

# Do the Benefits of Joint Models for Relation Extraction Extend to Document-level Tasks?

Pratik Saini and Tapas Nayak and Indrajit Bhattacharya

TCS Research, India

{pratik.saini,nayak.tapas,b.indrajit}@tcs.com

## Abstract

Two distinct approaches have been proposed for relational triple extraction - pipeline and joint. Joint models, which capture interactions across triples, are the more recent development, and have been shown to outperform pipeline models for sentence-level extraction tasks. Document-level extraction is a more challenging setting where interactions across triples can be long-range, and individual triples can also span across sentences. Joint models have not been applied for document-level tasks so far. In this paper, we benchmark state-of-the-art pipeline and joint extraction models on sentence-level as well as document-level datasets. Our experiments show that while joint models outperform pipeline models significantly for sentence-level extraction, their performance drops sharply below that of pipeline models for the document-level dataset.

## 1 Introduction

Relation extraction is a crucial NLP task for constructing and enriching knowledge bases. Traditional pipeline approaches (Riedel et al., 2010; Hoffmann et al., 2011; Zeng et al., 2014, 2015; Nayak and Ng, 2019; Jat et al., 2017) first identify entities followed by relation identification one entity pair at a time. In contrast, more recent joint approaches (Zeng et al., 2018; Takanobu et al., 2019; Nayak and Ng, 2020; Wei et al., 2020; Wang et al., 2020b; Zhong and Chen, 2021; Zheng et al., 2021; Li et al., 2021; Wei et al., 2020; Yan et al., 2021; Shang et al., 2022) not only identify entities and relations for the same triple together but also extract all relational triples together. Thus, these recent approaches are better suited for capturing complex interactions.

Joint models for relation extraction outperform traditional pipeline models for sentence-level datasets such as NYT (Riedel et al., 2010). A more natural and complex setting for relation extraction is at the document-level. In the document-level

task, relational triples may also span across sentences. Further, there may be long range interactions between different triples across sentences. As a result, the search space for joint models blows up with document size. So far, research for document-level datasets such as DocRED (Yao et al., 2019) has used pipeline approaches and avoided the joint approach.

In this paper, we investigate if the benefits of the joint approach extrapolate from sentence-level to document-level tasks. We benchmark 5 SOTA joint models and 3 SOTA pipeline models on sentence-level (NYT) and document-level (DocRED) datasets. We observe that the benefits of the SOTA joint models do not extend to document-level tasks. While performance of both classes of models drop sharply, joint models fall significantly below that of pipeline models. We perform extensive analysis to identify the short-comings of the two classes of models highlighting areas of improvement.

## 2 Relation Extraction Approaches

**Pipeline RE approaches** solve the RE task in two sequential steps. In Step 1, they use an NER model to identify the entities and entity mentions in the input text. In Step 2, they take the predicted entities and entity mentions as input, and predict all possible relations from a pre-defined relation set between pairs of entities. We use PL-Marker (Ye et al., 2022) as the NER module and KD-DocRE (Tan et al., 2022), SSAN (Xu et al., 2021a), and experiment with SAIS (Xiao et al., 2022) as relation classification models for our experiments, and train these for specific datasets. Following standard practice, we use gold standard entity mentions for training and validation of the relation classification models, while for inference of the test instances, we naturally use the predicted entity mentions as input. Note that in both steps, these models perform *independent* classification for each entity mention

Original	<a href="#">Kiato</a> ( , <a href="#">Sidirodromikos Stathmos Kiatou</a> ) is a railway station in <a href="#">Kiato</a> in the northern <a href="#">Peloponnese</a> , Greece . The station is located a kilometre west of the town , near the Greek National Road 8A ( <a href="#">Patras</a> – Corinth highway ) . It opened on 9 July 2007 as the western terminus of the line from Athens Airport . Initially the station served as an exchange point for passengers to <a href="#">Patras</a> on the old metre gauge SPAP line to <a href="#">Patras</a> , but all traffic was suspended indefinitely in December 2010 for cost reasons . The nearby old <a href="#">Kiato</a> station was also closed . Passengers for <a href="#">Patras</a> must now change to bus services at <a href="#">Kiato</a> . The station is served by one train per hour to <a href="#">Piraeus</a> .
Processed	<a href="#">Kiato</a> ( , <a href="#">Kiato</a> ) is a railway station in <a href="#">Kiato</a> in the northern <a href="#">Peloponnese</a> , Greece . The station is located a kilometre west of the town , near the Greek National Road 8A ( <a href="#">Peloponnese</a> – Corinth highway ) . It opened on 9 July 2007 as the western terminus of the line from Athens Airport . Initially the station served as an exchange point for passengers to <a href="#">Peloponnese</a> on the old metre gauge SPAP line to <a href="#">Peloponnese</a> , but all traffic was suspended indefinitely in December 2010 for cost reasons . The nearby old <a href="#">Kiato</a> station was also closed . Passengers for <a href="#">Peloponnese</a> must now change to bus services at <a href="#">Kiato</a> . The station is served by one train per hour to <a href="#">Peloponnese</a> .

