# MedDistant19: Towards an Accurate Benchmark for Broad-Coverage Biomedical Relation Extraction

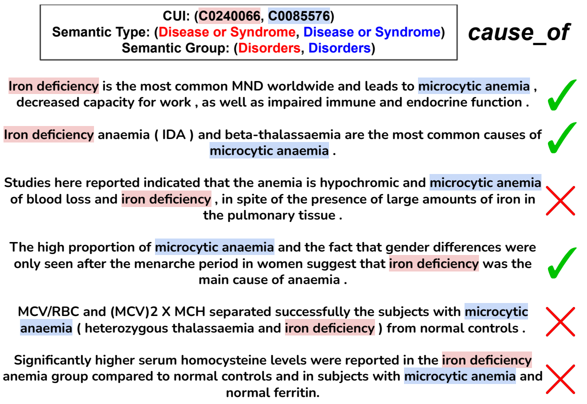
**Saadullah Amin**[♣,△] [*] **Pasquale Minervini**[♠] [*] **David Chang**[♢]
**Pontus Stenetorp**[♠] **Günter Neumann**[♣,△]
[♣]German Research Center for Artificial Intelligence [♠]UCL Centre for Artificial Intelligence
[△]Saarland Informatics Campus, Saarland University [♢]Yale Center for Medical Informatics
{saadullah.amin,guenter.neumann}@dfki.de {p.minervini,p.stenetorp}@cs.ucl.ac.uk
david.chang@yale.edu

## Abstract

Relation extraction in the biomedical domain is challenging due to the lack of labeled data and high annotation costs, needing domain experts. Distant supervision is commonly used to tackle the scarcity of annotated data by automatically pairing knowledge graph relationships with raw texts. Such a pipeline is prone to noise and has added challenges to scale for covering a large number of biomedical concepts. We investigated existing broad-coverage distantly supervised biomedical relation extraction benchmarks and found a significant overlap between training and test relationships ranging from 26% to 86%. Furthermore, we noticed several inconsistencies in the data construction process of these benchmarks, and where there is no train-test leakage, the focus is on interactions between narrower entity types. This work presents a more accurate benchmark MedDistant19 for broad-coverage distantly supervised biomedical relation extraction that addresses these shortcomings and is obtained by aligning the MEDLINE abstracts with the widely used SNOMED Clinical Terms knowledge base. Lacking thorough evaluation with domain-specific language models, we also conduct experiments validating general domain relation extraction findings to biomedical relation extraction.

## 1 Introduction

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



CUI: (C0240066, C0085576)
Semantic Type: (Disease or Syndrome, Disease or Syndrome)
Semantic Group: (Disorders, Disorders)

*cause_of*

Iron deficiency is the most common MND worldwide and leads to microcytic anemia , decreased capacity for work , as well as impaired immune and endocrine function . ✓

Iron deficiency anaemia ( IDA ) and beta-thalassaemia are the most common causes of microcytic anaemia . ✓

Studies here reported indicated that the anemia is hypochromic and microcytic anemia of blood loss and iron deficiency , in spite of the presence of large amounts of iron in the pulmonary tissue . ✗

The high proportion of microcytic anaemia and the fact that gender differences were only seen after the menarche period in women suggest that iron deficiency was the main cause of anaemia . ✓

MCV/RBC and (MCV)2 X MCH separated successfully the subjects with microcytic anaemia ( heterozygous thalassaemia and iron deficiency ) from normal controls . ✗

Significantly higher serum homocysteine levels were reported in the iron deficiency anemia group compared to normal controls and in subjects with microcytic anemia and normal ferritin. ✗

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 incorrect labels resulting from the distant supervision.

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).

To scale to a large number of biomedical entities, recent works have focused on broad-coverage relation extraction (Amin et al., 2020a; Xing et al., 2020; Hogan et al., 2021), where we investigated these benchmarks for possible train-test leakage of knowledge graph triples and found significant portions overlapping (Table 2). Such leakage impacts the model performance as it allows to score higher by simply memorizing the training relations rather than generalizing to new, previously unknown ones. We identify the sources of these issues as normalizing the textual form of concept mentions to their unique identifiers and improper handling of inverse relations. In contrast, more ac-

[*] *Equal contribution.*

| Benchmark | Relations | No Train-Test Overlap | Broad-Coverage | Ontology |
|---|---|---|---|---|
| UMLS.v1 (Roller and Stevenson, 2014) | 7 | - | ✗ | UMLS |
| DTI (Hong et al., 2020) | 6 | ✓ | ✗ | DrugBank |
| UMLS.v2 (Amin et al., 2020a) | 355 | ✗ | ✓ | UMLS |
| BioRel (Xing et al., 2020) | 125 | ✗ | ✓ | NDFRT, NCI |
| UMLS.v3 (Hogan et al., 2021) | 275 | ✗ | ✓ | UMLS |
| TBGA (Marchesin and Silvello, 2022) | 4 | ✓ | ✗ | DisGeNET |
| MedDistant19 | 22 | ✓ | ✓ | SNOMED CT |

Table 1: The landscape of distantly supervised biomedical relation extraction (Bio-DSRE) benchmarks: all the existing broad-coverage datasets have corpus-level triples overlap between the train and test splits (Table 2), where the knowledge graph (KG) is also extracted from multiple ontologies. The DTI and TBGA benchmarks focus on harmonized ontology but are limited to drug-target interactions and gene-disease associations. In contrast, MEDDISTANT19 has a broader coverage of entities and their semantic types and is normalized to a single ontology, SNOMED CT, which has significant clinical relevance. We named the datasets from (Roller and Stevenson, 2014; Amin et al., 2020a; Hogan et al., 2021) to UMLS.v1/2/3 since the original works had no names. For UMLS.v1, there is no publicly available code to reconstruct the dataset; thus, the overlap information is missing.

curate benchmarks exist (Hong et al., 2020; Marchesin and Silvello, 2022) but focus on narrower types of interactions. To alleviate the broad-coverage benchmark issues and bridge this gap, we present a new benchmark MEDDISTANT19 which draws its knowledge graph from the widely used healthcare ontology SNOMED CT (Chang et al., 2020). Further, with the success of domain-specific pre-trained language models for biomedical and clinical tasks (Gu et al., 2021), and inspired by existing thorough relation extraction studies in the general domain (Peng et al., 2020; Alt et al., 2020; Gao et al., 2021), we conduct an extensive evaluation using MEDDISTANT19 for the biomedical domain.

