From Memorization to Discovery: A Novel Benchmark for Relational Triple Extraction

Aoran Gan[1], Ye Liu[1], Hongbo Gang[1], Kai Zhang[1], Qi Liu[1](✉), Enhong Chen[1], Xin Li[2], and Guoping Hu[1,2]
{gar,liuyer,elysiumghb}@mail.ustc.edu.cn
{kkzhang08,qiliuql(✉),cheneh,leexin}@ustc.edu.cn
gphu@iflytek.cn

[1] State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei, China
[2] Artificial Intelligence Research Institute, iFLYTEK Co., Ltd, Hefei, China

Abstract. Relational Triple Extraction (RTE) mines factual knowledge components as relational triples from unstructured text. However, most triples tested in current datasets are already duplicated in the training set, leading past studies to rely more on memorization than on genuine discovery. In response to this, we suggest a novel benchmark **ENT** to assess the model's capability to **E**xtract **N**ew **T**riples, which aligns more closely with the practical objective of RTE such as automatic knowledge graph construction. [3] We developed the dataset by instructing the large language model to perform text expansion based on preprocessed knowledge graph segments, followed by rule-based and semantic check. The ENT dataset, boasting over 300,000 unique relational triples, encompasses a broad spectrum of knowledge. The proportion of new triples in the test set exceed 60%, and all the samples contain at least one unseen triples, highlighting a strong emphasis on discovering new knowledge. ENT is perceived by human annotators with a low level of hallucination, serving as a valid and valuable dataset. We re-evaluated 9 state-of-the-art RTE methods and found a generalized accuracy decrease on ENT, demonstrating that ENT is a more challenging and meaningful benchmark.

Keywords: Relational Triple Extraction · Large Language Model · Dataset and Benchmark

1 Introduction

Relational Triple Extraction (RTE), also called joint extraction of entities and relations or triple extraction, aims to extract the relational triples <subject, relation, object> from raw text [16]. Accurately mining and modeling valuable information from text data constitutes a significant challenge and plays a crucial role in the field of knowledge engineering [33, 34, 14, 28, 15]. Early researches

[3] The dataset is available at https://github.com/Kast-Nora/ENT-Dataset .

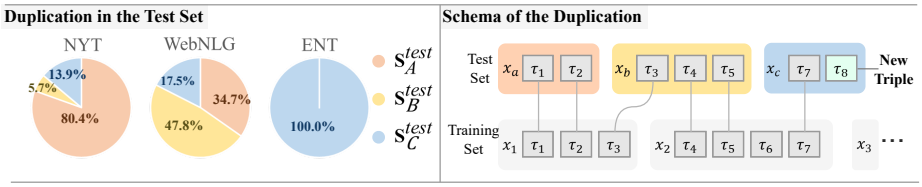

Fig. 1. Triple duplication in NYT, WebNLG and ENT. $\mathbf{S}_A^{test}$ exhibits the highest degree of triple duplication, followed by $\mathbf{S}_B^{test}$. $\mathbf{S}_C^{test}$ contains new triples. x_a, x_b, and x_c are schematic illustrations of the duplicated samples in $\mathbf{S}_A^{test}$, $\mathbf{S}_B^{test}$ and $\mathbf{S}_C^{test}$, respectively. Two triples τ_i, τ_j are considered duplicates only in the cases with all the identicalness between their subjects, relations, and objects, that is, $(s_i = s_j)\&(r_i = r_j)\&(o_i = o_j)$.

decomposed RTE into two components: entity identification and relation categorization [32, 3]. In recent years, researchers have increasingly focused on the deep connection between entities and relationships [30, 36, 22, 29, 27]. Among them, TPLinker [29] initially implemented a one-step triple extraction by conceptualizing RTE as a table-filling task. UniRel [27] proposed a unified entity-relation representation and interaction framework. These methods have made great strides in the development of RTE and achieved a high level of accuracy.

Despite the performance improvements of prior works, a considerable potential flaw as triple duplication exists within the current benchmark for RTE [12]. As a result of our calculation, more than 80% of the triples in the test set of NYT and WebNLG are duplicates, i.e., they appear at least once in the training set as is. A significant part of test set samples contain even completely duplicated triples to another sample in the training set (as $\mathbf{S}_A^{test}$ shown in Figure 1). This implies that the two benchmarks are biased towards evaluating the model's ability to memorize existing triples, rather than discover new ones. As new triples are considered more valuable in some of the real world requirements, such as the automatic or semi-automatic construction of KGs [6, 16], the existing benchmarks of RTE exhibit a significant gap due to the lack of adequate emphasis on them.

To broaden the scope of discovering new triples, we designed and implemented a KG-based automated dataset construction pipeline and developed a new benchmark dataset, ENT. The pipeline consists of four steps: i) *Preprocess* that performs irrelevant triple filtering in the collected and clustered knowledge base. ii) *Paragraph Generation* by prompting to the Large Language Model (LLM). iii) *Rule-based Check* that identifies and rectifies the unconforming paragraphs. iv) *Semantic Check* of the alignment between the relational triples and paragraphs. We finally obtained the ENT dataset with 62k samples and 347k unique triples. More than 60% of the test set triples are new, not found in the training set. Concurrently, each sample of the test set comprises at least one new triple. This indicates that ENT can represent the extraction capability of new knowledge more accurately.

We re-evaluated nine state-of-the-art RTE methods on the ENT benchmark and observed a generalized 10%+ and 7.5%+ accuracy decrease compared with

the two other most widely used benchmarks NYT [23] and WebNLG [8]. We conducted a more thorough analysis and revealed a lower propensity for bias towards duplicated triples of ENT. It demonstrates that ENT serves as a more challenging and meaningful benchmark from the perspective of discovering new triples. We also observed a positive correlation at a more granular level between the difficulty of triple extraction and the extent of new knowledge. We have open-sourced the dataset and hope that ENT will serve as a more challenging resource and lead to new directions for RTE.

