# Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction

**Patrick Verga, Emma Strubell, Andrew McCallum**

College of Information and Computer Sciences
University of Massachusetts Amherst
{pat, strubell, mccallum}@cs.umass.edu

## Abstract

Most work in relation extraction forms a prediction by looking at a short span of text within a single sentence containing a single entity pair mention. This approach often does not consider interactions across mentions, requires redundant computation for each mention pair, and ignores relationships expressed across sentence boundaries. These problems are exacerbated by the document- (rather than sentence-) level annotation common in biological text. In response, we propose a model which simultaneously predicts relationships between all mention pairs in a document. We form pairwise predictions over entire paper abstracts using an efficient self-attention encoder. All-pairs mention scores allow us to perform multi-instance learning by aggregating over mentions to form entity pair representations. We further adapt to settings without mention-level annotation by jointly training to predict named entities and adding a corpus of weakly labeled data. In experiments on two Biocreative benchmark datasets, we achieve state of the art performance on the Biocreative V Chemical Disease Relation dataset for models without external KB resources. We also introduce a new dataset an order of magnitude larger than existing human-annotated biological information extraction datasets and more accurate than distantly supervised alternatives.

## 1 Introduction

With few exceptions (Swampillai and Stevenson, 2011; Quirk and Poon, 2017; Peng et al., 2017), nearly all work in relation extraction focuses on classifying a short span of text within a single sentence containing a single entity pair mention. However, relationships between entities are often expressed across sentence boundaries or otherwise require a larger context to disambiguate. For example, 30% of relations in the Biocreative V CDR dataset (§3.1) are expressed across sentence boundaries, such as in the following excerpt expressing a relationship between the chemical **azathioprine** and the disease **fibrosis**:

> *Treatment of psoriasis with **azathioprine**.* **Azathioprine** treatment benefited 19 (66%) out of 29 patients suffering from severe psoriasis. Haematological complications were not troublesome and results of biochemical liver function tests remained normal. Minimal cholestasis was seen in two cases and portal **fibrosis** of a reversible degree in eight. Liver biopsies should be undertaken at regular intervals if **azathioprine** therapy is continued so that structural liver damage may be detected at an early and reversible stage.

Though the entities' mentions never occur in the same sentence, the above example expresses that the chemical entity *azathioprine* can cause the side effect *fibrosis*. Relation extraction models which consider only within-sentence relation pairs cannot extract this fact without knowledge of the complicated coreference relationship between *eight* and *azathioprine treatment*, which, without features from a complicated pre-processing pipeline, cannot be learned by a model which considers entity pairs in isolation. Making separate predictions for each mention pair also obstructs multi-instance learning (Riedel et al., 2010; Surdeanu et al., 2012), a technique which aggregates entity representations from mentions in order to improve robustness to noise in the data. Like the majority of relation extraction data, most annotation for biological relations is distantly supervised, and so we could benefit from a model which is amenable to multi-instance learning.

In addition to this loss of cross-sentence and cross-mention reasoning capability, traditional mention pair relation extraction models typically introduce computational inefficiencies by independently extracting features for and scoring every pair of mentions, even when those mentions occur in the same sentence and thus could share representations. In the CDR training set, this requires separately encoding and classifying each of the 5,318 candidate mention pairs independently, versus encoding each of the 500 abstracts once. Though abstracts



*Proceedings of NAACL-HLT 2018*, pages 872–884
New Orleans, Louisiana, June 1 - 6, 2018. 

are longer than e.g. the text between mentions, many sentences contain multiple mentions, leading to redundant computation.

However, encoding long sequences in a way which effectively incorporates long-distance context can be prohibitively expensive. Long Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) are among the most popular token encoders due to their capacity to learn high-quality representations of text, but their ability to leverage the fastest computing hardware is thwarted due to their computational dependence on the length of the sequence — each token's representation requires as input the representation of the previous token, limiting the extent to which computation can be parallelized. Convolutional neural networks (CNNs), in contrast, can be executed entirely in parallel across the sequence, but the amount of context incorporated into a single token's representation is limited by the depth of the network, and very deep networks can be difficult to learn (Hochreiter, 1998). These problems are exacerbated by longer sequences, limiting the extent to which previous work explored full-abstract relation extraction.

To facilitate efficient full-abstract relation extraction from biological text, we propose Bi-affine Relation Attention Networks (BRANs), a combination of network architecture, multi-instance and multi-task learning designed to extract relations between entities in biological text without requiring explicit mention-level annotation. We synthesize convolutions and self-attention, a modification of the Transformer encoder introduced by Vaswani et al. (2017), over sub-word tokens to efficiently incorporate into token representations rich context between distant mention pairs across the entire abstract. We score all pairs of mentions in parallel using a bi-affine operator, and aggregate over mention pairs using a soft approximation of the max function in order to perform multi-instance learning. We jointly train the model to predict relations and entities, further improving robustness to noise and lack of gold annotation at the mention level.

In extensive experiments on two benchmark biological relation extraction datasets, we achieve state of the art performance for a model using no external knowledge base resources in experiments on the Biocreative V CDR dataset, and outperform comparable baselines on the Biocreative VI ChemProt dataset. We also introduce a new dataset which is an order of magnitude larger than existing gold-annotated biological relation extraction datasets while covering a wider range of entity and relation types and with higher accuracy than distantly supervised datasets of the same size. We provide a strong baseline on this new dataset, and encourage its use as a benchmark for future biological relation extraction systems.[1]

## 2 Model

We designed our model to efficiently encode long contexts spanning multiple sentences while forming pairwise predictions without the need for mention pair-specific features. To do this, our model first encodes input token embeddings using self-attention. These embeddings are used to predict both entities and relations. The relation extraction module converts each token to a *head* and *tail* representation. These representations are used to form mention pair predictions using a bi-affine operation with respect to learned relation embeddings. Finally, these mention pair predictions are pooled to form entity pair predictions, expressing whether each relation type is expressed by each relation pair.

