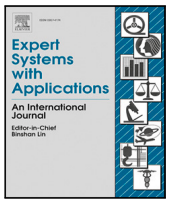
Expert Systems With Applications 284 (2025) 127609

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



SDZRE: A semantic distillation method for zero-shot relation extraction

Yuanjie Zhou [a], Ling Lu [a,*], Hengguang Li [b], Xiaoyang Liu [a], Dan Huang [c], Yinong Chen [d]

[a] School of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China
[b] Department of Mathematics, Wayne State University, Detroit, 48202, USA
[c] China Research and Development Academy of Machinery Equipment, Beijing, 100080, China
[d] School of Computing and Augmented Intelligence, Arizona State University, Tempe, 85281, USA

ARTICLE INFO

Keywords:
Relation extraction
Zero-shot
Semantic matching
Semantic distillation
Contrastive learning

ABSTRACT

Zero-shot relation extraction (ZeroSRE) task aims to identify and extract new relations that have not appeared in the training process. The semantic matching based method has gained significant attention in recent years, which predicts relationships by matching sentences with relational descriptions. However, while most research methods have improved sentence representation quality, they have not fully eliminated interference from complex semantics. Moreover, these methods are prone to misclassifying different relations as the same type when they share similar contextual and entity information. To solve these issues, we propose an efficient Semantic Distillation Method for Zero-Shot Relation Extraction (SDZRE) based on fine-grained semantic matching for ZeroSRE tasks. Specifically, We innovatively design a bidirectional semantic distiller to overcome the limitations of solely removing irrelevant features. This approach enables more effective extraction of the core semantics of sentences, thereby mitigating the interference of complex semantics. Additionally, we propose a novel contrastive learning framework that integrates the bidirectional semantic distiller and employs a combination of random masking and feature truncation strategies for data augmentation. This framework effectively amplifies the differences between similar relations, helping the model learn more meaningful feature representations and reducing the impact of relation similarity. Furthermore, we introduce a multi-negative sample selection and training strategy to further refine the relational feature space, thereby enhancing the discriminative ability of model. Extensive experimental results show that SDZRE, while maintaining efficient inference, significantly outperforms existing methods in extracting core semantics, reducing complex semantic interference, distinguishing similar relations, and enhancing the discriminative ability of model. It achieves state-of-the-art (SOTA) performance, providing a novel approach for the ZeroSRE task that balances both performance and efficiency.

1. Introduction

Relation Extraction (RE) task involves identifying and extracting relations between entities in a given context. It is an essential component of Information Extraction (IE) (Zhang et al., 2024) and serves as a crucial upstream process for many Natural Language Processing (NLP) tasks. Although relation extraction has made significant progress in supervised learning (Soares et al., 2019; Zheng et al., 2021), existing methods heavily rely on large-scale annotated data (He et al., 2023), and the labeling cost rises as the number of relations increases. Additionally, supervised extraction methods tend to have insufficient generalization capability when handling unseen relation types during training (Han et al., 2021). To address this issue, the zero-shot relation extraction (ZeroSRE) task has emerged.

The ZeroSRE aims to enable a model, through training on seen relations, to accurately identify and extract unseen relations (Wang et al., 2019). At present, the mainstream ZeroSRE methods include semantic matching (Obamuyide & Vlachos, 2018), classification network-based method (Liu et al., 2022), prompt learning (Zhang et al., 2022), and utilizing large language models (Huang et al., 2023), among which semantic matching has gained significant attention in recent years. Semantic matching can be categorized into coarse-grained and fine-grained methods. Coarse-grained semantic matching makes overall matching by stitching together different feature representations of sentence. For example, ZS-BERT (Chen & Te Li, 2021) projects sentences and relation descriptions into the same embedding space for semantic matching. However, since the method concatenating all features may introduce

* Corresponding author.
E-mail addresses: AC_RE_ZYJ@stu.cqut.edu.cn (Y. Zhou), ll@cqut.edu.cn (L. Lu), li@wayne.edu (H. Li), lxy3103@cqut.edu.cn (X. Liu), yinong@asu.edu (Y. Chen).

https://doi.org/10.1016/j.eswa.2025.127609
Received 7 December 2024; Received in revised form 6 March 2025; Accepted 4 April 2025
Available online 3 May 2025

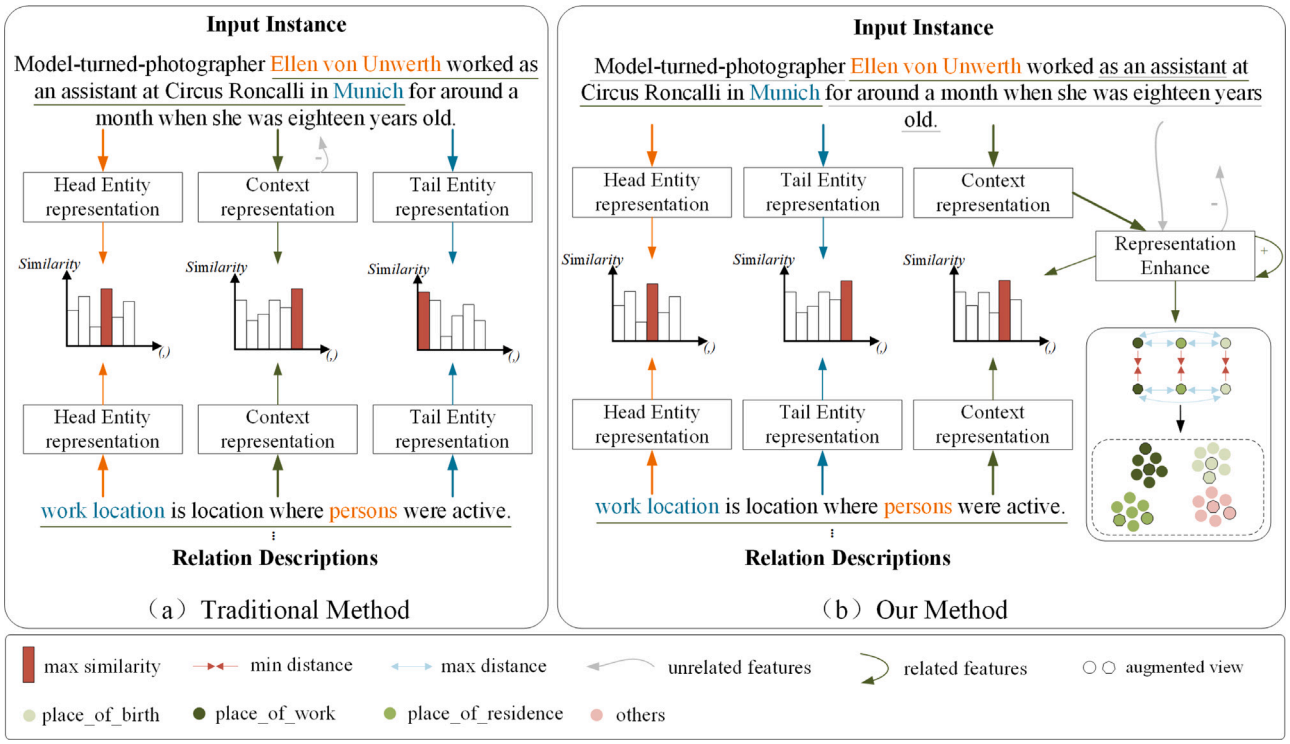


Fig. 1. Examples of different methods. (a) An example of the previous zero-shot relation extraction method based on semantic matching;.

noise, it leads to inaccurate matching, and the method performs poorly in handling similar relations. In contrast, fine-grained semantic matching decomposes sentences into smaller semantic units for matching, making it more sensitive to contextual variations, thus demonstrating a stronger capability in handling complex and similar relations. For example, RE-Matching (Zhao et al., 2023) takes ZeroSRE performance to a new level by using fine-grained semantic matching for the first time. However, although existing methods have improved the quality of sentence representations, they have not completely eliminated the interference of complex semantics and are prone to misclassifying different relations that share similar contextual and entity information as the same type. To this end, this paper proposes a semantic distillation method based on fine-grained semantic matching (SDZRE). There are three main points in this paper, and the corresponding examples are shown in Fig. 1(b).

First of all, the principle of compositionality in Fregean semantics holds that the meaning of a sentence is jointly determined by the meanings of its constituent parts and the way in which they are composed, which suggests that different parts of a text must carry different semantic weights and contribute differently to the construction of the overall semantics. This concept can also be explained by the way photographers accentuate the visual focus of their compositions. They usually enhance the visual presence of the subject while attenuating the visual interference of the background to highlight the main object. Therefore, when processing text, it is necessary to enhance its main features and suppress irrelevant features, so as to highlight its core semantic relations, and thus accurately understand the text. However, although existing methods have tried to deal with irrelevant features, they still cannot completely eliminate the interference caused by complex contexts. To this end, we design a bidirectional semantic distiller, which aims to enhance relevant features and weaken irrelevant features through a complementary approach, thus achieving higher-quality context representation.

Secondly, distinguishing similar relations during the matching process remains a challenge in the ZeroSRE task. It is often difficult for existing methods to accurately predict highly semantically similar relations. This is because these relations may overlap in the representation space, causing the model to be unable to distinguish effectively. For example, like the sentences “In 1978, he replaced Thomas Erdelyi in the Ramones, assuming the name Marky Ramone”. and “The Doctor tries to restore the universe with the help of River and the alternative universe versions of his companions Amy Pond(Karen Gillan) and Rory Williams(Arthur Darvill)”., where the semantics of the relations “member_of” and “part_of” are highly similar, making it difficult for the model to distinguish them correctly. In order to better solve this problem, we drew inspiration from the SimCLR contrastive learning framework (Nguyen et al., 2021) in visual tasks and designed a contrastive learning framework specifically suited for the ZeroSRE task. Building on the bidirectional semantic distiller, this framework incorporates data augmentation strategies based on random masking and feature truncation to generate diverse representations of known relations. These representations, along with the input representations, are treated as positive sample pairs for contrastive learning.

