Joint Biomedical Entity and Relation Extraction with Knowledge-Enhanced Collective Inference

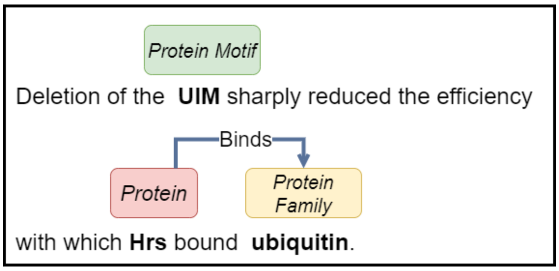
Tuan Lai [1], Heng Ji [1], ChengXiang Zhai [1], Quan Hung Tran [2]
[1]University of Illinois at Urbana-Champaign [2]Adobe Research
{tuanml2, hengji, czhai}@illinois.edu
qtran@adobe.com

Abstract

Compared to the general news domain, information extraction (IE) from biomedical text requires much broader domain knowledge. However, many previous IE methods do not utilize any external knowledge during inference. Due to the exponential growth of biomedical publications, models that do not go beyond their fixed set of parameters will likely fall behind. Inspired by how humans look up relevant information to comprehend a scientific text, we present a novel framework that utilizes external knowledge for joint entity and relation extraction named **KECI** (Knowledge-Enhanced Collective Inference). Given an input text, KECI first constructs an initial span graph representing its initial understanding of the text. It then uses an entity linker to form a knowledge graph containing relevant background knowledge for the the entity mentions in the text. To make the final predictions, KECI fuses the initial span graph and the knowledge graph into a more refined graph using an attention mechanism. KECI takes a collective approach to link mention spans to entities by integrating global relational information into local representations using graph convolutional networks. Our experimental results show that the framework is highly effective, achieving new state-of-the-art results in two different benchmark datasets: BioRelEx (binding interaction detection) and ADE (adverse drug event extraction). For example, KECI achieves absolute improvements of 4.59% and 4.91% in F1 scores over the state-of-the-art on the BioRelEx entity and relation extraction tasks [1].

1 Introduction

With the accelerating growth of biomedical publications, it has become increasingly challenging to manually keep up with all the latest articles. As a result, developing methods for automatic extraction of biomedical entities and their relations has attracted much research attention recently (Li et al., 2017; Fei et al., 2020; Luo et al., 2020). Many related tasks and datasets have been introduced, ranging from binding interaction detection (BioRelEx) (Khachatrian et al., 2019) to adverse drug event extraction (ADE) (Gurulingappa et al., 2012).



Deletion of the **UIM** sharply reduced the efficiency with which **Hrs** bound **ubiquitin**.

Figure 1: An example in the BioRelEx dataset. **UIM** is an abbreviation of "Ubiquitin-Interacting Motif". Our baseline SciBERT model incorrectly predicts the mention as a "DNA" instead of a "Protein Motif".

Many recent joint models for entity and relation extraction rely mainly on distributional representations and do not utilize any external knowledge source (Eberts and Ulges, 2020; Ji et al., 2020; Zhao et al., 2020). However, different from the general news domain, information extraction for the biomedical domain typically requires much broader domain-specific knowledge. Biomedical documents, either formal (e.g., scientific papers) or informal ones (e.g., clinical notes), are written for domain experts. As such, they contain many highly specialized terms, acronyms, and abbreviations. In the BioRelEx dataset, we find that about 65% of the annotated entity mentions are abbreviations of biological entities, and an example is shown in Figure 1. These unique characteristics bring great challenges to general-domain systems and even to existing scientific language models that do not use any external knowledge base during

[1] All programs, data and resources will be made publicly available for research purposes.

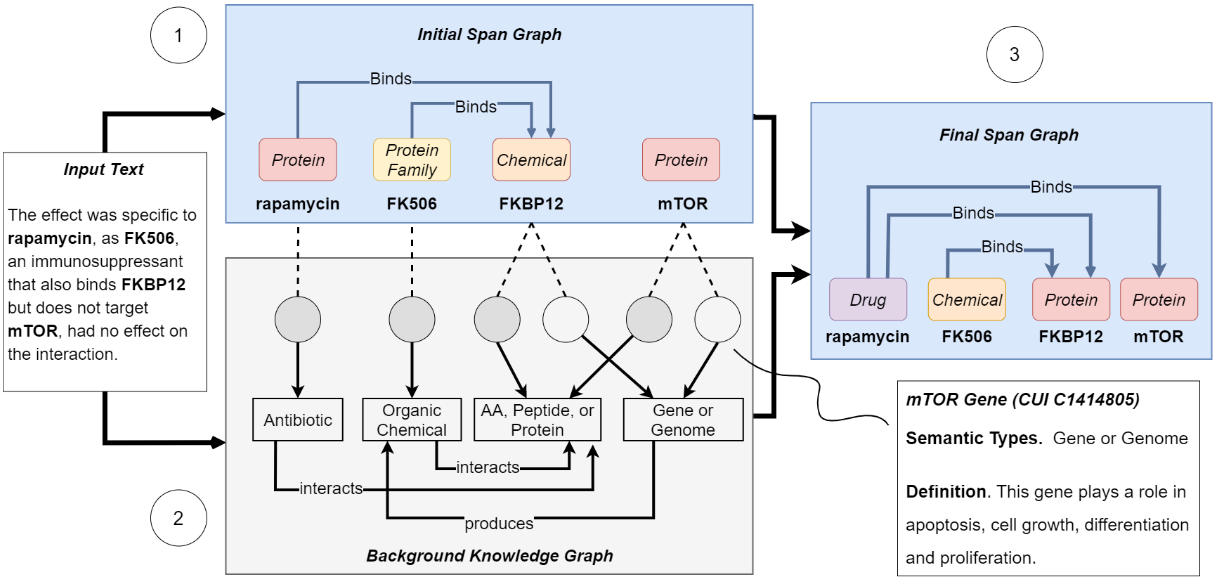

1
Initial Span Graph
Binds
Binds
Protein
Protein Family
Chemical
Protein
rapamycin
FK506
FKBP12
mTOR
Input Text
The effect was specific to **rapamycin**, as **FK506**, an immunosuppressant that also binds **FKBP12** but does not target **mTOR**, had no effect on the interaction.
Antibiotic
Organic Chemical
AA, Peptide, or Protein
Gene or Genome
interacts
interacts
produces
Background Knowledge Graph
2
3
Final Span Graph
Binds
Binds
Binds
Drug
Chemical
Protein
Protein
rapamycin
FK506
FKBP12
mTOR
mTOR Gene (CUI C1414805)
Semantic Types. Gene or Genome
Definition. This gene plays a role in apoptosis, cell growth, differentiation and proliferation.

