applied sciences

MDPI

Article

KGGCN: Knowledge-Guided Graph Convolutional Networks for Distantly Supervised Relation Extraction

Ningyi Mao [1], Wenti Huang [2,*] and Hai Zhong [2]

[1] School of Business, Central South University, Changsha 410083, China; 191611116@csu.edu.cn
[2] School of Computer Science and Engineering, Central South University, Changsha 410083, China; zhonghai@csu.edu.cn
* Correspondence: huangwenti@csu.edu.cn

Abstract: Distantly supervised relation extraction is the most popular technique for identifying semantic relation between two entities. Most prior models only focus on the supervision information present in training sentences. In addition to training sentences, external lexical resource and knowledge graphs often contain other relevant prior knowledge. However, relation extraction models usually ignore such readily available information. Moreover, previous works only utilize a selective attention mechanism over sentences to alleviate the impact of noise, they lack the consideration of the implicit interaction between sentences with relation facts. In this paper, (1) a knowledge-guided graph convolutional network is proposed based on the word-level attention mechanism to encode the sentences. It can capture the key words and cue phrases to generate expressive sentence-level features by attending to the relation indicators obtained from the external lexical resource. (2) A knowledge-guided sentence selector is proposed, which explores the semantic and structural information of triples from knowledge graph as sentence-level knowledge attention to distinguish the importance of each individual sentence. Experimental results on two widely used datasets, NYT-FB and GDS, show that our approach is able to efficiently use the prior knowledge from the external lexical resource and knowledge graph to enhance the performance of distantly supervised relation extraction.

Keywords: relation extraction; knowledge graph embedding; graph convolutional network; attention mechanism

check for updates

Citation: Mao, N.; Huang, W.; Zhong, H. KGGCN: Knowledge-Guided Graph Convolutional Networks for Distantly Supervised Relation Extraction. *Appl. Sci.* **2021**, *11*, 7734. https://doi.org/10.3390/app11167734

Academic Editor: Rafael Valencia-Garcia

Received: 2 July 2021
Accepted: 19 August 2021
Published: 22 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Relation extraction (RE) is a crucial task of natural language processing (NLP), which aims to recognize predefined semantic relations between two marked nominals in texts. Various relations extracted from texts are helpful for knowledge graph (KG) construction, as well as facilitating down-stream tasks that require relational understanding of texts, such as intelligent question–answer [1], biomedical knowledge discovery [2], and dialogue systems [3]. Accurate relation extraction results promote precise text interpretation, discourse processing and higher-level NLP systems. Given a sentence "*Bill Gates co-founded Microsoft with his childhood friend Paul Allen*", the goal of relation extraction is to automatically identify the relation "***founder***" between "***Bill Gates***" and "***Microsoft***" expressed in the sentence. In recent years, words and entities distribution representation learning have made significant progress. Thus, many works utilizing neural network models to deal with the relation extraction task have been proposed [4–6]. The most representative progresses are recurrent neural network (RNN), convolutional neural network (CNN), and other neural network architectures [7–9]. Existing approaches have achieved a great success based on the neural networks. However, most supervised relation extraction models require a large number of training data, which is usually expensive to obtain. To overcome this weakness, distant supervision is introduced to automatically construct large scale datasets [10]. It is under the assumption that if a pair of entities have a relationship in a KG, then all sentence mentioning these entities express this relation. For example, given a triple (e_1, r, e_2) in

the KG, all sentences that mention both entities e_1 and e_2 are regarded as the training instances of relation r. Despite distant supervision paradigm can automatically collect training data for relation extractor, it often suffers from the wrong labeling problem [11]. A pair of entities that appear in a sentence may not express the relation which links them in the KG, they are just related to a same topic. As a result, the distant supervision will inevitably bring noise into the generated training dataset, which drops the performance of relation extraction. Most existing methods alleviate the negative impact of noise by utilizing multi-instance leaning. The multi-instance based relation extraction can be regarded as a bag-level classification task, which allows different sentences to have at most one shared label.

Existing methods for distantly supervised relation extraction have made some progress. Nevertheless, they are still confronted with two challenges. (1) *How to design more effective sentence encoders to generate more expressive sentence-level features*. Previous works for relation extraction considered that all words in a sentence have equal contributions for predicting the relation between two marked entities. In fact, only a few words in a sentence are relevant for determining the relation expressed. Take the aforementioned sentence as an example, "*Bill Gates co-founded Microsoft with his childhood friend Paul Allen*". Obviously, the word "*co-founded*" is of great importance in predicting the relation ***founder*** between "*Bill Gates*" and "*Microsoft*". However, the words "*childhood friend*" have little relevance with that relation. Thus, encoding all words equally in the sentence without any distinction will confuse the feature extractor and degrade the performance of the relation extraction model. (2) *How to make full use of the informative sentences in a bag and integrate them to generate bag-level features for predicting the given relations*. The traditional multi-instance learning model only selects one sentence which has the maximum probability to be a valid candidate for representing the sentence bag [11]. This strategy does not make full use of the supervision information of the sentences in the bag, which may exacerbate the inadequate issue of training data. Calculating the average of all sentences in the bag to obtain the bag-level features is an improved method, but this method will introduce noise due to the existence of false positive instances. Moreover, previous methods for relation extraction only based on the individual semantic features in the textual sentences of the entity mentions. In fact, the human prior knowledge from external lexical resource is critical for reducing the reliance on training data, and it can improve the relation extraction performance.

To address the first challenge, a knowledge-guided graph convolutional network (KG-GCN) based on the word-level attention mechanism is proposed to encode the sentences, which attends to relation indicators that are useful in predicting relations. This is motivated by the fact that each word in a sentence has different importance for relation inference [12]. Specifically, the relation indicators are human prior knowledge obtained from the external lexical resource, which represent the key words and cue phrases of relations in sentences. The sentence encoder utilizes the knowledge attention to calculate the attention weight of each individual word and capture the informative linguistic clues of relations. In this way, the model increases the weights of critical words and cue phrases, while reduces the weights of trivial words. As a result, the critical words and cue phrases will contribute more to sentence encoding, which can form a purified representation for sentences and generate more informative sentence-level features.

To address the second challenge, we propose a knowledge-guided sentence-level attention model to select multiple valid sentences. For the distantly supervised relation extraction, the entities and relations are derived from existing knowledge graph. Thus, the knowledge about these entities and relations can be used as supervisory information to guide the selection of valid sentences. In previous methods, the relations just act as relation labels to specify the class of sentences in the training stage. The structural and semantic information between the entity pair and the relation is completely ignored, which can actually be used as additional knowledge for relation extraction. To this end, our work explores the structural and semantic information of triples from knowledge graph to guide the selection of valid sentences.

Contributions

In this paper, we combine the human prior knowledge obtained from the external lexical resource and the information learned from the training data to improve the performance of distant supervision based relation extraction model. Our main contributions in this paper are: (1) We propose a novel knowledge-guided graph convolutional network based on the word-level attention mechanism. It utilizes the relation indicators obtained from the FrameNet [13] to effectively capture the informative linguistic clues to generate more expressive sentence-level features; (2) We mine the semantic and structural information of triples from the knowledge graph as external knowledge to build a sentence-level attention model. This model can select multiple valid sentences in a bag and make full use of the supervision information of training instances; (3) A triplet embedding model is introduced to augment the interaction between entities and relations, which can help triplets to provide stronger supervisory information.

