Proceedings of Machine Learning Research 101:220–235, 2019 ACML 2019

A Model of Text-Enhanced Knowledge Graph Representation Learning with Collaborative Attention

Yashen Wang * YASHEN_WANG@126.COM
China Academy of Electronics and Information Technology of CETC, Beijing, China

Huanhuan Zhang HUANHUANZ_BIT@139.COM
China Academy of Electronics and Information Technology of CETC, Beijing, China

Haiyong Xie HAIYONG.XIE@IEEE.ORG
China Academy of Electronics and Information Technology of CETC, Beijing, China
University of Science and Technology of China, Hefei, Anhui, China

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

This paper proposes a novel collaborative attention mechanism, to fully utilize the mutually reinforcing relationship among the knowledge graph representation learning procedure (i.e., structure representation) and textual relation representation learning procedure (i.e., text representation). Based on this collaborative attention mechanism, a text-enhanced knowledge graph (KG) representation model is proposed, which could utilize textual information to enhance the knowledge representations and make the multi-direction signals to be fully integrated to learn more accurate textual representations for further improving structure representation and vice versa. Experimental results demonstrate the efficiency of the proposed model on both link prediction task and triple classification task.

Keywords: Knowledge Graph, Representation Learning, collaborative attention.

1. Introduction

A typical knowledge graph (KG) is usually a multiple relational directed graph, recorded as a set of relational triples (h, r, t), which indicate relation r between two entities h and t. Knowledge Graphs have become a crucial resource for many tasks in machine learning, data mining, and artificial intelligence applications including question answering Unger et al. (2012), entity linking/disambiguation Cucerzan (2007), fact checking Shi and Weninger (2016), short-text conceptualization Huang et al. (2018), information retrieval Wang et al. (2017) and link prediction Yi et al. (2017). KGs are widely used for many practical tasks, however, their completeness are not guaranteed. Therefore, it is necessary to develop Knowledge Graph Completion (KGC) methods to find missing or errant relationships with the goal of improving the general quality of KGs, which, in turn, can be used to improve or create interesting downstream applications.

Nowadays, a variety of low-dimensional representation-based methods Bordes et al. (2011)Bordes et al. (2012) have been developed to work on the KGC task. These methods usually learn continuous, low-dimensional vector representations (i.e., embeddings) for enti-

* The corresponding author.

ties and relationships by minimizing a margin-based pairwise ranking loss Lin et al. (2015a). Motivated by the linear translation phenomenon observed in well trained word embeddings Mikolov et al. (2013a), Many Representation Learning (RL) based algorithms Bordes et al. (2013)Wang et al. (2014a)Lin et al. (2015b)Xiao et al. (2016)Trouillon et al. (2016)Wang et al. (2019), have been proposed, aiming at embedding entities and relations into a vector space and predicting the missing element of triples. These models represents the head entity h, the relation r and the tail entity t with vectors $\mathbf{h}$, $\mathbf{r}$ and $\mathbf{t}$ respectively, which were trained so that $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$. To explore the instructive semantic signals from the plain text, recently it has gained lots of interests to jointly learn the embeddings of knowledge graph and text information Xu et al. (2016a) Toutanova et al. (2015), and there are several methods using textual information to help KG representation learning based on a jointly learning framework Socher et al. (2013)Wu et al. (2016)Wang et al. (2014a)Wang et al. (2014b)Riedel et al. (2013)Weston et al. (2013), different from the aforementioned work which reply only on structure information of knowledge graph itself. In these jointly-learning based models, text-based attention mechanism Wang et al. (2016)Tang et al. (2017)Kim et al. (2017)Shen et al. (2017)Shen et al. (2018)Lin et al. (2016) is widely used. However, attention values assigned for the knowledge graph representation learning (i.e., structure representation) and for the textual relation representation learning (i.e., text representation) haven't been fully integrated Mintz et al. (2009)Xie et al. (2016)Wang et al. (2014b)Verga and Mccallum (2016)Toutanova et al. (2015). Hence, the previous work fails to incorporate the complex structural signals from structure representation and semantic signals from text representation. To fully incorporate the multi-direction signals, this paper propose a novel collaborative attention mechanism, and therefore propose a text-enhanced knowledge graph representation with collaborative attention.

Actually, the main intuition behind the proposed collaborative attention is that there exists a mutually reinforcing relationship among the knowledge graph representation learning (i.e., structure representation) and textual relation representation learning (i.e., text representation), that could be reflected in the iterative training procedure, which is inspired by co-ranking strategy adopted in cooperative ranking over heterogeneous elements (e.g., entities and relations). However, our proposed adaptation of the collaborative attention mechanism to joint-learning task of knowledge graph and text is novel, and could make the multi-direction signals, i.e., signals from knowledge graph representation learning to textual relation representation learning and vice versa, to be fully integrated for deriving the solid joint-learning results for model the semantic embedded in the given knowledge graph.

In summary, the contributions of the proposed work are concluded as follows: (i) We propose a novel collaborative attention mechanism, which could mutually reinforce relationship among the knowledge graph representation learning and the textual relation representation learning; (ii) We propose a novel text-enhanced knowledge graph representation with collaborative attention; and (iii) We show the effectiveness of our model by outperforming baselines on benchmark datasets for knowledge graph representation learning task.

