Reasoning with Latent Structure Refinement for Document-Level Relation Extraction

Guoshun Nan[1*], **Zhijiang Guo**[1*], **Ivan Sekulić**[2†] and **Wei Lu**[1]
[1]StatNLP Research Group, Singapore University of Technology and Design
[2]Università della Svizzera italiana
guoshun_nan@sutd.edu.sg, zhijiang_guo@mymail.sutd.edu.sg
ivan.sekulic@usi.ch, luwei@sutd.edu.sg

Abstract

Document-level relation extraction requires integrating information within and across multiple sentences of a document and capturing complex interactions between inter-sentence entities. However, effective aggregation of relevant information in the document remains a challenging research question. Existing approaches construct static document-level graphs based on syntactic trees, co-references or heuristics from the unstructured text to model the dependencies. Unlike previous methods that may not be able to capture rich non-local interactions for inference, we propose a novel model that empowers the relational reasoning across sentences by automatically inducing the latent document-level graph. We further develop a refinement strategy, which enables the model to incrementally aggregate relevant information for multi-hop reasoning. Specifically, our model achieves an $F1$ score of 59.05 on a large-scale document-level dataset (DocRED), significantly improving over the previous results, and also yields new state-of-the-art results on the CDR and GDA dataset. Furthermore, extensive analyses show that the model is able to discover more accurate inter-sentence relations.

1 Introduction

Relation extraction aims to detect relations among entities in the text and plays a significant role in a variety of natural language processing applications. Early research efforts focus on predicting relations between entities within the sentence (Zeng et al., 2014; Xu et al., 2015a,b). However, valuable relational information between entities, such as biomedical findings, is expressed by multiple mentions across sentence boundaries in real-world scenarios (Peng et al., 2017). Therefore, the scope of extraction in biomedical domain has recently been expanded to cross-sentence level (Quirk and Poon, 2017; Gupta et al., 2018; Song et al., 2019).

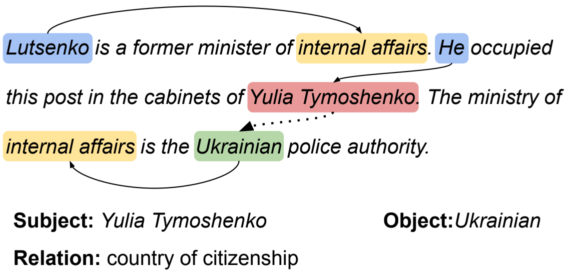
Lutsenko is a former minister of internal affairs. He occupied this post in the cabinets of Yulia Tymoshenko. The ministry of internal affairs is the Ukrainian police authority.
Subject: Yulia Tymoshenko
Object: Ukrainian
Relation: country of citizenship

Figure 1: An example adapted from the DocRED dataset. The example has four entities: *Lutsenko*, *internal affairs*, *Yulia Tymoshenko* and *Ukrainian*. Here entity *Lutsenko* has two mentions: *Lutsenko* and *He*. Mentions corresponding to the same entity are highlighted with the same color. The solid and dotted lines represent intra- and inter-sentence relations, respectively.

A more challenging, yet practical extension, is the document-level relation extraction, where a system needs to comprehend multiple sentences to infer the relations among entities by synthesizing relevant information from the entire document (Jia et al., 2019; Yao et al., 2019). Figure 1 shows an example adapted from the recently proposed document-level dataset DocRED (Yao et al., 2019). In order to infer the inter-sentence relation (i.e., country of citizenship) between *Yulia Tymoshenko* and *Ukrainian*, one first has to identify the fact that *Lutsenko* works with *Yulia Tymoshenko*. Next we identify that *Lutsenko* manages *internal affairs*, which is a *Ukrainian* authority. After incrementally connecting the evidence in the document and performing the step-by-step reasoning, we are able to infer that *Yulia Tymoshenko* is also a *Ukrainian*.

Prior efforts show that interactions between mentions of entities facilitate the reasoning process in the document-level relation extraction. Thus, Verga et al. (2018) and Jia et al. (2019) leverage Multi-Instance Learning (Riedel et al., 2010; Surdeanu

[*] Equally Contributed.
[†] Work done during internship at SUTD.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1546–1557
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

et al., 2012). On the other hand, structural information has been used to perform better reasoning since it models the non-local dependencies that are obscure from the surface form alone. Peng et al. (2017) construct dependency graph to capture interactions among n-ary entities for cross-sentence extraction. Sahu et al. (2019) extend this approach by using co-reference links to connect dependency trees of sentences to construct the document-level graph. Instead, Christopoulou et al. (2019) construct a heterogeneous graph based on a set of heuristics, and then apply an edge-oriented model (Christopoulou et al., 2018) to perform inference.

Unlike previous methods, where a document-level structure is constructed by co-references and rules, our proposed model treats the graph structure as a latent variable and induces it in an end-to-end fashion. Our model is built based on the structured attention (Kim et al., 2017; Liu and Lapata, 2018). Using a variant of Matrix-Tree Theorem (Tutte, 1984; Koo et al., 2007), our model is able to generate task-specific dependency structures for capturing non-local interactions between entities. We further develop an iterative refinement strategy, which enables our model to dynamically build the latent structure based on the last iteration, allowing the model to incrementally capture the complex interactions for better multi-hop reasoning (Welbl et al., 2018).

Experiments show that our model significantly outperforms the existing approaches on DocRED, a large-scale document-level relation extraction dataset with a large number of entities and relations, and also yields new state-of-the-art results on two popular document-level relation extraction datasets in the biomedical domain. The code and pretrained model are available at `https://github.com/nanguoshun/LSR` [1].

