# Discover rather than Memorize: A Novel Benchmark for Relational Triple Extraction

**Anonymous ACL submission**


## Abstract

Relational Triple Extraction (RTE), one of the crucial components of information extraction, has experienced rapid development in recent years. However, due to the triple duplication problem in existing datasets, previous methods often yield highly competitive results by simply memorizing the duplicated triples rather than discovering the new triples from raw text. Specifically, In the two most widely-used datasets (NYT and WebNLG), more than 80% of the triples from the test set are direct duplicates of triples already present in their training set. In response to this, we propose a new dataset, named **ENT**, to evaluate the model's ability to **E**xtract **N**ew **T**riples, which aligns more coherently with the objectives of the RTE task. Specifically, based on the Wikidata knowledge graph slices and Large Language Model Prompting, we design an RTE dataset construction pipeline. It consists of four steps, including: 1) Preprocess, 2) Paragraph Generation, 3) Rule-based Check and 4) Semantic Check. ENT comprises 300k+ unique triples with all the test set samples containing at least one new triple. We conduct a re-evaluation of nine existing state-of-the-art methods and observe a generalized 10%+ and 7.5%+ decrease in extraction accuracy on ENT compared to NYT and WebNLG respectively. This demonstrates that ENT is a more challenging and meaningful benchmark, and we hope it will lead to new directions in the study of the RTE task.


## 1 Introduction

Relation 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 (Nayak et al., 2021). In the field of information extraction, RTE is a crucial task and serves as a bridge between the unstructured human language and triple-structured explicit knowledge in knowledge graphs.

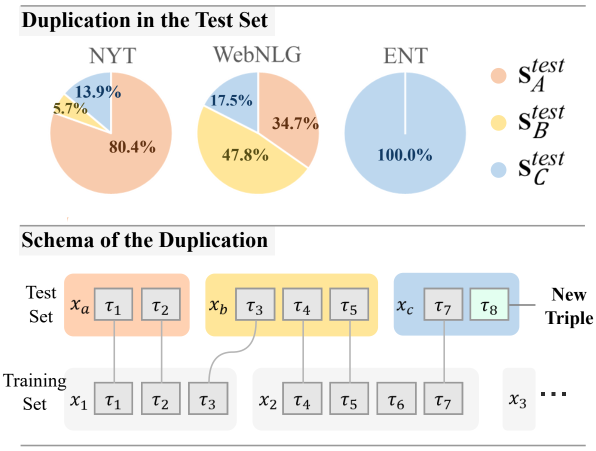


Figure 1: Triple duplication in NYT, WebNLG and ENT. $\mathbf{S}_A^{test}$ exhibits the highest degree of triple deplication, followed by $\mathbf{S}_B^{test}$. $\mathbf{S}_C^{test}$ contains new triples. $x_a$, $x_b$, and $x_c$ are schematic illustrations of the duplicated samples in $\mathbf{S}_A^{test}$,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

Early researches decomposed the RTE into two components: entity identification and relation categorization (Zelenko et al., 2002; Chan and Roth, 2011). Chan and Roth (2011) firstly recognized the entities, and then extracted the relation for each entity pair. In recent years, researchers have increasingly focused on the deep connection between entities and relationships (Wei et al., 2020; Zheng et al., 2021; Ren et al., 2021b; Wang et al., 2020; Tang et al., 2022). Among them, Wang et al. (2020) initially implemented a one-step triple extraction by conceptualizing RTE as a table-filling task; Tang et al. (2022) 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, we identify a significant potential flaw of triple duplication within the existing benchmark for

RTE. According to our calculation, more than 80% of the triples in the test set of NYT and WebNLG are duplicates. 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 primarily focus on evaluating the model's ability to memorize existing triples, rather than discovering 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 Knowledge Graphs (KGs) (Dong et al., 2014; Nayak et al., 2021), the existing benchmarks of RTE exhibit a significant gap due to their 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 develop a new benchmark dataset, ENT. The pipeline consists of four steps: 1) *Process* that performs irrelevant triple filtering in the collected and clustered knowledge base. 2) *Paragraph Generation* by prompting to the Large Language Model (LLM). 3) *Rule-based Check* that identifies and rectifies the unconforming paragraphs. 4) *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 (Riedel et al., 2010) and WebNLG (Gardent et al., 2017). We conducted a more thorough analysis on ENT and revealed a lower propensity for bias towards duplicated triples of ENT. It demostrates that ENT serves as a more challenging and meaningful benchmark from the perspective of discovering new triples. We plan to open-source the complete ENT dataset in the near future and hope it will lead to new directions of RTE study in the future.

## 2 Related Work

### 2.1 Relational Triple Extraction

Some of the RTE study 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 (Zheng et al., 2017; Fu et al., 2019; Liang et al., 2022; Zhao et al., 2021). Other works propose more realistic frameworks for exact-match extraction, which stipulate that all entities must be extracted in their entirety. Wei et al. (2020) proposed a two-stage triple extraction scheme, which successfully addressed a significant number of overlapped entities for the first time. Sui et al. (2023) treated RTE as an ensemble prediction problem and employed a non-autoregressive decoder. Wang et al. (2020) initially conceptualized the RTE task as a table filling problem. Ren et al. (2021b) proposed a straightforward and efficient approach to RTE by implementing a bi-directional extraction framework. Shang et al. (2022a) further simplified the labeling strategy and decoding method of table filling for RTE. Tang et al. (2022) proposed a novel unified entity-relation interaction modeling approach. Shang et al. (2022b) devised a method for entity extension matching, though at the cost of significantly increasing the text sequence length. Papaluca et al. (2023) attempted to utilize LLMs for direct few-shot triple extraction but observed that the LLM struggled to attain competitive performance with classical baseline models.