## 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). Distant Supervision (DS) allows for the automated collection of noisy training examples by aligning a given knowledge base (KB) with a collection of text sources (Mintz et al., 2009). DS was used in recent works (Alt et al., 2019; Amin et al., 2020a) with pre-trained language models using Multi-Instance Learning (MIL) by creating *bags* of instances (Riedel et al., 2010) for corpus-level triple extraction.[1] In biomedical domain, Roller and Stevenson (2014) first proposed 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.

For broad-coverage tasks, Dai et al. (2019) implemented a knowledge-based attention mechanism (Han et al., 2018) for mutual learning with knowledge graph completion and entity type classification. Xing et al. (2020) introduced a large-scale BioRel benchmark focusing on drug-disease and gene-cancer interactions and showed significant performance using a comprehensive selection of baselines. Recent works focused on using domain-specific pre-trained language models for distantly supervised biomedical relation extraction (Bio-DSRE). Amin et al. (2020a) extended relation enriched sentence-level BERT (Wu and He, 2019) to handle bag-level MIL and demonstrated that preserving the direction of the KB relationships can denoise the training signal. They also outlined the steps to create a broad-coverage benchmark from UMLS. Following this, Hogan et al. (2021) introduced the concept of *abstractified* MIL (AMIL), by including different argument pairs belonging to the same semantic type pair in one bag, boosting performance on rare triples.

For domain-specific Bio-DSRE, Hong et al. (2020) introduced the BERE framework for latent tree learning and self-attention to use the semantic and syntactic information in the sentence for MIL. They also introduced a drug-target interactions (DTI) Bio-DSRE benchmark, suitable for drug repositioning, drawn from DrugBank

[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 graph completion or link prediction task (Amin et al., 2020b).

| Triples | Train | Valid | Test |
|---|---|---|---|
| UMLS.v2 | 211,789 | 41,993 (26.7%) | 89,486 (26.5%) |
| BioRel | 39,969 | 17,815 (86.17%) | 17,927 (86.37%) |
| UMLS.v3 | 23,163 | 2,643 (44.38%) | 5,184 (40.12%) |

Table 2: Training-test leakage we identified in the existing broad-coverage benchmarks. Numbers between parentheses show the percentage overlap of CUI triples.

| Model and Data | Original | | Filtered | |
|---|---|---|---|---|
| | AUC | F1 | AUC | F1 |
| Amin et al. (2020a) | 68.4 | 64.9 | 50.8 | 53.1 |
| Hogan et al. (2021)$^{\dagger}$ | 82.6 | 77.6 | 11.8 | 19.8 |

Table 3: State-of-the-art Bio-DSRE language models were evaluated on the respective datasets before (Original) and after (Filtered) removing overlapping relationships. $^{\dagger}$ Our re-run of the AMIL (Type L) model; original scores are 87.2 (AUC) and 81.2 (F1).

(Wishart et al., 2018). Concurrent work of Marchesin and Silvello (2022) introduced a large-scale semi-automatically curated benchmark TGBA for gene-disease associations (GDA). TGBA uses DisGeNET (Piñero et al., 2020), which collects data on human genotype-phenotype relationships.

This work investigates recent results from the broad-coverage Bio-DSRE literature by probing the respective datasets for overlaps between training and test sets. Specifically, in UMLS, each concept is mapped to a *Concept Unique Identifier (CUI)*, and a given CUI might have different surface forms (Bodenreider, 2004), we thus probe for CUI-based KG triples leakage. Our results are shown in Table 2 for UMLS.v2 (Amin et al., 2020a), BioRel (Xing et al., 2020), and UMLS.v3 (Hogan et al., 2021). For UMLS.v2 and UMLS.v3, the triples use surface forms of CUIs rather than the CUIs themselves, which results in an overlap between training and test sets. For example, consider a relationship between a pair of UMLS entities (`C0013798`, `C0429028`). These two entities can appear in different forms within a text, 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. (2020a) claim no such text-based leakage, but when canonicalized to their CUIs, this results in leakage across the splits as reported in Table 2. In contrast, BioRel directly splits CUI triples without accounting for inverse relations that can also result in leakage (Chang et al., 2020). Since DSRE aims at corpus-level triples extraction, train-test triples leakage is problematic (see Table 3) compared to supervised sentence-level RE, where we aim to generalize to newer contexts.

We found no such overlap for DTI and TBGA, where the datasets used in (Roller and Stevenson, 2014; Dai et al., 2019) are not publicly available. Noting these shortcomings, we introduce a new and accurate benchmark MedDistant19 for broad-coverage Bio-DSRE. Our benchmark utilizes clinically relevant SNOMED CT Knowledge Graph (Chang et al., 2020), extracted from the UMLS, that offers a careful selection of the concept types and is suitable for large-scale biomedical relation extraction. Table 1 summarizes the current landscape of Bio-DSRE benchmarks.

In supervised RE, ChemProt (Krallinger et al., 2017) and DDI-2013 (Herrero-Zazo et al., 2013) focus on multi-class interactions between chemical-protein and drug-drug respectively. EU-ADR (van Mulligen et al., 2012) and GAD (Bravo et al., 2015) focus on binary relations between genes and diseases, while CDR (Li et al., 2016) focuses on binary relations between chemicals and diseases.