Our contributions are summarized as follows:

- We construct a document-level graph for inference in an end-to-end fashion without relying on co-references or rules, which may not always yield optimal structures. With the iterative refinement strategy, our model is able to dynamically construct a latent structure for improved information aggregation in the entire document.
- We perform quantitative and qualitative analyses to compare with the state-of-the-art models in various settings. We demonstrate that our model is capable of discovering more accurate inter-sentence relations by utilizing a multi-hop reasoning module.

2 Model

In this section, we present our proposed Latent Structure Refinement (LSR) model for the document-level relation extraction task. Our LSR model consists of three components: node constructor, dynamic reasoner, and classifier. The node constructor first encodes each sentence of an input document and outputs contextual representations. Representations that correspond to mentions and tokens on the shortest dependency path in a sentence are extracted as nodes. The dynamic reasoner is then applied to induce a document-level structure based on the extracted nodes. Representations of nodes are updated based on information propagation on the latent structure, which is iteratively refined. Final representations of nodes are used to calculate classification scores by the classifier.

2.1 Node Constructor

Node constructor encodes sentences in a document into contextual representations and constructs representations of mention nodes, entity nodes and meta dependency paths (MDP) nodes, as shown in Figure 2. Here MDP indicates a set of shortest dependency paths for all mentions in a sentence, and tokens in the MDP are extracted as MDP nodes.

2.1.1 Context Encoding

Given a document $\boldsymbol{d}$, each sentence d_i in it is fed to the context encoder, which outputs the contextualized representations of each word in d_i. The context encoder can be a bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997) or BERT (Devlin et al., 2019). Here we use the BiLSTM as an example:

$$\overleftarrow{\mathbf{h}_j^i} = \mathbf{LSTM}_l(\overleftarrow{\mathbf{h}^i}_{j+1}, \gamma_j^i) \tag{1}$$

$$\overrightarrow{\mathbf{h}_j^i} = \mathbf{LSTM}_r(\overrightarrow{\mathbf{h}^i}_{j-1}, \gamma_j^i) \tag{2}$$

where $\overleftarrow{\mathbf{h}_j^i}$, $\overleftarrow{\mathbf{h}^i}_{j+1}$, $\overrightarrow{\mathbf{h}_j^i}$ and $\overrightarrow{\mathbf{h}^i}_{j-1}$ represent the hidden representations of the j-th, (j+1)-th and (j-1)-th token in the sentence d_i of two directions, and γ_j^i denotes the word embedding of the j-th token. Contextual representation of each token in the sentence is represented as $\mathbf{h}_j^i = [\overleftarrow{\mathbf{h}_j^i}; \overrightarrow{\mathbf{h}_j^i}]$ by concatenating hidden states of two directions, where $\mathbf{h}_j^i \in \mathbb{R}^d$ and d is the dimension.

[1] Our model is implemented in PyTorch (Paszke et al., 2017)

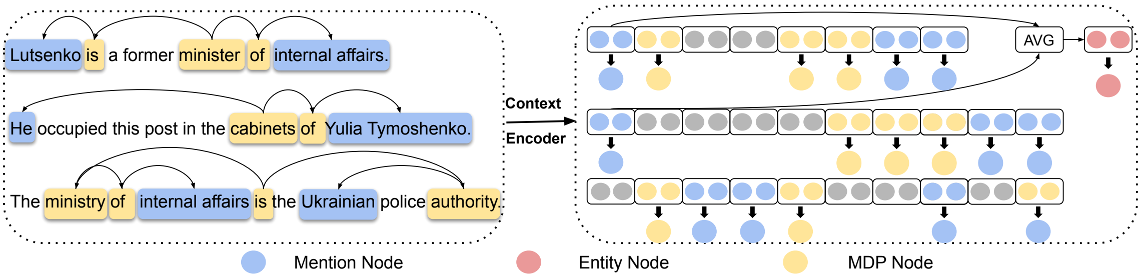
Lutsenko is a former minister of internal affairs.
He occupied this post in the cabinets of Yulia Tymoshenko.
The ministry of internal affairs is the Ukrainian police authority.
Context Encoder
AVG
Mention Node
Entity Node
MDP Node

Figure 2: Overview of the Node Constructor: A context encoder is applied to get the contextualized representations of sentences. The representations of mentions and words in the meta dependency paths are extracted as mention nodes and MDP nodes. An average pooling is used to construct the entity node from the mention nodes. For example, the entity node *Lutsenko* is constructed by averaging representations of its mentions *Lutsenko* and *He*. All figures best viewed in color.

2.1.2 Node Extraction

We construct three types of nodes for a document-level graph: mention nodes, entity nodes and meta dependency paths (MDP) nodes as shown in Figure 2. Mention nodes correspond to different mentions of entities in each sentence. The representation of an entity node is computed as the average of its mentions. To build a document-level graph, existing approaches use all nodes in the dependency tree of a sentence (Sahu et al., 2019) or one sentence-level node by averaging all token representations of the sentence (Christopoulou et al., 2019). Alternatively, we use tokens on the shortest dependency path between mentions in the sentence. The shortest dependency path has been widely used in the sentence-level relation extraction as it is able to effectively make use of relevant information while ignoring irrelevant information (Bunescu and Mooney, 2005; Xu et al., 2015a,b). Unlike sentence-level extraction, where each sentence only has two entities, each sentence here may involve multiple mentions.

2.2 Dynamic Reasoner

The dynamic reasoner has two modules, structure induction and multi-hop reasoning as shown in Figure 3. The structure induction module is used to learn a latent structure of a document-level graph. The multi-hop reasoning module is used to perform inference on the induced latent structure, where representations of each node will be updated based on the information aggregation scheme. We stack N blocks in order to iteratively refine the latent document-level graph for better reasoning.

