Federated Document-Level Biomedical Relation Extraction with Localized Context Contrast

Yan Xiao, Yaochu Jin*, Kuangrong Hao
Donghua University, Westlake University, Donghua University
Shanghai 201620, China, Hangzhou 310030, China, Shanghai 201620, China
xiaoyan@mail.dhu.edu.cn, jinyaochu@westlake.edu.cn, krhao@dhu.edu.cn

Abstract

Existing studies on relation extraction focus at the document level in a centralized training environment, requiring the collection of documents from various sources. However, this raises concerns about privacy protection, especially in sensitive domains such as finance and healthcare. For the first time, this work extends document-level relation extraction to a federated environment. The proposed federated framework, called FedLCC, is tailored for biomedical relation extraction that enables collaborative training without sharing raw medical texts. To fully exploit the models of all participating clients and improve the local training on individual clients, we propose a novel concept of localized context contrast on the basis of contrastive learning. By comparing and rectifying the similarity of localized context in documents between clients and the central server, the global model can better represent the documents on individual clients. Due to the lack of a widely accepted measure of non-IID text data, we introduce a novel non-IID scenario based on graph structural entropy. Experimental results on three document-level biomedical relation extraction datasets demonstrate the effectiveness of our method. Our code is available at https://github.com/xxxxyan/FedLCC.

Keywords: Federated learning, biomedical document-level relation extraction, contrastive learning

1. Introduction

Relation extraction (RE) aims to automatically determine the types of relationships that exist between pairs of entities. The task is crucial for understanding and extracting knowledge from large volumes of unstructured text, especially in domains like biomedicine where there is a vast amount of scientific text. In real-world applications, such as electronic health records and discharge summaries, data privacy is a significant concern and sharing or replicating medical texts is subject to stringent limitations and restrictions. However, conventional deep learning-based approaches to RE is data-hungry, demanding a significant amount of data to achieve high performance. These techniques typically rely on centralized data storage and data sharing across various organizations, giving rise to significant privacy concerns, particularly in sensitive domains like biomedicine or finance.

To solve the above problem, federated learning (FL) (Yang et al., 2019; Jin et al., 2022) has emerged as a promising approach, enabling collaborative training of models across multiple decentralized devices or institutions without sharing the raw data. In FL, as shown in Fig. 1, individual local models are situated on local platforms (also known as clients), and the training process is carried out locally, allowing each client to learn from its own data. After that, the model parameters or gradients on each client are aggregated by the central server to update the global model, which is then sent back to each client. Based on the updated global model, each client can further refine it using their local data. This iterative process involves local training and global model aggregation, which continues until the desired level of convergence or performance is achieved.

Further, real-world applications often involve information spanning beyond a single sentence, with numerous relations expressed across multiple sentences, necessitating the exploration of document-level RE. Consequently, federated document-level RE tasks are of great practical importance. In contrast to sentence-level RE, where only one sentence typically involves one entity pair for classification, document-level RE tasks are more challenging. Apart from the complex scenario where a document may involve multiple entity pairs, and an entity pair may have multiple relationships, there is also a common case where an entity appears in various forms throughout the document. These different forms include aliases, co-occurrence words, and coreferences, collectively referred to as "mentions" of the entity. An illustrative example is given in Fig. 1, where the target entities are the chemical "*Clotiazepam*" with five mentions, namely C_{11}, C_{12}, C_{13}, C_{14}, and C_{15}, and the disease "*hepatitis*" has four mentions, i.e., D_{11}, D_{12}, D_{13}, and D_{14}. Furthermore, in this example, it is clear that these entities are mentioned in almost every sentence, but only the fourth sentence explicitly states that the chemical "*Clotiazepam*" can interact with the disease "*hepatitis*". Such complex and diverse semantics emphasize the importance of capturing and understanding the context across sentences related to target entities in document-level RE tasks.

To address the above challenges, this work pro-

* Corresponding Author.

LREC-COLING 2024, pages 7163–7173

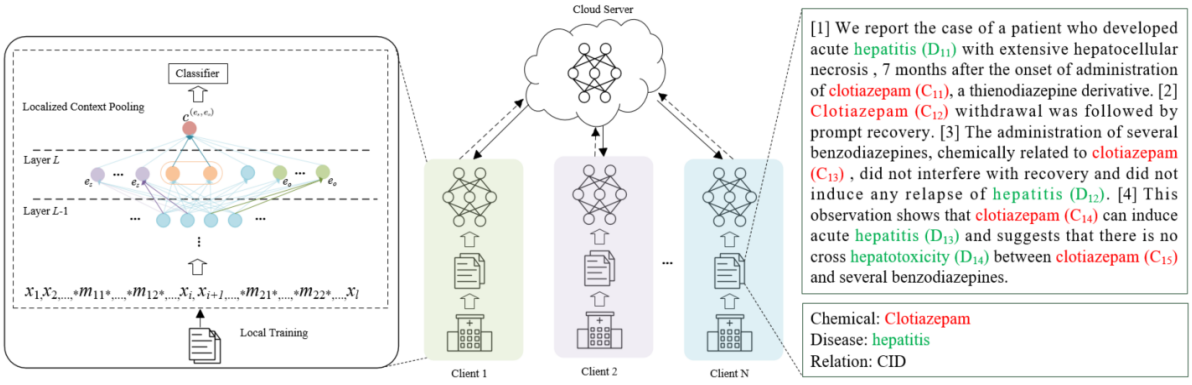

Figure 1: The overall framework of the proposed FedLCC for document-level biomedical relation extraction in a federated setting.

