AI

MDPI

Article

Sentence Interaction and Bag Feature Enhancement for Distant Supervised Relation Extraction

Wei Song and Qingchun Liu *

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: 6223112019@stu.jiangnan.edu.cn

Abstract: *Background:* Distant supervision employs external knowledge bases to automatically match with text, allowing for the automatic annotation of sentences. Although this method effectively tackles the challenge of manual labeling, it inevitably introduces noisy labels. Traditional approaches typically employ sentence-level attention mechanisms, assigning lower weights to noisy sentences to mitigate their impact. But this approach overlooks the critical importance of information flow between sentences. Additionally, previous approaches treated an entire bag as a single classification unit, giving equal importance to all features within the bag. However, they failed to recognize that different dimensions of features have varying levels of significance. *Method:* To overcome these challenges, this study introduces a novel network that incorporates sentence interaction and a bag-level feature enhancement (ESI-EBF) mechanism. We concatenate sentences within a bag into a continuous context, allowing information to flow freely between them during encoding. At the bag level, we partition the features into multiple groups based on dimensions, assigning an importance coefficient to each sub-feature within a group. This enhances critical features while diminishing the influence of less important ones. In the end, the enhanced features are utilized to construct high-quality bag representations, facilitating more accurate classification by the classification module. *Result:* The experimental findings from the New York Times (NYT) and Wiki-20m datasets confirm the efficacy of our suggested encoding approach and feature improvement module. Our method also outperforms state-of-the-art techniques on these datasets, achieving superior relation extraction accuracy.

Keywords: distant supervision; relation extraction; sentence interaction; feature enhancement

check for updates

Received: 18 December 2024
Revised: 18 February 2025
Accepted: 25 February 2025
Published: 4 March 2025

Citation: Song, W.; Liu, Q. Sentence Interaction and Bag Feature Enhancement for Distant Supervised Relation Extraction. *AI* **2025**, *6*, 51. https://doi.org/10.3390/ai6030051

Copyright: © 2025 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 fundamental task in natural language processing (NLP) that focuses on recognizing the relationships between pairs of entities within a sentence. The goal of RE is to derive relational triples formatted as <subject, relation, object> from the text. For instance, this is demonstrated in Table 1. Given a sentence, s_1, containing a pair of entities, <Jonathan Lethem, Brooklyn>, RE aims to automatically identify and extract relationships such as <Jonathan Lethem, place_of_birth, Brooklyn>. In the past decade, it has emerged as a focal point for research [1], and this task is both critical for the construction of knowledge graphs and fundamental for various NLP applications [2], including language inference [3], knowledge graph [4], and question answering systems [5]. By predicting entity relationships, this process enhances the enrichment of the knowledge base and accelerates the development of NLP technologies in a wide range of practical applications [6].

Early methods in the distant supervision for relation extraction (DSRE) field mainly depended on manually created features [7–9], including part-of-speech (POS) tags, called entity recognition (NER) tags, and dependency paths. These methods employed various strategies to transform feature extraction cues, including sequences, parse trees, and syntactic information, into high-dimensional feature vectors for DSRE tasks [10]. Although these methods achieved significant success, they rely heavily on fully supervised paradigms and demand large-scale manually annotated training corpora to achieve optimal performance. The main problem is that manual annotation is both time-consuming and expensive, making such methods less feasible for real-world applications. To resolve this limitation, Mintz et al. [7] presented the distant supervision method, a strategy that substantially enhances corpus annotation in NLP tasks through automated generation of training datasets. This is achieved by aligning entities from knowledge graphs with corresponding mentions in the text. As proof, in Table 1, sentences s_1, s_2, and s_3 all contain the entity pair <Jonathan Lethem, Brooklyn>, and are automatically labeled as expressing the relationship place_of_birth through this alignment process.

Table 1. Alignment between knowledge bases and plain text.

Relational Triple	Sentence	Weights
<Jonathan Lethem, place_of_birth, Brooklyn>	s_1: In **Brooklyn**, they ask when you're going on Charlie Rose and if you know **Jonathan Lethem**.	0.613
	s_2: You could say that only the dead, and **Jonathan Lethem**, know **Brooklyn**.	0.512
	s_3: It's unlikely that the . . . **Jonathan Lethem**'s novels -RRB- about old **Brooklyn**.	0.166

Traditional feature-based methods for identifying semantic relationships between entity pairs depend on NLP tools to obtain lexical and syntactic features [8]. Nevertheless, the limitations of human knowledge often result in errors that propagate through the extraction process [11]. To tackle this challenge, Zeng et al. [12] developed a deep convolutional neural network capable of automatically extracting features from datasets labeled at the sentence level. Although this method is effective, it faces challenges in scaling to large knowledge bases because of the lack of manually annotated data. To help mitigate the lack of labeled data, distant supervision was implemented to connect text with a knowledge base, enabling the automatic generation of large-scale training datasets [9,13]. This method introduces significant noise, as it fails to consider the contextual nuances essential for sentence-level relation extraction. This noise leads to suboptimal performance without robust noise-handling techniques. Recently, researchers have focused on enhancing the robustness of RE models to label noise. A widely used approach to reduce the effects of label noise involves integrating multi-instance learning (MIL) with attention mechanisms [14–16]. Researchers have utilized Multiple Instance Learning (MIL) to categorize sentences that refer to the same pair of entities into 'bags', based on the premise that at least one sentence in each bag accurately conveys the intended relationship. The relationship is subsequently applied to the whole collection, which enhances the method's resilience to noise. By selecting sentences from these bags, models can focus on higher-quality sentences while minimizing the impact of noisy ones. Techniques such as piecewise CNN (PCNN), proposed by Zeng et al. [17], incorporate MIL to extract perceptual features from sentences and classify them based on these features. But choosing only one sentence from a bag, as MIL does, risks losing valuable information found in other sentences within the same bag. Recent advancements [14,18–21] have substituted the rigid selection method with sentence-level

selective attention (ATT), which allocates distinct weights to each sentence in a collection, enabling the model to utilize the complete spectrum of available information. By way of example, Lin et al. [14] assigned weights to each instance, enabling the model to utilize the rich information from all sentences while avoiding information loss. These advancements notwithstanding, sentence-level attention mechanisms continue to face certain limitations, highlighting opportunities for further refinement.

Firstly, the aforementioned methods encode each sentence independently [20,22–24], neglecting the relationships between sentences within the same bag. This oversight can lead to inaccurate weight assignments during bag representation calculation, as the relationships between different sentences are not considered. For example, in Table 1, using the method proposed by Lin et al. [14], which only considers each sentence in isolation to compute the weights for sentences s_1, s_2, and s_3 in the bag, one might assign a low weight to s_3, implying that it does not express the relation place_of_birth. Conversely, if the correlations between s_3 and s_1 (as well as s_2) were taken into account, the result might differ, showing that s_3 indeed expresses the relation — which is the true label in this case. This demonstrates that evaluating s_3 based solely on the target relation is insufficient, as s_3 could still be a correctly labeled sentence.

Secondly, the aforementioned methods often treat features in different dimensions with equal importance, disregarding the fact that feature relevance can vary across dimensions. By ignoring this variability, they fail to fully consider how the quality of features impacts model performance. Given the noise introduced by distant supervision and the mislabeling it causes, designing an effective DSRE neural network remains a significant challenge. Therefore, it is crucial to improve feature quality to better represent entity relationships.

