EmRel: Joint Representation of Entities and *Em*bedded *Rel*ations for Multi-triple Extraction

Anonymous ACL submission

Abstract

Multi-triple extraction is a challenging task due to the existence of informative inter-triple correlations and consequently rich interactions across the constituent entities and relations. While existing works only explore cross-entity interactions, we propose to explicitly introduce *relation* representation, jointly represent it with entities, and novelly align them to identify valid triples. We perform comprehensive experiments[1] on document-level relation extraction and joint entity and relation extraction along with detailed ablations to demonstrate the advantage of the proposed method.

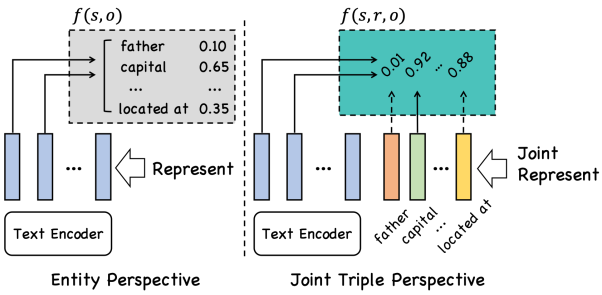

Figure 1: Different formulations for multi-triple extraction. 1) *entity perspective* constructs only entity representation and feed them into a relation-specific classifier. 2) *joint triple perspective* constructs both entity representation and relation representation to model comprehensive correlations across all components.

1 Introduction

Relation extraction aims at discovering structured knowledge in the form of <*subject-relation-object*> triples from plain text. It is an essential task towards constructing knowledge bases, which further supporting various applications such as search engines and question answering systems. Although a lot of efforts have been made in building advanced relation extraction systems, it is still a challenging problem under certain practical scenarios where multiple entities and relations are involved, *e.g.*, document-level relation extraction (Yao et al., 2019) and joint entity and relation extraction (Riedel et al., 2010; Gardent et al., 2017).

Existing works mostly take the *entity perspective* that focuses on exploring cross-entity interactions (Xu et al., 2021; Zeng et al., 2020). They either treat relations as atomic labels specified in a final classifier (Xu et al., 2021; Zeng et al., 2020; Wang et al., 2020), or simply search for each individual relation its possible subjects and objects(Wei et al., 2020; Zheng et al., 2021). However, as an essential component, relations also interact with entities and context, which jointly exhibit informative inter-triple correlations. e.g., the two relations *capital of* and *located at* often co-occur between the same pair of entities but with different probabilities conditioned on specific contextual clues. As a consequence, the capability to model and make full use of rich interactions across relations, entities, and context is crucial for the task.

In this paper, we advocate a novel *joint triple perspective* for relation extraction (see Figure 1 for illustration). Different from previous works that only seek to represent entities, we propose ***EmRel*** that creates, refines and leverages the ***Em***bedded representations of ***Rel***ations. Specifically, we first explicitly create relation representations as embedded vectors; then refine these relation (as well as entity) representations by modeling rich relation-entity-context interactions via an attention-based fusion module; and finally identify valid triples by aligning the representation of entities and relations in a joint space, where a novel alignment function based on Tucker Decomposition is designed to deliver such a purpose. This joint triple perspective actually considers entities along with relations as components of a small, context-dependent knowl-

[1]The code will be available at https://github.com/XXX/XXX

edge graph, and completes this graph by aligning and reasoning to extract multiple valid triples.

To demonstrate the advantage of the proposed *EmRel* framework, we conduct experiments on two specific scenarios of multi-triple extraction: document-level relation extraction(RE) and joint entity and relation extraction, with three popular datasets including DocRED (Yao et al., 2019), NYT (Riedel et al., 2010) and WebNLG (Gardent et al., 2017). The results verify the superiority of the joint triple perspective over the traditional entity perspective in multi-triple extraction. We also provide further ablation study to show the effectiveness of our fusion module and alignment function.