The rest of the paper is organized as follows. Section 2 covers the related works. In Sections 3 and 4, we formulate the problem formally, and provide our solution for relation extraction. We report the promising experiment results on real-world datasets in Section 5. Finally, we conclude the paper in Section 6.

2. Related Work

In recent years, relation extraction, like big data and cloud computing [14,15], has attracted considerable interest from researchers. The early works mainly focused on the handcrafted feature-based models [16–18] and kernel-based models [19,20]. These methods rely on the NLP tools, which unavoidably leads to error propagation or accumulation. With the development of neural network technology in recent years, a number of relation extraction researches have been proposed to utilize the neural network models [21]. These methods can alleviate the model's dependence on accurate feature matching and have achieved great progress for relation extraction. Liu et al. [22] utilized a simple CNN model that does not even have a pooling layer to extract the features of sentences, which is the first attempt that using a CNN model to extract relations of entity pairs. Zeng et al. [7] incorporated word embedding and position embedding as the input of a CNN model to generate the sentence features. Combining the word embedding and position embedding can obtain more semantic information of the relation mentions. Many works focus on improving the performances of the neural network methods. However, most of these supervised methods require large-scale labeled training data which is expensive to obtain.

In order to address this problem, distant supervision is proposed to automatically generate large-scale labeled training data [10]. The distant supervision is based on the assumption that there is a relation between an entity pair in a knowledge graph, all sentences in a corpus containing the entity pair express this relation. However, this assumption is too strong in practice due to the wrong label problem. To alleviate the impact of the wrong labeled instances in the distant supervision learning, some works have been proposed to use the multi-instance learning method [23,24], which gathers the sentences mentioning the same entity pair into a bag to share a label. The attention mechanism was also introduced in distantly supervised relation extraction in recent years, which can let the neural network models focus on the informative sentences. Guo et al. [25] proposed an attention guided graph convolutional networks to selectively attend to the relevant substructures of dependency trees useful for the relation extraction task. Lin et al. [26] proposed the sentence-level attention over instances in a bag to select the important sentences. Some works combine the word-level and sentence-level attention mechanism to further improve the performance of distantly supervised relation extraction [27,28].

In addition to utilizing the semantic information from sentences, some methods also introduce external information to augment existing relation extraction models. Such as the text descriptions of entities, which can provide helpful supplementary information for relation classification. Ren et al. [29] proposed a new neural relation classification method to integrate the text descriptions of the entities into a deep CNN model for relation

classification. Vashishth et al. [30] proposed to use the side information, such as entity type and relation alias, to enhance the performance of relation extraction. Han et al. [31] proposed a joint representation learning framework to generate the mutual attention between knowledge graphs and texts. This reciprocal attention mechanism can highlight the important features and perform better knowledge graph completion and relation extraction. Zeng et al. [32] modeled the relation path between two entities in a knowledge graph to encode the relational semantics from both direct sentences and inference chains.

Despite the previous works have achieved the state-of-the-art performances, most of these methods consider only the textual information of entity mentions or surface lexical features present in sentences. The semantic information present in external lexical resource is ignored, such as relation indicators in FrameNet and relation facts in existing knowledge graphs, which can provide additional auxiliary information for relation extraction. In this paper, the semantic information is used as prior knowledge to improve the performance of distantly supervised relation extraction.

3. Methodology

In this section, a novel framework is presented for distantly supervised relation extraction, which uses the prior knowledge from FrameNet and knowledge graph to provide hierarchical knowledge attention. We denote a KG as $G = \{(e_1, r, e_2)\}$, which consists of many triples (e_1, r, e_2). Each triple indicates a relation r between entity pair e_1 and e_2.

Given an entity pair (e_1, e_2) in a KG and a training set of sentence bags $D = \{B_1, B_2, ..., B_N\}$, a relation r is defined as a semantic property between the entity pair (e_1, e_2). For distantly supervised relation extraction, all sentences S_i refer to the entity pair (e_1, e_2) are regarded as instances of the relation r. They constitute a instance bag for this entity pair and relation type, denoted as $B_i = \{S_1, S_2, ..., S_n\}$. The target of the distantly supervised relation extraction is to predict the labels for unseen bags. To this end, we need to learn a relation extractor to capture features of the valid sentences in the bag and aggregate them to form the bag-level features. Then we use the bag-level features to train a classifier to predict the relations for the given entity pairs.

Our framework consists of two modules: Sentence embedding module and multi-instance selection module. The sentence embedding module utilizes the relation indicators from lexical resources as prior knowledge to guide the embedding of sentences, which consists of word-level knowledge attention layer and graph convolutional layer. The multi-instance selection module explores the structural and semantic information from KGs as prior knowledge to guide the selection of multiple valid sentences, which consists of knowledge graph embedding and sentence-level knowledge attention layer. The model leverages hierarchical knowledge attention to attend over instances to alleviate different levels of noise, which can generate more expressive relation representations to enhance the relation extraction.

3.1. Sentence Representation

In this paper, we employ a knowledge-guided GCN model to build the context encoder and transform the sentences into low-dimension vectors. The relation indicators extracted from lexical resource are prior knowledge for word-level attention, which guides the GCN model attend to the key words and cue phrases in the sentence embedding procedure.

3.1.1. Generation of Relation Indicator

The relation indicators represent the key words and cue phrases that reference to different relation types. They are prior knowledge for the GCN model to capture the linguistic clues of certain relation in texts, which can be obtained from lexical resource. In this paper, we collect the relation indicators from a large-scale lexical resource called FrameNet. The FrameNet is a publicly available lexical resource, which categorizes words and sentences into high level semantic frames to express different concepts. Each semantic

frame describes a type of relation, event, or object in the form of a conceptual structure, which consists of definition, frame elements, FE core sets, examples, lexical units (LUs), and frame–frame relations. The corresponding semantic frame of the relation ***founder*** is illustrated as Figure 1. There are over 1200 semantic frames and 13,000 LUs in the FrameNet, most of them describe different semantic relations.

Frame Name: Intentionally_Create	
Definition:	The Creator creates a created entity.
Frame Elements (FEs):	Place, Time, Purpose, Components, Participant, Created_entity ...
Lexical units (LUs):	Found, Create, Creation, Establish, Set up, Generate, Produce, Develop ...
Frame-frame relations:	Inherits from: Creating, Intentionally_act Is Inherited by: Building, Manufacturing, ...

Figure 1. Semantic frame of the relation *founder*. The relation indicators are extracted from the Lexical units.

In addition to FrameNet, there are some other popular lexical resources, such as PropBank [33] and VerbNet [34], which can also be utilized to extract the linguistic knowledge of entity relations. However, unlike FrameNet which is semantically motivated and contains lexical units with various part of speeches, PropBank and VerbNet are verb-oriented and focus more on syntactic level. Hence, many important linguistic clues of entity relations cannot be extracted, and many verbs with no relational meaning may be extracted unexpectedly, producing more noises to the relation extraction system.

For each relation type in our relation extraction, we first obtain the corresponding semantic frames by traversing the FrameNet. All the LUs involved in these semantic frames are relation indicators, which are actually the keywords that often used to express such relation. We eventually identify 62 semantic frames and 1136 LUs from the FrameNet. Each LU is a discrete word or phrase. In order to leverage the relation indicators to provide knowledge attention, we project each word and phrase of the corresponding LU into low-dimensional vector $\boldsymbol{u}_i \in \mathbb{R}^{d_w}$ by looking up the pre-trained word embedding matrix, where d_w is the size of the LU embeddings. If a LU consists of multiple words, such as the LU "*set up*" in Figure 1, we calculate the mean of the embeddings of these words to form the corresponding relation indicator. We aggregate all the relation indicators to form a indicator set $\boldsymbol{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_n\}$, where n is the number of LUs. This relation indicator set can be used as the prior knowledge for relation extraction.

