The State of Relation Extraction Data Quality: Is Bigger Always Better?

Erica Cai **Brendan O'Connor**
University of Massachusetts Amherst
{ecai,brenocon}@cs.umass.edu

Abstract

Relation extraction (RE) extracts structured tuples of relationships (e.g. *friend, enemy*) between entities (e.g. *Sherlock Holmes, John Watson*) from text, with exciting potential applications. Hundreds of RE papers have been published in recent years; do their evaluation practices inform these goals? We review recent surveys and a sample of recent RE methods papers, compiling 38 datasets currently being used. Unfortunately, many have frequent label errors, and ones with known problems continue to be used. Many datasets focus on producing labels for a large number of relation types, often through error-prone annotation methods (e.g. distant supervision or crowdsourcing), and many recent papers rely exclusively on such datasets. We draw attention to a promising alternative: datasets with a *small* number of relations, often in specific domains like chemistry, finance, or biomedicine, where it is possible to obtain high quality expert annotations; such data can more realistically evaluate RE performance. The research community should consider more often using such resources.

1 Introduction

Relation extraction (RE) methods extract tuple structures from unstructured text, where structures consist of types of relationships, e.g. *is a member of an organization*, and entities involved with them, e.g. *Samuel Gompers*, *American Federation of Labor*. Ding et al. (2021); Nadgeri et al. (2021); Xu and Barbosa (2019); Trisedya et al. (2019); Han et al. (2021) claim that applications of these methods include populating knowledge graphs, and we find that realistic downstream applications also benefit from these methods: for example, *automatic* extraction of relations could replace manual extraction of {*friend*, *enemy*} relations in Icelandic sagas (Mac Carron and Kenna, 2013), and could replace manual extraction of financial acquisition relations from news text in Gugler et al. (2003); Clougherty et al. (2014), helping to decrease labor in social science and humanities studies.

However, *significantly noisy dataset labels* and evaluation on datasets using *different performance metrics than intended* may impact estimation of RE method performance for a realistic setting. While some literature acknowledges the issues, various recent evaluations and efforts to construct datasets ignore them. We explore characteristics and application of 38 datasets for evaluating RE methods. While our review is preliminary, we find some persistent patterns:

- **Finding (F1):** Datasets with *larger numbers of relation types* often have ground truth labels assigned using distant supervision, which aligns entities in text to relation tuples in a knowledge base. Further, such datasets—whether their ground truth labels are assigned through distant supervision or through full manual annotation—are *vulnerable to significant labelling errors* (Tan et al., 2022; Stoica et al., 2021; Huang et al., 2022; Alt et al., 2020; Wang et al., 2022b).
- **Finding (F2):** Many recent studies (at least 40 in ACL/EMNLP/Findings, since 2021; details in §4, §A.1) *exclusively evaluate* RE on datasets with *larger numbers (24+) of relation types*.
- **Finding (F3):** We find that datasets *with fewer relation types*, although less widely used, are more likely to be *annotated by experts and have domain-specific text* (Herrero-Zazo et al., 2013; Luan et al., 2018).

Although many evaluations are exclusively on datasets with larger numbers of relation types (**F2**), we encourage more of the research community to broaden evaluation to datasets with fewer relation types, given **F3**. Evaluating on various datasets which are more likely to have expert-annotated labels and domain-specific text may have a higher chance to provide a more accurate estimation of RE performance for a realistic setting, given **F1**.

Findings of the Association for Computational Linguistics ACL 2024, pages 7893–7906

2 Potential of relation extraction methods for realistic applications

Instead of automatically extracting relations, studies in various applications *manually* extract relations from text for downstream analysis. Mac Carron and Kenna (2013) compare social network structure in Icelandic saga text and modern-day social networks by manually extracting *friend* and *enemy* relationships between saga characters. Gugler et al. (2003) and Clougherty et al. (2014) use manually extracted financial acquisition relations from various news texts, even though automatic extraction has been investigated for these relations (Freitag, 1998).

Why automatically extract relations? In addition to decreasing labor, relation extraction methods in the NLP literature seem increasingly promising for realistic applications, departing from the traditional fully supervised, sentence-level setting on general news and Wikipedia text. Recent work explores low resource methods which require few or no labelled instances for use (Han et al., 2018b; Sabo et al., 2021; Zhang and Lu, 2022), and new datasets and models target specific domains, e.g. financial, legal, and biomedical (Gurulingappa et al., 2012; Herrero-Zazo et al., 2013; Peng et al., 2019; Hendrycks et al., 2021; Sharma et al., 2022). Further, more recent methods extract relation instances across many sentences at the *document-level* instead of from single sentences (Yao et al., 2019; Jain et al., 2020; Xiao et al., 2020; Li et al., 2023).

3 Disconnect between datasets and realistic application

Given promise of RE for realistic applications (§2), we ask: *Does performance on the datasets that RE methods evaluate on give an accurate indication of how such methods will perform in a variety of realistic use cases?*

We explore, for datasets: (1) potential **sources of error** when assigning relation instance labels, (2) **literature about labelling errors and evaluation** on widely used datasets, (3) **persistence of ignoring literature** that addresses the issues.

