AI

MDPI

Article

Improving Distantly Supervised Relation Extraction with Multi-Level Noise Reduction

Wei Song and Zijiang Yang *

Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China; songwei@jiangnan.edu.cn
* Correspondence: 6223111073@stu.jiangnan.edu.cn

Abstract: *Background*: Distantly supervised relation extraction (DSRE) aims to identify semantic relations in large-scale texts automatically labeled via knowledge base alignment. It has garnered significant attention due to its high efficiency, but existing methods are plagued by noise at both the word and sentence level and fail to address these issues adequately. The former level of noise arises from the large proportion of irrelevant words within sentences, while noise at the latter level is caused by inaccurate relation labels for various sentences. *Method*: We propose a novel multi-level noise reduction neural network (MLNRNN) to tackle both issues by mitigating the impact of multi-level noise. We first build an iterative keyword semantic aggregator (IKSA) to remove noisy words, and capture distinctive features of sentences by aggregating the information of keywords. Next, we implement multi-objective multi-instance learning (MOMIL) to reduce the impact of incorrect labels in sentences by identifying the cluster of correctly labeled instances. Meanwhile, we leverage mislabeled sentences with cross-level contrastive learning (CCL) to further enhance the classification capability of the extractor. *Results*: Comprehensive experimental results on two DSRE benchmark datasets demonstrated that the MLNRNN outperformed state-of-the-art methods for distantly supervised relation extraction in almost all cases. *Conclusions*: The proposed MLNRNN effectively addresses both word- and sentence-level noise, providing a significant improvement in relation extraction performance under distant supervision.

Keywords: distant supervision; neural relation extraction; multi-instance learning; noise reduction

Citation: Song, W.; Yang, Z. Improving Distantly Supervised Relation Extraction with Multi-Level Noise Reduction. *AI* **2024**, *5*, 1709–1730. https://doi.org/10.3390/ai5030084

Academic Editor: Demos T. Tsahalis

Received: 24 August 2024
Revised: 14 September 2024
Accepted: 18 September 2024
Published: 23 September 2024

1. Introduction

Relation extraction (RE) aims to identify semantic relationships between entity pairs from sentences of large corpora, which is a crucial task in the fields of natural language processing (NLP). Since it has become a foundation for widespread downstream applications such as question answering [1], knowledge graph [2], and information retrieval [3], relation extraction has garnered significant attention over the past decade.

Conventional relation extraction methods adopt supervised training [4] and are constrained by a lack of large-scale manually labeled data. To reduce the high cost of human labor, the distant supervised method was proposed [5], which automatically generates training data by aligning a knowledge base (e.g., Freebase) with plain texts (e.g., New York Times). Thus, the objective of distant supervision relation extraction (DSRE) is to train an extractor with these automatically generated training data to identify semantic relations from large-scale text corpora. Specifically, this method relies on the strong assumption that, if an entity pair in the knowledge base participates in a relation, then all sentences mentioning these two entities will express that relation. Table 1 provides an example of the alignment process. As illustrated, alignment may lead to incorrect labeling problems, which refers to sentence-level noise. For example, sentences *S*2 and *S*3 do not express the relationship of 'founders' but are still in the bag. Apart from sentence-level noise, distantly supervised methods also suffer from word-level noise, which arises from the presence

of irrelevant words within a sentence. The noisy words inside a sentence diminish the importance of the keywords that help relation extractors discriminate relations. Therefore, it is necessary to address this multilevel noise to train an efficient relation extractor.

Table 1. The alignment process between KB and plain texts.

Relation triple in KB	**business/company/founders (Amazon, Jeffrey P. Bezos)**
Sentences in plain texts	***S*1:** Virtual world proponents, which include a roster of Linden Labs investors, among whom is **Jeffrey P. Bezos**, the founder of **Amazon** ***S*2: Jeffrey P. Bezos**, the **Amazon** chief executive, gave a speech about the early stages of technology. ***S*3:** Among the speakers will be **Jeffrey P. Bezos**, below, of **Amazon**, who is expected to talk about selling more Web services to business customers.

First, alleviating the impact of noisy words while identifying the significant words is challenging for relation extraction. For instance, as shown in Table 1, the relationship expressed in S1 corresponds to the label. However, the phrase 'Jeffrey P. Bezos, the founder of Amazon' alone suffices to denote the relation, which is much shorter than the sentence. Moreover, according to Liu et al. [6], word-level noise is widespread in datasets such as NYT-10, with approximately 12 noisy words present in each sentence. Several methods have been proposed for the sake of removing noisy words, such as dependency tree parser [6,7] and word-level attention [8–11] methods. The former is limited by fixed syntactic patterns, making it unable to handle sentences from large-scale corpora. The latter can be broadly divided into two categories. The first category applies the selective attention mechanism, as previous works demonstrated the effectiveness of relationship query vectors [9,12]. They calculate the relevance between words and entity pairs or single relation query vectors directly to determine the weight of each word. However, these approaches are somewhat coarse, as the initially input pretrained word vectors are insufficiently precise to reflect the semantics of the words in their specific sentence context. For example, Table 2 illustrates the correlation (each row sums to 1) between the embeddings of each word and the entity pair during the training of previous work using sentence S1 from Table 1. It can be observed that the weights of the keywords were not clearly highlighted, which implies that word noise has not been effectively removed, and it would be difficult to accurately enhance these weights during subsequent training processes without providing proper guidance to the extractor. The second category [13–15], in contrast, adopts BERT [16] as a sentence encoder, which applies the self-attention mechanism to each word in the sentence. Such a computationally intensive approach enables it to capture intricate dependencies and relationships between any two words in a sentence. However, in the relation extraction task, it is sufficient to capture only the features pertinent to the relationships between entity pairs, rather than other information within the sentence. Hence, the existing methods either introduce additional computational complexity or fail to accurately extract relation features.

Table 2. Weights (correlation between words and entity pair) of words in ***S*1** during training.

Words in *S*1	**Virtual World**	**Proponents**	…	**Jeffrey P. Bezos**	…	**Founder**	…	**Amazon**
Weights	0.058	0.052	…	0.105	…	0.062	…	0.101

Second, addressing the noise from the incorrect labels generated by distant supervision has long been a challenge for DSRE. Multi-instance learning (MIL) and its variants have been proposed to address the issue of sentence-level noise. For the sake of simplicity, sentences labeled with correct relations are called true instances, and false instances represent incorrectly labeled ones. They generate the representation of a bag by selecting the most likely sentence [17] or computing the weighted sum of the sentences in the bag [18],

and several recent works have utilized multi-level attention mechanisms to highlight high-quality features and de-emphasize false ones [19–21]. These methods either omit some true instances or neglect the positive effect of false instances. Thus, we contend that both problems should be considered in MIL. For true instances, when constructing bag-level representations, sentence-level features with similar semantics (true instances) should be emphasized, while those with different semantics should be suppressed. If the model can identify more true instances, this indicates that it has learned a broader range of syntactic patterns and richer features. Consequently, the bag representation refined from these instances will not only be accurate, but also more comprehensive. Otherwise, the information contained in bag-level representations may be insufficiently representative. As for the false instances, recent studies [22–24] have demonstrated that a considerable number of false instances are abandoned during the training process due to being treated as non-informative noise sentences. Specifically, no matter how many sentences a bag contains, the bag-level representation is formed by the weighted sum of all sentences, where the weights sum to 1. However, the majority of sentences have very low weights, meaning they hardly participate in the final training. Thus, the information contained in these numerous discarded sentences is squandered. If effectively utilized, it could significantly improve the performance of our model. Some efforts have been made to utilize false instances [23–25], but they focused solely on improving the generalization or robustness of the models, overlooking the first issue. We suggest that the information provided by false instances can assist the extractor in better identifying true instances, thereby allowing both issues to be addressed in an interactive manner. Therefore, an advanced multi-instance learning method ought to strive to thoroughly exploit true instances, while effectively utilizing the useful information provided by false instances.

In this article, both word-level and sentence-level noise are carefully considered, and a multi-level noise reduction neural network (MLNRNN) is proposed to jointly address the aforementioned challenges. For the first challenge, we design an iterative keyword semantic aggregator (IKSA), which adequately tackles word-level noise with reasonable computational cost. We leverage the entity pair information and the global context information of the sentence to capture a relation query vector for each sentence, as the relationship expressed by the same pair of entities can vary across different sentences. This vector is then used as an enhancement of the original input to help exploit the global dependency [9] of each word for initial denoising through context-to-relation attention. Subsequently, we take this relation vector as guidance to further integrate the word information that is crucial to relationships, to form the ultimate discriminative feature of a sentence. During this iterative process, we also incorporate self-attention layers to exploit pairwise dependencies [26], enabling a comprehensive extraction of relation features. For the second challenge, we propose multi-objective multi-instance learning to identify all true instances, while simultaneously properly utilizing false instances. In MOMIL, sentences within a bag are divided into true and false sets based on the representation assignment algorithm, and each set is then used for different training objectives. We designate the most likely sentence and its nearest neighbors in the relation space as true instances, while the remaining sentences are classified as false instances. Then, we use the true instances to train the extractor with the bag label after the semantic enhancement operation to refine the features. Meanwhile, the false instances are fed into our cross-level contrast learning (CCL) module, utilizing both sentence-level and bag-level information to construct positive and negative instance pairs, refining features that fully capture the sentence semantics and interactively contribute to the construction of a bag-level representation. These are the two objectives MOMIL aims to achieve. In particular, the goal of CCL is that the instances actually sharing the same relational triples (i.e., positive pairs) should be close in the relation space, while the representations of instances with different relational triples (i.e., negative pairs) should be far apart. Therefore, false instances, which have been treated inappropriately in previous works, can have a positive influence on identifying true instances in the MLNRNN, and both true and false instances contribute to strengthening the relation features collaboratively.

