

Connecting the Dots: Document-level Neural Relation Extraction with Edge-oriented Graphs

Fenia Christopoulou¹, Makoto Miwa^{2,3}, Sophia Ananiadou¹

¹National Centre for Text Mining,

School of Computer Science, The University of Manchester, United Kingdom

²Toyota Technological Institute, Nagoya, 468-8511, Japan

³Artificial Intelligence Research Center (AIRC),

National Institute of Advanced Industrial Science and Technology (AIST), Japan

{efstathia.christopoulou, sophia.ananiadou}@manchester.ac.uk
makoto-miwa@toyota-ti.ac.jp

Abstract

Document-level relation extraction is a complex human process that requires logical inference to extract relationships between named entities in text. Existing approaches use graph-based neural models with words as nodes and edges as relations between them, to encode relations across sentences. These models are node-based, i.e., they form pair representations based solely on the two target node representations. However, entity relations can be better expressed through unique edge representations formed as paths between nodes. We thus propose an edge-oriented graph neural model for document-level relation extraction. The model utilises different types of nodes and edges to create a document-level graph. An inference mechanism on the graph edges enables to learn intra- and inter-sentence relations using multi-instance learning internally. Experiments on two document-level biomedical datasets for chemical-disease and gene-disease associations show the usefulness of the proposed edge-oriented approach.¹

1 Introduction

The extraction of relations between named entities in text, known as Relation Extraction (RE), is an important task of Natural Language Processing (NLP). Lately, RE has attracted a lot of attention from the field, in an effort to improve the inference capability of current methods (Zeng et al., 2017; Christopoulou et al., 2018; Luan et al., 2019).

In real-world scenarios, a large amount of relations are expressed across sentences. The task of identifying these relations is named inter-sentence RE. Typically, inter-sentence relations occur in

¹Source code available at <https://github.com/fenchri/edge-oriented-graph>

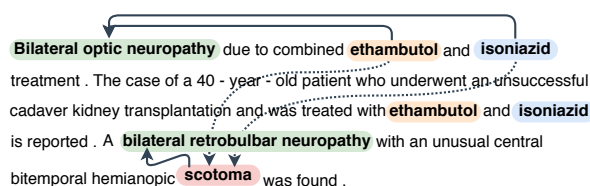


Figure 1: Example of document-level, inter-sentence relations adapted from the CDR dataset (Li et al., 2016a). The solid and dotted lines represent intra- and inter-sentence relations, respectively.

textual snippets with several sentences, such as documents. In these snippets, each entity is usually repeated with the same phrases or aliases, the occurrences of which are often named entity mentions and regarded as instances of the entity. The multiple mentions of the target entities in different sentences can be useful for the identification of inter-sentential relations, as these relations may depend on the interactions of their mentions with other entities in the same document.

As shown in the example of Figure 1, the entities *bilateral optic neuropathy*, *ethambutol* and *isoniazid* have two mentions each, while the entity *scotoma* has one mention. The relation between the chemical *ethambutol* and the disease *scotoma* is clearly inter-sentential. Their association can only be determined if we consider the interactions between the mentions of these entities in different sentences. A mention of *bilateral optic neuropathy* interacts with a mention of *ethambutol* in the first sentence. Another mention of the former interacts with the mention of *scotoma* in the third sentence. This chain of interactions can help us infer that the entity *ethambutol* has a relation with the entity *scotoma*.

The most common technique that is currently

used to deal with multiple mentions of named entities is Multi-Instance Learning (MIL). Initially, MIL was introduced by Riedel et al. (2010) in order to reduce noise in distantly supervised corpora (Mintz et al., 2009). In DS, training instances are created from large, raw corpora using Knowledge Base (KB) entity linking and automatic annotation with heuristic rules. MIL in this setting considers multiple sentences (bags) that contain a pair of entities serving as the multiple instances of this pair. Verga et al. (2018) introduced another MIL setting for relation extraction between named entities in a document. In this setting, entities mapped to the same KB ID are considered as mentions of an entity concept and pairs of mentions correspond to the pair’s multiple instances. However, document-level RE is not common in the general domain, as the entity types of interest can often be found in the same sentence (Banko et al., 2007). On the contrary, in the biomedical domain, document-level relations are particularly important given the numerous aliases that biomedical entities can have (Quirk and Poon, 2017).

To deal with document-level RE, recent approaches assume that only two mentions of the target entities reside in the document (Nguyen and Verspoor, 2018; Verga et al., 2018) or utilise different models for intra- and inter-sentence RE (Gu et al., 2016; Li et al., 2016b; Gu et al., 2017). In contrast with approaches that employ sequential models (Nguyen and Verspoor, 2018; Gu et al., 2017; Zhou et al., 2016), graph-based neural approaches have proven useful in encoding long-distance, inter-sentential information (Peng et al., 2017; Quirk and Poon, 2017; Gupta et al., 2019). These models interpret words as nodes and connections between them as edges. They typically perform on the nodes by updating the representations during training. However, a relation between two entities depends on different contexts. It could thus be better expressed with an edge connection that is unique for the pair. A straightforward way to address this is to create graph-based models that rely on edge representations rather focusing on node representations, which are shared between multiple entity pairs.

