IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 31, 2020, accepted February 24, 2020, date of publication March 16, 2020, date of current version March 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981212

A Model of Text-Enhanced Knowledge Graph Representation Learning With Mutual Attention

YASHEN WANG[1], HUANHUAN ZHANG[1], GE SHI[2], ZHIRUN LIU[2], AND QIANG ZHOU[2]

[1]National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC), China Academy of Electronics and Information Technology, Beijing 100041, China

[2]Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer, Beijing Institute of Technology, Beijing 100041, China

Corresponding author: Yashen Wang (yswang@bit.edu.cn)

This work was supported in part by the National Key Research and Development Project under Grant 2017YFC0820503, in part by the National Natural Science Foundation of China under Grant U19B2026, in part by the China Postdoctoral Science Foundation under Grant 2018M641436, in part by the New Generation of Artificial Intelligence Special Action Project under Grant AI20191125008, in part by the National Integrated Big Data Center Pilot Project under Grant 17111001 and Grant 17111002, and in part by the Joint Advanced Research Foundation of China Electronics Technology Group Corporation (CETC) under Grant 6141B08010102.

ABSTRACT Recently, it has gained lots of interests to jointly learn the embeddings of knowledge graph (KG) and text information. However, previous work fails to incorporate the complex structural signals (from structure representation) and semantic signals (from text representation). This paper proposes a novel text-enhanced knowledge graph representation model, which can utilize textual information to enhance the knowledge representations. Especially, a mutual attention mechanism between KG and text is proposed to learn more accurate textual representations for further improving knowledge graph representation, within a unified parameter sharing semantic space. Different from conventional joint models, no complicated linguistic analysis or strict alignments between KG and text are required to train our model. Besides, the proposed model could fully incorporate the multi-direction signals. Experimental results show that the proposed model achieves the state-of-the-art performance on both link prediction and triple classification tasks, and significantly outperforms previous text-enhanced knowledge representation models.

INDEX TERMS Knowledge graph representation, textual relation representation, mutual attention mechanism, representation learning.

I. INTRODUCTION

Knowledge Graphs (KGs) are graph-structured knowledge bases, wherein factual knowledge is represented in the form of relationships between entities, 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 [1], entity linking/disambiguation [2], text generation [3], fact checking [4], short-text conceptualization [5], information retrieval [6] and link prediction [7]. 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.

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Afzal.

Nowadays, a variety of low-dimensional representation-based methods [8], [9] have been developed to work on the KGC task. These methods usually learn continuous, low-dimensional vector representations (i.e., embeddings) for entities and relationships by minimizing a margin-based pairwise ranking loss [10]. Motivated by the linear translation phenomenon observed in well trained word embeddings [11], Many Representation Learning (RL) based algorithms [12]–[17], 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 $h + r \approx t$.

However, traditional knowledge graph model based on representation learning, only utilizes the structure information embedded in the given knowledge graph, and on the other hand textual information in plain text provides abundant semantic and contextual information, which could contribute to disambiguation and completion of the entity representation and relation representation of the given knowledge

graph. Hence, textual information could be regarded as an effective supplements for knowledge graph completion task. 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 [18], [19], and there are several methods using textual information to help KG representation learning based on a jointly learning framework [13], [20]–[24], 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 [25]–[29] 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 [19], [22], [30]–[32]. 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 mutual attention mechanism, and therefore propose a text-enhanced knowledge graph representation with collaborative attention.

Actually, the main intuition behind the proposed mutual 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 mutual 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 mutual 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 mutual attention; and (iii) We show the effectiveness of our model by outperforming baselines on benchmark datasets for knowledge graph representation learning task.

II. RELATED WORK

Many knowledge graphs have recently arisen, pushed by the W3C recommendation to use the resource description framework (RDF) for data representation. Examples of such knowledge graphs include DBPedia [33], Freebase [34] and the Google Knowledge Vault [35]. Motivating applications of knowledge graph completion include question answering [36] and more generally probabilistic querying of knowledge bases [37], [38]. First approaches to relational learning relied upon probabilistic graphical models [39], such as bayesian networks and markov logic networks. Then, asymmetry of relations was quickly seen as a problem and asymmetric extensions of tensors were studied, mostly by either considering independent embeddings [40] or considering relations as matrices instead of vectors in the RESCAL model [41]. Pairwise interaction models were also considered to improve prediction performances. For example, the Universal Schema approach [23] factorizes a 2D unfolding of the tensor (a matrix of entity pairs vs. relations).

Nowadays, a variety of low-dimensional representation-based methods have been developed to work on the KGC task, including Bilinear Model [42], Distance Model [8], Unstructured Model [9], and Single Layer Model [20]. And many translation-based methods are introduced, including TransE [12] and its extensions like TransH [13], TransD [43], TransR [14], TransG [15], ComplEx [16], and so on. These methods usually learn continuous, low-dimensional vector representations (i.e., embeddings) for entities and relationships by minimizing a margin-based pairwise ranking loss [10], [44]. The most widely used embedding model in this category is TransE [12], which views relationships as translations from a head entity to a tail entity on the same low-dimensional plane. Based on the initial idea of treating two entities as a translation of one another (via their relationship) in the same embedding plane, several models have been introduced to improve the initial TransE model. The newest contributions in this line of work focus primarily on the changes in how the embedding planes are computed and/or how the embeddings are combined. For example, the entity translations in TransH [13] are computed on a hyperplane that is perpendicular to the relationship embedding. In TransR [14] the entities and relationships are embedded on separate planes and then the entity-vectors are translated to the relationship's plane. Structured Embedding (SE) [8] creates two translation matrices for each relationship and applies them to head and tail entities separately. Knowledge Vault [35] and HolE [45], on the other hand, focus on learning a new combination operator instead of simply adding two entity embeddings element-wise. Take HolE as example, the circular correlation is used for combining entity embeddings, measuring the covariance between embeddings at different dimension shifts. Besides, [15] proposed a manifold-based embedding principle to deal with the overstrict geometric form of translation-based assumption. Reference [16] employed complex value embeddings to understand the structural information. However, only original structure information is utilized in these traditional knowledge graph representation learning methods, which directly affect the disambiguation and completion of the entity representation and relation representation of the given knowledge graph.

