Accurate Text-Enhanced Knowledge Graph Representation Learning

Bo An[1,2], **Bo Chen**[1,2], **Xianpei Han**[1], **Le Sun**[1]
[1]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{anbo, chenbo, xianpei, sunle}@iscas.ac.cn

Abstract

Previous representation learning techniques for knowledge graph representation usually represent the same entity or relation in different triples with the same representation, without considering the ambiguity of relations and entities. To appropriately handle the semantic variety of entities/relations in distinct triples, we propose an accurate text-enhanced knowledge graph representation learning method, which can represent a relation/entity with different representations in different triples by exploiting additional textual information. Specifically, our method enhances representations by exploiting the entity descriptions and triple-specific relation mention. And a mutual attention mechanism between relation mention and entity description is proposed to learn more accurate textual representations for further improving knowledge graph representation. Experimental results show that our method achieves the state-of-the-art performance on both link prediction and triple classification tasks, and significantly outperforms previous text-enhanced knowledge representation models.

1 Introduction

Knowledge graphs such as Freebase (Bollacker et al., 2008), YAGO (Suchanek et al., 2007) and WordNet (Miller, 1995) are among the most widely used resources in NLP applications. Typically, a knowledge graph consists of a set of triples $\{(h, r, t)\}$, where h, r, t stand for head entity, relation and tail entity respectively.

Learning distributional representation of knowledge graph has attracted many research attentions in recent years. By projecting all elements in a knowledge graph into a dense vector space, the semantic distance between all elements can be easily calculated, and thus enables many applications such as link prediction and triple classification (Socher et al., 2013).

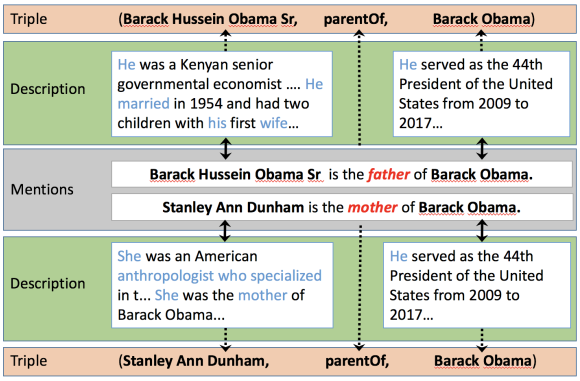
Triple
(Barack Hussein Obama Sr,
parentOf,
Barack Obama)
Description
He was a Kenyan senior governmental economist .... He married in 1954 and had two children with his first wife...
He served as the 44th President of the United States from 2009 to 2017...
Mentions
Barack Hussein Obama Sr is the father of Barack Obama.
Stanley Ann Dunham is the mother of Barack Obama.
Description
She was an American anthropologist who specialized in t... She was the mother of Barack Obama...
He served as the 44th President of the United States from 2009 to 2017...
Triple
(Stanley Ann Dunham,
parentOf,
Barack Obama)

Figure 1: A demonstration of our accurate text-enhanced model. The meanings of relation parentOf in different triples are distinguished by their entity descriptions, and the relation parentOf emphasizes words which describe its social relationship in entity descriptions.

Recently, translation-based models, including TransE (Bordes et al., 2013), TransH (Wang et al., 2014), TransD (Ji et al., 2015) and TransR (Lin et al., 2015b), have achieved promising results in distributional representation learning of knowledge graph. ComplEx (Trouillon et al., 2016) has achieved the state-of-the-art performance on multiple tasks, such as triple classification and link prediction. Unfortunately, all of these methods only utilize the structure information of knowledge graph, which inevitably suffer from the sparseness and incompleteness of knowledge graph. Even worse, structure information usually cannot distinguish the different meanings of relations and entities in different triples.

To address the above problem, additional information is introduced to enrich the knowledge representations, including entity types and logic rules. However, most researches of this line are limited by manually constructed logic rules, which

Proceedings of NAACL-HLT 2018, pages 745–755
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

are knowledge graph sensitive and require the expert knowledge. Another type of widely used resources is textual information, such as entity descriptions and words co-occurrence with entities (Socher et al., 2013; Wang et al., 2014; Zhong et al., 2015).

The main drawback of the above methods is that they represent the same entity/relation in different triples with a unique representation. Unfortunately, by detailed analyzing the triples in knowledge graph, we find two problems of the unique representation: (1) Relations are ambiguous, i.e., the accurate semantic meaning of a relation in a specific triple is related to the entities in the same triple. For example, the relation "parentOf" may refer to two different meanings of (i.e., "*father*" and "*mother*"), depending on the entities in triples. (2) Because different relations may concern different attributes of an entity, the same entity may express different aspects in different triples. For example, different words in the description of "*Barack Obama*" should be emphasized by relations "parentOf" and "professionOf". The ambiguity of entity/relation has been considered as one of the primary reasons why translation-based models cannot handle 1-to-N, N-to-1 and N-to-N categories of relations (Wang et al., 2014). Wang et al. (2016) tried to solve the two issues using words co-occurrence with the entities in the same sentences. Despite its apparent success, there remains a major drawback: this method suffers from noisy text, which reduces the value of textual information.