In summary, we make the following major contributions in this article:

- In this work, we propose a novel neural network named MLNRNN to jointly handle multi-level noise for relation extraction. The first module, the IKSA, continuously refines and aggregates keyword semantics within a specific sentence, while removing noisy words and ultimately capturing salient relation features.
- The next module, MOMIL, is designed to alleviate the influence of sentence-level noise by identifying all true instances and enhancing their features. Moreover, we leverage false instances with cross-level contrastive learning to further improve the classification capability of the extractor in MOMIL.
- Our experiments on two DSRE benchmark datasets (NYT-10 and NYT-16) demonstrated the improved performance of our approach compared to seven state-of-the-art methods.

2. Related Work

2.1. Neural Relation Extraction

Neural networks have proven to be useful methods for handling a wide range of NLP tasks [27,28], due to their ability to extract semantic meaning without hand-designed features. Convolutional neural networks (CNN) and recurrent neural networks (RNN) have been shown to be effective for relation extraction [29–31]. Built on a CNN, the classification by ranking CNN (CR-CNN) computed a distributed vector representation of text by minimizing the pairwise ranking loss function, and implemented relation classification by matching such a vector representation with the relation space [29]. Subsequently, a piece-wise convolution neural network (PCNN) [17] was proposed to maintain more discriminative features with a convolutional operation and piecewise max pooling. Moreover, a multi-level attention-based convolution neural network was proposed to capture both primary attention at the input level, centered on the target entities, and secondary pooling attention, centered on the target relations. Building on the RNN, bidirectional long-short-term memory (Bi-LSTM) leverages the inherent dependencies in each entity pair to predict relationships. The bidirectional gated recurrent unit (Bi-GRU) combined a gated recurrent unit with a word-level attention mechanism to acquire the semantic relation features of sentences [30]. Both of the aforementioned neural network architectures have inherent issues when modeling long sequences. CNNs fail to capture long-range dependencies and are ineffective in handling word noise. Recurrent network architectures have long been plagued by the problem of forgetting long-term information.

To better remove noisy words, the dependency path between entities was proposed and demonstrated to be effective in [32,33], but this was limited in handling a large number of relations, due to the constraints of fixed syntactic patterns. Existing methods that apply an attention mechanism at word-level treat the words in a sentence as isolated, without considering their specific semantic information within the sentence. Recently, several methods have taken the BERT model as their sentence encoder, which employs a transformer-based architecture that allows it to capture nuanced meanings and dependencies in text at word level; it can adapt to various NLP tasks but results in significant computational overheads. By contrast, we propose an iterative method that performs denoising and extraction of sentence relation features based on the interaction of word information, with the guidance of a sentence-specific relation vector, rather than simply using basic word-level attention and combining this with other complex models as auxiliary information, which tends to consume a large amount of computation.

2.2. Multi-Instance Learning

To mitigate the effects of mislabeled sentences in automatically generated datasets for distantly supervised relation extraction, neural relation extractors were combined with multi-instance learning algorithms in [17,18]. A PCNN selects the sentence that is most likely to match the relation of the bag and ignores the remaining sentences. Selective attention over instances (ATT) employs an attention mechanism to assign distinct weights

to individual sentences within a bag. Relation extraction with joint label embedding (RELE) leverages external knowledge to calculate the weights of sentences by utilizing advanced learned label embeddings [34]. Cross-relation cross-bag selective attention (C2SA) [12] and intra-bag and inter-bag attentions (Intra-Inter Bag) [19] tackle negative instances at both bag level and sentence level. To excavate potentially useful information, the soft-label (SL) model adopted a dynamic correction mechanism to adjust inaccurate labels throughout the training process [35]. Deep clustering-based relation extraction (DCRE) employed an unsupervised deep clustering approach to assign reliable labels to negative instances [36]. Collaborative adversarial training revitalizes false instances from MIL by leveraging virtual adversarial techniques, thus improving data utilization [24]. False negative adversarial networks (FAN) harmonizes false-negative samples into the cohesive feature space, creating a unified representation that allows the effective assignment of pseudo-labels [37].

Although the aforementioned methods can address both issues separately, they have not been considered simultaneously. Considering that both issues affect the model's performance to varying degrees and can be viewed as different facets of multi-instance learning, they should be addressed concurrently. Thus, we proposed MOMIL to tackle both issues in an interactive manner. We explore true instances as thoroughly as possible by selecting the most likely sentence and its nearest neighbors, and for the sentences that are not selected, we use cross-level contrastive learning to obtain more refined features to help the extractor identify more true instances.

In addition, artificial intelligence (AI) security issues are becoming increasingly critical, especially in the context of deep learning models in DSRE. Recent works [24,38,39] have shown that neural networks are vulnerable to adversarial attacks, where small perturbations in input data can cause significant misclassifications, either by misleading the model into classifying the input as a specific target class or as an incorrect class in untargeted attacks. Thus, adversarial training has emerged as an effective approach to enhance the robustness of neural networks against such perturbations. By incorporating adversarial examples during the training process, models can learn to resist these attacks and maintain a stable performance, even when subjected to adversarial interference.

3. Materials and Methods

In the distantly supervised relation extraction paradigm, a bag is formed by grouping all sentences annotated with the same relation triple, and each sentence is referred to as an instance. A relation triple is represented as [head, relation, tail], where head refers to the head entity and tail refers to the tail entity. There are M bags $\{B_1, \ldots, B_M\}$ in the training set and the ith bag contains N instances $B^i = \{S_1^i, \ldots, S_N^i\} (i = 1, \ldots, M)$. The purpose of relation extraction is to predict the label of unlabeled bags. As shown in the overall framework of the model in Figure 1, our method can be divided into two parts:

Iterative Keyword Semantic Aggregator. Given an instance S^* with two target entities, the IKSA encodes it into a high-quality sentence representation using the semantics of salient words with a novel neural network, which is more accurate and more and efficient than previous sentence encoders.

Multi-Objective Multi-instance Learning. Given a bag of instances B^* and two target entities, we handle true instances and false ones in different ways, to collaboratively form higher-quality bag-level representations for training the extractor.

3.1. Iterative Keyword Semantic Aggregator

The proposed iterative keyword semantic aggregator (IKSA) utilizes the information of the entity pair and global sentence information to capture a specific relation vector for each sentence, which is then used to remove noisy words and for semantic aggregation. Due to the fact that pairwise and global dependencies within the sentence ought to be jointly considered, we first utilize context-to-relation attention to explore the global dependency of each word on the entire sentence within the context of the relation vector and entity pair information, serving as an initial denoising step. Then, we aggregate the semantics of

the keywords guided by this information, to form the distinctive sentence representation. Furthermore, we apply standard token-to-token self-attention to produce a context-aware representation for each token in light of its syntactic dependencies on other tokens from the same sequence, which is computationally expensive but necessary. Finally, we repeat the last two steps to iteratively derive the refined semantic relation features, and the detailed structure of the IKSA is shown in Figure 2.

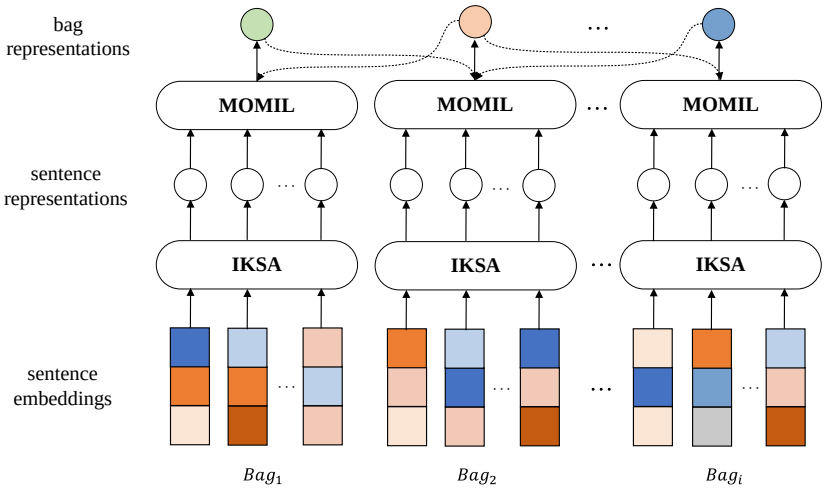

Figure 1. Overall framework of the proposed multi-level noise reduction model.The dashed arrows indicate the utilization of representations from other bags in CCL. Different colors are used to distinguish between the variations in the input sentence embeddings.

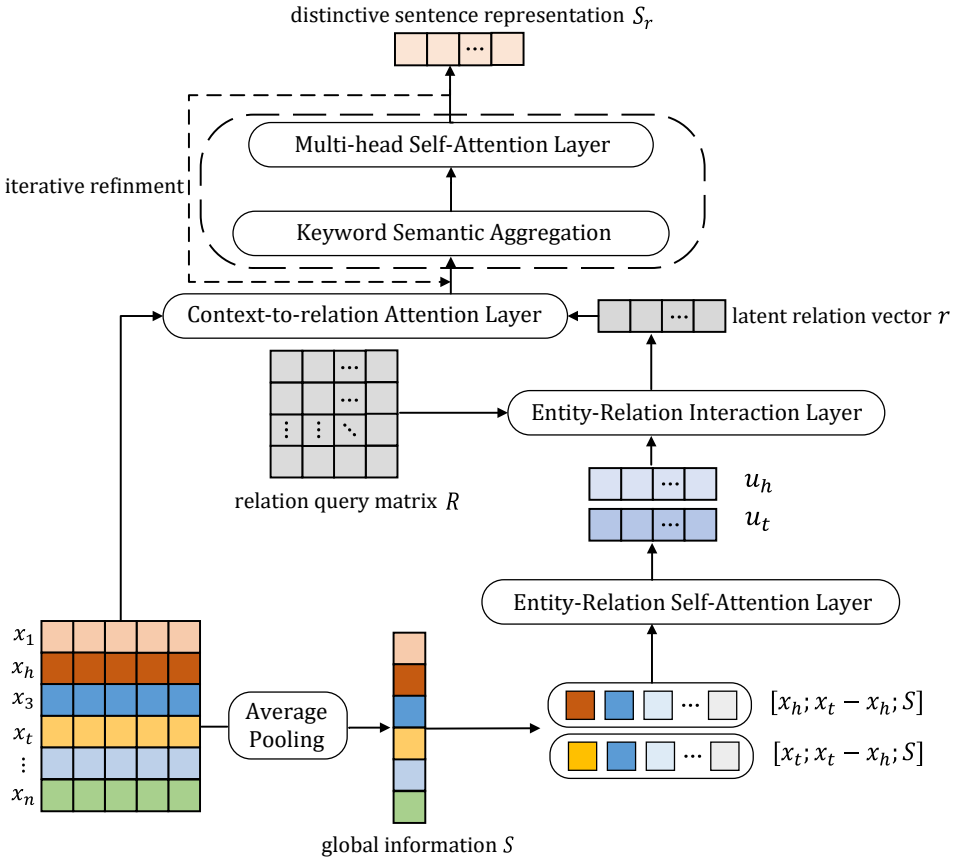

Figure 2. The detailed structure of the IKSA, illustrating the procedure for handling a sentence, and the dashed part represents the iterative step of this module.