poses a new framework called FedLCC that aims at extracting biomedical triplets based on text data stored on multiple devices, without the leakage of private data. Inspired by model-contrastive federated learning (Li et al., 2021a), we propose the concept of localized context contrast (LCC) in the local training processes to leverage the collective knowledge of all participating clients while ensuring that local updates are appropriately adjusted based on the localized context contrast among clients. At the same time, considering that the data distribution on each client often does not satisfy the independently and identically distributed (IID) condition in actual scenarios, we propose a partition method for non-independently and identically distributed (non-IID) data. This method is based on in graph structural entropy and is designed specifically for document-level RE datasets. We conduct extensive experiments on three different document-level biomedical relation extraction datasets. The results show that FedLCC achieves significant better performance under federated settings on both IID data and non-IID data, indicating the effectiveness of our framework, including the proposed concept of localized context contrast.

The main contributions of this work are summarized as follows:

- To ensure the protection of sensitive information, we investigate document-level biomedical relation extraction under the federated learning paradigm, enabling collaborative and privacy-preserving model training that only exchanges model parameters rather than raw data.
- To achieve and ensure the performance of federated document-level biomedical relation extraction, we proposed the FedLCC framework based on localized context contrast that conducts contrastive learning in document localized context to rectify the local traning of each client.
- Since there does no exist a widely accepted method for defining non-IID data partitions for RE tasks in a federated learning environment, we propose a measure for non-IID data partition RE tasks based on graph structural entropy, which is the first of its kind for such text datasets.

2. Related work

RE (Xiao et al., 2022) is a task of uncovering hidden knowledge and extracting structured information from unstructured text. Valuable relational information between entities often extends beyond individual sentences in real-world scenarios. For example, in the biomedical domain, important findings and underlying rules are frequently expressed through multiple mentions spread across sentence boundaries, which requires advanced techniques to handle this complex linguistic structures, namely the document-level RE technologies (Yao et al., 2019). The methods employed to tackle document-level RE tasks can be divided into two main streams. The first is graph-based approaches (Christopoulou et al., 2019; Sahu et al., 2019) that represent the document as a graph, where entities are nodes and relations are edges connecting the nodes. This representation enables the model to capture long-range dependencies and complex relational patterns, facilitating a more comprehensive understanding of documents. The second is transformer-based approaches (Xu et al., 2021; Zhou et al., 2021; Xie et al., 2022).They take advantage of pre-trained language models (Devlin et al., 2018) to learn rich contextual representations, which is simple yet highly effective and can yield state-of-the-art performance. However, it is important to note that no research attempts have been reported on deal-

ing with document-level RE tasks in a federated environment.

In contrast to traditional centralized machine learning, federated learning (Yang et al., 2019; Jin et al., 2022) is achieved by performing model training collaboratively across decentralized data sources while keeping the data localized. FedAvg (McMahan et al., 2017) is a baseline federated learning algorithm, which aggregates the local model parameters by weighted averaging. However, the effectiveness of FedAvg will significantly degrade when it is directly applied to a scenario where data distributions on different clients do not observe the IID assumption (Zhu et al., 2021), which remains an open challenge in federated learning. Recent studies have proposed several variants of FedAvg to address the non-IID problem (Zhu and Jin, 2019; Ji et al., 2019) and they mainly focus on solving heterogeneous label distribution over multiple clients.

Most existing federated learning algorithms on non-IID data are developed for computer vision. With the increasing awareness of privacy protection, federated learning has received increased attention also in natural language processing (Lin et al., 2021), such as named entity recognition (Ge et al., 2020) and knowledge graphs (Chen et al., 2021, 2022). Little research on federated learning for RE tasks has also been reported with few exceptions. For instance, a federated denoising framework (Sui et al., 2020a) is proposed to suppress label noise in distantly supervised RE. FedED (Sui et al., 2020b) is based on knowledge distillation to overcome the communication bottleneck in supervised RE, while a distributed joint extraction framework (Wang et al., 2023) is proposed for sedimentological entities and relations. To the best of our knowledge, these are the only studies on RE tasks in the federated learning setting, and all of which are developed for sentence-level RE tasks. Besides, there is a lack of further exploration of non-IID problems relevant to RE tasks. This poses an urgent demand for solutions to document-level RE tasks in a federated learning setting, particularly for addressing non-IID problems.