Table 1: A sample document from DocRED and its processed version. The different mentions of the entity ‘Kiato’ and ‘Peloponnese’ are marked with blue and red respectively. In the processed text, different mentions are normalized with the first mention of these two entities which are ‘Kiato’ and ‘Peloponnese’.

Dataset	# Relations	Train		Validation		Test	
		# Context	# Triples	# Context	# Triples	# Context	# Triples
NYT	24	56,196	94,222	5,000	8,489	5,000	8,616
DocRED	96	2,572	20,233	284	2,187	924	7,337

Table 2: Statistics of the NYT24 and DocRED datasets. Context refers to a sentence or document.

and relation.

**Joint RE approaches** identify the entities and relations in a relational triple in an end-to-end fashion. Further, they consider the entire input text (sentence or document) and output a set of relational triples together, thus, capturing complex interactions across triples in theory. The flip side, naturally, is that they need to explore a significantly larger space of candidates, which grows combinatorially with the length of the input text. We use 5 SOTA models for experiments. PtrNet (Nayak and Ng, 2020) and REBEL (Huguet Cabot and Navigli, 2021) use the Seq2Seq approach. PtrNet generates the index position of entities in text whereas REBEL generates the tokens for the triples. OneRel (Shang et al., 2022) uses a table-based tagging approach. The tagging approaches of BiRTE (Ren et al., 2022) and GRTE (Ren et al., 2021) have a separate entity extraction process in their end-to-end modeling. We train these models in an end-to-end fashion as described in the respective papers.

### 3 Extraction Settings and Datasets

The original and simpler setting for relation extraction is *sentence-level*. This setting consists of individual sentences containing one or more relations as context. NYT (Riedel et al., 2010) is a large-scale and popular benchmark for sentence-

level RE, and we use this dataset as it is for our sentence-level experiments.

This setting, however, is restrictive since a large fraction of relations in natural text spans across multiple sentences. This is captured in the *document-level* relation extraction setting. Here, relational triples may be intra-sentence or inter-sentence, meaning that the head and tail entities of the relational triple can span across multiple sentences and require reasoning across multiple sentences to identify them. The task is to predict all these relations given an entire document as context. DocRED (Yao et al., 2019) is a benchmark document-level dataset that we use for our experiments.

Contexts in DocRED (avg. number of tokens around 197) are much longer than in NYT (avg. number of tokens around 37). However, training data size is much larger for NYT. Since the relation labels of the DocRED test set are not released, we use the original validation set as test set and split the training data for training and validation. DocRED has mostly been used for pipeline models. We needed additional processing to make it tractable for joint models. We remove the documents with overlapping entity mentions in the training and validation set. We get 2,856 documents from the training set and 924 documents from the validation set. Then, we replace all entity mentions with the first occurring entity mention

for each entity so that co-reference resolution is not required. We include an example of document processing for DocRED in Table 1. The details of the dataset splits of the NYT and DocRED for our experiments are included in Table 2<sup>1</sup>.

**Evaluation Metric:** We use ‘strict’ criteria for evaluation. We consider an extracted relational triple as correct only if two entities and relation exactly match with a ground truth triple. We report triple level precision, recall and F1 scores for the models.

**Parameter Settings:** We use  $BERT_{BASE}$  (cased) (Devlin et al., 2019) for document encoding for all the models except REBEL for which we have used  $BART_{BASE}$  (Lewis et al., 2019). We used NVIDIA Tesla V100 32GB GPU to train the models. We used other hyper-parameters as per provided in their respective papers.

