# Distantly Supervised Biomedical Knowledge Acquisition via Knowledge Graph Based Attention

**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 in extracting scientific relation from the ever growing scientific publications. Distant supervision is widely applied approach to automatically generate large amounts of labelled data for Relation Extraction (RE). However, distant supervision inevitably accompanies the wrong labelling problem, which will negatively affect the RE performance. To address this issue, (Han et al., 2018) proposes a novel framework for jointly training RE model and Knowledge Graph Completion (KGC) model to extract structured knowledge from non-scientific dataset. In this work, we firstly investigate the feasibility of this framework on scientific dataset, specifically on biomedical dataset. Secondly, to achieve better performance on the biomedical dataset, we extend the framework with other competitive KGC models. Moreover, we proposed a new end-to-end KGC model to extend the framework. Experimental results not only show the feasibility of the framework on the biomedical dataset, but also indicate the effectiveness of our extensions, because our extended model achieves significant and consistent improvements on distantly supervised RE as compared with baselines.

## 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), Relation Extraction (RE), etc. 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, as with general KGs such as Freebase (Bollacker et al., 2008) and DBpedia (Lehmann et al., 2015), scientific KGs 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.

In 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 assumes that if two entities are connected by a relation in a KG, then all sentences that contain these entity pairs will express the relation. For instance, (*aspirin*, *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 evidences to support the *may_treat* relationship between *aspirin* and *pain*, but will still be annotated as positive instances by the distant supervision.

(1) *The clinical manifestations are generally typical nocturnal* ***pain*** *that prevents sleep and that is alleviated with* ***aspirin***.

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

(2) *The tumor was remarkably large in size , and* ***pain*** *unrelieved by* ***aspirin****.*

(3) *The level of* ***pain*** *did not change significantly with either* ***aspirin*** *or pentoxifylline , but the walking distance was farther with the pentoxifylline group .*

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. Despite the impressive achievement in RE, this model still has the limitation that it only selects the most informative sentence and ignores the rest, thereby loses the rich information stored in those neglected sentences, For instance, among Example 1, Example 2 and Example 3, Example 1 is undoubtedly the most informative one for detecting relation *may_treat*, but it unnecessarily means other sentences such as Example 3 could not contribute to the relation detection. In Example 3, entity *aspirin* and entity *pentoxifylline* have alternative relation, and the latter is a drug to treat muscle pain, therefore the former is also likely to be a pain-killing drug. To address this issue, recently, attention mechanism is applied to extract features from all collected sentences. (Lin et al., 2016) proposes a relation vector based attention mechanism for distantly supervised RE. (Han et al., 2018) proposes a novel joint model that leverages the KG-based attention mechanism and achieves better performance than (Lin et al., 2016) on distantly supervised RE from New York Times (NYT) corpus.

The success that the joint model (Han et al., 2018) has attained in the newswire domain (or non-scientific domain) inspires us to choose the strong model as our base model and assess its feasibility on biomedical domain. Specifically, the first question of this research is how the joint model behaves when the system is trained on biomedical KG (e.g., UMLS) and biomeical corpus (e.g., Medline corpus). (Han et al., 2018) indicates that the performance of the base model could be affected the representation ability of KGC model. The representation ability of a KGC model also varies with dataset (Wang et al., 2017). Therefore, given a new dataset (e.g., a biomedical dataset), it is necessary to extend the base model with other competitive KGC models, and choose the best fit for the given dataset. However, the base model only implements two KGC models, which are based on TransE (Bordes et al., 2013) and TransD (Ji et al., 2015) respectively. Thus, the second question of this work is how other competitive KGC models such as ComplEx (Trouillon et al., 2016) and SimplE (Kazemi and Poole, 2018) influence the performance of the base model on biomedical dataset. At last but not least, in biomedical KG, a relation is scientifically restricted by entity type (ET). For instance, in the relation ($h$, *may_treat*, $t$), the ET of $t$ should be `Disease or Syndrome`. Therefore, ET information is an important feature for biomedical RE and KGC. For leveraging the ET information, which the base model lacks, in this work, we propose an end-to-end KGC model to enhance the base model. The proposed KGC model is capable of identifying ET via the word embedding of target entity and incorporating the predicted ET into a state-of-to-art KGC model to evaluate the plausibility of potential fact triplets.

