

# OneRel: Joint Entity and Relation Extraction with One Module in One Step

Yu-Ming Shang<sup>1</sup>, Heyan Huang<sup>1,2</sup>, Xian-Ling Mao<sup>1\*</sup>

<sup>1</sup>School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China

<sup>2</sup>Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing, China  
{ymshang, hhy63, maoxl}@bit.edu.cn

## Abstract

Joint entity and relation extraction is an essential task in natural language processing and knowledge graph construction. Existing approaches usually decompose the joint extraction task into several basic modules or processing steps to make it easy to conduct. However, such a paradigm ignores the fact that the three elements of a triple are interdependent and indivisible. Therefore, previous joint methods suffer from the problems of cascading errors and redundant information. To address these issues, in this paper, we propose a novel joint entity and relation extraction model, named OneRel, which casts joint extraction as a fine-grained triple classification problem. Specifically, our model consists of a scoring-based classifier and a relation-specific horns tagging strategy. The former evaluates whether a token pair and a relation belong to a factual triple. The latter ensures a simple but effective decoding process. Extensive experimental results on two widely used datasets demonstrate that the proposed method performs better than the state-of-the-art baselines, and delivers consistent performance gain on complex scenarios of various overlapping patterns and multiple triples.

## Introduction

Extracting pairs of entities and their relations in the form of (head, relation, tail) or  $(h, r, t)$  from unstructured text is an important task in natural language processing and knowledge graph construction. Traditional pipeline approaches (Zelenko, Aone, and Richardella 2003; Zhou et al. 2005; Chan and Roth 2011) treat entity recognition and relation classification as two separate sub-tasks. Although flexible, pipeline methods ignore the interactions between the two sub-tasks and are susceptible for the problem of error propagation (Li and Ji 2014). Therefore, recent studies focus on building joint models to obtain entities together with their relations through a unified architecture.

To make the complex task easy to conduct, existing studies usually decompose the joint extraction into several basic modules or processing steps (Yu et al. 2020; Zhao et al. 2021b). As shown in Figure 1, according to the extraction procedure of triple elements, these approaches fall into two categories: *multi-module multi-step* and *multi-module one-step*. The first category utilizes different modules in the

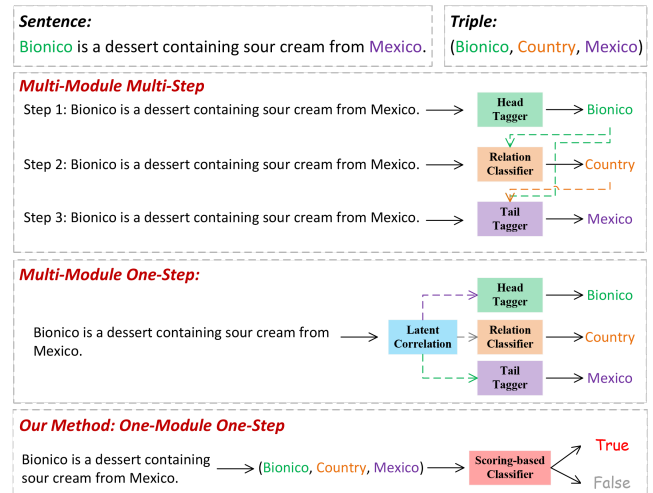


Figure 1: The extraction processes of existing approaches and our method. The dotted arrow indicates the dependencies between triple elements. Note that there are various extraction paradigms in *multi-module multi-step* approaches, e.g.,  $(h, t) \rightarrow r$ ,  $h \rightarrow r \rightarrow t$ ,  $r \rightarrow (h, t)$ . We just use  $h \rightarrow r \rightarrow t$  to illustrate the shortcomings of this kind of methods.

framework of cascading classification (Fu, Li, and Ma 2019; Yuan et al. 2020; Wei et al. 2020; Zheng et al. 2021; Zhao et al. 2021a,b) or text generation (Zeng et al. 2018; Zeng, Zhang, and Liu 2020; Ye et al. 2021) to obtain entities and relations step-by-step. Although promising, this kind of models suffers from the problem of cascading errors since mistakes in early steps may affect the prediction results of later steps. The second category attempts to identify entities and relations separately, and then combine them into triples based on their latent correlations (Wang et al. 2020; Sui et al. 2020; Wang et al. 2021). However, due to insufficient mutual constraints between entities and relations in the separate recognition process, such methods tend to produce redundant information, leading to errors when assembling triples (Zheng et al. 2017).

In fact, the fundamental reason for the above problems is that the decomposition-based paradigm ignores an important

\*Corresponding author.

property of a triple — its head entity, relation and tail entity are interdependent and indivisible. In other words, it is unreliable to extract one element without fully perceiving the information of the other two elements. To fill this gap, we try to accomplish the joint extraction task from the perspective of triple classification. For example, as shown in Figure 1, “Bionico” and “Mexico” are two words in the sentence and `Country` is a pre-defined relation, all of them are visible in training data. Intuitively, the triple (*Bionico*, `Country`, *Mexico*) can be directly identified by judging its correctness. The idea brings three advantages as follows. First, head entity, relation and tail entity are simultaneously fed into one classification module, making it possible to fully capture the dependencies between triple elements, thereby reducing redundant information. Second, only one-step classification is used, which is able to effectively avoid the cascading errors. Third, the simple architecture of *one-module one-step* empowers the network straightforward and easy to train.