## 4 Results and Discussion

Through our experiment, we try to find out the answers to following research questions (RQ).

### **RQ1: How do the two classes of models perform at sentence and documents scales?**

End-to-end performance of the joint models and the pipeline models on the NYT and DocRED datasets is shown in Table 3. On the sentence-level NYT dataset, both joint models and pipeline models achieve close to 0.90 F1 score. But, on DocRED, we see a huge drop in the F1 score for both categories. The pipeline models score below 0.60 whereas among the joint models GRTE and BiRTE perform around 0.45, the others drop to 0.20 or below. One reason for the drop for DocRED is the smaller training data size. But, it does not explain the gap of 10% F1 score between the pipeline and joint models when they performed almost at par for NYT. This suggests that joint models struggle with longer context and cross-sentence relations of documents. We investigate this in more detail next.

### **RQ2: Are some joint models better than others at document scale?**

Out of the 5 joint models, we see significantly higher drop in F1 score for OneRel and PtrNet than REBEL, GRTE, and BiRTE models. Given a document with  $L$  tokens and  $K$  predefined relations, OneRel maintains a three-dimensional matrix  $M_{L \times K \times L}$  and assigns tags for all possible triples.

When context length grows as in documents,  $M$  has many more negative tags and very few positive tags, which seems to affect OneRel performance. BiRTE and GRTE, on the other hand, extract the entities first separately and then classify the relations. While this is done in an end-to-end fashion, it is still similar to pipeline approach. This may be the reason for the smaller performance drop compared to pipeline models on DocRED.

PtrNet and REBEL are Seq2Seq models which use a decoder to extract the triples, so they possibly need more training data to learn from longer document contexts. Additionally, PtrNet extracts index positions for the entities. Since an entity may appear more than once in a document, we mark the first occurring index of the entity-mention to train this model. This very likely contributes to its poorer performance. On the other hand, REBEL outputs entities as text, and does not have this training issue.

### **RQ3: How different are performances for intra vs inter-sentence extraction?**

The fundamental difference between NYT and DocRED is that NYT contains only intra-sentence triples, whereas DocRED contains both intra and inter-sentence (cross-sentence) triples. In Table 5, we first show intra vs inter sentence relational triple distribution for the gold and model predictions on the DocRED dataset. Pipeline models have nearly the same distribution for gold and prediction. But, joint models are skewed towards intra-sentence relations. This suggests that joint models are very good at extracting intra-sentence triples but they struggle with inter-sentence triples. This is why joint models perform very well on the NYT dataset and fail to do so on DocRED.

In Table 4, we have reported the intra-sentence vs inter-sentence relations performance of the top-performing models on the DocRED test dataset. We see that all the models perform way better at intra-sentence extraction as compared to inter-sentence extraction, it demonstrates that inter-sentence extraction is significantly harder. We also see that pipeline models achieve around 10% higher F1 scores than the joint models for *both* intra and inter categories on DocRED. This shows that even for the familiar intra-sentence setting joint models face more difficulties compared to pipeline models when encountered with longer context and smaller training volume.

Lastly, we investigate the impact of the distance

<sup>1</sup>Our processed DocRED dataset is available at <https://github.com/pratiksaini4/nyt-docred-joint-pipeline-comparison>

	Model	NYT24			DocRED		
		P	R	F1	P	R	F1
Joint Models	OneRel	0.926	0.918	0.922	0.513	0.130	0.208
	BiRTE	0.914	0.920	0.917	0.522	0.402	0.454
	GRTE	0.929	0.924	0.926	0.586	0.373	0.456
	PtrNet	0.898	0.894	0.896	0.222	0.145	0.175
	Rebel	0.881	0.885	0.883	0.466	0.356	0.404
Pipeline Models	KD-DocRE	0.895	0.910	0.902	0.620	0.556	0.586
	SSAN	0.781	0.798	0.789	0.576	0.529	0.552
	SAIS	0.864	0.879	0.872	0.640	0.545	0.589

Table 3: Performance of the SOTA models on end-to-end relation extraction on NYT24 and DocRED datasets.