In this work, we tackle document-level, intra- and inter-sentence RE using MIL with a graph-based neural model. Our objective is to infer the relation between two entities by exploiting other interactions in the document. We construct a doc-

ument graph with heterogeneous types of nodes and edges to better capture different dependencies between nodes. In the proposed graph, a node corresponds to either entities, mentions, or sentences, instead of words. We connect distinct nodes based on simple heuristic rules and generate different edge representations for the connected nodes. To achieve our objective, we design the model to be edge-oriented in a sense that it learns edge representations (between the graph nodes) rather than node representations. An iterative algorithm over the graph edges is used to model dependencies between the nodes in the form of edge representations. The intra- and inter-sentence entity relations are predicted by employing these edges. Our contributions can be summarised as follows:

- We propose a novel edge-oriented graph neural model for document-level relation extraction. The model deviates from existing graph models as it focuses on constructing unique nodes and edges, encoding information into edge representations rather than node representations.
- The proposed model is independent of syntactic dependency tools and can achieve state-of-the-art performance on a manually annotated, document-level chemical-disease interaction dataset.
- Analysis of the model components indicates that the document-level graph can effectively encode document-level dependencies. Additionally, we show that inter-sentence associations can be beneficial for the detection of intra-sentence relations.

2 Proposed Model

We build our model as a significant extension of our previously proposed sentence-level model (Christopoulou et al., 2018) for document-level RE. The most critical difference between the two models is the introduction and construction of a partially-connected document graph, instead of a fully-connected sentence-level graph. Additionally, the document graph consists of heterogeneous types of nodes and edges in comparison with the sentence-level graph that contains only entity-nodes and single edge types among them. Furthermore, the proposed approach utilises multi-instance learning when mention-level annotations are available.

As illustrated in Figure 2, the proposed model consists of four layers: sentence encoding, graph

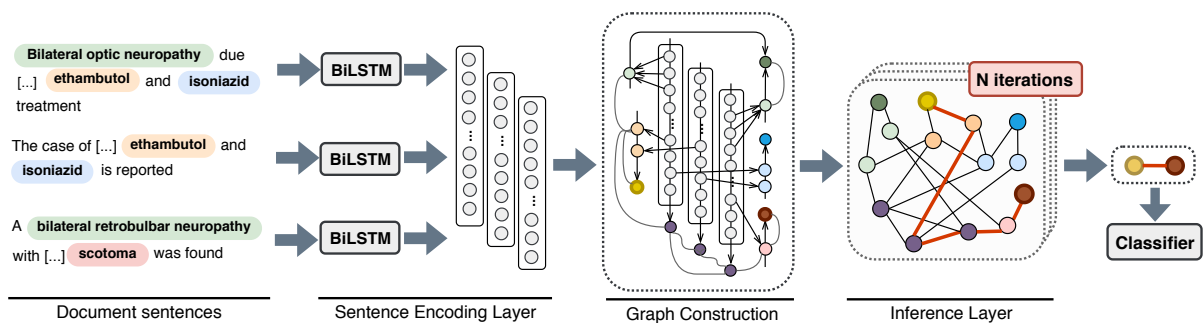


Figure 2: Abstract architecture of the proposed approach. The model receives a document and encodes each sentence separately. A document-level graph is constructed and fed into an iterative algorithm to generate edge representations between the target entity nodes. Some node connections are not shown for brevity.

construction, inference and classification layers.

The model receives as input a document with identified concept-level entities and their textual mentions. Next, a document-level graph with multiple types of nodes and edges is constructed. An inference algorithm is applied on the graph edges to generate concept-level pair representations. In the final layer, the edge representations between the target concept-entity nodes are classified into relation categories.

For the remainder of this section, we first briefly introduce the document-level RE task setting and then explain the four layers of the proposed model.

2.1 Task Setting

In concept, document-level RE the input is considered an annotated document. The annotations include concept-level entities (with assigned KB IDs), as well as multiple occurrences of each entity under the same phrase of alias, i.e., entity mentions. We consider the associations of mentions to concept entities given (also known as entity linking (Shen et al., 2014)). The objective of the task is given an annotated document, to identify all the related concept-level pairs in that document. In this work, we refer to concept-level annotations as *entities* and mention-level annotations as *mentions*.

2.2 Sentence Encoding Layer

First, each word in the sentences of the input document is transformed into a dense vector representation, i.e., a word embedding. The vectorised words of each sentence are then fed into a Bi-directional LSTM network (BiLSTM) (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997), named the encoder. The output of the encoder results in contextualised representations for each word of the input sentence.

2.3 Graph Layer

The contextualised word representations from the encoder are used to construct a document-level graph structure. The graph layer comprises of two sub-layers, a node construction layer and an edge construction layer. We compose the representations of the graph nodes in the first sub-layer and the representations of the edges in the second.

2.3.1 Node construction

We form three distinct types of nodes in the graph: mention nodes (M) n_m , entity nodes (E) n_e , and sentence nodes (S) n_s . Each node representation is computed as the average of the embeddings of different elements. Firstly, mention nodes correspond to different mentions of entities in the input document. The representation of a mention node is formed as the average of the words (w) that the mention contains. Secondly, entity nodes represent unique entity concepts. The representation of an entity node is computed as the average of the mention (m) representations associated with the entity. Finally, sentence nodes correspond to sentences. A sentence node is represented as the average of the word representations in the sentence. In order to distinguish different node types in the graph, we concatenate a node type (t) embedding to each node representation. The final node representations are then estimated as $\mathbf{n}_m = [\text{avg}_{w_i \in m}(\mathbf{w}_i); \mathbf{t}_m]$, $\mathbf{n}_e = [\text{avg}_{m_i \in e}(\mathbf{m}_i); \mathbf{t}_e]$, $\mathbf{n}_s = [\text{avg}_{w_i \in s}(\mathbf{w}_i); \mathbf{t}_s]$.