Input Representation. Tokens in sentences should be embedded into distributed representations for mathematical operations in neural networks [40]. For the input tokens $\{t_1, \ldots, t_h, \ldots, t_t, \ldots, t_m\}$ in a sentence, where t_h and t_t represent the head entity and tail entity, respectively, we train the token t_i to vector $x_i \in \mathbb{R}^{d_w}$ in a priori manner with the use of GloVe [40].The parameter d_w indicates the dimension of the word. In addition, to encode the sentence in an entity-aware manner, relative position embedding [17] is leveraged to represent the position information in the sentence. For example, the relative distances from the token "founder" to the head entity [Jeffery P. Bezos] and the tail entity [Amazon] are -2 and 2, respectively, in sentence S1 from Table 1. Finally, the representation of an input token is the concatenation of word embedding and position embedding. We denote all the input tokens in a sentence as an input matrix $X = \{x_1, \ldots, x_h, \ldots, x_t, \ldots, x_m\}$, where $x_i \in \mathbb{R}^{1 \times d_x} (d_x = d_w + 2 * d_p)$ and m is the number of tokens in a sentence.

Capture Sentence-specific Relation Vector. Inspired by TransE [41], which treats the embedding representation of the relationship between the two entities as a transformation of the embedding representations of the two entities: $h + r \approx t$, we argue that $x_t - x_h$ can only approximate part of the relation between the two entities. However, the same entity pair may correspond to different relationships in different contexts, and the embeddings of the entities are fixed. Therefore, the IKSA module considers the potential relationships of entities within this context, to obtain a latent vector r between the entity pair.

Specifically, we first perform a compression operation by leveraging the global average pooling, which retains the overall context information $S \in \mathbb{R}^{1 \times m}$,

$$S = [\frac{1}{d_x} \sum_{j=1}^{d_x} X_{1,j}, \ldots, \frac{1}{d_x} \sum_{j=1}^{d_x} X_{m,j}], \tag{1}$$

where $X_{i,j}$ indicates the j-th dimension of the i-th word's features in the sentence input. Subsequently, the rough relation $x_t - x_h$ and the global contextual information S are concatenated with the embeddings of the entity pair to be the input of the entity-relation self-attention layer:

$$u_h = tanh(W_p[x_h; x_t - x_h; S]^T), \tag{2}$$

$$u_t = tanh(W_p[x_t; x_t - x_h; S]^T), \tag{3}$$

where $tanh(\cdot)$ represents the tanh non-linear function, $W_p \in \mathbb{R}^{r_{hid} \times (2d_x + m)}$ is the weight matrix, and r_{hid} is the dimension of the relation vector. Next, this section designs an entity-relation interaction layer, which uses a learnable relation query matrix $R \in \mathbb{R}^{t \times r_{hid}}$ with t relations to interact with $[u_h, u_t]$, obtaining a weight matrix $A \in \mathbb{R}^{t \times 2}$ for the two entities regarding each relation:

$$E = R[u_h, u_t], \tag{4}$$

$$A_{ij} = \frac{exp(E_{ij})}{\sum_k^t exp(E_{kj})}. \tag{5}$$

In other words, each row of the matrix R is a query vector, which is a representation of a specific type of relation. Since the potential relationship ought to consider both entities, a weight α_j is assigned to each relation:

$$\alpha_j = \sum_{j=1}^{2} A_{ij}, \tag{6}$$

The latent relation vector r is obtained using the weighted sum of all relations, serving as a compact representation of relations for a specific sentence:

$$r = \sum_{i=1}^{t} \alpha_i R_i, \tag{7}$$

where R_i is the i-th row of matrix R. The resultant latent relation vector r corresponds to the relation features of its sentence, and is the latent representation of the relationship

expressed by the sentence, which will also be utilized in the forthcoming denoising and semantic aggregation processes.

Context-to-relation Attention. To exploit the global dependencies of each word in expressing a relation, we first utilize the acquired relation vector r and the information of the entity pair to enhance the original input X, transforming it into $\tilde{X} = \{\tilde{x_1}, \ldots, \tilde{x_h}, \ldots, \tilde{x_t}, \ldots, \tilde{x_m}\}$, $\tilde{x_i} = [x_i; x_h; x_t; r] \in \mathbb{R}^{1\times(3d_x+r_{hid})}$. Following this is the operation to obtain the dependency of each token from the enhanced input:

$$V = \sigma(W_v^2 \delta(W_v^1 \tilde{X}^T + b_v^1) + b_v^2), \tag{8}$$

where $W_v^1 \in \mathbb{R}^{d_{hid}\times(3d_x+r_{hid})}$ and $W_v^2 \in \mathbb{R}^{d_x\times d_{hid}}$ represent two weight matrices, and $b_v^1 \in \mathbb{R}^{d_{hid}}$ and $b_v^2 \in \mathbb{R}^{d_x}$ represent their bias terms, respectively, for calculating $V \in \mathbb{R}^{d_x\times m}$. Accordingly, we leverage a sigmoid active function $\sigma(\cdot)$ to set the output in the range between 0 and 1, and $\delta(\cdot)$ refers to the Gaussian error linear unit (GELU) function, as the input of neurons tends to follow a normal distribution. According to the dimension of V, it computes a score for each feature of each word, so it can select the features that can best describe the word's relational meaning in the enhanced sentence:

$$H = V^T \odot X, \tag{9}$$

where $\odot$ represents element-wise multiplication. Hence, such information is preserved in the output $H = \{h_1, \ldots, h_m\} \in \mathbb{R}^{m\times d_x}$ for further relation feature extraction.

Relation Semantic Aggregation and Word Interaction. To obtain accurate relation features, we perform keyword semantic extraction and aggregation in this section. Additionally, we consider the specific meanings of the same word in different contexts through word-level interactions. In other words, after the semantic extraction of each keyword, we update the word vectors using sequence self-attention to more comprehensively form a sentence representation. Preliminarily, we project $[x_h; r]$ and $[x_t; r]$ into the vector space of word embeddings to serve as the two target vectors $[q_h, q_t] \in \mathbb{R}^{2\times d_x}$ for information aggregation. As a result, we select words that are relatively relevant to the target vectors. The relevancy matrix $A^{'} \in \mathbb{R}^{2\times m}$ is computed with the following operation:

$$E^{'} = [q_h, q_t]H^T, \tag{10}$$

$$A_{ij}^{'} = \frac{exp(E_{ij}^{'})}{\sum_k^t exp(E_{kj}^{'})}. \tag{11}$$

For the sake of selecting the top k tokens, we compute the weights α for each token with a relevance matrix using equation $\alpha_j = \sum_{i=1}^{2} A_{ij}^{'}$. Meanwhile, the aggregated relation features are obtained:

$$[Q_h, Q_t] = A^{'}H. \tag{12}$$

It is necessary to note that the weights of the entity pair are naturally higher when calculating relevance, because an accurate relational feature must include the information of the entity pair along with the words describing the relationship. To avoid misleading the extractor, the aggregated $\tilde{Q} = [Q_h, Q_t]$ does not incorporate information from q_h or q_t in a weighted manner, as the relation vector contained is not inherently presented in the sentence itself. Then, the aggregated information is activated by two linear transformations with a ReLU activation in between:

$$Q = max(0, W_q^1 \tilde{Q} + b_q^1)W_q^2 + b_q^2, \tag{13}$$

where W_q^1 and W_q^2 are learnable parameters that keep Q the same shape of $\tilde{Q}$. To implement word interaction, multi-head self-attention (MHSA) is utilized to obtain the dependencies between every two tokens in H. For clarity, we first give the definition of the dot-product attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{14}$$

where Q, K, and V represent the query, key, and value, respectively. Note that they are all derived from H through three different transformation matrices in IKSA: $Q = HW^Q$, $K = HW^K$, $V = HW^V$. Subsequently, these three matrices can each be replaced by n matrices of the same shape to form n heads:

$$head_i = \text{Attention}(Q_i, K_i, V_i), \tag{15}$$

$$\text{MHSA}(Q, K, V) = Concat(head_1, \ldots, head_n)W^O. \tag{16}$$

Based on the fact that the MHSA keeps the input shape the same as the output shape, when the word interaction is insufficient, we repeat the above operations in this section to obtain more comprehensive semantics for each word and more refined relation features. For example, the input of the i-th relation semantic aggregation is H^{i-1}, $[Q_h^{i-1}; r]$, and $[Q_t^{i-1}; r]$. Correspondingly, the skip connection is taken into account of in the output of each operation:

$$Q^{j+1} = Q^j + A'^j H^j. \tag{17}$$

The residual information can facilitate the deeper layer training of the extractor, and i represents the i-th iterative process. In addition, layer normalization is also considered during different iterations to stabilize and accelerate the training process. Eventually, a distinctive sentence representation $S_r \in \mathbb{R}^{d_r}$ is fitted by a neural layer in following operation:

$$S_r = W_r^1(Q_h^i)^T + W_r^2(Q_t^i)^T + b_r, \tag{18}$$

where $W_r^1, W_r^2 \in \mathbb{R}^{d_r \times d_x}$, and Q_h^i and Q_t^i are the final keyword semantic aggregations for the head and tail entities, respectively.

