Federated Incremental Named Entity Recognition

Zesheng Liu[1,2], **Qiannan Zhu**[4,5*], **Cuiping Li**[1,2*], **Hong Chen**[1,3]
[1]School of Information, Renmin University of China, Beijing, China
[2]Key Laboratory of Data Engineering and Knowledge Engineering, MOE, China
[3]Engineering Research Center of Database and Business Intelligence, MOE, China
[4]School of Artificial Intelligence, Beijing Normal University, Beijing, China
[5]Engineering Research Center of Intelligent Technology and Educational Application, MOE, China
{lzs2022,licuiping,chong}@ruc.edu.cn, zhuqiannan@bnu.edu.cn

Abstract

Federated learning-based Named Entity Recognition (FNER) has attracted widespread attention through decentralized training on local clients. However, most FNER models assume that entity types are pre-fixed, so in practical applications, local clients constantly receive new entity types without enough storage to access old entity types, resulting in severe forgetting on previously learned knowledge. In addition, new clients collecting only new entity types may join the global training of FNER irregularly, further exacerbating catastrophic forgetting. To overcome the above challenges, we propose a Forgetting-Subdued Learning (FSL) model which solves the forgetting problem on old entity types from both intra-client and inter-client two aspects. Specifically, for intra-client aspect, we propose a prototype-guided adaptive pseudo labeling and a prototypical relation distillation loss to surmount catastrophic forgetting of old entity types with semantic shift. Furthermore, for inter-client aspect, we propose a task transfer detector. It can identify the arrival of new entity types that are protected by privacy and store the latest old global model for relation distillation. Qualitative experiments have shown that our model has made significant improvements compared to several baseline methods.

1 Introduction

Federated learning (FL) (Fallah et al., 2020; Wang et al., 2020; De Lange et al., 2020; Wen et al., 2023; Liu et al., 2024) is a decentralized training mode that can learn global models across distributed local clients without accessing their private data. Under privacy protection, it alleviates the limitations of data islands by training on multiple dispersed local clients and achieves rapid development in named entity recognition (NER) (Ma and Hovy, 2016; Lample et al., 2016; Li et al., 2020, 2022; Shen et al., 2023). Meanwhile, federated learning-based named entity recognition (FNER) (Ge et al., 2020) is also a popular research direction, which can significantly save annotation costs in data scarce scenarios by training global NER models on private data from different clients.

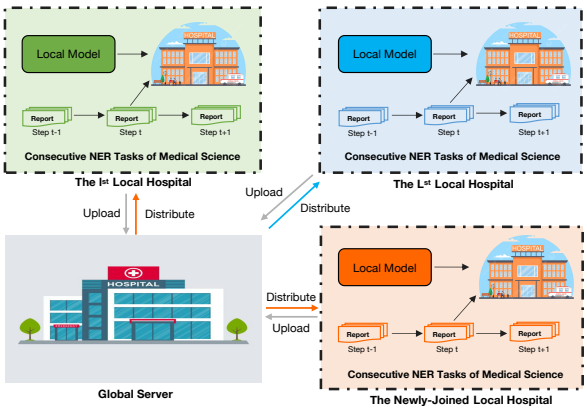

Figure 1: Typical FINER setting for medical science. Many hospitals including newly-joined ones receive new entity types incrementally according to their own needs. FINER aims to consecutively recognize new medical entities such as diseases and drugs in clinical reports via collaboratively learning a global NER model on private medical data of different hospitals.

Existing FNER methods (Ge et al., 2020; Zhao et al., 2021; Wang et al., 2023) unrealistically assume that entity types are static and fixed over time, because in real-world applications, local clients may continuously receive stream data of new entity types. A direct solution is to force local clients to store all samples of old entity types, and then learn a global model to continuously recognize new entity types through FL. But with the continuous arrival of new entity types, this requires a significant amount of computation and memory overhead, which limits the application capability of FNER. Even worse, if local clients have no memory to store old data of old entity types, existing FNER methods will significantly reduce their recognition

*Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 188–198
January 19–24, 2025.

ability on old entity types (i.e., catastrophic forgetting (Goodfellow et al., 2013; Kirkpatrick et al., 2017; Rebuffi et al., 2017; Wang et al., 2024)). Moreover, non-entity type in current task has the semantic shift (Douillard et al., 2021; Zhang et al., 2023a), which may belong to the entity types in old tasks or future tasks. This phenomenon seriously exacerbates the forgetting speed. More importantly, in practical scenarios, new local clients that incrementally receive new entity types may irregularly join global training, further intensifying catastrophic forgetting.

To overcome the problems in realistic scenarios mentioned above, we focus a novel practical problem called Federated Incremental Named Entity Recognition (FINER), where local clients continuously collect new entity types based on their preferences, and new local clients that collect unseen entity types may participate in global training irregularly. In the FINER setting, entity types are non-independent and identically distributed (Non-IID) across different clients, and training data of old entity types is not available for all local clients. FINER aims to train a global incremental NER model through collaborative FL training on local clients to address catastrophic forgetting. In this work, we use medical named entity recognition as an example to better illustrate FINER, as shown in Figure 1. A lot of hospitals collect unseen/new medical entity types continuously in clinical reports. Considering privacy protection, these hospitals hope to learn global entity recognition patterns through FL without accessing the data of each other (Zhang et al., 2022).

A simple solution for FINER is to directly integrate incremental named entity recognition (INER) (Monaikul et al., 2021; Zheng et al., 2022; Zhang et al., 2023b; Qiu et al., 2024) with FL. However, such a trivial solution requires that the global server needs to have strong manual prior about which and when local clients can collect new entity types, so that local clients can solve the forgetting issue on old entity types through knowledge distillation (Hinton et al., 2015; Wang et al., 2022; Asadi et al., 2023). Considering privacy protection in FINER, this sensitive information cannot be shared between local clients and global server. Consequently, from the intra-client perspective, due to lack of the signals for knowledge distillation, this simple solution suffers from serious forgetting issue caused by catastrophic forgetting with semantic shift. Also, from the inter-client perspective, it suffers from forgetting issue across different clients caused by Non-IID distributions, because the global server is unable to provide the above signals to local clients.

