Fine-grained Information Extraction from Biomedical Literature based on Knowledge-enriched Abstract Meaning Representation

Zixuan Zhang[1], **Nikolaus Parulian**[1], **Heng Ji**[1],
Ahmed S. Elsayed[2], **Skatje Myers**[2], **Martha Palmer**[2]
[1]University of Illinois at Urbana-Champaign
[2]University of Colorado Boulder
{zixuan11, nnp2, hengji}@illinois.edu
{ahmed.s.elsayed, skatje.myers, martha.palmer}@colorado.edu

Abstract

Biomedical Information Extraction from scientific literature presents two unique and non-trivial challenges. First, compared with general natural language texts, sentences from scientific papers usually possess wider contexts between knowledge elements. Moreover, comprehending the fine-grained scientific entities and events urgently requires domain-specific background knowledge. In this paper, we propose a novel biomedical Information Extraction (IE) model to tackle these two challenges and extract scientific entities and events from English research papers. We perform Abstract Meaning Representation (AMR) to compress the wide context to uncover a clear semantic structure for each complex sentence. Besides, we construct the sentence-level knowledge graph from an external knowledge base and use it to enrich the AMR graph to improve the model's understanding of complex scientific concepts. We use an edge-conditioned graph attention network to encode the knowledge-enriched AMR graph for biomedical IE tasks. Experiments on the GENIA 2011 dataset show that the AMR and external knowledge have contributed 1.8% and 3.0% absolute F-score gains respectively. In order to evaluate the impact of our approach on real-world problems that involve topic-specific fine-grained knowledge elements, we have also created a new ontology and annotated corpus for entity and event extraction for the COVID-19 scientific literature, which can serve as a new benchmark for the biomedical IE community.[1]

1 Introduction

The task of Biomedical Information Extraction (IE) aims to extract structured knowledge from biomedical literature, which is usually represented by an information network composed of scientific named entities, relations, and key events. It is an essential task for accelerating practical applications of the results and achievements from scientific research. For example, practical progress on combating COVID-19 depends highly on efficient transmission, assessment and extension of cutting-edge scientific research discovery (Wang et al., 2020a; Lybarger et al., 2020; Möller et al., 2020). In this scenario, a powerful biomedical IE system will be able to create a dynamic knowledge base from the surging number of relevant papers, making it more efficient to get access to the latest knowledge and use it for scientific discovery, as well as diagnosis and treatment of patients.

IE from biomedical scientific papers presents two unique and non-trivial challenges. First, the authors of scientific papers tend to compose long sentences, where the event triggers and entity mentions are usually located far away from each other within the sentence. As shown in Table 1, we can see that compared to the ACE05 dataset in news domain, the average distance between triggers and entities is much longer in biomedical scientific papers. Therefore, it is more difficult for IE models to capture the global context with only flat sequential sentence encoders such as BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019).

Dataset	Average distance	Maximal distance
ACE05-E	0.212 sentence	56 words
GENIA-2011	0.330 sentence	77 words

Table 1: Comparison of the average and maximum distance between each event-argument pair in news domain (ACE-05 dataset) and scientific papers (GENIA-2011 dataset) with the same sentence tokenizer.

Moreover, comprehending sentences from scientific papers urgently requires external knowledge, because there are a number of domain-specific un-

[1]Data and source code are publicly available at https://github.com/zhangzx-uiuc/Knowledge-AMR.

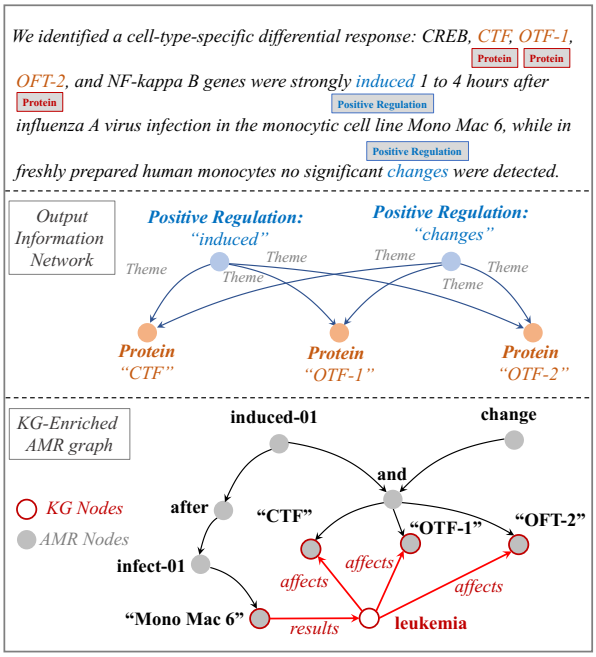

Figure 1: An illustrating example with IE annotations and the KG-enriched AMR graph from GENIA corpus. Note that we only include part of the KG-enriched AMR graph for conciseness.

explained common expressions, acronyms, and abbreviations that are difficult for the model to understand. For instance, as shown in Figure 1, it is nearly impossible for a typical end-to-end model, which only takes in the sentence as input, to get clear understanding of *CTF*, *OTF-1*, and *OTF-2* without background knowledge. Moreover, the complex biomedical and chemical interactions between multifarious chemicals, genes, and proteins are even harder to understand in addition to the entities themselves.