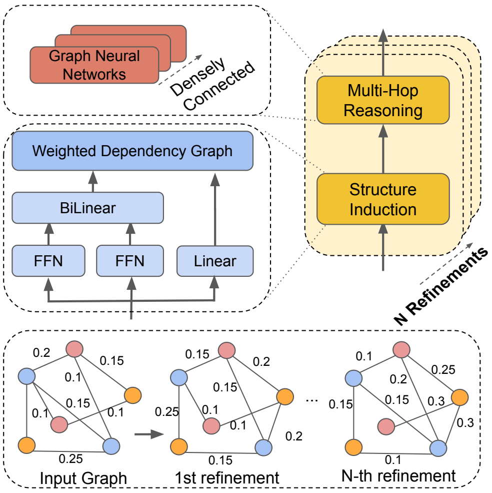

Figure 3: Overview of the Dynamic Reasoner. Each block consists of two sub-modules: structure induction and multi-hop reasoning. The first module takes the nodes constructed by the Node Constructor as inputs. Representations of nodes are fed into two feed-forward networks before the bilinear transformation. The latent document-level structure is computed by the Matrix-Tree Theorem. The second module takes the structure as input and updates representations of nodes by using the densely connected graph convolutional networks. We stack N blocks which correspond to N times of refinement. Each iteration outputs the latent structure for inference.

2.2.1 Structure Induction

Unlike existing models that use co-reference links (Sahu et al., 2019) or heuristics (Christopoulou et al., 2019) to construct a document-level graph for reasoning, our model treats the graph as a latent variable and induces it in an end-to-end fashion. The structure induction module is built based on the structured attention (Kim et al., 2017; Liu and Lapata, 2018). Inspired by Liu and Lapata (2018), we use a variant of Kirchhoff's Matrix-Tree Theorem (Tutte, 1984; Koo et al., 2007) to induce the latent dependency structure.

Let $\mathbf{u}_i$ denote the contextual representation of the i-th node, where $\mathbf{u}_i \in \mathbb{R}^d$, we first calculate the pair-wise unnormalized attention score $\mathbf{s}_{ij}$ between the i-th and the j-th node with the node representations $\mathbf{u}_i$ and $\mathbf{u}_j$. The score $\mathbf{s}_{ij}$ is calculated by two feed-forward neural networks and a bilinear transformation:

$$\mathbf{s}_{ij} = (\tanh(\mathbf{W}_p \mathbf{u}_i))^T \mathbf{W}_b (\tanh(\mathbf{W}_c \mathbf{u}_j)) \quad (3)$$

where $\mathbf{W}_p \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_c \in \mathbb{R}^{d \times d}$ are weights for two feed-forward neural networks, d is the dimension of the node representations, and $\tanh$ is applied as the activation function. $\mathbf{W}_b \in \mathbb{R}^{d \times d}$ are the weights for the bilinear transformation. Next we compute the root score $\mathbf{s}_i^r$ which represents the unnormalized probability of the i-th node to be selected as the root node of the structure:

$$\mathbf{s}_i^r = \mathbf{W}_r \mathbf{u}_i \quad (4)$$

where $\mathbf{W}_r \in \mathbb{R}^{1 \times d}$ is the weight for the linear transformation. Following Koo et al. (2007), we calculate the marginal probability of each dependency edge of the document-level graph. For a graph $\mathbf{G}$ with n nodes, we first assign non-negative weights $\mathbf{P} \in \mathbb{R}^{n \times n}$ to the edges of the graph:

$$\mathbf{P}_{ij} = \begin{cases} 0 & \text{if } i = j \\ \exp{(\mathbf{s}_{ij})} & \text{otherwise} \end{cases} \quad (5)$$

where $\mathbf{P}_{ij}$ is the weight of the edge between the i-th and the j-th node. We then define the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{n \times n}$ of $\mathbf{G}$ in Equation (6), and its variant $\hat{\mathbf{L}} \in \mathbb{R}^{n \times n}$ in Equation (7) for further computations (Koo et al., 2007).

$$\mathbf{L}_{ij} = \begin{cases} \sum_{i'=1}^{n} \mathbf{P}_{i'j} & \text{if } i = j \\ -\mathbf{P}_{ij} & \text{otherwise} \end{cases} \quad (6)$$

$$\hat{\mathbf{L}}_{ij} = \begin{cases} \exp(\mathbf{s}_i^r) & \text{if } i = 1 \\ \mathbf{L}_{ij} & \text{if } i > 1 \end{cases} \quad (7)$$

We use $\mathbf{A}_{ij}$ to denote the marginal probability of the dependency edge between the i-th and the j-th node. Then, $\mathbf{A}_{ij}$ can be derived based on Equation (8), where δ is the Kronecker delta (Koo et al., 2007).

$$\begin{aligned} \mathbf{A}_{ij} = & (1 - \delta_{1,j}) \mathbf{P}_{ij} [\hat{\mathbf{L}}^{-1}]_{ij} \\ & - (1 - \delta_{i,1}) \mathbf{P}_{ij} [\hat{\mathbf{L}}^{-1}]_{ji} \end{aligned} \quad (8)$$

Here, $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be interpreted as a weighted adjacency matrix of the document-level entity graph. Finally, we can feed $\mathbf{A} \in \mathbb{R}^{n \times n}$ into the multi-hop reasoning module to update the representations of nodes in the latent structure.

2.2.2 Multi-hop Reasoning

Graph neural networks have been widely used in different tasks to perform multi-hop reasoning (Song et al., 2018a; Yang et al., 2019; Tu et al., 2019; Lin et al., 2019), as they are able to effectively collect relevant evidence based on an information aggregation scheme. Specifically, our model is based on graph convolutional networks (GCNs) (Kipf and Welling, 2017) to perform reasoning.