To surmount the aforementioned challenges, we develop a novel Forgetting-Subdued Learning (FSL) model that alleviates the forgetting problem on old entity types from both intra-client and inter-client two aspects. Specifically, to address the intra-client forgetting issue caused by semantic shift and catastrophic forgetting, we first propose a prototype-guided adaptive pseudo labeling to adaptively generate confident pseudo labels for old entity types with semantic shift. We then design a prototypical relation distillation loss to maintain semantic consistency between old model and current local model, thereby overcoming catastrophic forgetting within the local client under the guidance of confident pseudo labels. Furthermore, considering solving the inter-client forgetting problem, we develop a task transfer detector that automatically recognizes new entity types without any human prior and generate signals to store the latest old model from a global perspective for relation distillation. Experiments on two NER datasets (i.e., I2B2 (Murphy et al., 2010) and OntoNotes5 (Hovy et al., 2006)) show that our model has significant improvements compared to baseline methods. We summarize the main contributions of this work as follows:

- We focus a novel practical problem called Federated Incremental Named Entity Recognition (FINER), where the major challenges are intra-client and inter-client forgetting problems on old entity types caused by intra-client catastrophic forgetting with semantic shift and inter-client Non-IID distributions.
- We propose a Forgetting-Subdued Learning (FSL) model to address the FINER problem via overcoming forgetting from intra-client and inter-client two aspects. As far as we know, this is the first work to explore a global continual NER model in the FL field.
- We develop a prototypical relation distillation loss to solve intra-client forgetting problem, under the guidance of confident pseudo labels generated via prototype-guided adaptive pseudo labeling.
- We design a task transfer detector to surmount inter-client forgetting by accurately recogniz-

ing new entity types under privacy protection and storing the latest old model from global aspect for relation distillation.

2 Related Work

Federated Learning-based Named Entity Recognition (FNER) is a secure distributed machine learning paradigm that aggregates model parameters of local-client to build a global NER model under the privacy protection. FedNER (Ge et al., 2020) proposes to decompose medical NER model on each client into shared and private modules to sufficiently utilize the knowledge from other clients and learn the features of local data in unison. FAL (Zhao et al., 2021) introduces the adversarial training technology to effectively improve the model robustness and generalization for FNER. (Wang et al., 2023) employs distillation with pseudo-complete annotation and an instance weighting mechanism to cope with the heterogeneous tag sets and facilitate knowledge transfer across different clients. However, the above-mentioned FNER methods cannot recognize new entity types continuously under the FINER settings.

Incremental Named Entity Recognition (INER) considers class-incremental learning in named entity recognition. ExtendNER (Monaikul et al., 2021) is the pioneer in applying knowledge distillation to INER task. CFNER (Zheng et al., 2022) introduces a causal framework for extracting new causal effects in entities and non-entities. L&R (Xia et al., 2022) proposes a learn and review framework by simultaneously training a backbone model and a generative model to generate samples of old entity types to be trained with new samples. DLD (Zhang et al., 2023b) improves the knowledge distillation method in ExtendNER via dividing it into negative terms and positive terms for a fine-grained knowledge distillation. CPFD (Zhang et al., 2023a) proposes a pooled features distillation loss and designs a confidence-based pseudo-labeling strategy for classification. Nevertheless, these INER methods cannot be effectively applied to address the FINER problem, due to their strong prior knowledge to access privately-sensitive information (i.e., when and which local clients receive new entity types).

3 Task Definition

As claimed in INER, some continual NER tasks are defined as $\mathcal{T} = \{\mathcal{T}^t\}_{t=1}^{T}$, where the t-th task $\mathcal{T}^t = \{\mathbf{X}_i^t, \mathbf{Y}_i^t\}_{i=1}^{N^t}$ is composed of N^t pairs of token sequences and labels. The label space $\mathcal{Y}^t$ of t-th task consists of $\mathcal{E}^t$ new entity types. Besides, $\mathcal{E}^t$ new entity types have no overlap with $\mathcal{E}^o = \sum_{i=1}^{t-1} \mathcal{E}^i$ old entity types ($\cup_{j=1}^{t-1}\mathcal{Y}^j$) learned from the $t-1$ old tasks. In the t-th task, we follow INER methods to annotate $\mathcal{E}^o$ old entity types as non-entity type e_o (*i.e.*, semantic shift), due to unavailable training data of $\mathcal{E}^o$ old entity types.

Then, we extend the settings from INER to FINER. Denote global server as $\mathcal{S}_g$ and L local clients as $\{\mathcal{S}_l\}_{l=1}^{L}$. In the FINER, at the r-th ($r = 1, \cdots, R$) global round, we randomly select some local clients to aggregate gradients. When we choose the l-th local client to learn the t-th NER task, the latest global model $\Theta^{r,t}$ is distributed to $\mathcal{S}_l$, and trained on private training data $\mathcal{T}_l^t = \{\mathbf{X}_{li}^t, \mathbf{Y}_{li}^t\}_{i=1}^{N_l^t} \sim \mathcal{P}_l$ of $\mathcal{S}_l$. $\mathbf{X}_{li}^t$ and $\mathbf{Y}_{li}^t \in \mathcal{Y}_l^t$ denote token sequences and labels of the l-th client. $\{\mathcal{P}_l\}_{l=1}^{L}$ are non-independent and identically distributed (*i.e.*, Non-IID) across local clients. The label space $\mathcal{Y}_l^t \subset \mathcal{Y}^t$ of $\mathcal{S}_l$ in the t-th task is composed of $\mathcal{E}_l^t$ new entity types ($\mathcal{E}_l^t \leq \mathcal{E}^t$) that belongs to a subset of $\mathcal{Y}^t = \cup_{l=1}^{L}\mathcal{Y}_l^t$. Following INER methods, we consider semantic shift in the FINER and also annotate $\mathcal{E}_l^o = \sum_{i=1}^{t-1} \mathcal{E}_l^i \subset \cup_{j=1}^{t-1}\mathcal{Y}_l^j$ old entity types from $t-1$ old tasks as non-entity type. After getting global model $\Theta^{r,t}$ and performing local training on $\mathcal{T}_l^t$, $\mathcal{S}_l$ obtains a updated local model $\Theta_l^{r,t}$. And global server $\mathcal{S}_g$ aggregates local models of selected clients as the global model $\Theta^{r+1,t}$ for training the next global round.

