Incorporating Chains of Reasoning over Knowledge Graph for Distantly Supervised Biomedical Knowledge Acquisition

Qin Dai[1], **Naoya Inoue**[1,2], **Paul Reisert**[2], **Ryo Takahashi**[1] **and Kentaro Inui**[1,2]
[1]Tohoku University, Japan
[2]RIKEN Center for Advanced Intelligence Project, Japan
{daiqin, naoya-i, preisert, ryo.t, inui}@ecei.tohoku.ac.jp

Abstract

The increased demand for structured scientific knowledge has attracted considerable attention on extracting scientific relation from the ever growing scientific publications. Distant supervision is a widely applied approach to automatically generate large amounts of labelled sentences for scientific Relation Extraction (RE). However, the brevity of the labelled sentences would hinder the performance of distantly supervised RE (DS-RE). Specifically, authors always omit the Background Knowledge (BK) that they assume is well known by readers, but would be essential for a machine to identify relationships. To address this issue, in this work, we assume that the reasoning paths between entity pairs over a knowledge graph could be utilized as BK to fill the "gaps" in text and thus facilitate DS-RE. Experimental results prove the effectiveness of the reasoning paths for DS-RE, because the proposed model that incorporates the reasoning paths achieves significant and consistent improvements as compared with a state-of-the-art DS-RE model.

1 Introduction

Scientific Knowledge Graph (KG), such as Unified Medical Language System (UMLS) [1], is extremely crucial for many scientific Natural Language Processing (NLP) tasks such as Question Answering (QA), Information Retrieval (IR) and Relation Extraction (RE). Scientific KG provides large collections of relations between entities, typically stored as (h, r, t) triplets, where h = *head entity*, r = relation and t = *tail entity*, e.g., (*acetaminophen*, *may_treat*, *pain*). However, KGs are often highly incomplete (Min et al., 2013). Scientific KGs, as with general KGs such as Freebase (Bollacker et al., 2008) and DBpedia (Lehmann et al., 2015), are far from complete and this would impede their usefulness in real-world applications. Scientific KGs, on the one hand, face the data sparsity problem. On the other hand, scientific publications have become the largest repository ever for scientific KGs and continue to increase at an unprecedented rate (Munroe, 2013). Therefore, it is an essential and fundamental task to turn the unstructured scientific publications into well organized KG, and it belongs to the task of RE.

One obstacle that is encountered when building a RE system is the generation of training instances. For coping with this difficulty, (Mintz et al., 2009) proposes distant supervision to automatically generate training samples via leveraging the alignment between KGs and texts. They assume that if two entities are connected by a relation in a KG, then all sentences that contain those entity pairs will express the relation. For instance, (*ketorolac_tromethamine*, *may_treat*, *pain*) is a fact triplet in UMLS. Distant supervision will automatically label all sentences, such as Example 1, Example 2 and Example 3, as positive instances for the relation *may_treat*. Although distant supervision could provide a large amount of training data at low cost, it always suffers from wrong labelling problem. For instance, comparing to Example 1, Example 2 and Example 3 should not be seen as the convincing evidences to support the *may_treat* relationship between *ketorolac_tromethamine* and *pain*, but will still be annotated as positive instances by the distant supervision.

[1]https://www.nlm.nih.gov/research/umls/

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 19-28, Hakodate, Japan, September 13-15, 2019

(1) *The analgesic effectiveness of **ketorolac_tromethamine** was compared with hydrocodone and acetaminophen for **pain** from an arthroscopically assisted patellar-tendon autograft anterior cruciate ligament reconstruction.*

(2) *This double-blind, split-mouth, and randomized study was aimed to compare the efficacy of dexamethasone and **ketorolac_tromethamine**, through the evaluation of **pain**, edema, and limitation of mouth_opening.*

(3) *A loading dose of parental **ketorolac_tromethamine** was administered and subjects were later given two staged doses of the same "unknown" drug with **pain** evaluations conducted after each dose.*

To automatically alleviate the wrong labelling problem, (Riedel et al., 2010; Hoffmann et al., 2011) apply multi-instance learning. In order to avoid the handcrafted features and errors propagated from NLP tools, (Zeng et al., 2015) proposes a Convolutional Neural Network (CNN), which incorporate mutli-instance learning with neural network model, and achieves significant improvement in distantly supervised RE (DS-RE). Recently, attention mechanism is applied to effectively extract features from all collected sentences, rather than from the most informative one that previous work has focused on. (Lin et al., 2016) proposes a relation vector based attention mechanism for DS-RE. (Han et al., 2018) proposes a novel joint model that leverages a KG-based attention mechanism and achieves significant improvement than (Lin et al., 2016).