To solve above problems, this paper proposes an accurate text-enhanced knowledge representation model, which can enhance the representations of entities and relations by incorporating accurate textual information for each triple. To learn the representation of a given triple, we first extract its accurate relation mentions from text corpus, which reflect the specific relationship between its head entity and tail entity. Then a mutual attention mechanism between relation mention and entity descriptions (extracted from knowledge graph), is introduced to enhance the representations of entities and relations. For example, the two triples in Figure 1 have the same "parentOf" relationship, but have different underlying semantics "*was the father of*" and "*was the mother of*" respectively. Besides, our mutual attention mechanism enables knowledge representation focusing more on related information from text information. For example, the "parentOf" relation will concern more about the social relations and gender attributes of a person, rather than his/her jobs, which are also contained in its descriptions. And such a relation-specific entity description will make an entity has more appropriate, relation-specific representations in different triples.

Concretely, we employ BiLSTM model (Schuster and Paliwal, 1997; Graves and Schmidhuber, 2005) with mutual attention mechanism (Zhou et al., 2016) to learn representations for relation mentions and entity descriptions. Specifically, in order to generate triple-specific textual representation of entities and relation, a mutual attention mechanism is proposed to model relation between entity descriptions and relation mention of one triple. Then the learned textual representations are incorporated with previous traditional transition-based representations, which are, learned from structural information of knowledge graph, directly to obtain enhanced triple specific representations of elements.

We evaluate our method on both link prediction task and triple classification task, using benchmark datasets from Freebase[1] and WordNet[2] with the text corpus. Experimental results show that, our model achieves the state-of-the-art performance, and significantly outperforms previous text-enhanced models.

The main contributions are threefold: (i) To the best of our knowledge, this is the first work which simultaneously exploits both relation mention and entity description to handle the ambiguity of relations and entities (Section 3). (ii) We propose a mutual attention mechanism which exploits the textual representations of relation and entity to enhance each other (Section 3.2). (iii) This paper achieves new state-of-the-art performances on triple classification tasks over two most widely used benchmarks (Section 4).

2 Related Work

Currently, a lot of structural-based knowledge representation learning methods have been proposed for knowledge graph completion, including Bilinear Model (Sutskever, 2009), Distance Model (Bordes et al., 2011), Unstructured Model (Bordes et al., 2012), Neural Tensor Network (Socher

[1] http://www.freebase.com

[2] http://www.princeton.edu/wordnet

et al., 2013), Single Layer Model (Socher et al., 2013). And many translation-based methods are introduced, including TransE (Bordes et al., 2013) and its extensions like TransH (Wang et al., 2014), TransD (Ji et al., 2015), TransR (Lin et al., 2015b). Xiao et al. (2016a) proposed a manifold-based embedding principle to deal with the overstrict geometric form of translation-based assumption. Trouillon et al. (2016) employed complex value embeddings to understand the structural information.

In recent years, many methods improve the knowledge representation by exploiting additional information. For example, both the path information and logic rules have been proved to be beneficial for knowledge representation (Lin et al., 2015a; Toutanova et al., 2016; Xiong et al., 2017; Xie et al., 2016; Xu et al., 2016).

One other direction to enhance knowledge representation is to utilize entity descriptions of entities and relations. Socher et al. (2013) proposed a neural tensor network model which enhances an entity's representation using the average of the word embeddings in its name. Wang et al. (2014) proposed a model which combines entity embeddings with word embeddings using its names and Wikipedia anchors. Zhong et al. (2015) further improved the model of Wang et al. (2014) by aligning entity and text using entity descriptions. Zhang et al. (2015) proposed to model entities with word embeddings of entity names or entity descriptions. Xie et al. (2016) proposed a model to learn the embeddings of a knowledge graph by modelling both knowledge triples and entity descriptions. Xu et al. (2016) learns different representations for entities based on the attention from relation. The textual mentions of relations are also explored by Fan et al. (2014). The universal schema based models (Riedel et al., 2013; Toutanova et al., 2015) enhance knowledge representation by incorporating textual triples, which assume that all the extracted triples express a relationship between the entities and they treat each pattern as a separate relation. The main drawback of these methods is that they assume all the relation mentions will express relationship between entity pairs, which inevitably introduces a lot of noisy information. For example, the sentence "`Miami Dolphins in 1966 and the Cincinnati Bengals in 1968`" does not express any relationship between "`miami_dolphins`" and "`cincinnati_bengals`". Even worse, the diversity of language often leads to the data sparsity problem.

