CETA: A Consensus Enhanced Training Approach for Denoising in Distantly Supervised Relation Extraction

Ruri Liu[1]*, **Shasha Mo**[1]†, **Jianwei Niu**[2], **Shengda Fan**[1]
[1] School of Cyber Science and Technology, Beihang University, Beijing 100191, China
[2] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{ruriliu, moshasha, niujianwei, fanshengda}@buaa.edu.cn

Abstract

Distantly supervised relation extraction aims to extract relational facts from texts but suffers from noisy instances. Existing methods usually select reliable sentences that rely on potential noisy labels, resulting in wrongly selecting many noisy training instances or underutilizing a large amount of valuable training data. This paper proposes a sentence-level DSRE method beyond typical instance selection approaches by preventing samples from falling into the wrong classification space on the feature space. Specifically, a theorem for denoising and the corresponding implementation, named Consensus Enhanced Training Approach (CETA), are proposed in this paper. By training the model with CETA, samples of different classes are separated, and samples of the same class are closely clustered in the feature space. Thus the model can easily establish the robust classification boundary to prevent noisy labels from biasing wrongly labeled samples into the wrong classification space. This process is achieved by enhancing the classification consensus between two discrepant classifiers and does not depend on any potential noisy labels, thus avoiding the above two limitations. Extensive experiments on widely-used benchmarks have demonstrated that CETA significantly outperforms the previous methods and achieves new state-of-the-art results.

1 Introduction

Relation Extraction (RE), which aims to identify the relation between two specific entities in the text, is a fundamental task in natural language processing. Most supervised RE methods demand large-scale labeled training data, which is difficult to acquire manually. To alleviate the problem, Distant Supervision (DS) is proposed by (Mintz et al., 2009) to automatically generate the labeled text corpus by aligning the plain texts and knowledge bases. For example, as shown in Figure 1, $[Steve\ jobs, born, America]$ is a relational triple in the knowledge base, DS annotates all sentences that contain the entity pair $(Steve\ jobs, America)$ as valid instances for relation $born$. However, DS inevitably introduces noisy labels when the sentences do not express the labeled relation (e.g., cases $S2$ and $S3$ in Figure 1). Hence, investigating a denoising method against noisy labels has become an urgent demand for Distantly Supervised Relation Extraction (DSRE), which aims to train the unbiased RE model under DS-built dataset.

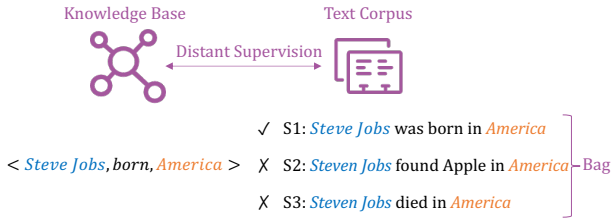

Figure 1: An example of annotating text corpus by distant supervision. $S2$ and $S3$ do not express the relation $born$ but still considered valid instances.

To alleviate the noise issue in DSRE, existing studies can be broadly classified into bag-level methods and sentence-level methods. The bag-level methods (Lin et al., 2016; Hu et al., 2019; Alt et al., 2019; Yuan et al., 2019) typically relax the relation label for each sentence to a bag, and then train the model by employing reliable bag-level representations. However, for bag-level methods, Feng et al. (2018); Jia et al. (2019) empirically verify that they cannot map each sentence to an explicit sentence label, resulting in inefficiency for sentence-level relation classification. From this perspective, several studies focus on sentence-level DSRE, which aims to select the reliable sentences for training and regard the sentence as a basic testing unit. Existing methods (Feng et al., 2018; Qin et al., 2018b; Han et al., 2018b; Zeng et al., 2018)

*The first two authors contributed equally.
†Corresponding authors.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2247–2258
October 12–17, 2022.

usually apply reinforcement learning or adversarial learning to train sentence selector by receiving feedback from the manually crafted reward function, or train the model to start with the reliable sentences selected by frequent patterns (Jia et al., 2019). Finally, these methods select trustable sentences whose predicted labels are consistent with DS-annotated labels. However, these methods may be trapped by some common noisy instances whose model-predicted labels and DS-annotated labels are both wrong (Li et al., 2020b). Besides, the patterns of many correct sentences do not match the frequent patterns, resulting in much valuable information being discarded, limiting the capability of the trained model.

This work proposes a sentence-level DSRE method beyond typical instance selection approaches by preventing samples from falling into the wrong classification space on the feature space. Specifically, a theorem for denoising and the corresponding implementation, named Consensus Enhanced Training Approach (CETA), are proposed in this paper. By training the model with CETA, samples of different classes are separated, and samples of the same class are closely clustered in the feature space. As a result, the robust classification boundary can be easily established to prevent noisy labels from biasing samples into the wrong classification space. Compared with existing sentence-level DSRE methods, CETA performs denoising by enhancing the classification consensus between two discrepant classifiers within the model and does not depend on any potential noisy labels. Therefore, when dealing with noisy labels, CETA does not get trapped by common noisy instances. In addition, CETA enables the model to be trained on all data, and the effect of noisy labels is eliminated in the feature space instead of directly filtering sentences as in previous methods.

Contributions of this paper can be summarized as follows:

- This paper proposes and proves a theorem for denoising that enhancing the prediction consistency between two different classifiers in a model can reduce the impact of noisy instances.
- With the support of the proposed theorem, our proposed CETA facilitates the model to separate the samples of the different classes and cluster the samples of the same class. As a result, a robust classification boundary can be established to reduce the impact of noisy labels.
- Evaluations on widely-used datasets of DSRE demonstrate that CETA significantly outperforms the previous state-of-the-art models.

2 Related Work

We discuss two lines of related work as follows.

DSRE. DS is an effective approach to annotate texts, but suffers from the noisy labels. Most existing studies of DSRE are bag-level DSRE methods, which apply multi-instance learning to handle noisy sentences in each bag and train models by exploiting the constructed reliable bag-level representations. These methods usually utilize attention mechanisms to assign small weights to the potential noisy sentences in the bag (Lin et al., 2016; Han et al., 2018c; Alt et al., 2019; Hu et al., 2019; Yuan et al., 2019; Li et al., 2020a), apply adversarial training or reinforcement learning to remove the noisy sentences from the bag. (Zeng et al., 2015; Qin et al., 2018b; Han et al., 2018b; Shang et al., 2020) However, the studies (Feng et al., 2018; Jia et al., 2019) indicate that the bag-level DSRE methods are ineffective for sentence-level prediction.