## 3 Constructing the MedDistant19 Benchmark

**Documents** We used PubMed MEDLINE abstracts published up to 2019[2] as our text source, containing 32,151,899 abstracts. Following Hogan et al. (2021), we used ScispaCy [3] (Neumann et al., 2019) for sentence tokenization, resulting in 150,173,169 unique sentences. We further introduce the use of ScispaCy for linking entity mentions to their UMLS CUIs and filtering disabled concepts from UMLS, which resulted in entity-linked mentions at the sentence-level.

Named entity recognition (NER) and normalization were two primary sources of errors in biomedical RE, as shown in Kilicoglu et al. (2020). While ScispaCy is reasonably performant among other options for biomedical entity linking, it remains quite noisy in practice; e.g., Vashishth et al. (2021) showed that ScispaCy had only about a 50% accuracy on extracting concepts in benchmark datasets. Despite this being a limitation, using ScispaCy is better than relying on string matching alone (Dai et al., 2019; Amin et al., 2020a; Hogan et al., 2021).

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

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

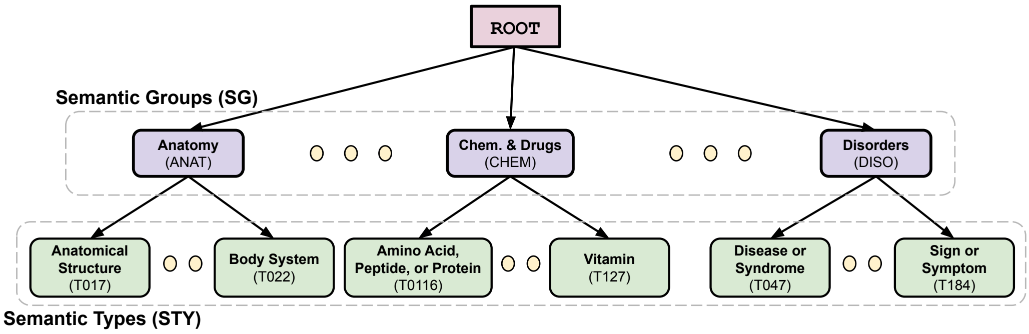


Figure 2: Type hierarchy in UMLS, where each concept is classified under a taxonomy. The *coarse-grained* and *fine-grained* entity types are referred to as Semantic Group (SG) and Semantic Type (STY) respectively.

**Knowledge Base** We use UMLS2019AB [4] as our primary knowledge source and apply a set of rules, resulting in a distilled and carefully reduced version of UMLS2019AB. The UMLS Metathesaurus (Bodenreider, 2004) covers concepts from 222 source vocabularies, thus being the most extensive 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 easier to have overlaps between training and test sets since the same fact may appear in both with different entity identifiers.

Furthermore, benchmark corpora for biomedical NER (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, the most widely used clinical terminology worldwide for documentation and reporting in healthcare (Chang et al., 2020).

Since UMLS classifies each entity in a type taxonomy of semantic types (STY) and semantic groups (SG) (Fig. 2), this allows for narrowing the concepts of interest. Following Chang et al. (2020), we first 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). We then remove CONC and PHEN as they are far too general to be informative for Bio-DSRE. For a complete list of semantic types covered in MEDDISTANT19, see Table A.4. Similarly, each relation is categorized into a type and has a reciprocal relation in UMLS (Table A.3), which can result in train-test leakage (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. 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) ensures 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, we are expected to extract relations between unseen entities in real-world applications of biomedical RE. To support this setup, we derive MEDDISTANT19 using an inductive KG split method proposed by Daza et al. (2021) (see Appendix A in their paper). Table 5 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.

### 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 UMLS CUIs and relation types, and let $\mathcal{G} \subseteq$

[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 4: MEDDISTANT19 (MD19) key data construction properties compared with the recent broad-coverage Bio-DSRE works.

| Facts | Training | Validation | Testing |
|---:|:---:|:---:|:---:|
| Inductive | 261,797 | 48,641 | 97,861 |
| Transductive | 318,524 | 28,370 | 56,812 |

Table 5: The number of raw inductive and transductive SNOMED KG triples used for alignment with text.

| Summary | Entities | Relations | STY | SG |
|:---:|:---:|:---:|:---:|:---:|
| | 20,256 | 22 | 51 | 6 |

| Split | Instances | Facts | Bags | Inst. per Bag | NA (%) |
|:---:|:---:|:---:|:---:|:---:|:---:|
| **Train** | 450,071 | 5,455 | 88,861 | 5.06 | 90.0% |
| **Valid** | 39,434 | 842 | 10,475 | 3.76 | 91.2% |
| **Test** | 91,568 | 1,663 | 22,606 | 4.05 | 91.1% |

Table 6: Summary statistics of the MEDDISTANT19 dataset using Inductive SNOMED KG split (Table 5). The number of relations includes the unknown relation type (NA).

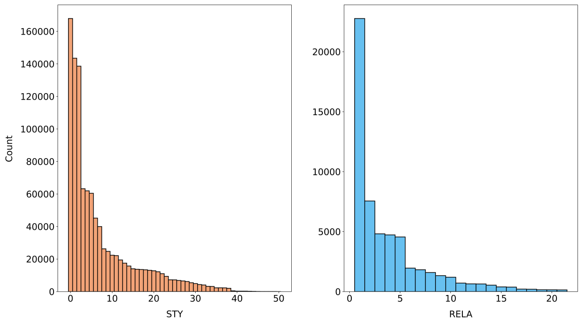


Figure 3: (*Left*) Entity distribution based on Semantic Types. (*Right*) Relations distribution.

$\mathcal{E} \times \mathcal{R} \times \mathcal{E}$ 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:

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

Following this, 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, defined as follows:

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

During the corruption process, we enforce two constraints: 1) *type constraint* – the two entities appearing in each negative pair in $\mathcal{G}^-$ should belong to an entity type pair from $\mathcal{G}^+$, and 2) *role constraint* – the noisy *head* (*tail*) entity in negative pair must have appeared in *head* (*tail*) role from a pair in $\mathcal{G}^+$.

