Label-aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition

Zhenghui Wang[†], Yanru Qu[†], Liheng Chen[†], Jian Shen[†], Weinan Zhang[†*]
Shaodian Zhang[†‡], Yimei Gao[‡], Gen Gu[‡], Ken Chen[‡], Yong Yu[†]
[†]APEX Data and Knowledge Management Lab, Shanghai Jiao Tong University
[‡]Synyi LLC.
{felixwzh,wnzhang,shaodian}@apex.sjtu.edu.cn
chen.ken@synyi.com

Abstract

We study the problem of named entity recognition (NER) from electronic medical records, which is one of the most fundamental and critical problems for medical text mining. Medical records which are written by clinicians from different specialties usually contain quite different terminologies and writing styles. The difference of specialties and the cost of human annotation makes it particularly difficult to train a universal medical NER system. In this paper, we propose a label-aware double transfer learning framework (La-DTL) for cross-specialty NER, so that a medical NER system designed for one specialty could be conveniently applied to another one with minimal annotation efforts. The transferability is guaranteed by two components: (i) we propose label-aware MMD for feature representation transfer, and (ii) we perform parameter transfer with a theoretical upper bound which is also label aware. We conduct extensive experiments on 12 cross-specialty NER tasks. The experimental results demonstrate that La-DTL provides consistent accuracy improvement over strong baselines. Besides, the promising experimental results on non-medical NER scenarios indicate that La-DTL is potential to be seamlessly adapted to a wide range of NER tasks.

1 Introduction

The development of hospital information system and medical informatics drives the leverage of various medical data for a more efficient and intelligent medical care service. Among many kinds of medical data, electronic health records (EHRs) are one of the most valuable and informative data as they contain detailed information about the patients and the clinical practices. EHRs are essential to many intelligent clinical applications, such as hospital quality control and clinical decision support systems (Wu et al., 2015). Most of EHRs are recorded in an unstructured form, i.e., natural language. Hence, extracting structured information from EHRs using natural language processing (NLP), e.g., named entity recognition (NER) and entity linking, plays a fundamental role in medical informatics (Zhang and Elhadad, 2013). In this paper, we focus on medical NER from EHRs, which is a fundamental task and is widely studied in the research community (Nadeau and Sekine, 2007; Uzuner et al., 2011).

In practice, the difficulty of building a universally robust and high-performance medical NER system lies in the variety of medical terminologies and expressions among different departments of specialties and hospitals. However, building separate NER systems for so many specialties comes with a prohibitively high cost. The data privacy issue further discourages the sharing of the data across departments or hospitals, making it more difficult to train a canonical NER system to be applied everywhere. This raises a natural question: if we have sufficient annotated EHRs data in one *source* specialty, can we distill the knowledge and transfer it to help training models in a related *target* specialty with few annotations? By transferring the knowledge we can achieve higher performance in target specialties with lower annotation cost and bypass the data sharing concerns. This is commonly referred to as *transfer learning* (Pan and Yang, 2010).

Current state-of-the-art transfer learning methods for NER are mainly based on deep neural networks, which perform an end-to-end training to distill sequential dependency patterns in the natural language (Ma and Hovy, 2016; Lample et al., 2016). These transfer learning methods include (i) feature representation transfer (Peng and Dredze, 2017; Kulkarni et al., 2016), which normally lever-

*Weinan Zhang is the corresponding author.

ages deep neural networks to learn a close feature mapping between the source and target domains, and (ii) parameter transfer (Murthy et al., 2016; Yang et al., 2017), which performs parameter sharing or joint training to get the target-domain model parameters close to those of the source-domain model. To the best of our knowledge, there is no previous literature working on transfer learning for NER in the medical domain, or even in a larger scope, i.e., medical natural language processing.