To tackle these two challenges, we propose a novel framework for biomedical IE that integrates Abstract Meaning Representation (AMR) (Banarescu et al., 2013) and external knowledge graphs. AMR is a semantic representation language that converts the meaning of each input sentence into a rooted, directed, labeled, acyclic graph structure. AMR semantic representation includes PropBank (Palmer et al., 2005) frames, non-core semantic roles, coreference, entity typing and linking, modality, and negation. The nodes in AMR are concepts instead of words, and the edge types are much more fine-grained compared with traditional semantic languages like dependency parsing and semantic role labeling. We train a transformer-based AMR semantic parser (Fernandez Astudillo et al., 2020) on biomedical scientific texts and use it in our biomedical IE model. To better handle long sentences with distant trigger and entity pairs, we use AMR parsing to compress each sentence and to better capture global interactions between tokens. For example, as shown in Figure 1, the *Positive Regulation* event trigger *"changes"* is located far away from its arguments *CTF*, *OTF-1*, *OTF-2* in the original sentence. However, in the AMR graph, such trigger-entity pairs are linked within two hops. Therefore, it will be much easier for the model to identify such kinds of events with the guidance of AMR parsing.

In addition, to make better use of the external knowledge, we extract a global knowledge graph from the Comparative Toxicogenomics Database (CTDB) that covers all biomedical entities in the corpus. For each sentence, we select a minimal connected subgraph as the sentence-level KG. We use this sentence KG to enrich AMR nodes and edges to give the model additional prior domain knowledge, especially the biomedical and chemical interactions between different genes and proteins. These fine-grained relations are important for biomedical event extraction. For example, as in Figure 1, the incorporation of the external KG can indicate that *Mono Mac 6* can result in leukemia, which will affect the expression of *CTF*, *OTF-1*, and *OFT-2* proteins. With this external knowledge, it will be much easier for the model to identify such proteins as the arguments of a *Positive Regulation* event. We encode the knowledge-enriched AMR graph using an edge-conditioned graph attention network (GAT) that is able to incorporate fine-grained edge features before conducting IE tasks. We evaluate our model on the existing benchmark GENIA-2011 dataset where our model greatly outperforms our baseline model by 4.8%. In addition to the existing GENIA-2011 benchmark, we also aim to evaluate the effectiveness of our framework on topic-specific literature. We develop a new ontology for entities and events with a large corpus from COVID-19 research papers, which is specifically annotated by medical professionals and can serve as a new benchmark for the biomedical IE community.

The major contributions of this paper are summarized as follows.

- We are the first to enrich the AMR graph with the external knowledge and use a graph neural network to incorporate the fine-grained edge features.
- We evaluate our model and create a new state-

of-the-art for biomedical event extraction on the GENIA-2011 corpus.

- We develop a new dataset from COVID-19 related research papers based on a new ontology that contains 25 fine-grained entity types and 14 event types.

2 Approach

2.1 Overview

As shown in Figure 2, our proposed biomedical information extraction framework mainly consists of four steps. First, we extract a global knowledge graph (KG) that contains all the entities from the corpus, and select out a sentence-level knowledge subgraph for the input sentence. Then, we perform AMR parsing and construct the sentence-level AMR graph, and use the sentence knowledge subgraph to enrich the AMR graph by adding additional nodes and edges. After that, given the contextualized word embeddings, we first identify entity and trigger spans, and then conduct message passing on the knowledge enriched AMR graph based on an edge-conditioned GAT. Finally, we use feed-forward neural networks based classifiers for trigger and argument labeling.

2.2 Knowledge Graph Construction

Global Knowledge Graph We use the Comparative Toxicogenomics Database (CTDB)[2] which contains fine-grained biomedical and chemical interactions between chemicals, genes, and diseases. We construct a global knowledge graph that involves all entities from the corpus with their pairwise chemical interactions. We extract these entity pairs with their biomedical interactions as triples, e.g., in Figure 1, (*Mono Mac 6, results, leukumia*) indicates that *Mono Mac 6* cell can *result in* the disease of *leukemia*. We merge all the extracted triples and form a global knowledge graph $G^g = (V^g, E^g)$. Our extracted global KG consists of 39,436 nodes and 590,235 edges.

Sentence-level Knowledge Graph Given an input sentence, we aim to generate a sentence-level KG by selecting out a subgraph from the gloabl KG, which contains the external knowledge between all entities within the sentence. Given an input sentence S, we use SciSpacy[3] to obtain all the related biomedical entities, including genes, chemicals, cells, and proteins. We then link each entity mention from the sentence to the nodes in global KG $G^g = (V^g, E^g)$. To select the sentence subgraph from the global KG, given the set of entity mentions $\mathcal{E} = \{\varepsilon_1, \cdots, \varepsilon_{|\mathcal{E}|}\}$ (where each ε_i is a word span), we select the *connected subgraph* that covers all entity mentions in $\mathcal{E}$ with the *minimal number of nodes* as the sentence KG. Note that such a sentence KG construction procedure can be accomplished in linear time complexity in terms of the number of nodes $|V^g|$. This can be done by first traversing all the nodes in the global KG using depth-first search and obtaining all connected subgraphs of G^g in linear time. After that, we select the set of subgraphs that can cover $\mathcal{E}$ and then choose the one $G^s = (V^s, E^s)$ with the minimal number of nodes as the sentence KG.

