Information Processing and Management 61 (2024) 103657

Contents lists available at ScienceDirect

Information Processing and Management

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

CoTea: Collaborative teaching for low-resource named entity recognition with a divide-and-conquer strategy ☆

Zhiwei Yang [a,b], Jing Ma [c], Kang Yang [d], Huiru Lin [e,*], Hechang Chen [d], Ruichao Yang [c], Yi Chang [d,f]

[a] Guangdong Institute of Smart Education, Jinan University, Guangzhou, China
[b] College of Computer Science and Technology, Jilin University, Changchun, China
[c] Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
[d] School of Artificial Intelligence, Jilin University, Changchun, China
[e] Institute of Physical Education, Jinan University, Guangzhou, China
[f] International Center of Future Science, Jilin University, Changchun, China

ARTICLE INFO

Keywords:
Low resource
Named entity recognition
Collaborative teaching
Divide-and-conquer

ABSTRACT

Low-resource named entity recognition (NER) aims to identify entity mentions when training data is scarce. Recent approaches resort to distant data with manual dictionaries for improvement, but such dictionaries are not always available for the target domain and have limited coverage of entities, which may introduce noise. In this paper, we propose a novel Collaborative Teaching (CoTea) framework for low-resource NER with a few supporting labeled examples, which can automatically augment training data and reduce label noise. Specifically, CoTea utilizes the entities in the supporting labeled examples to retrieve entity-related unlabeled data heuristically and then generates accurate distant labels with a novel mining-refining iterative mechanism. For optimizing distant labels, the mechanism mines potential entities from non-entity tokens with a recognition teacher and then refines entity labels with another prompt-based discrimination teacher in a divide-and-conquer manner. Experimental results on two benchmark datasets demonstrate that CoTea outperforms state-of-the-art baselines in low-resource settings and achieves 85% and 65% performance levels of the best high-resource baseline methods by merely utilizing about 2% of labeled data.

1. Introduction

Named entity recognition (NER) aims to identify entity mentions in sentences and assign them semantic categories such as a person (PER), organization (ORG), location (LOC), etc. It is one of the fundamental tasks preceding various natural language processing (NLP) applications, e.g., relation extraction (Sui, Zeng, Chen, Liu, & Zhao, 2023), question answering (Lan et al., 2022), knowledge graph construction (Zhu et al., 2022), etc. Existing supervised approaches for NER achieved superior performances in high-resource settings, i.e., training on a large amount of labeled data (Li, Sun, Han & Li, 2022). However, obtaining large-scale annotated data in a new or low-resource domain, e.g., the disease domain, is difficult and expensive for the NER task. Thus, these

☆ This work is partially supported by National Natural Science Foundation of China through grants (No. U2341229, No. 62206233), Key R&D Program of the Ministry of Science and Technology (2023YFF0905400), the International Cooperation Project of Jilin Province (20220402009GH), and Hong Kong RGC ECS (22200722).

* Corresponding author.: Huiru Lin, Hechang Chen, Yi Chang
E-mail addresses: yangzw18@mails.jlu.edu.cn (Z. Yang), majing@comp.hkbu.edu.hk (J. Ma), yangkang22@mails.jlu.edu.cn (K. Yang), linhuiru@jnu.edu.cn (H. Lin), chenhc@jlu.edu.cn (H. Chen), csrcyang@comp.hkbu.edu.hk (R. Yang), yichang@jlu.edu.cn (Y. Chang).

https://doi.org/10.1016/j.ipm.2024.103657
Received 17 July 2023; Received in revised form 27 December 2023; Accepted 11 January 2024
Available online 20 January 2024

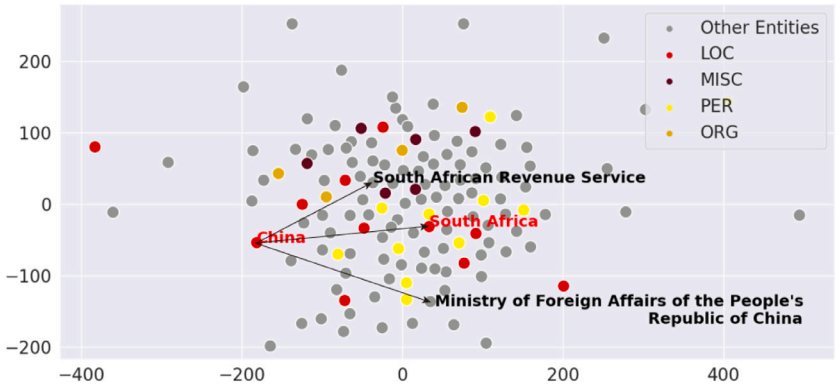

Fig. 1. t-SNE plot of entity embeddings from 100 sentences randomly selected from distant data-augmented CoNLL-2003. The entity labels are provided using dictionaries and "Other Entities" denotes the entities that are not matched.

methods may suffer from data scarcity and not easily identify entities in low-resource settings (Asghari, Sierra-Sosa, & Elmaghraby, 2022; Liu, Jiang, Hu, Shi & Fung, 2021; Liu, Xu et al., 2021). Therefore, the research on low-resource NER is challenging but in demand.

In recent years, low-resource NER has attracted increasing attention when only a small set of labeled data is available (Hedderich, Lange, Adel, Strötgen, & Klakow, 2021; Zevallos et al., 2022). There are three research branches for low-resource NER, including prototype-based methods, knowledge-transferring methods, and data-augmented methods. Prototype-based methods simply classify named entities based on their similarities with a few labeled samples (Ma, Ballesteros et al., 2022; Ma, Jiang, Wu, Zhao & and Lin, 2022), but they are not well-generalized on recognizing unseen low-resource entities. Knowledge-transferring methods attempt to utilize external knowledge from pre-trained language models (PLM) (Cui, Wu, Liu, Yang, & Zhang, 2021; Ma et al., 2023) or existing high-resource labeled datasets (Chen, Aguilar, Neves, & Solorio, 2021; Chen, Liu, Lin, Han, & Sun, 2022; Li, Hu et al., 2022; Nozza, Manchanda, Fersini, Palmonari, & Messina, 2021). However, utilizing data with different distributions and label sets for model training may introduce noise and knowledge mismatch between domains, e.g., "*America*" is annotated as *GPE* in OntoNotes (Weischedel et al., 2013) but *LOC* in CoNLL-2003 (Sang & De Meulder, 2003). Data-augmented methods propose to augment the limited labeled data by searching related samples from a wide range of resources, where dictionaries are applied for data annotation via distant supervision (Cheng, Zhang, Bu, Wu, & Song, 2023; Liang et al., 2020; Meng et al., 2021; Zhang, Yu et al., 2021). However, they may suffer from dictionary quality and thereby overfit to noise.

As shown in Fig. 1, dictionaries can be used to distantly map "*China*" and "*South Africa*" to the location (LOC) category, but there are still many entities that cannot be matched, e.g., "*South Africa Revenue Service*" and "*Ministry of Foreign Affairs of the People's Republic of China*" due to the limited coverage of the dictionary. Thus, we argue that the entity labels of distant data significantly rely on the quality of dictionaries. Although a larger dictionary can be a remedy for strengthening entity coverage and improving distant labels, it is prone to be labor-intensive and domain-specific (Lin et al., 2019; Rijhwani, Zhou, Neubig, & Carbonell, 2020). Besides, "*Ministry of Foreign Affairs of the People's Republic of China*" (ORG) may also be partially labeled with error types, i.e., "*People's Republic of China*" (LOC), because unseen entities were usually missed regarding the dictionary. This introduces label noise such as incomplete labels or incorrect types and thus affects model training. Therefore, to alleviate this issue, we are motivated to utilize the different knowledge from PLMs for refining distant labels, which could be helpful for low-resource NER.

To this end, we propose a novel collaborative teaching method (CoTea) for low-resource NER, enabling automatic data augmentation and label noise toleration. Specifically, CoTea utilizes the supporting entities as the query to retrieve relevant data from the knowledge graph and matches all entities in the distant sentences with low-resource and distant entities for initializing distant labels. Afterward, since there exists noise during retrieving distantly labeled data, a novel mining-refining iterative mechanism based on BERT (Kenton & Toutanova, 2019) and BART (Cui et al., 2021) is introduced to leverage extra knowledge from PLMs and generate refined distant labels. Unlike prior work relying on a well-prepared dictionary for distant labels, this mechanism uses a divide-and-conquer strategy to mine and check entities from the entity- and non-entity parts, respectively, without any dictionary. Extensive experimental results demonstrate that our method can effectively augment data and enhance the performance of recognition models for low-resource NER.

The main contributions can be summarized as follows:

- A novel collaborative teaching (CoTea) framework is proposed for low-resource NER, which enables automatic data augmentation and collaborative teacher models for enhancing model training. This contributes to alleviating data scarcity and distant label noise.
- In this framework, we introduce a novel mining-refining iterative mechanism, where prior knowledge from different pre-trained language models is integrated in a divide-and-conquer manner for label refinement, without manual gazetteers or dictionaries.
- Extensive validations on real-world benchmark datasets compared with the state-of-the-art methods demonstrate that our CoTea sets a new state-of-the-art performance in low-resource settings.

The rest of this paper is organized as follows: Related work is first summarized in Section 2. We then highlight the research implications and novelty of this work in Section 3 and formulate the research problem of this paper in Section 4. Next, our framework is detailed in Section 5. After that, the experimental setting and results are given in Section 6. Finally, Section 6 summarizes the conclusion and future work.

2. Related work

In this section, we provide a review of the research work that is related to our study. Existing methods on NER could be roughly categorized into two groups, i.e., high-resource NER and low-resource NER (Hedderich et al., 2021).