In this paper, we propose a novel NER transfer learning framework, namely label-aware double transfer learning (La-DTL): (i) We leverage bidirectional long-short term memory (Bi-LSTM) network (Graves and Schmidhuber, 2005) to automatically learn the text representations, based on which we perform a label-aware feature representation transfer. We propose a variant of maximum mean discrepancy (MMD) (Gretton et al., 2012), namely label-aware MMD (La-MMD), to explicitly reduce the domain discrepancy of feature representations of tokens with the same label between two domains. (ii) Based on the learned feature representations from Bi-LSTM, two conditional random field (CRF) models are performed for sequence labeling for source and target domain separately, where parameter transfer learning is performed. Specifically, an upper bound of KL divergence between the source and target domain's CRF label distributions is added over the emission and transition matrices across the source and target CRF models to explore the shareable parts of the parameters. Both (i) and (ii) have a label-aware characteristic, which will be discussed later. We further argue that label-aware characteristic is crucial for transfer learning in sequence labeling problems, e.g., NER, because only when the corresponding labels are matched, can the "similar" contexts (i.e. feature representation) and model parameters be efficiently borrowed to improve the label prediction.

Extensive experiments are conducted on 12 cross-specialty medical NER tasks with real-world EHRs. The experimental results demonstrate that La-DTL provides consistent accuracy improvement over strong baselines, with overall 2.62% to 6.70% absolute F1-score improvement over the state-of-the-art methods. Besides, the promising experimental results on other two non-medical NER scenarios indicate that La-DTL has the potential to be seamlessly adapted to a wide range of NER tasks.

2 Related Works

Named Entity Recognition (NER) is fundamental in information extraction area which aims at automatic detection of named entities (e.g., person, organization, location and geo-political) in free text (Marrero et al., 2013). Many high-level applications such as entity linking (Moro et al., 2014) and knowledge graph construction (Hachey et al., 2011) could be built on top of an NER system. Traditional high-performance approaches include conditional random fields models (CRFs) (Lafferty et al., 2001), maximum entropy Markov models (MEMMs) (McCallum et al., 2000) and hidden Markov models (HMMs). Recently, many neural network-based models have been proposed (Collobert et al., 2011; Chiu and Nichols, 2016; Ma and Hovy, 2016; Lample et al., 2016), in which few feature engineering works are needed to train a high-performance NER system. The architecture of those neural network-based models are similar, where different neural networks (LSTMs, CNNs) at different levels (char- and word-level) are applied to learn feature representations, and on top of neural networks, a CRF model is employed to make label predictions.

Transfer Learning distills knowledge from a source domain to help create a high-performance learner for a target domain. Transfer learning algorithms are mainly categorized into three types, namely instance transfer, feature representation transfer and parameter transfer (Pan and Yang, 2010). Instance transfer normally samples or reweights source-domain samples to match the distribution of the target domain (Chen et al., 2011; Chu et al., 2013). Feature representation transfer typically learns a feature mapping which projects source and target domain data simultaneously onto a common feature space following similar distributions (Zhuang et al., 2015; Long et al., 2015; Shen et al., 2017). Parameter transfer normally involves a joint or constrained training for the models on source and target domains, usually introduce connections between source target parameters via sharing (Srivastava and Salakhutdinov, 2013), initialization (Perlich et al., 2014), or inter-model parameter penalty schemes (Zhang et al., 2016).

Transfer Learning for NER Training a high-performance NER system requires expensive and

time-consuming manually annotated data. But sufficient labeled data is critical for the generalization of an NER system, especially for neural network-based models. Thus, transfer learning for NER is a practically important problem. The first group of methods focuses on sharing model parameters but they differ in the training schemes. He and Sun (2017) proposed to train the parameter-shared model with source and target data jointly, while the learning rates for sentences from source domain are re-weighted by the similarity with target domain corpus. Yang et al. (2017) proposed a family of frameworks which share model parameters in hierarchical recurrent networks to handle cross-application, cross-lingual, and cross-domain transfer in sequence labeling tasks. Differently, Lee et al. (2017) first trained the model with source domain data and then fine-tuned the model with little annotated target domain data.

