Exploiting Ubiquitous Mentions for Document-Level Relation Extraction

Ruoyu Zhang
Peking University
Beijing, China
ry_zhang@pku.edu.cn

Yanzeng Li
Peking University
Beijing, China
liyanzeng@stu.pku.edu.cn

Minhao Zhang
Peking University
Beijing, China
zhangminhao@pku.edu.cn

Lei Zou*
Peking University
Beijing, China
zoulei@pku.edu.cn

ABSTRACT

Recent years have witnessed the transition from sentence-level to document-level in relation extraction (RE), with new formulation, new methods and new insights. Yet, the fundamental concept, mention, is not well-considered and well-defined. Current datasets usually use automatically-detected named entities as mentions, which leads to the missing reference problem. We show that such phenomenon hinders models' reasoning abilities. To address it, we propose to incorporate coreferences (e.g. pronouns and common nouns) into mentions, based on which we refine and re-annotate the widely-used DocRED benchmark as $\mathcal{R}$-DocRED. We evaluate various methods and conduct thorough experiments to demonstrate the efficacy of our formula. Specifically, the results indicate that incorporating coreferences helps reduce the long-term dependencies, further improving models' robustness and generalization under adversarial and low-resource settings. The new dataset is made publicly available for future research.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

document-level relation extraction; information extraction; natural language processing

ACM Reference Format:
Ruoyu Zhang, Yanzeng Li, Minhao Zhang, and Lei Zou. 2023. Exploiting Ubiquitous Mentions for Document-Level Relation Extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591984

*Lei Zou is the corresponding author.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591984

1 INTRODUCTION

Relation extraction (RE) aims at extracting semantic relations among entities from plain texts, which plays a crucial role in information extraction and NLP research. Early studies mostly address this problem at sentence-level, which assumes that mentions of target entity pairs co-occurred within one single sentence [1, 10, 17]. However, this assumption strongly restricts the ability of knowledge discovery in real world scenarios, where valuable knowledge can be only inferred from multiple sentences [11, 16]. To this end, researchers have stepped forward to document-level relation extraction (DocRE) [6, 8, 14, 19].

DocRE naturally inherits and expands the setting from its sentence-level counterpart. Given a document D and a set of entities $\{e_i\}$, the task is to automatically predict relations among entities. The main difference between document-level RE and sentence-level RE is two-fold: (1) the number of potential entity pairs in DocRE is exponential in the number of entities, which poses extra burden for predictor; (2) since an entity might be mentioned multiple times in a single document, every entity e_i has a corresponding mention set $\{m_j\}$, which explicitly provides coreference information for later predictions.

However, the transition from sentence-level to document-level leaves a grey area: the ad hoc concept, *mention*, is created without being clearly defined. This causes further problems in practice. In the standard benchmark DocRED [16], a large amount of references (mainly pronouns and common nouns/noun phrases) are left out, and we refer this phenomenon as the *missing reference* problem. The omission of referential units hinders downstream models' ability to extract relations from documents. Our findings indicate that: (1) DocRE methods generally struggle with learning long-term dependencies. The *missing reference* problem increases the distance of entity pairs (shown by Figure 1), making it more difficult to perform inference; (2) besides, the *missing reference* problem reduces the robustness of RE performance facing different mention permutations.

To overcome these difficulties, we propose to incorporate coreferences into mentions, in addition to named entities. To unify these linguistic expressions, we formalize the concept of *mention* from both semantic and syntactic perspectives. Inspired by early named-entity recognition (NER) and coreference resolution (CR) works [4, 9], we provide a formal definition of mention, denoted as $\mathcal{R}$-mention, and further apply it to the widely-used DocRED dataset

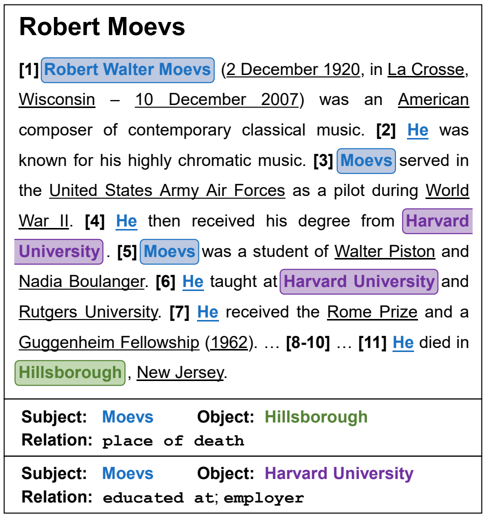

Figure 1: An example from the DocRED dataset. References such as pronoun "*He*"s are neglected in the original annotation. As a result, "*Moevs*" and "*Hillsborough*" are separated by several sentences; "*Moevs*" and "*Harvard University*"s fall in different sentences as well. These phenomena increase the difficulty to conduct reasoning within entity pairs.