3. Method

3.1. Task Definition

Given a document D containing a set of entities $\{e_i\}_{i=1}^{n}$, an RE task aims to extract the relations between each entity pairs $(e_s, e_o)_{s,o=1...n;s\neq o}$, where e_s and e_o are identified as subject and object entities, respectively. Generally, one entity e_i may occur multiple times in a document, so we define these mentioned entities as mentions $\{m_{ij}\}_{j=1}^{N_{e_i}}$. Furthermore, a federated document-level biomedical RE task can be defined as follows. Suppose there are K biomedical clients $\{C_1, ..., C_K\}$ with respective private document data $\{\mathcal{D}_1, ..., \mathcal{D}_K\}$, the goal is to implement collaborative training on $\mathcal{D} \triangleq \cup_{i\in[K]}\mathcal{D}_i$ via a central server, and guarantees that no data on any local devices is exposed to others.

3.2. Document-Level RE

As shown in Fig. 1, given a document $D = [x_1, x_2, ...x_l]$, we first mark the spans of the entity mentions by inserting a special entity markers "*" at the start and end of each mention. Then the document is fed into a pretrained language model (PLM) to obtain the contextual embedding of textual tokens:

$$H = [h_1, h_2, ..., h_l] = PLM\left([x_1, x_2, ..., x_l]\right) \quad (1)$$

where $H \in \mathbb{R}^{l\times d}$, l is the length of the input document, and d is the hidden dimension of the PLM. For each entity e_i with its mentions $\{m_{ij}\}_{j=1}^{N_{e_i}}$, N_{e_i} denotes the number of mentions for e_i, the embedding of the special token "*" at the start of one mention is taken as the mention embedding, and it is denoted as $h_{m_{ij}}$. Then the entity embedding is calculated by logsumexp pooling (Jia et al., 2019) the embedding of mentions corresponding to the same entity, which can be expressed as follows:

$$h_{e_i} = \log \sum_{j=1}^{N_{e_i}} \exp\left(h_{m_{ij}}\right) \quad (2)$$

To get a more condensed and precise representation that is useful to determine the relation for an entity pair, we adapt the localized context pooling (Zhou et al., 2021) to enhance the embedding of an entity pair with an additional local context embedding that is related to both entities. Specifically, since the pretrained transformer-based language models is used as our document encoder, the multi-head attention matrix in the last layer of PLM is denoted as $A \in \mathbb{R}^{h\times l\times l}$, where h is the number of attention heads, and A_{kij} represents the attention from token i to token j in the k-th attention head of the last layer. We first take the attention at the position of the "*" symbol as the mention-level attention, and then average the attention over mentions of the same entity to obtain the entity-level attention $A^{e_i} \in \mathbb{R}^{h\times l}$, which denotes attention from the i-th entity to all tokens. For each entity pair (e_s, e_o), we multiply the entity-level attention of e_s and e_o to obtain the local context that is important to both entities. Then the localized context embedding $c^{(e_s,e_o)} \in \mathbb{R}^d$ is calculated as follows from the

original contextual embedding H:

$$q^{(e_s,e_o)} = \sum_{j=1}^{h} \left(A_j^{e_s} \cdot A_j^{e_o}\right) \tag{3}$$

$$a^{(e_s,e_o)} = q^{(s,o)}/1^\top q^{(e_s,e_o)} \tag{4}$$

$$c^{(e_s,e_o)} = H^\top a^{(e_s,e_o)} \tag{5}$$

The above process is illustrated on the left panel in Fig. 1.

To get different representations for different entities, localized context embedding is then fused into the globally pooled entity embedding as follows. $z_s^{(e_s,e_o)}$ and $z_o^{(e_s,e_o)}$ represent the context-enhanced representations of subject e_s and object e_o, respectively.

$$z_s^{(e_s,e_o)} = \tan \mathrm{h}\left(W_s h_{e_s} + W_{c_1} c^{(e_s,e_o)}\right) \tag{6}$$

$$z_o^{(e_s,e_o)} = \tan \mathrm{h}\left(W_o h_{e_o} + W_{c_2} c^{(e_s,e_o)}\right) \tag{7}$$

To reduce the amount of parameter calculations, we also use a grouped bilinear function for feature combination. The entity embedding $z_s^{(e_s,e_o)}$ and $z_o^{(e_s,e_o)}$ are both split into k equal-sized groups, then the probability at which relation r is associated with the entity pair (e_s, e_o) is calculated as follows:

$$z_s^{(e_s,e_o)} = \left[z_s^1, z_s^2, ..., z_s^k\right] \tag{8}$$

$$z_o^{(e_s,e_o)} = \left[z_o^1, z_o^2, ..., z_o^k\right] \tag{9}$$

$$P\left(r \mid e_s, e_o\right) = \text{sigmoid}\left(\sum_{i=1}^{k} z_s^{i\intercal} W_r^i z_o^i + b_r\right) \tag{10}$$

where $W_r^i \in \mathbb{R}^{d/k \times d/k}$ for $i = 1, ..., k$ are the model parameters.