Although the KG-based model outperforms several state-of-the-art DS-RE models, the brevity of textual information would inevitably hinder its performance. Specifically, authors always leave out information that they assume is known to their readers. For instance, Example 2 omits the background connection between *ketorolac_tromethamine* and *pain* and implicitly conveys that the former *may_treat* the latter. Human readers could easily make this inference based on their Background Knowledge (BK) about the target entity pair. However, for a machine, it would be extremely difficult to identify the relationship just from the given sentence without the important BK.

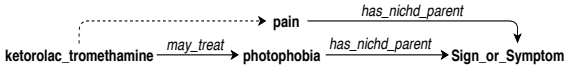

Figure 1: An example of reasoning path.

To address the issue of textual brevity, in this work, we assume that the paths (or reasoning paths) between an entity pair over a KG could be applied as the BK to fill the "gaps" and thereby improve the performance of DS-RE. For instance, one reasoning path between *ketorolac_tromethamine* and *pain* over UMLS is shown in Figure 1. By observing the path, we may infer with some likelihood that $(ketorolac\_tromethamine, may\_treat, pain)$, because *ketorolac_tromethamine* could be prescribed to treat some *Sign_or_Symptom* such as *photophobia*, and *pain* is a *Sign_or_Symptom*, therefore *ketorolac_tromethamine* might be used to treat *pain*. By comprehensively considering the path in Figure 1 and the sentence in Example 2, we could further prove the inference. To this end, we propose the DS-RE model that not only encodes the sentences containing target entity pairs, but also the reasoning paths between them over a KG.

We conduct evaluation on biomedical datasets in which KG is collected from UMLS and textual data is extracted from Medline corpus. The experimental results prove the effectiveness of the incorporation of reasoning paths for improving DS-RE from biomedical datasets.

2 Related Work

RE is a fundamental task in the NLP community. In recent years, Neural Network (NN)-based models have been the dominant approaches for non-scientific RE, which include Convolutional Neural Network (CNN)-based frameworks (Zeng et al., 2014; Xu et al., 2015; Santos et al., 2015) Recurrent Neural Network (RNN)-based frameworks (Zhang and Wang, 2015; Miwa and Bansal, 2016; Zhou et al., 2016). NN-based approaches are also used in scientific RE. For instance, (Gu et al., 2017) utilizes a CNN-based model for identifying *chemical-disease* relations from Medline corpus. (Hahn-

Powell et al., 2016) proposes an LSTM-based model for identifying *causal precedence* relationship between two event mentions in biomedical papers. (Ammar et al., 2017) applies (Miwa and Bansal, 2016)'s model for scientific RE.

Although remarkably good performances are achieved by the models mentioned above, they still train and extract relations on sentence-level and thus need a large amount of annotation data, which is expensive and time-consuming. To address this issue, distant supervision is proposed by (Mintz et al., 2009). To alleviate the noisy data from the distant supervision, many studies model DS-RE as a Multiple Instance Learning (MIL) problem (Riedel et al., 2010; Hoffmann et al., 2011; Zeng et al., 2015), in which all sentences containing a target entity pair (e.g., *ketorolac_tromethamine* and *pain*) are seen as a bag to be classified. To make full use of all the sentences in the bag, rather than just the most informative one in the bag, researchers apply attention mechanism in deep NN-based models for DS-RE. (Lin et al., 2016) proposes a relation vector based attention mechanism to extract feature from the entire bag and outperforms the prior approaches. (Du et al., 2018) proposes multi-level structured self-attention mechanism. (Han et al., 2018) proposes a joint model that adopts a KG-based attention mechanism and achieves significant improvement than (Lin et al., 2016) on DS-RE.

The attention mechanism in deep NN-based models has achieved significant progress on DS-RE. However, the brevity of input sentences could still negatively affect the performance. To address this issue, we assume that the reasoning paths between target entity pairs over a KG could be applied as BK to fill the "gaps" of input sentences and thus promote the efficiency of DS-RE. (Roller et al., 2015) uses some inference pattern learned from UMLS for eliminating potentially related entity pairs from negative training data for DS-RE. (Ji et al., 2017) applies entity descriptions generated form Freebase and Wikipedia as BK, (Lin et al., 2017) utilizes multilingual text as BK and (Vashishth et al., 2018) uses relation alias information (e.g., $founded$ and $co\text{-}founded$ are aliases for the relation $founderOfCompany$) as BK for DS-RE. However, none of these existing approaches mentioned above comprehensively consider multiple sentences containing entity pairs and multiple reasoning paths between them for DS-RE.