### 2.2 RTE Dataset

NYT (Riedel et al., 2010) and WebNLG (Gardent et al., 2017) are the most widely used datasets for RTE at present. NYT was constructed by remote supervised relation extraction. It is contains noisy samples and has a limited number of relations. WebNLG employed native English speakers to write text descriptions for relational triples and got a dataset of limited size. With the rapid development of deep learning techniques in the field of natural language processing, it is becoming increasingly acceptable to use machines for data annotation. For example, Hennig et al. (2023) used machine translation models to build a multilingual relation extraction dataset.

### 2.3 Large Language Model for Text Generation

LLMs, with increasing model parameters and extensive training corpora, have demonstrated extraordinary capabilities of text generation (Radford et al., 2019; Brown et al., 2020). By incorporating human feedback into large language models, it is possible to generate outputs that are more aligned with human preferences (Ouyang et al.,

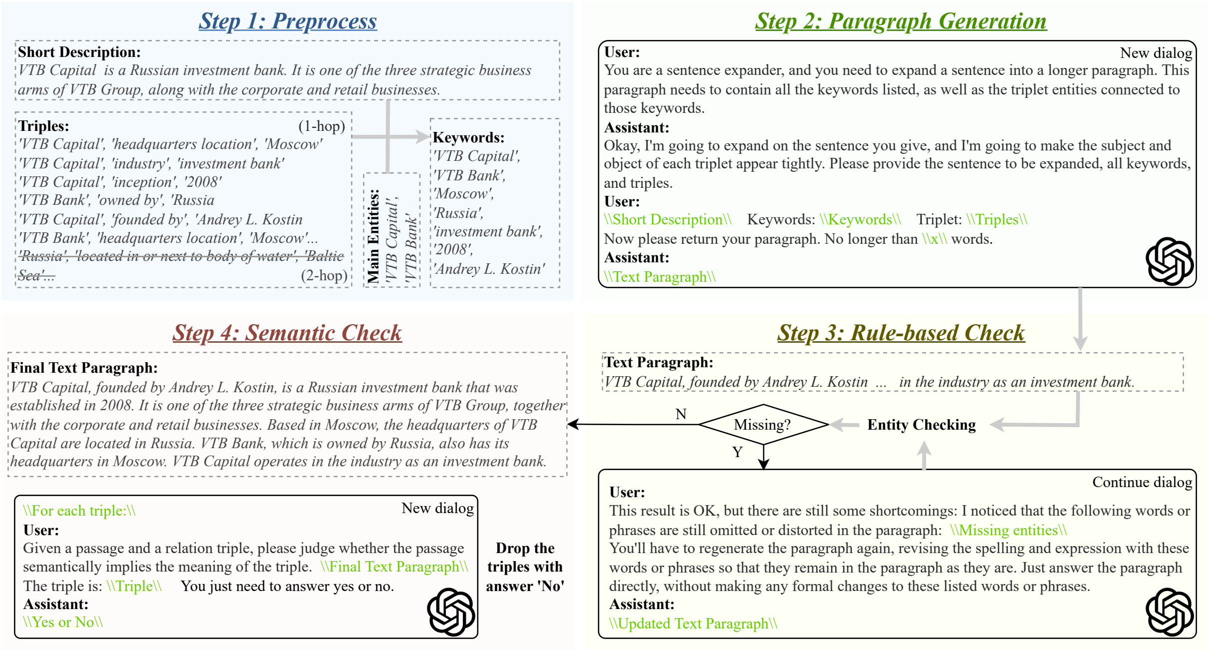


Figure 2: The process of constructing the ENT dataset with detailed content of prompt. The content between the pair of \\green slashes\\ are the comment for the variable transferred in the dialog with LLM.

2022). Concurrently, the content of the KG can significantly mitigate the hallucination issue of large language models (Guan et al., 2023; Yuan and Vlachos, 2023). Zero-shot automatic text generation via LLM with factual triples has demonstrated competitive performance (Axelsson and Skantze, 2023; Xu et al., 2023). In this work, we utilize the triples from a real-world KG to instruct the LLM for the development of an RTE dataset.

# 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 $[w_i, w_{i+1}, ..., w_j]$ $(1 \leqslant i \leqslant j \leqslant L)$ extracted from the input sentence. The number of triples $N$ per sentence may be greater than 1, while the exact number is not known in advance. The input consists of simple raw text, which does not contain explicit knowledge (e.g., entity information).

## 3.2 Dataset Construction Pipeline

Constructing an RTE dataset requires the collection of text-triples sample pairs. We notice that Cheng et al. (2020) has gathered a substantial number of entities from Wikipedia and Wikidata, along with the relational triples, to construct a dataset called ENT-DESC for KG-based concise national language generation. However, the original textual description is too short and insufficiently detailed for RTE, failing to encompass all the triples associated with the main entities. The original dataset is open-sourced for research.

To actualize the text construction, we utilized OpenAI's GPT-3.5-Turbo API [1] as the LLM for automatic text generation. The objective of the LLM is to generate a longer 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 divided into four steps :*Preprocess*, *Paragraph Generation*, *Rule-based Check* and *Semantic Check*.