Formally, given a graph $\mathbf{G}$ with n nodes, which can be represented with an $n \times n$ adjacency matrix $\mathbf{A}$ induced by the previous structure induction module, the convolution computation for the node i at the l-th layer, which takes the representation $\mathbf{u}_i^{l-1}$ from previous layer as input and outputs the updated representations $\mathbf{u}_i^l$, can be defined as:

$$\mathbf{u}_i^l = \sigma(\sum_{j=1}^{n} \mathbf{A}_{ij} \mathbf{W}^l \mathbf{u}_i^{l-1} + \mathbf{b}^l) \quad (9)$$

where $\mathbf{W}^l$ and $\mathbf{b}^l$ are the weight matrix and bias vector for the l-th layer, respectively. σ is the ReLU (Nair and Hinton, 2010) activation function. $\mathbf{u}_i^0 \in \mathbb{R}^d$ is the initial contextual representation of the i-th node constructed by the node constructor.

Following Guo et al. (2019b), we use dense connections to the GCNs in order to capture more structural information on a large document-level graph. With the help of dense connections, we are able to train a deeper model, allowing richer local and non-local information to be captured for learning a better graph representation. The computations on each graph convolution layer is similar to Equation (9).

2.2.3 Iterative Refinement

Though structured attention (Kim et al., 2017; Liu and Lapata, 2018) is able to automatically induce a latent structure, recent research efforts show that the induced structure is relatively shallow and may not be able to model the complex dependencies for document-level input (Liu et al., 2019b; Ferracane et al., 2019). Unlike previous work (Liu and Lapata, 2018) that only induces the latent structure once, we repeatedly refine the document-level graph based on the updated representations, allowing the model to infer a more informative structure that goes beyond simple parent-child relations.

As shown in Figure 3, we stack N blocks of the dynamic reasoner in order to induce the document-level structure N times. Intuitively, the reasoner induces a shallow structure at early iterations since the information propagates mostly between neighboring nodes. As the structure gets more refined by interactions with richer non-local information, the induction module is able to generate a more informative structure.

2.3 Classifier

After N times of refinement, we obtain representations of all the nodes. Following Yao et al. (2019), for each entity pair $(\mathbf{e}_i, \mathbf{e}_j)$, we use a bilinear function to compute the probability for each relation type r as:

$$P(r|\mathbf{e}_i, \mathbf{e}_j) = \sigma(\mathbf{e}_i^T \mathbf{W}_\mathbf{e} \mathbf{e}_j + \mathbf{b}_e)_r \qquad (10)$$

where $\mathbf{W}_e \in \mathbb{R}^{d \times k \times d}$ and $\mathbf{b}_e \in \mathbb{R}^k$ are trainable weights and bias, with k being the number of relation categories, σ is the *sigmoid* function, and the subscript r in the right side of the equation refers to the relation type.

3 Experiments

3.1 Data

We evaluate our model on DocRED (Yao et al., 2019), the largest human-annotated dataset for document-level relation extraction, and another two popular document-level relation extraction datasets in the biomedical domain, including Chemical-Disease Reactions (CDR) (Li et al., 2016a) and Gene-Disease Associations (GDA) (Wu et al., 2019). DocRED contains $3,053$ documents for training, $1,000$ for development and $1,000$ for test, totally with $132,375$ entities and $56,354$ relational facts. CDR consists of 500 training instances, 500 development instances, and 500 testing instances. GDA contains $29,192$ documents for training and $1,000$ for test. We follow (Christopoulou et al., 2019) to split training set of GDA into an 80/20 split for training and development.

With more than 40% of the relational facts requiring reading and reasoning over multiple sentences, DocRED significantly differs from previous sentence-level datasets (Doddington et al., 2004; Hendrickx et al., 2009; Zhang et al., 2018). Unlike existing document-level datasets (Li et al., 2016a; Quirk and Poon, 2017; Peng et al., 2017; Verga et al., 2018; Jia et al., 2019) that are in the specific biomedical domain considering only the drug-gene-disease relation, DocRED covers a broad range of categories with 96 relation types.

Batch size	20
Learning rate	0.001
Optimizer	Adam
Hidden size	120
Induction block number	2
GCN dropout	0.3

Table 1: Hyper-parameters of LSR.

3.2 Setup

We use spaCy[2] to get the meta dependency paths of sentences in a document. Following Yao et al. (2019) and Wang et al. (2019), we use the GloVe (Pennington et al., 2014) embedding with BiLSTM, and Uncased BERT-Base (Devlin et al., 2019) as the context encoder. All hyper-parameters are tuned based on the development set. We list some of the important hyper-parameters in Table 1.

Following Yao et al. (2019), we use F_1 and Ign F_1 as the evaluation metrics. Ign F_1 denotes F_1 scores excluding relational facts shared by the training and dev/test sets. F_1 scores for intra- and inter-sentence entity pairs are also reported. Evaluation on the test set is done through CodaLab[3].

3.3 Main Results

We compare our proposed LSR with the following three types of competitive models on the DocRED dataset, and show the main results in Table 2.