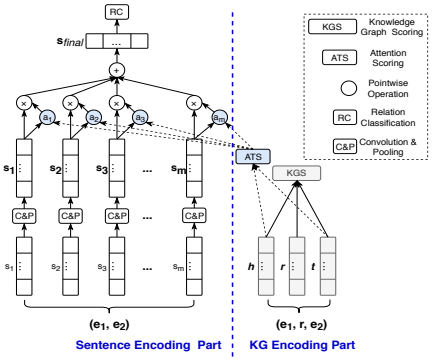

Figure 2: Overview of the base model.

3 Base Model

The success of the joint model proposed by (Han et al., 2018) inspires us to choose the strong model as our base model for biomedical DS-RE. The architecture of the base model is illustrated in Figure 2. In this section, we will introduce the base model proposed by (Han et al., 2018) in two main parts: KG Encoding part and Sentence Encoding part.

3.1 KG Encoding Part

Suppose we have a KG containing a set of fact triplets $\mathcal{O} = \{(e_1, r, e_2)\}$, where each fact triplet consists of two entities $e_1, e_2 \in \mathcal{E}$ and their relation $r \in \mathcal{R}$. Here $\mathcal{E}$ and $\mathcal{R}$ stand for the set of entities and relations respectively. KG Encoding Part then encodes $e_1, e_2 \in \mathcal{E}$ and their relation $r \in \mathcal{R}$ into low-dimensional vectors $\mathbf{h}, \mathbf{t} \in R^d$ and $\mathbf{r} \in R^d$ respectively, where d is the dimensionality of the embedding space. The base model adopts two Knowledge Graph Completion (KGC) models Prob-TransE and Prob-TransD, which are based on TransE (Bordes et al., 2013) and TransD (Ji et al., 2015) respectively, to score a given fact triplet. Specifically, given an entity pair (e_1, e_2), Prob-TransE defines its latent relation embedding $\mathbf{r}_{ht}$ via the Equation 1.

$$\mathbf{r}_{ht} = \mathbf{t} - \mathbf{h} \tag{1}$$

Prob-TransD is an extension of Prob-TransE and introduces additional mapping vectors $\mathbf{h}_p, \mathbf{t}_p \in R^d$ and $\mathbf{r}_p \in R^d$ for e_1, e_2 and r respectively. Prob-TransD encodes the latent relation embedding via

the Equation 2, where $\mathbf{M}_{rh}$ and $\mathbf{M}_{rt}$ are projection matrices for mapping entity embeddings into relation spaces.

$$\begin{aligned} \mathbf{r}_{ht} &= \mathbf{t}_r - \mathbf{h}_r, \\ \mathbf{h}_r &= \mathbf{M}_{rh}\mathbf{h}, \\ \mathbf{t}_r &= \mathbf{M}_{rt}\mathbf{t}, \\ \mathbf{M}_{rh} &= \mathbf{r}_p\mathbf{h}_p^\top + \mathbf{I}^{d\times d}, \\ \mathbf{M}_{rt} &= \mathbf{r}_p\mathbf{t}_p^\top + \mathbf{I}^{d\times d} \end{aligned} \tag{2}$$

The conditional probability can be formalized over all fact triplets $\mathcal{O}$ via the Equations 3 and 4, where $f_r(e_1, e_2)$ is the KG scoring function, which is used to evaluate the plausibility of a given fact triplet. For instance, the score for (*aspirin*, *may_treat*, *pain*) would be higher than that for (*aspirin*, *has_ingredient*, *pain*), because the former is more plausible than the latter. $\theta_\mathcal{E}$ and $\theta_\mathcal{R}$ are parameters for entities and relations respectively, b is a bias constant.

$$P(r|(e_1, e_2), \theta_\mathcal{E}, \theta_\mathcal{R}) = \frac{\exp(f_r(e_1, e_2))}{\sum_{r'\in\mathcal{R}} \exp(f_{r'}(e_1, e_2))} \tag{3}$$

$$f_r(e_1, e_2) = b - \|\mathbf{r}_{ht} - \mathbf{r}\| \tag{4}$$

3.2 Sentence Encoding Part

Sentence Representation Learning. Given a sentence s with n words $s = \{w_1, ..., w_n\}$ including a target entity pair (e_1, e_2), CNN is used to generate a distributed representation $\mathbf{s}$ for the sentence. Specifically, vector representation $\mathbf{v}_t$ for each word w_t is calculated via Equation 5, where $\mathbf{W}^w_{emb}$ is a word embedding projection matrix (Mikolov et al., 2013), $\mathbf{W}^{wp}_{emb}$ is a word position embedding projection matrix, $\mathbf{x}^w_t$ is a one-hot word representation and $\mathbf{x}^{wp}_t$ is a one-hot word position representation. The word position describes the relative distance between the current word and the target entity pair (Zeng et al., 2014). For instance, in the sentence *"Patients recorded pain$_{e_2}$ and aspirin$_{e_1}$ consumption in a daily diary"*, the relative distance of the word *"and"* is [1, -1].