### 3.2.1 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. For example, as shown

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

in Figure 2, the 2-hop triple <*'Russia', 'located in or next to body of water', 'Baltic Sea'*>, only connected to the entity *'Russia'*, is not directly related to either of the two main entities, *'VTB Capital'* or *'VTB Bank'*. We retain the 200 relationships with the highest frequency of occurrence. Each relation has at least 20 unique triples.

#### 3.2.2 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 information contained within 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 implemente a straightforward soft-limit policy by instructing the LLM to generate paragraphs no longer than $x$ words. $x = 8N + 4$, where $N$ means the number of triples in a sample. It is essential to highlight both the keywords and triples explicitly: the de-emphasis of keywords may result in more missing entities, while the de-emphasis of triples can lead to semantic distortion in the generated paragraph.

#### 3.2.3 Rule-based Check

Although the keyword- and triple-based prompt enables the LLM to generate more accurate paragraph, it runs the risk of syntactic reconstruction or entity content re-expression, potentially disrupting the original entity structure. In this step, we use a direct rule-based method to check if the original entity is missing from the paragraph. We use a BERT-base-cased (Devlin et al., 2018) tokenizer to tokenize the text paragraph and all the entities. If both the entity's string and token id sequence can be matched within the paragraph, we deem the entity to be rule-compliant for RTE extraction. 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 can be successfully matched. We discard the sample with the token [UNK] or $\geqslant$1 missing entities after the third dialog.

#### 3.2.4 Semantic Check

Not all the paragraphs that pass the entity matching check fully encompasses the entity and relationship information expressed by the triples. In this step, we reinitiate a new dialog with the LLM to ascertain 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 evidenced by the introduction of human opinions on a smaller subset of samples, which is introduced in Appendix A.

### 3.3 ENT Dataset

We collect all the samples that underwent the 4 steps and obtain 62,609 English paragraphs with 347,452 unique exact-match triples overall. The domains of the triples include humans, events, locations and organizations. We divide the entire dataset into the training set(~80%), validation set(~10%) and test set(~10%) in the original order. Note that the relational triples in the dataset are identified as generalized, potentially including attribute triples that also comply with the formulated definition of the RTE in Section 3.1. For instance, triples with relations such as *'date of birth'* and *'start time'* would be considered.

Futher more, every sample in the ENT test or validation set contains new triples (as shown in Figure 1). This feature is achieved without altering the distribution of data. The reason can be attributed to two factors. 1) The main entities of the original triple groups were derived and clustered from PageRank scores, demonstrating strong topic independence. 2) We discard the 2-hop triples, further reducing the triple duplication between different samples. ENT, with over 60% proportion of new triples in test set, is a more persuasive benchmark for evaluating the methods' ability to **E**xtract **N**ew **T**riples. We name this dataset **ENT**. In contrast, the new triples in the test sets of both NYT and WebNLG comprise only ~10%.

The detailed statistical information of ENT and the other existing datasets are presented in Table 1. ENT has a comparable sample size to NYT but contains a larger number of relations, longer text, and a greater quantity of triples in each sample. The mini-KG size is determined by counting the number of all the unique triples, which can serve as a rough representation of the scope of knowledge encompassed by the dataset. ENT has made significant strides in this metric.

The assessment of new knowledge discovery has not been clearly defined, particularly when con-

| Dataset | Train | Valid | Test | Relations | Mini-KG Size | $\mu_N$ | $\mu_{F(\tau)}$ | $N'_{test}/N_{test}$ |
|---|---|---|---|---|---|---|---|---|
| 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 1: ENT vs. NYT and WebNLG. $\mu_N$ 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 $\tau$ 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.

| | Category | Number |
|---|---|---|
| t1 | $N'/N < 0.2$ | 242 |
| t2 | $0.2 \leqslant N'/N < 0.4$ | 1127 |
| t3 | $0.4 \leqslant N'/N < 0.6$ | 1265 |
| t4 | $0.6 \leqslant N'/N < 0.8$ | 1147 |
| t5 | $0.8 \leqslant N'/N < 1.0$ | 796 |
| t6 | $N'/N = 1.0$ | 1481 |
| e1 | $E' = 0$ | 461 |
| e2 | $E' = 1$ | 2357 |
| e3 | $E' = 2$ | 1366 |
| e4 | $E' = 3$ | 837 |
| e5 | $E' = 4$ | 513 |
| e6 | $E' \geqslant 5$ | 524 |
| r1 | $R_m < 10$ | 1490 |
| r2 | $10 \leqslant R_m < 25$ | 1077 |
| r3 | $25 \leqslant R_m < 50$ | 1233 |
| r4 | $50 \leqslant R_m < 75$ | 667 |
| r5 | $75 \leqslant R_m < 100$ | 636 |
| r6 | $R_m \geqslant 100$ | 955 |

Table 2: Categories from different perspective of the intensity of the new knowledge for ENT test set. $N'$, $N$ and $E'$ denote the number of new triples, all triples and new unique entities in each sample. $R_m$ denotes the max oridinal number of the relations in each sample.

sidering the intensity of the new knowledge. Nevertheless, we endeavor to provide three intuitive perspectives for quantitative evaluation. Table 2 illustrates the three perspectives of the category. From the perspective of triples, we categorize the test set by the proportion of new triples in each sample (t1-t6), a significant intuitive indicator to gauge the intensity of new knowledge. For the entities, we implement the division in terms of the number of new unique entities of each sample (e1-e6). For the relations, we sorted all the relations by the frequency of occurrence in descending order and and assign a unique ordinal number to each relation (from 0 to 199). A higher ordinal number indicates a less common relationship. We perform the division on the test set in terms of the maximum relation ordinal number in each sample (r1-r6).