3.2. Multi-Objective Multi-Instance Learning

In this section, we present our proposed multi-objective multi-instance learning (MOMIL) module, as shown in Figure 3, to alleviate the influence of sentence-level noise. We focus on handling multiple relations within a bag of sentences, taking into account both the bag labels and the potential relations of false instances.

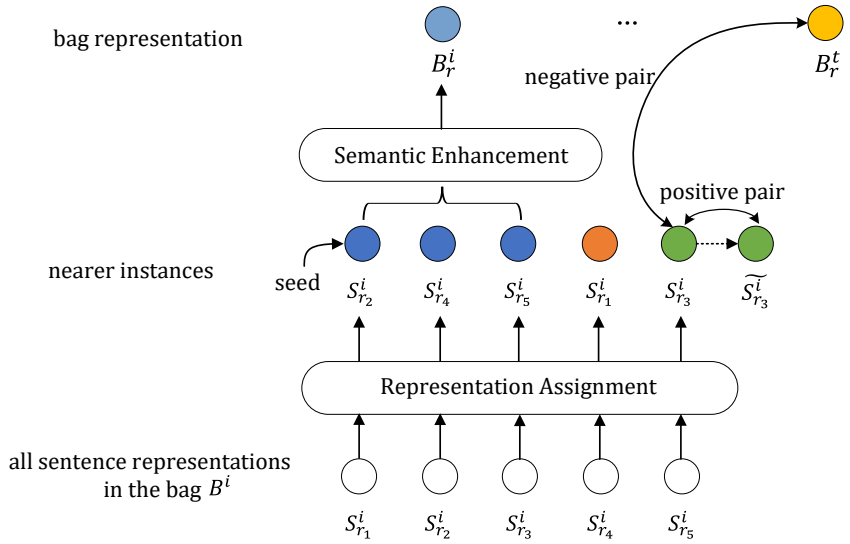

Figure 3. Overview of multi-objective multi-instance learning. The sentence representations highlighted in blue represent the selected true instances, while the others are false instances.

Given a sentence representation S_r produced by the IKSA, we select the instance that best matches the prediction relation r as the seed true instance. Intuitively, we argue that instances with a distance less than a certain threshold Th_d from the seed true instance express the same relation. In other words, instances that express the same relations can be clustered in a relation space. Under such an assumption, an appropriate threshold and a proper clustering algorithm are crucial. Otherwise, instances expressing different relationships might be clustered together, or other true instances might be missed. Therefore, we chose a tight threshold along with a greedy algorithm, as shown in Algorithm 1, to avoid omitting true instances.

Given two sentence representations S_r^a and S_r^b, we encode them into probability distributions p_a and p_b. We adopt a JS distance of (p_a, p_b) as the distance between p_a and p_b, which is computed using the Jensen–Shannon (JS) divergence:

$$D_{JS}(p_a \parallel p_b) = \frac{1}{2} D_{KL}(p_a \parallel p_b) + \frac{1}{2} D_{KL}(p_b \parallel p_a), \tag{19}$$

The JS divergence is the symmetrized and normalized version of the Kullback–Leibler (KL) divergence:

$$D_{KL}(p_a \parallel p_b) = \sum_{i}^{t} p_a(i) \log \frac{p_a(i)}{p_b(i)}, \tag{20}$$

whose advantages can make calculations and threshold adjustments more convenient. Specifically, the value of the JS divergence ranges from 0 to 1; the closer it is to 0, the more similar the two distributions are, while the value of the KL divergence has no upper limit.

With the representation assignment algorithm, all instances in a bag are categorized into true and false sets. Referring to the concept from Zhou et al. [42], semantic enhancement is applied to true instances to distill more similar relation features in these sentence representations, forming an ultimate accurate bag representation for training. The weight of each instance is determined based on its correlation with other instances:

$$e_i = \sum_{j=1,2,\ldots,k, j \neq i} \frac{S_r^i}{\parallel S_r^i \parallel_2} \left(\frac{S_r^j}{\parallel S_r^j \parallel_2} \right)^T, \tag{21}$$

where k is the size of set V, $\parallel \cdot \parallel_2$ denotes the Euclidean norm (or 2-norm) of a vector, and the bag representation B_r is obtained using the weighted sum of these true instances after softmax. In this way, features that are less relevant to the relation will be further filtered out. Unlike previous methods that directly assign weights to sentences in a bag, we refine the features after identifying the set of true instances. Our approach prevents any single sentence that best aligns with the label relationship from dominating with an excessively high weight. Instead, we focus on extracting similar features from the set of positive instances, leading to a more even distribution of weights and, consequently, a more comprehensive and accurate representation of the bag.

Relation Prediction. To make the use of comprehensive bag feature a fully connected layer, a $tanh(\cdot)$ activation function is adopted, which aims to perform a nonlinear transformation and map B_r to the relation prediction space o:

$$o = W_B tanh(B_r) + b_B \tag{22}$$

where $W_G \in \mathbb{R}^{t \times d_r}$ and $b_G \in \mathbb{R}^t$ represent the weight and the bias. Then, a softmax classifier is utilized to predict the entity relation $p(r|B_r; \theta)$:

$$p(r|B_r; \theta) = \frac{exp(o_i)}{\sum_{j=1}^{t} exp(o_j)}. \tag{23}$$

Algorithm 1 Representation assignment

Require: sentence representations in a bag B^i, threshold Th_d
Ensure: true instances set V
Add the most possible true instance S_r^* to set V
Initialize the queue Q with S_r^*
while $Q \neq \emptyset$ **do**
  Dequeue the first element S_r from Q
  Compute the distances d_i between S_r and instances in $[B - V]$
  if $d_i < Th_d$ **then**
    Add the corresponding S_r^i to V
    Enqueue S_r^i into Q
  end if
end while

We define the objective of classification using a cross-entropy function, as follows:

$$\mathcal{L}_{CE}(\theta) = -\frac{1}{M}\sum_{i=1}^{M} logp(r_i|B_r^i;\theta), \tag{24}$$

where M is the number of bags.

Cross-level Contrastive Learning. For false instances, we implement cross-level contrastive learning to exploit useful information within them. Previous works have shown the effectiveness of utilizing cross-level information. Thus, for the i-th false instance S_r^i, we chose the representation of other bags $B_r^t (t \neq i)$ as its negative pair, since is has been denoised and naturally has a different relation triple from false instances within the original bag. However, it is not clear to which other instance expresses the same relationship as the false instance S_r^i when constructing a positive pair. Therefore, we generate a positive instance with dynamic gradient perturbations pt_{adv} to solve this issue, since a previous work [43] proved its effectiveness in creating a pseudo-positive sample with minimal deviation from the original sample, making sure that they are sufficiently similar:

$$\tilde{S}_r^i = S_r^i + pt_{adv_i} = S_r^i + \varepsilon \frac{g_i}{\| g_i \|_2} \tag{25}$$

where $g_i = \nabla_{B_r^i} \mathcal{L}_{CE}(B_r^i;\theta)$ is the gradient from the loss function, ε is a hyperparameter that regulates the degree of disturbance. Figure 3 shows an example of how to construct a positive pair and negative pair in a bag. We define a objective using InfoNCE [23] loss for the representation S_r^i:

$$\mathcal{L}_{CL}(S_r^i;\theta) = -log\frac{e^{\text{sim}(S_r^i,\tilde{S}_r^i)}}{e^{\text{sim}(S_r^i,\tilde{S}_r^i)} + \sum_{t:t\neq i} e^{\text{sim}(S_r^i,\tilde{B}_r^t)}} \tag{26}$$

where $\text{sim}(a, b)$ indicates a cosine function to measure the similarity between two sentence representations. Through such a design, we can bring the sentence representations of the same relational triples closer together, while pushing the representations of different relational triples further apart.

3.3. Training Objective

Considering the initial stage of training, our extractor should first obtain an accurate encoder IKSA for predicting entity relations. Thus, we introduce a increasing function $\lambda(s)$ with respect to the training step s into the training objective:

$$\lambda(s) = \frac{2}{1 + e^{-s}} - 1. \tag{27}$$

Accordingly, the training objective is defined as

$$\mathcal{L}(\theta) = \mathcal{L}_{CE}(\theta) + \frac{\lambda(s)}{F} \sum_{i}^{F} \mathcal{L}_{CL}(S_r^i; \theta) \tag{28}$$

where F is the number of false instances in a batch. The value of $\lambda(s)$ progressively approaches 1 as the relative training steps s increases, thereby allocating greater emphasis to the CCL.

4. Results

In this section, we present a series of experiments to demonstrate the effectiveness of our proposed method. With this aim, we first introduce the datasets and evaluation metrics. Then, we discuss the parameter settings and compare the MLNRNN with seven DSRE competitors. Additionally, we carried out an ablation study to investigate the effectiveness of each component. Furthermore, a complexity analysis is provided to assess the computational efficiency of the method. Finally, we conclude with a case study, for a more specific evaluation.

4.1. Datasets and Evaluation Metrics

In our experiments, we adopted three DSRE benchmark datasets: NYT-10 [44], NYT-16 [17], and Wiki-20m [13]. Both NYT-10 and NYT-16 were generated by aligning the Freebase entity relation with the New York Times corpus and support 53 different relations, including NA. NYT-10 takes the corpus from 2005 to 2006 as a training set and employs the data of 2007 as a test set. In detail, the training set contains 522,611 sentences, 281,270 entity pairs, and 18,252 relation facts, while the testing set contains 172,448 sentences, 96,678 entity pairs, and 1950 relation facts. NYT-16 provides 112,941 sentences, 65,726 entity pairs, and 4266 relation facts as a training set, and 152,416 sentences, 93,574 entity pairs, and 1732 relation facts for testing. In contrast, Wiki-20m is a recently released and larger dataset for training DSRE models. It consists of 6,987,222 sentences, 304,870 entity pairs, and 157,740 relation facts for training, along with 137,986 sentences, 74,390 entity pairs, and 56,000 relation facts for testing.