	Model	Intra			Inter		
		P	R	F1	P	R	F1
Joint Models	BiRTE	0.600	0.425	0.497	0.420	0.366	0.391
	GRTE	0.677	0.407	0.508	0.460	0.320	0.378
Pipeline Models	KD-DocRE	0.666	0.601	0.631	0.545	0.485	0.513
	SAIS	0.697	0.594	0.641	0.548	0.467	0.504

Table 4: Performance of top performing SOTA models on Intra vs Inter relational triples on DocRED dataset.

	Model	Predict	
		Intra %	Inter %
-	Gold	61	39
Joint Models	BiRTE	56	44
	GRTE	58	42
Pipeline Models	KD-DocRE	62	38
	SAIS	62	38

Table 5: Intra vs Inter relational triple distribution of SOTA models predictions. We include top two models from the joint and pipeline class for this analysis.

# Hops	Pipeline Models		Joint Models	
	KD-DocRE	SAIS	BiRTE	GRTE
1	0.50	0.45	0.38	0.33
2	0.44	0.44	0.34	0.31
3	0.48	0.49	0.36	0.31
4	0.48	0.49	0.37	0.31
5	0.51	0.54	0.39	0.32
6	0.42	0.43	0.29	0.32

Table 6: Recall score of Pipeline and Joint SOTA models on inter-sentence relations with respect to distance between subject and object entities.

between subject and object mentions in the context on the performance of inter-sentence relations. In Table 6, we record recall of SOTA models on inter-sentence relations for different subject-object hop distances. Hop distance  $k$  refers to the minimum sentence-level distance between the subject and object entity of a triple within the document being  $k$ . Again, we see that pipeline models outperform joint models by  $\sim 12\%$  for all hop distances, and

not just for longer ones.

#### RQ4: How is performance affected by training data size?

Next, we analyze how training volume affects performance for the two model classes for the simpler intra-sentence extraction task. Note that NYT contains such relations exclusively. Since DocRED has both categories, we prepare DocRED-Intra including only intra-sentence triples and the corresponding sentences. The size of these datasets are significantly different. DocRED-Intra has only  $\sim 6.5K$  training instances compared to  $94K$  for NYT. We train all the models with these intra-sentence triples and record their performance for DocRED-Intra in Table 7. Corresponding NYT performance is in Table 3. We observe a big gap of  $\sim 42 - 50\%$  for joint models and  $\sim 33\%$  for pipeline models in the performance between NYT and DocRED-Intra. This is due to the smaller training volume associated with a larger number of relations in DocRED. The notable disparity between pipeline and joint models in the case of DocRED-Intra demonstrates that joint models are not as effective at generalization compared to pipeline models, particularly when working with limited training data volumes and longer contexts.

#### RQ5: How different are entity extraction performances at sentence and document scales?

Finally, we aim to analyze if the huge gap in the performance of pipeline and joint models on

	Model	P	R	F1
Joint Models	BiRTE	0.527	0.462	0.492
	GRTE	0.544	0.346	0.423
Pipeline Models	KD-DocRE	0.524	0.619	0.567
	SAIS	0.485	0.610	0.540

Table 7: Performance of top performing SOTA models on DocRED-Intra dataset.

DocRED is affected by their performance on NER subtask of relation extraction. In Table 8, we include the performance of these models on the entity extraction task. Pipeline models have a separate NER model. The performance of this model - PL-Marker - on NER task is similar to that of the BiRTE and GRTE models for NYT dataset. But, for DocRED, BiRTE and GRTE perform much worse than PL-Marker, the drop in F1 score being around 25%. This, in turn, hurts their performance on the relational triple extraction. Though training data volume for DocRED is smaller, note that the PL-Marker model is trained on the DocRED dataset itself as are the BiRTE/GRTE models. This shows that, aside from overall extraction performance, joint models struggle with the NER subtask as well when training data is limited. This suggests that a separate NER model may be more useful in such settings.

Dataset	Model	P	R	F1
NYT24	PL-Marker	0.948	0.955	0.952
	BiRTE	0.955	0.954	0.954
	GRTE	0.958	0.956	0.957
DocRED	PL-Marker	0.942	0.934	0.938
	BiRTE	0.73	0.647	0.686
	GRTE	0.757	0.576	0.654

Table 8: Performance of PL-Marker, BiRTE and GRTE on the NER task for NYT24 and DocRED dataset.