Thirdly, in baseline models, the matching process usually selects negative samples at random. Although this strategy is simple and effective, it may lead to too simple decision boundary of the model, or only learn some simple features with obvious differences while ignoring more important complex features, so that it is difficult to capture the subtle differences between positive and negative samples. For this reason, we select negative samples based on similarity and adopt a multi-negative sample strategy inspired by the work of Shuen Wang et al. (Wang, Duan et al., 2022). This strategy involves learning from both hard negative samples with high similarity to positive samples and semi-hard negative samples with lower similarity, helping the model enhance its ability to capture subtle differences between samples.

The proposed SDZRE method offers the following advantages:

(1) The bidirectional semantic distiller avoids the limitations of simply removing irrelevant features, effectively eliminating noise while enhancing the ability to extract the core semantics of a sentence.

(2) The proposed contrastive learning framework focuses more on core semantic features and employs various strategies to generate diverse representations. This not only mitigates the risk of semantic drift but also forces the model to attend to more discriminative local semantics.

(3) By leveraging a multi-negative sample strategy, the model learns subtle differences between positive samples and easily confused negative samples, further refining the decision boundaries of the relational feature space. This significantly enhances its robustness to ambiguous semantic boundaries.

In summary, our main contributions are the following five points:

(1) We propose a semantic distillation method based on fine-grained semantic matching, which significantly enhances the utilization of sentence semantics and effectively mitigates the interference of similar relations.

(2) We innovatively design a bidirectional semantic distiller that dynamically reinforces key information while suppressing irrelevant semantics, effectively improving the extraction of core sentence semantics and reducing the impact of complex semantics.

(3) We design a contrastive learning framework based on the bidirectional semantic distiller, enabling the model to learn more discriminative feature representations, further increasing the distinction between similar relations and effectively alleviating confusion among them.

(4) We explore and leverage a multi-negative sample strategy, constructing and learning from negative samples with varying similarity levels to further enhance the model's discrimination and generalization capabilities.

(5) Experimental results on FewRel and Wiki-ZSL datasets show that SDZRE achieves a new SOTA performance that significantly outperforms the existing SOTA methods, which fully proves the effectiveness and superiority of our method.

The rest of the paper is structured as follows. Section 2 discusses the related work, particularly on zero-shot relation extraction and contrastive learning. Section 3 presents the proposed model and its components. Section 4 outlines our experiments, results and analyses. Section 5 concludes the paper.

2. Related work

2.1. Zero-shot relation extraction

The purpose of the ZeroSRE is to identify and extract unseen relations between entities in scenarios without training instances. The task was initially regarded as a question-answering task (Levy et al., 2017), which inevitably requires manually defined question templates for new relations. The existing methods for the ZeroSRE task can be grouped into four main categories.

The first method is a classification network-based method, which treats relation extraction as a classification problem by introducing a classification network to predict relations (Lv et al., 2023). For example, MCMN (Liu et al., 2022) achieves ZeroSRE by multi-choice prompt and triple-interpreter learning pre-training. This type of method can effectively improve prediction accuracy and stability using classification techniques, but these methods overly relies on manual templates and the computational cost increases significantly with the increase of the number of unknown relations (Li, Zhang et al., 2024). The second method is a prompt learning-based method, which guides pre-trained language models to directly generate or identify relations by constructing appropriate prompts(Guo et al., 2024). For example, ZS-SKA (Gong & Eldardiry, 2024) effectively extracts unknown relational triples in a zero-shot setting through semantic knowledge augmentation and virtual label construction. NSP-RTE (Liao et al., 2024) transforms relation extraction into a next-sentence prediction task, eliminating the need for sample synthesis. This type of method can significantly reduce labeling costs in zero-shot scenarios without requiring a large amount of annotated data. However, such methods depends on synthetic data or external resources, and in some cases, similar relations with minor differences in description templates may cause confusion due to semantic overlap. Moreover, in complex semantic scenarios, the coverage of synthetic data is often insufficient, and the model's ability to comprehend intricate semantic structures is limited. As a result, the accuracy of relation extraction decreases, ultimately affecting task performance. The third method is based on large language models. Such methods typically utilize the pre-training knowledge of large models to identify complex relation types and make predictions without annotated data (Agrawal et al., 2022; Zhou et al., 2024). For example, ChatIE (Wei et al., 2023) effectively improved the performance of ZeroSRE task by using a large-scale language model. REA (Layegh et al., 2024) integrates external knowledge for prompt tuning, enabling the direct extraction of entities and their relations from unlabeled text. Such methods have brought new breakthroughs to the field (Li et al., 2023; Zhang et al., 2023), enabling efficient identification and extraction of unknown relations without requiring additional annotated data, thereby significantly reducing data annotation and development costs.but many methods have not been specifically optimized for similar relations, making it difficult to distinguish their subtle differences. Especially during the inference process, the model fails to effectively handle complex sentence structures and semantic dependencies, leading to a decline in relation extraction accuracy. Moreover, the high inference cost limits multiple optimization attempts, further impacting task performance.

The fourth method is a semantic matching based method that matches the relation descriptions with the input instances correspondingly and minimizes the distance between them. For example, ZS-BERT learns relation representations through relation descriptions and uses nearest neighbor search to predict unseen relations in new sentences. In recent years, most research have built upon this method, further improving the quality of semantic representations through techniques such as semantic alignment, contrastive learning, and the incorporation of external knowledge (Wang et al., 2019; Zhao et al., 2024). (Chen et al., 2023) enhanced the performance of zero-shot relation extraction by leveraging graph structures to provide additional semantic information for the model. Li, Zhang et al. (2024) were the first to propose aligning input text with relation descriptions through encoding and semantic alignment, mapping them into a shared semantic space. However, these methods focus only on sentence-level semantics and suffers from the problems of easily confusing similar relations and insensitivity to complex syntax and lexical polysemy. Zhao et al. (2023) proposed a fine-grained semantic matching method, enabling the model to focus more on the information between specific features within a sentence, thereby partially addressing the limitations of coarse-grained matching. However, there are still limitations in the method when facing complex relation types and highly similar semantics. Li, Bai et al. (2024) incorporated a multi-granularity matching mechanism to capture detailed features in entity pairs and relation descriptions, enhancing model performance and generalization ability through information fusion. Our method is based on fine-grained semantic matching, but it is also significantly different from existing methods. Without relying on explicitly labeled data, we innovatively design a more comprehensive semantic distillation method to overcome the limitations of existing approaches in extracting core semantics. Additionally, we propose a multi-negative sample strategy to construct diverse and highly distinguishable negative samples.

2.2. Contrastive learning

The core idea of the contrastive learning task is to learn more discriminative feature representations by constructing pairs of positive and negative samples to bring positive pairs closer together and push negative pairs farther apart in the potential space (Chen & Li, 2024; He et al., 2020; Zhu et al., 2022). In the field of Natural Language Processing (NLP), contrastive learning has been gradually applied to improve the ability of the models to distinguish semantic relations (Xu et al., 2024). For example, ConSERT (Yan et al., 2021) introduced contrastive learning in the representation layer of the pre-trained language model, which significantly improved the performance of the model in text

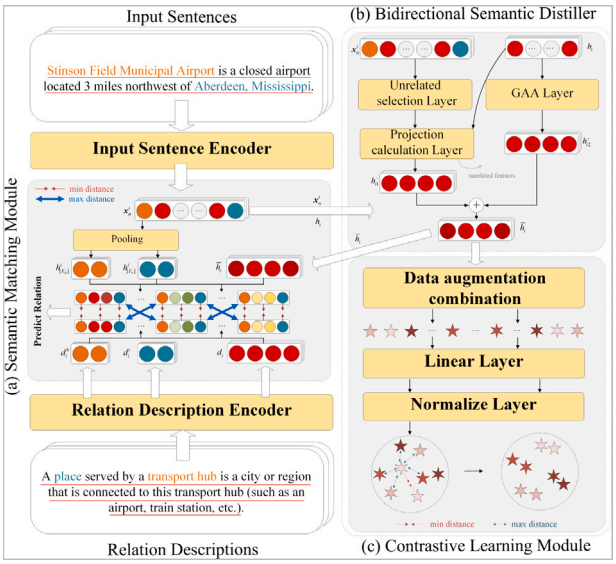


Fig. 2. The overall architecture of the proposed SDZRE.

similarity tasks. DeCLUTR (Giorgi et al., 2021) used an unsupervised contrastive learning method to capture fine-grained semantic relations in sentences th rough self-supervised training at the sentence level. RCL (Wang, Zhang et al., 2022) employs Dropout as a data augmentation technique to amplify semantic differences between similar instances without disrupting relation representations. CL&CD (Yang et al., 2024) adopts a two-stage contrastive learning approach, leveraging both pseudo-labeled and labeled data for training. However, these methods face challenges such as difficulty in capturing complex semantic variations and high computational complexity. Inspired by these contrastive learning tasks, we introduce a contrastive learning method in this study, combined with a bidirectional semantic distiller, to help the model focus more on core semantics, better learn and amplify differences between similar relations, and ultimately obtain more effective feature representations.

3. Methodology

3.1. Task formulation

The goal of the zero-shot relation extraction (ZeroSRE) task is to generalize to the unseen relations $R_u = \{r_s^1, r_s^2, \ldots, r_s^m\}$ by learning the samples in the seen relations $R_s = \{r_u^1, r_u^2, \ldots, r_u^n\}$. Moreover, the two relation sets are disjoint, and the model can only learn from the samples in the seen relations R_s during training. Following previous work (Chen & Te Li, 2021; Zhao et al., 2023), we formulate the ZeroSRE as a semantic matching task.

In the training phase, given a training set $D = \{(X_i, e_{ih}, e_{it}, y_i, p_i) \mid i = 1, 2, \ldots, N\}$ containing N sample data, where X_i represents input sentence, e_{ih} represents head entity, e_{it} represents tail entity, $y_i \in R_s$ represents relation and p_i represents relation description. We optimize a semantic matching model $\mathcal{M}(X, e_h, e_t, p) \rightarrow s \in \mathbb{R}$ on R_s by a designed semantic distillation method, where s represents the matching score between the input sentence X and the relation description p.

In the testing phase, given a sample (X_j, e_{jh}, e_{jt}) from the unseen relations R_u, we use the model $\mathcal{M}$ to calculate the matching score between the relation description and the input sentence, and select the relation with the highest matching score as the prediction result.