Inspired by the above idea, in this paper, we propose a novel joint entity and relation extraction model, named OneRel, which is capable of extracting all triples from unstructured text with one module in one step. Considering that an entity may consist of multiple tokens, we design a scoring-based classifier and cast the joint extraction task into a fine-grained triple classification problem. Specifically, for a token pair  $(w_i, w_j)$  and a pre-defined relation  $r_k$ , the scoring-based classifier measures the correctness of the combination  $(w_i, r_k, w_j)$ , which will be assigned with a meaningful tag if it is valid and “-” otherwise. To this end, for an input sentence, the output of OneRel is a three-dimensional matrix with each entry corresponding to the classification result of  $(w_i, r_k, w_j)$ . In order to decode entities and relations from the output matrix accurately and efficiently, we introduce a novel relation-specific horns tagging (Rel-Spec Horns Tagging for short) strategy to determine the boundary tokens of head entities and tail entities. Experimental results on two widely used benchmark datasets prove that the proposed method outperforms previous approaches and achieves the state-of-the-art performance.

In summary, our contributions are as follows:

- We provide a novel perspective to transform joint extraction into fine-grained triple classification, making it possible to capture the information of head entities, relations and tail entities at the same time.
- Following our perspective, we introduce a novel scoring-based classifier and a Rel-Spec Horns Tagging strategy. The former is responsible for parallel tagging, and the latter ensures efficient decoding.
- We evaluate our model on two public datasets, and the results indicate that our method performs better than state-of-the-art baselines, especially for complex scenarios of overlapping triples.

## Related Work

Existing joint methods can be roughly divided into two classes according to their extraction procedure of triple elements:

The first class is *multi-module multi-step*, which uses different modules and interrelated processing steps to extract entities and relations serially. For example, a line of works first identify all entities in a sentence, and then perform relation classification between every entity pairs (Katiyar and Cardie 2017; Tan et al. 2019; Fu, Li, and Ma 2019; Liu et al. 2020). The second line of works first detect the relations expressed by a sentence rather than preserve all redundant relations; then head entities and tail entities are predicted (Zeng et al. 2018; Yuan et al. 2020; Zheng et al. 2021; Ma, Ren, and Zhang 2021). The third line of works first distinguish all head entities, and then inference corresponding relations and tail entities via sequence labeling or question answering (Wei et al. 2020; Yu et al. 2020; Zhao et al. 2021a,b; Ye et al. 2021). Despite their success, the *multi-module multi-step* methods suffer from the problem of cascading errors, as the mistakes in early steps cannot be corrected in later steps.

The second class is *multi-module one-step*, which extracts entities and relations in parallel, and then combines them into triples. For example, Miwa and Bansal (2016); Zhang, Zhang, and Fu (2017); Wang et al. (2020, 2021) treat entity recognition and relation classification as a table-filling problem, where each entry represents the interaction between two individual words. Sui et al. (2020) formulate the joint extraction task as a set prediction problem, avoiding considering the prediction order of multiple triples. However, due to insufficient mutual constraints between entities and relations in the separate recognition process, such *multi-module one-step* approaches cannot fully capture the dependencies between predicted entities and relations, resulting in redundant information during triple construction.

Different from existing methods, in this paper, we propose to treat the joint extraction task as a fine-grained triple classification problem, which is able to extract triples from sentences in a *one-module one-step* manner. Therefore, the aforementioned cascading errors and redundant information can be greatly addressed. Besides, the classical model Novel-Tagging (Zhang, Zhang, and Fu 2017) designs a complex tagging strategy to establish connections between entities and relations, and can also identify triples from sentences in one-step. Nevertheless, this technique cannot handle overlapping cases because it assumes every entity pair holds at most one relation.

## Method

In this section, we first give the task definition and notations. Then, we introduce our Rel-Spec Horns Tagging strategy and its decoding algorithm. Finally, we provide a detailed formalization of the scoring-based classifier.

### Task Definition

Given a sentence  $\mathcal{S} = \{w_1, w_2, \dots, w_L\}$  with  $L$  tokens and  $K$  predefined relations  $\mathcal{R} = \{r_1, r_2, \dots, r_K\}$ . The purpose of joint entity and relation extraction is to identify all possible triples  $\mathcal{T} = \{(h_i, r_i, t_i)\}_{i=1}^N$  in  $\mathcal{S}$ , where  $N$  is the number of triples,  $h_i, t_i$  are the head entity and tail entity composed of several consecutive tokens, i.e.,  $\text{entity.span} = w_{p:q}$ , where  $w_{p:q}$  refers to the concatenation of  $w_p$  to  $w_q$ .