To overcome this problem, recent research has explored the instructive semantic signals from the plain text, and recently

TABLE 1. Overview of notations in this study.

Parameters	Definition
θ_{E_G}	parameters of learning entity vector representation from knowledge graph G (i.e., entity's structure representation).
θ_{R_G}	parameters of learning relation vector representation from knowledge graph G (i.e. relation's structure representation).
θ_{E_C}	parameters of learning entity vector representation from plain text in corpus C (i.e., entity's text representation).
θ_{R_C}	parameters of learning relation vector representation from plain text in corpus C (i.e., relation's text representation).
θ_V	parameters of word vector representation of words in vocabulary V.
Θ	$\Theta = \{\theta_{E_G}, \theta_{R_G}, \theta_{E_C}, \theta_{R_C}, \theta_V\}$

it has gained lots of interests to jointly learn the embeddings of knowledge graph and text information [14], [18], [19], [46] [18]. Several methods using textual information to help KG representation learning based on a jointly learning framework have been proposed [13], [20]–[24], different from the aforementioned traditional representation learning work which reply only on structure information of knowledge graph itself. Generally, these kinds of jointly-learning based models widely used the text-based attention mechanism [25]–[29]. Especially, [24] directly sum up knowledge and text ranking scores. References [46] and [47] used neural networks to embed text descriptions into knowledge graph embedding spaces. Reference [19] extracted textual relations using dependency parsing to incorporate text information. These models need well-aligned datasets and cannot be well generalized to most general cases of combining knowledge graph and text. Reference [22] trained words and entities together to let them share parameters. Reference [23] proposed universal schema to transmit information between relations of knowledge graph and textual patterns via their common entity pairs. Reference [32] further incorporated neural networks to relax constraints imposed by entity pairs in universal schema. These models have no need of strictly aligned datasets, but only take partial information into consideration. These works have achieved reasonable results by attempting to combine knowledge graph and text for knowledge acquisition. Unfortunately, in these models, 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 [19], [22], [30]–[32]. Hence, the previous work fails to incorporate the complex structural signals from structure representation and semantic signals from text representation. We argue 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). To fully incorporate the multi-direction signals, this paper propose a novel mutual attention mechanism, and therefore propose a text-enhanced knowledge graph representation with collaborative attention. Our proposed adaptation of the mutual 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.

III. METHODOLOGY

In this section, we introduce the model of text and knowledge graph jointly-learning with mutual attention, starting with notations and definitions.

A. NOTATIONS AND DEFINITIONS

We denote knowledge graph (KG) as $G = \{E, R, T\}$, where E, R and T indicate sets of entities, relations and facts respectively. Each fact triple $(h, r, t) \in T$ indicates a relation $r \in R$ between $h \in E$ and $t \in E$.

Accompanying with G, we denote the text corpus consisting of sentences as T. The vocabulary of T is denoted as V. Each sentence in T is a sequence with n words $s = \{w_1, \ldots, w_n\}$, $w_i \in V$. In each sentence, there are two annotated entity mentions along with a textual relation $r_s \in R$ between them. For each entity, relation and word $h, t \in E, r \in R$ and $w \in V$, we use the bold face $\mathbf{h}$, $\mathbf{t}$, $\mathbf{r}$, $\mathbf{w} \in R^{k_w}$ to indicate their low-dimensional vectors respectively, where k_w is the embedding dimension. The main parameters of the proposed model is described in details in the following Table 1.

B. OVERVIEW

Overall, the proposed model aims to find optimal parameters:

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

Which is denoted as the joint representations of entities, relationships and words. Wherein, $\mathcal{L}_\Theta(G, C)$ represents the loss function defined over the given knowledge graph G and the text corpus C, according to the parameters Θ. To closely coupling the process of knowledge graph representation learning and the process of textual representation learning, we could further reform $\mathcal{L}_\Theta(G, C)$ in the E.q. (1) as follows:

$$\mathcal{L}_\Theta(G, C) = \mathcal{L}_{\theta_{E_G}, \theta_{R_G}}(G) + \alpha \cdot \mathcal{L}_{\theta_{R_C}}(G, C) + \lambda \cdot \| \Theta \|_l \tag{2}$$

wherein, we usually use the l_2 norm of Θ to represent $\| \Theta \|_l$. $\mathcal{L}_{\theta_{E_G}, \theta_{R_G}}(G)$ is used to learn the vector representation for entities and relations from the given knowledge graph G, which will be described in details in following Section III-C, and $\mathcal{L}_{\theta_{R_C}}(G, C)$ is used to learn the vector