To solve the challenges of the above challenges, we propose a network named ESI-EBF, designed to enhance sentence interaction and improve bag feature representation. The model consists of four parts: the context construction components, the feature extraction components, the group-wise enhancement components, and the classification components. To address the first challenge, we introduced the context construction components as the initial component of our design. This module transforms the sentences within a bag into a coherent context, facilitating richer interaction between sentences. The group-wise enhancement components address the second challenge by dividing the bag features into multiple groups along the feature dimension, assigning importance coefficients to the sub-features in each group to enhance significant features and diminish less important ones, thereby improving the overall bag feature representation. Based on our understanding, this is the initial research that differentiates the significance of features across various dimensions, successfully reducing the performance decline associated with low-quality features. Our experimental findings indicate that the suggested model provides enhanced bag representation and greatly enhances RE performance. The key contributions of this study are summarized in the following points:

- For the part that enhances sentence interaction (ESI), we propose a context construction module that emphasizes the associations between sentences by constructing them into a unified context for joint encoding. Then, we conduct more than just simple sentence processing; we consider that the meaning of the same entity may vary across different sentences in the feature extraction module.
- To enhance bag feature (EBF) quality, we introduce a group-wise enhancement module that amplifies important sub-features within each group while attenuating less significant ones.
- Extensive testing on two standard DSRE datasets, NYT-10 and Wiki-20m, demonstrates that our model surpasses the state-of-the-art methods.

The remainder of this article is organized in the following manner: Section 2 discusses previous research. Section 3 provides a detailed description of the model we are suggesting. Section 4 showcases the outcomes of our experiments. Lastly, Section 5 concludes the paper.

2. Related Work

2.1. Conventional Feature-Based DSRE Methods

RE is a fundamental field in numerous natural language processing (NLP) applications. Early DSRE methods largely relied on supervised learning and kernel-based techniques [21,25], which demanded a considerable amount of handcrafted features for training. For example, Culotta et al. [26] leveraged syntactic parse trees to extract features from sentences and entity pairs, while Mooney et al. [21] improved this by segmenting sentences into three parts, focusing on the position of entities, and capturing more detailed features to enhance model performance. These strategies can definitely increase the efficiency of relation extraction; however, they have two main drawbacks. First, manually designed features often lack robustness due to the limitations of human knowledge, leading to error propagation. Second, identifying relevant features for relation extraction remains challenging, particularly when working with high-dimensional feature vectors. Conventional methods often struggle to directly select meaningful features from such representations, leading to inefficiencies. In contrast, our method can automatically model features without relying on manual design while enhancing the representational capacity of the features.

2.2. Deep Feature-Based DSRE Methods

Traditional supervised learning methods require large, accurately annotated training corpora, but manual annotation is both time-intensive and costly, making these methods impractical for real-world applications. To tackle the issue, Mintz et al. [7] introduced distant supervision for RE, where raw text is automatically aligned with a knowledge base to generate relational labels for entity pairs. It is unavoidable that the annotated data contain some inaccurate sentences. To alleviate the impact of mislabeled data, researchers [8,9,16] utilized a flexible distant supervision approach grounded in the multi-instance learning (MIL) framework. These techniques organize sentences that share the same entity pairs into groups. They operate on the premise that at least one sentence in each group is correctly labeled, which helps to minimize errors.

As deep learning has advanced, deep neural networks have proven to be very effective in natural language processing (NLP) tasks because of their excellent performance and ability to generalize. As a result, various deep learning models integrated with Multiple Instance Learning (MIL) have been created to tackle challenges in DSRE. For instance, Zeng et al. [17] introduced PCNN to generate sentence representations and determine the most dependable sentence to represent the entire bag. Miwa and Bansal [27] improved neural networks by integrating both sequential and structural information from dependency trees. Zhou et al. [28] utilized attention mechanisms in bidirectional long short-term memory networks (BLSTMs) to emphasize the key semantic components in sentences. Vashishth et al. [22] utilized external knowledge and used graph convolutional networks to capture syntactic information, improving the effectiveness of RE. Wen et al. [29] created a gated PCNN that integrates entity pairs with the context of sentences to improve word encoding, while Ye and Luo [30] proposed a multi-label learning framework that employs a CNN. Shi et al. [31] alleviated limitations in dependency tree-based RE methods by better pruning mislabeled data, resulting in improved model accuracy. Meng et al. [32] introduced a noise-aware approach for detecting and dynamically correcting noisy labels associated with entity information.

However, much of the previous work handles each sentence independently without facilitating information sharing between sentences, thereby overlooking relational information among them. In contrast, we propose a context construction module in our proposed model that allows for comprehensive interaction between sentences to enhance the interaction between them, resulting in richer sentence embeddings and enhanced relation extraction performance.

2.3. Attention and Gating Mechanism

In recent years, the attention mechanism has become increasingly popular due to its ability to effectively reduce noise in DSRE. Neural relation extraction with selective attention (NRESA) was the first to implement selective attention across instances to decrease noise in bag representations. Lin et al. [14] utilized PCNNs as sentence encoders and introduced intra-bag attention, which calculates the bag representation by taking a weighted sum of all sentence representations within the bag. The Hierarchical Selective Attention Network (HSAN) [33] further improved this approach by using a two-tier attention mechanism: sentence-level attention to identify relevant instances and word-level attention to concentrate on key terms essential for relation extraction.

In parallel, gating mechanisms have proven effective in controlling information flow. By utilizing the nonlinear mapping of gate units, they allow for differentiated pass rates based on the importance of the information. For instance, SeG integrated gating mechanisms with pooling functions to regulate bag-level representations [34]. This fact notwithstanding, existing attention mechanisms mainly emphasize distinguishing instance importance—amplifying correctly labeled instances while downplaying noisy ones—and they generally treat all feature dimensions equally, ignoring the varying importance of features across different dimensions.

Our proposed method tackles this limitation by dividing the feature map of bag instances into multiple groups and assigning weights to features within each group. This approach enhances important features and weakens less significant ones by generating importance coefficients for each sub-feature. Moreover, we incorporate the head and tail entity information from all instances in a bag to further improve the model's ability to predict entity relationships.

3. The Proposed Method

In this section, we introduce a relation extraction model that integrates sentence interaction with bag-level feature enhancement. This approach adheres to the multi-instance learning framework, where a set of sentences that share the same entity pair forms a bag. The training dataset is made up of N bags, each associated with a relation r, represented by a randomly initialized vector of size m. For each bag, which consists of n sentences related to the same entity pair, the goal is to develop effective bag representations B from the input sentences. The trained model can then predict relations for unlabeled bags containing entity pairs. Figure 1 depicts the structure of our model, which includes four main components:

- The context construction module concatenates the sentences within a bag, allowing sentences containing the same entity pair to form a coherent context, facilitating richer interaction between sentences.
- The feature extraction module aims to generate a bag representation. It sends the concatenated context to a BERT encoder for encoding, then splits it based on markers to obtain the embedding for each sentence. Additionally, logsumexp pooling is applied to aggregate all entity mentions to generate entity representations. Finally,

the sentence embeddings and entity representations are concatenated to form the bag representation.

- The group-wise enhancement module is designed to refine the quality of bag features. It partitions the bag feature into multiple semantic groups and assigns a significance coefficient to each sub-feature within these groups. This process aims to amplify the more important features while diminishing the influence of less significant ones.
- The Relation Classification Module uses an MLP and a Sigmoid activation function to process the bag representation and determine the probability of it belonging to relation i.