Figure 2: KECI operates in three main steps: (1) initial span graph construction (2) background knowledge graph construction (3) fusion of these two graphs into a final span graph. KECI takes a *collective* approach to link multiple mentions simultaneously to entities by incorporating *global* relational information using GCNs.

inference (Beltagy et al., 2019; Lee et al., 2019). For example, even though SciBERT (Beltagy et al., 2019) was pretrained on 1.14M scientific papers, our baseline SciBERT model still incorrectly predicts the type of the term *UIM* in Figure 1 to be "DNA", which should be a "Protein Motif" instead. Since the biomedical literature is expanding at an exponential rate, models that do not go beyond their fixed set of parameters will likely fall behind.

In this paper, we introduce **KECI** (Knowledge-Enhanced Collective Inference), a novel end-to-end framework that utilizes external domain knowledge for joint entity and relation extraction. Inspired by how humans comprehend a complex piece of scientific text, the framework operates in three main steps (Figure 2). KECI first reads the input text and constructs an initial *span graph* representing its initial understanding of the text. In a span graph, each node represents a (predicted) entity mention, and each edge represents a (predicted) relation between two entity mentions. KECI then uses an entity linker to form a background knowledge graph containing all potentially relevant biomedical entities from an external knowledge base (KB). For each entity, we extract its semantic types, its definition sentence, and its relational information from the external KB. Finally, KECI uses an attention mechanism to fuse the initial span graph and the background knowledge graph into a more refined graph representing the final output. Different from previous methods that link mentions to entities based solely on *local* contexts (Li et al., 2020), our framework takes a more collective approach to link multiple semantically related mentions simultaneously by leveraging global topical coherence. Our hypothesis is that if multiple mentions co-occur in the same discourse and they are probably semantically related, their reference entities should also be connected in the external KB. KECI integrates *global* relational information into mention and entity representations using graph convolutional networks (GCNs) before linking.

The benefit of collective inference can be illustrated by the example shown in Figure 2. The entity linker proposes two candidate entities for the mention *FKBP12*; one is of semantic type "AA, Peptide, or Protein" and the other is of semantic type "Gene or Genome". It can be tricky to select the correct candidate as *FKBP12* is already tagged with the wrong type in the initial span graph (i.e., it is predicted to be a "Chemical" instead of a "Protein"). However, because of the structural resemblance between the mention-pair ⟨*FK506*, *FKBP12*⟩ and the pair ⟨"Organic Chemical", "AA, Peptide, or Protein"⟩, KECI will link *FKBP12* to the entity of semantic type "AA, Peptide, or Protein". As a result, the final predicted type of *FKBP12* will also be corrected to "Protein" in the final span graph.

Our extensive experimental results show that the proposed framework is highly effective, achieving new state-of-the-art biomedical entity and relation extraction performance on two benchmark datasets: BioRelEx (Khachatrian et al., 2019) and ADE (Gurulingappa et al., 2012). For example, KECI achieves absolute improvements of 4.59% and 4.91% in F1 scores over the state-of-the-art on the BioRelEx entity and relation extraction tasks. Our analysis also shows that KECI can automatically learn to select relevant candidate entities without any explicit entity linking supervision during training. Furthermore, because KECI considers text spans as the basic units for prediction, it can extract nested entity mentions.

2 Methods

2.1 Overview

KECI considers text spans as the basic units for feature extraction and prediction. This design choice allows us to handle nested entity mentions (Sohrab and Miwa, 2018). Also, joint entity and relation extraction can be naturally formulated as the task of extracting a *span graph* from an input document (Luan et al., 2019). In a span graph, each node represents a (predicted) entity mention, and each edge represents a (predicted) relation between two entity mentions.

Given an input document D, KECI first enumerates all the spans (up to a certain length) and embeds them into feature vectors (Sec. 2.2). With these feature vectors, KECI predicts an initial span graph and applies a GCN to integrate initial relational information into each span representation (Sec. 2.3). KECI then uses an entity linker to build a background knowledge graph and applies another GCN to encode each node of the graph (Sec. 2.4). Finally, KECI aligns the nodes of the initial span graph and the background knowledge graph to make the final predictions (Sec. 2.5). We train KECI in an end-to-end manner without using any additional entity linking supervision (Sec. 2.6).

2.2 Span Encoder

Our model first constructs a contextualized representation for each input token using SciBERT (Beltagy et al., 2019). Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ be the output of the token-level encoder, where n denotes the number of tokens in D. Then, for each span s_i whose length is not more than L, we compute its span representation $\mathbf{s}_i \in \mathbb{R}^d$ as:

$$\mathbf{s}_i = \mathrm{FFNN_g}\big(\big[\mathbf{x}_{\mathrm{START}(i)}, \mathbf{x}_{\mathrm{END}(i)}, \hat{\mathbf{x}}_i, \phi(\mathbf{s}_i)\big]\big) \quad (1)$$

where $\mathrm{START}(i)$ and $\mathrm{END}(i)$ denote the start and end indices of s_i respectively. $\mathbf{x}_{\mathrm{START}(i)}$ and $\mathbf{x}_{\mathrm{END}(i)}$ are the boundary token representations. $\hat{\mathbf{x}}_i$ is an attention-weighted sum of the token representations in the span (Lee et al., 2017). $\phi(\mathbf{s}_i)$ is a feature vector denoting the span length. $\mathrm{FFNN_g}$ is a feedforward network with ReLU activations.