A naive choice for the negative group could be $\mathcal{G}^- = (\mathcal{E} \times \mathcal{E}) - \mathcal{G}^+$, for which the current approach is only a subset; however, enumerating all possible entity pairs can be infeasible if $|\mathcal{E}|$ is high. Furthermore, we do not assume the completeness of UMLS, and only derive a *fixed* sub-graph from the 2019 version subject to the constraints. This process is similar to Local-Closed World Assumption (LCWA, Dong et al., 2014; Nickel et al., 2016), in which a KG is assumed to be only locally complete: if we observed a triple for a specific entity $e_i \in \mathcal{E}$, then we assume that any non-existing relationship $(e_i, e_j)$ denotes a false fact and include them in $\mathcal{G}^-$. Therefore, it is likely that if a triple emerges in a new PubMed article such that it violates the negative sampling assumptions, it will be considered a false negative. However, this amount is negligible due to intractable search space that scales with the size of the KG.

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. (2020a), 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 relationship, but the considered set of relations does not cover it. Our design choices are summarized in Table 4.

We also remove mention-level overlap across the splits and apply type-based mention pruning. Specifically, we pool mentions by type and remove the sentences which have the mention appearing more than 10,000 times. We selected the threshold based on manual inspection of frequent mentions

| Model | Bag | Strategy | AUC | F1-micro | F1-macro | P@100 | P@200 | P@300 | P@1k | P@2k |
|---|---|---|---|---|---|---|---|---|---|---|
| CNN | - | AVG | 27.3 | 33.0 | 16.1 | 50.0 | 46.0 | 44.0 | 41.0 | 33.6 |
| | - | ONE | 30.4 | 36.7 | 18.2 | 67.0 | 58.5 | 52.6 | 43.5 | 34.4 |
| | ✓ | AVG | 30.4 | 36.2 | 19.8 | 70.0 | 58.0 | 56.0 | 46.0 | 35.5 |
| | ✓ | ONE | 34.6 | 40.4 | 17.8 | 77.0 | 72.5 | 67.6 | 50.0 | 37.3 |
| | ✓ | ATT | 35.0 | 40.1 | 19.8 | 78.0 | 73.5 | 68.6 | 51.4 | 36.4 |
| PCNN | - | AVG | 27.2 | 32.4 | 12.9 | 54.0 | 49.5 | 50.3 | 40.7 | 33.2 |
| | - | ONE | 29.8 | 36.7 | 16.2 | 66.0 | 55.5 | 52.3 | 44.4 | 34.2 |
| | ✓ | AVG | 29.6 | 37.3 | 20.5 | 59.0 | 50.5 | 50.0 | 47.0 | 35.9 |
| | ✓ | ONE | 28.6 | 36.5 | 18.1 | 66.0 | 65.0 | 62.0 | 44.7 | 33.7 |
| | ✓ | ATT | 32.5 | 38.2 | 14.4 | 71.0 | 71.0 | 67.3 | 49.0 | 35.2 |
| GRU | - | AVG | 42.7 | 47.4 | 27.8 | 78.0 | 74.0 | 76.0 | 59.2 | 42.7 |
| | - | ONE | 46.4 | 49.3 | 29.2 | 86.0 | 80.5 | 78.3 | 61.2 | 44.9 |
| | ✓ | AVG | 28.6 | 37.2 | 17.9 | 57.0 | 57.0 | 56.0 | 45.3 | 35.4 |
| | ✓ | ONE | 32.6 | 40.8 | 17.7 | 73.0 | 70.5 | 66.3 | 51.2 | 37.0 |
| | ✓ | ATT | 36.6 | 40.9 | 22.2 | 77.0 | 72.0 | 67.6 | 51.3 | 38.7 |
| BERT | - | AVG | **79.8** | **76.1** | **65.3** | 95.0 | 96.0 | 96.0 | **90.2** | 67.2 |
| | - | ONE | 79.3 | **76.1** | 64.7 | 93.0 | 94.0 | 94.0 | 89.2 | **67.4** |
| | ✓ | AVG | 78.3 | 73.1 | 51.1 | **99.0** | **97.5** | **96.6** | 87.8 | 66.0 |
| | ✓ | ONE | 67.0 | 55.7 | 44.4 | 89.0 | 90.5 | 91.0 | 78.7 | 57.8 |
| | ✓ | ATT | 64.6 | 56.4 | 42.7 | 89.0 | 87.5 | 85.6 | 75.4 | 57.9 |

Table 7: Baseline results for MEDDISTANT19.

in each semantic type, so the information loss is minimal. At the same time, we still removed generalized mentions such as *disease*, *drugs*, *temperature* etc. We provide a complete list of mentions removed by this step in Table A.2. Table 6 shows the final summary of MEDDISTANT19 using inductive split covering 20,256 entities with 51 types and 343 type pairs. Fig. 3 shows entity and relation plots, following a long-tail distribution.

## 4 Experiments

MEDDISTANT19 is released in a format that is compatible with the widely adopted RE framework OpenNRE (Han et al., 2019).[5] To report our results, we use the *corpus-level* Area Under the Precision-Recall (PR) curve (AUC), Micro-F1, Macro-F1, and Precision-at-$k$ (P@$k$) with $k \in \{100, 200, 300, 1k, 2k\}$, and the *sentence-level* Precision, Recall, and F1. Due to the imbalanced nature of relational instances, 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) with the addition of GRU models.[6] For sentence encoding, we use CNN (Liu et al., 2013), PCNN (Zeng et al., 2015), bidirectional GRU (Hong et al., 2020), and BERT (Devlin et al., 2019). We use GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013) for CNN/PCNN/GRU models and initialize BERT with BioBERT (Lee et al., 2020).