## 4 RTE Experiment Setups

We select 9 state-of-the-art RTE methods for our reassessment: **CasRel** (Wei et al., 2020), **SPN4RE** (Sui et al., 2023), **TPLinker** (Wang et al., 2020), **PRGC** (Zheng et al., 2021), **GRTE** (Ren et al., 2021a), **BiRTE** (Ren et al., 2021b), **OneRel** (Shang et al., 2022a), **UniRel** (Tang et al., 2022), and **OD-RTE** (Ning et al., 2023). For each method, we create and configure a specific miniconda environment based on the packages and their versions indicated in the respective source code. We initialize all the models with the pretrained BERT-base-cased weights, which are widely cited as beneficial. We test each model on the checkpoint with the highest validation F1 score and set batch size = 1 for inference. We uniformly evaluate the triples in the format of <subject, relation, object>.

For NYT and WebNLG benchmark, we focus on the exact-match version as it more closely aligns with the real-world RTE applications. In certain scenarios requiring model retraining, we utilize publicly available source code and the optimal hyperparameter configurations cited in the original paper to train the model.

ENT is also exact-matched. For the training of ENT, we separately utilize the optimal parameters of each method reported on NYT due to the comparable sample sizes of the two. Appendix 8 list some of them. We synchronize and pre-tune the data format for specific methods, given the separate code requirements. 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, we utilize a concise automatic tokenizing strategy: If a relation's first or last

| Method | NYT | | | WebNLG | | | ENT (Ours) | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| CasRel (Wei et al., 2020) | 89.8* | 88.2* | 89.0* | 88.3* | 84.6* | 86.4* | 73.8 | 54.2 | 62.2 |
| SPN4RE (Sui et al., 2023) | 92.5 | 92.2 | 92.3 | 85.7* | 82.9* | 84.3* | 78.3 | 76.6 | 77.4 |
| TPLinker (Wang et al., 2020) | 91.4 | 92.6 | 92.0 | 88.9 | 84.5 | 86.7 | 70.7 | 75.3 | 72.9 |
| PRGC (Zheng et al., 2021) | 93.5 | 91.9 | 92.7 | 89.9 | 87.2 | 88.5 | 72.4 | 74.2 | 73.3 |
| GRTE (Ren et al., 2021a) | 933.4 | 93.5 | 93.4 | 92.3 | 87.9 | 90.0 | **83.9** | 81.1 | **82.4** |
| BiRTE (Ren et al., 2021b) | 91.9 | 93.7 | 92.8 | 89.0 | 89.5 | 89.3 | 81.5 | 80.8 | 81.2 |
| OneRel (Shang et al., 2022a) | 93.2 | 92.6 | 92.9 | 91.8 | 90.3 | 91.0 | 81.9 | 79.7 | 80.8 |
| UniRel (Tang et al., 2022) | 93.7 | 93.2 | 93.4 | 91.8 | 90.5 | 91.1 | 78.9 | 80.8 | 79.8 |
| OD-RTE (Ning et al., 2023) | **94.2** | **93.6** | **93.9** | **92.8** | **92.1** | **92.5** | 78.7 | **81.9** | 80.3 |

Table 3: 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's are sourced from the respective original paper.

word can be tokenized into a single token that is not already occupied by another relation, it is used as the hint of the relation. Otherwise, the token is sequentially tokenized as [unuse x]. In addition, we set the maximum input length as 400 for all the methods.

## 5 Results and Analysis

### 5.1 Main Results

We present the overall accuracy of various RTE methods on ENT in Table 3, 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.

Furthermore, the performance of OD-RTE on ENT is slightly inferior to that of GRTE, despite the fact that OD-RTE was previously reported as a state-of-the-art method at present. We observe that OD-RTE, when performing tagging, training, and inference, identifies all the entities that appear multiple times in the text, regardless of their location. This lead to the aggressive decoding of a greater number of triples, notably enhanced by the larger quantity of triples contained by each sample in ENT (higher $\mu_N$ in Table 1). Besides, considering the data processing of CasRel is slightly outdated and lead to a bias in the content of the ENT entities, we only report its overall results just for general inference.

| Method | R-$\mathcal{T}^\circ$ | R-$\mathcal{T}'$ | R-$\mathcal{E}^\circ$ | R-$\mathcal{E}'$ |
|---|---|---|---|---|
| SPN4RE | 79.4 | 71.9 | 90.7 | 86.7 |
| TPLinker | 83.2 | 70.4 | 88.9 | 79.9 |
| PRGC | 78.2 | 71.6 | 90.2 | 84.8 |
| GRTE | 87.1 | 77.3 | 93.7 | 87.4 |
| BiRTE | 87.4 | 76.7 | 93.7 | 87.3 |
| OneRel | 86.0 | 75.9 | 93.4 | 86.3 |
| UniRel | 86.4 | 77.4 | 93.3 | 86.5 |
| OD-RTE | **87.9** | **78.2** | **94.2** | **88.1** |

Table 4: The recall (%) on the old ($\mathcal{T}^\circ$) and new ($\mathcal{T}'$) triples, as well as the old ($\mathcal{E}^\circ$) and new ($\mathcal{E}'$) entities.