2.3 Initial Span Graph Construction

With the extracted span representations, we predict the type of each span and also the relation between each span pair jointly. Let E denote the set of entity types (including non-entity), and R denote the set of relation types (including non-relation). We first classify each span s_i:

$$\mathbf{e}_i = \mathtt{Softmax}\big(\mathrm{FFNN}_e(\mathbf{s}_i)\big) \quad (2)$$

where FFNN_e is a feedforward network mapping from $\mathbb{R}^d \to \mathbb{R}^{|E|}$. We then employ another network to classify the relation of each span pair $\langle s_i, s_j \rangle$:

$$\mathbf{r}_{ij} = \mathtt{Softmax}\big(\mathrm{FFNN}_r\big(\big[\mathbf{s}_i, \mathbf{s}_j, \mathbf{s}_i \circ \mathbf{s}_j\big]\big)\big) \quad (3)$$

where $\circ$ denotes the element-wise multiplication, FFNN_r is a mapping from $\mathbb{R}^{3\times d} \to \mathbb{R}^{|R|}$. We will use the notation $\mathbf{r}_{ij}[k]$ to refer to the predicted probability of s_i and s_j having the relation k.

At this point, one can already obtain a valid output for the task from the predicted entity and relation scores. However, these predictions are based solely on the local document context, which can be difficult to understand without any external domain knowledge. Therefore, our framework uses these predictions only to construct an initial span graph that will be refined later based on information extracted from an external knowledge source.

To maintain computational efficiency, we first prune out spans of text that are unlikely to be entity mentions. We only keep up to λn spans with the lowest probability scores of being a non-entity. The value of λ is selected empirically and set to be 0.5. Spans that pass the filter are represented as nodes in the initial span graph. For every span pair $\langle s_i, s_j \rangle$, we create $|R|$ directed edges from the node representing s_i to the node representing s_j. Each edge represents one relation type and is weighted by the corresponding probability score in $\mathbf{r}_{ij}$.

Let $G_s = \{V_s, E_s\}$ denote the initial span graph. We use a bidirectional GCN (Marcheggiani and Titov, 2017; Fu et al., 2019) to recursively update each span representation:

$$\vec{\mathbf{h}}_i^l = \sum_{s_j \in V_s \setminus \{s_i\}} \sum_{k \in R} \mathbf{r}_{ij}[k] \Big(\vec{\mathbf{W}}_k^{(l)} \mathbf{h}_j^l + \vec{\mathbf{b}}_k^{(l)} \Big)$$
$$\overleftarrow{\mathbf{h}}_i^l = \sum_{s_j \in V_s \setminus \{s_i\}} \sum_{k \in R} \mathbf{r}_{ji}[k] \Big(\overleftarrow{\mathbf{W}}_k^{(l)} \mathbf{h}_j^l + \overleftarrow{\mathbf{b}}_k^{(l)} \Big)$$
$$\mathbf{h}_i^{l+1} = \mathbf{h}_i^l + \text{FFNN}_a^{(l)} \Big(\texttt{ReLU} \Big([\vec{\mathbf{h}}_i^l, \overleftarrow{\mathbf{h}}_i^l] \Big) \Big) \tag{4}$$

where $\mathbf{h}_i^l$ is the hidden feature vector of span s_i at layer l. We initialize $\mathbf{h}_i^0$ to be $\mathbf{s}_i$ (Eq. 1). $\text{FFNN}_a^{(l)}$ is a feedforward network whose output dimension is the same as the dimension of $\mathbf{h}_i^l$.

After multiple iterations of message passing, each span representation will contain the global relational information of G_s. Let $\mathbf{h}_i$ denote the feature vector at the final layer of the GCN. Note that the dimension of $\mathbf{h}_i$ is the same as the dimension of $\mathbf{s}_i$ (i.e., $\mathbf{h}_i \in \mathbb{R}^d$).

2.4 Background Knowledge Graph Construction

In this work, we utilize external knowledge from the Unified Medical Language System (UMLS) (Bodenreider, 2004). UMLS consists of three main components: Metathesaurus, Semantic Network, and Specialist Lexicon and Lexical Tools. The Metathesaurus provides information about millions of fine-grained biomedical concepts and relations between them. To be consistent with the existing literature on knowledge graphs, we will refer to UMLS concepts as entities. Each entity is annotated with one or more higher-level semantic types, such as *Anatomical Structure*, *Cell*, or *Virus*. In addition to relations between entities, there are also semantic relations between semantic types. For example, there is an *affects* relation from *Acquired Abnormality* to *Physiologic Function*. This information is provided by the Semantic Network.

We first extract UMLS biomedical entities from the input document D using MetaMap, an entity mapping tool for UMLS (Aronson and Lang, 2010). We then construct a background knowledge graph (KG) from the extracted information. More specifically, we first create a node for every extracted biomedical entity. The semantic types of each entity node are also modeled as type nodes that are linked with associated entity nodes. Finally, we create an edge for every relevant relation found in the Metathesaurus and the Semantic Network. An example KG is in the grey shaded region of Figure 2. Circles represent entity nodes, and rectangles represent nodes that correspond to semantic types.

Note that we simply run MetaMap with the default options and do not tune it. In our experiments, we found that MetaMap typically returns many candidate entities unrelated to the input text. However, as to be discussed in Section 3.4, we show that KECI can learn to ignore the irrelevant entities.

