Knowledge Triplets Derivation from Scientific Publications via Dual-Graph Resonance

by

Kai Zhang

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

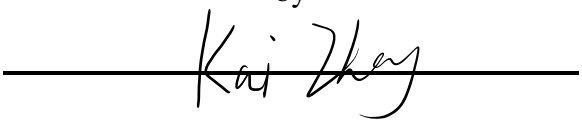
In partial fulfillment of the requirements for the

Degree of Master of Science

in

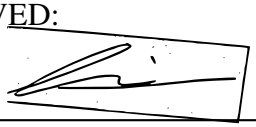
Computer Science

by

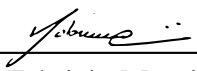


October 2022

APPROVED:



Professor Xiaozhong Liu, Thesis Advisor



Professor Fabricio Murai, Thesis Reader

Abstract

Scientific Information Extraction (SciIE) is a vital task and is increasingly being adopted in domain-specific (e.g., Biomedical) data mining to conceptualize and epitomize knowledge triplets from scientific literature. Existing relation extraction methods aim to extract explicit triplet knowledge from documents, however, they can hardly perceive unobserved factual relations. Recent generative methods have more flexibility, but their generated relations will encounter trustworthiness problems. In this paper, we propose a novel **E**xtraction-**C**ontextualization-**D**erivation (**ECD**) strategy to generate a document-specific and entity-expanded dynamic graph from a shared static knowledge graph. Then, we propose a novel **D**ual-**G**raph **R**esonance **N**etwork (**DGRN**) which can generate richer explicit and implicit relations under the guidance of static and dynamic knowledge topologies. Experiments conducted on a public PubMed corpus validate the superiority of our method against several state-of-the-art baselines. **We will release our code and data splits for reproducibility of results by the community.**

Acknowledgements

We gratefully acknowledge support from NSF, Award # 2122232 - SCISIPBIO: Constructing Heterogeneous Scholarly Graphs to Examine Social Capital DuringMentored K Awardees Transition to Research Independence: Explicating a Matthew Mechanism. Thank you to my advisors, Professor Xiaozhong Liu and to my reader, Professor Fabricio Murai. I would also like to thank Professor Xuhong Zhang of Indiana University and collaborators of Damo Academy for their support and help throughout the process of conducting this research. Furthermore, thank you to the Academic & Research Computing group at WPI, particularly James Kingsley. Finally, thank you to my friends and family for their continued support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As scientific literature grows at an expeditious pace, it becomes increasingly labor-intensive for scholars to curate the massive information and consume their interested knowledge. For example, PubMed, as one of the most commonly used searching biomedical publication databases, contains more than 34 million publications[1]. Even with a focused research interest, such as oncology, it is still very laborious to locate useful information from noisy retrieval results. To address this challenge, SciIE approaches can be employed to extract structural information from scientific articles, which has drawn great attentions from Natural Language Processing (NLP) community [11, 29].

Recently, great efforts have been made regarding SciIE tasks and obtained substantial achievements. [17] and [5] trained pre-trained language models with biomedical corpus for representation learning and downstream task fine-tuning. Based on these models, many extractive methods are proposed to extract triples explicitly present in the scientific documents [24, 38]. Despite some success, extractive methods can only obtain limited biomedical knowledge due to the limited searching space. To enable implicit biomedical triple extraction, generative extraction methods provide more flexibility, however they can not guarantee correctness of the generated knowledge and may cause trustworthiness

[1] https://pubmed.ncbi.nlm.nih.gov/about/

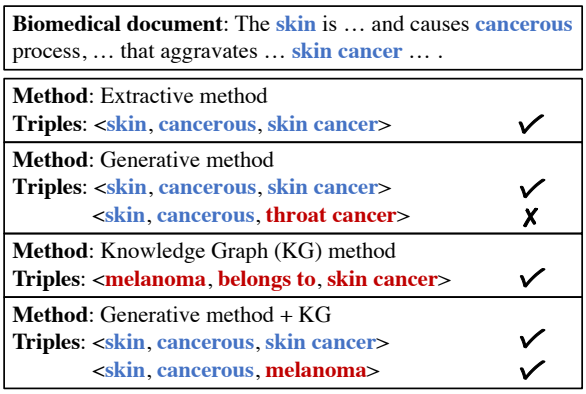


Figure 1.1: Comparison among different methods. Entities and relations observed in the document are denoted in red color, and unobserved ones are denoted in blue. Correct/wrong triplets are labeled as check/cross marks.

concerns [44]. For example in Figure 1.1, the generative method produces an incorrect triplet <*skin, cancerous, throat cancer*> because there exists no direct relation between entities "*skin*" and "*throat cancer*", while "*throat cancer*" is generated based on the trigger word "*cancerous*". Recent graph-based methods provide the possibility to produce trusted and unobserved biomedical triplets via multi-hop path-reasoning on graphs, such as <*melanoma, belongs to, skin cancer*> can be derived from <*skin, cancerous, skin cancer*> and <*skin, cancerous, melanoma*>. However, such methods only alleviate the problem and their extracted triplets are restricted by limited document entities.

High-quality and large-scale biomedical knowledge graphs (BKG) have been studied extensively and constructed [31], which provide possibilities to expand limited document entities and enable the generative methods to generate more trusted and unobserved biomedical triplets[2]. However, one may hesitate to adopt BKG directly due to the potential of massive noise. In this study, we propose a novel *Extraction-Contextualization-Derivation (ECD)* strategy to address this problem - using large-scale BKG as the *Static Graph* to encapsulate the biomedical domain knowledge and the derived knowledge sub-BKG as *Dynamic Graph* to characterize document-related knowledge. The interactions between static

[2]We resort to the biomedical domain since there exists rich data resources in both text and graph format.

and dynamic graphs ensure the comprehensiveness and trustworthiness of the knowledge generation.

