Large Language Models as Near-Threshold Triple Corrector for Document-Level Relation Extraction

Anonymous ACL submission

Abstract

Document-level relation extraction (DocRE) is crucial for large-scale mining of complex, cross-sentence relational knowledge, but it also presents challenges due to its multi-label nature. To address this issue, the adaptive thresholding technique for separating positive and negative triples has become a standard practice. However, by inspecting multiple advanced DocRE models, we find that they still assign scores close to the threshold for a small portion of triples, and the accuracy for these triples is significantly lower than the overall level. In this paper, we explore leveraging the exceptional in-context learning (ICL) capability of large language models (LLMs) to alleviate this performance bottleneck. We transform the binary classification problem of triples into a reading comprehension task, and propose an evidence-driven prompting method to improve LLM's predictions. Although this ICL paradigm on average underperforms fine-tuned DocRE models, we observe a better performance when restricted to near-threshold triples predicted by fine-tuned models. Based on this, we propose a simple yet effective framework that employs LLM to correct the prediction on near-threshold triples. Experiment results demonstrate that our framework achieves significant improvements across multiple representative DocRE models.

1 Introduction

Motivated by the demand for large-scale mining of complex, cross-sentence relational knowledge in real-world scenarios, document-level relation extraction (DocRE), which aims to identify the relations between each pair of entities within a document, has recently garnered increasing research interest (Tan et al., 2022a; Zhou and Lee, 2022; Xiao et al., 2022; Sun et al., 2023). Compared to sentence-level relation extraction, DocRE is more challenging due to its multi-label nature, where each entity pair may express multiple relations through different contextual clues. Additionally, the large number of entity pairs (which grows quadratically with the number of entities) and the rich types of pre-defined relations further increase the complexity of DocRE.

To better tackle the multi-label problem, existing studies typically reduce it into multiple binary classification problems and introduce a threshold class, whose score is used as an adaptive threshold to distinguish the positive and negative relations for an entity pair. Ideally, the predictions during inference should align with the training objective that the scores for positive and negative classes should be higher and lower enough[1] than the threshold. However, by inspecting three advanced DocRE models with distinct architectures or training losses, we discover that they all still assign scores close to the threshold for a small portion of triples, and the performance on these near-threshold triples is significantly lower than the overall level of each model. This common performance bottleneck reflects to some extent the inherent limitation of current approaches and reveals the potential for further improvement.

Inspired by the recent remarkable advances of large language models (LLMs) in text understanding, logical reasoning and in-context learning (ICL), we explore the possibility of pushing the performance boundary of DocRE with LLMs in this work. By expressing the candidate triple as a statement sentence via the relation verbalization template, we first reformulate the binary classification problem of triples into a reading comprehension task, utilizing the LLM to determine whether the statement can be inferred from the document. To further improve the prediction results of LLM, we propose an evidence-driven prompt-

[1] It has been shown that training the model to make "good enough" rather than "perfect" predictions by margin regularization helps avoid overfitting to the dominant negative triples (Zhou and Lee, 2022; Wei and Li, 2022; **?**).

ing method, guiding the LLM to conduct evidence retrieval rather than directly provide the answer. Through experiments, we find that although this ICL paradigm still lags behind fine-tuned DocRE models in effectiveness and is too costly if applied to all candidate triples, it demonstrates noticeable performance advantage when focused on the near-threshold triples in fine-tuned models' predictions. Based on these findings, we propose a simple yet effective framework that combines the complementary advantages of LLM and fine-tuned DocRE models in a collaborative manner. Specifically, we utilize the fine-tuned DocRE model to predict each triple in the document and calculate the margin between the predicted score and threshold score of each triple. For those near-threshold triples with small margins, we feed them into the LLM and adopt its predictions as the final output. Experiment results demonstrate that our framework achieves significant improvements across multiple representative DocRE models.