Let $G_k = \{V_k, E_k\}$ denote the constructed background KG, where V_k and E_k are the node and edge sets, respectively. We use a set of UMLS embeddings pretrained by Maldonado et al. (2019) to initialize the representation of each node in V_k. We also use SciBERT to encode the UMLS definition sentence of each node into a vector and concatenate it to the initial representation. After that, since G_k is a heterogeneous relational graph, we use a relational GCN (Schlichtkrull et al., 2018) to update the representation of each node v_i:

$$\mathbf{v}_i^{l+1} = \texttt{ReLU} \Big(\mathbf{U}^{(l)} \mathbf{v}_i^l + \sum_{k \in R} \sum_{v_j \in N_i^k} \Big(\frac{1}{c_{i,k}} \mathbf{U}_k^{(l)} \mathbf{v}_j^l \Big) \Big) \tag{5}$$

where $\mathbf{v}_i^l$ is the feature vector of v_i at layer l. N_i^k is the set of neighbors of v_i under relation $k \in R$. $c_{i,k}$ is a normalization constant and set to be $|N_i^k|$.

After multiple iterations of message passing are performed, the global relational information of the KG will be integrated into each node's representation. Let $\mathbf{v}_i$ denote the feature vector at the final layer of the relational GCN. We further project each vector $\mathbf{v}_i$ to another vector $\mathbf{n}_i$ using a simple feed-forward network, so that $\mathbf{n}_i$ has the same dimension as the span representations (i.e., $\mathbf{n}_i \in \mathbb{R}^d$).

2.5 Final Span Graph Prediction

At this point, we have two graphs: the initial span graph $G_s = \{V_s, E_s\}$ (Sec. 2.3) and the background knowledge graph $G_k = \{V_k, E_k\}$ (Sec. 2.4). We have also obtained a structure-aware representation for each node in each graph (i.e., $\mathbf{h}_i$ for each span $s_i \in V_s$ and $\mathbf{n}_j$ for each entity $v_j \in V_k$).

The next step is to soft-align the mentions and the candidate entities using an attention mechanism (Figure 3). Let $C(s_i)$ denote the set of candidate entities for a span $s_i \in V_s$. For example, in Figure 2, the mention *FKBP12* has two candidate entities, while *FK506* has only one candidate. For each candidate entity $v_j \in C(s_i)$, we calculate a scalar

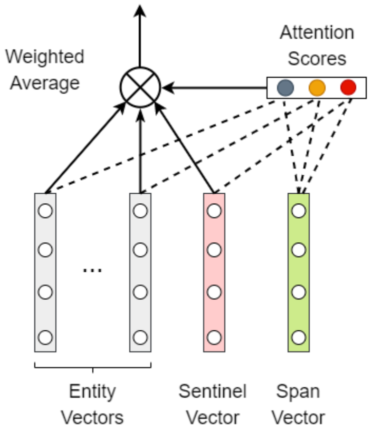


Figure 3: An illustration of the attention mechanism.

score α_{ij} indicating how relevant v_j is to s_i:

$$\alpha_{ij} = \text{FFNN}_\text{c}\big(\big[\mathbf{h}_i, \mathbf{n}_j\big]\big) \tag{6}$$

where FFNN_c is a feedforward network mapping from $\mathbb{R}^{2\times d} \rightarrow \mathbb{R}$. Then we compute an additional sentinel vector $\mathbf{c}_i$ (Yang and Mitchell, 2017; He et al., 2020) and also compute a score α_i for it:

$$\begin{aligned}\mathbf{c}_i &= \text{FFNN}_\text{s}\big(\mathbf{h}_i\big)\\ \alpha_i &= \text{FFNN}_\text{c}\big(\big[\mathbf{h}_i, \mathbf{c}_i\,\big]\big)\end{aligned} \tag{7}$$

where FFNN_s is another feedforward network mapping from $\mathbb{R}^d \rightarrow \mathbb{R}^d$. Intuitively, $\mathbf{c}_i$ records the information of the local context of s_i, and α_i measures the importance of such information. After that, we compute a final knowledge-aware representation $\mathbf{f}_i$ for each span s_i as follows:

$$\begin{aligned}&Z = \exp\left(\alpha_i\right) + \sum_{v_z \in C(s_i)} \exp\left(\alpha_{iz}\right)\\ &\beta_i = \exp\left(\alpha_i\right)/Z \quad \text{and} \quad \beta_{ij} = \exp\left(\alpha_{ij}\right)/Z\\ &\mathbf{f}_i = \beta_i\,\mathbf{c}_i + \sum_{v_j \in C(s_i)} \beta_{ij}\mathbf{n}_j\end{aligned} \tag{8}$$

The attention mechanism is illustrated in Figure 3.

With the extracted knowledge-aware span representations, we predict the final span graph in a way similar to Eq. 2 and Eq. 3:

$$\begin{aligned}\widehat{\mathbf{e}_i} &= \texttt{Softmax}\big(\text{FFNN}_{\widehat{e}}(\mathbf{f}_i)\big)\\ \widehat{\mathbf{r}_{ij}} &= \texttt{Softmax}\big(\text{FFNN}_{\widehat{r}}(\big[\mathbf{f}_i, \mathbf{f}_j, \mathbf{f}_i \circ \mathbf{f}_j\big]\big)\big)\end{aligned} \tag{9}$$

where $\text{FFNN}_{\widehat{e}}$ is a mapping from $\mathbb{R}^d \rightarrow \mathbb{R}^{|E|}$, and $\text{FFNN}_{\widehat{r}}$ is a mapping from $\mathbb{R}^{3\times d} \rightarrow \mathbb{R}^{|R|}$. $\widehat{\mathbf{e}_i}$ is the *final* predicted probability distribution over possible entity types for span s_i. $\widehat{\mathbf{r}_{ij}}$ is the *final* predicted probability distribution over possible relation types for span pair $\langle s_i, s_j\rangle$.

