MEDDISTANT19: A Challenging Benchmark for Distantly Supervised Biomedical Relation Extraction

Anonymous ACL submission

Abstract

Relation Extraction in the biomedical domain is a challenging task due to the lack of labeled data and the long-tail distribution of the entity mentions. Recent works propose distant supervision as a way to tackle the scarcity of annotated data by automatically pairing knowledge graph relationships with raw textual data. In several benchmarks, Distantly Supervised Biomedical Relation Extraction (Bio-DSRE) models can produce very accurate results. However, given the challenging nature of the task, we set out to investigate the validity of such impressive results. We probed the datasets used by Amin et al. (2020) and Hogan et al. (2021) and found a significant overlap between training and evaluation relationships that, once resolved, reduced the accuracy of the models by up to 71%. Furthermore, we noticed several inconsistencies along the data construction process, such as the creation of negative samples and improper handling of redundant relationships. To mitigate these issues we present MEDDISTANT19, a new benchmark dataset obtained by aligning the MEDLINE abstracts with the widely used SNOMED-Clinical Terms (SNOMED-CT) knowledge base. We experimented with several state-of-the-art models following our methodology, showing that there is still plenty of room for improvement for the task. We release our code and data for reproducibility.

1 Introduction

Extracting structured knowledge from unstructured text is an important task for knowledge discovery and management. Biomedical literature and clinical narratives offer rich interactions between entities mentioned in the text (Craven et al., 1999; Xu and Wang, 2014), which can be useful for applications such as bio-molecular information extraction, pharmacogenomics, and identifying drug-drug interactions (DDIs), among others (Luo et al., 2017).

Model and Data	Original		Filtered	
	AUC	F1	AUC	F1
Amin et al. (2020)	68.4	64.9	50.8	53.1
Hogan et al. (2021)	82.6	77.6	11.8	19.8

Table 1: Two state-of-the-art Bio-DSRE models evaluated on the respective datasets before (Original) and after (Filtered) removing test relationships also appearing in the training set. Both models were trained and evaluated at *bag-level*.

Manually annotating these relations for training supervised learning systems is an expensive and time-consuming process (Segura-Bedmar et al., 2011; Kilicoglu et al., 2011; Segura-Bedmar et al., 2013; Li et al., 2016), so the task often involves leveraging rule-based (Abacha and Zweigenbaum, 2011; Kilicoglu et al., 2020) and weakly supervised approaches (Peng et al., 2016; Dai et al., 2019).

More recently, Amin et al. (2020) and Hogan et al. (2021) used domain-specific language models (Gu et al., 2021) that were pre-trained explicitly on biomedical data for Bio-DSRE, producing disproportionately more accurate results when compared with recent results in the general domain (Gao et al., 2021; Christopoulou et al., 2021; Zhang et al., 2021).

In this work, we highlight that these results can be largely attributed to the overlap between the training and the test facts, which allows the model to score higher just by memorizing the training relations rather than generalizing to new, previously unknown ones. In Table 1, we show that removing this leakage (51.9%, Table 2) results in a significant decrease in predictive accuracy. For example, AMIL model with relation type embedding L proposed by Hogan et al. (2021) achieves an 82.6 AUC when evaluated on their Bio-DSRE dataset, while producing an 11.8 AUC when evaluated on the subset of the test set of relationships that do not overlap with the training set.

Triples	Train	Valid	Test
Textual	92,972	13,555 (51.9%)	33,888 (51.2%)
CUI	211,789	41,993 (26.7%)	89,486 (26.5%)

Table 2: Training-test leakage we identified in the data constructed and used by Amin et al. (2020) (see their Appendix A.4 in their *k-tag* setup). Numbers between parentheses show the percentage overlap, where the authors considered text-based instead of CUI-based triples.

The training-test overlap in the datasets proposed by Hogan et al. (2021) and Amin et al. (2020) is due to the same entities appearing with different names in multiple relationships, although two entity names are mapped to the same UMLS concept (Bodenreider, 2004), they are still treated as two distinct entities. Furthermore, we also identified other problems, such as redundant facts, and unclear coverage of UMLS concepts. To mitigate these issues, we follow the guidelines outlined by Chang et al. (2020) for benchmarking biomedical link prediction models, and propose a new Bio-DSRE benchmark.

2 Related Work

Relation Extraction (RE) is an important task in biomedical applications. Traditionally, supervised methods require large-scale annotated corpora, which is impractical to scale for broad-coverage biomedical relation extraction (Kilicoglu et al., 2011, 2020). In cases where such supervision is available, it is limited to protein-protein interactions (Peng and Lu, 2017), drug-drug interactions (Kavuluru et al., 2017), and chemical-disease interactions (Peng et al., 2016).

Distant Supervision (DS) allows for the automated collection of noisy training examples (Mintz et al., 2009) by aligning a given knowledge base (KB) with a collection of text sources. DS was used in recent works (Alt et al., 2019) using Multi-Instance Learning (MIL) by creating *bags* of instances (Riedel et al., 2010) for corpus-level triples extraction.[1]

Dai et al. (2019) introduced the use of the Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004) as a KB with PubMed (Canese and Weis, 2013) MEDLINE abstracts as text collection, and implemented a knowledge-based attention mechanism (Han et al., 2018) for joint learning with knowledge graph completion using SimplE (Kazemi and Poole, 2018) embeddings and entity type classification. Their pipeline was simplified by Amin et al. (2020), who extended R-BERT (Wu and He, 2019) to handle bag-level MIL, and demonstrated that preserving the direction of the relationships improves the accuracy of the model. Lacking benchmark corpora, Amin et al. (2020) also outlined the steps to create the dataset. Similar steps were followed by Hogan et al. (2021), who introduced the concept of *abstractified* MIL (AMIL), by absorbing different argument pairs belonging to the same semantic types (see Fig. 2) pair in one bag, boosting performance on rare-triples. They also proposed the use of SCISPACY (Neumann et al., 2019) for sentence tokenization, resulting in improved overall performance.

In this work, we investigate some recent results from the Bio-DSRE literature by probing the respective benchmarks for overlaps between training and test sets. We found a severe overlap between the training set and the held-out validation and test sets in the dataset constructed by Amin et al. (2020) and Hogan et al. (2021). An issue of entity linking or concept normalization. Where in UMLS, each concept is mapped to a UMLS *Concept Unique Identifier (CUI)*, where a given CUI might have different surface forms (Bodenreider, 2004). Table 2 shows the leakage statistics.