2. Notations

This paper represents vectors with lowercase letters and matrices with uppercase letters. Let $\mathbf{v} \in \mathrm{R}^k$ be vectors of length k, i.e., the embedding dimensionality is k. This paper

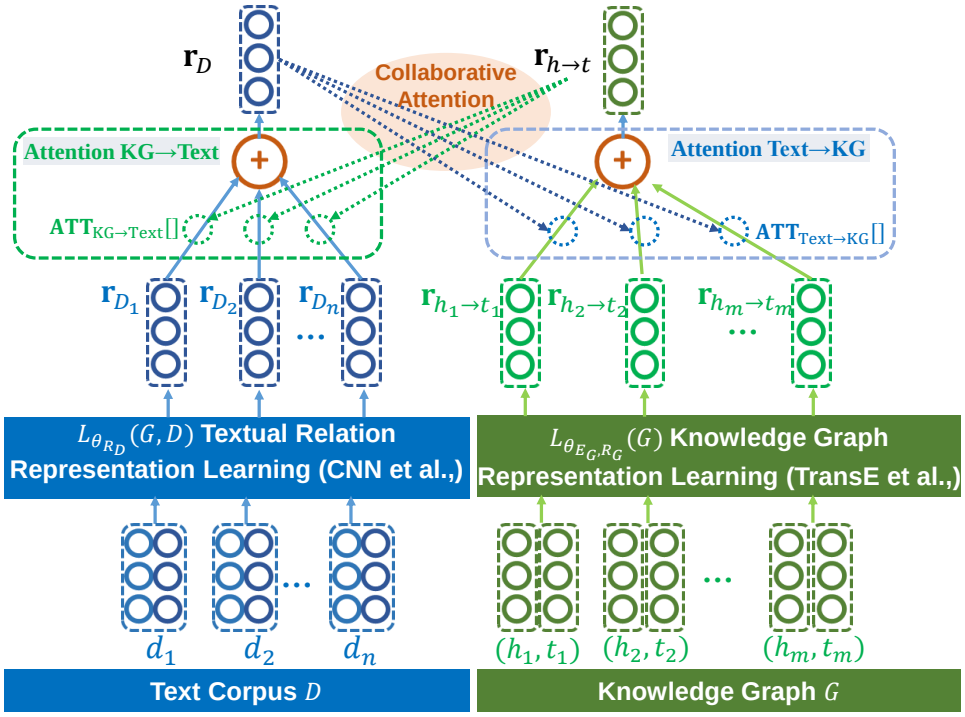

Figure 1: The framework of the proposed approach for text-enhanced knowledge graph representation with collaborative attention.

denote the text corpus consisting of sentences as D. Each sentence in D is denoted as a word sequence $d = \{w_1, \ldots, w_{|d|} | w_i \in V\}$[1], which contains two annotated mentions along with a textual relation $r_d \in R$ between them. V indicates the vocabulary, and accordingly $|V|$ indicates the number of all the words in the vocabulary. Let E and R represent the set of entities and relations respectively. A triple is represented as (h, r, t), where $h \in E$ is the head entity, $r \in R$ is the relation, and $t \in E$ is the tail entity of the triple. A knowledge graph (KG) is denoted as G here. Given a knowledge graph G, it contains $|E|$ entities and $|R|$ types of relations. The set of triples $T = \{(h, r, t)\}$ could be obtained by the representation learning models. An embedding is a function from an entity or a relation to one vector, more vectors or matrices of numbers. A representation learning model generally defines two aspects: (i) the embedding functions for entities and relations; and (ii) a function taking the embeddings for h, r and t as input and generating a prediction of whether (h, r, t) is "true" in a world or not. A representation learning model generally defines the embedding functions for entities and relations and the values of the embeddings are learned using the triples in a KG G.

1. The notation "word" (i.e., w_i) represents the word and phrase here.

3. Text-Enhanced Knowledge Graph Representation with Collaborative Attention

Knowledge Graphs (KGs) are graph-structured knowledge bases, where factual knowledge is represented in the form of relationships between entities. In our view, KGs are an example of a heterogeneous information network containing entity-nodes and relationship-edges corresponding to RDF-style triples (h, r, t) where h represents a head entity, and r is a relationship that connects h to a tail entity t. This paper introduces a novel model of text-enhanced knowledge graph representation with collaborative attention. We model each entity and each relation contained in the knowledge graph from both the structure signals from the knowledge graph itself and the textual signals from the plain text. Different from the traditional joint learning model for knowledge graph and text, the proposed model introduces a collaborative-attention mechanism for merge the aforementioned heterogeneous signals.

3.1. Sketch of the Overall Architecture

The joint representations of entities, relationships and words are recorded as model parameters $\Theta = \{\theta_{E_G}, \theta_{E_D}, \theta_{R_G}, \theta_{R_D}, \theta_V\}$. Wherein, θ_{E_G} denotes the parameters of learning entity vector representation from knowledge graph G (i.e., entity's structure representation), and θ_{E_D} denotes the parameters of learning entity vector representation from plain text in D (i.e., entity's text representation). θ_{R_G} denotes the parameters of learning relation vector representation from knowledge graph G (i.e. relation's structure representation), and θ_{R_D} denotes the parameters of learning relation vector representation from plain text in D (i.e., relation's text representation) . Besides, θ_V denotes the parameters of word vector representation of words in vocabulary V. Therefore, the proposed model aims to find optimal parameters:

$$\hat{\theta} = \arg\min \mathcal{L}_\Theta(G, D) \tag{1}$$

Wherein, $\mathcal{L}_\Theta(G, D)$ represents the loss function defined over the knowledge graph G and the text corpus D given the parameters Θ, which could be formed as follows:

$$\mathcal{L}_\Theta(G, D) = \mathcal{L}_{\theta_{E_G}, \theta_{R_G}}(G) + \alpha \cdot \mathcal{L}_{\theta_{R_D}}(G, D) + \lambda \cdot \parallel \Theta \parallel_2 \tag{2}$$

Wherein, α and λ indicate the harmonic factors, and α is also used to balance the learning ratio between knowledge graph G and plain text in corpus D. Besides, $\parallel \Theta \parallel_2$ represents the l_2 norm of Θ. $\mathcal{L}_{\theta_{E_G}, \theta_{R_G}}(G)$ is used to learn the vector representation for entities and relations from the given knowledge graph G (details in Section 3.2), and $\mathcal{L}_{\theta_{R_D}}(G, D)$ aims at learning the vector representation for relations from the plain text in corpus D (details in Section 3.3). From Eq. (2), we could observe that, with this definition of loss function, the process of knowledge graph representation learning and the process of textual representation learning could be closely coupled. Besides, Skip-Gram Mikolov et al. (2013b) based on negative sampling is leveraged for construct word vector $\mathbf{w} \in \mathrm{R}^k$.