In the t-th task, following (Dong et al., 2022, 2023), all local clients $\{\mathcal{S}_l\}_{l=1}^{L}$ are divided into three categories: $\{\mathcal{S}_l\}_{l=1}^{L} = \mathbf{S}_o \cup \mathbf{S}_c \cup \mathbf{S}_n$. Specifically, $\mathbf{S}_o$ is composed of L_o local clients that have accumulated experience of previous tasks but cannot collect new data of the t-th task; $\mathbf{S}_c$ consisting of L_c local clients can receive new data of current task and has experience of old tasks; $\mathbf{S}_n$ includes L_n new local clients with unseen new entity types but without experience of old entity types. These local clients are randomly determined in each incremental task. New clients $\mathbf{S}_n$ are added randomly at any global round in FINER, increasing $L = L_o + L_c + L_n$ gradually as continuous tasks. More importantly, we don't have any prior knowledge about the distributions $\{\mathcal{P}_l\}_{l=1}^{L}$, quantity and order of NER tasks, when and which local clients receive new entity types. In this paper, FINER aims to learn a global model $\Theta^{R,T}$ to recognize

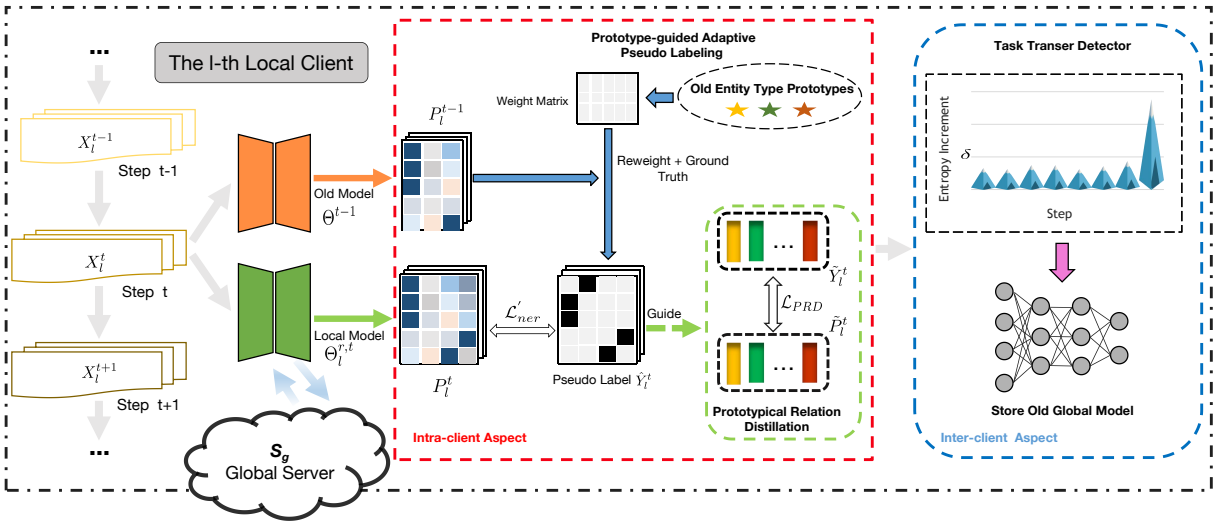

Figure 2: Overview of the proposed FSL model. It includes a *prototypical relation distillation loss* $\mathcal{L}_{\text{PRD}}$ to overcome intra-client catastrophic forgetting with semantic shift, under the guidance of *prototype-guided adaptive pseudo labeling*. At the same time, it makes use of a *task transfer detector* to tackle inter-client forgetting brought by Non-IID distributions.

new entity types continuously while surmounting forgetting on old entity types, under privacy preservation of local clients.

4 Methodology

Figure 2 presents the overview of our model to address the FINER problem. Our FSL model overcomes intra-client forgetting problem via a prototype-guided adaptive pseudo labeling (PAP, Section 4.1) to mine pseudo labels for old entity types with semantic shift, and a prototypical relation distillation loss (PRD, Section 4.2), collaborating with generated pseudo labels. Meanwhile, it addresses inter-client forgetting problem via a task transfer detector (TTD, Section 4.3) to recognize new entity types and store old model for relation distillation.

4.1 Prototype-guided Adaptive Pseudo Labeling

For the l-th local client $\mathcal{S}_l \in \mathbf{S}_c \cup \mathbf{S}_n$, the named entity recognition loss $\mathcal{L}_{\text{ner}}$ for a mini-batch $\{\mathbf{X}_{li}^t, \mathbf{Y}_{li}^t\}_{i=1}^{Bs} \subset \mathcal{T}_l^t$ sampled from the t-th incremental task is formulated as:

$$\mathcal{L}_{\text{ner}} = \frac{1}{Bs} \sum_{i=1}^{Bs} \sum_{j=1}^{|\mathbf{X}_{li}|} \mathcal{D}_{\text{CE}}\big(\mathbf{P}_l^t(\mathbf{X}_{li}^t, \Theta^{r,t})_j, (\mathbf{Y}_{li}^t)_j\big) \quad (1)$$

Algorithm 1 Determination of $\{(w_{lij}^t)_e\}_{e=0}^{\mathcal{E}^o}$ in Eq. (2).

Input: $\mathcal{T}_l^t = \{\mathbf{X}_{li}^t, \mathbf{Y}_{li}^t\}_{i=1}^{N_l^t}$ and number K;
for $i = 1, \cdots, N_l^t$ **do**
  $\mathbf{F}_{li}^t = \mathbf{F}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1})$;
  $\mathrm{L}_{li}^t = \arg\max \mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1}) \in \mathbb{R}^{|\mathbf{x}_{li}^t|}$;
  $\mathbf{F}_l^t = [\mathbf{F}_l^t; flatten(\mathbf{F}_{li}^t)]$;
  $\mathbf{L}_l^t = [\mathbf{L}_l^t; flatten(\mathbf{L}_{li}^t)]$;
for $e = 0, \cdots, \mathcal{E}^o$ **do**
  $F_{le} = \mathbf{F}_l^t[\mathbf{L}_l^t == e]$;
  $\eta_{l,e}^t = \text{mean}\{F_{le}\}$;
  Select K feature vectors closest to $\eta_{l,e}^t$ from F_{le} and recalculate $\eta_{l,e}^t$;
  Calculate $(w_{lij}^t)_e$ using Eq. (4);