Domain adaptation method has been well studied in NER scenarios such as using distributed word representations (Kulkarni et al., 2016) and leveraging rule-based annotators (Chiticariu et al., 2010). Multi-task learning has also been studied to improve performance in multiple NER tasks by transferring meaningful knowledge from other tasks (Collobert et al., 2011; Peng and Dredze, 2016). To take the advantages of both domain adaptation and multi-task learning, Peng and Dredze (2017) proposed a multi-task domain adaptation model.

3 Preliminaries

This section briefly introduces bidirectional LSTM, conditional random field and maximum mean discrepancy, which are the building blocks of our transfer learning framework.

Bidirectional LSTM Recurrent neural networks (RNNs) are widely used in NLP tasks for their great capability to capture contextual information in sequence data. A widely used variant of RNNs is long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), which incorporates input and forget gates to capture both long and short term dependencies. Furthermore, it will be beneficial if we process the sequence in not only a forward but also a backward way. Thus, bidirectional LSTM (Bi-LSTM) was employed in many previous works (Chiu and Nichols, 2016; Ma and Hovy, 2016; Lample et al., 2016) to capture bidirectional information in a sequence. More specifically, for token $\mathbf{x}_t$ (embedding vector) at timestep t in sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, the θ_b-parameterized Bi-LSTM recurrently updates hidden vectors $\mathbf{h}_t^{\rightarrow} = G^f_{\theta_b}(\mathbf{X}, \mathbf{h}_{t-1}^{\rightarrow})$ and $\mathbf{h}_t^{\leftarrow} = G^b_{\theta_b}(\mathbf{X}, \mathbf{h}_{t+1}^{\leftarrow})$ produced by a forward LSTM and a backward one, respectively. Then we concatenate $\mathbf{h}_t^{\rightarrow}$ and $\mathbf{h}_t^{\leftarrow}$ to $\mathbf{h}_t$ as the final hidden vector produced by Bi-LSTM:

$$\mathbf{h}_t = \mathbf{h}_t^{\rightarrow} \oplus \mathbf{h}_t^{\leftarrow}.$$

The representations learned from Bi-LSTM for sequence $\mathbf{X}$ is thus denoted as $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n)$.

Conditional Random Field The goal of NER is to detect named entities in a sequence $\mathbf{X}$ by predicting a sequence of labels $\mathbf{y} = (y_1, y_2, ..., y_n)$. Conditional random field (CRF) is widely used to make joint labeling of the tokens in a sequence (Lafferty et al., 2001).

Recently, Lample et al. (2016) proposed to build a CRF layer on top of a Bi-LSTM so that the automatically learned feature representation $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n)$ of the sequence can be directly fed into the CRF for sequence labeling. For a sequence of labels $\mathbf{y}$, given the hidden vector sequence $\mathbf{H}$, we define its θ_c-parametrized score function $s_{\theta_c}(\mathbf{H}, \mathbf{y})$ as:

$$s_{\theta_c}(\mathbf{H}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{E}_{i,y_i} + \sum_{i=1}^{n-1} \mathbf{A}_{y_i, y_{i+1}},$$

where $\mathbf{E}$ is the emission score matrix of size $n \times m$ (m is the number of unique labels), and is computed by $\mathbf{E} = \mathbf{HW}$ where $\mathbf{W}$ is the label emission parameter matrix; $\mathbf{A}$ is the label transition parameter matrix; thus $\theta_c = \{\mathbf{W}, \mathbf{A}\}$. We then define the conditional probability of label sequence $\mathbf{y}$ given $\mathbf{H}$ by a softmax over all possible label sequences in set $\mathcal{Y}(\mathbf{H})$ as:

$$\begin{aligned} p_{\theta_c}(\mathbf{y}|\mathbf{H}) &= \exp\{s_{\theta_c}(\mathbf{H}, \mathbf{y})\}/Z(\mathbf{H}) \\ &= \exp\{s_{\theta_c}(\mathbf{H}, \mathbf{y})\} \Big/ \sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{H})} \exp\{s_{\theta_c}(\mathbf{H}, \mathbf{y}')\}, \end{aligned} \quad (1)$$

where θ_c is omitted for simplification in the following part. The training objective in the CRF layer is to maximize the log-likelihood $\max_{\theta_c} \log p(\mathbf{y}|\mathbf{H})$. In the label prediction phase, we give the output label sequence $\mathbf{y}^*$ with the highest conditional probability $\mathbf{y}^* = \arg\max_{\mathbf{y}' \in \mathcal{Y}(\mathbf{H})} p(\mathbf{y}'|\mathbf{H})$ by dynamic programming (Sutton et al., 2012).

