Label-Perceptive Adversarial Domain Adaptation for Named Entity Recognition in Traditional Chinese Medicine: Dataset and Approach

Yu Tong TONGYU@STU.EDU.CN
Shantou University

Editors: Hung-yi Lee and Tongliang Liu

Abstract

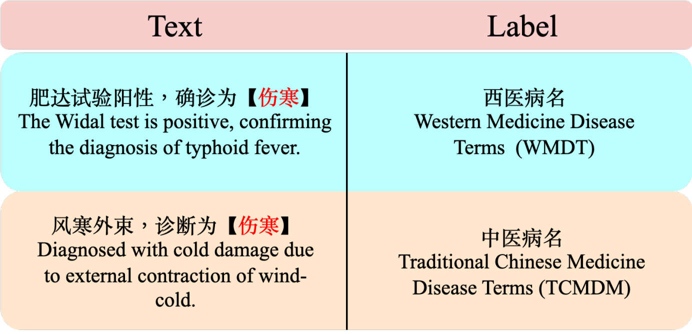
In the field of Traditional Chinese Medicine (TCM), Named Entity Recognition (NER) is a crucial task. However, the scarcity of NER datasets in TCM significantly hampers the performance of models in this domain. A promising approach to addressing this low-resource issue is through domain adaptation techniques. Current domain adaptation methods typically leverage large amounts of labeled data from a source domain to bridge the gap between the source and target domains, making the features of the generated target domain data as similar as possible to those of the source domain, thereby enhancing model performance in the target domain. However, existing methods primarily focus on aligning textual features and neglect the importance of label information. In the NER task, labels not only indicate categories but also carry important categorical information. Therefore, this paper proposes a Label-Perceptive Adversarial Domain Adaptation (LPADA) method that integrates label information with textual features, providing additional contextual information for the domain adaptation process, thus enhancing the model's performance in the TCM domain. Furthermore, we annotate medical case records to construct a dataset TCMNER2024 [1] and establish a baseline. The evaluation demonstrates that our approach significantly outperforms existing methods.

Keywords: Data Resource, Adversarial Domain Adaptation, Traditional Chinese Medicine, Named Entity Recognition

1. Introduction

TCM NER is crucial for extracting valuable information from traditional Chinese medicine texts, and supporting tasks such as knowledge graph construction, clinical decision-making, and automated question-answering systems. However, the development of accurate and efficient NER systems in the TCM domain faces significant challenges, primarily due to the lack of annotated datasets and resources. As a low-resource domain, TCM is plagued by the scarcity of high-quality data and the high costs associated with data annotation. Moreover, there have been few NER tasks specifically tailored to the TCM domain, and no systematic research framework or benchmark dataset has been established. Domain adaptation is an effective approach for addressing low-resource tasks by transferring models or knowledge trained on large-scale data to tasks in the target domain. This can significantly improve model performance and generalization. Adversarial training is often employed as a strategy for achieving domain adaptation. However, existing domain adaptation algorithms typically focus solely on aligning the feature distributions between the source and target domains to achieve adaptation. In the NER task, labels often carry crucial information about entity categories, which is highly valuable for domain adaptation. However, current methods frequently fail to fully leverage this label information, leading to inadequate differentiation of entity categories during the

[1] TCMNER2024 dataset can be accessed via https://github.com/TCMNER/TCMNER2024.



Text	Label
肥达试验阳性，确诊为【伤寒】 The Widal test is positive, confirming the diagnosis of typhoid fever.	西医病名 Western Medicine Disease Terms (WMDT)
风寒外束，诊断为【伤寒】 Diagnosed with cold damage due to external contraction of wind-cold.	中医病名 Traditional Chinese Medicine Disease Terms (TCMDM)

Figure 1: In the upper sentence, “The Widal test is positive, confirming the diagnosis of typhoid fever”, the entity “Typhoid Fever” falls under the “Western Medicine Disease Terms (WMDT)” category. In contrast, in the lower sentence, “Diagnosed with cold damage due to external contraction of wind-cold”, the entity “cold damage” is classified under the “Traditional Chinese Medicine Disease Terms (TCMDT)” category. Although both expressions are consistent in Chinese, they belong to entirely different entity categories.

transfer process. As shown in Figure 1, the entity “Typhoid Fever” in the upper sentence falls under the Western Medicine Disease Terms (WMDT) category. In contrast, the entity “cold damage” in the lower sentence is classified under the “Traditional Chinese Medicine Disease Terms (TCMDT)” category. Although both expressions are the same in Chinese, they belong to entirely different entity categories. Since the same text appears consistent in the feature space, focusing solely on text features while ignoring label information can lead to incorrect entity category classification.

Therefore, we propose a Label-Perceptive Adversarial Domain Adaptation (LPADA) framework for the NER task in the TCM domain via generative adversarial architecture. The workflow of the LPADA framework is illustrated in Figure 2, and it comprises two main components: the generator and the discriminator. The generator is built on the pre-trained model RoBERTa (Liu et al., 2019) combined with the typical sequence tagger CRF (Lafferty et al., 2001), producing intermediate outputs. It takes the target domain dataset as input and generates pseudo labels leveraging the pre-trained language model. The discriminator processes labeled data from the source domain, progressively minimizing the discrepancy between the source and target domains by aligning the joint feature distribution of the text and its corresponding labels. Simultaneously, it minimizes the gap between the predicted and true labels. Therefore, the overall loss consists of two components: one derived from supervised information and the other from the discrepancy between the source and target domains, which is backpropagated to the generator.

Furthermore, given the absence of systematic research and standardized evaluation benchmarks for the NER task in the TCM domain, we have developed the TCMNER2024 dataset to evaluate the LPADA framework and establish a baseline for future research. TCMNER2024 is a high-quality dataset derived from medical records within the TCM field. We evaluate the LPADA framework using the TCMNER2024 dataset, provide the evaluation results as a baseline, and make the dataset publicly available to support and advance future NER research in the TCM domain. The three key highlights of this work are as follows:

- We propose the LPADA framework, which leverages the strengths of adversarial mechanisms with label-perceptual. By integrating text features and label information through joint probability distribution, LPADA facilitates domain adaptation by aligning the feature spaces of both text and labels across source and target domains.