- **Sequence-based Models.** These models leverage different neural architectures to encode sentences in the document, including convolutional neural networks (CNN) (Zeng et al., 2014), LSTM, bidirectional LSTM (BiLSTM) (Cai et al., 2016) and attention-based LSTM (ContextAware) (Sorokin and Gurevych, 2017).
- **Graph-based Models.** These models construct task-specific graphs for inference. GCNN (Sahu et al., 2019) constructs a document-level graph by co-reference links, and then applies relational GCNs for reasoning. EoG (Christopoulou et al., 2019) is the state-of-the-art document-level relation extraction model in biomedical domain. EoG first uses heuristics to construct the graph, then leverages an edge-oriented model to perform inference. GCNN and EoG are based on static structures. GAT (Veličković et al., 2018) is able to learn the weighted graph structure based on a local attention mechanism. AGGCN (Guo

[2] https://spacy.io/

[3] https://competitions.codalab.org/competitions/20717

	Dev				Test	
Model	Ign $F1$	$F1$	Intra-$F1$	Inter-$F1$	Ign $F1$	$F1$
CNN (Yao et al., 2019)	41.58	43.45	51.87*	37.58*	40.33	42.26
LSTM (Yao et al., 2019)	48.44	50.68	56.57*	41.47*	47.71	50.07
BiLSTM (Yao et al., 2019)	48.87	50.94	57.05*	43.49*	48.78	51.06
ContexAware (Yao et al., 2019)	**48.94**	51.09	56.74*	42.26*	48.40	50.70
GCNN ♣ (Sahu et al., 2019)	46.22	51.52	57.78	44.11	49.59	51.62
EoG ♣ (Christopoulou et al., 2019)	45.94	52.15	58.90	44.60	49.48	51.82
GAT ♣ (Veličković et al., 2018)	45.17	51.44	58.14	43.94	47.36	49.51
AGGCN ♣ (Guo et al., 2019a)	46.29	52.47	58.76	45.45	48.89	51.45
GloVe+LSR	48.82	**55.17**	**60.83**	**48.35**	**52.15**	**54.18**
BERT (Wang et al., 2019)	-	54.16	61.61*	47.15*	-	53.20
Two-Phase BERT (Wang et al., 2019)	-	54.42	61.80*	47.28*	-	53.92
BERT+LSR	**52.43**	**59.00**	**65.26**	**52.05**	**56.97**	**59.05**

Table 2: Main results on the development and the test set of DocRED: Models with ♣ are adapted to DocRED based on their open implementations. Results with $*$ are computed based on re-trained models as we need to evaluate F_1 for both intra- and inter-sentence setting, which are not given in original papers.

et al., 2019a) is the state-of-the-art sentence-level relation extraction model, which constructs the latent structure by self-attention. These two models are able to dynamically construct task-specific structures.

- **BERT-based Models.** These models fine-tune BERT (Devlin et al., 2019) for DocRED. Specifically, Two-Phase BERT (Wang et al., 2019) is the best reported model. It is a pipeline model, which predicts if the relation exists between entity pairs in the first phase and predicts the type of the relation in the second phase.

As shown in Table 2, LSR with GloVe achieves 54.18 F_1 on the test set, which is the new state-of-the-art result for models with GloVe. In particular, our model consistently outperforms sequence-based models by a significant margin. For example, LSR improves upon the best sequence-based model BiLSTM by 3.1 points in terms of F_1. This suggests that models which directly encode the entire document are unable to capture the inter-sentence relations present in documents.

Under the same setting, our model consistently outperforms graph-based models based on static graphs or attention mechanisms. Compared with EoG, our LSR model achieves 3.0 and 2.4 higher F_1 on development and test set, respectively. We also have similar observations for the GCNN model, which shows that a static document-level graph may not be able to capture the complex interactions in a document. The dynamic latent structure induced by LSR captures richer non-local dependencies. Moreover, LSR also outperforms GAT and AGGCN. This empirically shows that compared to the models that use local attention and self-attention (Veličković et al., 2018; Guo et al., 2019a), LSR can induce more informative document-level structures for better reasoning. Our LSR model also shows its superiority under the setting of Ign F_1.

In addition, LSR with GloVe obtains better results than two BERT-based models. This empirically shows that our model is able to capture long-range dependencies even without using powerful context encoders. Following Wang et al. (2019), we leverage BERT as the context encoder. As shown in Table 2, our LSR model with BERT achieves a 59.05 F_1 score on DocRED, which is a new state-of-the-art result. As of the ACL deadline on the 9th of December 2019, we held the first position on the CodaLab scoreboard under the alias *diskorak*.

3.4 Intra- and inter-sentence performance

In this subsection, we analyze intra- and inter-sentence performance on the development set. An entity pair requires inter-sentence reasoning if the two entities from the same document have no mentions in the same sentence. In DocRED's development set, about 45% of entity pairs require information aggregation over multiple sentences.

Under the same setting, our LSR model outperforms all other models in both intra- and inter-sentence setting. The differences in F_1 scores between LSR and other models in the inter-sentence setting tend to be larger than the differences in the intra-sentence setting. These results demonstrate that the majority of LSR's superiority comes from the inter-sentence relational facts, suggesting that

Model	$F1$	Intra-$F1$	Inter-$F1$
Gu et al. (2017)	61.3	57.2	11.7
Nguyen and Verspoor (2018)	62.3	-	-
Verga et al. (2018)	62.1	-	-
Sahu et al. (2019)	58.6	-	-
Christopoulou et al. (2019)	63.6	68.2	50.9
LSR	61.2	66.2	50.3
LSR w/o MDP Nodes	**64.8**	**68.9**	**53.1**
Peng et al. (2016)	63.1	-	-
Li et al. (2016b)	67.3	58.9	-
Panyam et al. (2018)	60.3	65.1	45.7
Zheng et al. (2018)	61.5	-	-

Table 3: Results on the test set of the CDR dataset. The methods below the double line take advantage of additional training data and/or incorporate external tools.

the latent structure induced by our model is indeed capable of synthesizing the information across multiple sentences of a document.

Furthermore, LSR with GloVe also proves better in the inter-sentence setting compared with two BERT-based (Wang et al., 2019) models, indicating latent structure's superiority in resolving long-range dependencies across the whole document compared with the BERT encoder.