To resolve the ambiguity of entities and relations in different triples (i.e., a relation/entity may have different meanings in different triples), Xiao et al. (2016b) proposed a generative model to handle the ambiguous relations. Wang et al. (2016) extended the translation-based models by textual information, which assigns a relation with different representations for different entity pairs, using words co-occurred with both entities in a triple. However, the words co-occur with an entity pair nay also not express the meanings of the relation between them, which will inevitably introduce noisy information for the specific triple. Compared with these methods the main advantages of our methods are: (i) We filters out noisy textual information for accurate enrich knowledge representation. (ii) We simultaneously take the ambiguity of entities and relations in various triples into consideration.

3 Accurate Text-enhanced Knowledge Graph Representation

This section presents our accurate text-enhanced knowledge graph representation learning framework. We first describe how to extract accurate textual information for a given triple, and then we propose a textual representation learning model, which generates textual representations for both entities and relation in a specific triple. Finally, we describe how to enhance knowledge representations based on the textual representations.

The framework of the proposed approach is illustrated in Figure 2.

3.1 Text Information Extraction

Given a triple, our method will first extract accurate textual mentions of its relation from a text corpus. For example, we will extract the relation mention "*Barack Hussein Obama Sr was the father of Barack Obama.*" for the triple *(Barack Hussein Obama Sr, parentOf, Barack Obama)*]]. We collect relation mentions by two steps: (1) Entity linking: linking entity names in a text corpus to entities in a knowledge graph. (2) Relation mention extraction: collecting accurate relation mentions which express the meanings of the relation in a given triple.

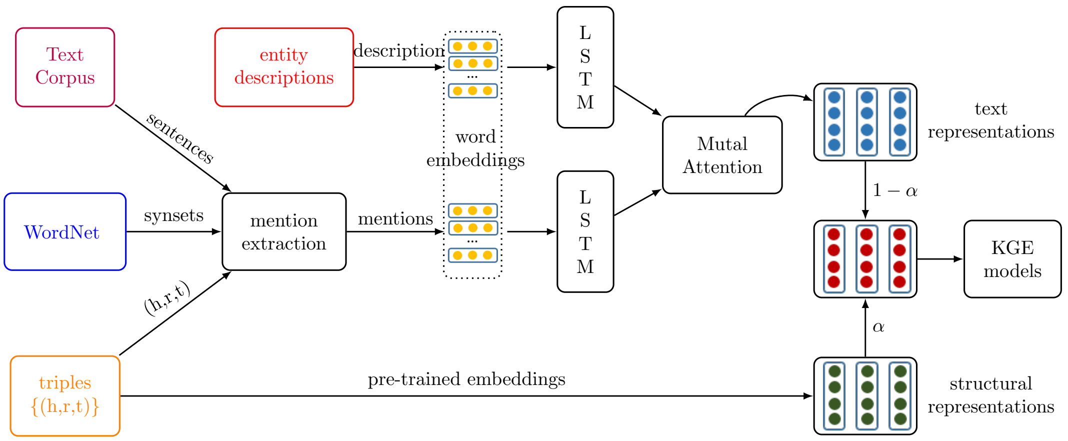

Figure 2: Framework of the proposed approach.

Entity Linking. Given a sentence $D = (w_1, w_2, ..., w_n)$, and an entity set $E = (e_1, e_2, ..., e_m)$, we first recognize entities of E in D to construct a new sentence $D' = (w_1, ..., e_1, ..., e_m, ..., w_n)$, where w_i represents the ith word in D and e_j corresponds to the jth entity in E. There are many general entity linking tools can be used for this purpose. The proposed method employs a simple and precise method to link entities of Freebase and WordNet as Wang et al. (2016). Concretely, we link a Wikipedia inner-link as an entity of Freebase if they have the same titles, and link a word in the corpus as a WordNet entity if the word belongs to one of its synsets.

Relation Mention Extraction. To extract accurate relation text mentions for a specific triple, we first collect all sentences containing both entities of the triple as candidate text mentions. And then, we calculate the similarity between a text mention and the relation based on WordNet. For example, for the triple of *(Steve Jobs, /people/person/parents, Paul Jobs)*, we treat a sentence as its accurate relation mention only if the sentence contains both of its entities and at least one hyponym/synonyms word of the relation. We collect accurate relation mentions for triples in WordNet in a similar way.

In this way, we can extract accurate relation mentions for triples with high precision. However, if a relation mention doesn't contain any hyponym/synonym words of the relation, our method would be unable to identify it. For example, the sentence "*In 1961 Obama was born in Hawaii, US*" expresses the meanings of `/people/person/nationality` in the triple `(Barack Obama, /people/person/nationality, USA )` but without any words belonging to the hyponym or synonyms of "*nationality*". For this, we further employ word embeddings to compute the similarity. Concretely, we represent a relation by averaging the pre-trained word embeddings of its last two words. Then we extract a sentence as an accurate relation mention of a given triple if the similarity between a word in the sentence and the relation representation is above a threshold, with the similarity between a word and a relation is calculated by the cosine similarity of their representations.