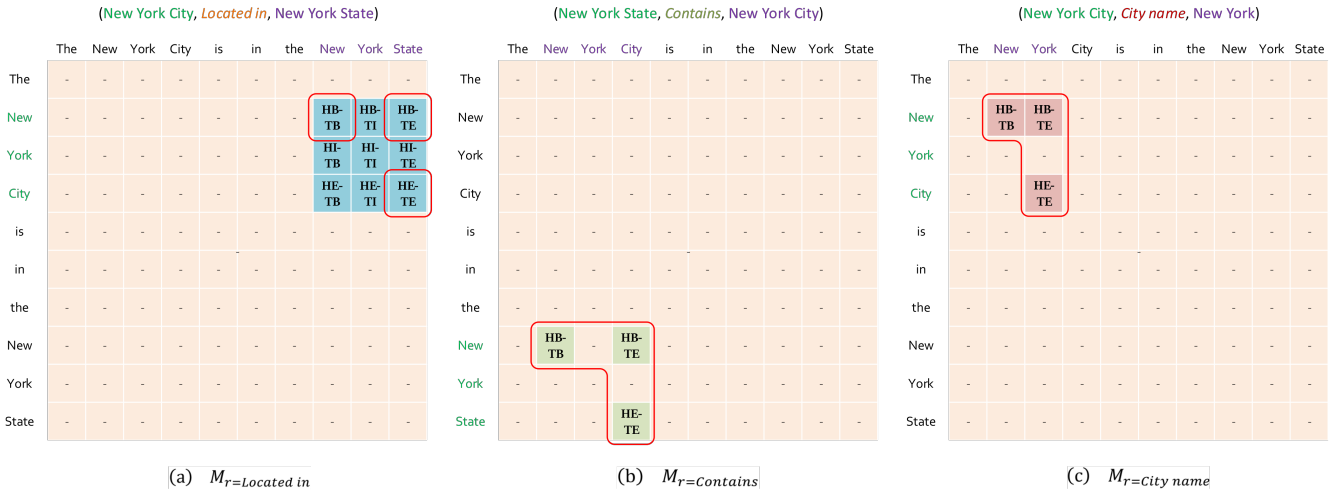


Figure 2: Examples of the Rel-Spec Horns Tagging. For the convenience to explanation, we illustrate the sub-matrix with a given relation, e.g., *Located in*. So, the matrix  $M$  degenerates into two dimensions, where the rows represent head entities and the columns represent tail entities.

Note that different triples may share overlapping entities, which poses a big challenge for joint extraction task (Zeng et al. 2018).

### Relation Specific Horns Tagging

For a sentence, we design a classifier to assign tags for all possible  $(w_i, r_k, w_j)$  combinations, where  $w_i, w_j \in \mathcal{S}, r_k \in \mathcal{R}$ . We maintain a three-dimensional matrix  $M^{L \times K \times L}$  to store the classification results (Tagging). Therefore, in the test phase, our task is to decode entities and relations from the matrix  $M$  (Decoding).

**Tagging** We employ the “BIE” (Begin, Inside, End) signs to indicate the position information of a token in entities. For example, “HB” means the beginning token of the head entity, “TE” means the end token of the tail entity. As shown in Figure 2 (a), for a sentence that expresses the triple (*New York City*, *Located in*, *New York State*), there are nine special tags (blue tags) in the relation-specific sub-matrix  $M_{r=Located\ in}$ .

According to the insight that an entity can be determined by detecting its boundary tokens (Wei et al. 2020), four types of tags are used in our tagging strategy: (1) **HB-TB**. This tag refers to that both positions are respectively the beginning tokens of a paired head and tail conditioned on a specific relation. For example, there is a relation *Located in* between the two entities “New York City” and “New York State”, therefore, the classification tag of the combination (“New”, *Located in*, “New”) is assigned with the tag “HB-TB”. (2) **HB-TE**. This tag means that the token corresponding to the row is the beginning of a head entity, meanwhile the token corresponding to the column is the end of a tail entity. For instance, “New” is the start token of “New York City” and “State” is the end token of “New York State”, so the combination of (“New”, *Located in*, “State”) is assigned with the tag “HB-TE”. (3) **HE-TE**. This tag shares a similar logic with “HB-TB”, which means two positions

are respectively the end tokens of a paired head entity and tail entity conditioned on a specific relation. For example, the combination of (“City”, *Located in*, “State”) is assigned with “HE-TE”. (4) “-”. All cells other than the above three cases will be marked as “-”. As we can see from Figure 2 (b) and (c), because only the three corners of the rectangle need to be labeled, we vividly name this method as Rel-Spec Horns Tagging.

Obviously, the tagged matrix  $M$  is sparse, which has the following advantages: First, using three instead of nine special tags can effectively narrow down the potential searching space when conducting classification. Second, a sparse  $M$  means that there are sufficient negative samples during the training process. Third, the sparsity of  $M$  ensures the simplicity and efficiency of triple elements decoding.

Furthermore, our Rel-Spec Horns Tagging can naturally address the complex scenarios with overlapping patterns. Specifically, for *EntityPairOverlap* (EPO) case, entity pairs will be labeled in different sub-matrices according to their relations. For example in Figure 2 (a) and (b), (*New York City*, *Located in*, *New York State*) and (*New York State*, *Contains*, *New York City*) are two EPO triples, thus, the two entity pairs are marked in  $M_{r=Located\ in}$  and  $M_{r=Contains}$ , respectively. For *SingleEntityOverlap* (SEO) scenario, if two triples contain the same relation, the two entity pairs will be marked in different parts of  $M_{r=i}$ , otherwise they will be labeled in different sub-matrices according to their relations. For the most complicated *HeadTailOverlap* (HTO<sup>1</sup>) pattern, e.g. the triple (*New York City*, *City Name*, *New York*) in Figure 2 (c), entity pairs are located near the diagonal of  $M_{r=City\ name}$  and can still be easily decoded.

**Decoding** The tagged matrix  $M^{L \times K \times L}$  marks the boundary tokens of paired head entities and tail entities, as well as

<sup>1</sup>HTO is also called SOO when a triple is represented by (subject, relation, object).

the relations between them. Therefore, decoding triples from  $M$  becomes easy and straightforward. That is, for each relation, the spans of head entities are spliced from “HB-TE” to “HE-TE”; the spans of tail entities are spliced from “HB-TB” to “HB-TE”; and two paired entities share the same “HB-TE”.

### Scoring-based Classifier

For an input sentence, we employ a pre-trained BERT (Devlin et al. 2019) as sentence encoder to capture the  $d$ -dimensional token embedding  $e_i$  for each token:

$$\{e_1, e_2, \dots, e_L\} = BERT(\{x_1, x_2, \dots, x_L\}), \quad (1)$$

where  $x_i$  is the input representation of each token. It is the summation over the corresponding token embedding and positional embedding.