to verify its effectiveness. We first investigate the wrong entity linking problem of the original dataset, and refine the data quality by correcting these wrong links. We then re-annotate the mention with our new definition and publish the new $\mathcal{R}$-DocRED dataset[1]. We evaluate various DocRE methods and conduct thorough experiments under different settings. Empirical results reveal that incorporating coreferences brings consistent advancements to multiple methods by reducing the long-term dependencies, and further improves models' robustness and generalization under adversarial and low-resource settings.

2 DEFINITION OF $\mathcal{R}$-MENTION

2.1 Formal Definition

In order to introduce coreferences into *mentions*, we start from an onomasiological perspective to clarify the concept. Given a specific entity of the real world, we say that a *mention* of it is a linguistic expression that designates the entity unambiguously in its context, including nouns, pronouns and noun phrases (NP).

Semantic relations. We shall first emphasize that our scheme only focus on the *identity* relation [4] between entities and their linguistic expressions, which is *symmetrical* (if A is identical to B, then B is identical to A) and *transitive* (if A is identical to B and B is identical to C, then A is identical to C). These properties induce an equivalent class of mentions with respect to a given entity.

Other similar semantic relationships, e.g. part-whole and set-subset, are out of our consideration. For instance, neither "*the Nobel Prize in Physics*" nor "*the Nobel Prize 2021*" is a mention of "*the Nobel Prize*".

[1]https://github.com/ridiculouz/r-docred

However, the nature of the *identity* relation poses a problem where an expression may refer to either of two contradictory entities [4]. Consider the following:

The stock price fell from \$10 to \$1.

If we assume that "*stock price*" is identical to "*\$10*" and "*stock price*" is identical to "*\$1*", we then come to the counter-intuitive conclusion that "*\$10*" is identical to "*\$1*". To prevent this collapsing reference chain phenomenon, we discard the referential links among mention and its conflicting entities to keep the integrity of each equivalent class. In the above example, "*stock price*" will link to neither of the two prices instead.

Parts of speech. Syntactically, we only focus on nominal groups (i.e. words that functions as nouns [3]) in text. Therefore, adjective forms of proper nouns or possessive adjectives are not markable (e.g. "American" is not a mention of "USA"). Additionally, we propose several heuristic rules:

(1) To avoid redundancy with subject, apposition and predicative nominal are not markable.
(2) Gerund and clause are not markable due to their complexity. This rule is also reasonable because few entities are mentioned by these expressions.
(3) To avoid ambiguity and nested annotation, nouns and NPs in modifier are not markable (e.g. "*America*" in "*Bank of America*").

Text spans. It is common in practice that a nominal group consists of complex syntactic structure. Therefore, it is necessary to determine the exact text span of mention. To keep concise and informative simultaneously, we stipulate that mention should be the shortest identifiable expression with respect to a given entity and document context.

For example:

…The bank was founded in San Francisco. …
…Einstein's former physics professor Hendrik Lorentz …

In the former example, the definite article "*the*" should be included in mention "*The bank*", because it denotes the entity aforementioned. While in the latter example, though "*Einstein's former physics professor Hendrik Lorentz*", "*physics professor Hendrik Lorentz*" and "*Hendrik Lorentz*" all refer to the same entity, one should use "*Hendrik Lorentz*" for conciseness.

2.2 Discussion

Comparing the vanilla mention and our formally-defined concept, there are two major differences: (1) many neglected referential units are recognized by our formula, mainly pronouns and common nouns/noun phrases (as shown in Figure 1); (2) our definition regularizes the concept of mention, and provides clear clues for human annotators to detect mentions from plain text. $\mathcal{R}$-mention considers both semantically and syntactically with respect to given context, and provides stronger clues for not only annotation but also downstream inference.

