Manual Evaluation Matters: Reviewing Test Protocols of Distantly Supervised Relation Extraction

Tianyu Gao[1], **Xu Han**[2], **Yuzhuo Bai**[2], **Keyue Qiu**[2], **Zhiyu Xie**[2], **Yankai Lin**[3]
Zhiyuan Liu[2*], **Peng Li**[3], **Maosong Sun**[2] **and Jie Zhou**[3]
[1]Department of Computer Science, Princeton University
[2]Department of Computer Science and Technology, Tsinghua University
[3]Pattern Recognition Center, WeChat AI, Tencent Inc
tianyug@princeton.edu
{hanxu17, qky18, byz18, xiezy19}@mails.tsinghua.edu.cn

Abstract

Distantly supervised (DS) relation extraction (RE) has attracted much attention in the past few years as it can utilize large-scale auto-labeled data. However, its evaluation has long been a problem: previous works either take costly and inconsistent methods to manually examine a small sample of model predictions, or directly test models on auto-labeled data—which, by our check, produce as much as 53% wrong labels at the entity pair level in the popular NYT10 dataset. This problem has not only led to inaccurate evaluation, but also made it hard to understand where we are and what's left to improve in the research of DS-RE. To evaluate DS-RE models more credibly, we build manually-annotated test sets for two DS-RE datasets, NYT10 and Wiki20, and thoroughly evaluate several competitive models, especially the latest pre-trained ones. The experimental results show that the manual evaluation can indicate very different conclusions from automatic ones, especially some unexpected observations, e.g., pre-trained models can achieve dominating performance while being more susceptible to false-positives compared with previous methods. We hope that both our manual test sets and observations can help advance future DS-RE research.[1]

1 Introduction

Relation extraction (RE) aims at extracting relational facts between entities from the text. One crucial challenge for building an effective RE system is how to obtain sufficient annotated data. To tackle this problem, Mintz et al. (2009) propose distant supervision (DS) to generate large-scale auto-labeled data by aligning relational facts in knowledge graphs (KGs) to text corpora, with the core assumption that one sentence mentioning two entities is likely to express the relational facts between the two entities from KGs.

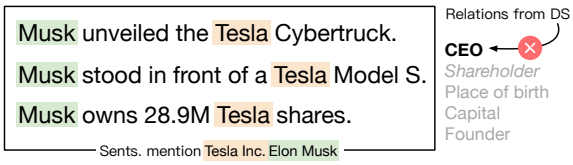

Figure 1: Typical errors made by DS evaluation. In the figure, DS labels the bag with only the relation `CEO`, while none of the sentences express the relation. Also, it misses a correct relation `shareholder` due to the incompleteness of the knowledge graphs.

As DS can bring hundreds of thousands of auto-labeled training instances for RE without any human labor, DS-RE has been widely explored in the past few years (Riedel et al., 2010; Hoffmann et al., 2011; Zeng et al., 2015; Lin et al., 2016; Feng et al., 2018; Vashishth et al., 2018) and has also been widely extended to other related domains, such as biomedical information extraction (Peng et al., 2017; Quirk and Poon, 2017) and question answering (Bordes et al., 2015; Chen et al., 2017).

Although DS-RE has achieved great success, we identify one severe problem for the current DS-RE research—its evaluation. Existing works usually take two kinds of evaluation methods following Mintz et al. (2009): *held-out evaluation*, which directly uses DS-generated test data to approximate the trend of model performance, and *human evaluation*, which manually checks the most confident relational facts predicted by DS-RE models. Since manually checking is costly, most works with human evaluation only examine a small proportion of the predictions. Moreover, different works may sample different splits of data, making human evaluation inconsistent across the literature. Most recent studies even skip the human evaluation for the above factors and only take the held-out one.

*Corresponding author e-mail: liuzy@tsinghua.edu.cn
[1]Our code and data are publicly available at https://github.com/thunlp/opennre.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1306–1318
August 1–6, 2021.

However, the held-out evaluation can be quite noisy: there are many false-positive cases, where the sentences do not express the auto-labeled relations at all; besides, due to the incompleteness of KGs, DS may miss some relations, just as shown in Figure 1. After checking $9,744$ sentences in the held-out set of NYT10 (Riedel et al., 2010), the most popular DS-RE dataset, we found that about 53% of the entity pairs are wrongly labeled, emphasizing the need for a more accurate evaluation.

To make DS-RE evaluation more credible and alleviate the trouble of manual checking for later work, we build human-labeled test sets for two DS-RE datasets: NYT10 (Riedel et al., 2010) and Wiki20 (Han et al., 2020). For NYT10, we manually annotate sentences with positive DS relations in its held-out test set. We also use a fine-tuned BERT-based (Devlin et al., 2019) RE model to predict all "N/A" (not applicable) sentences, and manually label the top $5,000$ sentences scored as having a relation. Additionally, we merge some unreasonably split relations and reduce the number of relation types from 53 to 25. For Wiki20 dataset, we utilize both the relation ontology and human-labeled instances of an existing supervised dataset Wiki80 (Han et al., 2019) for the test, and then re-organize the DS training data accordingly.