Then, we enumerate all possible  $(e_i, r_k, e_j)$  combinations and design a classifier to assign high-confidence tags, where  $r_k$  is the randomly initialized relation representation. Intuitively, we can employ a simple classification network whose input is  $(e_i, r_k, e_j)$  to achieve this goal. However, this intuition has two flaws: On the one hand, a simple classifier is not only unable to fully explore the interactions between entities and relations, but also difficult to model the inherent structural information of triples. On the other hand, using  $(e_i, r_k, e_j)$  as input means the model needs performing at least  $L \times K \times L$  calculations to classify all combinations, which is unacceptable in terms of time.

Inspired by knowledge graph embedding techniques, we borrow the idea from HOLE (Nickel, Rosasco, and Poggio 2016), whose score function is defined as:

$$f_r(h, t) = r^T(\mathbf{h} \star \mathbf{t}), \quad (2)$$

where  $\mathbf{h}$ ,  $\mathbf{t}$  are head and tail representations, respectively.  $\star$  means circular correlation, which is used to mine the latent dependencies between two entities. Here, we redefine the  $\star$  operator as a non-linear concatenation projection:

$$\mathbf{h} \star \mathbf{t} = \phi(\mathbf{W}[\mathbf{h}; \mathbf{t}]^T + \mathbf{b}), \quad (3)$$

where  $\mathbf{W} \in \mathbb{R}^{d_e \times 2d}$ ,  $\mathbf{b}$  are trainable weight and bias,  $d_e$  denotes the dimension of entity pair representations.  $[\cdot; \cdot]$  is the concatenation operation and  $\phi(\cdot)$  is the ReLU activation function. The new definition brings the following benefits: First, the score function of our classifier can be seamlessly connected with the output of sentence encoder. Second, the mapping function from entity features to entity pair representations can be learned adaptively via the matrix  $\mathbf{W}$ . Third, the concatenation between two entities is not commutative, i.e.,  $[\mathbf{h}; \mathbf{t}] \neq [\mathbf{t}; \mathbf{h}]$ , which is indispensable for modeling asymmetric relations.

Next, we use all relation representations  $\mathbf{R} \in \mathbb{R}^{d_e \times 4K}$  to simultaneously compute the salience of  $(w_i, r_k, w_j)_{k=1}^K$  for a token-pair  $(w_i, w_j)$  at once, where 4 is the number of classification tags. Therefore, the final score function of our method is defined as:

$$\mathbf{v}_{(w_i, r_k, w_j)_{k=1}^K} = \mathbf{R}^T \phi(\text{drop}(\mathbf{W}[e_i; e_j]^T + \mathbf{b})), \quad (4)$$

where  $\mathbf{v}$  is the score vector,  $\text{drop}(\cdot)$  denotes the dropout strategy (Srivastava et al. 2014) that is used to prevent overfitting. As a result, we achieve the parallel scoring with only

two layers of fully connected networks (the matrix  $\mathbf{R}$  can also be regarded as a trainable weight), and reduce the processing steps of actual implementation to  $L \times 1 \times L$ , even better than TPLinker (Wang et al. 2020). Moreover, the score function conforms the idea of HOLE and is capable of capturing the correlation and mutual exclusion between relations, which will be verified in the experiments.

Finally, we feed the score vector of  $(w_i, r_k, w_j)$  into a softmax function to predict corresponding tags:

$$P(y_{(w_i, r_k, w_j)} | \mathcal{S}) = \text{Softmax}(\mathbf{v}_{(w_i, r_k, w_j)}) \quad (5)$$

The objective function of OneRel is defined as:

$$\mathcal{L}_{\text{triple}} = - \frac{1}{L \times K \times L} \times \sum_{i=1}^L \sum_{k=1}^K \sum_{j=1}^L \log P(y_{(w_i, r_k, w_j)} = g_{(w_i, r_k, w_j)} | \mathcal{S}), \quad (6)$$

where  $g_{(w_i, r_k, w_j)}$  denotes the gold tag obtained from annotations.

## Experiments

In this section, extensive experiments are conducted to validate the effectiveness of the proposed OneRel and analyze its properties.

### Experimental Settings

**Datasets and Evaluation Metrics** Following previous works (Wei et al. 2020; Wang et al. 2020; Zheng et al. 2021), we evaluate our model and all baselines on two widely used datasets: NYT (Riedel, Yao, and McCallum 2010) and WebNLG (Gardent et al. 2017). The former is generated for distant supervised relation extraction. The latter is originally created for natural language generation (NLG). Both NYT and WebNLG have two versions: one version only annotates the last word of entities, and the other version annotates the whole span of entities. We denote the first version datasets as NYT\* and WebNLG\*, and the second version as NYT and WebNLG. To further study the capability of the proposed OneRel in handling complex scenarios, we split the test set by overlapping patterns and triple number. Detailed statistics of the two datasets are described in Table 1.

Three standard evaluation metrics are used in our experiments, i.e., micro Precision (Prec.), Recall (Rec.) and F1-score (F1). During evaluation, we adopt *Partial Match* for NYT\* and WebNLG\*: an extracted triple (head, relation, tail) is considered to be correct only if the relation and the last word of head and tail are all correct; and use *Exact Match* for NYT and WebNLG: a predicted triple is regarded to be correct only if the whole span of two entities and relation are all exactly matched.