2.3 KG-enriched AMR parsing

AMR Parsing After obtaining the sentence KG, we fuse it with the AMR graph as an external knowledge enrichment procedure. Given an input sentence $S = \{w_1, w_2, \cdots, w_N\}$, we first perform AMR parsing and obtain a sentence-level AMR graph $G^A = (V^A, E^A)$ with an alignment between AMR nodes and the spans in the original sentence. We employ the transformer-based AMR parser[4] (Fernandez Astudillo et al., 2020) pretrained on the Biomedical AMR corpus[5] released from the AMR official website. Each node $v_i^A = (m_i^A, n_i^A) \in V^a$ represents an AMR concept or predicate, and we use (m_i^A, n_i^A) to denote the corresponding span for such an AMR node. For AMR edges, we use $e_{i,j}^A$ to denote the specific relation type between nodes v_i^A and v_j^A in AMR annotations (e.g., *ARG-x*, *:time*, *:location*, etc.). We randomly initialize the edge embeddings as a look-up embedding matrix E^{AMR}, which is optimized in end-to-end training.

Enrich AMR with sentence KG Given a pair of AMR graph G^A and sentence KG G^S, we fuse them into an enriched AMR graph $G = (V, E)$ as the external reference for the subsequent information extraction tasks. In general, there are three cases for fusing each sentence's KG nodes $v_i^s \in V^s$ into the AMR graph. *First*, if v_i^s represents an entity within the sentence, and there is also an AMR

[2] http://ctdbase.org/

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

[4] https://github.com/IBM/transition-amr-parser

[5] https://amr.isi.edu/download/2018-01-25/amr-release-bio-v3.0.txt

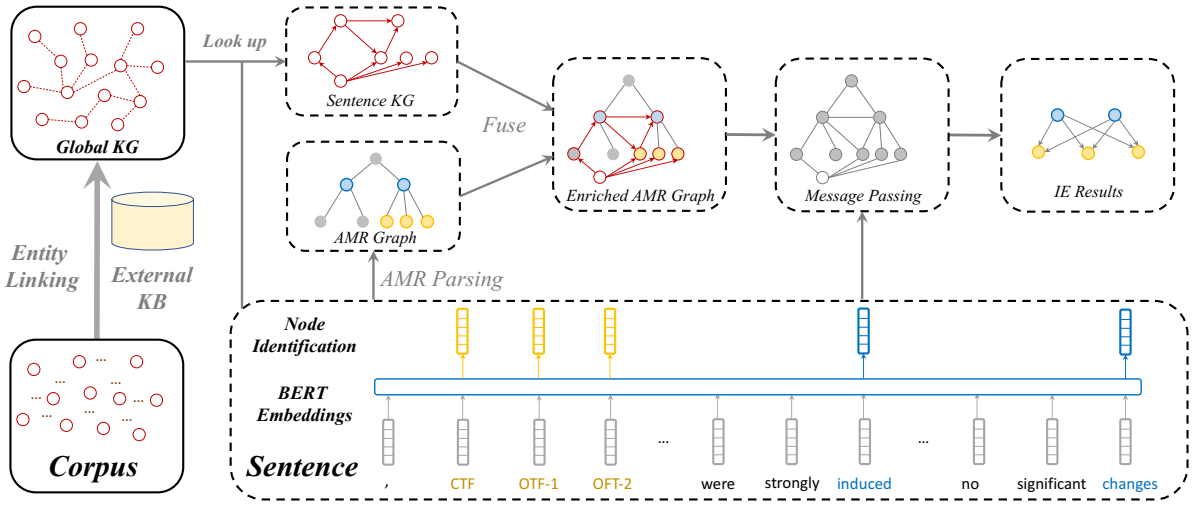

Figure 2: Overview of our proposed framework for biomedical information extraction.

node v_j^A with the same span, we then match v_i^s to v_j^A and add all KG edges linked to v_i^s into the AMR graph. *Second*, if v_i^s represents an entity within the sentence, but there is not any AMR node v_j^A with a matched span, we then add a new node (as well as all related edges) into the AMR graph. *Third*, if v_i^s is an additional KG node that does not represent any entity in the sentence, we directly add this node into the AMR graph with all related KG edges. After we match and link all the sentence KG nodes towards the AMR graph, we obtain the fused graph $G = (V, E)$. Note that such a graph fusion procedure could result in multiple edges between a pair of nodes. We keep all these edges with their embeddings for the subsequent message passing procedure. The illustration for the graph fusion procedure is shown in Figure 2.

2.4 Node Identification and Message Passing

Contextualized Encoder Given an input sentence S, we use the BERT model pretrained on biomedical scientific texts (Lee et al., 2020) to obtain the contextualized word representations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$. If one word is split into multiple pieces by the BERT tokenizer, we take the average of the representation vectors for all pieces as the final word representation.