This paper focus on sentence-level relation extraction. Some recent studies also regard the sentence as a basic training unit and perform denoising by applying reinforcement learning to select reliable instances based on the reward of noisy labels (Feng et al., 2018), building initial reliable sentences based on several manually defined frequent relation patterns (Jia et al., 2019), assigning the complementary labels cooperating with the negative training to filter noisy instances (Ma et al., 2021), and utilizing meta learning to exploit the extra clean reference data (Li et al., 2020b). Different from the previous works, our proposed method does not rely on the noisy labels, frequent relation patterns, and handles the noisy instances in the feature space without any extra clean reference data.

Supervised learning with noisy labels. In both computer vision and natural language processing, many methods have been proposed to train models with noisy labels and these methods can be broadly classified into: robust regularization (Krogh and Hertz, 1991; Müller et al., 2019; Qu et al., 2021; Zhou and Chen, 2021), robust loss function (Zhang and Sabuncu, 2018; Wang et al., 2019a), label reweighting (Chang et al., 2017; Wang et al., 2019b), noise filtering adaption layers (Goldberger and Ben-Reuven, 2016) and sample selection (Han et al., 2018a; Yu et al., 2019; Wei et al., 2020). In particu-

lar, The methods (Zhou and Chen, 2021; Wei et al., 2020) train two or more models simultaneously and regularize their predictions to be similar, which can be considered as consensus enhancement, but are computationally expensive and affected by the number of trained models. Different from these methods, our method achieves denoising by enhancing the consensus of predictions between two discrepant classifiers within a model, which is computationally friendly and guaranteed by a proven theorem.

3 Methodology

In this section, we start with the learning setup and present the objective function in conjunction with our proposed theorem: the generalization error bound for DSRE in subsection 3.1. Then, in subsection 3.2, we will introduce the details of our proposed CETA, which aims to implement the proposed theorem for denoising.

3.1 Generalization Error Bound for DSRE

Formally, the DS-built training set can be denoted as $\mathcal{D}^{cs} = \{(x_i, y_i^{cs})\}_{i=1}^{m} \in (\mathcal{X} \times \mathcal{Y})$, where $x_i \in \mathcal{X}$ and $y_i^{cs} \in \mathcal{Y}$. $\mathcal{X}$ represents the input instances. $\mathcal{Y}$ indicates the class labels. y^{cs} indicates that the label may be a clean label y^c or a noisy label y^s. The relation extraction model based on neural networks is usually composed of a sentence encoder and a classifier. The sentence encoder $g : \mathcal{X} \to \mathcal{Z}$ maps input instances $\mathcal{X}$ into feature space $\mathcal{Z}$. The classifier $f : \mathcal{Z} \to \mathcal{Y}$ establishes the classification boundary in the feature space and maps the features of the instances into labels $\mathcal{Y}$. In order to evaluate the performance of the model, we denote the clean test set as $\mathcal{D}^c = \{(x_i, y_i^c)\}_{i=1}^{n} \in (\mathcal{X} \times \mathcal{Y})$, where $x_i \in \mathcal{X}$ and $y_i^c \in \mathcal{Y}$. We use $\epsilon^c(f)$ to denote the expected error of the model. The goal of DSRE is to reduce $\epsilon^c(f)$ under the DS-built training set.

The supervised methods usually reduce the expected error based on the structural risk minimization whose basic theorem requires the training set and the test set to come from the same distribution. which is unsuitable for DSRE since the DS-built training set cannot be as clean as the test data. In this paper, the basic theorem is extended to DSRE, a generalization error bound to measure the expected error is theoretically proposed as follows[1].

Theorem 1. *Let g be a fixed representation function from $\mathcal{X}$ to $\mathcal{Z}$, $\mathcal{F}$ be the hypothesis class of Vapink Chervonenkis d. If a random sample of size m $\mathcal{Z}^{cs}$ is generated by applying g to a $\mathcal{D}^{cs}$ - i.i.d. for any $\sigma > 0$, with probability $1 - \sigma$, we have the following uniform generalization error bound for any classification functions $f \in \mathcal{F}$,*

$$\epsilon^c(f) \leq \hat{\epsilon^{cs}}(f) + \frac{1}{2} d_{F \triangle \mathcal{F}}(\mathcal{Z}^{cs}) + \lambda, \quad (1)$$

where

$$\hat{\epsilon^{cs}}(f) = \frac{1}{m} \sum_{i=1}^{m} \left| \hat{f}(z_i^{cs}) - y_i^{cs} \right|$$

$$d_{\mathcal{F}\Delta\mathcal{F}}(\mathcal{Z}^{cs}) = 2 \sup_{f', f'' \in \mathcal{F}} \left| \Pr\left[f'(z^{cs}) \neq f''(z^{cs}) \right] \right|$$

$$f^* = arg\min_{f \in \mathcal{F}} \epsilon^c(f) + \epsilon^{cs}(f)$$

$$\lambda = \epsilon^{cs}(f^*) + \epsilon^c(f^*) + \sqrt{\frac{4}{m}\left(d \log \frac{2em}{d} + \log \frac{4}{\delta} \right)}$$

Eq. (1) demonstrates that the expected error $\epsilon^c(f)$ can be bounded by using three terms. The corresponding explanations are as follows.

1. $\hat{\epsilon^{cs}}(f)$ is the empirical error of the DS-built training data $\mathcal{D}^{cs}$.

2. $d_{F \triangle \mathcal{F}}(\mathcal{Z}^{cs})$ is a key novelty of this theorem and can be viewed as the regularization term that represents the upper bound (sup) of the probability (Pr) that two classifiers f' and f'' category the feature z^{cs} into different classes.

3. λ indicates the shared error of the ideal joint hypothesis (f^*). λ is a constant and can be ignored during training stage.

Based on our proposed Theorem 1, a new denoising method is pointed out, that is, the expected error can be reduced by reducing the generalization error bound. In particular, if $d_{F \triangle \mathcal{F}}(\mathcal{Z}^{cs})$ is minimized, the feature z^{cs} will be classified into the same class by f' and f'' with a higher probability, which is equivalent to enhancing the consistency of model predictions by two discrepant classifiers.