3.1.2. Knowledge Attention over Words

Each sentence S in the sentence bag consists of a sequence of words, i.e., $S = \{w_1, w_2, ..., w_m\}$, where m is the length of the sentence. We first project the discrete words to low-dimensional word vectors by looking up the pre-trained word embeddings. Thus, these words can be processed and modeled by the knowledge attention layer and graph convolutional layer. The same word embedding matrix used in the LUs embedding procedure is employed to embedding the words in sentences. The word embedding of the i-th word in sentence S is denoted by $\boldsymbol{e}_i^w \in \mathbb{R}^{1 \times d_w}$, where d_w is the size of the word embeddings. The sentence can be expressed as the concatenation of the word embeddings, $\boldsymbol{S} = \{\boldsymbol{e}_1^w, \boldsymbol{e}_2^w, ..., \boldsymbol{e}_m^w\} \in \mathbb{R}^{m \times d_w}$.

Not all words in a sentence are equally important for relation extraction. In order to distinguish the importance of each word in a sentence, we adopt the recently-promoted self-attention mechanism [35,36] to measure the contribution (importance) of each word to the expression of a relevant relation. It helps in highlighting important relation words with respect to each of the relation indicators present in the LUs set. The generation of the word-level knowledge attention is illustrated in Figure 2.

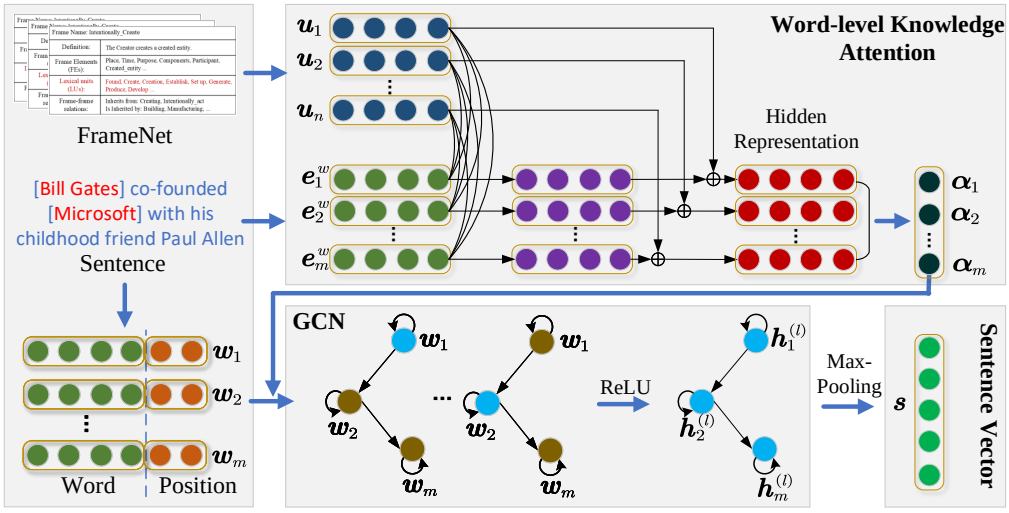


Figure 2. The framework of the sentence embedding module. The upper right part illustrates the generation of the word-level knowledge attention. The lower right part is the knowledge attention based graph convolutional network, which calculates the vector representation of input sentences.

Formally, the query($\boldsymbol{Q}$) is the word embeddings of a sentence, and the key($\boldsymbol{K}$)-value($\boldsymbol{V}$) pairs are both the relation indicator embeddings, i.e., $\boldsymbol{Q} = \boldsymbol{S} \in \mathbb{R}^{m \times d_w}, \boldsymbol{K} = \boldsymbol{V} = \boldsymbol{U} \in \mathbb{R}^{n \times d_w}$. Thus, the hidden representation of the input sentence can be obtained from

$$\boldsymbol{H} = Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_w}})\boldsymbol{V}. \tag{1}$$

d_w is a scaling factor, which is the dimension of word embeddings. Specifically, for each word $\boldsymbol{e}_i^w$ of the input sentence, the attention probability $\boldsymbol{p}^i$ is expressed as

$$\boldsymbol{p}^i = softmax(\frac{\boldsymbol{e}_i^w \boldsymbol{K}^T}{\sqrt{d_w}}). \tag{2}$$

Then, the hidden representation of each word can be calculated as a weighted sum of the values

$$\boldsymbol{h}_i = \sum_{j=1}^{n} \boldsymbol{p}_j^i \odot \boldsymbol{V}_j, \tag{3}$$

where $\boldsymbol{h}_i \in \mathbb{R}^{1 \times d_w}$ is the hidden representation of the i-th word in the sentence, $\odot$ means the element-wise multiplication, and $\sum$ performs along sequential dimension.

Eventually, the result of the knowledge attention can be calculated as

$$\alpha_i = \frac{\exp(\mu_i)}{\sum_{j=1}^{m} \exp(\mu_j)}, \mu_i = \boldsymbol{h}_i \boldsymbol{W}_1 \boldsymbol{r}, \tag{4}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d_w \times d_w}$ is a square matrix, and $\boldsymbol{r} \in \mathbb{R}^{d_w \times 1}$ is a random query vector. α_i is the knowledge attention score of the i-th word in the sentence, which is calculated by attending to the relation indicators. This attention score represents the importance of the word for relation extraction.

3.1.3. Knowledge Attention Based GCN

In this paper, we employ a knowledge-guided GCN model to build the context encoder and transform the sentences into low-dimension vectors, as shown in Figure 2. The GCN model is an adaptation of the convolutional neural network for encoding graphs, which encodes the dependency structure over the input sentence with efficient graph convolution operation to generate the vector representation of the sentence. Given a graph (dependency

tree of a input sentence) with k nodes, we can represent the graph with an $k \times k$ adjacency matrix $\boldsymbol{A}$. $A_{i,j} = 1$ if there is an edge between nodes i and j, otherwise $A_{i,j} = 0$. We denote the input vector as $\boldsymbol{h}_i^{(l-1)}$ and the output vector of node i at the l-th layer of the GCN model as $\boldsymbol{h}_i^{(l)}$. The graph convolution operation is expressed as:

$$\boldsymbol{h}_i^{(l)} = ReLU(\sum_{j=1}^{k} \boldsymbol{A}_{i,j} \boldsymbol{W}^l \boldsymbol{h}_j^{(l-1)} + \boldsymbol{b}^l), \tag{5}$$

where $ReLU$ is an activation function, $\boldsymbol{W}^l$ is the weight matrix at l-th layer, and $\boldsymbol{b}^l$ is the bias vector. The new hidden representation $\boldsymbol{h}_i^{(l)}$ of node i is obtained by considering only its immediate neighbors. Repeating this graph convolution operation L times forms a L layers GCN model. All these operations are conducted with matrix multiplications, which is suitable for batch computation over instances and running on GPUs. Since the information propagation between nodes is performed in parallel, the efficiency of the model is not affected by the depth of the dependency tree.