2.1. High-resource NER

There are large-scale, high-quality labeled data in high-resource domains, contributing to model training and thereby significantly enhancing the performance of NER. For example, BiLSTM-CRF combined bidirectional LSTMs with conditional random fields (CRF) for word-level and character-level features (Huang, Xu, & Yu, 2015; Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016), CSEmb used contextual string embeddings for sequence labeling (Akbik, Blythe, & Vollgraf, 2018), CrossNER extracted information by a joint cross-document BiLSTM and multi-task learning (Wang, Fan, & Liu, 2021), BERT-Linear/CRF became a popular paradigm with the pre-trained language model (PLM) for encoding (Kenton & Toutanova, 2019), ACE automated the process of finding better concatenations of different embeddings (Wang, Jiang et al., 2021), LinkBERT pre-trained a language model by leveraging links between documents (Yasunaga, Leskovec, & Liang, 2022), MINER improved out-of-vocabulary NER from an information theoretic perspective (Wang et al., 2022). In addition, channel attention (Xu et al., 2023), planarized sentence representation (Geng, Chen, Huang, Qin, & Zheng, 2023), and interactive networks (Tang, Zhang, Wu, He, & Song, 2022) inspired us for further improvement. Although traditional fully supervised methods have achieved promising performances, they significantly relied on large-scale labeled data for training (Li, Sun et al., 2022). This significantly impairs the effectiveness of these methods in low-resource domains without sufficient labeled data.

2.2. Low-resource NER

Recently, low-resource NER has attracted increasing attention when labeled data is scarce, including different theoretical paradigms, e.g., few-shot (Fritzler, Logacheva, & Kretov, 2019; Huang et al., 2021) or zero-shot learning (Pourpanah et al., 2022). They could be roughly classified into three groups as follows: (1) Prototype-based methods categorize unseen entities regarding their distances with only a small portion of labeled examples, e.g., MAML decomposed meta-learning for NER (Ma, Jiang et al., 2022), and metric-based NER encoded label to help determine entity categories (Ma, Ballesteros et al., 2022). Their strength lies in their ability to generalize from a few examples, making them effective in low-resource scenarios. However, their limitation is that their performance heavily depends on the quality and representativeness of the few labeled examples used. The performance may suffer if these examples do not represent the overall data distribution. (2) Knowledge-transferring methods adapt extra knowledge from PLMs or high-resource domains/languages to that of the low-resource, e.g., BART-NER utilized prompt-based learning for NER (Cui et al., 2021), demonstration-based NER integrated prompt into the input with task demonstrations (Lee et al., 2022), cross-lingual NER used teacher–student distillation training to align high-source languages and low-source languages (Li, Hu et al., 2022), cross-domain augmentation methods transformed the data representation from high-resource domains into the low-resource domains (Chen et al., 2021, 2022; Nozza et al., 2021). Their strength is that they can leverage existing resources and knowledge to improve performance. However, the limitation is that the effectiveness of the transfer might be compromised if there is a significant discrepancy between the source and target domains or languages. (3) Data-augmented methods utilize noisy unlabeled data to augment limited labeled data for better performance, e.g., BOND exploited distant supervision and self-training based on BERT (Liang et al., 2020), RoSTER combined noise-robust learning with augmented self-training for NER (Meng et al., 2021), NEEDLE continually pre-trained on large unlabeled open-domain data and target-domain data based on manual dictionaries (Jiang, Zhang, Cao, Yin, & Zhao, 2021), LADA adopted local additivity based data augmentation to create virtual samples (Chen, Wang, Tian, Yang, & Yang, 2020), and dictionaries are also used to enhance NER (Lin et al., 2019; Rijhwani et al., 2020). Their strength is their ability to leverage large amounts of unlabeled data or existing resources like dictionaries. However, they face the challenge of label noise introduced by the unlabeled data or the limitations in entity coverage by the dictionaries. Despite their effectiveness, these methods share a common limitation in entity coverage and seldom strive to devise a universal approach for low-resource NER, which may limit their applicability.

2.3. Semi-supervised learning

Semi-supervised Learning is a learning paradigm that leverages a large number of unlabeled data for improving the learning performance given a small number of labeled samples (Yang, Song, King, & Xu, 2022). For example, deep generative methods adopt GAN-based frameworks for learning the distribution of real data from unlabeled samples (Kang et al., 2023; Li, Yao et al., 2022; Yu et al., 2021), or VAE-based frameworks for combining deep autoencoders with generative latent-variable models (Cao, Luo, & Klabjan, 2021; Fang et al., 2022), respectively. Consistency regularization methods usually use a teacher–student structure to produce a more accurate model instead of directly using output predictions (Tian, Zhang, Sun, Yin, & Dong, 2022; Ye & Bors, 2021). Graph-based methods perform label inference on a constructed similarity graph to propagate the label information from the labeled

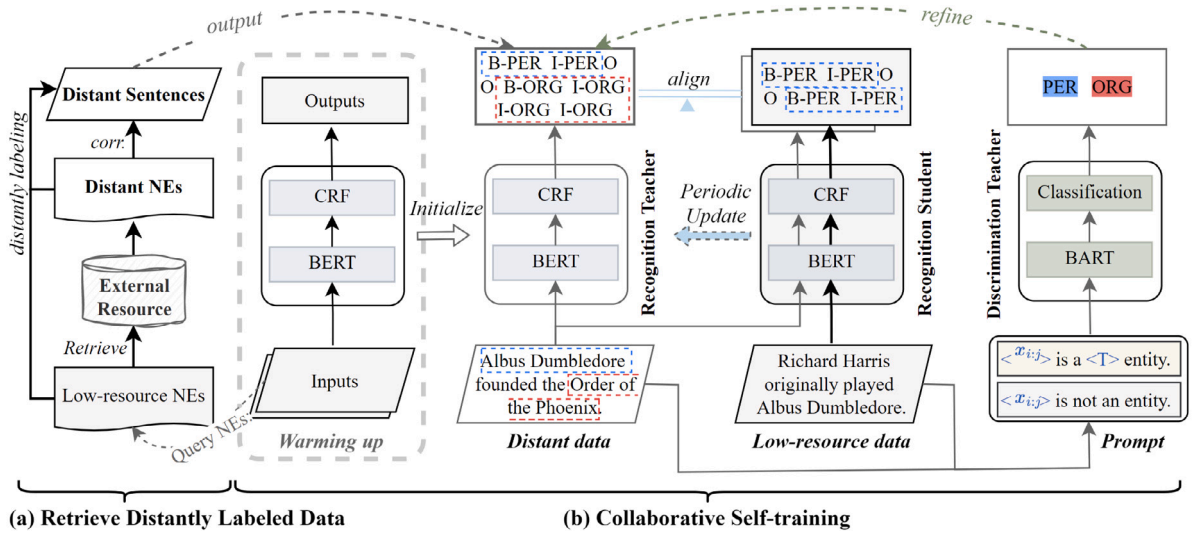

Fig. 2. An overview of CoTea. In the phase (a), distant data is obtained with initial labels. In the phase (b), the recognition teacher (RT) and discrimination teacher (DT) cooperate to refine these distant labels. RT is initialized by warm-up training on low-resource labeled data and updated by consecutive recognition student (RS) models periodically.

samples to the unlabeled ones by incorporating both topological and feature knowledge (Niu, Anitescu, & Chen, 2023; Wan et al., 2021). Pseudo-labeling methods improve the performance of the whole framework based on the disagreement of views or multiple networks, and the emerging hybrid methods reach state-of-the- art performances in most vision benchmarks (Li, Liu, & Song, 2022; Wang, Kihara, Luo & Qi, 2021; Zhang, Wang et al., 2021). Thus, this study explores a novel hybrid method named the collaborative teacher–student (CoTea) framework for consistency regularization and pseudo-labeled self-training, effectively reducing label noise and augmenting training data for better performance in low-resource settings.

Implication and novelty. Finally, we would like to highlight the research implications and emphasize its unique aspects compared to prior works regarding low-resource NER. (1) From a theoretical point of view, this study proposes a novel collaborative teaching framework, which contributes to alleviating data scarcity and distant label noise for low-resource NER. From a practical point of view, our proposed method enables automatic data augmentation without dictionaries and can be easily applied to other research tasks in low-resource settings, which can improve task performance and save annotation costs.

(2) Different from the previous studies, we not only focus on alleviating data scarcity, but we also hope to retrieve distantly labeled data from online resources heuristically (Section 4.1). In addition, we introduce a novel mining-refining iterative mechanism containing two teacher models, where a recognition teacher (Section 4.2) mines potential entities from non-entity tokens and another prompt-based discrimination teacher (Section 4.3) refines entity labels, respectively. Finally, we train neural NER networks using collaborative self-training that form a consensus prediction of distant labels on the ensemble output of these two teachers (Section 4.4) and joint optimization (Section 4.5), which can automatically augment training data and reduce label noise, without manual dictionaries.

3. Problem definition

Formally, a sentence is represented as $\mathbf{X} = \{x_1, \ldots, x_t, \ldots, x_{|X|}\}$, where $|\cdot|$ denotes the sequence length, each word x_t is associated with one entity label $y_t \in \mathbf{Y} = \{y_1, \ldots, y_t, \ldots, y_{|\mathbf{X}|}\}$ based on BIO schema (Li et al., 2012). Specifically, y_t could be $B-T$, $I-T$, and O, indicating the beginning ($B-$), inside ($I-$), and outside (O) of the pre-defined entity type T (e.g., PER and LOC), respectively. Thus, an entity is a span $x_{i:j} = \{x_i, \ldots, x_j\}(1 \le i \le j \le |\mathbf{X}|)$, which is determined by $B-T$ or ($B-T$ and $I-T$) in $y_{i:j} = \{y_i, \ldots, y_j\}$. The high-resource NER can be formulated as a sequence labeling problem, i.e., $f(\mathbf{X}) \rightarrow \hat{\mathbf{Y}}$.