Maximum Mean Discrepancy Maximum Mean Discrepancy (Gretton et al., 2012) is a non-

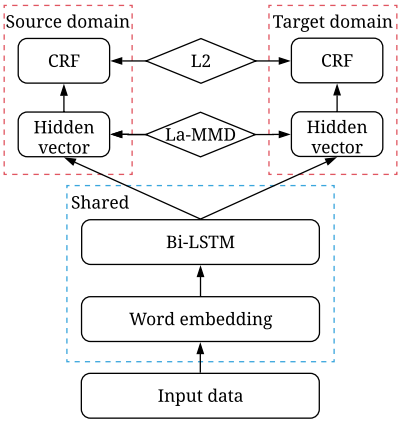

Figure 1: La-DTL framework overview: embedding and Bi-LSTM layers are shared across domains, predictors in red (upper) boxes are task-specific CRFs, with label-aware MMD and L2 constraints to perform feature representation transfer and parameter transfer.

parametric test statistic to measure the distribution discrepancy in terms of the distance between the kernel mean embeddings of two distributions p and q. The MMD is defined in particular function spaces that witness the difference in distributions

$$\text{MMD}(\mathcal{F}, p, q) = \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)]).$$

By defining the function class $\mathcal{F}$ as the unit ball in a universal Reproducing Kernel Hilbert Space (RKHS), denoted by $\mathcal{H}$, it holds that $\text{MMD}[\mathcal{F}, p, q] = 0$ if and only if $p = q$. And then given two sets of samples $X = \{x_1, ..., x_m\}$ and $Y = \{y_1, ..., y_n\}$ independently and identically distributed (i.i.d.) from p and q on the data space $\mathcal{X}$, the empirical estimate of MMD can be written as the distance between the empirical mean embeddings after mapping to RKHS

$$\text{MMD}(X, Y) = \left\| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(y_j) \right\|_{\mathcal{H}}, \quad (2)$$

where $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$ is the nonlinear feature mapping that induces $\mathcal{H}$.

4 Methodology

In this section, we present a label-aware double transfer learning (La-DTL) framework and discuss its rationale.

4.1 Framework Overview

Figure 1 gives an overview of La-DTL for NER. From bottom up, each input sentence is converted into a sequence of embedding vectors, which are then fed into a Bi-LSTM to sequentially encode contextual information into fixed-length hidden vectors. The embedding and Bi-LSTM layers are shared among source/target domains. With label-aware maximum mean discrepancy (La-MMD) to reduce the feature representation discrepancy between two domains, the hidden vectors are directly fed into source/target domain specific CRF layers to predict the label sequence. We use domain constrained CRF layers to enhance the target domain performance.

More formally, let $\mathcal{D}_s = \{(\mathbf{X}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N^s}$ be the training set of N^s samples from the source domain and $\mathcal{D}_t = \{(\mathbf{X}_i^t, \mathbf{y}_i^t)\}_{i=1}^{N^t}$ be the training set of N^t samples from the target domain, with $N^t \ll N^s$. Bi-LSTM encodes a sentence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ to hidden vectors $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n)$. We occasionally use $\mathbf{H}(\mathbf{X})$ to denote the corresponding hidden vectors when feeding $\mathbf{X}$ into the Bi-LSTM. CRF decodes hidden vectors $\mathbf{H}$ to a label sequence $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$. Our goal is to improve label prediction accuracy on the target domain $\mathcal{D}_t$ by utilizing the knowledge from the source domain $\mathcal{D}_s$:

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) =& p(\mathbf{y}|\mathbf{H}(\mathbf{X})), \\ \log p(\mathbf{y}|\mathbf{H}) =& \sum_{i=1}^{n} \mathbf{E}_{i,y_i} + \sum_{i=1}^{n-1} \mathbf{A}_{y_i,y_{i+1}} - \log Z(\mathbf{H}). \end{aligned} \quad (3)$$

Thus training a transferable model $p(\mathbf{y}|\mathbf{X})$ requires both $\mathbf{H}(\mathbf{X})$ and $p(\mathbf{y}|\mathbf{H})$ to be transferable.