**Implementation Details** In our experiments, all training process is completed on a work station with an AMD 7742 2.25G CPU, 256G memory, a single RTX 3090 GPU, and Ubuntu 20.04. For the pre-trained BERT, we reuse the base

Category	Dataset				Details of Test Set									
	Train	Valid	Test	Relations	Normal	SEO	EPO	HTO	N=1	N=2	N=3	N=4	N>5	Triples
NYT*	56,195	4,999	5,000	24	3,266	1,297	978	45	3,244	1,045	312	291	108	8,110
WebNLG*	5,019	500	703	171	245	457	26	84	266	171	131	90	45	1,591
NYT	56,195	5,000	5,000	24	3,222	1,273	969	117	3,240	1,047	314	290	109	8,120
WebNLG	5,019	500	703	216	239	448	6	85	256	175	138	93	41	1,607

Table 1: Statistics of datasets.  $N$  is the number of triples in a sentence.

cased English model released by Huggingface<sup>2</sup>, which contains 12 Transformer blocks and the hidden size  $d$  is 768. We tune our model on the valid set and use grid search to adjust important hyper-parameters. Specifically, the batch size is set to 8/6 on NYT/WebNLG, and all parameters are optimized by Adam algorithm (Kingma and Ba 2015) with a learning rate of  $1e-5$ . The dimension of the hidden layer  $d_e$  is set to  $3 \times d$ , the dropout probability in equation (4) is 0.1, the max sequence length is set to 100.

**Baselines** We compare our model with ten state-of-the-art baselines: **GraphRel** (Fu, Li, and Ma 2019), **RSAN** (Yuan et al. 2020), **MHSA** (Liu et al. 2020), **CasRel** (Wei et al. 2020), **TPLinker** (Wang et al. 2020), **SPN** (Sui et al. 2020), **CGT** (Ye et al. 2021), **CasDE** (Ma, Ren, and Zhang 2021), **RIFRE** (Zhao et al. 2021a), **PRGC** (Zheng et al. 2021).

Note that the sentence encoders used in GraphRel, RSAN and MHSA are LSTM networks, while other baselines employ a pre-trained BERT to obtain feature representations. For fair comparison, the reported results for all baselines are directly from the original literature. We also conduct ablation test: **OneRel<sup>-</sup>** is the model that replaces the classifier of OneRel with  $f(w_i, r_k, w_j) = \mathbf{W}[e_i; r_k; e_j] + \mathbf{b}$ .

## Results and Analysis

**Main Results** Table 2 shows the comparison results of our model against ten baselines on NYT and WebNLG in terms of *Partial Match* and *Exact Match*. It can be observed that our method, OneRel, outperforms all the ten baselines and achieves the state-of-the-art F1-score on all datasets. Especially on WebNLG\* and WebNLG, OneRel obtains the best performance in terms of all three evaluation metrics, and improves the three indicators on WebNLG to above 90% for the first time. We attribute the outstanding performance of OneRel to its two advantages: First, OneRel solves the joint extraction task from the perspective of fine-grained triple classification. Thus, the information of entities and relations can be combined at the same time during extraction, and the redundant information can also be reduced. Second, the combination of the scoring-based classifier and the Rel-Spec Horns Tagging accomplishes entity and relation extraction in a straightforward way, effectively avoiding the problem of cascading errors.

Compared with the representative *multi-module multi-step* method PRGC, OneRel achieves 1.3 and 2.5 absolute

gain in F1-score on WebNLG\* and WebNLG, respectively. This demonstrates that extracting entities and relations simultaneously in one step can effectively address the cascading errors problem. Besides, for another *multi-module one-step* model TPLinker which employs four independent tags to detect entities and relations, our OneRel outperforms it by 0.9, 2.4, 0.9 and 4.3 absolute gains on the four datasets, respectively. Such results confirm that using one module to extract all triple elements in one step is effective for exploring the interactions between entities and relations. These all indicate that *One-module One-step* is expected to become a new paradigm for completing joint extraction task.

We can also observe that among BERT-based models, OneRel<sup>-</sup> achieves a competitive performance with CasRel, TPLinker and CasDE; OneRel obtains the best performance on all datasets. In addition to BERT, OneRel<sup>-</sup> and OneRel only use one and two layers of fully connected network, respectively. Their architectures are much simpler than most baselines and easier to train. Besides, the performance of OneRel is much better than that of OneRel<sup>-</sup> on all datasets, which reveals that it is crucial to capture the dependencies between entities and relations when designing a classifier. This conclusion points the way for us to design stronger models in the future.

**Detailed Results on Complex Scenarios** To verify the ability of our method in handling overlapping patterns and multiple triples, we conduct two extended experiments on different subsets of NYT\* and WebNLG\*. We select four powerful models as baselines and the detailed results are shown in Table 3.

It can be observed that our OneRel achieves the best F1-score on 13 of the 18 subsets, especially for the most complex cases *HTO* and  $N \geq 5$ . The *HTO* pattern contains two situations: one is nested entities that most previous approaches cannot accurately identify, e.g., the triple (*Bruce Lee*, *Family name*, *Lee*). The other is that the head entity and tail entity share the same words, for example, the sentence “*Native Americans in the United States are one of the ethnic groups of the country.*” in WebNLG\* expresses the triple (*States*, *ethnicGroup*, *States*). Besides, a  $N \geq 5$  sentence may contain SEO, EPO and *HTO* patterns at the same time, which brings a big challenge to existing approaches. Nevertheless, our OneRel achieves the best performance on both *HTO* and  $N \geq 5$  of NYT\* and WebNLG\*, which adequately proves that our Rel-Spec Horns Tagging is able to address the overlapping triples problem from de-

