Knowledge Graph Construction by Triple and Event Extraction using Few-Shot CoT Techniques

MohammadJavad Ardestani
Ardestan@ualberta.ca
University of Alberta
Edmonton, CA

Jai Riley
jai.riley@ualberta.ca
University of Alberta
Edmonton, CA

1 Problem

Two problems have been recognized for tasks involved in knowlegde graph (KG) construction: over-fitting of large language models (LLMs) on KG data to achieve State-of-the-art (SOTA) performance and large performance gaps between zero-shot/one-shot models compared to SOTA models fine-tuned on the individual tasks. To address these problems our approach aims to leverage the robust reasoning capabilities of LLMs to enhance triple/event extraction for KG construction. We want to do this through few-shot Chain of Thought (CoT) prompting techniques with different LLMs.

Employing the CoT prompting techniques are a pivotal strategy to augment a model's reasoning capabilities in KG construction. CoT involves embedding demonstrations within prompts to guide the LLMs to generate better intermediate reasoning steps and results (Wei et al., 2022).

We hypothesize the use of CoT will enable the use of generalized methods in KG constuction tasks without extensive fine-tuning models for individual tasks. Our objective is to facilitate even small LLMs using few-shot CoT prompting to outperform larger ones in our given tasks. This approach offers efficiency and accessibility, benefiting users with limited access to high-resource LLMs or extensive datasets.

2 State-of-the-art

For our problem we will be using two datasets for triple extraction, Re-TACRED (Stoica et al., 2021) and SciERC (Luan et al., 2018), and one dataset, MAVEN (Wang et al., 2020), for event extraction.

2.1 Re-TACRED SOTA

(Park and Kim, 2021) used a Curriculum learning approach in which the difficulty of the samples was determined by a cross review method based on the predictions of a large RoBERTa model. Once they determined the best ordering of the samples to feed into the RoBERTa model, they were able to achieve 91.4 F1-score on the Re-TACRED dataset using this approach.

2.2 SciERC SOTA

(Ye et al., 2022) uses a packed levitated marker (PL-Marker) model with the SciBERT (Beltagy et al., 2019) encoder for triple extraction on the SciERC dataset. They received a SOTA F1-score of 53.2 for this approach.

2.3 MAVEN SOTA

(Wang et al., 2022) experimented with different types of prompting (continuous soft-prompting, APEX prompting) under various settings such as supervised, few-shot, and zero-shot event detection. Using a bert-base-uncased model they were able to achieve SOTA for the MAVEN dataset with APEX prompting under supervised conditions with an F1-score of 68.8.

2.4 Baseline

(Zhu et al., 2023) conducted experiments involving zero-shot and one-shot prompts for ext-davinci-003, ChatGPT, and GPT-4 (Achiam et al., 2023) on triple and event extraction using the previously mentioned datasets. Their results indicate GPT-4 outperforms other LLMs, with its best F1-scores being 22.5, 30.4, and 9.1 for Re-TACRED, MAVEN, and SciERC, respectively.

3 Draft Proposal

3.1 Artifacts

For triple extraction, our model aims to extract triples (head, relation, tail) from a given data sample. We will craft a suitable prompt for our LLM based on this data in which we anticipate the LLM to extract the target triple from the given input. Consider the following sentence:

> A verdict on the case of Wen Qiang, former justice department director in the southwestern city of Chongqing, would be announced at an undisclosed location.

This sentence will be incorporated into a well-defined prompt given to the LLM. We expect the LLM's response to include the desired triple: [Wen Qiang, title, director].

For event extraction, we will use a similar approach to create a suitable prompt for the task of extracting all events from a given sentence. For example consider the sentence:

> Unprepared for the attack, the Swedish attempted to save their ships by cutting their anchor ropes and to flee.

We expected the output to be: Removing, Rescuing, Escaping, Attack, Self_motion.

3.2 Methods

To implement CoT for triple extraction, we propose the following steps: (1) extract the entities from the given sentence, (2) identify potential relations between each pair of extracted entities, considering different head and tail combinations for each pair, (3) compare the predicted relation for each pair to the provided ground truths (including the option of "no-relation") to pinpoint its closest match, and (4) return the triples in the exact format of [head, relation, tail], separating them with commas.

For event extraction we will implement a similar set of steps: (1) analyze and predict possible events from the given sentence, (2) compare predicted events to the ground truths and pinpoint their closest match, and (3) return a list of the events in the format of [event, event,...].

We intend to design prompts that guide the model through these steps rather than simply asking it to extract a triple/event. By providing samples that illustrate the intermediate results of this process, we aim to facilitate the model's reasoning through a few-shot CoT approach.

Our experiments will entail applying the CoT prompting technique to LLMs of varying sizes, capabilities and training sources, including GPT-3.5, LLaMA (7B and 13B), Gemini, and GPT-4. Potential limitations include large expenses associated with prompting high-resource LLMs, especially GPT-3.5 and GPT-4. To address this, we plan to conduct experiments on smaller portions of our datasets to evaluate our assumptions, seeking resources to overcome any limitations encountered.

3.3 Available code

All the code involved in preprocessing the datasets and prompt creation is available through the Github provided by (Zhu et al., 2023). We have not made an attempt to run the code as it will be used as a guideline to accomplish our tasks. We plan to customize the prompt creation process, develop API calls for prompting the LLMs and design the evaluation components from scratch.

3.4 Available data

We have downloaded the Re-TACRED, SciERC, MAVEN datasets (both training and testing data) including example prompts for zero-shot and one-shot [1]. The test/training samples appear to be identical, only differing in the size of each set. All data points in the triple extraction datasets (Re-TACRED, SciERC) consist of a sentence and it's related triple. The data points in the MAVEN dataset for event extraction consist of a sentence and a list of events contained in the sentence.

3.5 Preprocessing

Since the code for preprocessing our datasets is provided through the previously mentioned Github, only slight modification of this code will need to take place for our modified prompting technique. This will severely decrease the amount of time needed to create our preprocessing guidelines, making most of the time reliant on how much of each dataset we want to use.

4 Draft Evaluation Protocol

A common metric used for the evaluation of our tasks is F1-score. We will use this metric to compare our method's extracted triples and events against the given SOTA benchmarks for the datasets as well as the zero-shot and one-shot approaches performed by (Zhu et al., 2023). Since the given datasets have been used as benchmarks for their respective tasks, an attempt will be made to use all testing data for evaluation. In the case where the model outputs more than one triple or multiple events that are not listed in the gold, a correct prediction will be granted if the output contains the gold triple/event.

5 Repository URL

`https://github.com/jai-riley/CMPUT656_Project`

[1] `https://github.com/zjunlp/AutoKG/tree/main`

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232. Association for Computational Linguistics, Brussels, Belgium.

Seongsik Park and Harksoo Kim. 2021. Improving sentence-level relation extraction through curriculum learning.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-TACRED: Addressing shortcomings of the TACRED dataset.

Sijia Wang, Mo Yu, and Lifu Huang. 2022. The art of prompting: Event detection based on type specific prompts.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A massive gen- eral domain event detection dataset. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020:1652–1671.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *arXiv preprint arXiv:2305.13168*.