Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1073–1082
Marseille, 20-25 June 2022

A Distant Supervision Corpus for Extracting Biomedical Relationships Between Chemicals, Diseases and Genes

Dongxu Zhang[*,1], **Sunil Mohan**[*,2], **Michaela Torkar**[2], **Andrew McCallum**[1]
[1] University of Massachusetts, Amherst, Massachusetts, USA
[2] Chan Zuckerberg Initiative, Redwood City, California, USA
{dongxuzhang, mccallum}@cs.umass.edu
smohan@chanzuckerberg.com, michaela.torkar@contractor.chanzuckerberg.com

Abstract

We introduce *ChemDisGene*, a new dataset for training and evaluating multi-class multi-label document-level biomedical relation extraction models. Our dataset contains 80k biomedical research abstracts labeled with mentions of chemicals, diseases, and genes, portions of which human experts labeled with 18 types of biomedical relationships between these entities (intended for evaluation), and the remainder of which (intended for training) has been distantly labeled via the CTD database with approximately 78% accuracy. In comparison to similar preexisting datasets, ours is both substantially larger and cleaner; it also includes annotations linking mentions to their entities. We also provide three baseline deep neural network relation extraction models trained and evaluated on our new dataset.

Keywords: Corpus, Information Extraction, Linked Data, Weakly-supervised Learning, Relation Extraction

1. Introduction

Biomedical researchers have used systems of experimentally confirmed interactions between chemicals, diseases, genes/proteins and other entities, for understanding disease mechanisms for diagnosis, e.g. Lee et al. (2019), for drug repurposing (Morselli Gysi et al., 2021), and even for understanding the health hazards associated with spaceflight (Nelson et al., 2021). These knowledge graphs (KGs) are often built by integrating manually curated databases like CTD[1] and DrugBank[2], who use domain experts to extract observed interactions from research publications and other sources. While the information in these databases is high in *precision*, with the growing publication rate their *recall* is low (Baumgartner et al., 2007). To improve coverage, researchers have resorted to automated mining of biomedical interactions from research texts, to supplement their KG (Himmelstein et al., 2017), or even to build the entire KG, e.g. (Crichton et al., 2020).

The bioinformatics community recognized that machine learning Relation Extraction (RE) models could help the manual curation task, and the BioCreative workshops introduced the first shared task and manually labeled 'gold standard' dataset for training and evaluating models for extracting protein-protein interactions from full text articles in 2006 (Krallinger et al., 2006). Several such labeled corpora have followed, primarily focusing on extracting relationships from abstracts. However, labeling of relationships requires domain experts and is slow and expensive. Consequently, most labeled corpora are small, and focus on a small number of entity types and relationships.

In this paper, we introduce *ChemDisGene*[3], a new dataset of biomedical research abstracts labeled with pairwise interactions between Chemicals, Diseases and Genes/Gene-products. It contains two sub-corpora:

- A large corpus of $\sim 80k$ abstracts with distant labeling of 14 relation types. This corpus is automatically derived from CTD (Davis et al., 2020), thus allowing for a larger size more suitable for training deep learning models. However, relationships are distantly labeled because relationships in CTD are associated only with a paper, and not with a specific text passage within the paper.
- A smaller corpus of 523 abstracts, manually annotated with relationships by domain experts. This corpus is aimed primarily for testing models trained on the CTD-derived corpus, and the relationships here are also distantly labeled.

A previous version of the CTD-derived corpus was introduced in (Verga et al., 2018). *ChemDisGene* adds a manually annotated component, and includes several improvements to the derivation process:

- More recent updates (2021 February) from CTD.
- Entity linking uses PubTator Central (Wei et al., 2019) with significantly improved models for recognizing Chemicals ($+66.3\%$ improvement in F1 score), Diseases ($+3.8\%$) and Genes/Proteins ($+8.2\%$) over the previous PubTator model.
- The previous dataset was randomly split into training, dev and test, while in *ChemDisGene* these splits are based on paper publication date, to better simulate a real world scenario.
- A cleaner extraction of binary relationships from complex nested relationships captured by CTD.

The rest of this paper describes how the labeled cor-

* Equal contribution
[1]CTD: Comparative Toxicogenomics Database
[2]https://go.drugbank.com
[3]https://github.com/chanzuckerberg/ChemDisGene

pus was developed (§2), corpus statistics (§3), baseline models trained and evaluated on the *ChemDisGene* (§4), and related work (§5).

2. Methodology

A note on terminology: we will use *relation* to refer to the predicate schema $r(T_s, T_o)$, where r is the *relation type*, and T_s, T_o are the argument entity types: *Chemical*, *Disease* or *Gene*. A *relationship* is a ground instance $r(e_s, e_o)$ of a *relation*, with argument entities $e_s \in T_s, e_o \in T_o$.

The *ChemDisGene* dataset comprises a large corpus automatically derived from CTD, and a smaller curated corpus manually labeled by domain experts.