where $\mathcal{D}_{\text{CE}}(\cdot,\cdot)$ denotes the cross-entropy loss. At the r-th global round, global model $\Theta^{r,t}$ is transmitted from global server $\mathcal{S}_g$ to $\mathcal{S}_l$. $\mathbf{P}_l^t(\mathbf{X}_{li}^t, \Theta^{r,t})_j \in \mathbb{R}^{1+E^o+E^t}$ is the probability at the j-th ($j = 1, \cdots, |\mathbf{X}_{li}|$) token predicted by $\Theta^{r,t}$, and it can predict non-entity type, $\mathcal{E}^o$ old entity types, and $\mathcal{E}^t$ new entity types for the j-th token. $(\mathbf{Y}_{li}^t)_j \in \mathcal{Y}_l^t$ is corresponding label of the j-th token. Bs denotes the batch size. $E = card(\mathcal{E})$ represents the cardinality of entity types.

As aforementioned, in the FINER settings, local client $\mathcal{S}_l$ has no memory to store $\mathcal{E}^o$ old entity types, while non-entity tokens may belong to $\mathcal{E}^o$ old entity types, entity types from future tasks or real non-entity type (*i.e.*, semantic shift). As a result, it enforces the updating of local model $\Theta_l^{r,t}$ (*i.e.*, Eq. (1)) to suffer from intra-client forgetting prob-

lem among different old entity types brought by semantic shift, after $\mathcal{S}_l$ receives the global model $\Theta^{r,t}$ from $\mathcal{S}_g$ for local training. To this end, as shown in Figure 2, we develop a prototype-guided adaptive pseudo labeling to adaptively mine high-confidence pseudo labels for old entity types marked as non-entity tokens in t-th incremental task. These pseudo labels based on dynamic weights of $\mathcal{E}^o$ old entity types are essential to alleviate semantic shift within local clients.

At the t-th learning step, as shown in Figure 2, given a sample $\{\mathbf{X}_{li}^t, \mathbf{Y}_{li}^t\} \subset \mathcal{T}_l^t$, we feed it into old global model Θ^{t-1} of the last task and current local model $\Theta_l^{r,t}$ to obtain the probabilities $\mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1}) \in \mathbb{R}^{|\mathbf{x}_{li}^t| \times (1+E^o)}$ and $\mathbf{P}_l^t(\mathbf{X}_{li}^t, \Theta_l^{r,t}) \in \mathbb{R}^{|\mathbf{X}_{li}^t| \times (1+E^o+E^t)}$ respectively. Then pseudo label $\hat{\mathbf{Y}}_{li}^t \in \mathbb{R}^{|\mathbf{X}_{li}^t|}$ of given token sequence $\mathbf{X}_{li}^t$ is defined as:

$$(\hat{\mathbf{Y}}_{li}^t)_j = \begin{cases} e, \text{ if } (\mathbf{Y}_{li}^t)_j \neq 0 \ \ \& \ \ e = (\mathbf{Y}_{li}^t)_j; \\ e, \text{ if } (\mathbf{Y}_{li}^t)_j = 0 \ \ \& \ e = \arg\max \\ \quad (\mathcal{W}_{li}^t)_j \odot \mathbf{P}_l^{t-1}(\mathbf{X}_{li}^t, \Theta^{t-1})_j; \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

where $(\hat{\mathbf{Y}}_{li}^t)_j$ is pseudo label of the j-th token from $\hat{\mathbf{Y}}_{li}^t$ and $\odot$ denotes the element-wise multiplication. $\mathbf{P}_l^{t-1}(\mathbf{X}_{li}^t, \Theta^{t-1})_j$ is softmax probability of the j-th token from $\mathbf{P}_l^{t-1}(\mathbf{X}_{li}^t, \Theta^{t-1})$. Additionally, $(\mathcal{W}_{li}^t)_j = \{(w_{lij}^t)_e\}_{e=0}^{\mathcal{E}^o}$ denotes the dynamic weights used to adaptively select pseudo labels with high confidence. As shown in Eq. (2), in the t-th task $\mathcal{T}_l^t$, when the j-th token belongs to non-entity type, its pseudo label is determined by $(\hat{\mathbf{Y}}_{li}^t)_j = \arg\max(\mathcal{W}_{li}^t)_j \odot \mathbf{P}_l^{t-1}(\mathbf{X}_{li}^t, \Theta^{t-1})_j$. If the j-th token is not labeled as e_o, we consider its pseudo label as current entity types: $(\hat{\mathbf{Y}}_{li}^t)_j = (\mathbf{Y}_{li}^t)_j$. Otherwise, $(\hat{\mathbf{Y}}_{li}^t)_j = 0$ denotes real non-entity type for the j-th token of $\hat{\mathbf{Y}}_{li}^t$. And classification loss $\mathcal{L}_{\text{ner}}$ can be rewritten in the following form:

$$\mathcal{L}_{\text{ner}}^{'} = \frac{1}{Bs} \sum_{i=1}^{Bs} \sum_{j=1}^{|\mathbf{X}_{li}|} \mathcal{D}_{\text{CE}}\big(\mathbf{P}_l^t(\mathbf{X}_{li}^t, \Theta^{r,t})_j, (\hat{\mathbf{Y}}_{li}^t)_j\big) \tag{3}$$

The calculation flow of $\{(w_{lij}^t)_e\}_{e=0}^{\mathcal{E}^o}$ is summarized in **Algorithm** 1. Firstly, we obtain the feature representation $\mathbf{F}_l^t$ with its corresponding position $\mathbf{L}_l^t$ (*i.e.*, entity type) for all training samples in the t-th task $\mathcal{T}_l^t$ based on old global model Θ^{t-1}. Nextly, we get the prototype $\eta_{l,e}^t$ for each entity type $e \in e_o \cup \mathcal{Y}^o$ with $\mathbf{F}_l^t$ and $\mathbf{L}_l^t$. Considering noise in F_{le}, we reselect K feature vectors closest to $\eta_{l,e}^t$ from F_{le} to recalculate $\eta_{l,e}^t$. Finally, $(w_{lij}^t)_e$ is determined via the following process:

$$(w_{lij}^t)_e = \frac{exp(-||\mathbf{F}_l^{t-1}(\mathbf{X}_{li}^t, \Theta^{t-1})_j - \eta_{l,e}^t||)}{\sum_{e'} exp(-||\mathbf{F}_l^{t-1}(\mathbf{X}_{li}^t, \Theta^{t-1})_j - \eta_{l,e'}^t||)} \tag{4}$$

where $e^{'}$ represents any previously seen old entity types and we set $K = 100$ in this work.