Some previous researchers attribute the implicit coreference resolution process as coreference reasoning, one kind of DocRE reasoning skills [15, 16]. However, one should notice that existing vanilla setting already provides some coreference information explicitly. As shown in Figure 1, both "*Robert Walter Moevs*" and "*Moevs*" are linked to the same person, while models need to recognize that vast remaining mention "*He*"s also link to this entity, without supervision signal. We argue that dividing coreference resolution into explicit and implicit parts is unnecessary. Instead, we show that this practice exaggerates the long-term dependency problem and loss of robustness (detailed analyses in § 4.3). Additionally, since our formula of mention can be naturally adapted to the task of coreference resolution, our $\mathcal{R}$-DocRED dataset (proposed in § 3) can serve as a new corpus for coreference resolution as well.

3 ANNOTATION

3.1 The DocRED Dataset

To further verify the effectiveness of $\mathcal{R}$-mention, we implement the new scheme on a commonly-used benchmark, DocRED [16]. It is a large-scale human-annotated document-level RE dataset that consists of 3,053/998/1,000 documents for training/development/testing, respectively. DocRED is constructed from English Wikipedia and Wikidata, involving 96 relations, 132,375 entities and 56,354 relational facts.

3.2 Refining Referential Links

The original annotation process conducted by Yao et al. [16] follows a "recommend-revise" pipeline, i.e. human annotators revise the machine-generated distantly supervised annotation. However, we observe that some mistakes made by the automatic recommendation are not corrected and retained in the output annotation. Therefore, we start our annotation from refining data quality.

Specifically, we focus on the entity linking problem, where the annotated mentions are not correctly linked to their corresponding entities. There are two major types of mistakes: (1) Duplicated entities. Two different annotated entities point to the same concept of real world. (2) Mismatches. An annotated mention is not a reference to its entity. We ask annotators to modify these wrong links, and merge duplicated entities. Our annotation follows the closed-world assumption [12], which means that no new entity is introduced.

3.3 Annotation Process and Quality Assurance

We adapt the FITAnnotator system [7] as our annotation platform, and recruit 9 educated annotators. The tasks of link refinement and mention annotation are conducted simultaneously during the annotation process.

To ensure high-quality annotations, we employ a two-step quality control and validation process during annotation: (1) Before annotation, a brief training session is conducted. Annotators first learn the guideline, and then annotate 30 passages randomly. A meta annotator checks the results with F_1 scores of mention annotation. If the F_1 score is below 90%, the meta annotator will explain to the annotators about their mistakes and repeat the process until the annotations are of high quality. (2) After annotation, we first perform auto-check and basic corrections with scripts. We then randomly sample from the annotations (30 passages from each annotator) and ask the meta annotator to manually check the results. The high agreement rate between meta annotator and recruited annotators indicates high annotation quality, as shown in Table 1.

	Merged Entity F_1	Mention F_1
Train	91.11 ± 9.770	84.87 ± 7.85
Dev	88.89 ± 11.48	84.51 ± 5.94
Test	85.56 ± 12.94	93.72 ± 3.64

Table 1: Agreement between meta annotator and recruited annotators, measured by F_1 scores and 95% confidence intervals of merged entity and mention annotation.

	Train	Dev	Test
# Documents	3,053	998	1,000
# Entities	56,440	18,530	18,539
# Merged	876	337	335
# Mentions	91,713	30,312	30,784
# Added	12,665	4,301	4,360
# Deleted	43	30	16
# Modified	142	52	50

Table 2: Statistics of $\mathcal{R}$-DocRED and new annotations.

3.4 The $\mathcal{R}$-DocRED Dataset

Our annotation results in the $\mathcal{R}$-DocRED dataset. Statistics of annotation are given in Table 2.

In general, $\mathcal{R}$-DocRED has slightly fewer entities and more mentions, and it well addresses the *missing reference* problem in original DocRED. We will then conduct in-depth analyses between DocRED and $\mathcal{R}$-DocRED from multiple perspectives in § 4.

4 EXPERIMENTS

4.1 Setup

Data. We take three versions of data for comparison. (1) **Vanilla**, i.e. original version of DocRED. (2) **Refined**, where original entity-mention links are refined according to § 3.2. (3) **Reference**, where mentions are further re-annotated under our formula with refined links. For the **reference** split, we report both performance with and without re-annotated mentions at test stage for thorough analysis[2].