$$\begin{aligned} \mathbf{v}_t &= [\mathbf{v}^w_t; \mathbf{v}^{wp1}_t; \mathbf{v}^{wp2}_t], \\ \mathbf{v}^w_t &= \mathbf{W}^w_{emb}\mathbf{x}^w_t, \\ \mathbf{v}^{wp1}_t &= \mathbf{W}^{wp}_{emb}\mathbf{x}^{wp1}_t, \\ \mathbf{v}^{wp2}_t &= \mathbf{W}^{wp}_{emb}\mathbf{x}^{wp2}_t \end{aligned} \tag{5}$$

The distributed representation $\mathbf{s}$ is formulated via the Equation 6, where, $[\mathbf{s}]_i$ and $[\mathbf{h}_t]_i$ are the i-th value of $\mathbf{s}$ and $\mathbf{h}_t$, M is the dimensionality of $\mathbf{s}$, $\mathbf{W}$ is the convolution kernal, $\mathbf{b}$ is a bias vector, and k is the convolutional window size.

$$[\mathbf{s}]_i = \max_t\{[\mathbf{h}_t]_i\},\ \forall i = 1, ..., M \tag{6}$$

$$\begin{aligned} \mathbf{h}_t &= \tanh(\mathbf{W}\mathbf{z}_t + \mathbf{b}), \\ \mathbf{z}_t &= [\mathbf{v}_{t-(k-1)/2}; ...; \mathbf{v}_{t+(k-1)/2}] \end{aligned}$$

KG-based Attention. Suppose for each fact triplet (e_1, r, e_2), there might be multiple sentences $S_r = \{s_1, ..., s_m\}$ in which each sentence contains the entity pair (e_1, e_2) and is assumed to imply the relation r, m is the size of S_r. As discussed before, the distant supervision inevitably collect noisy sentences, the base model adopts a KG-based attention mechanism to discriminate the informative sentences from the noisy ones. Specifically, the base model uses the latent relation embedding $\mathbf{r}_{ht}$ from Equation 1 (or Equation 2) as the attention over S_r to generate its final representation $\mathbf{s}_{final}$. $\mathbf{s}_{final}$ is calculated via Equation 7, where $\mathbf{W}_s$ is the weight matrix, $\mathbf{b}_s$ is the bias vector, a_i is the weight for $\mathbf{s}_i$, which is the distributed representation for the i-th sentence in S_r.

$$\mathbf{s}_{final} = \sum_{i=1}^{m} a_i\mathbf{s}_i, \tag{7}$$

$$\begin{aligned} a_i &= \frac{\exp(\langle\mathbf{r}_{ht}, \mathbf{x}_i\rangle)}{\sum_{k=1}^{m}\exp(\langle\mathbf{r}_{ht}, \mathbf{x}_k\rangle)}, \\ \mathbf{x}_i &= \tanh(\mathbf{W}_s\mathbf{s}_i + \mathbf{b}_s) \end{aligned}$$

Finally, the conditional probability $P(r|S_r, \theta_s)$ is formulated via Equation 8 and Equation 9, where, θ_S is the parameters in Sentence Encoding Part, $\mathbf{M}$ is the representation matrix of relations, $\mathbf{d}$ is a bias vector, $\mathbf{o}$ is the output vector containing the prediction probabilities of all target relations for the input sentences set S_r, and n_r is the total number of relations.

$$P(r|S_r, \theta) = \frac{\exp(\mathbf{o}_r)}{\sum_{c=1}^{n_r}\exp(\mathbf{o}_c)} \tag{8}$$

$$\mathbf{o} = \mathbf{M}\mathbf{s}_{final} + \mathbf{d} \tag{9}$$

3.3 Optimization

The base model defines the optimization function as the log-likelihood of the objective function in Equation 10.

$$P(G, D|\theta) = P(G|\theta_{\mathcal{E}}, \theta_{\mathcal{R}}) + P(D|\theta_S) \quad (10)$$

where, G and D are KG and textual data respectively. The base model applies Stochastic Gradient Descent (SGD) and L_2 regularization. In practice, the base model optimizes the KG Encoding Part and Sentence Encoding Part in parallel.

4 Proposed Model

As discussed before, the sentences containing the entity pairs of interest tend to omit the BK that the authors assume is known to the readers. However, the omitted BK would be extremely important for a machine to identify the relation between the entity pairs. To fill the "gaps" and improve the efficacy of DS-RE, we assume that the reasoning paths between the entity pairs over a KG could be utilized as BK to compensate for the brevity of the sentences. Motivated by this issue, we propose the DS-RE model that integrates both reasoning paths and sentences.