3.3. Local Training

During the local training, each client independently trains the model using its local data. Here we take the k-th client as an example to introduce this procedure, which is the same for the rest clients. In round t, client k first receives the global model parameters Θ_k from the master server and then uses the local document data to train the RE model introduced in Section 3.2, and its classification loss function $\mathcal{L}_{BCE}$ is calculated as follows:

$$\begin{aligned}\mathcal{L}_{BCE} = -\sum_{r} &\left(y_r \cdot \log\left(P\left(r \mid e_s, e_o\right)\right) + \right. \\ &\left.\left(1 - y_r\right) \cdot \log\left(1 - P\left(r \mid e_s, e_o\right)\right)\right)\end{aligned} \tag{11}$$

Since there is always feature drift among these local clients in federated training, we propose an additional loss called the localized context contrast loss $\mathcal{L}_{LCC}$ to update the parameters of local models. It enables the relevant context, crucial for determining the relations in the document, to be better located during each round of model update. Specifically, a loss function similar to the NT-Xent loss (Chen et al., 2020) is modified to rectify the local updates by adjusting the agreements of localized context learned from the local and global models. It is mainly based on the idea of contrastive learning. For each sample on client k, suppose that L^t is the localized context learned by the global model on the server during the t-th communication round, L_k^t is the localized context learned by the local model on client k, and L_k^{t-1} is the localized context learned by the same client in the $t-1$-th round, then the goal is to simultaneously reduce the distance between L^t and L_k^t and increase the distance between L_k^t and L_k^{t-1}, where $L = c^{(e_s,e_o)}$ is calculated by Eq. (5). The localized context contrast loss function is calculated as follows, where τ is the temperature parameter and sim$(\cdot, \cdot)$ is the cosine similarity function.

$$\mathcal{L}_{LCC} = -\log \frac{\exp\left(\text{sim}\left(L_k^t, L^t\right)/\tau\right)}{\exp\left(\text{sim}\left(L_k^t, L^t\right)/\tau\right) + \exp\left(\text{sim}\left(L_k^t, L_k^{t-1}\right)/\tau\right)} \tag{12}$$

Overall, the final local training loss is defined as the combination of the classification loss and the localized context contrast loss:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{BCE}\left(\Theta_k^t; (e_s, e_o; r)\right) + \\ &\mu \mathcal{L}_{LCC}\left(\Theta_k^t; \Theta_k^{t-1}; \Theta^t; (e_s, e_o)\right)\end{aligned} \tag{13}$$

where μ is a hyper-parameter to control the weight of localized context contrast loss.

3.4. Model Aggregation

Suppose $\mathcal{C}_{sel}$ is the set of participating clients in the t-th communication round. Once the clients have completed local training, the trained model parameters on each client will be uploaded to the server for global model aggregation. This aggregation process generates a new global model that captures the collective knowledge from the distributed clients by taking an average of all trained local model parameters. The model parameters are calculated as follows:

$$\Theta^{t+1} = \frac{1}{n} \sum\nolimits_{k \in \mathcal{C}_{sel}} \Theta_k^t \tag{14}$$

where n is the number of participating clients and Θ_k^t is the trained parameters by minimizing Eq. (13) on the local data of client k in round t. Subsequently, the global model finalizes the aggregation

process and send the updated model back to each client for the next round of model update. Algorithm 1 lists the main steps of the entire procedure.

Algorithm 1 The FedLCC Framework

Require: the number of local clients C;
the client selection fraction F;
the number of communication rounds T;
the local batch size B
the local epoch E;
the learning rate η;
the coefficient of comparison μ;

Initialize Θ_0 on the central server
for round $t = 0, 1, 2, ..., T-1$ **do**
  $n \leftarrow max(C \times F, 1)$
  $\mathcal{C}_{sel} \leftarrow$ (random set of n local clients)
  for each client $k \in \mathcal{C}_{sel}$ **in parallel do**
    $\Theta_k^t \leftarrow$ LocalTraning(k, Θ^t)
  end for
  $\Theta^{t+1} \leftarrow \frac{1}{n} \sum_{k \in \mathcal{C}_{sel}} \Theta_k^t$
end for

function LocalTraining(k, Θ^t)
  $\Theta_k^t \leftarrow \Theta^t$
  $\mathcal{B} \leftarrow$ (split $\mathcal{D}_k$ into batches of size B)
  for epoch $i = 0, 1, 2, ..., E$ **do**
    for batch $b \in \mathcal{B}$ **do**
      $\mathcal{L}_{BCE} \leftarrow Eq.(11)$
      $\mathcal{L}_{LCC} \leftarrow Eq.(12)$
      $\mathcal{L} \leftarrow \mathcal{L}_{BCE} + \mu \mathcal{L}_{LCC}$
      $\Theta_k^t \leftarrow \Theta_k^t - \eta \nabla \mathcal{L}\left(\Theta_k^t, b\right)$
    end for
  end for
  return Θ_k^t
end function