Figure 1 overview the architecture of the proposed model for text-enhanced knowledge graph representation with collaborative attention. The overall model consists of two main modules:

(i) Knowledge Graph Representation Learning module (details in Section 3.2): This module learning the embedded representation for entity vector representation from knowledge graph G and relation vector representation from knowledge graph G, as shown in the green part in Figure 1.

(ii) Textual Relation Representation Learning module (details in Section 3.3): This module learning the embedded representation for relation vector representation from plain text in D, as shown in the blue part in Figure 1.

During the training procedure, the proposed collaborative attention mechanism integrate the aforementioned modules, as shown in the orange part in Figure 1.

Finally, the stochastic gradient descent (SGD) strategy is applied to optimize the optimization function in the proposed algorithm.

3.2. Learning Knowledge Graph Representation: $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$

Recently, translation-based models, including TransE Bordes et al. (2013), TransH Wang et al. (2014a), and TransR Lin et al. (2015b), have achieved promising results in distributional representation learning of knowledge graph. Therefore, translation-based model is utilized here to learn the vector representation of entities and relations form knowledge graph G. In order to facilitate the description, only TransE model is taken here as an example to describe the modeling procedure. Note that, any other knowledge graph embedding methods could be adopted here, and the following experimental section (Section 4) compare the experimental results of different kinds of knowledge graph representation learning methods based on translation mechanisms (e.g., TransE Bordes et al. (2013),TransH Wang et al. (2014a), and TransR Lin et al. (2015b), etc.,).

Given a each couple of head entity and tail entity, (h, t), defined in knowledge graph G, we assume there exists an *implicit* relation vector $\mathbf{r}_{h\to t}$, representing the "translation" from head entity vector $\mathbf{h}_G$ (corresponding to head entity h) to tail entity vector $\mathbf{t}_G$ (corresponding to tail entity t), as follows:

$$\mathbf{r}_{h\to t} = \mathbf{t}_G - \mathbf{h}_G \tag{3}$$

On the other hand, for each triple $(h, r, t) \in T$ defined in knowledge graph G, there exists an *explicit* relation vector $\mathbf{r}_G$, representing the "translation" from $\mathbf{h}_G$ to $\mathbf{t}_G$. With efforts above, we could define score function for each triple (h, r, t), as follows:

$$\varphi_r(h,t) = \| \mathbf{r}_{h\to t} - \mathbf{r}_G \|_2 = \| (\mathbf{t}_G - \mathbf{h}_G) - \mathbf{r}_G \|_2 \tag{4}$$

Eq. (4) shows that, for each $(h, r, t) \in T$, we would like that $\mathbf{t}_G - \mathbf{h}_G \approx \mathbf{r}_G$. Based on the this score function, the loss function over all the triples in T, $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$ in Eq. (2), could be defined as follows:

$$\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G) = \sum_{(h,r,t)\in T} \sum_{(h',r,t')\in T'} [\mu + \varphi_r(h,t) - \varphi_r(h',t')]_+ \tag{5}$$

Accompanying with T, T' is the set of negative triple, i.e., we need to sample a negative triple (h', r, t') to compute loss, given a positive triple $(h, r, t) \in T$. Following previous work Lin et al. (2015b); Nguyen et al. (2016), we construct a set of negative triples by replacing the head entity h or tail entity t with a random entity uniformly sampled from the knowledge graph G, although many other negative sampling methods exist. Therefore, $T' = \{(h', r, t)|h' \in E\} \bigcup \{(h, r, t')|t' \in E\}$. Wherein, $h' \in E$ indicates the negative head entity obtained by random sampling, and $t' \in E$ indicates the negative tail entity obtained by the same way. Besides, $\mu > 0$ represents the margin parameter. $[x]_+ = x$ where $x > 0$, and $[x]_+ = 0$ where $x \leq 0$.

3.3. Learning Textual Relation Representation: $\mathcal{L}_{\theta_{R_D}}(G, D)$

Generally, the main goal of textual relation extraction is to determine a type of relation between two entities appearing together in a piece of text. Following Sorokin and Gurevych (2017), to each occurrence of the target entity pair h and t in some sentence d, we should assign a relation type $r \in R$, which could reveals the implicit semantic of the textual relation between these two entities. Previous work Xu et al. (2016b); Sorokin and Gurevych (2017) show that, textual relations can be captured by deep neural network model and projected into a low-dimensional semantic space. Compared with the traditional algorithm Mintz et al. (2009), the algorithm based on deep learning can accurately model the semantic relation between entities from text fragments without using explicit syntactic feature Xu et al. (2015); Xiao and Liu (2016). In this study, we utilize convolutional neural network (CNN) for textual relation representation learning.