4.1 Architecture

The proposed model consists of three parts: KG Encoding Part, Sentence Encoding Part and Path Encoding Part, as shown in Figure 3. The KG Encoding Part and Sentence Encoding Part are identical to the base model, except that the final input to the relation classification layer. The Path Encoding Part takes as input a set of reasoning paths, $P_r = \{p_1, ..., p_m\}$, between two entities of interest (e_1, e_2), and encodes them into the final representation of paths, $\mathbf{p}_{final}$. Specifically, let $p = \{e_1, r_1, e_{r_1}, r_2, e_{r_2}, ..., r_i, e_{r_i}..., e_2\}$ denote a path between (e_1, e_2). To express the semantic meaning of a relation in a path, we represent r_i by its component words, rather than treat it as an unit. Therefore, a path will be represented as $p = \{e_1, w_1^{r_1}, w_2^{r_1}, ..., e_{r_1}, , w_1^{r_2}, w_2^{r_2}, ..., e_{r_2}, ..., e_2\}$, where $w_2^{r_1}$ denotes the second word of r_1 (e.g., $treat$ in may_treat relation).

Since a path is represented as a sequence of words, or a special sentence, we apply the similar CNN model used in the Sentence Encoding Part

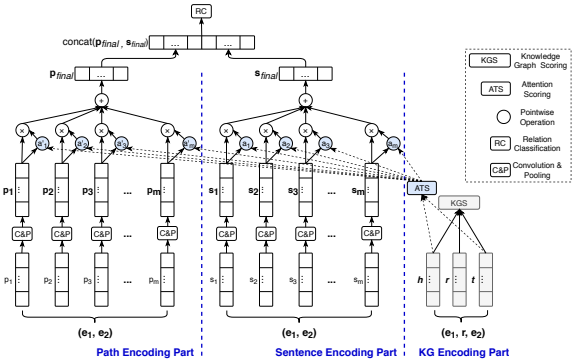

Figure 3: Overview of the proposed model.

to encode the path into vector representation $\mathbf{p}_i$. The Path Encoding Part and Sentence Encoding Part share the word embedding projection matrix $\mathbf{W}_{emb}^{w}$, and word position projection matrix $\mathbf{W}_{emb}^{wp}$ in Equation 5 except the convolutional kernal $\mathbf{W}$ and its corresponding bias vector $\mathbf{b}$ in Equation 6. To utilize evidence from all the paths between target entity pair, we also adopt the KG-based attention mechanism applied in Sentence Encoding Part to calculate the final representation of paths $\mathbf{p}_{final}$. We calculate $\mathbf{p}_{final}$ via Equation 11, where $\mathbf{W}_s$ is the weight matrix, $\mathbf{b}_s$ is the bias vector, a'_i is the weight for $\mathbf{p}_i$, which is the distributed representation for the i-th path in P_r.

$$\mathbf{p}_{final} = \sum_{i=1}^{m} a'_i \mathbf{p}_i, \quad (11)$$

$$a'_i = \frac{\exp(\langle \mathbf{r}_{ht}, \mathbf{x'}_i \rangle)}{\sum_{k=1}^{m} \exp(\langle \mathbf{r}_{ht}, \mathbf{x'}_k \rangle)},$$

$$\mathbf{x'}_i = \tanh(\mathbf{W}_s \mathbf{p}_i + \mathbf{b}_s)$$

Finally, we concatenate the resulting representation $\mathbf{s}_{final}$ and $\mathbf{p}_{final}$ for S_r (the set of input sentences) and P_r (the set of reasoning paths) respectively as the input to the relation classification layer. The conditional probability $P(r|S_r, P_r, \theta_S, \theta_P)$ is formulated via Equation 12 and Equation 13, where, θ_P is the parameters in Path Encoding Part, $\mathbf{M}$ is the representation matrix of relations, $\mathbf{d}$ is a bias vector, $\mathbf{o}$ is the output vector containing the prediction probabilities of all target relations for both input sentences set S_r and input paths set P_r. n_r is the total number of relations.

$$P(r|S_r, P_r, \theta_S, \theta_P) = \frac{\exp(\mathbf{o}_r)}{\sum_{c=1}^{n_r} \exp(\mathbf{o}_c)} \quad (12)$$

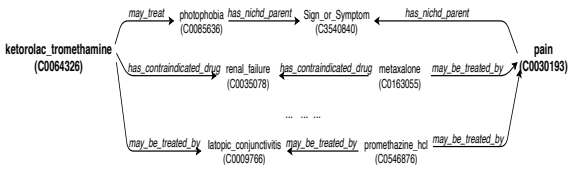

Figure 4: Multiple reasoning paths between *ketorolac_tromethamine* and *pain*.

