Collaborative Teaching for Low-resource Named Entity Recognition with A Divide-and-conquer Strategy

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

[a]*College of Computer Science and Technology, Jilin University, Changchun, China*
[b]*Department of Computer Science, Hong Kong Baptist University, Hong Kong, China*
[c]*School of Artificial Intelligence, Jilin University, Changchun, China*
[d]*Central China Normal University, Wuhan, China*
[e]*International Center of Future Science, Jilin University, Changchun, China*

Abstract

Low-resource named entity recognition (NER) aims to identify entity mentions when training data is scarce, which leads to inferior performances of most existing methods in low-resource domains such as disease. Since annotating more data is labor-intensive and time-consuming, recent approaches resort to distant data with manual dictionaries for training. However, 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. The mechanism mines potential entities from non-entity tokens with a recognition teacher and then refines entity labels with another prompt-based discrimination teacher. As a result, CoTea can foster model training by optimizing distant labels iteratively in a divide-and-conquer manner. Experimental results on two benchmark datasets demonstrate

*Corresponding author

Email addresses: yangzw18@mails.jlu.edu.cn (Zhiwei Yang), majing@comp.hkbu.edu.hk (Jing Ma), yangkang22@mails.jlu.edu.cn (Kang Yang), chenhc@jlu.edu.cn (Hechang Chen), huiru@mails.ccnu.edu.cn (Huiru Lin), csrcyang@comp.hkbu.edu.hk (Ruichao Yang), yichang@jlu.edu.cn (Yi Chang)

that CoTea outperforms a set of state-of-the-art baselines in low-resource settings.

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

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 [1], question answering [2], knowledge graph construction [3], etc. Existing supervised approaches for NER achieved superior performances in high-resource settings, i.e., training on a large amount of labeled data [4]. 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 methods may suffer from data scarcity and not easily identify entities in low-resource settings [5, 6, 7]. 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 [8, 9]. 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 [10, 11], 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) [12, 13] or existing high-resource labeled datasets [14, 15, 16]. 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 [17] but *LOC* in CoNLL-2003 [18]. 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 [19, 20, 21]. However, they may suffer from dictionary quality and thereby overfit to noise.

As shown in Figure 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

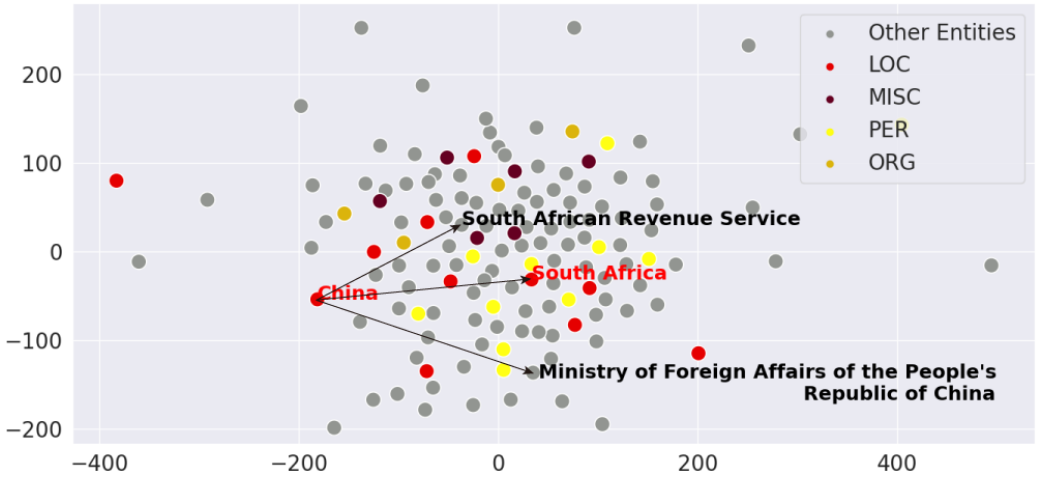

Figure 1: t-SNE plot of entity embeddings from 100 sentences randomly selected from distant data-augmented CoNLL-2003. The entity labels are provided using dictionaries and "Other Entities" denotes the entities that aren't matched.

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 [22, 23]. 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 [24] and BART [12] 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.

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 [8], as follows.