<sup>2</sup><https://huggingface.co/bert-base-cased>

Model	Partial Match						Exact Match					
	NYT*			WebNLG*			NYT			WebNLG		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
GraphRel (Fu, Li, and Ma 2019)	63.9	60.0	61.9	44.7	41.1	42.9	-	-	-	-	-	-
RSAN (Yuan et al. 2020)	-	-	-	-	-	-	85.7	83.6	84.6	80.5	83.8	82.1
MHSA (Liu et al. 2020)	88.1	78.5	83.0	89.5	86.0	87.7	-	-	-	-	-	-
CasRel (Wei et al. 2020)	89.7	89.5	89.6	93.4	90.1	91.8	-	-	-	-	-	-
TPLinker (Wang et al. 2020)	91.3	92.5	91.9	91.8	92.0	91.9	91.4	92.6	92.0	88.9	84.5	86.7
SPN (Sui et al. 2020)	93.3	91.7	92.5	93.1	93.6	93.4	92.5	92.2	92.3	-	-	-
CGT (Ye et al. 2021)	<b>94.7</b>	84.2	89.1	92.9	75.6	83.4	-	-	-	-	-	-
CasDE (Ma, Ren, and Zhang 2021)	90.2	90.9	90.5	90.3	91.5	90.9	89.9	91.4	90.6	88.0	88.9	88.4
RIFRE (Zhao et al. 2021a)	93.6	90.5	92.0	93.3	92.0	92.6	-	-	-	-	-	-
PRGC (Zheng et al. 2021)	93.3	91.9	92.6	94.0	92.1	93.0	<b>93.5</b>	91.9	92.7	89.9	87.2	88.5
OneRel <sup>-</sup>	91.3	90.5	90.9	93.8	91.4	92.6	91.1	90.4	90.8	90.5	88.2	89.4
OneRel	92.8	<b>92.9</b>	<b>92.8</b>	<b>94.1</b>	<b>94.4</b>	<b>94.3</b>	93.2	<b>92.6</b>	<b>92.9</b>	<b>91.8</b>	<b>90.3</b>	<b>91.0</b>

Table 2: Precision(%), Recall (%) and F1-score (%) of our proposed OneRel and baselines.

Model	NYT*									WebNLG*								
	Normal	EPO	SEO	HTO	N=1	N=2	N=3	N=4	N≥5	Normal	EPO	SEO	HTO	N=1	N=2	N=3	N=4	N≥5
CasRel	87.3	92.0	91.4	77.0 <sup>§</sup>	88.2	90.3	91.9	94.2	83.7	89.4	94.7	92.2	90.4 <sup>§</sup>	89.3	90.8	94.2	92.4	90.9
TPLinker	90.1	94.0	93.4	90.1 <sup>§</sup>	90.0	92.8	93.1	96.1	90.0	87.9	95.3	92.5	86.0 <sup>§</sup>	88.0	90.1	94.6	93.3	91.6
SPN	90.8	94.1	94.0	-	90.9	93.4	<b>94.2</b>	95.5	90.6	-	-	-	-	89.5	91.3	<b>96.4</b>	94.7	93.8
PRGC	<b>91.0</b>	94.5	94.0	81.8	<b>91.1</b>	93.0	93.5	95.5	93.0	90.4	<b>95.9</b>	93.6	94.6	89.9	91.6	95.0	94.8	92.8
OneRel	90.6	<b>95.1</b>	<b>94.8</b>	<b>90.8</b>	90.5	<b>93.4</b>	93.9	<b>96.5</b>	<b>94.2</b>	<b>91.9</b>	95.4	<b>94.7</b>	<b>94.9</b>	<b>91.4</b>	<b>93.0</b>	95.9	<b>95.7</b>	<b>94.5</b>

Table 3: F1-score (%) on sentences with different overlapping patterns and different triple numbers. § marks the results reported by (Zheng et al. 2021).

sign, and more robust than baselines when dealing with the complicated scenarios.

**Results on Different Sub-tasks** We further explore the performance of OneRel on different sub-tasks, i.e., entity pair recognition and relation classification. From Table 4, it can be found that our OneRel outperforms all the baselines on most test instances of NYT\* and all the indicators of WebNLG\*. These encouraging results once again verifies our motivation.

Interestingly, the four models show the same trend on the two datasets: for NYT\*, there is an obvious gap between the F1-score on  $(h, t)$  and  $r$ ; for WebNLG\*, the performance on  $(h, t)$  and  $r$  are much higher than that of  $(h, r, t)$ . Based on this phenomenon, previous researchers have analyzed that entity pair recognition and triple formation are two bottlenecks of the joint extraction task (Sui et al. 2020). We believe that in addition to the above reasons, the characteristic of the datasets is also an important factor. Concretely, NYT\* contains lots of EPO triples, which means that the impact of wrongly recognized entity pairs is much greater than the influence of mistakenly classified relations. Suppose a

sentence that expresses three triples (*Obama*, President of, *United States*), (*Obama*, Live in, *United States*), (*Obama*, Place of birth, *United States*). If one relation is wrongly classified, the recall on  $r$  may be 0.67. While if one entity pair is incorrectly recognized, the recall on  $(h, t)$  may drop to 0. In contrast, the proportion of EPO patterns in WebNLG\* is much smaller than that of NYT\* (3.6% vs 19.6%), thus, the performance of models on entity pair recognition is consistent with that on relation classification. Besides, WebNLG\* contains more relations than NYT\* (171 vs 24) and some of them are confusing, e.g., *Leader* and *LeaderName*. So, the triple formation on WebNLG\* is more difficult than NYT\*.