$$\mathbf{o} = \mathbf{M}[\mathbf{s}_{final}; \mathbf{p}_{final}] + \mathbf{d} \tag{13}$$

Similar to the base model, we define the optimization function as the log-likelihood of the objective function in Equation 14.

$$P(G, D|\theta) = P(G|\theta_{\mathcal{E}}, \theta_{\mathcal{R}}) + P(D|\theta_S, \theta_P) \tag{14}$$

4.2 Reasoning Paths Generation

Let (e_1, e_2) be an entity pair of interest. The set of reasoning paths P_r is obtained by computing all shortest paths in a KG starting from e_1 till e_2. For simulating the situation where the direct relation between a target entity pair is unavailable in a sparse KG, we remove the triplet that directly connect the target entity pair of interest from the KG. Each reasoning path, thus, is at least a two-hop path, namely $p = \{e_1, r_1, e_{r_1}, r_2, e_2\}$. However, if the shortest path is not found due to the sparsity of KG, we will use a padding path to represent the missing path $p = \{r_{padding}\}$. Figure 4 shows the generated paths between *ketorolac_tromethamine* and *pain*.

5 Experiments

Our experiments aim to demonstrate the effectiveness of the proposed model, which is discussed in Section 4, for DS-RE from biomedical datasets.

5.1 Data

The biomedical datasets used for evaluation consist of knowledge graph, textual data and reasoning path, which will be detailed as follows.

Knowledge Graph. We choose the UMLS as the KG. UMLS is a large biomedical knowledge base developed at the U.S. National Library of Medicine. UMLS contains millions of biomedical concepts and relations between them. We follow (Wang et al., 2014), and only collect the fact triplet with RO relation category (RO stands for "has Relationship Other than synonymous, narrower, or broader"), which covers the interesting relations such as *may_treat* and *my_prevent*. From the UMLS 2018 release, we extract about 50 thousand such RO fact triplets (i.e., (e_1, r, e_2)) under the restriction that their entity pairs (i.e., e_1 and e_2) should coexist within a sentence in Medline corpus. They are then randomly divided into training and testing sets for KGC. Following (Weston et al., 2013), we keep high entity overlap between training and testing set, but zero fact triplet overlap. The statistics of the extracted KG is shown in Table 1.

#Entity	#Relation	#Train (triplet)	#Test (triplet)
16,049	295	34,378	12,502

Table 1: Statistics of KG in this work.

Textual Data. Medline corpus is a collection of bimedical abstracts maintained by the National Library of Medicine. From the Medline corpus, by applying the UMLS entity recognizer, QuickUMLS (Soldaini and Goharian, 2016), we extract $682,093$ sentences that contain UMLS entity pairs as our textual data, in which $485,498$ sentences are for training and $196,595$ sentences for testing. For identifying the NA relation, besides the "related" sentences, we also extract the "unrelated" sentences based on a closed world assumption: pairs of entities not listed in the KG are regarded to have NA relation and sentences containing them considered to be the "unrelated" sentences. By this way, we extract $1,394,025$ "unrelated" sentences for the training data, and $598,154$ "unrelated" sentences for the testing data. Table 2 presents some sample sentences in the training data.

Reasoning Path. Following the Section 4.2, we extract $197,396$ paths for not NA triplets ($139,224$ / $58,172$ for training / testing) and $679,408$ for NA triplets ($474,263$ / $205,145$ for training / testing), under the restriction that each entity in a path should be observed in Medline corpus.

5.2 Parameter Settings

We base our work on (Han et al., 2018) and its implementation available at `https://github.com/thunlp/JointNRE`, and thus adopt identical optimization process. We use the default settings

Fact Triplet	Textual Data
(insulin, gene_product_plays_role_in_biological_process, energy_expenditure)	s_1 : *These results indicate that hyperglucagonemia during* $\underline{insulin}_{e_1}$ *deficiency results in an increase in* $\underline{energy\_expenditure}_{e_2}$*, which may contribute to the catabolic_state in many conditions.* s_2 : *It was hypothesized that the waxy maize treatment would result in a blunted and more sustained glucose and* $\underline{insulin}_{e_1}$ *response, as well as* $\underline{energy\_expenditure}_{e_2}$ *and appetitive responses.* s_3 : ...
(IRI, NA, insulin)	s_1 : *Plasma insulin immunoreactivity (*$\underline{IRI}_{e_1}$*) results from high molecular weight substances with insulin immunoreactivity (HWIRI), proinsulin (PI) and* $\underline{insulin}_{e_2}$ *(I).* s_2 : *The beads method demonstrated high* $\underline{IRI}_{e_1}$ *values in both* $\underline{insulin}_{e_2}$ *fractions and the fractions containing serum_proteins bigger than 40,000 molecular weight.* s_3 : ...