- We developed TCMNER2024, a comprehensive and robust benchmark dataset for the TCM NER task, derived from over 5,000 high-quality clinical records. This dataset provides a solid foundation for future research and advancements in the field.

- We evaluate the LPADA framework on the TCMNER2024 dataset, thoroughly comparing it with existing baseline models. The results show that LPADA significantly surpasses current methods, setting a new standard in the field.

2. Related Work

2.1. Application of Generative Adversarial Networks in NLP Domain

GANs had seen limited application in NLP due to the discrete nature of the text, which complicated the back-propagation of errors from the discriminator to the discrete data. While various solutions had been proposed to address this challenge, GANs had predominantly been used for text generation tasks (Zhang et al., 2016; Kusner and Hernández-Lobato, 2016; Lamb et al., 2016; Yu et al., 2017; Guo, 2015). For example, GANs had been successfully applied to dialogue generation by incorporating policy gradients from reinforcement learning (RL) (Li et al., 2017a). Additionally, RL had been instrumental in tasks such as poem, speech-language, and music generation through the use of stochastic policies (Yu et al., 2017). The actor-critic approach, another RL method, had been used to bridge the gap between training and testing models, enhancing language generation tasks (Brakel et al., 2017). Besides, researchers had also explored Deep Q-Networks (DQN), a popular RL variant, to iteratively decode output sequences, improving sequence-to-sequence learning and achieving a promising BLEU score when decoding unseen sentences (Guo, 2015). In summary, GANs had been primarily employed on text generation tasks, with limited application to sequence labeling tasks.

2.2. Domain Adaptation

2.2.1. Feature Distribution Alignment

Aligning feature distributions is a common strategy in domain adaptation, with metrics like Maximum Mean Discrepancy (MMD) (Borgwardt et al., 2006; Tzeng et al., 2014; Ma et al., 2019), Jensen-Shannon (JS) divergence (Chen and Cardie, 2018), and Wasserstein distance (Flamary et al., 2016; Shen et al., 2018) used to reduce differences and ensure the extraction of domain-invariant features. To improve sentiment classification, researchers developed the Structural Correspondence Learning (SCL) algorithm (Blitzer et al., 2007) to enhance sentiment classification by minimizing inter-domain error. Additionally, the Spectral Feature Alignment (SFA) algorithm (Pan et al., 2010) was introduced, clustering domain-specific words from different domains into unified groups using domain-independent words as a bridge. Other approaches included introducing large margins between classes within an embedding space, making it domain-agnostic by aligning data distributions across domains (Rostami and Galstyan, 2021).

2.2.2. ADVERSARIAL DOMAIN ADAPTATION

Researchers proposed to leverage domain adversarial training to select domain-discriminative data from the source domain and fine-tune BERT through curriculum learning (Ma et al., 2019). Other approaches, such as DANN (Ganin et al., 2016) focused on training features that are discriminative for the source domain but invariant to domain shifts, while MDAN (Zhao et al., 2018) extended this by incorporating multiple domain classifiers. The EADA framework employed adversarial learning in an unsupervised, energy-based setup, utilizing an autoencoder and feature extractor to align features with the source domain (Zou et al., 2021). Instead of manually selecting pivots for discriminative learning, the Adversarial Memory Network (AMN) provided an end-to-end solution for identifying pivots for cross-domain sentiment classification (Li et al., 2017b). Additionally, researchers combined masked language modeling with adversarial training to enhance domain-invariant features (Du et al., 2020). Frameworks such as WS-UDA and 2ST-UDA focused on transfer learning and unsupervised domain adaptation through weighting schemes and two-stage training, respectively (Dai et al., 2020).

2.3. Research on the NER task in the TCM Domain

The authors introduced a distantly supervised NER method designed to extract medical entities from TCM clinical records (Jia et al., 2021). Other authors built the domain dictionary leveraging a TCM knowledge graph (Zhang et al., 2017). Additionally, some leveraged web crawlers to assemble a TCM named entity corpus from sources such as KingNet [2] and CloudTCM [3], encompassing 1,097 articles, 1,412 disease names, and 38,714 TCM terms. Their evaluation revealed that the combination of RoBERTa with BiLSTM and CRF delivered the best performance among the models tested (Su et al., 2022). Moreover, a dataset of TCM medical cases related to chest discomfort was introduced, covering six entity types: “syndrome”, “symptom”, “etiology and pathogenesis”, “treatment method”, “medication”, and “prescription”. They utilized a word segmentation method to dynamically update the dictionary, combined with BiLSTM-CRF and IDCNN-CRF models which enhanced the recognition of both domain-specific and novel entities (Liu et al., 2023). Furthermore, a novel approach combining word and character information with a self-attention module was proposed, defining five distinct TCM entity classes. Evaluations on a comprehensive NER dataset, encompassing both published materials and clinical records, confirmed the method’s effectiveness (Liu et al., 2021a).

3. Proposed Method

Let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$ denote a corpus of annotated TCM texts, where $X_i = \{x_1, x_2, \ldots, x_T\}$ represents a tokenized input sequence and $Y_i = \{y_1, y_2, \ldots, y_T\}$ denotes the corresponding sequence of entity labels drawn from a predefined label set $\mathcal{L}$ (e.g., disease name, syndrome, symptom). The goal of TCM Named Entity Recognition (TCM-NER) is to learn a mapping function $f_\theta : X \to Y$, parameterized by θ, that accurately predicts the entity label sequence $\hat{Y} = f_\theta(X)$ for any unseen TCM text. Formally, the optimization objective is to minimize the expected loss between the predicted and gold label sequences: $\mathcal{L}(\theta) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[\ell(f_\theta(X), Y)\big]$, where $\ell(\cdot)$ denotes a sequence-level loss function such as the cross-entropy or conditional random field (CRF) loss.