However, for low-resource NER in this paper, only a few labeled data $\mathcal{D}^L = \{(\mathbf{X}^L, \mathbf{Y}^L)\}$ is available, so we further incorporate distant data $\mathcal{D}^D = \{\mathbf{X}^D\}$ for model training, where the superscripts L and D denote labeled data and distant data, respectively. In this study, we use the output of the teacher model as the pseudo labels of distant data, i.e., $\mathbf{Y}^D$, and denote the predicts of the student model as $\hat{\mathbf{Y}}^D$, we generate and refine such distant labels as an auxiliary means of low-resource NER. Thus, we formulate it as $f(\mathbf{X}^L, \mathbf{X}^D) \rightarrow (\hat{\mathbf{Y}}^L, \hat{\mathbf{Y}}^D)$. Ultimately, our goal is to minimize the objective functions $\mathcal{L}_{RT}, \mathcal{L}_{DT}$, and $\mathcal{L}_{TS}$, which can be written as:

$$\mathcal{L}_{RT} = \min \mathcal{L}(\mathbf{X}^L, \mathbf{Y}^L, \hat{\mathbf{Y}}^L) \tag{1}$$

$$\mathcal{L}_{DT} = \min \mathcal{L}(\mathbf{Y}^L, \hat{\mathbf{Y}}^L) \tag{2}$$

$$\mathcal{L}_{TS} = \min \mathcal{L}(\mathbf{Y}^D, \hat{\mathbf{Y}}^D) \tag{3}$$

This work provides an effective paradigm to alleviate data scarcity and reduce label noise, thereby improving the NER performance in low-resource settings. We will describe the details in the following sections.

4. Collaborative teaching framework

Our core idea is to utilize a few labeled examples to obtain a set of distantly labeled data and train NER models on such augmented data using two collaborative teachers in a divide-and-conquer manner, where one teacher is used to mine entities from non-entity parts, and the other teacher checks the predicted categories of entities separately. In the following subsections, we will introduce the details of our proposed framework, including (1) Retrieve distantly labeled data; and (2) Collaborative self-training, as shown in Fig. 2. After that, we will detail the optimization process.

4.1. Retrieve distantly labeled data

Neural NER methods achieved promising performances relying on large-scale labeled data, but manually annotating such training data is expensive and time-consuming. Fortunately, there are many external online resources, e.g., knowledge bases (Dong et al., 2014; Hoffart, Suchanek, Berberich, & Weikum, 2013; Speer, Havasi, et al., 2012), describing named entities. This provides a promising and cheap remedy to automate this process for low-resource NER.

Instead of constructing a large-scale dictionary for each domain (Lin et al., 2019; Rijhwani et al., 2020), our method heuristically searches highly related named entities with their descriptions from Google knowledge graph[1], which can significantly augment training data for deep learning, as shown in Fig. 2(a). Specifically, given a set of pre-defined entity types (*K-way*) and some labeled examples for each type (*N-shot*), we use each entity in a supporting labeled example as a query to search the knowledge graph for collecting more related entities and their description sentences via its API, and then leverage the entity type with string matching methods to obtain initial distant labels (Noise will be considered later). Thus, the retrieved description with initial labels can be used for self-training and thereby foster student model training. Taking the sentence in Table 6 as an example, "*Jean-François van Boxmeer*" and its description sentences are obtained from the knowledge of "*Van Boxmeer*" (PER) and then "*Jean-François van Boxmeer*" is distantly labeled as PER.

Text encoding. Given an input sentence $\mathbf{X} = \{x_1, \dots, x_t, \dots, x_{|\mathrm{X}|}\}$, we utilize pre-trained language models, e.g., BERT (Kenton & Toutanova, 2019) and BART (Lewis et al., 2020), to obtain the representations of the inputs to the following recognizer and discriminator models, respectively, i.e., $\mathbf{h}_{1:|\mathrm{X}|} = \mathrm{Emb}(x_{1:|\mathrm{X}|})$.

4.2. Recognition teacher and recognition student networks

As averaging model weights over training can obtain better model (Polyak & Juditsky, 1992), we adopt a collaborative teacher–student model for NER, where the recognition teacher is used to mine entities from the mismatching part and periodically updated by an average of consecutive student models to tolerate incorrect labels.

Specifically, for recognition teacher and student models in Fig. 2 (b, left), we use the conditional random field (CRF) on the BERT representations for sequence labeling, e.g., B-PER, I-PER, O, etc. CRF can jointly consider neighboring generic labels for decoding the best chain of labels (Ma & Hovy, 2016), e.g., $I-T$ may follow after $B-T$ rather than O in the token label sequence of input sentences based on BIO schema, e.g., $B-T, I-T$. For a given sentence $\mathbf{X}$, we denote its representations as $\mathbf{h}$ and its corresponding predicted label sequence as $\hat{\mathbf{Y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{|\mathrm{X}|})$. Thus, the output conditional probability of CRF is defined as follows:

$$p(\hat{\mathbf{Y}}|\mathbf{h}; \mathbf{W}, \mathbf{b}) = \frac{\prod_{t=1}^{|\mathrm{X}|} f_t(\hat{y}_{t-1}, \hat{y}_t, \mathbf{h})}{\sum_{y \in \mathbf{Y}_\mathbf{x}} \prod_{t=1}^{|\mathrm{X}|} f_t(y_{t-1}, y_t, \mathbf{h})} \tag{4}$$

where y represents a label chosen from all possible labels $\mathbf{Y}_\mathbf{x}$ and $f_t(y_{t-1}, y_t, \mathbf{h}) = exp(\mathbf{W}_{y_{t-1}, y_t} \mathbf{h}_t + \mathbf{b}_{y_{t-1}, y_t})$. Here $\mathbf{W}_{y_{t-1}, y_t}$ and $\mathbf{b}_{y_{t-1}, y_t}$ are the weight parameters corresponding to the label pair (y_{t-1}, y_t). Thus, the training objective of the CRF-based network is to minimize the negative log-likelihood of the correct label sequence as follows:

$$\mathcal{L}_{RT} = -\sum_t \log(p(\hat{\mathbf{Y}}|\mathbf{h}; \mathbf{W}, \mathbf{b})) \tag{5}$$

For decoding, we minimize the loss to search for the best label sequence $\hat{\mathbf{Y}}^*$ as follows:

$$\hat{\mathbf{Y}}^* = \arg\max_{\hat{y} \in \mathbf{Y}_\mathbf{x}} p(\hat{\mathbf{Y}}|\mathbf{h}; \mathbf{W}, \mathbf{b}) \tag{6}$$

where $\hat{\mathbf{Y}}_t^D$ denotes the predictions of recognition teacher on distant data, $\hat{\mathbf{Y}}_s^D$ and $\hat{\mathbf{Y}}_s^L$ denote the student predictions on distant and low-resource data, respectively. This can be computed by the widely-used Viterbi algorithm (Viterbi, 1967). For warming up,

[1] API: https://kgsearch.googleapis.com/v1/entities:search

supervised learning is used to initialize the recognition teacher and its parameters will be periodically updated by trained students during the following training.

Named entity mining. For given $\mathbf{X}^D$, distantly labeling provides initial labels for distant sentences, but there are still many unmatched entities, as shown in Fig. 1. For mining entities from non-entity tokens, we use the outputs of the latest recognition teacher to update their labels as follows:

$$\hat{y}_t = \begin{cases} \hat{y}_t^D & , \text{if} \quad \hat{y}_t^0 = \text{O} \\ \hat{y}_t^0 & , otherwise \end{cases} \tag{7}$$

where $\hat{y}_t^D \in \hat{\mathbf{Y}}_t^D$ denotes the output labels from the recognition teacher and $\hat{y}_t^0$ denotes the original distant labels. In this way, the recognition teacher will iteratively update the distant labels for finetuning the student model.

4.3. Discrimination teacher network

Distant labels generated from distantly labeling (Fig. 2(a)) or the outputs of the recognition teacher (Fig. 2 (b, left)) could inevitably introduce label noise such as incomplete labels or incorrect types. To alleviate this issue, we further design a discrimination teacher based on BART for label refinement. Specifically, for inputs with distant labels, a discrimination teacher network aims to revise the mismatching errors by (1) mapping entities of incorrect types to correct categories or non-entities, and (2) converting incomplete labels to non-entity labels for re-labeling.

To check dubious entities with the discrimination teacher, we need to create entity-centered templates following the formulation of BART, which is superior in classifying such inputs as true or not. Specifically, given an input sequence pair $(\mathbf{X}, \hat{\mathbf{Y}})$, the target sequence $\mathbf{T}_{y_c, x_{i:j}} = \{t_1, \ldots, t_{|\mathrm{T}|}\}$ is a template filled by a predicted entity $x_{i:j}$ and the natural language format of its entity category T, e.g., PER is mapped to "*person*". For example, $\langle Albus Dumbledore \rangle$ is a $\langle people \rangle$ entity. Denoting the one-hot label of entity type T as $y_c (c \in [0, \ldots, k])$. We also randomly select spans $x_{i:j}$ that contain non-entity labels or different categorical labels as negative examples for training, e.g., $\langle theOrder \rangle$ is not an entity. In this way, we create templates for entity $\mathbf{T}^+_{y_c, x_{i:j}}$ and non-entity $\mathbf{T}^-_{y_c, x_{i:j}}$, respectively, as follows:

$\mathbf{T}^+_{y_c, x_{i:j}}$: $\langle x_{i:j} \rangle$ *is a* $\langle y_c \rangle$ *entity.*

$\mathbf{T}^-_{y_c, x_{i:j}}$: $\langle x_{i:j} \rangle$ *is not an entity.*

Different from span classification, we explicitly combine the original entity context with the prompt for sentence classification, e.g., "$\langle Albus Dumbledore \rangle$ *is a* $\langle people \rangle$ *entity*" is appended to "*Richard Harris originally played Albus Dumbledore*", providing an informative context for refining entity labels. Thus, the pre-trained model can easily discriminate different categories based on the informative context. Specifically, given a sequence enhanced with a prompt, i.e., $x'_{i:j} = [x_{i:j}; \mathbf{T}_{y_c, x_{i:j}}]$, we feed $x'_{i:j}$ into BART and obtain the output last hidden state as the final representations, i.e., $\mathbf{h}'_{1:|\mathrm{X}'|}$. In this paper, we adopt the representation $\mathbf{h}'_0$ of classification token $\langle s \rangle$ of BART to represent the current sentence and then a linear layer is used to predict the probability of entity classes as follows:

$$p = \text{Softmax}(\mathbf{W}'\mathbf{h}'_0 + \mathbf{b}') \tag{8}$$

where $\mathbf{W}'$ and $\mathbf{b}'$ are trainable parameters. The cross-entropy between the discrimination predictions and gold labels on $\mathcal{D}^L$ is used as the loss function:

$$\mathcal{L}_{DT} = y_c \log(p) \tag{9}$$

Thus, we also warmed up the discrimination teacher similar to the recognition teacher so as to check dubious entities for the following distant label correction.