Table 2: Examples of textual data extracted from Medline corpus.

of parameters [2] provided by the base model. Since we address the DS-RE in biomedical domain, we use the Medline corpus to train the domain specific word embedding projection matrix $\mathbf{W}^{w}_{emb}$ in Equation 5.

5.3 Result and Discussion

We investigate the effectiveness of our proposed model with respect to enhancing the DS-RE from biomedical datasets. We follow (Mintz et al., 2009; Weston et al., 2013; Lin et al., 2016; Han et al., 2018) and conduct the held-out evaluation, in which the model for DS-RE is evaluated by comparing the fact triplets identified from textual data (i.e., the set of sentences containing the target entity pairs) with those in KG. Following the evaluation of previous works, we draw Precision-Recall curves and report the micro average precision (AP) score, which is a measure of the area under the Precision-Recall curve (higher is better), as well as Precision@N (P@N) metrics, which gives the percentage of correct triplets among top N ranked candidates.

Precision-Recall Curves. The Precision-Recall (PR) curves are shown in Figure 5, where "CNN+AVE" represents that the DS-RE model uses the average vector of sentences as $\mathbf{s}_{final}$ in Equation 7. "JointE+KATT" (or "JointD+KATT") represents that the DS-RE model applies Prob-TransE (or Prob-TransD) as its KG Encoding Part for attention calculation. "(TEXT)" indicates that the model only takes the textual data as input (i.e., the set of sentences containing target entity pairs). "(PATH)" indicates the DS-RE model only takes the reasoning paths between entity pairs as its input. "(TEXT+PATH)" indicates the DS-RE model takes both the textual data and reasoning paths as its input.

[2] As a preliminary study, we only adopt the default hyperparameters, but we will tune them for our task in the furture.

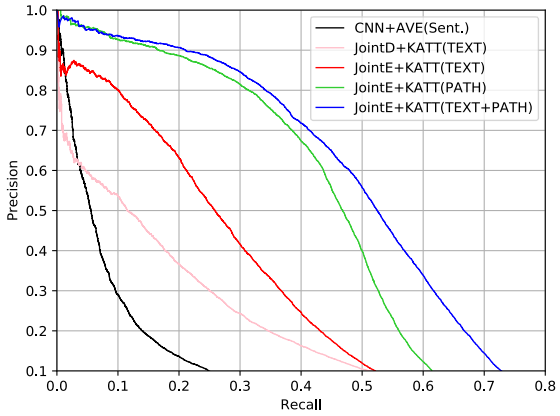

Figure 5: Precision-Recall curves.

The results show that:

(1) The proposed model (i.e., "JointE+KATT(PATH+TEXT)") significantly outperform the base model (i.e., "JointE+KATT(TEXT)"), proving that reasoning paths are useful BK for biomedical DS-RE. This result inspires us to explore other reasoning strategy such as by reasoning across multiple documents. (2) "JointE+KATT(PATH+TEXT)" achieves better overall performance than "JointE+KATT(PATH)", demonstrating the mutual complementary relationship between the sentences containing entity pairs and the reasoning paths between them. Specifically, on the one hand, as discussed in Section 1, reasoning paths could provide BK for interpreting the implicitly expressed relation in sentences. On the other hand, due to the sparsity of KG, it is by no means certain that all entity pairs are fully connected by plausible reasoning paths in the KG. In that case, the sentences could provide the informative evidence to identify the relation between them.

AP and P@N Evaluation. The results in terms of P@1k, P@2k, P@3k, P@4k, P@5k, the mean of them and AP are shown in Table 3. From the table, we have similar observation to the PR curves: (1) The proposed model (i.e., "JointE+KATT(TEXT+PATH)") significantly outperforms the base model for all measures. (2) "JointE+KATT(TEXT+PATH)" outperforms "JointE+KATT(PATH)" in most of the metrics.

Model	P@1k	P@2k	P@3k	P@4k	P@5k	Mean	AP
CNN+AVE	0.852	0.751	0.685	0.640	0.602	0.706	0.098
JointD+KATT(TEXT)	0.628	0.614	0.552	0.495	0.446	0.547	0.186
JointE+KATT(TEXT)	0.835	0.759	0.692	0.629	0.564	0.696	0.272
JointE+KATT(PATH)	**0.945**	0.911	0.881	0.842	0.796	0.875	0.432
JointE+KATT(TEXT+PATH)	0.941	**0.922**	**0.897**	**0.865**	**0.818**	**0.889**	**0.496**

Table 3: P@N and AP for different RE models, where k=1000.