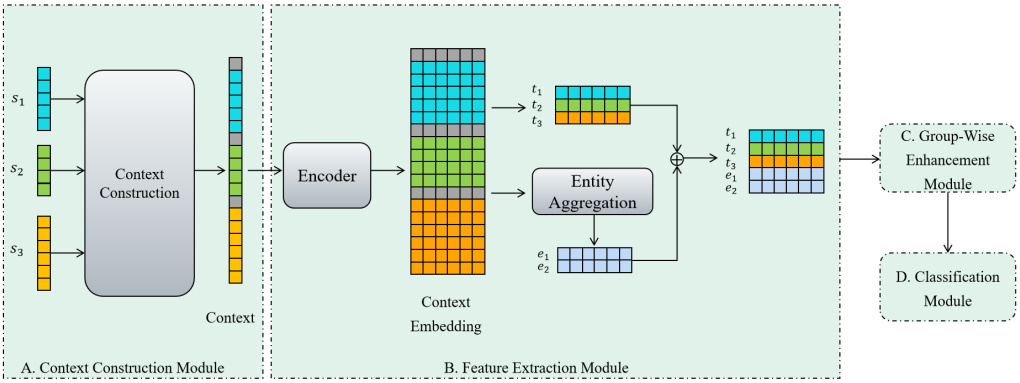


Figure 1. The general structure of the proposed method.

3.1. Context Construction Module

To fully leverage the data within a bag, we concatenate the sentences to form a cohesive context, allowing for rich information exchange between the sentences within the bag. Sentences are ordered sequentially, beginning with a [CLS] tag, separated by [SEP] tokens, and padded with [PAD] tokens for consistency. Additionally, following best practices for relation extraction [35,36], entity mentions in each sentence are indicated by specific tokens: $< e_1 >$ and $< /e_1 >$ for the first entity, and $< e_2 >$ and $< /e_2 >$ for the second entity. Inspired by [37] , this process continues until (a) including an additional sentence would surpass the encoder's maximum token capacity, or (b) every sentence in the collection has already been accounted for. An example of this construction is provided in Table 2.

Table 2. The context construction process.

Context: $[CLS]$ Many Christians have fled $< e_2 >$ Iraq $< /e_2 >$ since the fall of $< e_1 >$ Saddam Hussein $< /e_1 >$. $[SEP]$ A1 One of the most coveted jobs in $< e_2 >$ Iraq $< /e_2 >$ does not yet exist: the executioner for $< e_1 >$ Saddam Hussein $< /e_1 >$. $[SEP]$ ..... $[SEP]$ Toppling $< e_1 >$ Saddam Hussein $< /e_1 >$ did not automatically create a new and better $< e_2 >$ Iraq $< /e_2 >$. $[PAD][PAD]$
Subject: Saddam Hussein **Object:** Iraq **Relation:** nationality

3.2. Feature Extraction Module

The sentence encoder fed the constructed context into a BERT encoder to obtain contextual embeddings, x_i^j, for each token, w_i^j, which can be indicated as

$$C = [x_1^1, x_2^1, \ldots, x_j^i, \ldots, x_{d_n}^n] = Bert([w_1^1, w_2^1, \ldots, w_j^i, \ldots, w_{d_n}^n]) \quad (1)$$

where d_n represents the length of the n-th sentence. By encoding the concatenated sentences with BERT, we transform sentences that were originally processed independently into a cohesive context, allowing for comprehensive information exchange among them. To obtain the encoding for each individual sentence, we split the contextual embeddings based on the marker $[SEP]$ positions. The embedding for the i-th sentence after splitting is denoted as $\{x_1^i, x_2^i, \ldots, x_{d_i}^i\}$, reflecting the information exchange with other sentences in the bag. A weighted average is subsequently used on these embeddings to derive the sentence representation t_i, denoted as:

$$t_i = \frac{1}{d_i} \sum_{j=1}^{d_i} x_j^i \tag{2}$$

Once all the sentences are averaged, the resulting sentence embeddings are combined to create a comprehensive joint sentence representation, T. This process is depicted in Figure 2.

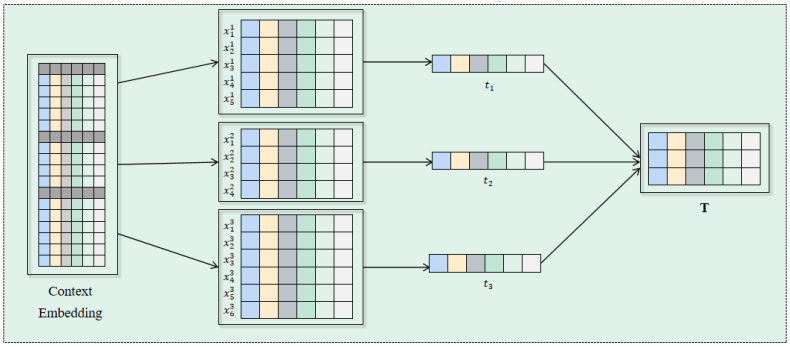


Figure 2. The process of the rich joint sentence representation T. Different colors represent different dimensional features.

Entity enhancement is utilized to gather worldwide entity information within a specific context. We represent the head entity in each sentence using the "$< e_1 >$" token positioned before the entity. To capture its overall contextual information, we aggregate the head entity representations from all sentences using logsumexp pooling [38], as shown below:

$$e_1 = log \sum_{j=1}^{n} exp(x_{e_1}^j) \tag{3}$$

A similar approach is applied to the tail entity e_2; we represent it using the "$< e_2 >$" token preceding the tail entity in each sentence:

$$e_2 = log \sum_{j=1}^{n} exp(x_{e_2}^j) \tag{4}$$

This pooling accumulates information from references in context. Experimental results show that, compared with average pooling, this method has better performance.

Joint mapping combines the sentence and entity embeddings to produce a richly informative bag-level embedding, $B = [T; E] = [t_1, t_2, \ldots, t_n, e_1, e_2] \in R^{L \times m}$, where n is the number of sentence in the context, L represents n+2, and m denotes the dimensionality of the word embeddings.

3.3. Group-Wise Enhancement Module

This module aims to enhance important features and suppress less relevant ones at the bag level, improving the quality of bag features to facilitate classification. The overall process is shown in Figure 3. We first divide the joint bag mapping, B, obtained from the previous module into p groups along the feature dimension. Then, we can achieve the bag

group feature mapping, $G = [g_1, g_2, \ldots, g_p] \in R^{L \times m}$, where p represents the number of groups. Let us examine how the process works for a given group in detail.

For a given group, its corresponding feature mapping is denoted as $g = [f_1, f_2, \ldots, f_L] \in R^{L \times k}$, where $k = \frac{m}{p}$, and the sub-feature vector $f_i = [a_1, a_2, \ldots, a_k] \in R^k$, where a_i is a feature in sub-feature f. We presume that each group contains a representative semantic feature during the network learning process. We obtain a representation that approximates the global semantic feature of the group learning representation $\bar{g} \in R^k$ by averaging the features within the group:

$$\bar{g} = \frac{1}{L} \sum_{i=1}^{L} f_i \tag{5}$$

Next, using this semantic vector, we use a simple dot product operation to generate an importance coefficient c for each sub-feature, representing the resemblance between the semantic feature $\bar{g}$ and each sub-feature vector f_i. Specifically, the formula is defined as follows:

$$c_i = \bar{g} \cdot f_i \tag{6}$$

To avoid bias in the coefficients across different samples, we apply normalization to c_i within the feature space:

$$\widehat{c_i} = \frac{c_i - \mu_c}{\sigma_c + \epsilon}, \mu_c = \frac{1}{k} \sum_{j}^{k} c_j, {\sigma_c}^2 = \frac{1}{k} \sum_{j}^{k} (c_j - \mu_c)^2 \tag{7}$$