In order to adapt the GCN model to encode the sentences, we convert each dependency tree of the sentences into its corresponding adjacency matrix $\boldsymbol{A}$. If there is a dependency edge between words w_i and w_j, $A_{i,j} = 1$. Motivated by Zeng et al. [7], the position features of words are also considered in our work, which can express the structural features of a sentence. Each word in a sentence has two relative distances PF_1 and PF_2 with entities e_1 and e_2. Take the sentence mentioned in the previous section as an example, the relative distances of word "co-founded" to "*Bill Gates*" and "*Microsoft*" are -1 and 1. The position embedding of each word in the sentence S is denoted by $\boldsymbol{e}_{i,1}^p \in \mathbb{R}^{d_p}$ and $\boldsymbol{e}_{i,2}^p \in \mathbb{R}^{d_p}$, where d_p is the size of the position embeddings. The i-th word in sentence S can then be projected to a low-dimensional vector $\boldsymbol{w}_i = [\boldsymbol{e}_i^w; \boldsymbol{e}_{i,1}^p; \boldsymbol{e}_{i,2}^p] \in \mathbb{R}^d$ by concatenating the word embedding with two position embeddings, where $d = d_w + 2d_p$. The initial representation of the sentence S can be expressed as $\boldsymbol{S} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_m\} \in \mathbb{R}^{m \times d}$. Then, each node of the dependency tree can be represented by it's corresponding word embedding, and the inputs of the GCN $\boldsymbol{h}^{(0)} = \{\boldsymbol{h}_1^{(0)}, \boldsymbol{h}_2^{(0)}, ..., \boldsymbol{h}_m^{(0)}\} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_m\}$ are obtained.

Since words never connect to themselves in the original dependency tree, the information of $\boldsymbol{h}_i^{l-1}$ can never be propagated to $\boldsymbol{h}_i^l$. To address this issue, we update the dependency structure by adding a self-loop for each word. Thus, the updated adjacency matrix is expressed as $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$, where $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ is an identity matrix. Furthermore, previous GCN-based methods for sentence encoding treat each node of the dependency tree equally without distinguishing the importance of them. In this paper, we introduce the prior knowledge obtained from FrameNet to highlight the key words and cue phrases of a sentence for relation extraction. During the graph convolution operation, we assign each word w_i a knowledge attention score α_i calculated by using the pseudo self-attention mechanism described in previous section. Then, we modify the calculation of each layer, and the knowledge attention guided graph convolution operation of node i at l-th layer is expressed as

$$\boldsymbol{h}_i^{(l)} = ReLU(\sum_{j=1}^{k} \alpha_j (\tilde{\boldsymbol{A}}_{i,j} \boldsymbol{W}^l \boldsymbol{h}_j^{(l-1)} + \boldsymbol{b}^l)), \tag{6}$$

where α_j is the knowledge attention score of node j. Using knowledge attention to selectively obtain information from neighboring nodes can effectively alleviate the negative impact of noisy nodes. As a result, the key words will contribute more to the sentence encoding in the graph convolution operation. After conducting a L layers knowledge-guided GCN for the word vectors, we obtain the hidden representations of each word that directly integrated information from neighbors no more than L edges apart in the dependency tree.

For each sentence in a instance bag, all the original word embeddings of the sentence are feed into the GCN model to obtain the representation of the whole sentence. The hidden representation of the sentence is obtained as follows:

$$\boldsymbol{h}^{(L)} = GCN(\boldsymbol{h}^{(0)}) = GCN(\boldsymbol{S}). \tag{7}$$

Finally, a max-pooling layer is used to capture the most important and relevant features from generated sequence and address the issue of variable sentence lengths.

$$\boldsymbol{s} = Max(\boldsymbol{h}^{(L)}) = Max\{\boldsymbol{h}_1^{(L)}, \boldsymbol{h}_2^{(L)}, ..., \boldsymbol{h}_m^{(L)}\}. \tag{8}$$

where $\boldsymbol{s} \in \mathbb{R}^d$ is the final representation of the input sentence, $Max()$ is the max-pooling operation that maps m output vectors to the finial sentence vector.

3.2. Knowledge Supervised Sentences Selection

As mentioned in above sections, the latent semantic information contained in KGs plays a vital role in distantly supervised relation extraction, since the training data are obtained by aligning the textual corpus with the existing KGs. These semantic information can provide additional supervision for selecting multiple valid sentences. Previous works use only the textual information to train the relation extractor. Additionally, the distant supervision simply incorporates the KG information as meaningless one-hot labels instead of treating it as a graph, which ignores the rich structure and semantic information present in KGs. In this paper, we extract the interactions between the entity pair and relations in KGs as prior knowledge to guide the selection of valid sentence.

3.2.1. Knowledge Graph Embedding

Knowledge graph embedding is an independent work that maps the entities and relations into low-dimensional vector space. In order to learn the vector representation of entities and relations of triples in KGs, the TransE [37] model is the natural choice. For each triple (e_1, r, e_2), the explicit relation r can be treated as the translation from e_1 to e_2, which is formalized as $\boldsymbol{r} = \boldsymbol{e}_2 - \boldsymbol{e}_1$. We denote the relations in knowledge graphs as KG-relations. For the encoding of each sentence bag, our model gives each sentence in the bag a confidence score by measuring the semantic distance between the sentence and the KG-relation. As a result, the model can selectively assign higher weights for valid sentences and reduce the impact of noisy sentences by assigning low weights to them.

We also propose a interactive model to learn the representation of KGs and mine the interactions between the entities and relations. We believe that the confidence of a triple depends on the interaction of the entities and relationships it contains. If the vector distribution of entities and relations in a triple are more similar in the semantic space, the triple has higher confidence. For each triple (e_1, r, e_2) in a knowledge graph G, we calculate the interactions between the entities and relation by using the following interaction scoring function:

$$S_{inter}(e_1, r, e_2) = \boldsymbol{e}_1 \cdot \boldsymbol{r} + \boldsymbol{r} \cdot \boldsymbol{e}_2, \tag{9}$$

where $\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2 \in \mathbb{R}^{d_w}$ are low-dimensional vectors of the entities and relation in the triple, $\boldsymbol{e}_1 \cdot \boldsymbol{r}$ and $\boldsymbol{r} \cdot \boldsymbol{e}_2$ are considered as the interactions between e_1 and e_2 with r, respectively. The interactive scoring function will assign higher scores to the fact triples than negative ones. Based on the above scoring function, we train a margin-based ranking loss function over all triples in G as follows,

$$\mathcal{L}_{inter} = \sum_{G} \sum_{G'} \max\{0, S_{inter}(\boldsymbol{e}_1', \boldsymbol{r}, \boldsymbol{e}_2') - S_{inter}(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2)\}, \tag{10}$$

where

$$G'_{(e_1,r,e_2)} = \{(e_1', r, e_2) | e_1' \in E\} \cup \{(e_1, r, e_2') | e_2' \in E\} \tag{11}$$

is the collection of negative triples. The negative triples is generated by using the random entities in the knowledge graph to replace the head entity or tail entity in the fact triples. We use stochastic gradient descent (SGD) as optimizer to optimize the loss function and we use a $L2$ regularization that helps prevent over-fitting. The constraint on the $L2$-norm of embeddings are defined as $\forall(e_1, r, e_2), \| \boldsymbol{e}_1 \|_2 \leqslant 1, \| \boldsymbol{r} \|_2 \leqslant 1, \| \boldsymbol{e}_2 \|_2 \leqslant 1$. The model learns an unique vector representation for each entity and relation in the knowledge graph G and maps it into a low-dimensional semantic space.

