

Biomedical Event Extraction with Hierarchical Knowledge Graphs

Kung-Hsiang Huang¹ Mu Yang¹ Nanyun Peng^{1,2}

¹ Information Sciences Institute, University of Southern California

² Computer Science Department, University of California, Los Angeles

{kunghsia, yangmu}@usc.edu

violetpeng@cs.ucla.edu

Abstract

Biomedical event extraction is critical in understanding biomolecular interactions described in scientific corpus. One of the main challenges is to identify nested structured events that are associated with non-indicative trigger words. We propose to incorporate domain knowledge from Unified Medical Language System (UMLS) to a pre-trained language model via a hierarchical graph representation encoded by a proposed Graph Edge-conditioned Attention Networks (GEANet). To better recognize the trigger words, each sentence is first grounded to a sentence graph based on a jointly modeled hierarchical knowledge graph from UMLS. The grounded graphs are then propagated by GEANet, a novel graph neural networks for enhanced capabilities in inferring complex events. On BioNLP 2011 GENIA Event Extraction task, our approach achieved 1.41% F_1 and 3.19% F_1 improvements on all events and complex events, respectively. Ablation studies confirm the importance of GEANet and hierarchical KG.

1 Introduction

Biomedical event extraction is a task that identifies a set of actions among proteins or genes that are associated with biological processes from natural language texts (Kim et al., 2009, 2011). Development of biomedical event extraction tools enables many downstream applications, such as domain-specific text mining (Ananiadou et al., 2015; Spangher et al., 2020), semantic search engines (Miyao et al., 2006) and automatic population and enrichment of database (Hirschman et al., 2012).

A typical event extraction system 1) finds triggers that most clearly demonstrate the presence of events, 2) recognizes the protein participants (arguments), and 3) associates the arguments with the corresponding event triggers. For instance, the

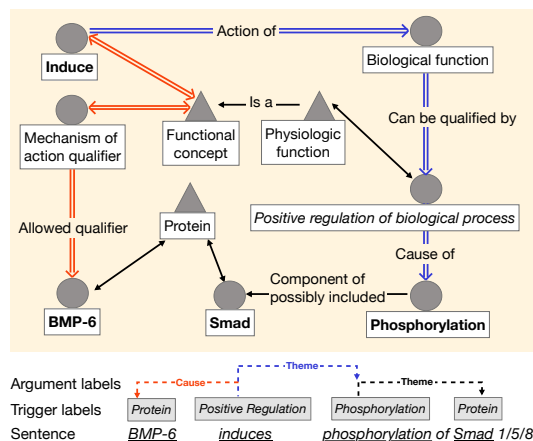


Figure 1: An example of a UMLS-based hierarchical KG assisting event extraction. Circles represent concept nodes and triangles represent semantic nodes. Nodes associated with the tokens in the example sentence are boldfaced. Bidirectional edges imply hierarchical relation between concept and semantic nodes. The word “induces” is a trigger of a *Positive regulation* event, whose trigger role and corresponding argument role cannot be easily determined with only textual input. The KG provides clues for identifying this trigger and its corresponding arguments given the red and blue double line reasoning paths connecting nodes **BMP-6**, **Induce**, **Phosphorylation**, and *Positive regulation of biological process*. We can infer that: 1) “induces” is an action of biological function, 2) a biological function can be quantified by positive regulation, and 3) positive regulation can result in phosphorylation.

sentence “Protein A inhibits the expression of Protein B” will be annotated with two nested events: *Gene expression*(Trigger: expression, Arg-Theme: Protein B) and *Negative Regulation*(Trigger: inhibits, Arg-Theme: *Gene expression*(Protein B), Arg-Cause: Protein A).

Early attempts on biomedical event extraction adopted hand-crafted features (Björne et al., 2009; Björne and Salakoski, 2011; Riedel and McCallum, 2011; Venugopal et al., 2014a). Recent advances have shown improvements using deep neural networks via distributional word representations in the

biomedical domain (Moen and Ananiadou, 2013; Rao et al., 2017a; Björne and Salakoski, 2018; ShafieiBavani et al., 2019). Li et al. (2019) further extends the word representations with embeddings of descriptive annotations from a knowledge base and demonstrates the importance of domain knowledge in biomedical event extraction.