[2] https://www.kingnet.com.tw/tcm/

[3] https://cloudtcm.com/

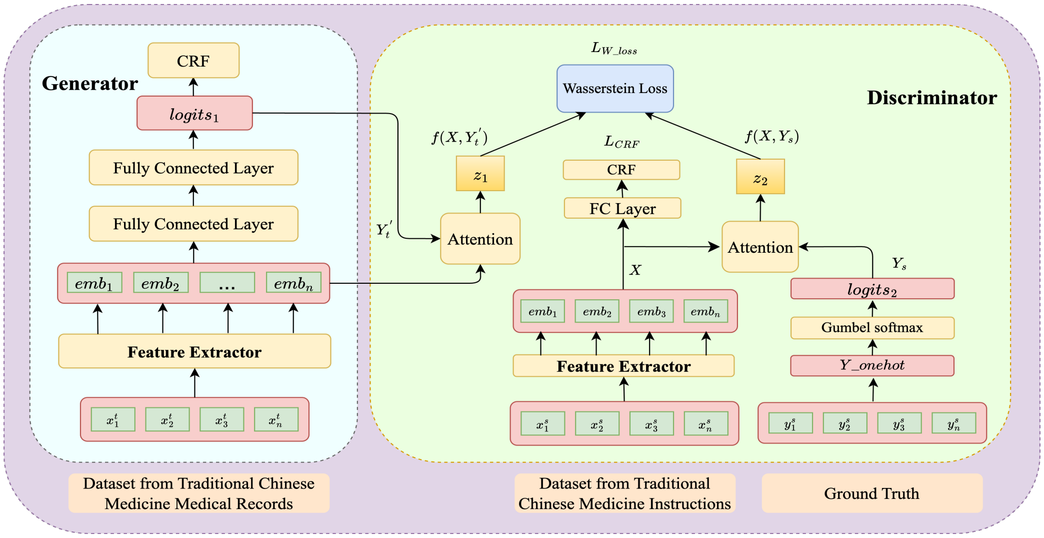


Figure 2: The training process of LPADA is based on a GAN framework, consisting of a generator and a discriminator. The generator (G) receives unlabeled data from the target domain and generates pseudo-labels using a pre-trained language model. Specifically, the pre-trained RoBERTa is employed as a feature extractor and leveraged to generate the pseudo label. Moreover, two Fully Connected (FC) layers are utilized to fine-tune the model for the specific target domain. The discriminator (D) simultaneously receives the intermediate output of G and labeled data from the source domain, and integrates the text and label information through a joint probability distribution. Here, an attention layer is incorporated to fuse the text and label features. On one hand, the features are fed into a CRF layer to calculate the CRF loss of the generated tag sequence. On the other hand, the discriminator (D) employs the Wasserstein loss to minimize the distribution gap between the two domains. The total loss consists of CRF loss and Wasserstein loss.

However, due to the scarcity of annotated TCM corpora, complex semantic compositions, and high lexical ambiguity, conventional sequence labeling models often struggle to achieve stable and domain-robust performance. Addressing these challenges requires a framework capable of enhancing feature representation and improving generalization under low-resource conditions. To this end, we propose the LPADA framework, based on the architecture of Generative Adversarial Networks (GANs), which consists primarily of two components, as illustrated in Figure 2:

- Generator: The generator's role is to create pseudo-labels for unlabeled, low-resource data. Specifically, it takes data from the target domain and leverages a pre-trained language model (such as RoBERTa, BERT) to generate these pseudo-labels.
- Discriminator: The discriminator receives labeled data from the source domain and pseudo-labeled data generated by the generator. Its task is to minimize the joint feature distribution discrepancy between the source domain data and the pseudo-labeled data, while also reducing the prediction loss for the source domain data. Through this adversarial training process, the generator continuously improves to produce more convincing pseudo-labels, thereby enhancing the quality of the pseudo-labeling.

The LPADA framework leverages adversarial training between the generator and discriminator to effectively utilize labeled data from the source domain, ultimately enhancing the model's performance in the target domain. As shown in Algorithm 1, we first initialize the parameters of the

Algorithm 1 The training process of LPADA

```
Input: Training dataset $S = (x_{1:N}, y_{1:N})$
Parameter: $G_\theta$, $D_\beta$
Output: $G_\theta$
Initialize $G_\theta$, $D_\beta$ .
while LPADA not converged do
    Sample mini_batch $(x^s, y^s) \sim (X^s, Y^s)$.
    Sample mini_batch $(x^t, y^t) \sim (X^t, Y^t)$.
    Get $logits1$ through $G(x^t)$.
    Encode $x^s$ as $emb_s^{'}$.
    Pre-process $y^s$ and get $logits2$.
    Get $z1$ by $Attention(logits_1, emb_t^{'})$.
    Get $z2$ by $Attention(logits_2, emb_s^{'})$.
    Compute $\mathcal{L}_{CRF}$ by (5).
    Compute $\mathcal{L}_{W\_loss}$ by (6).
    Update $D_\beta$.
    Update $G_\theta$.
end while
return $G_\theta$.
```

generator and the discriminator. Then, we sample source-domain data (x^s, y^s) and target-domain data (x^t, y^t), and train the generator (G) to obtain the $logits_1$. Subsequently, we train the discriminator by first encoding x^s into embeddings $emb_s^{'}$, preprocessing the source-domain labels y^s, and obtaining $logits_2$. Through the attention module, the textual and label representations from both the source and target domains $(logits_2, emb_s^{'})$ and $(logits_1, emb_t^{'})$ are integrated. Finally, the generator loss is computed according to Equation (5), and the discriminator loss is computed according to Equation (6). The parameters of the discriminator and generator are alternately updated until the model converges.

3.1. Generative Model

RoBERTa is employed as the encoder for the source domain data, projecting input sequences into a 1024-dimensional vector space. This is followed by two fully connected (FC) layers, which are used for fine-tuning and generating pseudo labels. Due to the discrete nature of the tags, the loss propagated back by the discriminator cannot directly update the labels. Therefore, we provide the intermediate continuous logits to the discriminator.