3.1 Sources of labelling error

To clarify sources of labelling issues, we first define the RE task as having inputs $\mathcal{R}$ of relation types, text $\mathcal{T}$ from which to extract and classify relations, and some labelled examples of relation instances in text for training. The output are relations $\langle e_1, r, e_2 \rangle$ where $r \in \mathcal{R}$ is a relation type (e.g. *part of*), and e_1, e_2 are head and tail entities respectively in a sentence or document of $\mathcal{T}$ (e.g. *Neolithic*, *Stone Age*) as in the example:

*Discovery of late **Stone Age** jugs suggest that intentionally fermented beverages existed at least as early as the **Neolithic** period (c. 10000 BC).* (Han et al., 2018b)

$$\langle e_1 = \text{Neolithic}, r = \textit{part-of}, e_2 = \text{Stone Age} \rangle$$

We selected 38 datasets to analyze (§4), finding that label error may occur from distant supervision and manual annotation as follows:

Missing labels from distant supervision. Distant supervision assigns labels to relation instances in text by aligning entity pairs e_1, e_2 in text to relation triples $\langle e_1, r, e_2 \rangle$ in a knowledge base (KB), e.g. Wikipedia text to Wikidata entries (Mintz et al., 2009; Bunescu and Mooney, 2007). While distant supervision is a common practice because it allows automatic assignment of labels on large amounts of data, it is prone to significant errors (§3.2). Many efforts aim to address false positive labels by manually verifying them (Han et al., 2018b; Huguet Cabot et al., 2023) or designing models to be more selective about assigning labels (Riedel et al., 2010; Bing et al., 2015; Xiao et al., 2020; Jia et al., 2019; Zeng et al., 2018; Feng et al., 2018; Surdeanu et al., 2012; Qin et al., 2018a,b). However, few efforts (Chen et al., 2021; Hao et al., 2021; Tan et al., 2022; Xu et al., 2013; Roller et al., 2015; Xie et al., 2021) aim to address **missing labels**, which occurs when related entities in text are not part of a KB relation triple, and therefore are not assigned a label. Missing labels may be common; for the DocRED dataset with distant supervision-assigned labels, Huang et al. (2022) and Tan et al. (2022) find up to $2/3$ of true labels are missing.

Ambiguous annotation guidelines. For full manual annotation, label quality depends on whether annotation guidelines specify relation type definitions clearly and unambiguously with respect to other relation types, and on annotator quality. For example, Stoica et al. (2021) and Alt et al. (2020) observe ambiguous documentation in the TACRED dataset for the pair of relation types "Person:Other_Family" and "No_Relation", which they find to be responsible for many label errors.

3.2 Discussion on widely-used datasets: Label and evaluation errors

We provide examples of discussion on labelling errors described in §3.1 that may affect estimation of RE performance for realistic applications. We review use of four popular datasets over years 2019-2024, chosen based on citation count from Semantic Scholar: NYT-FB (Riedel et al., 2010) (778 cit.) and TACRED (Zhang et al., 2017) (657 cit.) for sentence-level RE, where entities in relation triples belong to the same sentence, DocRED (Yao et al., 2019) (312 cit.) for document-level RE, where entities could be anywhere in a document, and FewRel 1.0 (Han et al., 2018b; Gao et al., 2019) (460 cit.) for few-shot RE, where the number of training examples is limited.

Discussion on errors in DocRED (312 citations since 2019). DocRED has distant supervision assigned labels and annotations for 96 relation types on 5053 Wikipedia documents. To reduce false positive labels assigned through distant supervision (§3.1), annotators review entities and relation types, filtering out incorrect labels. However, the missing label issue related to distant supervision approaches persists: Huang et al. (2022) identify that almost two-thirds of ground truth relations are not labelled from their re-annotation of 96 documents. Tan et al. (2022) independently find that approximately 64.6% of ground truth relations are missing. They further replace DocRED with Re-DocRED which has 4053 documents, by training RE models on the original distant supervised data and manually validating relation instances.

Discussion on errors in NYT-FB (778 citations since 2019). NYT-FB has distant supervision assigned labels for 24 relations, aligning New York Times article text (Sandhaus, 2008) over years 2005-2007 with Freebase relations (Bollacker et al., 2008). While precision of labels is 91%, their recall struggles, and many efforts aim to address this in diverse ways. Wang et al. (2022b) find issues on labels of 40 out of 100 randomly selected sentences. Hoffmann et al. (2011) add more labelled relations by joining tables in Freebase. Zeng et al. (2015) alternatively manually annotate the test set. Han et al. (2018a) add more ground truth labels by linking the text with another knowledge graph, FB60K. Zhu et al. (2020) manually annotate a larger test set as NYT-H.

Mismatch between intended versus actual use of FewRel 1/2 (460 citations since 2019). FewRel 1/2 has distant supervision assigned and manually verified labels for 80 relation types, and uses accuracy as a performance metric. Each relation type has 700 instances, with a one-to-one mapping of each instance to a sentence. Since a sentence may have more ground truth relation instances that are not labelled, precision and recall metrics are not able to accurately assess performance. However, among others, Zhao et al. (2023a); Lv et al. (2023); Zhao et al. (2023b); Wang et al. (2022a); Najafi and Fyshe (2023); Chen and Li (2021) use FewRel for computing precision and recall.

Discussion on errors in TACRED (667 citations since 2019). Fully manually annotated datasets are also vulnerable to labelling issues, such as TACRED, where Alt et al. (2020) found at least 50% of samples need to be relabelled and Stoica et al. (2021) found 23.9% of labels were incorrect. A revised version, Re-TACRED, has annotators from Amazon Mechanical Turk and improved annotation guidelines that remove ambiguity of relation definitions. On analysis using several RE methods, Stoica et al. (2021) found an average improvement of 14 F1 score on Re-TACRED, suggesting that label quality heavily impacts performance.