2 Related Work

Relational Triple Extraction. Some of the RTE studies are conducted on the simplified version of the existing datasets called partial-match, where the RTE model identify only the final word of the entities [37, 7, 13, 35]. Other works propose more realistic frameworks, that is, exact-match extraction, which stipulate that all entities must be extracted in their entirety. CasRel [30] proposed a two-stage triple extraction scheme, which successfully addressed a significant number of overlapped entities for the first time. SPN4RE [26] treated RTE as an ensemble prediction problem and employed a non-autoregressive decoder. UniRel [27] proposed a novel unified entity-relation interaction modeling approach. DirectRel [25] devised a method for entity extension matching, though at the cost of significantly increasing the text sequence length. [19] attempted to utilize LLMs for direct few-shot triple extraction but observed that the LLM struggled to attain competitive performance with classical baseline models.

RTE Dataset. NYT [23] and WebNLG [8] are the most widely used datasets for RTE at present. NYT was constructed by remote supervised relation extraction, containing noisy samples and having a limited number of relations. WebNLG employed native English speakers to write text descriptions for relational triples and got a dataset of limited size. [12] introduced for the first time the absence of model generalization assessment within these two datasets and proposed a few of improved strategies. To address this issue thoroughly, we suggest the development of a novel RTE benchmark that significantly and directly reduces knowledge duplication. While obtaining well-aligned and adequate RTE data was difficult in the past, it is fortunately becoming acceptable to use machines for data annotation with the rapid development of deep learning techniques [10, 11]. For example, [10] used machine translation models to build a multilingual relation extraction dataset.

LLM for Text Generation. LLMs have demonstrated extraordinary capabilities of text generation [20, 2]. By incorporating human feedback into LLMs, it is possible to generate outputs that are more aligned with human preferences [18]. Concurrently, the KG can significantly mitigate the hallucination issue of LLMs [9, 31]. Zero-shot automatic text generation via LLM with factual triples has demonstrated competitive performance [1]. In this work, we utilize the triples from a real-world KG to instruct the LLM for the development of an RTE dataset.

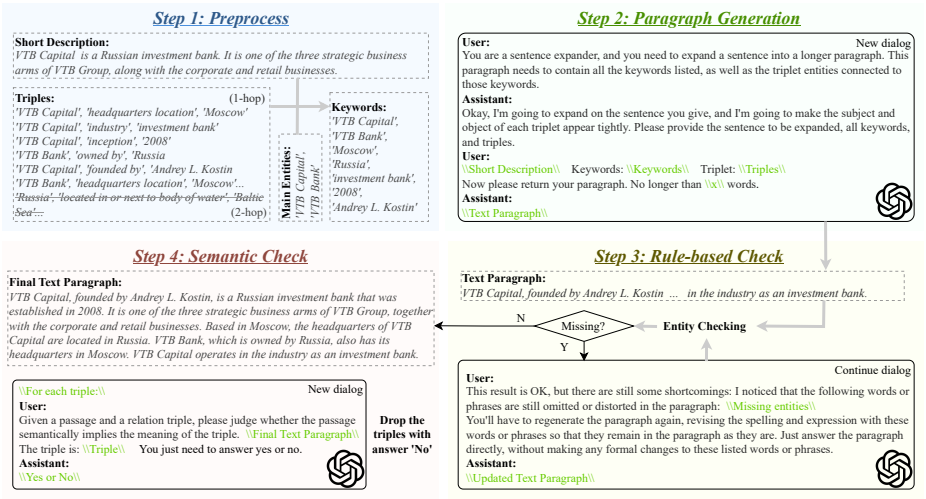

Fig. 2. The process of constructing the ENT dataset with detailed content of prompt.

3 Methodology

3.1 Formalized Definition of RTE

Given a text sequence input $W = [w_1, w_2, ..., w_L]$, RTE aims to predict the set of relational triples: $\mathcal{T} = \{\tau_n \mid n \in \{1, ...N\}\}$, $\tau_n = (s_n, r_n, o_n)$. Each relationship r_n of the triple belongs to a pre-defined relation set $\mathcal{R}$. All the subjects $\{s_n\}$ and the objects $\{o_n\}$ are consecutive segments $[w_i, w_{i+1}, ..., w_j]$ $(1 \leq i \leq j \leq L)$ extracted from the input sentence. The number of triples N per sentence may be greater than 1, and the exact number is unknown in advance. The input does not contain extra knowledge (e.g., entity information).

3.2 Dataset Construction

Constructing an RTE dataset requires the collection of text-triples sample pairs. Our work is predicated on ENT-DESC [4], a dataset for natural language generation that includes the triples set from Wikidata's KG slices and entity descriptions from Wikipedia pages as paired samples. However, the textual description is notably brief, omitting most entities and falling short of RTE. To actualize the text construction, we utilized OpenAI's GPT-3.5-Turbo-1106 API [4] as the LLM for automatic text generation. The objective of the LLM is to generate a textual paragraph incorporating all the specified entity keywords, which is both textually and semantically aligned with the relational triples. The entire text generation process is designed as a four-step procedure as follows: *Preprocess*, *Paragraph Generation*, *Rule-based Check* and *Semantic Check*.

[4] https://platform.openai.com/docs/api-reference

Preprocess. Each sample in ENT-DESC has several main entities and the relational triples within 2-hop paths. We retain the 1-hop triples, whose subject or object connected with the main entities directly, and discard the 2-hop ones. This is due to the fact that the 2-hop triples result in more verbose paragraphs, thereby making the expository focus of the paragraphs more ambiguous. We retain the 200 relationships with the highest frequency of occurrence. Each relation has at least 20 unique triples.