## 5 Conclusion

While joint models for relational triple extraction have been shown to outperform pipeline models for sentence-level extraction settings, in this paper we have demonstrated, with extensive experimentation, that these benefits do not extend to the more realistic and natural document-level extraction setting, which entails longer contexts and cross-sentence relations. Experimenting with 5 SOTA joint models and 3 SOTA pipeline models, we have shown that while performance of both classes of models drops significantly in the more complex document setting, joint models suffer significantly more with longer context, inter-sentence relations and limited train-

ing data for the overall task as well as for subtasks such as NER. This aids in establishing a research agenda for joint models to extend the promised benefits of joint entity identification, relation classification, and joint extraction of all triples from the context. This pertains to the more challenging yet natural and crucial setting for relation extraction.

## 6 Limitations

The major limitation of this work is that we could only analyze 3 pipeline models and 5 joint models. Recently, many models have been proposed for this task both in pipeline and joint class. Out of these, we chose different kinds of SOTA models to cover the different design choices made by these models. We chose PtrNet (Nayak and Ng, 2020) and REBEL (Huguet Cabot and Navigli, 2021) as they used Seq2Seq model for this task. OneRel (Shang et al., 2022) used table-filling method whereas BiRTE (Ren et al., 2022) and GRTE (Ren et al., 2021) used sequentially extracting entities and relations in their end-to-end model.

## References

- Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *EMNLP*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
- Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy. Association for Computational Linguistics.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.
- Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. 2021. Entity and evidence guided document-level relation extraction. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepLANLP-2021)*, pages 307–315, Online. Association for Computational Linguistics.
- Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language

- generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2017. Improving distantly supervised relation extraction using word and entity based attention. In *Proceedings of the 6th Workshop on Automated Knowledge Base Construction*.
- Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multi-scale representation learning. 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 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.
- Xianming Li, Xiaotian Luo, Cheng Jie Dong, Daichuan Yang, Beidi Luan, and Zhen He. 2021. TDEER: An efficient translating decoding schema for joint extraction of entities and relations. In *EMNLP*.
- Makoto Miwa and Mohit Bansal. 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*.
- 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.
- Tapas Nayak and Hwee Tou Ng. 2019. Effective attention modeling for neural relation extraction. In *Proceedings of the Conference on Computational Natural Language Learning*.
- Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- Xutan Peng, Chenghua Lin, and Mark Stevenson. 2021. Cross-lingual word embedding refinement by  $\ell_1$  norm optimisation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2690–2701, Online. Association for Computational Linguistics.
- Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain. Association for Computational Linguistics.
- Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. In *EMNLP*.
- Feiliang Ren, Longhui Zhang, Xiaofeng Zhao, Shujuan Yin, Shilei Liu, and Bochao Li. 2022. 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 *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*.
- Y. Shang, Heyan Huang, and Xian-Ling Mao. 2022. OneRel: Joint entity and relation extraction with one module in one step. In *AAAI*.
- Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. N-ary relation extraction using graph-state LSTM. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2226–2235, Brussels, Belgium. Association for Computational Linguistics.
- Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xianrong Zeng, and Shengping Liu. 2020. Joint entity and relation extraction with set prediction networks. *CoRR*, abs/2011.01675.
- Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence*.
- Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.
- Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. HIN: hierarchical inference network for document-level relation extraction. *CoRR*, abs/2003.12754.
- Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020a. Global-to-local neural networks for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3711–3721, Online. Association for Computational Linguistics.

- Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Yang Wang. 2019. [Fine-tune bert for docred with two-step process](#). *CoRR*, abs/1909.11898.
- Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021. [UniRE: A unified label space for entity relation 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 220–231, Online. Association for Computational Linguistics.
- Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020b. [TPLinker: Single-stage joint extraction of entities and relations through token pair linking](#). In *COLING*.
- Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. [A novel cascade binary tagging framework for relational triple extraction](#). In *ACL*.
- Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. [SAIS: Supervising and augmenting intermediate steps for document-level relation extraction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2395–2409, Seattle, United States. Association for Computational Linguistics.
- Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. [Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268, Dublin, Ireland. Association for Computational Linguistics.
- Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021a. [Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction](#). *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. [Discriminative reasoning for document-level relation extraction](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1653–1663, Online. Association for Computational Linguistics.
- Wang Xu, Kehai Chen, and Tiejun Zhao. 2021c. [Document-level relation extraction with reconstruction](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14167–14175.
- Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. [A partition filter network for joint entity and relation extraction](#). In *EMNLP*.
- 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 ACL 2019*.
- Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. [Packed levitated marker for entity and relation extraction](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.
- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. [Distant supervision for relation extraction via piecewise convolutional neural networks](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. [Relation classification via convolutional deep neural network](#). In *Proceedings of the 25th International Conference on Computational Linguistics*.
- Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. [SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 524–534, Online. Association for Computational Linguistics.
- 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.
- Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 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*.
- Meishan Zhang, Yue Zhang, and Guohong Fu. 2017. [End-to-end neural relation extraction with global optimization](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1730–1740, Copenhagen, Denmark. Association for Computational Linguistics.
- Heng Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Ming Xu, and Yefeng Zheng. 2021. [PRGC: Potential relation and global correspondence based joint relational triple extraction](#). In *ACL*.
- Zexuan Zhong and Danqi Chen. 2021. [A frustratingly easy approach for entity and relation extraction](#). In *North American Chapter of the Association for Computational Linguistics*.
- Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. [Document-level relation extraction](#)

with adaptive thresholding and localized context pooling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14612–14620.

## A Appendix

### A.1 Details of Joint Models

#### A.1.1 PtrNet (Nayak and Ng, 2020)

PtrNet utilizes a seq2seq approach along with the pointer network-based decoding for jointly extracting entities and relations. Each triple contains the start and end index of the subject and object entities, along with the relation class label. Their decoder has two-pointer networks to identify the start and end index of the two entities and a classifier to identify the relation between the two entities. Decoder extracts a relational triple at each time steps and continue the process till there is no triple to extract. To ensure parity with other SOTA models, their BiLSTM encoder is replaced with BERT encoder.

#### A.1.2 REBEL (Huguet Cabot and Navigli, 2021)

REBEL utilizes an auto-regressive seq2seq model that streamlines the process of relation extraction by presenting triples as a sequence of text and uses special separator tokens, as markers, to achieve the linearization.. WordDecoder model of (Nayak and Ng, 2020) uses a similar approach using LSTMs whereas REBEL is a BART-based Seq2Seq model that utilizes the advantages of transformer model and pre-training. REBEL uses a more compact representation for the relational triples over WordDecoder model.

#### A.1.3 GRTE (Ren et al., 2021)

GRTE utilizes individual tables for each relation. The cell entries of a table denote the presence or absence of relation between the associated token pairs. It uses enhanced table-filling methods by introducing two kinds of global features. The first global feature is for the association of entity pairs and the second is for relations. Firstly, a table feature is generated for each relation, which is then consolidated with the features of all relations. This integration produces two global features related to the subject and object, respectively. These two global features are refined multiple times. Finally, the filled tables are utilized to extract all relevant triples.

#### A.1.4 OneRel (Shang et al., 2022)

OneRel frames the joint entity and relation extraction task as a fine-grained triple classification problem. It uses a three-dimensional matrix with relation-specific horns tagging strategy. The rows in this matrix refer to the head entity tokens and the columns in this matrix refer to the tail entity tokens from the original text. The scoring-based classifier checks the accuracy of the decoded relational triples. It discards the triples with low confidence.

#### A.1.5 BiRTE (Ren et al., 2022)

In this paper, a bidirectional tagging approach with multiple stages is utilized. BiRTE first discovers the subject entities and then identifies the object entities based on the subject entities. It then does this in the reverse direction, first discovering the object entities and identifying subject entities for the object entities. The final stage involves the relation classification of subject-object pairs. They perform these tasks jointly in a single model.

### A.2 Details of Pipeline Models

#### A.2.1 SSAN (Xu et al., 2021a)

This paper frames the structure of entities as defined by the specific dependencies between the entity mention pairs in a document. The proposed approach, SSAN, integrates these structural dependencies with the self-attention mechanism at the encoding stage. To achieve this, two transformation modules are included in each self-attention building block, which generates attentive biases to regulate the attention flow adaptively. This approach achieved SOTA performance on the DocRED dataset.