Prior SciIE studies rarely equip generative methods with domain knowledge in an end-to-end fashion, and such investigation can be especially critical for biomedical knowledge mining which poses several challenges. Firstly, the input document and knowledge graph are information complementary and should be fully interactively modeled in the encoding process. Secondly, multi-hop path reasoning on graphs should be utilized for guiding copy mechanism to provide trusted biomedical triplets. To this end, we propose a novel Dual-Graph Resonance Network with a fourfold contribution:

• We propose an "Extraction-Contextualization-Derivation (ECD)" strategy to derive a document-related dynamic graph from a shared static graph which can be used for implicit entity expansion.

• We propose a "Dual-Graph Resonance Network (DGRN)" frame to generate both observed and unobserved knowledge triplets by jointly modeling the input document and the dual graphs.

• A new dataset (Bio-Sci), derived from open-source biomedical domain corpus, which contains 32,330 publications and each of them equipped with implicit and explicit triplets is released for biomedical NLP research community.

• Extensive experiments conducted on Bio-Sci show an average 4.91% improvement on F1 score against the best SOTA method which validate the effectiveness and superiority of our method.

Chapter 2

Related Work

Existing SciIE works extract or generate knowledge triplets from different parts of scientific publications, such as content [2, 20], abstract, introduction and citation sentences[1] [23]. The mainstream approaches can be concluded as follow.

2.1 Extractive Models

Extractive models have been extensively studied. [5, 33] integrated pre-trained language models into an encoder-decoder framework for performance improvements. [24, 46] introduced a joint learning framework to model connections between relations and their corresponding entity pairs. Moreover, [4, 28] utilized a hierarchical structure that featured connections among different content layers to find facts within the content summarization for relation extraction [19]. Further, with the emergence of powerful large language models, zero-shot fashion can be leveraged for triple extraction via chatting with such models [32]. However, these methods ignore fine-grained entity-level information interaction and integration.

[1] called citance in the following sections

2.2 Graph-based Models

Graph-based models can provide both dependencies among entities and path reasoning potential for inference. [41] proposed a mention-to-entity graph aggregation model which can capture the relation of entities across sentences. Instead of integrating graph structure into neural network models, [10,13,26,36] enhanced the mention-to-entity graph paradigm by introducing multi-hop path reasoning and reconstructing the graph based on the obtained path information. Another fashion is to transform sentences into a graph and perform multi-view GCN to obtain the relation [6,27]. Unfortunately, existing graph-based models can not synthesize new domain knowledge.

2.3 Generative Models

Generative models are recently proposed to generate triplets flexibly from input documents. [42] proposed a CopyRE model to select entities or relations via copy mechanism. [38,40] further improved this paradigm by introducing multi-task learning and contrastive learning frameworks. Other generative models utilized additional information. For example, [45] proposed a Knowledge-Graph (KG)-enriched Abstract Meaning Representation (AMR) framework which uses external information to enrich the AMR graph extracted from scientific papers. [8] leveraged transformers to refine semantic embedding of a given text for better generation [9,43]. However, [44] claimed the factual correctness and trustworthiness problems of these methods which ignore prior knowledge to ensure that generated knowledge triplets are more reliable.

While these studies have accomplished notable advancements, our DGRN steps further by leveraging the synergistic power of generative and graph-based methods within a cohesive framework.

Chapter 3

Dual-Graph Construction

Generative extraction methods may encounter trustworthiness issues. To address this, we propose enhancing these methods with guidance from biomedical knowledge. We first introduce a static background knowledge graph (BKG), which is built on public resources. We then generate a document-specific dynamic BKG using a dynamic graph generator, which enables the generation of unobserved biomedical triplets. The detail construction process is depicted in Figure 3.1.

3.1 Static Biomedical Knowledge Graph

Large-scale biomedical knowledge graphs are recently constructed, and some have been made publicly available. [31] constructed a public knowledge graph with 1.47 million triplets and 96,397 entities from multiple sources, such as PubMed, DrugCentral etc. Thus, the knowledge graph owns massive biomedical knowledge, which will expand document entities greatly. Given any BKG as an external public resource, we formulate it as a shared static graph $G_S = \{(e_i, r_{i,j}, e_j) | e_i, e_j \in \mathcal{E}, r_{i,j} \in \mathcal{R}\}$, where $\mathcal{E}$ and $\mathcal{R}$ represent an entity set and a relation set respectively. However, direct application of G_S may introduce noisy information into generative models and pollute the SciIE results. Thus, we propose a