Paragraph Generation. In this step, we instruct the LLM to expand the description and generate a longer paragraph. We meticulously outline the commands that the LLM needs to execute in the prompt. The LLM needs to expand the existing short description based on the relational triples and ensure that all the keywords are located within the expanded paragraph. In an effort to mitigate the verbosity of the LLM's statements, we implement a straightforward dynamic soft-limit policy by instructing the LLM to generate paragraphs of a certain word length or less. The specific length parameter is discussed in Section 4.3.

Rule-based Check. Although the keyword- and triple-based prompt enables the LLM to generate more accurate paragraph, it runs the risk of hallucination that the LLM does not execute instructions. Specifically, the generated paragraph may contain syntactic reconstruction or entity content re-expression. We thus introduce a direct rule-based operation by introducing a BERT-base-cased [5] tokenizer to check if the original entity is missing from the paragraph in this step. We consider the entity to be rule-compliant for extraction if it's token sequence can be actually matched in the paragraph. Otherwise, it is considered to be missing. If there are missing entities, we continue to identify such entities and instruct the LLM to regenerate a new paragraph until all the entities are matched. We discard the sample with the token [UNK] or ≥ 1 missing entities.

Semantic Check. LLMs might overlook the expression or comprehension of specific local information. In this step, we reinitiate a new dialog with the LLM to check whether the semantic meaning of the triple is conveyed within the paragraph. The LLM here does not have access to the previous dialog. We drop the triples with semantically negative response. We verify good semantic alignment between the triples and LLM-generated text passages and low level of hallucination evidenced by the introduction of human opinions. Section 4.5 provides an in-depth explanation.

3.3 ENT Dataset and Benchmark

We finally collected 62,609 English paragraphs with 347,452 unique exact-match triples with 200 relations overall. The domains of the triples include characters, locations and institutions. We divide the entire dataset into the training set(~80%), validation set(~10%) and test set(~10%). This is a wiki-style dataset, with each sample consisting of a textual paragraph and several relational triples that focus

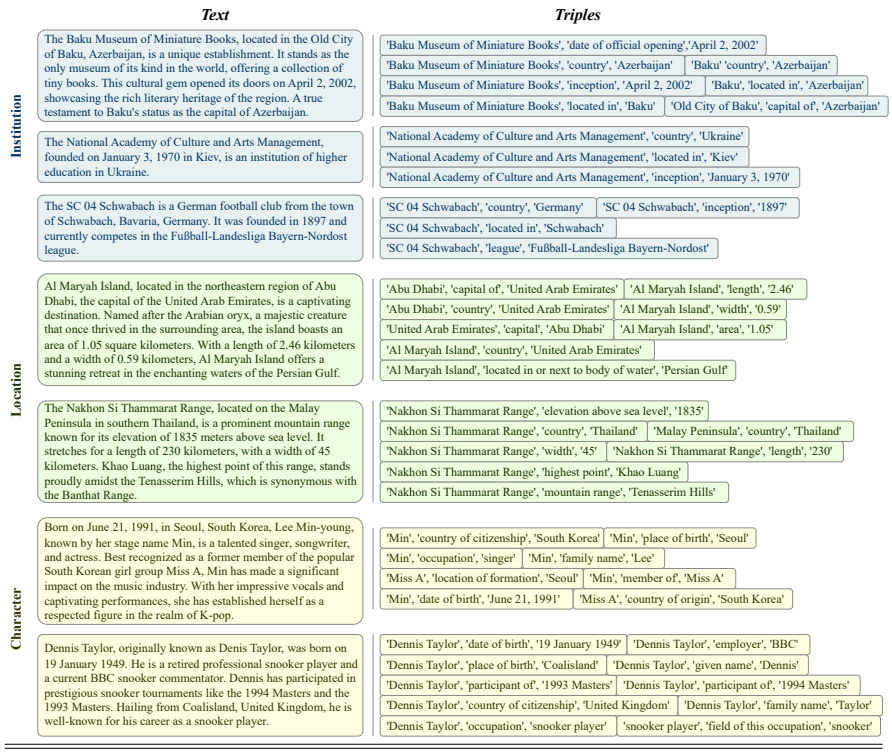

Fig. 3. Some of the samples of our dataset.

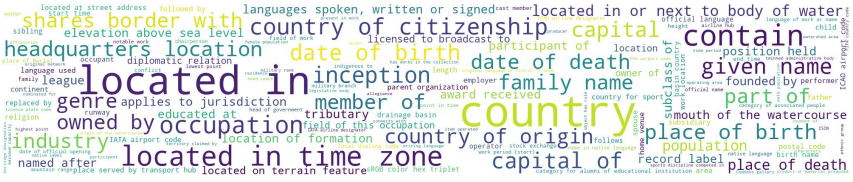

Fig. 4. Word cloud of all the 200 relationships of our dataset.

on one or several main entities. Each paragraph begins with a comprehensive overview that introduces the main entities. Subsequent sentences then provide more detailed descriptions according to the triples. The generated text paragraphs typically hold well-structured content and do not adhere rigidly to the listed triples. Figure 3 and 4 provide a detailed presentation of the samples and relationships.

The key improvement of this dataset is the introduction of new knowledge discovery test, both in terms of the volume and the proportion of previously unseen triples. Each sample in the test set contains at least one new triples (as shown in Figure 1). The proportion of new triples exceeds 60% in the test set

Table 1. ENT vs. NYT and WebNLG. μ_N denotes the average number of triples of each sample. $\mu_{F(\tau)}$ denotes the average frequency of each unique triple in the training set. $F(\tau) = 1$ means the triple τ appears only once in the training set. N'_{test} and N_{test} represent the number of new triples and all the triples in the test set, respectively.