Consider a relationship between a pair of UMLS entities (C0013798, C0429028). These two entities can appear in different forms within a text,

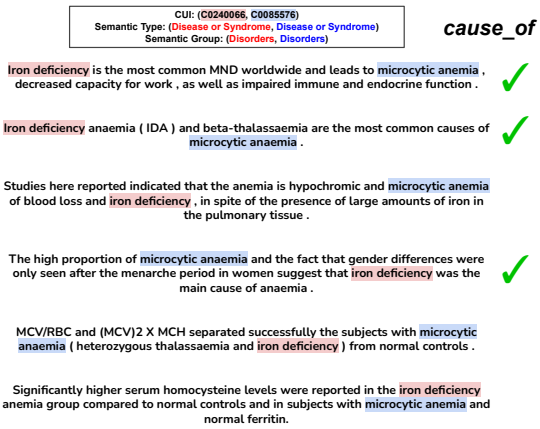

Figure 1: An example of a bag instance representing the UMLS concept pair (C0240066, C0085576) from the MEDDISTANT19 dataset, expressing the relation *cause_of*. In this example, three out of six sentences express the relation, while others are wrong labels from distant supervision.

[1] RE is used to refer to two different tasks: sentence-level detection of relational instances and corpus-level triples extraction, a kind of knowledge completion task (Ji et al., 2021).

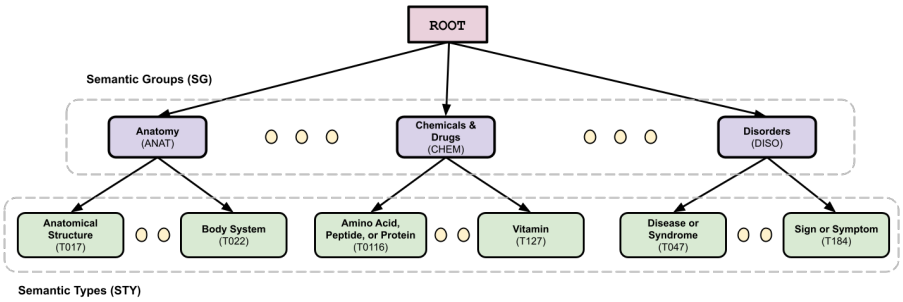

Figure 2: Type Hierarchy: each concept in the UMLS is classified under a type taxonomy. The *coarse-grained* entity type is called **Semantic Group (SG)** and the *fine-grained* entity type is called **Semantic Type (STY)**.

such as (*electrocardiography*, *Q-T interval*), (*ECG*, *Q-T interval*), and (*EKG*, *Q-T interval*); each of these distinct pairs still refers to the same original pair (C0013798, C0429028). Amin et al. (2020) claim no such text-based leakage, but when normalized this results in leakage across the splits as reported in Table 2.

Due to these inconsistencies and lack of benchmark and best practices, we introduce the MEDDISTANT19 dataset. Our work utilizes the SNOMED-CT Knowledge Graph (KG) extracted from the UMLS that offers a careful selection of the concept types, proper handling of the inverse relations, and highlights the need for downstream benchmarks (Chang et al., 2020). The dataset is particularly focused on rare-triples and considers a narrower subset of the relations.

3 Constructing the MedDistant19 Dataset

Documents We used PubMed MEDLINE abstracts from 2019[2] as our text source, containing 32,151,899 abstracts. Following Hogan et al. (2021), we used SCISPACY (Neumann et al., 2019) for sentence tokenization, resulting in 150,173,169 unique sentences.

Previous studies used Exact Match for entity linking (Amin et al., 2020; Hogan et al., 2021). In this work, we further introduce the use of a specialized UMLS entity linker from SCISPACY [3], since named entity recognition and normalization was shown to be the largest source of errors in biomedical RE (Kilicoglu et al., 2020). We used the default settings in SCISPACY for linking entity mentions to their UMLS CUIs, and filtering disabled concepts from UMLS. This resulted in the entity linked mentionsat the sentence level.

Knowledge Base We use UMLS2019AB [4] as our main knowledge source. The UMLS Metathesaurus (Bodenreider, 2004) covers concepts from 222 source ontologies, thus being the largest ontology of biomedical concepts. However, covering all ontologies can be challenging given the interchangeable nature of the concepts. For example, *programmed cell death 1 ligand 1* is an alias of concept C1540292 in the HUGO Gene Nomenclature Committee ontology (Povey et al., 2001), and it is an alias of concept C3272500 in the National Cancer Institute Thesaurus. This makes entity linking more challenging, since a surface form can be linked to multiple entity identifiers, and makes it easier to have overlaps between training and test set, since the same fact may appear in both with different entity identifiers.

Furthermore, benchmark corpora for biomedical Named Entity Recognition (Doğan et al., 2014; Li et al., 2016) and RE (Herrero-Zazo et al., 2013; Krallinger et al., 2017) focuses on specific entity types (e.g. diseases, chemicals, proteins), and are usually normalized to a single ontology (Kilicoglu et al., 2020). Following this trend, we also focus on a single vocabulary for Bio-DSRE. We use SNOMED-CT, which is the most widely used clinical terminology in the world for documentation and reporting in healthcare (Chang et al., 2020).

UMLS classifies each entity in a type taxonomy as shown in Fig. 2. This allows for narrowing the concepts of interest. Following (Chang et al., 2020), we consider 8 semantic groups in SNOMED-CT: Anatomy (ANAT), Chemicals & Drugs (CHEM), Concepts & Ideas (CONC), Devices (DEVI), Disorders (DISO), Phenomena (PHEN), Physiology (PHYS), and Procedures (PROC). For a complete list of semantic types covered in MEDDISTANT19,

[2] https://lhncbc.nlm.nih.gov/ii/information/MBR/Baselines/2019.html

[3] https://github.com/allenai/scispacy

[4] https://download.nlm.nih.gov/umls/kss/2019AB/umls-2019AB-full.zip

Properties	Prior	MD19
approximate entity linking		✓
unique NA sentences		✓
inductive		✓
triples leakage	✓	
NA-type constraint		✓
NA-argument role constraint		✓

Table 3: MEDDISTANT19 (MD19) properties in comparison with the prior works (Amin et al., 2020; Hogan et al., 2021).

Facts	Training	Validation	Testing
Inductive (I)	345,374	62,116	130,563
Transductive (T)	402,522	41,491	84,414

Table 4: Number of raw inductive and transductive SNOMED-KG triples used for alignment with text data.

see Appendix A.1. Similarly, each relation is categorized into a type and has a reciprocal relation in UMLS (Appendix A.2), which can result in an overlap between the training and test set if not addressed (Dettmers et al., 2018).

These steps follow Chang et al. (2020), with the difference that we only consider relations of type *has relationship other than synonymous, narrower, or broader* (RO); this is consistent with prior works in Bio-DSRE (Dai et al., 2019; Amin et al., 2020; Hogan et al., 2021). We also exclude uninformative relations, *same_as*, *possibly_equivalent_to*, *associated_with*, *temporally_related_to*, and ignore inverse relations as generally is the case in RE.

In addition, Chang et al. (2020) ensure that the validation and test set do not contain any new entities, making it a transductive learning setting where we assume all test entities are known beforehand. However, in real-world applications of biomedical RE, we are expected to extract relations between unseen entities. To support this setup, we also consider an inductive KG split proposed by Daza et al. (2021).