### 5.2 Detailed Results of ENT

We observe and illustrate the alterations in the accuracy with various intensity of new knowledge from different perspectives in this section. Among the three perspectives introduced in Table 2, the most obvious correlation with extraction difficulty is observed in the proportion of new triples (t1-t6). It can be noted that almost all the methods exhibit a decline in accuracy as the proportion of new triples increases. Appendix B elaborate the detailed demonstration. Furthermore, the r1 subset, as delineated based on the frequency of relation occurrence (r1-r6), yield the highest scores for each method. This implies an intuitive assumption that it is easier for the model to extract knowledge with more common relations. In contrast, when viewed from the perspective of new entities (e1-e6), the performance exhibits more fluctuations. This suggests that the new entities may not adequately represent the intensity of new knowledge.

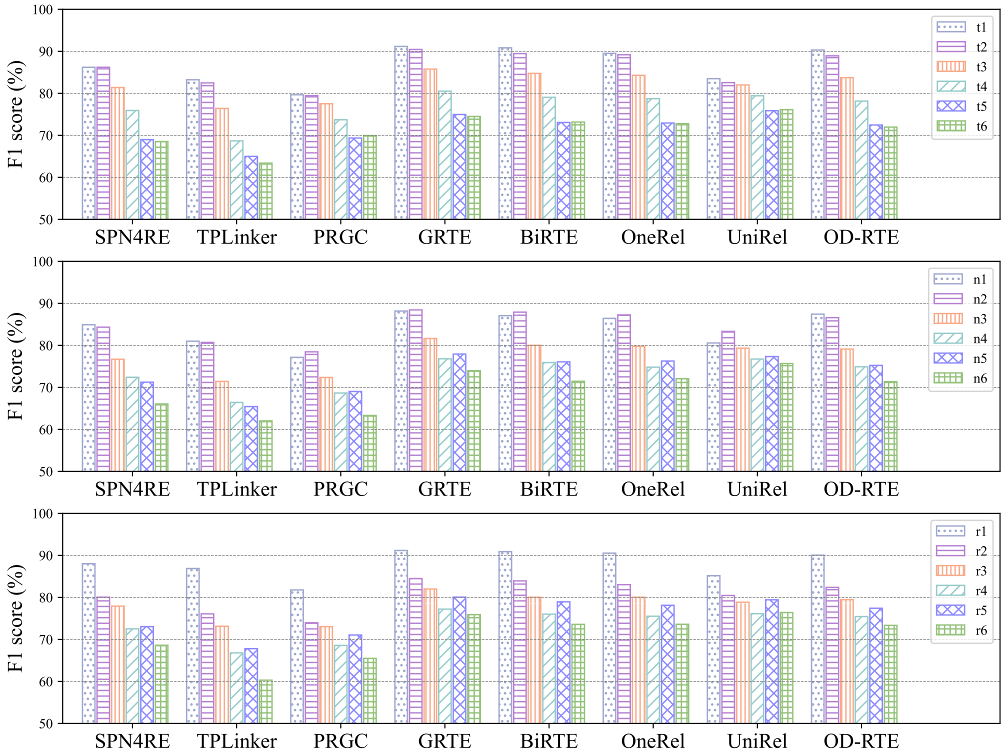


Figure 3: Specific triple micro F1 scores of RTE methods in three different perspectives of ENT test set. t1-t6 presents the different proportion of new triples in a sample. e1-e6 presents the different number of new unique entities. r1-r6 presents the max ordinal relation number in a sample. Category details are shown in Table 2.

We further report the recall of separate triples and entities within the test set as shown in Table 4. The recalls for new triples and entities are consistently lower than that of duplicate ones, which underscores the complexity of discovering new knowledge from another perspective. The significantly lower recall of triples compared to entities further indicates that accurately extracting entities accurately is insufficient for RTE, regardless of whether the knowledge is new or duplicated. We do not have precisions accurately reported from a similar perspective, as it is not feasible to categorize the error triples extracted.

### 5.3 Review on NYT and WebNLG

The review on NYT and WebNLG from the perspective of discovering new knowledge can similarly highlight the considerable difficulty in extracting new triples. Based on the degree of triple duplication shown in Figure 1, the NYT and WebNLG test sets can be sliced into three disjoint subsets, $\mathbf{S}_A^{test}$, $\mathbf{S}_B^{test}$ and $\mathbf{S}_C^{test}$. For each method, we conduct tests on each of the three subsets using the same checkpoints. Table 5 presents the performance on the separate three subsets. All the methods consistently demonstrate significantly low accuracy on $\mathbf{S}_C^{test}$, suggesting that this task is more challenging. In contrast, the highest accuracy is undoubtedly observed in the group $\mathbf{S}_A^{test}$ with the most duplication, where all the model jsut need to memorize the triples. $\mathbf{S}_B^{test}$ also exhibit high accuracy slightly trailing behind $\mathbf{S}_A^{test}$, implying that the arrangement and combination of knowledge present a lower degree of difficulty. As $\mathbf{S}_A^{test}$ and $\mathbf{S}_B^{test}$ hold an absolute majority in the test set, the model's ability to memorize duplicated triples primarily contributes to the high performance of existing benchmarks. In addition, although OD-RTE is currently reported as the overall state-of-the-art, it leads by a smaller margin and lags slightly on some indicators.