We trained our models both at *sentence-level* and at *bag-level*. In contrast, prior works only considered bag-level training for Bio-DSRE. The sentence-level setup is similar to standard RE (Wu and He, 2019), with the difference that the evaluation is conducted at the 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 7 presents our main results. In all the cases, the BERT sentence encoder performed better than

[5] https://github.com/suamin/MedDistant19

[6] https://github.com/pminervini/meddistant-baselines

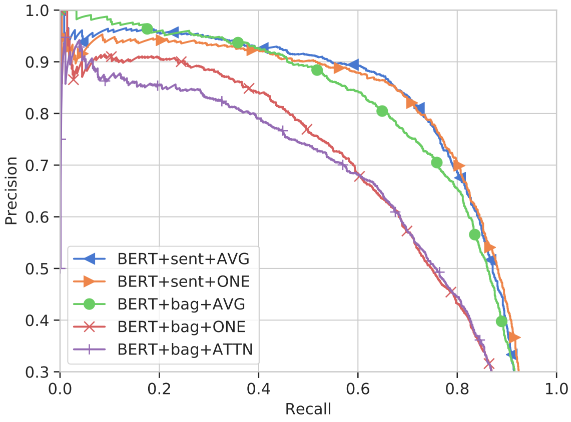


Figure 4: Precision-Recall curves for BERT baselines.

| Model | 1-1 | 1-M | M-1 |
|---|---|---|---|
| BERT+bag+AVG | **66.6** | **48.3** | **66.6** |
| BERT+bag+ONE | 52.6 | 33.2 | 47.1 |
| BERT+bag+ATT | 56.4 | 30.7 | 26.4 |

Table 8: Averaged F1-micro score on relation-specific category for *bag* pooling methods. The categories are defined using the *cardinality* of head and tail SGs.

| Model | P | R | F1 |
|---|---|---|---|
| **All Triples** | | | |
| BERT+sent+AVG | **0.79** | **0.65** | **0.71** |
| BERT+bag+AVG | 0.72 | 0.64 | 0.68 |
| **Common Triples** | | | |
| BERT+sent+AVG | **0.98** | **0.62** | **0.76** |
| BERT+bag+AVG | 0.96 | 0.60 | 0.74 |
| **Rare Triples** | | | |
| BERT+sent+AVG | **0.97** | 0.70 | 0.82 |
| BERT+bag+AVG | 0.95 | **0.73** | **0.83** |

Table 9: Sentence-level RE comparing BERT baselines trained at bag and sentence-level with AVG pooling on Rare and Common subsets of MEDDISTANT19. The triples include NA relational instances.

others since pre-trained language models are effective for entity-centric transfer learning (Amin and Neumann, 2021), domain-specific fine-tuning (Amin et al., 2019), and can implicitly store relational knowledge during pre-training (Petroni et al., 2019). This trend is similar to the general domain, and the BERT-based experiments provide consistent baselines lacking in the prior works. Similar to the general domain (Gao et al., 2021), we find sentence-level training to perform better than the bag-level. However, BERT+bag+AVG had much better precision for the top-scoring triples at the expense of long-tail performance. At the sentence-level, those instances that have been correctly labeled by distant supervision (e.g. Fig. 1) provide enough learning signal, given the generalization abilities of LMs. However, the model is supposed to jointly learn from clean and noisy samples in bag-level training, thus limiting its overall performance. But, we do not find this trend for CNN/PCNN. Instead, the bag-level models performed slightly better except for GRU. We further plot Precision-Recall (PR) curves for BERT-based baselines in Fig. 4.

**Pooling Strategies** In all cases, AVG proved to be a better pooling strategy; this finding is consistent with prior works. Both Amin et al. (2020a) and Gao et al. (2021) found ATT to produce less accurate results with LMs, which we also find to hold true for MEDDISTANT19. To further study the impact of bag-level pooling strategies, we analyze the relation category-specific results. 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 belongs to only one semantic group, then it has a cardinality one otherwise, M (many). The results are shown in Table 8 for bag-level BERT-based models with three pooling schemes. On average, models struggled the most with the 1-M category due to a lack of enough training signal to differentiate between heterogeneous entity types pooled over instances in a bag. While we would expect symmetric performance, to some extent, in 1-M and M-1 categories, the difference highlights that the KB-direction plays a role in Bio-DSRE, which previously has been used to de-noise the training signal (Amin et al., 2020a).

**Long-Tail Performance** Following Hogan et al. (2021), we also perform sentence-level triples 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 9 shows these results. We note that both training strategies performed comparably on rare triples with BERT+sent+AVG more precise than BERT+bag+AVG at the expense of low recall. However, we find a noticeable difference in com-

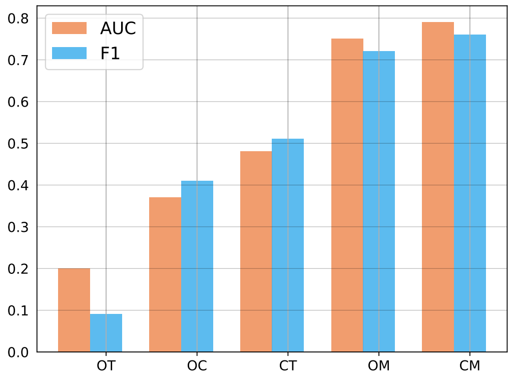


Figure 5: Ablation showing the effect of different text encoding methods with MedDistant19.

mon triples where BERT+sent+AVG performed better. At the bag level, the model can overfit to certain type and mention heuristics, whereas sentence-level training allows more focus on context. The current state-of-the-art model from Hogan et al. (2021) creates a bag of instances by abstracting entity pairs belonging to the same semantic type pair into a single bag, thus producing heterogeneous bags. Due to such bag creation, it is not suited for sentence-level models.

