MiDTD: A Simple and Effective Distillation Framework for Distantly Supervised Relation Extraction

RUI LI, CHENG YANG, TINGWEI LI, and SEN SU, Beijing University of Posts and Telecommunications

Relation extraction (RE), an important information extraction task, faced the great challenge brought by limited annotation data. To this end, distant supervision was proposed to automatically label RE data, and thus largely increased the number of annotated instances. Unfortunately, lots of noise relation annotations brought by automatic labeling become a new obstacle. Some recent studies have shown that the teacher-student framework of knowledge distillation can alleviate the interference of noise relation annotations via label softening. Nevertheless, we find that they still suffer from two problems: *propagation of inaccurate dark knowledge* and *constraint of a unified distillation temperature*. In this article, we propose a simple and effective Multi-instance Dynamic Temperature Distillation (MiDTD) framework, which is model-agnostic and mainly involves two modules: multi-instance target fusion (MiTF) and dynamic temperature regulation (DTR). MiTF combines the teacher's predictions for multiple sentences with the same entity pair to amend the inaccurate dark knowledge in each student's target. DTR allocates alterable distillation temperatures to different training instances to enable the softness of most student's targets to be regulated to a moderate range. In experiments, we construct three concrete MiDTD instantiations with BERT, PCNN, and BiLSTM-based RE models, and the distilled students significantly outperform their teachers and the state-of-the-art (SOTA) methods.

CCS Concepts: • **Computing methodologies** → **Information extraction**; **Neural networks;**

Additional Key Words and Phrases: Natural language processing, NLP, knowledge distillation, distant supervision, neural network, multi-instance learning, label softening

ACM Reference format:
Rui Li, Cheng Yang, Tingwei Li, and Sen Su. 2022. MiDTD: A Simple and Effective Distillation Framework for Distantly Supervised Relation Extraction. *ACM Trans. Inf. Syst.* 40, 4, Article 83 (January 2022), 32 pages.
https://doi.org/10.1145/3503917

1 INTRODUCTION

Information extraction, which focuses on automatically extracting structured information from unstructured or semi-structured texts or web pages, is an important technology for many applications [44], such as patent analysis [43], text mining [38, 53], financial analysis [37], and so on. As a

This work is supported by the National Natural Science Foundation of China (No. 62002029) and Innovation Research Group Project of NSFC (No. 61921003).

Authors' address: R. Li, C. Yang (corresponding author), T. Li, and S. Su, School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, 10 Xi Tucheng Road, Haidian District, Beijing 100876, China; emails: {lirui, yangcheng, litingwei2016, susen}@bupt.edu.cn.

1046-8188/2022/01-ART83 $15.00
https://doi.org/10.1145/3503917

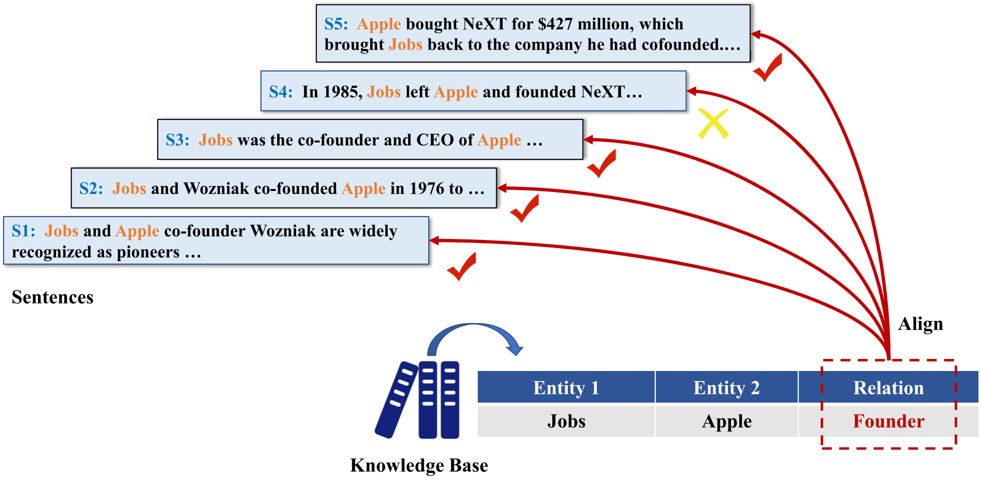

S5: Apple bought NeXT for $427 million, which brought Jobs back to the company he had cofounded.…
S4: In 1985, Jobs left Apple and founded NeXT…
S3: Jobs was the co-founder and CEO of Apple …
S2: Jobs and Wozniak co-founded Apple in 1976 to …
S1: Jobs and Apple co-founder Wozniak are widely recognized as pioneers …
Sentences
Align
Entity 1
Entity 2
Relation
Jobs
Apple
Founder
Knowledge Base

Fig. 1. An illustration of distantly supervised RE. Distant supervision automatically aligns the relation of "Jobs" and "Apple" in sentences S1-S5 with their relation in a knowledge base, i.e., "Founder". But this method will also bring noise instances, e.g., S4.

fundamental information extraction task in the **natural language processing** (**NLP**) and **information retrieval** (**IR**) fields, **relation extraction** (**RE**) aims at identifying the relation between the specified entity pair in a given text [9, 35, 41]. Although the studies of RE have made many breakthroughs under the promotion of deep learning technologies in recent years [16, 17, 69], the development of RE faced the great challenge brought by limited annotation data. To alleviate this problem, distant supervision [40] was proposed to automatically label each instance[1] by aligning the relation of the specified entity pair with their relation in a knowledge base (as illustrated in Figure 1), which largely increased the annotated RE instances and had a broad application prospect [4, 60]. Unfortunately, lots of noise annotations brought by automatic labeling (e.g., S4 in Figure 1) have become a new obstacle.

Recently, various instantiations of knowledge distillation framework [19] have been proposed for deep learning tasks with noisy data including distantly supervised RE by alleviating the interference of noise annotations in virtue of label softening abilities [25, 29, 57, 72]. Label softening can not only weaken the noises[2] [2, 64] but also provide more useful information than hard labels [15, 19].

Specifically, a knowledge distillation framework usually contains a teacher predictor, a student predictor, and a specific distillation approach. The teacher is pre-trained with one-hot hard annotations and outputs a set of probabilistic predictions as the soft targets (also known as "soft labels") of the student. As the link between teacher and student, the distillation approach completely determines the loss function for training the student, which is the core of the entire framework.

Although existing knowledge distillation instantiations for distantly supervised RE have achieved positive results [25, 63], they are based on the conventional distillation approach [19] and suffer from two problems in this task:

[1] In this article, we refer to a sentence and the specified entity pair appearing in it as an instance.

[2] For the one-hot labels annotated wrongly, label softening will reduce the probabilities of the wrong categories and increase those of the right ones.

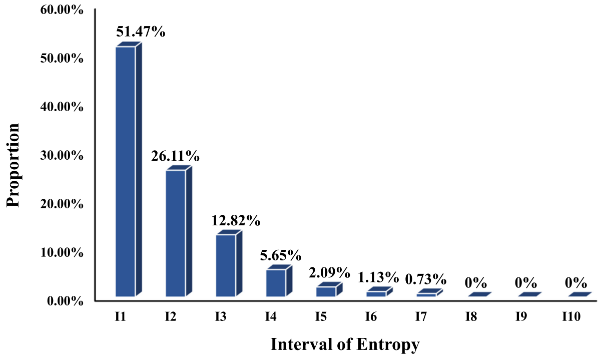


Fig. 2. The distribution of entropies of a PCNN [75]-based teacher model's predictions (temperature = 1) on NYT19-1.0 training set. We train a PCNN RE model on a distantly supervised dataset: NYT19-1.0 (NYT19-1.0 contains 25 relations, 370 k instances in the training set. Please see Section 4.1 for more details.), and plot the distribution of entropies of its probabilistic predictions for the training instances at temperature 1. Here, entropy is the index characterizing the softness: the larger the entropy, the softer the prediction. We evenly split the entire entropy range $[0, \log(25)]$ into 10 sub-intervals and denote them as I1–I10.

Propagation of inaccurate dark knowledge. As mentioned before, in addition to noise reduction, another key function of soft labels generated by the teacher is to provide more useful information than the one-hot clean labels,[3] i.e., dark knowledge. Ref. [19] pointed out that dark knowledge hides in the distribution of probabilities and reveals the similarity between different categories. For example, if the pre-specified relation categories of an RE task contain: "founder", "company", and "NA" three relations, the probabilities of "company" should be closer to those of "founder" rather than "NA" in the soft labels of different instances.[4] Because "founder" and "company" are more similar semantically. However, the training data of our teacher are very noisy, and thus the correctness of dark knowledge in each soft label generated by it will be damaged. The conventional distillation approach is unable to prevent such inaccurate dark knowledge from propagating to the student.

Constraint of a unified distillation temperature. In most cases, the teacher cannot guarantee a suitably soft set of probability distributions predicted by it, so the conventional distillation approach manually raises a unified temperature in the softmax of teacher and student to further soften the teacher's predictions.[5] However, the realistic RE datasets annotated by distant supervision could be very large, a unified temperature cannot adapt to all training instances. Because the teacher's predictions for different training instances are not equally soft at temperature 1. To validate this claim, we measure the softness of teacher's predictions by entropy in Figure 2: at initial temperature 1, the entropy of 51.47% teacher's predictions is in $[0, \frac{\log(25)}{10})$, and the rest 48.53% have larger entropy. Therefore, using a unified temperature cannot regulate the softness of all teacher's predictions to a moderate range, where a part of soft enough predictions (in I2–I7) may be softened excessively to lose important distinct information, or some hard predictions (in I1) cannot be softened sufficiently.

In this article, we propose a novel model-agnostic knowledge distillation framework, dubbed as **Multi-instance Dynamic Temperature Distillation (MiDTD)**, which aims at distilling

[3] We refer to the correct distantly supervised labels as clean labels.
[4] "NA" denotes that the two entities are unrelated, while "founder" and "company" are two sub-relations of business.
[5] The initial temperature of softmax is 1.

high-quality sentence-level RE models and involves two simple and effective modules: **multi-instance target fusion (MiTF)** and **dynamic temperature regulation (DTR)**. MiTF combines the teacher's predictions for multiple sentences within a "bag" to amend the inaccurate dark knowledge in each soft target of the student. In RE task, a bag refers to the set of sentences (instances) containing the same entity pair, and a well-known experience is that the sentences in a bag are likely to express the same relation. Therefore, a natural intuition is that the (hard or soft) labels of most instances in a bag should be consistent, which motivates the design of our MiTF. In fact, the success of distant supervision on RE also verifies this intuition [22, 42, 58]. To break the constraint of a unified temperature, DTR allocates alterable temperatures to the teacher according to the softness of its predictions at initial temperature 1 for different instances, i.e., the harder the predicted distribution of the instance, the greater the temperature assigned to the teacher. In this way, the extreme cases caused by a unified temperature can be effectively reduced and the softness of more teacher's predictions (that are also the student's targets) can be regulated to a moderate range. Furthermore, based on the same assumptions in Reference [19], we theoretically deduce that the distillation approach of MiDTD is more flexible and controllable than the conventional one.

Afterward, we construct three different MiDTD instantiations with BERT [8], PCNN [75], and BiLSTM [77]-based sentence-level RE models for evaluation, where the teachers and students have the same architectures. On two large-scale benchmark datasets: NYT19-1.0 and NYT19-2.0 [20], the students in the three instantiations significantly outperform their teachers, and the distilled BERT-RE model has 2.9% and 2.2% increments in micro-F1 score over the **state-of-the-art** (**SOTA**) method (Section 4). Furthermore, we take the BERT-based MiDTD instantiation as the representative and conduct systematic exploration experiments (Section 5). We find that the performance of the BERT-based MiDTD instantiation on bag-level evaluation is also very satisfactory (Section 5.1). Besides, MiDTD surpasses other label softening methods on both sentence-level and bag-level evaluations (Section 5.2). Finally, we use an elaborate statistical experiment to demonstrate the model distilled by MiDTD do learn more accurate dark knowledge than its teacher (Section 5.4).

Our contributions are as follows:

(1) As far as we know, this work is the first to take the advantage of multi-instance learning to train sentence-level distantly supervised RE models, and is the first to explore the knowledge distillation approach allocating alterable temperatures for different instances in the NLP field.
(2) We propose MiDTD, a model-agnostic knowledge distillation framework for distantly supervised RE tasks, which reduces the propagation of inaccurate dark knowledge and breaks the constraint of unified temperature with two modules: MiTF and DTR.
(3) Under the same assumptions in Reference [19], we deduce that the distillation approach of MiDTD can be converted into a more flexible and controllable form than the conventional approach, which provides theoretical bases for MiTF and DTR.
(4) We construct three MiDTD instantiations with BERT, PCNN, and BiLSTM-based sentence-level RE models on two large-scale distantly supervised datasets: NYT19-1.0 and NYT19-2.0, where the distilled students effectively surpass their teachers and the SOTA methods.

2 RELATED WORK

2.1 Distantly Supervised Relation Extraction

RE is a hot spot in the research on information extraction [6, 7, 24], which has been used for support passage ranking [21], precision medicine [11], access control [59], question answering [23], augmenting knowledge graphs [36], and so on. A body of supervised RE methods was proposed [55, 78], which heavily relied on manually annotated clean data. To reduce the cost of data annotation, Mintz et al. [40] proposed distant supervision to automatically annotate large-scale training data, which inevitably resulted in a mass of noisy annotations. There are mainly three kinds of methods for dealing with such noise annotations.

Firstly, pattern-based methods manually set relation patterns or automatically study relation patterns through neural networks [49, 54]. Takamatsu et al. [62] presented a generative model that directly modeled the pattern of heuristic labeling process of distant supervision and then predicted whether assigned labels are correct or wrong via its hidden variables. Wu et al. [71] introduced a question-answer pattern as an indirect source of supervision for relation extraction, and studied how to use such supervision to reduce noise induced from distant supervision. Feng et al. [12] proposed a model consisting of a top-level sentence selector, a low-level mention extractor, and a reward estimator, so as to extract patterns hidden in representative phrases for a particular relation from noisy sentences that are collected via distant supervision. Qu et al. [47] designed a fine-grained reward function, and modeled the sentence selection pattern as an auction where different relations for a bag needed to compete together to achieve the possession of a specific sentence based on its expressiveness. To capture the complex pattern of relation expression and tighten the correlated features, Qu et al. [48] discovered and used informative correlations between features with a generative method. Jia et al. [20] learned RE patterns from a manually annotated clean dataset, and then iteratively trained an interpretable model to select trustable instances.