3.5. Non-IID Data Partition for Documents

Considering that real-world data distributions can be remarkably diverse due to variations in context, language, or user behavior, we design a new non-IID data partition strategy based on the structural entropy of graphs (Solé and Valverde, 2004) to verify the proposed algorithm. To the best of our knowledge, this is the first of its kind for partitioning document data in a federated learning environment. The partition method proposed in this work is partly inspired by the idea of building an edge-oriented graph neural model for biomedical document-level RE (Christopoulou et al., 2019). Through the calculation of the graph structural entropy for the established graph, which somewhat measures the complexity of the graph and the distribution of structure, we can use it to assess the complexity of RE in each document sample:

$$H = -\sum_i p_i \log\left(p_i\right) \tag{15}$$

where H represents the calculated graph structural entropy, p_i is the probability of each different degree value in the graph, which can be calculated by the adjacency matrix of the graph. In contrast to most existing work focusing on heterogeneous label distribution over the clients, this work uses the above complexity to achieve the division of a given dataset into different data distributions.

More details of the proposed non-IID data partition method can be found in Section 4.2.

4. Experimental settings

4.1. Dataset

We conduct experiments on three publicly available document-level biomedical RE datasets and their statistics are listed in Table 1.

CDR (Chemical-Disease Reactions) dataset (Li et al., 2016) is a document-level dataset constructed by utilizing PubMed abstracts. It primarily revolves around a binary classification task, which aims to identify induced relation from chemical entity to disease entity.

CHR (CHemical Reactions) dataset (Sahu et al., 2019) is created by distant supervision. If two chemical entities have a relation in *Biochem4j*, they will be regarded as positive instances in the dataset; otherwise as negative.

GDA (Gene-Disease Associations) (Piñero et al., 2016) aims to identify gene and disease concepts interactions at the document level, but with a much more massive scale. It is constructed through distant supervision and we further divide the training set into a training set and a development set in a ratio of 8 to 2.

Data	Count	Train	Dev	Test
CDR	#Document	500	500	500
	#Pos pairs	1038	1012	1066
	#Neg pairs	4202	4075	4138
CHR	#Document	7298	1182	3614
	#Pos pairs	19652	3188	9584
	#Neg pairs	34713	5666	16580
GDA	#Document	23353	5839	1000
	#Pos pairs	36079	8762	1502
	#Neg pairs	96399	24362	3720

Table 1: Statistics of the document-level biomedical RE datasets.

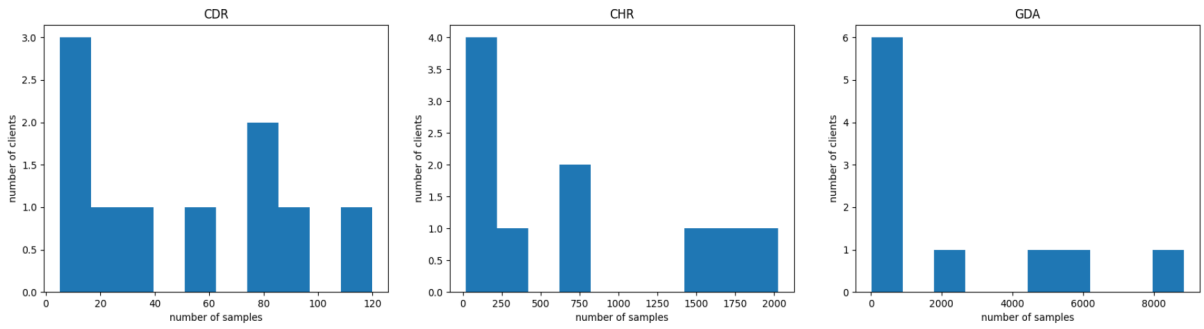

Figure 2: Statistical visualization of the partitioned data for each client under the non-IID scenario based on graph structural entropy.

Hyperparameter	Value
number of local clients C	10
client selection fraction F	1
number of communication rounds T	50
local batch size B	3
local epochs E	1
optimizer	Adam
learning rate η	0.000025
coefficient of comparison μ	0.1
temperate of comparison τ	0.5

Table 2: Hyperparameter configuration.

4.2. Implementation Details

In our experiments, we adopt the Huggingface's Transformer (Wolf et al., 2019), which is initialized with the pretrained cased SciBERT (Beltagy et al., 2019), a BERT model trained on multi-domain corpora of scientific publications. For a fair comparison, we implemented our method and all baselines in the same experimental settings and all hyperparameters are listed in Table 2. We set the number of local clients (C) to 10 for all datasets, other parameters are also set the same. Since the client selection fraction F, local batch size B and number of local epochs E directly influence the number of secure local updates per round, we tested different values for analysis in the experiments.