Therefore, given a sentence $d = \{w_1, \cdots, w_{|d|}\}$ containing entity pair (h, t), we assume that this sentence includes semantic signals about textual relation r_D. CNN is utilized here to expose implicit semantic of the textual relation between entity h and entity t. The procedure could be described as follows. There exists relation r defined in the knowledge graph between h and t, and the corresponding relation vector is denoted as $\mathbf{r}_G$. The concatenation of word vectors ($\mathbf{w}_i$) and their corresponding position vectors ($\mathbf{p}_i$), $\{[\mathbf{w}_1, \mathbf{p}_1], \cdots, [\mathbf{w}_{|d|}, \mathbf{p}_{|d|}]\}$, is utilized as CNN's input, and we could obtain the final embedded vector $\mathbf{r}_D$ for textual relation with the pooling layer and the convolution layer of CNN. With efforts above, we could define the following score function for the given sentence d:

$$\psi_r(d) = \| \mathbf{r}_D - \mathbf{r}_G \|_2 \tag{6}$$

Based on the this score function, the loss function over all the sentence in corpus D, $\mathcal{L}_{\theta_{R_D}}(G, D)$ in Eq. (2), could be defined as follows:

$$\mathcal{L}_{\theta_{R_D}}(G, D) = \sum_{d\in D} \sum_{r'\neq r} [\gamma + \psi_r(d) - \psi_{r'}(d)]_+ \tag{7}$$

Wherein, $\gamma > 0$ represents the margin parameter. Similar to Eq. (5), $[x]_+ = x$ where $x > 0$, and $[x]_+ = 0$ where $x \leq 0$.

3.4. Collaborative Attention between Knowledge Graph Representation Learning and Textual Relation Representation Learning

Actually, the main intuition behind the proposed collaborative attention is that there exists a mutually reinforcing relationship among the knowledge graph representation learning (i.e., structure representation) and textual relation representation learning (i.e., text representation), that could be reflected in the iterative training procedure, which is inspired by co-ranking strategy adopted in cooperative ranking over heterogeneous elements (e.g., entities and relations). However, our proposed adaptation of the collaborative attention mechanism to joint-learning task of knowledge graph and text is novel, and could make the multi-direction signals, i.e., signals from knowledge graph representation learning to textual relation representation learning (details in Section 3.4.1) and vice versa (details in Section 3.4.2), to be fully integrated for deriving the solid joint-learning results for model the semantic embedded in the given knowledge graph.

3.4.1. Text → KG: Attention From Textual Relation Representation Learning to Knowledge Graph Representation Learning

As discussed in Section 3.2, given relation r defined in the knowledge graph, we assume that there exist m pairs of entities which are eligible to relation r, $\{(h_1, t_1), \cdots, (h_m, t_m)\}$ (shown in Figure 1), and the corresponding implicit relation vectors are $\{\mathbf{r}_{h_1 \to r_1}, \cdots, \mathbf{r}_{h_m \to r_m}\}$, representing the "translation" from head entity vector $\mathbf{h}_{i_G}$ (corresponding to head entity h_i) to tail entity vector $\mathbf{t}_{i_G}$ (corresponding to tail entity t_i). Actually, there exists an explicit relation vector $\mathbf{r}_G$ corresponding to relation r after our iterative training procedure. However, not each $\mathbf{r}_{h_i \to r_i}$ contributes to $\mathbf{r}_G$ (the approximate fact) equally, let alone the noise. To overcome this problem, we attempt to leverage the beneficial semantic signal from textual relation representation learning for knowledge graph representation learning, by introducing a softmax-based attention mechanism (Text → KG), as follows:

$$\mathbf{ATT}_{\text{Text}\to\text{KG}}[i] = \text{Softmx}[\mathbf{r}_{h_i \to r_i} \cdot \tanh(\mathbf{M}_{\text{Text}\to\text{KG}} \cdot \mathbf{r}_D + \mathbf{b}_{\text{Text}\to\text{KG}})] \tag{8}$$

Wherein, $\mathbf{M}_{\text{Text}\to\text{KG}} \in \mathrm{R}^{k\times k}$ and $\mathbf{b}_{\text{Text}\to\text{KG}} \in \mathrm{R}^k$ are a part of parameters. With efforts above, the implicit relation representation $\mathbf{r}_{h\to t}$ (Eq. (3)) could be modeled as follows:

$$\mathbf{r}'_{h\to t} = \sum_{i=1}^{m} \mathbf{ATT}_{\text{Text}\to\text{KG}}[i] \cdot \mathbf{r}_{h_i \to r_i} \tag{9}$$

Wherein, $\mathbf{ATT}_{\text{Text}\to\text{KG}}[i]$ denotes the i-th attention for the corresponding implicit vector $\mathbf{r}_{h_i \to r_i}$, representing the importance (or weight) of the implicit vector. Therefore, we could redefine score function (Eq. (4)) for each triple $(h, r, t) \in \{(h_1, r, t_1), \cdots, (h_m, r, t_m)\}$, as follows:

$$\varphi_r(h, t) = \| \mathbf{r}'_{h\to t} - \mathbf{r}_G \|_2 \tag{10}$$

Accordingly, based on the this score function, the loss function over all the triples in $\{\mathbf{r}_{h_1 \to r_1}, \cdots, \mathbf{r}_{h_m \to r_m}\}$, $\mathcal{L}_{\theta_{E_G}, \theta_{R_G}}(G)$ in Eq. (2), could be defined as follows:

$$\mathcal{L}_{\theta_{E_G}, \theta_{R_G}}(G) = \sum_{(h,r,t) \in \{(h_1,r,t_1), \cdots, (h_m,r,t_m)\}} \sum_{(h',r,t') \notin \{(h_1,r,t_1), \cdots, (h_m,r,t_m)\}} [\mu + \varphi_r(h,t) - \varphi_r(h',t')]_+ \tag{11}$$

Wherein, $\mu > 0$ represents the margin parameter. $[x]_+ = x$ where $x > 0$, and $[x]_+ = 0$ where $x \leq 0$. Note that, negative sampling is utilized here for generating a negative triple $(h', r, t') \notin \{(h_1, r, t_1), \cdots, (h_m, r, t_m)\}$ to compute loss, given a positive triple (h, r, t), similar to Section 3.2.