3.2 Learning Textual Representation

As mentioned above, the underlying semantics of entities and relations vary from different triples, and different attributes of an entity are concerned by different relations. In this section, we first utilize BiLSTM to encode relation mentions and entity descriptions. And then, we propose a mutual attention mechanism to learn more accurate text representations of relations and entities. Our model contains four layers including Embedding layer, BiLSTM layer and Mutual Attention Layer, and the details of these layers are described as follows.

Embedding Layer. To learn the distributional representation of relation mentions and entity descriptions, we convert words into distributional representations based on lookup word embeddings matrix (Mikolov et al., 2013). Concretely, given a relation mention $m = \{w_1, w_2, w_3, ..., w_n\}$, we transform the word w_i into its distributional representation $\vec{e_i} \in d^w$ using a word embeddings matrix. We use the same pre-trained word embeddings as input for the BiLSTM networks of relation mentions and entity descriptions.

BiLSTM Layer. To learn the representation of text mentions, we utilize a BiLSTM (Long Short-Term Memory) (Hochreiter and Schmidhuber, 1997; Le and Zuidema, 2015; Zhou et al., 2016) model to compose the words in a sequence into the distributional representation. Concretely, we employ a two layer Bidirectional LSTM network to generate text representations. The detailed description of LSTM is presented in (Hochreiter and Schmidhuber, 1997). Two different BiLSTM networks are employed to encode relation mentions and entity descriptions respectively.

Mutual Attention Layer. Attention based neural networks have recently achieved success in a wide range of tasks, including machine translation, speech recognition and paraphrase detection (Luong et al., 2015; Yang et al., 2016; Yin et al., 2016; Vaswani et al., 2017). In this paper, we introduce a mutual attention to improve text representations. Given a triple, the goal of our mutual attention mechanism is two-fold. On one hand, our model wants to identify words in relation mention associated with the entity descriptions in the same triple. On the other hand, our model wants to recognize words in entity descriptions which are emphasized by its relation. To achieve the above goal, we first infer the representations of entity descriptions using relation representation as attention:

$$a_i(e) = \frac{exp(score(\vec{h_i}, \vec{r'}))}{\sum_{i'} exp(score(\vec{h_{i'}}, \vec{r'}))} \tag{1}$$

$$score(\vec{h_i}, \vec{r'}) = \vec{h_i}^T W_e \vec{r'} \tag{2}$$

where $\vec{r'} \in d^w$ is the representation of the relation mention by averaging all the hidden vectors of BiLSTM, $\vec{h_i}$ is the hidden representation of w_i, and $W_e \in d^{w \times 2 \times h}$ is a trained parameter matrix. The relation-sensitive representation of the entity description is generated as follows:

$$\vec{e^*} = tanh(\vec{a_e}^T H_e) \tag{3}$$

where $\vec{a_e} \in d^m$ is the relation-specific attention vector over the words in the entity description, d^m is the length of the description, $H_e \in d^{m \times h}$ is the hidden representation matrix generated by BiLSTM, and $\vec{e^*} \in d^h$ is the representation of the description. In this way, we learn the representations of entity descriptions of head entity $\vec{e_h^*} \in d^h$ and tail entity $\vec{e_t^*} \in d^h$ with the attention from relation representation.

The above two entity description representations are utilized as the attention for learning the triple-sensitive relation mention representation as follows:

$$\vec{e} = \vec{e_h^*} + \vec{e_t^*} \tag{4}$$

$$a_i(r) = \frac{exp(score(\vec{h_i}, \vec{e}))}{\sum_{i'} exp(score(\vec{h_{i'}}, \vec{e}))} \tag{5}$$

$$score(\vec{h_i}, \vec{e}) = \vec{h_i}^T W_r \vec{e} \tag{6}$$

where $\vec{e_h^*}$ and $\vec{e_t^*}$ are representations of head entity description and tail entity description respectively, $\vec{h_i}$ is the hidden vector of w_i for each word in the text mention, and $W_r \in d^{w \times 2 \times h}$ is a trained parameter matrix. The representation of the triple-sensitive relation mention is generated as Formula (7):

$$\vec{r^*} = tanh(\vec{a_r}^T H_r) \tag{7}$$

where $\vec{a_r}^T \in d^n$ is the triple-sensitive attention vector over the words in the relation mention, d^n is the length of the relation mention, $H_r \in d^{n \times h}$ is the hidden representation matrix generated by BiLSTM, and $\vec{r^*} \in d^h$ is the representation of the mention. In this way, we learn the triple-attention representation of all text mentions.

3.3 Text-Enhanced Representation Learning

In this section, we introduce how to incorporate the learned textual representations with representations learned from knowledge graph structure using previous methods.