Secondly, multi-instance learning methods take into account all sentences in a bag for judgment. Riedel et al. [52] used a factor graph to explicitly model the decisions: (i) whether two entities are related, (2) whether this relation is mentioned in a given sentence second, and then applied constraint-driven semi-supervision to train this model without any knowledge about which sentences express the relations in the knowledge base. Zeng et al. [75] used a piecewise Max pooling after a convolutional layer to extract relation features and proposed an objective function at the bag level, which can take into account the uncertainty of the label correctness in training. Lin et al. [31] developed a sentence-level attention mechanism over multiple instances to reduce the influence of wrong labelled instances. Vashishth et al. [66] employed graph convolution networks to encode syntactic information from text, and introduced the extra side information obtained by Open Information Extraction methods to improve the performance. Alt et al. [1] was based on the pre-trained language model GPT-1 [50], which was also demonstrated to possess rich general semantic knowledge like BERT. Li et al. [26] additionally used entity descriptions crawled from Wikipedia together with a knowledge base to construct an accurate dataset, and trained the RE model using a generative adversarial network framework. Li et al. [28] proposed a bag-level collaborating relation-augmented attention mechanism to handle both the wrong labeling and long-tail relations. Li et al. [27] used an entity-aware word embedding method to integrate relative position information and head/tail entity embeddings, and developed a self-attention mechanism to capture relation features. However, these methods were used to train bag-level RE models, which predicted a common relation for all sentences with the same entity pair. (We will introduce more about bag-level RE models in Section 3.1.) Feng et al. [13] demonstrated that the bag-level RE models often failed to perform well on sentence-level evaluation, and we also observe a similar phenomenon (Section 5.2). In our work, we propose MiTF (Section 3.3) to absorb the benefits of multi-instance learning to train sentence-level RE models and achieve satisfactory results. Besides, we also experimentally demonstrate that the sentence-level models' advantages will not be weakened in bag-level evaluation (Section 5.1).

The third research line relies on annotation recasting, which aims at generating a new set of better annotation distributions [30]. Qin et al. [45] introduced an adversarial learning framework to learn a sentence-level true-positive generator, and the optimal generator was obtained until the discrimination ability of the discriminator had the greatest decline. Qin et al. [46] explored a deep reinforcement learning strategy to generate a false-positive indicator, which automatically recognized false positives for each relation type without supervised information. Zeng et al. [76] firstly extracted the relation of every sentence independently after given a bag, and predicted the bag relation based on the sentence relations and compared it with the gold bag relation, then used the result of the comparison to guide the training of relation extractor. Yang et al. [73] firstly proposed an instance discriminator with reinforcement learning to split the noisy data into correctly labeled data and incorrectly labeled data, and then learned a robust relation classifier in a semi-supervised learning way. The above studies [45, 46, 73, 76] all resist the noises of distant supervision through selecting or redistributing the labels, but they are too "radical", and it is possible to leave the wrong annotations and remove the correct ones. In contrast, the label softening method is more "mild". Liu et al. [32] introduced an entity-pair level denoise method that exploited semantic information from correctly labeled entity pairs to amend wrong labels dynamically during training, so as to achieve a similar effect of using soft labels. Wang et al. [67] used TransE [5] to encode entities and relations of a knowledge graph into a continuous low-dimensional space, and then applied the learned relation embeddings as the soft targets of the RE models. References [32] and [67] can be classified into the third research line together with the existing knowledge distillation-based methods [25, 63]. Compared with References [32] and [67], knowledge distillation-based methods can flexibly control the softness of new annotation distributions by regulating the temperatures.

2.2 Knowledge Distillation for Noise Reduction in NLP

Knowledge Distillation is a representative technology in transfer learning, whose teacher–student framework was used for model compression in the early stage [19, 65]. At this time, the teacher was often a wider and deeper network with larger capacity, and the student was a lightweight model. Li et al. [29] used the label softening ability of knowledge distillation to alleviate the interference of noisy annotations to the image classification models. Hereafter, the new function of knowledge distillation was widely applied in various deep learning tasks to combat noisy labels [57, 72, 80]. Moreover, some studies observed that the students having the same capacity as their teachers were easier to be trained [15], so researchers are no longer limited to the use of lightweight students.

In the NLP field, Liu et al. [34] applied a self-knowledge distillation method to weaken the negative influence of noisy data on the training of text summarization models. Hahn and Choi [18] proposed a self-knowledge distillation method based on the soft target probabilities of the training model itself, and the multi-mode information was distilled from the word embedding space right below the softmax layer, so as to boost language models' robustness to noisy targets. To enhance the text-based relational reasoning models' ability against label interference, Dong et al. [10] first pre-trained a graph neural network on a reasoning task using structured inputs and then incorporated its knowledge into an NLP model (e.g., an LSTM) via knowledge distillation. Saiful Bari et al. [56] proposed a knowledge distillation-based data augmentation framework to reduce the damage of potential label noise to cross-lingual NLP tasks. Zhao et al. [79] leveraged a knowledge distillation method to enable the sentiment classification models to learn from noisy data.

Recently, some knowledge distillation instantiations have achieved certain results in distantly supervised RE tasks. Lei et al. [25] proposed a bag-level cooperative denoising framework to leverage the corpus-based and knowledge graph-based information, which employed a bi-directional knowledge distillation approach to train a corpus net and knowledge graph net. Tang et al. [63] proposed an External Neural Constraints Regularized distant supervision framework, which iteratively distilled information from external neural networks to an existing relation extraction model trained with clean data.

Although the existing distillation instantiations for distantly supervised RE are different in the architectures of their teachers and students [25, 63], their distillation approaches are essentially the variants of the conventional methods [19], whose performance will be limited by the propagation of inaccurate dark knowledge and the constraint of a unified distillation temperature (Section 1). Compared with them, our MiDTD is model-agnostic, which has no restrictions on the structure of teacher and student. More importantly, the distillation approach of MiDTD can effectively reduce the two problems by virtue of MiTF and DTR modules.

To the best of our knowledge, the knowledge distillation approach that dynamically regulates temperatures has not been studied in the NLP area before.

3 METHODOLOGY

In this section, we propose a model-agnostic MiDTD framework for distantly supervised RE, which aims at distilling high-quality sentence-level RE models. The remainder of this section is structured as follows.

Section 3.1 introduces the preliminaries about sentence-level and bag-level RE models, thus explaining the purpose of MiDTD. Section 3.2 formalizes the problem and outlines the overall framework. Sections 3.3 and 3.4, respectively, elaborate the two key modules, i.e., *multi-instance target fusion* and *dynamic temperature regulation*, Section 3.5 uses a concrete case to show that the distillation approach of MiDTD is more flexible and controllable than the conventional one, and Section 3.6 constructs three concrete MiDTD instantiations for experimental evaluation.

3.1 Preliminaries

There are two types of distantly supervised RE models, i.e., bag-level and sentence-level, where the former predicts a relation for the common entity pair of all sentences in a bag [31, 52, 66], and the latter makes decisions based on an individual sentence [13, 30, 71].

Correlations between Bag-level and Sentence-level RE Models. In fact, the two kinds of RE models can be applied to both bag-level and sentence-level evaluations. For bag-level RE models, we can test them at the sentence level by setting the sentence number of each bag to 1 (i.e., splitting the bag with multiple sentences into different bags). Accordingly, the sentence-level models can also participate in bag-level evaluation via some simple modifications to their prediction mechanisms (Section 5.1).

Differences between Bag-level and Sentence-level RE Models. The differences mainly reflect in the training pattern. The training of bag-level RE models is mainly based on multi-instance learning, where the models are required to learn to synthesize the information of multiple sentences in a bag to make a unified prediction. The benefit of this pattern is to weaken the influence of the instances wrongly labeled in each bag on model training. But in sentence-level evaluation, it is more difficult for bag-level models to make correct judgments because the information of input sentence is reduced [13]. On the other hand, the training of sentence-level models is for an individual sentence, so such models are more suitable for sentence-level evaluation. But without multi-instance learning, sentence-level RE models often need the assistance of other methods to alleviate the noise annotations from distant supervision. (In our work, we use the knowledge distillation's label softening ability.) In our experiments, we are surprised to find that the advantages of sentence-level RE models can be maintained in bag-level evaluation (Section 5.1).

Why We Choose Sentence-level Models? As mentioned above, on one hand, the performance of bag-level models will be limited in sentence-level evaluation as mentioned above. On the other hand, References [13] and [20] have pointed out that the RE models based on sentence-level prediction are more friendly to the tasks that need to comprehend sentences, e.g., question answering and semantic parsing. Therefore, we set the goal of MiDTD as distilling high-quality sentence-level models. Besides, we find that the performance of our distilled sentence-level RE model on bag-level evaluation is also very satisfactory (Section 5.1).

3.2 Overview of MiDTD Framework

Assume $\{x_i | x_i = (\hat{x}_i, v_i^1, v_i^2)\}_{i=1}^N$ is a raw RE dataset consisting of N instances, where $\hat{x}_i$ is the sentence of the i_{th} instance x_i, v_i^1 and v_i^2 are two specified entities mentioned in $\hat{x}_i$. For instance x_i, a sentence-level RE model is required to classify the relation between v_i^1 and v_i^2 into one of the pre-defined m types (including "NA" type for the entity pairs with no relations). In MiDTD, both teacher and student are sentence-level RE models, where the former is pre-trained with hard distantly supervised annotations, and the latter is asked to imitate the teacher's predicted soft distributions.

Specifically, we denote the teacher and student as $\mathcal{T}$-net and $\mathcal{S}$-net, and their predicted probability distributions for instance x_i as $p(z_i^{\mathcal{T}}, \tau_i^{\mathcal{T}})$ and $p(z_i^{\mathcal{S}}, \tau_i^{\mathcal{S}})$, where $z_i^{\mathcal{T}}, z_i^{\mathcal{S}} \in \mathbb{R}^m$ are logit vectors and $\tau_i^{\mathcal{T}}, \tau_i^{\mathcal{S}} \in \mathbb{R}$ are temperatures in the softmax layers of $\mathcal{T}$-net and $\mathcal{S}$-net. Here, $p(z, \tau)$ is the softmax function with temperature:

$$p(z, \tau) = \frac{\exp(z/\tau)}{\mathbf{1}^{tr} \cdot \exp(z/\tau)}, \tag{1}$$

where $\mathbf{1}^{tr} = (1, \ldots, 1) \in \mathbb{R}^m$, and "$tr$" denotes the transpose operation. In Equation (1), temperature τ is used to control the softness of $p(z, \tau)$: a larger τ indicates a softer (or flatter) prediction $p(z, \tau)$ [19].

To pre-train $\mathcal{T}$-net, we directly use cross entropy as the loss function:

$$\mathcal{L}^{\mathcal{T}} = -\sum_{i=1}^{N} q_i^{tr} \cdot \log\left(p\left(z_i^{\mathcal{T}}, 1\right)\right), \tag{2}$$

where q_i is the one-hot distantly supervised annotation vector for instance x_i.

After $\mathcal{T}$-net is pre-trained, the distillation approach of MiDTD allows $\mathcal{S}$-net to learn from both hard distantly supervised annotations (illustrated in Figure 3) and soft teacher predictions (illustrated in Figure 4) by setting the loss function as

$$\mathcal{L}^{\mathcal{S}} = (1 - \alpha) \cdot \mathcal{L}_{ds}^{\mathcal{S}} + \alpha \cdot \mathcal{L}_{kd}^{\mathcal{S}}, \tag{3}$$

where $\mathcal{L}_{ds}^{\mathcal{S}}$ and $\mathcal{L}_{kd}^{\mathcal{S}}$ are respectively the losses from distant supervision and knowledge distillation, and $0 \leq \alpha \leq 1$ is a trade-off hyper-parameter. $\mathcal{L}_{ds}^{\mathcal{S}}$ is consistent with $\mathcal{L}^{\mathcal{T}}$ in form

$$\mathcal{L}_{ds}^{\mathcal{S}} = -\sum_{i=1}^{N} q_i^{tr} \cdot \log\left(p\left(z_i^{\mathcal{S}}, 1\right)\right). \tag{4}$$

The knowledge distillation loss $\mathcal{L}_{kd}^{\mathcal{S}}$, which decides how we align the student's predictions with the teacher's, is the core of the entire MiDTD. Formally, $\mathcal{L}_{kd}^{\mathcal{S}}$ will be calculated as

$$\mathcal{L}_{kd}^{\mathcal{S}} = -\sum_{i=1}^{N} \left(p\left(\tilde{z}_i^{\mathcal{T}}, \tau_i^{\mathcal{T}}\right)\right)^{tr} \cdot \log\left(p\left(z_i^{\mathcal{S}}, \tau_i^{\mathcal{S}}\right)\right). \tag{5}$$

Here, $\tilde{z}_i^{\mathcal{T}}$ is an integrated logit vector produced by the multi-instance target fusion module (Section 3.3), which combines $\mathcal{T}$-net's logit vectors for multiple sentences in the bag to improve the dark knowledge in $\mathcal{S}$-net's soft target corresponding to x_i. Temperatures $\tau_i^{\mathcal{T}}, \tau_i^{\mathcal{S}}$ are generated by the dynamic temperature regulation module (Section 3.4), which controls the softness of probabilistic distributions predicted by $\mathcal{T}$-net and $\mathcal{S}$-net within a moderate range in a flexible way.

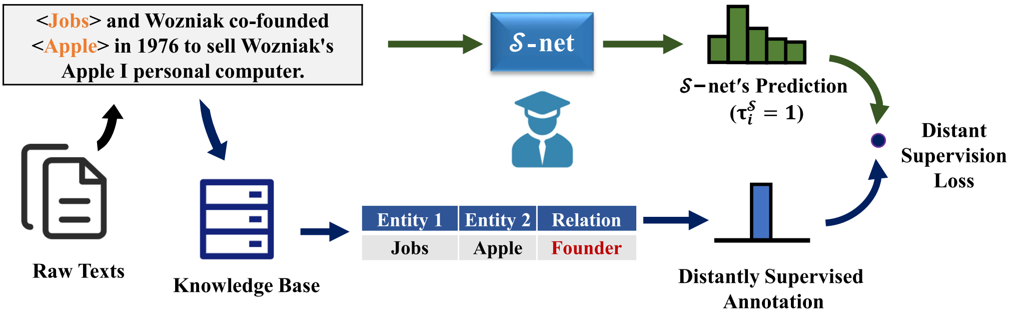

<Jobs> and Wozniak co-founded <Apple> in 1976 to sell Wozniak's Apple I personal computer.
$\mathcal{S}$-net
$\mathcal{S}$−net's Prediction ($\tau_i^S = 1$)
Distant Supervision Loss
Raw Texts
Knowledge Base
Entity 1
Entity 2
Relation
Jobs
Apple
Founder
Distantly Supervised Annotation

Fig. 3. An illustration of $\mathcal{S}$-net's distant supervision loss in MiDTD framework.