Based on the newly-constructed benchmarks, we carry out a thorough evaluation of existing DS-RE methods, as well as incorporating recently advanced pre-trained models like BERT (Devlin et al., 2019). We found that our manually-annotated test sets can indicate very different conclusions from the held-out one, especially with some surprising observations: (1) although pre-trained models lead to large improvements, they also suffer from false-positives more severely, probably due to the pre-encoded knowledge they have (Petroni et al., 2019); (2) existing DS-RE denoising strategies that have been proved to be effective generally do not work for pre-trained models, suggesting more efforts needed for DS-RE in the era of pre-training. To conclude, our main contributions in this work are:

- We provide large human-labeled test sets for two DS-RE benchmarks, making it possible for later work to evaluate in an accurate and efficient way.
- We thoroughly study previous DS-RE methods using both held-out and human-labeled test sets, and find that human-labeled data can reveal inconsistent results compared to the held-out ones.
- We discuss some novel and important observations revealed by manual evaluation, especially with respect to pre-trained models, which calls for more research in these directions.

2 Related Work

Relation extraction is an important NLP task and has gone through significant development in the past decades. In the early days, RE models mainly take statistical approaches, such as pattern-based methods (Huffman, 1995; Califf and Mooney, 1997), feature-based methods (Kambhatla, 2004; Zhou et al., 2005), graphical methods (Roth and Yih, 2002), etc. With the increasing computing power and the development of deep learning, neural RE methods have shown a great success (Liu et al., 2013; Zeng et al., 2014; Zhang and Wang, 2015; Zhang et al., 2017). Recently, pre-trained models like BERT (Devlin et al., 2019) have dominated various NLP benchmarks, including those in RE (Baldini Soares et al., 2019; Zhang et al., 2019b). All these RE methods focus on training models in a supervised setting and require large-scale sufficient human-annotated data.

To generate large-scale auto-labeled data without human effort, Mintz et al. (2009) first use DS to label sentences mentioning two entities with their relations in KGs, which inevitably brings wrongly labeled instances. To handle the noise problem, Riedel et al. (2010); Hoffmann et al. (2011); Surdeanu et al. (2012) apply multi-instance multi-label training in DS-RE. Following their settings, later research mainly takes on two paths: one aims at selecting informative sentences from the noisy dataset, using heuristics (Zeng et al., 2015), attention mechanisms (Lin et al., 2016; Han et al., 2018c; Zhu et al., 2019), adversarial training (Wu et al., 2017; Wang et al., 2018; Han et al., 2018a), and reinforcement learning (Feng et al., 2018; Qin et al., 2018); the other incorporates external information like KGs (Ji et al., 2017; Han et al., 2018b; Zhang et al., 2019a; Hu et al., 2019), multilingual corpora (Verga et al., 2016; Lin et al., 2017; Wang et al., 2018), as well as relation ontology and aliases (Vashishth et al., 2018). Recently, pre-trained DS-RE models have also been explored, including both domain-general (Alt et al., 2019; Xiao et al., 2020) and domain-specific (Amin et al., 2020) models. Some other latest works (Peng et al., 2020) utilize DS data for intermediate pre-training in order to boost supervised RE tasks.

As mentioned in our introduction, the evalua-

Dataset	#rel	Train			Validation			Test		
		#facts	#sents	N/A	#facts	#sents	N/A	#facts	#sents	N/A
NYT10†	53	18,409	522,611	74%	-	-	-	1,950	172,448	96%
NYT10	25	17,137	417,893	80%	4,062	46,422	80%	3,899 (1,940)	9,744 (157,859)	32% (96%)
Wiki20†	454	203,176	1,050,246	48%	4,333	29,145	48%	4,333	28,897	48%
Wiki20	81	157,740	698,721	59%	17,485	64,607	73%	56,000	137,986	25%

Table 1: The statistics of the datasets used for our benchmarks, including both the original (†) and our modified versions. We list the numbers of relations (#rel), relational facts (#facts) and sentences (#sents), and N/A rate (N/A) for these datasets. For NYT10, numbers in brackets are for the held-out test, otherwise for bag-level manual test.

tion of DS-RE has long been a problem, especially since many existing methods solely rely on auto-labeled test data. Some preliminaries have noticed this problem: Jiang et al. (2018); Zhu et al. (2020) also annotate the test set of NYT10, yet Jiang et al. (2018) only sample $2,040$ sentences from it, and Zhu et al. (2020) discard all N/A data from DS, which are an important part of DS evaluation, and assume that the original held-out data have either the DS relations or no relation at all, while we find that a large proportion of held-out data actually express some other relations; Li et al. (2020) propose active testing, an iterative method to correct the bias of DS evaluation. However, it still requires consistent human efforts during each evaluation phase. To the best of our knowledge, our work, building benchmarks with large-scale manually-labeled test data, conducts the most comprehensive human evaluations of DS-RE methods so far.

3 DS-RE Datasets

In this section, we introduce the way we build the manually-annotated test sets for NYT10 (Riedel et al., 2013) and Wiki20 (Han et al., 2020). We show the statistics of these datasets in Table 1.

3.1 NYT10 Dataset

NYT10 is constructed by aligning facts from the FreeBase (Bollacker et al., 2008) with the New York Times (NYT) corpus (Sandhaus, 2008). The original NYT10 dataset contains 53 relations (including N/A). After thoroughly examining the dataset, we found that (1) there are many duplicate instances in the dataset, (2) there is no public validation set, and some previous works directly take the test set to tune the model, and (3) the relation ontology is not reasonable for an RE task. Therefore, we first clean the dataset by removing duplicate sentences, split a validation set, and then merge some of the relations as described below.