For each given triple and its accurate textual information, we enhance the representations of the relation and entities based on the text representations of entities $\vec{e_h^*} \in d^h$, $\vec{e_t^*} \in d^h$ and relation $\vec{r^*} \in d^h$. Specifically, we enhance the relation and entity representations as follows:

$$Re(\vec{r}_{ate}) = \alpha \cdot Re(\vec{r_{kg}}) + (1-\alpha) \cdot \vec{r^*}, 0 \le \alpha \le 1 \tag{8}$$

$$Re(\vec{h}_{ate}) = \alpha \cdot Re(\vec{h_{kg}}) + (1-\alpha) \cdot \vec{e_h^*}, 0 \le \alpha \le 1 \tag{9}$$

$$Re(\vec{t}_{ate}) = \alpha \cdot Re(\vec{t_{kg}}) + (1-\alpha) \cdot \vec{e_t^*}, 0 \le \alpha \le 1 \tag{10}$$

where α represents the weight factor for the structural representations, $\vec{r_{kg}} \in d^h$, $\vec{h_{kg}} \in d^h$ and $\vec{t_{kg}} \in d^h$ represent the distributional representations of relation r head entity h and tail entity t

learned from structural information of knowledge graph, $\vec{r^*} \in d^h$, $\vec{e_h^*} \in d^h$ and $\vec{e_t^*} \in d^h$ represent the vectors of the text mention, head and tail entity descriptions for the triple, $\vec{r}_{ate} \in d^h$, $\vec{h}_{ate}$ and $\vec{t}_{ate}$ are the accurate text-enhanced representations of relation, head and tail entity, respectively. Note that, we enhance the real part vector of an entity with the textual representation of the entity as Formula (9) and (10), and treat the matrix representation of a relation as a vector with each element the same as the element in diagonal matrix, and then enhance its real part as Formula (8). In this way, we enhance the representation of knowledge graph, and calculate the plausibility of a triple based on their score functions.

If there is no accurate relation mention extracted for a triple, we only utilize the knowledge embeddings to estimate the plausibility of the triple, and the weight factor α is set to 1 in this case. For example, if there is no accurate relation mention extracted for triple `(Su Shi, /people/person/profession, Artist)`, then only its structural representations will be utilized to compute the plausibility of the triple. And α is set to 1 for the triples if none of the entities in it is linked.

3.4 Model Training

In the training process, the *(h, r, t, h_t, r_t, t_t)* tuples are used as supervision, where h_t, r_t and t_t are the description of head entity, relation text mention and the description of tail entity, respectively. Since there are only correct triples in the knowledge graph, following Lin et al. (2015a), we construct the corrupted tuples $(h', r, t', h_t, r_t, t_t) \in KG'$ for a $(h, r, t, h_t, r_t, t_t) \in KG$ by randomly replacing head/tail entity with entities from knowledge graph using Bernoulli Sampling Method (Wang et al., 2014). Furthermore, to train the model of text representation model, we construct the corrupted tuples $(h, r, t, h_t', r_t', t_t') \in KG'$ for a $(h, r, t, h_t, r_t, t_t) \in KG$ by random replacing the text information. We use the following margin-based ranking loss:

$$L = \sum_{q \in KG} \sum_{q' \in KG'} max(0, \gamma + f(q) - f(q')) \quad (11)$$

where f is the score function of our model, and $\gamma > 0$ is the margin between golden tuples and negative tuples, KG is the set of tuples in training dataset, and KG' is the corrupted set of tuples. The parameters of our model are optimized using the stochastic gradient descent (SGD) algorithm. To accelerate the training process and avoid overfitting, we initialize the representations of entities and relations using base models and initialize word representations with the pre-trained word embeddings, and all these embeddings are fine-tuned during training.

4 Experiments

In this section, we first describe the settings in our experiments, and then we conduct experiments of link prediction and triple classification tasks and compare our method with base models and the state-of-the-art baselines.

4.1 Experiment Settings

In this paper, we evaluate our model on four benchmark datasets: WN11, WN18, FB13 and FB15k (Bordes et al., 2013; Socher et al., 2013; Wang et al., 2014). For the text corpus, we use a snapshot of the English Wikipedia (Wiki) (Shaoul and Westbury, 2010) [3] dump in April 2016, which contains more than 1.2 billion tokens. We link entities in the text corpus to entities in Freebase and synsets in WordNet as described above, and replace entities with HEAD_TAG and TAIL_TAG. The text descriptions of entities are freely available[4]. In addition, we pre-process the word-entity corpus, including stemming, lowercasing and removing words with fewer than 5 occurrences. The statistics of the datasets and linked-entities in text corpus are shown in Table 1.