2 Related Works

Document-level Relation Extraction Extracting multi-triples from document-level text has recently aroused increasing interests (Yao et al., 2019). Existing methods take the entity perspective that proposes various techniques to model entity interactions. Nan et al. (2020) and Zeng et al. (2020) construct an entity graph, and perform graph-level reasoning to refine the entity node representations. Xu et al. (2021) introduces entity structure as useful prior, and models such information within the transformer attention layer. Zhang et al. (2021) utilizes a segmentation network to model the interdependency among entity pairs. Therefore, inter-triple correlations are only captured at the entity level while relation-based ones are neglected.

Joint Entity and Relation Extraction Joint entity and relation extraction is a popular task that extracts multi-triples along with their entities. Existing works can be concluded into two frameworks: one that searches for each individual relation its possible subjects and objects (Liu et al., 2020; Wang et al., 2020;Wei et al., 2020), and the other that directly see each word as a candidate entity and assign them with relation labels (Gupta et al., 2016; Zheng et al., 2021). Both formulations do not explicitly include inter-triple correlations. Very recently, Wu and Shi (2021) proposes to model the interdependencies between entity labels and relation labels, However, such correlation is constrained within a specific word position, while *EmRel* exploits the global correlations among all triples and across entities, context, and relations. Li et al. (2021) introduces a translation-based function that predicts object from subject and relation, while *EmRel* proposes a more expressive alignment func-

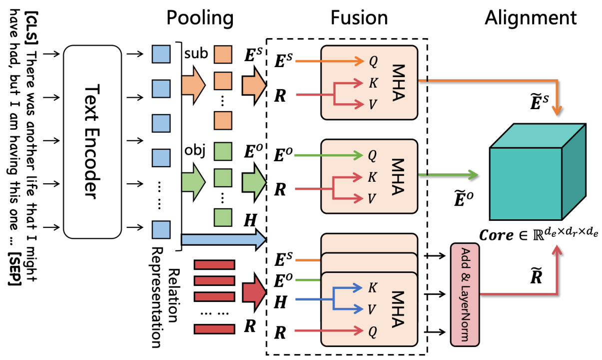

Figure 2: The overall framework of EmRel. It explicitly introduces relations embedding, and jointly represents it with entities to identify all valid triples.

tion that models the ternary interaction of subject, relation and object.

3 Methodology

3.1 Task Formulation

We first formulate the multi-triple extraction task to suitably contains both document-level RE and joint entity and relation extraction. Given a sequence of text $\{w_i\}$, a set of candidate entities $E = \{e_i\}$ and the pre-defined relations $R = \{r_i\}$, the candidate triples can be derived as:

$$T = \{< s, r, o > | s, o \in \{e_i\}, r \in \{r_i\}\} \quad (1)$$

the target is to assign each t in T a binary label that discriminates its validity. The candidate entities can either be pre-annotated, as in document-level relation extraction, or be jointly recognized, as in joint entity and relation extraction. In the latter scenario, one prevailing solution is to directly see each word as a candidate entity, such as tagging-based methods (Wang et al., 2020) or table filling methods (Gupta et al., 2016). Here we follow Wang et al. (2020) as our baseline, and thus formulate both tasks under a unified framework that extracts multi-triples from a given candidate entity set.

3.2 EmRel

EmRel consists of three modules: *Representation Construction* for both entities and relations, *Representation Fusion* that captures multi-triple correlations by modeling the informative interactions across entities, context and relations, and *Representation Alignment* that leverages these representations to extract triples by aligning their ternary structures (see Figure 2 for illustration).

Method	NYT*			WebNLG*			NYT			WebNLG		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
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
Baseline†	91.1	**92.5**	91.8	91.4	92.7	92.1	91.2	92.1	91.6	88.7	86.5	87.6
EmRel	**91.7**	**92.5**	**92.1**	92.7	**93.0**	**92.9**	**92.6**	**92.7**	**92.6**	**90.2**	**87.4**	**88.7**

Table 1: Results on NYT and WebNLG. * denotes task settings that only annotate the last word. † denotes our reproduced results of Wang et al. (2020) as the baseline. Best results in **bold**.