We conduct evaluation on biomedical datasets in which KG is collected from UMLS and textual data is extracted from Medline corpus. The experimental results not only show the feasibility of the base model on the biomedical domain, but also prove the effectiveness of our proposed extensions for the base model.

## 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 distant supervision for 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.,*aspirin* 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, (Lin et al., 2016) proposes a relation vector based attention mechanism to extract feature from the entire bag and outperforms the prior approaches. (Han et al., 2018) proposes a joint model that adopts a KG-based attention mechanism and achieves better performance than (Lin et al., 2016) on distantly supervised RE from NYT corpus.

In this work, we are primarily interested in applying distant supervision techniques to extract biomedical fact triplets from scientific publications. To validate and enhance the efficacy of the previous techniques in biomedical domain, we choose the strong joint model proposed by (Han et al., 2018) as the base model and make some necessary extension for our scientific RE task. Since from the two main groups of KGC models (Wang et al., 2017): translational distance models and semantic matching models, the base model only implements the translational distance models, TransE (Bordes et al., 2013) and TransD (Ji et al., 2015), we thus extend the base model with the semantic matching models, ComplEx (Trouillon et al., 2016) and SimplE (Kazemi and Poole, 2018), for selecting the best fit for our task. In addition, the base model has not incorporated the ET information, which we assume is crucial for scientific RE. Therefore, we propose an end-to-end KGC model to enhance the base model. Different from the work (Xie et al., 2016), which utilizes an ET look-up dictionary to obtain ET, the end-to-end KGC is capable of identifying ET via the word embedding of a target entity and thus is free of the attachment to an incomplete ET look-up dictionary.

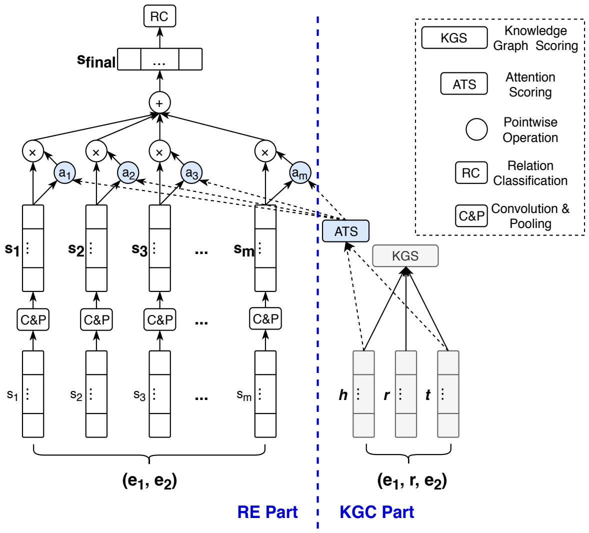

Figure 1: Overview of the base model.

## 3 Base Model

The architecture of the base model is illustrated in Figure 1. In this section, we will introduce the base model proposed by (Han et al., 2018) in two main parts: KGC part, RE part.

### 3.1 KGC 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. KGC model 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. As mentioned above, the base model adopts two representative translational distance models Prob-TransE and Prob-TransD, which are based on TransE (Bordes et al., 2013) and TransD (Ji et al., 2015) repectively, to score a 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.

$$\mathbf{r}_{ht} = \mathbf{t}_r - \mathbf{h}_r, \tag{2}$$

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

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 the one 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 RE 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* $\underline{\textit{pain}}_{e_2}$ *and* $\underline{\textit{aspirin}}_{e_1}$ *consumption in a daily diary"*, the relative distance of the word *"and"* is [1, -1].

$$\mathbf{v}_t = [\mathbf{v}^w_t; \mathbf{v}^{wp1}_t; \mathbf{v}^{wp2}_t], \tag{5}$$

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

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 use 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 $P(r|S_r, \theta)$ is formulated via Equation 8 and Equation 9, where, $\theta$ is the parameters for RE, which includes $\{\mathbf{W}^w_{emb}, \mathbf{W}^{wp}_{emb}, \mathbf{W}, \mathbf{b}, \mathbf{W}_s, \mathbf{b}_s, \mathbf{M}, \mathbf{d}\}$, $\mathbf{M}$ 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.