3.3 Propagation of errors despite discussion

Despite these multiple papers that reveal labelling and evaluation issues in relation extraction, we find that many recent works ignore and continue to propagate the issues. We investigate two categories of widespread evaluation issues in current work: (1) persistent use of original versions of datasets or of unintended evaluation metrics, and (2) continued introduction of datasets that face the same issues as previous ones (e.g., missing labels from distant supervision). In this section, the papers that we cite are from ACL/EMNLP/Findings venues.

Persistent use of original versions of datasets. However, recent evaluations still use original versions of these datasets or use unintended performance metrics. Many evaluations use TACRED (467 cit., since 2021) as opposed to Re-TACRED (72 cit., since 2021), without noting labelling issues (Wan et al., 2023; Zhao et al., 2023c; Chen et al., 2023b; Wang et al., 2022b; Sainz et al., 2021). Some methods still use FewRel, designed to measure performance using accuracy, to evaluate precision and recall, which are not appropriate performance metrics for the dataset Zhao et al. (2023a); Lv et al. (2023); Zhao et al. (2023b); Wang et al.

(2022a); Najafi and Fyshe (2023); Chen and Li (2021). Despite revised versions of NYT-FB, several evaluations still use original NYT-FB (Wu and Shi, 2021; Hao et al., 2021; Hu et al., 2020). For document-level relation extraction methods, some evaluations use original DocRED (Li et al., 2023) but luckily, many evaluations are using Re-DocRED.

New datasets are as vulnerable to missing labels. While more literature points out issues of various datasets, new datasets such as CodRED (Yao et al., 2021), T-rex (Elsahar et al., 2018), and RED-FM (Huguet Cabot et al., 2023) still do not consider the recall issue of distant supervision-assigned labels discussed in (§3.1) that has caused many labelling issues for other datasets, e.g. DocRED (Yao et al., 2019), NYT-FB (Riedel et al., 2010).

4 Which datasets are more susceptible to noise?

To help determine if any characteristics lead a dataset to be more susceptible to noise discussed in §3, we find:

- Datasets with *larger numbers of relation types*, which tend to have labels assigned using distant supervision, are *vulnerable to significant labelling and evaluation errors*.
- Many evaluations *exclusively use* datasets with *larger numbers (24+) of relation types* for evaluation (Zhang et al., 2023a; Wang et al., 2023a; Lu et al., 2023a), and 37 more papers in 2021-2023 ACL/EMNLP/Findings venues (§A.1).
- Datasets with labels for *fewer relation types* are *more likely to be annotated by experts and have domain-specific text* (Herrero-Zazo et al., 2013; Luan et al., 2018).

We compile a list of English datasets that have been used for RE evaluation from two sources. First, we search for papers at several NLP and machine learning venues[1] over years 2019-2023 that have the keyword "relation extraction" in their title, read a random sample of 100 such papers, and record all datasets used by each paper in evaluations. Second, we add all datasets mentioned in (Zhao et al., 2023e)'s relation extraction survey. This results in 38 datasets.

For each dataset, we read its original paper and/or documentation to record metadata including how labels were assigned, the type of annotator (if any), the domain, the dataset's size, and number of citations on Semantic Scholar (with some attempt to restrict to uses of the dataset).[2] See §A.2 for more details, including metadata for all 38 (Table 2). Table 1 shows a portion of this information for the 21 most-cited datasets.

Dataset	Labels?	Dom?	# rel
ADE (Gurulingappa et al., 2012)	Man-Exp	Bio	1
BC5CDR (Lin et al., 2016)	Man-Exp	Bio	1
CONLL04 (Roth and Yih, 2004)	Man	Gen	5
DDI (Herrero-Zazo et al., 2013)	Man-Exp	Bio	5
SciERC (Luan et al., 2018)	Man-Exp	Sci	6
i2b2 2010 (Uzuner et al., 2011)	Man-Exp	Bio	8
SemEval Task 8 (Hendrickx et al., 2010)	Man	Gen	9
ChemProt (Peng et al., 2019)	Man-Exp	Chem	14
REFinD (Kaur et al., 2023)	Man-Exp	Fin	22
ACE04 (Doddington et al., 2004)	Man-Exp	Gen	24
****NYT-FB (Riedel et al., 2010)**	**DS**	**Gen**	**24**
RED-FM (Huguet Cabot et al., 2023)	DS	Gen	32
DialogRE (Yu et al., 2020)	Man	Dia	37
****TACRED (Zhang et al., 2017)**	**Man**	**Gen**	**42**
****FewRel 1.0 (Han et al., 2018b)**	**DS**	**Gen**	**80**
****DocRED (Yao et al., 2019)**	**DS**	**Gen**	**96**
WikiZSL (Chen and Li, 2021)	DS	Gen	113
WebNLG (Gardent et al., 2017)	Oth	Gen	171
CodRED (Yao et al., 2021)	DS	Gen	276
SRED-FM (Huguet Cabot et al., 2023)	DS	Gen	400
T-rex (Elsahar et al., 2018)	DS	Gen	615

Table 1: Metadata on popular RE datasets by citation count (§A.2), where columns contain numbers of relation types for each dataset, method of assigning labels (*Man*ually annotated), by experts (*Man-Exp*), *D*istant *S*upervision (sometimes with subsequent manual filtering), *Oth*er), and *dom*ain of text and relation types (*Bio*medical, *Gen*eral-purpose (i.e. Wiki/news), *Sci*ence, *Chem*istry, *Dia*logue, *Fin*ancial). ** indicates the widely used datasets discussed in §3.2.