2.3.2 Edge construction

We initially construct non-directed edges between the graph nodes using heuristic rules that stem from the natural associations between the elements of a document, i.e., mentions, entities and sentences. As we cannot know in advance if two

entities are related, we do not directly connect entity nodes. Connections between nodes are based on pre-defined document-level interactions. The model objective is to generate entity-to-entity (EE) edge representations using other existing edges in the graph and consequently infer entity-to-entity relations. The different pre-defined edge types are described below.

Mention-Mention (MM): Co-occurrence of mentions in a sentence might be a weak indication of an interaction. For this reason, we create mention-to-mention edges only if the corresponding mentions reside in the same sentence. The edge representation between each mention pair m_i and m_j is generated by concatenating the representations of the nodes, the contexts c_{m_i, m_j} and a distance embedding associated with the distance between the two mentions d_{m_i, m_j} , in terms of intermediate words: $\mathbf{x}_{MM} = [\mathbf{n}_{m_i}; \mathbf{n}_{m_j}; \mathbf{c}_{m_i, m_j}; \mathbf{d}_{m_i, m_j}]$. Here, we generate the context representation for these pairs in order to encode local, pair-centric information. We use an argument-based attention mechanism (Wang et al., 2016), to measure the importance of other words in the sentence towards the mention, denoting $k \in \{1, 2\}$ as the mention arguments.

$$\begin{aligned} \alpha_{k,i} &= \mathbf{n}_{m_k}^\top \mathbf{w}_i, \\ \mathbf{a}_{k,i} &= \frac{\exp(\alpha_{k,i})}{\sum_{j \in [1, n], j \neq m_k} \exp(\alpha_{k,j})}, \\ \mathbf{a}_i &= (\mathbf{a}_{1,i} + \mathbf{a}_{2,i})/2, \\ \mathbf{c}_{m_1, m_2} &= \mathbf{H}^\top \mathbf{a}, \end{aligned} \quad (1)$$

where \mathbf{n}_{m_k} is a mention node representation, \mathbf{w}_i is a sentence word representation, \mathbf{a}_i is the attention weight of word i for mention pair m_1, m_2 , $\mathbf{H} \in \mathbb{R}^{w \times d}$ is a sentence word representations matrix, $\mathbf{a} \in \mathbb{R}^w$ is the attention weights vector for the pair and \mathbf{c}_{m_1, m_2} is the final context representation for the mention pair.

Mention-Sentence (MS): Mention-to-sentence nodes are connected only if the mention resides in the sentence. Their initial edge representation is constructed as a concatenation of the mention and sentence nodes, $\mathbf{x}_{MS} = [\mathbf{n}_m; \mathbf{n}_s]$.

Mention-Entity (ME): We connect a mention node to an entity node if the mention is associated with the entity, $\mathbf{x}_{ME} = [\mathbf{n}_m; \mathbf{n}_e]$.

Sentence-Sentence (SS): Motivated by Quirk and Poon (2017), we connect sentence nodes to encode non-local information. The main differences with prior work is that our edges are unlabelled,

non-directed and span multiple sentences. To encode the distance between sentences, we concatenate to the sentence node representations their distance in the form of an embedding: $\mathbf{x}_{SS} = [\mathbf{n}_{s_i}; \mathbf{n}_{s_j}; \mathbf{d}_{s_i, s_j}]$. We connect all sentence nodes in the graph. We consider $\mathbf{SS}_{\text{direct}}$ as direct, ordered edges (distance equal to 1) and $\mathbf{SS}_{\text{indirect}}$ as indirect, non-ordered edges (distance > 1) between S nodes, respectively. In our setting, \mathbf{SS} denotes the combination of $\mathbf{SS}_{\text{direct}}$ and $\mathbf{SS}_{\text{indirect}}$.

Entity-Sentence (ES): To directly model entity-to-sentence associations, we connect an entity node to a sentence node if at least one mention of the entity resides in this sentence, $\mathbf{x}_{ES} = [\mathbf{n}_e; \mathbf{n}_s]$.

In order to result in edge representations of equal dimensionality, we use different linear reduction layers for different edge representations,

$$\mathbf{e}_z^{(1)} = \mathbf{W}_z \mathbf{x}_z, \quad (2)$$

where $\mathbf{e}_z^{(1)}$ is an edge representation of length 1, $\mathbf{W}_z \in \mathbb{R}^{d_z \times d}$ corresponds to a learned matrix and $z \in [\text{MM}, \text{MS}, \text{ME}, \text{SS}, \text{ES}]$.

2.4 Inference Layer

We utilise an iterative algorithm to generate edges between different nodes in the graph, as well as to update existing edges. We initialise the graph only with the edges described in Section 2.3.2, meaning that direct entity-to-entity (EE) edges are absent. We can only generate EE edge representations by representing a path between their nodes. This implies that entities can be associated through an edge path of minimum length equal to 3².

For this purpose, we adapt our two-step inference mechanism, proposed in Christophoulou et al. (2018), to encode interactions between nodes and edges in the graph and hence model EE associations.