3.2. Model overview

Our proposed SDZRE is shown in Fig. 2, which consists of five major parts: input sentence encoder, relational description encoder, bidirectional semantic distiller, contrastive learning and fine-grained semantic matching.

First, BERT (Devlin et al., 2019) is used as a pre-training encoder for the input sentence encoder to generate representation vectors for the entity and context. Sentence-BERT (Reimers & Gurevych, 2019) is used as a pre-training encoder for the relation description encoder to generate entity and context representation vectors in the relation description. The bidirectional semantic distiller is used to weaken irrelevant features in context representation and strengthen relevant features, thereby improving the effect of context matching. Contrastive learning is performed by utilizing enhanced representations of different context representations in the same batch to further optimize the discriminant power of the model. Finally, the fine-grained semantic matching module matches the entity and context representation of the input sentence with the entity and context representation of the relation description, ensuring that the model can capture more accurate semantic information (see Fig. 3).

3.3. Pretraining encoder

3.3.1. Input sentence encoder

We choose BERT as a pre-trained encoder to generate entity and context representations of input sentences. Specifically, given an input sentence $X_i = \{x_1^i, x_2^i, \ldots, x_n^i\}$, four special tokens $[E_h]$, $[\backslash E_h]$, $[E_t]$ and $[\backslash E_t]$ are used to mark the head entity and tail entity in the sentence. After obtaining the output of the encoder, we use special tokens $[E_h]$ and $[E_t]$ to obtain the representation of the head entity, the tail entity, and the context representation. The specific formula is as follows:

$$\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_n^i = \text{BERT}\left(x_1^i, x_2^i, \ldots, x_n^i\right) \tag{1}$$

$$h_{[E_h]}^i = \text{WeightPooling}\left(\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_n^i\right) \tag{2}$$

$$h_{[E_t]}^i = \text{WeightPooling}\left(\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_n^i\right) \tag{3}$$

$$h_i = \mathbf{x}_{[E_h]}^i \oplus \mathbf{x}_{[E_t]}^i \tag{4}$$

where, $\mathbf{x}_n^i$ represents the hidden state of the token x_n^i, $h_{[E_h]}^i$ is the embedded representation of the head entity, and $h_{[E_t]}^i$ is the embedded representation of the tail entity. We concatenate the hidden states of two special tokens $[E_h]$, $[E_t]$ to form a context embedded representation h_i, and $\oplus$ represents the concatenation operator. We take a similar approach to Lin et al. (2017) and use a weight pool to extract the entity representation from $\mathbf{x}_n^i$.

3.3.2. Relation description encoder

Relation descriptions $p \in \{p_1, p_2, \ldots, p_n\}$ are brief textual descriptions of the type of relation. For example, the corresponding relation description for "born in" is "The city where someone was born". We use Sentence-BERT as a pre-trained encoder to generate the corresponding entity and context representations. Specifically, given a relation description p_i, the corresponding embedded representation is obtained by the head entity abbreviation s_h, the tail entity abbreviation s_t, and the relation description $des = \{w_1, w_2, \ldots, w_n\}$, the specific formula is as follows:

$$d^v = \text{Sentence} - \text{BERT}\,(des) \tag{5}$$

$$d^h = \text{Sentence} - \text{BERT}\left(s_h\right) \tag{6}$$

$$d^t = \text{Sentence} - \text{BERT}\left(s_t\right) \tag{7}$$

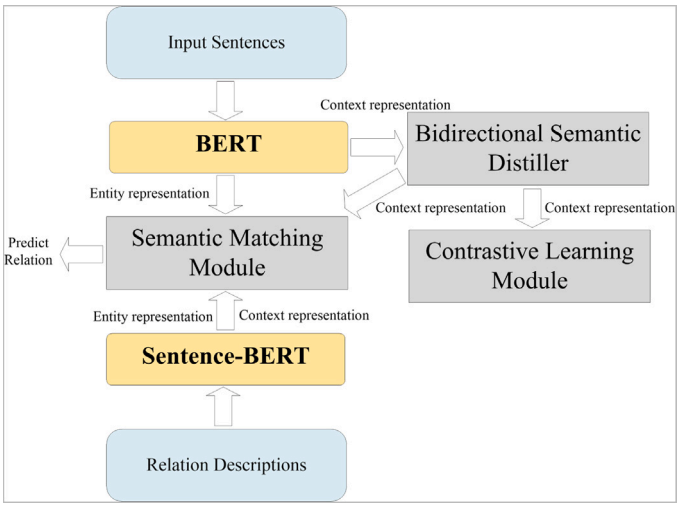


Fig. 3. SDZRE Overall Process Diagram. First, we use BERT to encode the input sentences and Sentence-BERT to encode the relation descriptions. The resulting representations contain both entity and contextual information. Next, the context representation of the input sentence is passed into the bidirectional semantic distiller. This module extracts the core semantics of the sentence by weakening irrelevant features and strengthening relevant features, thereby reducing interference from complex semantics. Then, the obtained results are fed into the contrastive learning module, where data is augmented using random masking and feature truncation strategies. This generates multiple different representations for the same relation to facilitate effective learning. Simultaneously, the results and the entity representations are combined with the relation description's entity and context representations and sent into the semantic matching module. The semantic matching module employs a fine-grained matching approach, combining marginal ranking loss and a multi-negative sample selection strategy to help the model better perform relation matching.

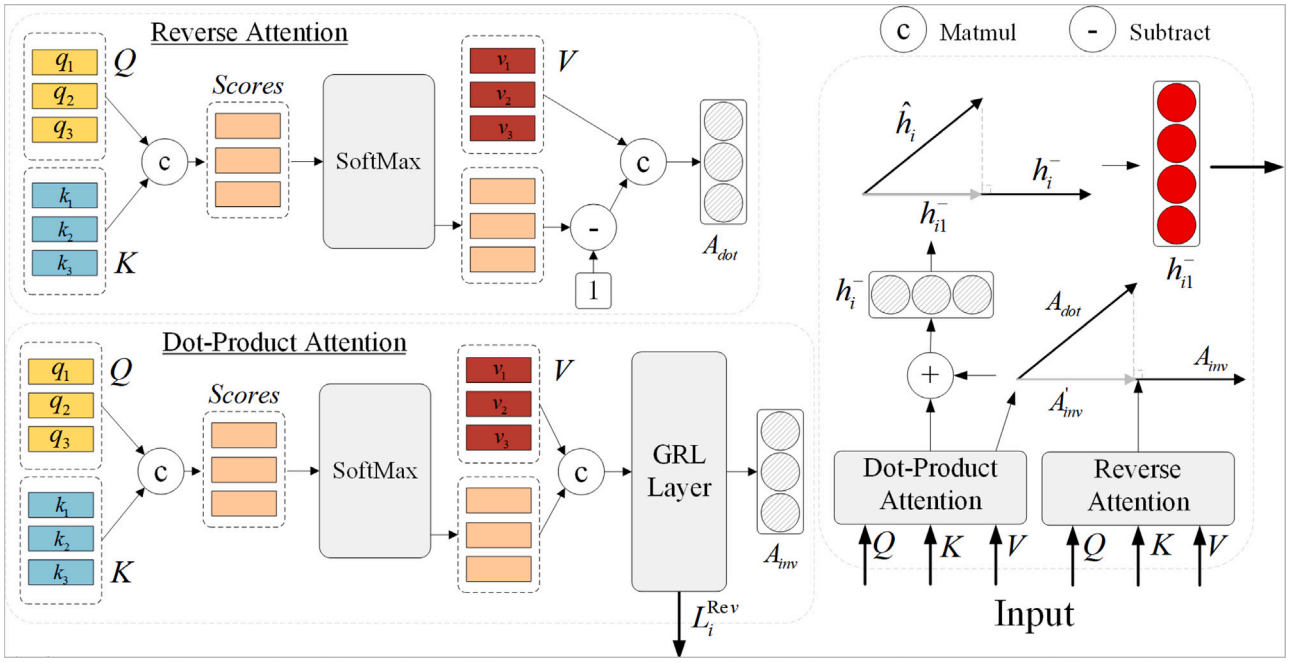


Fig. 4. The overall framework of the fusion projection distiller. The module effectively identifies irrelevant features through a fusion attention approach that combines dot-product attention with reverse attention. Then, these irrelevant features are removed from the representation by vector projection.

where, d^v is the context embedded representation of the relation description, d^h is the head entity embedded representation of the relation description, and d^t is the tail entity embedded representation of the relation description.

3.4. Bidirectional semantic distiller

Since irrelevant features in context embedded representations may interfere with model learning, the model needs to be able to identify and weaken these irrelevant features, resulting in a higher quality and more efficient representation. To achieve this goal, we design a bidirectional semantic distiller, which not only weakens the influence of irrelevant features, but also improves the influence of relevant features.