However, encoding knowledge with distributional embeddings does not provide adequate clues for identifying challenging events with non-indicative trigger words and nested structures. These embeddings do not contain *structural* or *relational* information about the biomedical entities. To overcome this challenge, we present a framework that incorporates knowledge from hierarchical knowledge graphs with graph neural networks (GNN) on top of a pre-trained language model.

Our first contribution is a novel representation of knowledge as hierarchical knowledge graphs containing both *conceptual* and *semantic* reasoning paths that enable better trigger and word identification based on Unified Medical Language System (UMLS), a biomedical knowledge base. Fig. 1 shows an example where the *Positive Regulation* event can be better identified with knowledge graphs and factual relational reasoning. **Our second contribution** is a new GNN, Graph Edge-conditioned Attention Networks (GEANet), that encodes complex domain knowledge. By integrating edge information into the attention mechanism, GEANet has greater capabilities in reasoning the plausibility of different event structure through factual relational paths in knowledge graphs (KGs).

Experiments show that our proposed method achieved state-of-the-art results on the BioNLP 2011 event extraction task (Kim et al., 2011).¹

2 Background

UMLS Knowledge Base. Unified Medical Language System (UMLS) is a knowledge base for biomedical terminology and standards, which includes three knowledge sources: *Metathesaurus*, *Semantic Network*, and *Specialist Lexicon and Lexical Tools* (Bodenreider, 2004). We use the former two sources to build hierarchical KGs. The *concept network* from Metathesaurus contains the relationship between each biomedical concept pairs, while each concept contains one or more semantic types

¹Our code for pre-processing, modeling, and evaluation is available at <https://github.com/PlusLabNLP/GEANet-BioMed-Event-Extraction>.

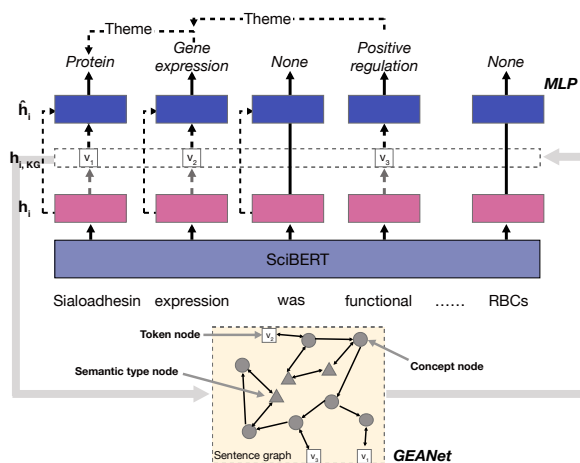


Figure 2: Overview of knowledge incorporation. Contextualized embeddings for each token are generated by SciBERT. GEANet updates node embeddings for v_1 , v_2 , and v_3 via corresponding sentence graph.

that can be found in the *semantic network*. The *concept network* provides direct definition lookup of the recognized biomedical terms, while the *semantic network* supports with additional knowledge in the semantic aspect. Example tuples can be found in Figure 1.² There are 3.35M concepts, 10 concept relations, 182 semantic types, and 49 semantic relations in total.

3 Proposed Approach

Our event extraction framework builds upon the pre-trained language model, SciBERT (Beltagy et al., 2019), and supplement it with a novel graph neural network model, GEANet, that encodes domain knowledge from hierarchical KGs. We will first illustrate each component and discuss how training and inference are done.

3.1 Hierarchical Knowledge Graph Modeling

The two knowledge sources discussed in Section 2 are jointly modeled as a *hierarchical graph* for each sentence, which we refer to as a *sentence graph*. Each sentence graph construction consists of three steps: concept mapping, concept network construction, and semantic type augmentation.