**Model Efficiency** We evaluate the model efficiency with respect to *Training Time* and *Inference Time* of the most similar baseline TPLinker in two datasets NYT\* and WebNLG\*, and the results are shown in Table 5. In this experiment, the batch size of the two models during training and testing are set to 6 and 1, respectively, and the max length of input sentence is 100. Although the theoretical complexity of the two models is  $\mathcal{O}(KL^2)$ , OneRel is better than TPLinker in

Model	Element	NYT*			WebNLG*		
		Prec.	Rec.	F1	Prec.	Rec.	F1
CasRel	$(h, t)$	89.2	90.1	89.7	95.3	91.7	93.5
	$r$	96.0	93.8	94.9	96.6	91.5	94.0
	$(h, r, t)$	89.7	89.5	89.6	93.4	90.1	91.8
SPN	$(h, t)$	93.2	92.7	92.9	95.0	95.4	95.2
	$r$	96.3	95.7	96.0	95.2	95.7	95.4
	$(h, r, t)$	93.3	91.7	92.5	93.1	93.6	93.4
PRGC	$(h, t)$	<b>94.0</b>	92.3	93.1	96.0	93.4	94.7
	$r$	95.3	96.3	95.8	92.8	96.2	94.5
	$(h, r, t)$	<b>93.3</b>	91.9	92.6	94.0	92.1	93.0
OneRel	$(h, t)$	93.3	<b>93.4</b>	<b>93.3</b>	<b>96.2</b>	<b>96.5</b>	<b>96.3</b>
	$r$	<b>96.7</b>	<b>96.9</b>	<b>96.8</b>	<b>96.7</b>	<b>97.0</b>	<b>96.8</b>
	$(h, r, t)$	92.8	<b>92.9</b>	<b>92.8</b>	<b>94.1</b>	<b>94.4</b>	<b>94.3</b>

Table 4: Results on triple elements.  $(h, t)$  denotes the entity pair and  $r$  means the relation.

Dataset	Model	Training Time	Inference Time	F1
NYT*	TPLinker	1592	46.2	91.9
	OneRel	<b>1195</b>	<b>41.5</b>	<b>92.9</b>
WebNLG*	TPLinker	599	<b>40.1</b>	91.9
	OneRel	<b>88</b>	45	<b>94.3</b>

Table 5: Comparison of the model efficiency. Training Time (s) means the time required to train one epoch, Inference Time (ms) is the time to predict triples of one sentence.

terms of parallel processing, i.e., OneRel processes  $K$  relations at a time, while TPLinker handles one relation at each step. Therefore, when the size of relations increases from 24 (NYT\*) to 171 (WebNLG\*), the training time of OneRel increases from  $1.3\times$  faster than TPLinker to an astonishing  $6.8\times$ . As opposed to what we observed in *Training Time*, the F1-score of OneRel is much better than that of TPLinker. This confirms the efficiency and the learning ability of our proposed classifier. Besides, the inference time of OneRel is similar to that of TPLinker on NYT\* and WebNLG\*, which illustrates the effectiveness and rationality of our proposed Rel-Spec Horns Tagging.

**Topology Structure of Relations** Our scoring-based classifier borrows the idea of HOLE (Nickel, Rosasco, and Poggio 2016), and theoretically, it should also be able to learn the correlation and mutual exclusion between relations. To verify the learning ability of our classifier, we visualize the relation representations of NYT by the  $t$ -SNE (Maaten and Hinton 2008), which is a nonlinear dimensionality reduction algorithm. We omit 6 long-tail relations that appear less than 50 times in whole training set, and the results are shown in Figure 3. It can be observed that the topology structure of relations reflects their inherent connections. For instance, the relations related to `people` are on the left region of the figure, while the relations related to `location` are on

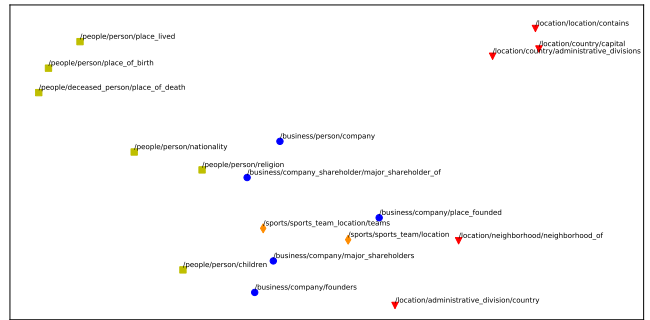


Figure 3: (Best viewed in color and zoom in.) Visualization of relations on NYT dataset.

the right. Especially, the spatial positions of `place_lived`, `place_of_birth` and `place_of_death` are very close, which is in consistent with our common sense. This property is critical for predicting EPO triples. That is, if the model predicts the relation `place_of_birth` for an entity pair, we can infer that the entity pair may hold another relation `place_lived` with a high-probability. These all suggest that our method has learned not only the features of a specific dataset, but also the general knowledge that conforms to the real world. Thus, our OneRel may have generalization capabilities.

## Conclusion

In this paper, we provide a novel perspective to transform the joint extraction task into a fine-grained triple classification problem, and propose a novel joint model with a scoring-based classifier and Rel-Spec Horns Tagging strategy to obtain triples with one module in one step, which greatly alleviates the problems of cascading errors and redundant information. Experiments on public datasets show that our model performs better than the state-of-the-art approaches on different scenarios.