Dataset	WN11	WN18	FB13	FB15K
#Train	112,581	141,442	316,232	483,142
#Valid	2,609	5,000	5,908	50,000
#Test	10,544	5,000	23,733	59,071
# Entities	38,696	40,943	75,043	14,951
# Relations	11	18	13	1,345
#1-to-1	0	0	0	247
#1-to-N	0	0	0	179
#N-to-1	0	0	1	225
#N-to-N	11	18	12	694
#Linked	31,432	34,159	66,328	13,567

Table 1: Statistics of different datasets and the number of entities linked in Wikipedia, *#Linked* represents the number of entities linked in the text corpus. *#N-to-1* is the number of N-to-1 type of relations.

As introduced above, we implement our framework using TransE, TransH, TransR and ComplEx as base models, and evaluate on two classi-

[3] https://www.wikipedia.org/

[4] https://github.com/xrb92/DKRL

cal tasks: link prediction and triple classification. We refer $AATE\_E$ as the proposed model which enhances TransE with accurate textual informations and mutual attention mechanism, and refer ATE_E as the proposed model without mutual attention mechanism to reveal the effect of our attention mechanism.

To speed up training and reduce overfitting, we employ the SkipGram model of word2vec (Mikolov et al., 2013) to pre-train the word embeddings with the dimension of word embeddings is $d^w = 200$, the windows size is 5, the number of iterations is 5, and the number of negative samples is 10. And we pre-train the representations of entities and relations of knowledge graph using the mentioned base models, and the parameters are empirically tuned as follows: the dimension of vectors is $d^{kg} = 200$, the number of epochs is 2000 and the margin is 1.0. We implement our model based on the OpenKE [5] framework.

In our experiments, the hyper-parameters of BiLSTM are empirically set as follows: the number of hidden units is $d^h = 200$, the learning rates for SGD are among $\{0.1, 0.001, 0.0001\}$, the margin λ values are among $\{0.5, 1.0, 2.0\}$ and the batch sizes are among $\{100, 500, 2000\}$. We employ two different BiLSTM networks with the same hyper-parameters to learn the representations of text mentions and entity descriptions. And all the parameters are learned jointly, including BiLSTM networks and knowledge representations.

4.2 Link Prediction

Link prediction aims to predict missing head or tail entity of a triple, which is a widely employed evaluation task for knowledge graph completion models (Bordes et al., 2011; Wang et al., 2016). Concretely, given a head entity `h` (or tail entity `t`) and a relation `r`, the system will return a rank list of candidate entities for tail entity. Following (Bordes et al., 2013; Lin et al., 2015b), we conduct the link prediction task on WN18 and FB15k datasets.

In the testing phase, for each triple (h, r, t), we replace its head/tail entity with all entities to construct candidate triples, and extract text mentions from the text corpus for each candidate triple. Then we rank all these entities in descending order of the scores, which are calculated by our score function. Based on the entity ranking list, we employ two evaluation metrics from (Bordes et al., 2013): (1) mean rank of correct entities (MR); and (2) proportion of correct entities in top-10 rank entities Hit@10 ($Hit10$). A good link predictor should achieve low MR and high Hit@10. We tuned model parameters using validate datasets. We implement our framework using TransE, TransH, TransR and ComplEx as base models, and treat these base models as baselines. Furthermore, we also compare our method with the state-of-the-art results from Unstructured, SME, TransD, TEKE , Jointly (Xu et al., 2016), TransG and Mainifold, and we report the results from their original papers. The overall results are presented in Table 2.

Models		**WN18**		**FB15K**	
		MR	Hit10	MR	Hit10
Others	UnS	304	38.2	154	40.8
	SME	533	74.1	979	6.3
	TransD	212	92.2	91	77.3
	TransG	345	94.7	**50**	**88.2**
	Mainifold	-	**94.9**	-	88.1
Jointly	LSTM	**95**	91.6	90	69.7
	A-LSTM	123	90.9	73	75.5
TransE	TransE	251	89.2	125	47.1
	TEKE_E	127	93.8	79	67.6
	ATE_E	158	91.7	89	57.1
	AATE_E	123	94.1	76	76.1
TransH	TransH	303	86.7	84	58.5
	TEKE_H	128	93.6	75	70.4
	ATE_H	167	92.5	80	68.2
	AATE_H	132	94.0	73	74.6
TransR	TransR	219	91.7	78	65.5
	TEKE_R	203	92.3	79	68.5
	ATE_R	210	92.1	80	67.2
	AATE_R	185	93.7	77	69.4
ComplEx	ComplEx	219	94.7	78	84.0
	ATE_C	217	94.7	61	86.2
	AATE_C	**179**	**94.9**	**52**	**88.0**

Table 2: Evaluation results of link prediction.

From Table 2, we can see that both ATE and AATE models surpass all base models (TransE, TransH, TransR and ComplEx) on all metrics. This result verifies that the textual information is beneficial for structure-based knowledge graph representation learning models. Compared with the ATE models, the AATE models achieve better results on link prediction task, which verifies that the mutual attention between entity description and relation mention is effect for selecting meaningful words and enhancing the learning of knowledge graph representation.