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

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

## 4 Extensions

The base model opens the possibility to jointly train RE models with KGC models for distantly supervised RE. The empirical results of the base model on NYT corpus indicate that the performance of distantly supervised RE varies with KGC models (Han et al., 2018). In addition, the performance of KGC models depends on a given dataset (Wang et al., 2017). Therefore, we assume that it is necessary to attempt multiple competitive KGC models for the joint framework so as to find the optimal combination for our biomedical dataset. However, the base model only implements translational distance models: TransE and TransD, but not the semantic matching models, and this, we assume, might hinder its performance in the new dataset. To address this, we select two representative semantic matching models: ComplEx (Trouillon et al., 2016) and SimplE (Kazemi and Poole, 2018) as the alternative KGC part.

As discussed in Section 1, in scientific KGs, a fact triplet is severely restricted by ET information (e.g., ET of $e_2$ should be `Disease or Syndrome` in the fact triplet $(e_1, may\_treat, e_2)$). Therefore, for leveraging ET information, which the base model lacks, we also propose an end-to-end KGC model to extend the base model. Since the proposed KGC model is build on SimplE and is capable of Named Entity Recognition (NER), we call it SimplE_NER.

### 4.1 ComplEx based Attention

Given a fact triplet $(e_1, r, e_2)$, ComplEx then encodes entities $e_1$, $e_2$ and relation $r$ into a complex-valued vector $\mathbf{e}_1 \in C^d$, $\mathbf{e}_2 \in C^d$ and $\mathbf{r} \in C^d$ respectively, where $d$ is the dimensionality of the embedding space. Since entities and relations are represented as complex-valued vector, each $\mathbf{x} \in C^d$ consists of a real vector component $Re(\mathbf{x})$ and imaginary vector component $Im(\mathbf{x})$, namely $\mathbf{x} = Re(\mathbf{x}) + iIm(\mathbf{x})$. The KG scoring function of ComplEx for a fact triplet $(e_1, r, e_2)$ is calculated via Equation 10, where $\bar{\mathbf{e}}_2$ is the conjugate of $\mathbf{e}_2$; $Re(\cdot)$ (or $Im(\cdot)$) means taking the real (or imaginary) part of a complex value. $\langle u, v, w\rangle$ is defined via Equation 11, where $[\cdot]_n$ is the $n$-th entry of a vector.

$$\begin{aligned} f_r(e_1, e_2) = Re(\langle \mathbf{e}_1, \mathbf{r}, \bar{\mathbf{e}}_2\rangle) = \\ \langle Re(\mathbf{r}), Re(\mathbf{e}_1), Re(\mathbf{e}_2)\rangle \\ +\langle Re(\mathbf{r}), Im(\mathbf{e}_1), Im(\mathbf{e}_2)\rangle \\ +\langle Im(\mathbf{r}), Re(\mathbf{e}_1), Im(\mathbf{e}_2)\rangle \\ -\langle Im(\mathbf{r}), Im(\mathbf{e}_1), Re(\mathbf{e}_2)\rangle \end{aligned} \tag{10}$$

$$\langle \mathbf{u}, \mathbf{v}, \mathbf{w}\rangle = \sum_{n=1}^{d} [\mathbf{u}]_n [\mathbf{v}]_n [\mathbf{w}]_n \tag{11}$$

Since the asymmetry of this scoring function, namely $f_r(e_1, e_2) \neq f_r(e_2, e_1)$, ComplEx can effectively encode asymmetric relations (Trouillon et al., 2016). For calculating the attention, the $\mathbf{r}_{ht}$ in Equation 7 is defined via Equation 12, where $\odot$ represents the element-wise multiplication.

$$\mathbf{r}_{ht} = Re(\mathbf{e}_1) \odot Re(\mathbf{e}_2) + Im(\mathbf{e}_1) \odot Im(\mathbf{e}_2) \tag{12}$$