3.4.2. KG → Text: Attention From Knowledge Graph Representation Learning to Textual Relation Representation Learning

As discussed in Section 3.3, for each relation r defined in the knowledge graph G, we could derive a set of sentences, $\{d_1, \cdots, d_n\}$ (shown in Figure 1), which reveal the implicit semantic of the textual relation r_D of relation r with the occurrence of the target entity pair h and t in this sentence while the triple (h, r, t) is defined in given knowledge graph. Besides, the corresponding output embedded relation vectors are $\{\mathbf{r}_{D_1}, \cdots, \mathbf{r}_{D_n}\}$. On the other hand, there exists an explicit relation vector $\mathbf{r}_G$ corresponding to relation r after our iterative training procedure. We aims at bridging the gap between $\mathbf{r}_D$ and $\mathbf{r}_G$ (as described in Eq. (6)), with the help of modeling $\{d_1, \cdots, d_n\}$ to generate $\{\mathbf{r}_{D_1}, \cdots, \mathbf{r}_{D_n}\}$. However, facing the same difficulties with representation learning of knowledge graph (i.e., structure representation learning in Section 3.2), not each d_j contributes to $\mathbf{r}_D$ equally, let alone the noise. To overcome this problem, we seek help from the beneficial semantic signal from knowledge graph representation learning for enhance the semantic robustness of textual relation representation learning, by introducing a softmax-based attention mechanism (Text → KG), as follows:

$$\mathbf{ATT}_{\text{KG}\to\text{Text}} = \text{Softmx}[\mathbf{r}_{h\to t} \cdot \tanh(\mathbf{M}_{\text{KG}\to\text{Text}} \cdot \mathbf{r}_{D_j} + \mathbf{b}_{\text{KG}\to\text{Text}})] \tag{12}$$

Wherein, $\mathbf{M}_{\text{KG}\to\text{Text}} \in \mathrm{R}^{k\times k}$ and $\mathbf{b}_{\text{KG}\to\text{Text}} \in \mathrm{R}^k$ are a part of parameters. With efforts above, the final embedded vector $\mathbf{r}_D$ for textual relation (in Eq. (6)) could be modeled as follows:

$$\mathbf{r}'_D = \sum_{j=1}^{n} \mathbf{ATT}_{\text{KG}\to\text{Text}}[j] \cdot \mathbf{r}_{D_j} \tag{13}$$

Wherein, $\mathbf{ATT}_{\text{KG}\to\text{Text}}[j]$ corresponds to the attention for j-th sentence with the occurrence of the target entity pair h and t in this sentence while the triple (h, r, t) is defined in given knowledge graph G, measuring the importance of the corresponding embedded vector $\mathbf{r}_{D_j}$. Therefore, we could redefine score function (Eq. (6)) for each sentence in $\{d_1, \cdots, d_n\}$, as follows:

Table 1: Statistics of dataset WN11, dataset WN18, dataset FB13 and dataset FB15k used in our experiments.

Dataset	$\|E\|$	$\|R\|$	#Train	#Valid	#Test
WN11	38,696	11	112,581	2,609	10,544
WN18	40,943	18	141,442	5,000	5,000
FB13	75,043	13	316,232	5,908	23,733
FB15k	14,951	1,345	483,142	50,000	59,071

$$\psi_r(d) = \| \mathbf{r}_D^{'} - \mathbf{r}_G \|_2 \tag{14}$$

Accordingly, based on the this score function, the loss function over all the sentences in $\{d_1, \cdots, d_n\}$, $\mathcal{L}_{\theta_{R_D}}(G, D)$ in Eq. (2), could be defined as follows:

$$\mathcal{L}_{\theta_{R_D}}(G, D) = \sum_{d \in \{d_1, \cdots, d_n\}} \sum_{r' \neq r} [\gamma + \psi_r(d) - \psi_{r'}(d)]_+ \tag{15}$$

Wherein, $\gamma > 0$ represents the margin parameter. Similar to Eq. (7) in Section 3.3, $[x]_+ = x$ where $x > 0$, and $[x]_+ = 0$ where $x \leq 0$.

4. Experiments

We evaluate our proposed text-enhanced knowledge graph representation model with collaborative attention for our comparative analysis, by leveraging representation learning based Knowledge Graph Completion (KGC) task. Generally, The Knowledge Graph Completion task includes two subtasks: (i) Link Prediction task, and (ii) Triple Classification task. We evaluate our model on both tasks with benchmark static datasets.

4.1. Datasets and Baselines

To evaluate the proposed text-enhanced knowledge graph representation model with collaborative attention, we conduct experiments on the dataset WN11 (WordNet), dataset WN18 (WordNet), dataset FB13 (Freebase) and dataset FB15k (Freebase) introduced by Bordes et al. (2013)Wang et al. (2014a)Wang et al. (2019) and use the same training\validation\test split as in previous work. The statistical information of the aforementioned datasets is sketched in Table 1. Wherein, $|E|$ and $|R|$ denote the number of entities and relation types respectively. $\#Train$, $\#Valid$ and $\#Test$ are the numbers of triple in the training, validation and test sets respectively. Moreover, we also construct a Wikipedia dataset for entity linking. We preprocess the Wikipedia articles with the following rules. First, we remove the articles less than 100 words, as well as the articles less than 10 links. Then we remove all the category pages and disambiguation pages. Moreover, we move the content to the right redirection pages. Finally we obtain about 3.74 million Wikipedia articles for indexing.