3.5 Results on the Biomedical Datasets

Table 3 depicts the comparisons with state-of-the-art models on the CDR dataset. Gu et al. (2017); Nguyen and Verspoor (2018); Verga et al. (2018) leverage sequence-based models. Convolutional neural networks and self-attention networks are used as the encoders. Sahu et al. (2019); Christopoulou et al. (2019) use graph-based models. As shown in Table 3, our LSR performs worse than the state-of-the-art models. It is challenging for an off-the-shelf parser to get high quality dependency trees in the biomedical domain, as we observe that the MDP nodes extracted by the spaCy parser from the CDR dataset contains much less informative context compared with the nodes from DocRED. Here we introduce a simplified LSR model indicated as "LSR w/o MDP Nodes" , which removes the MDP nodes and builds a fully-connected graph using all tokens of a document. It shows that "LSR w/o MDP Nodes" consistently outperforms sequence-based and graph-based models, indicating the effectiveness the latent structure. Moreover, the simplified LSR outperforms most of the models with external resources, except for Li et al. (2016b), which leverages co-training with additional unlabeled training data. We believe such a setting also benefits our LSR model.

Model	$F1$	Intra-$F1$	Inter-$F1$
NoInf (Christopoulou et al., 2019)	74.6	79.1	49.3
Full (Christopoulou et al., 2019)	80.8	84.1	**54.7**
EoG (Christopoulou et al., 2019)	81.5	85.2	50.0
LSR	79.6	83.1	49.6
LSR w/o MDP Nodes	**82.2**	**85.4**	51.1

Table 4: Results on the test set of the GDA dataset.

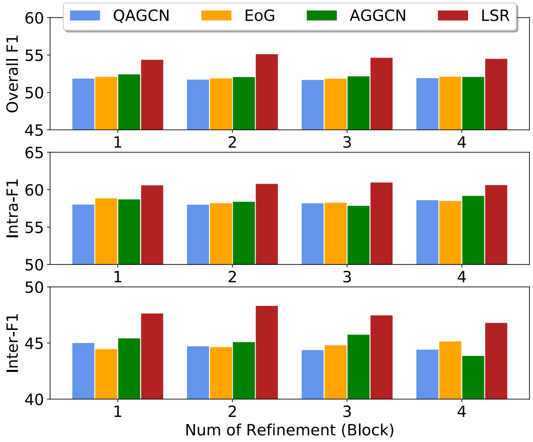

Figure 4: Intra- and inter-sentence $F1$ for different graph structures in QAGCN, EoG, AGGCN and LSR. The number of refinements is ranging from 1 to 4.

Table 4 shows the results on the distantly supervised GDA dataset. Here "Full" indicates EoG model with a fully connected graph as the inputs, while "NoInf" is a variant of EoG model without inference component (Christopoulou et al., 2018). The simplified LSR model achieves the new state-of-the-art result on GDA. The "Full" model (Christopoulou et al., 2019) yields a higher $F1$ score on the inter-sentence setting while having a relatively low score on the intra-sentence. It is likely because that this model neglects the differences between relations expressed within the sentence and across sentences.

3.6 Model Analysis

In this subsection, we use the development set of DocRED to demonstrate the effectiveness of the latent structure and refinements.

3.6.1 Does Latent Structure Matter?

We investigate the extent to which the latent structures, that are induced and iteratively refined by the proposed dynamic reasoner, help to improve the overall performance. We experiment with the three different structures defined below. For fair comparisons, we use the same GCN model to perform multi-hop reasoning for all these structures.

Rule-based Structure: We use the rule-based structure in EoG (Christopoulou et al., 2019). Also,

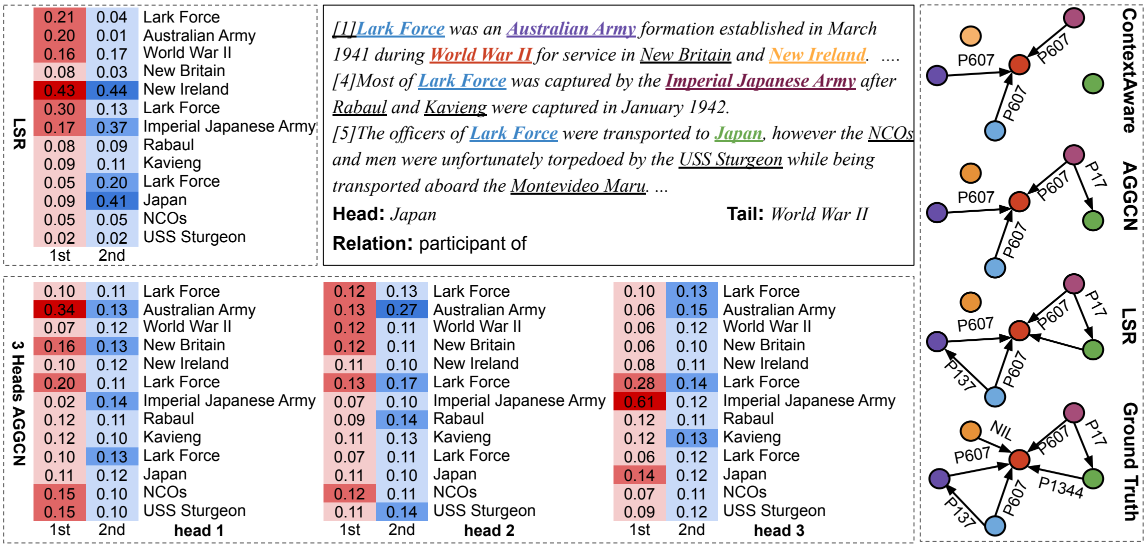

Figure 5: Case study of an example from the development set of DocRED. We visualize the reasoning process for predicting the relation of an entity pair ⟨*Japan*, *World War II*⟩ by LSR and AGGCN in two refinement steps, using the attention scores of the mention *World War II* in each step. We scale all attention scores by 1000 to illustrate them more clearly. Some sentences are omitted due to space limitation.