Since the true relations in the test set are unknown, for each entity pair (e_1, e_2), we simply define the KG relation embedding $\boldsymbol{r}_{kg} \in \mathbb{R}^{d_w}$ as a translation from $\boldsymbol{e}_1$ to $\boldsymbol{e}_2$, which is formalized as

$$\boldsymbol{r}_{kg} = \boldsymbol{e}_2 - \boldsymbol{e}_1. \tag{12}$$

Eventually, the embeddings of the KG relations $\boldsymbol{r}_{kg}$ can be used as prior knowledge to guide the selection of valid sentences during the aggregation of multiple instances to generate bag-level relation representation.

3.2.2. Knowledge Attention over Sentences

Given a sentence bag $B_i = \{S_1, S_2, ..., S_n\}$ with n sentences, all the sentences refer to a common entity pair. The embeddings of each sentence $\{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_n\}$ in the bag can be obtained by using the sentence representation module, as described in Section 3.1. However, the sentence bag are obtained by using the distant supervision algorithm, which contains some vague and wrong sematic components. Thus, we argue that some sentences may contribute more to the final textual relation representation. In order to discriminately aggregate sentence-level representations into bag-level representation, the multi-instance learning that use a selective attention mechanism is an intuitive choice. The selective attention algorithm generates a weight distribution over all sentences in the bag to alleviate the noise problem. However, when there is only one sentence in the bag, even the only sentence is a noise instance (wrong labeled instance), the selective attention mechanism will be useless. It is worth noting that in the commonly used distant supervision relation extraction corpus, almost 80% of the bags contain only one sentence, and many of the them are even wrong labeled. To address this problem, we use the KG relation embedding $\boldsymbol{r}_{kg}$ as knowledge attention over sentences to augment the contribution of positive instances and reduce the impact of negative instances, as shown in Figure 3. In this way, the wrongly labeled sentence will be dynamically assigned with low weight score to prevent the propagation of noise representation.

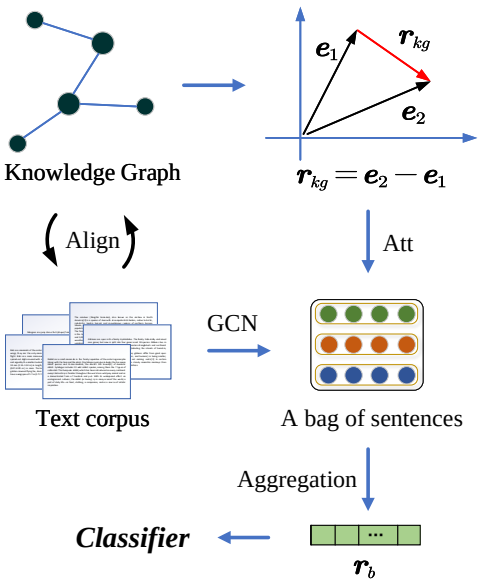


Figure 3. Knowledge attention guided selection of multiple valid sentences.

Instead of generating a weight probability distribution for all sentences in the bag, we calculate a confidence score for each sentence based on the prior knowledge $\boldsymbol{r}_{kg}$. For the i-th sentence in the bag, the scoring function is formally defined as

$$\begin{aligned} u_i &= \text{sigmoid}(\boldsymbol{v}_a^{\intercal}\tanh(\boldsymbol{W}_a[\boldsymbol{s}_i;\boldsymbol{r}_t])), \\ \boldsymbol{r}_t &= \boldsymbol{W}_t\boldsymbol{r}_{kg} + \boldsymbol{b}_t, \end{aligned} \tag{13}$$

where u_i is the confidence score for the i-th sentence, and $[;]$ denotes the concatenation operation. $\boldsymbol{W}_t \in \mathbb{R}^{d_w \times d_w}$, $\boldsymbol{W}_a \in \mathbb{R}^{d_a \times d_a}$, $\boldsymbol{v}_a \in \mathbb{R}^{d_a \times 1}$, $and d_a = d + d_w$, are parameters learned in the training stage. The mean aggregation operation is performed over sentence in the bag to form the bag-level vector representation for further relation classification, which is obtained by

$$\boldsymbol{r}_b = \frac{1}{n}\sum_{i}^{n} u_i \cdot \boldsymbol{s}_i, \tag{14}$$

where $\boldsymbol{r}_b \in \mathbb{R}^d$ is the final representation of the sentence bag.

3.3. Complexity Analysis

In our work, the time and space cost is mainly on the word-level knowledge attention computation and knowledge graph embedding module.

For word-level knowledge attention computation, the time and space cost is mainly on the self-attention operation. For the self-attention layer, the dimension of the input representation is d_w, and the length of the sentence is m. In Equation (1), the dot products of the query with all keys are implemented. We can obtain that the time and space complexities of each self-attention operation are $O(d_w m^2)$ and $O(d_w m)$, respectively.

In the knowledge graph embedding module, the time and space cost mainly depends on the calculation of the interaction between entity and relation in the triples, i.e., solving the Equation (9). The time complexity of Equation (9) is $O(d_k)$, and the space complexity is $O(1)$, where $d_k = d_w$ is the size of the entity embedding and relation embedding in the knowledge graph embedding space. The computational complexity of the knowledge graph embedding model is proportional to the dimension of the entity and relation embeddings. The time consumption in the training process is mainly determined by the number of entities and relations in the training set. Due to the low complexity of the knowledge graph embedding model, it can adapt well to the embedding of large-scale knowledge graphs.

4. Implementation for Relation Classification

Our approach introduces the semantic and structure information from FrameNet and knowledge graphs as prior knowledge to guide the distantly supervised relation extraction. In this section, we discuss how to train the knowledge guided relation extraction model. First, the word embedding (words in sentences and in LUs) and KG embedding are pre-trained by using the GloVe [38] tool and the knowledge graph embedding module, respectively. Then, the sentence embedding model can be trained by using the word-level attention mechanism based on the relation indicator of LUs. Finally, the bag-level representation of the sentence bag can be obtained by using the sentence-level attention mechanism. We adopt a pairwise margin-based ranking loss function [39] as the optimization target of our knowledge-guided graph convolutional network model.

Given a text corpus D and a knowledge graph with relation set R, the model aims to predict a relation type for each sentence bag in the textual corpus, which assigns a semantic matching score to each sentence bag as for how well the bag expresses a candidate relation. The vector representation of each bag $\boldsymbol{r}_b$ can be obtained by using the models proposed in aforementioned section. During the training stage, we learn the vector representation $[W^R]_r$

for each relation label r. To this end, we calculate the semantic matching score between each bag-level representation $\boldsymbol{r}_b$ with each relation type, which is formalized as follows,

$$S(x)_r = \boldsymbol{r}_b^{\mathsf{T}} \cdot [W^R]_r, \tag{15}$$

where $W^R \in \mathbb{R}^{d \times |R|}$ is a randomly initialized relation matrix whose columns represent different relation labels, and $|R|$ is the number of the predefined relation types. In order to train the model, we define a loss function and optimize it over all instance in the training set D,

$$\begin{aligned} \mathcal{L} =& log(1 + exp(\gamma(m^+ - S(x)_{r^+}))) \\ &+ log(1 + exp(\gamma(m^- + S(x)_{r^-}))), \end{aligned} \tag{16}$$

where γ is a scaling parameter, m^+ and m^- are hyper parameters. $S(x)_{r^+}$ and $S(x)_{r^-}$ represent the semantic matching score between $\boldsymbol{r}_b$ with the corresponding actual relation type r^+ and false relation type r^-, respectively. In the training stage, we choose r^- with the highest false score as the negative relation type. We employ SGD optimizer to optimize the loss function and utilize $L2$-norm $\beta\|\theta\|_2^2$ to prevent over-fitting, where θ is the parameter set.