2.6 Training

The total loss is computed as:

$$\mathcal{L}_{total} = (\mathcal{L}_1^e + \mathcal{L}_1^r) + 2(\mathcal{L}_2^e + \mathcal{L}_2^r) \tag{10}$$

where $\mathcal{L}_*^e$ denotes the cross-entropy loss of span classification. $\mathcal{L}_*^r$ denotes the binary cross-entropy loss of relation classification. $\mathcal{L}_1^e$ and $\mathcal{L}_1^r$ are loss terms for the initial span graph prediction (Eq. 2 and Eq. 3 of Section 2.3). $\mathcal{L}_2^e$ and $\mathcal{L}_2^r$ are loss terms for the final span graph prediction (Eq. 9 of Section 2.5). We apply a larger weight score to the loss terms $\mathcal{L}_2^e$ and $\mathcal{L}_2^r$. We train the framework using only ground-truth labels of the entity and relation extraction tasks. We do not make use of any entity linking supervision in this work.

3 Experiments and Results

3.1 Data and Experiments Setup

Datasets and evaluation metrics We evaluate KECI on two benchmark datasets: BioRelEx and ADE. The **BioRelEx** dataset (Khachatrian et al., 2019) consists of 2,010 sentences from biomedical literature that capture binding interactions between proteins and/or biomolecules. BioRelEx has annotations for 33 types of entities and relations for binding interactions. The training, development, and test splits contain 1,405, 201, and 404 sentences, respectively. The training and development sets are publicly available. The test set is unreleased and can only be evaluated against using CodaLab [2]. For BioRelEx, we report Micro-F1 scores. The **ADE** dataset (Gurulingappa et al., 2012) consists of 4,272 sentences extracted from medical reports that describe drug-related adverse effects. Two entity types (*Adverse-Effect* and *Drug*) and a single relation type (*Adverse-Effect*) are pre-defined. Similar to previous work (Eberts and Ulges, 2020; Ji et al., 2020), we conduct 10-fold cross-validation and report averaged Macro-F1 scores. All the reported results take overlapping entities into consideration.

Implementation details We implement KECI using PyTorch (Paszke et al., 2019) and Huggingface's Transformers (Wolf et al., 2020). KECI uses SciBERT as the Transformer encoder (Beltagy et al., 2019). All details about hyperparameters and reproducibility information are in the appendix.

[2] `https://competitions.codalab.org/competitions/20468`

Model	Entity (Micro-F1)	Relation (Micro-F1)
SciIE (2018)	77.90	49.60
DYGIEPP + ELMo (2020)	81.10	55.60
DYGIEPP + BioELMo (2020)	82.80	54.80
SentContextOnly	83.98	63.90
FlatAttention	84.32	64.23
KnowBertAttention	85.69	65.13
Full Model (KECI)	**87.42**	**66.09**

Table 1: Overall results (%) on the development set of BioRelEx.

Model	Entity (Micro-F1)	Relation (Micro-F1)
SciIE (2018)	73.56	50.15
Second Best Model	82.76	62.18
Full Model (KECI)	**87.35**	**67.09**

Table 2: Overall results (%) on the test set of BioRelEx (from the leaderboard as of January 20th, 2021).

Baselines for comparison In addition to comparing our method with state-of-the-art methods on the above two datasets, we implement the following baselines for further comparison and analysis:

1. **SentContextOnly**: This baseline does not use any *external knowledge*. It uses only the local sentence context for prediction. It extracts the final output directly from the predictions obtained using Eq. 2 and Eq. 3.
2. **FlatAttention**: This baseline does not rely on *collective inference*. It does not integrate any global relational information into mention and entity representations. Each $\mathbf{h}_i$ mentioned in Sec. 2.3 is set to be $\mathbf{s}_i$ (Eq. 1), and each $\mathbf{v}_i$ mentioned in Sec. 2.4 is set to be $\mathbf{v}_i^0$. Then, the prediction of the final span graph is the same as described in Sec. 2.5.
3. **KnowBertAttention**: This baseline uses the Knowledge Attention and Recontextualization (KAR) mechanism of KnowBert (Peters et al., 2019), *a state-of-the-art knowledge-enhanced language model*. The baseline first uses SciBERT to construct initial token-level representations. It then uses the KAR mechanism to inject external knowledge from UMLS into the token-level vectors. Finally, it embeds text spans into feature vectors (Eq. 1) and uses the span representations to extract entities and relations in one pass (similar to Eq. 9).

For fair comparison, all the baselines use SciBERT as the Transformer encoder.

Model	Entity (Macro-F1)	Relation (Macro-F1)
Relation-Metric (2019)	87.11	77.29
SpERT (2020)	89.28	78.84
SPAN$_{\text{Multi-Head}}$ (2020)	90.59	80.73
SentContextOnly	88.13	77.23
FlatAttention	89.16	78.81
KnowBertAttention	90.08	79.95
Full Model (KECI)	**90.67**	**81.74**

Table 3: Overall results (%) on the ADE dataset.

Ablation setting	Entity (Micro-F1)	Relation (Micro-F1)
Full Model (KECI)	**87.42**	**66.09**
• w/o external knowledge	83.98*	63.90*
• w/o collective inference	84.32*	64.23*
• w/o the bidirectional GCN	84.76*	64.25*
• w/o the relational GCN	85.14*	65.32*
• w/o the pretrained UMLS vectors	86.25†	65.29*
• w/o the UMLS definition vectors	86.76†	65.45†

Table 4: Results (%) of ablation experiments on the development set of BioRelEx. We use the symbols * and † to indicate statistical significance with 95% and 90% confidence levels respectively (compared to KECI).

3.2 Overall Results

Table 1 and Table 2 show the overall results on the development and test sets of BioRelEx, respectively. Compared to SentContextOnly, KECI achieves much higher performance. This demonstrates the importance of incorporating external knowledge for biomedical information extraction. KECI also outperforms the baseline FlatAttention by a large margin, which shows the benefit of collective inference. In addition, we see that our model performs better than the baseline KnowBertAttention. Finally, at the time of writing, KECI achieves the first position on the BioRelEx leaderboard [3].