Node Identification After encoding the input sentence using BERT, we first identify the entity and trigger spans as the candidate nodes. Similar to (Wadden et al., 2019), given the contextualized word representations, we first enumerate all possible spans up to a fixed length K, and calculate each span representation according to the concatenation of the left and right endpoints and a trainable feature vector characterizing the span length[6]. Specifically, given each span $s_i = [start(i), end(i)]$, the span representation vector is:

$$\boldsymbol{s}_i = \left[\boldsymbol{x}_{start(i)}, \boldsymbol{x}_{end(i)}, \boldsymbol{z}(s_i)\right], \quad (1)$$

where $\boldsymbol{z}(s_i)$ denotes a trainable feature vector that is only determined by the span length. We use separate binary classifiers for each specific entity and trigger type to handle the spans with multiple labels. Each binary classifier is a feed-forward neural network with ReLU activation in the hidden layer, which is trained with binary cross-entropy loss jointly with the whole model. Note that we do not use the specific entity and event types predicted by these classifiers as the final output, but only keep the identified word spans. In the diagnostic setting of using gold-standard entity mentions, we only employ span enumeration for event trigger identification, and use the gold-standard entity set for the following event extraction steps.

Edge-conditioned GAT To fully exploit the information of external knowledge and AMR semantic structure, similar to (Zhang and Ji, 2021), we use an L-layer graph attention network to let the model aggregate neighbor information from the fused graph $G = (V, E)$. We use $\boldsymbol{h}_i^l$ to denote the node feature for $v_i \in V$ in layer l, and $\boldsymbol{e}_{i,j}$ to represent the edge feature vector for $e_{i,j} \in E$. To update the node feature from l to $l+1$, we first calculate the attention score for each neighbor $j \in \mathcal{N}_i$ based on the concatenation of node features $\boldsymbol{h}_i^l$, $\boldsymbol{h}_j^l$ and

[6]We use different maximum span length K for entity and trigger spans.

edge features $\boldsymbol{e}_{i,j}$.

$$\alpha_{i,j}^{l} = \frac{\exp\left(\sigma\left(f^{l}[\mathbf{W}\boldsymbol{h}_{i}^{l} : \mathbf{W}_{e}\boldsymbol{e}_{i,j} : \mathbf{W}\boldsymbol{h}_{j}^{l}]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\sigma\left(f^{l}[\mathbf{W}\boldsymbol{h}_{i}^{l} : \mathbf{W}_{e}\boldsymbol{e}_{i,k} : \mathbf{W}\boldsymbol{h}_{k}^{l}]\right)\right)},$$

where $\mathbf{W}$, $\mathbf{W}_e$ are trainable parameters, and f^l and $\sigma(\cdot)$ are a single layer feed-forward neural network and *LeakyReLU* activation function respectively. Then we obtain the neighborhood information $\boldsymbol{h}_i^*$ by the weighted sum of all neighbor features:

$$\boldsymbol{h}_{i}^{*} = \sum_{k\in\mathcal{N}_i}\alpha_{i,j}^{l}\mathbf{W}^{*}h_{k}^{l},$$

where $\mathbf{W}^*$ is a trainable parameter. The updated node feature is calculated by a combination of the original node feature and its neighborhood information, where γ controls the level of message passing between neighbors.

$$\boldsymbol{h}_{i}^{l+1} = \boldsymbol{h}_{i}^{l} + \gamma\cdot\boldsymbol{h}_{i}^{*} \tag{2}$$

Note that our edge-conditioned GAT structure is similar to (Huang et al., 2020). The main difference is that (Huang et al., 2020) only uses edge features for calculating the attention score $\alpha_{i,j}^l$, while we use the concatenation of the feature vectors of each edge and its involved pair of nodes. Such a method can better characterize differing importance levels for neighbor nodes, and thus yield better model performance. We select the last layer $\boldsymbol{h}_i^L$ as the final representation for each entity or trigger.

Message Passing Given the knowledge enriched AMR graph $G = (V, E)$ and representation vectors of extracted trigger and entity spans, we initialize the feature vectors for nodes and edges as follows. For each KG node v_i^s which does not belong to any AMR node, we initialize its feature vectors $\boldsymbol{v}_i^s$ using KG embeddings pre-trained on the global KG using *TransE* (Bordes et al., 2013). For each original AMR node $v_i^A = (m_i^A, n_i^A)$, we first calculate its span representation $\boldsymbol{v}_i^A$ according to Eq. (1), and then use a linear transformation $\mathbf{W}^A\boldsymbol{v}_i^A + \mathbf{b}^A$ to initialize the node feature vector $\boldsymbol{h}_i^0$. For edge features, we use pre-trained TransE embeddings for KG edges, and use the trainable embedding matrix E^{AMR} for AMR relations. We use our proposed edge-conditioned GAT to conduct message passing and get the feature vectors from the final layer as the updated node representations. We obtain the final representation vectors for the trigger and entity nodes and denote them as $\{\boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{|\mathcal{T}|}\}$ and $\{\boldsymbol{\varepsilon}_1, \cdots, \boldsymbol{\varepsilon}_{|\mathcal{E}|}\}$ respectively.