3.4.1. Fusion projection distiller

The overall framework of the component is shown in Fig. 4. The module mainly consists of an irrelevant selection layer and a projection calculation layer, whose task is to first identify irrelevant features and then remove them. The purpose of the irrelevant selection layer is to identify features that are irrelevant to the relation. Specifically, the encoder output $\mathbf{x}_n^i$ is initially screened through the dot-product attention mechanism to capture potentially irrelevant features, and then the reverse attention mechanism (Huang et al., 2017) is introduced to assist the dot-product attention mechanism to further identify those features that are not highly attended to, which may be relevant to the relation. Through this fusion attention strategy, the model can more accurately

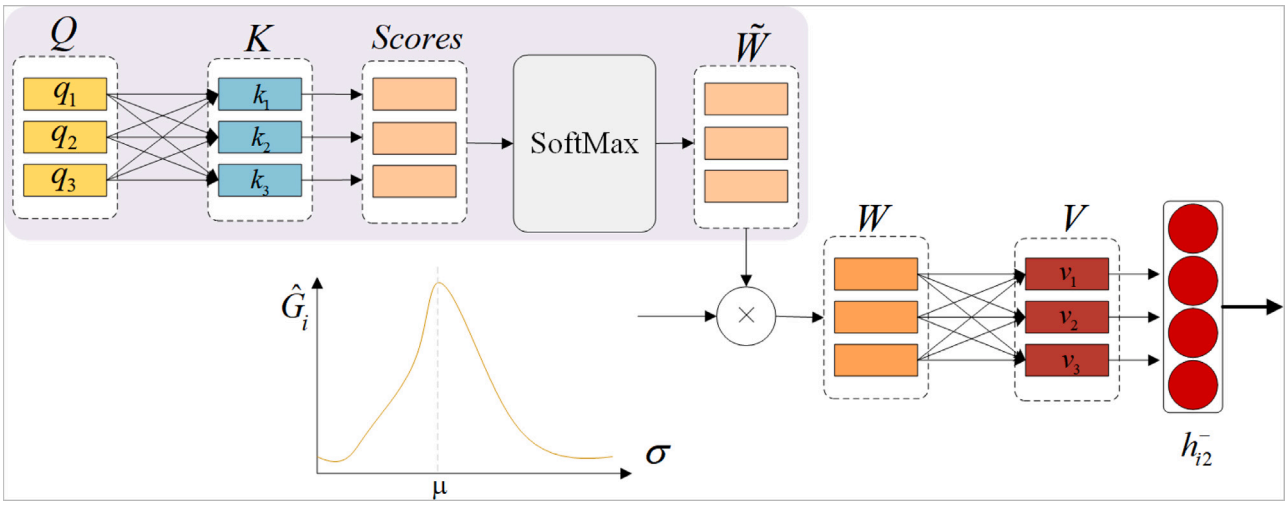


Fig. 5. The overall framework of Gaussian adaptive selector. The position weights are dynamically adjusted by the Gaussian-weighted attention mechanism, and the weighted results are combined with the traditional attention mechanism to accurately select the relevant features in each context representation.

identify irrelevant features, and the specific formula is as follows:

$$A_{dot} = \text{Softmax}\left(Q \cdot K^T\right) \cdot V \tag{8}$$

$$A_{inv} = \left(1 - \text{Softmax}\left(Q \cdot K^T\right)\right) \cdot V \tag{9}$$

$$A'_{inv} = A_{inv} - \left(\frac{A_{inv} \cdot A_{dot} \cdot A_{dot}}{\left|A_{dot}\right|^2}\right) \tag{10}$$

$$h_i^- = A_{dot} + A'_{inv} \tag{11}$$

where, Q is the query matrix composed of the query vector q, K and V are the key and the value matrix, respectively. A_{dot} and A_{inv} are the irrelevant features obtained through the dot-product attention and reverse attention, respectively. A'_{inv} is the irrelevant features computed by vector projection, i.e. those irrelevant features that are not paid attention to by the dot-product attention. $\left|A_{dot}\right|^2$ is the square of the length of A_{dot}, h_i^- represents the finally obtained irrelevant features. In addition, we introduce the Gradient Reverse Layer (GRL) (Ganin & Lempitsky, 2015), which is widely used in related research, to ensure that query vector q can identify and select irrelevant features. The function of GRL is to keep the input unchanged when propagating forward, and reverse the gradient when propagating backward. The specific formula is as follows:

$$Cp_i = \mathcal{E} \cdot GRL\left(A_{dot}\right) + b \tag{12}$$

$$L^{Rev} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\left(\text{Softmax}\left(Cp_i\right)\right) \tag{13}$$

where, Cp_i represents the output of the classifier, $\mathcal{E}$ and b represent the classification weight and bias, respectively, L^{Rev} represents the reverse gradient loss, and N represents the number of samples in the batch.

The purpose of the projection calculation layer is to minimize irrelevant features in context representation. Our method is to project the context representation h_i through the fully-connected layer transformation to the direction of irrelevant features h_i^- to find the irrelevant features in it. Then, these irrelevant features are removed to obtain a new context representation $\bar{h}_{i1}$, which is formulated as follows:

$$\bar{h}_{i1} = \hat{h}_i - \left(\frac{\hat{h}_i \cdot h_i^- \cdot h_i^-}{\left|h_i^-\right|^2}\right) \tag{14}$$

where, $\left|h_i^-\right|^2$ represents the square of the length of h_i^-, and $\hat{h}_i$ is the context representation after the fully connected layer.

3.4.2. Gaussian adaptive selector

Different from the fusion projection distiller, the task of the module is to select the most relevant features from the context representation. To this end, we introduce the Gaussian Adaptive Attention Mechanism (GAAM) (Ioannides et al., 2024), which aims to enable the model to focus on the most relevant features in the context by adaptively adjusting the distribution of attention.

The overall framework of the component is shown in Fig. 5. In the module, we use the dot-product attention mechanism to calculate the attention weight of context representation $\hat{h}_i$, and the specific formula is as follows:

$$\widetilde{W} = \text{Softmax}\left(Q \cdot K^T\right) \tag{15}$$

where, Q indicates the query matrix, K indicates the key matrix, and $\widetilde{W}$ indicates the normalized attention weight. In order to better capture the relation between various features in the context representation, we use the mean μ of input sequence k_i to determine the relative position relation between each element in $\hat{h}_i$ and the center position of the sequence, and μ can further adjust the position weight of features in the attention distribution through the learning offset μ_{offset}, the specific formula is as follows:

$$\mu = \frac{1}{n}\sum_{j=1}^{n} k_i + \mu_{offset} \tag{16}$$

$$D_i = \left|l_i - \mu\right| \tag{17}$$

where, n represents the sequence length, l_i represents the position sequence composed of each element position in k_i, and D_i represents the distance of each element position in k_i relative to the mean μ. Then, we use standard deviation σ and distance D_i to calculate the Gaussian distribution weight, with the standard deviation dynamically calculated. The specific formula as follows:

$$\sigma = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(k_i - \mu\right)^2} \tag{18}$$

$$G_i = \exp\left(-\frac{1}{2} \cdot \left(\frac{D_i}{\sigma}\right)^2\right) \tag{19}$$

$$\hat{G}_i = \frac{G_i}{\sum_{i=1}^{n} G_i} \tag{20}$$

where, G_i and $\hat{G}_i$ represent the original Gaussian distribution weight and the normalized Gaussian distribution weight, respectively. In the model, as μ and σ change, the Gaussian weights of different positions will also change accordingly, thus dynamically adjusting the attention

weights of each position. Specifically, when a position is closer to the central position μ, the greater the Gaussian weight, the higher the importance of the position; Conversely, the farther a position is from the central position μ, the smaller the Gaussian weight and the lower the importance of the position.

Finally, we multiply the Gaussian distribution weight $\hat{G}_i$ by the dot-product attention weight $\widetilde{W}$ to achieve dynamic calibration of the attention distribution, so that the features closer to the central position gain higher importance, as follows:

$$W = \hat{G}_i \cdot \widetilde{W} \tag{21}$$

$$\bar{h}_{i2} = W \cdot V \tag{22}$$

where, W is the final attention weight, $\bar{h}_{i2}$ is the final output representation, and V is the value matrix.

3.4.3. Bidirectional semantic distillation strategy

As shown in Algorithm 1, we input the context representation h_i into the fully connected layer and input the output into the fusion projection distiller and Gaussian adaptive selector respectively for further processing. In the fusion projection distiller, we use the fusion attention strategy and vector projection to maximize the recognition and remove the irrelevant features in the context representation h_i, thus obtaining an optimized context representation $\bar{h}_{i1}$. In the Gaussian adaptive selector, we mainly use the Gaussian adaptive attention mechanism to select the relevant features in h_i, so as to obtain the context representation $\bar{h}_{i2}$. Finally, we combine $\bar{h}_{i1}$ and $\bar{h}_{i2}$ to generate a higher quality context representation, as follows:

$$\bar{h}_i = \bar{h}_{i1} + \bar{h}_{i2} \tag{23}$$

where, $\bar{h}_i$ is the final context representation.

Algorithm 1: Bidirectional Semantic Distiller

Input: The sentence representation $\mathbf{x}_n^i$, the context representation h_i, and the context representation after the fully connected layer is denoted as $\hat{h}_i$.
Output: The higher-quality context representation $\bar{h}_i$

for r *in* $[\mathbf{x}_n^i, h_i]$ **do**
  A_{dot} <= get_dot_ind($h, \mathbf{x}$);
  A_{inv} <= get_inv_ind($h, \mathbf{x}$);
  A'_{inv} <= get_newinv_ind(A_{dot}, A'_{inv});
  h_i^- <= get_final_ind($h, \mathbf{x}$);
  for i *in* $\hat{h}_i$ **do**
    $\bar{h}_{i1}$ <= get_final_one($\hat{h}$);
  end
  for i *in* $\hat{h}_i$ **do**
    μ <= get_mean($\hat{h}_i$);
    σ <= get_sta($\hat{h}_i$);
    $\bar{h}_{i2}$ <= get_final_two($\mu, \sigma, \hat{h}_i$);
  end
  $\bar{h}_i$ <= $\bar{h}_{i1} + \bar{h}_{i2}$;
  return $\bar{h}_i$;
end

3.5. Contrastive learning

The contrastive learning framework we designed consists of a data augmentation layer and a feature transformation layer to help the model better learning and distinguishing between similar relations and highly abstract relations. Specifically, this goal is achieved by maximizing the similarity of enhanced representations from different data of the same representation vector and minimizing the similarity between the enhanced representations from other samples in the same batch.

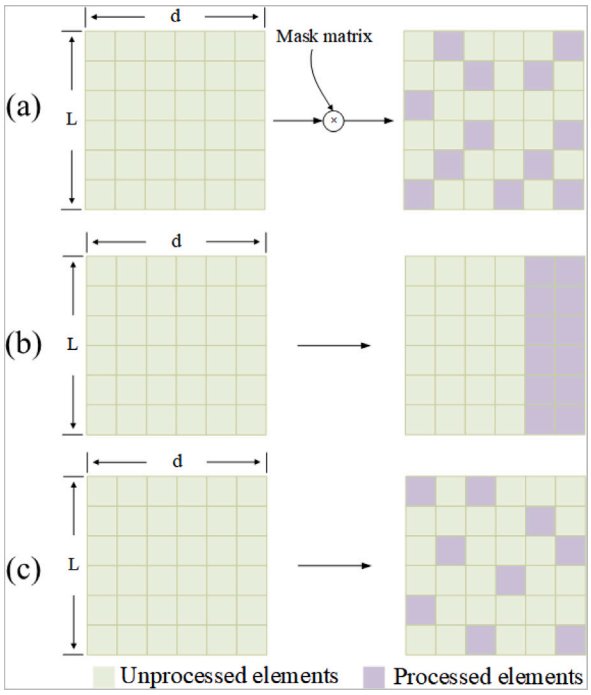


Fig. 6. Examples of three different data augmentation strategies. (a) The process of random mask strategy processing; (b) The process of feature cutoff strategy processing;(c) The process of Dropout strategy processing.