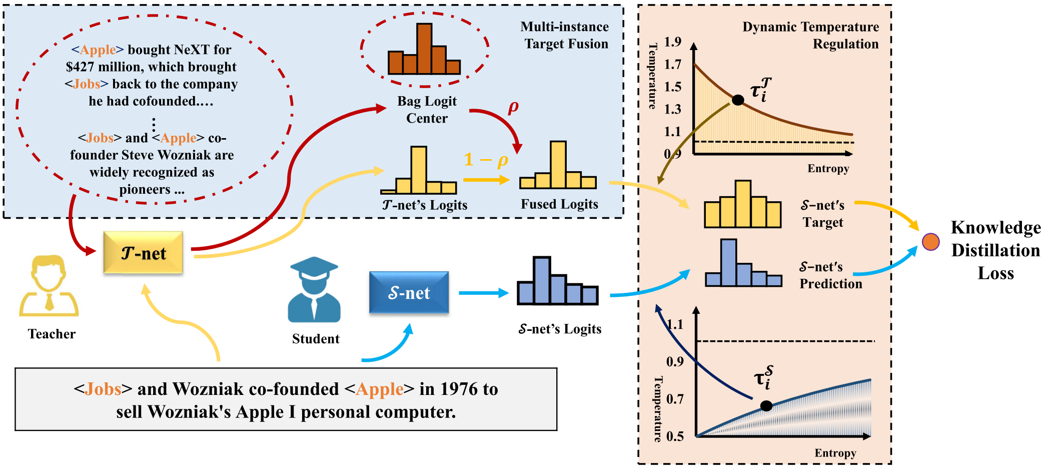

<Apple> bought NeXT for $427 million, which brought <Jobs> back to the company he had cofounded….
⋮
<Jobs> and <Apple> co-founder Steve Wozniak are widely recognized as pioneers ...
Bag Logit Center
Multi-instance Target Fusion
ρ
$1-\rho$
$\mathcal{T}$-net's Logits
Fused Logits
Dynamic Temperature Regulation
Temperature
1.9
1.7
1.5
1.3
1.1
0.9
$\tau_i^{\mathcal{T}}$
Entropy
$\mathcal{S}$-net's Target
$\mathcal{S}$-net's Prediction
Knowledge Distillation Loss
$\mathcal{T}$-net
Teacher
Student
$\mathcal{S}$-net
$\mathcal{S}$-net's Logits
1.1
0.9
0.7
0.5
$\tau_i^{\mathcal{S}}$
Temperature
Entropy
<Jobs> and Wozniak co-founded <Apple> in 1976 to sell Wozniak's Apple I personal computer.

Fig. 4. An illustration of $\mathcal{S}$-net's knowledge distillation loss in MiDTD framework, whose calculation involves two modules: MiTF and DTR. MiTF fuses $\mathcal{T}$-net's bag logit center into its logit vector for the current instance to generate an accurate soft label. DTR assigns alterable temperatures (computed by monotonic exponential functions) to $\mathcal{T}$-net and $\mathcal{S}$-net for different instances.

3.3 Multi-instance Target Fusion

As discussed in Section 1, lots of noise instances will interfere with the learning of $\mathcal{T}$-net and damage the correctness of dark knowledge in the soft labels generated by it. To reduce the inaccurate dark knowledge propagated to $\mathcal{S}$-net, we propose MiTF.

According to the typical experience on RE tasks and the successes of various distantly supervised RE methods, we conjecture that most sentences in a bag will express the same relation for their common entity pairs, i.e., distant supervision relation, so they ought to share the same soft labels and dark knowledge. In light of this, the core idea of MiTF is to combine $\mathcal{T}$-net's predictions for multiple sentences expressing the same relations in the bag to improve the robustness and correctness of the dark knowledge in their respective soft labels. But in fact, we have no way to know which sentences in the bag expressing the same relation, because the distant supervision annotations may be wrong. Therefore, we make a simplification and use all $\mathcal{T}$-net's predictions in the bag to influence the generation of each soft label. Moreover, since the dark knowledge is mainly embodied in the relative sizes between probabilities of different categories [19], the operation of MiTF is directly based on the logit vectors.

Concretely, MiTF uses an integrated logit vector $\tilde{z}_i^{\mathcal{T}}$ to produce $\mathcal{S}$-net's target distribution for x_i in Equation (5). $\tilde{z}_i^{\mathcal{T}}$ is calculated as

$$\tilde{z}_i^{\mathcal{T}} = \frac{\rho}{|b_i|} \cdot \sum_{j \in b_i} z_j^{\mathcal{T}} + (1 - \rho) \cdot z_i^{\mathcal{T}}, \tag{6}$$

where b_i is the index set of all training instances with (v_i^1, v_i^2) as the specified entity pairs, i.e., the instances from the same bag of x_i. $|b_i|$ is the number of indexes in b_i. $\frac{1}{|b_i|} \cdot \sum_{j \in b_i} z_j^{\mathcal{T}}$ is the center of $\mathcal{T}$-net's logit vectors for all sentences in the bag, which we call *bag logit center* for short; $0 \leq \rho \leq 1$ is a balancing hyper-parameter.

Theoretically, the bag logit center will be dominated by the majority of sentences expressing the distant supervision relation, and thus play a positive role for most sentences' soft labels. But we cannot avoid the fact that the soft labels of the sentences expressing other relations do not benefit from the bag logit center, and even be damaged. So we introduce the balancing hyper-parameter ρ.

To interpret Equation (6) from another perspective, we can understand the bag logit center and $z_i^{\mathcal{T}}$ as the representatives of the commonalities of all sentences in the bag and the particularity of the individual sentence. The preservation of each sentence's particularity is the key difference between our MiTF module and existing multi-instance learning-based methods mentioned in Section 2.1, which enables our MiDTD framework to train *sentence-level* $\mathcal{S}$-net while enjoying the benefits of general multi-instance learning.

3.4 Dynamic Temperature Regulation

As mentioned in Section 1, suitable softening can not only alleviate the interference of *noisy* distant supervision annotations, but also provide useful knowledge to enrich the information of *clean* hard annotations. To control the softness of $\mathcal{S}$-net's targets generated by $\mathcal{T}$-net, conventional distillation approaches adopt a unified temperature (i.e., $\tau_i^{\mathcal{T}} = \tau_i^{\mathcal{S}} = \tau$). However, as presented in Figure 2, on the large distantly supervised RE dataset, the softness of initial probabilistic distributions predicted by $\mathcal{T}$-net (temperature is 1) is uneven. A unified temperature is likely to lead to the cases where the relatively softer predictions (I2–I7 in Figure 2) are excessively softened to lose some distinct useful information, or the hard ones (I1 in Figure 2) are not softened sufficiently.

To ensure a moderate range of softness for all $\mathcal{S}$-net's targets, we propose DTR. The idea is to select an indicator quantitatively measuring the softness of $\mathcal{T}$-net's predicted distributions at base temperature 1 (i.e., $p(\tilde{z}_i^{\mathcal{T}}, 1)$), and then dynamically allocate temperatures for different instances according to the indicator.

Concretely, we employ entropy to measure the softness of $\mathcal{T}$-net's initial prediction $p(\tilde{z}_i^{\mathcal{T}}, 1)$

$$e_i = -p\left(\tilde{z}_i^{\mathcal{T}}, 1\right)^{tr} \cdot \log\left(p\left(\tilde{z}_i^{\mathcal{T}}, 1\right)\right). \tag{7}$$

According to Hinton et al. [19], a softer probability distribution will have a larger entropy.

Then $\mathcal{T}$-net's temperature is computed by an exponential function

$$\tau_i^{\mathcal{T}} = 1 + \mu_1 \cdot \exp(-k \cdot e_i), \tag{8}$$

where $\mu_1 \geq 0$ and $k \geq 0$ are two hyper-parameters. In Equation (8), a prediction $p(\tilde{z}_i^{\mathcal{T}}, 1)$ with larger entropy will correspond to a higher temperature $\tau_i^{\mathcal{T}}$. In this way, the harder $\mathcal{T}$-net's predictions will be softened to a greater extent, so that the softness of different $\mathcal{S}$-net's targets in Equation (5) can be controlled in an appropriate range.

Also, DTR does not force $\mathcal{S}$-net to keep the same temperatures as $\mathcal{T}$-net. In our opinion, the temperature suitable for softening the targets of $\mathcal{S}$-net is not necessarily suitable for its learning.

Therefore, we set the temperature of $\mathcal{S}$-net $\tau_i^{\mathcal{S}}$ as

$$\tau_i^{\mathcal{S}} = 1 + \mu_2 \cdot \exp(-k \cdot e_i), \tag{9}$$

where $\mu_2 > -1$ is another hyper-parameter independent of μ_1 in Equation (8). Compared with μ_1, we expand the range of μ_2–$(-1, \infty)$ to enable $\mathcal{S}$-net to either soften or harden its predictions.[6] Actually, it is not necessary for $\tau_i^{\mathcal{S}}$ and $\tau_i^{\mathcal{T}}$ to be consistent in form. But we find the settings of Equations (8) and (9) that keep the changing ratio between $\tau_i^{\mathcal{T}}$ and $\tau_i^{\mathcal{S}}$ at μ_1/μ_2 are more friendly for parameter tuning and training in practical use. After distilling, the temperature of $\mathcal{S}$-net will return back to 1 in the test stage.

Overall, DTR has two desired characteristics: (i) $\tau_i^{\mathcal{T}}$ is alterable across training instances; (ii) $\tau_i^{\mathcal{T}}$ and $\tau_i^{\mathcal{S}}$ are independent. Therefore, MiDTD not only provides a set of more appropriately soft targets for $\mathcal{S}$-net, but also allows $\mathcal{S}$-net to search the temperatures suitable for its learning.

3.5 Theoretical Analysis of MiDTD

From Sections 3.3 and 3.4, we can find that the distillation approach of MiDTD is actually an extension of the conventional one. Therefore, We can utilize the same assumptions in Reference [19] to deduce an interesting conclusion about MiDTD, so as to better explain the benefits of our approach.

In Reference [19], the authors deduced that the conventional distillation approach is equivalent to the optimization problem

$$\min \quad \sum_{i=1}^{N} \frac{1}{2} \left\| z_i^{\mathcal{T}} - z_i^{\mathcal{S}} \right\|^2, \tag{10}$$

under the conditions

$$\begin{aligned} &\tau_i^{\mathcal{T}} \gg \left\| z_i^{\mathcal{T}} \right\|, \\ &\tau_i^{\mathcal{S}} \gg \left\| z_i^{\mathcal{S}} \right\|, \\ &O\left(\tau_i^{\mathcal{S}}\right) = O\left(\tau_i^{\mathcal{T}}\right), \\ &\mathbf{1}^{tr} \cdot z_i^{\mathcal{T}} = \mathbf{1}^{tr} \cdot z_i^{\mathcal{S}} = 0, \\ &\forall\, i \in \{1, \ldots, N\}, \end{aligned} \tag{11}$$

where $|| \cdot ||$ is L_2-norm.

Conclusion. *With the same conditions in Equation (11), we can derive a more flexible form of our MiDTD*[7]*:*

$$\min \quad \sum_{i=1}^{N} \frac{1}{2} \left\| z_i' - z_i^{\mathcal{S}} \right\|^2, \tag{12}$$

where

$$z_i' = \frac{\tau_i^{\mathcal{S}}}{\tau_i^{\mathcal{T}}} \cdot \tilde{z}_i^{\mathcal{T}} = \frac{\tau_i^{\mathcal{S}}}{\tau_i^{\mathcal{T}}} \cdot \left(\frac{\rho}{|b_i|} \cdot \sum_{j \in b_i} z_j^{\mathcal{T}} + (1 - \rho) \cdot z_i^{\mathcal{T}} \right). \tag{13}$$

Proof. During distilling, the gradient of $\mathcal{L}_{kd}^{\mathcal{S}}$ about the j_{th} logit in $z_i^{\mathcal{S}}$ is

$$\frac{\partial \mathcal{L}_{kd}^{\mathcal{S}}}{\partial z_{i,j}^{\mathcal{S}}} = \frac{1}{\tau_i^{\mathcal{S}}} \left(p_j\left(z_i^{\mathcal{S}}, \tau_i^{\mathcal{S}}\right) - p_j\left(\tilde{z}_i^{\mathcal{T}}, \tau_i^{\mathcal{T}}\right) \right). \tag{14}$$

[6] In a knowledge distillation framework, $\mathcal{T}$-net's temperature is generally restricted to be greater than 1 to further soften $\mathcal{S}$-net's target, which is irrelevant to $\mathcal{S}$-net's temperature.

[7] Here, we do not consider the distant supervision loss $\mathcal{L}_{ds}^{\mathcal{S}}$, i.e., set $\alpha = 1$ in Equation (3).

Notice that

$$p_j\left(\tilde{z}_i^{\mathcal{T}}, \tau_i^{\mathcal{T}}\right) = \frac{\exp\left(\tilde{z}_{i,j}^{\mathcal{T}}/\tau_i^{\mathcal{T}}\right)}{\mathbf{1}^{tr} \cdot \exp\left(\tilde{z}_i^{\mathcal{T}}/\tau_i^{\mathcal{T}}\right)} = \frac{\exp\left(\frac{\tilde{z}_{i,j}^{\mathcal{T}} \cdot \tau_i^{\mathcal{S}}/\tau_i^{\mathcal{T}}}{\tau_i^{\mathcal{S}}}\right)}{\mathbf{1}^{tr} \cdot \exp\left(\frac{\tilde{z}_i^{\mathcal{T}} \cdot \tau_i^{\mathcal{S}}/\tau_i^{\mathcal{T}}}{\tau_i^{\mathcal{S}}}\right)} = p_j\left(z_i', \tau_i^{\mathcal{S}}\right). \tag{15}$$

Therefore, Equation (14) can be written as

$$\begin{aligned} \frac{\partial \mathcal{L}_{kd}^{\mathcal{S}}}{\partial z_{i,j}^{\mathcal{S}}} &= \frac{1}{\tau_i^{\mathcal{S}}}\left(p_j\left(z_i^{\mathcal{S}}, \tau_i^{\mathcal{S}}\right) - p_j\left(z_i', \tau_i^{\mathcal{S}}\right)\right) \\ &= \frac{1}{\tau_i^{\mathcal{S}}}\left(\frac{\exp\left(z_{i,j}^{\mathcal{S}}/\tau_i^{\mathcal{S}}\right)}{\mathbf{1}^{tr} \cdot \exp(z_i^{\mathcal{S}}/\tau_i^{\mathcal{S}})} - \frac{\exp\left(z_{i,j}'/\tau_i^{\mathcal{S}}\right)}{\mathbf{1}^{tr} \cdot \exp(z_i'/\tau_i^{\mathcal{S}})}\right). \end{aligned} \tag{16}$$

If $\tau_i^{\mathcal{T}} \gg ||z_i^{\mathcal{T}}||$, $\tau_i^{\mathcal{S}} \gg ||z_i^{\mathcal{S}}||$, and $O(\tau_i^{\mathcal{S}}) = O(\tau_i^{\mathcal{T}})$ for $\forall\ i \in \{1, \ldots, N\}$, we can deduce

$$z_{i,j}^{\mathcal{S}}/\tau_i^{\mathcal{S}} \to 0, \qquad z_{i,j}'/\tau_i^{\mathcal{S}} \to 0, \qquad \forall\ i \in \{1, \ldots, N\}. \tag{17}$$

So we can approximate

$$\frac{\partial \mathcal{L}_{kd}^{\mathcal{S}}}{\partial z_{i,j}^{\mathcal{S}}} \approx \frac{1}{\tau_i^{\mathcal{S}}}\left(\frac{1 + z_{i,j}^{\mathcal{S}}/\tau_i^{\mathcal{S}}}{m + \mathbf{1}^{tr} \cdot z_i^{\mathcal{S}}/\tau_i^{\mathcal{S}}} - \frac{1 + z_{i,j}'/\tau_i^{\mathcal{S}}}{m + \mathbf{1}^{tr} \cdot z_i'/\tau_i^{\mathcal{S}}}\right), \tag{18}$$

where m is the number of total relation categories (including "NA" for the entity pairs that have no relations).