2 Methodology

Problem Formulation. Given a document D containing a set of entities $\mathcal{E} = \{e_i\}_{i=1}^{N_e}$, the task of DocRE is to predict the set of all possible relations between each entity pair $(e_h, e_t) \in \{(e_i, e_j) \mid i, j = 1, \ldots, N_e; i \neq j\}$ from a pre-defined relation type set $\mathcal{R}$. The subscripts of e_h and e_t refer to the head and tail entity in an entity pair. To handle the multi-label problem, existing DocRE models typically introduce a learnable threshold class TH. During inference, the model predicts a score $s^{(h,r,t)}$ for each triple (e_h, r, e_t) where $r \in \mathcal{R}$, and also a score $s^{(h,TH,t)}$ which serves as the threshold for all triples with head entity e_h and tail entity e_t. A triple (e_h, r, e_t) is predicted as positive if $s^{(h,r,t)} > s^{(h,TH,t)}$ and negative otherwise. We denote the margin (in absolute value) between $s^{(h,r,t)}$ and $s^{(h,TH,t)}$ as $m^{(h,r,t)} = |s^{(h,r,t)} - s^{(h,TH,t)}|$.

2.1 Analysis of Existing DocRE Models

We first conduct an analytical experiment to investigate how well the triples are separated from the threshold in existing DocRE models. Specifically, given a test set and a DocRE model, we predict the labels for all candidate triples and group the triples based on the margin $m^{(h,r,t)}$. Then we count the number of triples in each group and evaluate the per-group performance of the model. We conduct the experiment on Re-DocRED dataset (Tan et al., 2022b) and select three DocRE models: ATLOP (Zhou et al., 2021), NCRL (Zhou and Lee, 2022) and DREEAM (Ma et al., 2023). More details on Re-DocRED and three models are listed in Appendix A.

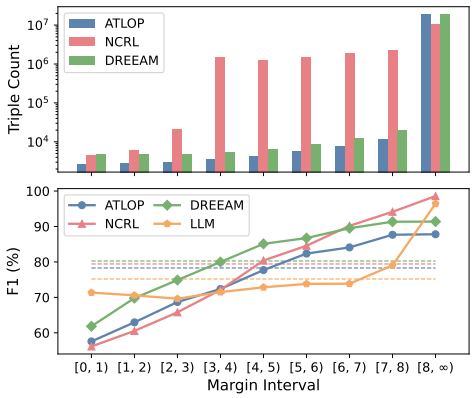

Figure 1: The number of predicted triples (in logarithmic scale) in different margin intervals and the corresponding F1 scores. For the F1 scores of LLM, the triples are grouped by the average margin across all training trials of ATLOP, NCRL and DREEAM. The dashed line represents the overall F1 on the test set.

Figure 1 shows the average results of each model across three training trials with different random seeds. We discover that although three models provide confident predictions with large margins for the vast majority of triples, they still assign scores close to the threshold for a small portion of triples. Meanwhile, the performance of each model on these near-threshold triples is significantly lower than the overall level. When the margin is less than 1, this performance gap can reach around 20% F1. This common performance bottleneck reflects partly the inherent limitation of current approaches, while on the other hand, also indicates a potential for target optimization on near-threshold triples.

2.2 Evidence-Driven ICL for DocRE

The ICL capability which learns directly from the instruction and demonstrations in contextual input has enabled LLMs to perform well on a series of tasks. We argue that the difference in learning paradigms between ICL and fine-tuned DocRE models may lead to ICL exhibiting a different predictive distribution, offering a new perspective for pushing the performance boundary of DocRE. To this end, we first reformulate the binary classification problem of triples into a reading comprehension task, as illustrated in Figure 2. The basic idea is to convert a candidate triple into a statement

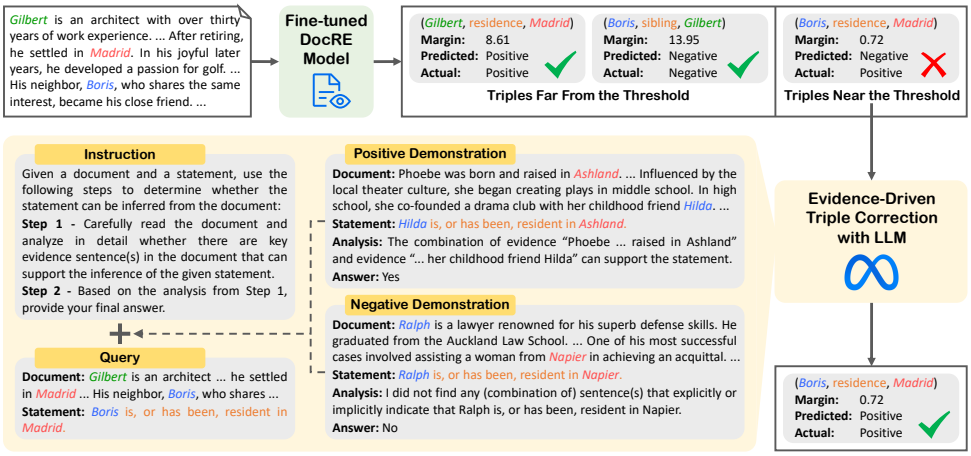

Figure 2: The proposed framework that functions LLM as a near-threshold triple corrector to collaborate with the fine-tuned DocRE model.