As introduced above, Following An et al. (2018), **TransE** Bordes et al. (2013), **TransH** Wang et al. (2014a), **TransR** Lin et al. (2015b) and **ComplEx** Trouillon et al. (2016) are utilized here for the baseline models, and we introduce two kinds of extended versions: (i) we denote their first kind of extended version as **JOINT+TransE**, **JOINT+TransH**, **JOINT+TransR**, and **JOINT+ComplEx**, which are enhanced text signals *without* collaborative attention (i.e., Eq. (4), Eq. (5), Eq. (6), and Eq. (7)); and (ii) we denote their second kind of extended version as **aJOINT+TransE**, **aJOINT+TransH**, **aJOINT+TransR**, and **aJOINT+ComplEx**, which are enhanced text signals *with* collaborative attention (i.e., Eq. (10), Eq. (11), Eq. (14), and Eq. (15)). Besides, **TransD** Ji et al. (2015) and **TransG** Xiao et al. (2016) are also introduced for the contrast experiments. Two widely-used measures are considered as evaluation metrics in our experiments: (i) Mean Rank (MR), indicating the mean rank of original triples in the corresponding probability ranks; and (ii) HITS@N, indicating the proportion of original triples whose rank is not larger than N ($N = 10$ is utilized here). Lower mean rank or higher HITS@10 mean better performance. What's more, we follow Bordes et al. (2013) to report the filter results, i.e., removing all other correct candidates h in ranking, which is called the "Filter" setting. In contrast to this stands the "Raw" setting.

4.2. Link Prediction

Link prediction aims at predicting the missing relation when given two entities, i.e., we predict r given $(h, ?, t)$. The dataset WN18 and dataset FB15k are the benchmark datasets for this task following Wang et al. (2019). For each triple (h, r, t) in the test set, we replace the relation r with every relation in the dataset. Overall, the original **TransE** Bordes et al. (2013), **TransH** Wang et al. (2014a), **TransR** Ji et al. (2015) and **ComplEx** Trouillon et al. (2016) are introduced here, and boosted by the proposed text-enhanced knowledge graph representation learning model, and furthermore compared with their enhanced variant with **TEKE** Wang and Li (2016). Besides, **TransD** Ji et al. (2015) and **TransG** Xiao et al. (2016) are also introduced for the contrast experiments. Mean Rank and HITS@10 are considered as evaluation metrics for this task. As the datasets are the same, we directly reuse the experimental results of several baselines from the previous literature Wang et al. (2019)An et al. (2018).

The optimal-parameter configurations are described as follows: For dataset WN18, (i) the learning rate for $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$ in Eq. (2) is 0.0005, (ii) the learning rate for $\mathcal{L}_{\theta_{R_D}}(G, D)$ in Eq. (2) is 0.0005, (iii) the vector dimension k is 300, (iv) the harmonic factors α and λ in Eq. (2) are set as 0.00005 and 0.0001 respectively, and (v) the margin parameters μ in Eq. (11) and γ in Eq. (15) are set as 5 and 3 respectively. We train the model until convergence. For dataset FB15K, (i) the learning rate for $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$ in Eq. (2) is 0.001, (ii) the learning rate for $\mathcal{L}_{\theta_{R_D}}(G, D)$ in Eq. (2) is 0.0005, (iii) the vector dimension k is 230, (iv) the harmonic factors α and λ in Eq. (2) are both set as 0.0001 respectively, and (v) the margin parameters μ in Eq. (11) and γ in Eq. (15) are set as 3 and 4 respectively. The overall link prediction results are presented in Table 2.

From the results, we observe that: (i) The proposed collaborative attention based joint-learning model outperforms previous text-enhanced knowledge representation models in the most cases; (ii) The enhancements from the proposed collaborative attention are more

Table 2: Evaluation results of link prediction task on dataset WN18 and dataset FB15K (MR and HIT@10).

Models		WN18		FB15K	
		MR	HIT@10	MR	HIT@10
	TransD Ji et al. (2015)	212	92.2	91	77.3
	TransG Xiao et al. (2016)	345	94.7	**50**	88.2
TransE	**TransE** Bordes et al. (2013)	251	89.2	125	47.1
	TEKE+TransE	**127**	93.8	79	67.6
	JOINT+TransE (Ours)	149	92.1	85	58.4
	aJOINT+TransE (Ours)	131	93.9	79	78.7
TransH	**TransH** Wang et al. (2014a)	303	86.7	84	58.5
	TEKE+TransH	128	93.6	75	70.4
	JOINT+TransH (Ours)	164	93.1	79	69.4
	aJONIT+TransH (Ours)	138	94.8	75	75.1
TransR	**TransR** Lin et al. (2015b)	219	91.7	78	65.5
	TEKE+TransR	203	92.3	79	68.5
	JOINT+TransR (Ours)	208	93.2	82	68.1
	aJONIT+TransR (Ours)	189	**95.1**	78	70.3
ComplEx	**ComplEx** Trouillon et al. (2016)	219	94.7	78	84.0
	JONIT+ComplEx (Ours)	206	94.9	63	86.7
	aJOINT+ComplEx (Ours)	184	**95.1**	54	**88.3**