### 4.2 SimplE based Attention

Given a fact triplet $(e_1, r, e_2)$, SimplE then encodes each entity $e \in \mathcal{E}$ into two vectors $\mathbf{h}_e$, $\mathbf{t}_e \in R^d$ and each relation $r \in \mathcal{R}$ into two vectors $\mathbf{v}_r$, $\mathbf{v}_{r^{-1}} \in R^d$ respectively, where $d$ is the dimensionality of the embedding space. $\mathbf{h}_e$ captures the entity $e$'s behaviour as the *head entity* of a fact triplet and $\mathbf{t}_e$ captures $e$'s behaviour as the *tail entity*. $\mathbf{v}_r$ represents $r$ in a fact triplet $(e_1, r, e_2)$, while $\mathbf{v}_{r^{-1}}$ represents its inverse relation $r^{-1}$ in the triplet $(e_2, r^{-1}, e_1)$. The KG scoring function of SimplE for a fact triplet $(e_1, r, e_2)$ is defined via Equation 13.

$$f_r(e_1, e_2) = \frac{1}{2}(\langle \mathbf{h}_{e_1}, \mathbf{v}_r, \mathbf{t}_{e_2}\rangle + \langle \mathbf{h}_{e_2}, \mathbf{v}_{r^{-1}}, \mathbf{t}_{e_1}\rangle) \tag{13}$$

Similar to the attention from ComplEx, the $\mathbf{r}_{ht}$ in Equation 7 is defined via Equation 14.

$$\mathbf{r}_{ht} = \frac{1}{2}(\mathbf{h}_{e_1} \odot \mathbf{h}_{e_2} + \mathbf{t}_{e_1} \odot \mathbf{t}_{e_2}) \tag{14}$$

### 4.3 SimplE_NER based Attention

The proposed end-to-end KGC model is based on SimplE, because SimplE outperforms several state-of-the-art models including ComplEx (Kazemi and Poole, 2018). The proposed model is illustrated in Figure 2. It includes ET classification part (below) and KG Scoring part (above). In ET classification part, a multi-layer perceptron (MLP) with two hidden layers are applied to identify ET based on word embedding of target entity. In KG Scoring part, *head entity* and *tail entity* along with their predicted ETs and their relation are projected into corresponding KG embeddings, which are then fed to a KG scoring function.

**ET Classification Part.** In this work, we use a MLP network to classify ET for *head entity* and *tail entity*. The architecture of our MLP network is as bellow:

$$\begin{aligned} \mathbf{h}_w &= \tanh(\mathbf{W}^w_{emb}\mathbf{x}^w), \\ \mathbf{h}_1 &= \mathrm{sigmoid}(\mathbf{W}_1\mathbf{h}_w + \mathbf{b}_1), \\ \mathbf{h}_2 &= \mathrm{sigmoid}(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2), \\ \mathbf{y} &= \mathrm{sigmoid}(\mathbf{W}_{ET}\mathbf{h}_2 + \mathbf{b}_{ET}) \end{aligned} \tag{15}$$

where $\mathbf{W}^w_{emb}$ is a word embedding projection matrix, which is initialized by the pre-trained word

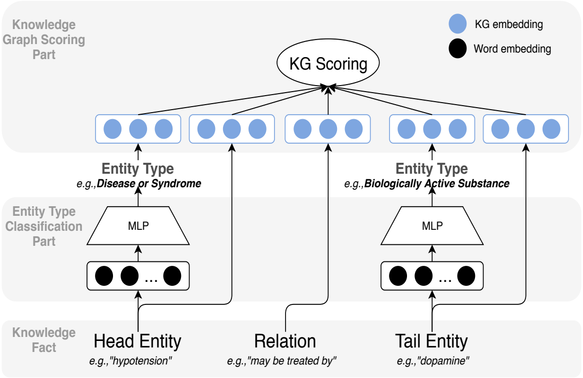

Figure 2: Overview of the proposed end-to-end KGC model.

embedding that is trained on Medline corpus via Gensim word2vec tool, $\mathbf{x}^w$ is a one-hot entity representation, $\mathbf{y}$ is the output vector containing the prediction probabilities of all target ETs. $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$, $\mathbf{b}_2$, $\mathbf{W}_{ET}$ and $\mathbf{b}_{ET}$ are parameters to optimize.