## 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 (Amin et al., 2020a) as the *Context + Mention* (CM) setting, which is common in RE with LMs. We then remove the context and only use mentions, the *Only Mention* (OM) setting, which reduces 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 train the models at the sentence-level and evaluate them at the bag-level.

We observe in Fig. 5 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). 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, more pronounced in noise-prone distant RE than in supervised RE. Our CT setup can be seen as a 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 constrain the space of the relations. Using only the type information had the least performance as the model fails to disambiguate between different entities belonging to the same type.

| Split | AUC | F1-micro | F1-macro |
|---|---|---|---|
| Inductive | **79.9** | **76.2** | 65.4 |
| Transductive | 79.6 | 73.3 | **65.9** |

Table 10: BERT+sent+AVG performance on corpora created with an inductive and transductive set of triples.

**Inductive or Transductive?** To study the impact of *transductive* and *inductive* splits (Table 5), we created another Bio-DSRE corpus using transductive train, validation, and test triples. The corpus generated differs from the inductive one, but it can offer insights into the model's ability to handle seen (*transductive*) and unseen (*inductive*) mentions. As shown in Table 10, the performance using inductive is slightly better than transductive for corpus-level extractions in terms of AUC. However, the F1-macro score is better for transductive. We conclude that the model can learn patterns that exploit mentions and type information to extrapolate to unseen mentions in the inductive setup.

**Does Expert Knowledge Help?** We now consider several pre-trained LMs with different knowledge capacities, specific to biomedical and clinical language understanding, to gain insights about the state-of-the-art encoders' performance and effectiveness on the MedDistant19 benchmark.

We use BERT (Devlin et al., 2019) as baseline. We next consider only those pre-trained models trained with masked language modeling (MLM) objectives using domain-specific corpora. This includes ClinicalBERT (Alsentzer et al., 2019), BlueBERT (Peng et al., 2019), BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), and PubMedBERT (Gu et al., 2021). We categorize these

| Encoder | Knowledge Type | | | | | AUC |
|---|---|---|---|---|---|---|
| | *Biomedical* | *Clinical* | *Type* | *Triples* | *Synonyms* | |
| | NON-EXPERT MODELS | | | | | |
| BERT | | | | | | 0.72 |
| ClinicalBERT | ✓ | ✓ | | | | 0.73 |
| BlueBERT | ✓ | | | | | 0.78 |
| SciBERT | ✓ | | | | | 0.78 |
| BioBERT | ✓ | | | | | 0.79 |
| PubMedBERT | ✓ | | | | | **0.80** |
| | EXPERT KNOWLEDGE MODELS | | | | | |
| MedType | ✓ | | ✓ | | | 0.77 |
| KeBioLM | ✓ | | | ✓ | | **0.80** |
| UmlsBERT | ✓ | ✓ | ✓ | | | 0.75 |
| SapBERT | ✓ | | | | ✓ | 0.78 |

Table 11: Expert and non-expert pre-trained language models performance on MEDDISTANT19.

models as non-experts.

Secondly, we consider expert models that 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 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 11 shows the results of these sentence encoders fine-tuned on the MEDDISTANT19 dataset at sentence-level with AVG pooling. Without domain-specific knowledge, BERT performs slightly worse than the lowest-performing biomedical model, highlighting the presence of shallow heuristics in the data 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 improvement, consistent with Gu et al. (2021).

For expert knowledge-based models, we noted a negative impact on performance. While we would expect type-based models, MedType and UmlsBERT, to bring improvement, their effect can be attributed to overfitting certain types and patterns. KeBioLM, initialized with PubMedBERT, has the same performance despite seeing the triples used in MEDDISTANT19 during pre-training, highlighting the difficulty of the Bio-DSRE. SapBERT, which uses the knowledge of synonyms, also hurt PubMedBERT's performance, suggesting that while synonyms can help in entity linking, RE is a more challenging task in noisy real-world scenarios.

## 5 Discussion

In the biomedical domain, health experts are often concerned with a particular type of interaction, for example, drug-target and gene-disease. However, the number of ontologies is constantly growing (222 in UMLS2019AB), thus a growing need for a more general purpose relation extraction benchmark. Broad-coverage benchmarks exist for biomedical entity linking, such as MedMentions (Mohan and Li, 2018), but they still lack many important concepts involved in relational learning. The research community has come up with several RE benchmarks (see Table 1), but the challenge remains as new entities, and relations emerge with the constant growth of biomedical literature. Hence, constructing a broad benchmark for biomedical RE is challenging due to domain requirements; nonetheless, having an accurate benchmark could offer a utility for future research. We supplement this discussion with Appendix D for a note on limitations.

Further, the train-test overlap highlights the need to systematically assess the proposed benchmarks for inconsistencies that can overestimate the model performance. Similar assessments have shown up in QA generalization where train-test overlap inflates the model performance (Liu et al., 2022). Related to RE generalization, Rosenman et al. (2020) exposed shallow heuristics while Taillé et al. (2021) showed that neural RE models could retain triples, primarily due to type hints. MEDDISTANT19 partially addresses these issues by an inductive setup that can offer insights into the generalization trend in biomedical RE using unseen entities.

## 6 Conclusion

In this work, we highlighted a need for an accurate broad-coverage benchmark for Bio-DSRE. We bridged this gap by utilizing SNOMED CT for constructing the benchmark and laying out the best practices. We thoroughly evaluated the benchmark with baselines and state-of-the-art, showing there is room to conduct further research.

## Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful feedback and William Hogan for assistance in providing data and code for AMIL experiments. The work was partially funded by the European Union (EU) Horizon 2020 research and innovation program through the projects Precise4Q (777107) and Clarify (875160), and the German Federal Ministry of Education and Research (BMBF) through the projects CoRA4NLP (01IW20010) and XAINES (01IW20005). The authors also acknowledge the computing resources provided by the DFKI and UCL.

## 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 domains 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 a thorough analysis to distill such information since it is in the public domain.