marked at metric HIT@10 compared with metric MR. The proposed collaborative attention has more evident effect on upgrading of performance on dataset FB15K, which contains more complex relationship types, than dataset WN18: (i) On datset WN18, Compared with **TransE** Bordes et al. (2013), our **aJOINT+TransE** improves the average accuracy by 5.27% for metric HIT@10. Similarly, compared with **TransH** Wang et al. (2014a) and **ComplEx** Trouillon et al. (2016), our **aJOINT+TransH** improves the average accuracy by 9.34% for metric HIT@10. (i) On datset FB15K, Compared with **TransE** Bordes et al. (2013), our **aJOINT+TransE** improves the average accuracy by 67.09% for metric HIT@10. Similarly, compared with **TransH** Wang et al. (2014a) and **ComplEx** Trouillon et al. (2016), our **aJOINT+TransH** and **aJOINT+ComplEx** improves the average accuracy by 28.38% and 5.12%, respectively for metric HIT@10. Interestingly, the performance of the original **TransR** Lin et al. (2015b) is not being as good as that of original **ComplEx** Trouillon et al. (2016), however our collaborative attention mechanism puts it in the same league as **ComplEx**, as both **aJONIT+TransR** and **aJOINT+ComplEx** have reached the best experimental results at metric HITS@10.

4.3. Triple Classification

Generally, the triple classification task could be reviewed as a binary classification task, which discriminate whether the given triple is correct or not Wang et al. (2019); Wang and Li (2016); Bordes et al. (2013). Following Socher et al. (2013); An et al. (2018) we evaluate

the comparative models on the benchmark datasets, i.e., dataset WN11 and dataset FB13, and utilize the binary classification accuracy as the evaluation metric.

We introduce the proposed text-enhanced knowledge graph representation learning model to boost the original **TransE** Bordes et al. (2013), **TransH** Wang et al. (2014a), **TransR** Ji et al. (2015) and **ComplEx** Trouillon et al. (2016), and compare with their enhanced variant with **TEKE** Wang and Li (2016). Besides, **TransD** Ji et al. (2015) and **TransG** Xiao et al. (2016) are also introduced for the contrast experiments. The measurement Accyract (%) is are considered as evaluation metrics for this task. We report the results from An et al. (2018), and the overall results of triple classification task are listed in Table 3.

Table 3: Evaluation results of triple classification task on dataset WN11 and dataset FB13 (Accuracy(%)).

Models		WN11	FB13	AVG.
	TransD Ji et al. (2015)	86.4	**89.1**	87.8
	TransG Xiao et al. (2016)	87.4	87.3	87.4
TransE	**TransE** Bordes et al. (2013)	75.9	81.5	78.7
	TEKE+TransE	84.1	75.1	79.6
	JOINT+TransE (Ours)	83.7	76.0	79.8
	aJOINT+TransE (Ours)	87.2	86.9	87.1
TransH	**TransH** Wang et al. (2014a)	78.8	83.3	81.1
	TEKE+TransH	84.8	84.2	84.5
	JOINT+TransH (Ours)	85.4	83.3	84.3
	aJOINT+TransH (Ours)	87.3	86.9	87.1
TransR	**TransR** Lin et al. (2015b)	85.9	82.5	84.2
	TEKE+TransR	86.1	81.6	83.7
	JOINT+TransR (Ours)	85.9	85.0	85.5
	aJOINT+TransR (Ours)	87.1	86.0	86.6
ComplEx	**ComplEx** Trouillon et al. (2016)	86.2	85.7	86.0
	JOINT+ComplEx (Ours)	86.5	87.7	87.1
	aJOINT+ComplEx (Ours)	**88.3**	87.9	**88.1**

The optimal-parameter configurations are described as follows: For dataset WN11, (i) the learning rate for $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$ in Eq. (2) is 0.0005, (ii) the learning rate for $\mathcal{L}_{\theta_{R_D}}(G,D)$ in Eq. (2) is 0.001, (iii) the vector dimension k is 200, (iv) the harmonic factors α and λ in Eq. (2) are set as 0.00005 and 0.0001 respectively, and (v) the margin parameters μ in Eq. (11) and γ in Eq. (15) are set as 3 and 4 respectively. We train the model until convergence. For dataset FB13, (i) the learning rate for $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$ in Eq. (2) is 0.001, (ii) the learning rate for $\mathcal{L}_{\theta_{R_D}}(G,D)$ in Eq. (2) is 0.001, (iii) the vector dimension k is 250, (iv) the harmonic factors α and λ in Eq. (2) are set as 0.00005 and 0.0001 respectively, and (v) the margin parameters μ in Eq. (11) and γ in Eq. (15) are set as 5 and 4 respectively.

From Table 3, we could get the similar conclusion that, the proposed text-enhanced knowledge graph representation learning model outperforms all baselines in most cases. Compared with **TransE** Bordes et al. (2013), our **aJOINT+TransE** improves the aver-

age accuracy by 10.67%. Similarly, compared with **TransH** Wang et al. (2014a) and **ComplEx** Trouillon et al. (2016), our **aJOINT+TransH** and **aJOINT+ComplEx** improves the average accuracy by 7.40% and 2.21%, respectively. We argue that, this phenomenon is rooted in the methodology that, the proposed text-enhanced knowledge graph representation model with collaborative attention could utilize accurate textual information enhance the knowledge representations of a triple. Furthermore, compared with **JOINT+TransE** and **JOINT+TransH**, the collaborative attention variants **aJOINT+TransE** and **aJOINT+TransH** improve the average accuracy by 9.15% and 3.32%, respectively. The results verifies that it is critical to introduce the collaborative attention mechanism (details in Section 3.4) to mutually reinforce the relationship among the knowledge graph representation learning (i.e., structure representation) and textual relation representation learning (i.e., text representation), which could be reflected in the iterative training procedure as described in Figure 1.