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 [25, 26], CSEmb used contextual string embeddings for sequence labeling [27], CrossNER extracted information by a joint cross-document BiLSTM and multi-task learning [28], BERT-Linear/CRF became a popular paradigm with the pre-trained language model (PLM) for encoding [24], ACE automated the process of finding better concatenations of different embeddings [29], and LinkBERT pre-trained a language model by leveraging links between documents [30]. Although traditional fully supervised methods have achieved promising performances, they significantly relied on large-scale labeled data for training [4]. 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 [31, 32] or zero-shot learning [33]. 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 [10], and metric-based NER encoded label to help determine entity categories [11]. 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 [12], demonstration based NER integrated prompt into the input with task demonstrations [34], cross-lingual NER used teacher-student distillation training to align high-source languages and low-source languages [14], cross-domain augmentation methods

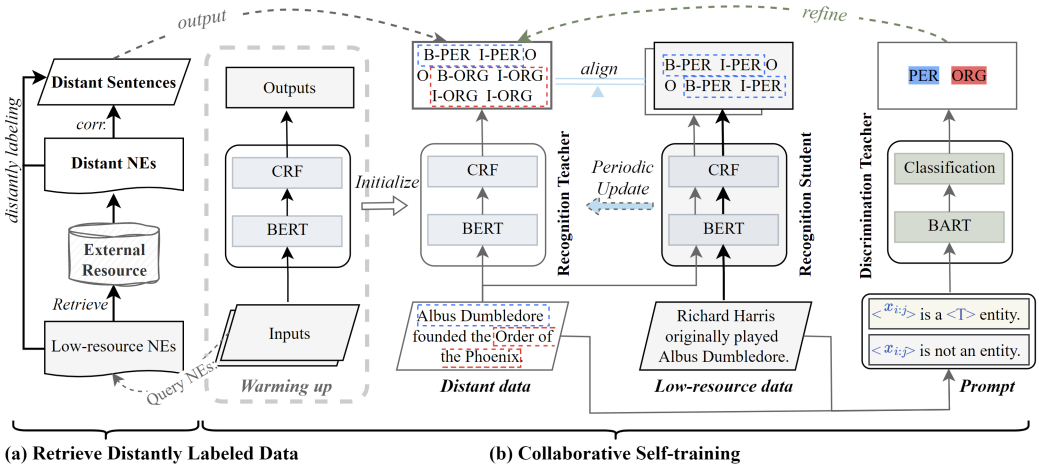

Figure 2: An overview of CoTea. In the phase (a), distant data is obtained with initial labels. In the phase (b), the recognition teacher (RT) and discrimination teacher (DT) cooperate to refine these distant labels. RT is initialized by warm-up training on low-resource labeled data and updated by consecutive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

transformed the data representation from high-resource domains into the low-resource domains [15, 16]. 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 [19], RoSTER combined noise-robust learning with augmented self-training for NER [21], NEEDLE continually pre-trained on large unlabeled open-domain data and target-domain data based on manual dictionaries [35], LADA adopted local additivity based data augmentation to create virtual samples [36], and dictionaries are also used to enhance NER [22, 23]. Although being effective, they are limited in entity coverage and rarely attempt to develop a generalized method for low-resource NER.

3. Problem statement

Formally, a sentence is represented as $\mathbf{X} = \{x_1, ..., x_t, ..., x_{|X|}\}$, where $|\cdot|$ denotes the sequence length, each word x_i is associated with one entity label $y_t \in \mathbf{Y} = \{y_1, ..., y_t, ..., y_{|\mathbf{X}|}\}$ based on BIO schema [37]. 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 $x_{i:j} = \{x_i, ..., x_j\}(1 \le i \le j \le |\mathbf{X}|)$, which is determined by $B-T$ or

($B-T$ and $I-T$) in $y_{i:j} = \{y_i, ..., y_j\}$. 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. By denoting the distant labels of $\mathbf{X}^D$ 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 $f(\mathbf{X}^L, \mathbf{X}^D) \rightarrow (\hat{\mathbf{Y}}^L, \hat{\mathbf{Y}}^D)$.

4. Collaborative teaching framework

Our core idea is to utilize a few labeled examples to obtain a set of distantly labeled data and then train NER models on such augmented datasets 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 Figure 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 [38, 39, 40], 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 [22, 23], 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 Figure 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

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

labels can be used for self-training and thereby foster student model training. Taking the sentence in Table 5 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 $\mathbf{X} = \{x_1, ..., x_t, ..., x_{|\mathrm{X}|}\}$, we utilize pre-trained language models, e.g., BERT [24] and BART [41], 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 student networks