### 2.1 Inputs

Our model takes in a sequence of $N$ token embeddings in $\mathbb{R}^d$. Because the Transformer has no innate notion of token position, the model relies on positional embeddings which are added to the input token embeddings.[2] We learn the position embedding matrix $P^{m \times d}$ which contains a separate $d$ dimensional embedding for each position, limited to $m$ possible positions. Our final input representation for token $x_i$ is:

$$x_i = s_i + p_i$$

where $s_i$ is the token embedding for $x_i$ and $p_i$ is the positional embedding for the $i$th position. If $i$ exceeds $m$, we use a randomly initialized vector in place of $p_i$.

We tokenize the text using byte pair encoding (BPE) (Gage, 1994; Sennrich et al., 2015). The BPE algorithm constructs a vocabulary of sub-word pieces, beginning with single characters. Then, the algorithm iteratively merges the most frequent co-occurring tokens into a new token, which is added to the vocabulary. This procedure continues until a pre-defined vocabulary size is met.

BPE is well suited for biological data for the following reasons. First, biological entities often have unique mentions made up of meaningful subcomponents, such as *1,2-dimethylhydrazine*. Additionally, tokenization of chemical entities is challenging, lacking a universally agreed upon algorithm (Krallinger et al., 2015). As we demonstrate in §3.3.2, the sub-word representations produced by BPE allow the model to formulate better predictions, likely due to better modeling of rare and unknown words.

[1] Our code and data are publicly available at: https://github.com/patverga/bran.

[2] Though our final model incorporates some convolutions, we retain the position embeddings.

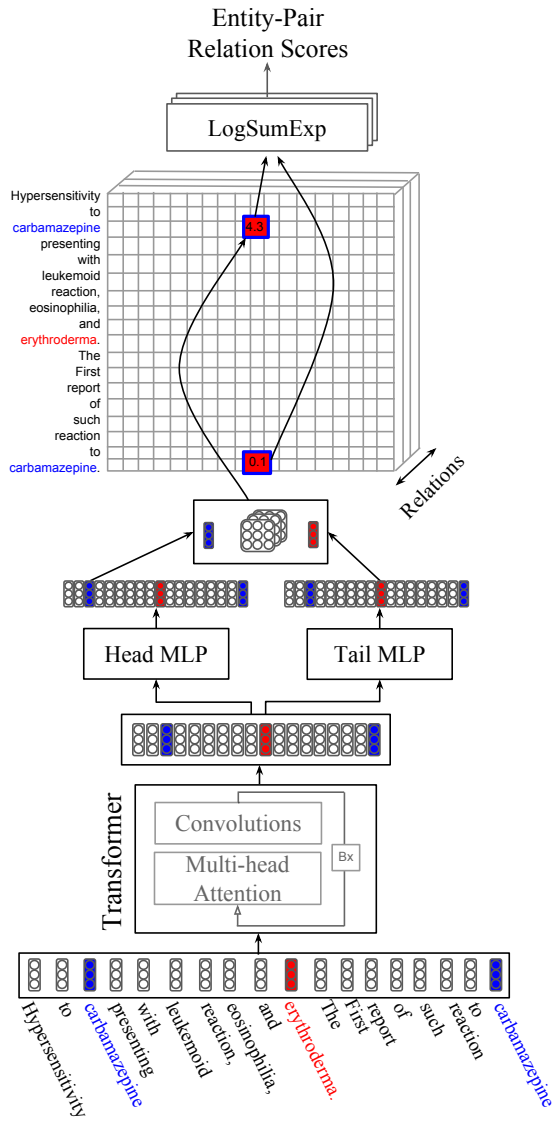


Figure 1: The relation extraction architecture. Inputs are contextually encoded using the Transformer(Vaswani et al., 2017), made up of $B$ layers of multi-head attention and convolution subcomponents. Each transformed token is then passed through a *head* and *tail* MLP to produce two position-specific representations. A bi-affine operation is performed between each *head* and *tail* representation with respect to each relation's embedding matrix, producing a pair-wise relation affinity tensor. Finally, the scores for cells corresponding to the same entity pair are pooled with a separate LogSumExp operation for each relation to get a final score. The colored tokens illustrate calculating the score for a given pair of entities; the model is only given entity information when pooling over mentions.

## 2.2 Transformer

We base our token encoder on the Transformer self-attention model (Vaswani et al., 2017). The Transformer is made up of $B$ blocks. Each Transformer block, which we denote $\text{Transformer}_k$, has its own set of parameters and is made up of two subcomponents: multi-head attention and a series of convolutions[3]. The output for token $i$ of block $k$, $b_i^{(k)}$, is connected to its input $b_i^{(k-1)}$ with a residual connection (He et al., 2016). Starting with $b_i^{(0)} = x_i$:

$$b_i^{(k)} = b_i^{(k-1)} + \text{Transformer}_k(b_i^{(k-1)})$$

### 2.2.1 Multi-head Attention

Multi-head attention applies self-attention multiple times over the same inputs using separately normalized parameters (attention heads) and combines the results, as an alternative to applying one pass of attention with more parameters. The intuition behind this modeling decision is that dividing the attention into multiple heads make it easier for the model to learn to attend to different types of relevant information with each head. The self-attention updates input $b_i^{(k-1)}$ by performing a weighted sum over all tokens in the sequence, weighted by their importance for modeling token $i$.