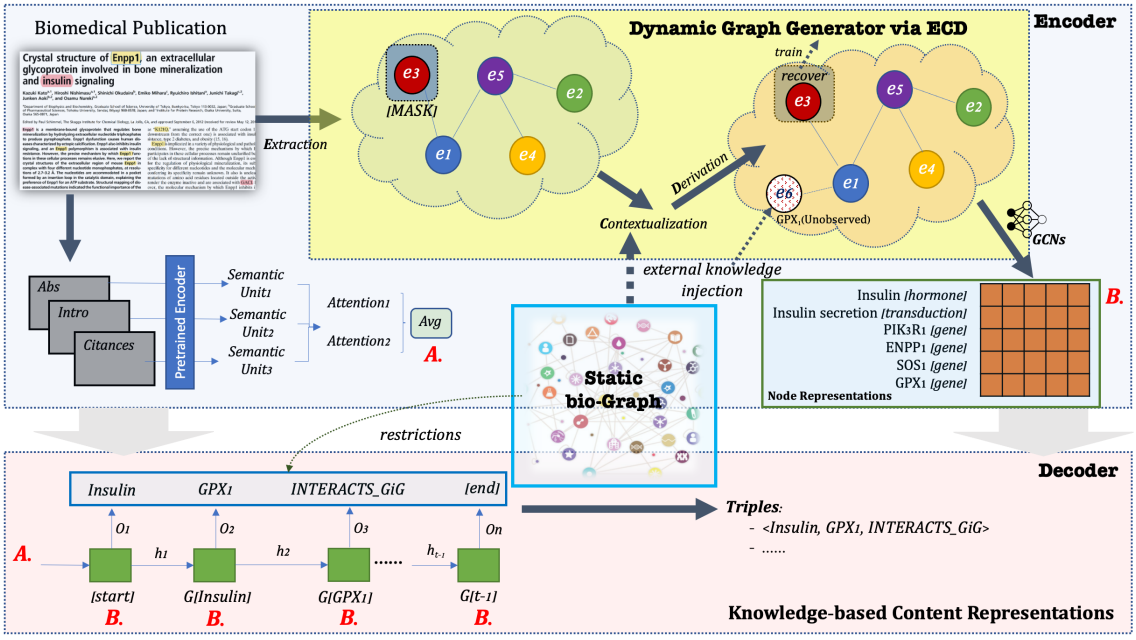


Figure 3.1: The DGRN architecture and ECD strategy. Given any biomedical publication, a pretrained language model and the ECD strategy can be used for content information modeling and dynamic graph generation. Then, both content and expanded entity-level information (denoted in red alphabets) can be used for decoder to generate observed and unobserved under the guidance of dual-graph.

dynamic biomedical knowledge graph construction strategy.

3.2 Dynamic Sub-Graph

Inspired by [25, 29], we let $D = \{S_{abs}, S_{int}, S_{cit}\}$ be a biomedical document, which consists of three important sections, i.e., *abstract*, *introduction* and *citance*. Each section $S_* = \{w_i\}_{i=1}^{N_*}$ refers to a sequence of words of length N_*. We aim to derive a dynamic graph automatically by enriching consisted document relations under the prior guidance of the static graph G_S. *Note that we consider the three sections together when constructing the dynamic graph*. Particularly, we propose an **E**xtraction-**C**ontextualization-**D**erivation (**ECD**) framework which consists of three key steps:

3.2.1 Extraction

Extraction. We first resort to the tool of SciSpacy[1] [25], which extracts all the biomedical entities from the document D and obtain an entity (i.e., node) set $\mathcal{E}^\circ$. Then, we retrieve pre-defined relations from the static graph G_S and produce an edge set $\mathcal{R}^\circ$. Finally, we construct an initial dynamic graph $G_D^\circ = \left\{(e_i, r_{ij}, e_j) | e_i, e_j \in \mathcal{E}^\circ, r_{ij} \in \mathcal{R}^\circ\right\}$, where e_i denotes a head entity, e_j is a tail entity, and r_{ij} denotes the relation. G_D° contains limited document entities and it should be expanded to cover richer biomedical knowledge.

3.2.2 Contextualization

Contextualization. Intuitively, we can choose k-hop expansion (*we use one-hop in our experiments*) for contextualization because the directly connected entities are always similar. However, such a simple expansion will bring ungovernable noisy entities and relations, which can pollute the dynamic graph generation quality. Instead, we propose a Dynamic Graph Generator (DGG) which produces triplets and expands G_D° as $G_D =$

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

$\{(e_i, r_{i,j}, e_j)|e_i, e_j \in \mathcal{E}_D, r_{i,j} \in \mathcal{R}_D\}$, where $\mathcal{E}_D$ and $\mathcal{R}_D$ denote the expanded entity set and relation set respectively, $\mathcal{E}^\circ \subseteq \mathcal{E}_D$ and $\mathcal{R}^\circ \subseteq \mathcal{R}_D$.

Inspired by the masked language model [7], we can randomly mask entities in the graph G_D° and train a DGG to recover the original topology. The trained generator, then, can detect the unobserved triplets while avoiding introducing noise. The training procedure is detailed below:

We mask 10% entities randomly from G_D° and label the masked entities $\mathcal{E}_+^\circ$ as positive instances, and label the remaining unmasked entities $\mathcal{E}_-^\circ$ as negative instances, and $\mathcal{E}^\circ = \mathcal{E}_+^\circ \cup \mathcal{E}_-^\circ$.

Then, we let $\mathbf{e}_i$ be the representation of any entity $\mathbf{e}_i \in \mathcal{E}_-^\circ$ and $\mathbf{e}_j$ be the representation of any entity $\mathbf{e}_j \in \mathcal{E}_+^\circ$. The initial entity representations are obtained based on pre-trained language models, like BERT and its variants [7, 17] as can be seen in Section 4.2. For each entity e_j to be recovered, we can calculate the selected probability based on an attention mechanism as below:

$$\begin{aligned} p(e_j|\mathcal{E}_-^\circ) &= \delta\Big(\text{Dense}\Big(\sum_{e_i \in \mathcal{E}_-^\circ} \alpha_i \mathbf{e}_i\Big)\Big) \\ \alpha_i &= \sigma\big([\mathbf{e}_i; \mathbf{e}_j]\mathbf{W}_0\big) \end{aligned} \tag{3.1}$$

where $\delta(\cdot)$ is a sigmoid function, $\text{Dense}(\cdot)$ is a fully-connected layer, $[\ ;\ ]$ denotes a vector concatenation operation, $\sigma(\cdot)$ denotes the ReLU function and $\mathbf{W}_0$ represents a trainable weight matrix.

To train the DGG, we use the cross-entropy function to minimize the loss between the masked entities and their corresponding recovered ones:

$$\mathcal{L}_{DGG} = \frac{1}{|\mathcal{E}_+^\circ|} \sum_{e_j \in \mathcal{E}_+^\circ} -\log\Big(p\big(e_j|\mathcal{E}_-^\circ\big)\Big) \tag{3.2}$$

3.2.3 Derivation

Derivation In the derivation step, we feed the initial dynamic graph G_D° to the trained DGG which considers one-hop entities as positive instances and the initial entities $\mathcal{E}^\circ$ as negative instances. Then, we choose those entities with high probabilities[2] and produce the expanded entity set $\mathcal{E}_D = \mathcal{E}^\circ \cup \{e_j\}_{\geq Threshold}$ and the expanded dynamic graph G_D.

[2]The hyper-parameter of $Threshold$ is set as 0.8.