A New Relation Ontology One example of NYT10's improper relation ontology is the relations related to state/province capitals. There are 12 such relations in the original dataset, representing region capitals of different countries, while some of these relations even do not have instances in the test set. We merge these 12 relations as one unified relation `/location/region/capital`, and we also merge 3 relations related to organization locations as `/business/location`. Besides, we delete relations that only show up in either the training set or the test set (most of which have very few sentences). In the end, we get 25 relations in the new dataset. We provide the detailed relation list of the new dataset in Appendix A.

Interestingly, we found that training models with the new dataset leads to a slight performance degeneration, as shown in Table 2, which is counter-intuitive (since merging classes usually makes the task easier). We suspect that the original relation ontology provides heuristics for the model. For example, models can learn from the original ontology that every sentence with a "US state" as the head entity expresses the fine-grained relation `/location/us_state/capital`, which is a shortcut that cannot be acquired with the merged relation `/location/region/capital`. This shows the bias of the original NYT10 dataset being inappropriately exploited by models.

Annotated Test Set The original NYT10 dataset only provides an auto-labeled test set for the held-out evaluation. Based on the original test set, we manually annotate all sentences that have a positive DS label. In addition to that, we also fine-tune a BERT model (as described in §4.2) to predict the relations of all sentences originally labeled as N/A, and annotate the $5,000$ sentences with the highest predicted scores of non-N/A relations. For each sentence, we ask 4 annotators to decide whether it expresses one or more relations among 25 rela-

Model	Original		New	
	AUC	F1	AUC	F1
PCNN+bag+AVG	31.8	38.6	28.4	35.8
PCNN+bag+ATT	33.6	40.2	32.2	39.1

Table 2: AUC and micro F1 (%) of model predictions on the original NYT10 relation ontology and our new NYT10 relation ontology (using held-out test).

tions. Note that one sentence may have multiple relations. Specially, we do not provide any relation suggestions using DS labels or model predictions to annotators, in order to avoid the annotators being biased by the external information.

When aggregating the final annotation results, we consider each relation for each sentence independently. The sentence is regarded to express one relation if more than half of the annotators labeled it with this relation. If one sentence gets no votes for any positive relations in the above process, then it is labeled as N/A. For annotation conflicts (i.e., no candidate relation gets more than half votes), the authors manually annotate these sentences.

Finally, we obtain the human-labeled test set with $9,744$ sentences, 32% of which are N/A instances. It contains $5,174$ entity pairs and $3,899$ manually-verified relational facts in total. After comparing it with the corresponding original DS-generated labels, we found that at the fact level, the DS annotations only have a precision of 69.1% and a recall of 33.9%. At the entity pair level, the accuracy of DS labels is only 47.1%. This emphasizes the need to take the human-annotated test set for more accurate evaluation in DS-RE.

3.2 Wiki20 Dataset

Han et al. (2020) construct the Wiki20 dataset by aligning the English Wikipedia corpus with Wikidata (Vrandečić and Krötzsch, 2014). To provide an annotated test set for it, we utilize Wiki80 (Han et al., 2019), a supervised RE dataset with 80 relations from Wikidata. We re-organize Wiki20 by adopting the same relation ontology as Wiki80 and re-splitting the train/validation/test sets, while taking sentences in Wiki80 as the test set. We make sure that there is no overlap of entity pairs among the three splits, to avoid any information leakage.

Note that Wiki20 is quite different from NYT10: NYT10 labels a sentence as "N/A" if the entity pair in the sentence does not have a relation in the KG. On the contrary, Wiki20 labels entity pairs with a relation outside its relation ontology as "N/A". In other words, "N/A" instances in Wiki20 express *unknown* relations instead of *no* relation.

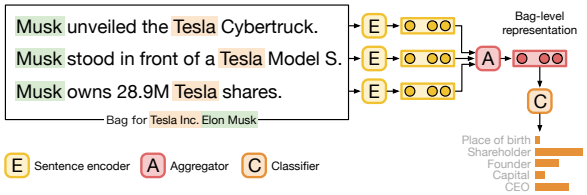

Figure 2: An illustration for a typical multi-instance multi-label (bag-level) model. The model aims to predict relation probabilities for entity pairs, instead of sentences, which is usually accomplished by aggregating a bag representation and doing classification over it.

4 DS-RE Models

In this section, we elaborate the multi-instance multi-label framework for DS-RE, and introduce models we evaluate, including both previous methods and the latest pre-trained ones.

4.1 Multi-instance Multi-label Evaluation

Unlike the supervised RE tasks which usually evaluate models at the sentence level, DS-RE evaluates *how well models can extract relational facts* from the corpus, i.e. measuring the precision and recall of extracted relational facts (a *fact* is an entity pair and a relation between them). It is named as **multi-instance multi-label**, since each entity pair might be mentioned in multiple sentences, and one entity pair can have more than one relation. Under the framework, models are required to predict the potential relations for each entity pair—according to all sentences mentioning the pair—during the evaluation, as shown in Figure 2. Sentences correlated with the same entity pair are also named as a *bag*, and thus we interchangeably refer to multi-instance multi-label framework as **bag-level** framework.