2.1. Derivation from CTD

Comparative Toxicogenomics Database (CTD) is a public knowledge base containing manually curated interactions between chemicals, genes, diseases and phenotypes (Davis et al., 2020). CTD curators regularly scan new research publications to identify those interactions that are the primary contributions of each paper (Davis et al., 2011). These are then encoded using a hierarchical ontology of $\sim$ 50 Chemical–Gene interaction classes, and two types each for Chemical–Disease and Gene–Disease interactions (phenotypes are not covered in our dataset). Each interaction is expressed using relation types from these interaction classes, along with the argument entities, and recorded with a reference to the paper from which it was extracted (but no reference to any text within the paper). Entities are also identified using public ontologies: MeSH for Chemicals, MeSH and OMIM for Diseases, and NCBI Gene for Genes and Gene-products[4].

While CTD curators scan full papers to extract these relationships, we limited the text in *ChemDisGene* to only the title and abstract. Starting with the February 2021 dump of CTD, we obtained abstracts for all referenced articles from PubMed[5]. Each abstract was processed through PubTator Central[6] (PTC) to identify and link mentions of chemical, disease and gene/gene-product entities. We then performed a 'distant alignment' of the annotated abstracts with the relationships linked to each paper in CTD: relationships whose entities were not detected in the abstract were discarded. This yielded a dataset of abstracts with linked entity mentions, and distantly linked relationships.

This distant linking of relationships to aligned abstracts is noisy due to the following sources of error: (i) entity recognition models in PTC, whose F1 scores for each entity type are in the range 0.84–0.90 (Wei et al., 2019), (ii) even if the entities of a relationship are correctly identified in the abstract, the corresponding interaction may not have been mentioned in the abstract text, and (iii) an abstract may mention some relationships that are not extracted by CTD. To measure these sources of error, we selected a subset of aligned abstracts for manual curation (see §2.2, §3.2).

Relations in CTD are organized into a class hierarchy, with some relation classes qualified by a 'degree'. *ChemDisGene* includes 10 of these classes[7], which combined with the degrees defines 18 relation types:

- *Chemical-Disease*:
  - *marker/mechanism*: A chemical that correlates with a disease.
  - *therapeutic*: A chemical that has a known or potential therapeutic role in a disease.
- *Chemical-Gene*: Each qualified by a *degree*.
  - *activity*: An elemental function of a molecule. Degrees: *increases*, *decreases*, or *affects* when the direction is not indicated.
  - *binding*: A molecular interaction (*affects*).
  - *expression*: Expression of a gene product (*increases*, *decreases*, *affects*).
  - *localization*: Part of the cell where a molecule resides (*affects*).
  - *metabolic_processing*: The biochemical alteration of a molecule's structure, not including changes in expression, stability, folding, localization, splicing, or transport (*increases*, *decreases*, *affects*).
  - *transport*: The movement of a molecule into or out of a cell (*increases*, *decreases*, *affects*).
- *Gene-Disease*:
  - *marker/mechanism*: A gene that may be a biomarker of a disease or play a role in the etiology of a disease.
  - *therapeutic*: A gene that is or may be a therapeutic target in the treatment of a disease.

In some cases, CTD defines a finer granularity of Chemical-Gene interactions. Because their occurrence is rare, they would be harder for a model to recognize, so we abstracted them to the levels described above.

The relationships in CTD also include complex and nested biomedical interactions involving multiple entities. For *ChemDisGene* we only extracted binary relationships. In particular, (a) we omitted CTD's "*cotreatment*" relation type because it is non-binary, and (b) we implemented a cleaner extraction of binary relationships from nested interactions (see example in fig. 1).

The previous CTD-derived dataset in (Verga et al., 2018) used the same relation types for Chemical-Disease and Gene-Disease interactions, but a different set of 10 relation types for Chemical-Gene. With three years of new publications, the distribution of relation types in CTD has changed, affecting our selection.

The derivation of relationships from CTD in (Verga et al., 2018) did not take into account nesting levels in complex interactions: in the example in fig. 1, the previous dataset would also extract *reaction-decreases* between the chemical '24-hydroxycholesterol' and the

[4]Links: MeSH, OMIM, NCBI Gene

[5]https://pubmed.ncbi.nlm.nih.gov

[6]https://www.ncbi.nlm.nih.gov/research/pubtator/

[7]Definitions are from the CTD glossary.

Complex nested interaction in CTD:

Quercetin$_{\text{Disease}}$ *inhibits the reaction*
[[24-hydroxycholesterol$_{\text{Chemical}}$ *co-treated with*
27-hydroxycholesterol$_{\text{Chemical}}$ *co-treated with*
cholest-5-en-3 beta,7 alpha-diol$_{\text{Chemical}}$]
results in increased expression of ITGB1$_{\text{Gene}}$ mRNA]

Extracted binary relationships:

expression-increases(24-hydroxycholesterol, ITGB1)
expression-increases(27-hydroxycholesterol, ITGB1)
expression-increases(cholest-5-en-3 beta,7 alpha-diol, ITGB1)

Figure 1: An example showing extraction of binary relationships from a complex nested interaction in CTD.

gene 'ITGB1', even though the corresponding indicator "*inhibits the reaction*" is at a different nesting level. As a final step, we added some randomly sampled abstracts that did not align with any CTD relationships as 'null' documents with no relationships. This forms 10% of the CTD-derived corpus, which was then split into *train*, *development* (*dev*) and *test* sets by publication year (2018 as dev and years 2019, 2020 as test).