Chapter 4

Dual-Graph Resonance Network

In this section, we propose an innovative Dual-Graph Resonance Network (DGRN) which combines the dual graphs and generative method into a unified framework. The method is depicted in Figure 3.1, which consists of three modules: *Text Encoder*, *Graph Encoder* and *Triplet Decoder*.

4.1 Text Encoder Module

Given any document D, its constituent section S_* is fed into a BioBERT [17] to produce token-level representations ($\mathbf{H}^{abs} = \{\mathbf{h}_i^{abs}\}$, $\mathbf{H}^{int} = \{\mathbf{h}_i^{int}\}$ and $\mathbf{H}^{cit} = \{\mathbf{h}_i^{cit}\}$) and section-level representations ($\mathbf{h}_{CLS}^{abs}$, $\mathbf{h}_{CLS}^{int}$ and $\mathbf{h}_{CLS}^{cit}$). [29] found that citation sentences are more relevant to the document topics in comparison with abstract and introduction. Based on this, we opt for abstract-aware attention to measure the importance of each token representation $\mathbf{h}_i^{cit}$ through a scoring function using a feed-forward neural network:

$$\begin{aligned} \alpha_i^{abs} &= \text{softmax}\big((\mathbf{h}_{CLS}^{abs})^{\mathrm{T}}\sigma(\mathbf{W}_1\mathbf{h}_i^{cit} + \mathbf{b}_1)\big) \\ \mathbf{p}^{abs-cit} &= \sum_{i\in[1,N_{cit}]} \alpha_i^{abs} \times \mathbf{h}_i^{cit} \end{aligned} \tag{4.1}$$

Similarly, we also use introduction-aware attention to measure the importance of each

token representation $\mathbf{h}_i^{cit}$ through a scoring function as below:

$$\begin{aligned} \alpha_i^{int} &= \text{softmax}\big((\mathbf{h}_{CLS}^{int})^{\mathrm{T}}\sigma(\mathbf{W}_2\mathbf{h}_i^{cit} + \mathbf{b}_2)\big) \\ \mathbf{p}^{int-cit} &= \sum_{i\in[1,N_{cit}]} \alpha_i^{int} \times \mathbf{h}_i^{cit} \end{aligned} \tag{4.2}$$

where $\mathbf{W}_*$ and $\mathbf{b}_*$ are trainable model parameters, $\mathbf{p}^{abs-cit}$ and $\mathbf{p}^{int-cit}$ represent the abstract-aware citance representation and introduction-aware citance representation, respectively. Finally, we take the average of the sum of the two vectors and obtain the document representation $\mathbf{p}^{doc}$ as below:

$$\mathbf{p}^{doc} = \frac{1}{2}(\mathbf{p}^{abs-cit} + \mathbf{p}^{int-cit}) \tag{4.3}$$

4.2 Graph Encoder Module

We apply multi-layer Graph Convolutional Network (GCN) [15] on the dynamic graph G_D to aggregate the features from neighbors to obtain node representation. For any entity e_i at the l-th layer, the graph convolutional operation can be applied by the formula as below:

$$\mathbf{e}_i^{(l+1)} = \sigma\Big(\sum_{e_j\in\mathcal{N}_{e_i}} \mathbf{W}_3^{(l)}\mathbf{e}_j^{(l)} + \mathbf{b}_3^{(l)}\Big) \tag{4.4}$$

where $\mathcal{N}_{e_i}$ denotes the neighbors for the node e_i, $\mathbf{W}_3^{(l)}$ and $\mathbf{b}_3^{(l)}$ are trainable model parameters. In the initial stage, the $\mathbf{e}_i^{(0)} = \frac{1}{t-s+1}\sum_{i\in[s,t]}\mathbf{h}_i^*$ indicates entity i ranges from t-th token to s-th token in any section S_*. Intuitively, the GCN layers can encapsulate rich topological information.

4.3 Triplet Decoder Module

A decoder is adopted to generate knowledge triplets. Given training data, the decoder can copy an entity from the graph G_D as the head entity of the triplet, and then generate a relation for the triplet. Then, it can copy the tail entity from G_D. Repeating this process, the decoder could generate multiple triplets. In time step t $(1 \leq t)$, we can calculate the decoder output $\mathbf{o}_t$ and hidden state $\mathbf{h}'_t$ as follows:

$$\mathbf{o}_t, \mathbf{h}'_t = f(\mathbf{x}_t, \mathbf{h}'_{t-1}) \tag{4.5}$$

where $f(\cdot)$ represents the RNN-based decoder function, $\mathbf{h}'_{t-1}$ indicates the hidden state of time step $t-1$, and $\mathbf{x}_t$ is the input representation of time step t and defined as below:

$$\mathbf{x}_t = [\mathbf{o}_{t-1}; \mathbf{c}_t]\mathbf{W}_4 \tag{4.6}$$

where $\mathbf{o}_{t-1}$ denotes the entity or relation representation copied from dynamic graph in time step $t-1$, $\mathbf{c}_t$ is the attention vector [3] and $\mathbf{W}_4$ is a trainable weight matrix. In the initial step, input representation $\mathbf{x}_0 = \mathbf{p}^{doc}$ (see Eq. 4.3).

4.3.1 Attention Vector

Attention Vector Entities and relations are generated and treated differently based on their different positions. In the time step t $(t\%3 = 0, 1)$ (generating head or tail), the attention vector $\mathbf{c}_t$ is calculated by copying entities from the entity set $\mathcal{E}_D$ by the following formula:

$$\begin{aligned} \mathbf{c}_t &= \sum_{e_i \in \mathcal{E}_D} \beta_i \times \mathbf{e}_i \\ \beta_i &= \text{softmax}\Big(\sigma\big([\mathbf{h}'_{t-1}; \mathbf{e}_i]\mathbf{W}_5\big)\Big) \end{aligned} \tag{4.7}$$

where $\mathbf{h}'_{t-1}$ is the hidden state of the decoder in the $t-1$ time step, and $\mathbf{W}_5$ are trainable parameters.