3.2 Consensus Enhanced Training Approach

CETA aims to reduce the expected error $\epsilon^c(f)$ by minimizing $d_{F \triangle \mathcal{F}}(\mathcal{Z}^{cs})$ and $\hat{\epsilon^{cs}}(f)$ in the generalization error bound proposed in Theorem 1. We introduce the architecture of CETA and the optimization strategy of CETA in sequence.

[1]The detailed proof can be found in Appendix A.

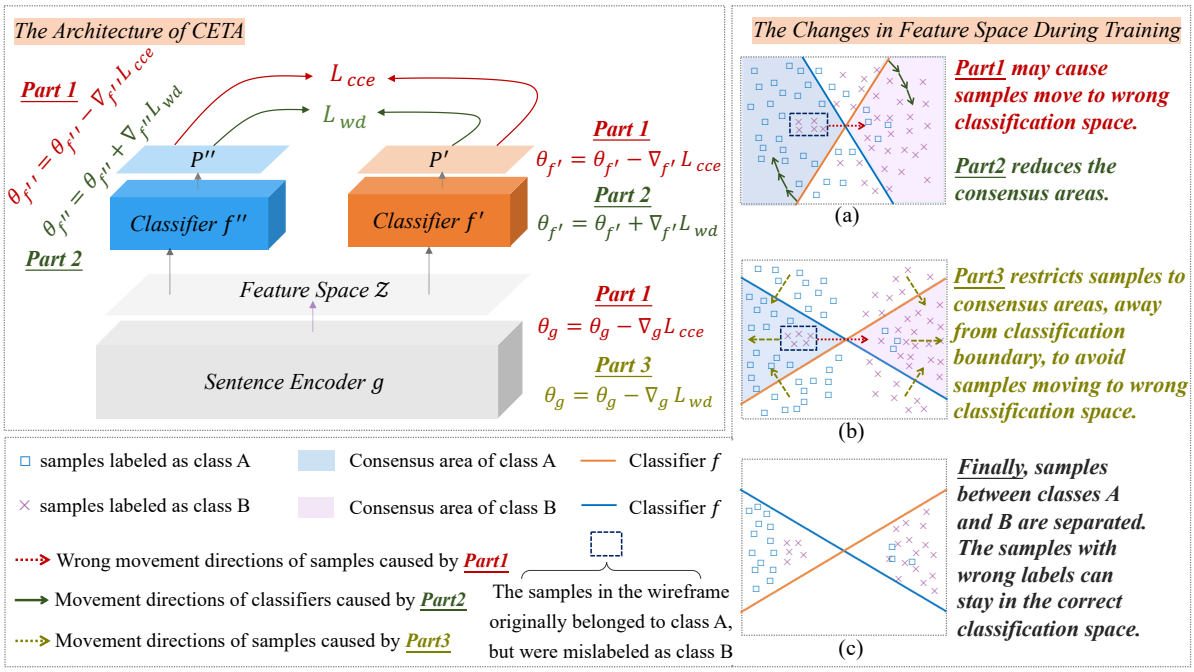

Figure 2: The left part is the architecture and optimization strategy of CETA. The right part is the corresponding changes in the feature space when the optimization algorithm acts on different components. $\mathcal{L}_{cce}$ is the categorical cross-entropy loss function defined in Eq. (2). $\mathcal{L}_{wd}$ is the divergence loss based on wasserstein distance defined in Eq. (3). The symbol Θ indicates the component's parameters of the model. The symbol ∇ denotes the component's gradient calculated by the corresponding loss. The consensus area refers to the area where two classifiers classify the sample into the same class.

.

3.2.1 Architecture of CETA

As shown in Figure 2, the architecture of CETA consists of two classifiers sharing an encoder. Given the input instance x, the encoder g transforms the x from the instance space $\mathcal{X}$ to the feature space $\mathcal{Z}$, the corresponding hidden feature vector is denoted as: $(z_1, z_2, \ldots z_{e_1} \ldots z_{e_2} \ldots z_L)$, where z_{e_1} and z_{e_2} are the feature vectors corresponding to the entities e_1 and e_2. We can obtain the instance representation $z = [z_{e_1}; z_{e_2}]$ for classification by concatenating z_{e_1} and z_{e_2}. In particular, CETA adds an auxiliary classifier, which is not only used to reduce the empirical loss $\hat{\epsilon^{cs}}(f)$, but also aims to use two classifiers to approximate $f^{'}$ and $f^{''}$ in $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$, and then combine the proposed optimization strategy to reduce $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$.

3.2.2 Optimization Strategy of CETA

The optimization strategy of CETA can be broadly divided into three parts. The first part aims to reduce the empirical loss $\hat{\epsilon^{cs}}(f)$. The second part and the third part are combined to reduce the $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$. The details are as follows.

Part 1. To reduce $\hat{\epsilon^{cs}}(f)$, we adopt the categorical cross-etropy function to calculate the classification loss $\mathcal{L}_{cce}$ of the noisy training set.

$$\mathcal{L}_{cce} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K}\mathbb{I}(y_i = k)\left[\log(p'_{ik}) + \log(p''_{ik})\right] \tag{2}$$

where M is the number of training instances, K is the number of relation classes. $\mathbb{I}(y_i = k)$ is an indicator function, which returns 1 when $y_i = k$, and 0 otherwise. The p'_{ik} and p''_{ik} are two probabilities that the instance i belongs to class k predicted by two classifiers f' and f'', respectively. As shown in Figure 2, we use $\mathcal{L}_{cce}$ to calculate the gradient on each component of the model. Then we update the parameters of each component to reduce the empirical error $\hat{\epsilon^{cs}}(f)$. In feature space, reducing $\hat{\epsilon^{cs}}(f)$ is equivalent to forcing the wrongly labeled samples to move to the wrong classification space, and the direction of the movement is indicated by the red arrow in the right part of Figure 2.