As averaging model weights over training can obtain better model [42], 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 Figure 2 (b, left), we use the conditional random field (CRF) on the BERT representations for sequence labeling. CRF can jointly consider neighboring generic labels for decoding the best chain of labels [43], e.g., $I-T$ can not follow O in the token label sequence of input sentences based on BIO schema. For a given sentence $\mathbf{X}$, we denote its representations as $\mathbf{h}$ and its corresponding predicted label sequence as $\hat{\mathbf{Y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{|\mathrm{X}|})$. Thus, the output conditional probability of CRF is defined as follows:

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

where y represents a label chosen from all possible labels $\mathbf{Y_x}$ 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:

$$\mathcal{L}_{RT} = -\sum_{t} \log(p(\hat{\mathbf{Y}}|\mathbf{h}; \mathbf{W}, \mathbf{b})) \tag{2}$$

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

$$\hat{\mathbf{Y}}^* = \arg\max_{\hat{y} \in \mathbf{Y_x}} p(\hat{\mathbf{Y}}|\mathbf{h}; \mathbf{W}, \mathbf{b}) \tag{3}$$

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 [44]. For warming up, 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 Figure 1. For mining entities from non-entity tokens, we use the outputs of the latest recognition teacher to update their labels as follows:

$$\hat{y}_t = \begin{cases} \hat{y}_t^D & \text{, if} \quad \hat{y}_t^0 = \text{O} \\ \hat{y}_t^0 & \text{, } otherwise \end{cases} \tag{4}$$

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 (Figure 2 (a)) or the outputs of the recognition teacher (Figure 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 relabeling.

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, i.e., given an input sequence pair $(\mathbf{X}, \hat{\mathbf{Y}})$, the target sequence $\mathbf{T}_{y_c, x_{i:j}} = \{t_1, ..., t_{|\mathrm{T}|}\}$ 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*". Denoting the one-hot label of entity type T as $y_c (c \in [0, ..., k])$. We also randomly select spans $x_{i:j}$ that contain non-entity labels or different categorical labels as negative examples for training. 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., "$\langle AlbusDumbledore \rangle$ *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:

$$p = \mathrm{Softmax}(\mathbf{W}'\mathbf{h}'_0 + \mathbf{b}') \tag{5}$$

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:

$$\mathcal{L}_{DT} = y_c \log(p) \tag{6}$$

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 $\mathrm{Max}(p)$. For each span $x_{i:j}$ in $\mathbf{X}^D$, the discrimination teacher network g_{θ_t} generates its refined distant labels as follows:

$$\hat{y}'_{i:j} = g_{\theta_t}(x'_{i:j}) \tag{7}$$

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:

$$\hat{y}_t = \begin{cases} \hat{y}'_t & , \text{if} \quad \hat{y}^0_{i:j} = \text{NE} \ \& \ t \in [i:j] \\ \hat{y}^0_t & , otherwise \end{cases} \tag{8}$$

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. Figure 2 (b) illustrates the overview of collaborative self-training, where CoTea combines distant labels from named entity mining and label refinement using Equation (4) and Equation (8), 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:

$$\mathcal{L}_{TS} = -\frac{1}{|\mathrm{X}|}\sum_{t}^{|\mathrm{X}|}(\hat{y}_t^s - \hat{y}_t)^2 \tag{9}$$

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) [45] of the student and teacher weights with a ramp-up strategy for promotion as follows:

$$\theta_t^{'} = \beta\theta_t + (1-\beta)\theta_s \tag{10}$$

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 [46, 45]. 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 Equation (2) and Equation (6), 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:

$$\mathcal{L}_{all} = \mathcal{L}_{RT} + \alpha\mathcal{L}_{TS} \tag{11}$$

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.

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 [18] and NCBI-disease [47], 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 [15, 35, 21]. 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 [18] 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 [47] 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 1.

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. (2) and Eq. (6); *#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. (3) ;

  {*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. (4) *#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. (5) 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. (8); *#refining*

  $\hat{\mathbf{Y}}^D \leftarrow \{\hat{y}_t\}_{t=1}^{|\mathbf{X}^D|}$;

  {*Collaborative self-training*}

  $\mathcal{L}_{TS} \leftarrow$ Eq. (4, 8) for alignment;

  if *global_steps % 20 == 0* **then**

    $\beta = \exp(-5 * (1 - \frac{\epsilon}{\tau})^2)$;

    $\theta'_t \leftarrow \beta$, Eq. (10) using EMA;

{*Jointly Optimization*}

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

Table 1: 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" denotes the automatically retrieved data.