The first step is to map each sentence in the corpus to *UMLS biomedical concepts* with MetaMap, an entity mapping tool for UMLS concepts (Aronson, 2001). There are 7903 concepts (entities) being mapped from the corpus, denoted as K . The next step is *concept network construction*, where a minimum spanning tree (MST) that connects

²There are several bi-directional relations between some concepts. We only show one direction for simplicity.

mapped concepts in the previous step is identified, forming concept reasoning paths. This step is NP-complete.³ We adopt a 2-approximate solution that constructs a global MST for the corpora GE' 1.1 by running breadth-first search, assuming all edges are of unit distance. To prune out less relevant nodes and to improve computation efficiency, concept nodes that are not in K with less than T neighbors in K are removed.⁴ The spanning tree for each sentence is then obtained by depth-first search on the global MST. Each matched token in the corpus is also included as a token node in the sentence graph, connecting with corresponding concept node. Finally, the semantic types for each concept node are modeled as nodes that are linked with associated concept nodes in the sentence graph. Two semantic type nodes will also be linked if they have known relationships in the *semantic network*.

3.2 GEANet

The majority of existing graph neural networks (GNN) consider only hidden states of nodes and adjacency matrix without modeling edge information. To properly model the hierarchy of the graph, it is essential for the message passing function of a GNN to consider edge features. We propose Graph Edge Conditioned Attention Networks (GEANet) to integrate edge features into the attention mechanism for message propagation. The node embeddings update of GEANet at the l -th layer can be expressed as follows:

$$\mathbf{x}_i^{(l)} = \text{MLP}_\theta \mathbf{x}_i^{(l-1)} + \sum_{j \in \mathcal{N}(i)} a_{i,j} \cdot \mathbf{x}_j^{(l-1)} \quad (1)$$

$$a_{i,j} = \frac{\exp(\text{MLP}_\psi(\mathbf{e}_{i,j}))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{MLP}_\psi(\mathbf{e}_{i,k}))} \quad (2)$$

where $\mathbf{x}_i^{(l)}$ denotes the node embeddings at layer l , $\mathbf{e}_{i,j}$ denotes the embedding for edge (i, j) , and MLP_ψ and MLP_θ are two multi-layer perceptrons.

GEANet is inspired by Edge Conditioned Convolution (ECC), where convolution operation depends on edge type (Simonovsky and Komodakis, 2017),

$$\mathbf{x}_i^{(l)} = \text{MLP}_\theta \mathbf{x}_i^{(l-1)} + \sum_{j \in \mathcal{N}(i)} \mathbf{x}_j^{(l-1)} \cdot \text{MLP}_\psi(\mathbf{e}_{i,j}) \quad (3)$$

Compared to ECC, GEANet is able to determine the relative importance of neighboring nodes with attention mechanism.

³Finding a MST on a subset of nodes (K) is known as a Steiner tree problem.

⁴ T is empirically set to be 35.

Knowledge Incorporation. We build GEANet on top of SciBERT (Peters et al., 2019) to incorporate domain knowledge into rich contextualized representations. Specifically, we take the contextual embeddings $\{\mathbf{h}_1, \dots, \mathbf{h}_n\}$ produced by SciBERT as inputs and produces knowledge-aware embeddings $\{\hat{\mathbf{h}}_1, \dots, \hat{\mathbf{h}}_n\}$ as outputs. To initialize the embeddings for a sentence graph, for a mapped token, we project its SciBERT contextual embedding to initialize its corresponding node embedding $\mathbf{h}_{i,\text{KG}} = \mathbf{h}_i \mathbf{W}_{\text{KG}} + \mathbf{b}_{\text{KG}}$. Other nodes and edges are initialized by pretrained KG embeddings (details in Section 4.1). To accommodate multiple relations between two entities in UMLS, edge embeddings $\mathbf{e}_{i,j}$ are initialized by summing the embeddings of each relation between the nodes i and j . Then we apply layers of GEANet to encode the graph $\mathbf{h}_{i,\text{KG}}^l = \text{GEANet}(\mathbf{h}_{i,\text{KG}})$. The knowledge-aware representation is obtained by aggregating SciBERT representations and KG representations, $\hat{\mathbf{h}}_i = \mathbf{h}_{i,\text{KG}}^l \mathbf{W}_{\text{LM}} + \mathbf{b}_{\text{LM}} + \mathbf{h}_i$.⁵ The process is illustrated in Figure 2 GEANet layer.