2.2. Curation

As described above, the relationship labels in the CTD-derived corpus are noisy. To perform more reliable testing of RE models, we selected some documents for manual annotation: 303 sampled from the *test* split, and an additional 252 documents from CTD that were also included in the DrugProt corpus (Martin Krallinger and Valencia, 2021), to enable future comparative analyses. These were distributed for annotation by five biologists, each document assigned randomly to three curators.

We developed a web-based annotation tool which displayed for each document, the title and abstract, all the linked chemicals, diseases and genes/gene-products, and their mentions in the text, and all the relationships derived from CTD. Annotating a document involved two tasks: (i) review each relationship derived from CTD, and either reject or approve it, and (ii) add all other established relationships expressed in the document. Relationships mentioned in the abstract without any conclusions were excluded from annotation. In keeping with our goal of a realistic dataset, 44 of these documents had no CTD-derived relationships.

We developed annotation guidelines (published with the dataset) that describe the steps in the curation process and the types of pairwise interactions curated in this dataset, including brief definitions and real-world example statements that do or do not support a specific relation type. These guidelines underwent multiple rounds of revisions through 4 iterations of practice annotations. During the practice phase, all 5 curators were given the same set of documents to curate (15–30 per cycle); annotation disagreements and questions were clarified during multiple workshops, and feedback and suggestions from the curators were used to improve the guidelines. See fig. 2 for some examples.

A: *Don't record investigated or motivating relationships that remain unknown and hypothetical.*

"Gene A is a therapeutic target for treatment of Disease X; it may therefore have a potential role in treatment of Disease Z."

Record a relationship between Gene A and Disease X; but not between Gene A and Disease Z.

B: *Inferring a relationship across sentences.*

"We have previously identified a panel of fusion genes in aggressive prostate cancers. In this study, we showed that . . . CCNH-C5orf30 and TRMT11-GRIK2 gene fusions were found in breast cancer, colon cancer, . . . "

Record a 'Gene-Disease: marker/mechanism' *relationship between C5orf30 and prostate cancers.*

Figure 2: Two examples from the curation guidelines. Colors identify disease and gene mentions.

Some interactions were easy to identify, like *Chemical-disease: marker/mechanism*, and were labeled with high consistency. Other relation types required more interpretation and created more disagreement; e.g., the upregulation of a gene product by a chemical can be described by the types "*expression*" or "*activity*", depending on the context. A number of edge cases were identified during the practice phase and added to the guidelines, such as how to record opposite effects, e.g. when both an increase and a decrease in expression of a gene product is mentioned under different experimental conditions, or how to label relationships between two entities that depend on the presence of a third entity.

Only entities correctly detected by PubTator Central and linked to the right ontology record were considered in an interaction pair; the annotation guidelines therefore also included instructions for accepting or rejecting detected entities that did not unambiguously match the text mention, such as a detected entity that is broader than the mention in the abstract.

At the end of the annotation period, we observed that 30 of the documents had been annotated by only two biologists. We also observed that a number of new relationships added to each document had only been added by one annotator. This was not unexpected, as scanning through text to identify all relationships is much harder than verifying whether a specified relationship occurs. We refer to these relationships as '*singletons*', and marked as '*approved*' all relationships that a majority of the annotators had approved. From the documents annotated by two curators, we also added all relationships derived from CTD that were approved by only one of the two annotators to the list of singletons. We then discarded documents with more than 10 singletons, while keeping all 252 DrugProt documents,

	Train	Dev	Test
Nbr. abstracts	76,942	1,521	1,939
. . . with no relationships	7,244	397	436
Nbr. relationships	167,005	3,290	5,116
. . . unique relationships	93,801	3,127	4,801
Total Entity mentions	1,532,117	36,114	49,839
Chemicals	686,102	13,986	19,895
Diseases	478,397	8,962	11,750
Genes	367,618	13,166	18,194
Unique Entities in relns.	14,991	1,894	2,345
Chemicals	7,187	759	999
Diseases	2,413	283	287
Genes	5,391	852	1,059

Table 1: General statistics for the CTD-derived corpus.

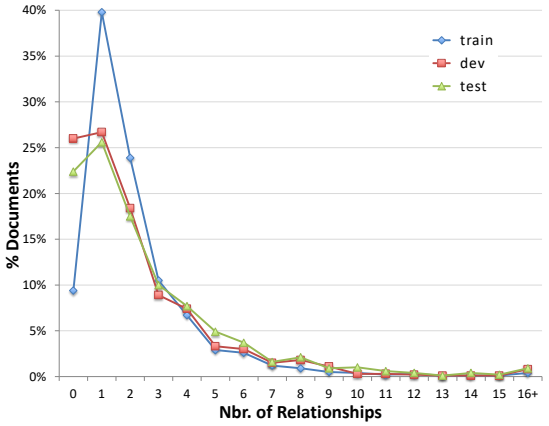

Figure 3: Relationships per doc., CTD-derived corpus.

yielding a total of 523 annotated documents.

On analyzing the singletons, we noticed that some of these differed only in degree from an approved relationship in the same document: 45 were abstractions (degree *affects*) and 7 refinements (degree *increases* or *decreases*) of an approved relationship. These singletons were then automatically rejected.