Trends: On larger numbers of relation types. Tables 1 and 2 show that the more relation types a dataset has labels for, the more likely that labels are assigned through distant supervision. Such datasets are susceptible to various labelling issues—§3.1 discusses sources of potential issues and §3.2 provides examples of significance of the issues. Further, we find at least 40 papers in ACL/EMNLP/Findings since 2021 (§A.1) that exclusively evaluate on such (24+ rel. types) datasets.

Trends: On fewer relation types. Datasets with fewer relation types are more likely to avoid distant supervision labelling issues and be manually annotated by experts (Herrero-Zazo et al., 2013; Luan et al., 2018; Hendrycks et al., 2021; Gurulingappa et al., 2012; Peng et al., 2019). Originally, such datasets contain general-purpose text, e.g. ACE (Doddington et al., 2004), SemEval Task 8 (Hendrickx et al., 2010), and Conll04 (Roth and Yih,

[1]Proceedings of ACL, Proceedings of EMNLP, AAAI Conference Proceedings, or Findings of the ACL.

[2]`www.semanticscholar.org`, accessed February 2024.

2004). Increasingly, new datasets cover other domains such as DDI (biomedical, 5 rel types), where text is from the DrugBank database and Medline abstracts and relations involve drug-drug interactions (Herrero-Zazo et al., 2013), and SciERC (scientific, 6 rel types), where text is 500 scientific abstracts (Luan et al., 2018).

5 Recommendations

Despite data quality challenges that may affect estimation of RE method performance on realistic applications (§3), we find potential for using RE methods in real-world applications such as those described in §2. Based on findings in §4, we provide two types of recommendations: (1) on selecting datasets to use for evaluation, and (2) on constructing future datasets to use for evaluation.

On selecting datasets for evaluation. We encourage the research community to broaden evaluation to include datasets with smaller numbers of relation types, which are more likely to be annotated by experts, and ideally to use multiple such datasets. This helps test the flexibility of a method across diverse relation types and domains. To further strengthen confidence in dataset quality, researchers can also manually check correctness of a sample of labels on familiar relation types (if any), and check the literature for potential revised versions of a dataset.

While we advocate for smaller and higher quality data, we note a counterargument that larger datasets—even with label noise—are crucial for training many relation extraction methods. However, the point of relation extraction research is to support *applications*, and we believe training data will be sparse and very expensive to obtain in most realistic settings, since annotations require significant domain expertise—heavy supervision is not a feasible modeling approach. Therefore evaluation ought to be the primary role of relation extraction annotation, where more accurate labels are ideal, even if there are fewer of them.

On future construction of datasets. While evaluation on noisy datasets may help to provide a rough indication of RE method performance, noisy labels render datasets unhelpful for accurately estimating performance in a real world application. Therefore, considering strategies to increase recall of labels assigned through distant supervision using approaches such as Xie et al. (2021) could help to label larger datasets more accurately and more efficiently—we find recent datasets (Huguet Cabot et al., 2023) aim to increase precision, but do not check or address recall of labels. Further, defining relation types unambiguously is helpful for avoiding manual labelling issues such as in Zhang et al. (2017).

6 Conclusion

Relation extraction is a popular task with hundreds of relevant papers published in recent years. Methods continue to improve and are becoming more promising for real world applications. First, we examined factors that potentially undermine relation extraction evaluation. Next, we provided recommendations to overcome these challenges, involving broadening evaluation to include smaller and higher quality datasets, and considering strategies to increase the recall of labels in new datasets, to improve estimation of relation extraction performance for realistic settings.

7 Limitations

This paper reviews 38 English datasets for evaluating RE methods, but will have missed datasets if they did not appear in the random sample of 100 papers from the four NLP/AI venues, or the survey paper (§4,§A.2). In particular, this data collection procedure may be more likely to miss datasets that are less frequently used.

In our final 38 datasets reported in the metadata tables, we restrict entries to original versions of datasets, and as mentioned in §3, do not list or analyze revised variants such as Re-TACRED for TACRED, Re-DocRED for DocRED, and NYT-H for NYT-FB; these three are instead described within §3.2). We found that revised versions of datasets tend to be not widely cited, and leave further analysis for future work.

The paper points out several patterns—that datasets with annotations for smaller numbers of relation types are more likely to be annotated by experts and be domain-specific, and that datasets that have annotations for larger numbers of relation types are more likely to have labels assigned from distant supervision. These statements are often, but not always, true.

Acknowledgements

We thank the UMass NLP group for feedback. This material is based upon work supported by National Science Foundation award 1845576. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Abhyuday Bhartiya, Kartikeya Badola, and Mausam. 2022. DiS-ReX: A multilingual dataset for distantly supervised relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 849–863, Dublin, Ireland. Association for Computational Linguistics.