3.3 Event Extraction

The entire framework is trained with a multitask learning pipeline consisting of trigger classification and argument classification, following (Han et al., 2019a,b). Trigger classification predicts the trigger type for each token. The predicted score of each token is computed as $\hat{\mathbf{y}}_i^{\text{tri}} = \text{MLP}^{\text{tri}}(\hat{\mathbf{h}}_i)$. In the argument classification stage, each possible pair of gold trigger and gold entity is gathered and labeled with corresponding argument role.⁶ The argument scores between the i -th token and j -th token are computed as $\hat{\mathbf{y}}_{i,j}^{\text{arg}} = \text{MLP}^{\text{arg}}(\hat{\mathbf{h}}_i; \hat{\mathbf{h}}_j)$, where $(;)$ denotes concatenation. Cross Entropy loss $\mathcal{L}^t = -\frac{1}{N^t} \sum_{i=1}^{N^t} \mathbf{y}_i^t \cdot \log \hat{\mathbf{y}}_i^t$, is used for both tasks, where t denotes task, N^t denotes the number of training instances of task t , \mathbf{y}_i^t denotes the ground truth label, and $\hat{\mathbf{y}}_i^t$ denotes the predicted label. The multitask learning minimizes the sum of the two losses $\mathcal{L} = \mathcal{L}^{\text{tri}} + \mathcal{L}^{\text{arg}}$ in the training stage. During inference, unmerging is conducted to combine identified triggers and arguments for multiple arguments events (Björne and Salakoski, 2011). We adopted similar unmerging heuristics. For *Regulation* events, we use the same heuristics as Björne et al. (2009). For *Binding* events, we subsume all *Theme* arguments associated with a trigger

⁵ $\hat{\mathbf{h}}_i = \mathbf{h}_i$ for each token i without mapped concept.

⁶During inference, predicted triggers are used instead.

	Model	Recall	Prec.	F1
Prior	TEES	49.56	57.65	53.30
	Stacked Gen.	48.96	66.46	56.38
	TEES CNN	49.94	69.45	58.10
	KB-driven T-LSTM	52.14	67.01	58.65
Ours	SciBERT-FT	53.89	63.97	58.50
	GEANet-SciBERT	56.11	64.61	60.06

Table 1: Model comparison on GE' 11 test set.

Model	Recall	Prec.	F1
KB-driven T-LSTM	41.73	55.73	47.72
SciBERT-FT	45.39	54.48	49.52
GEANet- SciBERT	47.23	55.21	50.91

Table 2: Performance comparison on the *Regulation* events of the test set (including *Regulation*, *Positive Regulation*, and *Negative Regulation* sub-events).

into one event such that every trigger corresponds to only one single *Binding* event.

4 Experiments

4.1 Experimental Setup

Our models are evaluated on BioNLP 11 GENIA event extraction task (GE' 11). All models were trained on the training set, validated on the dev set, and tested on the test set. A separate evaluation on *Regulation* events is conducted to validate the effectiveness of our framework on nested events with non-indicative trigger word. Reported results are obtained from the official evaluator under approximate span and recursive criteria.

In the preprocessing step, the GE' 11 corpora were parsed with TEES preprocessing pipeline (Björne and Salakoski, 2018). Tokenization is done by the SciBERT tokenizer. Biomedical concepts in each sentence are then recognized with MetaMap and aligned with their corresponding tokens. The best performing model was found by grid search conducted on the dev set. The edge and node representation in KGs were initialized with 300 dimensional pre-trained embeddings using TransE (Wang et al., 2014). The entire framework is optimized with BERTAdam optimizer for a maximum of 100 epochs with batch size of 4. Training is stopped if the dev set F_1 does not improve for 5 consecutive epochs (more details see Appendix).

4.2 Results and Analysis

Comparison with existing methods We compare our method with the following prior works: **TEES** and **Stacked Gen.** use SVM-based models with token and sentence-level features (Björne and Salakoski, 2011; Majumder et al., 2016);

Model	Dev F1	Test F1
GEANet-SciBERT	60.38	60.06
- GEANet	59.33	58.50
- STY nodes	60.12	59.34
GEANet→ ECC	58.50	58.27
GEANet→ GAT	59.55	59.87

Table 3: Ablation study over different components.

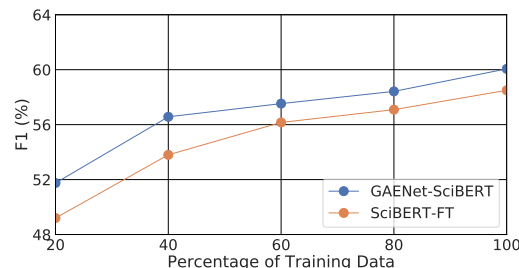


Figure 3: Performance comparison on the test set w.r.t. different amount of training data.