Table A.3 summarizes the statistics of the KGs used for alignment with the text. We use split ratios of 70%, 10%, and 20%. Relationships are defined between CUIs, and have no overlap between training, validation, and test set.

3.1 Knowledge-to-Text Alignment

We now describe the procedure for searching fact triples to match relational instances in text.

Let $\mathcal{E}$ and $\mathcal{R}$ respectively denote the set of

Summary	Entities	Relations	STY	SG
	25,028	39	65	8

Split	Instances	Facts	Rare (%)	Bags	NA (%)
Train	251,558	2,366	92.3%	80,668	96.9%
Valid	179,393	806	87.8%	31,805	98.2%
Test	213,602	1,138	91.3%	50,375	98.1%

Table 5: Summary statistics of the MEDDISTANT19 dataset using Inductive SNOMED-KG split (Table A.3). The number of relations include the unknown relation type (NA). Rare represents the proportion of the fact triples which have 8 or fewer instances in a given split as defined by Hogan et al. (2021). MEDDISTANT19 focuses on rare triples with high NA proportions, making it a challenging benchmark.

UMLS CUIs and relation types, and let $\mathcal{G} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ denote the set of relationships contained in UMLS. For producing a training-test split, we first create a set $\mathcal{G}^+ \subseteq \mathcal{E} \times \mathcal{E}$ of related entity pairs, as follows:

$$\mathcal{G}^+ = \{(e_i, e_j) \mid \langle e_i, p, e_j \rangle \in \mathcal{G} \vee \langle e_j, p, e_i \rangle \in \mathcal{G}\}.$$

Following the Local-Closed World Assumption (LCWA, Dong et al., 2014; Nickel et al., 2016), we obtain a set of unrelated entity pairs by corrupting one of the entities in each pair in $\mathcal{G}^+$ and making sure it does not appear in $\mathcal{G}^+$, obtaining a new set $\mathcal{G}^- \subseteq \mathcal{E} \times \mathcal{E}$ of unrelated entities:

$$\mathcal{G}^- = \{(\overline{e_i}, e_j) \mid (e_i, e_j) \in \mathcal{G}^+ \wedge (\overline{e_i}, e_j) \notin \mathcal{G}^+\} \cup \{(e_i, \overline{e_j}) \mid (e_i, e_j) \in \mathcal{G}^+ \wedge (e_i, \overline{e_j}) \notin \mathcal{G}^+\}.$$

During the corruption process, we enforce two constraints 1) the two entities appearing in each negative pair in $\mathcal{G}^-$ should belong to the same entity types as the entities in the initial positive pair, and 2) that the entities used in the negative pair must have appeared in one or more positive pairs.

For each entity linked sentence, we only consider those sentences that have SNOMED-CT entities and have pairs in $\mathcal{G}^+$ and $\mathcal{G}^-$. Selected positive and negative pairs are mutually exclusive and have no overlap across splits.

Since we only consider unique sentences associated with a pair, this makes for unique negative training instances, in contrast to Amin et al. (2020) who considered generating positive and negative pairs from the same sentence. We define negative examples as relational sentences mentioning argument pairs with *unknown relation type* (NA), i.e. there might be a relation but the considered set of relations do not cover it. Our design choices are summarized in Table 3.

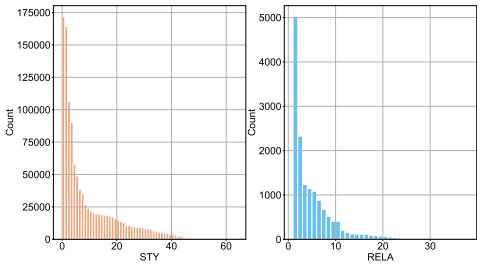

Figure 3: (*Left*) Entity distribution based on Semantic Types. (*Right*) Relations distribution.

We prune high-frequency positive and negative pairs, remove mention-level overlap across the splits and apply type-based mention pruning. Specifically, we pool mentions by types and remove the sentences which have the mention appearing more than 1000 times. This step was crucial in removing highly non-informative mentions, such as *increased* (STY: *Qualitative Concept*), mentioned over 449951 times compared to *malignant tumor* (STY: *Neoplastic Process*) mentioned 473 times. Table 5 shows the final summary of MEDDISTANT19 using inductive split. Fig. 3 shows entity and relation plots, following a long-tail.

4 Experiments

MEDDISTANT19 is released in a format that is compatible with the widely adopted RE framework OpenNRE (Han et al., 2019). To report our results, we use the *corpus-level RE metrics* Area Under the Precision-Recall (PR) curve (AUC), Micro-F1, Macro-F1, and Precision-at-k (P@k) with $k \in \{100, 200, 300, 1000, 2000\}$, and the *sentence-level RE metrics* Precision, Recall, and F1. Due to imbalanced nature of relational instances (Fig. 3), following Gao et al. (2021), we report Macro-F1 values, and following Hogan et al. (2021) we report sentence-level RE results on relationships including frequent and rare triples.

4.1 Baselines

Our baseline experiments largely follow the setup of Gao et al. (2021). For sentence encoding, we use CNN (Liu et al., 2013), PCNN (Zeng et al., 2015), and BERT (Devlin et al., 2019). We used GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013)[5] for CNN/PCNN models, and initialized BERT with BioBERT (Lee et al., 2020).

[5]DRE baselines using CNN/PCNN models use 50-dimensional word embeddings from GloVe. Therefore, we trained 50-dim Word2Vec embeddings on PubMed abstracts.

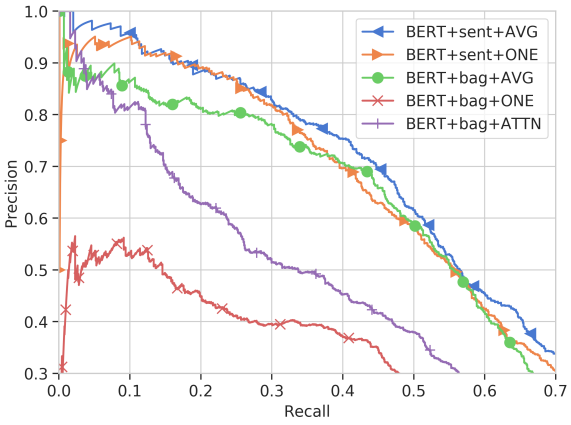

Figure 4: Precision-Recall (PR) curves for BERT-based baselines initialized with BioBERT on MEDDISTANT19. The trends largely follow the general-domain (Gao et al., 2021) with exception of BERT+bag+ONE.