In the time step t $(t\%3 = 2)$ (generating relation), $\mathbf{c}_t$ can be calculated by copying relations from relation set $\mathcal{R}_D$ by the following formula:

$$\begin{aligned} \mathbf{c}_t &= \sum_{r_{ij} \in \mathcal{R}_D} \gamma_{i,j} \times \mathbf{r}_{i,j} \\ \gamma_{i,j} &= \text{softmax}\Big(\sigma\big([\mathbf{h}'_{t-1}; \mathbf{r}_{i,j}; \mathbf{p}_{i,j}]\mathbf{W}_6\big)\Big) \end{aligned} \tag{4.8}$$

where $\mathbf{W}_6$ are trainable parameters, $\mathbf{r}_{i,j}$ is the relation representation, $\mathbf{p}_{i,j}$ is the representation of edge between entity e_i and e_j via path reasoning.

4.3.2 Path Reasoning

Path Reasoning In the dynamic graph G_D, the head entity and tail entity in a triplet are not always directly connected. Thus, we introduce a path reasoning method that can model dependency among entities with multi-hop distances in the graph. Similar to [41], given the head entity and tail entity, we can define the representation of directed edge from entity e_i to entity e_j as below:

$$\mathbf{e}_{ij} = \sigma(\mathbf{W}_7[\mathbf{e}_i; \mathbf{e}_j] + \mathbf{b}_7) \tag{4.9}$$

where $\mathbf{W}_7$ and $\mathbf{b}_7$ are trainable parameters, $\mathbf{e}_i$ and $\mathbf{e}_j$ denote representations of entity e_i and entity e_j respectively (see Eq. 4.4).

Based on the vectorized edge representation, the path between head entity e_i and tail entity e_j passing through entity e_o is represented as follow:

$$\mathbf{p}_{i,j} = [\mathbf{e}_{i,o}; \mathbf{e}_{o,j}; \mathbf{e}_{j,o}; \mathbf{e}_{o,i}] \tag{4.10}$$

For computation efficiency, we choose one-hop path, while it can be extended to multi-hop

paths.

4.3.3 Entity Prediction

Entity Prediction To copy a head/tail entity, we calculate the confidence vector $\mathbf{q} = [q_1, ..., q_{|\mathcal{E}_D|}]$ for all the entities in $\mathcal{E}_D$. We also apply a softmax on $\mathbf{q}$ to obtain the probability distribution $\mathbf{p}^{entity} = [p_1^{entity}, ..., p_{|\mathcal{E}_D|}^{entity}]$ by the formulas as below:

$$\begin{aligned} q_t &= \sigma(\mathbf{o}_t \mathbf{W}_8 + \mathbf{b}_8) \\ \mathbf{p}^{entity} &= \text{softmax}(\mathbf{q}) \end{aligned} \tag{4.11}$$

where $\mathbf{W}_8$ and $\mathbf{b}_8$ are trainable parameters. We select the entity with the highest probability as the predicted entity and use its embedding to produce the next time step input $\mathbf{c}_{t+1}$. *Note that the tail entity can not be the same as the head entity*.

4.3.4 Relation Prediction

Relation Prediction We now use a fully connected layer to calculate a confidence vector $\mathbf{q}' = [q'_1, ..., q'_{|\mathcal{R}_D|}]$ of all the valid relations. Specifically, we apply a softmax on $\mathbf{q}'$ to obtain the probability distribution $\mathbf{p}^{relation} = [p_1^{relation}, ..., p_{|\mathcal{R}_D|}^{relation}]$ by the formulas as below:

$$\begin{aligned} \mathbf{q}' &= \sigma(\mathbf{o}_t \mathbf{W}_9 + \mathbf{b}_9) \\ \mathbf{p}^{relation} &= \text{softmax}(\mathbf{q}') \end{aligned} \tag{4.12}$$

where $\mathbf{W}_9$ and $\mathbf{b}_9$ are trainable parameters. We select the relation with the highest probability as the prediction relation and use its embedding to produce input $\mathbf{c}_{t+1}$ in the next time step.

Objective Function Our DGRN is trained with the negative log-likelihood loss func-

tion. Suppose $Y = \{y_1, y_2, ..., y_T\}$ is the target result for triplets generation, the loss function is defined as:

$$\mathcal{L}_{SEQ} = \frac{1}{T} \sum_{t=1}^{T} -log\big(p(y_t|y_{<t})\big) \tag{4.13}$$

where T is the maximum steps of the decoder, $p(y_t|y_{<t})$ denotes the conditional probability of target y_t given previous output sequence $y_{<t}$.

There are two negative log-likelihood loss functions used for training our DGRN; one is DGG loss (see Eq. 3.2) and one is the seq2seq loss (see Eq. 4.13). We optimize $\mathcal{L}_{DGG}$ and $\mathcal{L}_{SEQ}$ iteratively. We use backpropagation to calculate the gradients of all the trainable parameters and update them with Adam optimizer [18].

Chapter 5

Experiments

In this section, we describe the detailed experimental settings include dataset usage, parameters setting, results analysis and etc.

5.1 Dataset

5.1.1 Motivation

Motivation One of our contributions in this work is that we construct a new dataset, Bio-Sci, that derived from open-source PubMed corpus. The motivation behind Bio-Sci is that existing datasets rarely contain implicit relation triples and most relations are derived from sentence level [21]. Our assumption is that some applications such as paper recommendation reason generation requires not only knowledge from query words but also from cross-sentence/document relations and thus we propose to inject both explicit and implicit triples for PubMed publications for a further investigation towards real-world applications such as paper recommendation.

5.1.2 Construction

Construction Following [31], we use their public triplet dataset built from multiple public datasets[1] as the bases to construct our static biomedical knowledge graph. This dataset consists of 1,426,025 triplets, 41,078 entities, and 27 relation types. Besides, we also construct a dataset with 32,330 biomedical publications collected from PubMed Central[2], and split the dataset into training, development and testing sets with a split of 22,330/5,000/5,000. "To generate knowledge triplets for every training/development document, we draw inspiration from the work of [35] and [21] and employ the pubmed_parser [1] to extract biomedical concepts for establishing connections between the entities in the Static Graph and each document. Subsequently, we annotate the respective triples for each document. In the case of the testing set, we deliberately omit 10% of the entities and their corresponding sentences from each document, allowing us to assess the model's capability in entity recovery.