Here, ϵ (for example, 1×10^{-5}) is a constant included to ensure numerical stability, μ_c is the mean of all correlation coefficient data in the current group, and σ_c is its standard deviation. To ensure that the normalization within the network accurately reflects the identity transformation, we incorporate two parameters, γ and β, for each coefficient $\widehat{c_i}$. These parameters adjust the scaling and shifting of the normalized value. As a result, a new importance coefficient $\bar{c_i}$ is produced through the sigmoid function $\sigma(\bullet)$:

$$\bar{c_i} = \sigma(\gamma \widehat{c_i} + \beta) \tag{8}$$

Finally, the generated normalized importance coefficients are used to scale, enhancing the important features while weakening the less significant ones, resulting in an enhanced sub-feature vector $\bar{f_i}$:

$$\bar{f_i} = \bar{c_i} \cdot f_i \tag{9}$$

In this way, after all the features have been enhanced, a more optimal feature group, $\widehat{g} = [\bar{f_1}, \bar{f_2}, \ldots, \bar{f_L}] \in R^{L \times k}$, is formed. Once the features in all groups have been enhanced, we can obtain the group-wise feature mapping $\widehat{G} = [\widehat{g_1}, \widehat{g_2}, \ldots, \widehat{g_p}] \in R^{L \times m}$. Thus, we obtain an improved mapping of the features of the joint bag $\bar{B} = [\bar{t_1}, \bar{t_2}, \ldots, \bar{t_n}, \bar{e_1}, \bar{e_2}] \in R^{L \times m}$.

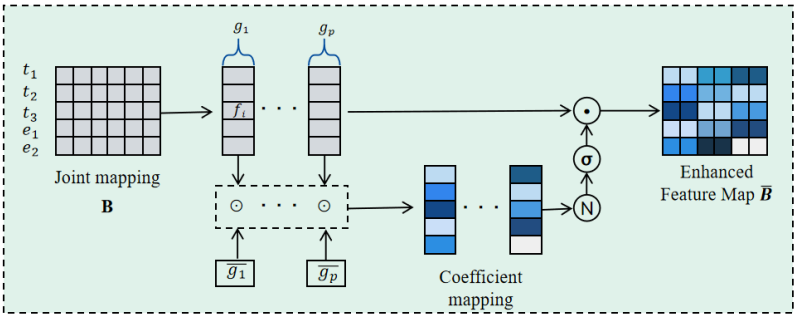


Figure 3. The overall process of the group-wise enhancement module. Different colors represent different levels of attention.

3.4. Relation Classification Module

In this final module, the enhanced bag representation, $\bar{B}$, obtained from the group-wise enhancement module is used to perform relation classification. The purpose of this module is to translate the enhanced features into a probability distribution regarding the potential relationships, enabling us to forecast the most probable relationship for each entity pair within the bag.

First, we assign a weight to each row of the bag $\bar{B}$ (where each row represents the embedding of a sentence), thus minimizing the effect of irrelevant sentences in the collection [39], as shown below:

$$b_k = \sum_{i=1}^{L} \alpha_{ki} t_i \tag{10}$$

Here, k is an index for the relation, taking values from 1 to h, and α_{ki} represents the attention weight that connects the k-th relation to the i-th sentence in the bag B. For the uniform representation, where t_{n+1} is the head entity e_1 and t_{n+2} is the tail entity e_2, α_{ki} can be further defined as

$$\alpha_{ki} = \frac{exp(e_{ki})}{\sum_{j=1}^{L} exp(e_{kj})} \tag{11}$$

where e_{ki} represents the degree of correspondence between the i-th sentence and the k-th relation in the bag, which is specifically computed using the dot product of the vectors to assess the extent of their correspondence as follows:

$$e_{ki} = r_k t_i^T \tag{12}$$

where r_k represents the k-th row of the relation embedding matrix R^2.

Next, the score o_k for classifying bag $\bar{B}$ into relation k is determined using b_k and the relation embedding r_k as follows:

$$o_k = r_k b_k^T + d_k \tag{13}$$

where d_k is a bias. Next, we apply a softmax activation function to transform the score vector into probability values:

$$P(k|\bar{B};\theta) = softmax(o_k) = \frac{exp(o_k)}{\sum_{j=1}^{h} exp(o_j)} \tag{14}$$

where $P(k|\bar{B};\theta)$ indicates the likelihood that the bag is associated with a particular relation k. This allows the model to predict the likelihood of the bag being associated with each relation in the label space.

To enhance the model's performance, we use the negative log-likelihood as the objective function, which is defined in the following way:

$$J(\theta) = -\sum_{q=1}^{N} \log p(r_q|\bar{B}_r^i;\theta) \tag{15}$$

where N refers to the total number of bags in a batch, r_q represents the label for a bag b_q, and θ represents the set of parameters for the model. During the model training, we utilize mini-batch stochastic gradient descent (SGD) to reduce the loss function $J(\theta)$.

4. Experiment

In this section, we present several experiments to demonstrate the effectiveness of our suggested method. We begin by detailing the datasets and the experimental framework. Next, we explore various versions of the method (which can serve as ablation experiments) and compare the ESI-EBF against seven competing DSRE methods. We also conduct an ablation study to evaluate the impact of each component. Furthermore, we examine how the number of feature groups affects performance. Lastly, we present a case study to demonstrate the effectiveness of our proposed method.

4.1. Datasets

We assessed the model by using the NYT [8] and Wiki-20m [37] datasets in our experiments. The details are described in Table 3.

The **NYT dataset** was initially introduced by Riedel et al. [8]. The New York Times employs data from 2005 to 2006 for training purposes and uses data from 2007 for testing. In NYT, sentences are organized into bags, and the labels are determined based on whether they contain specific entity pairs, which are automatically created through distant supervision. NYT is the most popular dataset in DSRE, and is widely used for evaluating relation extraction models. Both the training and testing splits are supervised from a distance, ensuring consistency in the dataset's automatic labeling process. Additionally, the NYT dataset exemplifies the primary purpose of DSRE, which is to reduce the cost of labeling. Therefore, given its widespread adoption and reliability, we chose this dataset for experimental validation. The dataset includes over 52 relation types, with "1" indicating N/A, and is divided into training and testing sets. The dataset details are as follows: (1) The training set consists of 522,611 sentences, 281,270 entity pairs, and 18,252 triples. (2) The test set includes 172,448 sentences, 96,678 entity pairs, and 1950 triples.

The **Wiki-20m dataset** was constructed by integrating Wikipedia articles [40], leveraging them as the corpus to ensure high-quality annotations and fewer noisy labels compared to automatically labeled datasets like NYT. Wiki-20m is a newly launched dataset created for training DSRE models and assessing their performance using a manually annotated test set. In this dataset, sentences (instances) are grouped into bags based on entity pairs, with labels derived from the relationships between the entities. The primary purpose of Wiki-20m is to provide a benchmark for evaluating relation extraction models under a more realistic and less noisy setting. The dataset comprises 30 different types of relationships, excluding "N/A". The particulars of the dataset are outlined as follows: (1) The training dataset includes 6,987,222 sentences, 304,870 entity pairs, and 157,740 triples. (2) The testing dataset includes 137,986 sentences, 74,390 pairs of entities, and 56,000 triples.

Table 3. Distant supervision datasets (NYT and Wiki-20m).

Dataset	NYT			Wiki-20m		
	Sentences	**Entity Pair**	**Triples**	**Sentences**	**Entity Pair**	**Triples**
Train	22,611	281,270	18,252	6,987,222	30,487	157,740
Test	172,448	96,678	1950	137,986	74,390	56,000

4.2. Experimental Setup

Evaluation metrics. In line with earlier techniques, each DSRE method was assessed using held-out testing, where precision and recall are determined by comparing the predictions against the relational facts in the dataset. For our experiments, we utilized the Precision–Recall (PR) curve for visual analysis, the Area Under the ROC Curve (AUC) for numerical evaluation, and Precision@N (P@N) to assess the accuracy of the top-N

outcomes [21,39,41]. For Precision@N, we employed the P@N metrics for the top 100, 200, and 300 results for these datasets. In comparison, Wiki-20m is a dataset annotated by humans, so it provides better performance for DSRE models because of its excellence labels. To facilitate a more precise evaluation, we utilized P@N metrics for the top 30,000, 40,000, and 50,000 results from the Wiki-20m dataset.