Dataset	Train	Valid	Test	Relations	Unique Triples	μ_N	$\mu_{F(\tau)}$	N'_{test}/N_{test}
NYT	56,196	5,000	5,000	24	17,621	1.6	5.5	0.104
WebNLG	5,019	500	703	216	2,661	2.3	4.6	0.089
ENT	49,968	6,043	6,058	200	**347,452**	8.6	**1.5**	**0.617**

Table 2. Categories and the number of samples from different perspective of the degree of the new knowledge for ENT / NYT / WebNLG test set. N', N and E' denote the number of new triples, all triples and unique new entities in each sample. R_m denotes the ordinal number of the frequencies of the globally rarest relationship in each sample. For example, group t1 means $N'/N \in [0, 0.2)$, and NYT's t1 group has 4316 samples.

Perspective	Metric	1	2	3	4	5	6	Group
New Triples	N'/N	[0,0.2)	[0.2,0.4)	[0.4,0.6)	[0.6,0.8)	[0.8,1.0)	{1.0}	t1-t6
New Entity	E'	{0}	{1}	{2}	{3}	{4}	[5,+∞)	e1-e6
Rare Relation	R_m	[1,10)	[10,25)	[25,50)	[50,75)	[75,100)	[100,+∞)	r1-r6

Dataset	Group	Number of Samples						KRC
NYT	t1-t6	4316	42	53	5	2	582	t-e 0.380
	e1-e6	4886	106	8	0	0	0	t-r 0.103
	r1-r6	2105	2895	0	0	0	0	e-r 0.040
WebNLG	t1-t6	581	37	21	4	0	60	t-e 0.310
	e1-e6	689	14	0	0	0	0	t-r 0.031
	r1-r6	191	139	184	68	52	69	e-r 0.022
ENT	t1-t6	242	1127	1265	1147	796	1481	t-e **0.450**
	e1-e6	461	2357	1366	837	513	524	t-r **0.277**
	r1-r6	1490	1077	1233	667	636	955	e-r **0.429**

without manually changing the data distribution. In contrast, NYT and WebNLG comprise only ~10%. The cause can be attributed to two factors. i) The main entities of the original triple groups were derived and clustered from PageRank scores, demonstrating strong topic independence. ii) We discard the 2-hop triples, further reducing the triple duplication between different samples. We name this dataset **ENT** and believe it serves as a more compelling and appropriate benchmark for evaluating the methods' ability to **E**xtract **N**ew **T**riples.

The detailed statistical metrics are listed in Table 1. ENT has a sufficient sample size as NYT and contains a greater number of relationships. Furthermore, the unique triples within ENT are significantly more numerous, encompassing a wider range of knowledge. Additionally, the average frequency of each unique triples in the training set is much lower, which is experimentally validated as beneficial for the model's generalization to new knowledge in Section 4.4.

The assessment of new knowledge discovery of RTE has not been clearly defined before. Nevertheless, we endeavor to assess the degree from three intuitive perspectives: i) *The proportion of new triples.* ii) *The number of unique unseen entities.* iii) *The rarity of the relationships.* Based on these three perspectives, we divided the test set into six distinct groups separately, as shown in Table 2. The sample sizes of the diverse test subsets of ENT are more uniform and adequate across all of the categories. We further compute the Kendall's Rank Correlation (KRC, tau-a) between each categories within the three dataset. The categories of ENT exhibit a moderate degree of correlation, particularly in regard to relationships (KRC on t-r and e-r). This aligns more closely with the intuitive understanding of human cognitive processes. Specifically, the new, previously unseen entities are more likely to exhibit uncommon relationships compared to the more familiar entities, as humans acquire new knowledge. In contrast, this correlation is very weak in the other two datasets. This suggests that ENT benchmark holds more value and more practical significance.

Noted that the samples of ENT have typically longer text compared to NYT and WebNLG. It is primarily due to the larger number of triples associated with the main entities (higher μ_N in Table 1). As expressing richer knowledge relies on the longer texts, we cannot further compress the text while simultaneously keeping the triple as complete as possible. Additionally, the challenge of ENT does not arise from the longer paragraph, which is discussed in Section 4.3.

4 Experiment and Analysis

4.1 RTE Experiment Setups

We focus on the exact-match format of NYT and WebNLG, that is, evaluating the triples as the whole <subject, relation, object>, which more closely aligns with the real-world RTE applications. ENT is also exact-matched. For NYT, we choose the most commonly used version for RTE task, also known as NYT24. We select 9 state-of-the-art RTE methods for our reassessment: **CasRel** [30], **SPN4RE** [26], **TPLinker** [29], **PRGC** [36], **GRTE** [21], **BiRTE** [22], **OneRel** [24], **UniRel** [27], and **OD-RTE** [17]. For each method, we create and configure a specific anaconda environment based on the packages and their versions indicated in the source code respectively. We initialize all the models with the pretrained BERT-base-cased weights, and finally test each model on the checkpoint with the highest validation F1 score and set the batch size as 1 for inference. We utilize publicly available code and the optimal hyperparameter configurations cited in the original paper to retrain the model. For CasRel, we preprocess ENT in the same manner as Wiki-KBP. For OneRel, we insert spaces between the text and punctuation and record the entity mapping for inference. For the relation hint in UniRel, each relationship will use its first or last word as the hint if the token is available, or occupy an exclusive unused token. Table 3 list the key hyperparameters when training models with ENT dataset. All the RTE experiments are conducted on a computer equipped with an Intel(R) Xeon(R) Platinum 8350C CPU, 56 GB of RAM, and one NVIDIA GeForce RTX 3090.