Representation Construction The entity representation is constructed similar to existing practices. We employ a text encoder, *e.g.*, pretrained language models like BERT (Devlin et al., 2019), to obtain the contextualized representation:

$$(h_1, h_2, ...h_n) = \texttt{encoder}(w_1, w_2, ...w_n) \tag{2}$$

which we denote as $\mathbf{H}$. Then we construct each entity representation $\mathbf{e}_i \in \mathbb{R}^{d_e}$ by applying a pooling operation on its corresponding mention positions, and further map it into respective subject and object representation $\mathbf{e}_i^s$, $\mathbf{e}_i^o$.

We embed the target relations R into an embedding matrix $\mathbf{R} \in \mathbb{R}^{|R|\times d_r}$, where each row $\mathbf{R}_{i,:}$ represents a vectorized relation r_i. This matrix is maintained as part of the model parameter and trained accordingly.

Representation Fusion In order to jointly represent entities and relations in a shared knowledge representation space, we fuse them to be aware of each other. We adopt the attention network (Bahdanau et al., 2015) to model inter-component interactions, which has proven to be very successful in modeling rich interactions across contexts (Yu et al., 2018) or modalities (Lu et al., 2016). Specifically, we employ the canonical multi-head attention (MHA) network (Vaswani et al., 2017). Given the target representation $\mathbf{X}_Q$ and the source representation $\mathbf{X}_S$, each head of MHA operates them as:

$$\begin{aligned}\widehat{\mathbf{X}}_Q =& \texttt{Att}(\mathbf{X}_Q W^Q, \mathbf{X}_S W^K, \mathbf{X}_S W^V) \\ =& \texttt{softmax}(\frac{(\mathbf{X}_Q W^Q)(\mathbf{X}_S W^K)^T}{\sqrt{d_k}})\mathbf{X}_S W^V\end{aligned} \tag{3}$$

where $\widehat{\mathbf{X}}_Q$ is the updated representation of $\mathbf{X}_Q$ w.r.t. $\mathbf{X}_S$, all heads operate in parallel and will be concatenated together.

In *EmRel*, to exploit the comprehensive interactions across all components, we first construct entity/context-aware relation representation:

$$\begin{aligned}\widehat{\mathbf{R}}^s =& \texttt{Att}_{s2r}(\mathbf{R}W^Q, \mathbf{E}^s W^K, \mathbf{E}^s W^V) \\ \widehat{\mathbf{R}}^o =& \texttt{Att}_{o2r}(\mathbf{R}W^Q, \mathbf{E}^o W^K, \mathbf{E}^o W^V) \\ \widehat{\mathbf{R}}^c =& \texttt{Att}_{c2r}(\mathbf{R}W^Q, \mathbf{H}W^K, \mathbf{H}^s W^V)\end{aligned} \tag{4}$$

which are then aggregated together using layer normalization:

$$\widehat{\mathbf{R}} = \texttt{LayerNorm}(\widehat{\mathbf{R}}^s + \widehat{\mathbf{R}}^o + \widehat{\mathbf{R}}^c) \tag{5}$$

we symmetrically construct relation-aware entity representation:

$$\begin{aligned}\widehat{\mathbf{E}}^s =& \texttt{Att}_{r2s}(\mathbf{E}^s W^Q, \mathbf{R}W^K, \mathbf{R}W^V) \\ \widehat{\mathbf{E}}^o =& \texttt{Att}_{r2o}(\mathbf{E}^o W^Q, \mathbf{R}W^K, \mathbf{R}W^V)\end{aligned} \tag{6}$$

s, o, c are abbreviations for $subject$, $object$ and $context$. Each attention module is wrapped with residual connection, feedforward layer, layer normalization, and is instantiated with different parameters of W_Q, W_K, W_V to model distinguished attending patterns. The outputs of fusion module are refined representations $\widehat{\mathbf{R}}$, $\widehat{\mathbf{E}}^s$, $\widehat{\mathbf{E}}^o$ for relations, subjects and objects.