Parameter Settings. In our experiment, the word vector dimension was 300. The specific parameter configurations are presented in Table 4. To optimize hyperparameters, we conducted a grid search using learning rates of 1×10^{-5} and 2×10^{-5}, batch sizes of 8, 16, and 32, and feature group sizes of 10, 20, and 30, ultimately choosing the configuration that yielded the best performance for each dataset. For the baseline methods used for comparison, we adopted the parameter configurations that achieved the best performance according to the original studies.

Comparative Experiments. We carried out two types of comparative experiments. The first involved comparing three different versions of our proposed method (refer to Section 4.3), which served as an ablation study. The second set compared our proposed method with state-of-the-art approaches using two different datasets (see Section 4.4). Alongside these comparative experiments, we also conducted ablation studies (see Section 4.5). Lastly, we analyzed the impact of varying sizes of different feature groups.

Table 4. Parameter settings for our proposed method.

Parameters	Value
word embedding	300
batch size	16
learning rate	2×10^{-5}
max_length	512
dropout	0.5
feature groups	20

4.3. Different Versions of the Proposed Method

Because our suggested framework can be applied in various manners, we will present three different versions of the proposed method. (1) ESI + EBF: This is the original approach we came up with, which constructs the sentences in a bag into a coherent context for encoding, enhances the information interaction between the sentences (ESI), and the features of the bag are grouped to enhance the features (EBF) separately. (2) PCNN + EBF: The method of feature extraction is replaced by a piecewise CNN encoder (Zeng et al. [12]). (3) ESI + ATT: The variations in bag-level features are not taken into account, with ATT being the attention technique introduced by Lin et al. [14]. Furthermore, we included PCNN + ATT as part of the baseline models. The evaluation of these four approaches on the NYT dataset acted as an ablation study.

The experimental findings are presented in Figure 4 and Table 5. Our observations are as follows: (1) Regardless of the sentence encoders employed, our EBF outperforms ATT. This indicates the importance of taking bag feature quality into account. EBF evaluates feature importance from the perspective of sentence feature dimensions, ensuring that an unimportant sentence does not lead to the suppression of all its features. Even though some sentences may be noisy, certain feature dimensions within those noisy sentences could still play a crucial role. (2) In both sentence encoder modes, the models that utilize EBF perform better than those that do not include EBF. This further illustrates the importance of features in the different dimensions. (3) From an encoding standpoint, models that incorporate ESI perform better than those that do not. This indicates that it is important to take into account the relationships between sentences within the collection. However, a model with only the

ESI module performs worse than only the EBF module, highlighting the importance of considering feature significance from the perspective of feature dimensions. (4) ESI + EBF reaches the highest AUC of 0.440, which is an improvement of 5.8% over PCNN + ATT.

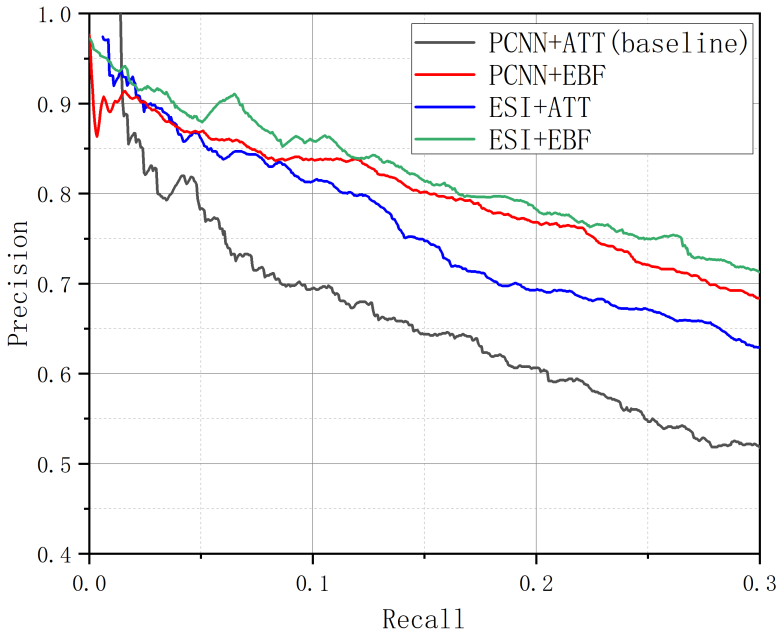


Figure 4. Precision–Recall curves for the different versions of the suggested model using the NYT dataset.

Table 5. The AUC for our three variants and one baseline on the NYT dataset.

Model	AUC
PCNN + ATT (baseline)	0.382
PCNN + EBF	0.415
ESI + ATT	0.403
ESI + EBF	0.440

4.4. Comparison with Previous Work

We compared our advanced ESI-EBF model with eight state-of-the-art models on the NYT dataset and five top models on the Wiki-20m dataset. A brief overview of the models being compared is provided below:

- **PCNN + ATT** [34] PCNN + ATT uses a selective attention mechanism to reduce the influence of noisy sentences, serving as a strong baseline for RE models.
- **PCNN+RL** [19] introduces a reinforcement learning (RL) framework designed to handle false positives by optimizing instance selection.
- **Intra–Inter Bag** [39] utilizes intra-bag methods to reduce noise within individual sentences, while inter-bag attention is applied to address noise at the bag level.
- **C2SA** [21] enhances relation extraction by applying a contextualized attention mechanism that considers all potential relations, assigning higher importance to higher-quality entity pairs.
- **SeG** [34] uses an Self-Attention Enhanced Selective Gate to overcome problems occurring in selective attention, which is caused by one-sentence bags.
- **CIL** [24] is a strategy for relation extraction that enhances the model's ability to perceive subtle differences during training through a contrastive learning framework, thereby improving overall relation extraction performance.

- **HiCLRE** [41] employs a multi-layer structure to extract features and model associations at both the instance level and the concept level, thus enhancing the accuracy and robustness of relation extraction.
- **PARE** [37] provides a simple and powerful baseline for both monolingual and multilingual DSRE.
- **FAN** [42] employs adversarial training to consolidate false negative cases into a common feature space and apply pseudo labels, using a PCNN combined with a transformer layer as its encoder.

Performance Measured by the PR Curve. The PR curve comparison for eight models on the NYT dataset is presented in Figure 5. Since the authors of PCNN + ATT and PCNN + RL only plotted the first 2000 points of their PR curves, we also limited our plot to the first 2000 points to ensure a fair comparison. The following conclusions can be obtained: (1) Our proposed ESI-EBF model demonstrates superior performance compared to the other models, validating the effectiveness of incorporating sentence-level interactions and accounting for feature-level distinctions. (2) Our model demonstrates more stable performance. Our approach demonstrates minor fluctuations at lower recall rates, but maintains overall strong performance. When the recall rate exceeds 0.09, our model consistently performs the best.