To ensure fair comparisons, each DSRE method was evaluated using a held-out test, where the precision and recall were determined by comparing the predictions of models with the relational facts in the datasets. In our experiments, we employed precision–recall (PR) curves, P@N(s), and the area under the precision–recall curve (AUC) as metrics for evaluation. P@N(s) was calculated from the top N predictions that were sorted by the confidence score of each bag. It is well-established that the NYT-10 and NYT-16 datasets, both of which are auto-labeled, contain numerous incorrect labels and duplicate instances, hurting the performance of DSRE methods. In line with previous studies [12,14,19], we used the metrics Top-100, Top-200, and Top-300 P@N(s) for these datasets. However, Wiki-20m is a human-labeled dataset, which results in the better performance of DSRE models, due to the high-quality annotations it provides. We leveraged Top-30,000, -40,000, and -50,000 P@N(s) for Wiki-20m to enable a more accurate performance assessment.

4.2. Parameter Settings

In the experiments, word embeddings were trained a priori using GloVe [40] on the two datasets. In our work, we concatenated the words of an entity horizontally if it consisted of multiple words. We employed the Adam optimizer to train the objective function. The dimension for word embeddings d_w and position embeddings d_p were set to 50 and 5, respectively. Correspondingly, the dimensions of the relation query vector r_{hid} and the sentence representation d_r were both 60. We adjusted parameters through cross-validation and grid search, including a distance threshold Th_d, disturbance regulator ε, head count of MHSA l, dropout rate DR to avoid overfitting, learning rate LR, and batch size b_s. Table 3 lists all the parameter settings for our model.

Table 3. Model parameters and their values.

Parameter	Value
d_w	50
d_p	5
r_{hid}	60
d_r	60
Th_d	0.018
ε	2
l	4
LR	0.0005
DR	0.1
b_s	45

4.3. Comparative Experiments with Competitors

In this section, the MLNRNN was compared with seven state-of-the-art methods to demonstrate its effectiveness, including

- PCNN + ATT [18]: A method that integrates a piece-wise CNN with selective attention over instances.
- PARE [13]: A method based on BERT concatenates all sentences in a bag to explore more information.
- Intra-Inter Bag [19]: A method with a multi-level attention mechanism to refine features at sentence-level (ATT_RA) and bag-level (BAG_ATT), which employs a PCNN as its encoder.
- FAN [37]: A method utilizing adversarial training to align false negative instances into a unified feature space and generate pseudo labels for them, which employs a PCNN with a transformer layer as its encoder.
- HiCLRE [14]: A method leveraging interactive information between entity level, sentence level, and bag level with an MHSA to learn more intra-level information with adversarial contrastive learning, which employs BERT as its encoder.
- Multicast [24]: A method employing adversarial training at sentence-level and bag-level to improve data utilization, which employs a PCNN as its encoder.
- CIL [23]: A method proposing contrastive instance learning to construct a positive pair with TF-IDF, which employs BERT as its encoder.

Figures 4–6 show the PR curves of the MLNRNN and the competitors on the NYT-10, NYT-16, and Wiki-20m datasets. It can be observed that the MLNRNN achieved the highest precision on the majority of recalls on the three datasets. Since the NYT-16 dataset is significantly larger than the NYT-10 dataset, the PR curves for all models showed a greater decline on the NYT16 dataset. Although the Wiki-20m dataset was much larger than the previous two, PR curves of all models tended to decline less sharply, because it was manually annotated, with fewer incorrect labels. Consequently, all competitors, including our method, generally performed better on it.

We adopted the AUC in our experiments because it can show the difference in performance more clearly between the competitors and the MLNRNN. Tables 4–6 present a comprehensive comparison of the MLNRNN with the other competitors, detailing the Top-N P@N(s), Mean P@N, and AUC for NYT-10, NYT-16, and Wiki-20m, respectively. The highest values are indicated in bold. From Tables 4–6, we can see that the MLNRNN outperformed all competitors across all evaluation metrics. Among the competitors, PARE performed better than the others, while the MLNRNN demonstrated a superior performance, showing improvements of 3.2% in Mean P@N and 1.2% in AUC from Table 4, 2.3% in Mean P@N and 1.9% in AUC from Table 5, and 0.6% in Mean P@N and 1.1% in AUC from Table 6.

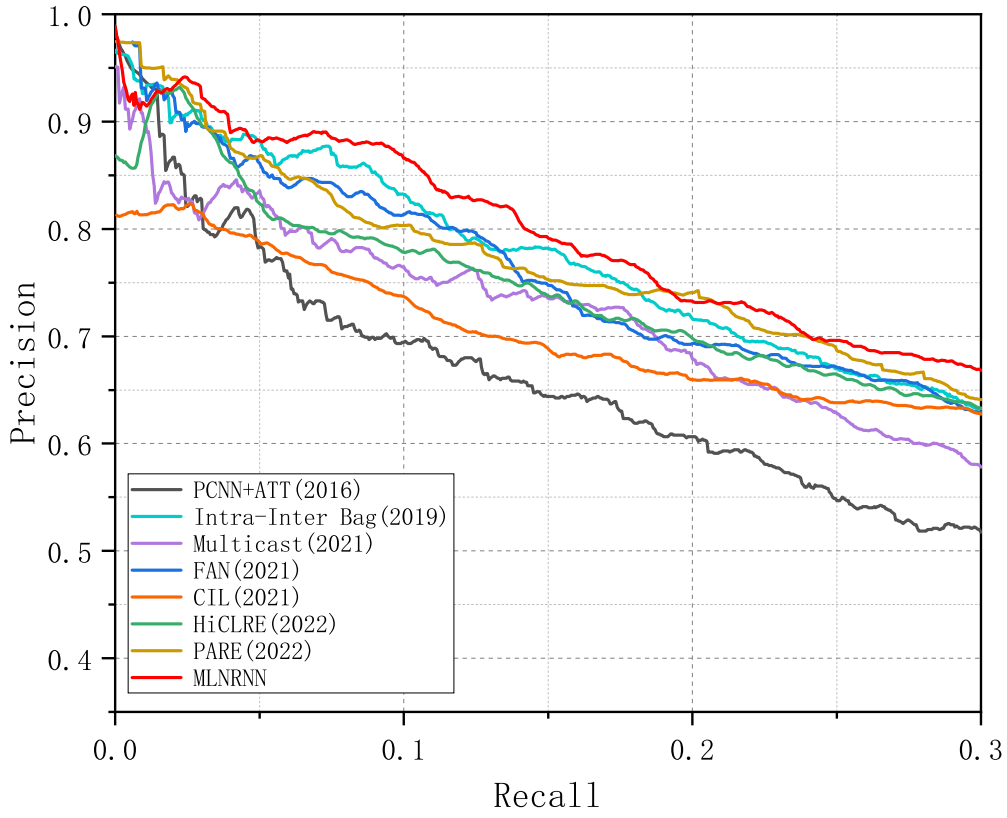

Figure 4. PR curves of the MLNRNN and the competitors for the dataset NYT-10.

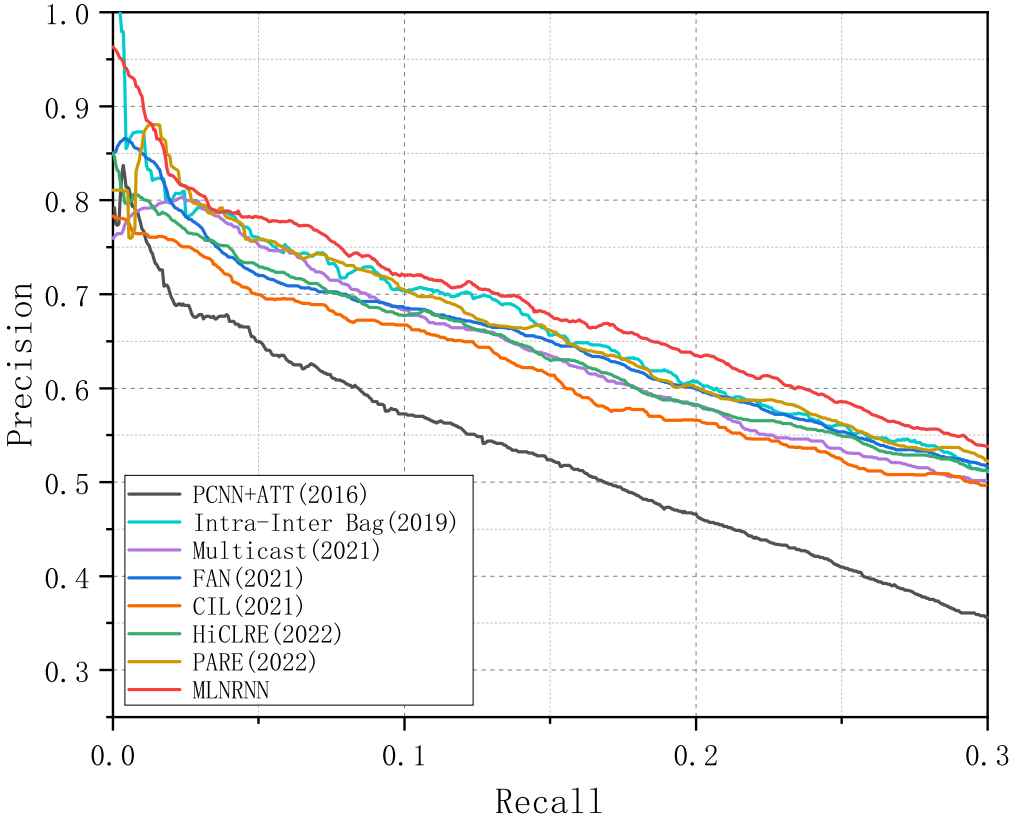

Figure 5. PR curves of the MLNRNN and the competitors for the dataset NYT-16.