TEES CNN leverages Convolutional Neural Networks and dependency parsing graph (Björne and Salakoski, 2018); **KB-driven T-LSTM** adopts an external knowledge base with type and sentence embeddings, into a Tree-LSTM model (Li et al., 2019). **SciBERT-FT** is a fine-tuned SciBERT without external resources, the knowledge-agnostic counterpart of **GEANet-SciBERT**. According to Table 1, **SciBERT-FT** achieves similar performance to **KB-driven T-LSTM**, implying that SciBERT may have stored domain knowledge implicitly during pre-training. Similar hypothesis has also been studied in commonsense reasoning (Wang et al., 2019). **GEANet-SciBERT** achieves an absolute improvement of 1.41% in F_1 on the test data compared to the previous state-of-the-art method. In terms of *Regulation* events, Table 2 shows that **GEANet-SciBERT** outperforms the previous system and fine-tuned SciBERT by 3.19% and 1.39% in F_1 .

Ablation study To better understand the importance of different model components, ablation study is conducted and summarized in Table 3. GEANet achieves the highest F_1 when compared to two other GNN variants, ECC and GAT (Veličković et al., 2018), demonstrating its stronger knowledge incorporation capacity. Hierarchical knowledge graph representation is also shown to be critical. Removing semantic type (STY) nodes from hierarchical KGs leads to performance drop.

Impact of amount of training data Model performance on different amount of randomly sampled training data is shown in Fig. 3. **GEANet-**

SciBERT shows consistent improvement over fine-tuned SciBERT across different fractions. The performance gain is slightly larger with less training data. This illustrates the robustness of GEANet in integrating domain knowledge and its particular advantage under low-resource setting.

Error Analysis By comparing the predictions from **GEANet-SciBERT** and gold events in the dev set, two major failed cases are identified:

- **Adjective Trigger:** Most events are associated with a verb or noun trigger. Adjective triggers are scarce in the training set ($\sim 7\%$), which poses a challenge to identify this type of trigger. Although knowledge-aware methods should be able to resolve these errors theoretically, these adjective triggers often cannot be linked with UMLS concepts. Without proper grounding, it is hard for our model to recognize these triggers.
- **Misleading Trigger:** Triggers providing “clues” about incorrect events can be misleading. For instance,

*Furthermore, **expression** of an activated PKD1 mutant enhances HPK1-mediated NFkappaB activation.*

Our model predicts **expression** as a trigger of type *Gene expression*, while the gold label is *Positive regulation*. Despite that fact that our model is capable of handling such scenarios sometimes given grounded biomedical concepts and factual reasoning paths, there is still room for improvement.

5 Related Works

Event Extraction Most existing event extraction systems focus on extracting events in news. Early attempts relied on hand-crafted features and a pipeline architecture (Gupta and Ji, 2009; Li et al., 2013). Later studies gained significant improvement from neural architectures, such as convolutional neural networks (Chen et al., 2015; Nguyen and Grishman, 2015), and recurrent neural networks (Nguyen et al., 2016). More recent studies leverages large pre-trained language models to obtain richer contextual information (Wadden et al., 2019; Lin et al., 2020). Another line of works utilized GNN to enhance event extraction performance. Liu et al. (2018) applied attention-based

graph convolution networks on dependency parsing trees. We instead propose a GNN, GEANet, for integrating domain knowledge into contextualized embeddings from pre-trained language models.

Biomedical Event Extraction Event extraction for biomedicine is more challenging due to higher demand for domain knowledge. BioNLP 11 GENIA event extraction task (GE'11) is the major benchmark for measuring the quality of biomedical event extraction system (Kim et al., 2011). Similar to event extraction in news domain, initial studies tackle biomedical event extraction with human-engineered features and pipeline approaches (Miwa et al., 2012; Björne and Salakoski, 2011). Great portion of recent works observed significant gains from neural models (Venugopal et al., 2014b; Rao et al., 2017b; Jagannatha and Yu, 2016; Björne and Salakoski, 2018). Li et al. (2019) incorporated information from Gene Ontology, a biomedical knowledge base, into tree-LSTM models with distributional representations. Instead, our strategy is to model two knowledge graphs from UMLS hierarchically with conceptual and semantic reasoning paths, providing stronger clues for identifying challenging events in biomedical corpus.