We use share word embedding and Bi-LSTM by approaching the feature representation distributions $p(\mathbf{h}|\mathcal{D}_s)$ and $p(\mathbf{h}|\mathcal{D}_t)$, i.e., the distributions of Bi-LSTM hidden vectors at each timestep of the sentences from the source and target domains respectively. The rationale behind it lies on the insufficiency of labeled target data. Even though LSTM has high capacity, its generalization ability highly relies on viewing "sufficient" data. Otherwise, LSTM is very likely to overfit the data. Training on both source and target data, the Bi-LSTM is expected to learn feature representations with high quality. Yosinski et al. (2014) provided a justification of this solution that sharing bottom layers is promising for transfer learning in practice.

With the sentences projected onto the same hidden space, the conditional distribution $p(\mathbf{h}^s|\mathcal{D}_s)$ and $p(\mathbf{h}^t|\mathcal{D}_t)$, however, may be distant because

LSTM hidden vectors contain contextual information which is different across domains. In order to reduce source/target discrepancy, we refine MMD (Gretton et al., 2012) with label constraints, i.e., label-aware MMD (La-MMD). Using La-MMD, the source/target hidden states are pushed to similar distributions to make the feature representation $\mathbf{H}(\mathbf{X})$ transfer feasible.

Based on the hidden vectors from Bi-LSTM, we adopt independent CRF layers for each domain. The rationale lies in the hypotheses that (i) the target domain predictor can better capture target data distribution which could be very unique; (ii) a good predictor trained on the source domain directly could be leveraged to assist the target domain predictor without directly borrowing the source domain training data to bypass the data privacy issue. With respect to the emission and transition score matrices $\sum \mathbf{E}_{i,y_i}$ and $\sum \mathbf{A}_{y_i,y_{i+1}}$, we adopt an upper bound between source/target domains, which helps the target domain predictor to be guided by the source domain predictor. Thus $p(\mathbf{y}|\mathbf{H})$ is also transferable.

There are also other transfer methods, including fine-tuning, sharing parameter directly (without constraints) (He and Sun, 2017; Lee et al., 2017; Yang et al., 2017), etc. However, simply sharing models may dismiss target specific instances.

4.2 Learning Objective

The learning objective is to minimize the following loss $\mathcal{L}$ with respect to parameters $\mathbf{\Theta} = \{\theta_b, \theta_c\}$:

$$\mathcal{L} = \mathcal{L}_c + \alpha\, \mathcal{L}_{\text{La-MMD}} + \beta\, \mathcal{L}_p + \gamma\, \mathcal{L}_r,$$

where $\mathcal{L}_c$ is the CRF loss, $\mathcal{L}_{\text{La-MMD}}$ is the La-MMD loss, $\mathcal{L}_p$ is the parameter similarity loss on CRF layers, and $\mathcal{L}_r$ is the regularization term, with α, β, γ as hyperparameters to balance loss terms.

The CRF loss is our ultimate objective predicting the label sequence given the input sentence, i.e., we minimize the negative log-likelihood of training samples from both source/target domains:

$$\mathcal{L}_c = -\frac{\varepsilon}{N^s}\sum_{i=1}^{N^s}\log p(\mathbf{y}_i^s|\mathbf{H}_i^s) - \frac{1-\varepsilon}{N^t}\sum_{i=1}^{N^t}\log p(\mathbf{y}_i^t|\mathbf{H}_i^t),$$

where $\mathbf{H}$ are hidden vectors obtained from Bi-LSTM, ε is the balance coefficient. The La-MMD loss $\mathcal{L}_{\text{La-MMD}}$ and parameter similarity loss $\mathcal{L}_p$ are discussed in Section 4.3 and 4.4, respectively. The regularization term is to generally control overfitting:

$$\mathcal{L}_r = \|\theta_b\|_2^2 + \|\theta_c\|_2^2.$$

We will provide the model convergence and hyperparameter study in Section 5.1.