Table 3. Key hyperparameters for model training on the ENT dataset. *lr*, *bsz*, and *epc* represent learning rate, batch size, and training epochs, respectively.

	CasRel	SPN4RE	TPLinker	PRGC	GRTE	BiRTE	OneRel	UniRel	OD-RTE
lr	1e-5	1e-5	1e-5	1e-3	3e-5	3e-5	1e-5	3e-5	5e-5
bsz	6	8	6	64	6	18	8	12	6
epc	100	100	100	100	50	100	200	100	20

Table 4. Precision (P), recall (R) and micro F1 score (F1)(%) on NYT, WebNLG and ENT. Except for the data with '*' reported by GRTE, the other metrics of NYT and WebNLG are sourced from the original papers.

Method	NYT			WebNLG			ENT		
	P	R	F1	P	R	F1	P	R	F1
CasRel [30]	89.8*	88.2*	89.0*	88.3*	84.6*	86.4*	73.8	54.2	62.2
SPN4RE [26]	92.5	92.2	92.3	85.7*	82.9*	84.3*	78.3	76.6	77.4
TPLinker [29]	91.4	92.6	92.0	88.9	84.5	86.7	70.7	75.3	72.9
PRGC [36]	93.5	91.9	92.7	89.9	87.2	88.5	72.4	74.2	73.3
GRTE [21]	93.4	93.5	93.4	92.3	87.9	90.0	**83.9**	81.1	**82.4**
BiRTE [22]	91.9	93.7	92.8	89.0	89.5	89.3	81.5	80.8	81.2
OneRel [24]	93.2	92.6	92.9	91.8	90.3	91.0	81.9	79.7	80.8
UniRel [27]	93.7	93.2	93.4	91.8	90.5	91.1	78.9	80.8	79.8
OD-RTE [17]	**94.2**	**93.6**	**93.9**	**92.8**	**92.1**	**92.5**	78.7	**81.9**	80.3

4.2 Main Results

We present the overall accuracy of various RTE methods on ENT in Table 4, contrasting them with NYT and WebNLG. The accuracy of existing methods on ENT is typically 10%+ lower than that on NYT, which has a comparable data volume to the former. The ENT accuracy is also generally 7.5%+ lower than WebNLG, whose data volume is approximately 0.1x. This suggests that our dataset presents a greater challenge. We observe that OD-RTE, reported as the latest state-of-the-art, identifies all the entities that appear multiple times in the text, regardless of their location when performing tagging, training, and inference. This may lead to an aggressive decoding of more triples and lower accuracy, notably enhanced by the larger quantity of triples in ENT.

Figure 5 presents more detailed benchmark results. As the degree of new knowledge increases, the accuracy of extraction tends to decrease. Additionally, in the group with less new knowledge, the accuracy of the existing RTE methods are comparable to that of both NYT and WebNLG, suggesting that the principal challenge associated with ENT primarily originate from the incorporation of assessments aimed at discovering new knowledge. The proportion of new triples serves as the most intuitive and convincing metric for measuring the degree of new knowledge, as all the RTE methods consistently demonstrate a decline in extraction accuracy as it increases. This indicates that our dataset and benchmark are valid and valuable.

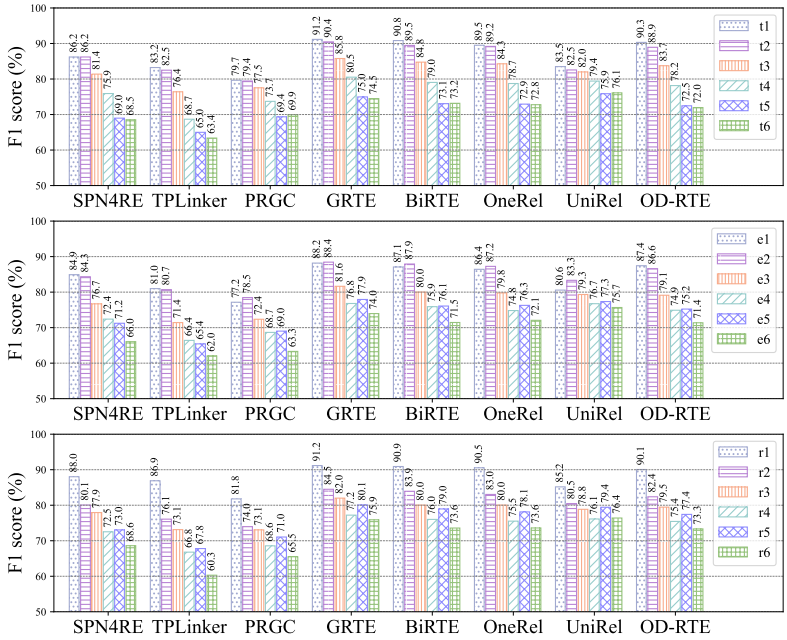

Fig. 5. Specific triple micro F1 scores of RTE methods in three different perspectives of ENT test set. t1-t6 presents the different proportion of new triples in a sample. e1-e6 presents the different number of new entities. r1-r6 presents the max ordinal relation number in a sample. Category details are shown in Table 2.

4.3 Text Length Analysis

The text of ENT is typically long, but not as long as a document. This is due to our desire to retain as much information as possible within the text paragraph during the data construction procedure. We let the LLM not only has a relatively sufficient amount of words to express the knowledge, but also simplify the expression as much as possible by dynamically setting the text length limit based on the number of triples. We conducted a preliminary experiment on 200 samples to select the appropriate statement in the prompt. Table 5 shows the trade-off between the triple recall and text length. We finally chose $x = 8N + 4$ in the prompt for the large scale API calling.