Part 2. To reduce $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$, the goal of this part is to increase the discrepancy between the two classifiers, so as to approximate the $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$ with two discrepant classifiers. Specifically, CETA first adopts the wasserstein distance (Kantorovich, 2006) to capture the discrepancy $\mathcal{L}_{wd}$ between $\mathcal{P}'$

and $\mathcal{P}''$ for measuring the classification discrepancy between two classifiers f' and f''. $\mathcal{P}'$ and $\mathcal{P}''$ are two probability distributions representing the probabilities of the samples being divided into different classes by f' and f'', respectively.

$$\mathcal{L}_{wd} = \underset{\gamma^* \in \Pi[\mathcal{P}', \mathcal{P}'']}{argmin} \mathbb{E}_{(p', p'') \sim \gamma} \|p\prime - p\prime\prime\| \quad (3)$$

$$\mathcal{P}' = f'(\mathcal{Z}^{cs}), \mathcal{P}'' = f''(\mathcal{Z}^{cs}) \quad (4)$$

where $\Pi[\mathcal{P}', \mathcal{P}'']$ indicates the set of all jointed distributions whose marginals are $\mathcal{P}'$ and $\mathcal{P}''$. Calculating $\mathcal{L}_{wd}$ can be regarded as finding the optimal $\gamma*$. $\gamma*$ can transform $\mathcal{P}'$ into $\mathcal{P}''$ with minimal modification.

Then, CETA uses $\mathcal{L}_{wd}$ to calculate the gradient $\nabla_{f'}\mathcal{L}_{wd}$, $\nabla_{f''}\mathcal{L}_{wd}$, and $\nabla_g\mathcal{L}_{wd}$ of two classifiers f', f'' and the sentence encoder g, respectively.

$$\theta_{f'} = \theta_{f'} + \nabla_{f'}\mathcal{L}_{wd} \quad (5)$$

$$\theta_{f''} = \theta_{f''} + \nabla_{f''}\mathcal{L}_{wd} \quad (6)$$

Since the $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$ refers to the *upper bound* of the probability that two classifiers divide the sample into different classes. In order to approximate the *upper bound of classification discrepancy* of $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$, we update the parameters $\theta_{f'}$ and $\theta_{f''}$ by executing Eq. (5) and Eq. (6). It can increase the classification discrepancy between the two classifiers. As shown in the right side of Figure 2, it can reduce the consensus area in the feature space. The consensus area refers to the area where two classifiers classify the sample into the same class. The smaller the consensus area, the greater the discrepancy between the classifiers. When the discrepancy between the two classifiers reaches a threshold value, the discrepancy between these two classifiers is approximately equal to $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$.

Part 3. On the basis that the discrepancy between the two classifiers can well approximate $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$, executing Eq. (7) can reduce the discrepancy between the two classifiers, which is equivalent to reducing $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$. As shown in the right side of Figure 2, Eq. (7) is applied to the sentence encoder g, which can change the distribution of samples in the feature space. The samples encoded by g will enter the narrow consensus area and be pulled away from each other, only in this way the classification discrepancy between two classifiers can be reduced. After three parts of the optimization strategy, the samples of different classes are separated in the feature space, and the distance between clusters of different classes is enlarged. So that the wrongly labeled samples cannot be easily moved to the wrong classification space and stay in the original correct classification space.

$$\theta_g = \theta_g - \nabla_g\mathcal{L}_{wd} \quad (7)$$

The complete training steps of CETA are summarized in Algorithm 1. Besides, only sentence encoder g and the classifier f'' are adopted to predict the relation type during the inference procedure.

Algorithm 1 CETA Algorithm

Input: training sets $\mathcal{D}^{cs}$, β, learning rate η, sentence encoder θ_g , full connected layers as classifier $\theta_{f'}$ and $\theta_{f''}$, epoch T, iteration N
Output:$\theta_{f'}$, $\theta_{f''}$, θ_g

for $t = 1, 2, 3, ..., T$ **do**
  Shuffle training set $\mathcal{D}^{cs}$
  for $n = 1, 2, 3, ..., N$ **do**
    Fetch mini-batch $\bar{cs}$ from $\mathcal{D}^{cs}$
    Calculate $\mathcal{L}_{cce}$ and $\mathcal{L}_{wd}$ on $\bar{cs}$
    Update $\theta_{f'} \leftarrow \theta_{f'} - \nabla_{f'}\mathcal{L}_{cce}$
    Update $\theta_{f''} \leftarrow \theta_{f''} - \nabla_{f''}\mathcal{L}_{cce}$
    Update $\theta_g \leftarrow \theta_g - \nabla_g\mathcal{L}_{cce}$
    Update $\theta_{f'} \leftarrow \theta_{f'} + \nabla_{f'}\mathcal{L}_{wd}$
    Update $\theta_{f''} \leftarrow \theta_{f''} + \nabla_{f''}\mathcal{L}_{wd}$
    Update $\theta_g \leftarrow \theta_g - \beta\nabla_g\mathcal{L}_{wd}$
  end for
end for

4 Experiments

The experiments in this work are divided into two part. (1) The first part is the effectiveness study on sentence-level evaluation for our method and the compared methods. Many previous bag-level DSRE methods adopt held-out evaluation, where both training set and test sets are DS-built. However, the studies (Gao et al., 2021; Feng et al., 2018) have demonstrated that the bias is inevitably introduced into held-out evaluations since the DS-build test set is noisy. To provide more accurate and credible evaluations, this part of experiment follows most sentence-level DSRE methods that conduct sentence-level evaluations on benchmarks with clean test sets. (2) The second part is the ablation experiments, which adopts feature visualization to better illustrate the behaviors of our proposed CETA.

Benchmarks		NYT	KBP
#Label num		24	6
Train	Instances	371,461	151,091
	Positive	110,518	38,922
Test	Instances	2,164	4,168
	Positive	323	1,075

Table 1: Statistics of benchmarks. "Positive" means positive instances that are not labeled as "NA". "NA" indicates that the sample does not belong to any of the predefined relation labels.

4.1 Benchmarks

We evaluate our method on two widely-used DSRE benchmarks: NYT and KBP, and the dataset statistics are shown in Table 1.

NYT. This dataset is developed by Riedel et al. (2010) aligning New York Times corpus with the relation facts in Freebase. The origin training set and test set are both DS-built. To make the evaluation more precisely, we adopt the original training set and a widely-used manually annotated test set provided by Jia et al. (2019).

KBP. This dataset is constructed by Ling and Weld (2012) aligning English Wikipedia corpus with the relation facts in Freebase as training set. and the test set is built by utilizing the manually-annotated sentences from 2013 KBP (Ellis et al., 2012). However, some test relation types have no or only one training instance. Besides, this test set only contains 165 positive instances. To reduce the bias of evaluation, we utilize the other refined version of KBP proposed by Li et al. (2020b) to avoid the above problems. Our adopted test set has the same relation types with the training data, contains more positive instances, and keep the same proportion of positive instances as the training set.