If $\mathbf{1}^{tr} \cdot z_i^{\mathcal{T}} = \mathbf{1}^{tr} \cdot z_i^{\mathcal{S}} = 0$ for $\forall\ i \in \{1, \ldots, N\}$, Equation (18) can be simplified to

$$\frac{\partial \mathcal{L}_{kd}^{\mathcal{S}}}{\partial z_{i,j}^{\mathcal{S}}} \approx \frac{1}{m\left(\tau_i^{\mathcal{S}}\right)^2}\left(z_{i,j}^{\mathcal{S}} - z_{i,j}'\right). \tag{19}$$

Using a simple integration technique, we can find $\mathcal{L}_{kd}^{\mathcal{S}}$ of MiDTD for individual instance x_i is equivalent to $\frac{1}{2}||z_i^{\mathcal{S}} - z_i'||^2$, and the overall $\mathcal{L}_{kd}^{\mathcal{S}}$ is equivalent to $\frac{1}{2}\sum_{i=1}^{N}||z_i' - z_i^{\mathcal{S}}||^2$. The conclusion is verified.

Analysis. Comparing the two optimization problems in Equation (10) and Equations (12)–(13), we can see the effects of MiTF and DTR more intuitively. Under the conditions in Equation (11), the two distillation approaches are transformed to directly match the logit vectors of $\mathcal{S}$-net with that of $\mathcal{T}$-net. The differences are (i) MiDTD can determine the distance between $\mathcal{S}$-net's target logit vector and $\mathcal{T}$-net's bag logit center by adjusting ρ in Equation (13), (ii) MiDTD can control the norm of $\mathcal{S}$-net's target logit vector by regulating the temperature ratio $\frac{\tau_i^{\mathcal{S}}}{\tau_i^{\mathcal{T}}}$ in Equation (13). Therefore, MiTF and DTR make the distillation approach of MiDTD more flexible and controllable than the conventional one.

3.6 Instantiations of MiDTD

In this subsection, we respectively use BERT, PCNN, and BiLSTM-based sentence-level RE models to construct three different MiDTD instantiations for evaluation. To directly observe the effect of MiDTD, we use the same architecture for $\mathcal{T}$-net and $\mathcal{S}$-net in each instantiation.

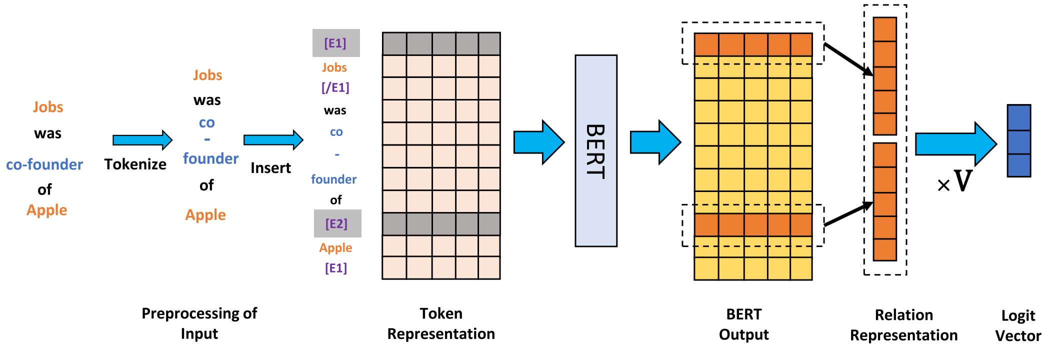


Fig. 5. An illustration of the BERT-based sentence-level RE Model. In preprocessing, the words are tokenized into corresponding word-pieces, and special tokens "[E1]", "[/E1]" and "[E2]", and "[/E2]" are inserted on both sides of the two entities. Then, the input of each token is inputted into a BERT. The BERT outputs corresponding to "[E1]" and "[E2]" are concatenated as the relation representation of the two entities. The relation representation is converted into logit vector by the relation projection matrix V.

3.6.1 A BERT-based MiDTD Instantiation. First of all, we build a BERT-based MiDTD instantiation[8] as the main test object in Sections 4 and 5.

The $\mathcal{T}$-net is a BERT$_{base}$ [8] model for sentence-level RE, whose architecture is mainly borrowed from Reference [61] and is illustrated in Figure 5. For instance x_i, we first use WordPiece Tokenization [70] to decompose the words in $\hat{x}_i$ into a set of tokens. Then, we respectively insert the special tokens "[E1]", "[/E1]" and "[E2]", and "[/E2]" on both sides of the tokenized entities v_i^1 and v_i^2 and denote the new token sequence as $\bar{x}_i$. In the BERT, the input of each token in $\bar{x}_i$ is the sum of the token embedding, position embedding and segment embedding.[9] After processed by BERT, the outputs corresponding to tokens "[E1]" and "[E2]" are concatenated as the relation representation of the entity pair, denoted as $d_i^{\mathcal{T}}$. At last, $d_i^{\mathcal{T}}$ is fed into the softmax classifier of $\mathcal{T}$-net and generate a logit vector as $z_i^{\mathcal{T}} = V^{\mathcal{T}} \cdot d_i^{\mathcal{T}}$, where $V^{\mathcal{T}}$ is the relation projection matrix. The architecture of $\mathcal{S}$-net is the same as $\mathcal{T}$-net.

3.6.2 PCNN and BiLSTM-based MiDTD Instantiations. To demonstrate the robustness of our MiDTD, we further leverage a pair of PCNN/BiLSTM-based sentence-level RE models to build MiDTD instantiation.

For the PCNN-based MiDTD instantiation, $\mathcal{T}$-net and $\mathcal{S}$-net are played by two PCNN RE models [75] as illustrated in Figure 6. Concretely, the word embeddings are trained by Skip-gram models [39]. For instance x_i, the concatenations of word embeddings and position embeddings are fed into a convolution layer, and then the feature maps output from the convolution layer are transformed into the relation representation $d_i^{\mathcal{T}}$ or $d_i^{\mathcal{S}}$ by a piecewise max pooling. Concretely, each feature map is divided into three parts according to the positions of entities v_i^1 and v_i^2, and the three parts will be pooled by max pooling, respectively. Then, the pooling results of each feature map are concatenated as the relation representation $d_i^{\mathcal{T}}$ or $d_i^{\mathcal{S}}$. The relation representation will be converted into logit vector by the relation projection matrix V. Moreover, we remove the multi-instance learning technique in Reference [75] for sentence-level prediction.

For the BiLSTM-based MiDTD instantiation, $\mathcal{T}$-net and $\mathcal{S}$-net are two BiLSTM RE models with the similar architectures as Reference [77], which is illustrated in Figure 7. For instance x_i, the

[8] The code is available at https://github.com/nadineAug/MiDTD.

[9] Since the input sequence contains only one segment, the segment embeddings are the same across all tokens.

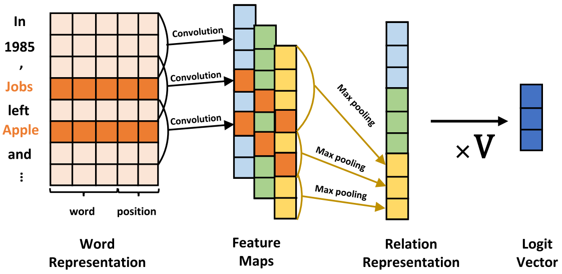


Fig. 6. An illustration of the PCNN-based sentence-level RE Model. The inputs are the concatenation of word embeddings and position embeddings. In convolution layer, the length and number of filters are both 3, and the feature maps will be transformed into a set of 3-dimension vectors after a piecewise max pooling. The concatenation of such vectors serves as the relation representation. The relation representation is converted into logit vector by the relation projection matrix V.

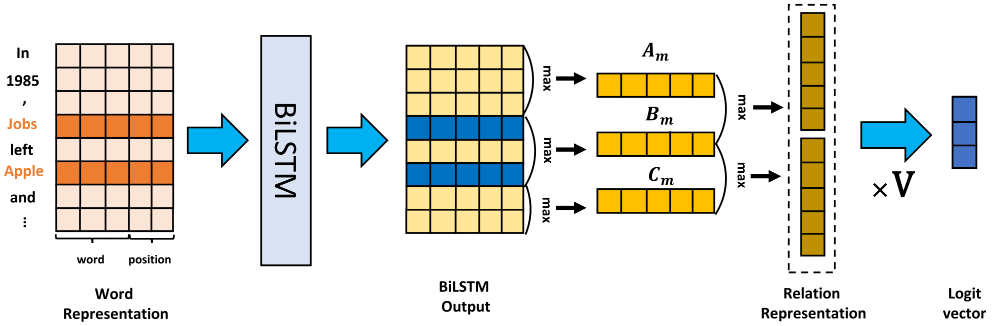


Fig. 7. An illustration of the BiLSTM-based sentence-level RE Model. The input of each word is the concatenation of word embedding and relative position embedding. The two entities split the BiLSTM outputs into three parts, and the results of their respective max pooling enter the next max pooling, then the pooling results will be concatenated as the relation representation. The relation representation is converted into logit vector by relation projection matrix V.

input of a word is the concatenation of its word embedding and the relative position embedding. The inputs enter a BiLSTM. The outputs of the BiLSTM can be divided into three parts by the two entities v_i^1 and v_i^2, and we denote them A, B, and C. In order to obtain the final relation representation $d_i^{\mathcal{T}}$ or $d_i^{\mathcal{S}}$, Reference [77] adopts two max pooling on A, B, and C: (i) respective max pooling on A, B, and C: $A_m = \text{MaxPooling}(A)$, $B_m = \text{MaxPooling}(B)$, and $C_m = \text{MaxPooling}(C)$; (ii) relative max pooling between A_m and B_m: $AB_m = \text{MaxPooling}(A_m, B_m)$ and relative max pooling between B_m and C_m: $BC_m = \text{MaxPooling}(B_m, C_m)$. Then, the concatenation of AB_m and BC_m will serve as $d_i^{\mathcal{T}}$ or $d_i^{\mathcal{S}}$. The relation representation is converted into logit vector by the relation projection matrix V.

4 MAIN EXPERIMENTS

As stated in Section 3, our MiDTD aims at distilling high-quality sentence-level RE models. In this subsection, we test the three MiDTD instantiations on two large-scale benchmark datasets

Table 1. The Statistics of NYT19-1.0 and NYT19-2.0 Datasets, where #Instances, #Entity-Pairs, and #Relations are the Numbers of Instances, Entity-pairs, and Training/test Set's Relations

Dataset		#Instances	#Entity-Pairs
NYT19-1.0	Train	371,461	110,323
#Relations: 25/11	Valid	2,361	1,581
(including "NA")	Test	4,543	2,769
NYT19-2.0	Train	367,892	110,232
#Relations: 11/11	Valid	4,551	2,815
(including "NA")	Test	4,461	2,613

Table 2. HPs of the Teachers and Students in the BERT, PCNN, and BiLSTM-based MiDTD Instantiations on NYT19-1.0/NYT19-2.0

Hyper-Parameter	BERT-$\mathcal{T}$	PCNN-$\mathcal{T}$	BiLSTM-$\mathcal{T}$	BERT-$\mathcal{S}$	PCNN-$\mathcal{S}$	BiLSTM-$\mathcal{S}$
Max length	512	512	512	512	512	512
Word dim	768	50	50	768	50	50
Position dim	-	5	10	-	5	10
Batch size	32/16	16/16	16/20	32/16	16/16	16/20
Learning rate	5e-4/**2.5e-4**	7e-4/7e-4	5e-4/1e-4	1e-4/1e-4	5e-4/2e-4	5e-4/8e-4
Feature maps	-	230	-	-	230	-
Window size	-	3	-	-	3	-
Dropout rate	-	0.2/0.25	0.5/0.5	-	0.25/0.25	0.15/0.25
Hidden size	768	-	300	768	-	300

The Manually Fine-tuned Values are in Bold.

released by Jia et al. [20],[10] and we denote them as NYT19-1.0 and NYT19-2.0. By convention, we use precision, recall, and micro-F1 score as the metrics.

4.1 Datasets

NYT19-1.0 and NYT19-2.0 are annotated on the basis of Ren et al. [51], whose test sets are annotated manually, and training sets are labeled by distant supervision, i.e., aligning entity pairs of sentences from New York Times corpus with the Freebase relations. The statistics of NYT19-1.0 and NYT19-2.0 are presented in Table 1. The main differences between the datasets of Jia et al. [20] (NYT19-1.0 and NYT19-2.0) and Ren et al. [51] are in their test sets, where the former revises the mislabeled instances of the latter and removes some of the relation types, which are overlapping or ambiguous. In the training set of NYT19-2.0, the relations unseen in the test set are relabeled as "NA", and the relation "administrative divisions" (unseen in the test set) are reclassified into "contains" to avoid ambiguity.

4.2 Hyper-Parameter Settings

We denote the teachers of the three instantiations as *BERT-$\mathcal{T}$*, *PCNN-$\mathcal{T}$*, and *BiLSTM-$\mathcal{T}$*, and the students are *BERT-$\mathcal{S}$*, *PCNN-$\mathcal{S}$*, and *BiLSTM-$\mathcal{S}$*. All hyper-parameters of each MiDTD instantiation are divided into two parts: the hyper-parameters of teacher/student, and the hyper-parameters of our distillation approach, which are presented in Tables 2 and 3.

[10] The official dataset website is: https://github.com/PaddlePaddle/Research/tree/master/NLP.

Table 3. HPs of Distillation Approaches in the BERT, PCNN, and BiLSTM-based MiDTD Instantiations on NYT19-1.0 and NYT19-2.0

HPs	**NYT19-1.0**			**NYT19-2.0**		
	BERT	PCNN	BiLSTM	BERT	PCNN	BiLSTM
α	0.6	0.5	0.8	0.8	0.5	0.8
ρ	0.5	0.75	0.5	**0.68**	0.75	0.5
μ_1	0.3	0.25	0.25	0.5	0.35	0.5
μ_2	−0.75	−0.25	0.0	−0.55	0.10	−0.3
k	**45**	10	5	**12**	10	5

The Manually Fine-tuned Values are in Bold.

Moreover, we provide our hyper-parameter adjustment strategies. According to different functions, we divide all **hyper-parameters (HPs)** into six groups, and use performance on the validation set to select their values, which include training HPs: {batch size (BS), learning rate (LR)}, regularization HP[11]: {**dropout rate (DR)**}, MiTF HP: $\{\rho\}$, soft target HPs: $\{\mu_1, k\}$, $\mathcal{S}$-net's temperature HP: $\{\mu_2\}$, and loss balance HP: $\{\alpha\}$.

We sequentially search these HPs in order: (1) training HPs of $\mathcal{T}$-net (2) regularization HP of $\mathcal{T}$-net, (3) training HPs of $\mathcal{S}$-net, (4) regularization HP of $\mathcal{S}$-net, (5) MiTF-HP, (6) soft target HPs, (7) $\mathcal{S}$-net's temperature HP, and (8) loss balance HP. *The value of each HP is set to 0 before adjusting (except that α is initialized to 1).* For the training HPs and soft target HPs (they have two HPs that need to be adjusted at the same time), we tentatively set BS/k to the middle value of its initial range (i.e., BS = 24/k = 10) to adjust LR/μ_1 firstly, and then regulate BS/k.