sentence by filling the head and tail entity into a pre-defined verbalization template of the relation. Then the input to LLM can be constructed as a concatenation of the task instruction, document and statement, where the instruction explicitly prompts LLM to determine whether the statement can be inferred from the document. Moreover, we can add positive and negative demonstrations, where the underlying triples have the same relation type as the candidate triple.

Recent studies have shown that training the DocRE model to jointly extract relation and the evidence sentences supporting the relation can enhance each other. Inspired by this, as well as the importance of generating intermediate reasoning steps to improve LLMs' reasoning ability, we further propose an evidence-driven prompting method, transferring the idea of evidence retrieval into ICL. As shown in Figure 2, instead of instructing the LLM to provide the answer directly, we prompt it to first analyze whether there are key evidence sentences that can support the inference of the statement, then derive an answer based on the analysis. In few-shot setting, we also design separate templates for positive and negative demonstrations to automatically generate the evidence-based analysis. The complete prompt template is provided in B.

2.3 LLM as Near-Threshold Triple Corrector

We evaluate the performance of the proposed evidence-driven ICL paradigm on Re-DocRED. We experiment with Llama-3.1-70B (Dubey et al., 2024) and provide five positive and negative demonstrations respectively for enhancement. For the test data, we group all candidate triples based on the average margin across all training trials of three fine-tuned models, and sample 500 positive and 500 negative triples from each group for testing. As shown in Figure 1, although this ICL paradigm has achieved an F1 over 75%, it still lags behind the fine-tuned DocRE models. Additionally, due to the vast number of candidate triples (nearly 20 million in the test set of Re-DocRED), the higher time cost (and financial cost if using proprietary LLM like GPT-4o) of LLM also makes it hard to apply to all candidate triples. Nevertheless, we find that LLM exhibits a different trend of per-group performance compared to fine-tuned DocRE models. With a certain range, the LLM does not show significant performance fluctuations. And more importantly, when focusing on near-threshold triples (with a margin less than 2) predicted by the fine-tuned models, the ICL paradigm outperforms fine-tuned models notably.

Base on these findings, we propose a simple yet effective framework that functions LLM as a near-threshold triple corrector to collaborate with the fine-tuned DocRE model. As illustrated in Figure 2, we leverage the fine-tuned DocRE model to make initial predictions on and calculate margins for all candidate triples. For each triple, if the corresponding margin is below a certain value, we feed it to the LLM and adopt its prediction as the final output.

3 Experiments

Setup. In this section we evaluate the overall performance of our proposed framework when integrated with existing DocRE models. Besides the fully-supervised training data setting, we also

Model	Fully-Supervised Setting				Low-Resource Setting				Incompletely-Labeled Setting			
	P	R	F1	ΔF1	P	R	F1	ΔF1	P	R	F1	ΔF1
ATLOP	89.01	69.90	78.30	-	81.40	63.08	71.05	-	87.17	36.34	51.11	-
+ Zero-Shot NTC	89.06	71.92	79.56	↑ 1.26	82.03	64.75	72.36	↑ 1.31	88.03	38.47	53.33	↑ 2.22
+ Few-Shot NTC	87.26	74.93	80.60	↑ 2.30	80.11	66.86	72.86	↑ 1.81	86.72	41.52	55.87	↑ 4.76
NCRL	88.12	72.30	79.43	-	81.38	58.02	67.73	-	93.97	34.24	50.19	-
+ Zero-Shot NTC	87.91	74.53	80.67	↑ 1.24	81.72	63.64	71.54	↑ 3.81	91.00	41.44	56.94	↑ 6.75
+ Few-Shot NTC	86.34	76.79	**81.28**	↑ 1.85	78.83	67.32	72.60	↑ 4.87	88.90	46.99	61.46	↑ 11.27
DREEAM	82.38	78.25	80.26	-	76.48	70.84	73.55	-	67.89	54.67	60.53	-
+ Zero-Shot NTC	83.66	78.10	80.78	↑ 0.52	78.69	70.95	74.62	↑ 1.07	69.83	55.16	61.61	↑ 1.08
+ Few-Shot NTC	81.79	80.39	81.08	↑ 0.82	76.01	74.08	**75.03**	↑ 1.48	66.88	58.93	**62.61**	↑ 2.08

Table 1: Main results on the test set of Re-DocRED in three training data settings.