It is important to note that while each sample

| Dataset | Method | $S_A^{test}$ | | | $S_B^{test}$ | | | $S_C^{test}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 | P | R | F1 |
| NYT | CasRel | 93.3 | 95.2 | 94.2 | 82.9 | 77.0 | 79.8 | 68.8 | 56.3 | 61.9 |
| | SPN4RE | 94.6 | 97.2 | 95.9 | 90.6 | 85.6 | 88.0 | 71.8 | 66.4 | 69.0 |
| | TPLinker | 94.9 | 97.8 | 96.4 | 90.8 | 88.4 | 89.6 | 71.7 | 64.1 | 67.7 |
| | PRGC | 95.5 | 97.0 | 96.3 | 90.8 | 85.7 | 88.2 | 71.2 | 62.6 | 66.6 |
| | GRTE | **96.2** | **98.4** | **97.2** | 93.9 | **90.2** | 92.0 | 72.7 | 65.7 | 69.0 |
| | BiRTE | 95.7 | 96.8 | 96.3 | 92.9 | 89.2 | 91.0 | 71.2 | 63.1 | 66.9 |
| | OneRel | 94.7 | 97.8 | 96.2 | 92.6 | 89.8 | 91.2 | 68.2 | 64.2 | 66.1 |
| | UniRel | 96.0 | 98.3 | 97.1 | **94.3** | 90.0 | **92.1** | **73.2** | 64.5 | 68.6 |
| | OD-RTE | 96.1 | 98.0 | 97.0 | 92.7 | 89.6 | 91.2 | 71.4 | **68.4** | **69.9** |
| WebNLG | CasRel | 92.8 | 94.2 | 93.5 | 89.4 | 88.4 | 88.9 | 67.7 | 55.3 | 60.9 |
| | SPN4RE | 92.5 | 94.2 | 93.3 | 93.3 | 92.2 | 92.8 | 72.7 | 63.2 | 67.6 |
| | TPLinker | 90.6 | 95.3 | 92.9 | 90.9 | 89.6 | 90.3 | 78.9 | 65.8 | 71.8 |
| | PRGC | 92.5 | 93.9 | 93.2 | 93.5 | 88.9 | 91.1 | 71.8 | 65.8 | 68.7 |
| | GRTE | 94.5 | 96.7 | 95.6 | 93.4 | 91.9 | 92.7 | **83.4** | 65.5 | 73.5 |
| | BiRTE | 92.5 | 95.7 | 94.1 | 92.9 | 91.3 | 92.1 | 74.6 | 68.7 | 71.5 |
| | OneRel | 93.1 | 96.0 | 94.5 | 93.3 | 91.5 | 92.4 | 77.7 | 66.5 | 71.7 |
| | UniRel | **94.6** | 95.5 | 95.1 | 93.0 | 93.0 | 93.0 | 79.4 | 70.5 | 74.7 |
| | OD-RTE | 94.5 | **97.9** | **96.2** | **94.4** | **94.2** | **94.3** | 80.4 | **71.1** | **75.5** |

Table 5: The precision (P), recall (R), and micro F1 scores (F1) (%) for the three divided subsets of test from NYT and WebNLG. The metrics are colored for ease of comparison.

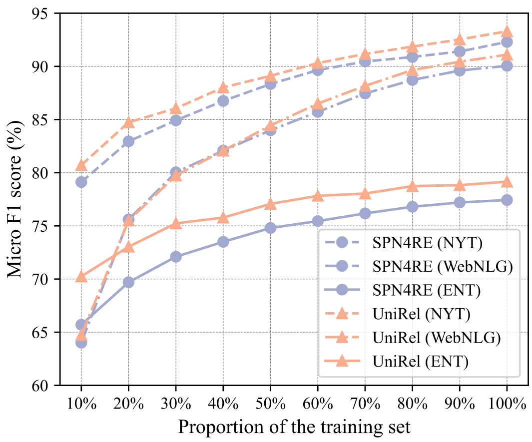


Figure 4: Accuracy of different training data volume.

in $S_C^{test}$ also consist of new triples, $S_C^{test}$ of NYT and WebNLG cannot serve as a direct benchmark for new knowledge discovery evaluation, as the slicing process of the three subsets results in different distributions. A higher degree of duplication can hinder the discovery of new knowledge by the model, which is discussed in Appendix C.

### 5.4 Data Volume

We also tested the adequacy of the data volume by randomly reducing the size of the training set for NYT/WebNLG/ENT to 10%-90% and executing training operations separately. The results in Figure 4 demonstrates that the marginal impact of increasing the size of the training set on model performance enhancement is already apparent. NYT and ENT grow more gradually than WebNLG. It implies a sufficient volume of the ENT dataset.

## 6 Conclusion

In this paper, we propose a new benchmark, ENT, for Relation Triple Extraction. The dataset is developed based on factual Knowledge Graph slices and Large Language Model Prompting. ENT offers a more accurate representation of the model's ability to discover new triples compared to the existing benchmarks. Following extensive experiments on 9 advanced prior works, ENT is found to be more challenging than the other two benchmarks. Besides, we have identified a positive correlation between extraction difficulty and the intensity of new knowledge. We will open-source the complete ENT dataset in the near future.

## 7 Limitations

We discuss the limitations of this work in two aspects.

- Despite the significant improvement in authenticity achieved through the KG & LLM-based national language generation, the word usage patterns of LLM may differ from those of humans. LLM may lead to convergence of language styles for the paragraphs as well. This may result in stylistic shifts in the generated text of ENT. Furthermore, although we conduct close triple accuracy checks on the generated passages, there may be unanticipated triples in the paragraph, leading to a degree of noise. We intend to implement language style evaluation strategies and continue to identify potential triples in the future.