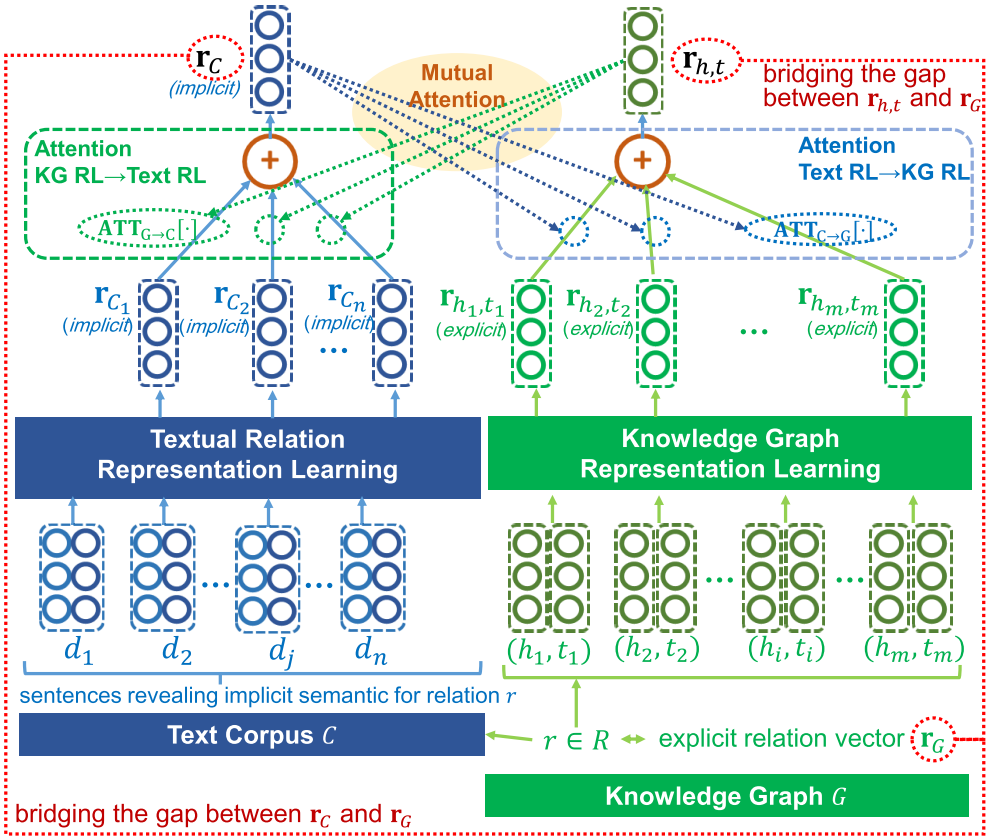

FIGURE 1. The overall framework of the proposed jointly-learning model with mutual attention.

representation for relations from the plain text in corpus C, which will be discussed in Section III-D. Besides, the widely-used Skip-Gram mechanism [11] based on negative sampling is leveraged for constructing word vector $w \in R^k$. Figure 1 overview the architecture of the proposed model for jointly-learning text and knowledge graph with mutual attention (Section III-E), combing knowledge graph representation learning (Section III-C) and textual relation representation learning (Section III-D). Finally, the stochastic gradient descent (SGD) strategy is applied to optimize the optimization function in the proposed algorithm.

Besides, Figure 2 sketches the key modules of the overall model:

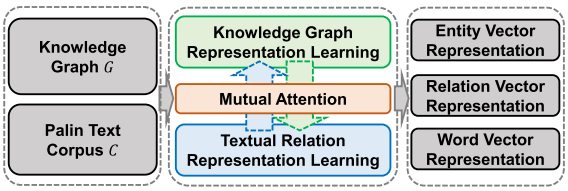

FIGURE 2. The flow chart of the proposed approach for text-enhanced knowledge graph representation with mutual attention.

(i) Knowledge Graph Representation Learning Module (details in Section III-C): 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 III-D): This module learning the embedded representation for relation vector representation from plain text in corpus C, as shown in the blue part in Figure 1.

(iii) Mutual Attention Module (details in Section III-E): During the training procedure, the proposed mutual attention mechanism integrate the aforementioned modules, as shown in the orange part in Figure 1 (Especially, green/blue crossed dashed lines in Figure 1 represent the proposed mutual mechanism between knowledge graph representation learning and textual relation representation learning).

In the following parts, we will describe each component in detail.

C. KNOWLEDGE GRAPH REPRESENTATION LEARNING

Recently, to complete or predict the missing relation element of triples, translation-based knowledge graph representation learning (RL) is widely deployed. RL embeds entities and relations into a vector space, and has produced many successful translation models including models, including TransE [12], TransH [13], and TransR [14]. Therefore, translation-based model is utilized here to learn the distributional vector representation of entities and relations form

knowledge graph G. In order to facilitate the description, only TransE model [12], which aims to generate precise vectors of entities and relations following the principle $h + r \approx t$, which means t is "translated" from h by r, is taken here as an example to describe the modeling procedure. Note that, any other knowledge graph embedding methods could be adopted here, because these methods usually learn continuous, low-dimensional vector representations (i.e., embeddings) for entities and relationships by minimizing a margin-based pairwise ranking loss. The following experimental section (Section IV) compare the experimental results of different kinds of knowledge graph representation learning methods based on translation mechanisms (e.g., TransE [12], TransH [13], TransA [48], HolE [45] and TransR [14], etc.,).

For a each couple of head entity h and tail entity t defined in given knowledge graph G, we assume that there is an *implicit* relation vector $r_{h,t}$, indicating the "translation" from head entity vector h_G (respect to the head entity h) to tail entity vector t_G (respect to the tail entity t), as follows:

$$r_{h,t} = t_G - h_G \tag{3}$$

Moreover, we could denote r_G as the *explicit* relation vector for each triple $(h, r, t) \in T$ defined in knowledge graph G, representing the "translation" from h_G to t_G. E.q. (4) shows that, for each $(h, r, t) \in T$, We would like that $t_G - h_G \approx r_G$, hence the score function for each triple (h, r, t) could be defined as follows:

$$\varphi_r(h, t) = \| r_{h,t} - r_G \|_l = \| (t_G - h_G) - r_G \|_l \tag{4}$$

Usually, l_2 norm is utilized in E.q. (4). 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 E.q. (2), could be defined as follows:

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

wherein, $\mu > 0$ represents the margin parameter. $[x]_+ = x$ where $x > 0$, and $[x]_+ = 0$ where $x \leq 0$. $\mu > 0$ represents the margin parameter. T' is the set of negative triple respect to T defined above, i.e., we need to sample a negative triple (h', r, t') to compute loss, given a positive triple $(h, r, t) \in T$. 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, following previous work [14], [49], [50], which is widely used in many research. Therefore, $h' \in E$ and $t' \in E$ indicate the negative head entity and tail entity obtained by random sampling respectively.

D. TEXTUAL RELATION REPRESENTATION LEARNING

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 [51], to each occurrence of the target entity pair h and t in the given sentence d, we should assign a relation type $r \in R$, representing the implicit semantic of the textual relation between these two entities, for h and t. Recently, the deep neural network models and representation learning strategy, are widely-used in previous work [51], [52] for capturing textual relations, which are projected into a low-dimensional semantic space in these models. Compared with the traditional algorithm [30], the algorithm based on deep learning can accurately model the semantic relation between entities from context fragments without using extra and explicit syntactic feature [53], [54]. The convolutional neural network (CNN) is adopted in this study 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_C. 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 r_G. The concatenation of word vectors (w_i) and their corresponding position vectors (p_i), is utilized as CNN's input, and we could obtain the final embedded vector r_C for textual relation with the pooling layer and the convolution layer of CNN. The score function for the given sentence d could be defined as follows:

$$\psi_r(d) = \| r_C - r_G \|_2 \tag{6}$$

With efforts above, the loss function over all the sentence in corpus C, $\mathcal{L}_{\theta_{R_C}}(G, C)$ in E.q. (2), could be defined as follows:

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

E. JOINTLY-LEARNING WITH MUTUAL ATTENTION

The proposed mutual attention mechanism for text and knowledge graph jointly-learning could make the multi-direction signals, i.e., signals from KG representation learning to textual relation representation learning (Section III-E.1) and vice versa (Section III-E.2), to be fully integrated for deriving the solid joint-learning results for model the semantic embedded in the given knowledge graph.