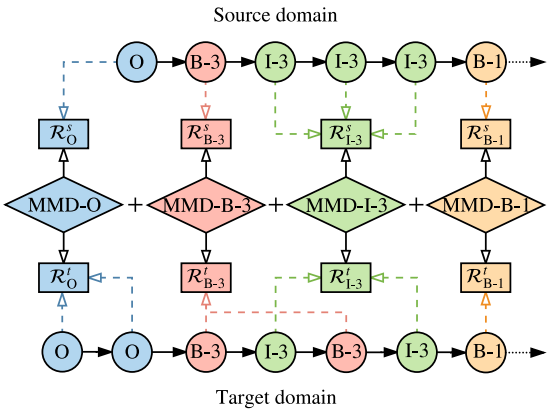

Figure 2: Illustration for La-MMD. MMD-y is computed between two domains' hidden representations with the same ground truth label y. A linear combination is then applied to each label-wise MMD to form La-MMD and the coefficient is set as $\mu_y = 1$.

4.3 Bi-LSTM Feature Representation Transfer

To learn transferable feature representations, the maximum mean discrepancy (MMD) which measures the distance between two distributions, has been widely used in domain adaptation scenarios (Long et al., 2015; Rozantsev et al., 2016). Almost all these works focus on reducing the *marginal* distribution distance between different domain features in an unsupervised manner to make them indistinguishable. However, considering a word is not evenly distributed conditioning on different labels, it may result in that the discriminative property of features from different domains may not be similar, which means that close source and target samples may not have the same label. Different from previous works, we propose label-aware MMD (La-MMD) in Eq. (5) to explicitly reduce the discrepancy between hidden representations with the same label, i.e., the linear combination of the MMD for each label. For each label class $y \in \mathcal{Y}_v$, where $\mathcal{Y}_v$ is the set of matched labels in two domains, we compute the squared population MMD between the hidden representations of source/target samples with the same label y:

$$\text{MMD}^2(\mathcal{R}_y^s, \mathcal{R}_y^t) = \frac{1}{(N_y^s)^2} \sum_{i,j=1}^{N_y^s} k(\mathbf{h}_i^s, \mathbf{h}_j^s) + \frac{1}{(N_y^t)^2} \sum_{i,j=1}^{N_y^t} k(\mathbf{h}_i^t, \mathbf{h}_j^t) - \frac{2}{N_y^s N_y^t} \sum_{i,j=1}^{N_y^s, N_y^t} k(\mathbf{h}_i^s, \mathbf{h}_j^t), \quad (4)$$

where $\mathcal{R}_y^s$ and $\mathcal{R}_y^t$ are sets of hidden representation $\mathbf{h}^s$ and $\mathbf{h}^t$ with corresponding number N_y^s and N_y^t. Eq. (4) can be easily derived by casting Eq. (2) into inner product form and applying $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = k(x, y)$ where k is the reproducing kernel function (Gretton et al., 2012). For each label class, we compute the MMD loss in a normal manner. After that, we define the La-MMD loss as:

$$\mathcal{L}_{\text{La-MMD}} = \sum_{y \in \mathcal{Y}_v} \mu_y \cdot \text{MMD}^2(\mathcal{R}_y^s, \mathcal{R}_y^t), \quad (5)$$

where μ_y is the corresponding coefficient. The illustration of La-MMD is shown in Figure 2.

Once we have applied this La-MMD to our representations learned from Bi-LSTM, the representation distribution of instances with the same label from different domains should be close. Then the standard CRF layer which has a simple linear structure takes these similar representations as input and is likely to give a more transferable label decision for instances with the same label.