5. Experimental Setup

5.1. Dataset and Evaluation Metrics

The datasets of our experiment contain two parts, knowledge graphs and text corpus. We use FB60K as KG to learn the representation of entities and KG relations. The FB60K is extracted from Freebase (FB) and extended from the dataset developed by Riedel et al. [24]. There are 1324 relations, 69,512 entities and 335,350 facts in this dataset. We adopt two widely used datasets of distantly supervised relation extraction as text corpus to demonstrate the effectiveness of our method and baselines. They are NYT-FB [24] and GDS [40] datasets, where the statistical comparison of them are illustrated in Table 1.

NYT-FB: The NYT-FB dataset is developed by Riedel et al. [24], which is constructed by aligning the New York Times (NYT) corpus with Freebase facts. The association between the NYT and FB is built by performing a string match between entity mentions in NYT and canonical names of entities in FB. The entity mentions are fined by using the Stanford named entity recognizer. NYT-FB is a standard benchmark for distantly supervised relation extraction in most of the previous works [26,40], which contains 52 predefined relation types and a null class *NA* relation (no relation between two entities). The most common relations in this dataset are *location*, *nationality*, *capital*, *place_lived*, and *neighborhood_of*. The training instances are obtained by aligning the sentences from the NYT corpus of years 2005–2006. The test instances are obtained by aligning sentences from 2007. There are 570,088 sentences, 291,699 entity pairs in the training set, and 172,488 sentences, 96,678 entity pairs in the testing set. Since this dataset does not have a validation set, we split the training set into 80% for training, and 20% for validation. This dataset is available at: https://drive.google.com/file/d/1UD86c_6O_NSBn2DYirk6ygaHy_fTL-hN/view?usp=sharing, accessed on 20 May 2021.

GDS: The Google distant supervision (GDS) dataset is developed by Jat et al. [40], which is extended from the Google relation extraction corpus. There are 5 relation types, including *perGraduatedInstitution*, *perHasDegree*, *perPlaceOfBirth*, *perPlaceOfDeath*, and a *NA* relation. Each instance bag in this dataset is guaranteed to contain at least one sentence which expresses the relation type assigned to that instance bag, which alleviates the noise in distant supervision setting. This makes automatic evaluation more reliable. The dataset is divided into three parts, 60% for training, 10% for validation, and 30% for testing. There are 11,297 sentences and 6498 entity pairs in the training set, 1864 sentences and 1082 entity pairs in the validation set, and 5663 sentences and 3247 entity pairs in the testing set. This dataset is available at: https://drive.google.com/file/d/1UMS4EmWv5SWXfaSl_ZC4DcT3dk3JyHeq/view?usp=sharing, accessed on 20 May 2021.

Following previous works [26], we evaluate our model on the held-out test set from the datasets. It evaluates our model by comparing the relation facts recognized from the test sentences with those in Freebase. To show the performance of our model, we use precision-recall (PR) curves and top-N precision (P@N) as metrics in our experiments. The PR curves are constructed using the model predictions on all entity pairs in the test set for all relation types sorted by the confidence scores from the highest to lowest.

Table 1. A statistical comparison of the used datasets.

Datasets		Sentences	Entity Pairs	Relations
NYT-FB	Train	455,771	233,064	53
	Valid	114,317	58,635	
	Test	172,448	96,678	
GDS	Train	11,297	6498	5
	Valid	1864	1082	
	Test	5663	3247	

5.2. Baselines

We compare our proposed model with extensive previous works, including feature-based methods and state-of-the-art neural-based methods. The baselines are listed in following.

5.2.1. Feature-Based

Distant supervision for relation extraction without labeled data (**Mintz**) [10]. The original distantly supervised approach for relation extraction, which is a multi-class logistic regression model.

Multi-instance learning with overlapping relations (**MultiR**) [23]. A probabilistic graphical model for multi-instance learning, which is able to handle problems with overlapping relations.

5.2.2. Neural-Based

Piece-wise convolutional neural network (**PCNN**) [11]. A convolutional neural network based distantly supervised relation extraction approach, which employs the piecewise max-pooling operation to generate the vector representation of sentences.

Piece-wise convolutional neural network with sentence-level attention (**PCNN-ATT**) [26]. A improved approach based on the PCNN model, which employs a selective attention over multiple instances to alleviate the wrongly labeled problem.

Bi-directional gated recurrent unit based word attention model (**BGWA**) [40]. A Bi-GRU based method for relation extraction, which employs the word-level and sentence-level attention mechanism to enhance the representation of instance bags.

Relation extraction with side information (**RESIDE**) [30]. A distantly supervised neural relation extraction approach which uses relevant side information and employs graph convolutional networks to encode the syntactic information of instances.

5.3. Parameter Settings

The initial word and entity embeddings in our experiment are pre-trained by using the 50 dimensional GloVe embeddings on a 6 billion corpus [38]. For multiple words nominal, we average the embeddings of its subcomponents. For the out of vocabulary words, we assign random vectors to them. The dimension of the position embedding d_p is 5. In our experiment, we use two layers GCN architecture to encode the sentences. The hyper-parameters are determined on the development dataset. The scaling parameter $\gamma = 2$ and margins $m^+ = 2.5$, $m^- = 0.5$ of the ranking loss function are set according to the parameter settings reported in the works of Santos et al. [39]. We employ mini-batch mechanism to train our model with 50 instances in each mini-batch to. The initial learning rate λ is 0.5, we gradually reduce the learning rate according to the training epoch. Additionally,

we apply a dropout strategy with a dropout rate of 0.5 to all but the last GCN layer. The hyper-parameter values of our model are shown in Table 2.

Table 2. Hyper-parameters used in our experiments.

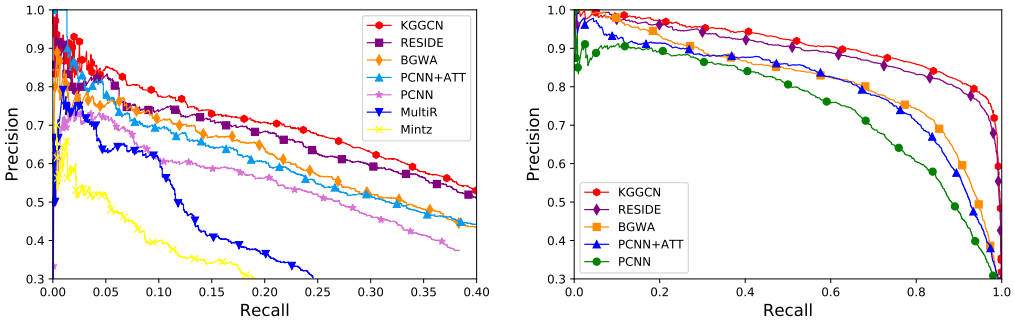
Para	Description	Value
d_w	Word\entity\relation Embedding	50
d_p	Position Embedding	5
β	Regularization Factor	0.001
λ	Learning Rate	0.5
p	Dropout probability	0.5
L	GCN Layers	2

5.4. Comparison with Baselines

To verify the effectiveness of our model, we compare it against the baselines on the NYT-FB and GDS datasets. Since GDS is a recently proposed dataset, we only compare our model with neural-based methods on this dataset. Figure 4 summarizes the comparison results in terms of PR curves on the datasets. From the comparison results illustrated in Figure 4, we can observe that:

(1) The neural-based methods significantly outperform the feature based methods Mintz and MultiR on the NYT-FB dataset. The results demonstrate that the human-designed features are limited in relation extraction, and the use of NLP tools to generate features often leads to the propagation and accumulation of errors. It also demonstrates the robustness and effectiveness of the neural models for relation extraction;

(2) The BGWA and PCNN+ATT outperform the PCNN model over the entire range of recall on both datasets, which indicates the attention mechanism is helpful for distantly supervised relation extraction. The higher performance of the RESIDE model over BGWA and PCNN+ATT demonstrates that the additional side information (relation alias and entity types) from knowledge graphs helps in improving the performance of the model;



(**a**) (**b**)

Figure 4. Performance comparison for proposed model and previous baselines in terms of precision-recall curves. (**a**) Comparison of precision-recall curves on NYT-FB dataset. (**b**) Comparison of precision-recall curves on GDS dataset.