Specifically, the adjustment process of each HP involves *normal adjustment* and *fine-tuning* two parts. In normal adjustment, BS | LR | DR | ρ | μ_1 | μ_2 | α is selected with binary search from the candidate values in range $[16, 32]$ | $[1e{-}4,\ 1e{-}3]$ | $[0, 0.5]$ | $[0, 1.0]$ | $[0, 2.0]$ | $(-1.0, 2.0]$ | $[0, 1.0]$ with interval 1 | 1e−4 | 0.05 | 0.25 | 0.05 | 0.05 | 0.1.[12] k is searched from $[0, 5, 10, 30, 50]$. For achieving a highest possible result to compare with other existing models, we further add a fine-tuning stage after each BERT-based instantiation's HP is normally adjusted (except for BS). For an HP in fine-tuning stage, we manually try three values around the normally adjusted one. Concretely, k is fine-tuned in positive integers around its normally adjusted value. For each of LR, ρ, μ_1, μ_2, and α, the fine-tuning range is the real values in $(\max\{\mathrm{B}_L,\ \tilde{\theta} - \epsilon\},\ \min\{\mathrm{B}_R,\ \tilde{\theta} + \epsilon\})$, where $\tilde{\theta}$ is the normally adjusted value, B_L and B_R are left and right boundaries of the HP's initial range, and $\epsilon = \frac{B_R - B_L}{8}$. Our fine-tuning is a heuristic HP adjustment strategy, whose core idea is to manually try three values in one quarter of the initial range with the normally adjusted value as center. The functions $\max\{B_L, \cdot\}$ and $\min\{B_R, \cdot\}$ are used to prevent exceeding the boundaries of the initial range. Tables 2 and 3 are the optimal value of each HP, and the manually fine-tuned values are in bold. To save resources, we omit fine-tuning for PCNN and BiLSTM-based MiDTD instantiations.

All HPs are adjusted on 8 GTX 1080 Ti and 4 GTX 2080 Ti GPUs. We empirically train PCNN and BiLSTM for 20 epochs, and BERT models for 15 epochs.

[11] BERT has its own regularization, so we do not apply dropout on it.

[12] Binary search can locate a fine value for each HP with very few attempts, although the value may not be optimal. The initial range and interval of candidate values for each HP are set according to some preliminary experiments. For the sensitive HPs, we will set more dense candidate values, while the interval between candidate values will be larger if the model is insensitive to the HP.

Table 4. Precision (P), Recall (R), and Micro-F1 (F1) of Each Method on the Validation and Test Sets of NYT19-1.0 and NYT19-2.0

Method	**Valid 1.0 (%)**			**Test 1.0 (%)**			**Valid 2.0 (%)**			**Test 2.0 (%)**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PCNN*	36.1	63.7	46.1	36.1	64.9	46.4	39.1	74.7	51.3	42.2	77.5	54.6
BiLSTM*	36.7	66.5	47.3	35.5	67.4	46.5	41.2	70.2	52.1	44.1	71.1	54.5
PCNN+SeATT*	46.0	30.4	36.6	45.4	30.0	36.2	82.4	34.10	48.2	81.0	35.5	49.4
BiLSTM+ATT*	37.6	64.9	47.6	34.9	65.2	45.5	40.8	70.4	51.7	42.8	71.6	53.6
CNN+RL*	40.0	59.2	47.7	40.2	63.8	49.3	42.7	72.6	53.8	44.5	73.4	55.4
ATT+RL*	37.7	52.7	44.0	39.4	61.6	48.1	42.5	71.6	53.3	43.7	72.3	54.5
ARNOR* (SOTA)	**62.5**	58.5	60.4	**65.2**	56.8	60.9	**78.1**	59.8	67.8	**79.7**	62.3	69.9
PCNN-$\mathcal{T}$	47.9	51.3	49.5	47.1	53.3	50.0	43.4	67.1	52.7	43.6	71.7	54.2
PCNN-$\mathcal{S}$	52.4	56.1	54.2	52.0	55.8	53.8	46.1	70.6	55.8	47.1	75.1	57.9
BiLSTM-$\mathcal{T}$	50.1	50.5	50.3	49.9	54.1	51.9	43.4	66.2	52.4	46.6	67.3	55.1
BiLSTM-$\mathcal{S}$	52.5	54.9	53.7	53.6	56.4	55.0	45.5	68.0	54.5	48.5	70.2	57.4
BERT-$\mathcal{T}$	50.8	69.6	58.7	50.5	69.2	58.4	60.0	75.3	66.8	60.2	78.9	68.3
BERT-$\mathcal{S}$	55.0	**75.7**	**63.7**	55.4	**75.3**	**63.8**	63.2	**79.9**	**70.6**	64.4	**81.8**	**72.1**

"*" denotes the result retrieved from Reference [20] an official GitHub website. To facilitate analysis of Section 4.4, the better results between our BERT-$\mathcal{S}$ and the SOTA model ARNOR are in bold.

4.3 Baselines

We use the following SOTA models on NYT19-1.0 and NYT19-2.0 as baselines:

(1) *PCNN* [75] used a piecewise max-pooling after a convolution layer to extract more relation features, which is the base model of our PCNN-$\mathcal{T}$ and PCNN-$\mathcal{S}$.

(2) *BiLSTM* [77] used a bidirectional LSTM to process the concatenation of lexical embedding and relative position embedding of each word, which is the base model of our BiLSTM-$\mathcal{T}$ and BiLSTM-$\mathcal{S}$.

(3) *PCNN+SeATT* [31] adopted a selective attention mechanism over the relation representations of all sentences in a bag to reduce noise, and the relation representations are generated by a PCNN.

(4) *BiLSTM+ATT* [81] added an attention mechanism into BiLSTM to capture the most important features for identifying relations.

(5) *CNN+RL* [13] used a reinforcement learning method to remove the noise distantly supervised instances, which jointly trained a CNN model for RE as well as an instance selector.

(6) *ATT+RL* [46] redistributed false positive instances into negative instances with a reinforcement learning method.

(7) *ARNOR* [20] firstly learned RE patterns from a manually annotated clean dataset. Then, it iteratively trained an interpretable model and utilized it to select trustable instances from the distantly supervised dataset, which is the SOTA method on NYT19-1.0 and NYT19-2.0.

4.4 Results and Analysis

The precision, recall, and F1 score of each method on NYT19-1.0 and NYT19-2.0 are presented in Table 4. From Table 4, we can find that MiDTD is effective and robust.

Comparison between our distilled models and the baselines. Our PCNN-$\mathcal{S}$ and BiLSTM-$\mathcal{S}$ exceed the methods having much more complicated architectures, namely, PCNN+SeATT, BiLSTM+ATT, CNN+RL, and ATT+RL. Moreover, BERT-$\mathcal{S}$ significantly surpasses the SOTA method ARNOR by 3.3%/2.9% on the validation/test set of NYT19-1.0 and 2.8%/2.2% on the validation/test set of NYT19-2.0. In fact, our BERT-based MiDTD instantiation and ARNOR both use additional knowledge resources, where the former resorts to the rich general semantic knowledge of the pre-trained language model, and the latter requires the knowledge from the manually annotated clean RE data. From the comparison between ARNOR and BERT-$\mathcal{T}$, we can conclude that the specialized knowledge from clean RE data is more advantageous than the general semantic knowledge, but MiDTD makes the BERT-RE model out of its disadvantage. Recently, more and more pre-trained language models are opened. Applying MiDTD to these accessible models can even outperform the methods equipping with costly specialized RE knowledge, which fully demonstrates the value of MiDTD.

Another interesting thing is that compared with other methods, the superiority of ARNOR is mainly reflected in precision, while our BERT-$\mathcal{S}$ is better in recall. We think this is because we use soft labels, while ARNOR uses the hard ones. More than 70% sentences in both NYT19-1.0 and NYT19-2.0 do not express relations, i.e., their labels are “NA”. Although ARNOR learns some correct knowledge from the manually annotated RE data, the sentences with relations are still the minority. Therefore, ARNOR learns more about “NA”, and its main mistakes are to predict too much “NA”, which lowers the recall. But in MiDTD, the soft targets enable BERT-$\mathcal{S}$ to learn more knowledge about different relations, i.e., each target has probabilities of different relations. As a result, BERT-$\mathcal{S}$ is more daring to predict different relations, rather than “NA”. The main mistakes of BERT-$\mathcal{S}$ are to confuse different relations (including “NA”), which restricts the precision. We will try to combine the characteristics of the two methods in our future work.

Comparison between the students and teachers. On the validation/test set of NYT19-1.0, PCNN-$\mathcal{S}$, BiLSTM-$\mathcal{S}$, and BERT-$\mathcal{S}$ respectively surpass their teachers by 4.7%/3.8%, 3.4%/3.1%, and 5.0%/5.4% in micro-F1 score. On the validation/test set of NYT19-2.0, they outperform their teachers by 3.1%/3.7%, 2.1%/2.3%, and 3.8%/3.8%. These results verify the robustness of MiDTD, which can significantly improve the performance of different distantly supervised RE models.

Stability of BERT models' performance. Overall, the models engined by BERT outperform other ones in Table 4. To ensure the credibility of their results, we further train BERT-$\mathcal{T}$ and BERT-$\mathcal{S}$ for 4 times with different random seeds, which shows that the averages of F1 scores for BERT-$\mathcal{T}$ and BERT-$\mathcal{S}$ are 57.7%/67.8% and 63.6%/72.0% on the test sets of NYT19-1.0/NYT19-2.0, and the standard deviations are 0.71%/0.44% and 0.28%/0.20%. Therefore, the advantage of BERT models and the effect of our MiDTD on them are stable. We also notice that the results of BERT-$\mathcal{S}$ are more stable than its teacher on both datasets. We think this is because the soft labels are more informative than the hard ones, which reduces the uncertainty of BERT-$\mathcal{S}$' learning.

5 EXPLORATION EXPERIMENTS

We have verified the effectiveness and robustness of our MiDTD instantiations on sentence-level evaluation in Section 4, where the distilled BERT-RE model (BERT-$\mathcal{S}$) significantly surpasses its teacher BERT-$\mathcal{T}$ and the SOTA method ARNOR on both NYT19-1.0 and NYT19-2.0. In this section, we take the BERT-based MiDTD instantiation as the representative and conduct more experiments to further explore the properties of MiDTD.

5.1 Performance on Bag-Level Evaluation

Although the main purpose of MiDTD is to distill sentence-level RE models, there are many distantly supervised RE methods are for bag-level models. So a natural question is: how effective is our MiDTD instantiation on bag-level evaluation? To answer this question, we use a simple modification to change the sentence-level prediction mechanisms of BERT-$\mathcal{T}$ and BERT-$\mathcal{S}$ to the bag-level ones and compare them with specialized bag-level models. Concretely, after distillation, BERT-$\mathcal{T}$ and BERT-$\mathcal{S}$ are asked to predict all instances in a bag, and use their most confident predictions as the predicted results for the entire bag.

In particular, there is no bag-level evaluation result on the two datasets we used in Section 4. Most existing bag-level methods test on another dataset: NYT10 [52], and use a new set of

Table 5. The Statistics of NYT10 Dataset

Dataset		#Instances	#Entity-Pairs
NYT10	Train	399,016	217,298
#Relation: 53/32	Valid	171,008	106,264
(including "NA")	Test	172,448	96,678

There was no validation set originally, so we randomly selected 30% from the training set. #Instances, #Entity-Pairs, and #Relations are the numbers of instances, entity-pairs, and training/test set's relations.

evaluation metrics. In order to compare with them, we additionally introduce NYT10 as a new benchmark dataset and apply the evaluation metrics they used for testing.

Evaluation Metrics. Different from sentence-level evaluation based on precision, recall, and F1 score, the existing literatures employ other metrics for bag-level evaluation, which generally contain the top-N (N=100, 200, and 300) precision and the area under P–R curve [33, 68, 74].

Introduction of NYT10. NYT10 is specifically used for bag-level evaluation by most existing studies [33, 68, 74], which is developed by Riedel et al. [52] via aligning Freebase relations with

New York Times (**NYT**) corpus, where sentences from the year 2005–2006 are used for creating the training set and from the year 2007 for the test set. The entity mentions are annotated by Stanford NER [14], and are linked to Freebase. Although the corpora of NYT10, NYT19-1.0, and NYT19-2.0 are all NYT, their relation numbers, instances are different. The statistics of NYT10 are shown in Table 5.

Hyper-Parameters Settings. For the BERT-based MiDTD instantiation on NYT10, we also adopt normal adjustment + finetuning strategy (Section 4.2) to search each hyper-parameter. Since we empirically find the BERT models on this dataset are more suitable for small learning rates, the initial range of learning rate is expanded as $[5e{-}5, 1e{-}3]$. Finally, the batch sizes of BERT-$\mathcal{T}$ and BERT-$\mathcal{S}$ are adjusted to 32, and their optimal learning rates are both 7e−5. The hyper-parameters of distillation approach $(\alpha, \rho, \mu_1, \mu_2, k)$ are (0.9, 0.5, 0.7, −0.45, 30). Only the learning rates of BERT-$\mathcal{T}$ and BERT-$\mathcal{S}$ are changed by fine-tuning, and the other hyper-parameters are more suitable for their normally adjusted values.

Baselines. The recent SOTA methods on NYT10 include

(1) *RESIDE* [66] employed Graph Convolution Networks to encode syntactic information from texts, and introduced the extra side information obtained by Open Information Extraction methods [3] to improve the performance.

(2) *CNNRL* [47] proposed two assumptions and crafted reinforcement learning to capture the expressive sentence for each relation mentioned in a bag.

(3) *Multir-Cor* [48] discovered and used informative correlations between features with a generative method to capture the complex pattern of relation expression.

(4) *DISTRE* [1] was based on the pre-trained language model GPT-1 [50], which was also demonstrated to possess rich general semantic knowledge like BERT.

(5) *REGAN* [26] additionally used entity descriptions crawled from Wikipedia together with a knowledge base to construct an accurate dataset, and trained the RE model using a generative adversarial network framework. REGAN keeps the SOTA results on NYT10 so far.

(6) *CoRA* [28] proposed a bag-level collaborating relation-augmented attention mechanism to handle both the wrong labeling and long-tail relations.

(7) *SeG* [27] used an entity-aware word embedding method to integrate relative position information and head/tail entity embeddings, and developed a self-attention mechanism to capture relation features.

Table 6. P@100, 200, 300, and the Area under P–R Curve of Each Method on NYT-10 Test Set

Method	P@100	p@200	P@300	Average	AUC-PR
RESIDE	84.5%	76.0%	66.7%	76.1%	0.42
CNNRL	82.0%	76.5%	72.0%	76.8%	-
Multir-Cor	71.3%	66.2%	61.1%	66.2%	-
DISTRE	68.0%	67.0%	65.3%	66.8%	0.42
REGAN (SOTA)	96.0%	93.5%	93.0%	94.2%	0.56
CoRA	98.0%	92.5%	88.3%	92.9%	0.53
SeG	93.0%	90.0%	86.0%	89.3%	0.51
BERT-$\mathcal{T}$	91.0%	87.5%	84.0%	87.5%	0.50
BERT-$\mathcal{S}$	**99.0%**	**95.5%**	**93.7%**	**96.1%**	**0.62**

All results of other existing methods are retrieved from their original papers.
The best results are in bold.