Moreover, the difficulty with our dataset isn't due to the length of the text. We segment all the paragraphs in the training set, validation set, and test set into short texts with 1~2 sentences. We conduct two approaches: sequential segmentation that concatenate the adjacent sentences, and tree segmentation that concatenate the first sentence with the other ones in the paragraph respectively. Only those triples that comply with the task definition in Section 3.1 are retained. As illustrated in Table 6, the models underperform when training and testing with

Table 5. Triple recall and average text length with different statement of the 200 samples. x refers to the 'No longer than x words.' in Figure 2.

Prompt Keyword	$x = 4N + 4$	$x = 6N + 4$	$x = 8N + 4$	$x = 10N + 4$	No Statement
Triple Recall (%)	84.0	89.7	92.2	92.2	92.4
Text Length (token)	91.9	106.5	123.0	137.7	160.4

Table 6. The accuracy (%) of the RTE model in reducing text length. ENT-seq and ENT-tree are the two versions with shorter text obtained by the sequential and tree segmentation, respectively. We choose SPN4RE, BiRTE and UniRel because they represent diverse technical routes of RTE.

Method	SPN4RE			BiRTE			UniRel		
Accuracy	P	R	F1	P	R	F1	P	R	F1
ENT	78.6	76.6	77.4	81.5	80.8	81.2	78.9	80.8	79.8
ENT-seq	$76.0_{\downarrow}$	$75.2_{\downarrow}$	$75.5_{\downarrow}$	$78.3_{\downarrow}$	$80.5_{\downarrow}$	$79.4_{\downarrow}$	$76.3_{\downarrow}$	$79.1_{\downarrow}$	$77.6_{\downarrow}$
ENT-tree	$75.4_{\downarrow}$	$75.0_{\downarrow}$	$75.2_{\downarrow}$	$77.8_{\downarrow}$	$79.0_{\downarrow}$	$78.4_{\downarrow}$	$76.7_{\downarrow}$	$77.8_{\downarrow}$	$77.2_{\downarrow}$

shorter text. We posit that the longer paragraphs in the ENT offer a wider range of syntactic structures across sentences with more beneficial knowledge expression. It is clear that the longer text does not inherently make ENT challenging.

4.4 The Effect of Internal Duplication

We have previously highlighted the issue that triple duplication hinders the assessment of new knowledge. We developed the new ENT dataset and benchmark to solve the problem, instead of opting for a potentially simplistic approach, that is, eliminating the samples with duplicated triples in the test set. We undertook this operation following the discovery of another issue related to the duplication of the training set (independent of the test set) within the existing dataset: a considerable number of identical relational triples are expressed with a high frequency across different samples in the training set. Our experiments reveal that this problem, we call *internal duplication*, significantly impaired the model's ability to generalize and discover new knowledge.

Specifically, we slice two training subset and retrain the RTE model on NYT or WebNLG by different strategies for comparison. We firstly filter the samples by detecting internal duplicate triples within the training set and obtain a subset f such that it can just include all the unique triples. The samples with internal duplicate triples are discarded as much as possible. The second subset d is randomly sliced form the whole training set with the same sample size as the first one. The average frequency of each unique triple $\mu_{F(\tau)}$ in group f is much lower than that in group d. Meanwhile, we set the test set as $\mathbf{S}_{C*}^{test}$, the subset of $\mathbf{S}_{C}^{test}$ with $N' = N$ for each sample. In this manner, all the triples of the test set will be new ones, regardless of the slicing strategy. For subset d, we use three different random seeds and correspondingly get different versions as $d1, d2, d3$. The experimental results are shown in Table 7. The accuracy is significantly lower

Table 7. Comparison of the micro F1 score (%) on $\mathbf{S}_{C*}^{test}$ of the RTE methods with different training set slices. *nyt* and *web* denotes the training subset slices from NYT and WebNLG, respectively. SPN is short for SPN4RE.

Subset	Size	$\mu_{F(\tau)}$	SPN	BiRTE	UniRel	Subset	Size	$\mu_{F(\tau)}$	SPN	BiRTE	UniRel
nyt_f	11,925	**1.1**	**65.4**	**65.8**	**65.1**	web_f	1,463	**1.4**	**53.4**	**56.7**	**56.9**
nyt_{d1}	11,925	3.0	61.8	61.3	59.2	web_{d1}	1,463	2.3	45.4	42.6	48.0
nyt_{d2}	11,925	3.1	61.6	61.2	58.9	web_{d2}	1,463	2.5	44.3	41.7	46.6
nyt_{d3}	11,925	3.0	61.0	61.5	59.9	web_{d3}	1,463	2.2	45.5	43.0	48.2

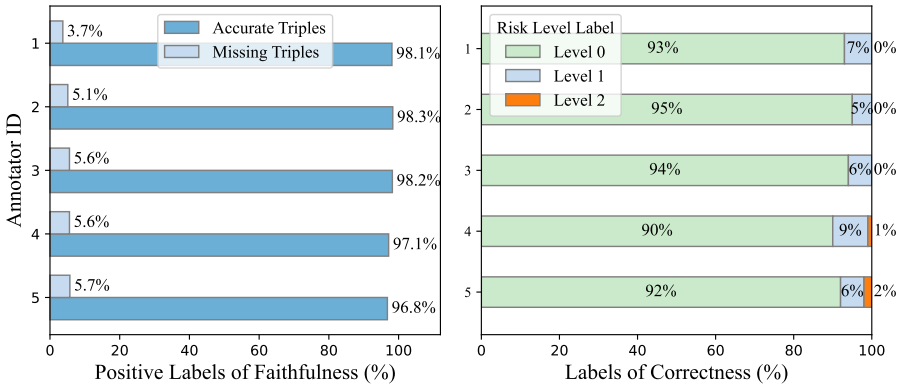

Fig. 6. Statistics on faithfulness (left) and correctness (right) derived from manual annotation. None of the samples were labeled as risk Level 3. Annotator 1-3 live in Asia, 4-5 live in North America. All of the annotators hold a bachelor's degree or higher and are compensated above the local minimum wage.

in the random slicing group, which implies that the internal duplicate triples do not facilitate the model's generalization at equivalent sample sizes. The minimal internal duplication in our ENT dataset (lower $\mu_{F(\tau)}$ listed in Table 1) further corroborates its value for assessing new knowledge discovery.