**KG Scoring Part.** Given fact triplet and predicted ET pair $ET_1$ (for $e_1$) and $ET_2$ (for $e_2$), the proposed model project them into their corresponding KG embeddings namely $\mathbf{h}_{e_1}$, $\mathbf{t}_{e_1}$, $\mathbf{v}_r$, $\mathbf{v}_{r^{-1}}$, $\mathbf{h}_{e_2}$, $\mathbf{t}_{e_2}$, $\mathbf{h}_{ET_1}$, $\mathbf{t}_{ET_1}$, $\mathbf{h}_{ET_2}$ and $\mathbf{t}_{ET_2}$ respectively, where $\mathbf{h}_{ET_1}$ (or $\mathbf{t}_{ET_1}$) represents the KG embedding of ET for $e_1$ when $e_1$ acts as the *head entity* (or *tail entity*) in a fact triplet. The KG scoring function is defined via Equation 16. Since the proposed KGC model is build on SimplE, we apply Equation 14 to calculate $\mathbf{r}_{ht}$.

$$\begin{aligned} f_r(e_1, e_2) = \frac{1}{4}(&\langle \mathbf{h}_{e_1}, \mathbf{v}_r, \mathbf{t}_{e_2} \rangle \\ &+ \langle \mathbf{h}_{e_2}, \mathbf{v}_{r^{-1}}, \mathbf{t}_{e_1} \rangle \\ &+ \langle \mathbf{h}_{ET_1}, \mathbf{v}_r, \mathbf{t}_{ET_2} \rangle \\ &+ \langle \mathbf{h}_{ET_2}, \mathbf{v}_{r^{-1}}, \mathbf{t}_{ET_1} \rangle) \end{aligned} \tag{16}$$

## 5 Experiments

Our experiments aim to demonstrate that, (1) the base model proposed by (Han et al., 2018) is feasible for biomedical dataset, such as UMLS and Medline corpus, and (2) in order to improve the performance on the given biomedical dataset, it is necessary to extend the base model with other competitive KGC models, such as ComplEx and SimplE, and (3) the proposed end-to-end KGC model is effective for distantly supervised RE from biomedical dataset.

| #Entity | #Relation | #Train | #Test |
|---|---|---|---|
| 25,080 | 360 | 53,036 | 11,810 |

Table 1: Statistics of KG in this work.

### 5.1 Data

The biomedical datasets used for evaluation consist of biomedical knowledge graph and biomedical textual data, 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 like *may_treat*, *my_prevent*, etc. From the UMLS 2018 release, we extract about 60 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. For training the ET Classification Part in Section 4.3, we also collect about 35 thousand entity-ET pairs (e.g., *heart rates*-`Clinical Attribute`) from the UMLS 2018 release.

**Textual Data.** Medline corpus is a collection of bimedical abstracts maintained by the National Library of Medicine. From the Medline corpus, by applying a string matching model [2], we extract $732,771$ sentences that contain the entity pairs (i.e., $e_1$ and $e_2$) in the KG mentioned above as our textual data, in which $592,605$ sentences are for training and $140,166$ 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 $1,738,801$ "unrelated" sentences for the training data, and $431,212$ "unrelated" sentences for the testing data. Table 2 presents some

[2] We adopt the NER model that is available at `https://github.com/mpuig/spacy-lookup`.

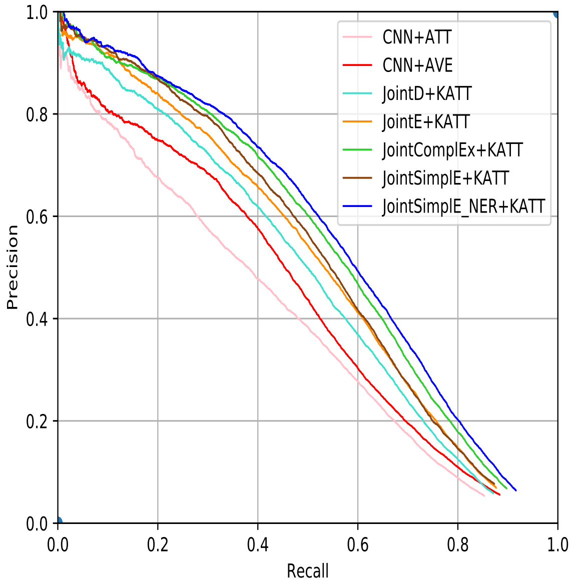

Figure 3: Aggregate precision/recall curves for different RE models.

sample sentences in the training data.