Method	Dev		Test	
	IgnF1	**F1**	**IgnF1**	**F1**
BERT-TS	-	54.42	-	53.92
CorefBERT	55.32	57.51	54.54	56.96
LSR	52.43	59.00	56.97	59.05
SSAN	57.03	59.19	55.84	58.16
Baseline†	$56.45_{\pm 0.47}$	$58.56_{\pm 0.44}$	55.84	58.15
EmRel	**57.23**$_{\pm 0.15}$	**59.30**$_{\pm 0.10}$	**57.27**	**59.66**

Table 2: Results on DocRED. † denotes our reproduced results of the baseline implementation in Xu et al. (2021). All results are produced with multiple runs using different random seeds. Best results in **bold**.

Representation Alignment *EmRel* extracts triples by aligning their ternary components $\widehat{\mathbf{R}}$, $\widehat{\mathbf{E}}^s$, and $\widehat{\mathbf{E}}^o$. In order to fully leverage their expressiveness, we propose factorization-based alignment

Method	Dev		Test	
	IgnF1	F1	IgnF1	F1
EmRel	$57.23_{\pm 0.15}$	$59.30_{\pm 0.10}$	57.27	59.66
−Fusion	$57.02_{\pm 0.20}$	$59.12_{\pm 0.19}$	56.66	58.92
−Alignment	$56.45_{\pm 0.47}$	$58.56_{\pm 0.44}$	55.84	58.15

Table 3: Ablation results on *EmRel* modules.

using Tucker decomposition (Tucker et al., 1964). We introduce a core tensor $\mathcal{Z} \in \mathbb{R}^{d_e * d_r * d_e}$, and the validity for each $< s_i, r_k, o_j >$ is scored as:

$$\phi(s_i, r_k, o_j) = \sigma(\mathcal{Z} \times_1 \hat{\mathbf{e}}_i^s \times_2 \hat{\mathbf{r}}_k \times_3 \hat{\mathbf{e}}_j^o + b_k) \quad (7)$$

where $\hat{\mathbf{e}}_i^s = \widehat{\mathbf{E}}_{i,:,}^s$, $\hat{\mathbf{r}}_k = \widehat{\mathbf{R}}_{k,:,}$, $\hat{\mathbf{e}}_i^o = \widehat{\mathbf{E}}_{j,:,}^o$, and $\times_n$ indicates tensor product along the n-th mode, σ denotes sigmoid function. We compute ϕ for all triples in parallel using batched tensor product, and train them using cross-entropy loss:

$$\begin{aligned} L = \sum_{<s_i, r_k, o_j>}^{T} & [-\mathbb{1}^{True}(< s_i, r_k, o_j >) \log \phi(s_i, r_k, o_j) \\ & - \mathbb{1}^{False}(< s_i, r_k, o_j >) \log(1 - \phi(s_i, r_k, o_j))] \end{aligned} \quad (8)$$

where $\mathbb{1}$ indicates the ground truth validity.

4 Experiments

4.1 Main Results

We conduct comprehensive experiments on document-level RE dataset DocRED (Yao et al., 2019) and joint entity and relation extraction dataset NYT (Riedel et al., 2010) and WebNLG (Gardent et al., 2017). The specifics about these datasets and our implementation details can be referred to Appendix. We provide our reproduced results of TPLinker (Wang et al., 2020) and the baseline system of Xu et al. (2021). Both are competitive baselines based on the entity perspective, and are directly comparable with *EmRel*.