Inference for label refinement. We can obtain the categorical label $\hat{y}'_{i:j}$ via Max(p). For each span $x_{i:j}$ in $\mathbf{X}^D$, the discrimination teacher network g_{θ_t} generates its refined distant labels as follows:

$$\hat{y}'_{i:j} = g_{\theta_t}(x'_{i:j}) \tag{10}$$

where $x'_{i:j}$ denotes the enhanced format of $x_{i:j}$. In this way, the discrimination teacher can verify the labels of $x'_{i:j}$ and map it to the correct labels $\hat{y}'_{i:j}$. Thus, we can use these labels to update the entity part of labels $\hat{y}_{i:j} \in \hat{\mathbf{Y}}$, finally obtaining $\hat{\mathbf{Y}}'$ for given $\mathbf{X}^D$, as follows:

$$\hat{y}_t = \begin{cases} \hat{y}'_t & , \text{if} \quad \hat{y}^0_{i:j} = \text{NE} \ \& \ t \in [i : j] \\ \hat{y}^0_t & , otherwise \end{cases} \tag{11}$$

where NE denotes the named entity, and "B" and "I" are not presented for convenience. Besides, the non-entity parts will be explored by the recognition teacher.

4.4. Collaborative self-training

To train neural NER networks using collaborative self-training, we propose to form a consensus prediction of distant labels using the ensemble output of these two teachers, which is expected to be a better predictor. Fig. 2(b) illustrates the overview of collaborative self-training, where CoTea combines distant labels from named entity mining and label refinement using Eq. (7) and Eq. (11), respectively. In this case, CoTea finally reaches a superior balance and refines distant labels for student model training.

During training, to align the outputs of student and teacher models and mitigate the influence of noise, we leverage mean square error (MSE) as the loss function to measure the consistency of the refined outputs from the teachers and the student model as follows:

$$\mathcal{L}_{TS} = -\frac{1}{|\mathrm{X}|}\sum_{t}^{|\mathrm{X}|}(\hat{y}_t^s - \hat{y}_t)^2 \tag{12}$$

where $\hat{y}_t^s$ denotes the predicted label of the student model, $\hat{y}_t$ denotes the refined output of the teacher models.

Periodic update. We assume that a greater teacher produces a better student, which can sometimes be even more excellent than the teacher. After the weights of the recognition student model have been updated with stochastic gradient descent for fixed steps, the weights of the recognition teacher are updated as an exponential moving average (EMA) (Tarvainen & Valpola, 2017) of the student and teacher weights with a ramp-up strategy for promotion as follows:

$$\theta_t' = \beta\theta_t + (1-\beta)\theta_s \tag{13}$$

where β is a smoothing parameter. θ_s and θ_t denote the student and teacher weights, respectively. We use $\beta = \exp(-5 * (1 - \frac{\epsilon}{\tau})^2)$ for calculating β, where τ is a fixed temperature for ramping up and ϵ denotes the training step. This is because the student updates quickly, and the teacher benefits from noisy student training in doing so. The update process is conducted periodically. Different from using $\theta_t' = \theta_s$, periodic EMA for ensembling the student and teacher weights is more smooth and robust for making a better teacher model.

4.5. Model optimization

Overall, we relaxed the limit of using the unlabeled data that distributes similarly to test data for model training, e.g., removing the labels of existing labeled datasets as unlabeled data (Laine & Aila, 2016; Tarvainen & Valpola, 2017). We argue that this process potentially reduces the difficulty of low-resource NER, because unlabeled data with the same distribution is not always available for the target domain. Actually, if we simply mix them up for supervised learning, the divergence or gap between unlabeled and labeled data could impair the final performance.

Thus, we introduce a novel divide-and-conquer collaborative teaching framework for low-resource NER, aiming to train a better recognition model by reducing label noise and improving the teacher model. For warming up, we first train recognition teacher and discrimination teacher networks in supervised learning using Eq. (5) and Eq. (9), respectively. Then, we jointly optimize the recognition training loss $\mathcal{L}_{RT}$ and the teacher–student consistency loss $\mathcal{L}_{TS}$ by minimizing the overall loss $\mathcal{L}_{all}$ as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{RT} + \alpha\mathcal{L}_{TS} \tag{14}$$

where α is the hyperparameter for trading off these two loss terms. In this way, CoTea can effectively train a better model for low-resource NER, and cast new light on distantly supervised learning. Note that we pre-trained the discrimination teacher model on target labeled data so as to check dubious entities for distant label correction, as shown in Algorithm 1. As shown in Fig. 2, our proposed method includes two phases, i.e., (a) retrieving distantly labeled data and (b) collaborative self-training. In the phase (a) iteration, we utilize the entities in the supporting labeled examples as a query to search the knowledge graph for collecting more data and initialize them with distant labels. In the phase (b) iteration, we introduce a mining-refining mechanism based on a divide-and-conquer strategy, where the recognition teacher mines entities from non-entity tokens and updates the non-entity part of distant labels, while the discrimination teacher checks the corresponding entity labels and refines the entity part of distant labels. Finally, we align their outputs and use collaborative self-training for joint optimization.

5. Experiments

In this section, We conduct comprehensive experiments to evaluate the effectiveness of our method compared with the state-of-the-art baseline methods and provide fine-grained analysis for low-resource NER.

5.1. Datasets

We conduct experiments on two datasets, i.e., CoNLL-2003 (Sang & De Meulder, 2003) and NCBI-disease (Doğan, Leaman, & Lu, 2014), corresponding to news and disease domain, respectively. Since there is no public low-resource dataset, existing methods used a small set of fully labeled data to fulfill the requirement of low-resource settings (Chen et al., 2022; Jiang et al., 2021; Meng et al., 2021). Thus, we randomly sample N examples for each class (N-shot) from training and valid sets, respectively. We use N-shot examples and a full test set for evaluation.

The datasets are described as follows: (1) CoNLL-2003 (Sang & De Meulder, 2003) is an English news dataset collected from the Reuters Corpus including 4 entity types, i.e., LOC, MISC, ORG, and PER labels. (2) NCBI-disease (Doğan et al., 2014) is a collection of 793 PubMed abstracts with mentions and concepts annotated as DISEASE or not (one entity class). The dataset statistics are presented in Table 2.

Algorithm 1: CoTea for Low-resource NER

Input: Low-resource training set $\mathcal{D}^L$; Distant data $\mathbf{X}^D$;
The maximal number of retrieved results M; The ramp-up temperature τ.
Output: Predicted labels $\hat{\mathbf{Y}}_s^L, \hat{\mathbf{Y}}^D$;

Initialize $M = 20, \tau = 80$;
for *each example* $\{(x_t, y_t)\}_{t=1}^{|\mathcal{D}^L|}$ **do**
  $\mathbf{X}^D \leftarrow M$, Searching KG based on NEs in $\mathcal{D}^L$;
  $\hat{\mathbf{Y}}^D \leftarrow$ Distant labeling using NEs in $\mathcal{D}^L$ and KG;
$\theta_t, \phi_t \leftarrow \mathcal{D}^L$, Eq. (5) and Eq. (9); *#warm up*
for *each epoch* **do**
  { *Recognition teacher* θ_t *and student* θ_s }
  $\mathbf{h}_{1:|\mathrm{X}|} \leftarrow x_{1:|\mathrm{X}|}$, using BERT;
  $\hat{\mathbf{Y}}_s^L, \hat{\mathbf{Y}}_s^D, \hat{\mathbf{Y}}_t^D \leftarrow$ Eq. (6) ;
    {*Recognition teaching*}
  for $\hat{y}_t^0 \in \hat{\mathbf{Y}}^D, \hat{y}_t^D \in \hat{\mathbf{Y}}_t^D$ **do**
    $\hat{y}_t \leftarrow$ Eq. (7) *#mining*
  $\hat{\mathbf{Y}}^D \leftarrow \{\hat{y}_t\}_{t=1}^{|\mathbf{X}^D|}$;
  {*Discrimination teacher* ϕ_t}
  $\mathbf{X}'_{i:j} = [\mathbf{X}^D_{i:j}; \mathbf{T}_{y_c, x_{i:j}}] \leftarrow$ Template generation $\mathbf{T}_{y_c, x_{i:j}}$;
  $\mathbf{h}'_{1:|\mathrm{X}'|} \leftarrow \mathbf{X}'_{i:j}$, using BART;
  $\hat{y}'_{i:j} \leftarrow \mathbf{h}'_{1:|\mathrm{X}'|}$, Eq. (8) for entity discrimination;
    {*Discrimination teaching*}
  for $\hat{y}_t^0 \in \hat{\mathbf{Y}}^D, \hat{y}'_t \in \hat{\mathbf{Y}}'$ **do**
    $\hat{y}_t \leftarrow$ Eq. (11); *#refining*
  $\hat{\mathbf{Y}}^D \leftarrow \{\hat{y}_t\}_{t=1}^{|\mathbf{X}^D|}$;
  {*Collaborative self-training*}
  $\mathcal{L}_{TS} \leftarrow$ Eq. (7), (11) for alignment;
  if *global_steps % 20 == 0* **then**
    $\beta = \exp(-5 * (1 - \frac{\epsilon}{\tau})^2)$;
    $\theta'_t \leftarrow \beta$, Eq. (13) using EMA;
{*Jointly Optimization*}
Optimize $\mathcal{L}_{all} = \mathcal{L}_{RT} + \alpha\mathcal{L}_{TS}$