We adapt rules from De Cao et al. (2019) for multi-hop question answering, i.e., each mention node is connected to its entity node and to the same mention nodes across sentences, while mention nodes and MDP nodes which reside in the same sentence are fully connected. The model is termed QAGCN.

Attention-based Structure: This structure is induced by AGGCN (Guo et al., 2019a) with multi-head attention (Vaswani et al., 2017). We extend the model from sentence-level to document-level.

We explore multiple settings of these models with different block numbers ranging from 1 to 4, where a block is composed of a graph construction component and a densely connected GCN component. As shown in Figure 4, LSR outperforms QAGCN, EoG and AGGCN in terms of overall F_1. This empirically confirms our hypothesis that the latent structure induced by LSR is able to capture a more informative context for the entire document.

3.6.2 Does Refinement Matter?

As shown in Figure 4, our LSR yields the best performance in the second refinement, outperforming the first induction by 0.72% in terms of overall F_1. This indicates that the proposed LSR is able to induce more accurate structures by iterative refinement. However, too many iterations may lead to an F_1 drop due to over-fitting.

3.7 Ablation Study

Table 5 shows F_1 scores of the full LSR model and with different components turned off one at a time. We observe that most of the components contribute to the main model, as the performance deteriorates with any of the components missing. The most significant difference is visible in the structure induction module. Removal of structure induction part leads to a 3.26 drop in terms of F_1 score. This result indicates that the latent structure plays a key role in the overall performance.

Model	$F1$	Intra-$F1$	Inter-$F1$
Full model	55.17	60.83	48.35
- 1 Refinement	54.42	60.46	47.67
- 2 Structure Induction	51.91	58.08	45.04
- 1 Multi-hop Reasoning	54.49	59.75	47.49
- 2 Multi-hop Reasoning	54.24	60.58	47.15
- MDP nodes	54.20	60.54	47.12

Table 5: Ablation study of LSR on DocRED.

3.8 Case Study

In Figure 5, we present a case study to analyze why the latent structure induced by our proposed LSR performs better than the structures learned by AGGCN. We use the entity *World War II* to illustrate the reasoning process and our goal here is to predict the relation of the entity pair ⟨*Japan*, *World War II*⟩. As shown in Figure 5, in the first refinement of LSR, *Word War II* interacts with several local mentions with higher attention scores, e.g., *0.43* for the mention *Lake Force*, which will be used as a bridge between the mention *Japan* and *World War II*. In the second refinement, the attention scores of several non-local mentions, such as *Japan* and *Imperial Japanese Army*, significantly increase from

0.09 to *0.41*, and *0.17* to *0.37*, respectively, indicating that information is propagated globally at this step. With such intra- and inter-sentence structures, the relation of the entity pair ⟨*Japan*, *World War II*⟩ can be predicted as *"participant of"*, which is denoted by *P1344*. Compared with LSR, the attention scores learned by AGGCN are much more balanced, indicating that the model may not be able to construct an informative structure for inference, e.g., the highest score is *0.27* in the second head, and most of the scores are near *0.11*.

We also depict the predicted relations of ContextAware, AGGCN and LSR on the graph on the right side of the Figure 5. Interested reader could refer to (Yao et al., 2019) for the definition of a relation, such as *P607*, *P17*, *etc.* The LSR model proves capable of filling out the missing relation for ⟨*Japan*, *World War II*⟩ that requires reasoning across sentences. However, LSR also attends to the mention *New Ireland* with a high score, thus failing to predict that the entity pair ⟨*New Ireland*, *World War II*⟩ actually has no relation (*NIL* type).

4 Related Work

Document-level relation extraction. Early efforts focus on predicting relations between entities within a single sentence by modeling interactions in the input sequence (Zeng et al., 2014; Wang et al., 2016; Zhou et al., 2016; Zhang et al., 2017; Guo et al., 2020) or the corresponding dependency tree (Xu et al., 2015a,b; Liu et al., 2015; Miwa and Bansal, 2016; Zhang et al., 2018). These approaches do not consider interactions across mentions and ignore relations expressed across sentence boundaries. Recent work begins to explore cross-sentence extraction (Quirk and Poon, 2017; Peng et al., 2017; Gupta et al., 2018; Song et al., 2018c, 2019). Instead of using discourse structure understanding techniques (Liu et al., 2019a; Lei et al., 2017, 2018), these approaches leverage the dependency graph to capture inter-sentence interactions, and their scope is still limited to several sentences. More recently, the extraction scope has been expanded to the entire document (Verga et al., 2018; Jia et al., 2019; Sahu et al., 2019; Christopoulou et al., 2019) in the biomedical domain by only considering a few relations among chemicals. Unlike previous work, we focus on document-level relation extraction datasets (Yao et al., 2019; Li et al., 2016a; Wu et al., 2019) from different domains with a large number of relations and entities, which require understanding a document and performing multi-hop reasoning.

Structure-based relational reasoning. Structural information has been widely used for relational reasoning in various NLP applications including question answering (Dhingra et al., 2018; De Cao et al., 2019; Song et al., 2018a) and relation extraction (Sahu et al., 2019; Christopoulou et al., 2019). Song et al. (2018a) and (De Cao et al., 2019) leverage co-reference information and set of rules to construct document-level entity graph. GCNs (Kipf and Welling, 2017) or GRNs (Song et al., 2018b) are applied to perform reasoning for multi-hop question answering (Welbl et al., 2018). Sahu et al. (2019) also utilize co-reference links to construct the dependency graph and use labelled edge GCNs (Marcheggiani and Titov, 2017) for document-level relation extraction. Instead of using GNNs, Christopoulou et al. (2019) use the edge-oriented model (Christopoulou et al., 2018) for logical inference based on a heterogeneous graph constructed by heuristics. Unlike previous approaches that use syntactic trees, co-references or heuristics, LSR model treats the document-level structure as a latent variable and induces it in an iteratively refined fashion, allowing the model to dynamically construct the graph for better relational reasoning.