3.5.1. Data augmentation layer

The data augmentation layer is designed to magnify the semantic differences between similar contexts and help the model distinguishing between difficult relation types without destroying relational semantics. To this end, we compare three different combinations of data enhancement strategies: Dropout (Gao et al., 2021) and random mask (Hinton, 2012), Dropout and feature cutoff (Shen et al., 2020), and random mask and feature cutoff, and the specific examples are shown in Fig. 6.

The effectiveness of the random mask strategy has been fully verified (Tian et al., 2022). The method is similar to dropout, which randomly masks a part of the feature elements in the context representation by a preseting probability, thus preventing the model from over-relying on them. In our experiment, we generate a random mask matrix with the same shape as the embedding matrix by setting a specific mask probability. We then multiply it with the embedding matrix element by element, and set the element at the corresponding positions to zero.

The feature cutoff strategy can effectively reduce the interference to the overall semantic of the sentence, maintain the consistency of the context semantics to the maximum extent, and weaken the influence of some features in the representation. In the experiment, we truncate some dimensions of input features by setting certain feature dimensions to zero in the embedding matrix, which avoids the model from relying on some specific features.

The effectiveness of Dropout as one of the simplest and most effective data augmentation strategies has been proven by various studies (Hinton, 2012). In the experiment, we set specific probabilities and randomly set certain elements in the embedding matrix to zero to weaken the model from relying on specific features and avoid overfitting.

We use a combination strategy for different data augmentation strategies, which involves simultaneously applying different data augmentation strategies to the same input to generate two different enhanced representations of the same input. The experimental results

show that the combination strategy can improve the diversity of data and the robustness of the model. In this paper, a combination strategy of random mask and feature cutoff is chosen.

3.5.2. Contrastive learning strategy

As shown in Algorithm 2. For input representation $\bar{h}_i$, we apply random mask and feature truncation strategies to it respectively in the data augmentation layer to generate its enhanced representation z_i^r and z_i^d. Then, we linearly transform the enhanced representation with a fully connected layer and transform it into a uniform length representation $\hat{z}_i^r$, $\hat{z}_i^d$ by a normalization layer, making it better suited for similarity measurement. Finally, following the SimCLR contrastive learning framework, we choose to use the InfoNCE loss function for presentation optimization. The specific formula is as follows:

$$m_{ij} = \frac{z_i^r \cdot \left(z_j^d\right)^T}{\tau} \tag{24}$$

$$L^{Ctr} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(m_{ij}\right)}{\sum_{j=1}^{N} \exp\left(m_{ij}\right)} \tag{25}$$

where, τ represents the temperature parameter, m_{ij} represents the similarity matrix between the two enhanced representations, and N represents the batch size.

Algorithm 2: Contrastive Learning

Input: The context representation $\bar{h}_i$ processed by the bidirectional semantic distiller, temperature τ, mask probability, feature truncation ratio

Output: Contrastive learning loss L^{Ctr}

```
for i in h̄_i do
    z_i^r <= mark_aug(h̄_i, mask probability);
    z_i^d <= truncation_aug(h̄_i, feature truncation ratio);
    ẑ_i^r <= process_layer(z_i^r);
    ẑ_i^d <= process_layer z_i^r);
    L^Ctr <= com_loss(ẑ_i^r, ẑ_i^d, h̄_i);
    return L^Ctr;
end
```

3.6. Semantic matching

3.6.1. Fine-grained matching score

We divide the complete input sentence into context, head entity and tail entity, and match them with the relation description accordingly. Specifically, first we encode the input sentence X_i to obtain the corresponding representation $\left(h^i_{[E_h]}, h_i, h^i_{[E_t]}\right)$. Then, we input the context representation h_i into the context bidirectional semantic distiller for processing, resulting in a new sentence representation $\left(h^i_{[E_h]}, \overline{h}_i, h^i_{[E_t]}\right)$. At the same time, relation description p is encoded to obtain relation description representation $\left(d^h, d^v.d^t\right)$. Finally, the formula for calculating the matching score between input sentence X_i and relation description p is as follows:

$$f\left(X_i, p\right) = \alpha \cdot \left[\mathrm{sim}\left(h^i_{[E_h]}, d^h\right) + \mathrm{sim}\left(h^i_{[E_t]}, d^t\right)\right] + (1-2\alpha)\cdot \mathrm{sim}\left(\overline{h}_i, d^v\right) \tag{26}$$

where, α indicates the balance parameter, and the default setting is 0.33. $\mathrm{sim}(,)$ represents the cosine similarity calculation function, and $f\left(X_i, p\right)$ represents the final match score.

3.6.2. Marginal rank loss

In order to further refine the relation feature space and improve the accuracy of matching, we introduce an efficient multi-negative sample strategy to achieve joint optimization based on the original marginal ranking loss. Specifically, we select negative samples based on similarity, adopt top-k method to treat the first k negative samples as hard negative samples, and treat the negative samples whose similarity is lower than hard negative samples as semi-hard negative samples. In the matching process, let the model not only pays attention to the hard negative sample p_j^{negt} with the highest similarity to the positive sample, but also pays attention to the semi-hard negative sample p_j^{negb} with slightly lower similarity to the positive sample, but still has certain interference. The specific ranking loss calculation formula is as follows:

$$L_i^{Mrt} = \max\left(0, \max_{i\neq j}\left(f\left(X_i, p_j^{negt}\right)\right) - f\left(\left(X_i, p_{y_i}\right)\right) + gam_t\right) \tag{27}$$

$$L_i^{Mrb} = \max\left(0, \max_{i\neq j}\left(f\left(X_i, p_j^{negb}\right)\right) - f\left(\left(X_i, p_{y_i}\right)\right) + gam_b\right) \tag{28}$$

$$L^{Mr} = \frac{1}{N}\sum_{i=1}^{N}\left(L_i^{Mrt} + L_i^{Mrb}\right) \tag{29}$$

where, N denotes the batch size, and gam_t and gam_b are boundary parameters for marginal ranking loss. After L^{Mr} optimization, the model can make the input sentence closer to the correct relation description, while moving away from the most similar but incorrect relation description, and slightly less similar relation description. The final loss function of the whole model is as follows:

$$L = L^{Rev} + L^{Ctr} + L^{Mr} \tag{30}$$

4. Experiments

4.1. Datasets

To evaluate the performance of different methods on zero-shot relation extraction (ZeroSRE) task, we use two datasets commonly used in this field: FewRel (Han et al., 2018) and Wiki-ZSL (Chen & Te Li, 2021).

FewRel is a Wikipedia-based human-annotated dataset specifically designed for being used with few-shot task. However, as long as the training set and the test set do not contain the same relation type, it can also be applied to zero-shot task. It contains 80 classes of relations, each consisting of 700 sentence instances.

Wiki-ZSL is a dataset derived from Wiki-KB and generated by remote supervision. Wiki-ZSL has more data noise, but its relation types are richer than FewRel. It contains 113 classes of relations and 93 383 sentence instances.

Following the method of Li, Bai et al. (2024), our experiments randomly select $m \in \{5, 10, 15\}$ class relations as the test set, 5 class relations as the validation set, and the remaining relations as the training set. All experiments were conducted on datasets with 5 different random seed partitions, and the average results of each experiment were reported.

4.2. Evaluation metrics

In our experiment, F1 score is used as the main evaluation metrics. The F1 score is a harmonic average of Precision and Recall, which can effectively reflect the performance of the model in tasks with an unbalanced number of class instances. In addition, we also reported the accuracy and recall corresponding to F1 scores to more comprehensively evaluate the performance of the model.

4.3. Baselines

We contrast our method with the following methods:

AlignRE (Li, Zhang et al., 2024) improves the performance of the ZeroSRE task through coding pattern alignment and semantic alignment, reducing manual intervention in prototype construction.

RE-Matching (Zhao et al., 2023) realizes the matching of entity and context separation for the first time, and effectively filters out irrelevant information in the context.

SUMASK (Li et al., 2023) uses large language models to improve the performance of ZeroSRE tasks through recursive text summary and question answering framework.

REA (Layegh et al., 2024) proposes a novel "Refine-Estimate-Answer" prompting strategy, leveraging pre-trained large models to progressively optimize information processing in the zero-shot relation extraction task.

NSP-RTE (Liao et al., 2024) transforms zero-shot relational triple extraction into a next-sentence prediction task, eliminating the need for sample synthesis and effectively enhancing the generalization ability of model.

ZRCM (Zhu et al., 2022) is a ZeroSRE method based on contrastive learning, which improves model generalization ability by designing negative sample generator and multi-task learning structure.

RelationPrompt (Chia et al., 2022) completes the zero-shot relation triple extraction task by generating synthetic data with a prompt language model, and it designs a triplet search and decoding method to improve the effect of extracting multiple relational triples from a single sentence.

NoGen (Chia et al., 2022) method has the same setting as RelationPrompt, but it does not use the generated synthetic sample for training.

PromptMatch (Sainz et al., 2021) is a ZeroSRE model based on the most advanced full encoding technology, which concatenates input pairs through BERT and models their fine-grained semantic interactions in depth.

ZS-BERT (Chen & Te Li, 2021) is a Siamese network ZeroSRE model using BERT as an encoder. By combining classification loss and metrics-based loss, the representation space is optimized to improve the effect of nearest neighbor search.

We select these specific baseline methods primarily for their innovation and representativeness in the ZeroSRE task. They encompass diverse technical approaches and implementations, including semantic matching, prompt learning, classification networks, and contrastive learning. These methods provide a comprehensive comparative perspective, allowing us to more effectively demonstrate the advantages and innovations of our proposed approach.