The results (see Table 1 and Table 2) show that *EmRel* universally outperforms its baselines on all datasets. Respectively, **+0.3 F1** for NYT*, **+0.8 F1** for WebNLG*, **+1.0 F1** for NYT and **+1.1 F1** for WebNLG. On DocRED, *EmRel* improves the baseline by **+0.95 Dev F1**, **+1.47 Test F1**, and also outperforms several previous studies including BERT-TS (Wang et al., 2019), CorefBERT (Ye et al., 2020), LSR (Nan et al., 2020), and SSAN (Xu et al., 2021).

4.2 Ablation Studies

This section gives ablation studies on DocRED.

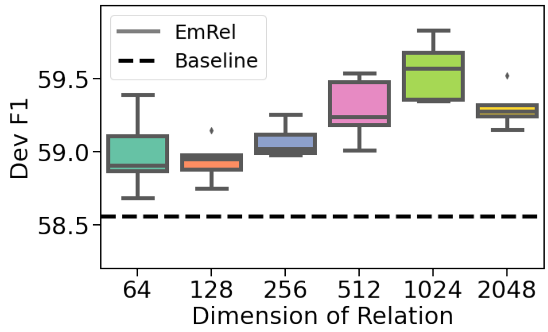

Figure 3: Ablation on dimensions of relation representation.

On EmRel Modules We first varify the design of *EmRel* modules. Table 3 shows that both fusion and alignment module contribute to the improvements. We also observe that *EmRel* has more robust performance across multiple runs. This can be attributed to our alignment function, which, once removed, would result in an increased standard deviation from ±0.20 to ±0.47.

On the Dimensionality of Relation Representations We investigate the effects of choices for d_r in Fig 3. First of all, the advantage of *EmRel* is general across variant choices comparing to the baseline. As we gradually set a higher d_r from 64 to 1024, we get improved performance for its stronger expressive capability. While we further increase d_r to 2048, the performance starting to degrades, which might attribute to overfitting. Overall, the optimal dimension lies within [512, 2048], which is quite robust and also computationally acceptable.

5 Conclusion

In this paper, we propose *EmRel* for multi-triple extraction. Distinguished from existing works, *EmRel* explicitly creates, refines, and leverages the embedded representation of relations. Notably, we design a novel alignment function that discriminates triple validity by aligning its components in a joint representation space. We conduct experiments on both document-level relation extraction and joint entity and relation extraction, to demonstrate the advantage of *EmRel* over its baselines.

EmRel also provides a new joint triple perspective, where multi-triple extraction is formulated as completion of a small, context-dependent knowledge graph, with candidate entities and relations as its components. In the future, we think more intricate techniques *e.g.*, graph-based reasoning, can be explored following such formulation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015 : International Conference on Learning Representations 2015*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planning. In *55th annual meeting of the Association for Computational Linguistics (ACL)*.

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2537–2547, Osaka, Japan. The COLING 2016 Organizing Committee.

Xianming Li, Xiaotian Luo, Chenghao Dong, Daichuan Yang, Beidi Luan, and Zhen He. 2021. TDEER: An efficient translating decoding schema for joint extraction of entities and relations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8055–8064, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jie Liu, Shaowei Chen, Bingquan Wang, Jiaxin Zhang, Na Li, and Tong Xu. 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-20*, pages 3787–3793. International Joint Conferences on Artificial Intelligence Organization. Main track.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29:289–297.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Ledyard R Tucker et al. 1964. The extension of factor analysis to three-dimensional matrices. *Contributions to mathematical psychology*, 110119.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *arXiv preprint arXiv:1909.11898*.

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.

Hui Wu and Xiaodong Shi. 2021. Synchronous dual network with cross-type attention for joint entity and relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2769–2779, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14149–14157.