At the first step, we aim to generate a path between two nodes i and j using intermediate nodes k . We thus combine the representations of two consecutive edges e_{ik} and e_{kj} , using a modified bilinear transformation. This action generates an edge representation of double length. We combine all existing paths between i and j through k . The i , j , and k nodes can be any of the three node types E, M, or S. Intermediate nodes without adjacent

²Length 2 for an intra-sentence pair (E-S-E) or length 3 for an inter-sentence pair (E-S-S-E)

edges to the target nodes are ignored.

$$f\left(\mathbf{e}_{ik}^{(l)}, \mathbf{e}_{kj}^{(l)}\right) = \sigma\left(\mathbf{e}_{ik}^{(l)} \odot \left(\mathbf{W} \mathbf{e}_{kj}^{(l)}\right)\right), \quad (3)$$

where σ is the sigmoid non-linear function, $\mathbf{W} \in \mathbb{R}^{d_z \times d_z}$ is a learned parameter matrix, \odot refers to element-wise multiplication, l is the length of the edge and \mathbf{e}_{ik} corresponds to the representation of the edge between nodes i and k .

During the second step, we aggregate the original (short) edge representation and the new (longer) edge representation resulted from Equation (3) with linear interpolation as follows:

$$\mathbf{e}_{ij}^{(2l)} = \beta \mathbf{e}_{ij}^{(l)} + (1 - \beta) \sum_{k \neq i, j} f\left(\mathbf{e}_{ik}^{(l)}, \mathbf{e}_{kj}^{(l)}\right), \quad (4)$$

where $\beta \in [0, 1]$ is a scalar that controls the contribution of the shorter edge presentation. In general β is larger for shorter edges as we expect that the relation between two nodes is better expressed through the shortest path between them (Xu et al., 2015; Borgwardt and Kriegel, 2005).

The two steps are repeated a finite number of times N . The number of iterations is correlated with the final length of the edge representations. With initial edge length l equal to 1, the first iteration results in edges of length up-to 2. The second iteration results in edges of length up-to 4. Similarly, after N iterations, the length of edges will be up-to 2^N .

2.5 Classification Layer

To classify the concept-level entity pairs of interest, we incorporate a softmax classifier, using the entity-to-entity edges (EE) of the document graph that correspond to the concept-level entity pairs.

$$\mathbf{y} = \text{softmax}\left(\mathbf{W}_c \mathbf{e}_{EE} + \mathbf{b}_c\right), \quad (5)$$

where $\mathbf{W}_c \in \mathbb{R}^{r \times d_z}$ and $\mathbf{b}_c \in \mathbb{R}^r$ are learned parameters of the classification layer and r is the number of relation categories.

3 Experimental Settings

The model was developed using PyTorch (Paszke et al., 2017). We incorporated early stopping to identify the best training epoch and used Adam (Kingma and Ba, 2015) as the model optimiser.

3.1 Data and Task Settings

We evaluated the proposed model on two datasets: **CDR** (BioCreative V): The Chemical-Disease Reactions dataset was created by Li et al. (2016a) for document-level RE. It consists of 1,500 PubMed abstracts, which are split into three equally sized sets for training, development and testing. The dataset was manually annotated with binary interactions between Chemical and Disease concepts. For this dataset, we utilised PubMed pre-trained word embeddings (Chiu et al., 2016).

GDA (DisGeNet): The Gene-Disease Associations dataset was introduced by Wu et al. (2019), containing 30,192 MEDLINE abstracts, split into 29,192 articles for training and 1,000 for testing. The dataset was annotated with binary interactions between Gene and Disease concepts at the document-level, using distant supervision. Associations between concepts were generated by aligning the DisGeNet (Piñero et al., 2016) platform with PubMed³ abstracts. We further split the training set into a 80/20 percentage split as training and development sets. For the GDA dataset, we used randomly initialized word embeddings.

3.2 Model Settings

We explore multiple settings of the proposed graph using different edges (MM, ME, MS, ES, SS) and enhancements (node type embeddings, mention-pairs context embeddings, distance embeddings). We name our model EoG, an abbreviation of Edge-oriented Graph. We briefly describe the model settings in this section.

EoG refers to our main model with edges {MM, ME, MS, ES, SS}. The EoG (Full) setting refers to a model with a *fully connected* graph, where the graph nodes are all connected to each other, including E nodes. For this purpose, we introduce an additional linear layer for the EE edges as in Equation (2). The EoG (NoInf) setting refers to a *no inference* model, where the iterative inference algorithm (Section 2.4) is ignored. The concatenation of the entity node embeddings is used to represent the target pair. In this case, we also make use of an additional EE linear layer for EE edges. Finally, the EoG (Sent) setting refers to a model that was trained on *sentences* instead of documents. For each entity-level pair we merge the predictions of the mention-level pairs in different sentences using a maximum assumption: if at

³<https://www.ncbi.nlm.nih.gov/pubmed/>

Method	Overall (%)			Intra (%)			Inter (%)		
	P	R	F1	P	R	F1	P	R	F1
Gu et al. (2017)	55.7	68.1	61.3	59.7	55.0	57.2	51.9	7.0	11.7
Verga et al. (2018)	55.6	70.8	62.1	-	-	-	-	-	-
Nguyen and Verspoor (2018)	57.0	68.6	62.3	-	-	-	-	-	-
EoG	62.1	65.2	63.6	64.0	73.0	68.2	56.0	46.7	50.9
EoG (Full)	59.1	56.2	57.6	71.2	62.3	66.5	37.1	42.0	39.4
EoG (NoInf)	48.2	50.2	49.2	65.8	55.2	60.2	25.4	38.5	30.6
EoG (Sent)	56.9	53.5	55.2	56.9	76.4	65.2	-	-	-
Zhou et al. (2016)	55.6	68.4	61.3	-	-	-	-	-	-
Peng et al. (2016)	62.1	64.2	63.1	-	-	-	-	-	-
Li et al. (2016b)	60.8	76.4	67.7	67.3	52.4	58.9	-	-	-
Panyam et al. (2018)	53.2	69.7	60.3	54.7	80.6	65.1	47.8	43.8	45.7
Zheng et al. (2018)	56.2	67.9	61.5	-	-	-	-	-	-