In this annotation task, when one annotator does not identify a particular relationship that was found by another, it could be for one of two reasons: (i) both annotators noticed the same text passage but disagreed on whether it expressed the relationship, or (ii) the first annotator did not notice the passage that the second annotator used to identify the relationship. To resolve this ambiguity, all the singleton relationships were reviewed by an annotator not originally assigned to that document, followed by a second review by the curation manager to ensure consistency. Relationships approved in this phase were added to the curated data.

3. *ChemDisGene* Corpus Statistics

3.1. The CTD-derived Corpus

The median of number of tokens (split by space) in abstracts is 214. And about 99.8% of abstracts have less than 512 tokens. Other statistics for the CTD-derived corpus are shown in table 1, and the distribution of the number of relationships per document in fig. 3. About 80% of the documents have 3 or fewer relationships, followed by a long thin tail. The *dev* and *test* splits have a higher proportion of documents with no relationships. There are an average of 2.2 relationships per document, with over 9,000 entity pair occurrences with multiple relation type labels in the same document.

Counts for each relation type are shown in table 2. Unique numbers count unique argument-entity pairs. Four Chemical-Gene relation types (*activity-affects*, *metabolic_processing-affects*, *transport-affects*, and *transport-increases*) were omitted from the CTD-derived corpus because of their low incidence. However they are included in the curated corpus for completeness, making the annotation task a little easier.

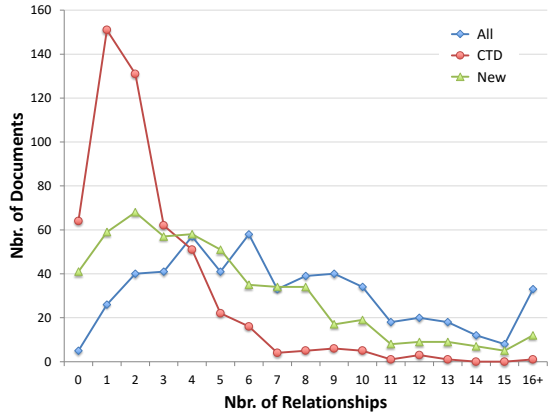

Figure 4: Distribution of the nbr. of approved relationships per document in the Curated corpus.

3.2. The Curated Corpus

The curated corpus contains 523 documents: 271 from CTD-derived's *test* split, and an additional 252 documents taken from DrugProt, that are not in the CTD-derived corpus. Twenty seven of these documents had no relationships derived from CTD. Manual annotation rejected 22% of all CTD-derived relationships, leaving 64 documents with no approved CTD-derived relationships. This indicates a fairly high 78% confidence in the automatically derived relationships.

There are a total of 1,279 approved CTD-derived relationships (avg. 2.4/doc), 2,632 approved new relationships (5.0/doc). The distribution of the 18 types of relations in this corpus is shown in table 3.

The 3,911 approved relationships (3,806 unique) in the curated corpus involve 1,875 unique entities: 670 unique Chemicals, 318 Diseases and 887 Genes. Figure 4 shows the distribution of number of approved relationships in each document. As expected, the CTD-derived approved relationships are more skewed to the left than the new added relationships.

#	Relation type	*Total*			*Unique*		
		Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
1	Chemical-Disease : marker/mechanism	66,155	559	754	27,706	486	602
2	Chemical-Disease : therapeutic	34,775	250	410	16,093	245	398
3	Chemical-Gene : activity - decreases	5,555	101	232	4,128	97	232
4	Chemical-Gene : activity - increases	6,152	127	174	4,133	120	157
5	Chemical-Gene : binding - affects	3,123	67	77	2,024	65	73
6	Chemical-Gene : expression - affects	1,247	51	160	1,206	51	158
7	Chemical-Gene : expression - decreases	10,204	480	923	8,487	467	905
8	Chemical-Gene : expression - increases	19,810	919	1,570	14,685	878	1,491
9	Chemical-Gene : localization - affects	1,448	50	73	1,216	50	70
10	Chemical-Gene : metabolic_processing - decreases	1,653	101	116	1,313	100	111
11	Chemical-Gene : metabolic_processing - increases	4,640	175	293	3,507	172	283
12	Chemical-Gene : transport - increases	1,962	92	108	1,405	88	96
13	Gene-Disease : marker/mechanism	9,388	301	219	7,384	292	218
14	Gene-Disease : therapeutic	893	17	7	514	16	7

Table 2: Nbr. of relationships (instances) for each relation type in the CTD-derived corpus.

#	Relation type	*Distribution* (%)	
		Approved, New	**Approved, CTD**
1	Chemical-Disease : marker/mechanism	16.6	16.4
2	Chemical-Disease : therapeutic	10.4	12.0
3	Chemical-Gene : activity - affects	1.2	
4	Chemical-Gene : activity - decreases	8.3	7.3
5	Chemical-Gene : activity - increases	8.7	7.8
6	Chemical-Gene : binding - affects	4.3	6.7
7	Chemical-Gene : expression - affects	2.8	0.6
8	Chemical-Gene : expression - decreases	10.4	13.1
9	Chemical-Gene : expression - increases	11.8	18.4
10	Chemical-Gene : localization - affects	0.8	1.5
11	Chemical-Gene : metabolic_processing - affects	0.8	
12	Chemical-Gene : metabolic_processing - decreases	1.7	1.5
13	Chemical-Gene : metabolic_processing - increases	3.0	4.0
14	Chemical-Gene : transport - affects	0.3	
15	Chemical-Gene : transport - decreases	0.6	
16	Chemical-Gene : transport - increases	1.1	0.9
17	Gene-Disease : marker/mechanism	14.1	9.3
18	Gene-Disease : therapeutic	2.9	0.5

Table 3: Frequency distribution of relation types in curated corpus (each column sums to 100%). Empty frequencies indicate some relations are rare in CTD.