As for the data partitioning, in the IID scenario, the training data of each dataset is randomly shuffled and evenly divided into C portions. In the non-IID scenario, for each dataset, we first calculate the graph structural entropy of all samples for the training data to obtain the upper and lower bounds of its value, and the divide the value interval into C equal parts. The corresponding document samples are divided according to the division of the numerical intervals, and the number of samples contained on each client may be different. Figure 2 shows the client data distribution and statistical information of each dataset. These dataset partitioning operations simulate the scenario where each hospital or institution is treated as a local client and the central server is located in a trusted third party. The non-IID setting also simulates the imbalance in the number of training samples for each client in reality.

4.3. Compared Algorithms

In centralized training, we compare our model as depicted in Section 3.2 with the following state-of-the-art models:

- GCNN (Sahu et al., 2019): A labeled edge graph convolutional neural network model that leverages multi-instance learning with bi-affine pairwise scoring to predict the relations in a document.
- LSR (Nan et al., 2020): A latent structure refinement model empowers the relational reasoning across sentences by automatically inducing the latent document-level graph.
- DocRE-HGNN (Shi et al., 2021): A heterogeneous GNN-based framework that encodes the document with temporal convolutional networks and utilizes graph transformer networks to generate semantic paths.
- MGSN(Liu et al., 2021): A multi-granularity sequential network based on the accumulation of both document-level information and entity-level information.

In the federated training manner, we compare the proposed FedLCC with the following algorithms:

- FedAvg (McMahan et al., 2017): In FedAvg, the parameters of local models are combined by taking their weighted average, with each weight being determined by the size of the corresponding local dataset.
- FedAtt (Ji et al., 2019): A layer-wise attention mechanism is utilized during model aggregation, enabling it to automatically attend to the

weights of the relation between the central global model and the local models of different clients.

5. Results and analysis

5.1. Comparative Results

Method	P	R	F1
Centralized Training			
GCNN	52.8	66.0	58.6
LSR	-	-	64.8
DocRE-HGNN	-	-	64.4
MGSN	69.0	66.7	67.8
Our	63.00	74.11	**68.10**
Federated Training (IID)			
FedAvg	69.57	60.23	64.55
FedAtt	66.83	64.63	65.71
FedLCC	69.56	65.85	**67.66**
Federated Training (Non-IID)			
FedAvg	64.05	62.01	63.01
FedAtt	65.93	61.35	63.56
FedLCC	65.54	65.67	**65.60**

Table 3: Results on CDR.

Method	P	R	F1
Centralized Training			
MGSN	78.5	73.4	75.9
GCNN	84.7	90.5	87.5
Our	89.03	91.13	**90.07**
Federated Training (IID)			
FedAvg	90.78	90.82	90.80
FedAtt	90.88	90.47	90.68
FedLCC	91.84	92.14	**91.99**
Federated Training (Non-IID)			
FedAvg	88.62	90.91	90.23
FedAtt	89.44	90.93	90.18
FedLCC	89.18	93.22	**91.15**

Table 4: Results on CHR.

Table 3, Table 4 and Table 5 present the results of FedLCC against baselines on three real-world biomedical document-level datasets. By comparing the results between centralized training and federated training, we can see that FedLCC effectively guarantees the performance in federated scenarios and mostly achieves competitive results against other baselines. In Table 3, the results of centralized training are better than those of federated training on the CDR dataset. This can be attributed to the fact that there is a relatively limited amount of data in each client due to the small size of the CDR dataset, making it vulnerable to overfitting by the local models. Nonetheless, federated training offers

Method	P	R	F1
Centralized Training			
DocRE-HGNN	-	-	81.6
LSR	-	-	82.2
Our	82.26	87.35	**84.73**
Federated Training (IID)			
FedAvg	81.37	86.95	84.07
FedAtt	82.97	87.55	**85.20**
FedLCC	83.17	86.55	84.83
Federated Training (Non-IID)			
FedAvg	78.83	86.02	82.27
FedAtt	80.27	85.89	82.98
FedLCC	82.62	86.09	**84.32**

Table 5: Results on GDA.

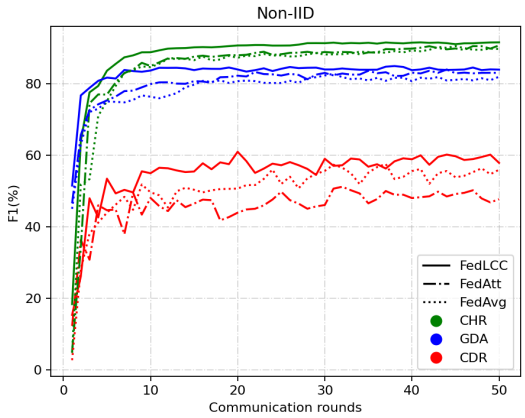

Figure 3: Convergence plots for FedLCC and other baselines for each dataset under non-IID scenario that use the structural entropy of the graph.

the distinctive benefit of safeguarding privacy, and our approach effectively reduces the performance gap compared to other methods. In Table 4 and Table 5, where the dataset size and the number of samples per client are relatively large, the results of federated training are unexpectedly better than those of centralized training. Actually, in some cases where each client has sufficient data, distributed training under the federated setting can be seen as an ensemble, which can potentially outperform centralized learning with a single model. Additionally, it is worth noting that in our experiments, the parameters for centralized training directly follow those used in the federated settings without specific fine-tuning. We surmise that this may also contribute to what we observed.