Evaluation metrics. Following Yao et al. [16], we use F_1 as evaluation metric. We also report Ign F_1 when analyzing the **vanilla** and **refined** data. Ign F_1 denotes F_1 scores excluding relational facts shared by the training and dev/test sets, based on mention pair co-occurrence. Because the **reference** data involves a large amount of pronouns and pronoun pairs, we believe that Ign F_1 is uncomparable on this split.

Methods. We evaluate various methods for comprehensive analysis. **CNN** [17], **LSTM** [5], **Bi-LSTM** [2] and the **Context**-Aware model [13] are baselines used by Yao et al. [16]. We also report the results of **ATLOP** [20] and **GAIN** [18], two state-of-the-art representatives of sequence-based and graph-based methods. In all experiments, we conduct 3 runs with different random seeds and report the mean results.

[2]Without re-annotated mentions here is equivalent to use reference version at training stage and refined at test stage.

	Method	Vanilla	Refined	Reference (-/+)
I	**CNN**	43.45	47.44	47.91 / **48.61**
	LSTM	50.68	50.79	50.75 / **51.47**
	BiLSTM	50.94	51.57	51.43 / **52.37**
	Context	51.09	51.41	51.82 / **52.62**
II	**ATLOP**	61.09	61.25	61.62 / **61.73**
	GAIN	$60.21^{\dagger}$	60.53	60.26 / **60.80**

Table 3: F_1 scores on dev set (%)[3]. † is our re-implemented result. -/+ denotes the performance without and with re-annotated mentions at test stage. I/II stands for using GloVe/BERT$_{\text{base}}$ as embedding, respectively.

	Original F_1	Permute F_1
Refined	61.25 ± 0.07	58.45 ± 0.09
Reference (+)	61.73 ± 0.06	61.35 ± 0.07

Table 4: Results for permute operation on dev set.

	10%	20%	50%	100%
Refined	51.72	55.77	59.80	61.25
Reference (+)	52.51	56.19	60.03	61.73

Table 5: F_1 scores with different amounts of training data.

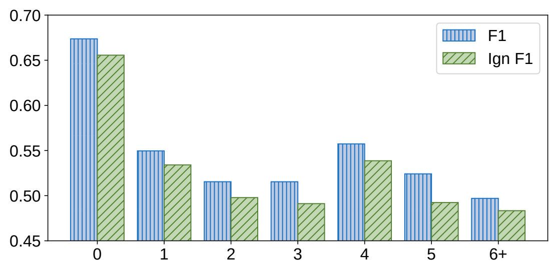

Figure 2: F_1/Ign F_1 scores under different distances. In general, performance drops as distance increases. Specifically, model performs significantly better when mention pairs fall in the same sentence (distance = 0).

4.2 Overall Performance

Table 3 shows the experiment results under different settings, from which we can observe that: (1) performance on refined data is consistently better than its vanilla counterpart, showing that data quality advances with link refinement; (2) F_1 scores under reference (+) setting further improve, which demonstrates the efficacy of $\mathcal{R}$-mention. Moreover, all models suffer a performance decay when removing the coreferences at test stage, which indicates that incorporating coreferences into mentions is beneficial for both training and inference. We will further analyse the advantages of $\mathcal{R}$-mention below.

4.3 Understanding the Merits of $\mathcal{R}$-Mention

In this section, we seek to answer the following questions: (1) How does coreferences help improve RE performance? (2) Can $\mathcal{R}$-mention perform better in real world scenarios, where the training data may be limited or noisy?

For the rest of this paper, we conduct studies with **ATLOP**, due to its tractable size and strong overall performance.

Long-term dependencies. We say that the distance of two entities is the minimal distance of their mention pairs, counted by sentence, where 0 indicates that there exists one mention pair falling in the same sentence. DocRE methods are sensitive to the distance of head and tail entities, and perform relatively poor learning long-term dependencies, as shown by Figure 2. The *missing reference* problem under vanilla setting increases the distance between entities, which hinders models' ability to reason. However, with $\mathcal{R}$-mention, more elements (e.g. pronouns) are annotated, making the distance distribution monotonically decrease, which alleviates this issue.

[3]The gold labels of test set are not publicly available. Since link refinement includes merging entities which changes entity indices in labels, we are unable to report results on test set.

Robustness with different permutations. We conduct a permutation experiment on dev set to test model's robustness.