3.3. Inter-Annotator Agreement

Commonly used measures of inter-annotator agreement are defined for tasks where the units being classified or measured are precisely specified. As noted in (Kilicoglu et al., 2011), identifying all relationships expressed in a text does not match this paradigm. This task could be decomposed into the following steps: (i) find relationship indicators in the text, (ii) identify the entity mentions each indicator refers to, and (iii) map the expressed relationship to the appropriate ontological term. Here the space of possible annotations is clearly defined only for step (iii). In step (ii) the space would be clearly specified only if we presented the annotators with every pair of linked mentions. The set of possible relationship indicators in a document, in step (i), is also not presented to the annotators. When a relationship is identified by only one of two curators reviewing the same text, it could be because either the first one did not 'notice' the same sentence, or actually saw it and rejected it. This inherent ambiguity causes a problem even for measures that allow varying number of annotations per unit.

Similar to (Kilicoglu et al., 2011), we evaluated each curator's annotations against a reference, using precision, recall and F1 scores, as more feasible and intuitively understandable metrics for our use case. We used the 'majority approved' relationships (§2.2) as the reference dataset. The annotator agreement metrics (ta-

Relationships	A	B	C	D	E
All	0.85	0.84	0.83	0.88	0.88
CTD-derived only	0.99	0.96	0.97	0.99	0.96
New only	0.76	0.77	0.69	0.82	0.83

Table 4: Agreement F1 scores for the 5 annotators (A-E) against the 'approved' reference subset.

ble 4) are fairly high, indicating a high confidence in the approved subset. As expected, agreement levels on prompted relationships (those from CTD) in the annotator UI is higher than for relationships that the annotator has to find and add (new relationships).

4. The Relationship Extraction Task

4.1. Task definition

The document-level relation extraction (RE) task in *ChemDisGene* is to identify all relationships $r(e_s, e_o)$ expressed in a document, comprised of the title and abstract texts, that are the primary contributions of that article. We consider 14 binary relation types (from the CTD-derived corpus) among chemical, disease and gene/gene-product entities. All mentions of these entities in the text are identified and linked to the corresponding ontologies. This is a *distant supervision* (relationships are associated with documents, but not specific entity mentions) *multi-label* (a document, and a pair of entities, may have more than one relationship) classification task. For evaluation, we use Micro/Macro F1 scores where per-relation thresholds are tuned on the dev set, and average precision where thresholds are not required.

4.2. Models

In our experiments, we trained and evaluated three baseline methods on *ChemDisGene*.

BRAN Bi-affine relation attention networks (Verga et al., 2018) is one of the first papers to tackle document level (distant supervision) relation extraction in the biomedical domain. It uses multiple self-attention + convolutional neural network (NN) layers to encode the text input, then leverages per-relation biaffine transformation to calculate mention level scores of the query $r(e_s, e_o)$, and a $logsumexp$ layer to capture the most significant signal among mention pairs. In our experiment, we omitted BRAN's NER joint loss in order to analyze its core RE module.

PubmedBert (Gu et al., 2021) is a BERT-based pretrained language model (Devlin et al., 2019) trained from scratch on PubMed abstracts. For relation extraction, we first get each entity's embeddings by max-pooling over PubmedBert's encoding of all the entity's mentions. Then concatenated embeddings of candidate argument entity pairs are processed through a feed-forward NN to predict scores for each relation type.

PubmedBert + BRAN. This model combines the stronger text encoder of PubmedBert with the relation detection layers of BRAN. The model structure is: *Input* → *PubMedBertEncoder* → *Biaffine* → *logsumexp* → *logits*.

4.3. Empirical Results

Table 5 shows overall performance of the three baseline models on *ChemDisGene*[8]. Performance metrics are shown for the test split of the CTD-derived corpus, and separate metrics on the curated corpus for approved relationships derived from CTD, and for all approved relationships, which also includes new relationships added by the curators.

From the table, our best model PubmedBert + BRAN has 43.8 Micro F1 and 50.6 average precision on the '*all relationships*' curated test set, indicating the difficulty of this task. The pretrained language model adds significant improvement over BRAN. And the biaffine transformation and $logsumexp$ layer are also complementary to the pretrained language model.

Compared with the CTD-derived test set, the performance decreases significantly on the curated test set, indicating the necessity of evaluation on expert-labeled data. We also observe that Macro results are lower than Micro, indicating that performance varies across different relation types. In table 6 we see that relation types with low frequencies in the training data tend to perform poorly. The particularly bad performance of our model on *Chemical-Gene: expression-affects* is also caused by distraction from two similar but common *Chemical-Gene* relation types: *expression-increases* and *expression-decreases*.