Results and Analysis. The bag-level evaluation results are presented in Table 6. The results of distilled BERT-$\mathcal{S}$ are surprisingly good. On the one hand, BERT-$\mathcal{S}$ outperforms BERT-$\mathcal{T}$ by 8.0%, 8.0%, and 9.3% of P@100, 200, and 300, and has a 24% relative growth over BERT-$\mathcal{T}$ in terms of AUC-PR. On the other hand, the SOTA method REGAN and our BERT RE model respectively use different external knowledge, where REGAN utilizes the knowledge of entity descriptions crawled from Wikipedia, and our BERT model possesses rich general semantic knowledge. From the results between REGAN and BERT-$\mathcal{T}$, the former is better than the latter. But the BERT-$\mathcal{S}$ distilled by MiDTD has a 10.7% relative increase over REGAN in terms of AUC-PR. Therefore, the improvements over existing methods are mostly attributed to our MiDTD distillation approach rather than the general semantic knowledge of BERT.

In addition, compared with the sentence-level evaluation results in Section 4, the superiority of BERT-$\mathcal{S}$ over its teacher BERT-$\mathcal{T}$ is more obvious in bag-level evaluation. We think this is related to our modification on the prediction mechanism. The results in Section 4 have demonstrated that the performance of BERT-$\mathcal{S}$ on single sentences is better than BERT-$\mathcal{T}$. Therefore, the most confident choice of BERT-$\mathcal{S}$ in multiple sentences of each bag is more likely to be correct, and thus its superiority is further expanded in bag-level evaluation. We also tried another modification, where we randomly selected a sentence's prediction as the result of the entire bag. Then, the AUC-PR growth of BERT-$\mathcal{S}$ over BERT-$\mathcal{T}$ was halved (i.e., 0.06). These results also reveal that the advantages between sentence-level RE models will not be restricted in bag-level evaluation, and even can be further improved.

In this experiment, we also train BERT-$\mathcal{T}$ and BERT-$\mathcal{S}$ for 4 times with different random seeds as in Section 4.4. Since only 300 test samples participate in the calculation of P@100, 200, and 300, the variations of their values are relatively large, whose averages' standard deviations are 4.26% and 2.90% for BERT-$\mathcal{T}$ and BERT-$\mathcal{S}$ on NYT10, and averages are 87.8 and 93.9. In contrast, there are sufficient samples for computing AUC-PR, so the standard deviations are obviously smaller and both less than 0.01, and the averages are 0.50 and 0.62.

5.2 Comparing with Other Label Softening Methods

In this subsection, we compare MiDTD with other label softening methods. For a fair comparison, we uniformly use our BERT-RE model in Section 3.6.1 as the base model, and remove the parts unrelated to label softening in each selected method. We respectively evaluate on the three datasets, namely NYT19-1.0, NYT19-2.0, and NYT10, for sentence-level and bag-level evaluations.

Baselines. In order to ensure fairness and the effect of our implementation, we avoid the complex methods that require additional knowledge sources (e.g.,extra supervised RE data). The baselines can be divided into two groups: label softening methods for distantly supervised RE tasks, and general label softening methods, and they are respectively bag-level and sentence-level methods.

The label softening methods for this task include:

(1) *Bi-directional distillation framework* [25] is for bag-level distantly supervised RE models. We first feed all sentences in each bag into a BERT RE model to generate their respective sentence-level relation representations: $\{d_i\}$ as in Section 3.6.1, and then we follow Reference [25] to generate the bag-level relation representation by the attention mechanism

$$d = \sum_i a_i \cdot x_i, \qquad a_i = \frac{\exp(d_i A e)}{\sum_j \exp(d_j A e)}, \tag{20}$$

where A and e are respectively trainable parameter diagonal matrix and vector. d will be transformed into the probabilistic distribution of the bag's relation as

$$z = V \cdot d, \qquad p(z) = \frac{\exp(z)}{\mathbf{1}^{tr} \cdot \exp(z)}, \tag{21}$$

where V is the relation projection matrix. The temperature regulation is abandoned in Reference [25], so we fix the softmax temperatures to 1. For convenience, we remove the corner marks representing teacher or student in Equation (21), because the operations of the teacher and the student are symmetrical. The characteristic of the bi-directional distillation approach is to let two RE models learn from each other, and their loss functions are

$$\mathcal{L}^{te} = \sum (CE(q||p(z^{te})) + \pi^{te} \cdot CE(p(z^{st}||p(z^{te}))), \tag{22}$$

and

$$\mathcal{L}^{st} = \sum (CE(q||p(z^{st})) + \pi^{st} \cdot CE(p(z^{te})||p(z^{st}))), \tag{23}$$

where $CE(\cdot||\cdot)$ is cross entropy, z^{te} and z^{st} are logit vectors of the teacher and student. $\pi^{\mathcal{T}}$ and $\pi^{\mathcal{S}}$ are calculated as

$$\pi^{\mathcal{T}} = \frac{CE(q||p(z^{st}))}{CE(q||p(z^{st})) + CE(q||p(z^{te}))}, \tag{24}$$

and

$$\pi^{\mathcal{S}} = \frac{CE(q||p(z^{te}))}{CE(q||p(z^{st})) + CE(q||p(z^{te}))}. \tag{25}$$

After distillation, we respectively evaluate the teacher and student.

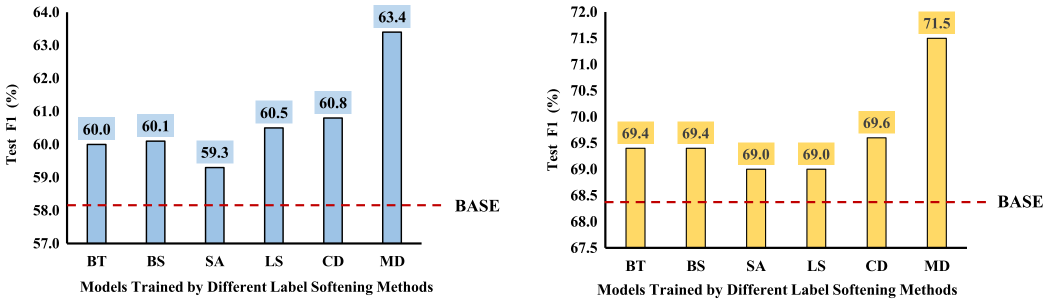
(2) *Soft label adjustment method* [32] is also a bag-level method, which only needs a single model. We first input all sentences in each bag into the BERT-RE model to generate their respective sentence-level relation representations: $\{d_i\}$. Following Reference [32], the bag relation representation d is also generated by the attention mechanism in Equation (20), and the predicted distribution $p(z)$ is also calculated as Equation (21). The target relation of the bag is determined as

$$r = arg\max(p(z) + \max(p(z)) \cdot G \odot q), \tag{26}$$

where G is an m-dimension trainable parameter vector. The loss function is

$$\mathcal{L} = CE(I_r||p(z)), \tag{27}$$

where I_r is a one-hot indicator vector, whose element corresponding to relation r is 1, and the others are 0. Liu et al. [32] call this method soft label adjustment because the bag relation r in Equation (26) is not fixed but will change with the training. In this way, the trained model can achieve a similar effect of training with soft targets.



(a) NYT19-1.0 for Sentence Level

(b) NYT-19-2.0 for Sentence Level

Fig. 8. Sentence-level evaluation results of BT, BS, SA, LS, CD, and MD on NYT19-1.0 and NYT19-2.0. (a) and (b) respective present the micro-F1 scores of the models on the test set of NYT19-1.0 and NYT19-2.0.

The general label softening methods are sentence-level, which contain

(3) *Label smoothing* is a widely-used label softening method, which does not require another model. Label smoothing replaces the hard one-hot target q with $(1-\gamma) \cdot q + \frac{\gamma}{m} \cdot \mathbf{1}$, where $0 \leq \gamma < 1$ is a hyper-parameter, and m is the total relation number, which is also the size of q. For bag-level evaluation, we still use the most confident prediction within all sentences in the bag.

(4) *The conventional distillation framework* [19] applies a unified temperature and directly uses the probabilistic distribution predicted by BERT-$\mathcal{T}$ for each single sentence as BERT-$\mathcal{S}$'s target. This approach is equivalent to forcing $\rho = k = 0$ and $\mu_1 = \mu_2$ in our MiDTD framework.

Hyper-Parameter Settings. In this experiment, the HPs are divided into common HPs (i.e., batch size, learning rate) and private HPs (e.g., γ in label smoothing). Following the adjustment process of Section 4.2, we first determine common HPs and then select private HPs. Before adjusting, each HP is set to 0 (except that α of MiDTD and the conventional distillation is initialized to 1).

According to the previous experience of adjusting BERT-$\mathcal{T}$ and BERT-$\mathcal{S}$, we can directly determine the values of common HPs. The batch size of each model is set to 32/16/32 on NYT19-1.0/NYT19-2.0/NYT10. The **learning rates** (**LRs**) of the models in soft label adjustment and label smoothing methods and the teachers in MiDTD and the conventional distillation framework are 5e-4/2.5e-4/7e-5; the LRs of students in MiDTD, the conventional distillation framework, and bi-directional distillation framework[13] are 1e-4/1e-4/7e-5.

For the private HPs of these methods, we use normal adjustment (i.e., binary search or grid search) to determine their values (*no fine-tuning*). γ in label smoothing is selected by binary search from candidate values in $[0, 0.3]$ with interval 0.05, which is 0.15/0.2/0.2. For our MiDTD and the conventional distillation framework, $(\alpha, \rho, \mu_1, \mu_2, k)$ are searched with normal adjustment as in Section 4.2, which are (0.6, 0.5, 0.25, −0.75, 50)/(0.6, 0.75, 0.45, −0.55, 10)/(0.9, 0.5, 0.7, −0.45, 30) for MiDTD and (0.6, 0, 0.25, 0.25, 0)/(0.8, 0, 0.5, 0.5, 0)/(0.9, 0, 0.25, 0.25, 0) for the conventional distillation framework. The bi-directional distillation framework and soft label adjustment method have no private HPs need to be adjusted.

Results and Analysis. For convenience, we denote the two models in the bi-directional distillation framework as BT and BS. The models trained by soft label adjustment method and label smoothing are denoted as SA and LS. The models distilled by the conventional distillation framework and our MiDTD framework are denoted as CD and MD. The results of sentence-level and

[13] In bi-directional distillation framework, the two models are both students that need to learn from both distant supervision RE data and each of other.

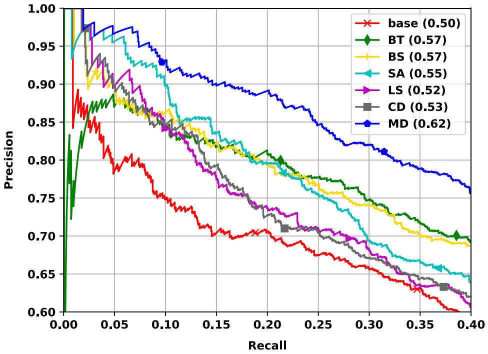


Fig. 9. Bag-level evaluation results (P–R curves) of BT, BS, SA, LS, CD, and MD on NYT10. AUC-PR of each trained model is presented in the parentheses next to the legend name.

bag-level evaluations are plotted in Figures 8 and 9. Overall, all BERT RE models trained with label softening methods exceed the base model trained with one-hot distantly supervised annotations, so label softening is helpful for noise reduction.

Comparing the results in Figures 8 and 9, we can find that the bag-level methods: bi-directional distillation framework and soft label adjustment effectively outperform the sentence-level methods: label smoothing and the conventional distillation on NYT10, but their strengths are obviously weakened and even become disadvantages on NYT19-1.0 and NYT19-2.0. Such results are consistent with References [13] and [20], i.e., the performance of bag-level models is limited in sentence-level evaluation.

Even on NYT10, MD exceeds SA and BT (BS) significantly, and we think the reasons include (i) MiDTD can flexibly control the the softness of MD's targets, (ii) MiDTD enables the student to learn more accurate dark knowledge. About SA, on one hand, soft label adjustment cannot control its targets' softness, so the role of label softening cannot be fully played. On the other hand, the targets of SA are one-hot at each training iteration, which cannot explicitly reflect useful dark knowledge. For BT and BS, although the bi-directional distillation approach theoretically can change the softness of the targets of BT and BS, it is hard to find a suitable temperature because the softness of their targets is constantly changing in the training process. Therefore, the temperatures of BT and BS are restricted to 1 during distilling. Moreover, although BT and BS can learn from each other in distilling, their fundamental RE knowledge resources is the noisy distantly supervised training data, which will inevitably interfere with the dark knowledge in their predictions. As in Equations (22) and (23), the bi-directional distillation approach does not further process the predictions of BT and BS, but directly asks them to learn. Thus, the inaccurate dark knowledge will be directly propagated to BT and BS. But from the fact that the results of BT and BS obviously surpass SA, LS, and CD on NYT10, we can infer that the mechanism of letting two models learn from each other has a certain effect. CD surpasses LS on the three datasets, because label smoothing softens the labels in a more rigid way.

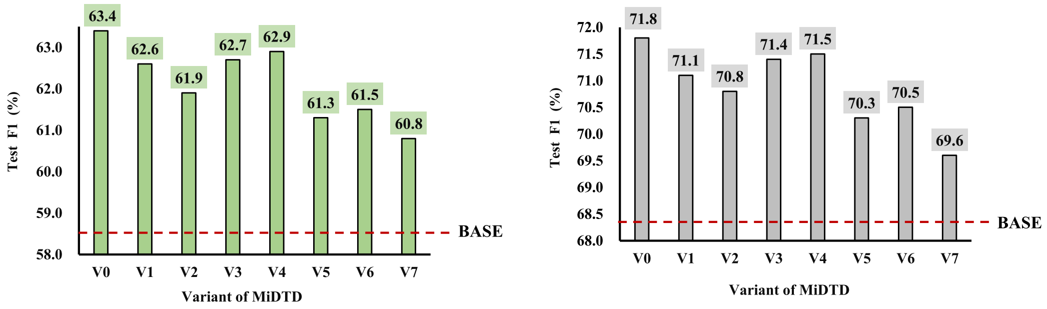
5.3 Ablation Study

We have demonstrated the remarkable performance of the BERT-based MiDTD instantiations. In this subsection, we will dig into the effects of multi-instance target fusion (denoted as M) and dynamic temperature regulation (denoted as D) on the three datasets through ablation study.