5. Conclusions

Knowledge Graphs (KGs) are graph-structured knowledge bases, where factual knowledge is represented in the form of relationships between entities. To fully utilize the mutually reinforcing relationship among the knowledge graph representation learning (i.e., structure representation) and textual relation representation learning (i.e., text representation), this paper proposes a novel collaborative attention mechanism to enhance the knowledge graph representation by text semantic signals, which could make the multi-direction signals, i.e., signals from knowledge graph representation learning to textual relation representation learning and vice versa, to be fully integrated. Empirically, we show the proposed text-enhanced knowledge graph representation with collaborative attention can improve the performance of the current translation-based knowledge representation models on several benchmark datasets.

Acknowledgements

The authors are very grateful to the editors and reviewers for their helpful comments. This work is funded by: (i) the China Postdoctoral Science Foundation (No.2018M641436); (ii) the Joint Advanced Research Foundation of China Electronics Technology Group Corporation (CETC) (No.6141B08010102); (iii) 2018 Culture and tourism think tank project (No.18ZK01); (iv) the Financial Support from Beijing Science and Technology Plan (Z181100009818020); and (v) the Capital's Funds for Health Improvement and Research (2018-1-2121).

References

Bo An, Bo Chen, Xianpei Han, and Le Sun. Accurate text-enhanced knowledge graph representation learning. In *NAACL-HLT*, 2018.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *AAAI'11*, pages 301–306, 2011.

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. Joint learning of words and meaning representations for open-text semantic parsing. In *AISTATS*, 2012.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS'13*, pages 2787–2795, 2013.

Silviu Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 708–716, 2007.

Heyan Huang, Yashen Wang, Chong Feng, Zhirun Liu, and Qiang Zhou. Leveraging conceptualization for short-text embedding. *IEEE Transactions on Knowledge and Data Engineering*, 30(7):1282–1295, 2018.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jian Zhao. Knowledge graph embedding via dynamic mapping matrix. In *ACL*, 2015.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. 2017.

Yankai Lin, Zhiyuan Liu, Huan-Bo Luan, Maosong Sun, Siwei Rao, and Song Liu. Modeling relation paths for representation learning of knowledge bases. In *EMNLP*, 2015a.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI'15*, pages 2181–2187, 2015b.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, volume 1, pages 2124–2133, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119, 2013b.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Joint Conference of the Meeting of the Acl and the International Joint Conference on Natural Language Processing of the Afnlp: Volume*, 2009.

Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. Stranse: a novel embedding model of entities and relationships in knowledge bases. In *HLT-NAACL*, 2016.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. Relation extraction with matrix factorization and universal schemas. In *HLT-NAACL*, 2013.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. 2017.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Bi-directional block self-attention for fast and memory-efficient sequence modeling. 2018.

Baoxu Shi and Tim Weninger. Fact checking in heterogeneous information networks. In *International Conference Companion on World Wide Web*, pages 101–102, 2016.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *NIPS'13*, pages 926–934, 2013.

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

Yujin Tang, Jianfeng Xu, Kazunori Matsumoto, and Chihiro Ono. Sequence-to-sequence model with attention for time series classification. In *IEEE International Conference on Data Mining Workshops*, pages 503–510, 2017.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *EMNLP'15*, pages 1499–1509, 2015.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. 2016.

Christina Unger, Jens Lehmann, Axel Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. Template-based question answering over rdf data. In *International Conference on World Wide Web*, pages 639–648, 2012.

Patrick Verga and Andrew Mccallum. Row-less universal schema. 2016.

Yashen Wang, Heyan Huang, Chong Feng, Qiang Zhou, Jiahui Gu, and Xiong Gao. Cse: Conceptual sentence embeddings based on attention model. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 505–515, 2016.

Yashen Wang, Heyan Huang, and Chong Feng. Query expansion based on a feedback concept model for microblog retrieval. In *International Conference on World Wide Web*, pages 559–568, 2017.

Yashen Wang, Yifeng Liu, Huanhuan Zhang, and Haiyong Xie. Leveraging lexical semantic information for learning concept-based multiple embedding representations for knowledge graph completion. In *APWeb/WAIM*, 2019.

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

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph and text jointly embedding. In *EMNLP'14*, pages 1591–1601, 2014b.

Zhigang Wang and Juanzi Li. Text-enhanced representation learning for knowledge graph. In *International Joint Conference on Artificial Intelligence*, 2016.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. Connecting language and knowledge bases with embedding models for relation extraction. In *EMNLP*, 2013.

Jiawei Wu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Knowledge representation via joint learning of sequential text and knowledge graphs. *ArXiv*, abs/1609.07075, 2016.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. Transg : A generative model for knowledge graph embedding. In *Meeting of the Association for Computational Linguistics*, pages 2316–2325, 2016.

Minguang Xiao and Cong Liu. Semantic relation classification via hierarchical recurrent neural network with attention. In *COLING*, 2016.

Ruobing Xie, Zhiyuan Liu, J. J. Jia, Huanbo Luan, and Maosong Sun. Representation learning of knowledge graphs with entity descriptions. In *AAAI*, 2016.

Jiacheng Xu, Chen Kan, Xipeng Qiu, and Xuanjing Huang. Knowledge graph representation with jointly structural and textual encoding. 2016a.

Jiacheng Xu, Xipeng Qiu, Kan Chen, and Xuanjing Huang. Knowledge graph representation with jointly structural and textual encoding. In *IJCAI*, 2016b.

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

Tay Yi, Anh Tuan Luu, and Siu Cheung Hui. Non-parametric estimation of multiple embeddings for link prediction on dynamic knowledge graphs. In *Thirty First Conference on Artificial Intelligence*, 2017.