Performance of baseline models on 'BRAN' dataset. We also trained and tested the baseline models on the CTD-derived dataset from (Verga et al., 2018), referred to as the 'BRAN' dataset (table 7). As described above (§2.1), there are several differences in this dataset that account for the different performance results from those on *ChemDisGene* (table 5). Perhaps the most important one is that in the BRAN dataset, abstracts from *test* and *dev* splits are randomly selected, whereas in *ChemDisGene* 271 abstracts are assigned based on publication date. The *ChemDisGene* test set also includes more documents with no relationships. While this makes *ChemDisGene* more challenging, it also reflects a more realistic scenario for applying such RE models. The relative order of performance of the three baselines is the same on both datasets.

Comparing the performance on CTD-derived and All relationships in curated corpora. From Table 5 we can see that while model precision increases when tested on *all approved relationships* from the curated corpus, compared to the performance on just *CTD-derived approved relationships*, the recall of all models drops significantly. A main reason is that the training

[8]We trained all three baselines on $ChemDisGene$ training set with hidden dimension 128, and we tuned the hyperparameters such as learning rate [1e-5, 1e-4] and weight decay = [0, 1e-4] over the distant supervision dev set.

data only includes CTD-derived relationships, which are selected by CTD to be the 'primary' contributions of the paper. While this is mostly determined within the context of other publications, there might be a signal in the wording (an area for further investigation).
Curators were asked to reject CTD-derived relationships when the entities involved were incorrectly linked. This probably accounts for the small difference in models' performance between the CTD-derived and curated corpora.

5. Related Work

5.1. Distant Supervision Biomedical Corpora

As described above (see §1,2.1), *ChemDisGene* offers a reworking of the derived corpus introduced in (Verga et al., 2018), focusing on a cleaner derivation from an updated CTD with better entity linking. The number of abstracts also increased by $\sim 20k$.
A well known manually labeled biomedical corpus is BC5-CDR (Li et al., 2016), which identifies a single relation type between Chemicals and Diseases, distantly labeled in 1,500 abstracts. BC6-PM (Islamaj Doğan et al., 2019) is another manually annotated distant supervision corpus, for Protein-Protein interactions. It has a total of 1,232 abstracts, but only one relation type.
The GDA dataset (Wu et al., 2019) takes a similar approach to CTD-derived, to derive a Gene-Disease associations dataset from the DisGeNET[9] database, using PubTator to link entity mentions. Abstracts are distantly labeled with a single relation type.

5.2. Direct Supervision Biomedical Corpora

DrugProt (Miranda et al., 2021), (Martin Krallinger and Valencia, 2021), is the most recent manually annotated corpus of biomedical research abstracts covering multiple (13) relation types between Chemicals and Genes/Gene-products. *ChemDisGene* uses a different set of 14 relation types between Chemicals and Genes, derived from CTD. These relations generally describe the observed effect of an interaction. For example, a *Localization* relation is recorded when the interaction between a Chemical and Gene product affects the part of the cell where the molecule resides. In contrast, the DrugProt relation classes are defined by the specific type of interaction between a Chemical and Gene/Gene-product: they distinguish between 'Direct' and 'Indirect' regulation (where possible), and the subclasses focus on the direction of the interaction ('Upregulator', 'Downregulator'). The subclasses for direct regulation are highly granular, differentiating between 'Activator', 'Agonist', 'Antagonist', etc.
As an example, the relationship expressed in the text
"bisphenol P_{Chemical} showed estrogen receptor$_{\text{Gene}}$ antagonistic activities"
would be annotated as *Chemical-Gene: activity-decreases* in *ChemDisGene*, whereas DrugProt would record it as *Chemical-Gene: antagonist*.

[9]https://www.disgenet.org

DrugProt contains a larger number of abstracts (3500 in training, 750 in dev), with $\sim$5 relationships per abstract. All mentions of Chemicals and Gene-related entities are identified, but not linked. Relationships are *directly supervised* by identifying the actual pair of mentions expressing each relationship.
Our *ChemDisGene* manually curated corpus is smaller, but also includes relationships between Chemicals and Diseases, and Diseases and Genes. All entity mentions are identified and linked by the models in Pubtator Central, and relationships are distantly labeled, associated with a document but not specific entity mentions. The curated corpus contains $\sim$6 approved relationships per abstract, distinguishing between primary contributions (derived from CTD) and other ('new') relationships.
Most other manually annotated corpora used in biomedical RE tasks are also directly supervised, and cover fewer relation types, typically between fewer types of entities. As another example, Drug-drug interaction (DDI) (Herrero-Zazo et al., 2013) specifies 4 relation types among drugs, on sentences extracted from 1025 documents.

5.3. Other RE Corpora

In the general domain, there exist several RE benchmarks for sentence level, document level and few-shot scenarios. SemEval 2010 task 8 (Hendrickx et al., 2010) includes ten semantic relation types between nouns over $\sim 11k$ sentences. The TAC relation extraction dataset (TACRED) (Zhang et al., 2017), as used in the TAC KBP challenges, contains 106k sentences from newswire and web text covering 41 relation types. TACREV (Alt et al., 2020) and Re-TACRED (Stoica et al., 2021) provides cleane versions of TACRED. DOCRED (Yao et al., 2019) is a document level relation extraction dataset on the WikiPedia domain, with 5053 manually annotated documents and 100 relation types. FewRel (Han et al., 2018)is a relation extraction benchmark for few-shot scenario, based on WikiPedia. A newer version (Gao et al., 2019) includes Biomedical relations as a domain adaptation task.