Specifically, dynamic temperature regulation further contains two parts: allocating alterable temperatures for different training instances (denoted as D1) and using independent temperatures

Table 7. The Constraints and AHPs of Different MiDTD Variants on NYT19-1.0, NYT19-2.0, and NYT-10

No.	Variants	Constraints	AHPs	NYT19-1.0	NYT19-2.0	NYT10
V0	MiDTD	-	$(\rho, \mu_1, \mu_2, \mathrm{k})$	$(0.5, 0.25, -0.75, 50)$	$(0.75, 0.5, -0.5, 10)$	$(0.5, 0.75, -0.5, 25)$
V1	-w/o M	$\rho = 0$	$(\mu_1, \mu_2, \mathrm{k})$	$(0.25, -0.75, 50)$	$(0.75, -0.5, 15)$	$(0.75, -0.5, 25)$
V2	-w/o D	$\mathrm{k} = 0, \mu_2 = \mu_1$	(ρ, μ_1)	$(0.5, 0.25)$	$(0.75, 0.5)$	$(0.5, 0.25)$
V3	-w/o D1	$\mathrm{k} = 0$	(ρ, μ_1, μ_2)	$(0.5, 0.25, -0.25)$	$(0.75, 0.25, 0)$	$(0.5, 0.25, -0.25)$
V4	-w/o D2	$\mu_2 = \mu_1$	$(\rho, \mu_1, \mathrm{k})$	$(0.5, 0.25, 50)$	$(0.75, 0.5, 10)$	$(0.5, 0.75, 25)$
V5	-w/o M-D1	$\rho = \mathrm{k} = 0$	(μ_1, μ_2)	$(0.25, -0.25)$	$(0.5, 0)$	$(0.25, -0.25)$
V6	-w/o M-D2	$\rho = 0, \mu_1 = \mu_2$	(μ_1, k)	$(0.25, 50)$	$(0.5, 10)$	$(0.5, 25)$
V7	-w/o M-D	$\rho = \mathrm{k} = 0, \mu_1 = \mu_2$	μ_1	0.25	0.5	0.25



(a) NYT19-1.0 for Sentence-Level Evaluation

(b) NYT19-2.0 for Sentence-Level Evaluation

Fig. 10. Sentence-level evaluation results of different MiDTD variants on NYT19-1.0 and NYT19-2.0. (a) and (b) plot the micro-F1 scores of the variants on the test set of NYT19-1.0 and NYT19-2.0.

for $\mathcal{T}$-net and $\mathcal{S}$-net (denoted as D2). The use of M, D1, and D2 is controlled by the values of ρ, μ_1, μ_2, and k in Equations (6), (8), and (9). Table 7 shows the variants of the ablation study, the corresponding constraints, the **adjustable hyper-parameters** (**AHPs**), and the values of AHPs on the three datasets. Therefore, in addition to exploring the effects of M, D1, and D2, this experiment can also point out which hyper-parameters are more important. It is worth noting that the variant **MiDTD-w/o M-D** is essentially the conventional knowledge distillation approach as in Section 5.2.

For a fair comparison, we ensure (i) all variants of MiDTD share a common $\mathcal{T}$-net; (ii) the AHPs of each variant (containing our MiDTD) are selected from: $\rho \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$, $\mu_1 \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, 3, 5, 9\}$, $\mu_2 \in \{0, \pm\frac{1}{4}, \pm\frac{1}{2}, \pm\frac{3}{4}, 1, 3, 5, 9\}$, $k \in \{0, 5, 10, 15, 25, 50\}$[14]; (iii) the AHPs of different variants are adjusted independently.

Results and Analysis. The results of BERT-$\mathcal{S}$ distilled by these MiDTD variants on NYT19-1.0, NYT19-2.0, and NYT-10 are plotted in Figures 10 and 11.

From Figures 10 and 11, we can draw three conclusions. (i) The order of the results of V0, V1, V3, and V4 is: V0 > V3 ≥ V4 > V1, which means that all three modules M, D1, and D2 have positive effects on the overall performance, and their importance order is: M > D1 ≥ D2. (ii) Overall, the performance of each variant is positively correlated to the number of its adjustable hyper-parameters, which reflects the degree of its function's completeness. (iii) In the variants with two

[14] In ablation study, we try to search larger scopes for μ_1, μ_2, and k. To balance the computing resources and the size of search scope, we remove a part of candidate values from the initial search scope (Section 4.2) of each AHP, and add new ones. For the equidistant candidate values of μ_1/μ_2, we use binary search to select one of them and compare with the newly added candidate values 3, 5, and 9. k is still selected with grid search.

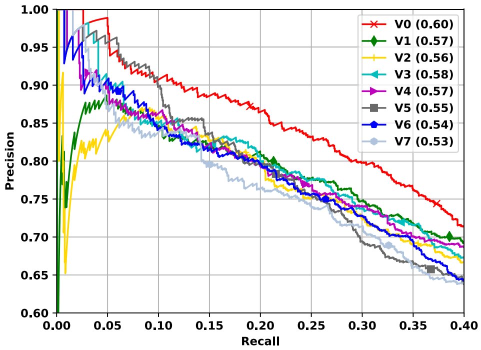


Fig. 11. Bag-level evaluation results (P–R curves) of MiDTD variants on NYT10. AUC-PR of each distilled model is presented in the parentheses next to the legend name.

AHPs, V2 effectively surpasses V5 and V6 on all datasets, which reveals that adjusting ρ is more meaningful than regulating μ_2 and k. If the computing resources are limited in practical use, we can give preference to V2, which can effectively improve the performance of the conventional distillation approach with only one more hyper-parameter needs to be adjusted.

5.4 Statistical Study

In this subsection, we design a statistical experiments to further observe (i) whether MiTF helps to improve the dark knowledge in student's soft targets and (ii) whether the student learns better dark knowledge.

Experimental Settings. Summarizing the discussions in existing literatures, we know dark knowledge reflects the similarity between different categories and mainly embodied in the relative sizes between probabilities [15, 19]. Therefore, we design the following statistical experiment:

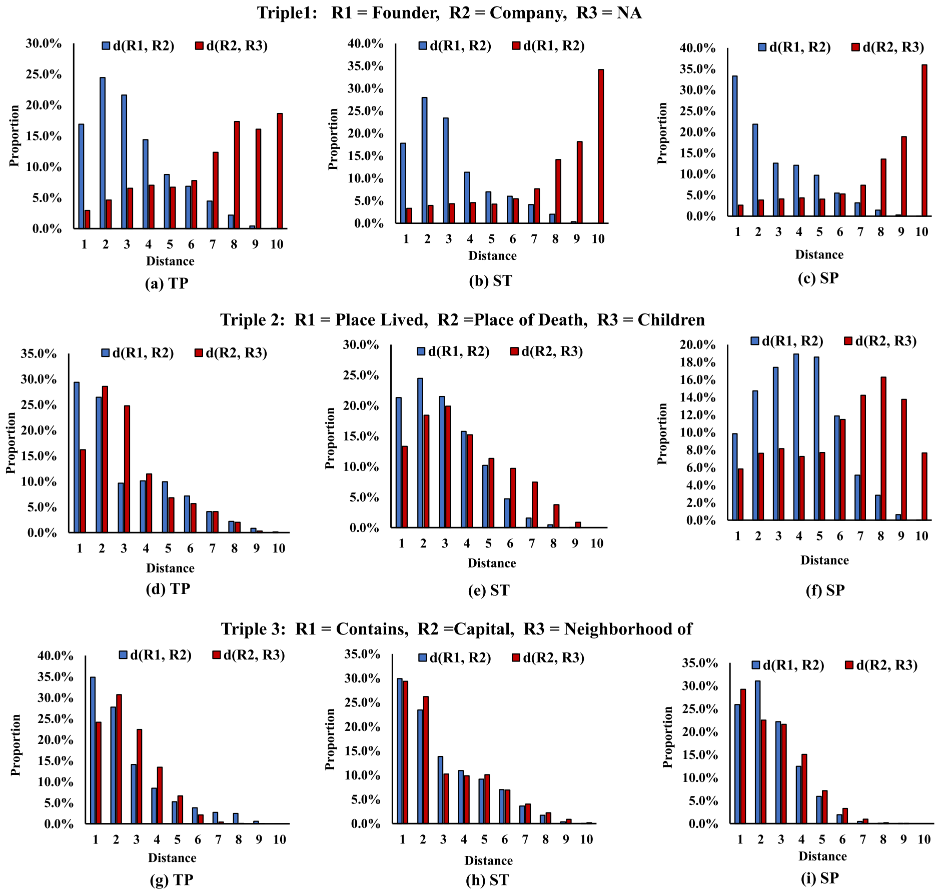
(1) Select a triple of relations (R1, R2, and R3), where relations R1 and R2 are semantically more similar than R2 and R3. For example, as mentioned in Section 1, we can set R1 = "founder", R2 = "company", and R3 = "NA".

(2) Define a distance measuring the similarity between different relation categories in a soft label. We use $r(R)$ to denote the ranking of relation R's probability in the soft label, and the distance between R1 and R2 is defined as $d(R1, R2) =: |r(R1) - r(R2)|$.

(3) Compare $d(R1, R2)$ and $d(R2, R3)$ in three kinds of soft labels: (i) the predictions of BERT-$\mathcal{T}$ on the *training* set of NYT19-2.0, i.e., $p(z_i^{\mathcal{T}}, 1)$, (ii) the targets of BERT-$\mathcal{S}$ on the training set of NYT19-2.0, i.e., $p(\tilde{z}_i^{\mathcal{T}}, 1)$, (iii) the predictions of BERT-$\mathcal{S}$ on the training set of NYT19-2.0, i.e., $p(z_i^{\mathcal{S}}, 1)$. Since $d(R1, R2)$ and $d(R2, R3)$ are unrelated to temperatures, we uniformly set all temperatures to 1.

In experiments, we set three relations triples (R1, R2, and R3) = (founder, company, NA), (Place lived, place of death, and children), and (contains, capital, and neighborhood of). For each relation triple, we plot the distributions of $d(R1, R2)$ and $d(R2, R3)$ on each kind of soft labels in Figure 12. Besides, we count the averages of $d(R1, R2)$ and $d(R2, R3)$ (denoted as $\bar{d}(R1, R2)$ and $\bar{d}(R2, R3)$), and their differences $\Delta\bar{d} = \bar{d}(R2, R3) - \bar{d}(R1, R2)$ in Table 8.

Results and Analysis. Through the preliminary observation of Figure 12 and Table 8, we can infer all the three soft labels have some correct dark knowledge and they have "realized" in most cases that R1 and R2 are more similar than R2 and R3. Actually, in the three relation triples, the difference between the semantic similarity of R1 and R2 and that of R2 and R3 is gradually decreasing. Therefore, it will be more difficult for the soft labels to reflect the correct size relationship between $d(R1, R2)$ and $d(R2, R3)$.



Triple1: R1 = Founder, R2 = Company, R3 = NA

(a) TP (b) ST (c) SP

Triple 2: R1 = Place Lived, R2 =Place of Death, R3 = Children

(d) TP (e) ST (f) SP

Triple 3: R1 = Contains, R2 =Capital, R3 = Neighborhood of

(g) TP (h) ST (i) SP

Fig. 12. Distributions of distances between R1 and R2 ($d(R1, R2)$), R2 and R3 ($d(R2, R3)$) in the soft labels of TP, ST, and SP on NYT19-2.0 training set. In each subplot, the x-axis is the distance of relation (NYT19-2.0 has 11 relation categories in total, so the range of relation distance is [1, 10]), and the y-axis is the proportion of soft labels with the same relation distances. The averages of D1-2 and D2-3 on the entire training set are in the parentheses next to their legend names.

In triple (founder, company, and NA), "founder" and "company" are two sub-relations of business, while "company" and "NA" respectively represent two entities are "related" and "unrelated". So R1 and R2 are obviously more similar in semantics. Therefore, the distributions of $d(R1, R2)$ and $d(R2, R3)$ are very distinct in subplots (a), (b), and (c) of Figure 12, and the values of $\Delta\bar{d}$ for Triple 1 in Table 8 are consistently larger than 5. Besides, $\Delta\bar{d}$ for Triple 1 in Table 8 is increasing across the three kinds of soft labels, which means MiTF further improves the dark knowledge of student's targets, and the student finally masters better dark knowledge.

In the second triple (Place lived, place of death, and children), "place lived" and "place of death" are both relations between people and location, while "children" is a relation between people. The semantic similarity of the R1 and R2 is still higher than that of R2 and R3. But the semantic similarity difference of this triple is less than that of the last triple (i.e., Triple 1), so it is more difficult to judge the relation similarity for the models. Accordingly, the distributions of $d(R1, R2)$ and $d(R2, R3)$ become more similar in subplots (d), (e), and (f) of Figure 12, especially (d) and (e). Moreover, $\Delta\bar{d}$ of BERT-$\mathcal{T}$'s Predictions is only 0.1 in Table 8. At this time, the correctness of the

Table 8. Statistical Results of $\bar{d}(R1, R2)$, $\bar{d}(R2, R3)$, and $\Delta\bar{d}$ in Each Kind of Soft Label

Triple 1: **R1** = Founder, **R2** = Company, **R3** = NA			
Soft Label Type	$\bar{d}(R1, R2)$	$\bar{d}(R2, R3)$	$\Delta\bar{d} = \bar{d}(R2, R3) - \bar{d}(R1, R2)$
BERT-$\mathcal{T}$'s Predictions	3.25	6.96	3.71
BERT-$\mathcal{S}$'s Targets	3.08	7.68	4.60
BERT-$\mathcal{S}$'s Predictions	2.81	7.82	5.01
Triple 2: **R1** = Place Lived, **R2** = Place of Death, **R3** = Children			
Soft Label Type	$\bar{d}(R1, R2)$	$\bar{d}(R2, R3)$	$\Delta\bar{d} = \bar{d}(R2, R3) - \bar{d}(R1, R2)$
BERT-$\mathcal{T}$'s Predictions	3.00	3.10	0.10
BERT-$\mathcal{S}$'s Targets	2.92	3.76	0.84
BERT-$\mathcal{S}$'s Predictions	3.96	6.12	2.16
Triple 3: **R1** = Contains, **R2** = Capital, **R3** = Neighborhood of			
Soft Label Type	$\bar{d}(R1, R2)$	$\bar{d}(R2, R3)$	$\Delta\bar{d} = \bar{d}(R2, R3) - \bar{d}(R1, R2)$
BERT-$\mathcal{T}$'s Predictions	2.60	2.56	−0.04
BERT-$\mathcal{S}$'s Targets	2.93	3.00	0.07
BERT-$\mathcal{S}$'s Predictions	2.49	2.63	0.14

dark knowledge in BERT-$\mathcal{T}$'s Predictions has become vague, but $\Delta\bar{d}$ is still increasing across the three soft labels in Table 8, which verifies the effect of MiTF and the quality of the dark knowledge learnt by BERT-$\mathcal{S}$ again.

In the last triple (contains, capital, and neighborhood of), all of them are relations between locations. Semantically, if one location is the "capital" of another, they naturally have the relation "contain", and they cannot be neighborhoods of each other. Compared with the last two triples, the relation in this one is more likely to be confused. As a result, the differences between distributions of $d(R1, R2)$ and $d(R2, R3)$ are tiny in subplots (g), (h), and (i) of Figure 12, and $\Delta\bar{d}$ of BERT-$\mathcal{T}$'s Predictions is even less than 0 in Table 8, which means most BERT-$\mathcal{T}$'s predictions have confused the three relations. Fortunately, the other two soft labels still guarantee $\Delta\bar{d} > 0$, which verifies our MiTF indeed amend the inaccurate dark knowledge of student's soft targets.