The most popular way to compare DS-RE models is to plot precision-recall (P-R) curve and calculate the area-under-the-curve (AUC). We also report micro F1 and macro F1 in our experiments. Since the numbers of instances for different relations are extremely imbalanced, macro F1 can better demonstrate model performance while avoiding the bias brought by those major relations.

4.2 Model Details

Sentence-level Training Models are trained in a sentence-level fashion (as in supervised RE), but used as a bag-level model during evaluation. As shown in Figure 2, a typical bag-level model takes

Model	AUC	Micro	Macro
PCNN+bag+AVG (full)	28.4	35.8	13.3
PCNN+bag+ATT (full)	32.2	39.1	9.5
PCNN+bag+ATT (sample)	31.8	39.6	11.8

Table 3: Comparison between full and sampled bag training with NYT10 held-out test. For PCNN+bag+ATT, the sampled bag performs similarly to the full bag, and it still outperforms PCNN+bag+AVG (full). "Micro" and "Macro" represent micro and macro F1 (%).

an *aggregator* to fuse embeddings of all sentences in the bag, and then feeds the bag-level representation to the classifier. Here we take two aggregation strategies: average (**AVG**), which averages the representations of all the sentences in the bag; and at-least-one (**ONE**) (Zeng et al., 2015), which first predicts relation scores for each sentence in the bag, and then takes the highest score for each relation.

Bag-level Training Directly deploying sentence-level training for DS-RE suffers from the wrong labeling problem: one sentence mentioning two entities may not express its auto-labeled relation. To alleviate this problem, models can also take bag-level framework in the training, based on the *expressed-at-least-one* assumption (Riedel et al., 2010): at least one sentence in the bag expresses the auto-labeled relation. Besides the AVG and ONE[2] strategies mentioned above that can be directly deployed for bag-level training, Lin et al. (2016) also propose to use sentence-level attention (**ATT**) for aggregation: It produces bag-level representation as a weighted average over embeddings of sentences, and determines weights by attention scores between sentences and relations.

For our experiments, we take CNN (Liu et al., 2013), PCNN (Zeng et al., 2015) and BERT (Devlin et al., 2019) as options of sentence encoders, which are all common choices for neural RE models. We evaluate combinations of different sentence encoders, training policies, and aggregation strategies, e.g., bag-level trained PCNN with ATT aggregator (PCNN+bag+ATT) or sentence-level trained BERT with ONE aggregator (BERT+sent+ONE). Besides, we evaluate several representative DS-RE models from literature, namely RL-DSRE (Qin et al., 2018), which takes deep reinforcement learning for denoising training instances, BGWA (Jat et al., 2017), which takes both word-level and sentence-level attention, and RESIDE (Vashishth et al., 2018), which introduces side information like relation aliases to put soft constraints on prediction.

For BERT-based sentence encoder, there are some practical challenges when adopting bag-level training: in the worst cases, one bag can contain thousands of sentences, which are beyond the capacity of most computing devices due to the large size of pre-trained models. To address this issue, we take a random sampling strategy during training: for each bag, we randomly sample b sentences, instead of taking all of them. For evaluation, we use the same routine as other non-pre-trained encoders, taking all of the sentences into account (because back propagation is not needed here so the bag can be split into several batches). Since this is different from the original bag-level training, we carry out a pilot experiment to examine the effect of the sampled training. From Table 3, we can see that our sampling strategy does not significantly hurt the performance of the bag-level training.

We also add another variant, BERT-M, in our evaluation. We observe from the top predictions of BERT models (Figure 3) that BERT tends to make false-positive errors for entity pairs that express a relation in the KG but do not have any sentence truly expressing the relation in the data, probably due to that model learns shallow cues solely from entities. Thus, following Peng et al. (2020), we mask entity mentions during training and inference to avoid learning biased heuristics from entities.

5 Experiment

5.1 Implementation Details

We use the OpenNRE toolkit (Han et al., 2019) for most of our experiments, including both sentence-level and bag-level training. For CNN and PCNN, we follow the hyper-parameters of Han et al. (2019). For BERT, we use pre-trained checkpoint `bert-base-uncased` for initialization, take a batch size of 64, a bag size of 4 and a learning rate of 2×10^{-5},[3] and train the model for 3 epochs. For RL-DSRE, RESIDE and BGWA, we directly use their original implementation.

5.2 Evaluation Settings

We take three different settings in our experiments:

[2]During training, since we have DS label r for the bag, we directly take the sentence embedding that has the highest score for r as the bag-level representation.

[3]This is determined by a grid search over batch sizes in {16, 32, 64} and learning rates in {1e-5, 2e-5, 5e-5}.