#### A.2.2 KD-DocRE (Tan et al., 2022)

This paper suggests a semi-supervised framework for extracting document-level relations. To achieve this, they exploit the inter-dependency among the relational triples through the implementation of an axial attention module. This approach leads to enhanced performance when dealing with two-hop relations. In addition to this, an adaptive focal loss is proposed as a means of resolving the issue of imbalanced label distribution for long-tail classes. Finally, to account for the difference between human-annotated data and distantly supervised data, knowledge distillation is utilized. Their experiments on the DocRED show the effectiveness of the approach.



### A.2.3 SAIS (Xiao et al., 2022)

The objective of this paper is to train the model to identify relevant contexts and entity types by using the Supervising and Augmenting Intermediate Steps (SAIS) approach for relation extraction. The SAIS framework proposed in this paper results in the extraction of relations that are of superior quality, owing to its more efficient supervision. By utilizing evidence-based data augmentation and ensemble inference, SAIS also improves the accuracy of the supporting evidence retrieval process while minimizing the computational cost.

## A.3 Details of NER Model

### A.3.1 PL-Marker (Ye et al., 2022)

In their approach for span representation, PL-Marker strategically uses levitated markers to consider the interrelation between pairs of spans. They propose a packing strategy that factors in neighbouring spans to improve the modeling of entity boundary information. Additionally, they utilize a subject-oriented packing approach, which groups each subject with its objects to effectively model the interrelations between the same subject span pairs.

## A.4 Related Work

**Sentence-level Relation Extraction:** Early approaches for relation extraction use two steps pipeline approach. The first step, Named Entity Recognition (NER), extracts entities from the text. The second step, Relation Classification (RC), identifies pairwise relations between the extracted entities (Zeng et al., 2014, 2015; Jat et al., 2017; Nayak and Ng, 2019). Pipeline methods fail to capture the implicit correlation between the two sub-tasks. They suffer from error propagation between the two stages. They cannot model the interaction among the relational triples.

To mitigate the drawbacks of pipeline approaches, recent works have focused on Joint entities and relation extraction. Joint approaches referred to as End-to-End Relation Extraction (RE) accomplish both tasks jointly. Training simultaneously on both NER and RC tasks allows for capturing more complex interactions among the multiple relational triples present in the context. Miwa and Bansal (2016) proposed a model that trained the NER and RC module in a single model. Nayak and Ng (2020); Cabot and Navigli (2021) propose seq2seq models for extracting the triples

in a sequence. Sui et al. (2020) casts the joint extraction task as a set prediction problem rather than a sequence extraction problem. Zhang et al. (2017); Wang et al. (2020b, 2021); Shang et al. (2022) formulate the NER and RC tasks as table filling problem where each cell of the table represents the interaction between two tokens. Ren et al. (2022); Zheng et al. (2021); Li et al. (2021); Yan et al. (2021); Wei et al. (2020) have separate NER and RC modules in the same model trained in an end-to-end fashion.

**Document-level Relation Extraction:** Recently, there has been a shift of interest towards document-level RE (Yao et al., 2019). Document-level relation extraction (DocRE) is known to be a more complex and realistic task compared to the sentence-level counterpart. DocRE typically involves large volumes of data, which can be computationally expensive to handle using joint models. Recent work in DocRE has avoided using joint models for this task as joint models are not scalable for long documents. In DocRE, there can be multiple mentions of an entity with different surface forms across the document and the evidence of the relations can spread across multiple sentences. In document-level RE, mostly pipeline approaches are proposed, joint extraction approaches are not explored for this task. Earlier works (Peng et al., 2021; Quirk and Poon, 2017; Song et al., 2018; Jia et al., 2019) used dependency graph between the two entities to find the relations. Recent works (Guo et al., 2019; Nan et al., 2020; Wang et al., 2020a; Zeng et al., 2020, 2021; Xu et al., 2021c,b) proposed graph-based approaches that use advanced neural techniques to do multi-hop reasoning. More recent Transformer-based approaches (Wang et al., 2019; Tang et al., 2020; Huang et al., 2021; Xu et al., 2021a; Zhou et al., 2021; Xie et al., 2022) use pre-trained language models to encode long-range contextual dependencies in the documents. Huang et al. (2021); Xie et al. (2022); Xiao et al. (2022); Tan et al. (2022) use neural classifier to identify the evidences for relations along with relation classification for performance improvement.