Lidong Bing, Sneha Chaudhari, Richard Wang, and William Cohen. 2015. Improving distant supervision for information extraction using label propagation through lists. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 524–529, Lisbon, Portugal. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Sam Brody, Sichao Wu, and Adrian Benton. 2021. Towards realistic few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5338–5345, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 576–583, Prague, Czech Republic. Association for Computational Linguistics.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online. Association for Computational Linguistics.

Haotian Chen, Bingsheng Chen, and Xiangdong Zhou. 2023a. Did the models understand documents? benchmarking models for language understanding in document-level relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6418–6435, Toronto, Canada. Association for Computational Linguistics.

Jhih-wei Chen, Tsu-Jui Fu, Chen-Kang Lee, and Wei-Yun Ma. 2021. H-FND: Hierarchical false-negative denoising for distant supervision relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2579–2593, Online. Association for Computational Linguistics.

Xiudi Chen, Hui Wu, and Xiaodong Shi. 2023b. Consistent prototype learning for few-shot continual relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7409–7422, Toronto, Canada. Association for Computational Linguistics.

Nancy Chinchor and Elaine Marsh. 1998. Appendix D: MUC-7 information extraction task definition (version 5.1). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998.*

Joseph A Clougherty, Klaus Gugler, Lars Sørgard, and Florian W Szücs. 2014. Cross-border mergers and domestic-firm wages: Integrating "spillover effects" and "bargaining effects". *Journal of International Business Studies*, 45(4):450–470.

Li Cui, Deqing Yang, Jiaxin Yu, Chengwei Hu, Jiayang Cheng, Jingjie Yi, and Yanghua Xiao. 2021. Refining sample embeddings with relation prototypes to enhance continual 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 232–243, Online. Association for Computational Linguistics.

Ning Ding, Xiaobin Wang, Yao Fu, Guangwei Xu, Rui Wang, Pengjun Xie, Ying Shen, Fei Huang, Haitao Zheng, and Rui Zhang. 2021. Prototypical representation learning for relation extraction. *ICLR*.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *AAAI Conference on Artificial Intelligence*.

Dayne Freitag. 1998. Toward general-purpose learning for information extraction. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 404–408, Montreal, Quebec, Canada. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

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

Klaus Gugler, Dennis C Mueller, B.Burcin Yurtoglu, and Christine Zulehner. 2003. The effects of mergers: an international comparison. *International Journal of Industrial Organization*, 21(5):625–653.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. of Biomedical Informatics*, 45(5):885–892.

Jiale Han, Bo Cheng, and Wei Lu. 2021. Exploring task difficulty for few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2616, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018a. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018b. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2018:1884–1895.

Kailong Hao, Botao Yu, and Wei Hu. 2021. Knowing false negatives: An adversarial training method for distantly supervised relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9661–9672, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. In *NeurIPS Datasets and Benchmarks*.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus. *J. of Biomedical Informatics*, 46(5):914–920.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A novel large-scale language understanding task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545, Berlin, Germany. 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: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. 2020. SelfORE: Self-supervised relational feature learning for open relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3673–3682, Online. Association for Computational Linguistics.

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2022. Does recommend-revise produce reliable annotations? an analysis on missing instances in DocRED. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6241–6252, Dublin, Ireland. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. RED$^{\rm fm}$: a filtered and multilingual relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4326–4343, Toronto, Canada. Association for Computational Linguistics.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. ARNOR: Attention regularization based noise reduction for distant supervision relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1399–1408, Florence, Italy. Association for Computational Linguistics.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Prakash Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. Refind: Relation extraction financial dataset. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180–2.

Ruben Kruiper, Julian Vincent, Jessica Chen-Burger, Marc Desmulliez, and Ioannis Konstas. 2020. In layman's terms: Semi-open relation extraction from scientific texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1500, Online. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Jing Li, Yequan Wang, Shuai Zhang, and Min Zhang. 2023. Rethinking document-level relation extraction: A reality check. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5715–5730, Toronto, Canada. Association for Computational Linguistics.

Xiangyu Lin, Tianyi Liu, Weijia Jia, and Zhiguo Gong. 2021. Distantly supervised relation extraction using multi-layer revision network and confidence-based multi-instance learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 165–174, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Fangchao Liu, Hongyu Lin, Xianpei Han, Boxi Cao, and Le Sun. 2022a. Pre-training to match for unified low-shot relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5785–5795, Dublin, Ireland. Association for Computational Linguistics.

Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. 2021. Element intervention for open 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 4683–4693, Online. Association for Computational Linguistics.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022b. A simple yet effective relation information guided approach for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 757–763, Dublin, Ireland. Association for Computational Linguistics.

Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. 2023a. Anaphor assisted document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15453–15464, Singapore. Association for Computational Linguistics.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2023b. Multi-hop evidence retrieval for cross-document relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10336–10351, Toronto, Canada. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Bo Lv, Xin Liu, Shaojie Dai, Nayu Liu, Fan Yang, Ping Luo, and Yue Yu. 2023. DSP: Discriminative soft prompts for zero-shot entity and relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5491–5505, Toronto, Canada. Association for Computational Linguistics.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, and Yaqian Zhou. 2021. SENT: Sentence-level distant relation extraction via negative training. 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 6201–6213, Online. Association for Computational Linguistics.

P. Mac Carron and Ralph Kenna. 2013. Viking sagas: Six degrees of icelandic separation social networks from the viking era. *Significance*, 10(6):12–17.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

Abhishek Nadgeri, Anson Bastos, Kuldeep Singh, Isaiah Onando Mulang', Johannes Hoffart, Saeedeh Shekarpour, and Vijay Saraswat. 2021. KGPool: Dynamic knowledge graph context selection for relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 535–548, Online. Association for Computational Linguistics.