2.5 Biomedical Event Extraction

Model Training Given the event trigger set $\mathcal{T}$ with the representation vectors $\boldsymbol{\tau}_i$ for each trigger, we use a feed-forward neural network (*FFN*) to classify each event trigger into pre-defined event type categories, where we have $\boldsymbol{y}_i^t = \mathit{FFN}_t(\boldsymbol{\tau}_i)$. For event argument role labeling, we concatenate candidate trigger-entity pairs or trigger-trigger pairs (for nested events) and feed them into two other separate FFNs for role type classification, where we have $\boldsymbol{y}_{i,j}^{tt} = \mathit{FFN}_{tt}([\boldsymbol{\tau}_i : \boldsymbol{\tau}_j])$ or $\boldsymbol{y}_{i,j}^{te} = \mathit{FFN}_{te}([\boldsymbol{\tau}_i : \boldsymbol{\varepsilon}_j])$. The overall training objective is defined in a multi-task setting, which includes the cross-entropy loss for trigger and argument classification, as well as the binary classification loss $\mathcal{L}_I$ for the binary classifiers in the node identification step.

$$\begin{aligned}\mathcal{L} = \mathcal{L}_I &- \sum_{i=1}^{|\mathcal{T}|}\boldsymbol{y}_i^t\log\hat{\boldsymbol{y}}_i^t \\ &- \sum_{i,j}\boldsymbol{y}_{i,j}^{tt}\log\hat{\boldsymbol{y}}_{i,j}^{tt} - \sum_{i,j}\boldsymbol{y}_{i,j}^{te}\log\hat{\boldsymbol{y}}_{i,j}^{te}.\end{aligned} \tag{3}$$

3 Experiments

3.1 Experimental Setup

Data Similarly to the recent work (Li et al., 2019; Huang et al., 2020; Ramponi et al., 2020), we also conduct experiments on the BioNLP GENIA 2011 (Kim et al., 2011) dataset consisting of both abstracts and main body texts from biomedical scientific papers. Similarly to previous work (Li et al., 2019; Huang et al., 2020; Ramponi et al., 2020), we only focus on extracting the core events, which involves *Protein* entities, 9 fine-grained event types, and 2 event argument types. We do not incorporate event ontology or training data from the newer versions of the BioNLP GENIA shared tasks (e.g., GENIA 2013) to ensure fair comparisons with previous models. The statistics of this dataset are shown in Table 2. The original GENIA dataset

Data Split	Train Set	Dev Set	Test Set
# Documents	908	259	231
# Sentences	8,620	2,846	3,348
# Proteins	11,625	4,690	5,301
# Events	10,310	3,250	4,487

Table 2: GENIA 2011 Dataset Statistics.

is annotated in paragraphs. Following (Li et al., 2019), we focus on sentence-level event extraction and only keep events and argument roles within each sentence (around 94% of the events).

3.2 Baselines and Ablation Variants

We consider the most recent models on biomedical event extraction: KB-Tree-LSTM (Li et al., 2019), GEANet (Huang et al., 2020), BEESL (Ramponi et al., 2020), and DeepEventMine (Trieu et al., 2020) for comparison in our experiments, and we report the precision, recall, and F1 score from the GENIA 2011 online test set evaluation service[7]. In addition to the previous models, we also conduct ablation studies to evaluate the contributions of different parts in our model. We adopt the model variants *BERT-Flat* and *BERT-AMR*, where *BERT-Flat* only uses the BERT representations without any help from AMR and KG, and *BERT-AMR* denotes the model with an edge-conditioned GAT to encode the AMR graph without incorporating external knowledge. More details about implementation and hyperparameter choices are described in Appendix A.1.

3.3 Overall Performance

We report the performance of our model and compare it with the most recent biomedical IE models KB-Tree-LSTM (Li et al., 2019), GEANet (Huang et al., 2020), BEESL (Ramponi et al., 2020), and DeepEventMine (Trieu et al., 2020) in Table 3. In general, our KG enriched AMR model can achieve slightly higher performance compared with the state-of-the-art model DeepEventMine. Besides, our model greatly outperforms all other previous models for biomedical event extraction. To further measure the impact of each individual part in our model, we also introduce two model variants for the ablation study. We can see that compared with simply finetuning a flat BERT model, the AMR parsing contributes a 1.84% absolute gain on F1-Score, while the incorporation of external knowledge graph contributes 2.95%. We also report the overall development set F1 scores with and without using gold-standard entities, and compare the performance with BEESL in Table 4. We can discover that our model performs significantly better than the BEESL model in both settings, which proves that our model can better handle practical scenarios without gold-standard entities.

[7] http://bionlp-st.dbcls.jp/GE/2011/eval-test/

Model	Prec	Rec	F1
String Matching	43.92	21.82	29.16
Tree-LSTM (Li et al., 2019)	67.01	52.14	58.65
GEANet (Huang et al., 2020)	64.61	56.11	60.06
BEESL (Ramponi et al., 2020)	69.72	53.00	60.22
DeepEM (Trieu et al., 2020)	71.71	56.20	**63.02**
BERT-Flat	64.68	52.98	58.25
BERT-AMR	68.39	53.58	60.09
BERT-AMR-KG (Ours)	72.74	55.62	**63.04**

Table 3: Overall test F-score (%) of biomedical extraction on GENIA 2011 dataset.