By comparing the performance across datasets and taking into account the sample size of each dataset, we can conclude that our algorithm is more advantageous on small datasets. Meanwhile, it demonstrates that the incorporation of localized context contrast is highly beneficial for federated

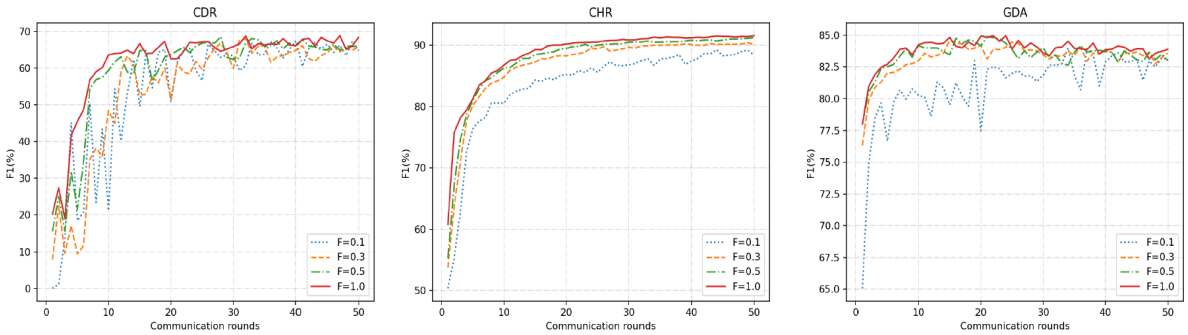

Figure 4: F1 score vs. communication rounds on the three datasets with various F (client selection fraction).

training, enabling efficient fusion of sample information from each client. From the convergence profiles of the algorithms under comparison on the three non-IID datasets shown in Fig. 3, we can see that our algorithm maintains high performance while having a relatively fast convergence speed.

5.2. Multi-client Parallelism

We further conduct experiments to analyze how varying the fraction values will influence the generalization performance of the proposed FedLCC algorithm. The fraction indicates the proportion of the local clients participating in federated training in each round of communication. In our experiments, we configure it to take on values of 0.1, 0.3, 0.5, and 1.0, respectively. With a fixed count of 10 local clients for all datasets, we calculate the specific number of selected clients by multiplying the total number of local clients by the given fraction.

In Fig. 4, we visualize the F1 score performance for each dataset in the IID scenario. Different curves represent different configurations of FedLCC, each employing various client selection fractions. We find that an increased number of clients participating in the training process leads to better F1 scores, despite that some datasets exhibit fluctuations in their curves. Moreover, increasing the degree of multi-client parallelism can help accelerate convergence, as it allows for the utilization of a greater amount of data in each training round. Nonetheless, it is important to highlight that as the proportion of clients increases, the consumption of time and computing resources also increase. Consequently, when determining the client fraction, it is crucial to consider these factors, including the expected performance.

5.3. Various Local Client Computation

Another critical factor that should be investigated in federated training is the local client computation, i.e., the number of epochs in local updating, since the actual computing power of each client may be limited. Actually, local client computation is calculated by $\frac{|\mathcal{D}_k|}{B}E$, which means it is controlled by the local epoch number E and the local batch size B. $\mathcal{D}_k$ is the size of private data in client k. That is to say, a large number of local epochs, a smaller batch size, or both will incur more client computation. We conduct experiments in the IID scenario on the CHR dataset, and fix the client selection fraction F to 1.0 while varying the local batch size B with {3, 9, 15} and local epochs E with {1, 5}. Table 6 lists the number of communication rounds required for each algorithm to achieve an F1 value of 80%. The results reveal that optimizing computation per local client by adjusting both B and E yields favorable results for all methods. Notably, our method exhibits faster convergence towards the targeted F1 value compared to other baseline methods.

B	E	FedAvg	FedAtt	FedLCC
3	1	8	9	5
9	1	10	10	7
15	1	13	12	9
3	5	7	5	2
9	5	9	8	3
15	5	11	10	4

Table 6: The number of communication rounds required to achieve an F1 value of 80% on the CHR dataset, fixing C to 1.0.