Saeed Najafi and Alona Fyshe. 2023. Weakly-supervised questions for zero-shot relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3075–3087, Dubrovnik, Croatia. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Gabriele Picco, Marcos Martinez Galindo, Alberto Purpura, Leopold Fuchs, Vanessa Lopez, and Thanh Lam Hoang. 2023. Zshot: An open-source framework for zero-shot named entity recognition and relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 357–368, Toronto, Canada. Association for Computational Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018a. DSGAN: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Melbourne, Australia. Association for Computational Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147, Melbourne, Australia. Association for Computational Linguistics.

Vipul Rathore, Kartikeya Badola, Parag Singla, and Mausam. 2022. PARE: A simple and strong baseline for monolingual and multilingual distantly supervised relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–354, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML/PKDD*.

Roland Roller, Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2015. Improving distant supervision using inference learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 273–278, Beijing, China. Association for Computational Linguistics.

Shachar Rosenman, Alon Jacovi, and Yoav Goldberg. 2020. Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3702–3710, Online. Association for Computational Linguistics.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium*, LDC2008T19.

Alessandro Seganti, Klaudia Firląg, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1946–1955, Online. Association for Computational Linguistics.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2022. Finred: A dataset for relation extraction in financial domain. In *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 595–597. ACM.

George Stoica, Emmanouil Antonios Platanios, and Barnab'as P'oczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *AAAI Conference on Artificial Intelligence*.

Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang, and Soujanya Poria. 2023. Uncertainty guided label denoising for document-level distant relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15960–15973, Toronto, Canada. Association for Computational Linguistics.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240, Florence, Italy. Association for Computational Linguistics.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L Duvall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association: JAMIA*, 18 5:552–6.

Erik M. van Mulligen, Annie Fourrier-Réglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifirò, Jan A. Kors, and Laura Inés Furlong. 2012. The eu-adr corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45 5:879–84.

Severine Verlinden, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2021. Injecting knowledge base information into end-to-end joint entity and relation extraction and coreference resolution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1952–1957, Online. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Jiaxin Wang, Lingling Zhang, Jun Liu, Xi Liang, Yujie Zhong, and Yaqiang Wu. 2022a. MatchPrompt: Prompt-based open relation extraction with semantic consistency guided clustering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7875–7888, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peiyi Wang, Yifan Song, Tianyu Liu, Binghuai Lin, Yunbo Cao, Sujian Li, and Zhifang Sui. 2022b. Learning robust representations for continual relation extraction via adversarial class augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6264–6278, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. 2023a. Improving unsupervised relation extraction by augmenting diverse sentence pairs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12136–12147, Singapore. Association for Computational Linguistics.

Xinyi Wang, Zitao Wang, and Wei Hu. 2023b. Serial contrastive knowledge distillation for continual few-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12693–12706, Toronto, Canada. 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.

Heming Xia, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. 2023. Enhancing continual relation extraction via classifier decomposition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10053–10062, Toronto, Canada. Association for Computational Linguistics.

Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. 2020. Denoising relation extraction from document-level distant supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3683–3688, Online. Association for Computational Linguistics.

Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. 2021. Revisiting the negative data of distantly supervised 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 3572–3581, Online. Association for Computational Linguistics.

Peng Xu and Denilson Barbosa. 2019. Connecting language and knowledge with heterogeneous representations for neural relation extraction. 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 3201–3206, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 665–670, Sofia, Bulgaria. Association for Computational Linguistics.

Jianhao Yan, Lin He, Ruqin Huang, Jian Li, and Ying Liu. 2019. Relation extraction with temporal reasoning based on memory augmented distant supervision. 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 1019–1030, Minneapolis, Minnesota. Association for Computational Linguistics.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. Entity concept-enhanced few-shot 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)*, pages 987–991, Online. Association for Computational Linguistics.

Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. 2021. CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4452–4472, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Jonathan Yellin and Omri Abend. 2021. Paths to relation extraction through semantic structure. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2614–2626, Online. Association for Computational Linguistics.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2020. Dwie: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58:102563.

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*, pages 1753–1762, Lisbon, Portugal. 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, Shizhu He, Kang Liu, and Jun Zhao. 2018. Large scaled relation extraction with reinforcement learning. In *AAAI Conference on Artificial Intelligence*.

Dongxu Zhang, Sunil Mohan, Michaela Torkar, and Andrew McCallum. 2022. A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1073–1082, Marseille, France. European Language Resources Association.

Liang Zhang, Chulun Zhou, Fandong Meng, Jinsong Su, Yidong Chen, and Jie Zhou. 2023a. HyperNetwork-based decoupling to improve model generalization for few-shot relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6213–6223, Singapore. Association for Computational Linguistics.

Peiyuan Zhang and Wei Lu. 2022. Better few-shot relation extraction with label prompt dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6996–7006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, and Lei Zou. 2023b. A novel table-to-graph generation approach for document-level joint entity and relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10853–10865, Toronto, Canada. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Xue Mengge, Tingwen Liu, and Li Guo. 2021. From what to why: Improving relation extraction with rationale graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 86–95, Online. Association for Computational Linguistics.

Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. 2021. A relation-oriented clustering method for open relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9707–9718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jun Zhao, WenYu Zhan, Xin Zhao, Qi Zhang, Tao Gui, Zhongyu Wei, Junzhe Wang, Minlong Peng, and Mingming Sun. 2023a. RE-matching: A fine-grained semantic matching method for zero-shot relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6680–6691, Toronto, Canada. Association for Computational Linguistics.

Jun Zhao, Yongxin Zhang, Qi Zhang, Tao Gui, Zhongyu Wei, Minlong Peng, and Mingming Sun. 2023b. Actively supervised clustering for open relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4985–4997, Toronto, Canada. Association for Computational Linguistics.

Jun Zhao, Xin Zhao, WenYu Zhan, Qi Zhang, Tao Gui, Zhongyu Wei, Yun Wen Chen, Xiang Gao, and Xuanjing Huang. 2023c. Open set relation extraction via unknown-aware training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9453–9467, Toronto, Canada. Association for Computational Linguistics.

Kang Zhao, Hua Xu, Jiangong Yang, and Kai Gao. 2022. Consistent representation learning for continual relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3402–3411, Dublin, Ireland. Association for Computational Linguistics.

Wenzheng Zhao, Yuanning Cui, and Wei Hu. 2023d. Improving continual relation extraction by distinguishing analogous semantics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1162–1175, Toronto, Canada. Association for Computational Linguistics.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2023e. A comprehensive survey on deep learning for relation extraction: Recent advances and new frontiers.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021a. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021b. 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.

Tong Zhu, Haitao Wang, Junjie Yu, Xiabing Zhou, Wenliang Chen, Wei Zhang, and Min Zhang. 2020. Towards accurate and consistent evaluation: A dataset for distantly-supervised relation extraction. In *International Conference on Computational Linguistics*.

A Appendix

A.1 List of 40 papers that exclusively evaluate on datasets with 24+ relation types

The following papers exclusively evaluate RE methods on datasets with 24+ relation types; these papers are drawn from *Proceedings of ACL*, *Proceedings of EMNLP*, or *Findings of the ACL*, between 2021 and 2023: Wang et al. (2023a); Lu et al. (2023a); Wang et al. (2022b); Zhang et al. (2023a); Zhang and Lu (2022); Wang et al. (2022a); Lin et al. (2021); Sainz et al. (2021); Han et al. (2021); Wu and Shi (2021); Brody et al. (2021); Zhao et al.

(2021); Zeng et al. (2020); Hu et al. (2020); Rosenman et al. (2020); Zhao et al. (2023d,b,a); Chen et al. (2023a); Zhao et al. (2023c); Zhang et al. (2023b); Sun et al. (2023); Picco et al. (2023); Lv et al. (2023); Xia et al. (2023); Lu et al. (2023b); Wang et al. (2023b); Liu et al. (2022a); Rathore et al. (2022); Liu et al. (2022b); Zhao et al. (2022); Cui et al. (2021); Liu et al. (2021); Ma et al. (2021); Zheng et al. (2021b); Yang et al. (2021); Zhang et al. (2021); Nadgeri et al. (2021); Verlinden et al. (2021); Yellin and Abend (2021).

The papers involve various types of RE methods, including sentence-level, which could be evaluated using many datasets, document-level, for which there are fewer datasets to evaluate on yet still some datasets with small and some with large numbers of relation types, and low-resource, which could be evaluated using many datasets. This is not a comprehensive list of papers within 2021–2023 that exclusively evaluate on 24+ relation type datasets.

A.2 Preliminary analysis details on 38 datasets

Table 2 column description. Each row of Table 2 corresponds to a dataset, listing its name, year of introduction, method of construction, domain of text and relation types, number of relations that it has labels for, size (with units of number of sentences, abstracts, or documents), and citation count from Semantic Scholar since 2019 of all papers that are not a “background" citation according to Semantic Scholar and have the word “relation" in their title. To verify that these filtering rules extract papers that use the dataset to evaluate an RE method, we manually checked 20 such papers that matched these filtering requirements and found that all of them use the dataset to evaluate an RE method.

Datasets in Table 1. To select a subset of datasets for Table 1, we chose all datasets that have 12+ citation counts, where the citation counts pass the filtering rules above. We also added datasets that were published in 2023 to Table 1 since they do not have time to accumulate citations yet.

On selected datasets for tables 1 and 2. Tables 1 and 2 show original versions of datasets; as mentioned in §3, some datasets have multiple revised versions such as TACRED with Re-TACRED, DocRED with Re-DocRED, and NYT-FB with NYT-H. Although revised versions of datasets are not in the tables, we discussed well known ones in §3.2. Many revised versions are not well-cited.

On trends in Table 2. The trends in Table 1 of §4 are also in Table 2: datasets with labels for larger numbers of relation types tend to have labels assigned using distant supervision, and tend to have general purpose text.

We observe a stronger correlation between the number of relation types and the method of assigning labels than between the size of a dataset and the method of assigning labels, but we observe a correlation for both comparisons.

On exceptions in Table 2. Two datasets have labels assigned in other ways than full manual annotation and distant supervision, noted with an ‘Oth’ entry in Table 2: WebNLG (Gardent et al., 2017), where human annotators manually convert one or several sets of triples from a KB into sentences, and where other annotators verify if each resulting sentence is faithful to the triple/s and seems natural; and FOBIE (Kruiper et al., 2020), which uses the Journal of Experimental Biology and Biomed Central Journal as text, and where annotators manually correct all initial annotations that pass an automated search for relations through trigger word keyword-matching. Additionally, DWIE (Zaporojets et al., 2020), noted with a ‘Both’ entry, uses distant supervision to assign some labels and performs full manual annotation to assign others.