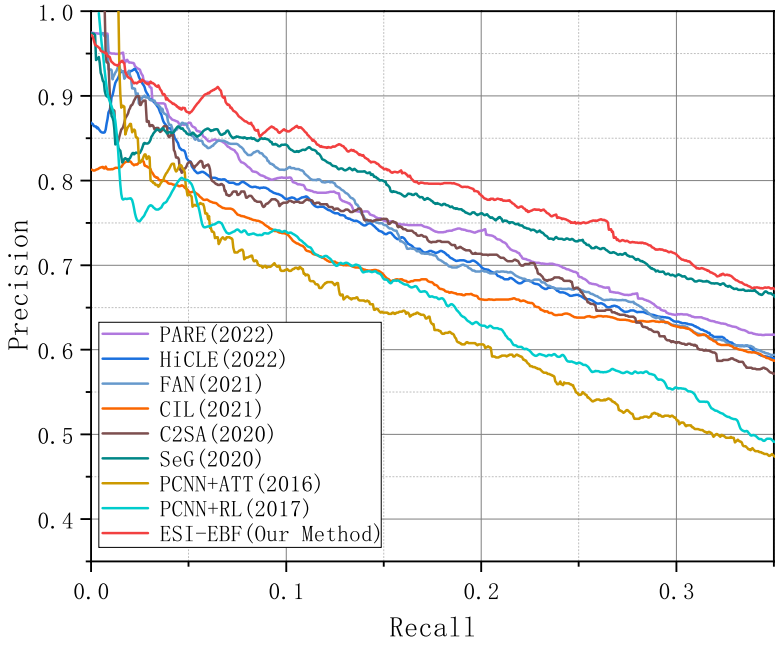


Figure 5. Precision–Recall curves for eight models evaluated on the NYT dataset.

The outcomes of evaluating five models on the Wiki-20m dataset are presented in Figure 6. The following is evident: (1) Our model outperformed all other models in the comparison. (2) With manually annotated labels, Wiki-20m contains far fewer mislabeled instances. Despite its significantly larger size compared to NYT, all models demonstrate strong performance on this dataset, with our method achieving superior results.

Performance Measured by the AUC. In our experiments, AUC was chosen as the evaluation metric because it provides a clearer distinction in performance between our proposed ESI-EBF model and the competing methods. Tables 6 and 7 offer a detailed comparison of ESI-EBF with other approaches, presenting the Top-N P@N(s), Mean P@N, and AUC results for the NYT and Wiki-20m datasets, respectively. Similarly, our proposed ESI-EBF model consistently outperforms all other models across both datasets, demonstrat-

ing improvements of 1.1% in average P@N and 0.9% in AUC according to Table 6, and 1.0% in average P@N and 1.4% in AUC as seen in Table 7.

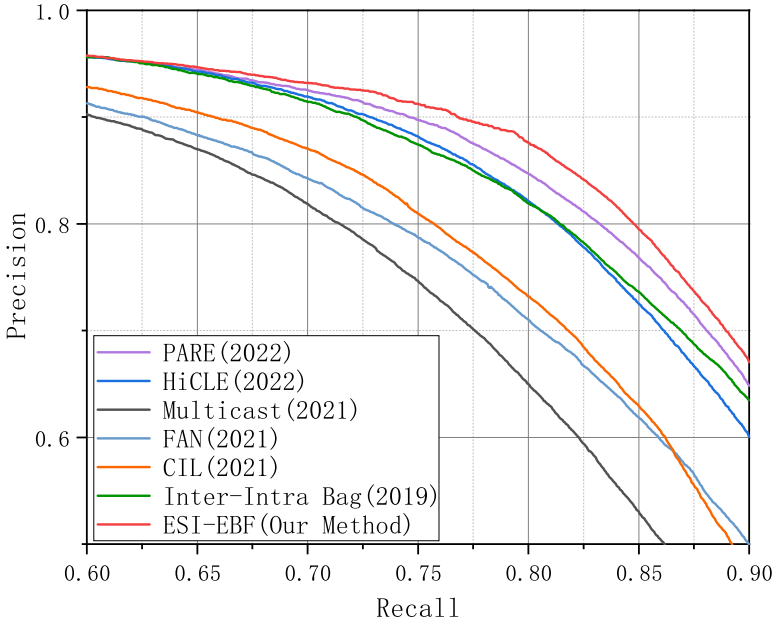


Figure 6. Comparison of PR curves for six models using the Wiki-20m dataset.

Table 6. Comparison of the AUC for eight models using the NYT dataset.

DSRE Methods	P@N (%)				AUC
	100	200	300	Mean	
PCNN + ATT	76.0	71.0	67.3	71.4	36.3
PCNN + RL	81.8	79.4	73.9	78.4	40.8
CIL	82.0	74.3	72.3	76.2	41.7
Intra–Inter Bag	91.2	84.3	78.6	84.7	42.3
HICLRE	82.4	76.9	74.4	77.9	44.8
C2SA	83.1	79.5	74.1	78.9	45.1
FAN	85.7	83.4	76.5	81.9	45.5
SeG	92.8	85.0	82.7	86.8	47.2
ESI-EBF(Ours)	**93.5**	**87.2**	**83.1**	**87.9**	**48.1**

Bold font indicates the best performance data, and the following table is consistent.

Table 7. Comparison of the AUC for five models using the Wiki-20m dataset.

DSRE Methods	P@N (%)				AUC
	30,000	40,000	50,000	Mean	
Multicast	94.3	89.1	82.3	88.6	82.3
FAN	95.8	89.7	83.9	89.8	84.8
CIL	96.0	90.2	84.3	90.2	85.2
HICLRE	97.4	94.3	84.4	92.0	87.8
Intra–Inter Bag	97.6	92.3	87.6	92.5	89.3
ESI-EBF(Ours)	**97.8**	**94.5**	**88.1**	**93.5**	**90.7**

4.5. Examination and Evaluation of Ablation Studies

As these two modules are essential parts, they cannot be eliminated for ablation studies. Instead, we initially substituted them with various existing methods for comparison purposes. Afterward, we selected the best-performing alternatives for in-depth comparative experiments.

To determine the performance of ESI, we compared it with LSTM, CNN, and BERT-based relation extractors using the NYT dataset, as NYT is the most commonly utilized benchmark in DSRE. As shown in Table 8, ESI + ATT consistently outperforms other neural network architectures. Furthermore, we compared EBF with alternative approaches that utilize the same sentence encoder for sentence- or bag-level representations. The results in Table 8 indicate that EBF also achieves superior performance over other models.

Table 8. Comparisons between replaced methods on NYT to select appropriate ablation methods.

Replaced Methods	P@N (%)			
	100	**200**	**300**	**Mean**
PCNN + ONE	75.1	70.6	63.9	69.9
PCNN + ATT	76.0	71.0	67.3	71.4
PCNN + ATT_RA + BAG_ATT	91.2	84.3	78.6	84.7
PCNN + EBF	**91.8**	**84.9**	**79.0**	**85.2**
Bi-LSTM + ATT	71.6	67.3	64.1	67.7
BERT + ATT	87.8	83.7	76.9	82.8
ESI + ATT	**91.1**	**85.0**	**80.7**	**85.6**

To further validate the effectiveness of ESI-EBF, we compared it with two ablation variants. In particular, we swapped out ESI for BERT to create ESI-EBF without ESI, and we replaced EBF with ATT_RA + BAG_ATT to develop ESI-EBF without EBF. As shown in Figure 7, the accuracy curve of ESI-EBF surpasses those of the ablation models, indicating that each module contributes to performance improvement.