5.2. Baseline methods

For comparison purposes, we extensively evaluate our proposed model with a set of low-resource state-of-the-art baseline methods as follows:

- **BOND** (Liang et al., 2020): This leveraged multi-source gazetteers and BERT to improve open-domain NER with distant supervision. The model was also trained using low-resource data.
- **LADA** (Chen et al., 2020): This proposed a local additivity-based method for semi-supervised NER, which generated augmentations with different sentence structures.
- **RoSTER** (Meng et al., 2021) : This proposed a self-training method based on a generalized noise-robust loss using distantly-labeled data. Additionally, the model was trained using low-resource data for improvement.
- **NEEDLE** (Jiang et al., 2021): This trained deep NER models over a weighted combination of manually labeled and distantly labeled data.
- **LabelSem** (Ma, Ballesteros et al., 2022): This learned to match the representations of named entities and labels based on two BERT encoders.
- **Demons** (Lee et al., 2022): This proposed demonstration-based NER with in-context learning based on example sampling and template construction.
- **SDNet** (Chen et al., 2022): This proposed to describe mentions using a universal concept set for few-shot NER.

Table 1
The symbols and their respective meanings.

Symbol	Meaning
$\mathbf{X}$	The input Sentence
$\mathbf{Y}$	The entity labels
x_t	A token in $\mathbf{X}$
y_t	A entity label in $\mathbf{Y}$
$\hat{\mathbf{Y}}$	The predicted entity labels
$\mathbf{h}$	The representation of $\mathbf{X}$
$\hat{\mathbf{Y}}^*$	* is D or L, denoting the predictions on distant and low-resource data, respectively.
$\mathcal{L}_{RT}$	The loss of the recognition teacher
$\mathcal{L}_{DT}$	The loss of the discrimination teacher
$\mathcal{L}_{TS}$	The loss of the collaborative alignment of the teacher and student.
$\mathbf{W}, \mathbf{b}, \theta$	The trainable parameters
α	The trading-off hyper-parameter
β	The auto-increment variable
$\mathbf{T}^+, \mathbf{T}^-$	The templates for entity and non-entity, respectively.

Table 2
Dataset statistics with the number of sequences. #Per denotes the percentage of labeled data. "HIGH" and "LOW" denote the high-resource setting and low-resource setting, respectively. "Distant" is the automatically retrieved data.

Dataset	Train	Valid	Test	Distant	#Per
CoNLL-2003 (HIGH)	14,041	3250	3453	–	2.31%
CoNLL-2003 (LOW)	200	200	3453	5475	
NCBI-disease (HIGH)	5,424	923	940	–	1.58%
NCBI-disease (LOW)	50	50	940	409	

- **MAML** (Ma, Jiang et al., 2022): This presented a decomposed meta-learning method for span detection and entity typing.

In addition, high-resource state-of-the-art baselines are compared as follows:

- **BiLSTM-CRF** (Lample et al., 2016): This proposed a neural architecture based on a bidirectional LSTM with a sequential conditional random field.
- **BERT-Linear/CRF** (Kenton & Toutanova, 2019): This used BERT with a linear or CRF classifier.
- **ACE** (Wang, Jiang et al., 2021): This work automated the process of finding better concatenations of different level embeddings to enhance NER performance.
- **LinkBERT** (Yasunaga et al., 2022): This pre-trained a language model by leveraging links between documents for NER.

5.3. Experimental settings

As shown in Table 2, we use the full data in the standard supervised setting. For the low-resource settings, N is set to 50 to sample supporting examples for each class of K classes, and M is set to 20 to retrieve maximal top-M results for each entity. The related distant entities and low-resource entities are utilized to provide initial labels for retrieved sentences. For fair comparison, distantly supervised baselines, e.g., BOND and RoSTER, are provided with the initial distant data. We integrate different PLMs, i.e., BERT (Kenton & Toutanova, 2019) and BART (Lewis et al., 2020), for collaborative self-training. We train the recognition and discrimination teacher networks on the low-resource data, while training the recognition student both on low-resource and distant data. The outputs of dual teachers are combined to refine labels for distant data. We initialize words as 768-dimensional embeddings with the base uncased BERT and the max length of word sequences is empirically set to 400. We use AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 3e-5 for training our framework. The ramp-up temperature is set to 80 and the updating period is empirically set to 20 for EMA. The hyperparameter α is set to 1. For the discrimination network, we pre-finetuned BART (Lewis et al., 2020) for prompt learning. The batch sizes of low-resource and distant data are set to 2 and 16, respectively. This work tries Masked language modeling (MLM) to warm up BERT with a fill-in-the-blank task (Devlin, Chang, Lee, & Toutanova, 2018) and Mixup to directly combine low-resource data with distant data in supervised learning, respectively. We train the framework for 3 epochs and employ entity-level and token-level precision (P), recall (R), and macro F1-score (F1) for evaluation. The primary symbols and their respective meanings are provided as shown in Table 1.

5.4. Overall performance analysis

To verify the effectiveness of CoTea, we comprehensively compare our method with existing state-of-the-art baselines on CoNLL-2003 (*news* domain) and NCBI-disease (*disease* domain) in high-resource settings and low-resource settings, respectively. Note that high-resource settings denote using fully supervised learning with full data for NER as a reference in contrast to low-resource ones.

Table 3
Experimental results on test sets compared to the state-of-the-art methods using low-resource data ($p < 0.05$ under t-test). + denotes using distant labels for original datasets rather than manual labels, but it failed in low-resource NER. ++ denotes SDNet used 53M external sentences while the other baselines used 1.5M sentences (2.8%).

	Model	CoNLL-2003			NCBI-disease		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
HIGH	BiLSTM-CRF (Huang et al., 2015)	92.78	87.43	90.02	85.47	74.32	79.51
	BERT-Linear (Kenton & Toutanova, 2019)	90.67	90.10	91.25	83.39	88.31	86.06
	BERT-CRF (Kenton & Toutanova, 2019)	89.67	90.85	90.26	85.25	87.29	86.27
	ACE (Wang, Jiang et al., 2021)	–	–	94.60	–	–	–
	LinkBERT (Yasunaga et al., 2022)	–	–	–	–	–	88.18
LOW	BERT-Linear (Kenton & Toutanova, 2019)	68.14	78.22	72.83	48.20	61.66	54.11
	BERT-CRF (Kenton & Toutanova, 2019)	73.45	77.17	75.31	51.75	47.71	49.73
	BOND (Liang et al., 2020) +	60.51	51.63	55.72	20.17	27.71	23.34
	LADA (Chen et al., 2020)	76.39	**83.63**	79.39	45.94	65.42	53.97
	RoSTER (Meng et al., 2021)	53.72	30.42	38.67	**60.02**	51.15	55.23
	NEEDLE (Jiang et al., 2021)	76.11	82.41	79.27	53.41	54.58	54.00
	LabelSem (Ma, Ballesteros et al., 2022)	16.49	14.00	15.14	2.27	1.69	1.94
	SDNet (Chen et al., 2022) ++	79.40	79.66	79.53	44.43	61.46	51.57
	Demons (Lee et al., 2022)	76.41	77.35	76.88	44.28	49.58	46.78
	MAML (Ma, Jiang et al., 2022)	72.88	75.90	74.36	48.62	52.97	51.24
	CoTea	**79.47**	81.31	**80.39**	45.24	**69.37**	**57.31**

Overall, CoTea significantly outperforms the baselines in low-resource settings and achieves competitive performance compared with high-resource baselines, as shown in Table 3. This demonstrates the effectiveness of CoTea, which improves data augmentation with collaborative teaching in low-resource settings. Specifically, our CoTea uses about 2.31% and 1.58% labeled data to obtain nearly 85% and 65% of the state-of-the-art high-resource performances on these datasets, respectively, which significantly improves the data efficiency in low-resource domains. Specifically, BOND and RoSTER achieve inferior performances than basic baselines in the extremely low-resource setting, which is due to the negative effect of external noisy data and their significant dependence on target-domain unlabeled data for self-training. LADA and NEEDLE obtain similarly competitive performances on CoNLL-2003 and NCBI-disease, demonstrating data augmentations with virtual examples and re-weighted external examples significantly contribute to the final performance. Demons also achieves promising results, showing that good demonstrations can save a lot of labor in low-resource environments. LabelSem and SDNet further incorporated label semantics and a unified label set for NER, but LabelSem, unfortunately, failed to enhance NER and achieved extremely low performances, which is probably because LabelSem cannot effectively learn knowledge from such low-resource data. In contrast, SDNet[2] consistently outperforms LabelSem by a large margin, which should be credited to additional training data (53M) and a unified multi-domain label set. Besides, MAML classified unseen entities regarding their distances with supporting examples, achieving competitive performances in these domains. This demonstrates the effectiveness of prototype-based methods and motivates us to explore task-specific metrics for this task.