4.4. Implementation details

We used BERT-base-uncased and Sentence-BERT as pre-trained encoders for the input sentence and relation descriptions, respectively, and fine-tuned them to the task requirements. We use the AdamW optimizer with the learning rate set to 2e−6, epochs of 5, temperature τ of 0.05, and mask probability of 0.15. In order to achieve the best performance for the model on both datasets, we optimize the other parameters: the feature truncation ratios for FewRel and Wiki-ZSL are set to 0.2 and 0.1, respectively, and the batch sizes were 32 and 128, respectively. All experiments were performed in the NVIDIA GeForce RTX 3090.

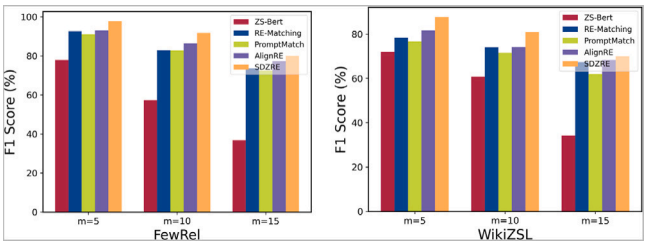


Fig. 7. Change in Matching F1 score for different number of unseen relations.

4.5. Experimental results

4.5.1. Main results

In Table 1, we compare the results of different methods in predicting different m unknown relations. The experimental results show that SDZRE significantly outperforms the previous best baseline model on both datasets, achieving the current SOTA results. On predicting three different numbers of unknown relations, SDZRE increases the F1 values in FewRel by 4.70%, 5.25%, and 2.75%, and in Wiki-ZSL by 1.05%, 6.58%, and 0.7%, respectively, compared to the State-of-the-Art methods. In addition, our method provides a significant improvement over baseline RE-Matching. In predicting three different quantity unknown relations, F1 values in FewRel are increased by 5.21%, 8.73% and 6.40%, respectively, and F1 values in Wiki-ZSL are increased by 9.37%, 7.06% and 2.63%, respectively. These results demonstrate the excellent performance of SDZRE in terms of its effectiveness in ZeroSRE tasks and its ability to predict more unknown relations.

Fig. 7 shows the F1 score changes of some baseline models and SDZRE under different numbers of unknown relations. As the value of m increases (from 5 to 15), the F1 scores of all models gradually decrease on both datasets. Analyzing the reason for this result, we believe that with the increase of m

value, it means that the model needs to identify and distinguish more kinds of relations. How to effectively alleviate this problem is also a worthy direction for future research. In addition, the F1 scores of all models on FewRel dataset are generally higher than those on Wiki-ZSL dataset, indicating that Wiki-ZSL dataset (remotely supervised generation) brings greater challenges to model learning.

4.5.2. Ablation study

To further verify the contribution of SDZRE modules, we have conducted ablation study, and the results are shown in Table 2.

After removing the contrastive learning module (w/o Ctr.) from the model, the performance of the model decreases significantly: 1.83% on the FewRel dataset and 3.00% on the Wiki-ZSL dataset, indicating that the contrastive learning framework we design effectively improves the ability of the mode to recognize similar relations and deal with complex semantics.

After removing the bidirectional semantic distiller (w/o Bid.), the model directly uses the original context representation to make match prediction. We can notice a significant decrease in the performance of the experimental results: 2.37% on the FewRel dataset and 3.99% on the Wiki-ZSL dataset, indicating that the bidirectional semantic distiller plays an important role in weakening the interference of irrelevant features and strengthening relevant features.

After removing the multi-negative sample strategy (w/o Mult.), the model uses only randomly selected negative samples in the matching process. The experimental results show that the performance of the model decreases by 1.75% on the FewRel dataset and 0.29% on the Wiki-ZSL dataset, indicating the importance of adopting multi-negative sample strategy in improving the robustness of the model and the ability to handle complex matching tasks.

After removing the semi-hard negative sample strategy (w/o Semi.), the model uses hard negative samples only by the Top-K method in the

Table 1
Results (%) of REDBM and baseline methods on the FewRel and Wiki-ZSL datasets. m represents the number of unknown relations, and bold represent the best results.

Unseen relation	Method	FewRel			Wiki-ZSL		
		Prec.	Rec.	F1	Prec.	Rec.	F1
m = 5	ZS-BERT (Chen & Te Li, 2021)	76.96	78.86	77.90	71.54	72.39	71.96
	PromptMatch (Sainz et al., 2021)	91.14	90.86	91.00	77.39	75.90	76.63
	RelationPrompt (Chia et al., 2022)	90.15	88.50	89.30	70.66	83.75	76.63
	NoGen (Chia et al., 2022)	72.36	58.61	64.57	51.78	46.76	48.93
	ZRCM (Zhu et al., 2022)	86.70	84.51	85.60	76.15	77.1	76.6
	SUMASK (Li et al., 2023)	78.27	72.55	75.30	75.64	70.96	73.23
	RE-Matching (Zhao et al., 2023)	92.82	92.34	92.58	78.19	78.41	78.30
	AlignRE (Li, Zhang et al., 2024)	93.30	92.90	93.09	83.11	80.30	81.64
	REA (Layegh et al., 2024)	92.57	84.7	88.46	78.88	68.8	73.38
	NSP-RTE (Liao et al., 2024)	87.98	86.06	87.00	87.35	86.01	86.62
	SDZRE	**97.79**	**97.80**	**97.79**	**88.93**	**86.44**	**87.67**
m = 10	ZS-BERT (Chen & Te Li, 2021)	56.92	57.59	57.25	60.51	60.98	60.74
	PromptMatch (Sainz et al., 2021)	83.05	82.55	82.80	71.86	71.14	71.50
	RelationPrompt (Chia et al., 2022)	80.33	79.62	79.96	68.51	74.76	71.50
	NoGen (Chia et al., 2022)	66.47	48.28	55.61	54.87	36.52	43.80
	ZRCM (Zhu et al., 2022)	53.67	53.96	53.81	62.41	64.16	63.27
	SUMASK (Li et al., 2023)	64.77	60.94	62.80	62.31	61.08	61.69
	RE-Matching (Zhao et al., 2023)	83.21	82.64	82.93	74.39	73.54	73.96
	AlignRE (Li, Zhang et al., 2024)	86.41	85.14	86.41	75.00	73.26	74.10
	REA (Layegh et al., 2024)	82.26	79.47	80.85	73.15	61.2	66.64
	NSP-RTE (Liao et al., 2024)	82.59	80.68	81.62	76.05	72.81	74.32
	SDZRE	**91.64**	**91.68**	**91.66**	**81.43**	**80.38**	**80.90**
m = 15	ZS-BERT (Chen & Te Li, 2021)	35.54	38.19	36.82	34.12	34.38	34.25
	PromptMatch (Sainz et al., 2021)	72.83	72.10	72.46	62.13	61.76	61.95
	RelationPrompt (Chia et al., 2022)	74.33	72.51	73.40	63.69	67.93	65.74
	NoGen (Chia et al., 2022)	66.49	40.05	49.38	54.45	29.43	37.45
	ZRCM (Zhu et al., 2022)	40.27	40.72	40.50	33.47	33.47	35.01
	SUMASK (Li et al., 2023)	44.76	41.13	42.87	43.55	40.27	41.85
	RE-Matching (Zhao et al., 2023)	73.80	73.52	73.66	67.31	67.33	67.32
	AlignRE (Li, Zhang et al., 2024)	77.63	77.00	77.31	69.01	67.52	68.26
	REA (Layegh et al., 2024)	64.34	68.68	66.44	58.2	52.6	55.25
	NSP-RTE (Liao et al., 2024)	76.24	74.83	75.51	69.85	68.71	69.25
	SDZRE	**80.68**	**79.46**	**80.06**	**69.60**	**70.30**	**69.95**

matching process. The experimental results show that the performance of the model decreases by 1.17% on the FewRel dataset and 0.39% on the Wiki-ZSL dataset, indicating the importance of using Top-K method to select effective negative samples and these negative samples in improving the generalization ability of the model.

The results of the ablation study show that all of our proposed methods play an important role, which underscores the effectiveness of the overall framework we designed to enhance the model to handle complex tasks and improve accuracy.

Table 2
Ablation study (%) of our method, we uniformly choose the experimental results when m = 10.

Dataset	Method	Prec.	Rec.	F1
FewRel	w/o Ctr.	90.00	87.65	89.83
	w/o Bid.	89.42	89.16	89.29
	w/o Mult.	90.08	89.75	89.91
	w/o Semi.	90.50	90.49	90.49
	Our	**91.64**	**91.68**	**91.66**
Wiki-ZSL	w/o Ctr.	79.47	76.40	77.90
	w/o Bid.	79.48	79.61	76.91
	w/o Mult.	81.97	79.29	80.61
	w/o Semi.	81.30	79.74	80.51
	Our	**81.43**	**80.38**	**80.90**

4.6. Qualitative analysis

4.6.1. Combination of different data augmentation strategies

To demonstrate the impact of different combinations of data augmentation strategies, we experimentally observe the experimental results under the pairwise combinations of Dropout, random mask, and feature cutoff, and the results are shown in Table 3. It can be seen that the combination of random mask and feature cutoff performs the best among all strategy combinations. We believe that this is due to the combination can generate more representative and robust features in contrastive learning, thereby improving the performance of the model on different datasets. Especially when dealing with complex and noisy data sets, the combination shows the potential to significantly enhance model generalization. In this paper, the combination of random mask and feature cutoff is selected for data enhancement.

4.6.2. Visualization of relation representation

In order to further observe how our method learns better relation representations, we randomly select 5 types of relations as unknown relations in FewRel dataset, and use t-SNE (Van der maaten & hinton, 2008) to reduce dimension visualization of the representation of unknown relations. The results are shown in Fig. 8. As can be seen from Fig. 8(a), the data points are relatively mixed and dense, especially for blue and yellow data points, indicating that there are similar relations in these instances, and RE-Matching fails to distinguish these relations effectively. However, as shown in Fig. 8(b), the data points of the same color are relatively concentrated and the distribution among different categories is scattered. We believe that by introducing contrastive learning combined with bidirectional semantic distiller, SDZRE focuses more on core semantic features and generates diverse representations through different strategies. This enables more effective differentiation between instances and helps mitigate the interference caused by similar relations.