5.1.3 Human Agreement Test

Human Agreement Test To evaluate the quality of our dataset, we conduct human agreement tests on a randomly selected small-scale and human-annotated testing set. Specifically, we first randomly selected 50 documents and provide 8 triplets as label candidates for each document. Two annotators with professional biomedical knowledge participated in the annotation task and chose the most appropriate triplets. The Kappa value of the manual annotation is 0.88. Besides, we conduct a paired t-test between the best performance baseline and our DGRN, the p-value is less than 0.05.

[1] https://synlethdb.sist.shanghaitech.edu.cn/#/download
[2] https://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/oa_comm

5.2 Experimental Settings

Since our experiments are conducted on biomedical data, we choose BioBert as our text encoder and implement DGRN with Pytorch and DGL [30]. The hyper-parameter settings are detailed in Table 5.1. We follow the commonly used *precision*, *recall* and *F1-score* as the evaluation metrics [37]. All the methods run on a server configured with 4 Tesla P100 GPUs, 16 CPUs, and 512G memory.

5.3 Baselines

Extractive models can extract triplets directly from documents based on sequence classification models. Generative models can generate triplets based on the Seq2Seq framework. Graph-based models usually use graph structure to assist with extractive or generative models. In Table 5.2, we compare our DGRN with several state-of-the-art baselines:

HRL is a hierarchical extraction paradigm which approaches relation extraction via hierarchical reinforcement learning [28].

CASREL is a cascade binary tagging framework, which models relations as functions that map subjects to objects in a sentence [33].

CopyRE is a Seq2Seq model which leverages copy mechanism through textual similarity for relation copy. [42].

CopyMTL is a multi-task learning framework equipped with copy mechanism to allow the model to predict multi-token entities [40].

GAIN is a double graph reasoning network that aggregates mentions and their paths for better triplets extraction. [41].

AGGCN is a soft-pruning approach which automatically selects the relevant substructures of text for the relation extraction. [10].

For fair comparisons, we also utilize some state-of-the-art works that use pre-trained

GCN for Dual-Graph		DGRN Training	
Parameter	**Value**	**Parameter**	**Value**
Number of layers	1, **2**, 3	learning rate	$\mathbf{10^{-3}}$, 10^{-5}
emb size	100	dropout	0.2, **0.5**, 0.8
hidden size	808	batch_size	10, **20**, 50
		weight decay	$\mathbf{10^{-3}}$

Table 5.1: The experimental settings of our method. The best parameter settings are highlighted.

models:

KECI is an end-to-end model which utilizes external domain knowledge graph for joint entity and relation extraction [16].

TEMPGEN is a a cross-attention guided model for triplet template generation [12].

KeBioLM is a biomedical pretrained language model which uses external knowledge graph for relation extraction [39].

UmlsBert integrates domain knowledge during the pre-training process for knowledge augmentation [22].

Note that for fair comparisons, we retain their best performance settings reported in their paper and use BioBert as the base for all the baselines that use pre-trained model.

5.3.1 Comparative Study

Table 5.2 provides the main experiment results. One can witness the DGRN superiority compared with the best-performed baseline (KECI) with 4.90% improvement in F1. Interestingly, we find that generative methods can not compete with extractive approaches since generative methods require more searching space in the decoding stage than extractive models which only need to select target tokens among candidates. Meanwhile, the

Model	Model Type	P	R	F1
HRL	Extractive	61.17	21.81	32.16
CASREL	Extractive	71.12	32.94	45.03
CopyRE	Generative	54.73	25.72	34.99
CopyMTL	Generative	56.91	29.64	38.98
GAIN	Graph	56.01	21.43	31.00
AGGCN	Graph	61.43	33.91	43.70
KeBioLM	KG	61.18	32.88	42.77
UmlsBert	KG	59.61	29.17	39.17
KECI-BioBert	Extractive	61.93	**40.81**	49.20
TEMPGEN-BioBert	Generative	63.13	33.71	43.95
GAIN-BioBert	Graph	56.61	21.87	31.55
DGRN	Generative+Graph	**73.77***	39.69*	**51.61***

Table 5.2: Comparison among different models. Models above the double line do not use pre-trained model. Superscript * indicates statistical significance at $p < 0.05$ level compared to the best performance of baselines.

graph-based models achieve convincing performance, which indicates that graphs can provide essential topological information for knowledge extraction. We further investigate the performance of other state-of-the-art works using pre-trained model (BioBert). As can be observed in Table 5.2, DGRN achieves a 3.95% F1 improvement than TEMPGEN-BioBert and a 14.42% improvement in precision. However, we notice that KECI-BioBert gets the highest recall among all models. This can be explained by the fact that KECI-BioBert first constructs a span graph to filter the noisy tokens before encoding the input while our method first encodes the sentence and can only process 1024 tokens in each iteration. However, our DGRN still performs the best which can be attributed to the aggregation of interactions between two graph. Moreover, we also compared our DGRN with knowledge-based methods of which KeBioLM achieves a 42.77 score on F1. We

Model	ECD Step	P	R	F1
Extract	1	47.07	23.02	30.92
Extract+Text	1	59.03	38.21	46.39
Extract+Context+Text	1,2	41.17	20.81	27.65
DGRN (Full)	1,2,3	**73.77**	**39.69**	**51.61**

Table 5.3: Ablation study on DGRN components.

attributed the incomparable results to the fact that KeBioLM is designed for sentence-level relation extraction. Overall, DGRN outperforms all baselines because it utilizes topology information with dual-graph resonance to guide knowledge triplet generation while avoids disturbance from noisy information, which assures its effectiveness.

5.3.2 Ablation Study

We also explored how different configurations can impact our model's performance. **Extract** only uses the initial graph. **Extract+Text** considers text modeling additionally. **Extract+Context+Text** considers one-hop path expansion by adding up to 20 entities. **DGRN** is our fully configured model. The experimental results are displayed in Table 5.3.