1) ATTENTION FROM TEXTUAL RELATION REPRESENTATION LEARNING TO KG REPRESENTATION LEARNING

As discussed above, given relation r defined in the knowledge graph (KG), there exist two kinds of relation vectors for r: (i) Implicit relation vectors: 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 $\{r_{h_1,r_1}, \cdots, r_{h_m,r_m}\}$, representing the translation from head entity vector h_{i_G} (cor responding to head entity h_i) to tail entity vector t_{i_G} (corresponding to tail entity t_i). (ii) Explicit relation vector: there exists an explicit relation vector r_G corresponding to relation r defined in knowledge graph. However, not each

r_{h_i,r_i} contributes to r_G 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 (notation "$C \rightarrow G$" indicating the attention direction from corpus to knowledge graph), as follows:

$$ATT_{C\rightarrow G}[i] = Softmx[r_{h_i,r_i} \cdot \tanh(M_{C\rightarrow G} \cdot r_C + b_{C\rightarrow G})] \tag{8}$$

wherein, $\{M_{C\rightarrow G} \in R^{k\times k}, b_{C\rightarrow G} \in R^k\}$ are parts of parameters to be trained. With efforts above, the implicit relation representation $r_{h,t}$ (E.q. (3)) could be modeled as follows:

$$r_{h,t} = \sum_{i=1}^{m} ATT_{C\rightarrow G}[i] \cdot r_{h_i,r_i} \tag{9}$$

wherein, $ATT_{C\rightarrow G}[i]$ denotes the i-th attention for the corresponding implicit vector r_{h_i,r_i}, representing the Confidence of the implicit vector. Therefore, we could redefine score function (E.q. (4)) for each triple $(h, r, t) \in \{(h_1, r, t_1), \cdots, (h_m, r, t_m)\}$, as follows:

$$\varphi_r(h, t) = \| r_{h,t} - r_G \|_l \tag{10}$$

Intuitively, this score function is respect to the red dashed line in Figure 1, marked with "bridging the gap between $r_{h,t}$ and r_G". Based on the this score function, the loss function over all the triples in $\{r_{h_1,r_1}, \cdots, r_{h_m,r_m}\}$, $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$ in E.q. (2), could be defined as following E.q. (11), shown at the bottom of the next page. Wherein, the definition of $\mu > 0$ and $[\cdot]_+$ are the same with E.q. 5.

2) ATTENTION FROM KG REPRESENTATION LEARNING TO TEXTUAL RELATION REPRESENTATION LEARNING

As discussed in Section III-D, for each relation $r \in R$ 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_C 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 $\{r_{C_1}, \cdots, r_{C_n}\}$. On the other hand, there exists an explicit relation vector r_G corresponding to relation r. We aims at bridging the gap between r_C and r_G (as described in E.q. (6)), with the help of modeling $\{d_1, \cdots, d_n\}$ to generate $\{r_{C_1}, \cdots, r_{C_n}\}$. However, facing the same difficulties with representation learning of knowledge graph (i.e., structure representation learning in Section III-C), not each d_j contributes to r_C equally. 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 (notation "$G \rightarrow C$" representing the attention direction from knowledge graph to corpus), as follows:

$$ATT_{G\rightarrow C} = Softmx[r_{h,t} \cdot \tanh(M_{G\rightarrow C} \cdot r_{C_j} + b_{G\rightarrow C})] \tag{12}$$

wherein, $M_{G\rightarrow C} \in R^{k\times k}$ and $b_{G\rightarrow C} \in R^k$ are a part of parameters to be trained. With efforts above, the final embedded vector r_C for textual relation (in E.q. (6)) could be modeled as follows:

$$r_C = \sum_{j=1}^{n} ATT_{G\rightarrow C}[j] \cdot r_{C_j} \tag{13}$$

wherein, $ATT_{G\rightarrow C}[j]$ 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 r_{C_j}. Therefore, we could redefine score function (E.q. (6)) for each sentence in $\{d_1, \cdots, d_n\}$, as follows:

$$\psi_r(d) = \| r_C - r_G \|_l \tag{14}$$

This score function is respect to the red dashed line in Figure 1, marked with "bridging the gap between r_C and r_G". With efforts above, based on the this score function, we form the loss function over all the sentences in $\{d_1, \cdots, d_n\}$, $\mathcal{L}_{\theta_{R_C}}(G, C)$ in E.q. (2), as follows:

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

IV. EXPERIMENTS

We evaluate our proposed text-enhanced knowledge graph representation model with mutual attention based on Knowledge Graph Completion (KGC) task, mainly consists of: Link Prediction (Section IV-C), and (ii) Triple Classification (Section IV-D). Besides, the statistical t-test [55], [56] is employed here: To decide whether the improvement by algorithm A over algorithm B is significant, the t-test calculates a value p based on the performance of A and B. The smaller p is, the more significant the improvement is. If the p is small enough ($p < 0.05$), we conclude that the improvement is statistically significant.

A. DATASETS

This paper conducts experiments on the dataset WN11 (WordNet), dataset WN18 (WordNet), dataset WN18RR (WordNet), dataset FB13 (Freebase), dataset FB15k (Freebase) and dataset FB15k-237 introduced by [12], [13], [17], [57], [58], to evaluate the proposed text-enhanced knowledge graph representation model with mutual attention. Note that, we use the same training\validation\test split as in previous work. The statistic information of the aforementioned datasets is sketched in Table 2. Wherein, $|E|$ and $|R|$ represent the number of entities and relation types respectively. #*Train*, #*Valid* and #*Test* indicate the numbers of triple in the training, validation and test sets, respectively. Moreover, a Wikipedia

TABLE 2. Statistics of dataset WN11, dataset WN18, dataset WN18RR, dataset FB13 and dataset FB15k and dataset FB15k-237.