We trained our models both at *sentence-level* and at *bag-level*. In contrast, prior works only considered bag-level training for Bio-DSRE (Dai et al., 2019; Amin et al., 2020; Hogan et al., 2021). The sentence-level setup is similar to standard RE (Wu and He, 2019), with the difference that the evaluation is conducted at bag-level. We also consider different pooling strategies, namely average (AVG), which averages the representations of sentences in a bag, at-least-one (ONE, Zeng et al., 2015), which generates relation scores for each sentence in a bag and then selects the top scoring sentence, and attention (ATT), which learns an attention mechanism over the sentences within a bag.

Table 6 presents our main results. In all the cases, BERT sentence encoder performed better than CNN and PCNN. This trend is similar to the general-domain. We also validate the finding that sentence-level training in pre-trained language models (LMs) performs better than the bag-level (Gao et al., 2021; Zhang et al., 2021). We argue that when trained at sentence-level, those sentences that have been correctly labeled by distant supervision (e.g. Fig. 1) provides enough learning signal, given the generalization abilities of LMs. However, in bag-level training, we force the model to jointly learn from clean and noisy samples, thus limiting its overall performance. This raises further questions into using MIL with LMs. But, we do not find this trend to hold for CNN/PCNN, instead the bag-level models performed slightly better. We also find GloVe to be a better initialization for sentence-level training and Word2Vec for bag-level. We further plot PR curves for BERT-based baselines in Fig. 4.

Model	Bag	Strategy	AUC	F1-micro	F1-macro	P@100	P@200	P@300	P@1k	P@2k
CNN	-	AVG	5.8	10.5	3.2	32.0	28.0	23.3	15.6	10.8
	-	ONE	6.4	10.7	2.8	33.0	26.5	23.6	15.5	11.2
	✓	AVG	10.3	12.8	4.7	48.0	37.0	32.0	19.5	13.6
	✓	ONE	8.5	17.9	3.7	39.0	32.5	27.3	18.7	13.3
	✓	ATT	6.0	13.5	2.6	31.0	28.5	22.3	15.7	10.7
PCNN	-	AVG	6.3	12.8	4.8	37.0	30.0	26.6	16.8	10.6
	-	ONE	6.6	9.7	2.9	34.0	26.0	22.3	16.3	11.7
	✓	AVG	9.5	15.2	5.5	48.0	36.0	31.3	19.1	13.8
	✓	ONE	6.8	15.3	2.6	34.0	27.0	26.0	16.3	12.3
	✓	ATT	5.7	13.7	2.4	36.0	24.0	23.6	14.9	10.8
BERT	-	AVG	**55.4**	**55.1**	23.3	**97.0**	90.0	87.3	**58.8**	**37.8**
	-	ONE	53.0	52.1	**23.6**	94.0	**92.0**	**87.6**	57.5	36.4
	✓	AVG	49.8	53.5	20.3	89.0	82.0	80.3	58.1	36.1
	✓	ONE	25.2	27.8	12.3	52.0	53.5	50.6	39.3	28.0
	✓	ATT	36.9	40.3	12.7	84.0	73.5	66.0	45.3	31.4
JointSimplE_NER+KATT (Dai et al., 2019)			-	-	-	-	-	-	-	91.3
BERT+bag+AVG (Amin et al., 2020)			68.4	64.9	-	97.4	98.3	98.6	-	98.3
AMIL (Rel. Type L) (Hogan et al., 2021)			87.2	81.2	-	-	-	-	-	100.0
AMIL (Rel. Type L)*			82.6	77.6	-	100.0	100.0	100.0	-	99.7

Table 6: Baselines adopted from Gao et al. (2021) for MEDDISTANT19. CNN and PCNN models at sentence-level are reported with GloVe, while bag-level models are reported with Word2Vec. BERT-based models are initialized with BioBERT. We also include previously published results for completeness. The results are not directly comparable due to differences in the corpora used. All the previously published results were trained at bag-level. The symbol * marks our re-run of the best model reported by Hogan et al. (2021).

In all cases, AVG proved to be a better pooling strategy; this finding is consistent with prior works. Both Amin et al. (2020) and Gao et al. (2021) found ATT to produce less accurate results with LMs, however, contrary to general-domain, in MEDDISTANT19, BERT+bag+ONE had lower performance than BERT+bag+ATT. We attribute this to the challenging nature of the benchmark, since it is focused on long-tail relations and therefore, the signal to learn from is insufficient when picking the optimal example in the bag for BERT+bag+ONE. This results in sparse gradients and longer training time.[6]

The current state-of-the-art model AMIL (Rel. Type L) from Hogan et al. (2021) creates bags of instances by abstracting entity pairs belonging to the same semantic type pair into a single bag, thus producing heterogeneous bags. Due to the nature of their methodology, it is not suited for sentence-level models, which already produce more accurate results than bag models.

To further study the impact of bag-level and sentence-level training on MEDDISTANT19, we analyse the relation category-specific results as in Chang et al. (2020), and the results on rare and frequent triples as in Hogan et al. (2021). Following Chang et al. (2020), we grouped the relations based on cardinality, where the cardinality is defined as: for a given relation type, if the set of head or tail entities belong to only one semantic group, then it has cardinality 1 otherwise M (many). The results are shown in Table 8 for sentence- and bag-level training with average pooling. We note that both training strategies perform comparably on 1-1 category but the bag-level training suffers a huge performance drop in M-1 and 1-M settings. We reason that this could be due to the lack of enough training signal to differentiate between heterogeneous entity types pooled over instances in a bag.

Following Hogan et al. (2021), we also perform sentence-level evaluation of BERT-based encoders trained at sentence-level and bag-level. The authors divided the triples (including "NA" instances) into two categories, those with 8 or more sentences are defined as common triples and others as rare triples. Table 7 shows these results. We note that both training strategies performed comparably on rare-triples with BERT+sent+AVG more precise than BERT+bag+AVG. However, we find noticeable differences on common triples where BERT+bag+AVG had higher recall but still low precision. This could be explained because of over-

[6]While we trained all BERT-based models for 3 epochs, BERT+bag+ONE was trained for 50 epochs

Model	P	R	F1
All Triples			
BERT+sent+AVG	**0.44**	0.49	**0.46**
BERT+bag+AVG	0.36	**0.52**	0.42
Common Triples			
BERT+sent+AVG	**0.35**	0.47	**0.40**
BERT+bag+AVG	0.28	**0.53**	0.37
Rare Triples			
BERT+sent+AVG	**0.57**	**0.52**	**0.55**
BERT+bag+AVG	0.52	0.50	0.51

Table 7: Sentence-level RE metrics comparing BERT baselines trained at bag and sentence-level with AVG pooling on Rare, Common and All triples. The triples also include NA relational instances.

Model	1-1	1-M	M-1
BERT+sent+AVG	**21.3**	**26.1**	**30.7**
BERT+bag+AVG	19.4	9.4	3.0

Table 8: Averaged F1-micro score on relation specific category. The categories are defined using the *cardinality* of head and tail semantic group types.

fitting to type and mention heuristics at bag-level, where sentence-level training allows to have more focus on context.