Table 1: Overall, intra- and inter-sentence pairs performance comparison with the state-of-the-art on the CDR test set. The methods below the double line take advantage of additional training data and/or incorporate external tools.

least one mention-level prediction indicates a relation then we predict the entity pair as related, similarly to Gu et al. (2017). All of the settings incorporate node type embeddings, contextual embeddings for MM edges and distance embeddings for MM and SS edges, unless otherwise stated.

4 Results

Table 1 depicts the performance of our proposed model on the CDR test set, in comparison with the state-of-the-art. We directly compare our model with models that do not incorporate external knowledge. Verga et al. (2018) and Nguyen and Verspoor (2018) consider a single pair per document, while Gu et al. (2017) develops separate models for intra- and inter-sentence pairs. As it can be observed, the proposed model outperforms the state-of-the-art in CDR dataset by 1.3 percentage points of overall performance. We also show the methods that take advantage of syntactic dependency tools. Li et al. (2016b) uses co-training with additional unlabeled training data. Our model performs significantly better on intra- and inter-sentential pairs, even compared to most of the models with external knowledge, except for Li et al. (2016b).

In addition, we report the performance of three baseline models. The EoG model outperforms all baselines for all pair types. In particular, for the inter-sentence pairs, performance significantly drops with a fully connected graph (Full) or without inference (NoInf). The former might indicate the existence of certain reasoning paths that should be followed in order to relate entities residing in different sentences. It is also important

Model	Dev		Test		F1 (%)	
	Overall		Intra		Inter	
EoG	78.7	81.5	82.5	85.2	48.8	50.0
EoG (Full)	78.6	80.8	82.4	84.1	52.3	54.7
EoG (NoInf)	71.8	74.6	76.8	79.1	45.5	49.3
EoG (Sent)	73.8	73.8	78.1	78.8	-	-

Table 2: Performance comparison on the GDA development and test sets.

to note that the intra-sentence pairs substantially benefit from the document-level information, as EoG surpasses the performance of training on single sentences (Sent) by 3%. Finally, the performance drop in intra-sentence pairs, as a result of the inference algorithm removal (NoInf), suggests that multiple entity associations exist in sentences (Christopoulou et al., 2018). Their interactions can be beneficial in cases of lack of word context information.

We also apply our model on the distantly supervised GDA dataset. As shown in Table 2 results for intra-sentence pairs are consistent with the findings of the CDR dataset for both development and test sets. This indicates that document-level information is helpful. However, performance differs for inter-sentence pairs and in particular for the fully connected graph (Full) baseline. We partially attribute this behavior to the small number of inter-sentence pairs in the GDA dataset (only 13% compared to 30% in the CDR dataset) that results in inadequate learning patterns for EoG. We leave further investigation as part of future work.

Embeddings	F1 (%)		
	Overall	Intra	Inter
EoG (PubMed)	63.62	68.25	50.94
EoG (GloVe)	63.01	67.52	50.26
EoG (random)	61.41	66.80	46.51

Table 3: Performance of EoG on the CDR test set with different pre-trained word embeddings.

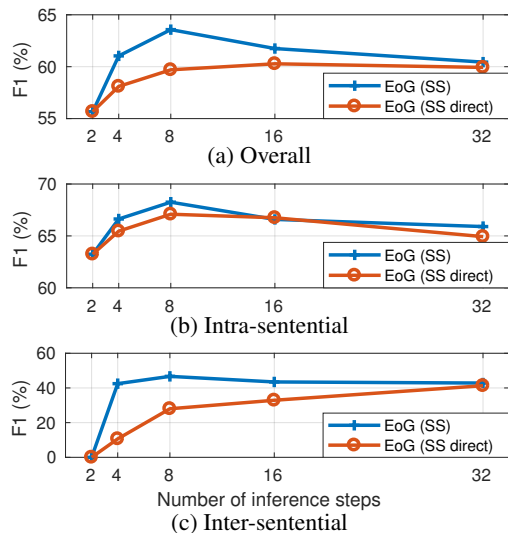


Figure 3: Performance as a function of the number of inference steps when using direct (SS_{direct}) or direct and indirect (SS) sentence-to-sentence edges, on the CDR development set.

5 Analysis & Discussion

We first analyse the performance of our main model (EoG) using different pre-trained word embeddings. Table 3 shows the performance difference between domain-specific (PubMed) (Chiu et al., 2016), general-domain (GloVe) (Pennington et al., 2014) and randomly initialized (random) word embeddings. As observed, our proposed model performs consistently with both in-domain and out-of-domain pre-trained word embeddings. The low performance of random embeddings is due to the small size of the dataset, which results in lower quality embeddings.

For further analysis, we choose the CDR dataset as it is manually annotated. To better analyse the behaviour of our model, we conduct analysis on the effect of direct and indirect sentence-to-sentence edges as a function of the inference steps. Figures 3a, 3b and 3c illustrate the performance of both graphs for overall, intra- and inter-sentence pairs respectively.