Each input is projected to a key $k$, value $v$, and query $q$, using separate affine transformations with ReLU activations (Glorot et al., 2011). Here, $k$, $v$, and $q$ are each in $\mathbb{R}^{\frac{d}{H}}$ where $H$ is the number of heads. The attention weights $a_{ijh}$ for head $h$ between tokens $i$ and $j$ are computed using scaled dot-product attention:

$$a_{ijh} = \sigma\left(\frac{q_{ih}^T k_{jh}}{\sqrt{d}}\right)$$
$$o_{ih} = \sum_j v_{jh} \odot a_{ijh}$$

with $\odot$ denoting element-wise multiplication and $\sigma$ indicating a softmax along the $j$th dimension. The scaled attention is meant to aid optimization by flattening the softmax and better distributing the gradients (Vaswani et al., 2017).

The outputs of the individual attention heads are concatenated, denoted $[\cdot;\cdot]$, into $o_i$. All layers in the network use residual connections between the output of the multi-headed attention and its input. Layer normalization (Ba et al., 2016), denoted $\text{LN}(\cdot)$, is then applied to the output.

$$o_i = [o_1; ...; o_h]$$
$$m_i = \text{LN}(b_i^{(k-1)} + o_i)$$

### 2.2.2 Convolutions

The second part of our Transformer block is a stack of convolutional layers. The sub-network used in

[3]The original Transformer uses feed-forward connections, i.e. width-1 convolutions, whereas we use convolutions with width $> 1$.

Vaswani et al. (2017) uses two width-1 convolutions. We add a third middle layer with kernel width 5, which we found to perform better. Many relations are expressed concisely by the immediate local context, e.g. *Michele's husband Barack*, or *labetalol-induced hypotension*. Adding this explicit n-gram modeling is meant to ease the burden on the model to learn to attend to local features. We use $C_w(\cdot)$ to denote a convolutional operator with kernel width $w$. Then the convolutional portion of the transformer block is given by:

$$t_i^{(0)} = \text{ReLU}(C_1(m_i))$$
$$t_i^{(1)} = \text{ReLU}(C_5(t_i^{(0)}))$$
$$t_i^{(2)} = C_1(t_i^{(1)})$$

Where the dimensions of $t_i^{(0)}$ and $t_i^{(1)}$ are in $\mathbb{R}^{4d}$ and that of $t_i^{(2)}$ is in $\mathbb{R}^d$.

### 2.3 Bi-affine Pairwise Scores

We project each contextually encoded token $b_i^{(B)}$ through two separate MLPs to generate two new versions of each token corresponding to whether it will serve as the first (head) or second (tail) argument of a relation:

$$e_i^{head} = W_{head}^{(1)}(\text{ReLU}(W_{head}^{(0)} b_i^{(B)}))$$
$$e_i^{tail} = W_{tail}^{(1)}(\text{ReLU}(W_{tail}^{(0)} b_i^{(B)}))$$

We use a bi-affine operator to calculate an $N \times L \times N$ tensor $A$ of pairwise affinity scores, scoring each (head, relation, tail) triple:

$$A_{ilj} = (e_i^{head} L) e_j^{tail}$$

where $L$ is a $d \times L \times d$ tensor, a learned embedding matrix for each of the $L$ relations. In subsequent sections we will assume we have transposed the dimensions of $A$ as $d \times d \times L$ for ease of indexing.

### 2.4 Entity Level Prediction

Our data is weakly labeled in that there are labels at the entity level but not the mention level, making the problem a form of strong-distant supervision (Mintz et al., 2009). In distant supervision, edges in a knowledge graph are heuristically applied to sentences in an auxiliary unstructured text corpus — often applying the edge label to all sentences containing the subject and object of the relation. Because this process is imprecise and introduces noise into the training data, methods like multi-instance learning were introduced (Riedel et al., 2010; Surdeanu et al., 2012). In multi-instance learning, rather than looking at each distantly labeled mention pair in isolation, the model is trained over the aggregate of these mentions and a single update is made. More recently, the weighting function of the instances has been expressed as neural network attention (Verga and McCallum, 2016; Lin et al., 2016; Yaghoobzadeh et al., 2017).

We aggregate over all representations for each mention pair in order to produce per-relation scores for each entity pair. For each entity pair $(p^{head}, p^{tail})$, let $P^{head}$ denote the set of indices of mentions of the entity $p^{head}$, and let $P^{tail}$ denote the indices of mentions of the entity $p^{tail}$. Then we use the LogSumExp function to aggregate the relation scores from $A$ across all pairs of mentions of $p^{head}$ and $p^{tail}$:

$$scores(p^{head}, p^{tail}) = \log \sum_{\substack{i \in P^{head} \\ j \in P^{tail}}} \exp(A_{ij})$$

The LogSumExp scoring function is a smooth approximation to the max function and has the benefits of aggregating information from multiple predictions and propagating dense gradients as opposed to the sparse gradient updates of the max (Das et al., 2017).

### 2.5 Named Entity Recognition

In addition to pairwise relation predictions, we use the Transformer output $b_i^{(B)}$ to make entity type predictions. We feed $b_i^{(B)}$ as input to a linear classifier which predicts the entity label for each token with per-class scores $c_i$:

$$c_i = W^{(3)} b_i^{(B)}$$

We augment the entity type labels with the BIO encoding to denote entity spans. We apply tags to the byte-pair tokenization by treating each subword within a mention span as an additional token with a corresponding B- or I- label.