Prompting Method	F1
Zero-Shot Vanilla Prompting	64.04
Zero-Shot Evidence-Driven Prompting	66.60 (↑ 2.56)
Few-Shot Vanilla Prompting	70.06
Few-Shot Evidence-Driven Prompting	71.78 (↑ 1.72)

Table 2: Ablation study results of evidence-driven prompting method.

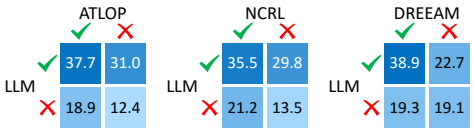

Figure 3: The cross matrix of prediction results between fine-tuned DocRE models and LLM.

experiment on two the low-resource setting and incompletely-labeled setting, where we use only 10% training data of Re-DocRED and the incompletely annotated training data of DocRED as the training documents. For the ICL of LLM, we test on zero-shot and few-shot settings. We provide five positive and negative demonstrations in the few-shot setting.

3.1 Main Results

We conduct experiments on Re-DocRED benchmark under various training data settings. We instantiate the LLM in the proposed framework with Llama-3.1-70B. Evaluation results in the fully-supervised setting show that, even though current fine-tuned DocRE models have obtained an F1 of around 80%, our framework still achieves an average improvement of 1.66% F1 across three advanced DocRE models, while maintaining efficiency for only predicting less than 0.02% of candidate triples using LLM. The best result outperforms the state-of-the-art DocRE model by 1.02% F1. In low-resource and incompletely-labeled settings, our framework demonstrates even more significant advantages.

3.2 Analysis

Effectiveness of Evidence-Driven Prompting. We further conduct an ablation study by comparing the vanilla prompting and the proposed evidence-driven prompting method. We random select 1000 triples from the predicted triples by LLM in fully-supervised setting of main experiments. As shown in Table 2, in both zero-shot and few-shot setting, the proposed evidence-driven prompting method shows a significant advantage.

Comparison of Predictive Distribution between Fine-tuned DocRE models and LLM. Considering that the LLM demonstrates noticeable performance advantage compared to fine-tuned DocRE models on their predicted near-threshold triples, an interesting question is whether the LLM only introduces some new correct predictions or has more distinct predictive distribution with fine-tuned models. We group all predicted triples by few-shot ICL in fully-supervised setting of main experiments based on whether fine-tuned models and LLM give a correct prediction. The cross matrix is show in Figure 3. We can observe that the predictive distribution of LLM is different with fine-tuned models in a certain degree. This suggests that it might be possible to group triples at a finer granularity (e.g., relation by relation), thereby better integrating the predictions of both paradigms, rather than overriding with the results of LLM.

4 Conclusion

In this work, we propose a simple, effective framework that serves LLM as a near-threshold triple corrector to collaborate with fine-tuned DocRE models, significantly improving upon existing models and achieving SOTA performance. Our framework can also transfer to other document-level multi-label tasks like document-level event extraction.

Limitations

We acknowledge the following limitations of this study. Firstly, the ICL paradigm adopted in the proposed framework still has room for further optimization, such as designing specific demonstration selection strategies tailored for the DocRE task. Secondly, the experiments are not conducted on domain-specific datasets, such as those in biomedical field, leaving a meaningful direction for future exploration. Lastly, due to the budget constraints, we do not experiment with proprietary LLMs like GPT-4o, where their stronger capabilities may further enhance the effectiveness of our framework.

References

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983.

Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang, and Soujanya Poria. 2023. Uncertainty guided label denoising for document-level distant relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15960–15973.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487.

Ying Wei and Qi Li. 2022. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 2000–2008.

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. SAIS: Supervising and augmenting intermediate steps for document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2395–2409.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, pages 14612–14620.