Dataset	Train	Valid	Test	Distant	#Per
CoNLL-2003 (HIGH)	14,041	3,250	3,453	-	2.31%
CoNLL-2003 (LOW)	200	200	3,453	5,475	
NCBI-disease (HIGH)	5,424	923	940	-	1.58%
NCBI-disease (LOW)	50	50	940	409	

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** [19]: This leveraged multi-source gazetteers and BERT to improve open-domain NER with distant supervision.
- **LADA** [36]: This proposed a local additivity-based method for semi-supervised NER, which generated augmentations with different sentence structures.
- **RoSTER** [21] : This proposed a self-training method based on a generalized noise-robust loss using only distantly-labeled data.
- **NEEDLE** [35]: This trained deep NER models over a weighted combination of manually labeled and distantly labeled data.
- **LabelSem** [11]: This learned to match the representations of named entities and labels based on two BERT encoders.
- **Demons** [34]: This proposed demonstration-based NER with in-context learning based on example sampling and template construction.
- **SDNet** [15]: This proposed to describe mentions using a universal concept set for few-shot NER.
- **MAML** [10]: This presented a decomposed meta-learning approach for span detection and entity typing.

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

- **BiLSTM-CRF** [26]: This proposed a neural architecture based on a bidirectional LSTM with a sequential conditional random field.
- **BERT-Linear/CRF** [24]: This used BERT with a linear or CRF classifier.
- **ACE** [29]: This work automated the process of finding better concatenations of different level embeddings to enhance NER performance.

Table 2: 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 [25]	92.78	87.43	90.02	85.47	74.32	79.51
	BERT-Linear [24]	90.67	90.10	91.25	83.39	88.31	86.06
	BERT-CRF [24]	89.67	90.85	90.26	85.25	87.29	86.27
	ACE [29]	-	-	94.60	-	-	-
	LinkBERT [30]	-	-	-	-	-	88.18
LOW	BERT-Linear [24]	68.14	78.22	72.83	48.20	61.66	54.11
	BERT-CRF [24]	73.45	77.17	75.31	51.75	47.71	49.73
	BOND [19] +	60.51	51.63	55.72	20.17	27.71	23.34
	LADA [36]	76.39	**83.63**	79.39	45.94	65.42	53.97
	RoSTER [21]	53.72	30.42	38.67	**60.02**	51.15	55.23
	NEEDLE [35]	76.11	82.41	79.27	53.41	54.58	54.00
	LabelSem [11]	16.49	14.00	15.14	2.27	1.69	1.94
	SDNet [15] ++	79.40	79.66	79.53	44.43	61.46	51.57
	Demons [34]	76.41	77.35	76.88	44.28	49.58	46.78
	MAML [10]	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**

- **LinkBERT** [30]: This pre-trained a language model by leveraging links between documents for NER.

5.3. Experimental settings

As shown in Table 1, 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 [24] and BART [41], 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 [48] 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 [41] 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 fine-tune BERT with a fill-in-the-blank task [49] and Mixup to fuse 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.

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.

Overall, CoTea significantly outperforms the baselines in low-resource settings and achieves competitive performance compared with high-resource baselines, as shown in Table 2. 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 effec-

Table 3: Fine-grained results of our 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**

tively 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 3, 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.

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

Table 4: 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**

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 4, 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 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 also 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. In summary, CoTea can benefit low-resource NER through collaborative teaching with a divide-and-conquer strategy, obtaining better performance and robustness.

Table 5: 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 south-eastern Poland, is a brewery founded in 1845. (...)	(...)

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 5. 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.

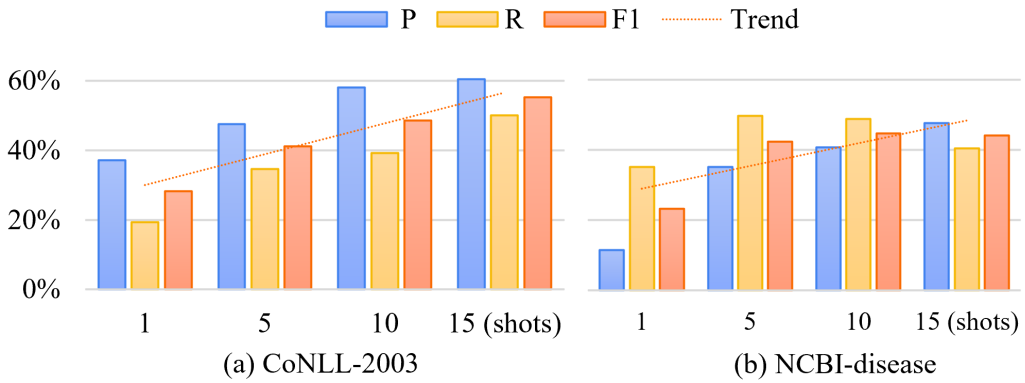

Figure 3: The impact of N-shot supporting labeled data for CoTea.