Thus, given a mini-batch $\{\mathbf{X}_{li}^t, \mathbf{Y}_{li}^t\}_{i=1}^{Bs} \subset \mathcal{T}_l^t$, we can generate pseudo labels $\{\mathbf{X}_{li}^t, \hat{\mathbf{Y}}_{li}^t\}_{i=1}^{Bs} \subset \mathcal{T}_l^t$ adaptively via considering dynamic weights $\mathcal{W}_{li}^t$ in Eq. (2) for all old entity types. These high-confident pseudo labels can provide strong guidance for the local training to surmount intra-client forgetting problem.

4.2 Prototypical Relation Distillation

To address catastrophic forgetting within local client $\mathcal{S}_l \in \mathbf{S}_c \cup \mathbf{S}_n$, we propose a prototypical relation distillation loss $\mathcal{L}_{\text{PRD}}$, as shown in Figure 2. It considers that the relationships between different steps should remain constant. In conformity to this, distilling inter-task relations from old global model Θ^{t-1} to current local model $\Theta_l^{r,t}$ can address forgetting problem on old entity types. In the meantime, considering that relying solely on the prediction of a single sample to perform semantic consistency between Θ^{t-1} and $\Theta_l^{r,t}$ may introduce noisy relations, so we construct type-wise prototypes for task relation distillation, also known as prototypical relation distillation.

Specifically, for a given sample $\{\mathbf{X}_{li}^t, \mathbf{Y}_{li}^t\} \subset \mathcal{T}_l^t$ with generated pseudo label $\hat{\mathbf{Y}}_{li}^t$, we first obtain its current probability $\mathbf{P}_l^t(\mathbf{X}_{li}^t, \Theta_l^{r,t})$ predicted via local model $\Theta_l^{r,t}$ and old probability $\mathbf{P}_l^{t-1}(\mathbf{X}_{li}^t, \Theta^{t-1})$ predicted via old global model Θ^{t-1}. We then replace the first $1 + E^o$ dimensions of $\mathbf{Y}_{li}^t$ with $\mathbf{P}_l^{t-1}(\mathbf{X}_{li}^t, \Theta^{t-1})$ and get its new representation $\bar{\mathbf{Y}}_l^t(\mathbf{X}_{li}^t, \Theta_l^{r,t})$. Next, under the guidance of pseudo labels $\hat{\mathbf{Y}}_{li}^t$, we separately construct new type-wise relation prototype $\widetilde{\mathbf{P}}_{l,k}^t$ and its relation groundtruth $\widetilde{\mathbf{Y}}_{l,k}^t$ for the k-th entity type in $\mathcal{Y}^o \cup \mathcal{Y}^t$ as follows:

$$\widetilde{\mathbf{P}}_{l,k}^t = \frac{1}{\Delta_k} \sum_{i=1}^{Bs} \sum_{j=1}^{|\mathbf{X}_{li}|} \mathbf{P}_l^t(\mathbf{X}_{li}^t, \Theta_l^{r,t}) \cdot \mathbb{I}_{(\hat{\mathbf{Y}}_{li}^t)_j = k} \tag{5}$$

$$\widetilde{\mathbf{Y}}_{l,k}^t = \frac{1}{\Delta_k} \sum_{i=1}^{Bs} \sum_{j=1}^{|\mathbf{X}_{li}|} \bar{\mathbf{Y}}_l^t(\mathbf{X}_{li}^t, \Theta_l^{r,t}) \cdot \mathbb{I}_{(\hat{\mathbf{Y}}_{li}^t)_j = k} \tag{6}$$

where $\Delta_k = \sum_{i=1}^{Bs} \sum_{j=1}^{|\mathbf{x}_{li}|} \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j=k}$ and $\mathbb{I}$ is the indicator function. Finally, the proposed $\mathcal{L}_{\text{PRD}}$ is formulated as:

$$\mathcal{L}_{\text{PRD}} = \frac{1}{\mathcal{E}^o + \mathcal{E}^t} \sum_{k=1}^{\mathcal{E}^o+\mathcal{E}^t} \mathcal{D}_{\text{KL}}(\widetilde{\mathbf{P}}_{l,k}^t, \widetilde{\mathbf{Y}}_{l,k}^t) \quad (7)$$

where $\mathcal{D}_{\text{KL}}(\cdot||\cdot)$ indicates Kullback-Leibler divergence. Consequently, $\mathcal{L}_{\text{PRD}}$ can address intra-client catastrophic forgetting problem via maintaining consistent semantic relations between old model Θ^{t-1} and current local model $\Theta_l^{r,t}$.

Overall, the objective formulation of the l-th local client $\mathcal{S}_l$ to learn the t-th NER task $\mathcal{T}_l^t$ is expressed as follows:

$$\mathcal{L}_{\text{obj}} = \mathcal{L}_{\text{ner}}^{'} + \alpha\mathcal{L}_{\text{PRD}} + \beta\mathcal{L}_{\text{KD}} \quad (8)$$

where $\mathcal{L}_{\text{KD}}$ inherits from (Monaikul et al., 2021), α and β are trade-off parameters. When $t \geq 2$, we set $\alpha = 0.5$ and $\beta = 2$ in Eq. (8) to train local model $\Theta_l^{r,t}$; otherwise, we use $\mathcal{L}_{\text{ner}}$ in Eq. (1) to optimize $\Theta_l^{r,t}$.