Table 3 shows the overall results on ADE. KECI again outperforms all the baselines and state-of-the-art models such as SpERT (Eberts and Ulges, 2020) and SPAN$_{\text{Multi-Head}}$ (Ji et al., 2020). This further confirms the effectiveness of our framework.

3.3 Ablation Study

Table 4 shows the results of ablation studies we did on the development set of the BioRelEx benchmark. We compare our full model against several partial variants. The variant [w/o external knowledge] is the same as the baseline SentContextOnly, and the variant [w/o collective inference] is the same as the

[3] https://competitions.codalab.org/competitions/20468

Datasets	Top 3 types with the lowest avg. attention scores	Top 3 types with the highest avg. attention scores
BioRelEx	Diagnostic Procedure (0.04); Activity (0.05); Plant (0.05)	Amino Acid, Peptide, or Protein (0.32); Enzyme (0.32); Molecular Function (0.36)
ADE	Intellectual Product (0.15); Idea or Concept (0.19); Temporal Concept (0.19)	Antibiotic (0.78); Organic Chemical (0.79); Nucleic Acid, Nucleoside, or Nucleotide (0.87)

Table 5: Average attention scores of different UMLS semantic types.

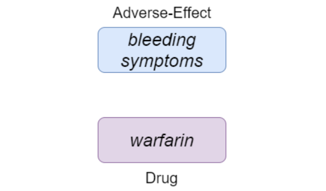
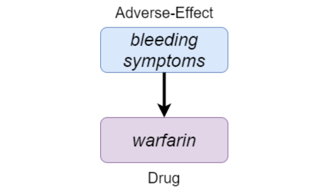
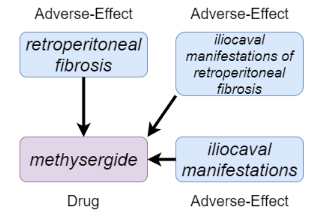
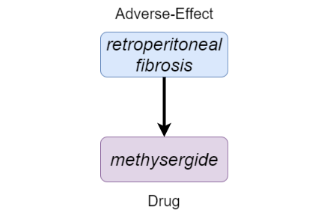
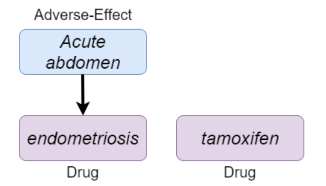
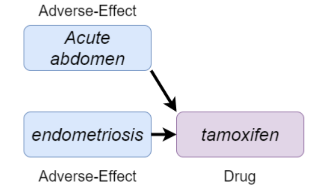
Input Sentence	**Initial Span Graph**	**Final Span Graph**
#1: *Despite the low dosage of warfarin, international normalized ratio (INR) was markedly elevated from 1.15 to 11.28 for only 4 days, and bleeding symptoms concurrently developed.*		
#2: *A 25-year-old woman sought medical attention because of iliocaval manifestations of retroperitoneal fibrosis while she was taking methysergide.*	 Adverse-Effect Adverse-Effect retroperitoneal fibrosis iliocaval manifestations of retroperitoneal fibrosis methysergide iliocaval manifestations Drug Adverse-Effect	 Adverse-Effect retroperitoneal fibrosis methysergide Drug
#3: TITLE: *Acute abdomen due to endometriosis in a premenopausal woman taking tamoxifen.*	 Adverse-Effect Acute abdomen endometriosis tamoxifen Drug Drug	 Adverse-Effect Acute abdomen endometriosis tamoxifen Adverse-Effect Drug

Table 6: Examples showing how external knowledge improves the quality of extracted span graphs. Edges represent relations of type *Adverse-Effect*. Only relations with predicted probabilities of at least 0.5 are shown.

baseline FlatAttention (Section 3.1). For the variant [w/o the bidirectional GCN], we simply set each $\mathbf{h}_i$ mentioned in Section 2.3 to be $\mathbf{s}_i$. Similarly, for the variant [w/o the relational GCN], we set each $\mathbf{v}_i$ in Section 2.4 to be $\mathbf{v}_i^0$. The last two variants are related to the initialization of each vector $\mathbf{v}_i^0$.

We see that all the partial variants perform worse than our full model. This shows that each component of KECI plays an important role.

3.4 Attention Pattern Analysis

There is no gold-standard set of correspondences between the entity mentions in the datasets and the UMLS entities. Therefore, we cannot directly evaluate the entity linking performance of KECI. However, for each UMLS semantic type, we compute the average attention weight that an entity of that type gets assigned (Table 5). Overall, we see that KECI typically pays the most attention to the relevant informative entities while ignoring the irrelevant ones.

3.5 Qualitative Analysis

Table 6 shows some examples from the ADE dataset that illustrate how incorporating external knowledge can improve the performance of joint biomedical entity and relation extraction.

In the first example, initially, there is no edge between the node “bleeding symptoms” and the node “warfarin”, probably because of the distance between their corresponding spans in the original input sentence. However, KECI can link the term “warfarin” to a UMLS entity (CUI: C0043031), and the definition in UMLS says that warfarin is a type of anticoagulant that prevents the formation of blood clots. As the initial feature vector of each entity contains the representation of its definition (Sec. 2.4), KECI can recover the missing edge.

In the second example, the initial span graph is predicted to have three entities of type *Adverse-Effect*, which correspond to three different overlapping text spans. Among these three, only “retroperitoneal fibrosis" can be linked to a UMLS entity. It is also evident from the input sentence that one of these spans is related to "methysergide". As a result, KECI successfully removes the other two unlinked span nodes to create the final span graph.

In the third example, probably because of the phrase "due to", the node "endometriosis" is initially predicted to be of type *Drug*, and the node "acute abdomen" is predicted to be its *Adverse-Effect*. However, KECI can link the term "endometriosis" to a UMLS entity of semantic type *Disease or Syndrome*. As a result, the system can correct the term's type and also predict the right edges for the final span graph.