Model	F1-Score
BEESL	65.04
Ours	**67.23**
BEESL (w/o gold-standard entities)	59.51
Ours (w/o gold-standard entities)	60.16

Table 4: Overall dev F-score (%) of biomedical extraction on GENIA 2011 dataset, and we also report the scores without using gold-standard entities.

3.4 Case Study on COVID-19 Dataset

COVID-19 Dataset In order to evaluate the impact of our approach on real-world problems, besides the GENIA dataset, we also develop a new dataset specifically labeled by medical professionals from research papers related to COVID-19. We select out 186 full-text articles with 12,916 sentences from PubMed and PMC. Three experienced annotators who are biomedical domain experts have participated in the annotation, and the Cohen's Kappa scores for pairwise agreement between the annotators are 0.79, 0.84, and 0.74 respectively. The pre-defined entity and event type distributions in this dataset are shown in Table 6.

Results We evaluate our proposed model by removing the event argument labeling procedure to accommodate a scenario limited to entity and event trigger labeling, that is, we remove the argument role classifiers FFN_{tt} and FFN_{te} while the overall training loss in Eq. (3) only contains the first two terms for span identification and event trigger classification. As shown in Table 5, our model achieves 78.05% overall F1 score with 83.60% F1 on entity extraction task and 72.37% F1 on event extraction. The entity extraction performance on the COVID dataset is lower than typical coarse-grained entity extraction model performance for BERT-like models on other datasets (e.g., our model can get around 90% F1 score for entity extraction on GENIA-2011 development set). This is probably because our proposed COVID-19 dataset is challenging with more

find-grained biomedical entity and event types.

Model	Prec	Rec	F1
BERT-AMR-KG (entities)	83.89	83.32	83.60
BERT-AMR-KG (events)	72.47	72.27	72.37
BERT-AMR-KG (overall)	78.11	78.00	78.05

Table 5: F-scores (%) on COVID-19 test dataset for entity and event extraction.

Entities	*# Labels*	**Events**	*# Labels*
Disease	6,231	BiologicalProcess	6,737
MedicalDrug	3,901	ResearchActivity	5,177
Patients	3,430	LabOrTestResult	4,637
Chemical	2,719	TherapeuticProcedure	3,819
Human	2,146	SymptomOrSign	2,585
Country	1,610	InfectionControl	1,937
Gene/Protein	1,501	PharmacologicalAction	1,567
SARS-CoV-2	1,452	Epidemic	1,180
Organization	1,182	DiagnosticProcedure	1,014
AnatomicalStructure	1,130	DiseaseTransmission	807
MedicalExpert	1,105	LaboratoryTechniques	527
UrbanArea	794	BiologicalProcess	262
Organism	666	EnvironmentalExposure	249
Coronavirus	567	GeneticEvolution	83
Animal	460		
FluidsAndSecretions	406		
SARS-CoV	341		
CellularComponent	289		
Gene	272		
MERS-CoV	185		
ProtectiveEquip	184		
ViralParticle	146		
CellLine	66		
Antigen	32		
Species	28		

Table 6: Our new COVID-19 ontology with 24 fine-grained entity types and 15 biomedical event types.

3.5 Qualitative Analysis

We select two typical examples in Table 7 to show how KG enriched AMR parsing helps to improve the performance of biomedical IE.

In the first example, we can see that the flat model fails to identify *CAII* as an entity of the *bind* event, which is probably due to the long distance between the trigger *bind* and the argument *CAII* (the model successfully detects the other two arguments *V-erbA* and *C-erbA* because they are much nearer). With the help of AMR parsing, the model successfully links *CAII* to the *bind* event since in the AMR graph, the three entities *C-erbA*, *V-erbA*, and *CAII* are located within the same number of hops from the *bind* trigger. But the model still cannot recognize *CAII* as the theme of *transcription*. This is probably because the model is not clear what *whose* refers to in the sentence. However, with the help of external knowledge, the model knows in advance that *V-erbA* could inhibit the trascription of *CAII*, thus it is able to identify *CAII* as the theme of the transcription event.

In the second example, the flat model is confused about which entity belongs to which event between two binding events in the same sentence. Here, the AMR parsing provides a clear tree structure and guides the model to correctly link the event-entity pairs (i.e., *heterodimers* with *RAR beta*, *binding* with *VDR*). However, the BERT-AMR model still fails to identify *heterodimers* as the theme of *stimulated*. With the further help of the external KG, the model knows in advance that *RA* can stimulate the generation of *RAR beta* heterodimers, and thus it is able to correctly identify a positive regulation between these two triggers.

3.6 Remaining Challenges

We compare the predictions from our model with the gold-standard annotations on the development set and discover the following typical remaining error cases.

Non-verb Event Triggers Most of the biomedical events are triggered by verbs (*bind*, *express*, etc.) or their noun forms (*binding*, *expression*, etc.). However, there are also events triggered by adjectives (e.g., *subsequent*), proper nouns (e.g., *mRNA, SiRNA*), and even prepositions (e.g., *from*) and conjunctions (e.g., *rather than*). Our model misses a lot of these non-verb event triggers due to the insufficient training examples.