- The relationships within our dataset do not align semantically with existing datasets, hindering the sharing or transfer of knowledge across different RTE datasets. In fact, there is often a lack of semantic alignment in the relations between different pre-existing datasets. We are currently exploring methods for semantic alignment across datasets in RTE tasks.

## 8 Ethics Statement

We use the data of the ENT-DESC dataset “as is”. Although we regarded some of the samples during the construction of the dataset, we did not inplement a specialized bias filtering mechanism. The new dataset may thus reflect biases of the original dataset. The authors of the original dataset (Cheng et al., 2020) have not stated measures that prevent collecting sensitive text. Throughout the dialog with the LLM API, we did not coerce, induce, or suggest that LLM generated harmful or biased content. However, we did not implement a specialized detection component to manage the content of conversations returned by the LLM. Therefore, we do not rule out the possible risk of sensitive content in the data.

The RTE experiments were conducted on a computer equipped with an Intel(R) Xeon(R) Platinum 8350C CPU, 56 GB of RAM, and one NVIDIA GeForce RTX 3090. The average time required for a complete training and testing process on the ENT dataset is approximately 35 hours. For each method’s experiments on ENT, we set 5 different random seeds to train the model five times. We choose the group with the median micro F1 score for accuracy report.

## References


Agnes Axelsson and Gabriel Skantze. 2023. 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)*, pages 39–54, Prague, Czech Republic. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 551–560, Portland, Oregon, USA. Association for Computational Linguistics.

Liying Cheng, Dekun Wu, Lidong Bing, Yan Zhang, Zhanming Jie, Wei Lu, and Luo Si. 2020. ENT-DESC: Entity description generation by exploring knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1187–1197, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. 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*, pages 601–610.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. 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*, pages 1409–1418, Florence, Italy. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2023. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. *CoRR*, abs/2311.13314.

Leonhard Hennig, Philippe Thomas, and Sebastian Möller. 2023. 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)*, pages 3785–3801, Toronto, Canada. Association for Computational Linguistics.

Jianming Liang, Qing He, Damin Zhang, and Shuangshuang Fan. 2022. Extraction of joint entity and relationships with soft pruning and globalpointer. *Applied Sciences*, 12(13).

Tapas Nayak, Navonil Majumder, Pawan Goyal, and Soujanya Poria. 2021. Deep neural approaches to relation triplets extraction: A comprehensive survey. *Cognitive Computation*, 13:1215–1232.

Jinzhong Ning, Zhihao Yang, Yuanyuan Sun, Zhizheng Wang, and Hongfei Lin. 2023. 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)*, pages 11120–11135, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Andrea Papaluca, Daniel Krefl, Sergio Mendez Rodriguez, Artem Lensky, and Hanna Suominen. 2023. Zero-and few-shots knowledge graph triplet extraction with large language models. *arXiv preprint arXiv:2312.01954*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021a. 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*, pages 2646–2656, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Feiliang Ren, Longhui Zhang, Xiaofeng Zhao, Shujuan Yin, Shilei Liu, and Bochao Li. 2021b. A simple but effective bidirectional framework for relational triple extraction. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. 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*, pages 148–163. Springer.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022a. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11285–11293.

Yu-Ming Shang, Heyan Huang, Xin Sun, Wei Wei, and Xian-Ling Mao. 2022b. Relational triple extraction: One step is enough. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4360–4366. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Dianbo Sui, Xiangrong Zeng, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Joint entity and relation extraction with set prediction networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12.

Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. 2022. UniRel: Unified representation and interaction for joint relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7087–7099, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online. Association for Computational Linguistics.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 190–200, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Zhangdie Yuan and Andreas Vlachos. 2023. Zero-shot fact-checking with semantic triples and knowledge graphs. *arXiv preprint arXiv:2312.11785*.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 71–78. Association for Computational Linguistics.

Kang Zhao, Hua Xu, Yue Cheng, Xiaoteng Li, and Kai Gao. 2021. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction. *Knowledge-Based Systems*, page 106888.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. 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)*, pages 6225–6235, Online. Association for Computational Linguistics.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. 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)*, pages 1227–1236, Vancouver, Canada. Association for Computational Linguistics.

## A Human Verification

We conducted multiple manual validations on a random sample of 100 entries from the final sample set in Section 4.1. The objective was to verify whether the triples were semantically accurately incorporated into the returned text passages. We hired five part-time annotators to provide five distinct feedbacks on the 100 samples. All of them hold a bachelor's degree or higher and don't know the full extent of this work. Each annotator was tasked with verifying each triple in each sample. A triple is considered semantically accurate from an artificial perspective when its meaning is accurately reflected in the text, as shown in Figure 5.

All the annotator were told the data would be collected for evaluating the quality of a machine-generated dataset. We remunerated the annotators at an amount higher than the local minimum income standard.

Based on the human feedback, our data construction process yielded an average semantic accuracy of 94.8%. This suggests that our dataset exhibits low semantic noise.

| Annotator | Corr. Triples (%) |
|---|---|
| 1 | 96.6 |
| 2 | 92.5 |
| 3 | 93.3 |
| 4 | 95.8 |
| 5 | 95.9 |
| Avg. | **94.8** |

Table 6: Human verification accuracy of the triples. Annotator 1-3 live in Asia, 4-5 live in North America. All the remunerations exceed the local minimum wage.