Case Study. Table 4 shows the comparison of the attention distribution between "JointE+KATT(TEXT)" (Base) and "JointE+KATT(TEXT+PATH)" (Proposed). The first and second columns represent the attention distribution (the highest and the lowest) over input sentences. From the Table 4, we can see that the proposed model that incorporates reasoning paths is more capable of selecting informative sentences than the base model, because it "focuses" on the second sentence that explicitly describes the may_treat relation via the word "reduction", in contrast, the base model "ignores" such informative sentence. Table 5 shows the attention allocated by our proposed model for given reasoning paths. The first path generally means if two chemicals should not be used in the case of (or contraindicated with) drug_allergy, they will treat lung_tumor. In contrast, the second path generally means if two chemicals treat Histiocytoses (an excessive number of cells), they will also treat lung_tumor. Apparently the second one that our proposed model focused on is more plausible. This indicates that our proposed model has the capacity of identifying the plausible reasoning path.

6 Conclusion and Future Work

In this work, we tackle the task of DS-RE from biomedical datasets. However, the biomedical DS-RE could be negatively affected by the brevity of text. Specifically, authors always omit the BK that would be important for a machine to identify relationships between entities. To address this issue, in this work, we assume that the reasoning paths over a KG could be utilized as the BK to fill the "gaps" in text and thus facilitate DS-RE. Experimental results prove the effectiveness of the combination, because our proposed model achieves significant and consistent improvements as compared with a state-of-the-art DS-RE model. Although the reasoning paths over KG are useful for DS-RE, the sparsity of KG would hinder their effectiveness. Therefore, in the future, beside the reasoning paths over KG, we will also utilize the reasoning paths across multiple documents for our task. For instance, reasoning across Document1 and Document2, shown below, would facilitate the relation identification between "Aspirin" and "inflammation".

Document1: "***Aspirin*** *and other* ***nonsteroidal anti-inflammatory drugs*** *(NSAID) show ...*"

Document2: "***Nonsteroidal anti-inflammatory drugs*** *reduce* ***inflammation*** *by ...*"

Base	Proposed	Sentences for (**Mitomycin_C (MCC)**, may_treat, **stomach/gastric_tumor**)
High	Low	The additive effect in the combination of TNF and Mitomycin_C was observed against two **Mitomycin_C** resistant **gastric_tumors**.
Low	High	One-quarter or one-half maximum tolerated doses (MTDs) of 5-FU or **MMC** resulted in a significant reduction of **stomach_tumor** growth, ...

Table 4: Comparison of attention between base model and proposed model, where High (or Low) represents the highest (or lowest) attention.

Attention	Paths for (**etoposide**, may_treat, **lung_tumor**)
Low	**etoposide** $\overleftarrow{has\_contraindicated\_drug}$ drug_allergy $\overrightarrow{has\_contraindicated\_drug}$ S-Liposomal Doxorubicin $\overrightarrow{may\_treat}$ **lung_tumor**
High	**etoposide** $\overrightarrow{may\_be\_treated\_by}$ Histiocytoses $\overleftarrow{may\_be\_treated\_by}$ Vinblastine $\overrightarrow{may\_treat}$ **lung_tumor**

Table 5: Some examples of attention distribution over reasoning paths from "JointE+KATT(TEXT+PATH)".

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR1513, Japan and KAKENHI Grant Number 16H06614.

References

Waleed Ammar, Matthew Peters, Chandra Bhagavatula, and Russell Power. 2017. The ai2 system at semeval-2017 task 10 (scienceie): semi-supervised end-to-end entity and relation extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 592–596.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2216–2225.

Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2017. Chemical-induced disease relation extraction via convolutional neural network. *Database*, 2017.

Gus Hahn-Powell, Dane Bell, Marco A Valenzuela-Escárcega, and Mihai Surdeanu. 2016. This before that: Causal precedence in the biomedical domain. *arXiv preprint arXiv:1606.08089*.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 687–696.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2124–2133.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–43.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782.

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 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-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. *arXiv preprint arXiv:1601.00770*.

Randall Munroe. 2013. The rise of open access. *Science*, 342(6154):58–59.

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.

RA Roller, E Agirre, A Sorora, and M Stevenson. 2015. Improving distant supervision using inference learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th*

International Joint Conference on Natural Language Processing. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580*.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*.

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 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, volume 14, pages 1112–1119.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. *arXiv preprint arXiv:1307.7973*.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015. Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv preprint arXiv:1506.07650*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

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 *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212.