4.3 Task Transfer Detector

When local clients recognize new entity types consecutively, global sever S_g requires to automatically identify when and which local clients collect new entity types, and then store the latest old global model Θ^{t-1} to perform $\mathcal{L}_{PRD}$. As a result, the accurate selection of the latest old model Θ^{t-1} is essential to address inter-client forgetting across different local clients brought by Non-IID distributions, when new entity types arrive. However, considering privacy preservation, we don't have human prior about when to obtain new entity types in local clients under the FINER settings. To address this issue, a naive way is to detect whether the labels of current training data have been observed before. Nevertheless, the Non-IID distributions across local clients make it impossible to identify whether the collected data belongs to old entity types seen by other clients or new entity types. Therefore, we design a task transfer detector to automatically discover when and which local clients collect new entity types. At the r-th round, when S_l receives global model $\Theta^{r,t}$, it evaluates the average entropy $\mathcal{Q}_l^{r,t}$ on $\mathcal{T}_l^t$:

$$\mathcal{Q}_l^{r,t} = \frac{1}{N_l^t} \sum_{i=1}^{N_l^t} \sum_{j=1}^{|\mathbf{X}_{li}^t|} \mathcal{Z}(P_l^t(\mathbf{X}_{li}^t, \Theta^{r,t})_j) \quad (9)$$

Datasets	#Entity Type	#Sample	Entity Type Sequence
I2B2	16	141k	AGE, CITY, COUNTRY, DATE, DOCTOR, HOSPITAL,IDNUM, MEDICALRECORD, ORGANIZATION, PATIENT, PHONE, PROFESSION, STATE, STREET, USERNAME, ZIP
OntoNotes5	18	77k	CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE,LAW, LOC, MONEY, NORP, ORDINAL, ORG,PERCENT, PERSON, PRODUCT, QUANTITY, TIME, WORK_OF_ART

Table 1: The statistical information for each NER dataset.

where $\mathcal{Z}(\cdot,\cdot) = \sum_i p_i \log p_i$ is entropy function. If there is a sudden rise for averaged entropy $\mathcal{Q}_l^{r,t}$: $\mathcal{Q}_l^{r,t} - \mathcal{Q}_l^{r-1,t} \geq \delta$, we believe this can serve as a signal that local clients are collecting new entity types. Then, we update t via $t \leftarrow t+1$, and automatically store the latest global model $\Theta^{r-1,t}$ as old model Θ^{t-1} to optimize local model $\Theta_l^{r,t}$ via Eq. (8). We set $\delta = 1.0$ empirically in this paper. This automatic selection of old model Θ^{t-1} from global aspect is essential to tackle inter-client forgetting problem via considering Non-IID distributions across local clients.

4.4 Optimization Procedure

At the beginning of each global round in each incremental task, all local clients employ Eq. (9) to calculate the average relative entropy of local data, and then some of local clients are randomly selected by global server $\mathcal{S}_g$ to conduct local training at each round. After these chosen clients utilize task transfer detector to accurately recognize new entity types, they automatically store the global model learned at the last global round as the old model Θ^{t-1} to generate confident pseudo labels for old entity types via Eq. (2), and optimize local model $\Theta_l^{r,t}$ via Eq. (8). Finally, the updated local models $\Theta_l^{r,t}$ of selected local clients are aggregated as $\Theta^{r+1,t}$ by $\mathcal{S}_g$ for the next round training.

5 Experiments

5.1 Implementation Details

We utilize two benchmark datasets: I2B2 (Murphy et al., 2010) and OntoNotes5 (Hovy et al., 2006) under various experimental settings to analyze effectiveness of our FSL model. We summarized the statistical data of them in Table 1. Meanwhile, we compare our FSL with recent INER methods under the FINER settings, namely ExtendNER (Mon-

Method	I2B2				OntoNotes5			
	FG-8-PG-1		FG-8-PG-2		FG-8-PG-1		FG-8-PG-2	
	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1
Finetuning + FL	15.69	12.95	27.03	22.06	12.58	8.59	18.51	12.02
PODNet (Douillard et al., 2020) + FL	20.80	12.35	47.21	25.21	11.12	7.64	13.10	10.33
LUCIR (Hou et al., 2019) + FL	25.38	17.37	51.40	27.51	19.23	12.46	21.04	14.92
ST (De Lange et al., 2019) + FL	41.82	23.93	55.25	32.29	45.14	22.61	48.66	26.65
ExtendNER (Monaikul et al., 2021) + FL	43.38	26.80	55.71	30.16	44.69	26.15	50.02	31.84
CFNER (Zheng et al., 2022) + FL	43.07	27.08	56.55	32.86	43.61	25.77	51.01	<u>32.82</u>
DLD (Zhang et al., 2023b) + FL	<u>44.12</u>	<u>27.19</u>	55.55	30.09	42.47	25.88	50.31	31.91
CPFD (Zhang et al., 2023a) + FL	43.52	26.60	<u>57.49</u>	<u>34.94</u>	<u>50.81</u>	<u>29.31</u>	<u>51.29</u>	31.28
FSL (Ours)	**47.07**	**29.88**	**58.11**	**35.50**	**52.13**	**34.71**	**53.12**	**33.35**
Imp.	**⇑2.95**	**⇑2.69**	**⇑0.62**	**⇑0.56**	**⇑1.32**	**⇑5.40**	**⇑1.83**	**⇑0.53**

Table 2: Comparisons with baselines on I2B2 and OntoNotes5 two datasets. The **bold** denotes the highest result, and the **<u>underline</u>** denotes the second highest result.

aikul et al., 2021), CFNER (Zheng et al., 2022) and CPFD (Zhang et al., 2023a). Furthermore, we also introduce incremental learning methods used in the field of computer vision as baseline methods, including Self-Training (ST) (De Lange et al., 2019; Rosenberg et al.), LUCIR (Hou et al., 2019), and PODNet (Douillard et al., 2020). Additionally, Finetuning method is directly employed as the lower bound.

For fair comparisons with these INER baseline methods, we follow them to set exactly the same incremental tasks and entity type order, and adopt BIO labeling scheme across all datasets. Besides, *a* entity types are used to train the base model, and we use *b* entity types for each incremental learning step, represented as FG-*a*-PG-*b*. And for the I2B2 and OntoNotes5 datasets, we both use two FINER settings: FG-8-PG-1 and FG-8-PG-2.