The first observation is that usage of direct edges only, reduces the overall performance al-

Edge Types	F1 (%)		
	Overall	Intra	Inter
EE	55.14	61.31	40.34
EoG	63.57	68.25	46.68
-MM	62.77	67.93	46.65
-ME	61.57	66.39	45.40
-MS	62.92	67.55	44.74
-ES	61.41	66.44	43.04
- SS_{indirect}	59.70	67.09	28.00
-SS	57.41	65.45	1.59
-MM, ME, MS	60.46	66.07	39.56
-ES, MS, SS	56.86	64.63	0.00

Table 4: Ablation analysis for different edge and node types on the CDR development set.

most by 4%, for inference step $l = 8$. This drop mostly affects inter-sentence pairs, where a 18% point drop is observed. In fact, ordered edges (SS_{direct}) need longer inference to perform better, in comparison with additional indirect edges (SS) for which less steps are required. The superiority of SS edges, for all inference steps, compared to SS_{direct} edges on inter-sentence pairs detection, indicates that in a narrative, some intermediate information is not important. The observation that indirect edges perform slightly better than direct for intra-sentence pairs ($l \leq 16$) agrees with the results of Table 1 where we showed that inter-sentence information can act as complementary evidence for intra-sentence pairs.

We additionally conduct ablation analysis on the graph edges and nodes, as shown in Table 4. Usage of EE edges only results in poor performance across pairs. Removal of MM and ME edges does not significantly affect the performance as ES edges can replace their impact. Complete removal of connections to M nodes results in low inter-sentence performance. This behaviour pinpoints the importance of some local dependencies in identifying cross-sentence relations.

Removal of ES edges reduces the performance of all pairs, as encoding of EE edges becomes more difficult⁴. We further observe very poor identification of inter-sentence pairs without sentence-to-sentence connections. This is complementary with the inability of the model to identify any inter-sentence pairs without connections to S nodes. In this scenario, we enable identification of pairs across sentences only through MM and ME edges, as shown in Figure 4a. In the CDR dataset,

⁴Length 3 (E-M-M-E) for intra- and length 5 (E-M-S-S-M-E) for inter-sentence pairs.

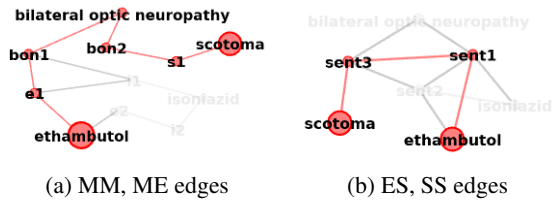


Figure 4: Relation paths with different types of edges.

Model	F1 (%)		
	Overall	Intra	Inter
EoG	63.57	68.25	46.68
-node types (T)	62.31	67.50	44.80
-MM context (C)	62.88	67.67	46.59
-distances (D)	62.53	68.00	41.53
-T, C, D	63.10	68.44	43.48

Table 5: Ablation analysis of edge enhancements on the CDR development set.

78% of inter-sentential pairs have at least one argument that is mentioned only once in the document. The identification of these pairs, without S nodes, requires very long inference paths⁵. As shown in Figure 4b, the introduction of S nodes results in a path with half the length, which we expect to better represent the relation. Longer inference representations are much weaker than shorter ones. This suggests that the inference mechanism has limited capability in identifying very complex associations.

We then investigate the additional enhancements of the graph edges in Table 5. In general, intra-sentence pairs are not affected by these settings. However, for inter-sentence pairs, removal of node type embeddings and distance embeddings results in a 2% and 5% drop in terms of F1-score. These results indicate that the interactions between different elements in a document, along with the distance between sentences and mentions, play an important role in inter-sentence pair inference. Removing all of these settings does not perform worse than removing one of them, which might indicate model overfitting. We plan to further investigate this as part of future work.

We examine the performance of different models on inter-sentence pairs, based on their sentence-level distances. Figure 5 illustrates that for long-distanced pairs, EoG has lower performance, indicating the difficulty in predicting them and a possible requirement for other, latent document-level information (EoG (Full)).

⁵Minimum inference length 6 (E-M-M-E-M-M-E).

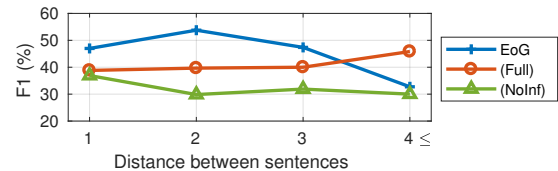


Figure 5: Performance of inter-sentence pairs on the CDR development set as a function of their sentence distance.

Following short exposure to oral **prednisone** [...]. Both presented in the emergency room with profound **coma**, **hypotension**, severe **hyperglycemia**, and acidosis.

The etiology of **pyeloureteritis cystica** has long been [...] The *disease* occurred subsequent to the initiation of **heparin** therapy [...]

Time trends in **warfarin**-associated **hemorrhage**. [...] The proportion of patients with major and **intracranial bleeding** increased [...]

Table 6: Inter-sentence pairs from the CDR development set that EoG fails to detect.

