

E1 Elaboration For Yes Or No: 4

Note From EiCs: These are the confidential supplementary materials of the submission. If you see no entries in this comment, this means there haven't been submitted any.