4.2 Analysis

Context, Mention, or Type? RE models are known to heavily rely on information from entity mentions, most of which is type information, and existing datasets may leak shallow heuristics via entity mentions that can inflate the prediction results (Peng et al., 2020). To study the importance of mentions, contexts, and entity types in MEDDISTANT19, we take inspiration from Peng et al. (2020); Han et al. (2020) and conduct an ablation of different text encoding methods. We consider entity mentions with special entity markers (Wu and He, 2019; Amin et al., 2020) as the *Context + Mention* (CM) setting, which is common in RE with LMs. We then remove the context and only use mentions, and we refer to this as the *Only Mention* (OM) setting. This is similar to KG-BERT (Yao et al., 2019) for relation prediction. We then only consider the context by replacing subject and object entities with special tokens, resulting in the *Only Context* (OC) setting. Lastly, we consider two type-based (STY) variations as *Only Type* (OT) and *Context + Type* (CT). We conduct these experiments with BioBERT trained at sentence-level and evaluated at bag-level. The results are shown in Fig. 5.

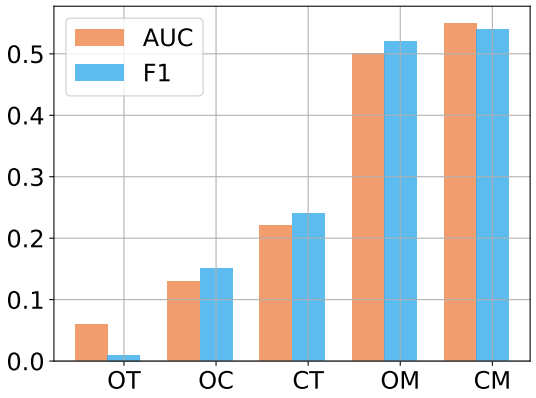

Figure 5: Ablation showing the effect of different text encoding methods following the general-domain trends.

We observe that the CM method had the highest performance but surprisingly, OM performed quite well. This highlights the ability of LMs to memorize the facts and act as soft KBs (Petroni et al., 2019; Safavi and Koutra, 2021). This trend is also consistent with general-domain (Peng et al., 2020). The poor performance in the OC setting shows that the model struggles to understand the context, which is more pronounced in noisy-prone distant RE compared to supervised RE. Our CT setup can be seen as sentence-level extrapolation of the AMIL model (Hogan et al., 2021), which struggles to perform better than the baseline (OM). However, comparing OC with CT, it is clear that the model benefits from type information as it can help constraint the relations space. Using only the type information had the least performance as the model fails to disambiguate between different entities belonging to the same type.

Inductive or Transductive? To study the impact of *transductive* and *inductive* splits (Table A.3), we created another Bio-DSRE corpus using transductive train, validation, and test triples. The corpus generated is different than the inductive one, but it can offer insights into the model's ability to handle unseen mentions. As shown in Table 9, the performance using transductive is slightly better than inductive for corpus-level extractions, in terms of AUC, however, the F1-micro score is slightly better in inductive than transductive. We conclude from this that the model is able to learn patterns that exploit mention and type information to extrapolate to unseen mentions.

Split	AUC	F1-micro	F1-macro
Inductive (I)	55.5	**56.5**	**24.8**
Transductive (T)	**57.4**	53.0	24.1

Table 9: BERT+sent+AVG performance on two corpora, one created with inductive set of triples and the other with transductive set of triples.

Encoder	Knowledge					AUC
	Biomedical	*Clinical*	*Type*	*Triples*	*Synonyms*	
BERT						0.42
ClinicalBERT	✓	✓				0.47
BlueBERT	✓					0.55
SciBERT	✓					0.55
BioBERT	✓					0.55
PubMedBERT	✓					**0.62**
MedType	✓		✓			0.54
KeBioLM	✓			✓		**0.61**
UmlsBERT	✓	✓	✓			0.53
SapBERT	✓				✓	0.57

Table 10: Fine-tuning different biomedical and clinical domain LMs on MEDDISTANT19.

Does Expert Knowledge Help? We now consider several pre-trained LMs with different knowledge capacities, specific to biomedical and clinical language understanding, with the aim to better understand MEDDISTANT19 challenges and gain insights into models behavior.

We consider BERT (Devlin et al., 2019) as a baseline model. Next, we consider domain-specific models: ClinicalBERT (Alsentzer et al., 2019) which is pre-trained on the clinical notes (Johnson et al., 2016), BlueBERT (Peng et al., 2019) and BioBERT (Lee et al., 2020) which are pre-trained on PubMed, and SciBERT (Beltagy et al., 2019), which is pre-trained on PubMed and Computer Science papers. The recently introduced PubMedBERT (Gu et al., 2021) is trained on PubMed from *scratch*, showing state-of-the-art performance on several biomedical tasks. We categorize these models as non-expert since they are only trained with Masked Language Modeling (MLM) objective.

In the second category, we consider expert models which either modify the MLM objective or introduce new pre-training tasks using external knowledge, such as UMLS. MedType (Vashishth et al., 2021), initialized with BioBERT, is pre-trained to predict semantic types. KeBioLM (Yuan et al., 2021), initialized with PubMedBERT, uses relational knowledge by initializing the entity embeddings with TransE (Bordes et al., 2013), improving downstream entity-centric tasks, including RE. UmlsBERT (Michalopoulos et al., 2021), initialized with ClinicalBERT, modifies MLM to mask words belonging to the same CUI and further introduces semantic type embeddings. SapBERT (Liu et al., 2021), initialized with PubMedBERT, introduces a metric learning task for clustering synonyms together in an embedding space.

Table 10 shows the results of these sentence encoders fine-tuned on the MEDDISTANT19 dataset at sentence-level with AVG pooling. Without any domain-specific knowledge, BERT performs slightly worse than the lowest-performing biomedical model, highlighting the presence of shallow heuristics in the data that are common to the general and biomedical domains. While domain-specific pre-training improves the results, similar to Gu et al. (2021), we find clinical LMs underperform on the biomedical RE task. There was no performance gap between BlueBERT, SciBERT and BioBERT. However, PubMedBERT brought significant improvement which is consistent with Gu et al. (2021). In terms of expert knowledge-based models, we do not notice any improvements instead, all of them had a negative impact. While we would expect type-based models, MedType and UmlsBERT, to bring improvement, their negative effect can be attributed to overfitting certain types and their patterns. KeBioLM, which is initialized with PubMedBERT, slightly degrades the performance despite having seen the triples used in MEDDISTANT19 during pre-training, highlighting the difficulty of the MEDDISTANT19 dataset. SapBERT which uses the synonyms knowledge also hurt PubMedBERT's performance, suggesting that while synonyms can help for entity linking, RE is a much more elusive task in noisy real-world scenarios.