4.6.3. Different context semantic distillation methods

To further observe the influence of the contextual semantic distillation method, we split the bidirectional semantic distiller into a fusion projection distiller and a Gaussian adaptive selector, while keeping other conditions unchanged. We conduct comparative experiments,

Table 3
Comparison of F1 scores of different combinations of data augmentation strategies (m = 10).

Data augmentation strategy	FewRel	Wiki-ZSL
Dropout+ Random mask	90.30	80.56
Dropout+ Feature cutoff	90.20	80.38
Random mask+ Feature cutoff	**91.66**	**80.90**

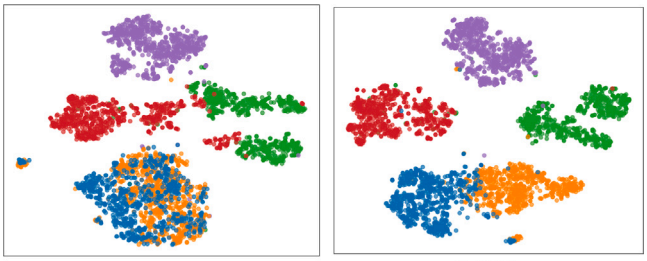


Fig. 8. Visualization of relation representation on FewRel datasets. The figure compares the clustering effect of RE-Matching and SDZRE methods, with different colors representing different categories.

Table 4
Performance of SDZRE under different contextual semantic distillation methods (FewRel, m = 15).

Method	Prec.	Rec.	F1
Fusion projection distiller	80.13	78.53	79.32
Gaussian adaptive selector	79.61	78.49	79.05
Both	**81.75**	**79.46**	**80.06**

and the results are shown in Table 4. It can be seen that although each module can bring some improvement separately, their synergy is more significant. By analyzing the reasons, we believe that the combination of fusion projection distiller and Gaussian adaptive selector can not only complement each other in model performance, but also improve context awareness and information extraction capabilities through architecture optimization, so as to achieve the best experimental results.

4.6.4. Inference efficiency analysis

To further analyze the efficiency of the proposed method, we compare the inference time and the corresponding F1 score for m = 10 on two datasets FewRel and Wiki-ZSL, and the results are shown in Fig. 9. It can be seen that SDZRE significantly improves F1 scores on both datasets, while simultaneously achieving faster inference speeds. We believe that this is primarily due to the fine-grained semantic matching mechanism, which enables the model to more accurately identify relations, thereby reducing unnecessary computation steps and improving inference efficiency. Additionally, the designed semantic distiller effectively eliminates irrelevant features, allowing the model to focus more on the most relevant information for relation matching, further reducing computational complexity and enhancing inference efficiency. Although SDZRE incurs some additional inference time compared to RE-Matching and AlignRE, its overall performance advantage remains significant. Our analysis suggests that this is mainly because the introduction of contrastive learning and the multi-negative sample strategy, while enhancing the generalization ability of model, also increases computational load and inference time to some extent.

We conducted a complexity comparison analysis of SDZRE and the previous advanced methods, AlignRE, NSP-RTE and RE-Matching. Table 5. shows the resource and training time consumption of the three methods under the same experimental conditions. We can observe that GPU memory usage is roughly comparable across the methods, with no significant resource wastage. however, our method demonstrates

Table 5
Complexity Analysis of SDZRE (FewRel m = 15).

Method	FewRel (m = 15)			
	GPU	CPU	Parameter quantity	Train time
RE-Matching	1.76 MB	1.4%	111.30 million	**10.52 min**
AlignRE	**1.73 MB**	1.2%	110.48 million	12.02 min
NSP-RTE	1.83 MB	1.4%	**108.50 million**	12.34 min
SDZRE	1.79 MB	**1.1%**	113.66 million	13.54 min

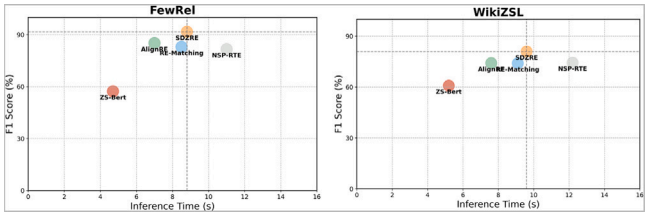


Fig. 9. Performance and inference speed of different methods on different data sets (m = 10).

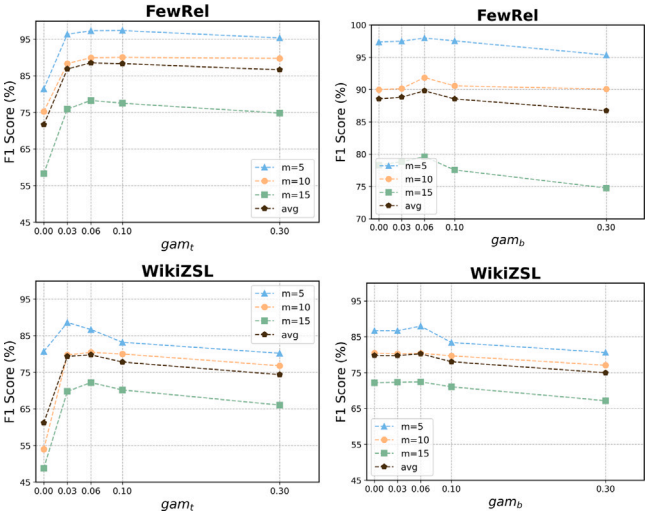


Fig. 10. Effect of different boundary parameters on model performance.

greater efficiency in CPU resource utilization. Although SDZRE incorporates modules such as a bidirectional semantic distiller and contrastive learning, which results in a slightly higher number of parameters compared to the other methods, which does not significantly increase the overall computational burden. Overall, our method is more resource efficient. Due to the inclusion of mechanisms like contrastive learning and a multi-negative sample strategy, the training time of SDZRE is slightly longer than that of the other methods, but the difference is not substantial, making its training time consumption acceptable.

Overall, compared to previous advanced methods, SDZRE significantly enhances model performance by introducing more complex mechanisms, while maintaining efficient inference and lower computational resource consumption. This ultimately achieves the best performance to date.

4.6.5. Hyper-parameter analysis

We also experimentally observe the effect of hyper-parameters on model performance, especially for two key boundary parameters. gam_t and gam_b are the key parameters in the marginal ranking loss function, which are used to control the minimum value of the similarity gap between positive and negative samples, so as to ensure that the model can effectively distinguish between positive and negative samples, so as to improve the discrimination ability and robustness of the model.

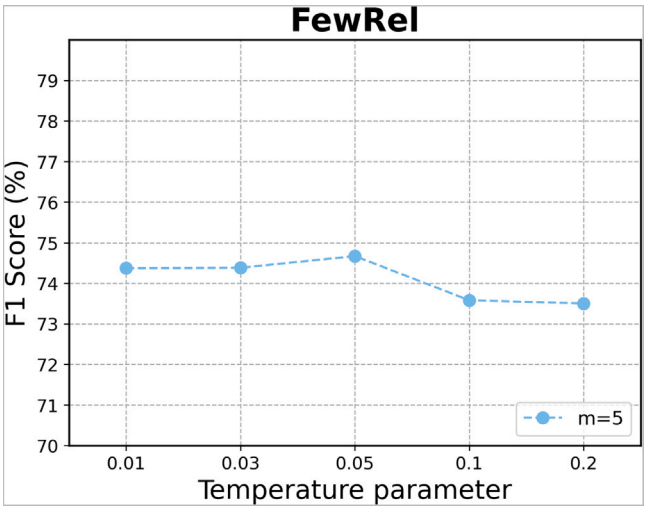


Fig. 11. The impact of temperature parameter τ on model performance.

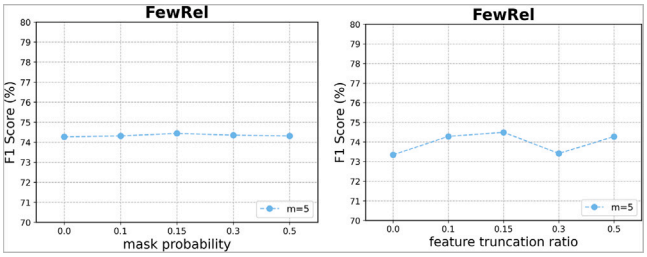


Fig. 12. The impact of data augmentation parameters on model performance.

In the experiment, we observe the effect of hyper-parameters gam_t and gam_b on the performance of the model under the condition of a fixed selection process. To ensure the comparability of the experiments, gam_b is set to 0.00 by default in the left charts of Fig. 10, while the optimal value of gam_t in the right charts is fixed. As can be seen from Fig. 10, the performance curves on the two data sets show a similar trend. As the hyper-parameter increases from 0 to 0.06, the model performance begins to decline, indicating that the optimal values of the two hyper-parameters on different data sets are the same. In addition, model matching does not break down even if gam_t and gam_b keep increasing, showing the good robustness of our method.

In our experiments, under the condition of a fixed selection process, we observed the impact of the temperature parameter τ on model performance. As shown in Fig. 11, when τ is too small (e.g., 0.01) or too large (e.g., 0.2), the model performs worse on the FewRel dataset. Our analysis suggests that when τ is too small, the similarity differences are amplified, making the model less sensitive to noise. Conversely, when τ is too large, the similarity distribution becomes overly smooth, weakening the ability of model to distinguish negative samples. Based on our experiments, we selected $\tau = 0.05$ as the optimal value.

Under the condition of a fixed selection process, we observed the impact of data augmentation parameters on model performance. As shown in Fig. 12, as the masking probability and truncation ratio increase, the performance of model on the FewRel dataset generally follows a trend of initial improvement followed by decline. Our analysis suggests that a higher masking probability may excessively disrupt the core semantics, while an increased truncation ratio may lead to the loss of critical features to some extent. Therefore, based on our experiments, we selected appropriate masking probabilities and truncation ratios to effectively control the augmentation strength.