### 5.2 Parameter Settings

We base our work on (Han et al., 2018) and extend their implementation available at `https://github.com/thunlp/JointNRE`, and thus adopt identical optimization process. We use the default settings of parameters [3] provided by the base model. Since we address the distantly supervised RE in biomedical domain, we use the Medline corpus to train the domain specific word embedding projection matrix $\mathbf{W}^w_{emb}$.

### 5.3 Result and Discussion

(Han et al., 2018) evaluates the base model on non-scientific dataset. In this work, we firstly plan to assess its feasibility on scientific dataset, and secondly, to investigate the effectiveness of our extensions, which is discussed in Section 4, with respect to enhancing the distantly supervised RE from scientific dataset.

**Relation Extraction** 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 distantly supervised RE is evaluated by comparing the fact triplets identified from textual data (i.e., the bag of sentences containing the target entity pairs) with those in KG. We report precision-recall curves and Precision@N (P@N) as well in our evaluation.

The precision-recall curves are shown in Figure 3, where "JointD+KATT" and "JointE+KATT" represent the RE model with the KG-based attention obtained from Prob-TransD and Prob-TransE respectively, which are our base models and trained on both KG and textual data. Similarly, "JointComplEx+KATT", "JointSimplE+KATT" and "JointSimplE_NER+KATT" represent the RE model with the KG-based attention obtained from ComplEx, SimplE and SimplE_NER respectively, which are our extensions. "CNN+AVE" and "CNN+ATT" represent the RE model with average attention and relation vector based attention (Lin et al., 2016) respectively, which are not joint models and only trained on textual data. The results show that:

(1) All RE models with KG-based attention, such as "JointE+KATT", outperform those models without it, such as "CNN+ATT". This observation is in line with (Han et al., 2018). This demonstrates that not just for non-scientific dataset , jointly training a KGC model with a RE model is also an effective approach to improve the performance of distantly supervised RE for biomedical dataset. In other words, the outperformance proves the feasibility of the base model proposed by (Han et al., 2018) on biomedical dataset. The comparison between (Han et al., 2018)'s results on non-scientific dataset and ours on scientific dataset also indicates that the performance of base model could differ according to the dataset. Specifically, on scientific dataset, "JointE+KATT" performs better than "JointD+KATT" but in non-scientific dataset the latter outperforms the former.

(2) Our extended models, "JointComplEx+KATT", "JointSimplE+KATT" and "JointSimplE_NER+KATT", achieve better precision than the base model over the major range of recall. It could be attributed to their better capability of modeling asymmetric relations (e.g., $may\_treat$ and $may\_prevent$), because their KG scoring functions are asymmetry (i.e., $f_r(e_1, e_2) \neq f_r(e_2, e_1)$). The superior performance indicates the necessity of our extensions on the base model. Specifically, given the frequently used biomedical dataset, UMLS and Medline corpus, it would be an effective method to switch the translational distance models, such as TransE and TransD, with the semantic matching models,

[3] As a preliminary study, we only adopt the default hyper-parameters, but we will tune them in the furture.

| **Fact Triplet** | **Textual Data** |
|---|---|
| (insulin,<br>gene_plays_role_in_process,<br>lipid_metabolism) | $s_1$ : *It is unknown whether short - term angiotensin_receptor blocker therapy can improve glucose and lipid_metabolism*$_{e_2}$ *in insulin*$_{e_1}$ *- resistant subjects.*<br>$s_2$ : *Adipocyte lipid_metabolism*$_{e_2}$ *is primarily regulated by insulin*$_{e_1}$ *and the catecholamines norepinephrine and epinephrine.*<br>$s_3$ : ... |
| (insulin, NA, TPA) | $s_1$ : *M wortmannin resulted in 80% and 20% decreases of glucose uptake stimulated by insulin*$_{e_1}$ *and TPA*$_{e_2}$ *, respectively.*<br>$s_2$ : *The effects of insulin*$_{e_1}$ *, IGF1 and TPA*$_{e_2}$ *were also observed in the presence of cycloheximide.*<br>$s_3$ : ... |

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

such as ComplEx and SimplE, for increasing the performance of distantly supervised RE. The effect of different KGC models on the distantly supervised RE will be discussed later.