As final analysis, we investigate some of the cases where the graph models are unable to identify inter-sentence related pairs. For this purpose, we randomly check some of the common false negative errors among the EoG models. We identify three frequent cases of errors, as shown in Table 6. In the first case, when multiple entities reside in the same sentence and are connected with conjunctions (e.g., ‘and’) or commas, the model often failed to find associations with all of them. The second error derives from missing coreference connections. For instance, *pyeloureteritis cystica* is referred to as *disease*. Although our model cannot directly create these edges, S nodes potentially simulate such links, by encoding the co-referring entities into the sentence representation. Finally, incomplete entity linking results into additional model errors. For instance, in the third example, *hemorrhage* and *intracranial bleeding* are synonymous terms. However, they are assigned different KB IDs, hence treated as different entities. The model can find the intra-sentential relation but not the inter-sentential one.

6 Related Work

Traditional approaches focus on intra-sentence supervised RE, utilising CNN or RNN, ignoring multiple entities in a sentence (Zeng et al., 2014; Nguyen and Grishman, 2015) as well as incorporating external syntactic tools (Miwa and Bansal,

2016; Zhang et al., 2018). Christopoulou et al. (2018) considered intra-sentence entity interactions without domain dependencies by modelling long dependencies between the entities of a sentence.

Other approaches deal with distantly-supervised datasets but are also limited to intra-sentential relations. They utilise Piecewise Convolutional Neural Networks (PCNN) (Zeng et al., 2015), attention mechanisms (Lin et al., 2016; Zhou et al., 2018), entity descriptors (Jiang et al., 2016) and graph CNNs (Vashishth et al., 2018) to perform MIL on bags-of-sentences that contain multiple mentions of an entity pair. Recently, Zeng et al. (2017) proposed a method for extracting paths between entities using the target entities' mentions in several different sentences (in possibly different documents) as intermediate connectors. They allow mention-mention edges only if these mentions belong to the same entity and consider that a single mention pair exists in a sentence. On the contrary, we not only allow interactions between all mentions in the same sentence, but also consider multiple edges between mentions, entities and sentences in a document.

Current approaches that try to deal with document-level RE are mostly graph-based. Quirk and Poon (2017) introduced the notion of a document graph, where nodes are words and edges represent intra- and inter-sentential relations between the words. They connected words with different dependency edges and trained a binary logistic regression classifier. They evaluated their model on distantly supervised full-text articles from PubMed for Gene-Drug associations, restricting pairs within a window of consecutive sentences. Following this work, other approaches incorporated graphical models for document-level RE such as graph LSTM (Peng et al., 2017), graph CNN (Song et al., 2018) or RNNs on dependency tree structures (Gupta et al., 2019). Recently, Jia et al. (2019) improved n -ary RE using information from multiple sentences and paragraphs in a document. Similar to our approach, they choose to directly classify concept-level pairs rather than multiple mention-level pairs. Although they consider sub-relations to model related tuples, they ignore interactions with other entities outside of the target tuple in the discourse units.

Non-graph-based approaches utilise different

intra- and inter-sentence models and merge the resulted predictions (Gu et al., 2016, 2017). Other approaches extract document-level representations for each candidate entity pair (Zheng et al., 2018; Li et al., 2018; Wu et al., 2019), or use syntactic dependency structures (Zhou et al., 2016; Peng et al., 2016). Verga et al. (2018) proposed a Transformer-based model for document-level relation extraction with multi-instance learning, merging multiple mention pairs. Nguyen and Verspoor (2018) used a CNN with additional character-level embeddings. Singh and Bhatia (2019) also utilised Transformer and connected two target entities by combining them directly and via a contextual token. However, they consider a single target entity pair per document.

7 Conclusion

We presented a novel edge-oriented graph neural model for document-level relation extraction using multi-instance learning. The proposed model constructs a document-level graph with heterogeneous types of nodes and edges, modelling intra- and inter-sentence pairs simultaneously with an iterative algorithm over the graph edges. To the best of our knowledge, this is the first approach to utilise an edge-oriented model for document-level RE.

Analysis on intra- and inter-sentence pairs indicated that the proposed, partially-connected, document graph structure can effectively encode dependencies between document elements. Additionally, we deduce that document-level information can contribute to the identification of intra-sentence pairs leading to higher precision and F1-score.

As future work, we plan to improve the inference mechanism and potentially incorporate additional information in the document-graph structure. We hope that this study will inspire the community to further investigate the usage of edge-oriented models on RE and other related tasks.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments and AIST/AIRC for providing the computational resources for this study. Research was funded by the University of Manchester James Elson Studentship Award and the BBSRC Japan Partnering Award BB/P025684/1.

References

- Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2670–2676. Morgan Kaufmann Publishers Inc.
- Karsten M Borgwardt and Hans-Peter Kriegel. 2005. Shortest-path kernels on graphs. In *Proceedings of the IEEE International Conference on Data Mining*, pages 74–81. IEEE Computer Society.
- Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proceedings of the BioNLP workshop*, pages 166–174.
- Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2018. A walk-based model on entity graphs for relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 2)*, pages 81–88. Association for Computational Linguistics.
- Jinghang Gu, Longhua Qian, and Guodong Zhou. 2016. Chemical-induced disease relation extraction with various linguistic features. *Database*.
- Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2017. Chemical-induced disease relation extraction via convolutional neural network. *Database*.
- Pankaj Gupta, Subburam Rajaram, Hinrich Schtze, and Thomas Runkler. 2019. Neural relation extraction within and across sentence boundaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6513–6520.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multi-scale representation learning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1)*, pages 3693–3704. Association for Computational Linguistics.
- Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. 2016. Relation extraction with multi-instance multi-label convolutional neural networks. In *Proceedings of the International Conference on Computational Linguistics: Technical Papers*, pages 1471–1480. The COLING Organizing Committee.
- Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.
- Haodi Li, Ming Yang, Qingcai Chen, Buzhou Tang, Xiaolong Wang, and Jun Yan. 2018. Chemical-induced disease extraction via recurrent piecewise convolutional neural networks. *BMC medical informatics and decision making*, 18(2):60.
- Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wieggers, and Zhiyong Lu. 2016a. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*.
- 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 *IEEE International Conference on Bioinformatics and Biomedicine*, pages 994–1001. IEEE.
- Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1)*, pages 2124–2133.
- Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1)*, pages 3036–3046. Association for Computational Linguistics.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (Volume 2)*, pages 1003–1011. Association for Computational Linguistics.
- Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1)*, pages 1105–1116. Association for Computational Linguistics.
- Dat Quoc Nguyen and Karin Verspoor. 2018. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. In *Proceedings of the BioNLP workshop*, pages 129–136. Association for Computational Linguistics.
- Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48.