Finally, we also examined the errors made by KECI. One major issue is that MetaMap sometimes fails to return any candidate entity from UMLS for an entity mention. We leave the extension of this work to using multiple KBs as future work.

4 Related Work

Traditional pipelined methods typically treat entity extraction and relation extraction as two separate tasks (e.g., (Zelenko et al., 2002; Zhou et al., 2005; Chan and Roth, 2011)). Such approaches ignore the close interaction between named entities and their relation information and typically suffer from the error propagation problem. To overcome these limitations, many studies have proposed joint models that perform entity extraction and relation extraction simultaneously (e.g., (Roth and Yih, 2007; Li and Ji, 2014; Li et al., 2017; Zheng et al., 2017; Bekoulis et al., 2018a,b; Wadden et al., 2019; Fu et al., 2019; Luan et al., 2019; Zhao et al., 2020; Wang and Lu, 2020; Li et al., 2020; Lin et al., 2020)). Particularly, span-based joint extraction methods have gained much popularity lately because of their ability to detect overlapping entities. For example, Eberts and Ulges (2020) propose SpERT, a simple but effective span-based model that utilizes BERT as its core. The recent work of Ji et al. (2020) also closely follows the overall architecture of SpERT but differs in span-specific and contextual semantic representations. Despite their impressive performance, these methods are not designed specifically for the biomedical domain, and they do not utilize any external knowledge base. To the best of our knowledge, our work is the first span-based framework that utilizes external knowledge for joint entity and relation extraction from biomedical text.

Biomedical event extraction is a closely related task that has also received a lot of attention from the research community (e.g., (Poon and Vanderwende, 2010; Kim et al., 2013; V S S Patchigolla et al., 2017; Rao et al., 2017; Espinosa et al., 2019; Li et al., 2019; Wang et al., 2020; Huang et al., 2020; Ramponi et al., 2020; Yadav et al., 2020)). Several studies have proposed to incorporate external knowledge from domain-specific KBs into neural models for biomedical event extraction. For example, Li et al. (2019) incorporate entity information from Gene Ontology into tree-LSTM models. However, their approach does not explicitly use any external relational information. Recently, Huang et al. (2020) introduce a framework that uses a novel Graph Edge conditioned Attention Network (GEANet) to utilize domain knowledge from UMLS. In the framework, a global KG for the entire corpus is first constructed, and then a sentence-level KG is created for each individual sentence in the corpus. Our method of KG construction is more flexible as we directly create a KG for each input text. Furthermore, the work of Huang et al. (2020) only deals with event extraction and assumes that gold-standard entity mentions are provided at inference time.

Some previous work has focused on integrating external knowledge into neural architectures for other tasks, such as reading comprehension (Mihaylov and Frank, 2018), natural language inference (Sharma et al., 2019), and conversational modeling (Parthasarathi and Pineau, 2018). Different from these studies, our work explicitly emphasizes the benefit of collective inference using global relational information.

5 Conclusions and Future Work

In this work, we propose a novel span-based framework named KECI that utilizes external domain knowledge for joint entity and relation extraction from biomedical text. Experimental results show that KECI is highly effective, achieving new state-of-the-art results on two datasets: BioRelEx and ADE. Theoretically, KECI can take an entire document as input; however, the tested datasets are only sentence-level datasets. In the future, we plan to evaluate our framework on more document-level datasets. We also plan to explore a broader range of properties and information that can be extracted from external KBs to facilitate biomedical IE tasks. Finally, we also plan to apply KECI to other tasks.

References

A. Aronson and F. Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17 3:229–36.

Giannis Bekoulis, J. Deleu, Thomas Demeester, and Chris Develder. 2018a. Joint entity recognition and relation extraction as a multi-head selection problem. *ArXiv*, abs/1804.07847.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018b. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Abhinav Bhatt and Kaustubh D. Dhole. 2020. Benchmarking biorelex for entity tagging and relation extraction. *ArXiv*, abs/2006.00533.

O. Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 551–560, Portland, Oregon, USA. Association for Computational Linguistics.

Markus Eberts and A. Ulges. 2020. Span-based joint entity and relation extraction with transformer pretraining. In *European Conference on Artificial Intelligence*.

Kurt Junshean Espinosa, Makoto Miwa, and Sophia Ananiadou. 2019. A search-based neural model for biomedical nested and overlapping event detection. 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 3679–3686, Hong Kong, China. Association for Computational Linguistics.

Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. A span-graph neural model for overlapping entity relation extraction in biomedical texts. *Bioinformatics*. Btaa993.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy. Association for Computational Linguistics.

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.

C. Harris, K. J. Millman, S. Walt, Ralf Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, Nathaniel J. Smith, R. Kern, Matti Picus, S. Hoyer, M. Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'io, Mark Wiebe, P. Peterson, Pierre G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, Christoph Gohlke, and T. E. Oliphant. 2020. Array programming with numpy. *Nature*, 585 7825:357–362.

Keqing He, Yuanmeng Yan, and Weiran Xu. 2020. Learning to tag OOV tokens by integrating contextual representation and background knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 619–624, Online. Association for Computational Linguistics.

Kung-Hsiang Huang, Mu Yang, and Nanyun Peng. 2020. Biomedical event extraction with hierarchical knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1277–1285, Online. Association for Computational Linguistics.

Bin Ji, Jie Yu, Shasha Li, Jun Ma, Qingbo Wu, Yusong Tan, and Huijun Liu. 2020. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 88–99, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hrant Khachatrian, Lilit Nersisyan, Karen Hambardzumyan, Tigran Galstyan, Anna Hakobyan, Arsen Arakelyan, Andrey Rzhetsky, and Aram Galstyan. 2019. BioRelEx 1.0: Biological relation extraction benchmark. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 176–190, Florence, Italy. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The Genia event extraction shared task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Diya Li, Lifu Huang, Heng Ji, and Jiawei Han. 2019. Biomedical event extraction based on knowledge-driven tree-LSTM. 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 1421–1430, Minneapolis, Minnesota. Association for Computational Linguistics.