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
WN18RR	40,943	11	86,835	3,034	3,134
FB13	75,043	13	316,232	5,908	23,733
FB15k	14,951	1,345	483,142	50,000	59,071
FB15k-237	14,541	237	272.115	17,535	40,466

dataset is constructed with help fo the following strategies [59]: 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.

B. COMPARATIVE MODELS

Following [60], **TransE** [12], **TransH** [13], **TransR** [14] and **ComplEx** [16] are utilized here for the baseline models, and we introduce two kinds of extended versions:

(i) **JOINT+TransE**, **JOINT+TransH**, **JOINT+TransR**, and **JOINT+ComplEx**, which are enhanced text signals with E.q. (4), E.q. (5), E.q. (6), and E.q. (7);

(ii) **maJOINT+TransE**, **maJOINT+TransH**, **maJOINT+TransR**, and **maJOINT+ComplEx**, which are enhanced text signals with E.q. (10), E.q. (11), E.q. (14), and E.q. (15).

The prefix ''ma-'' (respect to ''mutual attention''), distinguishes these two kinds of extended versions, indicating whether the mutual attention is used. Besides, **TransD** [43], **TransA** [48], **HolE** [45], **TransG** [15] are also introduced for the contrast experiments.

C. LINK PREDICTION TASK

Link prediction aims at predicting the missing relation when given two entities, i.e., we predict r given $(h, ?, t)$. Following [17], [60]–[62], the dataset WN18, dataset WN18RR, dataset FB15k and dataset FB15k-237 are the benchmark datasets for this task. According to the previous work, 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** [12], **TransH** [13], **TransR** [43], **TransD** [43] and **TransG** [15] and **ComplEx** [16] are introduced here, and boosted by the proposed jointly-learning model, and furthermore compared with their enhanced variant with **TEKE** [47]. 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. As the datasets are the same, we directly reuse the experimental results of several baselines from the previous literature [17], [60].

For dataset WN18 and dataset WN18RR, the optimal-parameter configurations are described as follows: (i) the learning rate for $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$ in E.q. (2) is 0.0005, (ii) the learning rate for $\mathcal{L}_{\theta_{R_D}}(G, D)$ in E.q. (2) is 0.0005, (iii) the vector dimension k is 300, (iv) the harmonic factors α and λ in E.q. (2) are set as 0.00005 and 0.0001 respectively, and (v) the margin parameters μ in E.q. (11) and γ in E.q. (15) are set as 5 and 3 respectively. We train the model until convergence. For dataset FB15k and dataset FB15k-237, the optimal-parameter configurations are described as follows: (i) the learning rate for $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$ in E.q. (2) is 0.001, (ii) the learning rate for $\mathcal{L}_{\theta_{R_D}}(G, D)$ in E.q. (2) is 0.0005, (iii) the vector dimension k is 230, (iv) the harmonic factors α and λ in E.q. (2) are both set as 0.0001 respectively, and (v) the margin parameters μ in E.q. (11) and γ in E.q. (15) are set as 3 and 4 respectively.

The overall link prediction results are presented in Table 3 sketches the overall evaluation results of link prediction task on several datasets. Both **maJOINT+TransR** and **maJOINT+ComplEx** have reached the best experimental results at metric HITS@10, in most cases. The proposed text and knowledge graph jointly-learning model with mutual attention outperforms previous text-enhanced knowledge representation models in the most cases, and the enhancements from the proposed mutual attention are more marked at metric HIT@10 compared with metric MR. E.g., for daset FB15K-237, compared with conventional **TransE** [12], our **maJOINT+TransE** improves the average accuracy by 5.16% for metric HIT@10, while the improvement over metric MR is 57.93%; compared with **TransR** [14], our jointly-learning enhanced model **maJOINT+TransR** improves the average accuracy by 3.56% for metric HIT@10, while the improvement over metric MR is 13.89%; The proposed mutual attention methodology among text and KG 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** [12], our **maJOINT+TransE** improves the average accuracy by 5.27% for metric HIT@10. Similarly, compared with **TransH** [13] and **ComplEx** [16], our **maJOINT+TransH** improves the average accuracy by 9.34% for metric HIT@10. (ii) On datset FB15K, compared with **TransE** [12], our **maJOINT+TransE** improves the average accuracy by 67.09% for metric HIT@10. Similarly, compared with

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

TABLE 3. Performance of link prediction task on dataset WN18, dataset WN18RR, dataset FB15K and dataset FB15K-237 (MR and HIT@10). The superscript §, ⋆, † and ‡ respectively denote statistically significant improvements over **TransE** [12], **TransR** [14], **TransG** [15] and **ComplEx** [16] ($p < 0.05$).

Models		WN18		WN18RR		FB15K		FB15K-237	
		MR	HIT@10	MR	HIT@10	MR	HIT@10	MR	HIT@10
	TransD [43]	212	92.2	541	45.1	91	77.3	244	48.0
	HolE [45]	211	92.5	398	43.2	67	73.9	243	48.2
	TransA [49]	392	94.3	440	47.0	74	80.4	451	49.1
	TransG [15]	345	94.7	355	51.5	**50**	88.2	397	49.3
TransE	**TransE** [12]	251	89.2	743	56.0	125	47.1	347	46.5
	TEKE+TransE [47]	**127**	93.8	470	59.5	79	67.6	171	48.9
	JOINT+TransE (Ours)	149	92.1	505	64.1	85	58.4	151	48.0
	maJOINT+TransE (Ours)	$131^{§\dagger}$	$93.9^{§}$	$470^{§}$	$\mathbf{66.0}^{§\dagger}$	$79^{§}$	$78.7^{§}$	$\mathbf{146}^{§\dagger}$	$48.9^{§}$
TransH	**TransH** [13]	303	86.7	499	34.2	84	58.5	348	45.2
	TEKE+TransH [47]	128	93.6	446	41.1	75	70.4	259	48.8
	JOINT+TransH (Ours)	164	93.1	470	40.5	79	69.4	289	48.5
	maJOINT+TransH (Ours)	138	94.8	446	43.9	75	75.1	247	49.4
TransR	**TransR** [14]	219	91.7	464	38.3	78	65.5	252	47.8
	TEKE+TransR [47]	203	92.3	470	40.0	79	68.5	233	48.1
	JOINT+TransR (Ours)	208	93.2	374	39.8	82	68.1	239	48.6
	maJOINT+TransR (Ours)	$189^{\star\dagger}$	$\mathbf{95.1}^{\star\dagger}$	$\mathbf{321}^{\star\dagger}$	$41.1^{\star}$	78	$70.3^{\star}$	$217^{\star\dagger}$	$\mathbf{49.5}^{\star}$
ComplEx	**ComplEx** [16]	219	94.7	5261	49.1	78	84.0	254	41.9
	JONIT+ComplEx (Ours)	206	94.9	4687	50.6	63	86.7	237	49.4
	maJOINT+ComplEx (Ours)	$184^{‡\dagger}$	$\mathbf{95.1}^{‡\dagger}$	$4474^{‡}$	$59.6^{‡\dagger}$	54	$\mathbf{88.3}^{‡\dagger}$	$212^{‡\dagger}$	$\mathbf{49.5}^{‡\dagger}$