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

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 Günter Neumann. 2020a. 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.

Saadullah Amin and Günter Neumann. 2021. T2NER: Transformers based Transfer Learning Framework for Named Entity Recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 212–220. Association for Computational Linguistics.

Saadullah Amin, Günter Neumann, Katherine Ann Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. 2019. MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In *Conference and Labs of the Evaluation Forum (Working Notes)*, pages 1–15.

Saadullah Amin, Stalin Varanasi, Katherine Ann Dunfield, and Günter Neumann. 2020b. LowFER: Low-rank Bilinear Pooling for Link Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 257–268. PMLR.

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,*

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

*Lake Tahoe, Nevada, United States*, pages 2787–2795.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I. Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1):1–17.

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.

The Gene Ontology Consortium. 2018. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, page 77–86. AAAI Press.

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*, WWW '21, page 798–808, New York, NY, USA. Association for Computing Machinery.

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*, 3(1).

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*.

Lixiang Hong, Jinjian Lin, Shuya Li, Fangping Wan, Hui Yang, Tao Jiang, Dan Zhao, and Jianyang Zeng. 2020. A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. *Nature Machine Intelligence*, 2:347–355.

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):486.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Dongwook Shin. 2020. Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinformatics*, 21(1):188.

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. Baw068.

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

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.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2022. Challenges in generalization in open domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2014–2029, Seattle, United States. 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.

Stefano Marchesin and Gianmaria Silvello. 2022. TBGA: a large-scale gene-disease association dataset for biomedical relation extraction. *BMC Bioinformatics*, 23(1):1–16.

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.

Sunil Mohan and Donghui Li. 2018. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *Automated Knowledge Base Construction (AKBC)*.

Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. 2011. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4):441–448.

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, 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.

Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I. Furlong. 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855.

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, Berlin, Heidelberg. Springer, Springer Berlin Heidelberg.

Roland Roller and Mark Stevenson. 2014. Self-supervised relation extraction using UMLS. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 116–127. Springer.

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

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. *CEUR Workshop Proceedings 761*.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Separating retention from extraction in the evaluation of end-to-end Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10438–10449, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. 2012. The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884.

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.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

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.

Rui Xing, Jie Luo, and Tengwei Song. 2020. BioRel: towards large-scale biomedical relation extraction. *BMC Bioinformatics*, 21-S(16):543.

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.

# A UMLS

This section presents 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 (Xing et al., 2020; Amin et al., 2020a; Hogan et al., 2021) fail to account the inverse relations, and with naive split, it 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). We used UMLS2019AB to be consistent with the prior works.

## 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.3.

Note that we only consider relations belonging to the *RO* (*has a 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.

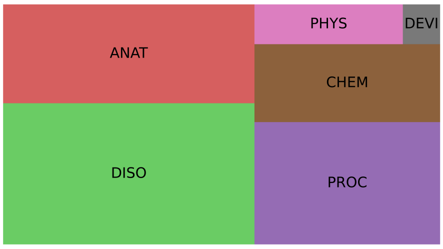


Figure A.1: Relative proportions of the entities present in MedDistant19, based on the semantic groups.

## 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 example, in the 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.

Further, we used an inductive split set with 70, 10, and 20 proportions of train, validation, and test splits for constructing MedDistant19. Below is an example instance from the dataset 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"
}

/-----------------------------/
```

[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

```
{
  "text": "Severe heart disease
  may result in cardiac cirrhosis
  in the elderly , with ascites
  and hepatomegaly .",
  "h": {
    "id": "C0018799",
    "pos": [7, 20],
    "name": "heart disease"
  },
  "t": {
    "id": "C0085699",
    "pos": [35, 52],
    "name": "cardiac cirrhosis"
  },
  "relation": "cause_of"
}

/------------------------------/

{
  "text": "Complications
  closely associated to the
  osteosynthesis appeared
  only in instable
  fractures ( 7 % ) .",
  "h": {
    "id": "C0016658",
    "pos": [81, 90],
    "name": "fractures"
  },
  "t": {
    "id": "C0016642",
    "pos": [40, 54],
    "name": "osteosynthesis"
   },
   "relation":
   "direct_morphology_of"
}

/------------------------------/

{
  "text": "Gluten proteins ,
  the culprits in celiac
  disease ( CD ) , show
  striking similarities in
  primary structure with
  human salivary proline-rich
  proteins ( PRPs ) .",
  "h": {
    "id": "C2362561",
    "pos": [0, 15],
    "name": "Gluten proteins"
  },
  "t": {
    "id": "C0007570",
    "pos": [34, 48],
    "name": "celiac disease"
  },
  "relation":
  "causative_agent_of"
}