(3) The model enhanced by our proposed KGC model, "JointSimplE_NER+KATT", achieves the highest precision over almost entire range of recall compared with the models that apply the existing KGC models. This proves the effectiveness of our proposed KGC model for the distantly supervised RE. Additionally, different from the exiting KGC models, the proposed end-to-end KGC model is capable of identifying ET information from word embedding of target entity. This indicates that the incorporation of semantic information of entity, such as ET, is a promising approach for enhancing the base model.

**Effect of KGC on RE.** (Han et al., 2018) indicates that KGC models could affect the performance of distantly supervised RE. For investigating the influence of KGC models on our specific RE task, we compare their link prediction results on our KG with their corresponding Precision@N (P@N) results on our RE task. Link prediction is the task that predicts *tail entity* $t$ given both *head entity* $h$ and relation $r$, e.g., $(h, r, *)$, or predict *head entity* $h$ given $(*, r, t)$. We report the mean reciprocal rank (MRR) and mean Hit@N scores for evaluating the KGC models. MRR is defined as: $MRR = \frac{1}{2*|tt|}\sum_{(h,r,t)\in tt}(\frac{1}{rank_h} + \frac{1}{rank_t})$, where $tt$ represents the test triplets. Hit@N is the proportion of the correctly predicted entities ($h$ or $t$) in top N ranked entities. Table 3 and Table 4 represent the RE precision@N and link prediction results respectively. This comparison indicates that given a biomedical dataset, the performance of a KGC model on the link prediction task could predict its effectiveness on its corresponding distantly supervised RE task. This observation also instruct us how to select the best KGC model for the base model. In addition, Table 3 and Table 4 indicate that ET is not only effective for distantly supervised RE task, but also for KGC task, and this observation will inspire us to explore other useful semantic feature of entity, such as the definition of entity, for our task.

| Model | P@2k | P@4k | P@6k | Mean |
|---|---|---|---|---|
| JointE+KATT | 0.876 | 0.786 | 0.698 | 0.786 |
| JointD+KATT | 0.848 | 0.725 | 0.528 | 0.700 |
| JointComplEx+KATT | 0.892 | 0.819 | 0.741 | 0.817 |
| JointSimplE+KATT | 0.900 | 0.808 | 0.721 | 0.809 |
| JointSimplE_NER+KATT | **0.913** | **0.829** | **0.753** | **0.831** |

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

| | **MRR** | | **Hit@** | | |
|---|---|---|---|---|---|
| **Model** | **Raw** | **Filter** | **1** | **3** | **10** |
| TransE | 0.156 | 0.200 | 0.113 | 0.244 | 0.356 |
| TransD | 0.138 | 0.149 | 0.098 | 0.160 | 0.245 |
| ComplEx | 0.278 | 0.457 | 0.380 | 0.507 | 0.587 |
| SimplE | 0.273 | 0.455 | 0.368 | 0.516 | 0.598 |
| SimplE_NER | **0.339** | **0.538** | **0.473** | **0.578** | **0.651** |

Table 4: Link prediction results for different KGC models.

## 6 Conclusion and Future Work

In this work, we tackle the task of distantly supervised RE from biomedical publications. To this end, we apply the strong joint framework proposed by (Han et al., 2018) as the base model. For enhancing its performance on our specific task, we extend the base model with other competitive KGC models. What is more, we also propose a new end-to-end KGC model, which incorporates word embedding based entity type information into a sate-of-the-art KGC model. Experimental results not only show the feasibility of the base model on the biomedical domain, but also indicate the effectiveness of our extensions. Our extended model achieves significant and consistent improvements on the biomedical dataset as compared with baselines. Since the semantic information of target entity, such as ET information, is effective for our task, in the future, we will explore other useful semantic features, such as the definition of target entity and fact triplet chain between entities (e.g., cancer→disease_has_associated_gene→ Ku86→gene_plays_role_in_process→NHEJ), for our task.

## 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.

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.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems*, pages 4289–4300.

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.

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.

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.

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

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

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*.

Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2016. Representation learning of knowledge graphs with hierarchical types. In *IJCAI*, pages 2965–2971.

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, 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.

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.

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.