TransH [13] and **ComplEx** [16], our **maJOINT+TransH** and **maJOINT+ComplEx** improves the average accuracy by 28.38% and 5.12%, respectively for metric HIT@10.

Interestingly, the performance of the conventional **TransR** [14] is not being as good as that of conventional **ComplEx** [16], however our jointly-learning mechanism make it be a match for **ComplEx**. Actually, the main intuition behind the proposed mutual 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).

Moreover, note that, we use two recent benchmark datasets WN18RR and FB15k-237 here. These two datasets are created to avoid reversible relation problems, thus the prediction task becomes more realistic and hence more challenging. The experimental results support this phenomenon: (i) the performance of most of the comparative models degrades on these dataset; (ii) the proposed text-enhanced model achieves the optimal results on all the metric (e.g., **maJOINT+TransR** and **maJOINT+TransE** on dataset WN18RR).

D. TRIPLE CLASSIFICATION TASK

Generally, the triple classification is a classical task in knowledge base embedding, which aims at predicting whether a given triple (h, r, t) is correct or not [12], [17], [47]. Following [20], [60], our evaluation protocol is the same as prior studies. Besides, WN11 and FB13 are the benchmark datasets for this task, and binary classification accuracy (%) is used as the evaluation metric here. Evaluation of classification needs negative labels. Especially, the datasets above mentioned have already been built with negative triples, where each correct triple is corrupted to get one negative triple.

TABLE 4. Performance of triple classification task on dataset WN11 and dataset FB13 (Accuracy(%)). The superscript §, ⋆, † and ‡ respectively denote statistically significant improvements over **TransE** [12], **TransR** [14], **TransG** [15] and **ComplEx** [16] ($p < 0.05$).

Models		WN11	FB13	AVG.
	TransD [43]	86.4	**89.1**	87.8
	HolE [45]	86.1	81.9	84.0
	TransA [49]	83.2	87.3	85.3
	TransG [15]	87.4	87.3	87.4
TransE	**TransE** [12]	75.9	81.5	78.7
	TEKE+TransE [47]	84.1	75.1	79.6
	JOINT+TransE (Ours)	83.7	76.0	79.8
	maJOINT+TransE (Ours)	$87.2^{§}$	$86.9^{§}$	$87.1^{§}$
TransH	**TransH** [13]	78.8	83.3	81.1
	TEKE+TransH [47]	84.8	84.2	84.5
	JOINT+TransH (Ours)	85.4	83.3	84.3
	maJOINT+TransH (Ours)	87.3	86.9	87.1
TransR	**TransR** [14]	85.9	82.5	84.2
	TEKE+TransR [47]	86.1	81.6	83.7
	JOINT+TransR (Ours)	85.9	85.0	85.5
	maJOINT+TransR (Ours)	$87.1^{\star}$	$86.0^{\star}$	$86.6^{\star}$
ComplEx	**ComplEx** [16]	86.2	85.7	86.0
	JOINT+ComplEx (Ours)	86.5	87.7	87.1
	maJOINT+ComplEx (Ours)	$\mathbf{88.3}^{‡\dagger}$	$87.9^{‡}$	$\mathbf{88.1}^{‡\dagger}$

We leverage the proposed jointly-learning model with mutual attention mechanism, for boosting the original **TransE** [12], **TransH** [13], **TransR** [43] and **ComplEx** [16], and compare with their enhanced variant with **TEKE** [47]. Besides, similar to previous evaluation section, **TransD** [43], **TransA** [48], **HolE** [45] and **TransG** [15] are also introduced for the contrast experiments.

As all methods use the same datasets, we directly copy the results of different methods from the previous literature (such as [60]), and the overall results of triple classification task are listed in Table 4. We have tried several settings on the validation dataset to get the best configuration, the optimal configurations are: For dataset WN11, (i) the learning rate for $\mathcal{L}_{\theta_{E_G},\theta_{R_G}}(G)$ in E.q. (2) is 0.0005, (ii) the learning rate for $\mathcal{L}_{\theta_{R_C}}(G, C)$ in E.q. (2) is 0.001, (iii) the vector dimension k is 200, (iv) the harmonic factors α and λ in E.q. (2) are set as 0.00005 and 0.0001 respectively, and (v) the margin

parameters μ in E.q. (11) and γ in E.q. (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 E.q. (2) is 0.001, (ii) the learning rate for $\mathcal{L}_{\theta_{R_C}}(G,C)$ in E.q. (2) is 0.001, (iii) the vector dimension k is 250, (iv) the harmonic factors α and λ in E.q. (2) are set as 0.00005 and 0.0001 respectively, and (v) the margin parameters μ in E.q. (11) and γ in E.q. (15) are set as 5 and 4 respectively.