We employ SGD optimizer with initial learning rate as 2×10^{-3} to train the base task and 4×10^{-4} to learn incremental tasks. Our model utilizes a BERT-based encoder (Devlin et al., 2018) and employs a fully connected layer as the classifier. We use the PyTorch (Paszke et al., 2019) framework to implement the model, which is built on top of the Huggingface (Wolf et al., 2019) implementation. Considering the limitation of GPU overhead, we set initial local clients as 10, and add 4 new local clients for each task. We choose 4 local clients randomly to perform local training with 8 epochs if PG $= 2$ else 4 epochs. We randomly select 30% samples for each client in each task if PG $= 1$; otherwise, we randomly sample 50% entity types from current label set $\mathcal{Y}^t$, and assign 60% samples from these entity types to selected local clients.

Following baseline INER methods, we employ Micro-F1 (Mi-F1) and Macro-F1 (Ma-F1) as metric and serve the mean value across all steps as the final performance, including the base task. This two metrics evaluate the effectiveness to address forgetting problem and the ability to recognize new entity types continually.

5.2 Comparisons with Baselines

Experiments on I2b2 and OntoNotes5 two datasets are introduced to analyze superiority of our FSL under various settings of FINER, as shown in Table 2. Our FSL achieves a certain improvement over existing INER methods under various FINER settings. Specifically, as depicted in the left half of Table 2, our FSL achieves improvements over the best results of previous INER methods mean 1.79% in Mi-F1, and 1.63% in Ma-F1, under the two FINER settings of I2B2. In the right half of Table 2, our FSL achieves improvements over the best results of other INER methods mean 1.58% in Mi-F1, and 2.97% in Ma-F1, under the two FINER settings of OntoNotes5.

These results quantitatively illustrate the effectiveness of our model against other INER methods to learn a global continual NER model via collaboratively training local models under privacy preservation. Except for this, they also validate superiority of the proposed prototype-guided adaptive pseudo labeling and prototypical relation distillation loss to address intra-client and inter-client forgetting problem under the FINER settings.

Input Sentence	Record	date	:	2097	-	03	-	25	Patient	Name	:	Whitaker	,	Vincent
No PL	[O]	[O]	[O]	[O]	[O]	[O]	[O]	[O]	[O]	[O]	[O]	[B-PATIENT]	[I-PATIENT]	[I-PATIENT]
PL	[O]	[O]	[O]	[B-DATE]	[O]	[I-DATE]	[I-DATE]	[I-AGE]	[O]	[O]	[O]	[B-PATIENT]	[I-PATIENT]	[I-PATIENT]
PAP	[O]	[O]	[O]	[B-DATE]	[I-DATE]	[I-DATE]	[I-DATE]	[I-DATE]	[O]	[O]	[O]	[B-PATIENT]	[I-PATIENT]	[I-PATIENT]
Golden Labels	[O]	[O]	[O]	[B-DATE]	[I-DATE]	[I-DATE]	[I-DATE]	[I-DATE]	[O]	[O]	[O]	[B-PATIENT]	[I-PATIENT]	[I-PATIENT]

Figure 3: A real visualization example of some pseudo labels on I2B2 dataset under the FG-8-PG-2 setting.

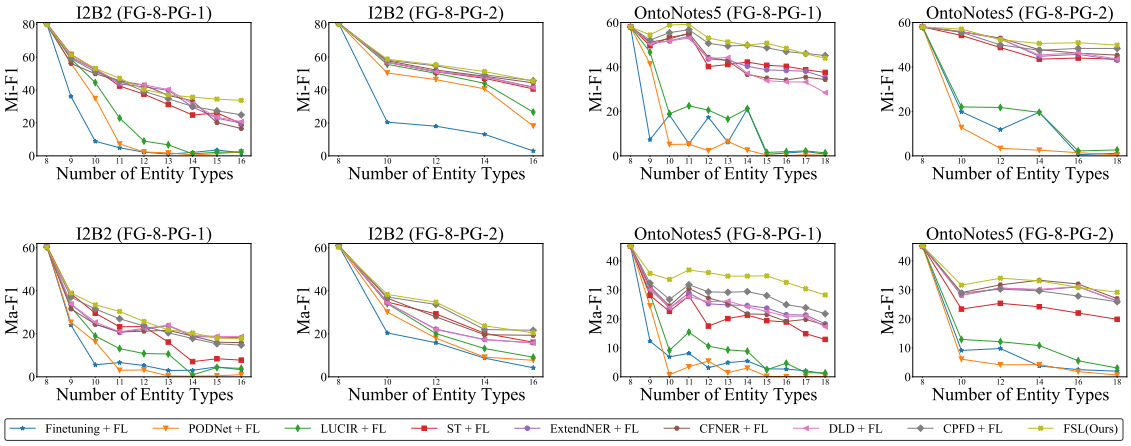

Figure 4: Comparisons of the step-wise Micro-F1 and Macro-F1 on I2B2 and OntoNotes5 two datasets.

Method	I2B2		OntoNotes5	
	Mi-F1	Ma-F1	Mi-F1	Ma-F1
Ours w/o PL	48.17	31.81	38.71	22.97
Ours w/o PAP	55.64	33.37	50.14	28.78
Ours w/o PRD	54.46	32.74	52.32	32.43
FSL (Ours)	**58.11**	**35.50**	**53.12**	**33.35**

Table 3: Ablation studies on I2B2 and OntoNotes5 under the FG-8-PG-2 setting of FINER.

5.3 Ablation Studies

To analyze effectiveness of each module in our model, Table 3 presents ablation experiments under various FINER settings. Ours w/o PL, Ours w/o PAP and Ours w/o PRD indicate the results of our model without pseudo labeling (denoted as PL), prototype-guided adaptive pseudo labeling (denoted as PAP) and prototypical relation distillation (denoted as PRD), where Ours w/o PAP directly use the prediction results of the old model as pseudo labels for the tokens marked as non-entity type. Compared to our FSL, the effectiveness of all ablation variants has significantly degraded.