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4538–4544.

A Dataset and Models

We conduct the experiment on Re-DocRED (Tan et al., 2022b), a revised version of the most popular public DocRE dataset DocRED (Yao et al., 2019) by resolving the overwhelming false negative issue in DocRED's annotations. And we choose three public DocRE models which are representative for their strong performance and high popularity: ATLOP (Zhou et al., 2021) proposes a localized context pooling technique and an adaptive thresholding loss to tackle the multi-entity and multi-label problem in DocRE; NCRL (Zhou and Lee, 2022) improves upon ATLOP by introducing a novel none class ranking loss which is theoretically proven to satisfy multi-label consistency; DREEAM (Ma et al., 2023) guides the attention modules with evidence information and further enhances the inference stage with the extracted evidence, achieving state-of-the-art performance. We use F1 score as the metric for performance evaluation.

B Prompt Template

In Table 3, we list the complete prompt template for the ICL of LLM used in this work.

C Relation Verbalization Templates

In Table 4 and Table 5, we list the verbalization templates for relation types in Re-DocRED.

You are an expert in reading comprehension. You will be provided with a document and a statement. Your task is to determine whether the statement can be inferred from the information contained in the document. Use the following step-by-step instructions to respond to user inputs.

Step 1 - Carefully read the document, comprehensively understand its content, and analyze in detail whether there are key evidence sentence(s) in the document that can support the inference of the given statement. Output all your work for this step starting with "Analysis: ".

Step 2 - Based on the analysis from Step 1, provide your final answer (either "Yes" or "No") starting with "Answer: ".

Document: <document of positive demonstration>

Statement: <statement of positive demonstration>

Analysis: Upon carefully reading the document, I found that the document mentions that "<evidence 1>" "<evidence 2>" and "<evidence 3>" These sentences combined indicate that "<statement>" supporting the inference of the given statement.

Answer: Yes

Document: <document of negative demonstration>

Statement: <statement of negative demonstration>

Analysis: Upon carefully reading the document, I did not find any sentence or combination of sentences that explicitly or implicitly indicate that "<statement>" There is no evidence in the document to support the inference of the given statement.

Answer: No

Document: <query document>

Statement: <query statement>

Table 3: The complete prompt template for the ICL of LLM used in this work. The slots enclosed in angle brackets are filled with the document, statement or evidence sentence.

Wikidata ID	Relation	Verbalization Template
P6	head of government	<tail> is the head of government of <head>.
P17	country	The sovereign state of <head> is <tail>.
P19	place of birth	The birth location of <head> is <tail>.
P20	place of death	The death location of <head> is <tail>.
P22	father	<tail> is the father of <head>.
P25	mother	<tail> is the mother of <head>.
P26	spouse	<tail> is the spouse of <head>.
P27	country of citizenship	The country of citizenship of <head> is <tail>.
P30	continent	<head> is a part of the continent <tail>.
P31	instance of	<head> is a particular example/member of the class <tail>.
P35	head of state	<tail> is the head of state of <head>.
P36	capital	The capital of <head> is <tail>.
P37	official language	The official language of <head> is <tail>.
P39	position held	<head> currently or formerly holds the <tail> position or public office.
P40	child	<tail> is the child of <head>.
P50	author	<tail> is the main author/creator of the work <head>.
P54	member of sports team	<head> currently represents or formerly represented the sports team <tail>.
P57	director	<tail> is the director of <head>.
P58	screenwriter	<tail> is the screenwriter of <head>.
P69	educated at	<head> is educated at <tail>.
P86	composer	<tail> is the composer of <head>.
P102	member of political party	<head> is or has been a member of the political party <tail>.
P108	employer	<head> works or worked for <tail>.
P112	founded by	<tail> is the founder or co-founder of <head>.
P118	league	<head> plays or has played in the league <tail>.
P123	publisher	<tail> is the publisher of <head>.
P127	owned by	<tail> is the owner of <head>.
P131	located in the administrative territorial entity	<head> is located on the territory of the administrative entity <tail>.
P136	genre	The genre (or field of work) of <head> is <tail>.
P137	operator	<tail> is the operator of <head>.
P140	religion	The religion of <head> is <tail>.
P150	contains administrative territorial entity	<head> contains the administrative territorial entity <tail>.
P155	follows	<head> follows <tail>.
P156	followed by	<head> is followed by <tail>.
P159	headquarters location	<head>'s headquarters is or has been situated at <tail>.
P161	cast member	<tail> is a cast member of <head>.
P162	producer	<tail> is the producer of <head>.
P166	award received	<head> received the award <tail>.
P170	creator	<tail> is the creator/maker of <head>.
P171	parent taxon	<tail> is the parent taxon of <head>.
P172	ethnic group	<head>'s ethnicity is <tail>.
P175	performer	<tail> is a performer involved in the performance or the recording of <head>.
P176	manufacturer	<tail> is the manufacturer of <head>.
P178	developer	<tail> is the developer of <head>.
P179	series	<head> is part of the series <tail>.
P190	sister city	<head> and <tail> are twin towns, or sister cities.
P194	legislative body	<tail> is the legislative body governing <head>.
P205	basin country	<tail> is a country bordering the <head> waters.