(3) Our proposed model KGGCN achieves the best performance compared with all the baselines on both datasets. Especially in contrast to feature-based methods, our model increases by more than 40% when the recall is larger than 0.25. Compared with other neural-based methods, our model also has a significant improvement. All these demonstrate that the prior knowledge from FrameNet and knowledge graphs can effectively guided the encoding of the sentence-level and bag-level features;

(4) The overall performance of the neural-based models on GDS dataset is better than that on NYT-FB dataset. This is because that the NYT-FB dataset contains many negative instance bags. All sentences in these bags are wrongly labeled, which is very noisy for

relation extraction. However, the GDS dataset is constructed based on the at-least-one restriction, i.e., each bag contains at least one sentence that exactly expresses the relation of the corresponding entity pair. Moreover, the GDS dataset only has 5 relation types, which is easier for classification.

For the proposed model and the neural-based methods, we also evaluate the performances of them in terms of P@N on the NYT-FB dataset. Following the experiment settings in the previous works [26], we randomly select one, two, and all sentences for each entity pair to form the instance bags. We evaluate the P@N of the models on each circumstance. From Table 3, we can observe that:

(1) The models with attention mechanism (PCNN+ATT, BGWA, RESIDE, KGGCN) achieve better performance than the PCNN (without attention mechanism) on each case. The training sentences generated by using the distant supervision may be wrongly labeled, because not all sentences contain the entity pair exactly express the corresponding textual relations, and not each sentence contributes equally for the encoding of a bag. Thus, the attention mechanism can effectively highlight the meaningful sentences and alleviate the impact of the wrong label problem;

(2) The RESIDE performs much better than BGWA and PCNN+ATT, which demonstrates that the side information (relation alias and entity type) extracted from external corpus is beneficial for the improving of relation extraction;

(3) Our proposed model outperforms all the baselines, which demonstrates that the traditional data-driven attention mechanism is limited for distantly supervised relation extraction. The same relation may have subtle semantic differences between different entity pairs, and a simple global attention can not distinguish these semantic differences. The knowledge-based attention of our model is obtained by embedding the interaction between the entity pairs and relations, which is more discriminative than the data-driven attention.

Table 3. Performances comparison of KGGCN with neural-based baseline models in terms of P@N using different number of sentences in bags on NYT-FB dataset. One, two, All mean the number of sentences randomly selected from a instance bag.

Models	One				Two				All			
	P@100	**P@200**	**P@300**	**Mean**	**P@100**	**P@200**	**P@300**	**Mean**	**P@100**	**P@200**	**P@300**	**Mean**
PCNN	73.3	64.8	56.8	65.0	70.3	67.2	63.1	66.9	72.3	69.7	64.1	68.7
PCNN+ATT	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
BGWA	78.0	71.0	63.3	70.8	81.0	73.0	64.0	72.7	82.0	75.0	72.0	76.3
RESIDE	80.0	75.5	69.3	74.9	83.0	73.5	70.6	75.7	84.0	78.5	75.6	79.4
KGGCN (ours)	**87.6**	**78.3**	**72.8**	**79.6**	**85.1**	**76.2**	**73.8**	**78.4**	**86.7**	**81.4**	**76.3**	**81.5**

It is worth noting that the RESIDE model also introduces information from knowledge graphs. This information is generated by directly extracting the names of relations from existing KGs, which does not model the interaction and structure information between entities and relations in the KGs. However, the core of distantly supervised relation extraction is to align the entity pairs and their relations of knowledge graph with these mentioned in textual corpus. Thus, the interaction between the entity pairs and the relations are important for distantly supervision relation extraction. The knowledge graph embedding module of our model focuses on the interaction between the entity pairs and relations, which can effectively extract the semantic and structure information of the KGs as additional knowledge attention to provide supervision for valid sentences selection. Thus, our proposed method achieves better performance.

5.5. Ablation Study

In order to analyze the effect of various components of the proposed KGGCN on its performance, we conduct an ablation study on the NYT-FB validation set. We define four variant models with cumulatively removed components, including *KGGCN w/o KG*, *KGGCN w/o LU*, *KGGCN w/o ALL* and *KGGCN w TransE*.

Specifically, *KGGCN w/o KG* denotes removing the knowledge attention from the knowledge graph and generating the bag representation by calculating the mean of all sentences in the bag. *KGGCN w/o LU* denotes removing the knowledge attention from lexical units, and *KGGCN w/o ALL* denotes removing both the knowledge attention from the knowledge graph and lexical units. *KGGCN w TransE* denotes replacing the knowledge graph embedding module with the TransE. The experimental results in terms of P@N are shown in Table 4. According to the results, we can observe that when removing different components from KGGCN, the performance of the variant models drops drastically. Particularly, by removing the word-level knowledge attention (i.e., KGGCN w/o LU) and sentence-level attention (i.e., KGGCN w/o KG), the performance decreases 2.5 and 7.2, respectively, in terms of P@N mean for all sentences. When removing both modules above (i.e., KGGCN w/o ALL), the performance of the variant model drops 13% in terms of P@N mean for all sentences. These demonstrate the effectiveness of the prior knowledge from knowledge graph and lexical units. In addition, in order to make an in-depth evaluation of the knowledge graph embedding module of our proposed method, we replace it with TransE to generate the representation of the entities and relations in the knowledge graph (i.e., KGGCN w TransE). The results show 1.4 drops in terms of P@N mean for all sentences. It demonstrates that our knowledge graph embedding module can effectively extract the semantic and structure information, as well as the interaction between entities and relations in the knowledge graph.

Table 4. Ablation study on the NYT-FB dataset.

Models	p@100	p@200	p@300	Mean
KGGCN(ours)	86.7	81.4	76.3	81.5
KGGCN w TransE	85.1	80.2	74.9	80.1
KGGCN w/o LU	84.5	78.7	73.7	79.0
KGGCN w/o KG	78.2	75.4	69.3	74.3
KGGCN w/o ALL	73.2	69.5	62.9	68.5

Outperforming these variant models highlights our model's ability to capture sentence-level and bag-level features. All these experimental results demonstrate that the external information from knowledge graph and FrameNet can be the prior knowledge to guide the extraction of textual features, which helps to improve the distantly supervised relation extraction.

6. Conclusions and Future Work

In this paper, we propose a novel method for distantly supervised relation extraction task by using a knowledge attention guided graph convolutional network. We aim at exploring the information from FrameNet and knowledge graphs as knowledge attention to improve the performance of graph convolutional networks. Extensive experiments are conducted to evaluate the proposed method. The experimental results show that our method can efficiently use the prior knowledge from the FrameNet and knowledge graph to enhance the performance of distantly supervised relation extraction, and it outperforms all the compared baselines.