In the future, we would like to explore the following directions:

- To improve the efficiency of the model, we design a simplified version of HOLE as the score function. We will next try to design a more efficient and powerful score function to further strengthen its ability of capturing the connections between entities and relations.
- We would like to explore the idea of triple classification in other information extraction problems, such as event extraction.

## Acknowledgments

The work is supported by National Key R&D Plan (No. 2018YFB1005100), National Natural Science Foundation of China (No. 61751201, 61772076, and 61732005), Natural Science Fund of Beijing (No. Z181100008918002), Central Leading Local Project(No. 2020L3024), BDAS National Engineering Laboratory Open Project (No. CAS-NDST202006) and Fujian Provincial DSTLST Project (No. 2019H0026).

## References

- Chan, Y. S.; and Roth, D. 2011. Exploiting Syntactico-Semantic Structures for Relation Extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 551–560. Portland, Oregon, USA: Association for Computational Linguistics.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.
- Fu, T.-J.; Li, P.-H.; and Ma, W.-Y. 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*, 1409–1418. Florence, Italy: Association for Computational Linguistics.
- Gardent, C.; Shimorina, A.; Narayan, S.; and Perez-Beltrachini, L. 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)*, 179–188. Association for Computational Linguistics.
- Katiyar, A.; and Cardie, C. 2017. Going out on a limb: Joint Extraction of Entity Mentions and Relations without Dependency Trees. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 917–928. Vancouver, Canada: Association for Computational Linguistics.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Li, Q.; and Ji, H. 2014. Incremental Joint Extraction of Entity Mentions and Relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 402–412. Baltimore, Maryland: Association for Computational Linguistics.
- Liu, J.; Chen, S.; Wang, B.; Zhang, J.; Li, N.; and Xu, T. 2020. Attention as Relation: Learning Supervised Multi-head Self-Attention for Relation Extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 3787–3793.
- Ma, L.; Ren, H.; and Zhang, X. 2021. Effective Cascade Dual-Decoder Model for Joint Entity and Relation Extraction. *CoRR*, abs/2106.14163.
- Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.
- Miwa, M.; and Bansal, M. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1105–1116. Association for Computational Linguistics.
- Nickel, M.; Rosasco, L.; and Poggio, T. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Riedel, S.; Yao, L.; and McCallum, A. 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*, volume 6323, 148–163. Springer.
- Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958.
- Sui, D.; Chen, Y.; Liu, K.; Zhao, J.; Zeng, X.; and Liu, S. 2020. Joint Entity and Relation Extraction with Set Prediction Networks. *arXiv preprint arXiv:2011.01675*.
- Tan, Z.; Zhao, X.; Wang, W.; and Xiao, W. 2019. Jointly Extracting Multiple Triplets with Multilayer Translation Constraints. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, 7080–7087. AAAI Press.
- Wang, Y.; Sun, C.; Wu, Y.; Zhou, H.; Li, L.; and Yan, J. 2021. UniRE: A Unified Label Space for Entity Relation Extraction. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 220–231.
- Wang, Y.; Yu, B.; Zhang, Y.; Liu, T.; Zhu, H.; and Sun, L. 2020. TPLinker: Single-stage Joint Extraction of Entities and Relations Through Token Pair Linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1572–1582. International Committee on Computational Linguistics.
- Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; and Chang, Y. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1476–1488. Association for Computational Linguistics.
- Ye, H.; Zhang, N.; Deng, S.; Chen, M.; Tan, C.; Huang, F.; and Chen, H. 2021. Contrastive Triple Extraction with Generative Transformer. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021*, 14257–14265. AAAI Press.
- Yu, B.; Zhang, Z.; Shu, X.; Liu, T.; Wang, Y.; Wang, B.; and Li, S. 2020. Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, 2282–2289. IOS Press.
- Yuan, Y.; Zhou, X.; Pan, S.; Zhu, Q.; Song, Z.; and Guo, L. 2020. A Relation-Specific Attention Network for Joint Entity and Relation Extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 4054–4060.



- Zelenko, D.; Aone, C.; and Richardella, A. 2003. Kernel Methods for Relation Extraction. *J. Mach. Learn. Res.*, 3: 1083–1106.
- Zeng, D.; Zhang, H.; and Liu, Q. 2020. CopyMTL: Copy Mechanism for Joint Extraction of Entities and Relations with Multi-Task Learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, 9507–9514. AAAI Press.
- Zeng, X.; Zeng, D.; He, S.; Liu, K.; and Zhao, J. 2018. Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 506–514. Association for Computational Linguistics.
- Zhang, M.; Zhang, Y.; and Fu, G. 2017. End-to-End Neural Relation Extraction with Global Optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 1730–1740. Association for Computational Linguistics.
- Zhao, K.; Xu, H.; Cheng, Y.; Li, X.; and Gao, K. 2021a. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction. *Knowledge-Based Systems*, 106888.
- Zhao, T.; Yan, Z.; Cao, Y.; and Li, Z. 2021b. A Unified Multi-Task Learning Framework for Joint Extraction of Entities and Relations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16): 14524–14531.
- Zheng, H.; Wen, R.; Chen, X.; Yang, Y.; Zhang, Y.; Zhang, Z.; Zhang, N.; Qin, B.; Ming, X.; and Zheng, Y. 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*, 6225–6235. Association for Computational Linguistics.
- Zheng, S.; Wang, F.; Bao, H.; Hao, Y.; Zhou, P.; and Xu, B. 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)*, 1227–1236. Vancouver, Canada: Association for Computational Linguistics.
- Zhou, G.; Su, J.; Zhang, J.; and Zhang, M. 2005. Exploring Various Knowledge in Relation Extraction. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, 427–434.