3.2. Discriminative Model

The discriminator consists of a pre-trained language model used as a feature extractor, complemented by a Gumbel softmax layer, an attention layer, a fully connected (FC) layer, and a CRF layer. It simultaneously processes the $logits_1$ from the target domain, $logits_2$, and the input sequence representation X^t and X^s from the target and source domains respectively. Here, $logits_1$ represents the intermediate output from the generator, while $logits_2$ is derived from the ground truth Y which is from the source domain. The feature dimension of x_i^s and x_i^t are both 1024, and the feature size of the entire sentence, denoted as X, is $(1024 * n)$, where n denotes the sequence length. The attention layer is employed to integrate the joint data distribution of text feature X and label Y. The outputs of these attention layers are referred to as z_1 and z_2. Here, $X^s = x_1^s, x_2^s, \ldots, x_n^s$, $X^t = x_1^t, x_2^t, \ldots, x_n^t$, and $Y \in logits_1, logits_2$. Before the attention operation, both X and Y are resized to middle_size for compatibility. After the attention mechanism is applied, the shape of z remains $(middle\_size \times n)$. Wasserstein distance is utilized to assess the difference in data distribution between the source and target domains. Moreover, the features from the FC layer are simultaneously fed into the CRF layer. The CRF layer then predicts the tag sequences and uses the ground truth Y to compute the CRF loss. We employ the "BISO" (Begin, Inside, Single, Other) tagging schema and the corresponding tags outputted from the CRF layer are defined as: $Y^{'} = \{y_1^{'}, y_2^{'}, ...y_n^{'}\}$ and $y_i^{'} \in \{B\_type, I\_type, S\_type, O\}$, $\forall i \leq n, i \in N^+$. The prefix indicates the related position of the current character in an entity, and the suffix "type" here stands for the entity type. The total number of tags is $n * m + 1$, where n is the entity type and m is the tag category. $Y^{'} \in L^n$, where L^n represents all the potential tag sequences. In summary, the decoder CRF attempts to search the optimal tag sequence $Y^{'}$ according to the input X where:

$$Y^{'} = \underset{Y \in L^n}{\arg\max}\, P(Y|X) \tag{1}$$

3.2.1. Gumbel Softmax Layer

Since the labels are discrete while the features are continuous, to ensure that both the labels and text features are in a continuous feature space, the labels Y_truth are first converted to a one-hot format. Then, Y_onehot is transformed into continuous $logits\_2$ using Gumbel softmax as described in Eq. 2.

$$logits\_2 = softmax(1/\tau(h + g * \alpha)) \tag{2}$$

here, h represents Y_onehot and g denotes the Gumbel-Softmax. The temperature τ is fixed at 0.5, with the parameter α is set to 0.1.

3.2.2. Attention Layer

A scaled dot-product attention mechanism is employed to model the joint data distribution of text features and labels by associating $(X^t, logits\_1)$ and $(X^s, logits\_2)$.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

$Q = X$, $K = V = Y$. The matrix X, representing the features generated by RoBERTa, has a size of $(1024 * n)$, while the matrix Y has a shape of $(label\_size * n)$. Here, $label\_size$ denotes the number of tag categories, and n represents the sequence length. To facilitate the attention operation, the matrix X is first reduced to a dense size matching Y. Both X and Y are then resized to a $middle\_size$, which is set to 200, with the dimension d_k equal to this middle_size.

3.3. Loss Function

The generator's primary role is to generate pseudo-labels for the unlabeled target domain data. The discriminator serves two functions. First, since it receives labeled source domain data, it calculates the loss between the predicted and true labels using the standard approach for the NER task, applying CRF loss in this process. Second, the discriminator assesses the difference in data distribution between the source and target domains, utilizing the Wasserstein loss for this purpose. Overall, the LPADA framework aims to minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{CRF} + \lambda \mathcal{L}_{W\_loss} \tag{4}$$

In Eq. (4), λ is a parameter that balances the weight of two kinds of loss and is set as 1. $\mathcal{L}_{CRF}$ and $\mathcal{L}_{W\_loss}$ are the CRF and Wasserstein loss (Arjovsky et al., 2017).

$$\mathcal{L}_{CRF} = -\log p(y|X) = -s(X, y) + \log(\sum_{y \in Y_x} e^{s(X,y)}) \tag{5}$$

here, s is the score function and Y_x are all potential tag sequences. As is well known, GAN architectures often encounter issues such as unstable training, vanishing gradients, and mode collapse. To address these inherent challenges, Wasserstein loss is employed to provide a more stable training process by measuring the discrepancy between data distributions from the source and target domains. This approach helps to mitigate the issues of gradient instability and improves the overall robustness of the training process.

$$\mathcal{L}_{WLoss} = \mathbb{E}_{x \sim \mathbb{P}_\gamma}[f_w(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f_w(x)] \tag{6}$$

4. Experiments

We evaluate the NER task within the Traditional Chinese Medicine (TCM) domain. Due to the lack of systematic research and the absence of a standard evaluation dataset for NER tasks in this domain, we constructed a specialized NER dataset for the low-resource TCM domain. This dataset serves as the target domain data for assessing the robustness and effectiveness of the proposed method. The source domain data we selected also comes from the field of traditional Chinese medicine, but the entity annotation was performed based on instructional texts. While the entity categories in the source and target domains overlap, they are not entirely identical. Table 1 compares the entity types present in the source and target domain datasets.