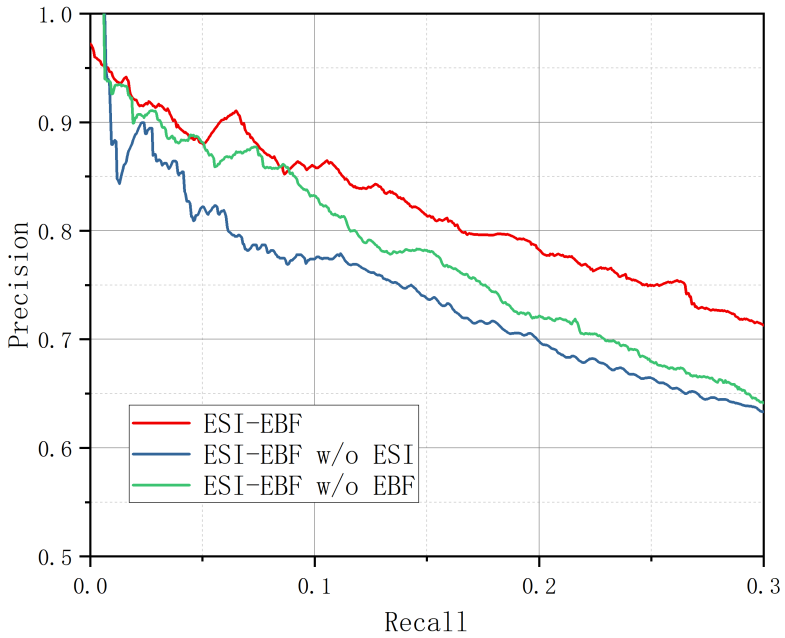


Figure 7. PR curves for ESI-EBF and two ablation methods using the NYT dataset.

Table 9 shows that ESI-EBF consistently surpasses both ablation variants in various Top-N configurations and average P@N scores. In contrast, the ablation models show variable performance across P@N values. Specifically, ESI-EBF without EBF has the lowest performance, indicating that EBF is essential for improving the overall effectiveness of the model. This is because EBF captures feature correlations across different dimensions while considering feature quality, preventing all dimensions from being treated equally.

Table 9. P@N(s) of ESI-EBF along with two ablation methods on NYT.

Replaced Methods	P@N (%)			
	100	**200**	**300**	**Mean**
ESI-EBF w/o EBF	91.1	85.0	80.7	85.6
ESI-EBF w/o ESI	91.8	84.9	79.0	85.2
ESI-EBF	**97.8**	**94.5**	**88.1**	**93.5**

Although ATT_RA + BAG_ATT also evaluates sentence-level features, it weakens sentence representations as a whole rather than adjusting feature importance at different dimensions. In contrast, EBF selectively suppresses irrelevant features while emphasizing essential ones. Even if a sentence is less informative overall, certain feature dimensions within it might still be valuable. By preserving these important feature dimensions, EBF avoids the loss of critical information caused by disregarding entire sentences.

4.6. The Influence of the Bag Feature Group-Wise Size

By applying group-wise enhancement to the features at the bag level, the capabilities of the bag are enhanced, allowing the model to better capture essential patterns in relation extraction. However, determining the optimal number of groups is crucial for achieving the best performance.

In this section, we examine the impact of the number of feature groups on the model's performance. Figure 8 illustrates that the model performs better with an increasing number of groups. This improvement occurs because having more groups allows for more detailed feature differentiation, enabling the model to concentrate on distinct aspects of the data. The performance peaks when the number of groups reaches 20, suggesting that this setting provides the best balance between feature refinement and noise reduction. Beyond this point, further increasing the number of groups leads to a decline in performance, likely due to excessive fragmentation of features, which introduces noise and reduces the model's ability to generalize effectively. In summary, if there are too few group sizes, the model's ability to improve is restricted because it cannot take full advantage of the interactions between features. Conversely, setting the group size too large disrupts meaningful patterns and adds unnecessary complexity.

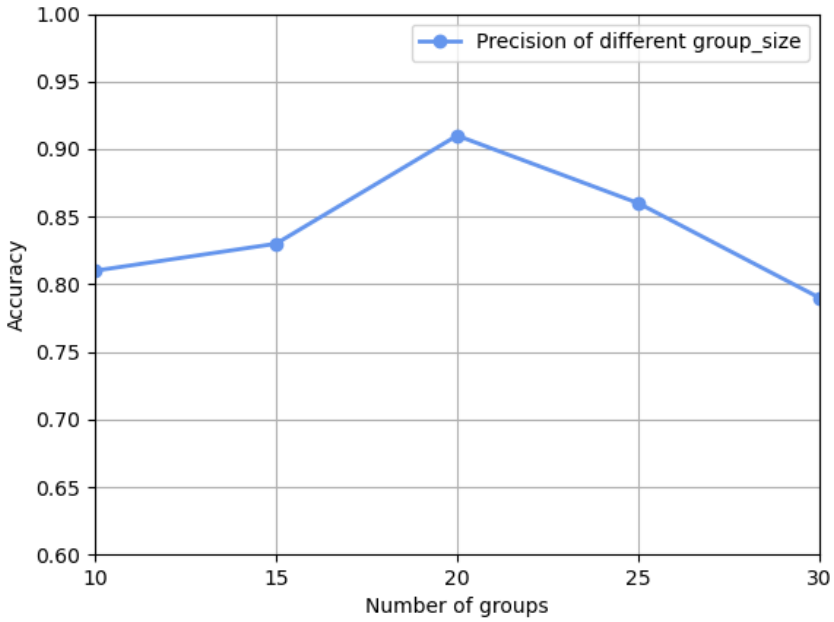


Figure 8. Precision of different group-wise sizes.

Therefore, in practical applications, selecting an appropriate number of groups is essential to maximizing performance while minimizing noise, ensuring a more robust and effective relation extraction model.

4.7. A Case Study

To further demonstrate the performance of our proposed method, we conducted a case study comparing the ESI-EBF model with PCNN + ATT [14] and Intra–Inter Bag [39]. Table 8 presents several sample cases selected from the NYT-10 dataset, consisting of two bags. Each bag is accompanied by its respective label and a collection of sentences that are included within the bag. Additionally, "True?" indicates whether the entity pair in the corresponding sentence aligns with the label of the bag. A "No" indicates that the sentence's label does not align with the bag's label, implying that the sentence can be viewed as irrelevant or extraneous within the bag. Conversely, a "Yes" signifies alignment. Furthermore, the table includes experimental results for the three models, featuring sentence weights and predictions (i.e., the anticipated label of the bag). The designation "NA" signifies a specific label indicating no relation.

For the input bag B_1, both ESI-EBF and Intra–Inter Bag correctly predict the relation. On the other hand, PCNN + ATT mistakenly assigns an "NA" label to the bag. For the input bag B_2, only ESI-EBF makes the correct prediction. As shown in Table 10, for all noise sentences that are consistent with the bag's label, PCNN + ATT assigns significantly higher weights to these sentences compared to the other two methods, while our proposed method assigns much lower weights. This demonstrates that fully leveraging the information exchange between sentences proves beneficial for the task. Therefore, it is evident that our model enhances the performance of relation extraction.

Table 10. An examination of a collection of bags that includes the following two bags.

Bags	Label	Sentences	True?	PCNN + ATT		Intra–Inter Bag		ESI-EBF	
				Weight	Pre	Weight	Pre	Weight	Pre
B_1	contains	*S1:* "Instead, a funeral took place at St. Francis de Sales Roman Catholic Church in **Belle Harbor**, **Queens**, which is the parish where he was born."	Y	0.621	**NA**	0.835	**contains**	0.863	**contains**
		S2: "One was for St. Francis de Sales Roman Catholic Church in **Belle Harbor**; ...of Christ in Rosedale, **Queens**."	No	0.618		0.326		0.257	
B_2	nationality	*S1:* "Under Thorne 's autocratic rule, Everycountry becomes a pastiche of Communist China, North Korea, **Cambodia** under **Pol Pot** and Iran since the Islamic Revolution."	Y	0.574	**NA**	0.827	**NA**	0.895	**nationality**
		S2: "...in France, Russia, China, **Cambodia** and ... that brought **Pol Pot** to power , kept him there briefly and then brought him down."	No	0.621		0.476		0.137	

"Pre" represents the predicted result.