Model	Bag	Strategy	Held-out			Bag-level Manual			Sentence-level Manual		
			AUC	Micro	Macro	AUC	Micro	Macro	AUC	Micro	Macro
CNN	-	AVG	20.0	30.3	6.5	48.7	50.4	21.3	52.0	53.3	22.1
	-	ONE	21.2	31.8	7.2	50.5	51.6	19.8	52.0	53.3	22.1
PCNN	-	AVG	20.4	30.9	9.0	49.4	51.6	22.6	52.2	54.3	23.2
	-	ONE	21.4	31.9	7.8	51.1	52.6	23.8	52.2	54.3	23.2
	✓	AVG	28.4	35.8	13.3	52.9	53.6	23.5	56.0	55.9	22.9
	✓	ONE	28.4	36.0	8.0	53.4	54.8	24.5	55.5	56.7	22.2
	✓	ATT	32.2	39.1	9.5	56.8	56.5	25.5	57.1	56.1	23.6
RL-DSRE	✓	-	32.6	39.5	13.4	55.1	55.9	26.4	55.6	56.1	23.9
BGWA	✓	-	31.0	37.4	11.6	47.8	54.0	14.1	42.2	48.9	7.2
RESIDE	✓	-	33.4	40.5	16.9	35.8	43.3	10.2	43.2	47.9	19.8
BERT	-	AVG	50.5	51.2	21.6	60.3	62.4	35.3	63.2	64.3	34.1
	-	ONE	50.5	51.6	21.2	61.3	62.9	36.1	63.2	64.3	34.1
	✓	AVG	43.0	47.4	22.4	56.7	60.4	35.7	60.4	63.9	34.6
	✓	ONE	38.5	46.1	10.9	58.1	61.9	33.9	61.5	65.1	32.1
	✓	ATT	27.8	37.4	13.4	51.2	54.1	25.8	54.2	57.2	26.4

Table 4: Results (%) on NYT10, including the held-out evaluation, bag-level manual evaluation, and sentence-level manual evaluation. The "bag" column indicates whether the model uses bag-level training, and the "strategy" column shows the bag aggregation policy. We report the AUC, micro F1 (Micro) and macro F1 (Macro) scores.

Held-out evaluation: We take the test data of the original DS datasets for evaluation. The trend of this evaluation should be consistent with the reported results in most DS-RE literature.

Bag-level manual evaluation: We take our human-labeled test data for bag-level evaluation. Since annotated data are at the sentence-level, we construct bag-level annotations in the following way: For each bag, if one sentence in the bag has a human-labeled relation, this bag is labeled with this relation; if no sentence in the bag is annotated with any relation, this bag is labeled as N/A.

Sentence-level manual evaluation: As we wonder how well bag-level-trained models can handle sentence-level predictions, our human-labeled test set is also used for a sentence-level evaluation.

We report AUC, micro F1 and macro F1 for all above evaluation settings. We take the best micro F1 on the P-R curves and use the corresponding threshold for calculating macro F1. Considering the difference between NYT10 and Wiki20, we evaluate models on the two datasets respectively.

5.3 The Results on NYT10

Table 4 shows the main results on NYT10. We also plot P-R curves of selected models in Appendix B. Overall, for all three settings, training pre-trained models in a sentence-level style always perform the best, while applying bag-level training strategies can significantly boost the performance when taking other non-pre-trained encoders. Feng et al. (2018) observe that bag-level training is not helpful

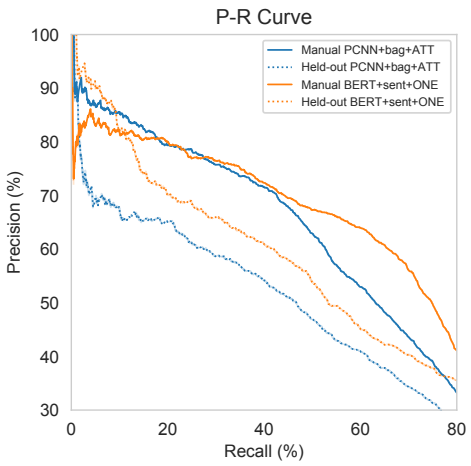

Figure 3: Comparisons of using bag-level manual test (solid lines) and corresponding held-out test (dotted lines). Both absolute and relative scores of models significantly change when taking on human-labeled data.

in the sentence-level evaluation, which contradicts our observation. We suspect that it is because Feng et al. (2018) only manually check a small proportion of test data, leading to a biased result.

More importantly, by comparing the held-out test results to the manual ones, we come to the conclusion that **manual evaluation matters**: auto-labeled and human-labeled test data lead to very different observations. For example, the comparisons between PCNN and RL-DSRE, and BGWA and RESIDE are reversed when taking different evaluations. Also, the performance gaps between different models become much smaller when it

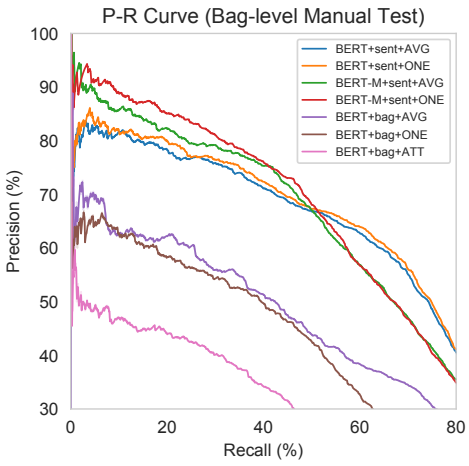

Figure 4: Comparisons of different DS-RE models with BERT on NYT10 bag-level manual test set.

comes to the manual test. Since our manual test set is smaller than the original held-out one (because we did not annotate all N/A sentences), to make a clearer comparison, we evaluate two selected models on the bag-level manual test set and on the corresponding instances in the held-out test set, respectively, and we plot the P-R curves in Figure 3. It shows that not only the absolute values of the two measurements differ a lot, but it also affects the relative performances between the models. For instance, BERT+sent+ONE shows a considerable advantage over PCNN+bag+ATT at the top predictions on the held-out test set, but it is completely the opposite case at the manual test, where BERT+sent+ONE is even significantly worse than PCNN+bag+ATT. It clearly suggests that using the held-out test set cannot well demonstrate the real pros and cons of the models.