Misleading Verb Prefix We also find that the prefix of a verb can sometimes be misleading for event trigger classification, especially for *Negative Regulation* events. Many *Negative Regulation* events are triggered by words with certain styles of prefix (*in-* or *de-*), e.g., *inactivation, inactivated, decrease, degradation*, etc., representing some negative interactions. As a result, the model mistakenly labels many other words with the same prefixes as *Negative Regulation* event triggers. For example, in the sentence: *Dephosphorylation of 4E-BP1 was also observed ...*, the word *dephosphorylation* should not be classified as a *Negative Regulation* event although it has a *de-* prefix. Because dephosphorylation denotes an inverse chemical process of phosphorylation rather than negative regulation between different events or proteins. This is probably because the BERT tokenizer breaks these words into pieces *de*, *phosphorylation*, encouraging BERT models to learn misleading patterns.

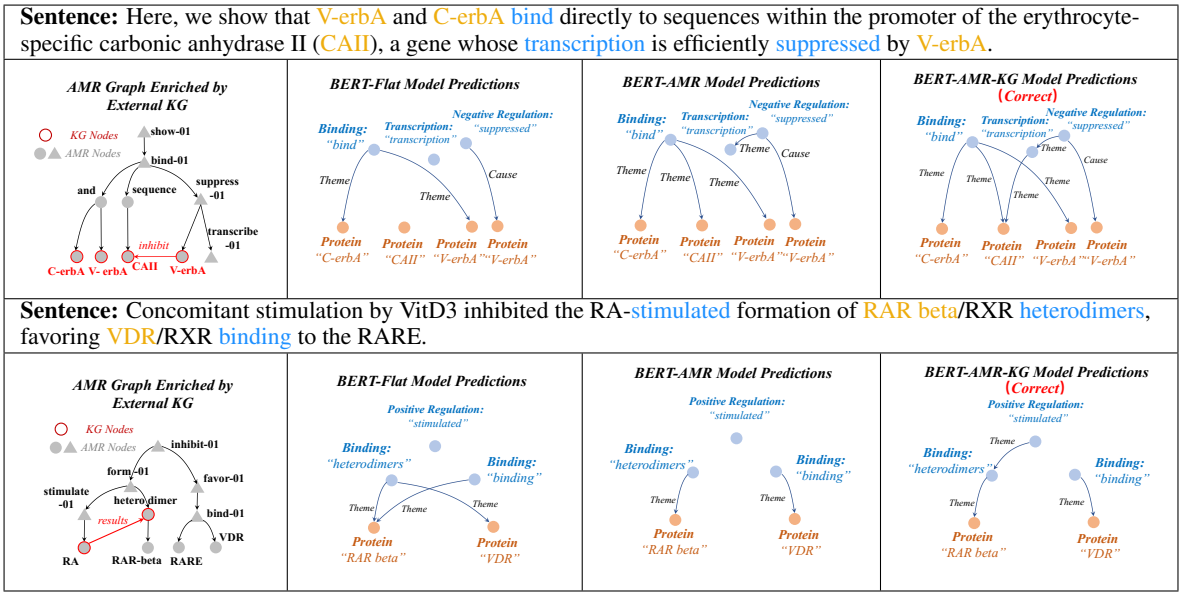

Table 7: Examples from development set showing how KG enriched AMR graph improves the model performance.

4 Related Work

Biomedical Information Extraction A number of previous studies contribute to biomedical event extraction with various techniques, such as dependency parsing (McClosky et al., 2011; Li et al., 2019), external knowledge base (Li et al., 2019; Huang et al., 2020), joint inference of triggers and arguments (Poon and Vanderwende, 2010; Ramponi et al., 2020), Abstract Meaning Representation (Rao et al., 2017), search based neural models (Espinosa et al., 2019), and multi-turn question answering (Wang et al., 2020b). Recently, to handle the nested biomedical events, BEESL (Ramponi et al., 2020) models biomedical event extraction as a unified sequence labeling problem for end-to-end training. DeepEventMine (Trieu et al., 2020) proposes to use a neural network based classifier to decide the structure of complex nested events. Our model is also in an end-to-end training pipeline, but additionally utilizes fine-grained AMR semantic parsing and external knowledge to improve the performance.

Utilization of External Knowledge In terms of utilization of external knowledge, (Li et al., 2019) proposes a knowledge-driven Tree-LSTM framework to capture dependency structures and entity properties from an external knowledge base. More recently, GEANet (Huang et al., 2020) introduces a Graph Edge conditioned Attention Network (GEANet) that incorporates domain knowledge from the Unified Medical Language System (UMLS) into the IE framework. The main difference of our model is that we use fine-grained AMR parsing to compress the wide context, and manage to use an external KG to enrich AMRs to better incorporate domain knowledge. Incorporating external knowledge is also widely used in other tasks such as relation extraction (Chan and Roth, 2010; Cheng and Roth, 2013), and QA for domain-specific (science) questions (Pan et al., 2019).

Biomedical Benchmarks for COVID-19 (Lo et al., 2020) releases a dataset containing open-access biomedical papers related to COVID-19. A lot of research has been done based on this dataset, including Information Retrieval (Wise et al., 2020), Entity Recognition (Wang et al., 2020b), distant supervision on fine-grained biomedical name entity recognition to support automatic information retrieval indexing or evidence mining (Wang et al., 2020c), and end-to-end Question Answering (QA) system for COVID-19 with domain adaptive synthetic QA training (Reddy et al., 2020). Our COVID-19 dataset will further advance the field in developing effective IE techniques specifically for the COVID-19 domain.