4.5 Hallucination Evaluation

Hallucination, a prevalent issue in LLMs, can be observed in the form of suboptimal instruction execution or the production of inaccurate or detrimental content. To assess the LLM's hallucinatory effect on our dataset, we hired five human annotators to evaluate the quality of 100 random samples in ENT, focusing on both *Faithfulness* and *Correctness*. For the faithfulness, the annotators were tasked with labeling the accurate triples and the missing triples respectively. They assessed whether the triples in each sample were accurately represented and whether any were missing, drawing on their comprehension of triples and textual content. For the correctness, we set 4 levels to signify the risk of hallucination and asked the annotators to label the text with the corresponding level number:

Table 8. Some of the bad cases with hallucination by human annotation. The similar examples constitute a minuscule proportion.

1-Text: *Born on June 21, 1991, in Seoul, South Korea, Lee Min-young, known by her stage name Min, is a talented singer, songwriter, and actress. Best recognized as a former member of the popular South Korean girl group Miss A, Min has made a significant impact on the music industry. With her impressive vocals and captivating performances, she has established herself as a respected figure in the realm of K-pop.* **1-Triples:** *<"Min", "country of citizenship", "South Korea"> <"Min", "place of birth", "Seoul"> <"Min", "occupation", "singer"> <"Min", "family name", "Lee"> <"Miss A", "location of formation", "Seoul"> <"Min", "member of", "Miss A"> <"Min", "date of birth", "June 21, 1991"> <"Miss A", "country of origin", "South Korea">* **Drawback:** Missing triples: *<"Min", "occupation", "songwriter"> <"Min", "occupation", "actress">*. All the 5 annotators recognized this case.
2-Text: *Unknown Mortal Orchestra is a New Zealand psychedelic rock band formed in Auckland in 2010. The band, primarily composed of Ruban Nielson and Jake Portrait, gained popularity under the record label Fat Possum Records in the United States of America.* **2-Triples:** *<"Unknown Mortal Orchestra", "inception", "2010"> < "Unknown Mortal Orchestra", "country of origin", "United States of America"> <"Unknown Mortal Orchestra", "record label", "Fat Possum Records"> <"Unknown Mortal Orchestra", "location of formation", "Auckland">* **Drawback:** Error triple description: *< "Unknown Mortal Orchestra", "country of origin", "United States of America">* Annotator 1, 3, 4, and 5 recognized this case.

- Level 0: It is fluent, safe, clear, and correct, with no anomalies discovered.
- Level 1: The content is safe and correct, but the word usage patterns slightly deviate from the normative human habits.
- Level 2: The content is safe to use, but the text lacks fluency or the descriptions can be confusing.
- Level 3: The content appears to be insecure, or there seems to be errors.

The annotation results are illustrated in Figure 6, where the Fleiss' Kappa values of the labels given by the annotators are 0.718, 0.656, and 0.621 for accurate triples, missing triples, and text hallucination risk levels, respectively. The results demonstrate that the annotation exhibit good consistency, and the prevalence of hallucinations in the samples within ENT is minimal. We present several instances of bad cases with hallucination in Table 8.

5 Conclusion

In this paper, we propose a new benchmark, ENT, with a wide range of knowledge for Relation Triple Extraction. To develop the dataset, we designed a pipeline based on LLM prompting to realize the automatic construction of massive data without large scale human labeling. The manual feedback indicates a low level of hallucination of the dataset construction. ENT not only encompasses 60k+ samples with 300k+ unique relational triples, but also has more new triples and lower duplication, thus offers a more meaningful benchmark of the model's generalization to the new knowledge. Following extensive experiments, ENT is found to be more challenging with 7%-20% accuracy decline compared to the other benchmarks. Our future research is anticipated to address the issue of grammatical convention gap between LLMs and humans, as well as the relational

alignment across different datasets. We hope that more researchers will focus on the discovery of new knowledge in RTE in the future.

Acknowledgments. This research was partially supported by the National Natural Science Foundation of China (Grants No.62406303), Anhui Provincial Natural Science Foundation (No. 2308085QF229), the Fundamental Research Funds for the Central Universities and the Iflytek joint research program.