5 Conclusion

In this work, we highlighted a severe training-test overlap in the corpus used by previous studies in Bio-DSRE, causing inflated performance. We noted other inconsistencies including the KGs used and lack of standard baselines. To mitigate these issues, we introduce a new benchmark MEDDISTANT19, which derives its KG from SNOMED-CT (Chang et al., 2020) and is particularly focused on long-tail relations. The benchmark can directly be used with standard RE frameworks, such as OpenNRE (Han et al., 2019). We conducted a thorough set of experiments and provided baselines showing both the quality of the dataset and the need for better models.

6 Legal & Ethical Considerations

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information) We use PubMed MEDLINE abstracts (Canese and Weis, 2013)[7] that are publicly available and is distributed by National Library of Medicine (NLM). These texts are in the biomedical and clinical domain, and are almost entirely in English. It is standard to use this corpus as a text source in several biomedical LMs (Gu et al., 2021). We cannot claim the guarantee that it does not contain any confidential or sensitive information e.g, it has clinical findings mentioned throughout the abstracts such as *A twenty six year old male presented with high grade fever*, which identifies the age and gender of a patient but not the identity. We did not perform thorough analysis to distill such information since it is in public domain. For other concerns, see Appendix section B and D.

[7] https://lhncbc.nlm.nih.gov/ii/information/MBR/Baselines/2019.html

References

Asma Ben Abacha and Pierre Zweigenbaum. 2011. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of biomedical semantics*, 2(5):1–11.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy. Association for Computational Linguistics.

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 the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 187–194, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI Handbook*, 2:1.

David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Andrew Taylor. 2020. Benchmark and best practices for biomedical knowledge graph embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 167–176, Online. Association for Computational Linguistics.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2021. Distantly supervised relation extraction with sentence reconstruction and knowledge base priors. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–26, Online. Association for Computational Linguistics.

Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

Qin Dai, Naoya Inoue, Paul Reisert, Ryo Takahashi, and Kentaro Inui. 2019. Distantly supervised biomedical knowledge acquisition via knowledge graph based attention. In *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, pages 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Daza, Michael Cochez, and Paul Groth. 2021. Inductive entity representations from text via link prediction. In *Proceedings of The Web Conference 2021*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1811–1818. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610. ACM.

Tianyu Gao, Xu Han, Yuzhuo Bai, Keyue Qiu, Zhiyu Xie, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Manual evaluation matters: Reviewing test protocols of distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1306–1318, Online. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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 the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 745–758, Suzhou, China. Association for Computational Linguistics.

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 the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 169–174, Hong Kong, China. Association for Computational Linguistics.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4832–4839. AAAI Press.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

William P Hogan, Molly Huang, Yannis Katsis, Tyler Baldwin, Ho-Cheol Kim, Yoshiki Baeza, Andrew Bartko, and Chun-Nan Hsu. 2021. Abstractified multi-instance learning (amil) for biomedical relation extraction. In *3rd Conference on Automated Knowledge Base Construction*.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Ramakanth Kavuluru, Anthony Rios, and Tung Tran. 2017. Extracting drug-drug interactions with word and character-level recurrent neural networks. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 5–12. IEEE.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4289–4300.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C Rindflesch. 2011. Constructing a semantic predication gold standard from the biomedical literature. *BMC bioinformatics*, 12(1):1–17.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Dongwook Shin. 2020. Broad-coverage biomedical relation extraction with semrep. *BMC bioinformatics*, 21:1–28.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, and Ander Intxaurrondo. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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*, 2016.

ChunYang Liu, WenBo Sun, WenHan Chao, and Wanxiang Che. 2013. Convolution neural network for relation extraction. In *International Conference on Advanced Data Mining and Applications*, pages 231–242. Springer.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Yuan Luo, Özlem Uzuner, and Peter Szolovits. 2017. Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. *Briefings in bioinformatics*, 18(1):160–178.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

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.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proc. IEEE*, 104(1):11–33.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.

Yifan Peng and Zhiyong Lu. 2017. Deep learning for extracting protein-protein interactions from biomedical literature. In *BioNLP 2017*, pages 29–38, Vancouver, Canada,. Association for Computational Linguistics.

Yifan Peng, Chih-Hsuan Wei, and Zhiyong Lu. 2016. Improving chemical disease relation extraction with rich features and weakly labeled data. *Journal of cheminformatics*, 8(1):1–12.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

S Povey, R Lovering, E Bruford, M Wright, M Lush, and H Wain. 2001. The hugo gene nomenclature committee (hgnc). *Hum Genet*, 109(6):678–680.

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

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Tara Safavi and Danai Koutra. 2021. Relational World Knowledge Representation in Contextual Language Models: A Review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1053–1067, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez Fernández, and Daniel Sánchez Cisneros. 2011. The 1st ddiextraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts.

Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of Biomedical Informatics*, 121:103880.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2361–2364. ACM.

Rong Xu and QuanQiu Wang. 2014. Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. *Journal of biomedical informatics*, 51:191–199.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *ArXiv preprint*, abs/1909.03193.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190, Online. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

Yue Zhang, Hongliang Fei, and Ping Li. 2021. *ReadsRE: Retrieval-Augmented Distantly Supervised Relation Extraction*, page 2257–2262. Association for Computing Machinery, New York, NY, USA.

A UMLS

In this section we present additional details about UMLS, including the final set of relations considered in MEDDISTANT19 (with their inverses obtained from the UMLS) and a complete list of semantic types (STY). Since in relation extraction (RE), we are not interested in bidirectional extractions, therefore it is sufficient to only model one direction. Previous studies (Dai et al., 2019; Amin et al., 2020; Hogan et al., 2021) fail to take into account these inverse relations and with naive split, can lead to train-test leakages. For more discussion on the relations in UMLS, including transitive closures, see Section 3.1 in Chang et al. (2020).

A.1 UMLS Files

In UMLS (Bodenreider, 2004), a concept is provided with a unique identifier called Concept Unique Identifier (CUI), a term status (TS), and whether or not the term is preferred (TTY) in a given vocabulary e.g., SNOMED-CT. The concepts are stored in a file distributed by UMLS called `MRCONSO.RRF`.[8] Each concept further belongs to one or more semantic types (STY), provided in a file called `MRSTY.RRF`, with a type identifier TUI. There are 127 STY[9] in the UMLS2019AB version, which are mapped to 15 semantic groups (SG).[10]. The relationships between the concepts are organized in a multi-relational graph distributed in a file called `MRREL.RRF`[11]. The final set of relations considered in MEDDISTANT19 is presented in Table A.1.

Note that we only consider relations belonging to the *RO* (*has relationship other than synonymous, narrower, or broader*) type, which is consistent with prior works. This consideration ignores relations such as *isa*, which defines hierarchy among relations.