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

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. Association for Computational Linguistics.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. 2019. Learning deep bilinear transformation for fine-grained image representation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

A Benchmarks

We introduce the benchmarks used in this work. Table 4 gives their detailed statistics. **DocRED** is constructed from Wikipedia document. It provides comprehensive human annotations for entity mentions, entity types, relational triples, along with their supporting evidences. Each document is a semantically integrate unit that centers in one concept (the title of the wiki page), resulting multiple triples with rich correlations. **NYT** is constructed from New York Times news articles and annotated through distant supervision. **WebNLG** is originally created for natural language generation task, and the sentences are written by humans to cover given triples. Both datasets have the other version denoted as NYT* and WebNLG*. The texts in NYT and WebNLG are much shorter than DocRED documents. These two datasets also feature in multiple triples. Specifically, previous studies have concluded multi-triples into three specific cases: *EPO*: triples overlap with both subjects and objects, *SEO*: triples overlap with subject or object, and *Normal*: without any overlapping. In this paper, we solve all three datasets under a unified multi-triple extraction formulation with *EmRel*.

B Implementation Details

To provide comparable results, we set hyper-parameters following previous works (Wang et al., 2020; Xu et al., 2021). On NYT / WebNLG, we set learning rate as 5e-5, batch size as 24 / 6, and epoch as 100. On DocRED, we set learning rate as 3e-5, batch size as 4, and search epochs in {40, 60, 80, 100}. To produce more robust results, we further perform multiple searches using different seeds, resulting a grid search on both epochs and random seeds. The mean and standard deviation results across different seed are reported on development set. We also provide our reproduced baseline results, *i.e.*, TPLinker Wang et al. (2020) and the baseline system of Xu et al. (2021). The former further adopts a hand-shaking strategy to decode entity spans. The dimension of embedded relation representation is set as 768 for DocRED, 128 for NYT / WebNLG[2], and the number of attention heads in the fusion module is set as 4. BERT-Base-Cased (Devlin et al., 2019) is used as the context encoder. All experiments are conducted on a single NIVDIA V100 or A100 GPU machine.

C Grouped Alignment

The WebNLG dataset has up to 216 relations, which requires increased computational cost. Inspired by (Zheng et al., 2019), we split the alignment tensors into N groups across its dimensions to reduce the computational overhead, and re-write Eq. 7 as:

$$\phi(s_i, p_k, o_j) = \sum_{n=1}^{N} \mathcal{Z}^n \times_1 \hat{\mathbf{e}}_i^{s,n} \times_2 \hat{\mathbf{r}}_k^n \times_3 \hat{\mathbf{e}}_j^{o,n} + b_k \quad (9)$$

$$\begin{aligned} \hat{\mathbf{e}}_i^{s,n} &= \widehat{\mathbf{E}}^s_{i,[(n-1)\frac{d_e}{N}:n\frac{d_e}{N}]} \\ \hat{\mathbf{r}}_k^n &= \widehat{\mathbf{R}}_{k,[(n-1)\frac{d_r}{N}:n\frac{d_r}{N}]} \\ \hat{\mathbf{e}}_i^{o,n} &= \widehat{\mathbf{E}}^o_{j,[(n-1)\frac{d_e}{N}:n\frac{d_e}{N}]} \end{aligned} \quad (10)$$

We set group N to 4 for WebNLG, and 1 for other datasets (that is, without further spliting).

[2] 768 is identical with BERT Base hidden size, and we use 128 in NYT/WebNLG to reduce computational footprints because these two datasets involve more candidate entities.

Dataset	No. of Instances w.r.t. Split			Entities (Avg.)	Relations	No. of Instances w.r.t. Multi-triples			
	Train	Dev	Test			$N = 1$	$1 < N <= 5$	$5 < N <= 25$	$N > 25$
DocRED	3053	1000	1000	19.5	96	48	561	3171	234
NYT*	56195	4999	5000	2.15	24	43397	22207	590	NA
WebNLG*	5019	500	703	3.15	171	2189	3969	64	NA
NYT	56196	5000	5000	2.16	24	43358	22237	601	NA
WebNLG	5019	500	703	3.26	216	2277	3862	83	NA

Table 4: Statistics of used datasets. * denotes task settings that only annotate the last word. N denotes the number of valid triples within an instance. We can see that these selected benchmarks all involve multiple triples, thus pose significant challenge for relation extraction systems.