Fei Li, Meishan Zhang, G. Fu, and D. Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, Baltimore, Maryland. Association for Computational Linguistics.

Zhijing Li, Yuchen Lian, Xiaoyong Ma, Xiangrong Zhang, and Chen Li. 2020. Bio-semantic relation extraction with attention-based external knowledge reinforcement. *BMC Bioinformatics*, 21.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. 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 3036–3046, Minneapolis, Minnesota. Association for Computational Linguistics.

Ling Luo, Zhihao Yang, M. Cao, Lei Wang, Y. Zhang, and Hongfei Lin. 2020. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of biomedical informatics*, page 103384.

R. Maldonado, Meliha Yetisgen, and Sanda M. Harabagiu. 2019. Adversarial learning of knowledge embeddings for the unified medical language system. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2019:543–552.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark. Association for Computational Linguistics.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.

Adam Paszke, S. Gross, Francisco Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Alban Desmaison, Andreas Köpf, E. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, B. Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Hoifung Poon and Lucy Vanderwende. 2010. Joint inference for knowledge extraction from biomedical literature. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 813–821, Los Angeles, California. Association for Computational Linguistics.

Alan Ramponi, Rob van der Goot, Rosario Lombardo, and Barbara Plank. 2020. Biomedical event extraction as sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5357–5367, Online. Association for Computational Linguistics.

Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. Biomedical event extraction using Abstract Meaning Representation. In *BioNLP 2017*, pages 126–135, Vancouver, Canada,. Association for Computational Linguistics.

Dan Roth and Wen-tau Yih. 2007. Global inference for entity and relation identification via a linear programming formulation. *Introduction to statistical relational learning*, pages 553–580.

M. Schlichtkrull, Thomas Kipf, P. Bloem, R. V. Berg, Ivan Titov, and M. Welling. 2018. Modeling relational data with graph convolutional networks. *ArXiv*, abs/1703.06103.

Soumya Sharma, Bishal Santra, Abhik Jana, Santosh Tokala, Niloy Ganguly, and Pawan Goyal. 2019. Incorporating domain knowledge into medical NLI using knowledge 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 6092–6097, Hong Kong, China. Association for Computational Linguistics.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium. Association for Computational Linguistics.

T. Tran and Ramakanth Kavuluru. 2019. Neural metric learning for fast end-to-end relation extraction. *ArXiv*, abs/1905.07458.

Rahul V S S Patchigolla, Sunil Sahu, and Ashish Anand. 2017. Biomedical event trigger identification using bidirectional recurrent neural network based models. In *BioNLP 2017*, pages 316–321, Vancouver, Canada,. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. 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 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online. Association for Computational Linguistics.

Xing David Wang, Leon Weber, and Ulf Leser. 2020. Biomedical event extraction as multi-turn question answering. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 88–96, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shweta Yadav, Pralay Ramteke, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Exploring disorder-aware attention for clinical event extraction. 16(1s).

Bishan Yang and Tom Mitchell. 2017. Leveraging knowledge bases in LSTMs for improving machine reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1436–1446, Vancouver, Canada. Association for Computational Linguistics.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 71–78. Association for Computational Linguistics.

Shan Zhao, Minghao Hu, Zhiping Cai, and Fang Liu. 2020. Modeling dense cross-modal interactions for joint entity-relation extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4032–4038. International Joint Conferences on Artificial Intelligence Organization. Main track.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada. Association for Computational Linguistics.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan. Association for Computational Linguistics.

A Reproducibility Checklist

In this section, we present the reproducibility information of the paper. We are planning to make the code publicly available after the paper is reviewed.

Implementation Dependencies Libraries Pytorch 1.6.0 (Paszke et al., 2019), Transformers 4.0.0 (Wolf et al., 2020), DGL 0.5.3[4], Numpy 1.19.1 (Harris et al., 2020), CUDA 10.2.

[4] https://www.dgl.ai/

Computing Infrastructure The experiments were conducted on a server with Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz and NVIDIA Tesla V100 GPUs. The allocated RAM is 187G. GPU memory is 16G.

Datasets The BioRelEx dataset (Khachatrian et al., 2019) is available at `https://github.com/YerevaNN/BioRelEx`. The ADE dataset (Gurulingappa et al., 2012) can be downloaded by using the script at `https://github.com/markus-eberts/spert`.

Average Runtime Table 7 shows the estimated average run time of our full model.

Number of Model Parameters The number of parameters in a full model trained on BioRelEx is about 121.0M parameters. The number of parameters in a full model trained on ADE is about 119.9M parameters.

Hyperparameters of Best-Performing Models The span length limit L is set to be 20 tokens. The pruning parameter λ is set to be 0.5. All of our models use SciBERT as the Transformer encoder (Beltagy et al., 2019). We use two different learning rates, one for the lower pretrained Transformer encoder and one for the upper layers. Table 8 summarizes the hyperparameter configurations of best-performing models.

Expected Validation Performance The main paper has the results on the dev set of BioRelEx. For ADE, as in previous work, we conduct a 10-fold cross validation.

Hyperparameter Tuning Process We experimented with the following range of possible values: {16, 32} for batch size, {2e-5, 3e-5, 4e-5, 5e-5} for lower learning rate, {1e-4, 2e-4, 5e-4} for upper learning rate, and {50, 100} for number of training epochs. For each particular set of hyperparameters, we repeat training for 3 times and compute the average performance.

Dataset	One Training Epoch	Evaluation (Dev Set)
BioRelEx	337.51 seconds	35.38 seconds
ADE	712.89 seconds	52.39 seconds

Table 7: Estimated average runtime of our full model.

Hyperparameters	BioRelEx	ADE
Lower Learning Rate	5e-05	5e-05
Upper Learning Rate	2e-04	1e-04
Batch Size	32	32
Number Epochs	50	50

Table 8: Hyperparameters for best-performing models.