4.2 Baseline Models

Our proposed CETA is a sentence-level DSRE method. For the fairness of the comparison, we compare with several strong DSRE methods. These compared methods can be categorized as: bag-level DSRE methods, sentence-level DSRE methods, sentence-level RE methods without denoising.

- **PCNN+ATT** (Lin et al., 2016) A bag-level DSRE method which employs the selective attention to alleviate noise.
- **PCNN+RA_BAG_ATT** (Ye and Ling, 2019) A bag-level DSRE method which utilizes inter-bag and intra-bag attentions to reduce the impact of noisy instances.
- **CNN+RL$_1$** (Qin et al., 2018b) A bag-level DSRE method which applies reinforcement learning to generate the false-positive indicator to recognize false positives, and redistribute the filtered data into the negative examples.
- **CNN+RL$_2$** (Feng et al., 2018) A sentence-level DSRE model which employs reinforcement learning to jointly train a RE model for relation classification and an instance selector for filtering the potential noisy instances.
- **PCNN+DSGAN** (Qin et al., 2018a) A sentence-level DSRE model which adopts adversarial learning to train a generator to recognize true positive instances, and then redistributes the remaining false positives to the negative set to obtain a new cleaned dataset.
- **ARNOR** (Jia et al., 2019) A sentence-level DSRE model which selects the reliable instances based on the reward of attention score on the selected patterns.
- **SENT** (Ma et al., 2021) A sentence-level DSRE model which filters noisy instances and relabeling based on negative training. It is the state-of-the-art method in sentence level.
- **CNN** (Zeng et al., 2014), **PCNN** (Zeng et al., 2015), **BiLSTM** (Zhang et al., 2015) and **BERT** (Devlin et al., 2019) are commonly-used models for RE without denoising methods.

4.3 Implementation Details

Our proposed CETA is a model-agnostic sentence-level DSRE method. We implement CETA using BiLSTM, PCNN, and BERT as sentence encoder, respectively[2].

When implemented with BiLSTM, our adopted word embedding are 50-dimensional Glove word embedding published by Lin et al. (2016). Besides, we utilize 50-dimension randomly initialized position and entity type embedding. The BiLSTM is single layer with hidden size 256 and optimized by Adam optimizer with the learning rate of 5e-4. All the adopted word embedding, position and

[2]The code and training scripts will be released at https://github.com/Ethan-RR/CETA

Model	NYT		
	Prec.	Rec.	F1
CNN*	35.75	64.54	46.01
PCNN*	36.06	64.86	46.35
BiLSTM*	35.52	67.41	46.53
BERT*	36.21	70.41	47.82
PCNN+ATT*	45.41	30.03	36.15
PCNN+RA_BAG_ATT*	56.76	50.60	53.50
CNN+RL_1*	39.41	61.61	48.07
CNN+RL_2*	40.23	63.78	49.34
BiLSTM+ARNOR*	65.23	56.79	60.90
BiLSTM+SENT*	71.22	59.75	64.99
BERT+BiLSTM+SENT*	76.34	63.66	69.42
BiLSTM+CETA	**71.34**	61.12	**65.83**
BERT+CETA	**63.98**	**69.13**	**66.45**
BERT+BiLSTM+CETA	76.29	**64.63**	**69.98**

Table 2: Main results of the sentence-level evaluation on NYT. Compared baselines include normal RE model (the first part of the table) and models for distant RE (the second part of the table). We run our experiment 5 times and report the average result. The results with * are reported in Ma et al. (2021).

Model	KBP		
	Prec.	Rec.	F1
PCNN*	56.12	33.38	41.75
BiLSTM †	57.10	47.06	51.48
PCNN+ATT*	72.65	29.24	41.69
PCNN+RL_1*	57.64	38.79	46.32
PCNN+DSGAN*	59.86	38.54	46.65
BiLSTM+ARNOR*	54.83	34.59	42.35
PCNN+SENT	59.98	32.78	42.39
BiLSTM+CETA	**59.74**	**56.15**	**57.89**
PCNN+CETA	58.72	**39.83**	**47.46**

Table 3: Main results of the sentence-level evaluation on KBP. We run our experiment 5 times and report the average result of our proposed method. The results with * and † are reported in Li et al. (2020b) and Li et al. (2022), respectively.

entity type embedding, the hyperparameters of BiLSTM and the optimizer are consistent with SENT. When implemented with PCNN, the size of position embeddings are 30 dimensions. The number of convolution filter for PCNN model is 230, and the filter window size is 3, which keeps the same with Li et al. (2020b). When implemented with BERT, we use *bert-base-uncased* as sentence encoder and apply AdamW optimizer with a learning rate of 2e-5. The above experimental setup is also applied to the compared RE model that utilizes BERT without denoising method.

We determine the best hyperparameters by grid search. Specifically, when training on the NYT and KBP datasets, we train the model for 10 epochs with a batch size of 256 when using BiLSTM and a batch size of 16 when using BERT. The optimal values of the scalar β for scaling gradients on the sentence encoder mentioned in Algorithm 1 are $\beta = 2.1$ for BiLSTM and $\beta = 4.7$ for BERT.

4.4 Sentence-Level Evaluation

we adopt the same evaluation metrics as the previous sentence-level DSRE method (Jia et al., 2019; Li et al., 2020b; Ma et al., 2021): Micro-Precision (Prec.), Micro-Recall (Rec.) and Micro-F1 (F1). The sentence-level evaluation results of our proposed CETA and other baselines are on NYT and KBP are shown in Table 2 and Table 3, respectively. We can observe that: (1) Our proposed CETA surpasses all baselines in F1 metrics on both KBP and NYT datasets when applying the same sentence encoder. (2) The results on the NYT dataset show that CETA has higher Rec. than both ARNOR and the current state-of-the-art SENT when LSTM is used as the basic sentence encoder. In addition, the results on the KBP dataset show that when PCNN is used as the basic sentence encoder, the Rec. of CETA is higher than that of other baselines, which shows that CETA can facilitate the model to fully exploit the training data. (3) The F1 metrics of ARNOR on the NYT dataset is significantly higher than that of RL_1, but on the KBP dataset, the F1 metrics of ARNOR is lower than that of RL_1. It shows that the performance of ARNOR is susceptible to different data distributions. Our proposed method consistently outperforms on both NYT and KBP datasets. We believe that the stability of the method is important for practical scenarios.