4.1. Datasets

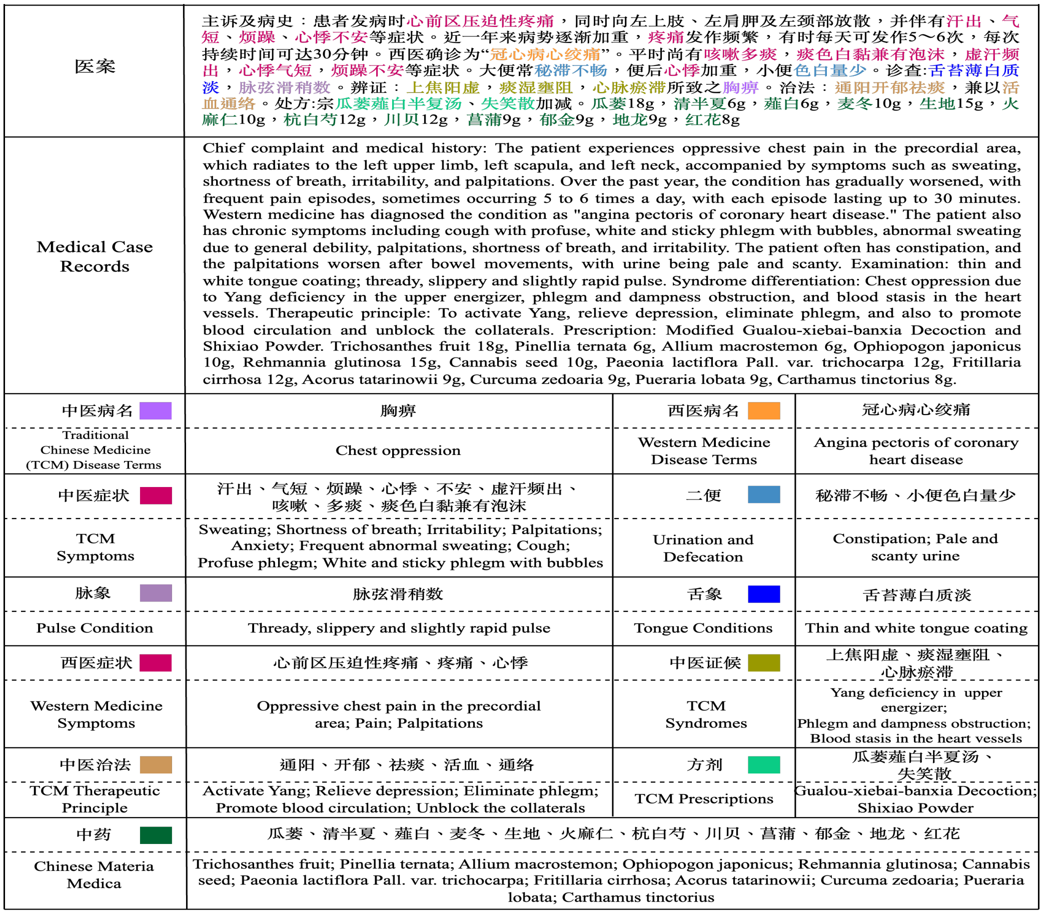
4.1.1. Source Domain

The source domain dataset we utilize comes from the "Wan Chuang Cup" TCM Tianchi Big Data Competition [4], specifically the Traditional Chinese Medicine Instructions Named Entity Recognition dataset (TCMI-NER), which includes 13 entity categories. The data is derived from Chinese medicine instructions, and the knowledge base is enriched by identifying and extracting entities from these instructions. The first row of Table 1 describes the various entity categories of the dataset.

4.1.2. Target Domain

The target domain data is annotated with entities extracted from medical records. We present the Traditional Chinese Medicine Named Entity Recognition 2024 (TCMNER2024) dataset, which contains 264,798 entities categorized into 11 distinct types.

[4] https://tianchi.aliyun.com/competition/entrance/531824/introduction



医案	主诉及病史：患者发病时心前区压迫性疼痛，同时向左上肢、左肩胛及左颈部放散，并伴有汗出、气短、烦躁、心悸不安等症状。近一年来病势逐渐加重，疼痛发作频繁，有时每天可发作5～6次，每次持续时间可达30分钟。西医确诊为“冠心病心绞痛”。平时尚有咳嗽多痰，痰色白黏兼有泡沫，虚汗频出，心悸气短，烦躁不安等症状。大便常秘滞不畅，便后心悸加重，小便色白量少。诊查:舌苔薄白质淡，脉弦滑稍数。辨证：上焦阳虚，痰湿壅阻，心脉瘀滞所致之胸痹。治法：通阳开郁祛痰，兼以活血通络。处方:宗瓜蒌薤白半夏汤、失笑散加减。瓜蒌18g，清半夏6g，薤白6g，麦冬10g，生地15g，火麻仁10g，杭白芍12g，川贝12g，菖蒲9g，郁金9g，地龙9g，红花8g		
Medical Case Records	Chief complaint and medical history: The patient experiences oppressive chest pain in the precordial area, which radiates to the left upper limb, left scapula, and left neck, accompanied by symptoms such as sweating, shortness of breath, irritability, and palpitations. Over the past year, the condition has gradually worsened, with frequent pain episodes, sometimes occurring 5 to 6 times a day, with each episode lasting up to 30 minutes. Western medicine has diagnosed the condition as "angina pectoris of coronary heart disease." The patient also has chronic symptoms including cough with profuse, white and sticky phlegm with bubbles, abnormal sweating due to general debility, palpitations, shortness of breath, and irritability. The patient often has constipation, and the palpitations worsen after bowel movements, with urine being pale and scanty. Examination: thin and white tongue coating; thready, slippery and slightly rapid pulse. Syndrome differentiation: Chest oppression due to Yang deficiency in the upper energizer, phlegm and dampness obstruction, and blood stasis in the heart vessels. Therapeutic principle: To activate Yang, relieve depression, eliminate phlegm, and also to promote blood circulation and unblock the collaterals. Prescription: Modified Gualou-xiebai-banxia Decoction and Shixiao Powder. Trichosanthes fruit 18g, Pinellia ternata 6g, Allium macrostemon 6g, Ophiopogon japonicus 10g, Rehmannia glutinosa 15g, Cannabis seed 10g, Paeonia lactiflora Pall. var. trichocarpa 12g, Fritillaria cirrhosa 12g, Acorus tatarinowii 9g, Curcuma zedoaria 9g, Pueraria lobata 9g, Carthamus tinctorius 8g.		
中医病名 🟪	胸痹	西医病名 🟧	冠心病心绞痛
Traditional Chinese Medicine (TCM) Disease Terms	Chest oppression	Western Medicine Disease Terms	Angina pectoris of coronary heart disease
中医症状 🟥	汗出、气短、烦躁、心悸、不安、虚汗频出、咳嗽、多痰、痰色白黏兼有泡沫	二便 🟦	秘滞不畅、小便色白量少
TCM Symptoms	Sweating; Shortness of breath; Irritability; Palpitations; Anxiety; Frequent abnormal sweating; Cough; Profuse phlegm; White and sticky phlegm with bubbles	Urination and Defecation	Constipation; Pale and scanty urine
脉象 🟪	脉弦滑稍数	舌象 🟦	舌苔薄白质淡
Pulse Condition	Thready, slippery and slightly rapid pulse	Tongue Conditions	Thin and white tongue coating
西医症状 🟥	心前区压迫性疼痛、疼痛、心悸	中医证候 🟨	上焦阳虚、痰湿壅阻、心脉瘀滞
Western Medicine Symptoms	Oppressive chest pain in the precordial area; Pain; Palpitations	TCM Syndromes	Yang deficiency in upper energizer; Phlegm and dampness obstruction; Blood stasis in the heart vessels
中医治法 🟫	通阳、开郁、祛痰、活血、通络	方剂 🟩	瓜蒌薤白半夏汤、失笑散
TCM Therapeutic Principle	Activate Yang; Relieve depression; Eliminate phlegm; Promote blood circulation; Unblock the collaterals	TCM Prescriptions	Gualou-xiebai-banxia Decoction; Shixiao Powder
中药 🟩	瓜蒌、清半夏、薤白、麦冬、生地、火麻仁、杭白芍、川贝、菖蒲、郁金、地龙、红花		
Chinese Materia Medica	Trichosanthes fruit; Pinellia ternata; Allium macrostemon; Ophiopogon japonicus; Rehmannia glutinosa; Cannabis seed; Paeonia lactiflora Pall. var. trichocarpa; Fritillaria cirrhosa; Acorus tatarinowii; Curcuma zedoaria; Pueraria lobata; Carthamus tinctorius		