Compared to others, BGWA and RESIDE suffer an extreme change in performance between the held-out and manual evaluations, and we suspect that it is due to the fact that they use entity types as extra information, which leads to overfitting biased heuristics of entities. This further emphasizes the need of using manually-labeled test data in DS-RE.

After checking the manual results, we further identify some interesting observations that have not been clearly demonstrated with the DS evaluation:

Pre-trained Models First of all, BERT-based models have achieved supreme performance across all three metrics. To thoroughly examine BERT and its variants in the DS-RE scenario, we further plot their P-R curves with the bag-level manual test in Figure 4. It is surprising to see that all bag-level training strategies, especially the ATT strategy which brings significant improvements for PCNN-based models, do not help or even degenerate the performance with pre-trained ones. This observation is also consistent with that in Amin et al. (2020), though they only compare BERT+bag+AVG and BERT+bag+ATT. We hypothesize the reasons are that solely using pre-trained models already makes a strong baseline, since they exploit more parameters and they have gained pre-encoded knowledge from pre-training (Petroni et al., 2019), all of which make them easier to directly capture relational patterns from noisy data; and bag-level training, which essentially increases the batch size, may raise the optimization difficulty for these large models.

Another unexpected observation is that, though the P-R curve of BERT is far above other models in the held-out test, we identify a significant drop of that in the manual test, as shown in Figure 3 and Appendix B. By manually checking those errors, we find that most of them are models predicting facts that exist in the KG but are not supported by the text (i.e., false-positive). For example, *Arthur Schnitzler* was indeed born in *Vienna*, but it is wrong for the model to infer the relation *place of birth* from sentence "*Authur Schnitzler wrote a story set in Vienna.*" We assume that it is not only because of the prior knowledge of pre-trained models, but is also due to that BERT can better learn heuristics from entity themselves, as shown in the study of Peng et al. (2020) with supervised RE. Considering the data of DS-RE are noisy and in many cases the text does not support the labeled facts, this overfitting-to-heuristic phenomenon can only be more severe.

To verify the assumption and try out a simple solution to alleviate the problem, we take a BERT-M variant (as described in §4.2) and show its results in Figure 4. We can see that the P-R curves of BERT-M are above those of BERT at the beginning, demonstrating that BERT-M models have higher precisions at those top predictions. Later on, since BERT can extract more facts than BERT-M by fully utilizing information from entity names, BERT-M reduces below BERT. From these results, we highlight that **how to handle the false-positives and denoise DS-RE data for pre-trained models** still remains an open and challenging problem.

Imbalanced Classes Previous works of DS-RE usually take AUC, micro F1, or P-R curves to measure the abilities of models, which show the over-

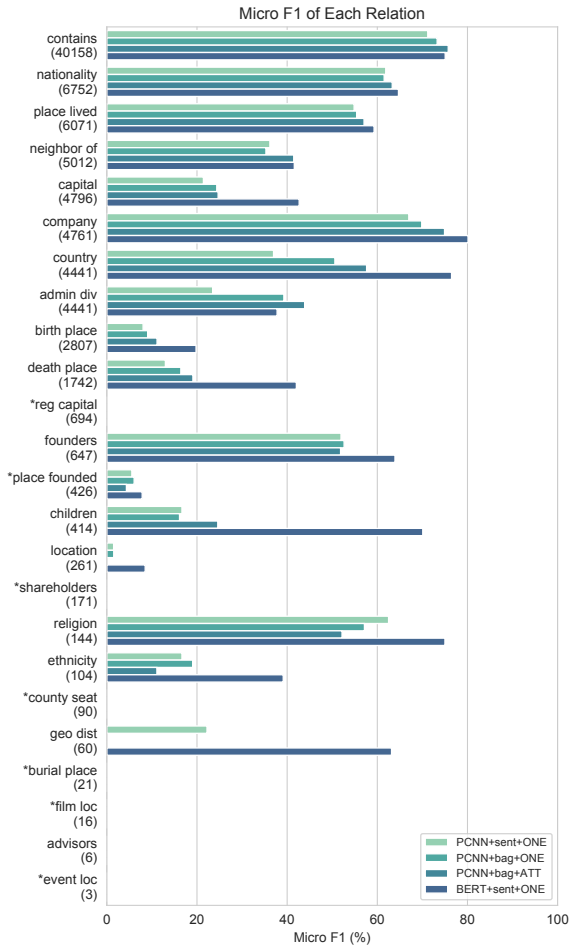

Figure 5: Micro F1 (%) of different relations on the bag-level manual test of NYT10. * represents the relation has less than 20 sentences in the test sets. Numbers in brackets are numbers of training sentences.

all performance trend averaged on relational facts. However, the distribution of training instances across relations is extremely uneven. For example, in NYT10 , almost half of the positive instances are `/location/location/contains`. On the contrary, half of the relations have fewer than $1,000$ sentences. In this case, macro F1 can better show the averaged performance across different relations, without being biased by the majority class. Table 4 demonstrates that even though in most times conclusions of different metrics are consistent, there are cases when models improve micro F1 but degenerate macro F1.