5.4. Relation Extraction Models

Traditional RE models have focused on classifying the entity interaction in a sentence. For example, Zeng et al. (2014) encoded sentences and entity pairs with convolutional neural networks and position embeddings. Soares et al. (2019) finetuned Bert with self-supervised signals from entity linking, and applied the model to downstream RE tasks. There is also previous work targeting longer text passages such as cross sentence RE (Quirk and Poon, 2017), or document level distant supervision RE (Verga et al., 2018; Sahu et al., 2019; Christopoulou et al., 2019).
Sahu et al. (2019) and Christopoulou et al. (2019) encode graphs generated from each document for RE. In contrast, BRAN (Verga et al., 2018) uses transformers to encode the text sequence and then evaluates each mention pair of candidate argument entities. All these

Model	Micro P	Micro R	Micro F1	Micro Avg. P	Macro P	Macro R	Macro F1
CTD-derived corpus: 'dev' split / 'test' split							
BRAN	32.1 / 31.7	46.3 / 44.2	37.9 / 36.9	28.4 / 27.9	25.9 / 23.6	32.3 / 30.1	28.2 / 26.0
PubmedBert	50.3 / 49.6	59.3 / 56.1	54.5 / 52.6	50.3 / 50.1	43.6 / 39.0	50.3 / 48.4	44.9 / 41.7
PubmedBert + BRAN	53.9 / 53.9	61.0 / 57.3	57.3 / 55.6	54.0 / 54.3	45.0 / 42.7	54.1 / 50.4	48.7 / 44.4
Curated corpus: CTD-derived relationships only / All relationships							
BRAN	24.4 / 41.8	45.8 / 26.6	31.8 / 32.5	28.1 / 33.5	20.3 / 37.2	35.7 / 22.5	24.5 / 25.8
PubmedBert	43.0 / 64.3	61.7 / 31.3	50.7 / 42.1	50.7 / 46.9	34.7 / 53.7	53.4 / 32.0	39.6 / 37.0
PubmedBert + BRAN	46.5 / 70.9	61.1 / 31.6	52.8 / 43.8	53.0 / 50.6	45.8 / 69.8	59.0 / 32.5	47.0 / 40.5

Table 5: Performance of baseline models on *ChemDisGene* CTD-derived 'dev', 'test' and curated corpora.

Relation Type	F1
Chemical-Disease : marker/mechanism	54.1
Chemical-Disease : therapeutic	45.5
Chemical-Gene : expression - increases	58.2
Chemical-Gene : expression - decreases	61.6
Gene-Disease : marker/mechanism	47.1
Chemical-Gene : activity - increases	52.4
Chemical-Gene : activity - decreases	56.3
Chemical-Gene : metabolic_processing - increases	36.4
Chemical-Gene : binding - affects	58.1
Chemical-Gene : transport - increases	36.1
Chemical-Gene : metabolic_processing - decreases	34.4
Chemical-Gene : localization - affects	48.9
Chemical-Gene : expression - affects	0.4
Gene-Disease : therapeutic	28.6

Table 6: 'PubmedBert + BRAN' model metrics for each relation type in the curated corpus, sorted on decreasing relation frequency in the training data.

Model	Micro F1	Macro F1
BRAN	43.5	30.1
PubmedBert	58.9	44.6
PubmedBert + BRAN	60.0	46.0

Table 7: Evaluation on data from (Verga et al., 2018).

document level RE models showed comparable performance on BC5CDR. We chose BRAN as a baseline because it does not require a graph generation step.

In addition to RE, there is triple extraction work (Bansal et al., 2020) that recognizes entities and relationships simultaneously.

There has also been recent work on extracting n-ary biomedical relationships across sentences, e.g. (Ernst et al., 2018) learns dependency parse tree patterns from seed facts, and (Peng et al., 2017) applies graph LSTMs to dependency parses, trained on a noisy distant labeled dataset. In this paper we focus on binary relations.

6. Conclusion

We introduced *ChemDisGene*, a new dataset of research abstracts labeled with biomedical entity mentions and distance-labeled with biomedical relationships, for training and evaluating multi-type multi-label biomedical RE models. The dataset includes a large automatically derived corpus with noisy relationship labels ($\sim$ 22% noise based on manual curation), and a cleaner manually curated dataset of 523 abstracts. We also provided three baseline ML models for RE, trained and evaluated on the *ChemDisGene* dataset. We believe this is the first dataset for biomedical relation extraction tasks that addresses multiple entity (more than 2) and relation types, and includes both a large automatically derived corpus (useful for model training), as well as a smaller corpus labeled by human experts.

Manually annotating raw text with biomedical relationships is a hard and time consuming task, even for domain experts. We facilitated the curation with high quality models for entity recognition.

Future refinements to this dataset could include verifying the linked entities in the curated corpus, and adding *Protein-Protein* interactions, useful for understanding disease mechanisms and drug repurposing.