### 2.6 Training

We train both the NER and relation extraction components of our network to perform multi-class classification using maximum likelihood, where NER classes $y_i$ or relation classes $r_i$ are conditionally independent given deep features produced by our model with probabilities given by the softmax function. In the case of NER, features are given by the per-token output of the transformer:

$$\frac{1}{N} \sum_{i=1}^{N} \log P(y_i \mid b_i^{(B)})$$

In the case of relation extraction, the features for each entity pair are given by the LogSumExp over pairwise scores described in § 2.4. For $E$ entity pairs, the relation $r_i$ is given by:

$$\frac{1}{E} \sum_{i=1}^{E} \log P(r_i \mid scores(p^{head}, p^{tail}))$$

We train the NER and relation objectives jointly, sharing all embeddings and Transformer parameters. To trade off the two objectives, we penalize the named entity updates with a hyperparameter $\lambda$.

## 3 Results

We evaluate our model on three datasets: The Biocreative V Chemical Disease Relation benchmark (CDR), which models relations between chemicals and diseases (§3.1); the Biocreative VI ChemProt benchmark (CPR), which models relations between chemicals and proteins (§3.2); and a new, large and accurate dataset we describe in §3.3 based on the human curation in the Chemical Toxicology Database (CTD), which models relationships between chemicals, proteins and genes.

The CDR dataset is annotated at the level of paper abstracts, requiring consideration of long-range, cross sentence relationships, thus evaluation on this dataset demonstrates that our model is capable of such reasoning. We also evaluate our model's performance in the more traditional setting which does not require cross-sentence modeling by performing experiments on the CPR dataset, for which all annotations are between two entity mentions in a single sentence. Finally, we present a new dataset constructed using strong-distant supervision (§2.4), with annotations at the document level. This dataset is significantly larger than the others, contains more relation types, and requires reasoning across sentences.

### 3.1 Chemical Disease Relations Dataset

The Biocreative V chemical disease relation extraction (CDR) dataset[4] (Li et al., 2016a; Wei et al., 2016) was derived from the Comparative Toxicogenomics Database (CTD), which curates interactions between genes, chemicals, and diseases (Davis et al., 2008). CTD annotations are only at the document level and do not contain mention annotations. The CDR dataset is a subset of these original annotations, supplemented with human annotated, entity linked mention annotations. The relation annotations in this dataset are also at the document level only.

#### 3.1.1 Data Preprocessing

The CDR dataset is concerned with extracting only chemically-induced disease relationships (drug-related side effects and adverse reactions) concerning the most specific entity in the document. For example *tobacco causes cancer* could be marked as false if the document contained the more specific *lung cancer*. This can cause true relations to be labeled as false, harming evaluation performance. To address this we follow (Gu et al., 2016, 2017) and filter hypernyms according to the hierarchy in the MESH controlled vocabulary[5]. All entity pairs within the same abstract that do not have an annotated relation are assigned the NULL label.

In addition to the gold CDR data, Peng et al. (2016) add 15,448 PubMed abstracts annotated in the CTD dataset. We consider this same set of abstracts as additional training data (which we subsequently denote +Data). Since this data does not contain entity annotations, we take the annotations from Pubtator (Wei et al., 2013), a state of the art biological named entity tagger and entity linker. See §A.1 for additional data processing details. In our experiments we only evaluate our relation extraction performance and all models (including baselines) use gold entity annotations for predictions.

The byte pair vocabulary is generated over the training dataset — we use a budget of 2500 tokens when training on the gold CDR data, and a larger budget of 10,000 tokens when including extra data described above Additional implementation details are included in Appendix A.

| Data split | Docs | Pos | Neg |
|---|---|---|---|
| Train | 500 | 1,038 | 4,280 |
| Development | 500 | 1,012 | 4,136 |
| Test | 500 | 1,066 | 4,270 |
| CTD | 15,448 | 26,657 | 146,057 |

Table 1: Data statistics for the CDR Dataset and additional data from CTD. Shows the total number of abstracts, positive examples, and negative examples for each of the data set splits.

#### 3.1.2 Baselines

We compare against the previous best reported results on this dataset not using knowledge base features.[6] Each of the baselines are ensemble methods for within- and cross-sentence relations that make use of additional linguistic features (syntactic parse and part-of-speech). Gu et al. (2017) encode mention pairs using a CNN while Zhou et al. (2016a) use an LSTM. Both make cross-sentence predictions with featurized classifiers.

#### 3.1.3 Results

In Table 2 we show results outperforming the baselines despite using no linguistic features. We show performance averaged over 20 runs with 20 random seeds as well as an ensemble of their averaged predictions. We see a further boost in performance by adding weakly labeled data. Table 3 shows the

[4] http://www.biocreative.org/

[5] https://www.nlm.nih.gov/mesh/download/2017MeshTree.txt

[6] The highest reported score is from (Peng et al., 2016), but they use explicit lookups into the CTD knowledge base for the existence of the test entity pair.

| Model | P | R | F1 |
|---|---|---|---|
| Gu et al. (2016) | 62.0 | 55.1 | 58.3 |
| Zhou et al. (2016a) | 55.6 | 68.4 | 61.3 |
| Gu et al. (2017) | 55.7 | 68.1 | 61.3 |
| BRAN | 55.6 | 70.8 | **62.1** ± 0.8 |
| + Data | 64.0 | 69.2 | **66.2** ± 0.8 |
| BRAN(ensemble) | 63.3 | 67.1 | 65.1 |
| + Data | 65.4 | 71.8 | **68.4** |

Table 2: Precision, recall, and F1 results on the Biocreative V CDR Dataset.

| Model | P | R | F1 |
|---|---|---|---|
| BRAN (Full) | 55.6 | 70.8 | **62.1** ± 0.8 |
| – CNN only | 43.9 | 65.5 | 52.4 ± 1.3 |
| – no width-5 | 48.2 | 67.2 | 55.7 ± 0.9 |
| – no NER | 49.9 | 63.8 | 55.5 ± 1.8 |

Table 3: Results on the Biocreative V CDR Dataset showing precision, recall, and F1 for various model ablations.

effects of ablating pieces of our model. 'CNN only' removes the multi-head attention component from the transformer block, 'no width-5' replaces the width-5 convolution of the feed-forward component of the transformer with a width-1 convolution and 'no NER' removes the named entity recognition multi-task objective (§2.5).