As can be seen from Table 5.3, **Extract+Text** outperforms **Extract** due to the semantic richness provided by text content. However, interestingly, **Extract+Context+Text** performs the worst which we attribute to the fact that a simple expansion strategy will introduce noisy entities/relations from the static knowledge graph. Beyond the encoder part, we also conduct experiments on by removing the Path Reasoning within the decoder. The results are 67.83, 38.87, 49.61 for precision, recall and f1, respectively. Our fully configured **DGRN** considers all the necessary components and achieves the best result, which again validates the effectiveness of our *Extraction-Contextualization-Derivation* strategy and the complete DGRN frame.

Model	ECD Step	P	R	F1
Extract	1	45.61	19.97	27.78
Extract+Text	1	51.53	37.62	43.49
Extract+Context+Text	1,2	39.01	19.22	25.75
DGRN (Full)	1,2,3	**73.77**	**39.69**	**51.61**

Table 5.4: Robustness study on dynamic graph generation with training instances containing more than 10 triplets.

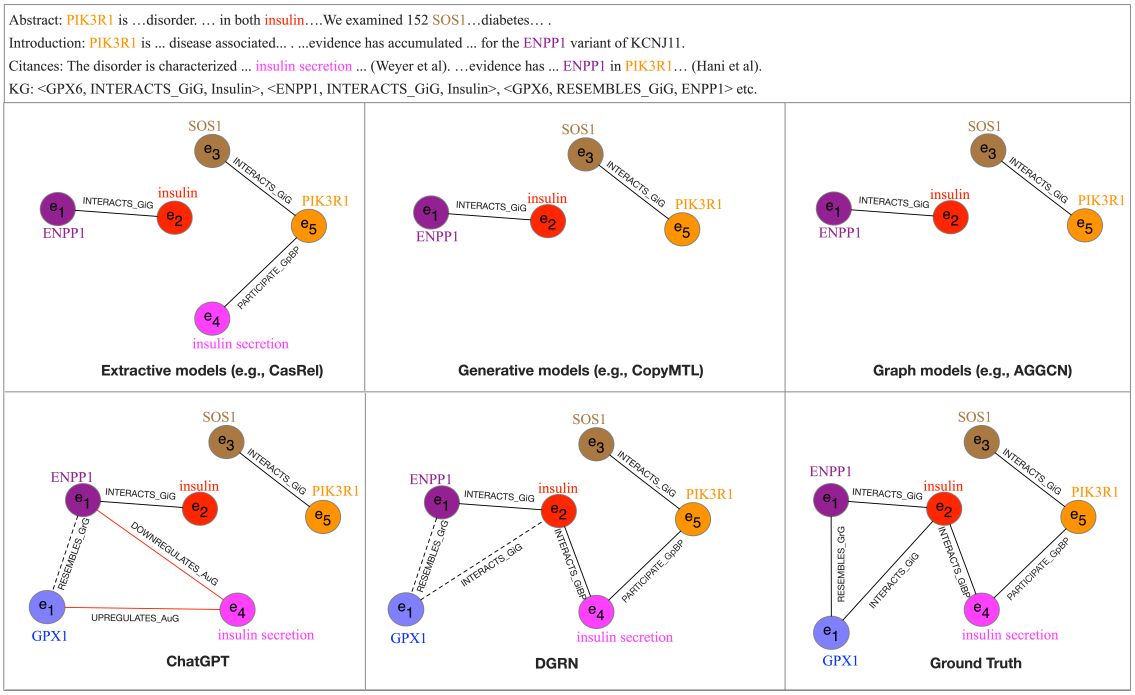


Figure 5.1: The exemplar generation results of baseline models (CASREL, CopyMTL, AGGCN) and DGRN.

Model	Mask	Precision	Recall	F1-score
DGRN (Full)	**10%**	**73.77**	**39.69**	**51.61**
	15%	70.91	38.02	49.50
	20%	69.21	37.17	48.37
	25%	60.71	32.03	41.94

Table 5.5: Results on different percent of masked entities.

5.3.3 Robustness Study

In contrast to shorter documents, intricate and lengthy documents present a greater challenge for SciIE. Furthermore, the number of triplets within each document significantly influences the model's performance. Consequently, we have established an additional training set comprising instances with more than 10 triplets. The results in Table 5.4 clearly illustrate that complex documents have a detrimental impact on the performance of less sophisticated models. Nevertheless, the fully configured DGRN remains largely unaffected due to the synergistic relationship between the document content and the dynamically generated graph. This observation underscores the robustness of our model.

5.3.4 Masking Study

To validate the masking efficiency, we mask different percentages of nodes for DGG training. The experimental results are presented in Table 5.5.

Table 5.5 reveals a notable trend: as the percentage of masked entities in the dynamic graph increases, the model's performance experiences a decline. This phenomenon is rooted in the dynamic graph's initiation process, which starts from a single document. As a result, the graphs constructed are typically rather small in scale, consisting of merely 10 to 20 nodes. Consequently, a higher masking percentage introduces more unseen

entities, thereby amplifying the challenge of successful recovery. In summary, masking approximately 10% of nodes proves beneficial in augmenting the DGG's ability to acquire valuable ontological information.

5.3.5 Case Study

As displayed in Figure 5.1, while other baselines successfully produce correct triplets, they can not compete with DGRN since they can't learn well from both rich BKG and the latent content. DGRN also shows its inference ability by generating "<GPX1, RESEMBLES_GiG, ENPP1>" "<GPX1, INTERACTS_GiG, Insulin>" based on the evidence derived from BKG. We further leverage in-context learning to let ChatGPT, one of the current most powerful AI tools, tackle this task. Despite the wrong relations it predicts for "insulin secretion", it still surprises us by knowing the "RESEMABLE_GrG" relation between "GPX1" and "ENPP1" since this requires inference from external BKG. However, the better results of DGRN again support the superiority of our model.