/------------------------------/

{
  "text": "Postherpetic
  neuralgia is an unfortunate
  aftermath of shingles ,
  and is most likely to
  develop , and most
  persistent , in elderly
  patients .",
  "h": {
    "id": "C0032768",
    "pos": [0, 22],
    "name": "Postherpetic
    neuralgia"
  },
  "t": {
    "id": "C0019360",
    "pos": [54, 62],
    "name": "shingles"
  },
  "relation": "occurs_after"
}
```

## B UMLS License Agreement

To use the MEDDISTANT19 benchmark, 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 acknowledging that the UMLS contains copyrighted material and that those restrictions are respected. The UMLS agreement requires users to agree to obtain agreements for *each* copyrighted source before its use within a commercial or production application.

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

## C Risks

While our work does not have direct risk, we 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 the hope to accelerate reproducible research in Bio-DSRE by having ready-to-use corpora, with only the condition that the user has obtained the license. We provide users with this note and hope this will be respected. However, there is a risk that users may download the data and redistribute it 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 a UMLS API key, which will be validated, and only then will the data be allowed to be downloaded.

## D 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, drug side-effects, and relations involving genes as provided by RxNorm (Nelson et al., 2011), Gene Ontology (Consortium, 2018), etc.

MEDDISTANT19 is automatically-created and susceptible to noise and thus needs to be approached carefully as a potential source for biomedical knowledge. While the dataset was not created to represent *true* biomedical knowledge, it has the potential to be treated as a reliable reference.

## E Experimental Setup and Hyperparameters

We followed the experimental setup of Gao et al. (2021) for BERT-based experiments. Specifically, we used batch size 64, with a learning rate of 2e-5, maximum sequence length 128, and bag size 4. We used a single NVIDIA Tesla V100-32GB for BERT-based experiments. Each experiment took about 1.5hrs, with half an hour per epoch. We also attempted to perform a 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.

| Encoder | Bag Size | Batch Size | Embedding |
|---:|:---:|:---:|:---:|
| CNN+sent+AVG | - | 128 | biowordvec |
| CNN+sent+ONE | - | 128 | biowordvec |
| CNN+bag+AVG | 8 | 128 | GloVe |
| CNN+bag+ONE | 16 | 256 | GloVe |
| CNN+bag+ATT | 8 | 256 | GloVe |
| PCNN+sent+AVG | - | 128 | biowordvec |
| PCNN+sent+ONE | - | 128 | biowordvec |
| PCNN+bag+AVG | 4 | 128 | GloVe |
| PCNN+bag+ONE | 8 | 128 | GloVe |
| PCNN+bag+ATT | 8 | 128 | GloVe |
| GRU+sent+AVG | - | 128 | biowordvec |
| GRU+sent+ONE | - | 128 | biowordvec |
| GRU+bag+AVG | 8 | 128 | biow2v |
| GRU+bag+ONE | 16 | 256 | GloVe |
| GRU+bag+ATT | 16 | 128 | GloVe |

Table A.1: Best hyperparameters for CNN, PCNN, and GRU sentence encoders.

For CNN and PCNN, we performed grid search with Adam (Kingma and Ba, 2015) optimizer using learning rate 0.001 for 20 epochs with: batch size $\in \{128, 256\}$, bag size $\in \{4, 8, 16, 32\}$, 200-d word embeddings $\in$ {Word2Vec (Mikolov et al., 2013)[13], GloVe (Pennington et al., 2014)}, and with (test-time) pooling $\in \{\text{ONE}, \text{AVG}\}$ when using sentence-level training and pooling in $\{\text{ONE}, \text{AVG}, \text{ATT}\}$ 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). The hyperparameters that had the most influence were batch size, bag size, and pre-trained word embeddings. All the experiments reported in this work are with a single run.

For sentence tokenization with ScispaCy, it took 9hrs with 32 CPUs (4GB each) and a batch size of 1024 to extract 151M sentences. Further, the ScispaCy entity linking job took about half TB of RAM with 72 CPUs (6GB each) with a batch size of 4096 with 40hrs of run-time to link 145M unique sentences.

[13] We used domain-specific word embeddings *biowordvec* and *biow2v* following Marchesin and Silvello (2022).

| **Semantic Type** | **10k-20k** | **20k-30k** | **$\geq$ 30k** |
|---|---|---|---|
| Body Part, Organ, or Organ Component | *bladder, heart, retinal, lungs, spinal, kidneys, colon* | *eyes, lung, kidney, intestinal* | *liver, brain* |
| Organism Function | *death* | *period, blood pressure* | - |
| Body Location or Region | *head* | - | - |
| Therapeutic or Preventive Procedure | *injection, prevention, chemotherapy, application*<br>*resection, infusion, treatments, therapeutic*<br>*surgical treatment, CT, surgical, transplantation* | *stimulation, delivery* | *intervention, procedure, removal, operation* |
| Neoplastic Process | *cancer* | - | *tumor, tumors* |
| Disease or Syndrome | *obesity, disorder, disorders* | *diseases, stroke* | *disease, infection, condition, hypertension* |
| Laboratory Procedure | *test, erythrocytes* | - | *cells* |
| Diagnostic Procedure | *US, biopsy, ultrasound* | *MRI* | - |
| Finding | *lesion, interaction, mass, difficulty, dependent* | *abnormal* | *presence, positive, negative, severe, lesions* |
| Hormone | *insulin* | - | - |
| Biologically Active Substance | *amino acids, glucose, ATP* | *protein, proteins* | |
| Pharmacologic Substance | *medication* | - | *drugs, drug* |
| Injury or Poisoning | *strains* | *injury, exposure* | *damage* |
| Tissue | *tissue, bone marrow, tissues* | - | - |
| Organism Attribute | *male* | - | *temperature, age* |
| Immunologic Factor | *antibody, antibodies* | - | - |
| Health Care Activity | *investigations* | *examination* | *assessment* |
| Body Substance | *plasma, blood, skin* | - | - |
| Body System | - | *cardiovascular* | - |
| Mental Process | - | - | *concentrations, concentration* |
| Congenital Abnormality | - | *abnormalities* | - |

Table A.2: Semantic types affected by type-based mention pruning with removed mentions placed in their respective frequency bins as discussed in Section 3.1.

| **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* |
| *interpretation_of* | *has_interpretation* |
| *component_of* | *has_component* |
| *indirect_procedure_site_of* | *has_indirect_procedure_site* |
| *direct_morphology_of* | *has_direct_morphology* |
| *cause_of* | *due_to* |
| *direct_substance_of* | *has_direct_substance* |
| *uses_device* | *device_used_by* |
| *focus_of* | *has_focus* |
| *direct_device_of* | *has_direct_device* |
| *procedure_site_of* | *has_procedure_site* |
| *uses_substance* | *substance_used_by* |
| *associated_finding_of* | *has_associated_finding* |
| *occurs_after* | *occurs_before* |
| *is_modification_of* | *has_modification* |

Table A.3: *(Left)* 21 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 |
| 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 |
| 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.4: 51 semantic types (STY) along with their TUIs and semantic groups (SG) covered in MEDDISTANT19.