Figure 3: Example of entity labeling for the TCMNER2024 dataset.

Dataset	Entity Categories
TCMI_NER	Drug, Drug_Ingredient, Disease, Symptom, Syndrome, Disease_Group, Food, Food_Group, Person_Group, Drug_Group, Drug_Dosage, Drug_Taste, Drug_Efficacy.
TCMNER2024	Traditional Chinese Medicine Disease Terms, Western Medicine Disease Terms, TCM Symptoms, Urination and Defecation, Pulse Conditions, Tongue Conditions, Western Medicine Symptoms, TCM Syndromes, TCM Therapeutic Methods, TCM Prescriptions, Chinese Materia Medica.

Table 1: Entity Types of TCMI_NER and TCMNER2024 Datasets.

4.1.3. Dataset Construction

The TCMNER2024 dataset is constructed from 5,172 medical case records, with contributions from over 400 renowned Chinese medicine practitioners. The dataset draws extensively from a diverse range of literature, including classical texts and published works in TCM. Figure 3 provides an ex-

Entity Category	Train	Dev	Test
TCM Disease Terms	1151	370	366
Western Medicine Disease Terms	2929	969	1025
TCM Symptoms	20826	6972	7466
Urination and Defecation	2391	774	961
Pulse Conditions	2948	959	981
Tongue Conditions	4346	1452	1492
Western Medicine Symptoms	12849	4864	4759
TCM Syndromes	6799	2161	2398
TCM Therapeutic Principle	5920	1843	1954
TCM Prescriptions	1674	492	949
Chinese Materia Medica	32025	10368	10653

Table 2: Data statistics of entity distribution for train, development, and test set in the TCMNER2024 Dataset.

Entities	**LSTM+CRF**	**BERT**	**RoBERTa**	**SLGAN**	**CrossNER**	**LPADA**	
	F1	F1	F1	F1	F1	F1	SD
TCM Disease Terms	27.42	35.92	40.64	41.03	42.06	**42.33**	0.19
Western Medicine Disease Terms	71.27	73.72	74.84	75.02	74.97	**75.66**	0.14
TCM Symptoms	45.18	49.26	50.04	50.12	50.03	**50.88**	0.16
Urination and Defecation	76.39	80.14	80.42	80.93	81.08	**81.86**	0.15
Pulse Conditions	93.04	95.77	95.92	96.04	95.91	**96.10**	0.12
Tongue Conditions	93.97	94.89	95.33	95.31	95.05	**95.39**	0.13
Western Medicine Symptoms	36.33	46.13	**48.43**	48.07	48.23	48.41	0.16
TCM Syndromes	82.88	84.18	84.54	85.38	84.97	**85.48**	0.17
TCM Therapeutic Principle	84.52	87.52	88.55	88.92	88.49	**89.27**	0.14
TCM Prescriptions	75.61	75.18	78.17	78.21	78.03	**79.42**	0.16
Chinese Materia	95.99	97.22	97.58	97.47	97.37	**97.61**	0.08
Avg	71.15	74.54	75.86	76.05	76.02	**76.58**	0.15

Table 3: Evaluation results of various models on TCMNER2024. F1 is the metrics. SD represents the Standard Deviation obtained from five experiments.

ample of entity labeling from the TCMNER2024 dataset, demonstrating the meticulous annotation process applied to ensure the accuracy and comprehensiveness of the dataset.

4.1.4. Data Collection and Annotation

We access only publicly available records and ensure all data are thoroughly anonymized to protect patient privacy. The text is preprocessed, focusing exclusively on the initial consultation descriptions, which are central to TCM diagnosis. To guarantee the accuracy and reliability of the annotated dataset, leading experts in the field of TCM established comprehensive annotation guidelines and category classifications. These guidelines ensure consistency throughout the annotation process, allowing the dataset to serve as a reliable resource for the TCM NER task. Based on traditional Chinese medical case records, we classify and annotate entities into 11 categories, as outlined in Table 2. This table displays the count of entities for each category. The reason for choosing these 11 categories is that they are the most commonly mentioned basic elements in medical case records and are most relevant to the descriptions within the records. To maintain consistency and accuracy in the annotation process, we develop annotation tools and accompanying usage instructions. Annotators are trained and calibrated to ensure a thorough understanding of the guidelines. The dataset is annotated by two independent annotators, with only the agreed-upon annotations retained after consistency checks. In cases of discrepancies, unclear guidelines, or ambiguous annotations, the annotators offer feedback, which experts use to refine and modify the guidelines. This process ensures that the final annotated data is consistent and reliable. More annotation details are presented in supplementary material.