From Table 4, we observe that: the proposed text-enhanced knowledge graph representation learning model yields the best average accuracy in most cases, illustrating the effectiveness of the mutual attention mechanism. Compared with the conventional **TransE** [12], **TransH** [13] and **ComplEx** [16], our **maJOINT+TransH** and **maJOINT+ComplEx** improves the average accuracy by 10.67%, 7.40% and 2.21%, respectively. This proves the theoretical analysis and the effectiveness of the proposed jointly-learning representation model with with mutual attention, which could utilize accurate textual information enhance the knowledge representations of a triple. Furthermore, compared with **JOINT+TransE** and **JOINT+TransH**, the mutual attention variants **maJOINT+TransE** and **maJOINT+TransH** perform better and especially improve the average accuracy by 9.15% and 3.32%, respectively. The results verifies that it is critical to introduce the mutual attention mechanism (details in Section III-E) to mutually reinforce the relationship among the knowledge graph representation learning (details in Section III-C) and textual relation representation learning (details in Section III-D), as described in Figure 1.

V. DISCUSSION

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. Recently, it has gained lots of interests to jointly learn the embeddings of knowledge graph (KG) and text information. previous work fails to incorporate the complex structural signals from structure representation and semantic signals from text representation. We argue 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). Hence, to fully utilize the mutually reinforcing relationship among the knowledge graph representation learning (i.e., structure representation in Section III-C) and textual relation representation learning (i.e., text representation in Section III-D), this paper proposes a novel mutual 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 (details in Section III-E), to be fully integrated. Empirically, we show the proposed text-enhanced knowledge graph representation with mutual attention can improve the performance of the current translation-based knowledge representation models on several benchmark datasets (details in Section IV).

VI. CONCLUSION

In this paper, we propose an accurate text-enhanced knowledge graph representation framework, which can utilize accurate textual information enhance the knowledge representations of a triple, and can work well with non-strictly aligned data through a mutual attention model between KG and text. Experiment results show that our method can achieve the state-of-the-art performance, and significantly outperforms previous text-enhanced knowledge representation models. And the mutual attention between relation mentions and entity descriptions can significantly improve the performance of knowledge representation. This paper achieves new state-of-the-art performances on link prediction and triple classification tasks over most widely used benchmarks. For future work, we want to further exploit entity types and logic rules as constraints to further improve knowledge representations.

REFERENCES

[1] C. Unger, L. Bühmann, J. Lehmann, A.-C. Ngonga Ngomo, D. Gerber, and P. Cimiano, "Template-based question answering over RDF data," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, 2012, pp. 639–648.

[2] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, Prague, Czech Republic, Jun. 2007, pp. 708–716.

[3] Y. Wang, H. Zhang, Y. Liu, and H. Xie, "KG-to-text generation with slot-attention and link-attention," in *Proc. Natural Language Process. Chinese Comput. (NLPCC)*. 2019, pp. 223–234.

[4] B. Shi and T. Weninger, "Fact checking in heterogeneous information networks," in *Proc. 25th Int. Conf. Companion World Wide Web (WWW Companion)*, 2016, pp. 101–102.

[5] H. Huang, Y. Wang, C. Feng, Z. Liu, and Q. Zhou, "Leveraging conceptualization for short-text embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1282–1295, Jul. 2018.

[6] Y. Wang, H. Huang, and C. Feng, "Query expansion based on a feedback concept model for microblog retrieval," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 559–568.

[7] T. Yi, A. T. Luu, and S. C. Hui, "Non-parametric estimation of multiple embeddings for link prediction on dynamic knowledge graphs," in *Proc. 31st Conf. Artif. Intell.*, 2017, pp. 1243–1249.

[8] A. Bordes, J. Weston, R. Collobert, and Y. Bengio, "Learning structured embeddings of knowledge bases," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, San Francisco, CA, USA, Aug. 2011, pp. 301–306.

[9] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in *Proc. AISTATS*, 2012, pp. 127–135.

[10] Y. Lin, Z. Liu, H. Luan, M. Sun, S. Rao, and S. Liu, "Modeling relation paths for representation learning of knowledge bases," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1–10.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 3111–3119.

[12] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2787–2795.

[13] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1112–1119.