- 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. In *NIPS-W*.
- 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(1):53.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Empirical Methods in Natural Language Processing*, pages 1532–1543. Association for Computational Linguistics.
- Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. 2016. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, pages D833–D839.
- Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (Volume 1)*, pages 1171–1182. Association for Computational Linguistics.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.
- Mike Schuster and Kuldeep K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.
- Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.
- Gaurav Singh and Parminder Bhatia. 2019. Relation extraction using explicit context conditioning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1442–1447. Association for Computational Linguistics.
- Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. N-ary relation extraction using graph-state lstm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2226–2235. Association for Computational Linguistics.
- Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266. Association for Computational Linguistics.
- Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1)*, pages 872–884. Association for Computational Linguistics.
- Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the Association for Computational Linguistics (Volume 1)*, pages 1298–1307. Association for Computational Linguistics.
- 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 *International Conference on Research in Computational Molecular Biology*, pages 272–284. Springer.
- Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 1785–1794. Association for Computational Linguistics.
- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344. Dublin City University and Association for Computational Linguistics.
- Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Incorporating relation paths in neural relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1768–1777. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215. Association for Computational Linguistics.

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.

Huiwei Zhou, Huijie Deng, Long Chen, Yunlong Yang, Chen Jia, and Degen Huang. 2016. Exploiting syntactic and semantics information for chemical-disease relation extraction. *Database*.

Peng Zhou, Jiaming Xu, Zhenyu Qi, Hongyun Bao, Zhineng Chen, and Bo Xu. 2018. Distant supervision for relation extraction with hierarchical selective attention. *Neural Networks*, 108:240–247.

A Datasets

In Tables 7-8 we summarise the statistics for the CDR and GDA datasets, respectively. For all datasets, we used the GENIA Sentence Splitter⁶ and GENIA Tagger⁷ for sentence splitting and word tokenisation respectively. We additionally removed mentions in the given abstracts that were not grounded to a Knowledge Base ID (ID equal to -1).

Due to the small size of the CDR dataset, some approaches create a new split from the union of train and development sets (Verga et al., 2018; Zhou et al., 2018). We select to merge the train and development sets and re-train our model on their entire union for evaluation on the test set following Lin et al. (2016) and Zhou et al. (2016). To compare with related work, we followed Verga et al. (2018) and Gu et al. (2016) and ignored non-related pairs that correspond to general concepts (MeSH vocabulary hypernym filtering).

B Hyper-parameter Setting

We used the development set to identify the stopping training epoch and tune the number of inference iterations. Except from these parameters, all experiments used the same hyperparameters, with a fixed initialisation seed. For the CDR dataset EoG, (Full) and (Sent) models performed best with $l = 8, 2, 4$ inference steps, respectively. The

⁶<http://www.nactem.ac.uk/y-matsu/geniass/>

⁷<http://www.nactem.ac.uk/GENIA/tagger/>

	Train	Dev	Test
Documents	500	500	500
Positive pairs	1,038	1,012	1,066
Intra	754	766	747
Inter	284	246	319
Negative pairs	4,202	4,075	4,138
Entities			
Chemical	1,467	1,507	1,434
Disease	1,965	1,864	1,988
Mentions			
Chemical	5,162	5,307	5,370
Disease	4,252	4,328	4,430

Table 7: CDR (BioCreative V) dataset statistics.

	Train	Dev	Test
Documents	23,353	5,839	1,000
Positive pairs	36,079	8,762	1,502
Intra	30,905	7,558	1,305
Inter	5,174	1,204	197
Negative pairs	96,399	24,362	3,720
Entities			
Gene	46,151	11,406	1,903
Disease	67,257	16,703	2,778
Mentions			
Gene	205,457	51,410	8,404
Disease	226,015	56,318	9,524

Table 8: GDA (DisGeNet) dataset statistics.

chosen batchsize was equal to 2. For the GDA dataset, EoG and EoG (Full) performed best with $l = 16$ and EoG (Sent) with $l = 8$ inference steps. The chosen batchsize was equal to 3. For all experiments performance was measured in terms of micro precision (P), recall (R) and F1-score (F1). We list the hyper-parameters used to train the proposed model in Table 9.

Parameter	Value
Batch size	[2, 3]
Learning rate	0.002
Gradient clipping	10
Early stop patience	10
Regularization	10^{-4}
Dropout word embedding layer	0.5
Dropout classification layer	0.3
Word dimension	200
Node type dimension	10
Distance dimension	10
Edge dimension	100
β	0.8
Optimizer	Adam
Inference iterations	[0, 5]

Table 9: Hyper-parameter values used in the reported experiments.