5. Conclusions

This paper proposed a semantic distillation method SDZRE for ZeroSRE task. This method used a bidirectional semantic distiller and a contrastive learning strategy to construct a semantic distillation framework, which highlighted important semantic relations by strengthening major features and weakening irrelevant features. It amplified the subtle differences between similar relations, so as to effectively deal with complex context interference. A fine-grained semantic matching method was used to introduce the multi-negative sample strategy including hard negative samples and semi-hard negative samples based on similarity, which further improved the generalization ability of the model to different semantic boundaries. The experimental results showed that SDZRE could achieve significantly higher SOTA results than the current methods, while maintaining faster inference speed and achieving double improvement of performance and efficiency.

Although our method significantly have improved the performance of the ZeroSRE task, it still suffers from noise data on the Wiki-ZSL dataset. Through manual inspection of error cases, we found that some of the noise in Wiki-ZSL originates from entity linking errors in distant supervision, which causes the model to confuse entities with contextual information, thereby affecting the accuracy of relation extraction. Additionally, although we have optimized the inference speed, the issue of decreased inference speed still persists as the number of unknown relations increases. This is a common bottleneck faced in the current zero-shot relation extraction field. Our future work will focus on addressing these two issues.

CRediT authorship contribution statement

Yuanjie Zhou: Conceptualization, Data curation, Formal Analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Ling Lu:** Conceptualization, Funding acquistion, Supervision, Validation, Writing – review & editing. **Hengguang Li:** Writing – review & editing. **Xiaoyang Liu:** Writing – review & editing. **Dan Huang:** Writing – review & editing. **Yinong Chen:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported in part by Key Project of Chongqing Municipal Education Commission (KJZD-K202401101, 23SKGH247).

Appendix A. Supplementary data

Supplementary data will be made available on request.

Data availability

Data will be made available on request.

References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). Large language models are zero-shot clinical information extractors. ArXiv preprint arXiv:2205.12689.

Chen, J., Geng, Y., Chen, Z., Pan, J. Z., He, Y., Zhang, W., Horrocks, I., & Chen, H. (2023). Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey. *Proceedings of the IEEE*, *111*(6), 653–685.

Chen, S., & Li, Z. (2024). Hierarchically coupled view-crossing contrastive learning for knowledge enhanced recommendation. *IEEE Access*.

Chen, C. Y., & Te Li, C. (2021). ZS-BERT: towards zero-shot relation extraction with attribute representation learning. In *2021 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2021* (pp. 3470–3479). Association for Computational Linguistics (ACL).

Chia, Y. K., Bing, L., Poria, S., & Si, L. (2022). RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the association for computational linguistics: ACL 2022* (pp. 45–57).

Devlin, M.-W., Chang, K., Lee, K., & Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning* (pp. 1180–1189). PMLR.

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP 2021-2021 conference on empirical methods in natural language processing, proceedings* (pp. 6894–6910).

Giorgi, J., Nitski, O., Wang, B., & Bader, G. (2021). Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 879–895).

Gong, J., & Eldardiry, H. (2024). Prompt-based zero-shot relation extraction with semantic knowledge augmentation. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-cOLING 2024)* (pp. 13143–13156).

Guo, Q., Guo, Y., & Zhao, J. (2024). KBPT: knowledge-based prompt tuning for zero-shot relation triplet extraction. *PeerJ Computer Science*, *10*, Article e2014.

Han, J., Cheng, B., & Lu, W. (2021). Exploring task difficulty for few-shot relation extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 2605–2616).

Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., & Sun, M. (2018). FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4803–4809).

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729–9738).

He, K., Huang, Y., Mao, R., Gong, T., Li, C., & Cambria, E. (2023). Virtual prompt pre-training for prototype-based few-shot relation extraction. *Expert Systems with Applications*, *213*, Article 118927.

Hinton, G. (2012). Improving neural networks by preventing co-adaptation of feature detectors. ArXiv preprint arXiv:1207.0580.

Huang, J., Gu, S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (2023). Large language models can self-improve. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 1051–1068).

Huang, Q., Xia, C., Wu, C., Li, S., Wang, Y., Song, Y., & Kuo, C.-C. J. (2017). Semantic segmentation with reverse attention. ArXiv preprint arXiv:1707.06426.

Ioannides, G., Chadha, A., & Elkins, A. (2024). Gaussian adaptive attention is all you need: Robust contextual representations across multiple modalities. ArXiv preprint arXiv:2401.11143.

Layegh, A., Payberah, A. H., & Matskin, M. (2024). REA: Refine-estimate-answer prompting for zero-shot relation extraction. In *International conference on applications of natural language to information systems* (pp. 301–316). Springer.

Levy, O., Seo, M., Choi, E., & Zettlemoyer, L. (2017). Zero-shot relation extraction via reading comprehension. In *21st conference on computational natural language learning* (pp. 333–342). Association for Computational Linguistics (ACL).

Li, S., Bai, G., Zhang, Z., Liu, Y., Lu, C., Guo, D., Liu, R., & Yong, S. (2024). Fusion makes perfection: An efficient multi-frained matching approach for zero-shot relation extraction. In *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 2: short papers)* (pp. 79–85).

Li, G., Wang, P., & Ke, W. (2023). Revisiting large language models as zero-shot relation extractors. ArXiv preprint arXiv:2310.05028.

Li, Z., Zhang, F., & Cheng, J. (2024). AlignRE: An encoding and semantic alignment approach for zero-shot relation extraction. In *Findings of the association for computational linguistics ACL 2024* (pp. 2957–2966).

Liao, W., Liu, Z., Zhang, Y., Huang, X., Liu, N., Liu, T., Li, Q., Li, X., & Cai, H. (2024). Zero-shot relation triplet extraction as next-sentence prediction. *Knowledge-Based Systems*, *304*, Article 112507.

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. ArXiv preprint arXiv:1703.03130.

Liu, F., Lin, H., Han, X., Cao, B., & Sun, L. (2022). Pre-training to match for unified low-shot relation extraction. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 5785–5795).

Lv, B., Liu, X., Dai, S., Liu, N., Yang, F., Luo, P., & Yu, Y. (2023). DSP: discriminative soft prompts for zero-shot entity and relation extraction. In *Findings of the association for computational linguistics: ACL 2023* (pp. 5491–5505).

Nguyen, K., Nguyen, Y., & Le, B. (2021). Semi-supervising learning, transfer learning, and knowledge distillation with simclr. ArXiv preprint arXiv:2108.00587.

Obamuyide, A., & Vlachos, A. (2018). Zero-shot relation classification as textual entailment. In *Proceedings of the first workshop on fact extraction and vERification* (pp. 72–78).

Reimers, I., & Gurevych (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. 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)* (pp. 3980–3990).

Sainz, O., de Lacalle, O. L., Labaka, G., Barrena, A., & Agirre, E. (2021). Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1199–1212).

Shen, D., Zheng, M., Shen, Y., Qu, Y., & Chen, W. (2020). A simple but tough-to-beat data augmentation approach for natural language understanding and generation. ArXiv preprint arXiv:2009.13818.

Soares, L. B., Fitzgerald, N., Ling, J., & Kwiatkowski, T. (2019). Matching the blanks: distributional similarity for relation learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2895–2905).

Tian, L., Cheng, Y., & Li, Z. (2022). Pseudo random masked autoencoder for self-supervised learning. In *Proceedings of the 2022 6th international conference on video and image processing* (pp. 140–143).

Van der maaten, l., & hinton, g. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Wang, S., Duan, B., Wu, Y., & Xu, Y. (2022). Learning discriminative representations for open relation extraction with instance ranking and label calibration. In *Findings of the association for computational linguistics: NAACL 2022* (pp. 2433–2438).

Wang, S., Zhang, B., Xu, Y., Wu, Y., & Xiao, B. (2022). RCL: Relation contrastive learning for zero-shot relation extraction. In *Findings of the association for computational linguistics: NAACL 2022* (pp. 2456–2468).

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1–37.

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., et al. (2023). Chatie: Zero-shot information extraction via chatting with chatgpt. ArXiv preprint arXiv:2302.10205.

Xu, J., Zhanyi, C. S., Xu, L., & Chen, L. (2024). BlendCSE: Blend contrastive learnings for sentence embeddings with rich semantics and transferability. *Expert Systems with Applications*, *238*, Article 121909.

Yan, Y., Li, R., Wang, S., Zhang, F., Wu, W., & Xu, W. (2021). ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*. Association for Computational Linguistics.

Yang, Z., Fei, J., Tan, Z., Tang, J., & Zhao, X. (2024). CL&CD: Contrastive learning and cluster description for zero-shot relation extraction. *Knowledge-Based Systems*, *293*, Article 111652.

Zhang, K., Gutierrez, B. J., & Su, Y. (2023). Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *The 61st annual meeting of the association for computational linguistics*.

Zhang, H., Zhang, X., Huang, H., & Yu, L. (2022). Prompt-based meta-learning for few-shot text classification. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 1342–1357).

Zhang, Z., Zhang, H., Wan, Q., & Liu, J. (2024). Entity–relation triple extraction based on relation sequence information. *Expert Systems with Applications*, *238*, Article 121561.

Zhao, X., Deng, Y., Yang, M., Wang, L., Zhang, R., Cheng, H., Lam, W., Shen, Y., & Xu, R. (2024). A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, *56*(11), 1–39.

Zhao, J., Zhan, W., Zhao, X., Zhang, Q., Gui, T., Wei, Z., Wang, J., Peng, M., & Sun, M. (2023). RE-matching: A fine-grained semantic matching method for zero-shot relation extraction. In *The 61st annual meeting of the association for computational linguistics*.

Zheng, H., Wen, R., Chen, X., Yang, Y., Zhang, Y., Zhang, Z., Zhang, N., Qin, B., Xu, M., & Zheng, Y. (2021). PRGC: Potential relation and global correspondence based joint relational triple extraction. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (ACL/IJCNLP)* (pp. 6225–6235).

Zhou, S., Meng, Y., Jin, B., & Han, J. (2024). Grasping the essentials: tailoring large language models for zero-shot relation extraction. ArXiv preprint arXiv:2402.11142.

Zhu, H., Zeng, J., Yang, Y., & Wu, Y. (2022). A zero-shot relation extraction approach based on contrast learning. In *SEKE* (pp. 293–299).