For translation-based models, the proposed method achieves the best result based on TransE.

[5] http://openke.thunlp.org/

This is probably because TransH and TransR have tried to project the entity embeddings into the space of relation space, which may lead to the fact that the text information could not enhance the entity representation directly. In addition, our method implemented based on ComplEx has achieve better performances w.r.t TEKE (Wang et al., 2016) on all metrics, that verifies the importance of filtering out the noisy information.

4.2.1 Analysis on 1-to-N, N-to-1 and N-to-N Relations

To better analyse the effect of textual information for knowledge graph representation learning, this section presents the results of our model on different categories of relations including 1-N, N-1 and N-N on link prediction task. We present the results of our models based on TransE and of all baselines.

From Table 3, we can see that, both of our proposed methods have achieved higher performance over the base model on all types of relations (1-to-N, N-to-1 and N-to-N). In addition, our AATE model achieves better results than the Jointly(A-LSTM) model. Since both of AATE and Joint (A-LSTM) are implemented based on TransE, we verify that the triple-specific relation mention is valuable to improving the knowledge representation. Another reason why our proposed model achieves better results is that the attention from textual representation of relation and entity is more effective than the attention using structural representation for textual representation.

4.2.2 Fault Analysis

To gain more insight, we present a failure analysis to explore possible limitations and weaknesses of our model. In particular, several illustrative triples from the test set of FB15K are listed in Table 4. The tail entities of those triples are failed to be ranked in the top-10 candidates.

It can be seen from Table 4 that, the failures are mostly caused by the data sparsity problem, which results in relatively limited occurrences of entities and relations. All of "`Elementary school`", "`Abugida`", "`interests/collection_category /sub_categories`" and "`martial_arts/ martial_artist/martial_art`" appear less than 4 times in training data. It must also be mentioned that the triple "`(Abugida, language /language_writing_system/ languages, Khmer language)`" is included in the training data. Therefore, we can infer the first triple in Table 4 based on the above triple due to the general logic that "`language/human_language/writing _system`" and "`/language/language _writing_system/languages`" are a pair of inverse relations. Consequently, we believe it is important to incorporate the logic rules into knowledge embeddings, especially for the entities and relations with limited occurrences.

4.3 Triple Classification

In this section, we assess different models on the triple classification task. Triple classification aims to judge whether a given triple (h, r, t) is true fact or not, and it is usually modeled as a binary classification task (Socher et al., 2013; Bordes et al., 2013; Wang et al., 2016). Following Socher et al. (2013) we evaluate different systems on WN11 and FB13 datasets.

Given a triple (h, r, t) and all its accurate relation mentions and entity descriptions of this triple, In our experiments, a triple will be classified as a true fact if the score obtained by function f is below the relation-specific threshold δ_r, otherwise it will be classified as a false fact. The δ_r and the weight factor of α are optimized by maximizing classification accuracy on validation dataset, and different values of δ_r will be set for different relations. We use the same settings as link prediction task, all parameters are optimized on the validation datasets to obtain the best accuracies. We compare our method with all base models and the state-of-the-art performances from TransD, TEKE (Wang et al., 2016), TransG, Mainfold, and we report the best results from their original papers. The results are listed in Table 5.

From Table 5, we can see that: (1) The accurate textual information can consistently increase the accuracies on triple classification task. In all of the four base models, our model achieves significant improvements over TransE, TransH, TransR and ComplEx. This results verify that our method is a useful framework for exploiting textual information to enhance structure-based models; (2) Our method achieves better results on all datasets than TEKE. This result reveals that it is important to filter out the noisy data for knowledge graph representation learning. (3) Compared with the ATE model, our relation-sensitive attention

	Prediction Head (Hits@10)			Prediction Tail (Hits@10)		
Relation Category	1-to-N	N-to-1	N-to-N	1-to-N	N-to-1	N-to-N
#Triples in Test	2,078	6,084	109,526	2,078	6,084	109,526
Jointly(A-LSTM)	95.1	21.1	47.9	30.8	94.7	53.1
TransE	65.7	18.2	47.2	19.7	66.7	50.0
ATE_E	80.2	22.1	47.6	20.3	67.7	60.0
AATE_E	**96.1**	**35.2**	**49.1**	**32.2**	**98.3**	**60.3**

Table 3: Hit@10 of link prediction on different type of relations on FB15k dataset.

No	Head Entity (#)	Relation (#)	Tail Entity (#)
1	Upper Canada College (16)	education/educational_institution/school_type (728)	Elementary school (1)
2	Khmer language (9)	language/human_language/writing_system (41)	Abugida (1)
3	Film (255)	interests/collection_category/sub_categories (3)	Star Wars (31)
4	Jean-Claude Van Damme (28)	martial_arts/martial_artist/martial_art (1)	Taekwondo (155)

Table 4: The triples whose tail entities were failed to be ranked in top 10 candidates, # is the number of occurrences of the entity/relation in the training data.