In addition, we conduct a hyperparameter-tuning study about the number of classifiers of CETA on NYT in Appendix B.

4.5 Ablation Study

The section 3.2 has demonstrated that training the model with CETA allows the model to separate

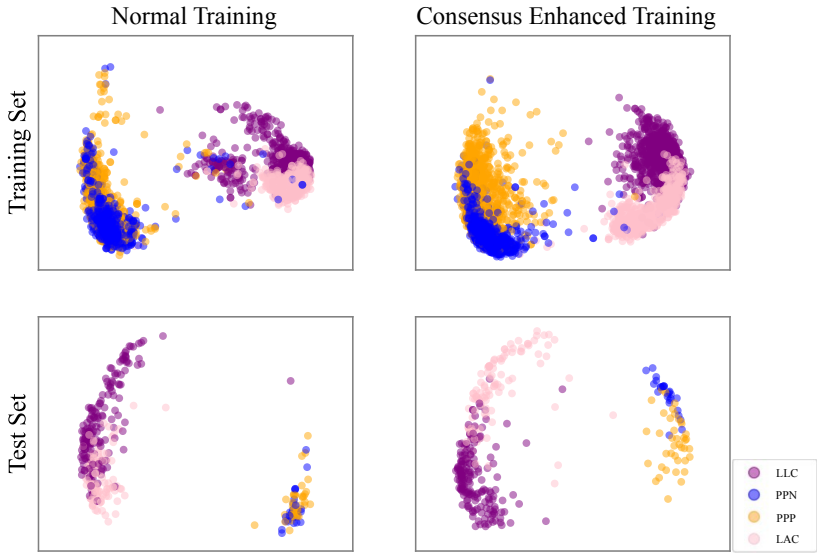

Figure 3: The visualization of instances' representations. The first column and second column show the representations produced by the sentence encoders trained by the *normal training method* and the *consensus enhanced training method*, respectively. The first row and second row show the results of 3750 and 410 random sampled instances from the noisy training set of and the clean test set of our adopted NYT, respectively.

samples between different classes, thereby reducing the possibility of mislabeled samples entering the wrong classification space, making it easier for the model to establish robust classification boundaries. To further verify this proposal, we conduct an ablation experiment that training our proposed CETA with two different training methods on the NYT dataset. The specific steps of this experiment are as follows.

The first training method only reduces the empirical error $\hat{\epsilon^{cs}}(f)$ by performing steps 6, 7, and 8 in our proposed Algorithm 1. We call the first training method *normal training method*. The second training method reduces both the empirical error $\hat{\epsilon^{cs}}(f)$ and the classification- consensus-related term $d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs})$ proposed in Theorem 1 by performing all steps in Algorithm 1. We call the second training method *consensus enhanced training method*. The *normal training method* and the *consensus enhanced training method* adopt the same experimental environment, relation extraction model (We adopt BERT model as sentence encoder), and hyperparameters. Second, we pick four main classes[3] of instances from the noisy training set and the clean test set of NYT. The randomly picked instances are mapped into representations by two sentence encoders that are trained by *normal training method* and *consensus enhanced training method*, respectively. Third, we adopt Principal Component Analysis (PCA) to reduce dimension of the representations and visualize these 2-dimension representations in Figure 3.

From the visualization of instances' representations plotted in Figure 3, we can observe that our proposed *consensus enhanced training method* facilitates model to closely cluster the representations of the same class's samples and clearly separate the representations of different classes' samples compared with the *normal training method*.

In addition, we perform sentence-level evaluation on the clean test set of NYT, and the model trained by *normal training method* achieves the following results: $Prec. = 36.21$, $Rec. = 70.41$, $F1 = 47.82$. The result of the model trained by *consensus enhanced training method* is as follows: $Prec. = 63.98$, $Rec. = 69.13$, $F1 = 66.45$. We can observe that the results obtained by the model trained by *consensus enhanced training method* lead across the board, strongly demonstrating the effectiveness of CETA.

5 Conclusion

This paper goes beyond the typical instance selection approaches, and focuses on handling the noisy labels in the feature space. A theorem for denoising and the corresponding implementation, named Consensus Enhanced Training Approach (CETA),

[3]Four selected classes are: (1) /Location/Location/Contains (LLC), (2) /People/Person/Nationality (PPN), (3) /People/Person/Place lived (PPP) and (4) /Location/Administrative division/Country (LAC), respectively.

are proposed in this paper. By training the model with CETA, samples of different classes are separated in the feature space. Thus the model can easily establish the robust classification boundary to prevent noisy labels from biasing wrongly labeled samples into the wrong classification space. Besides, CETA achieves denoising does not depend on any potential noisy labels. Therefore, CETA is not affected by common noisy instances. Extensive experiments on the widely-used benchmarks have demonstrated that our proposed CETA outperforms previous methods.

Acknowledgements

We thank anonymous reviewers for their responsible attitude and helpful comments. This work was supported by National Natural Science Foundation of China (62106013).

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy. Association for Computational Linguistics.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175.

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joe Ellis, Jeremy Getman, Justin Mott, Xuansong Li, Kira Griffitt, Stephanie Strassel, and Jonathan Wright. 2012. Linguistic resources for 2013 knowledge base population evaluations. *Theory and Applications of Categories*.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Tianyu Gao, Xu Han, Yuzhuo Bai, Keyue Qiu, Zhiyu Xie, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Manual evaluation matters: Reviewing test protocols of distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1306–1318, Online. Association for Computational Linguistics.

Jacob Goldberger and Ehud Ben-Reuven. 2016. Training deep neural-networks using a noise adaptation layer.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018a. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018b. Denoising distant supervision for relation extraction via instance-level adversarial training. *arXiv preprint arXiv:1805.10959*.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018c. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Linmei Hu, Luhao Zhang, Chuan Shi, Liqiang Nie, Weili Guan, and Cheng Yang. 2019. Improving distantly-supervised relation extraction with joint label embedding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3812–3820.

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. Arnor: Attention regularization based noise reduction for distant supervision relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1399–1408.

L. V. Kantorovich. 2006. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4):1381–1382.

Anders Krogh and John Hertz. 1991. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020a. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8269–8276.