Table 4: Verbalization templates for relation types in Re-DocRED. The slot <head> and <tail> in the template are filled with the name of head entity and tail entity, respectively. (continued on next page)

Wikidata ID	Relation	Verbalization Template
P206	located in or next to body of water	<head> is located in or next to the <tail> waters.
P241	military branch	<head> belongs to the military branch <tail>.
P264	record label	The record label of <head> is <tail>.
P272	production company	<tail> is the production company of <head>.
P276	location	The location of <head> is within <tail>.
P279	subclass of	<head> is a subclass of <tail>.
P355	subsidiary	<tail> is a subsidiary of <head>.
P361	part of	<head> is a part of <tail>.
P364	original language of work	<head> was originally created in the <tail> language.
P400	platform	<head> has been developed or released for the platform <tail>.
P403	mouth of the watercourse	The watercourse <head> flows into the <tail> waters.
P449	original network	<tail> is the channel that originally broadcasts the program <head>.
P463	member of	<head> is a member of <tail>.
P488	chairperson	<tail> is the chairperson of <head>.
P495	country of origin	The country of origin of <head> is <tail>.
P527	has part	<tail> is a part of <head>.
P551	residence	<head> is, or has been, resident in <tail>.
P569	date of birth	The birth date of <head> is <tail>.
P570	date of death	The death date of <head> is <tail>.
P571	inception	The date or point in time when <head> was founded/created is <tail>.
P576	dissolved, abolished or demolished	The date or point in time on which <head> was dissolved, abolished or demolished is <tail>.
P577	publication date	The publication date of <head> is <tail>.
P580	start time	The start time of <head> is <tail>.
P582	end time	The end time of <head> is <tail>.
P585	point in time	The time point of <head> is <tail>.
P607	conflict	<head> participated in the military conflict <tail>.
P674	characters	<tail> is a character that appears in <head>.
P676	lyrics by	<tail> is the lyricist of <head>.
P706	located on terrain feature	<head> is located on the specified landform <tail>.
P710	participant	<tail> takes part in the event <head>.
P737	influenced by	<head> is influenced by <tail>.
P740	location of formation	The formation location of <head> is <tail>.
P749	parent organization	<tail> is the parent organization of <head>.
P800	notable work	<tail> is a notable work of <head>.
P807	separated from	<head> was founded or started by separating from <tail>.
P840	narrative location	The narrative of the work <head> is set in <tail>.
P937	work location	<head>'s workplace is or has been located in <tail>.
P1001	applies to jurisdiction	<head> belongs to or has power over or applies to <tail>.
P1056	product or material produced	<tail> is a product or material produced by <head>.
P1198	unemployment rate	The unemployment rate of <head> is <tail>.
P1336	territory claimed by	<head>'s territory is claimed by <tail>.
P1344	participant of	<head> takes part in the event <tail>.
P1365	replaces	<head> replaces <tail>.
P1366	replaced by	<head> is replaced by <tail>.
P1376	capital of	The capital of <tail> is <head>.
P1412	languages spoken, written or signed	<head> speaks or writes the <tail> language.
P1441	present in work	<head> is present in the work <tail>.
P3373	sibling	<tail> is the sibling of <head>.

Table 5: Verbalization templates for relation types in Re-DocRED. The slot <head> and <tail> in the template are filled with the name of head entity and tail entity, respectively. (continued)