Models		WN11	FB13	AVG.
Others	TransD	86.4	89.1	87.8
	TransG	87.4	**87.3**	87.4
	Mainfold	87.5	87.2	87.4
TransE	TransE	75.9	81.5	78.7
	TEKE_E	84.1	75.1	79.6
	ATE_E	84.3	75.4	79.9
	AATE_E	86.1	86.4	86.3
TransH	TransH	78.8	83.3	81.1
	TEKE_H	84.8	84.2	84.5
	ATE_H	85.1	83.9	84.5
	AATE_H	86.7	86.2	86.5
TransR	TransR	85.9	82.5	84.2
	TEKE_R	86.1	81.6	83.7
	ATE_R	86.2	84.4	85.3
	AATE_R	86.4	85.2	85.8
ComplEx	ComplEx	86.2	85.7	86.0
	ATE_C	87.2	87.1	87.2
	AATE_C	**88.0**	87.2	**87.6**

Table 5: Evaluation results of triple classification.

model improves the accuracies on all the datasets. We believe this is because mutual attention mechanism can better identify the relation-sensitive words from entity descriptions and extract entity-sensitive words from relation mention.

The results demonstrate that, our method has achieved the best performances on the triple classification task, which verifies that it is critical to filter out noisy text information to determine whether a triple should be added into knowledge graph or not.

5 Conclusions

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 effectively handle the ambiguity of relations and entities through a mutual attention model between relation mentions and entity descriptions. 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. For future work, we want to further exploit entity types and logic rules as constraints to further improve knowledge representations.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants no. 61433015, 61772505 and 61572477, and the Young Elite Scientists Sponsorship Program no. YESS20160177. Moreover, we sincerely thank the reviewers for their valuable comments.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase:a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, Bc, Canada, June*. pages 1247–1250.

Antoine Bordes, Xavier Glorot, and Jason Weston. 2012. Joint learning of words and meaning representations for open-text semantic parsing. In *International Conference on Artificial Intelligence & Statistics*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko.

2013. Translating embeddings for modeling multi-relational data. *NIPS* pages 2787–2795.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning structured embeddings of knowledge bases. In *AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, Usa, August*.

Miao Fan, Qiang Zhou, Emily Chang, and Thomas Fang Zheng. 2014. Transition-based knowledge graph embedding with relational mapping properties. In *PACLIC*. pages 328–337.

A Graves and J Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* 18(56):602–610.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*. pages 687–696.

Bhushan Kotnis and Vivi Nastase. 2017. Learning knowledge graph embeddings with type regularizer. *arXiv preprint arXiv:1706.09278* .

Phong Le and Willem Zuidema. 2015. Compositional distributional semantics with long short term memory. *arXiv preprint arXiv:1503.02510* .

Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015a. Modeling relation paths for representation learning of knowledge bases. *arXiv preprint arXiv:1506.00379* .

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

Minh Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *Computer Science* .

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

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the Acm* 38(11):39–41.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *HLT-NAACL*. pages 74–84.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

R. Socher, D. Chen, C. D. Manning, and A. Y. Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *International Conference on Intelligent Control & Information Processing*. pages 464–469.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago:a core of semantic knowledge. In *International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May*. pages 697–706.

Ilya Sutskever. 2009. Modelling relational data using bayesian clustered tensor factorization. *Advances in Neural Information Processing Systems* pages 1821–1828.

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

Kristina Toutanova, Xi Victoria Lin, Wen-tau Yih, Hoifung Poon, and Chris Quirk. 2016. Compositional learning of embeddings for relation paths in knowledge bases and text. In *ACL2016*. volume 1, pages 1434–1444.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* .

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

Zhigang Wang, Juanzi Li, Zhiyuan LIU, and Jie TANG. 2016. Text-enhanced representation learning for knowledge graph. *To appear in IJCAI* 2016:04–17.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016a. From one point to a manifold: knowledge graph embedding for precise link prediction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, pages 1315–1321.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016b. Transg: A generative model for knowledge graph embedding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 2316–2325.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions. In *AAAI*. pages 2659–2665.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. *arXiv preprint arXiv:1707.0669* .

Jiacheng Xu, Kan Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Knowledge graph representation with jointly structural and textual encoding. *arXiv preprint arXiv:1611.08661* .

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. pages 21–29.

Wenpeng Yin, Hinrich Schtze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Computer Science* .

Huaping Zhong, Jianwen Zhang, Zhen Wang, Hai Wan, and Zheng Chen. 2015. Aligning knowledge and text embeddings by entity descriptions. In *EMNLP*. pages 267–272.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *ACL*. pages 207–212.