References

1. Axelsson, A., Skantze, G.: Using large language models for zero-shot natural language generation from knowledge graphs. In: Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023). pp. 39–54. Association for Computational Linguistics (2023)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Chan, Y.S., Roth, D.: Exploiting syntactico-semantic structures for relation extraction. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 551–560. Association for Computational Linguistics (2011)
4. Cheng, L., Wu, D., Bing, L., Zhang, Y., Jie, Z., Lu, W., Si, L.: ENT-DESC: Entity description generation by exploring knowledge graph. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1187–1197. Association for Computational Linguistics (2020)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 601–610 (2014)
7. Fu, T.J., Li, P.H., Ma, W.Y.: GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1409–1418. Association for Computational Linguistics (2019)
8. Gardent, C., Shimorina, A., Narayan, S., Perez-Beltrachini, L.: Creating training corpora for NLG micro-planners. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 179–188. Association for Computational Linguistics (2017)
9. Guan, X., Liu, Y., Lin, H., Lu, Y., He, B., Han, X., Sun, L.: Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. CoRR **abs/2311.13314** (2023), https://doi.org/10.48550/arXiv.2311.13314
10. Hennig, L., Thomas, P., Möller, S.: MultiTACRED: A multilingual version of the TAC relation extraction dataset. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3785–3801. Association for Computational Linguistics (2023)

11. Josifoski, M., Sakota, M., Peyrard, M., West, R.: Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 1555–1574. Association for Computational Linguistics (Dec 2023)
12. Lee, J., Lee, M.J., Yang, J.Y., Yang, E.: Does it really generalize well on unseen data? systematic evaluation of relational triple extraction methods. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3849–3858. Association for Computational Linguistics (2022)
13. Liang, J., He, Q., Zhang, D., Fan, S.: Extraction of joint entity and relationships with soft pruning and globalpointer. Applied Sciences **12**(13) (2022). https://doi.org/10.3390/app12136361, https://www.mdpi.com/2076-3417/12/13/6361
14. Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., Hu, G.: Ekt: Exercise-aware knowledge tracing for student performance prediction. IEEE Transactions on Knowledge and Data Engineering **33**(1), 100–115 (2019)
15. Liu, Y., Zhang, K., Gan, A., Yue, L., Hu, F., Liu, Q., Chen, E.: Empowering few-shot relation extraction with the integration of traditional re methods and large language models. In: International Conference on Database Systems for Advanced Applications. pp. 349–359. Springer (2024)
16. Nayak, T., Majumder, N., Goyal, P., Poria, S.: Deep neural approaches to relation triplets extraction: A comprehensive survey. Cognitive Computation **13**, 1215–1232 (2021)
17. Ning, J., Yang, Z., Sun, Y., Wang, Z., Lin, H.: OD-RTE: A one-stage object detection framework for relational triple extraction. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11120–11135. Association for Computational Linguistics (2023)
18. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
19. Papaluca, A., Krefl, D., Rodríguez Méndez, S., Lensky, A., Suominen, H.: Zero- and few-shots knowledge graph triplet extraction with large language models. In: Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024). pp. 12–23. Association for Computational Linguistics (2024)
20. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
21. Ren, F., Zhang, L., Yin, S., Zhao, X., Liu, S., Li, B., Liu, Y.: A novel global feature-oriented relational triple extraction model based on table filling. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 2646–2656. Association for Computational Linguistics (2021)
22. Ren, F., Zhang, L., Zhao, X., Yin, S., Liu, S., Li, B.: A simple but effective bidirectional framework for relational triple extraction. Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (2021)
23. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21. pp. 148–163. Springer (2010)
24. Shang, Y.M., Huang, H., Mao, X.: Onerel: Joint entity and relation extraction with one module in one step. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 11285–11293 (2022)

25. Shang, Y.M., Huang, H., Sun, X., Wei, W., Mao, X.L.: Relational triple extraction: One step is enough. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 4360–4366. International Joint Conferences on Artificial Intelligence Organization (2022), main Track
26. Sui, D., Zeng, X., Chen, Y., Liu, K., Zhao, J.: Joint entity and relation extraction with set prediction networks. IEEE Transactions on Neural Networks and Learning Systems pp. 1–12 (2023). https://doi.org/10.1109/TNNLS.2023.3264735
27. Tang, W., Xu, B., Zhao, Y., Mao, Z., Liu, Y., Liao, Y., Xie, H.: UniRel: Unified representation and interaction for joint relational triple extraction. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 7087–7099. Association for Computational Linguistics (2022)
28. Wang, F., Liu, Q., Chen, E., Huang, Z., Yin, Y., Wang, S., Su, Y.: Neuralcd: a general framework for cognitive diagnosis. IEEE Transactions on Knowledge and Data Engineering **35**(8), 8312–8327 (2022)
29. Wang, Y., Yu, B., Zhang, Y., Liu, T., Zhu, H., Sun, L.: TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 1572–1582. International Committee on Computational Linguistics (2020)
30. Wei, Z., Su, J., Wang, Y., Tian, Y., Chang, Y.: A novel cascade binary tagging framework for relational triple extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1476–1488. Association for Computational Linguistics (2020)
31. Yuan, Z., Vlachos, A.: Zero-shot fact-checking with semantic triples and knowledge graphs. arXiv preprint arXiv:2312.11785 (2023)
32. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). pp. 71–78. Association for Computational Linguistics (2002)
33. Zhang, K., Liu, Q., Qian, H., Xiang, B., Cui, Q., Zhou, J., Chen, E.: Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering **35**(1), 377–389 (2021)
34. Zhang, K., Zhang, H., Liu, Q., Zhao, H., Zhu, H., Chen, E.: Interactive attention transfer network for cross-domain sentiment classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5773–5780 (2019)
35. Zhao, K., Xu, H., Cheng, Y., Li, X., Gao, K.: Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction. Knowledge-Based Systems p. 106888 (2021). https://doi.org/https://doi.org/10.1016/j.knosys.2021.106888, https://www.sciencedirect.com/science/article/pii/S0950705121001519
36. Zheng, H., Wen, R., Chen, X., Yang, Y., Zhang, Y., Zhang, Z., Zhang, N., Qin, B., Ming, X., Zheng, Y.: PRGC: Potential relation and global correspondence based joint relational triple extraction. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 6225–6235. Association for Computational Linguistics (2021)
37. Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., Xu, B.: Joint extraction of entities and relations based on a novel tagging scheme. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1227–1236. Association for Computational Linguistics (2017)