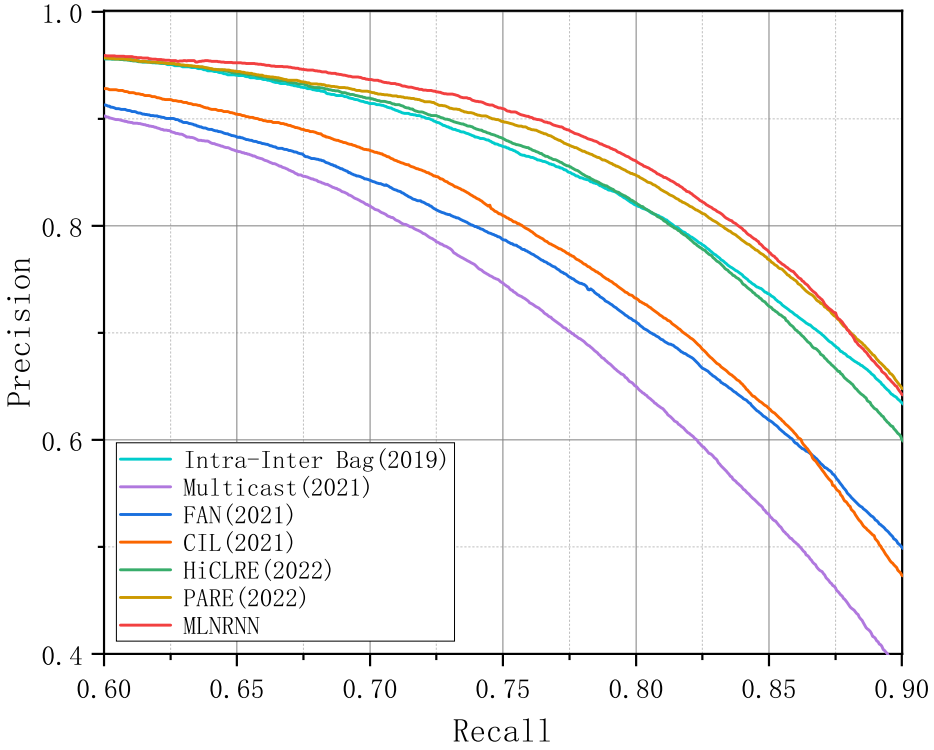

Figure 6. PR curves of the MLNRNN and the competitors for the dataset Wiki-20m.

Table 4. Comparisons of the MLNRNN with the competitors in terms of P@N(s) and AUC on NYT 10.

DSRE Methods	P@N(s) (%)				AUC
	Top 100	Top 200	Top 300	Mean	
PCNN + ATT	76.0	71.0	67.3	71.4	36.3
Multicast	83.7	79.2	74.2	79.0	40.2
CIL	81.5	75.5	72.1	76.4	42.1
FAN	85.8	83.4	79.9	83.0	44.8
Intra-Inter Bag	91.8	84.0	78.7	84.4	42.2
HiCLRE	82.0	78.5	74.0	78.2	45.3
PARE	90.0	84.3	82.3	85.5	47.5
MLNRNN(Ours)	**94.2**	**88.4**	**83.4**	**88.7**	**48.9**

The best P@N(s) or AUC values are marked in bold and this formatting is consistent across all tables below.

Table 5. Comparisons of the MLNRNN with the competitors in terms of P@N(s) and AUC on NYT 16.

DSRE Methods	P@N(s) (%)				AUC
	Top 100	Top 200	Top 300	Mean	
PCNN + ATT	68.7	64.3	61.1	64.7	26.9
Multicast	75.0	70.5	65.3	70.3	31.3
CIL	70.2	68.8	64.3	67.8	29.5
FAN	76.2	71.0	66.2	71.1	33.9
Intra-Inter Bag	76.0	72.0	67.3	71.8	33.1
HiCLRE	75.6	71.2	65.8	70.9	33.7
PARE	77.0	72.0	68.6	72.5	35.9
MLNRNN(Ours)	**79.2**	**74.0**	**71.2**	**74.8**	**37.8**

Table 6. Comparisons of the MLNRNN with competitors in terms of P@N(s) and AUC on Wiki-20m.

DSRE Methods	P@N(s) (%)				AUC
	Top 30,000	Top 40,000	Top 50,000	Mean	
Multicast	94.7	88.8	80.4	88.0	82.1
CIL	96.2	89.4	83.6	89.7	85.1
FAN	95.4	89.6	84.2	89.7	84.0
Intra-Inter Bag	97.0	92.8	86.7	92.2	88.7
HiCLRE	97.2	94.0	86.6	92.6	87.9
PARE	97.4	94.0	87.6	93.0	89.9
MLNRNN(Ours)	**97.9**	**94.4**	**88.5**	**93.6**	**91.0**

4.4. Ablation Study and Analysis

Since the two proposed modules are essential components, they could not be directly removed for ablation experiments. However, to fully demonstrate the effectiveness of each module, we first replaced each our module with several existing methods for comparison. Then, we selected the best-performing replacements for an overall comparison experiment.

To verify the effectiveness of the IKSA, we compared it with RNN/CNN/BERT relation extractors on the NYT-10 dataset for P@Ns, as NYT-10 is the most widely used dataset in DSRE. As noted in Table 7, IKSA+ATT outperformed the neural networks with the other structures. We can also see that the transformer-based model obtained better results than the CNN-based and RNN-based models. Accordingly, we compared MOMIL with other algorithms that deal with sentence representations or bag representations based on the same sentence encoder. From Table 7, we can also infer that MOMIL failed to achieve the best result for the Top 100, as this module relies on obtaining accurate sentence representations first and convolutional operations are unable to model long-distance dependencies.

Table 7. Comparisons between replaced methods on NYT-10 to select appropriate ablation methods.

Replaced Methods	P@N(s) (%)			
	Top 100	Top 200	Top 300	Mean
PCNN+ONE	75.0	70.0	64.7	69.9
PCNN+ATT	76.0	71.0	67.3	71.4
PCNN+ATT_RA+BAG_ATT	**91.8**	84.0	78.7	84.4
PCNN+MOMIL	90.8	**84.6**	**79.8**	**84.7**
Bi-LSTM+ATT	72.4	66.7	63.8	67.6
BERT+ATT	88.2	82.7	79.3	83.4
IKSA+ATT	**91.4**	**85.3**	**81.6**	**86.1**

Therefore, we compared the MLNRNN with three ablation methods. We could prune the CCL and replace the IKSA and MOMIL with their strongest competitors, respectively, that we selected from Table 7. Specifically, we replaced the IKSA with BERT to form a MLNRNN w/o IKSA, and replaced MOMIL with the ATT_RA+BAG_ATT to form a MLNRNN w/o MOMIL. Figure 7 depicts the precision curves of the MLNRNN and the three ablation methods on NYT-10, where the curve of the MLNRNN is better than the others, meaning that each module in the MLNRNN provided varying degrees of improvement to the model. We can observe that the MLNRNN obviously performed better than the three ablation methods across the different Top-N and mean P@N values from Table 8. Additionally, we can observe that the three ablation methods showed inconsistent performance across the different P@N values. For instance, the MLNRNN w/o IKSA performed the worst of the three ablation methods. Through this observation, we can infer that the IKSA plays a crucial role in the overall performance of the model, significantly contributing to its effectiveness, due to the fact that the IKSA can capture the semantic features that express relationships more comprehensively and accurately. Although BERT is also capable of effectively modeling the overall semantics of a sentence, it is inevitably

influenced by other features rather than relation features within the sentence. In contrast, the IKSA, through an interaction between denoised word information, only extracts the features that are more relevant to the entity pairs and the sentence-specific latent relation, thereby eliminating the influence of unrelated but frequently occurring words. From the perspective of computational cost, the IKSA achieved the optimal performance after six iterations, while even the BERT-based models consisted of twelve transformer encoder modules. The IKSA's parameter count was only 60% of that of the BERT-based models.

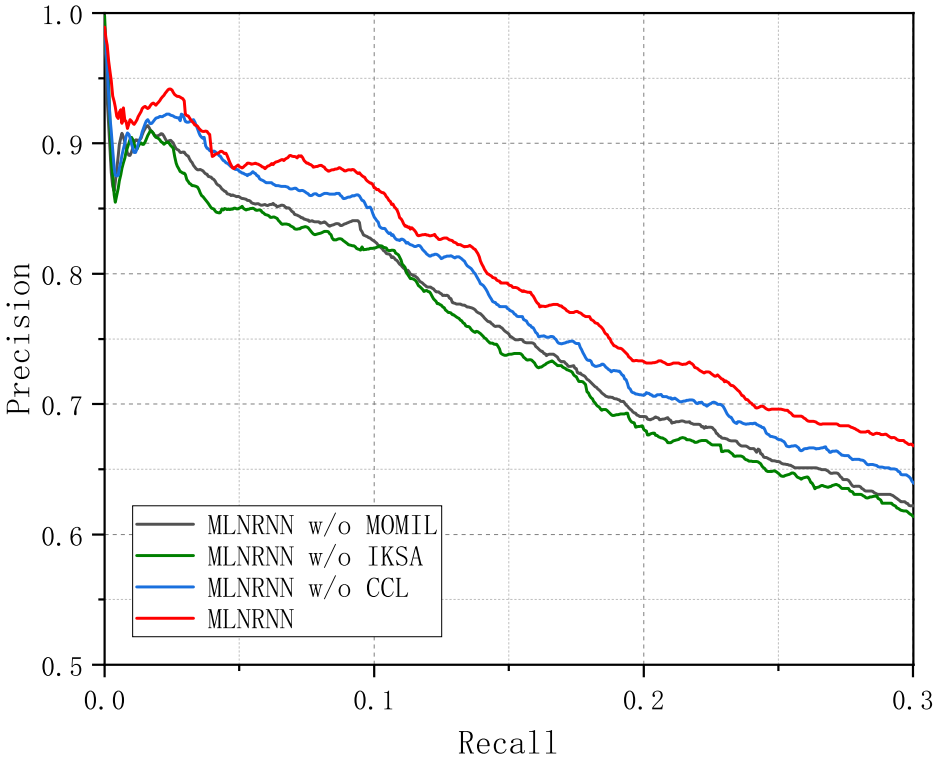

Figure 7. PR curves of the MLNRNN and three ablation methods on the NYT-10 dataset.