[14] Y. Lin, Z. Liu, X. Zhu, X. Zhu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2181–2187.
[15] H. Xiao, M. Huang, and X. Zhu, "TransG: A generative model for knowledge graph embedding," in *Proc. Meeting Assoc. Comput. Linguistics*, 2016, pp. 2316–2325.
[16] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1–10.
[17] Y. Wang, Y. Liu, H. Zhang, and H. Xie, "Leveraging lexical semantic information for learning concept-based multiple embedding representations for knowledge graph completion," in *Proc. APWeb/WAIM*, 2019, pp. 382–397.
[18] J. Xu, C. Kan, X. Qiu, and X. Huang, "Knowledge graph representation with jointly structural and textual encoding," 2016, *arXiv:1611.086610*. [Online]. Available: https://arxiv.org/abs/1611.08661
[19] K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon, "Representing text for joint embedding of text and knowledge bases," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1499–1509.
[20] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 926–934.
[21] J. Wu, R. Xie, Z. Liu, and M. Sun, "Knowledge representation via joint learning of sequential text and knowledge graphs," 2016, *arXiv:1609.07075*. [Online]. Available: http://arxiv.org/abs/1609.07075
[22] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph and text jointly embedding," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1591–1601.
[23] S. Riedel, L. Yao, A. McCallum, and B. M. Marlin, "Relation extraction with matrix factorization and universal schemas," in *Proc. HLT-NAACL*, 2013, pp. 1–11.
[24] J. Weston, A. Bordes, O. Yakhnenko, and N. Usunier, "Connecting language and knowledge bases with embedding models for relation extraction," in *Proc. EMNLP*, 2013, pp. 1–6.
[25] Y. Wang, H. Huang, C. Feng, Q. Zhou, J. Gu, and X. Gao, "CSE: Conceptual sentence embeddings based on attention model," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 505–515.
[26] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," 2017, *arXiv:1702.00887*. [Online]. Available: https://arxiv.org/abs/1702.00887
[27] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for RNN/CNN-free language understanding," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2017, pp. 5446–5455.
[28] T. Shen, T. Zhou, G. Long, J. Jiang, and C. Zhang, "Bi-directional block self-attention for fast and memory-efficient sequence modeling," 2018, *arXiv:1804.00857*. [Online]. Available: https://arxiv.org/abs/1804.00857
[29] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 2124–2133.
[30] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. Joint Conf. 47th Annu. Meeting (ACL), 4th Int. Joint Conf. Natural Lang. Process. (AFNLP)*, vol. 2, 2009, pp. 1003–1011.
[31] R. Xie, Z. Liu, J. J. Jia, H. Luan, and M. Sun, "Representation learning of knowledge graphs with entity descriptions," in *Proc. AAAI*, 2016, pp. 2659–2665.
[32] P. Verga and A. McCallum, "Row-less universal schema," in *Proc. 5th Workshop Automated Knowl. Base Construct.*, 2016, pp. 1–6
[33] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia—A crystallization point for the Web of data," *J. Web Semantics*, vol. 7, no. 3, pp. 154–165, Sep. 2009.
[34] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2008, pp. 1247–1250.
[35] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A Web-scale approach to probabilistic knowledge fusion," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 601–610.
[36] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in *Proc. ECML-PKDD*, 2014, pp. 165–180.
[37] H. Hai and C. Liu, "Query evaluation on probabilistic RDF databases," in *Proc. Int. Conf. Web Inf. Syst. Eng.*, 2009.
[38] M. Nickel and V. Tresp, "Querying factorized probabilistic triple databases," in *Proc. Int. Semantic Web Conf.*, 2014, pp. 114–129.
[39] L. Getoor and B. Taskar, "Introduction to statistical relational learning," *J. Royal Stat. Soc.*, vol. 173, no. 1, pp. 934–935, 2007.
[40] T. Franz, A. Schultz, S. Sizov, and S. Staab, "Triplerank: Ranking semantic Web data by tensor decomposition," in *Proc. Int. Semantic Web Conf.*, 2009, pp. 213–228.
[41] M. Nickel, V. Tresp, and H. P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2011, pp. 809–816.
[42] I. Sutskever, R. Salakhutdinov, and J. B. Tenenbaum, "Modelling relational data using Bayesian clustered tensor factorization," in *Proc. NIPS*, 2009, pp. 1821–1828.
[43] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2015, pp. 687–696.
[44] Y. Wang, H. Zhang, and H. Xie, "Geography-enhanced link prediction framework for knowledge graph completion," in *Proc. CCKS*, 2019, pp. 198–210.
[45] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1955–1961.
[46] R. Xie, Z. Liu, and M. Sun, "Representation learning of knowledge graphs with hierarchical types," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2965–2971.
[47] Z. Wang and J. Li, "Text-enhanced representation learning for knowledge graph," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 4–17.
[48] H. Xiao, M. Huang, Y. Hao, and X. Zhu, "TransA: An adaptive approach for knowledge graph embedding," 2015, *arXiv:1509.05490*. [Online]. Available: http://arxiv.org/abs/1509.05490
[49] D. Q. Nguyen, K. Sirts, L. Qu, and M. Johnson, "STransE: A novel embedding model of entities and relationships in knowledge bases," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1–8.
[50] Y. Wang, H. Zhang, Y. Li, and H. Xie, "Simplified representation learning model based on parameter-sharing for knowledge graph completion," in *Proc. CCIR*, 2019, p. 67–78.
[51] D. Sorokin and I. Gurevych, "Context-aware representations for knowledge base relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1784–1789.
[52] J. Xu, X. Qiu, K. Chen, and X. Huang, "Knowledge graph representation with jointly structural and textual encoding," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1–7.
[53] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin, "Classifying relations via long short term memory networks along shortest dependency paths," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1785–1794.
[54] M. Xiao and C. Liu, "Semantic relation classification via hierarchical recurrent neural network with attention," in *Proc. COLING*, 2016, pp. 1254–1263.
[55] A. A. Ding, C. Chen, and T. Eisenbarth, "Simpler, faster, and more robust t-test based leakage detection," in *Proc. COSADE*, 2015, pp. 163–183.
[56] K. R. B. Jankowski, K. J. Flannelly, and L. T. Flannelly, "The *t*-test: An influential inferential tool in chaplaincy and other healthcare research," *J. Health Care Chaplaincy*, vol. 24, no. 1, p. 30, 2018.
[57] K. Toutanova and D. Chen, "Observed versus latent features for knowledge base and text inference," in *Proc. 3rd Workshop Continuous Vector Space Models Compositionality*, 2015, pp. 57–66.
[58] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D knowledge graph embeddings," in *Proc. AAAI*, 2017, pp. 1811–1818.
[59] Y. Wang, "Short-text conceptualization based on a co-ranking framework via lexical knowledge base," in *Proc. CCL*, 2019, pp. 281–293.
[60] B. An, B. Chen, X. Han, and L. Sun, "Accurate text-enhanced knowledge graph representation learning," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 1, 2018, pp. 745–755.
[61] D. Q. Nguyen, T. Vu, T. D. Nguyen, D. Q. Nguyen, and D. Phung, "A capsule network-based embedding model for knowledge graph completion and search personalization," in *Proc. Conf. North (NAACL-HLT)*, 2019, pp. 2180–2189.
[62] J. Mei, R. Zhang, Y. Mao, and T. Deng, "On link prediction in knowledge bases: Max-K criterion and prediction protocols," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2018, pp. 755–764.

YASHEN WANG received the B.E. and Ph.D. degrees in computer science and technology from the Beijing Institute of Technology, Beijing. He is currently an Associate Professor with the National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC). His research interests include natural language processing, knowledge graph, social media analysis, and information retrieval.

HUANHUAN ZHANG received the master's degree from the Beijing Institute of Technology, in 2016. She works in the fields of natural language processing and data mining. Her current research interests include knowledge graph and information extraction.

GE SHI received the M.S. degree in computer science from the Beijing Institute of Technology, China, in 2016, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Technology. His research interests include few-shot learning and domain adaptation and their application in relation extraction.

ZHIRUN LIU received the master's degree from the Beijing Institute of Technology, in 2016. He works in the fields of machine learning and natural language processing. His current research interests include knowledge graph and question answering.

QIANG ZHOU received the master's degree from the Beijing Institute of Technology, in 2016. He works in the field of search advertising. His current research interests include natural language processing and deep learning.