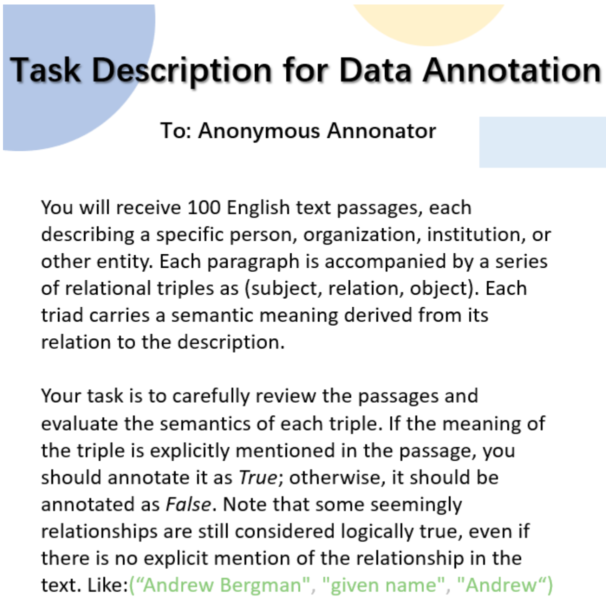


Figure 5: Task description to the anonymous annonator.

## B View of the New Triple Proportion in ENT

In this section, we illustrate more intuitively how the accuracy of each sample correlates with the proportion of new triples contained within them. Each subplot in Figure 6 represents a distinct RTE method. As the percentage of new triples continues to increase, more samples with lower extraction accuracy rates appear, while samples with high accuracy remains.

## C Detailed Analysis for Triple Duplication

In this section, we conduct an experiment to examine the impact of duplication on the model's ability to discover new triples. We slice two training subset on NYT or WebNLG by different ap-

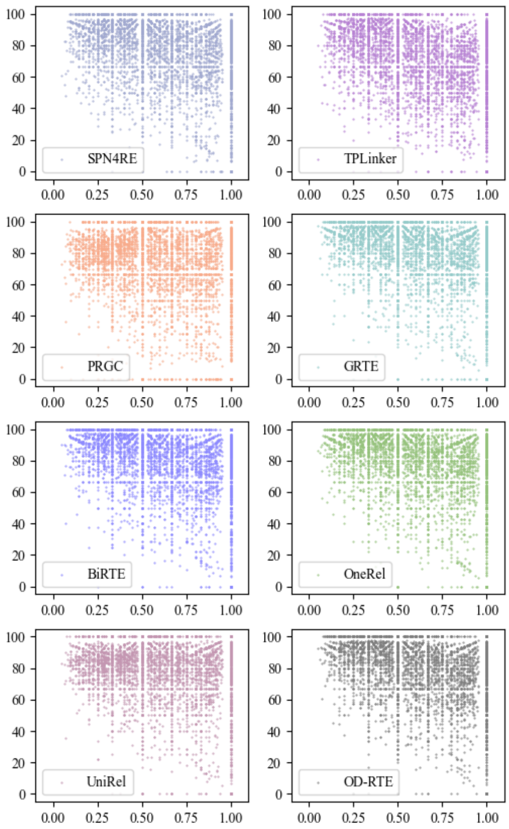


Figure 6: The overall view of triple extraction accuracy of the methods. The horizontal axis of each subplot represents the proportion of new triples in the sample, and the vertical axis represents the micro F1 score (%).

proach. We firstly filter the samples by detecting duplicate triples within the training set and obtain a subset $f$ such that it can just include all the unique triples. Samples with duplicate triples are discarded as much as possible. The second subset is randomly sliced to the same number of samples as the first one. We then randomly divide the training set into another subset $d$ with equal-sized samples as $f$. We set validation and test set as $\mathbf{S}_{C*}^{val}$ and $\mathbf{S}_{C*}^{test}$ respectively, where $\mathbf{S}_{C*}$ is the subset of $\mathbf{S}_C$ with $N' = N$ for each sample. The average number of occurrences ($\mu_{F(\tau)}$) of each unique triple in group $f$ is lower than that in group $d$. In this manner, regardless of how the training set is sliced, all the triples of the test set will be new ones.

Table 7 shows the accuracy with different training subset slices. For subset $d$, we use three different random seeds and get three different versions $d1, d2, d3$. It can be found that the accuracy is significantly higher in the group that we deliberately reduce the triple duplication. This implies that duplicated triples, even in the training set only, can diminish the model's tendency to uncover new triples.

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

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.

| | Learning Rate | Batch Size | Epoch |
|---|---|---|---|
| CasRel | 1e-5 | 6 | 100 |
| SPN4RE | 2e-5 for decoder<br>1e-5 doe encoder | 8 | 100 |
| TPLinker | 1e-5 | 6 | 100 |
| PRGC | 1e-3 | 64 | 100 |
| GRTE | 3e-5 | 6 | 50 |
| BiRTE | 3e-5 | 18 | 100 |
| OneRel | 1e-5 | 8 | 200 |
| UniRel | 3e-5 | 12 | 100 |
| OD-RTE | 5e-5 | 6 | 20 |

Table 8: Hyperparameters for model training on the ENT dataset

ENT dataset has a much lower $\mu_{F(\tau)}$ than NYT and WebNLG (as shown in Figure 2), which further enhances the effectiveness of our benchmark in assessing the discovery of new knowledge.

## D Hyperparameters for ENT Training

In this section, we list some of the hyperparameters for model training on the ENT dataset for all methods in Table 8. More details for each method can be found in the original paper and source code.