### 3.2 Chemical Protein Relations Dataset

To assess our model's performance in settings where cross-sentence relationships are not explicitly evaluated, we perform experiments on the Biocreative VI ChemProt dataset (CDR) (Krallinger et al., 2017). This dataset is concerned with classifying into six relation types between chemicals and proteins, with nearly all annotated relationships occurring within the same sentence.

#### 3.2.1 Baselines

We compare our models against those competing in the official Biocreative VI competition (Liu et al., 2017). We compare to the top performing team whose model is directly comparable with ours — i.e. used a single (non-ensemble) model trained only on the training data (many teams use the development set as additional training data). The baseline models are standard state of the art relation extraction models: CNNs and Gated RNNs with attention. Each of these baselines uses mention-specific features encoding relative position of each token to the two target entities being classified, whereas our model aggregates over all mention pairs in each sentence. It is also worth noting that these models use a large vocabulary of pre-trained word embeddings, giving their models the advantage of far more model parameters, as well as additional information from unsupervised pre-training.

| Model | P | R | F1 |
|---|---|---|---|
| CNN† | 50.7 | 43.0 | 46.5 |
| GRU+Attention† | 53.0 | 46.3 | 49.5 |
| BRAN | 48.0 | 54.1 | **50.8** ± .01 |

Table 4: Precision, recall, and F1 results on the Biocreative VI Chem-Prot Dataset. † denotes results from Liu et al. (2017)

#### 3.2.2 Results

In Table 4 we see that even though our model forms all predictions simultaneously between all pairs of entities within the sentence, we are able to outperform state of the art models classifying each mention pair independently. The scores shown are averaged across 10 runs with 10 random seeds. Interestingly, our model appears to have higher recall and lower precision, while the baseline models are both precision-biased, with lower recall. This suggests that combining these styles of model could lead to further gains on this task.

### 3.3 New CTD Dataset

#### 3.3.1 Data

Existing biological relation extraction datasets including both CDR (§3.1) and CPR (§3.2) are relatively small, typically consisting of hundreds or a few thousand annotated examples. Distant supervision datasets apply document-independent, entity-level annotations to all sentences leading to a large proportion of incorrect labels. Evaluations on this data involve either very small (a few hundred) gold annotated examples or cross validation to predict the noisy, distantly applied labels (Mallory et al., 2015; Quirk and Poon, 2017; Peng et al., 2017).

We address these issues by constructing a new dataset using strong-distant supervision containing document-level annotations. The Comparative Toxicogenomics Database (CTD) curates interactions between genes, chemicals, and diseases. Each relation in the CTD is associated with a disambiguated entity pair and a PubMed article where the relation was observed.

To construct this dataset, we collect the abstracts for each of the PubMed articles with at least one curated relation in the CTD database. As in §3.1, we use PubTator to automatically tag and disambiguate the entities in each of these abstracts. If both entities in the relation are found in the abstract, we take the (abstract, relation) pair as a positive example. The evidence for the curated relation could occur anywhere in the full text article, not just the abstract. Abstracts with no recovered relations are discarded. All other entity pairs with valid types and without an annotated relation that

| Types | Docs | Pos | Neg |
|---|---|---|---|
| Total | 68,400 | 166,474 | 1198,493 |
| Chemical/Disease | 64,139 | 93,940 | 571,932 |
| Chemical/Gene | 34,883 | 63,463 | 360,100 |
| Gene/Disease | 32,286 | 9,071 | 266,461 |

Table 5: Data statistics for the new CTD dataset.

occur in the remaining abstracts are considered negative examples and assigned the NULL label. We additionally remove abstracts containing greater than 500 tokens[7]. This limit removed about 10% of the total data including numerous extremely long abstracts. The average token length of the remaining data was $\tilde{2}30$ tokens. With this procedure, we are able to collect 166,474 positive examples over 13 relation types, with more detailed statistics of the dataset listed in Table 5.

We consider relations between chemical-disease, chemical-gene, and gene-disease entity pairs downloaded from CTD[8]. We remove inferred relations (those without an associated PubMed ID) and consider only human curated relationships. Some chemical-gene entity pairs were associated with multiple relation types in the same document. We consider each of these relation types as a separate positive example.

The chemical-gene relation data contains over 100 types organized in a shallow hierarchy. Many of these types are extremely infrequent, so we map all relations to the highest parent in the hierarchy, resulting in 13 relation types. Most of these chemical-gene relations have an increase and decrease version such as increase_expression and decrease_expression. In some cases, there is also an affects relation (affects_expression) which is used when the directionality is unknown. If the affects version is more common, we map decrease and increase to affects. If affects is less common, we drop the affects examples and keep the increase and decrease examples as distinct relations, resulting in the final set of 10 chemical-gene relation types.

#### 3.3.2 Results

In Table 7 we list precision, recall and F1 achieved by our model on the CTD dataset, both overall and by relation type. Our model predicts each of the relation types effectively, with higher performance on relations with more support.