Definition (Permutation). A permutation on a given entity is to shuffle its mentions' positions, while keeping their linguistic expressions unchanged.

Intuitively, for a specific entity, different permutations of mentions does not change the textual context, therefore should not affect the result of RE. Surprisingly, we find that this simple operation significantly decreases the F_1 score under vanilla setting (~3%), while with $\mathcal{R}$-mention the performance is much more consistent, as reported in Table 4. The above difference indicates that $\mathcal{R}$-mention can notably improve the robustness during inference stage.

Generalization under low-resource. Table 5 reports the F_1 scores with 10%, 20% and 50% of training data. $\mathcal{R}$-mention further shows advantages in low-resource scenarios, by consistently outperforms its refined counterpart at all levels. The promising results imply that $\mathcal{R}$-mention provides strong reasoning paths for inference and helps generalization when training data is limited.

5 CONCLUSION

In this paper, we delve into the *missing reference* problem in DocRE, and we propose to incorporate coreferences into mentions to solve the problem. We unify both named entities and coreferences with $\mathcal{R}$-mention. We then refine and re-annotate the DocRED benchmark as $\mathcal{R}$-DocRED. Comprehensive experiments and analyses demonstrate the effectiveness of our approach, especially under adversarial and low-resource settings.

ACKNOWLEDGMENTS

We would like to thank anonymous reviewers for their constructive comments. This work is partially supported by NSFC under grant 61932001 and U20A20174.

REFERENCES

[1] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2895–2905. https://doi.org/10.18653/v1/P19-1279

[2] Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional Recurrent Convolutional Neural Network for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 756–765. https://doi.org/10.18653/v1/P16-1072

[3] M.A.K. Halliday, C.M.I.M. Matthiessen, M. Halliday, and C. Matthiessen. 2014. *An Introduction to Functional Grammar*. Taylor & Francis.

[4] Lynette Hirschman and Nancy Chinchor. 1998. Appendix F: MUC-7 Coreference Task Definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. https://aclanthology.org/M98-1029

[5] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (11 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf

[6] Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three Sentences Are All You Need: Local Path Enhanced Document 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 2: Short Papers)*. Association for Computational Linguistics, Online, 998–1004. https://doi.org/10.18653/v1/2021.acl-short.126

[7] Yanzeng Li, Bowen Yu, Li Quangang, and Tingwen Liu. 2021. FITAnnotator: A Flexible and Intelligent Text Annotation System. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*. Association for Computational Linguistics, Online, 35–41. https://doi.org/10.18653/v1/2021.naacl-demos.5

[8] 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*. Association for Computational Linguistics, Online, 1546–1557. https://doi.org/10.18653/v1/2020.acl-main.141

[9] D. Nouvel, M. Ehrmann, and S. Rosset. 2016. *Named Entities for Computational Linguistics*. Wiley.

[10] Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3661–3672. https://doi.org/10.18653/v1/2020.emnlp-main.298

[11] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics* 5 (2017), 101–115. https://doi.org/10.1162/tacl_a_00049

[12] Raymond Reiter. 1978. *On Closed World Data Bases*. Springer US, Boston, MA, 55–76. https://doi.org/10.1007/978-1-4684-3384-5_3

[13] Daniil Sorokin and Iryna Gurevych. 2017. Context-Aware Representations for Knowledge Base Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1784–1789. https://doi.org/10.18653/v1/D17-1188

[14] 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*. Association for Computational Linguistics, Dublin, Ireland, 1672–1681. https://doi.org/10.18653/v1/2022.findings-acl.132

[15] 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*. Association for Computational Linguistics, Dublin, Ireland, 257–268. https://doi.org/10.18653/v1/2022.findings-acl.23

[16] 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*. Association for Computational Linguistics, Florence, Italy, 764–777. https://doi.org/10.18653/v1/P19-1074

[17] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation Classification via Convolutional Deep Neural Network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2335–2344. https://aclanthology.org/C14-1220

[18] 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)*. Association for Computational Linguistics, Online, 1630–1640. https://doi.org/10.18653/v1/2020.emnlp-main.127

[19] Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level Relation Extraction with Dual-tier Heterogeneous Graph. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 1630–1641. https://doi.org/10.18653/v1/2020.coling-main.143

[20] 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 (May 2021), 14612–14620. https://doi.org/10.1609/aaai.v35i16.17717