To further study how models perform on each relation, we plot several representative models and their micro F1 scores for each relation in Figure 5. We can see that: (1) The top-4 relations, which account for 80% test instances, do not vary much in performance with different models, while the difference of performance takes place mostly outside the top-4; (2) Some relations even have zero F1 scores, mostly because they have very few training or test sentences. These results further underscore the importance to look into per-relation scores for DS-RE, and we advocate that later works should **include macro F1** for more comprehensive comparisons.

Model	Bag	Strategy	AUC	Micro	Macro
PCNN	-	AVG	74.1	69.1	67.1
	-	ONE	74.0	69.1	66.9
	✓	AVG	78.1	71.8	69.5
	✓	ONE	76.6	70.3	67.7
	✓	ATT	77.5	71.2	68.6
BERT	-	AVG	90.0	83.5	82.9
	-	ONE	89.8	83.3	82.6
	✓	AVG	89.9	82.7	82.0
	✓	ONE	88.9	81.6	81.1
	✓	ATT	70.9	66.8	64.3

Table 5: Results (%) on Wiki20 of representative models. "Bag" indicates bag-level training and "Micro" and "Macro" represent micro and macro F1 respectively.

5.4 The Results on Wiki20

We choose several representative models and further evaluate them on Wiki20, as shown in Table 5. The main observation of results on Wiki20 is consistent with that of NYT10—sentence-level pre-trained models perform the best, and using bag-level training helps with non-pre-trained ones—though the overall performance is much higher. Another difference is that, in Wiki20, AVG performs better than ONE and ATT. We think that it is due to the inherent difference in how the two datasets are constructed, especially the difference in the aspect of determining N/A sentences. Compared to NYT10, part of the N/A instances in Wiki20, instead of indicating no relation between the entities, may correspond to a specific relation that is outside of the dataset ontology. It suggests that when dealing with N/A instances, **considering their latent semantics, rather than simply treating them as one abstract class**, may further benefit RE models.

6 Conclusion

In this paper, we study the problem of test protocols in DS-RE and build large manually-annotated test sets for two DS-RE datasets, to enable a more accurate and efficient evaluation. We not only demonstrate that our manual test sets show different observations from previous held-out ones, but also capture some interesting reflections by using the

manual test, e.g., pre-trained models suffer false-positives more and bag-level training strategies generally do not help with pre-trained models.

We hope that our manual test sets can mark a new starting line for DS-RE, while these observations can motivate novel research directions towards better DS-RE models, e.g., studying denoising methods for pre-trained models or processing N/A relations in a more fine-grained way.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106501) and Beijing Academy of Artificial Intelligence (BAAI). This work is also supported by the Pattern Recognition Center, WeChat AI, Tencent Inc.

Ethical Considerations

Our work mainly focuses on two parts: the construction of two manually-labeled test sets and the analyses of models based on the manual test. Regarding the annotation, we first approximate the workload by annotating a few examples on our own, and then determine the wages for annotators according to local standards. The two datasets are based on NYT and Wikipedia, and we did not identify any unethical content during annotation.

Concerning the analyses, we find that models tend to utilize some shallow clues for classification, such as learning heuristics from entities. This behavior can potentially create biased extraction results based on the distributions of entities in the training set and is worth further investigating.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of ACL*, pages 1388–1398.

Saadullah Amin, Katherine Ann Dunfield, Anna Vechkaeva, and Guenter Neumann. 2020. A data-driven approach for noise reduction in distantly supervised biomedical relation extraction. In *Proceedings of SIGBioMed*, pages 187–194.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of ACL*, pages 2895–2905.

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 SIGMOD*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Mary Elaine Califf and Raymond J. Mooney. 1997. Relational learning of pattern-match rules for information extraction. In *Proceedings of CoNLL*, pages 9–15.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of ACL*, pages 1870–1879.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of AAAI*, pages 5779–5786.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of AACL-IJCNLP*, pages 745–758.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of EMNLP-IJCNLP*, pages 169–174.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018a. Denoising distant supervision for relation extraction via instance-level adversarial training. *arXiv preprint arXiv:1805.10959*.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018b. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of AAAI*, pages 4832–4839.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018c. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of EMNLP*, pages 2236–2245.

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 ACL*, pages 541–550.

Linmei Hu, Luhao Zhang, Chuan Shi, Liqiang Nie, Weili Guan, and Cheng Yang. 2019. Improving distantly-supervised relation extraction with joint label embedding. In *Proceedings of EMNLP-IJCNLP*, pages 3812–3820.

Scott B Huffman. 1995. Learning information extraction patterns from examples. In *Proceedings of IJCAI*, pages 246–260.

Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2017. Improving distantly supervised relation extraction using word and entity based attention. In *Proceedings of the 6th Workshop on AKBC*.

Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of AAAI*, pages 3060–3066.

Tingsong Jiang, Jing Liu, Chin-Yew Lin, and Zhifang Sui. 2018. Revisiting distant supervision for relation extraction. In *Proceedings of LREC*, pages 3580–3585.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of ACL*, pages 178–181.

Pengshuai Li, Xinsong Zhang, Weijia Jia, and Wei Zhao. 2020. Active testing: An unbiased evaluation method for distantly supervised relation extraction. In *Findings of EMNLP 2020*, pages 204–211.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of ACL*, pages 34–43.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, pages 2124–2133.