Table 8. P@N(s) of MLNRNN and three ablation methods on NYT-10.

DSRE Methods	P@N(s) (%)			
	Top 100	Top 200	Top 300	Mean
MLNRNN w/o MOMIL	91.4	85.3	81.6	86.1
MLNRNN w/o CCL	92.0	86.4	81.7	86.7
MLNRNN w/o IKSA	90.8	84.6	79.8	84.7
MLNRNN	**94.2**	**88.4**	**83.4**	**88.7**

Subsequently, the MLNRNN w/o MOMIL failed to outperform the MLNRNN w/o CCL, which indicates that MOMIL is effective even without CCL. This effectiveness can be attributed to the greedy algorithm of MOMIL that can exploit as many true instances as possible, with a proper threshold and semantic enhancement to form the accurate bag representation for classification. From another perspective, combining MOMIL with other encoders could also lead to improvements, showing that MOMIL can be well integrated with other encoders such as PCNN and Bi-LSTM. Moreover, MOMIL approaches the multiple-instance learning problem from a new perspective, by first distinguishing between positive and negative instances and then processing them separately, rather than treating all instances uniformly.

Last but not least, the MLNRNN w/o CCL performed the best among the three ablation methods but worse than the MLNRNN, showing that false instances indeed contain abundant useful information to improve the performance of models. With the

contrastive learning framework that we designed, sentences containing the same relation triple bring their semantic features closer together, whereas the differences between the semantic features of sentences with different relation triples are magnified in the semantic space. This approach ensures that the learned features are more discriminative, thus improving the ability to effectively distinguish between sentence features. Thus, the results of the MLNRNN w/o CCL decreased in P@N(s).

Since the threshold Th_d is a critical parameter of MOMIL, we also conducted experiments to verify our assumption. Figure 8 shows the performance of tow models with different thresholds. When the threshold is 0, this means that the model selects only the most likely instance as the true instance. We can see that the performance of both models declined when the threshold was either below or above the optimal value. This is because a too low threshold results in missing true instances, while a too high threshold causes false instances to be recognized as true ones, introducing noise, and thereby hurting the performance. However, a higher threshold is less detrimental to the model compared to a lower threshold. This is due to the effect of the semantic enhancement mechanism, which can filter out some noise, while filtering out dissimilar semantics.

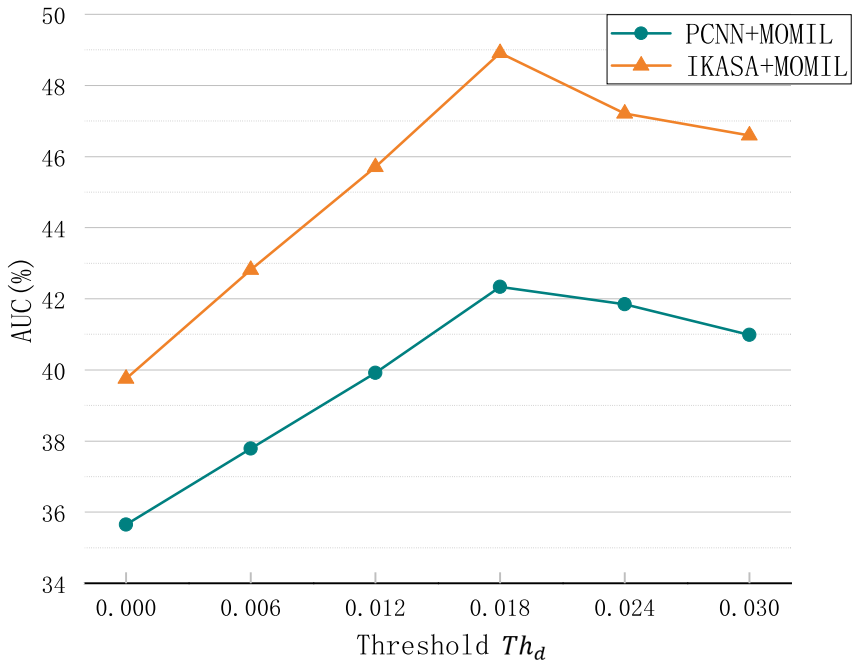

Figure 8. The performance with different thresholds.

4.5. Computational Complexity Analysis

To demonstrate the practical applicability of the proposed MLNRNN, we analyze its time and space complexity in this section. For the time cost, the primary computation of the MLNRNN is concentrated in the multi-head self-attention (MHSA) layer of the IKSA module. Therefore, we considered its time complexity to be $O(m^2 \times d_x^2)$, where m represents the sequence length, and d_x is the word embedding dimension. For comparison, we chose the strongest competitor, the PARE [13] model, which also focuses its computation on the multi-head self-attention layer and shares the same time complexity. To further compare, we analyzed the floating point operations (FLOPs) of both models to evaluate their real-time runtime performance. The FLOPs for PARE [13] were 348.21 G, while for our proposed MLNRNN, they were 220.41 G. This demonstrates that our method achieved a superior performance, while maintaining a lower time complexity. For space complexity, we analyzed the parameter numbers of the MLNRNN. Our method contains 69.73 M parameters, whereas the methods based on BERT, such as HiCLRE [14], typically have over 80 M parameters. Although convolutional neural-network-based methods have less

parameters, such as Intra-Inter Bag [19] with 78.25 K parameters and Multicast [24] with 81.75 K parameters, their ability to extract key features is limited. Despite having fewer parameters, these methods did not outperform MLNRNN on the various datasets.

4.6. Case Study

We present several sample cases from the NYT-10 dataset to show the realistic performance of the IKSA and MOMIL, respectively. We selected three typical sentences for the IKSA, and words with weights higher than the overall average are highlighted in bold in Table 9. We can see that the selected words contain salient relation features and other words are completely irrelevant. It is worth noting that these words are distributed discretely throughout the sentence and do not follow any fixed pattern. In addition, we provide three bags to exhibit the performance of MOMIL and compare it with three multi-instance learning algorithms, as shown in Table 10. The four algorithms were implemented on the basis of the same encoder. The first bag had two correct sentences, and the other two bags had only one true instance. MOMIL could effectively remove the noise and was not misled by sentences with ambiguous semantics. This was not only due to the accurate sentence representations provided by the IKSA, but also to an appropriate threshold Th_d.

Table 9. Cases study for keywords that IKSA selected.

Relations	Sentences
LC	In **[New York]**, a new **downtown center** is planned for **Brookhaven** on **[Long Island]**, and a village development designed by Robert, called Tuxedo Reserve, is planned for Tuxedo.
PN	... said the new foreign minister of **[Albania]** during an interview in his office, decorated with an elegant portrait of **[Faik Konica]**, who **became the first Albanian ambassador** to the United States in 1926.
CF	The most visible and one of the most outspoken is **[Vinod Khosla]**, **a founder of [Sun Microsystems]** and now a **partner** at Khosla Ventures.

Words in brackets are entities and the remaining bold text represents the keywords selected by IKSA. 'LC', 'PN', and 'CF' are relation labels in the dataset, which are 'location/contains', 'person/nationality', and 'company/founders', respectively.

Table 10. Case study for bags processed by MOMIL and other multi-instance learning algorithms.

Relations	Bags	Weights			
		ONE	ATT	ATT_RA	MOMIL
LC	***S*1:** The effort has proved increasingly contentious, involving everything from local financing for the arts to the future of democracy in [Hong Kong], which Britain returned to [China] in 1997.	1	0.105	0.127	0.376
	***S*2:** The Fountain Set Group, a fabric maker that has headquarters in [Hong Kong] and is [China]'s second-biggest textile exporter, was named in a study by Morgan Stanley.	0	0.889	0.597	0.624
	***S*3:** Chinese customs officials count the same goods as exports to [Hong Kong], a simpler approach that allows [China] to release trade statistics quickly.	0	0.006	0.276	0
NA	***S*1:** That is the sum of his old Congressional district, which was mostly in [Brooklyn], and the one drawn after the 2000 census, which added more [Queens] neighborhoods.	0	0.792	0.344	1
	***S*2:** He supervised Patrol, Administrative and Plainclothes duties in Manhattan, [Brooklyn] and [Queens].	1	0.173	0.437	0
	***S*3:** The suit has been joined by City Councilman Eric Gioia, who represents the [Queens] neighborhoods on the north side of Newtown Creek; and Marty Markowitz, the [Brooklyn] borough president.	0	0.035	0.219	0
CC	***S*1:** The months of tense negotiations between [Tokyo] and Washington to reopen the market of [Japan] received extensive coverage in their media.	0	0.013	0.16	0
	***S*2:** In practice, the subject matter of his prints was the urban life of Edo, the old name for [Tokyo] and de facto capital of [Japan] after 1603.	1	0.978	0.66	1
	***S*3:** In a measure of changing Japanese attitudes to Russia, Shintaro Ishihara, governor of [Tokyo], said in an interview on Thursday: [Japan], the U.S. and Russia should jointly work on the pipeline project.	0	0.009	0.18	0

'NA' represents 'non-relation' and 'CC' represents the relation of 'country/capital'.

5. Discussion

From the comparative experimental results in Section 4.3, we can observe that the BERT-based methods [13,14,23] generally outperformed the PCNN-based methods [18,19,37] (with the exception of Multicast [24], which primarily applies adversarial training (AT) and virtual adversarial training (VAT) to enhance model robustness, rather than designing methods specifically for classification accuracy). This demonstrates the effectiveness of self-attention mechanisms for modeling long sequences. Building on this foundation, our proposed MLNRNN introduced a keyword feature extraction approach tailored for relation extraction, achieving improved performance, while also reducing the number of parameters. In addition, the PCNN-based method Intra-Inter Bag [19] also achieved commendable results, indicating that its proposed multi-level attention mechanism is capable of refining representative relation features from the features extracted by the PCNN.