5 Conclusion

We introduce a novel latent structure refinement (LSR) model for better reasoning in the document-level relation extraction task. Unlike previous approaches that rely on syntactic trees, co-references or heuristics, LSR dynamically learns a document-level structure and makes predictions in an end-to-end fashion. There are multiple avenues for future work. One possible direction is to extend the scope of structure induction for constructions of nodes without relying on an external parser.

Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful and constructive comments. This research is supported by Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 2 Programme (MOE AcRF Tier 2 Award No: MOE2017-T2-1-156). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

References

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proc. of EMNLP*.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proc. of ACL*.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2018. A walk-based model on entity graphs for relation extraction. In *Proc. of ACL*.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proc. of EMNLP*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *Proc. of NAACL-HLT*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*.

Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2018. Neural models for reasoning over multiple mentions using coreference. In *Proc. of NAACL-HLT*.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ace) program - tasks, data, and evaluation. In *Proc. of LREC*.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2019. Evaluating discourse in structured text representations. In *Proc. of ACL*.

Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2017. Chemical-induced disease relation extraction via convolutional neural network. *Database: The Journal of Biological Databases and Curation*, 2017.

Zhijiang Guo, Guoshun Nan, Wei Lu, and Shay B. Cohen. 2020. Learning latent forests for medical relation extraction. In *Proc. of IJCAI*.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019a. Attention guided graph convolutional networks for relation extraction. In *Proc. of ACL*.

Zhijiang Guo, Yan Zhang, Zhiyang Teng, and Wei Lu. 2019b. Densely connected graph convolutional networks for graph-to-sequence learning. *Transactions of the Association for Computational Linguistics*, 7:297–312.

Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, Bernt Andrassy, and Thomas A. Runkler. 2018. Neural relation extraction within and across sentence boundaries. In *Proc. of AAAI*.

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

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *Proc. of NAACL-HLT*.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. 2017. Structured attention networks. In *Proc. of ICLR*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proc. of ICLR*.

Terry K Koo, Amir Globerson, Xavier Carreras, and Michael Collins. 2007. Structured prediction models via the matrix-tree theorem. In *Proc. of EMNLP-CoNLL*.

Wenqiang Lei, Xuancong Wang, Meichun Liu, Ilija Ilievski, Xiangnan He, and Min-Yen Kan. 2017. Swim: A simple word interaction model for implicit discourse relation recognition. In *Proc of. IJCAI*.

Wenqiang Lei, Yuanxin Xiang, Yuwei Wang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *Proc of. AAAI*.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016a. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database : the journal of biological databases and curation*, 2016.

Zhiheng Li, Zhihao Yang, Hongfei Lin, Jian Wang, Yingyi Gui, Yin Zhang, and Lei Wang. 2016b. Cidextractor: A chemical-induced disease relation extraction system for biomedical literature. In *Proc. of BIBM*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proc. of EMNLP*.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2019a. Discourse representation parsing for sentences and documents. In *Proc of. ACL*.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019b. Single document summarization as tree induction. In *Proc. of NAACL-HLT*.

Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. 2015. A dependency-based neural network for relation classification. In *Proc. of ACL*.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proc. of EMNLP*.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proc. of ACL*.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proc. of ICML*.

Dat Quoc Nguyen and Karin M. Verspoor. 2018. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. In *Proc. of BioNLP*.

Nagesh C Panyam, Karin Verspoor, Trevor Cohn, and Kotagiri Ramamohanarao. 2018. Exploiting graph kernels for high performance biomedical relation extraction. *Journal of Biomedical Semantics*, 9(1):7.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary Devito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

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

Yifan Peng, Chih-Hsuan Wei, and Zhiyong Lu. 2016. Improving chemical disease relation extraction with rich features and weakly labeled data. *Journal of Cheminformatics*, 8.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proc. of EACL*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proc. of ECML/PKDD*.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proc. of ACL*.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*, 45:2673–2681.

Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. 2018a. Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. *arXiv preprint arXiv:1809.02040*.

Linfeng Song, Yue Zhang, Daniel Gildea, Mo Yu, Zhiguo Wang, and Jinsong Su. 2019. Leveraging dependency forest for neural medical relation extraction. In *Proc. of EMNLP-IJCNLP*.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018b. A graph-to-sequence model for AMR-to-text generation. In *Proc. of ACL*.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018c. N-ary relation extraction using graph state lstm. In *Proc. of EMNLP*.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proc. of EMNLP*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proc. of EMNLP-CoNLL*.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proc. of ACL*.

William Thomas Tutte. 1984. *Graph theory*. Clarendon Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *Proc. of ICLR*.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proc. of NAACL-HLT*.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *arXiv preprint arXiv:1909.11898*.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proc. of ACL*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *Proc. of RECOMB*.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015a. Semantic relation classification via convolutional neural networks with simple negative sampling. In *Proc. of EMNLP*.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015b. Classifying relations via long short term memory networks along shortest dependency paths. In *Proc. of EMNLP*.

Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning cross-lingual entities with multi-aspect information. In *Proc. of EMNLP*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proc. of ACL*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jian Zhao. 2014. Relation classification via convolutional deep neural network. In *Proc. of COLING*.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proc. of EMNLP*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proc. of EMNLP*.

Wei Zheng, Hongfei Lin, Zhiheng Li, Xiaoxia Liu, Zhengguang Li, Bo Xu, Yijia Zhang, Zhihao Yang, and Jian Wang. 2018. An effective neural model extracting document level chemical-induced disease relations from biomedical literature. *Journal of biomedical informatics*, 83:1–9.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proc. of ACL*.