5.8. 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.

Figure 3 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.

5.9. 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 Figure 4, 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

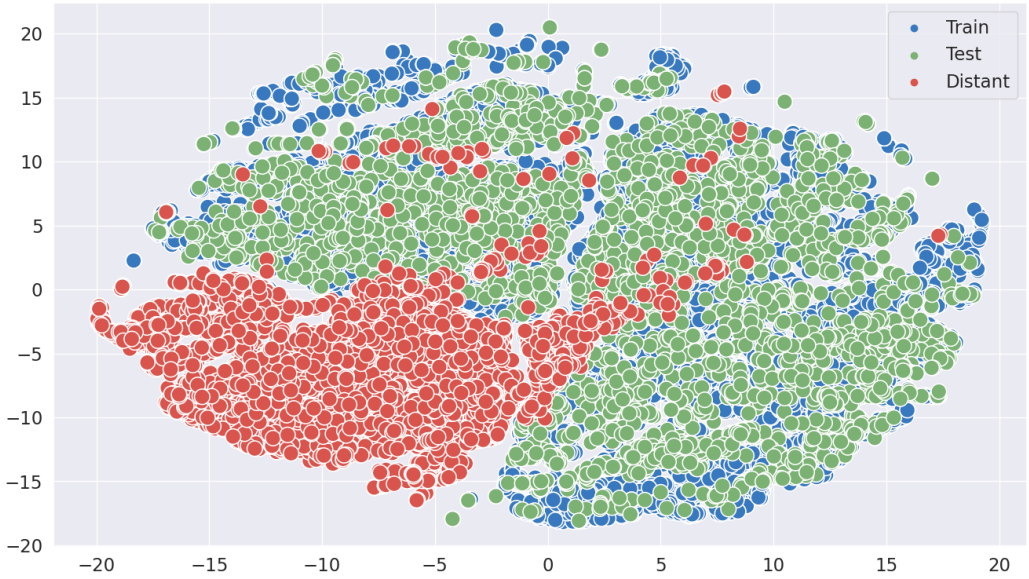

Figure 4: t-SNE plot of sentence embeddings for the target (i.e., Train and Test) and open (i.e., Distant) domains, respectively. Distant data is retrieved from different sources.

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.

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

Declaration of Completing 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.

Acknowledgement

We truly thank the reviewers and editors for their great effort in our submission. This work was partially supported by National Natural Science Foundation of China through grants No.61976102, No.U19A2065, No.61902144, and No.61902145. This work was partly supported by HKBU One-off Tier 2 Start-up Grant (RCOFSGT2/20-21/SCI/004), Hong Kong RGC ECS (22200722).

References

[1] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 50–61.

[2] S. Shah, A. Mishra, N. Yadati, P. P. Talukdar, Kvqa: Knowledge-aware visual question answering, in: Proceedings of the AAAI conference on artificial intelligence, 2019, pp. 8876–8884.

[3] X. Zheng, B. Wang, Y. Zhao, S. Mao, Y. Tang, A knowledge graph method for hazardous chemical management: Ontology design and entity identification, Neurocomputing 430 (2021) 104–111.

[4] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Transactions on Knowledge and Data Engineering 34 (1) (2020) 50–70.

[5] Z. Liu, F. Jiang, Y. Hu, C. Shi, P. Fung, Ner-bert: a pre-trained model for low-resource entity tagging, arXiv preprint arXiv:2112.00405 (2021).

[6] Z. Liu, Y. Xu, T. Yu, W. Dai, Z. Ji, S. Cahyawijaya, A. Madotto, P. Fung, Crossner: Evaluating cross-domain named entity recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 13452–13460.

[7] M. Asghari, D. Sierra-Sosa, A. S. Elmaghraby, Biner: A low-cost biomedical named entity recognition, Information Sciences 602 (2022) 184–200.

[8] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, D. Klakow, 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, 2021, pp. 2545–2568.

[9] R. Zevallos, J. Ortega, W. Chen, R. Castro, N. Bel, C. Toshio, R. Venturas, H. Aradiel, N. Melgarejo, 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, 2022, pp. 1–13.