Moreover, we acknowledge the limitations of this article, such as not considering the impact of different entities on the relational sentence structures. For example, sentences expressing the same relation may have significantly different representations depending on the entity pair involved. In addition, we did not account for interactions between sentences or the interactions between sentence-level and bag-level information. Consequently, much work remains to be done in neural relation extraction. (1) In relation extraction, greater emphasis could be placed on the types of entity pairs, rather than the entities themselves. For instance, when expressing the relationship of "place of birth", the phrase most used is "born in". By focusing on keywords and eliminating representation differences caused by different entity pairs, more comprehensive representations could be obtained. (2) Interactions between sentences can provide additional information for relation classification, as the relationship between entity pairs is not always explicitly expressed in the sentences where they are mentioned.

6. Conclusions

In this article, we proposed a novel neural network, MLNRNN, which focuses on both word-level and sentence-level noise, challenges that previous works have not adequately addressed. The first module, the IKSA, was designed to capture the salient discriminative features of sentences, while removing noisy words to address word-level noise. Moreover, we introduced a different multi-instance learning algorithm, MOMIL, to alleviate sentence-level noise, while leveraging information from false instances through CCL. Based on such a framework, the MLNRNN can obtain more accurate sentence representations by innovatively using latent relation vectors for guidance, and it takes into account both the global dependencies and pairwise dependencies of words. To prove the effectiveness of the MLNRNN, we performed extensive experiments on three DSRE benchmark datasets: NYT-10, NYT-16, and Wiki-20m. Specifically, the MLNRNN achieved significant improvements in AUC metrics compared to the state-of-the-art (SOTA) methods, with a 1.4% enhancement on the NYT-10 dataset, a 1.9% improvement on the NYT-16 dataset, and a 1.1% boost on the Wiki-20m dataset. We anticipate that this insight may be useful for advancing future research and applications, such as knowledge graph construction.

Author Contributions: Conceptualization, Z.Y.; methodology, Z.Y.; software, Z.Y.; resources, W.S.; writing—original draft preparation, Z.Y.; writing—review and editing, W.S.; visualization, Z.Y.; supervision, W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

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

References

1. Ma, X.; Zhu, Q.; Zhou, Y.; Li, X. Improving question generation with sentence-level semantic matching and answer position inferring. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8464–8471. [CrossRef]
2. Yao, L.; Mao, C.; Luo, Y. KG-BERT: BERT for knowledge graph completion. *arXiv* **2019**, arXiv:1909.03193.
3. Lin, X.; Chen, L. Canonicalization of open knowledge bases with side information from the source text. In Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE), Macao, China, 8–11 April 2019; pp. 950–961.
4. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. Relation classification via convolutional deep neural network. In Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Ireland, 23–29 August 2014; pp. 2335–2344.
5. 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, Singapore, 2–7 August 2009; pp. 1003–1011.
6. Liu, T.; Zhang, X.; Zhou, W.; Jia, W. Neural relation extraction via inner-sentence noise reduction and transfer learning. *arXiv* **2018**, arXiv:1808.06738.
7. Kádár, Á.; Xiao, L.; Kemertas, M.; Fancellu, F.; Jepson, A.; Fazly, A. Dependency parsing with structure preserving embeddings. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Online, 19–23 April 2021; pp. 1684–1697.
8. Jang, B.; Kim, M.; Harerimana, G.; Kang, S.U.; Kim, J.W. Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. *Appl. Sci.* **2020**, *10*, 5841. [CrossRef]
9. Li, Y.; Long, G.; Shen, T.; Zhou, T.; Yao, L.; Huo, H.; Jiang, J. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8269–8276.
10. Zhang, Y.; Chen, Y.; Yu, S.; Gu, X.; Song, M.; Peng, Y.; Chen, J.; Liu, Q. Bi-GRU relation extraction model based on keywords attention. *Data Intell.* **2022**, *4*, 552–572. [CrossRef]
11. Qu, J.; Hua, W.; Ouyang, D.; Zhou, X. A noise-aware method with type constraint pattern for neural relation extraction. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 1134–1148. [CrossRef]
12. Yuan, Y.; Liu, L.; Tang, S.; Zhang, Z.; Zhuang, Y.; Pu, S.; Wu, F.; Ren, X. Cross-relation cross-bag attention for distantly-supervised relation extraction. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 419–426.
13. Rathore, V.; Badola, K.; Singla, P. PARE: A simple and strong baseline for monolingual and multilingual distantly supervised relation extraction. *arXiv* **2021**, arXiv:2110.07415.
14. Li, D.; Zhang, T.; Hu, N.; Wang, C.; He, X. HiCLRE: A hierarchical contrastive learning framework for distantly supervised relation extraction. *arXiv* **2022**, arXiv:2202.13352.
15. Papaluca, A.; Krefl, D.; Suominen, H.; Lenskiy, A. Pretrained knowledge base embeddings for improved sentential relation extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Dublin, Ireland, 22–27 May 2022; pp. 373–382.
16. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
17. 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.
18. 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, Berlin, Germany, 7–12 August 2016; Volume 1, pp. 2124–2133.
19. Ye, Z.X.; Ling, Z.H. Distant supervision relation extraction with intra-bag and inter-bag attentions. *arXiv* **2019**, arXiv:1904.00143.
20. Song, W.; Gu, W. Hierarchical Knowledge Transfer Network for Distantly Supervised Relation Extraction. In Proceedings of the IEEE 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, QLD, Australia, 18–23 June 2023; pp. 01–09.
21. Lin, X.; Liu, T.; Jia, W.; Gong, Z. Distantly supervised relation extraction using multi-layer revision network and confidence-based multi-instance learning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 165–174.
22. Chen, J.W.; Fu, T.J.; Lee, C.K.; Ma, W.Y. H-FND: Hierarchical false-negative denoising for distant supervision relation extraction. *arXiv* **2020**, arXiv:2012.03536.
23. Chen, T.; Shi, H.; Tang, S.; Chen, Z.; Wu, F.; Zhuang, Y. CIL: Contrastive instance learning framework for distantly supervised relation extraction. *arXiv* **2021**, arXiv:2106.10855.
24. Chen, T.; Shi, H.; Liu, L.; Tang, S.; Shao, J.; Chen, Z.; Zhuang, Y. Empower distantly supervised relation extraction with collaborative adversarial training. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35, pp. 12675–12682.
25. Ma, R.; Gui, T.; Li, L.; Zhang, Q.; Zhou, Y.; Huang, X. SENT: Sentence-level distant relation extraction via negative training. *arXiv* **2021**, arXiv:2106.11566.

26. Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Zhang, C. Tensorized self-attention: Efficiently modeling pairwise and global dependencies together. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 3–5 June 2019; Volume 1.
27. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624. [CrossRef]
28. Alshemali, B.; Kalita, J. Improving the reliability of deep neural networks in NLP: A review. *Knowl.-Based Syst.* **2020**, *191*, 105210. [CrossRef]
29. Santos, C.N.d.; Xiang, B.; Zhou, B. Classifying relations by ranking with convolutional neural networks. *arXiv* **2015**, arXiv:1504.06580.
30. Luo, X.; Zhou, W.; Wang, W.; Zhu, Y.; Deng, J. Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data. *IEEE Access* **2017**, *6*, 5705–5715. [CrossRef]
31. Geng, Z.; Chen, G.; Han, Y.; Lu, G.; Li, F. Semantic relation extraction using sequential and tree-structured LSTM with attention. *Inf. Sci.* **2020**, *509*, 183–192. [CrossRef]
32. Miwa, M.; Bansal, M. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv* **2016**, arXiv:1601.00770.
33. Sun, H.; Grishman, R. Lexicalized Dependency Paths Based Supervised Learning for Relation Extraction. *Comput. Syst. Sci. Eng.* **2022**, *43*, 861–870. [CrossRef]
34. Hu, L.; Zhang, L.; Shi, C.; Nie, L.; Guan, W.; Yang, C. Improving distantly-supervised relation extraction with joint label embedding. 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. 3821–3829.
35. Liu, T.; Wang, K.; Chang, B.; Sui, Z. A soft-label method for noise-tolerant distantly supervised relation extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 1790–1795.
36. Shang, Y.; Huang, H.Y.; Mao, X.L.; Sun, X.; Wei, W. Are noisy sentences useless for distant supervised relation extraction? In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8799–8806.
37. Hao, K.; Yu, B.; Hu, W. Knowing false negatives: An adversarial training method for distantly supervised relation extraction. *arXiv* **2021**, arXiv:2109.02099.
38. Kwon, H.; Kim, Y.; Yoon, H.; Choi, D. Fooling a neural network in military environments: Random untargeted adversarial example. In Proceedings of the 2018 IEEE Military Communications Conference (MILCOM 2018), Los Angeles, CA, USA, 29–31 October 2018; pp. 456–461.
39. Kwon, H. Adversarial image perturbations with distortions weighted by color on deep neural networks. *Multimed. Tools Appl.* **2023**, *82*, 13779–13795. [CrossRef]
40. 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.
41. 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, 5–8 December 2013; Volume 26.
42. Zhou, Y.; Pan, L.; Bai, C.; Luo, S.; Wu, Z. Self-selective attention using correlation between instances for distant supervision relation extraction. *Neural Netw.* **2021**, *142*, 213–220. [CrossRef] [PubMed]
43. Miao, D.; Zhang, J.; Xie, W.; Song, J.; Li, X.; Jia, L.; Guo, N. Simple contrastive representation adversarial learning for NLP tasks. *arXiv* **2021**, arXiv:2111.13301.
44. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference (ECML PKDD 2010), Barcelona, Spain, 20–24 September 2010; Proceedings, Part III 21; Springer: Cham, Switzerland, 2010; pp. 148–163.