Some of the datasets in Table 2 do not provide labels for the exact relation extraction task defined in the paper (with output $\langle e_1, r, e_2\rangle$), but for a similar task — WIKITIME (Yan et al., 2019) includes a “time" component in its output relation tuple $\langle e_1, r, e_2, t\rangle$. The WikiReading (Hewlett et al., 2016) task is to predict entities and properties of text given the text and a relation type such as *original language of work* or *country*. The CUAD (Hendrycks et al., 2021) task is not binary, but n-ary, outputting tuples of the form $\langle r, e_1, e_2, e_3 ...\rangle$ where more than two entities could be part of a tuple.

Dataset	Labels?	Domain?	# rel	Size	# filt. cit.
ADE (Gurulingappa et al., 2012)	Man-Exp	Bio	1	21k	22
BC5CDR (Li et al., 2016)	Man-Exp	Bio	1	1500 articles	51
Spouse (Hancock et al., 2018)	Man	Gen	1	27.7k	6
Disease (Hancock et al., 2018)	Man	Gen	1	11.5k	6
GENIA (Kim et al., 2003)	Man-Exp	Bio	2	2000 abstracts	4
FOBIE (Kruiper et al., 2020)	Oth	Bio	3	1.5k	0
EU ADR (van Mulligen et al., 2012)	Man-Exp	Bio	3	100 abstracts	4
MUC 7 (Chinchor and Marsh, 1998)	Man	Gen	3	-	0
CONLL04 (Roth and Yih, 2004)	Man	Gen	5	1.4k	24
DDI (Herrero-Zazo et al., 2013)	Man-Exp	Bio	5	31k	21
SciERC (Luan et al., 2018)	Man-Exp	Sci	6	4716 relations	61
i2b2 2010 (Uzuner et al., 2011)	Man-Exp	Bio	8	877 reports	12
SemEval Task 8 (Hendrickx et al., 2010)	Man	Gen	9	10.7k	136
Materials Science Procedural Text (Mysore et al., 2019)	Man-Exp	Mat	14	2.1k	0
ChemProt (Peng et al., 2019)	Man-Exp	Chem	14	36.4k	22
ChemDisGene (Zhang et al., 2022)	Man-Exp	Chem	18	523 abstracts	5
SciREX (Jain et al., 2020)	DS	Sci	21	438 documents	4
REFinD (Kaur et al., 2023)	Man	Fin	22	6.8k	4
MNRE (Zheng et al., 2021a)	Man	Gen	23	15.4k	10
ACE04 (Doddington et al., 2004)	Man-Exp	Gen	24	30.9k	26
****NYT-FB** (Riedel et al., 2010)	**DS**	**Gen**	**24**	**66.2k**	**237**
CUAD (Hendrycks et al., 2021)	Man-Exp	Leg	25	13.1k	1
FinRED (Sharma et al., 2022)	DS	Fin	29	6.8k	-
RED-FM (Huguet Cabot et al., 2023)	DS	Gen	32	43.7K	3
SMiLER (Seganti et al., 2021)	DS	Gen	36	1.1M	3
DialogRE (Yu et al., 2020)	Man	Dia	37	7.9k	26
DiS-ReX (Bhartiya et al., 2022)	DS	Gen	37	1.8M	3
****TACRED** (Zhang et al., 2017)	**Man**	**Gen**	**42**	**119.4k**	**203**
WIKITIME (Yan et al., 2019)	DS	Gen	57	137.6k	-
DWIE (Zaporojets et al., 2020)	Both	Gen	65	-	11
****FewRel 1.0** (Han et al., 2018b)	**DS**	**Gen**	**80**	**70k**	**132**
****DocRED** (Yao et al., 2019)	**DS**	**Gen**	**96**	**5k documents**	**137**
WikiZSL (Chen and Li, 2021)	DS	Gen	113	94.4k	20
WebNLG (Gardent et al., 2017)	Oth	Gen	171	5.7k	26
CodRED (Yao et al., 2021)	DS	Gen	276	-	5
SRED-FM (Huguet Cabot et al., 2023)	DS	Gen	400	46.6M	3
T-rex (Elsahar et al., 2018)	DS	Gen	615	6.2M	8
WikiReading (Hewlett et al., 2016)	DS	Gen	884	18.6M	4

Table 2: Metadata on 38 RE datasets, where columns contain numbers of relation types for each dataset, method of assigning labels (*Man*ually annotated), by experts (*Man-Exp*), *D*istant *S*upervision (sometimes with subsequent manual filtering), *Oth*er), and domain of text and relation types (*Bio*medical, *Gen*eral-purpose (i.e. Wiki/news), *Sci*ence, *Chem*istry, *Dia*logue, *Fin*ancial, *Leg*al, *Mat*erials science), numbers of sentences in the dataset (sometimes unavailable, -), and numbers of citations that the dataset has according to Semantic Scholar (sometimes unavailable, -) after applying the filters that "relation" must be in the title, that the citation cannot be in the background section, and that time range is since 2019. ** indicates the widely used datasets discussed in §3.2.