4.2. Experimental Settings

The TCMNER2024 dataset is divided into training, development, and test sets in a 60%-20%-20% ratio. Table 4 provides the statistics for both the source and target domain datasets, while Table 2 outlines the entity distribution for each category within the TCMNER2024 dataset. The "BISO" tagging schema is employed, with the sequence length configured to 256, the batch size set to 16, and the learning rate adjusted to 2e-5. Training is conducted over 10 epochs, incorporating an early stopping mechanism to prevent overfitting. To evaluate the proposed LPADA, we use precision, recall, and micro F1 score as performance metrics. We present the average results obtained from 10 independent experimental runs to ensure robustness and reduce variance. This approach provides a more reliable assessment of the model's performance across multiple trials. We train the model on a machine equipped with an NVIDIA 3090 GPU to efficiently handle large-scale data and accelerate the training process. We acknowledge that GANs typically need large datasets for stable convergence. To mitigate this, we initialize the GAN with a task-specific pretrained encoder–decoder, stabilizing training and reducing data demands during fine-tuning.

4.3. Experimental Results

Since no prior research has been conducted on the TCMNER2024 dataset, we evaluate several representative models to establish baseline performance metrics. The results of models like RoBERTa represent the evaluation outcomes based on training and testing solely with target domain data, without employing any domain adaptation techniques. As shown in Table 3, LPADA consistently outperforms the baseline methods to varying degrees. Through domain adaptation, even when using data from slightly different domains, strong results can be achieved, sometimes even outperforming models trained on data from the same domain. This approach effectively addresses the challenge of acquiring sufficient data in low-resource domains, offering a practical solution for improving model performance in such settings. Moreover, we observe that the performance on categories such as TCMDM, TCMSY, and WMS falls noticeably below the average F1 score, underscoring the persistent challenges faced by modern models in handling the NER task across a diverse range of categories in low-resource domains. These results highlight the limitations of current state-of-the-art models, which, while powerful, still struggle to fully capture the complexities in these specific categories. This gap suggests considerable potential for further improvement, especially in domain adaptation and fine-grained entity recognition in specialized fields like Traditional Chinese Medicine. Furthermore, to ensure more objective and reliable experimental results while assessing the model's reliability and consistency, we analyze the Standard Deviation (SD) values across multiple experiments. This evaluation reinforces the robustness and stability of the LPADA framework. In addition, we select domain adaptation techniques specifically designed for NER tasks (Liu et al., 2021b) for comparison. The results further highlight the significant advantages of LPADA.

4.4. Validation on General-Domain Datasets

The framework and experiments primarily focus on TCM domain, which may be too specific for the general NLP community. To further verify the generalizability of the proposed algorithm, we

Dataset	Domain	Train	Dev	Test	Classes
Source	TCMI_NER	1,000	500	497	13
Target	TCMNER2024	2,846	949	949	11

Table 4: Statistical Analysis of the TCMI_Tianchi and TCMNER2024 Dataset.

	OnNote	**Weibo**	**MSRA**	**Resume**	**CoNLL**
Li et al., 2020	**84.47**	-	**96.72**	-	93.33
Wu et al., 2021	82.57	70.43	96.24	95.98	-
Zhu and Li, 2022	82.83	**72.66**	96.26	96.66	93.65
Xiong et al., 2023	81.47	68.23	95.42	-	-
Yang et al., 2023	82.66	71.94	-	96.2	-
LPADA	82.91	72.05	96.60	**96.75**	**93.74**
LPADA (SD)	0.19	0.15	0.14	0.21	0.20

Table 5: LPADA evaluated on general domain NER datasets v.s. existing primary methods.

followed previous studies and conducted additional validation on general-domain NER datasets. Table 5 presents the evaluation results, showing that LPADA achieves near SOTA performance on general domain, thereby demonstrating the framework's versatility and generalizability.

4.5. Ablation Study and Qualitative Analysis

To further analyze the impact of the proposed LPADA, we conduct a component-wise analysis to validate each part, aiming to clarify its specific contribution. We conduct a series of ablation studies under the TCM NER task setting, focusing respectively on adversarial training, the attention mechanism, the sequence tagging module, and the metric employed for evaluating distribution alignment. The experimental design is as follows:

- Removal of adversarial training: to assess its role in mitigating domain shift.
- Exclusion of the attention mechanism: to examine its impact on entity boundary recognition.
- Replacement of the sequence tagger: to test how different decoding architectures affect performance.
- Substituting the Wasserstein distance with a naive alternative (e.g., L1 distance): to test whether the proposed Wasserstein-based distribution alignment contributes significantly beyond simple distance-based losses.

Due to page limitations, we provide more detailed descriptions in the supplementary material. Through ablation experiments, we observed that the performance gains of the proposed LPADA framework in NER tasks primarily stem from the synergistic effect of domain-adaptive adversarial training and the attention mechanism. In contrast, variations in the distribution alignment metric and the sequence labeling module exert relatively minor influence on the overall framework performance.

5. Conclusion

We proposed the LPADA framework, which effectively leverages GANs to tackle the NER task in the TCM domain. LPADA introduces a domain adaptation approach by minimizing the joint probability distribution of text features and labels. Evaluation results confirm the effectiveness of LPADA in handling data in low-resource domains. Additionally, we developed TCMNER2024, a

Named Entity Recognition dataset specifically designed to address the unique challenges within the TCM domain. This dataset establishes a benchmark standard for future research.