5. Conclusions

This paper proposes a distantly supervised network based on sentence interaction enhancement and bag feature enhancement, which facilitates communication between sentences to mitigate the noise problem. Simultaneously, it strengthens useful bag features while suppressing less relevant ones, thereby improving the overall quality of bag representations.

For sentence representation, we first concatenated all sentences within a bag to construct a shared context and encoded them as a whole. Then, we split the encoded representation based on specific markers to obtain individual sentence embeddings. Additionally, we applied logsumexp pooling to aggregate all entity mentions, generating more robust entity representations. The sentence embeddings and entity representations were combined

to create the final bag representation. At the bag level, we enhanced the feature representations by organizing bag features across various dimensions. This grouping mechanism selectively enhances critical features while attenuating those of lesser importance, allowing the model to focus on the most informative aspects of the data.

To assess the performance of ESI-EBF, we carried out extensive experiments using two popular DSRE benchmark datasets: NYT and Wiki-20m. The results indicate that ESI-EBF surpasses current leading methods, with a 0.9% enhancement in AUC for the NYT dataset and a 1.4% rise for the Wiki-20m dataset. These findings highlight the potential of ESI-EBF to contribute to future advancements in knowledge graph construction, information extraction, and various natural language processing applications.

Author Contributions: Conceptualization, Q.L.; methodology, Q.L.; software, Q.L.; resources, W.S.; writing—original draft preparation, Q.L.; writing—review and editing, W.S.; visualization, Q.L.; 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. Li, W.; Qian, T.; Zhong, M.; Chen, X. Interactive Lexical and Semantic Graphs for Semisupervised Relation Extraction. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 7158–7169. [CrossRef] [PubMed]
2. Yao, L.; Mao, C.; Luo, Y. KG-BERT: BERT for knowledge graph completion. *arXiv* **2019**, arXiv:1909.03193. [CrossRef]
3. Luo, F.; Nagesh, A.; Sharp, R.; Surdeanu, M. Semi-Supervised Teacher-Student Architecture for Relation Extraction. In Proceedings of the SPNLP@NAACL-HLT, 2019, Minneapolis, MN, USA, 2–7 June 2019.
4. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Yu, P.S. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 494–514. [CrossRef] [PubMed]
5. Ma, X.; Zhu, Q.; Zhou, Y.; Li, X. Improving question generation with sentence-level semantic matching and answer position inferring. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8464–8471.
6. Xiao, Y.; Jin, Y.; Hao, K. Adaptive prototypical networks with label words and joint representation learning for few-shot relation classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 1406–1417. [CrossRef]
7. 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.
8. 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; Springer: Berlin/Heidelberg, Germany, 2010; pp. 148–163.
9. 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, Portland, OR, USA, 19–24 June 2011; pp. 541–550.
10. Zelenko, D.; Aone, C.; Richardella, A.R. Kernel Methods for Relation Extraction. *J. Mach. Learn. Res.* **2002**, *3*, 1083–1106.
11. Huang, W.; Mao, Y.; Yang, L.; Yang, Z.; Long, J. Local-to-global GCN with knowledge-aware representation for distantly supervised relation extraction. *Knowl.-Based Syst.* **2021**, *234*, 107565. [CrossRef]
12. 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.
13. Santos, C.N.d.; Xiang, B.; Zhou, B. Classifying relations by ranking with convolutional neural networks. *arXiv* **2015**, arXiv:1504.06580.

14. 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; pp. 2124–2133.
15. Liang, T.; Liu, Y.; Liu, X.; Zhang, H.; Sharma, G.; Guo, M. Distantly-Supervised Long-Tailed Relation Extraction Using Constraint Graphs. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 6852–6865. [CrossRef]
16. Surdeanu, M.; Tibshirani, J.; Nallapati, R.; Manning, C.D. Multi-instance multi-label learning for relation extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Republic of Korea, 12–14 July 2012; pp. 455–465.
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. Ji, G.; Liu, K.; He, S.; Zhao, J. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
19. 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.
20. Alt, C.; Hübner, M.; Hennig, L. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. *arXiv* **2019**, arXiv:1906.08646.
21. 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.
22. 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.
23. Christou, D.; Tsoumakas, G. Improving Distantly-Supervised Relation Extraction Through BERT-Based Label and Instance Embeddings. *IEEE Access* **2021**, *9*, 62574–62582. [CrossRef]
24. 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.
25. Mooney, R.; Bunescu, R. Subsequence kernels for relation extraction. In Proceedings of the Advances in Neural Information Processing Systems 18 (NIPS 2005), Vancouver, BC, Canada, 5–8 December 2005.
26. Culotta, A.; Sorensen, J. Dependency tree kernels for relation extraction. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 21–26 July 2004; pp. 423–429.
27. Miwa, M.; Bansal, M. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv* **2016**, arXiv:1601.00770.
28. 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, Berlin, Germany, 7–12 August 2016; pp. 207–212.
29. Wen, H.; Zhu, X.; Zhang, L.; Li, F. A gated piecewise CNN with entity-aware enhancement for distantly supervised relation extraction. *Inf. Process. Manag.* **2020**, *57*, 102373. [CrossRef]
30. Ye, H.; Luo, Z. Deep ranking based cost-sensitive multi-label learning for distant supervision relation extraction. *Inf. Process. Manag.* **2020**, *57*, 102096. [CrossRef]
31. Shi, Y.; Xiao, Y.; Quan, P.; Lei, M.; Niu, L. Distant supervision relation extraction via adaptive dependency-path and additional knowledge graph supervision. *Neural Netw.* **2021**, *134*, 42–53. [CrossRef]
32. Meng, X.W.; Jiang, T.; Zhou, X.; Ma, B.; Wang, Y.; Zhao, F. Improving Distant Supervised Relation Extraction with Noise Detection Strategy. *Appl. Sci.* **2021**, *11*, 2046. [CrossRef]
33. Zhou, P.; Xu, J.; Qi, Z.; Bao, H.; Chen, Z.; Xu, B. Distant supervision for relation extraction with hierarchical selective attention. *Neural Netw.* **2018**, *108*, 240–247. [CrossRef]
34. 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.
35. Han, X.; Gao, T.; Yao, Y.; Ye, D.; Liu, Z.; Sun, M. OpenNRE: An open and extensible toolkit for neural relation extraction. *arXiv* **2019**, arXiv:1909.13078.
36. Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; Manning, C.D. Position-aware attention and supervised data improve slot filling. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017.
37. Rathore, V.; Badola, K.; Mausam; Singla, P. PARE: A simple and strong baseline for monolingual and multilingual distantly supervised relation extraction. *arXiv* **2021**, arXiv:2110.07415.

38. Jia, R.; Wong, C.; Poon, H. Document-Level *N*-ary Relation Extraction with Multiscale Representation Learning. *arXiv* **2019**, arXiv:1904.02347.
39. Ye, Z.X.; Ling, Z.H. Distant supervision relation extraction with intra-bag and inter-bag attentions. *arXiv* **2019**, arXiv:1904.00143.
40. Gao, T.; Han, X.; Qiu, K.; Bai, Y.; Xie, Z.; Lin, Y.; Liu, Z.; Li, P.; Sun, M.; Zhou, J. Manual Evaluation Matters: Reviewing Test Protocols of Distantly Supervised Relation Extraction. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online Event, 1–6 August 2021.
41. 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.
42. Hao, K.; Yu, B.; Hu, W. Knowing false negatives: An adversarial training method for distantly supervised relation extraction. *arXiv* **2021**, arXiv:2109.02099.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.