5 Conclusions and Future Work

In this paper, we propose a novel biomedical Information Extraction framework to effectively tackle two unique challenges for scientific domain IE: complex sentence structure and unexplained concepts. We utilize AMR parsing to compress wide contexts, and incorporate external knowledge into the AMR. Our proposed model produces signifi-

cant performance gains compared with most state-of-the-art methods. In the future, we intend to exploit tables and figures in the scientific literature for multimedia representation. We also plan to further incorporate coreference graphs among sentences to further enrich contexts. We will also continue exploring the use of richer information from an external knowledge base to further improve the model's performance.

Acknowledgement

This research is based upon work supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, NSF No. 2034562, U.S. DARPA KAIROS Program No. FA8750-19-2-1004, the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract No. FA8650-17-C-9116, and Air Force No. FA8650-17-C-7715. Any opinions, findings and conclusions or recommendations expressed in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, LAW-ID@ACL 2013, August 8-9, 2013, Sofia, Bulgaria*, pages 178–186.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

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

Yee Seng Chan and Dan Roth. 2010. Exploiting background knowledge for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 152–160.

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Seattle, Washington, USA. Association for Computational Linguistics.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3677–3684.

Ramón Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodgett, and Radu Florian. 2020. Transition-based parsing with stack-transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1001–1007, 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.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of Genia event task in BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

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 Lan-*

guage Technologies, Volume 1 (Long and Short Papers), pages 1421–1430, Minneapolis, Minnesota. Association for Computational Linguistics.

K Lo, Y Chandrasekhar, R Reas, J Yang, D Eide, K Funk, R Kinney, Z Liu, W Merrill, P Mooney, et al. 2020. Cord-19: The covid-19 open research dataset. *Arxiv*.

Kevin Lybarger, Mari Ostendorf, Matthew Thompson, and Meliha Yetisgen. 2020. Extracting covid-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework.

David McClosky, Mihai Surdeanu, and Christopher D. Manning. 2011. Event extraction as dependency parsing. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 1626–1635.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31(1).

Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie, and Dong Yu. 2019. Improving question answering with external knowledge. *arXiv preprint arXiv:1902.00993*.

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.

Revanth Gangi Reddy, Bhavani Iyer, Md Arafat Sultan, Rong Zhang, Avi Sil, Vittorio Castelli, Radu Florian, and Salim Roukos. 2020. End-to-end qa on covid-19: Domain adaptation with synthetic training. *arXiv preprint arXiv:2012.01414*.

Hai-Long Trieu, Thy Thy Tran, Anh-Khoa Duong Nguyen, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. Deepeventmine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917.

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.

Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, David Liem, Ahmed Elsayed, Martha Palmer, Jasmine Rah, Cynthia Schneider, and Boyan Onyshkevych. 2020a. Covid-19 literature knowledge graph construction and drug repurposing report generation. In *arXiv:2007.00576*.

Xing David Wang, Leon Weber, and Ulf Leser. 2020b. 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.

Xuan Wang, Weili Liu, Aabhas Chauhan, Yingjun Guan, and Jiawei Han. 2020c. Automatic textual evidence mining in covid-19 literature. *arXiv preprint arXiv:2004.12563*.

Colby Wise, Vassilis N Ioannidis, Miguel Romero Calvo, Xiang Song, George Price, Ninad Kulkarni, Ryan Brand, Parminder Bhatia, and George Karypis. 2020. Covid-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. *arXiv preprint arXiv:2007.12731*.

Zixuan Zhang and Heng Ji. 2021. Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49, Online. Association for Computational Linguistics.

A Appendices

A.1 Implementation Details

For pretrained KG embeddings, we use 600-dim embedding vectors pre-trained on the global knowledge graph using TransE. We use a two-layer edge-conditioned GAT and the feature dimensions are 2048 for nodes and 256 for edges. Specifically, the FFNs consist of two layers with a dropout rate of 0.4, where the numbers of hidden units are 150 for entity extraction and 600 for event extraction. We train our model with Adam (Kingma and Ba, 2015) on NVIDIA Tesla V100 GPUs for 80 epochs (approximately takes 4 minutes for 1 training epoch) with learning rate 1e-5 for BERT parameters and 5e-3 for other parameters. We select the model checkpoint with optimal F1-Score on the development set to evaluation on the test set from the official website. The detailed hyper-parameter settings are shown in Table 8.

Hyper-parameters	Values
Number of model parameters (except BERT)	3.25M
KG embedding dimensions	600
Num of features for each node	2,048
Num of features for AMR relation	256
Num of GAT layers	2
Message passsing level γ	0.003
Num of layers for FFNNs	2
FFNN hidden dimensions for entity extraction	150
FFNN hidden dimensions for event extraction	600
Dropout rate	0.4
Activation function	*ReLU*
Learning rate for BERT params	1e-5
Learning rate for other params	1e-3
Batch size	16

Table 8: Detailed settings for model hyper-parameters.