More specifically, after removing PL, the results show 9.94% Mi-F1 and 3.69% Ma-F1 drop of I2B2, and 14.41% Mi-F1 and 10.38% Ma-F1 drop of OntoNotes5 compared to the full model. At the same time, after removing PAP from the full model, the results show 2.47% Mi-F1 and 2.13% Ma-F1 drop of I2B2, and 2.98% Mi-F1 and 4.57% Ma-F1 drop of OntoNotes5. Meanwhile, we can refer to an example in Figure 3. Without PL, the old entity type *DATE* is labeled as non-entity type, which can lead to semantic shift and exacerbate forgetting. Moreover, the error rate of conventional PL is relatively high compared to PAP (such as marking old entity type *DATE* as entity type *AGE* or non-entity type in Figure 3), so the effect will also be relatively poor, which is consistent with the previous experimental results. These results indicate that the proposed PAP module can effectively tackle semantic shift via confident pseudo labels.

And after removing PRD, the results show 3.65% Mi-F1 and 2.76% Ma-F1 drop of I2B2, and 0.80% Mi-F1 and 0.92% Ma-F1 drop of OntoNotes5 compared to the full model. This proves that the proposed PRD module can alleviate catastrophic forgetting of old entity types under the guidance of generated pseudo labels. As a consequence, the above results verify the importance of all modules to address the forgetting problem under the FINER settings.

5.4 Analysis of Step-Wise Comparisons

As shown in Figure 4, we introduce step-wise comparisons to analyze the validity of our model under FINER settings. Our model outperforms baseline INER methods (Douillard et al., 2020; Hou et al.,

2019; De Lange et al., 2019; Monaikul et al., 2021; Zheng et al., 2022; Zhang et al., 2023b,a) combined with FL for comparisons on I2B2 and OntoNotes5 under two FINER settings. Therefore, the proposed FSL model can encourage local clients to learn a global incremental NER model cooperatively under privacy preservation. Comparisons in Figure 4 show significant improvements of our model to address the FINER problem over other INER methods. When continuously recognizing new entity types, our FSL can effectively solve intra-client and inter-client forgetting problem.

6 Conclusion

In this paper, we propose a Federated Incremental Named Entity Recognition (FINER) problem, and develop a novel Forgetting–Subdued Learning (FBL) model to address intra-client and inter-client forgetting problem on old entity types. To tackle intra-client forgetting problem, we design a prototypical relation distillation loss, under the guidance of prototype-guided adaptive pseudo labeling. At the same time, we propose a task transfer detector to overcome inter-client forgetting problem. It can automatically recognize new entity types and store the latest old global model for relation distillation. Comparison results demonstrate the superiority of our FSL to tackle the FINER problem.

7 Limitations

Our PAP, which employs prototypes to calculate confidence, necessitates pre-calculation for each old entity type based on the current training data and the old model, thus extending training duration. Furthermore, it still have some mislabeled samples which will be introduced as noise into PRD. Furthermore, our PRD needs extra computational effort to align with the semantic relation between the new local model and the old model.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U23A20299, U24B20144, 62172424, 62276270, 62322214, 62472038, 62437001), National Key Research Develop Plan (2023YFB4503600), Fundamental Research Funds for the Central Universities (2233100004), and Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, China.

References

Nader Asadi, MohammadReza Davari, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. 2023. Prototype-sample relation distillation: towards replay-free continual learning. In *International Conference on Machine Learning*, pages 1093–1106. PMLR.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2019. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2(6):2.

Matthias De Lange, Xu Jia, Sarah Parisot, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2020. Unsupervised model personalization while preserving privacy and scalability: An open problem. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14451–14460. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. 2022. Federated class-incremental learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10154–10163. IEEE.

Jiahua Dong, Duzhen Zhang, Yang Cong, Wei Cong, Henghui Ding, and Dengxin Dai. 2023. Federated incremental semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3934–3943.

Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. 2021. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4040–4050.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16*, pages 86–102. Springer.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 3557–3568.

Suyu Ge, Fangzhao Wu, Chuhan Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2020. Fedner: Privacy-preserving medical named entity recognition with federated learning. *arXiv e-prints*, pages arXiv–2003.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *stat*, 1050:9.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. 2024. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

Xuezhe Ma and Eduard Hovy. 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 (Volume 1: Long Papers)*, pages 1064–1074.

Natawut Monaikul, Giuseppe Castellucci, Simone Filice, and Oleg Rokhlenko. 2021. Continual learning for named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13570–13577.

Shawn N. Murphy, Griffin M. Weber, Michael Mendis, Vivian S. Gainer, Henry C. Chueh, Susanne E. Churchill, and Isaac S. Kohane. 2010. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association : JAMIA*, pages 124–130.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Shengjie Qiu, Junhao Zheng, Zhen Liu, Yicheng Luo, and Qianli Ma. 2024. Incremental sequence labeling: A tale of two shifts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 777–791. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Diffusionner: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3875–3890.

FY Wang, DW Zhou, HJ Ye, and DC Zhan Foster. 2022. Feature boosting and compression for class-incremental learning. *ECCV FOSTER*.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. 2020. Federated learning with matched averaging. In *International Conference on Learning Representations*.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rui Wang, Tong Yu, Handong Wu, Junda andx Zhao, Sungchul Kim, Ruiyi Zhang, Subrata Mitra, and Ricardo Henao. 2023. Federated domain adaptation for named entity recognition via distilling with heterogeneous tag sets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7449–7463. Association for Computational Linguistics.

Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. 2023. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513–535.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yu Xia, Quan Wang, Yajuan Lyu, Yong Zhu, Wenhao Wu, Sujian Li, and Dai Dai. 2022. Learn and review: Enhancing continual named entity recognition via reviewing synthetic samples. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2291–2300.

Duzhen Zhang, Wei Cong, Jiahua Dong, Yahan Yu, Xiuyi Chen, Yonggang Zhang, and Zhen Fang. 2023a. Continual named entity recognition without catastrophic forgetting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8186–8197.

Duzhen Zhang, Yahan Yu, Feilong Chen, and Xiuyi Chen. 2023b. Decomposing logits distillation for incremental named entity recognition. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1919–1923.

Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. 2022. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125.

Hanyu Zhao, Sha Yuan, Niantao Xie, Jiahong Leng, and Guoqiang Wang. 2021. A federated adversarial learning method for biomedical named entity recognition. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2962–2969.

Junhao Zheng, Zhanxian Liang, Haibin Chen, and Qianli Ma. 2022. Distilling causal effect from miscellaneous other-class for continual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3602–3615.