6. Limitations

This study primarily addresses the NER task in the field of TCM. To more comprehensively demonstrate the effectiveness of the proposed method, we plan to extend our validation to cross-domain classification tasks. In future work, we will validate the proposed method across tasks in different domains to more robustly demonstrate its effectiveness in addressing low-resource issues. Moreover, the potential limitations of the proposed TCMNER2024 dataset are listed below:

- Limited Data Sources: The NER dataset used in this study is primarily derived from medical case records, which may limit its diversity and consequently restrict the model's generalization ability.
- Annotation Quality and Subjectivity: TCM NER involves complex entities such as diseases, prescriptions, herbs, and syndromes, where entity boundaries and naming conventions may depend on annotators' expertise. Due to the ambiguity and polysemy of TCM terminology, different annotators may provide inconsistent annotations, introducing subjectivity bias.
- Limited Dataset Scale: The acquisition and annotation of TCM-related texts are costly, resulting in relatively small dataset sizes, which may not be sufficient for deep learning models. The dataset may lack enough low-frequency entities, leading to poor model performance on rare terms.

Moreover, as LPADA is based on a generative adversarial mechanism, it inherits GAN's training instability and requires nearly twice the training time of a single-module framework.

7. Ethical Statement

Our study adheres to ethical principles in the following aspects:

- Data Privacy and Compliance: The NER datasets used in this study are either publicly available or properly authorized, with no personal identifiable information involved. If medical records or sensitive healthcare data are used, appropriate anonymization measures have been taken to ensure compliance with relevant regulations.
- Fairness and Bias: We are committed to developing fair and unbiased NER models. We analyzed the data distribution and implemented measures to mitigate potential biases, improving model fairness and interpretability.
- Potential Impact and Risks: The goal of this research is to enhance the automated processing of TCM texts and improve knowledge utilization, rather than replace professional medical decision-making. While the research outcomes may support medical information extraction and auxiliary diagnosis, final decisions should be made by professionals to avoid potential risks.
- Sustainability and Responsible AI: We encourage the responsible use of our research findings in accordance with ethical AI principles. The study's code and (if permissible) data will be shared under appropriate licenses to ensure transparency and reproducibility.

Acknowledgments

We sincerely appreciate and thank all the reviewers for their valuable feedback. This research was supported by the Scientific Research Starting Foundation of Shantou University under Project No.NTF23003.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 214–223, 2017.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447, 2007.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Dzmitry Bahdanau Philemon Brakel, Kelvin Xu Anirudh Goyal, Ryan Lowe Joelle Pineau, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. In *Conference ICLR*, 2017.

Xilun Chen and Claire Cardie. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1226–1240, 2018.

Yong Dai, Jian Liu, Xiancong Ren, and Zenglin Xu. Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7618–7625, 2020.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 4019–4028, 2020.

Rémi Flamary, Nicholas Courty, Davis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1(1-40):2, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Hongyu Guo. Generating text with deep reinforcement learning. *arXiv preprint arXiv:1510.09202*, 2015.

Qi Jia, Dezheng Zhang, Haifeng Xu, and Yonghong Xie. Extraction of traditional chinese medicine entity: Design of a novel span-level named entity recognition method with distant supervision. *JMIR Med Inform*, 9(6):e28219, Jun 2021. ISSN 2291-9694. doi: 10.2196/28219.

Matt J Kusner and José Miguel Hernández-Lobato. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*, 2016.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Alex M Lamb, Anirudh Goyal Alias Parth Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances in neural information processing systems*, pages 4601–4609, 2016.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark, September 2017a. Association for Computational Linguistics. doi: 10.18653/v1/D17-1230.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, 2020.

Zheng Li, Yun Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, pages 2237–2243, 2017b.

Qingping Liu, Lunlun Zhang, Gao Ren, and Beiji Zou. Research on named entity recognition of traditional chinese medicine chest discomfort cases incorporating domain vocabulary features. *Comput. Biol. Med.*, 166(C), nov 2023. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2023.107466.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zhi Liu, Changyong Luo, Zeyu Zheng, Yan Li, Dianzheng Fu, Xinzhu Yu, and Jiawei Zhao. Tcmner and pubmed: a novel chinese character-level-based model and a dataset for tcm named entity recognition. *Journal of Healthcare Engineering*, 2021:1–10, 2021a.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460, 2021b.

Xiaofei Ma, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. Domain adaptation with bert-based domain classification and data selection. 2019.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760, 2010.

Mohammad Rostami and Aram Galstyan. Domain adaptation for sentiment analysis using increased intraclass separation. *arXiv preprint arXiv:2107.01598*, 2021.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ming-Hsiang Su, Chin-Wei Lee, Chi-Lun Hsu, and Ruei-Cyuan Su. RoBERTa-based traditional Chinese medicine named entity recognition model. In Yung-Chun Chang and Yi-Chin Huang, editors, *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 61–66, Taipei, Taiwan, November 2022. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Shuang Wu, Xiaoning Song, and Zhenhua Feng. Mect: Multi-metadata embedding based cross-transformer for chinese named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1529–1539, 2021.

Limao Xiong, Jie Zhou, Qunxi Zhu, Xiao Wang, Yuanbin Wu, Qi Zhang, Tao Gui, Xuan-Jing Huang, Jin Ma, and Ying Shan. A confidence-based partial label learning model for crowd-annotated named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1375–1386, 2023.

Jiuding Yang, Jinwen Luo, Weidong Guo, Di Niu, and Yu Xu. Exploiting hierarchically structured categories in fine-grained chinese named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3407–3421, 2023.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Dezheng Zhang, Yonghong Xie, Man Li, and Chuan Shi. Construction of knowledge graph of traditional chinese medicine based on the ontology. *Technology intelligence engineering*, 3(1): 35–42, 2017.

Yizhe Zhang, Zhe Gan, and Lawrence Carin. Generating text via adversarial training. In *NIPS workshop on Adversarial Training*, volume 21, 2016.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Enwei Zhu and Jinpeng Li. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, 2022.

Han Zou, Jianfei Yang, and Xiaojian Wu. Unsupervised energy-based adversarial domain adaptation for cross-domain text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1208–1218, 2021.