A.2 Semantic Groups and Semantic Types

As we noted in Fig. 3, entities and relations follow a long-tail distribution. This has a major impact on the quality of the dataset created. For

[8] https://www.ncbi.nlm.nih.gov/books/NBK9685/table/ch03.T.concept_names_and_sources_file_mr/

[9] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/Docs/SemanticTypes_2018AB.txt

[10] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/Docs/SemGroups_2018.txt

[11] https://www.ncbi.nlm.nih.gov/books/NBK9685/table/ch03.T.related_concepts_file_mrrel_rrf/?report=objectonly

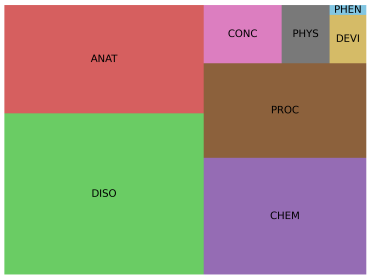

Figure A.1: Relative proportions of the entities present in MEDDISTANT19, based on the semantic groups.

```
python create_kb_aligned_text_corpora.py
        --medline_entities_linked_fname
    MEDLINE/medline_pubmed_2019_entity_linked.jsonl
        --triples_dir UMLS
        --split ind
        --sample 0.1
        --train_size 0.7
        --dev_size 0.1
        --raw_neg_sample_size 500
        --corrupt_arg
        --remove_multimentions_sents
        --use_type_constraint
        --use_arg_constraint
        --remove_mention_overlaps
        --canonical_or_aliases_only
        --prune_frequent_mentions
        --max_mention_freq 1000
        --min_rel_freq 1
        --prune_frequent_mentions
        --prune_frequent_bags
        --max_bag_size 500
```

Figure A.2: Relative proportions of the entities present in MEDDISTANT19, based on the semantic groups.

example in general-domain, the standard benchmark, NYT10 (Riedel et al., 2010), has more than half of the positive instances belonging to one relation type `/location/location/contains`. Fig. A.1 shows the relative proportions of the semantic groups in MEDDISTANT19.

Since MEDDISTANT19 aims to focus on rare triples, we prune the mentions by their types, to avoid creating and learning a biased data and model respectively. Below we provide a list of top-5 mentions for selected semantic types showing the presence of highly-frequent mentions, often picked by Bio-DSRE corpora. We remove such mentions by type-based pruning, setting the minimum mention frequency to be 1000.

- **Body Part, Organ, or Organ Component**: *(liver, 67264), (brain, 63234), (eyes, 25927), (lung, 25464), (kidney, 20825)*
- **Organism Function**: *(period, 29499), (blood pressure, 20868), (death, 12935), (BP, 9789), (died, 7905)*
- **Body Location or Region**: *(head, 16458), (neck, 6645), (face, 6480), (chest, 3919), (shoulder, 3338)*
- **Therapeutic or Preventive Procedure**: *(intervention, 59944), (procedure, 54594), (removal, 35543), (operation, 30961), (stimulation, 24058)*
- **Pathologic Function**: *(sensitivity, 49697), (sensitive, 25696), (inflammation, 18993), (blocked, 18138), (bleeding, 15292)*
- **Qualitative Concept**: *(increased, 449951), (effective, 48317), (effect, 44070), (normal, 43133), (reduced, 37787)*
- **Neoplastic Process**: *(tumor, 44632), (tumors, 34157), (cancer, 14314), (neck cancer, 8376), (tumour, 8288)*
- **Disease or Syndrome**: *(disease, 90345), (infection, 68763), (condition, 33060), (hypertension, 32197), (diseases, 25850)*
- **Functional Concept**: *(changes, 88517), (absence, 39080), (impaired, 30194), (progressive, 24817), (functions, 24678)*
- **Laboratory Procedure**: *(cells, 45314), (test, 12502), (erythrocytes, 11916), (tests, 9907), (RBC, 7020)*
- **Diagnostic Procedure**: *(MRI, 26224), (US, 17279), (biopsy, 14352), (ultrasound, 11663), (imaging, 9635)*
- **Finding**: *(presence, 176771), (positive, 88797), (negative, 42464), (severe, 37334), (lesions, 31747)*
- **Hormone**: *(insulin, 12365), (LH, 5738), (cortisol, 5223), (estradiol, 4144), (TSH, 3319)*
- **Biologically Active Substance**: *(protein, 23232), (proteins, 20662), (amino acids, 19187), (glucose, 13968), (ATP, 13228)*

This was the most important pruning method that removed a major portion of noisy sentences (removed / original): train (3,576,637 / 3,828,374), validation (561,176 / 740,576), and test (1,616,412 / 1,830,024).

Fig. A.2 shows the final command that was used to create MEDDISTANT19 benchmark with the inductive split set at 70, 10 and 20 proportions of train, validation and test splits.

Relation	Inverse Relation
finding_site_of	*has_finding_site*
associated_morphology_of	*has_associated_morphology*
method_of	*has_method*
interprets	*is_interpreted_by*
direct_procedure_site_of	*has_direct_procedure_site*
causative_agent_of	*has_causative_agent*
active_ingredient_of	*has_active_ingredient*
pathological_process_of	*has_pathological_process*
entire_anatomy_structure_of	*has_entire_anatomy_structure*
interpretation_of	*has_interpretation*
laterality_of	*has_laterality*
component_of	*has_component*
indirect_procedure_site_of	*has_indirect_procedure_site*
direct_morphology_of	*has_direct_morphology*
cause_of	*due_to*
intent_of	*has_intent*
direct_substance_of	*has_direct_substance*
uses_device	*device_used_by*
clinical_course_of	*has_clinical_course*
focus_of	*has_focus*
direct_device_of	*has_direct_device*
finding_method_of	*has_finding_method*
procedure_site_of	*has_procedure_site*
uses_substance	*substance_used_by*
associated_finding_of	*has_associated_finding*
associated_procedure_of	*has_associated_procedure*
occurs_after	*occurs_before*
is_modification_of	*has_modification*
uses_access_device	*access_device_used_by*
specimen_source_topography_of	*has_specimen_source_topography*
plays_role	*role_played_by*
specimen_procedure_of	*has_specimen_substance*
indirect_morphology_of	*has_indirect_morphology*
part_anatomy_structure_of	*has_part_anatomy_structure*
specimen_source_morphology_of	*has_specimen_source_morphology*
specimen_source_identity_of	*has_specimen_source_identity*
during	*inverse_during*
direct_site_of	*has_direct_site*

Table A.1: *(Left)* 38 relations included in MEDDISTANT19, excluding NA relation. *(Right)* For completeness, we also include their inverse relations.