Chunyang Liu, Wenbo Sun, Wenhan Chao, and Wanxiang Che. 2013. Convolution neural network for relation extraction. In *Proceedings of ICDM*, pages 231–242.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP*, pages 1003–1011.

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 EMNLP*, pages 3661–3672.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *TACL*, 5:101–115.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*, pages 2463–2473.

Pengda Qin, XU Weiran, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of ACL*, pages 2137–2147.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of EACL*, pages 1171–1182.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML-PKDD*, pages 148–163.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of NAACL*, pages 74–84.

Dan Roth and Wen-tau Yih. 2002. Probabilistic reasoning for entity & relation recognition. In *Proceedings of COLING*.

Evan Sandhaus. 2008. The new york times annotated corpus ldc2008t19. In *Philadelphia: Linguistic Data Consortium*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP*, pages 455–465.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of EMNLP*, pages 1257–1266.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of NAACL*, pages 886–896.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Proceedings of CACM*, pages 78–85.

Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018. Adversarial multi-lingual neural relation extraction. In *Proceedings of COLING*, pages 1156–1166.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of EMNLP*, pages 1778–1783.

Ya Xiao, Chengxiang Tan, Zhijie Fan, Qian Xu, and Wenye Zhu. 2020. Joint entity and relation extraction with a hybrid transformer and reinforcement learning based model. In *Proceedings of AAAI*, pages 9314–9321.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, pages 2335–2344.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019a. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In *Proceedings of NAACL-HLT*, pages 3016–3025.

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 EMNLP*, pages 35–45.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019b. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL*, pages 1441–1451.

Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of ACL*, pages 427–434.

Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. Graph neural networks with generated parameters for relation extraction. In *Proceedings of ACL*, pages 1331–1339.

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 *Proceedings of COLING*, pages 6436–6447.

A Relation Ontology Changes of NYT10

/location/country/administrative_divisions
/location/administrative_division/country
/location/country/capital
(merge) /location/region/capital
/location/fr_region/capital
/location/cn_province/capital
/location/in_state/administrative_capital
/location/in_state/legislative_capital
/location/in_state/judicial_capital
/location/it_region/capital
/location/br_state/capital
/location/mx_state/capital
/location/province/capital
/location/us_state/capital
/location/jp_prefecture/capital
/location/de_state/capital
/location/us_county/county_seat
/location/neighborhood/neighborhood_of
/location/location/contains
(merge) /business/location
/business/company/locations
/sports/sports_team/location
/broadcast/producer/location
/business/company/founders
/business/company/place_founded
/business/company/major_shareholders
/business/company/advisors
/business/person/company
/people/person/place_of_birth
/people/person/religion
/people/person/nationality
/people/person/place_lived
/people/person/ethnicity
/people/person/children
/people/deceased_person/place_of_death
/people/deceased_person/place_of_burial
/people/ethnicity/geographic_distribution
/time/event/locations
/film/film/featured_film_locations
N/A
(delete) location/country/languages_spoken
(delete) base/locations/countries/states_provinces_within
(delete) business/shopping_center_owner/shopping_centers_owned
(delete) business/shopping_center/owner
(delete) business/business_location/parent_company
(delete) business/company_advisor/companies_advised
(delete) people/profession/people_with_this_profession
(delete) people/person/profession
(delete) people/place_of_interment/interred_here
(delete) people/ethnicity/included_in_group
(delete) people/family/members
(delete) people/family/country
(delete) broadcast/content/location
(delete) film/film_festival/location
(delete) film/film_location/featured_in_films

Table A.1: NYT10 relation ontology. All those relations are from FreeBase. "(merge)" means a merge relation in our version of NYT10 and it is followed by relations merged from the original dataset. "(delete)" indicates that this relation is discarded in our version because there are no instances in the training or the test sets.

B P-R Curves for NYT10

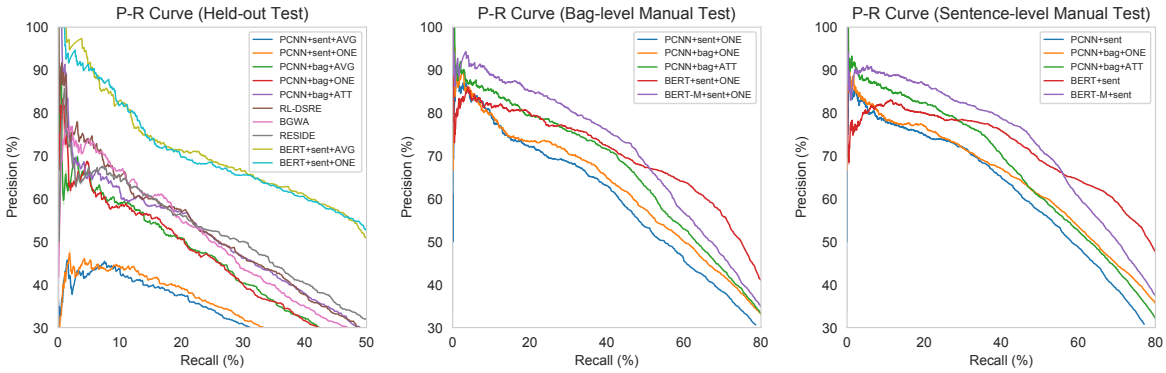

Figure B.1: P-R curves of representative models in held-out test, bag-level manual test and sentence-level manual test of NYT10. Note that the scales of X-axis are not the same in the three figures.