Zhenzhen Li, Jian-Yun Nie, Yiping Song, Pan Du, and Dongsheng Li. 2022. Learning to classify relations between entities from noisy data - a meta instance reweighting approach. *Expert Systems with Applications*, 202:117113.

Zhenzhen Li, Jian-Yun Nie, Benyou Wang, Pan Du, Yuhan Zhang, Lixin Zou, and Dongsheng Li. 2020b. Meta-learning for neural relation classification with distant supervision. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 815–824.

Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Yaqian Zhou, and Xuanjing Huang. 2021. Sent: Sentence-level distant relation extraction via negative training. *arXiv preprint arXiv:2106.11566*.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018a. DSGAN: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Melbourne, Australia. Association for Computational Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147.

Yuntao Qu, Shasha Mo, and Jianwei Niu. 2021. Dat: Training deep networks robust to label-noise by matching the feature distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6821–6829.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

B Schölkopf, J. Platt, and T. Hofmann. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*.

Yuming Shang, He-Yan Huang, Xian-Ling Mao, Xin Sun, and Wei Wei. 2020. Are noisy sentences useless for distant supervised relation extraction? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8799–8806.

Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019a. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019b. Crossweigh: Training named entity tagger from imperfect annotations. 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)*, pages 5154–5163.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13726–13735.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *NAACL-HLT (1)*.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR.

Yujin Yuan, Liyuan Liu, Siliang Tang, Zhongfei Zhang, Yueting Zhuang, Shiliang Pu, Fei Wu, and Xiang Ren. 2019. Cross-relation cross-bag attention for distantly-supervised relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 419–426.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Large scaled relation extraction with reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 73–78.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Wenxuan Zhou and Muhao Chen. 2021. Learning from noisy labels for entity-centric information extraction. *arXiv preprint arXiv:2104.08656*.

A Appendix

In this section, we proof our proposed Theorem 1. For ease of reference, we restate Theorem 1.

Theorem 1: Let g be a fixed representation function from $\mathcal{X}$ to $\mathcal{Z}$, $\mathcal{F}$ be the hypothesis class of Vapink Chervonenkis d. If a random sample of size m $\mathcal{Z}^{cs}$ is generated by applying g to a $\mathcal{D}^{cs}$ - i.i.d. for any $\sigma > 0$, with probability $1-\sigma$, we have the following uniform generalization error bound for any feature classification functions $f \in \mathcal{F}$,

$$\epsilon^{c}(f) \leq \hat{\epsilon^{cs}}(f) + \frac{1}{2} d_{F\triangle\mathcal{F}}\left(\mathcal{Z}^{cs}\right) + \lambda \tag{8}$$

where

$$\hat{\epsilon^{cs}}(f) = \frac{1}{m}\sum_{i=1}^{m}\left|\hat{f}(z_i^{cs}) - y_i^{cs}\right| \tag{9}$$

$$d_{\mathcal{F}\Delta\mathcal{F}}(\mathcal{Z}^{cs}) = 2\sup_{f^{'},f^{''}\in\mathcal{F}}\left|\Pr[f^{'}(z^{cs}) \neq f^{''}(z^{cs})]\right| \tag{10}$$

$$f^{*} = arg\min_{f\in\mathcal{F}}\epsilon^{c}(f) + \epsilon^{cs}(f) \tag{11}$$

$$\lambda = \epsilon^{cs}(f^{*}) + \epsilon^{c}(f^{*}) + \sqrt{\frac{4}{m}\left(d\log\frac{2em}{d} + \log\frac{4}{\delta}\right)} \tag{12}$$

Eq. (8) indicates the generalization error bound of $\epsilon^{c}(f)$. It demonstrates that the expected error $\epsilon^{c}(f)$ of the clean test set can be bounded by using three terms ($\hat{\epsilon^{cs}}(f)$, $d_{\mathcal{F}\Delta\mathcal{F}}(\mathcal{Z}^{cs})$ and λ), which are defined in Eq. (9), Eq. (10) and Eq. (12), respectively. The corresponding explanations for these three terms are as follows.

1. $\hat{\epsilon^{cs}}(f)$. This term is the empirical error of $\mathcal{D}^{cs}$.

2. $d_{F\triangle\mathcal{F}}\left(\mathcal{Z}^{cs}\right)$. This term represents the upper bound (sup) of the probability (Pr) that two classification functions $f^{'}$ and $f^{''}$ divide the feature z^{cs} of the same sample into different classes.

3. λ. This term is the shared error of the ideal joint hypothesis (f^{*}) proposed in Eq. (11). This term is a constant.

We begin with the following lemmas to prove the Theorem 1.

Lemma A. *Definition 1: Given two feature distribution $\mathcal{Z}^{s}$ and $\mathcal{Z}^{c}$ extracted by a fixed g, and a hypothesis class $\mathcal{F}$, a set of classifiers. Through a given classifier f, the divergence$\mathcal{F}\Delta\mathcal{F}$ between $\mathcal{Z}^{s}$ and $\mathcal{Z}^{c}$ is:*

$$\begin{aligned} & d_{\mathcal{F}\triangle\mathcal{F}}\left(\mathcal{Z}^{s},\mathcal{Z}^{c}\right) \\ &= 2\sup_{\eta\in\mathcal{F}\Delta\mathcal{F}}\left|\Pr_{z\sim\mathcal{Z}^{c}}[f^{'}(z)\neq f^{''}(z)] - \Pr_{z\sim\mathcal{Z}^{s}}[f^{'}(z)\neq f^{''}(z)]\right| \\ &= 2\sup_{\eta\in\mathcal{F}\Delta\mathcal{F}}\left|\Pr_{z\sim\mathcal{Z}^{s}}[z:\eta(z)=1] - \Pr_{z\sim\mathcal{Z}^{c}}[z:\eta(z)=1]\right| \end{aligned}$$

$$\begin{aligned} & \mathcal{F}\Delta\mathcal{F} = \{\eta : \eta\left(z^{*}\right) = 1\}, \oplus: \text{ XOR operator} \\ & z^{*} = \{z : f_1(z) \oplus f_2(z), f_1, f_2 \in \mathcal{F}\} \end{aligned}$$

where $\mathcal{Z}^{s} \subseteq \mathcal{Z}^{cs}$ and $\mathcal{Z}^{c} \subseteq \mathcal{Z}^{cs}$ are feature distribution of noisy data and clean data, respectively. Lemma A has been proposed and proved in (Ben-David et al., 2010).