[10] T. Ma, H. Jiang, Q. Wu, T. Zhao, C.-Y. Lin, Decomposed meta-learning for few-shot named entity recognition, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 1584–1596.

[11] J. Ma, M. Ballesteros, S. Doss, R. Anubhai, S. Mallya, Y. Al-Onaizan, D. Roth, Label semantics for few shot named entity recognition, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 1956–1971.

[12] L. Cui, Y. Wu, J. Liu, S. Yang, Y. Zhang, Template-based named entity recognition using bart, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 1835–1845.

[13] H. Yan, T. Gui, J. Dai, Q. Guo, Z. Zhang, X. Qiu, A unified generative framework for various ner subtasks, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 5808–5822.

[14] Z. Li, C. Hu, X. Guo, J. Chen, W. Qin, R. Zhang, 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, 2022, pp. 170–179.

[15] J. Chen, Q. Liu, H. Lin, X. Han, L. Sun, Few-shot named entity recognition with self-describing networks, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 5711–5722.

[16] S. Chen, G. Aguilar, L. Neves, T. Solorio, Data augmentation for cross-domain named entity recognition, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 5346–5356.

[17] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, Ontonotes release 5.0 ldc2013t19, in: Linguistic Data Consortium, Philadelphia, PA, 2013.

[18] E. T. K. Sang, F. De Meulder, 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, 2003, pp. 142–147.

[19] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, C. Zhang, 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, 2020, pp. 1054–1064.

[20] X. Zhang, B. Yu, T. Liu, Z. Zhang, J. Sheng, X. Mengge, H. Xu, Improving distantly-supervised named entity recognition with self-collaborative denoising learning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 10746–10757.

[21] Y. Meng, Y. Zhang, J. Huang, X. Wang, Y. Zhang, H. Ji, J. Han, 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, 2021, pp. 10367–10378.

[22] S. Rijhwani, S. Zhou, G. Neubig, J. G. Carbonell, Soft gazetteers for low-resource named entity recognition, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8118–8123.

[23] H. Lin, Y. Lu, X. Han, L. Sun, B. Dong, S. Jiang, 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), 2019, pp. 6232–6237.

[24] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[25] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).

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

[27] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: Proceedings of the 27th international conference on computational linguistics, 2018, pp. 1638–1649.

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

[29] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, K. Tu, 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, 2021, pp. 2643–2660.

[30] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 8003–8016.

[31] A. Fritzler, V. Logacheva, M. Kretov, Few-shot classification in named entity recognition task, in: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, 2019, pp. 993–1000.

[32] J. Huang, C. Li, K. Subudhi, D. Jose, S. Balakrishnan, W. Chen, B. Peng, J. Gao, J. Han, Few-shot named entity recognition: A comprehensive study, arXiv preprint arXiv:2012.14978 (2020).

[33] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, Q. J. Wu, A review of generalized zero-shot learning methods, IEEE transactions on pattern analysis and machine intelligence (2022).

[34] D.-H. Lee, A. Kadakia, K. Tan, M. Agarwal, X. Feng, T. Shibuya, R. Mitani, T. Sekiya, J. Pujara, X. Ren, 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, 2022, pp. 2687–2700.

[35] H. Jiang, D. Zhang, T. Cao, B. Yin, T. Zhao, 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, 2021, pp. 1775–1789.

[36] J. Chen, Z. Wang, R. Tian, Z. Yang, D. Yang, Local additivity based data augmentation for semi-supervised ner, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1241–1251.

[37] Q. Li, H. Li, H. Ji, W. Wang, J. Zheng, F. Huang, Joint bilingual name tagging for parallel corpora, in: Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 1727–1731.

[38] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, W. Zhang, 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, 2014, pp. 601–610.

[39] R. Speer, C. Havasi, et al., Representing general relational knowledge in conceptnet 5., in: LREC, Vol. 2012, 2012, pp. 3679–86.

[40] J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum, Yago2: A spatially and temporally enhanced knowledge base from wikipedia, Artificial intelligence 194 (2013) 28–61.

[41] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, 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, 2020, pp. 7871–7880.

[42] B. T. Polyak, A. B. Juditsky, Acceleration of stochastic approximation by averaging, SIAM journal on control and optimization 30 (4) (1992) 838–855.

[43] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, arXiv preprint:1603.01354 (2016).

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

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

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

[47] R. I. Doğan, R. Leaman, Z. Lu, Ncbi disease corpus: a resource for disease name recognition and concept normalization, Journal of biomedical informatics 47 (2014) 1–10.

[48] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

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