In Table 8 we see that our sub-word BPE model out-performs the model using the Genia tokenizer (Kulick et al., 2012) even though our vocabulary size is one-fifth as large. We see a 1.7 F1 point boost in predicting Pubtator NER labels for BPE. This could be explained by the increased out-of-vocabulary (OOV) rate for named entities. Word training data has 3.01 percent OOV rate for tokens with an entity. The byte pair-encoded data has an OOV rate of 2.48 percent. Note that in both the word-tokenized and byte pair-tokenized data, we replace tokens that occur less than five times with a learned UNK token.

[7]We include scripts to generate the unfiltered set of data as well to encourage future research

[8]http://ctdbase.org/downloads/

| | Train | Dev | Test |
|---|---|---|---|
| **Total** | 120k | 15k | 15k |
| **Chemical/Disease** | | | |
| marker/mechanism | 41,562 | 5,126 | 5,167 |
| therapeutic | 24,151 | 2,929 | 3,059 |
| **Gene/Disease** | | | |
| marker/mechanism | 5,930 | 825 | 819 |
| therapeutic | 560 | 77 | 75 |
| **Chemical/Gene** | | | |
| increase_expression | 15,851 | 1,958 | 2,137 |
| increase_MP | 5,986 | 740 | 638 |
| decrease_expression | 5,870 | 698 | 783 |
| increase_activity | 4,154 | 467 | 497 |
| affects_response | 3,834 | 475 | 508 |
| decrease_activity | 3,124 | 396 | 434 |
| affects_transport | 3,009 | 333 | 361 |
| increase_reaction | 2,881 | 367 | 353 |
| decrease_reaction | 2,221 | 247 | 269 |
| decrease_MP | 798 | 100 | 120 |

Table 6: Data statistics for the new CTD dataset broken down by relation type. The first column lists relation types separated by the types of the entities. Columns 2–4 show the number of positive examples of that relation type. MP stands for metabolic processing.

Figure 2 depicts the model's performance on relation extraction as a function of distance between entities. For example, the blue bar depicts performance when removing all entity pair candidates (positive and negative) whose closest mentions are more than 11 tokens apart. We consider removing entity pair candidates with distances of 11, 25, 50, 100 and 500 (the maximum document length). The average sentence length is 22 tokens. We see that the model is not simply relying on short range relationships, but is leveraging information about distant entity pairs, with accuracy increasing as the maximum distance considered increases. Note that all results are taken from the same model trained on the full unfiltered training set.

## 4 Related work

Relation extraction is a heavily studied area in the NLP community. Most work focuses on news and web data (Doddington et al., 2004; Riedel et al., 2010; Hendrickx et al., 2009).[9] Recent neural net-

[9]And TAC KBP: https://tac.nist.gov

| | P | R | F1 |
|---|---|---|---|
| **Total** | | | |
| Micro F1 | 44.8 | 50.2 | 47.3 |
| Macro F1 | 34.0 | 29.8 | 31.7 |
| **Chemical/Disease** | | | |
| marker/mechanism | 46.2 | 57.9 | 51.3 |
| therapeutic | 55.7 | 67.1 | 60.8 |
| **Gene/Disease** | | | |
| marker/mechanism | 42.2 | 44.4 | 43.0 |
| therapeutic | 52.6 | 10.1 | 15.8 |
| **Chemical/Gene** | | | |
| increases_expression | 39.7 | 48.0 | 43.3 |
| increases_MP | 26.3 | 35.5 | 29.9 |
| decreases_expression | 34.4 | 32.9 | 33.4 |
| increases_activity | 24.5 | 24.7 | 24.4 |
| affects_response | 40.9 | 35.5 | 37.4 |
| decreases_activity | 30.8 | 19.4 | 23.5 |
| affects_transport | 28.7 | 23.8 | 25.8 |
| increases_reaction | 12.8 | 5.6 | 7.4 |
| decreases_reaction | 12.3 | 5.7 | 7.4 |
| decreases_MP | 28.9 | 7.0 | 11.0 |

Table 7: BRAN precision, recall and F1 results for the full CTD dataset by relation type. The model is optimized for micro F1 score across all types.

| Model | P | R | F1 |
|---|---|---|---|
| **Relation extraction** | | | |
| Words | 44.9 | 48.8 | 46.7 ± 0.39 |
| BPE | 44.8 | 50.2 | **47.3** ± 0.19 |
| **NER** | | | |
| Words | 91.0 | 90.7 | 90.9 ± 0.13 |
| BPE | 91.5 | 93.6 | **92.6** ± 0.12 |

Table 8: Precision, recall, and F1 results for CTD named entity recognition and relation extraction, comparing BPE to word-level tokenization.

work approaches to relation extraction have focused on CNNs (dos Santos et al., 2015; Zeng et al., 2015) or LSTMs (Miwa and Bansal, 2016; Verga et al., 2016a; Zhou et al., 2016b) and replacing stage-wise information extraction pipelines with a single end-to-end model (Miwa and Bansal, 2016; Ammar et al., 2017; Li et al., 2017). These models all consider mention pairs separately.

There is also a considerable body of work specifically geared towards supervised biological relation extraction including protein-protein (Pyysalo et al., 2007; Poon et al., 2014; Mallory et al., 2015), drug-drug (Segura-Bedmar et al., 2013), and chemical-disease (Gurulingappa et al., 2012; Li et al., 2016a) interactions, and more complex events (Kim et al., 2008; Riedel et al., 2011). Our work focuses on modeling relations between chemicals, diseases, genes and proteins, where available annotation is often at the document- or abstract-level, rather than the

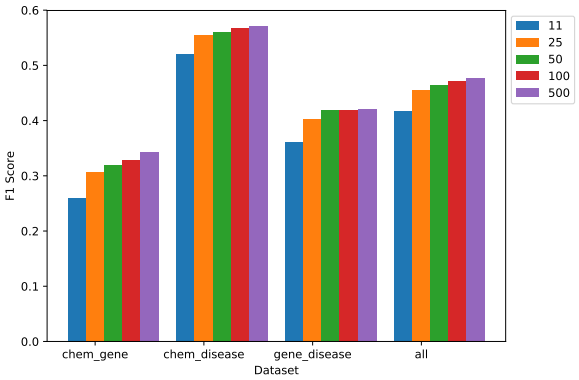


Figure 2: Performance on the CTD dataset when restricting candidate entity pairs by distance. The x-axis shows the coarse-grained relation type. The y-axis shows F1 score. Different colors denote maximum distance cutoffs.

sentence level.