SG	TUI	Semantic Type
ANAT	T017	Anatomical Structure
	T029	Body Location or Region
	T023	Body Part, Organ, or Organ Component
	T030	Body Space or Junction
	T031	Body Substance
	T022	Body System
	T021	Fully Formed Anatomical Structure
	T024	Tissue
CHEM	T116	Amino Acid, Peptide, or Protein
	T195	Antibiotic
	T123	Biologically Active Substance
	T103	Chemical
	T200	Clinical Drug
	T196	Element, Ion, or Isotope
	T126	Enzyme
	T131	Hazardous or Poisonous Substance
	T125	Hormone
	T129	Immunologic Factor
	T130	Indicator, Reagent, or Diagnostic Aid
	T197	Inorganic Chemical
	T114	Nucleic Acid, Nucleoside, or Nucleotide
	T109	Organic Chemical
	T121	Pharmacologic Substance
	T192	Receptor
	T127	Vitamin
CONC	T185	Classification
	T169	Functional Concept
	T102	Group Attribute
	T078	Idea or Concept
	T170	Intellectual Product
	T080	Qualitative Concept
	T081	Quantitative Concept
	T082	Spatial Concept
	T079	Temporal Concept
DEVI	T074	Medical Device
	T075	Research Device
DISO	T020	Acquired Abnormality
	T190	Anatomical Abnormality
	T049	Cell or Molecular Dysfunction
	T019	Congenital Abnormality
	T047	Disease or Syndrome
	T033	Finding
	T037	Injury or Poisoning
	T048	Mental or Behavioral Dysfunction
	T191	Neoplastic Process
	T046	Pathologic Function
	T184	Sign or Symptom
PHEN	T038	Biologic Function
	T068	Human-caused Phenomenon or Process
	T034	Laboratory or Test Result
	T070	Natural Phenomenon or Process
	T067	Phenomenon or Process
PHYS	T201	Clinical Attribute
	T041	Mental Process
	T032	Organism Attribute
	T040	Organism Function
	T042	Organ or Tissue Function
	T039	Physiologic Function
PROC	T060	Diagnostic Procedure
	T065	Educational Activity
	T058	Health Care Activity
	T059	Laboratory Procedure
	T063	Molecular Biology Research Technique
	T062	Research Activity
	T061	Therapeutic or Preventive Procedure

Table A.2: 65 semantic types (STY) along with their TUIs and semantic groups (SG) covered in MEDDISTANT19.

Below is an example instance from MEDDISTANT19 in OpenNRE (Han et al., 2019) format:

```
{
  "text": "In one patient who
  showed an increase of plasma
  prolactin level , associated
  with low testosterone and
  LH , a microadenoma
  of the pituitary gland
  ( prolactinoma ) was
  detected .",
  "h": {
    "id": "C0032005",
    "pos": [130, 145],
    "name": "pituitary gland"
  },
  "t": {
    "id": "C0033375",
    "pos": [148, 160],
    "name": "prolactinoma"
  },
  "relation": "finding_site_of"
}
```

B UMLS License Agreement

To use this MEDDISTANT19, the user must have signed the UMLS agreement[12]. The UMLS agreement requires those who use the UMLS (Bodenreider, 2004) to file a brief report once a year to summarize their use of the UMLS. It also requires the acknowledgment that the UMLS contains copyrighted material and that those copyright restrictions be respected. The UMLS agreement requires users to agree to obtain agreements for EACH copyrighted source prior to its use within a commercial or production application.

C Limitations

We provide several limitations of our work as presented in its current form. MEDDISTANT19 aims to introduce a new benchmark with good practices, however, it is still limited in its scope of ontologies considered. It also has a limited subset of relation types provided by UMLS. For example, the current benchmark does not include an important relation *may_treat*, because it is outside SNOMED-CT. Since, MEDDISTANT19 is focused on SNOMED-CT, it lacks coverage of important protein-protein interactions, drugs side effects, and relations involving genes as provided by RxNorm, Gene Ontology etc. It is also smaller in size compared to the benchmark in general-domain (Riedel et al., 2010). Despite these limitations, MEDDISTANT19 still offers a challenging and focused benchmark that can help improve the weakly supervised broad-coverage biomedical RE.

[12] https://uts.nlm.nih.gov/license.html

D Risks

While our work does not have direct risk, we do provide the dataset while asking users to respect the UMLS license before downloading it. This user agreement is needed to use our benchmark and to respect the source ontologies licenses. We provide this with hope to accelerate reproducible research in Bio-DSRE by having a ready-to-use corpora, with only the condition that the license has been obtained by the user. We provide users with this note and hope this will be respected. However, there is a risk that users may download the data and re-distribute without respecting the UMLS license. In case of such exploitation, we will add the UMLS authentication layer to protect data where the user will be required to provide UMLS api-key, which will be validated and only then the data will be allowed to be downloaded.

E Experimental Setup and Hyperparameters

We followed the experimental setup of Gao et al. (2021) for BERT-based experiments. Specifically we used the batch size 64, with learning rate 2e-5, maximum sequence length 128, bag size 4 where applicable. We used a single NVIDIA Tesla V100-32GB for BERT-based experiments. Each experiment took about 1.5 hrs with half an hour per epoch. We also attempted to perform grid search for BERT experiments but it was too expensive to continue, therefore we abandoned those jobs. Since we only used the `base` models, they amount to 110 million parameters. During fine-tuning, we do not freeze any parts of the model.

For CNN and PCNN, we performed grid search with optimizers $\in$ {Adam (Kingma and Ba, 2015), SGD (Ruder, 2016)}, learning rate $\in \{0.01, 0.001\}$, batch size $\in \{64, 160\}$, bag size $\in \{4, 8, 12, 16, 32, 64\}$, embeddings $\in$ {Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014)}, and with (test-time) pooling $\in$ {ONE, AVG} when using sentence-level train-

Encoder	Bag Size	Embedding
CNN+sent+AVG	16	GloVe
CNN+sent+ONE	16	GloVe
CNN+bag+AVG	32	Word2Vec
CNN+bag+ONE	4	Word2Vec
CNN+bag+ATT	12	Word2Vec
PCNN+sent+AVG	4	GloVe
PCNN+sent+ONE	4	GloVe
PCNN+bag+AVG	32	Word2Vec
PCNN+bag+ONE	16	GloVe
PCNN+bag+ATT	4	GloVe

Table A.3: Best hyperparameters for CNN and PCNN sentence encoders.

ing and pooling in {ONE, AVG, ATT} when using bag-level training. We ran this job on a cluster with support for array jobs. These amounted to over 700 experiments and took 3 days. We fixed other hyperparameters from literature (Han et al., 2018), with position dimension set to 5, kernel size set to 3, and dropout set to 0.5. These are also default in OpenNRE (Han et al., 2019). We found Adam to be the better optimizer in all configurations along with batch size 160 and learning rate 0.001 except in PCNN+sent+AVG, where 0.01 was better learning rate. The hyperparameters that had the most influence were bag size and pre-trained word embeddings. All the experiments reported in the paper are with a single run.

We also needed heavy compute budget for SciSpacy-based sentence tokenization and entity linking jobs. It took 9hrs with 32 CPUs (4GB each) and a batch size of 1024 for spaCy to extract 151M sentences. The entity linking job took about half TB of RAM with 72 CPUs (6GB each) with a batch size 4096. It took 40hrs to link 145M unique sentences.