Chapter 6

Discussion and Conclusion

In this paper, we discuss the challenges of existing extractive and generative methods and make two efforts to deal with them. First, we propose a novel Extraction-Contextualization-Derivation (ECD) strategy to generate a document-specific dynamic graph. Then, we propose a novel Dual-Graph Resonance Network (DGRN) to generate richer triplets under the guidance of dual-graph. Extensive experiments validate the effectiveness of our proposed method. We also release a new dataset comes from Biomedical domain to encourage the community further explore this task.

Chapter 7

Future Work

In the future, we will integrate heterogeneous graph simplification [34] and sub-graph mining [14] into our frame for better performance. Also, along with the representation learned from each paper and its citation network, we could use the triple extracted by our DGRN as backbone for paper recommendation reasons generation.

Chapter 8

Supplementary Material

Appendix A. SciSpacy Details

For SciSpacy usage, we use them to obtain the entities with "en_core_sci_sm" and "en_ner_bionlp13cg_md" corpus because these two settings are trained in large-scale biomedical domain corpus. The f1 score is 70.87% on the entity extraction task for "en_core_sci_sm", and 86.75% for "en_ner_bionlp13cg_md". We mainly leverage "en_ner_bionlp13cg_md" in our experiments. What we do for linking entities is we first extract the entities and retrieve the BKG to see if there are matches for them.

Appendix B. Metrics

Following the partial match strategy, a generated triple is regarded as correct if the predicted relation and both the subject and object entity are correct. In this context, TP is the number of correctly generated/extracted triples, FN is the number of triples that the model failed to generate/extract, and FP is the number of incorrectly generated/extracted triples.

Bibliography

[1] Titipat Achakulvisut, Daniel Acuna, and Konrad Kording. Pubmed parser: A python parser for pubmed open-access xml subset and medline xml dataset xml dataset. *Journal of Open Source Software*, 5(46):1979, 2020.

[2] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Hongxiao Bai and Hai Zhao. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

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

[6] Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985, 2022.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Krishna Garg, Jishnu Ray Chowdhury, and Cornelia Caragea. Keyphrase generation beyond the boundaries of title and abstract. *arXiv preprint arXiv:2112.06776*, 2021.

[9] John Giorgi, Gary D Bader, and Bo Wang. A sequence-to-sequence approach for document-level relation extraction. *arXiv preprint arXiv:2204.01098*, 2022.

[10] Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. *arXiv preprint arXiv:1906.07510*, 2019.

[11] Sonal Gupta and Christopher Manning. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1–9, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

[12] Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. Document-level entity-based extraction as template generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5257–5269, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[13] Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. Three sentences are all you need: Local path enhanced document relation extraction. *arXiv preprint arXiv:2106.01793*, 2021.

[14] Zhuoren Jiang, Zheng Gao, Jinjiong Lan, Hongxia Yang, Yao Lu, and Xiaozhong Liu. Task-oriented genetic activation for large-scale complex heterogeneous graph embedding. In *Proceedings of The Web Conference 2020*, pages 1581–1591, 2020.

[15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[16] Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6248–6260, Online, August 2021. Association for Computational Linguistics.

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

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[19] Keming Lu, I Hsu, Wenxuan Zhou, Mingyu Derek Ma, Muhao Chen, et al. Summarization as indirect supervision for relation extraction. *arXiv preprint arXiv:2205.09837*, 2022.

[20] Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 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, June 2019. Association for Computational Linguistics.

[21] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282, 2022.

[22] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online, June 2021. Association for Computational Linguistics.

[23] Preslav I Nakov, Ariel S Schwartz, Marti Hearst, et al. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR*, volume 4, pages 81–88. Citeseer, 2004.

[24] Tapas Nayak and Hwee Tou Ng. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8528–8535, 2020.

[25] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

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

[27] Kashif Rabbani, Matteo Lissandrini, and Katja Hose. Extraction of validating shapes from very large knowledge graphs. *Proceedings of the VLDB Endowment*, 16(5):1023–1032, 2023.

[28] Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. A hierarchical framework for relation extraction with reinforcement learning. In *AAAI*, pages 7072–7079, 2019.

[29] Vijay Viswanathan, Graham Neubig, and Pengfei Liu. Citationie: Leveraging the citation graph for scientific information extraction. *arXiv preprint arXiv:2106.01560*, 2021.

[30] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li,

and Zheng Zhang. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019.

[31] Shike Wang, Fan Xu, Yunyang Li, Jie Wang, Ke Zhang, Yong Liu, Min Wu, and Jie Zheng. Kg4sl: knowledge graph neural network for synthetic lethality prediction in human cancers. *Bioinformatics*, 37(Supplement_1):i418–i425, 2021.

[32] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*, 2023.

[33] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, 2020.

[34] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR, 2019.

[35] Rui Xing, Jie Luo, and Tengwei Song. Biorel: towards large-scale biomedical relation extraction. *BMC bioinformatics*, 21(16):1–13, 2020.

[36] Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with reconstruction. In *The 35th AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

[37] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy, July 2019. Association for Computational Linguistics.

[38] Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. Contrastive triple extraction with generative transformer. *arXiv preprint arXiv:2009.06207*, 2020.

[39] Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190, Online, June 2021. Association for Computational Linguistics.

[40] Daojian Zeng, Haoran Zhang, and Qianying Liu. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9507–9514, 2020.

[41] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online, November 2020. Association for Computational Linguistics.

[42] Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[43] Sheng Zhang, Patrick Ng, Zhiguo Wang, and Bing Xiang. Reknow: Enhanced knowledge for joint entity and relation extraction. *arXiv preprint arXiv:2206.05123*, 2022.

[44] Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online, July 2020. Association for Computational Linguistics.

[45] Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6261–6270, Online, August 2021. Association for Computational Linguistics.

[46] Tianyang Zhao, Zhao Yan, Yunbo Cao, and Zhoujun Li. Asking effective and diverse questions: a machine reading comprehension based framework for joint entity-relation extraction. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3948–3954, 2021.