Acknowledgments

We thank the CZI Curation team for their work on the curated corpus, and members of UMass IESL and NLP groups for helpful discussion and feedback. This work is based upon work supported in part by the Center for Data Science and the Center for Intelligent Information Retrieval, and in part by the National Science Foundation under Grant Nos. 1763618 and 1514053, and in part by the Chan Zuckerberg Initiative under the project "Scientific Knowledge Base Construction". Some of the work reported here was performed using high performance computing equipment obtained under a grant from the Collaborative R&D Fund managed by the Massachusetts Technology Collaborative. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

7. Bibliographical References

Alt, C., Gabryszak, A., and Hennig, L. (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.

Bansal, T., Verga, P., Choudhary, N., and McCallum, A. (2020). Simultaneously linking entities and extracting relations from biomedical text without mention-level supervision. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, AAAI-20, pages 7407–7414, New York, NY, USA, February. AAAI Press.

Baumgartner, William A., J., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–i48, 07.

Christopoulou, F., Miwa, M., and Ananiadou, S. (2019). Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. 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 4925–4936.

Crichton, G., Baker, S., Guo, Y., and Korhonen, A. (2020). Neural networks for open and closed literature-based discovery. *PLOS ONE*, 15(5):1–16, 05.

Davis, A. P., Wiegers, T. C., Murphy, C. G., and Mattingly, C. J. (2011). The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database*, 2011, 09. bar034.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J., Wiegers, T. C., and Mattingly, C. J. (2020). Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*, 49(D1):D1138–D1143, 10.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Ernst, P., Siu, A., and Weikum, G. (2018). Highlife: Higher-arity fact harvesting. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1013–1022, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Gao, T., Han, X., Zhu, H., Liu, Z., Li, P., Sun, M., and Zhou, J. (2019). FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, EMNLP-IJCNLP 2019, pages 6250–6255, Hong Kong, China, November. Association for Computational Linguistics.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., and Sun, M. (2018). Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July. Association for Computational Linguistics.

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

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726. PMID: 28936969.

Islamaj Doğan, R., Kim, S., Chatr-aryamontri, A., Wei, C.-H., Comeau, D. C., Antunes, R., Matos, S., Chen, Q., Elangovan, A., Panyam, N. C., Verspoor, K., Liu, H., Wang, Y., Liu, Z., Altınel, B., Hüsünbeyi, Z. M., Özgür, A., Fergadis, A., Wang, C.-K., Dai, H.-J., Tran, T., Kavuluru, R., Luo, L., Steppi, A., Zhang, J., Qu, J., and Lu, Z. (2019). Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine. *Database*, 2019, 01. bay147.

Kilicoglu, H., Rosemblat, G., Fiszman, M., and Rindflesch, T. C. (2011). Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinformatics*, 12(486), December.

Krallinger, M., Leitner, F., and Valencia, A. (2006). Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proceedings of the second BioCreative Challenge Evaluation Workshop*. CNIO Centro Nacional de Investigaciones Oncológicas, November.

Lee, Y.-S., Krishnan, A., Oughtred, R., Rust, J., Chang,

C. S., Ryu, J., Kristensen, V. N., Dolinski, K., Theesfeld, C. L., and Troyanskaya, O. G. (2019). A computational framework for genome-wide characterization of the human disease landscape. *Cell systems*, 8(2):152–162.e6.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Miranda, A., Mehryary, F., Luoma, J., Pyysalo, S., Valencia, A., and Krallinger, M. (2021). Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the BioCreative VII Challenge Evaluation Workshop*, November.

Morselli Gysi, D., do Valle, Í., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S. D., Patten, J. J., Davey, R. A., Loscalzo, J., and Barabási, A.-L. (2021). Network medicine framework for identifying drug-repurposing opportunities for covid-19. *Proceedings of the National Academy of Sciences*, 118(19).

Nelson, C. A., Acuna, A. U., Paul, A. M., Scott, R. T., Butte, A. J., Cekanaviciute, E., Baranzini, S. E., and Costes, S. V. (2021). Knowledge network embedding of transcriptomic data from spaceflown mice uncovers signs and symptoms associated with terrestrial diseases. *Life*, 11(1).

Peng, N., Poon, H., Quirk, C., Toutanova, K., and Yih, W.-t. (2017). Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Quirk, C. and Poon, H. (2017). Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182.

Sahu, S. K., Christopoulou, F., Miwa, M., and Ananiadou, S. (2019). Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4309–4316.

Soares, L. B., FitzGerald, N., Ling, J., and Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.

Stoica, G., Platanios, E. A., and Poczos, B. (2021). Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13843–13850.

Verga, P., Strubell, E., and McCallum, A. (2018). Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, NAACL 2018, pages 872–884, New Orleans, Louisiana, June. Association for Computational Linguistics.

Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator Central: Automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593, 05.

Wu, Y., Luo, R., Leung, H. C. M., Ting, H.-F., and Lam, T.-W. (2019). RENET: A deep learning approach for extracting gene-disease associations from literature. In Lenore J. Cowen, editor, *Research in Computational Molecular Biology*, RECOMB 2019, pages 272–284, Cham. Springer International Publishing.

Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., and Sun, M. (2019). DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy, July. Association for Computational Linguistics.

Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Zhang, Y., Zhong, V., Chen, D., Angeli, G., and Manning, C. D. (2017). Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

8. Language Resource References

Martin Krallinger, Obdulia Rabal, Antonio Miranda-Escalada and Alfonso Valencia. (2021). *DrugProt corpus: Biocreative VII Track 1 – Text mining drug and chemical-protein interactions*. BioCreative, BioCreative VII.