Lemma B. *The upper bound of the probability $d_{F\triangle F}(\mathcal{Z}^{cs})$ that two classification functions $f^{'}$ and $f^{''}$ divide the feature z^{cs} of the same sample into different classes is equal to the upper bound of $d_{\mathcal{F}\triangle\mathcal{F}}(\mathcal{Z}^{s},\mathcal{Z}^{c})$.*

Proof. *Now we proof the Lemma B.*

$$\begin{aligned} & d_{\mathcal{F}\triangle\mathcal{F}}\left(\mathcal{Z}^{s},\mathcal{Z}^{c}\right) \\ &= 2\sup_{\eta\in\mathcal{F}\Delta\mathcal{F}}\left|\Pr_{z\sim\mathcal{Z}^{s}}[z:\eta(z)=1] - \Pr_{z\sim\mathcal{Z}^{c}}[z:\eta(z)=1]\right| \\ &\leq 2\sup_{\eta\in\mathcal{F}\Delta\mathcal{F}}\left|\Pr_{z\sim\mathcal{Z}^{s}}[z:\eta(z)=1] + \Pr_{z\sim\mathcal{Z}^{c}}[z:\eta(z)=1]\right| \\ &= 2\sup_{\eta\in\mathcal{F}\Delta\mathcal{F}}\left|\Pr_{z\sim\mathcal{Z}^{cs}}[z:\eta(z)=1]\right| \qquad (13) \\ &= 2\sup_{f^{'},f^{''}\in\mathcal{F}}\left|\Pr_{z\sim\mathcal{Z}^{cs}}[f^{'}(z)\neq f^{''}(z)]\right| \qquad (14) \end{aligned}$$

The Eq. (14) is equal to Eq. (10), which is the $d_{\mathcal{F}\triangle\mathcal{F}}(\mathcal{Z}^{cs})$, thus the $d_{F\triangle F}(\mathcal{Z}^{cs})$ is the upper bound of $d_{\mathcal{F}\triangle\mathcal{F}}(\mathcal{Z}^{s}, \mathcal{Z}^{c})$. Lemma B has been proved.

Proof. *Now we proof the Theorem 1.*

For a classifier f, let $\mathcal{Z}_f \subseteq \mathcal{Z}$ be the feature subset for whose characteristic function is f. The parallel notation $\mathcal{Z}_{f^*}$ and $\mathcal{Z}_f$ are used for classifier f^* and f. Through the feature subset, we make $\Pr_c[\mathcal{Z}_f \triangle \mathcal{Z}_{f^*}] = \Pr_{z\sim \mathrm{Z}^c}[f(z) \neq f^*(z)]$, and the parallel notation $\Pr_{cs}$ is used.

$$\begin{aligned}
\epsilon^c(f) &\leq \epsilon^c(f^*) + \Pr_c[\mathcal{Z}_f \Delta \mathcal{Z}_{f^*}] && (15)\\
&\leq \epsilon^c(f^*) + \Pr_{cs}[\mathcal{Z}_f \triangle \mathcal{Z}_{f^*}] \\
&\quad + |\Pr_c[\mathcal{Z}_f \Delta \mathcal{Z}_{f^*}] - \Pr_{cs}[\mathcal{Z}_f \triangle \mathcal{Z}_{f^*}]| \\
&\leq \epsilon^c(f^*) + \epsilon^{cs}(f^*) + \epsilon^{cs}(f) \\
&\quad + |\Pr_c[\mathcal{Z}_f \Delta \mathcal{Z}_{f^*}] - \Pr_{cs}[\mathcal{Z}_f \Delta \mathcal{Z}_{f^*}]| && (16)\\
&= \epsilon^c(f^*) + \epsilon^{cs}(f^*) + \epsilon^{cs}(f) \\
&\quad + |\Pr_s[\mathcal{Z}_f \Delta \mathcal{Z}_{f^*}]| \\
&\leq \epsilon^c(f^*) + \epsilon^{cs}(f^*) + \epsilon^{cs}(f) \\
&\quad + \sup_{\hat{f}\in\mathcal{F}} |\Pr_s[\mathcal{Z}_f \Delta \mathcal{Z}_{\hat{f}}] + \Pr_c[\mathcal{Z}_f \Delta \mathcal{Z}_{\hat{f}}]| \\
&\leq \epsilon^c(f^*) + \epsilon^{cs}(f) + \frac{1}{2} d_{\mathcal{F}\Delta\mathcal{H}}(\mathcal{Z}^{cs})
\end{aligned}$$

Eq. (15) and Eq. (16) rely on the triangle inequality for classification error (Schölkopf et al.). Besides, according to the standard Vapnik-Chervonenkis theorem (Vapnik, 1999), the $\epsilon^{cs}(f)$ can be bounded by its empirical estimate:

$$\epsilon^{cs}(f) \leq \sqrt{\frac{4}{m}\left(d\log\frac{2em}{d} + \log\frac{4}{\delta}\right)} + \hat{\epsilon^{cs}}(f) \quad (17)$$

in summary:

$$\epsilon^c(f) \leq \hat{\epsilon^{cs}}(f) + \frac{1}{2} d_{F\triangle\mathcal{F}}(\mathcal{Z}^{cs}) + \lambda \quad (18)$$

Theorem 1 has been proved.

B Appendix

In this section, we conduct a hyperparameter tuning study on the number of classifiers for CETA. CETA is designed according to the denoising theorem, which states that the decomposition of the classifier helps to reduce the generalization error. To explore whether more classifiers can further improve the performance, we conduct the experiment of CETA implemented with two, three, and four classifiers on NYT based on BiLSTM, and the results are shown in Table 4.

Model	**NYT**			
	#Classifier	**Prec.**	**Rec.**	**F1**
BiLSTM	1	35.52	67.41	46.53
CETA+BiLSTM	2	71.34	61.12	65.83
CETA+BiLSTM	3	71.46	61.37	66.03
CETA+BiLSTM	4	71.49	61.39	66.05

Table 4: Main results of the sentence-level evaluation on NYT. #Classifier indicates the number of classifiers used to implement CETA.

The results demonstrate that CETA implemented by four classifiers is 0.34% higher than that of the two classifiers in F1 metrics. Besides, CETA implemented by three classifiers is 0.3% higher than that of the two classifiers in F1 metrics, indicating more classifiers can further improve the performance of CETA.