6. Conclusion and Future Work

In this paper, we propose a privacy-preserving biomedical document-level RE framework through federated learning, namely FedLCC, which is based on contrastive learning within the localized context of documents. In addition, we design a non-IID data partition strategy for document datasets based on graph structural entropy. We perform comprehensive experiments and analysis on three benchmark

datasets under IID and non-IID scenarios. The experimental results demonstrate the effectiveness of our proposed framework. We also analyze the impact of multi-client parallelism and client computation on the training of our federated framework. Increasing the number of participating clients in each training round can improve performance but also increase communication costs. Similarly, reducing the local batch size and increasing local epochs accelerates model convergence but also incurs additional communication costs. In this work, we only explore these influencing factors. In the future, we aim to explore federated RE in more realistic scenario with less communication cost.

7. Acknowledgements

This work was supported in part by the Fundamental Research Funds for the Central Universities (2232021A-10), Shanghai Pujiang Program (22PJ1423400), Shanghai Sailing Program (no. 22YF1401300), Natural Science Foundation of Shanghai (20ZR1400400), and the National Natural Science Foundation of China (62136003).

8. Bibliographical References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Mingyang Chen, Wen Zhang, Zonggang Yuan, Yantao Jia, and Huajun Chen. 2021. Fede: Embedding knowledge graphs in federated setting. In *Proceedings of the 10th International Joint Conference on Knowledge Graphs*, pages 80–88.

Mingyang Chen, Wen Zhang, Zonggang Yuan, Yantao Jia, and Huajun Chen. 2022. Federated knowledge graph completion via embedding-contrastive learning. *Knowledge-Based Systems*, 252:109459.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. *arXiv preprint arXiv:1909.00228*.

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*.

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 preprint arXiv:2003.09288*.

Shaoxiong Ji, Shirui Pan, Guodong Long, Xue Li, Jing Jiang, and Zi Huang. 2019. Learning private neural language modeling with attentive aggregation. In *2019 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. *arXiv preprint arXiv:1904.02347*.

Yaochu Jin, Hangyu Zhu, Jinjin Xu, and Yang Chen. 2022. *Federated Learning: Fundamentals and Advances*. Springer Nature.

Qinbin Li, Bingsheng He, and Dawn Song. 2021a. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. 2021b. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*.

Bill Yuchen Lin, Chaoyang He, Zihang Zeng, Hulin Wang, Yufen Huang, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. 2021. Fednlp: Benchmarking federated learning methods for natural language processing tasks. *arXiv preprint arXiv:2104.08815*.

Xiaofeng Liu, Kaiwen Tan, and Shoubin Dong. 2021. Multi-granularity sequential neural network for document-level biomedical relation extraction. *Information Processing & Management*, 58(6):102718.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR.

Guoshun Nan, Zhijiang Guo, Ivan Sekulić, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. *arXiv preprint arXiv:2005.06312*.

Yong Shi, Yang Xiao, Pei Quan, MingLong Lei, and Lingfeng Niu. 2021. Document-level relation extraction via graph transformer networks

and temporal convolutional networks. *Pattern Recognition Letters*, 149:150–156.

Ricard V Solé and Sergi Valverde. 2004. Information theory of complex networks: on evolution and architectural constraints. In *Complex Networks*, pages 189–207. Springer.

Dianbo Sui, Yubo Chen, Kang Liu, and Jun Zhao. 2020a. Distantly supervised relation extraction in federated settings. *arXiv preprint arXiv:2008.05049*.

Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, and Weijian Sun. 2020b. Feded: Federated learning via ensemble distillation for medical relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2118–2128.

Tianheng Wang, Ling Zheng, Hairong Lv, Chenghu Zhou, Yunheng Shen, Qinjun Qiu, Yan Li, Pufan Li, and Guorui Wang. 2023. A distributed joint extraction framework for sedimentological entities and relations with federated learning. *Expert Systems with Applications*, 213:119216.

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*.

Yan Xiao, Yaochu Jin, Ran Cheng, and Kuangrong Hao. 2022. Hybrid attention-based transformer block model for distant supervision relation extraction. *Neurocomputing*, 470:29–39.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14149–14157.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.

Hangyu Zhu and Yaochu Jin. 2019. Multi-objective evolutionary federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4):1310–1322.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. 2021. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390.

9. Language Resource References

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I. Furlong. 2016. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, pages D833–D839.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. *arXiv preprint arXiv:1906.04684*.

10. Appendices

Method	**Dev-Ign F1**	**Dev-F1**
Centralized Training		
Our	59.55	61.64
Federated Training (IID)		
FedAvg	50.82	51.95
FedAtt	52.03	53.22
FedLCC	55.56	56.34

Table 7: Results on DocRED.

To further verify the generalization capability of the framework, we conducted additional experiments on the document-level relation extraction dataset, DocRE, as suggested by one of the reviewers. As the DocRE dataset format differs from that of medical datasets, and since the test data is not publicly available, we conducted direct verification of the IID scenario on the validation set of the dataset. The results, illustrated in Table 7, demonstrate that the proposed FedLCC algorithm exhibits less performance degradation compared to other algorithms.