5.5. Fine-grained performance analysis

As shown in Table 4, we further investigate the performances for fine-grained categories from token-level and entity-level perspectives, respectively. The token-level overall performance of CoTea slightly surpasses its entity-level overall performance, denoting that a few tokens of entity mentions are still incorrectly labeled and thus cause performance degradation. The multi-level fine-grained results of CoTea reflect a similar trend where the overall performance of CoTea is significantly associated with the category "MISC", which is actually composed of various implicit sub-categories except "PER", "ORG", and "LOC". Another potential limitation is that the result of "ORG" is slightly worse than the final performance regarding F1 score, which should be credited to some hard examples associated with "LOC" and thus may confuse model training. Although limited to the performances of some sub-categories, CoTea still contributes to the performance of low-resource NER, because it can fully leverage limited supporting examples for entity-related data and label refinement for collaborative teaching.

5.6. Ablation study

To explore each component's contribution, we conduct an extensive ablation study for CoTea. As shown in the bottom part of Table 5, CoTea significantly outperforms its ablation counterparts on these two datasets, demonstrating that all its components contribute to the final performances for NER in low-resource settings. Specifically, (1) "CoTea w/o RT&DT" achieves sub-optimal results than CoTea, demonstrating that our collaborative teachers can effectively guide models for low-resource NER; (2) The results of "CoTea w/o RT" significantly reduced, indicating mining entities with the recognition teacher is very critical for model training using noisy data; Besides, it performs worse than "CoTea RT&DT" on CoNLL-2003 because the discrimination teacher without the

[2] It only provides the pre-trained checkpoint based on additional training data (53M).

Table 4
Fine-grained results of CoTea on CoNLL-2003 test set, including token-level and entity-level performance.

Category	Token-level			Entity-level		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
PER	94.84	96.72	**95.78**	93.33	94.31	**93.82**
ORG	80.88	75.24	**78.06**	75.82	72.67	**74.24**
LOC	80.94	84.26	**82.60**	80.55	85.91	**83.23**
MISC	68.29	71.79	**70.04**	68.19	72.36	**70.28**
overall	81.24	82.00	**81.62**	79.47	81.31	**80.39**

Table 5
Ablation study on test sets; "w/" denotes "with" and "w/o" denotes "without". "MLM" denotes masked language modeling and "Mixup" denotes mixing supporting labeled data with distant data. "DT" and "RT" are the recognition teacher and the discrimination teacher, respectively.

Model	CoNLL-2003			NCBI-disease		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CoTea w/ MLM	74.72	78.21	76.47	18.93	10.31	14.62
CoTea w/ Mixup	73.67	76.29	74.98	44.79	46.15	45.47
CoTea w/o RT&DT	76.09	78.33	77.21	53.14	55.52	54.33
CoTea w/o DT	79.41	81.31	80.35	**45.67**	68.65	57.16
CoTea w/o RT	72.04	67.39	69.71	36.65	54.06	45.36
CoTea	**79.47**	**81.31**	**80.39**	45.24	**69.37**	**57.31**

Table 6
Case study with CoTea for low-resource NER. ▲ NEs and ▼ NEs denote distant and low-resource named entities, respectively. () denotes the similarity score of the retrieved NEs. We initialize the distant labels by matching all existing entities including low-resource NEs and for the augmented corpus. Other potential (underlined) entities are recognized by our method.

Van Boxmeer said Zywiec had its eye on Okocim.
PER ORG LOC MISC

▲ NEs (Score)	Related Sentences	▼ NEs
Jean-François van Boxmeer (194.80)	Jean-François van Boxmeer is a Belgian businessman.	
John Van Boxmeer (167.67)	John Martin Van Boxmeer is a Canadian former professional ice hockey player. He has also served extensively as a hockey coach with various teams from 1984 to the present.	Canadian
(...)		
Żywiec Beskids (58.81)	The Żywiec Beskids is a mountain range in the Outer Western Carpathians in southern Poland.	Poland
Okocim Brewery (42.37)	Okocim Brewery, in Brzesko in southeastern Poland, is a brewery founded in 1845. (...)	(...)
(...)		

recognition teacher may also introduce label noise, e.g., making wrong relabeling, thus reducing robustness; (3) "CoTea w/o DT" achieves sub-optimal results than CoTea, demonstrating that the discrimination teacher contributes to the final performance of our recognition models. Besides, it performs better than "CoTea w/o RT", indicating that mining potential entities from distant data provides much informative guidance for training NER models; (4) We explored masked language modeling (MLM) and mixup strategies, but with poor performance. A potential reason is that the distribution of open-domain remote data is different from that of target-domain data and thus causes the model overlap to noise. This highlights the need to refine training methods for addressing distribution disparities. In summary, CoTea benefits low-resource NER through collaborative teaching with a divide-and-conquer strategy, obtaining better performance and robustness.

5.7. Case study

To analyze how CoTea augments NER in the low-resource domain, we conduct a case study on our framework, as shown in Table 6. Given a labeled sentence "*Van Boxmeer said Zywiec had its eye on Okocim*" containing three named entities, i.e., "*Van Boxmeer*" (PER), "*Zywiec*" (ORG) and "*Okocim*" (ORG), we use these entities to recall more related entities and descriptions. For example, the entity "*Van Boxmeer*" in red is used as a query to find a set of distant entities and their descriptions for data augmentation, e.g., the entity "*Jean-François van Boxmeer* (194.80)" (The higher the entity score, the more relevant it is) and its description "*Jean-François van Boxmeer is a Belgian businessman*". In addition, distant entities (▲ NEs) and low-resource entities (▼ NEs) are combined to initialize the distant labels for description sentences and the other (underlined) potential entities are recognized and refined using our recognition and teacher networks. This verifies the effectiveness of CoTea to obtain distant data for the model training.

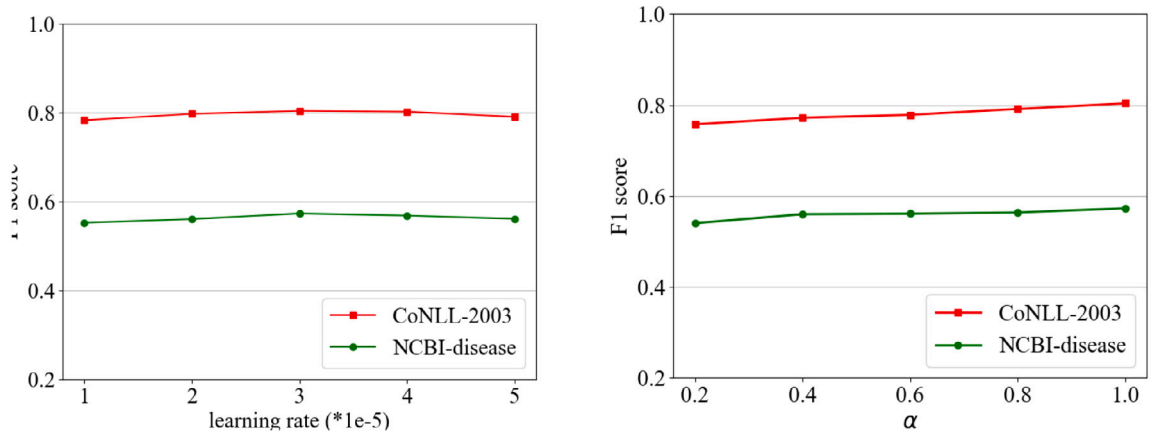

Fig. 3. Parameter sensitivity analysis of CoTea.

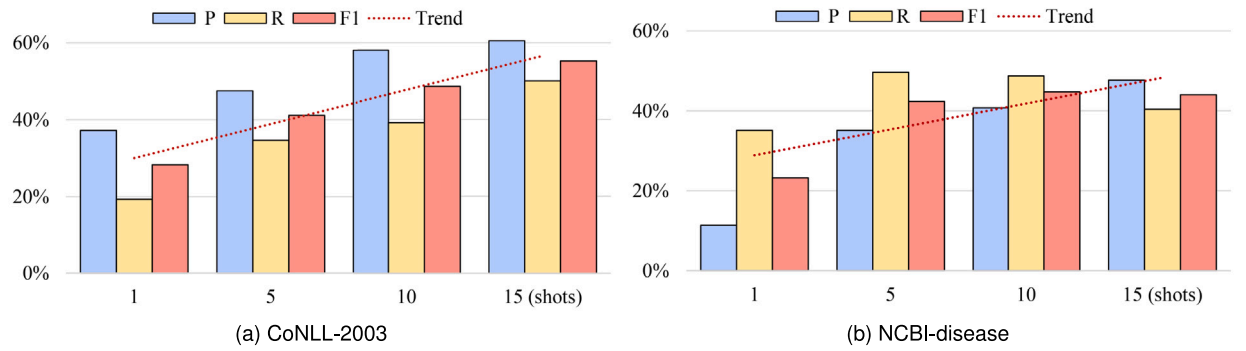

Fig. 4. The impact of N-shot supporting labeled data for CoTea.

5.8. Parameter sensitivity analysis

To verify the effect of parameters on the effectiveness of the CoTea, we conduct sensitivity analysis by varying the primary parameters, i.e., learning rate and trade-off parameter α. The learning rate is a hyperparameter that determines the step size at each iteration while moving toward a minimum of a loss function in algorithms, and the trade-off parameter α controls the contributions of loss terms in the overall loss. To study uncertainty in the output of our proposed model, we employ single-parameter sensitivity analysis by varying one parameter while fixing the others each time. As shown in Fig. 3, CoTea keeps high performance while the parameter varies on the two benchmark datasets, demonstrating that collaborative teaching contributes to low-resource NER, and further verifies the effectiveness and robustness of our proposed framework. The results also open avenues for investigating the potential of collaborative teaching in other tasks. As demonstrated by the analysis, the stability and effectiveness of CoTea make it a promising tool for tackling challenges in low-resource conditions, contributing to advancing the field.