6 Conclusion

We have proposed a framework to incorporate domain knowledge for biomedical event extraction. Evaluation results on GE'11 demonstrated the efficacy of GEANet and hierarchical KG representation in improving extraction of non-indicative trigger words associated nested events. We also show that our method is robust when applied to different amount of training data, while being advantageous in low-resource scenarios. Future works include grounding adjective triggers to knowledge bases, better biomedical knowledge representation and extracting biomedical events at document level.

Acknowledgements

We thank Rujun Han for helpful advice during the development of our model. We also appreciate insightful feedback from PLUSLab members, and the anonymous reviewers. This research was sponsored by an NIH R01 (LM012592) and the Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600007. The views and conclusions of this paper are those of the authors and do not reflect the official policy or position of NIH, IARPA, or the US government.

References

- Sophia Ananiadou, Paul Thompson, Raheel Nawaz, John McNaught, and Douglas B Kell. 2015. Event-based text mining for biology and functional genomics. *Briefings in functional genomics*, 14(3):213–230.
- Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [SciBERT: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. [Extracting complex biological events with rich graph-based feature sets](#). In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.
- Jari Björne and Tapio Salakoski. 2011. [Generalizing biomedical event extraction](#). In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 183–191, Portland, Oregon, USA. Association for Computational Linguistics.
- Jari Björne and Tapio Salakoski. 2018. [Biomedical event extraction using convolutional neural networks and dependency parsing](#). In *Proceedings of the BioNLP 2018 workshop*, pages 98–108, Melbourne, Australia. Association for Computational Linguistics.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl.1):D267–D270.
- Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. [Event extraction via dynamic multi-pooling convolutional neural networks](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.
- Prashant Gupta and Heng Ji. 2009. [Predicting unknown time arguments based on cross-event propagation](#). In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 369–372, Suntec, Singapore. Association for Computational Linguistics.
- Rujun Han, I Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019a. Deep structured neural network for event temporal relation extraction. In *The 2019 SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.
- Rujun Han, Qiang Ning, and Nanyun Peng. 2019b. [Joint event and temporal relation extraction with shared representations and structured prediction](#). 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 434–444, Hong Kong, China. Association for Computational Linguistics.
- Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *Proceedings of EMNLP*.
- Lynette Hirschman, Gully AP Burns, Martin Krallinger, Cecilia Arighi, K Bretonnel Cohen, Alfonso Valencia, Cathy H Wu, Andrew Chatr-Aryamontri, Karen G Dowell, Eva Huala, et al. 2012. Text mining for the biocuration workflow. *Database*, 2012.
- Abhyuday N Jagannatha and Hong Yu. 2016. [Bidirectional RNN for medical event detection in electronic health records](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, San Diego, California. Association for Computational Linguistics.
- Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun’ichi Tsujii. 2009. [Overview of BioNLP’09 shared task on event extraction](#). In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado. Association for Computational Linguistics.
- Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun’ichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of the BioNLP shared task 2011 workshop*, pages 1–6. Association for Computational Linguistics.
- Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.
- 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.
- Qi Li, Heng Ji, and Liang Huang. 2013. [Joint event extraction via structured prediction with global features](#). In *Proceedings of the 51st Annual Meeting of*