Overall, the results of the first two triples show that MiTF can improve the dark knowledge of student's targets, and the results of the last triple demonstrate that MiTF can correct the inaccurate dark knowledge. The results of all three relation triples verify that the distilled BERT-$\mathcal{S}$ has learnt better dark knowledge than its teacher, which also includes the contribution of DTR.

6 CONCLUSION AND FUTURE WORK

Distantly supervised RE is a practical information extraction task in the field of NLP. In this article, we propose a simple but novel model-agnostic knowledge distillation framework for this task, namely, MiDTD, which employs MiTF and DTR to reduce the propagation of inaccurate dark knowledge and break the constraint of a unified temperature. We theoretically deduce that the distillation approach of MiDTD is more flexible and controllable than the conventional one, and construct three MiDTD instantiations with BERT, PCNN, and BiLSTM-based sentence-level RE models on two large-scale distantly supervised datasets, where the distilled students all significantly outperform their teachers, and even surpass the SOTA methods. In exploration experiments, we also test the BERT-based MiDTD instantiation at the bag level, and the results are still satisfactory. Moreover, we implement four existing label softening methods based on the BERT-RE models, and compare them with our method at both sentence and bag levels, where our MiDTD achieves the best results in both evaluations. Further, we try eight variants of MiDTD in the ablation study and comprehensively explore the effect of each part of MiDTD. Finally, we design a statistical experiment and draw the conclusions: (1) MiTF can improve and amend the dark knowledge in student's soft targets; (2) The model distilled by MiDTD indeed learns better dark knowledge than its teacher.

For further work, as discussed in Section 4.4, an intriguing direction is to combine a part of manually annotated RE data to further improve our method (just like Jia et al. [20]). During exploring the current MiDTD, we found that the noisy distantly supervised data always hindered the neural networks from learning complex mechanisms. Thus, introducing additional neural network structures into MiTF and DTR could not achieve ideal results. But with the participation of manually annotated instances, it is possible to consider more factors and design some specific neural structure for the two modules for better distillation effect, especially precision. Furthermore, we will migrate our proposed MiDTD to other distantly supervised NLP tasks as well.

ACKNOWLEDGMENTS

We thank the EIC and the AE handling our article and the reviewers for their constructive comments and suggestions.

REFERENCES

[1] 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*. 1388–1398.

[2] Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, and Yushi Aono. 2018. Soft-target training with ambiguous emotional utterances for DNN-based speech emotion classification. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4964–4968.

[3] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 344–354.

[4] Andre Blessing and Hinrich Schütze. 2012. Crosslingual distant supervision for extracting relations of different complexity. Association for Computing Machinery, New York, NY, 1123–1132. DOI: https://doi.org/10.1145/2396761.2398411

[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the Neural Information Processing Systems*. 1–9.

[6] Yixuan Cao, Dian Chen, Hongwei Li, and Ping Luo. 2019. Nested relation extraction with iterative neural Network. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, New York, NY, 1001–1010. DOI: https://doi.org/10.1145/3357384.3358003

[7] Maengsik Choi, Harksoo Kim, and Bruce W. Croft. 2012. Dependency trigram model for social relation extraction from news articles. Association for Computing Machinery, New York, NY, 1047–1048. DOI: https://doi.org/10.1145/2348283.2348462

[8] 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)*. Association for Computational Linguistics, 4171–4186. https://aclanthology.org/N19-1423.

[9] Shimin Di, Yanyan Shen, and Lei Chen. 2019. Relation extraction via domain-aware transfer learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, 1348–1357. DOI: https://doi.org/10.1145/3292500.3330890

[10] Jin Dong, Marc-Antoine Rondeau, and William L Hamilton. 2020. Distilling structured knowledge for text-based relational reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 6782–6791.

[11] Ziling Fan, Luca Soldaini, Arman Cohan, and Nazli Goharian. 2018. Relation extraction for protein-protein interactions affected by mutations. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Washington, DC). Association for Computing Machinery, New York, NY, 506–507. DOI: https://doi.org/10.1145/3233547.3233617

[12] Jun Feng, Minlie Huang, Yijie Zhang, Yang Yang, and Xiaoyan Zhu. 2018. Relation mention extraction from noisy data with hierarchical reinforcement learning. arXiv:1811.01237. Retrieved from https://arxiv.org/abs/1811.01237.

[13] 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*.

[14] Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. 363–370.

[15] Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1607–1616. https://proceedings.mlr.press/v80/furlanello18a.html.

[16] Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1774–1784.

[17] Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 241–251.

[18] Sangchul Hahn and Heeyoul Choi. 2019. Self-knowledge distillation in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. 423–430.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems (Workshop)*. Curran Associates, Inc.

[20] 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*. 1399–1408.

[21] Amina Kadry and Laura Dietz. 2017. Open relation extraction for support passage retrieval: Merit and open issues. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo). Association for Computing Machinery, New York, NY, 1149–1152. DOI: https://doi.org/10.1145/3077136.3080744

[22] Jun Kuang, Yixin Cao, Jianbing Zheng, Xiangnan He, Ming Gao, and Aoying Zhou. 2020. Improving neural relation extraction with implicit mutual relations. In *Proceedings of the 2020 IEEE 36th International Conference on Data Engineering*. IEEE, 1021–1032.

[23] Changki Lee, Yi-Gyu Hwang, and Myung-Gil Jang. 2007. Fine-grained named entity recognition and relation extraction for question answering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands). Association for Computing Machinery, New York, NY, 799–800. DOI: https://doi.org/10.1145/1277741.1277915

[24] Sanghak Lee, Seungmin Seo, Byungkook Oh, Kyong-Ho Lee, Donghoon Shin, and Yeonsoo Lee. 2020. Cross-sentence n-ary relation extraction using entity link and discourse relation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (Virtual Event). Association for Computing Machinery, New York, NY, 705–714. DOI: https://doi.org/10.1145/3340531.3412011

[25] Kai Lei, Daoyuan Chen, Yaliang Li, Nan Du, Min Yang, Wei Fan, and Ying Shen. 2018. Cooperative denoising for distantly supervised relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*. 426–436.

[26] Pengshuai Li, Xinsong Zhang, Weijia Jia, and Hai Zhao. 2019. GAN driven semi-distant supervision for relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3026–3035.

[27] Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of the AAAI*. 8269–8276.

[28] Yang Li, Tao Shen, Guodong Long, Jing Jiang, Tianyi Zhou, and Chengqi Zhang. 2020. Improving long-tail relation extraction with collaborating relation-augmented attention. In *Proceedings of the 28th International Conference on Computational Linguistics*. 1653–1664.

[29] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. 2017. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1910–1918.

[30] Zhenzhen Li, Jian-Yun Nie, Benyou Wang, Pan Du, Yuhan Zhang, Lixin Zou, and Dongsheng Li. 2020. Meta-learning for neural relation classification with distant supervision. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (Virtual Event). Association for Computing Machinery, New York, NY, 815–824. DOI: https://doi.org/10.1145/3340531.3412039

[31] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2124–2133.

[32] Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1790–1795.

[33] Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia. 2018. Neural relation extraction via inner-sentence noise reduction and transfer learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2195–2204.

[34] Yang Liu, Sheng Shen, and Mirella Lapata. 2021. Noisy self-knowledge distillation for text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 692–703.

[35] Ismini Lourentzou, Daniel Gruhl, and Steve Welch. 2018. Exploring the efficiency of batch active learning for human-in-the-loop relation extraction. In *Companion Proceedings of the The Web Conference 2018* (Lyon). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1131–1138. DOI: https://doi.org/10.1145/3184558.3191546

[36] Erin Macdonald and Denilson Barbosa. 2020. Neural relation extraction on wikipedia tables for augmenting knowledge graphs. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (Virtual Event). Association for Computing Machinery, New York, NY, 2133–2136. DOI: https://doi.org/10.1145/3340531.3412164

[37] Hassan H. Malik, Vikas S. Bhardwaj, and Huascar Fiorletta. 2011. Accurate information extraction for quantitative financial events. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (Glasgow, Scotland). Association for Computing Machinery, New York, NY, 2497–2500. DOI: https://doi.org/10.1145/2063576.2064001

[38] Mstislav Maslennikov and Tat-Seng Chua. 2010. Combining relations for information extraction from free text. *ACM Transactions on Information Systems* 28, 3 (July 2010), 35 pages. DOI: https://doi.org/10.1145/1777432.1777437

[39] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceeding of the 1st International Conference on Learning Representations, Scottsdale, Arizona, USA, May 2-4, 2013 (ICLR'13)*, Yoshua Bengio and Yann LeCun (Eds.).

[40] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 1003–1011.

[41] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys* 54, 1 (Feb. 2021), 39 pages. DOI: https://doi.org/10.1145/3445965

[42] Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8528–8535.

[43] Masayuki Okamoto, Zifei Shan, and Ryohei Orihara. 2017. Applying information extraction for patent structure analysis. Association for Computing Machinery, New York, NY, 989–992. DOI: https://doi.org/10.1145/3077136.3080698

[44] Sinno Jialin Pan, Zhiqiang Toh, and Jian Su. 2013. Transfer joint embedding for cross-domain named entity recognition. *ACM Transactions on Information Systems* 31, 2 (May 2013), 27 pages. DOI: https://doi.org/10.1145/2457465.2457467

[45] Pengda Qin, Weiran Xu, and William Yang Wang. 2018. 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)*. 496–505.

[46] Pengda Qin, Weiran Xu, and William Yang Wang. 2018. 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)*. 2137–2147.

[47] Jianfeng Qu, Wen Hua, Dantong Ouyang, Xiaofang Zhou, and Ximing Li. 2019. A fine-grained and noise-aware method for neural relation extraction. Association for Computing Machinery, New York, NY, 659–668. DOI: https://doi.org/10.1145/3357384.3357997

[48] Jianfeng Qu, Dantong Ouyang, Wen Hua, Yuxin Ye, and Xiaofang Zhou. 2019. Discovering correlations between sparse features in distant supervision for relation extraction. Association for Computing Machinery, New York, NY, 726–734. DOI: https://doi.org/10.1145/3289600.3291004

[49] Meng Qu, Xiang Ren, Yu Zhang, and Jiawei Han. 2018. Weakly-supervised relation extraction by pattern-enhanced embedding learning. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1257–1266. DOI: https://doi.org/10.1145/3178876.3186024

[50] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[51] Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*. 1015–1024.

[52] 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*. Springer, 148–163.

[53] Ellen Riloff and Wendy Lehnert. 1994. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems* 12, 3 (July 1994), 296–333. DOI: https://doi.org/10.1145/183422.183428

[54] Benjamin Roth and Dietrich Klakow. 2013. Feature-based models for improving the quality of noisy training data for relation extraction. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management* (San Francisco, California). Association for Computing Machinery, New York, NY, 1181–1184. DOI: https://doi.org/10.1145/2505515.2507850

[55] Sunil Kumar Sahu, Derek Thomas, Billy Chiu, Neha Sengupta, and Mohammady Mahdy. 2020. Relation extraction with self-determined graph convolutional Network. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (Virtual Event). Association for Computing Machinery, New York, NY, 2205–2208. DOI: https://doi.org/10.1145/3340531.3412072

[56] M. Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. 2020. MultiMix: A robust data augmentation framework for cross-lingual NLP. arXiv–2004. Retrieved from https://arxiv.org/abs/2004.

[57] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. 2021. Knowledge distillation beyond model compression. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR'21).* IEEE, 6136–6143.

[58] 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*, Vol. 34. 8799–8806.

[59] John Slankas, Xusheng Xiao, Laurie Williams, and Tao Xie. 2014. Relation extraction for inferring access control rules from natural language artifacts. In *Proceedings of the 30th Annual Computer Security Applications Conference* (New Orleans, Louisiana). Association for Computing Machinery, New York, NY, USA, 366–375. DOI: https://doi.org/10.1145/2664243.2664280

[60] Alisa Smirnova and Philippe Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *ACM Computing Surveys* 51, 5 (Nov. 2018), 35 pages. DOI: https://doi.org/10.1145/3241741

[61] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 2895–2905.

[62] Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 721–729.

[63] Siliang Tang, Jinjian Zhang, Ning Zhang, Fei Wu, Jun Xiao, and Yueting Zhuang. 2017. ENCORE: External neural constraints regularized distant supervision for relation extraction. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1113–1116.

[64] Christian Thiel. 2008. Classification on soft labels is robust against label noise. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems.* Springer, 65–73.

[65] Gregor Urban, Krzysztof J. Geras, Samira Ebrahimi Kahou, Ozlem Aslan, Shengjie Wang, Rich Caruana, Abdelrahman Mohamed, Matthai Philipose, and Matt Richardson. 2016. Do deep convolutional nets really need to be deep and convolutional? In *Proceedings of the 5th International Conference on Learning Representations, Toulon, France, April 24-26, 2017 (ICLR'17).* OpenReview.net.

[66] Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* 1257–1266.

[67] Guanying Wang, Wen Zhang, Ruoxu Wang, Yalin Zhou, Xi Chen, Wei Zhang, Hai Zhu, and Huajun Chen. 2018. Label-free distant supervision for relation extraction via knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* 2246–2255.

[68] Shanchan Wu, Kai Fan, and Qiong Zhang. 2019. Improving distantly supervised relation extraction with neural noise converter and conditional optimal selector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7273–7280.

[69] Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management.* 2361–2364.

[70] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144. Retrieved from https://arxiv.org/abs/1609.08144.

[71] Zeqiu Wu, Xiang Ren, Frank F. Xu, Ji Li, and Jiawei Han. 2018. Indirect supervision for relation extraction using question-answer pairs. Association for Computing Machinery, New York, NY, 646–654. DOI: https://doi.org/10.1145/3159652.3159709

[72] Kunran Xu, Lai Rui, Yishi Li, and Lin Gu. 2020. Feature normalized knowledge distillation for image classification. In *Proceedings of the European Conference on Computer Vision*, Vol. 1.

[73] Kaijia Yang, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Exploiting noisy data in distant supervision relation classification. 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)*. 3216–3225.

[74] Changsen Yuan, Heyan Huang, Chong Feng, Xiao Liu, and Xiaochi Wei. 2019. Distant supervision for relation extraction with linear attenuation simulation and non-iid relevance embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7418–7425.

[75] 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*. 1753–1762.

[76] Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Large Scaled Relation Extraction with Reinforcement Learning. In *Proceedings of the AAAI*. 5658–5665.

[77] 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*. Shanghai, 73–78. Retrieved from https://www.aclweb.org/anthology/Y15-1009.

[78] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 35–45.

[79] Pinlong Zhao, Zefeng Han, Qing Yin, Shuxiao Li, and Ou Wu. 2020. Sentiment analysis via dually-born-again network and sample selection. *Intelligent Data Analysis* 24, 6 (2020), 1257–1271.

[80] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. 2021. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. In *Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria, May 3-7, 2021 (ICLR'21)*. OpenReview.net.

[81] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers)*. 207–212.

Received May 2021; revised November 2021; accepted December 2021