Some previous work exists on cross-sentence relation extraction. Swampillai and Stevenson (2011) and Quirk and Poon (2017) consider featurized classifiers over cross-sentence syntactic parses. Most similar to our work is that of Peng et al. (2017), which uses a variant of an LSTM to encode document-level syntactic parse trees. Our work differs in three key ways. First, we operate over raw tokens negating the need for part-of-speech or syntactic parse features which can lead to cascading errors. We also use a feed-forward neural architecture which encodes long sequences far more efficiently compared to the graph LSTM network of Peng et al. (2017). Finally, our model considers all mention pairs simultaneously rather than a single mention pair at a time.

We employ a bi-affine function to form pairwise predictions between mentions. Such models have also been used for knowledge graph link prediction (Nickel et al., 2011; Li et al., 2016b), with variations such as restricting the bilinear relation matrix to be diagonal (Yang et al., 2015) or diagonal and complex (Trouillon et al., 2016). Our model is similar to recent approaches to graph-based dependency parsing, where bilinear parameters are used to score head-dependent compatibility (Kiperwasser and Goldberg, 2016; Dozat and Manning, 2017).

## 5 Conclusion

We present a bi-affine relation attention network that simultaneously scores all mention pairs within a document. Our model performs well on three datasets, including two standard benchmark biological relation extraction datasets and a new, large and high-quality dataset introduced in this work. Our model out-performs the previous state of the art on the Biocreative V CDR dataset despite us-

ing no additional linguistic resources or mention pair-specific features.

Our current model predicts only into a fixed schema of relations given by the data. However, this could be ameliorated by integrating our model into open relation extraction architectures such as Universal Schema (Riedel et al., 2013; Verga et al., 2016b). Our model also lends itself to other pairwise scoring tasks such as hypernym prediction, co-reference resolution, and entity resolution. We will investigate these directions in future work.

## Acknowledgments

We thank Ofer Shai and the Chan Zuckerberg Initiative / Meta data science team for helpful discussions. We also thank Timothy Dozat and Kyubyong Park for releasing their code.

## References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. http://tensorflow.org/.

Waleed Ammar, Matthew E. Peters, Chandra Bhagavatula, and Russell Power. 2017. The ai2 system at semeval-2017 task 10 (scienceie): semi-supervised end-to-end entity and relation extraction. *nucleus* 2(e2):e2.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* .

Rajarshi Das, Arvind Neelakantan, David Belanger, and Andrew McCallum. 2017. Chains of reasoning over entities, relations, and text using recurrent neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 132–141. http://www.aclweb.org/anthology/E17-1013.

Allan Peter Davis, Cynthia G Murphy, Cynthia A Saraceni-Richards, Michael C Rosenstein, Thomas C Wiegers, and Carolyn J Mattingly. 2008. Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic acids research* 37(suppl_1):D786–D792.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ace) program tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 626–634. http://www.aclweb.org/anthology/P15-1061.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. *5th International Conference on Learning Representations* .

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal* 12(2):23–38.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. pages 315–323.

Jinghang Gu, Longhua Qian, and Guodong Zhou. 2016. Chemical-induced disease relation extraction with various linguistic features. *Database* 2016.

Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2017. Chemical-induced disease relation extraction via convolutional neural network. *Database* 2017.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 770–778.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of*

*the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Association for Computational Linguistics, pages 94–99.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6(2):107–116.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC bioinformatics* 9(1):10.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations (ICLR)*. San Diego, California, USA.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics* 4:313–327. https://transacl.org/ojs/index.php/tacl/article/view/885.

Martin Krallinger, Obdulia Rabal, Saber A. Akhondi, Martín Pérez Pérez, Jesús Santamaría, Pérez Gael Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, Erin Van Buel, Akileshwari Chandrasekhar, Marleen Rodenburg, Astrid Laegreid, Marius Doornenbal, Julen Oyarzabal, Analia Lourenço, and Alfonso Valencia. 2017. Overview of the biocreative vi chemical-protein interaction track. *Proceedings of the BioCreative VI Workshop* page 140.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics* 7(S1):S2.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Scott Winters, and Pete White. 2012. Integrated annotation for biomedical information extraction. *HLT/NAACL Workshop: Biolink* .

Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics* 18(1):198.

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. 2016a. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database* 2016.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016b. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1445–1455. http://www.aclweb.org/anthology/P16-1137.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 2124–2133. http://www.aclweb.org/anthology/P16-1200.

Sijia Liu, Feichen Shen, Yanshan Wang, Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Vipin Chaundary, and Hongfang Liu. 2017. Attention-based neural networks for chemical protein relation extraction. *Proceedings of the BioCreative VI Workshop* .

Emily K Mallory, Ce Zhang, Christopher Ré, and Russ B Altman. 2015. Large-scale extraction of gene interactions from full-text literature using deepdive. *Bioinformatics* 32(1):106–113.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

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*. Association for Computational Linguistics, Suntec, Singapore, pages 1003–1011. http://www.aclweb.org/anthology/P/P09/P09-1113.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1105–1116. http://www.aclweb.org/anthology/P16-1105.

Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. 2015. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807* .

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th international conference on machine*

*learning (ICML-11)*. Bellevue, Washington, USA, pages 809–816.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics* 5:101–115.

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):53.

Hoifung Poon, Kristina Toutanova, and Chris Quirk. 2014. Distant supervision for cancer pathway extraction from text. In *Pacific Symposium on Biocomputing Co-Chairs*. pages 120–131.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics* 8(1):50.

Chris Quirk and Hoifung Poon. 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*. Association for Computational Linguistics, Valencia, Spain, pages 1171–1182.

Sebastian Riedel, David McClosky, Mihai Surdeanu, Andrew McCallum, and Christopher D. Manning. 2011. Model combination for event extraction in bionlp 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*. Association for Computational Linguistics, Portland, Oregon, USA, pages 51–55. http://www.aclweb.org/anthology/W11-1808.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. *Machine learning and knowledge discovery in databases* pages 148–163.

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

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)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 341–350. http://www.aclweb.org/anthology/S13-2056.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* .

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15(1):1929–1958.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, pages 455–465. http://www.aclweb.org/anthology/D12-1042.

Kumutha Swampillai and Mark Stevenson. 2011. Extracting relations within and across sentences. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. RANLP 2011 Organising Committee, Hissar, Bulgaria, pages 25–32. http://aclweb.org/anthology/R11-1004.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*. pages 2071–2080.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* .

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016a. Multilingual relation extraction using compositional universal schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 886–896. http://www.aclweb.org/anthology/N16-1103.

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

Patrick Verga and Andrew McCallum. 2016. Row-less universal schema. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*. Association for Computational Linguistics, San Diego, CA, pages 63–68. http://www.aclweb.org/anthology/W16-1312.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research* 41. https://doi.org/10.1093/nar/gkt441.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database* 2016.

Yadollah Yaghoobzadeh, Heike Adel, and Hinrich Schütze. 2017. Noise mitigation for neural entity typing and relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 1183–1194. http://www.aclweb.org/anthology/E17-1111.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference for Learning Representations (ICLR)*. San Diego, California, USA.

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*. Association for Computational Linguistics, Lisbon, Portugal, pages 1753–1762. http://aclweb.org/anthology/D15-1203.

Huiwei Zhou, Huijie Deng, Long Chen, Yunlong Yang, Chen Jia, and Degen Huang. 2016a. Exploiting syntactic and semantics information for chemical–disease relation extraction. *Database* 2016.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016b. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 207–212. http://anthology.aclweb.org/P16-2034.

# A Implementation Details

The model is implemented in Tensorflow (Abadi et al., 2015) and trained on a single TitanX gpu. The number of transformer block repeats is $B = 2$ . We optimize the model using Adam (Kingma and Ba, 2015) with best parameters chosen for $\epsilon$, $\beta_1$, $\beta_2$ chosen from the development set. The learning rate is set to 0.0005 and batch size 32. In all of our experiments we set the number of attention heads to $h = 4$.

We clip the gradients to norm 10 and apply noise to the gradients (Neelakantan et al., 2015). We tune the decision threshold for each relation type separately and perform early stopping on the development set. We apply dropout (Srivastava et al., 2014) to the input layer randomly replacing words with a special UNK token with keep probability .85. We additionally apply dropout to the input $T$ (word embedding + position embedding), interior layers, and final state. At each step, we randomly sample a positive or negative (NULL class) minibatch with probability 0.5.

## A.1 Chemical Disease Relations Dataset

Token embeddings are pre-trained using skipgram (Mikolov et al., 2013) over a random subset of 10% of all PubMed abstracts with window size 10 and 20 negative samples. We merge the train and development sets and randomly take 850 abstracts for training and 150 for early stopping. Our reported results are averaged over 10 runs and using different splits. All baselines train on both the train and development set. Models took between 4 and 8 hours to train.

$\epsilon$ was set to 1e-4, $\beta_1$ to .1, and $\beta_2$ to 0.9. Gradient noise $\eta = .1$. Dropout was applied to the word embeddings with keep probability 0.85, internal layers with 0.95 and final bilinear projection with 0.35 for the standard CRD dataset experiments. When adding the additional weakly labeled data: word embeddings with keep probability 0.95, internal layers with 0.95 and final bilinear projection with 0.5.

## A.2 Chemical Protein Relations Dataset

We construct our byte-pair encoding vocabulary using a budget of 7500. The dataset contains annotations for a larger set of relation types than are used in evaluation. We train on only the relation types in the evaluation set and set the remaining types to the Null relation. The embedding dimension is set to 200 and all embeddings are randomly initialized. $\epsilon$ was set to 1e-8, $\beta_1$ to .1, and $\beta_2$ to 0.9. Gradient noise $\eta = 1.0$. Dropout was applied to the word embeddings with keep probability 0.5, internal layers with 1.0 and final bilinear projection with 0.85 for the standard CRD dataset experiments.

## A.3 Full CTD Dataset

We tune separate decision boundaries for each relation type on the development set. For each prediction, the relation type with the maximum probability is assigned. If the probability is below the relation specific threshold, the prediction is set to NULL. We use embedding dimension 128 with all embeddings randomly initialized. Our byte pair encoding vocabulary is constructed with a budget of 50,000. Models took 1 to 2 days to train.

$\epsilon$ was set to 1e-4, $\beta_1$ to .1, and $\beta_2$ to 0.9. Gradient noise $\eta = .1$.Dropout was applied to the word embeddings with keep probability 0.95, internal layers with 0.95 and final bilinear projection with 0.5