In future work, we will investigate the automatic selection of the relation indicator for relation identification. We will try to apply the prior knowledge to enhance the word representations, and explore potential methods to capture the semantic connection between words and relation facts. Furthermore, we will try to apply the knowledge attention in other domain-specific tasks.

Author Contributions: Conceptualization, W.H.; methodology, W.H.; formal analysis, H.Z.; investigation, N.M.; writing—original draft preparation, N.M.; writing—review and editing, N.M.; funding acquisition, W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Plan of Hunan grant number 2016TP1003 and the Key Technology R&D Program of Hunan Province grant number 2018GK2052.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RE	Relation extraction
NLP	Natural language processing
KG	Knowledge graph
CNN	Convolutional neural network
RNN	Recurrent neural network
GCN	Graph convolutional network
KGGCN	Knowledge-guided graph convolutional network
LUs	Lexical units
SGD	Stochastic gradient descent
NYT-FB	New York Times with freebase facts
GDS	Google distant supervision
PR	Precision-recall
P@N	Top-N precision

References

1. Cui, W.; Xiao, Y.; Wang, H.; Song, Y.; Hwang, S.W.; Wang, W. KBQA: Learning Question Answering over QA Corpora and Knowledge Bases. *Proc. VLDB Endow.* **2017**, *10*. [CrossRef]
2. Quirk, C.; Poon, H. Distant Supervision for Relation Extraction beyond the Sentence Boundary. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, 3–7 April 2017; pp. 1171–1182.
3. Young, T.; Cambria, E.; Chaturvedi, I.; Huang, M.; Zhou, H.; Biswas, S. Augmenting end-to-end dialog systems with commonsense knowledge. *arXiv* **2017**, arXiv:1709.05453.
4. Zhang, Z. Weakly-supervised relation classification for information extraction. In Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, Washington, DC, USA, 8–13 November 2004; pp. 581–588.
5. Qian, L.; Zhou, G.; Kong, F.; Zhu, Q. Semi-supervised learning for semantic relation classification using stratified sampling strategy. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; pp. 1437–1445.
6. Rink, B.; Harabagiu, S. Utd: Classifying semantic relations by combining lexical and semantic resources. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010; pp. 256–259.
7. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. Relation Classification via Convolutional Deep Neural Network. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 2335–2344.
8. Zhang, D.; Wang, D. Relation classification via recurrent neural network. *arXiv* **2015**, arXiv:1508.01006.
9. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, 7–12 August 2016; pp. 207–212.
10. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, 2–7 August 2009; Volume 2, pp. 1003–1011.
11. Zeng, D.; Liu, K.; Chen, Y.; Zhao, J. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1753–1762.
12. Li, P.; Mao, K.; Yang, X.; Li, Q. Improving Relation Extraction with Knowledge-attention. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 229–239.
13. Ruppenhofer, J.; Ellsworth, M.; Schwarzer-Petruck, M.; Johnson, C.R.; Scheffczyk, J. FrameNet II: Extended Theory and Practice. Available online: http://framenet.icsi.berkeley.edu/fndrupal (accessed on 2 May 2021).
14. Khan, A.W.; Khan, M.U.; Khan, J.A.; Ahmad, A.; Khan, K.; Zamir, M.; Kim, W.; Ijaz, M.F. Analyzing and Evaluating critical challenges and practices for software vendor organizations to secure Big Data on Cloud Computing: An AHP-based Systematic Approach. *IEEE Access* **2021**, 9, 2617–2633. [CrossRef]

15. Kaur, J.; Ahmed, S.; Kumar, Y.; Alaboudi, A.; Jhanjhi, N.; Ijaz, M.F. Packet Optimization of Software Defined Network Using Lion Optimization. *CMC-COMPUTERS MATERIALS & CONTINUA* **2021**, 69, 2617–2633.
16. Kambhatla, N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions, Barcelona, Spain, 21–26 July 2004; p. 22.
17. Suchanek, F.M.; Ifrim, G.; Weikum, G. Combining linguistic and statistical analysis to extract relations from web documents. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 712–717.
18. Alicante, A.; Corazza, A. Barrier features for classification of semantic relations. In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, Hissar, Bulgaria, 12–14 September 2011; pp. 509–514.
19. Bunescu, R.C.; Mooney, R.J. A shortest path dependency kernel for relation extraction. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 724–731.
20. Zhang, M.; Zhang, J.; Su, J.; Zhou, G. A composite kernel to extract relations between entities with both flat and structured features. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–21 July 2006; pp. 825–832.
21. Socher, R.; Huval, B.; Manning, C.D.; Ng, A.Y. Semantic compositionality through recursive matrix-vector spaces. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 12–14 July 2012; pp. 1201–1211.
22. Liu, C.; Sun, W.; Chao, W.; Che, W. Convolution neural network for relation extraction. In Proceedings of the International Conference on Advanced Data Mining and Applications, Hangzhou, China, 14–16 December 2013; pp. 231–242.
23. Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; Weld, D.S. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, OR, USA, 19–24 June 2011; pp. 541–550.
24. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Athens, Greece, 4–8 September 2010; pp. 148–163.
25. Guo, Z.; Zhang, Y.; Lu, W. Attention Guided Graph Convolutional Networks for Relation Extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 241–251.
26. Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; Sun, M. Neural relation extraction with selective attention over instances. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 2124–2133.
27. Liu, T.; Zhang, X.; Zhou, W.; Jia, W. Neural Relation Extraction via Inner-Sentence Noise Reduction and Transfer Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2195–2204.
28. Du, J.; Han, J.; Way, A.; Wan, D. Multi-Level Structured Self-Attentions for Distantly Supervised Relation Extraction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2216–2225.
29. Ren, F.; Zhou, D.; Liu, Z.; Li, Y.; Zhao, R.; Liu, Y.; Liang, X. Neural relation classification with text descriptions. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1167–1177.
30. Vashishth, S.; Joshi, R.; Prayaga, S.S.; Bhattacharyya, C.; Talukdar, P. Reside: Improving distantly-supervised neural relation extraction using side information. *arXiv* **2018**, arXiv:1812.04361.
31. Han, X.; Liu, Z.; Sun, M. Neural Knowledge Acquisition via Mutual Attention Between Knowledge Graph and Text. In Proceedings of the AAAI, New Orleans, LA, USA, 2–7 February 2018; pp. 4832–4839.
32. Zeng, W.; Lin, Y.; Liu, Z.; Sun, M. Incorporating relation paths in neural relation extraction. *arXiv* **2016**, arXiv:1609.07479.
33. Palmer, M.; Gildea, D.; Kingsbury, P. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.* **2005**, *31*, 71–106. [CrossRef]
34. Schuler, K.K. *VerbNet: A Broad-Coverage, Comprehensive verb Lexicon*; University of Pennsylvania: Philadelphia, PA, USA, 2005.
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
36. Xiao, Y.; Jina, Y.; Cheng, R.; Hao, K. Hybrid Attention-Based Transformer Block Model for Distant Supervision Relation Extraction. *arXiv* **2020**, arXiv:2003.11518.
37. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 2787–2795.
38. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.

39. dos Santos, C.; Xiang, B.; Zhou, B. Classifying Relations by Ranking with Convolutional Neural Networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 26-3 July 1 2015; pp. 626–634.
40. Jat, S.; Khandelwal, S.; Talukdar, P. Improving distantly supervised relation extraction using word and entity based attention. *arXiv* **2018**, arXiv:1804.06987.