5.9. Parameter analysis on few-shot settings

To analyze the impact of N shots supporting labeled data on CoTea and explore the causes behind characteristics, we further conduct few-shot experiments on these datasets, where only N-shot labeled data is available for training without additional valid data in such extreme settings.

Fig. 4 shows the trend that the entity-level F1 score (dotted line) of CoTea consistently increases with more supporting labeled data on these datasets, indicating that supporting data is important for model training and that more labeled data contributes to the final performance. Note that the token-level performance of CoTea shows a similar trend regarding F1 score. Specifically, CoTea achieves inferior performance when the labeled data is extremely rare, i.e., 1 shot. This indicates that we should provide a certain amount of supporting labeled data for better performance. In addition, CoTea can achieve much better performances with more than 5 shots supporting examples, which demonstrates the effectiveness and robustness of CoTea in few-shot settings.

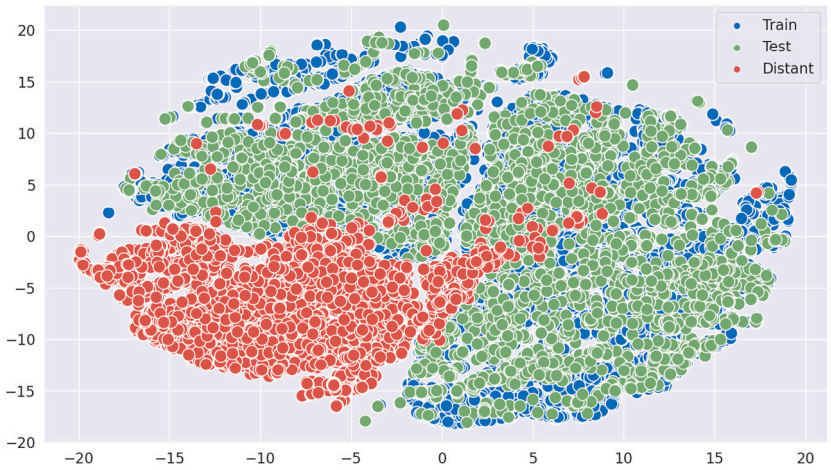

Fig. 5. t-SNE plot of sentence embeddings for the target (i.e., Train and Test) and open (i.e., Distant) domains, respectively. Distant data is retrieved from different sources. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

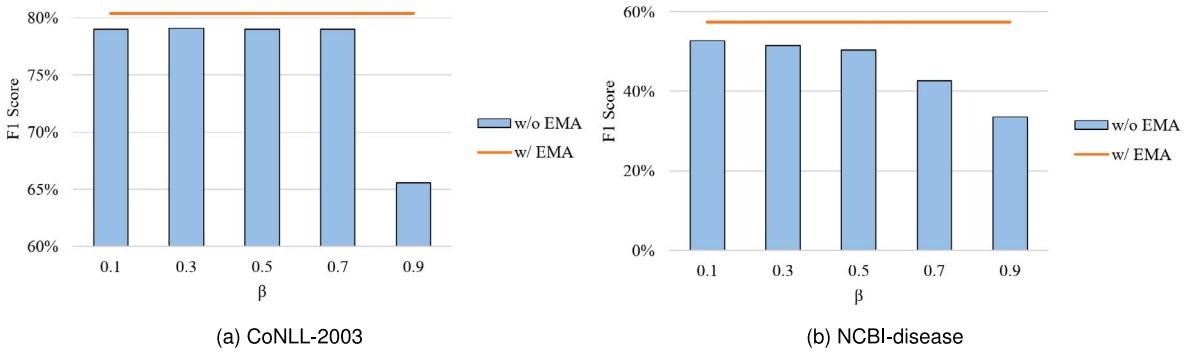

Fig. 6. Quantitative analysis of EMA in CoTea. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.10. Visualization of data distribution

To further analyze the superiority of the proposed method in low-resource settings, we visualize the data distribution as shown in Fig. 5, including the sentence embeddings of target-domain ("Train" and "Test") and distant data ("Distant"). We can conclude that (1) The distributions of the train (blue) and test (green) data are overlapped with each other, indicating that they share more similar features so as to train models on the target domain; (2) The distant data (red) shows a different distribution in the space comparing with target-domain data, i.e., the train and test data. This suggests that we should use them to learn general entity pattern features, rather than simply obfuscating them; (3) Furthermore, although not always available, more labeled data on the target domain is likely to help the final performance. Therefore, we introduce a more general divide-and-conquer framework to deal with the low-resource and external data, which can utilize external knowledge and alleviate the impact of noise.

5.11. Quantitative analysis of EMA

To investigate the influence of EMA in our proposed method, quantitative experiments were conducted on the CoNLL-2003 and NCBI-disease datasets. we employed single-parameter sensitivity analysis by systematically selecting fixed values from the range [0.1, 0.3, …, 0.9] for the parameter β in Eq. (13). As shown in Fig. 6, the blue bars represent the performances of CoTea with different fixed β but without EMA, denoted as 'w/o EMA.' In contrast, the orange line depicts the performance of CoTea with EMA, denoted as 'w/ EMA,' facilitating analysis and comparison. The experimental results indicate that CoTea with EMA consistently outperforms its variants with fixed parameters regarding F1 score. This superiority is attributed to the noise tolerance provided by EMA, thereby enhancing the overall performance of CoTea. Furthermore, CoTea's optimal performance is achieved through a careful balance between the student and teacher models, as high β values (e.g., $\beta = 0.9$) yield poor results. In summary, the incorporation of EMA enhances the effectiveness and robustness of the teacher model, with a ramp-up strategy automatically determining an appropriate β for effective ensemble weighting.

6. Conclusion

This paper presents a novel collaborative teaching for low-resource NER, which provides an effective paradigm to alleviate data scarcity in low-resource settings and improve the performance of this task. Specifically, this automatically retrieves entity-related data using existing knowledge and unifies the different pre-trained language models as collaborative teachers to generate refined labels for distant data. In addition, we explicitly take a divide-and-conquer strategy to re-label the entity- and non-entity parts, respectively and eventually reach the optimal equilibrium point of teachers. Extensive experimental results demonstrate that our CoTea outperforms existing baselines in low-source settings and also achieves comparable results with the state-of-the-art baselines in standard supervised settings. For future work, we consider utilizing the proposed framework for other tasks in low-resource settings, which can significantly save the annotation cost and improve the task performance in a new domain.

CRediT authorship contribution statement

Zhiwei Yang: Conceptualization, Methodology, Software, Writing – original draft. **Jing Ma:** Conceptualization, Resources, Project administration. **Kang Yang:** Software, Validation. **Huiru Lin:** Investigation, Proof-reading, Resources. **Hechang Chen:** Supervision, Funding acquisition. **Ruichao Yang:** Review, Proof-reading. **Yi Chang:** Supervision, Funding acquisition.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

We truly thank the reviewers and editors for their great effort in our submission. This work is partially supported by National Natural Science Foundation of China through grants (No. U2341229, No. 62206233), Key R&D Program of the Ministry of Science and Technology, China (2023YFF0905400), the International Cooperation Project of Jilin Province, China (20220402009GH), and Hong Kong RGC ECS (22200722).

References

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638–1649).

Asghari, M., Sierra-Sosa, D., & Elmaghraby, A. S. (2022). BINER: A low-cost biomedical named entity recognition. *Information Sciences*, *602*, 184–200.

Cao, A., Luo, Y., & Klabjan, D. (2021). Open-set recognition with Gaussian mixture variational autoencoders. In *Proceedings of the AAAI conference on artificial intelligence. Vol. 35, no. 8* (pp. 6877–6884).

Chen, S., Aguilar, G., Neves, L., & Solorio, T. (2021). Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 5346–5356).

Chen, J., Liu, Q., Lin, H., Han, X., & Sun, L. (2022). Few-shot named entity recognition with self-describing networks. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 5711–5722).

Chen, J., Wang, Z., Tian, R., Yang, Z., & Yang, D. (2020). Local additivity based data augmentation for semi-supervised NER. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 1241–1251).

Cheng, D., Zhang, L., Bu, C., Wu, H., & Song, A. (2023). Learning hierarchical time series data augmentation invariances via contrastive supervision for human activity recognition. *Knowledge-Based Systems*, *276*, Article 110789.

Cui, L., Wu, Y., Liu, J., Yang, S., & Zhang, Y. (2021). Template-based named entity recognition using BART. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 1835–1845).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, *47*, 1–10.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., et al. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 601–610).

Fang, X., Li, J., Shang, L., Jiang, X., Liu, Q., & Yeung, D.-Y. (2022). Controlled text generation using dictionary prior in variational autoencoders. In *Findings of the association for computational linguistics: ACL 2022* (pp. 97–111).

Fritzler, A., Logacheva, V., & Kretov, M. (2019). Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing* (pp. 993–1000).

Geng, R., Chen, Y., Huang, R., Qin, Y., & Zheng, Q. (2023). Planarized sentence representation for nested named entity recognition. *Information Processing & Management*, *60*(4), Article 103352.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 2545–2568).

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, *194*, 28–61.

Huang, J., Li, C., Subudhi, K., Jose, D., Balakrishnan, S., Chen, W., et al. (2021). Few-shot named entity recognition: A comprehensive study. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10408–10423).

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Jiang, H., Zhang, D., Cao, T., Yin, B., & Zhao, T. (2021). Named entity recognition with small strongly labeled and large weakly labeled data. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 1775–1789).

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., et al. (2023). Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10124–10134).

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).

Laine, S., & Aila, T. (2016). Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint:1603.01360.

Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W. X., & Wen, J.-R. (2022). Complex knowledge base question answering: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Lee, D.-H., Kadakia, A., Tan, K., Agarwal, M., Feng, X., Shibuya, T., et al. (2022). Good examples make a faster learner: Simple demonstration-based learning for low-resource NER. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 2687–2700).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880).

Li, Z., Hu, C., Guo, X., Chen, J., Qin, W., & Zhang, R. (2022). An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 170–179).

Li, Q., Li, H., Ji, H., Wang, W., Zheng, J., & Huang, F. (2012). Joint bilingual name tagging for parallel corpora. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 1727–1731).

Li, D., Liu, Y., & Song, L. (2022). Adaptive weighted losses with distribution approximation for efficient consistency-based semi-supervised learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(11), 7832–7842.

Li, J., Sun, A., Han, J., & Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70.

Li, C., Yao, K., Wang, J., Diao, B., Xu, Y., & Zhang, Q. (2022). Interpretable generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence. Vol. 36, no. 2* (pp. 1280–1288).

Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., et al. (2020). Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1054–1064).

Lin, H., Lu, Y., Han, X., Sun, L., Dong, B., & Jiang, S. (2019). Gazetteer-enhanced attentive neural networks for named entity recognition. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* EMNLP-IJCNLP, (pp. 6232–6237).

Liu, Z., Jiang, F., Hu, Y., Shi, C., & Fung, P. (2021). NER-BERT: A pre-trained model for low-resource entity tagging. arXiv preprint arXiv:2112.00405.

Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., et al. (2021). Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 13452–13460).

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of 7th international conference on learning representations* (pp. 1–8).

Ma, J., Ballesteros, M., Doss, S., Anubhai, R., Mallya, S., Al-Onaizan, Y., et al. (2022). Label semantics for few shot named entity recognition. In *Findings of the association for computational linguistics: ACL 2022* (pp. 1956–1971).

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (pp. 1064–1074).

Ma, T., Jiang, H., Wu, Q., Zhao, T., & Lin, C.-Y. (2022). Decomposed meta-learning for few-shot named entity recognition. In *Findings of the association for computational linguistics: ACL 2022* (pp. 1584–1596).

Ma, Y., Zhang, Y., Sangaiah, A. K., Yan, M., Li, G., & Wang, T. (2023). Active learning for name entity recognition with external knowledge. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Meng, Y., Zhang, Y., Huang, J., Wang, X., Zhang, Y., Ji, H., et al. (2021). Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10367–10378).

Niu, Z., Anitescu, M., & Chen, J. (2023). Graph neural network-inspired kernels for Gaussian processes in semi-supervised learning. In *International conference on learning representations*.

Nozza, D., Manchanda, P., Fersini, E., Palmonari, M., & Messina, E. (2021). LearningToAdapt with word embeddings: Domain adaptation of named entity recognition systems. *Information Processing & Management*, *58*(3), Article 102537.

Polyak, B. T., & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, *30*(4), 838–855.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., et al. (2022). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rijhwani, S., Zhou, S., Neubig, G., & Carbonell, J. G. (2020). Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8118–8123).

Sang, E. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on natural language learning At HLT-NAACL 2003* (pp. 142–147).

Speer, R., Havasi, C., et al. (2012). Representing general relational knowledge in conceptnet 5. In *LREC. Vol. 2012* (pp. 3679–3686).

Sui, D., Zeng, X., Chen, Y., Liu, K., & Zhao, J. (2023). Joint entity and relation extraction with set prediction networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Tang, Y., Zhang, L., Wu, H., He, J., & Song, A. (2022). Dual-branch interactive networks on multichannel time series for human activity recognition. *IEEE Journal of Biomedical and Health Informatics*, *26*(10), 5223–5234.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*: *vol.30*.

Tian, Y., Zhang, L., Sun, J., Yin, G., & Dong, Y. (2022). Consistency regularization teacher–student semi-supervised learning method for target recognition in SAR images. *The Visual Computer*, *38*(12), 4179–4192.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269.

Wan, S., Zhan, Y., Liu, L., Yu, B., Pan, S., & Gong, C. (2021). Contrastive graph poisson networks: Semi-supervised learning with extremely limited labels. *Advances in Neural Information Processing Systems*, *34*, 6316–6327.

Wang, X., Dou, S., Xiong, L., Zou, Y., Zhang, Q., Gui, T., et al. (2022). MINER: Improving out-of-vocabulary named entity recognition from an information theoretic perspective. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 5590–5600).

Wang, D., Fan, H., & Liu, J. (2021). Learning with joint cross-document information via multi-task learning for named entity recognition. *Information Sciences*, *579*, 454–467.

Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., et al. (2021). Automated concatenation of embeddings for structured prediction. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 2643–2660).

Wang, X., Kihara, D., Luo, J., & Qi, G.-J. (2021). EnAET: A self-trained framework for semi-supervised and supervised learning with ensemble transformations. *IEEE Transactions on Image Processing*, *30*, 1639–1647.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., et al. (2013). Ontonotes release 5.0 ldc2013t19. In *Linguistic data consortium*. Philadelphia, PA.

Xu, S., Zhang, L., Tang, Y., Han, C., Wu, H., & Song, A. (2023). Channel attention for sensor-based activity recognition: Embedding features into all frequencies in DCT domain. *IEEE Transactions on Knowledge and Data Engineering*.

Yang, X., Song, Z., King, I., & Xu, Z. (2022). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering (Early Access)*.

Yasunaga, M., Leskovec, J., & Liang, P. (2022). LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 8003–8016).

Ye, F., & Bors, A. G. (2021). Lifelong teacher-student network learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6280–6296.

Yu, N., Liu, G., Dundar, A., Tao, A., Catanzaro, B., Davis, L. S., et al. (2021). Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6731–6742).

Zevallos, R., Ortega, J., Chen, W., Castro, R., Bel, N., Toshio, C., et al. (2022). Introducing QuBERT: A large monolingual corpus and BERT model for southern quechua. In *Proceedings of the 3rd workshop on deep learning for low-resource natural language processing* (pp. 1–13).

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., et al. (2021). Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, *34*, 18408–18419.

Zhang, X., Yu, B., Liu, T., Zhang, Z., Sheng, J., Mengge, X., et al. (2021). Improving distantly-supervised named entity recognition with self-collaborative denoising learning. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10746–10757).

Zhu, X., Li, Z., Wang, X., Jiang, X., Sun, P., Wang, X., et al. (2022). Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Zhiwei Yang is currently an assistant professor at Jinan University. Before that, he was a Ph.D. student at the College of Computer Science and Technology, Jilin University (JLU), Changchun, Jilin Province, China, and a full-time exchange Ph.D. student in Department of Computer Science, Hong Kong Baptist University (HKBU), Hong Kong. His research interests include information extraction, rumor detection, and artificial intelligence. His publications include AAAI, IJCAI, EMNLP, COLING, TNNLS, IP&M, Neurocomputing and et al.. He has been serving as a reviewer for Neural Networks, Neurocomputing, KSEM, and et al.

Jing Ma received the Ph.D. degree from The Chinese University of Hong Kong (CUHK) in 2020. She is currently an Assistant Professor at the department of Computer Science, Hong Kong Baptist University (HKBU). Her current research interests include Natural Language Processing, Information Verification, and Social Media Analytics. She has been serving on the program committee of several international conferences, including: IJCAI, AAAI, WWW, CIKM, ACL, and EMNLP.

Kang Yang is presently a first-year master student at the School of Artificial Intelligence, Jilin University (JLU), Changchun, Jilin Province, China. His research fields include information extraction and artificial intelligence.

Huiru Lin is currently a lecturer at the Institute of Physical Education, Jinan University. Before that, she was a Ph.D. student at Central China Normal University, Wuhan, Hubei Province, China. Her research interests include data analysis and cognitive neuroscience.

Hechang Chen is an associate professor at the School of Artificial Intelligence, Jilin University (JLU), China. He received his PhD degree from the College of Computer Science and Technology, Jilin University, in December 2018. From November 2015 to December 2016, he was a joint training Ph.D. student with the Department of Computer Science, University of Illinois at Chicago (UIC), Chicago, IL, USA. From July 2017 to January 2018, he was a visiting Ph.D. student with the Department of Computer Science, Hong Kong Baptist University (HKBU), Hong Kong, China. He has published over 60 articles in many top international conferences and journals, e.g., IEEE TPAMI, TKDE, TNNLS, TKDD, NeurIPS, AAAI, IJCAI, SIGIR, WWW, ICDE, etc. His research interests include machine learning, data mining, reinforcement learning, complex systems, and knowledge engineering.

Ruichao Yang received the bachelor's degree from Jinlin University in 2015 and master's degree from Peking University in 2018 respectively. She is currently pursuing the Ph.D. degree at the department of Computer Science, Hong Kong Baptist University. Her current research interests include Natural Language Processing, Rumor Verification, Fake News Detection, Misinformation Detection and Social Media Analytics.

Yi Chang is currently the Dean of the School of Artificial Intelligence, Jilin University, Changchun, China. He became a Chinese National Distinguished Professor in 2017 and the ACM Distinguished Scientist in 2018. Before joining academia, he was the Technical Vice President at Huawei Research America, in charge of knowledge graph, question answering, and vertical search projects. Before that, he was the Research Director of Yahoo Labs/Research, USA, from 2006 to 2016, in charge of search relevance of Yahoo's web search engine and vertical search engines. He is the author of two books and more than 100 papers in top conferences or journals. His research interests include information retrieval, data mining, machine learning, natural language processing, and artificial intelligence.

Dr. Chang won the Best Paper Award on ACM KDD 2016 and ACM International WSDM Conference 2016. He has served as one of the Conference General Chairs for ACM International WSDM Conference 2018 and International ACM SIGIR Conference on Research and Development in Information Retrieval 2020. He is an Associate Editor of IEEE Transactions on Knowledge and Data Engineering.