- the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. Association for Computational Linguistics.
- 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.
- Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. [Jointly multiple events extraction via attention-based graph information aggregation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Brussels, Belgium. Association for Computational Linguistics.
- Amit Majumder, Asif Ekbal, and Sudip Kumar Naskar. 2016. [Biomolecular event extraction using a stacked generalization based classifier](#). In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 55–64, Varanasi, India. NLP Association of India.
- Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.
- Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Nomiya, and Jun'ichi Tsujii. 2006. [Semantic retrieval for the accurate identification of relational concepts in massive textbases](#). In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1017–1024, Sydney, Australia. Association for Computational Linguistics.
- SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44.
- Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. [Joint event extraction via recurrent neural networks](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.
- Thien Huu Nguyen and Ralph Grishman. 2015. [Event detection and domain adaptation with convolutional neural networks](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371, Beijing, China. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.
- 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.
- Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017a. Biomedical event extraction using abstract meaning representation. In *BioNLP 2017*, pages 126–135.
- Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017b. [Biomedical event extraction using Abstract Meaning Representation](#). In *BioNLP 2017*, pages 126–135, Vancouver, Canada,. Association for Computational Linguistics.
- Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1–12.
- Elaheh ShafieiBavani, Antonio Jimeno Yepes, and Xu Zhong. 2019. Global locality in event extraction. *arXiv preprint arXiv:1909.04822*.
- Martin Simonovsky and Nikos Komodakis. 2017. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3693–3702.
- Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2020. Enabling low-resource transfer learning across covid-19 corpora by combining event-extraction and co-training. In *ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID)*.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. [Graph attention networks](#). In *International Conference on Learning Representations*.
- Deepak Venugopal, Chen Chen, Vibhav Gogate, and Vincent Ng. 2014a. Relieving the computational bottleneck: Joint inference for event extraction with high-dimensional features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 831–843.

- Deepak Venugopal, Chen Chen, Vibhav Gogate, and Vincent Ng. 2014b. [Relieving the computational bottleneck: Joint inference for event extraction with high-dimensional features](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 831–843, Doha, Qatar. Association for Computational Linguistics.
- David Wadden, Ulme Wennberg, Yi Luan, and Hananeh 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.
- Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. [Does it make sense? and why? a pilot study for sense making and explanation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.
- Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Twenty-Eighth AAAI conference on artificial intelligence*.

A Implementation Details

Our models are implemented in PyTorch (Paszke et al., 2019). Hyper-parameters are found by grid search within search range listed in Table 4. The hyper-parameters of the best performing model are summarized in 5. All experiments are conducted on a 12-CPU machine running CentOS Linux 7 (Core) and NVIDIA RTX 2080 with CUDA 10.1.

To pre-train KGE, we leverage the TransE implementation from OpenKE (Han et al., 2018). All tuples associated with the selected nodes described in Section 3.1 are used for pre-training with margin loss and negative sampling,

$$\mathcal{L} = \sum_{(h,\ell,t) \in S} \sum_{(h',\ell,t') \notin S} \max(0, d(h,\ell,t) - d(h',\ell,t') + \gamma)$$

where γ denotes margin, and $d(x, x')$ denotes the $\ell - 1$ distance between x and x' . h and t are embeddings of head and tail entities from the gold training sets S with relation ℓ . (h', ℓ, t') denotes a corrupted tuple with either the head or tail entity replaced by a random entity. TransE is optimized using Adam (Kingma and Ba, 2015) with hyper-parameters illustrated in Table 6. Every 50 epochs, the model checkpoint is saved if the mean reciprocal rank on the development set improve from the last checkpoint; otherwise, training will be stopped.

B Dataset

The statistics of GE' 11 is shown in 7. The corpus contains 14496 events with 37.2% containing nested structure (Björne and Salakoski, 2011).⁷ We use the official dataset split for all the results reported.

Hyper-parameter	Range
Relation MLP dim.	{300, 500, 700, 1000}
Trigger MLP dim.	{300, 500, 700, 1000}
Learning rate	{ 1×10^{-5} , 3×10^{-5} , 5×10^{-5} }

Table 4: Hyper-paramter search range for fine-tuning SciBERT.

Hyper-parameter	Value
Relation MLP dim.	300
Trigger MLP dim.	300
Learning rate	3×10^{-5}
GEANet node dim.	300
GEANet edge dim.	300
GEANet layers	2
Dropout rate	0.2

Table 5: Hyper-paramters of the best performing GEANet-SciBERT model.

Hyper-parameter	Value
Learning rate	0.5
Margin	3
Batch size	128
# corrupted tuples / # gold tuples	25
# Epochs	500

Table 6: Hyper-paramters for pre-training KGE.

Metric	Number
events	14496
sentences	11581
nested events	37.2%
intersentence events	6.0%

Table 7: GE' 11 dataset statistics

⁷The dataset can be downloaded from <http://bionlp-st.dbcls.jp/GE/2011/downloads/>.