4.4 CRF Parameter Transfer

Simply sharing the CRF layer is non-promising when source/target data are diversely distributed. According to probability decomposition in Eq. (3), in order to transfer on source/target CRF layers, more specifically, $p(\mathbf{y}|\mathbf{H})$, we reduce the KL divergence from $p^t(\mathbf{y}|\mathbf{H})$ to $p^s(\mathbf{y}|\mathbf{H})$. But directly reducing $D_{\text{KL}}(p^s(\mathbf{y}|\mathbf{H})||p^t(\mathbf{y}|\mathbf{H}))$ is intractable, we tend to reduce its upper bound:

$$\begin{aligned} & D_{\text{KL}}(p^s(\mathbf{y}|\mathbf{H})||p^t(\mathbf{y}|\mathbf{H})) \\ =& \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H}) \log(\frac{p^s(\mathbf{y}|\mathbf{H})}{p^t(\mathbf{y}|\mathbf{H})}) \\ =& -H(p^s(\mathbf{y}|\mathbf{H})) - \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H}) \log p^t(\mathbf{y}|\mathbf{H}) \\ \leq& c(\|\mathbf{W}^s - \mathbf{W}^t\|_2^2 + \|\mathbf{A}^s - \mathbf{A}^t\|_2^2)^{\frac{1}{2}}, \end{aligned} \quad (6)$$

where $H(\cdot)$ is the entropy of distribution $(\cdot)$ and c is a constant. The detailed proof is provided in Appendix A.1. Since $c(\|\mathbf{W}^s - \mathbf{W}^t\|_2^2 + \|\mathbf{A}^s - \mathbf{A}^t\|_2^2)$ is the upper bound of $D_{\text{KL}}(p^s(\mathbf{y}|\mathbf{H})\|p^t(\mathbf{y}|\mathbf{H}))$,

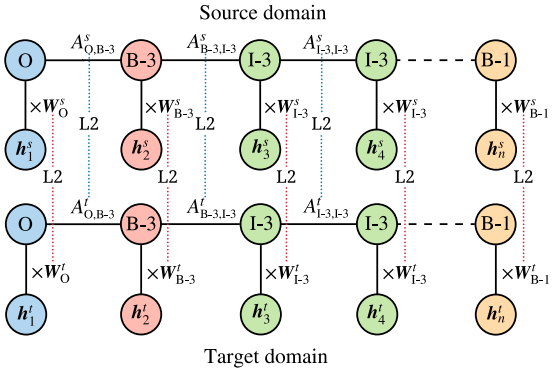

Figure 3: Illustration for CRF parameter transfer.

we conduct CRF parameter transfer by minimizing

$$\mathcal{L}_p = \|\mathbf{W}^s - \mathbf{W}^t\|_2^2 + \|\mathbf{A}^s - \mathbf{A}^t\|_2^2.$$

It turns out that a similar regularization term is applied in our CRF parameter transfer method and the regularization framework (RF) for domain adaptation (Lu et al., 2016). However, RF is proposed to generalize the feature augmentation method in (Daume III, 2007), and these two methods are only discussed from a perspective of the parameter. There is no guarantee that two models having similar parameters yields similar output distributions. In this work, we discuss the model behavior in CRF conditions, and we successfully prove that two CRF models having similar parameters (in Euclidean space) yields similar output distributions. In another word, our method guarantees transferability in the model behavior level, while previous works are limited in parameter level.

The CRF parameter transfer is illustrated in Figure 3, which is also label-aware since the L2 constraint is added over parameters corresponding to the same label in two domains, e.g., $\mathbf{W}_O^s$ and $\mathbf{W}_O^t$.

4.5 Training

We train La-DTL in an end-to-end manner with mini-batch AdaGrad (Duchi et al., 2011). One mini-batch contains training samples from both domains, otherwise the computation of $\mathcal{L}_{\text{La-MMD}}$ can not be performed. During training, word (and character) embeddings are fine-tuned to adjust real data distribution. During both training and decoding (testing) of CRF layers, we use dynamic programming to compute the normalizer in Eq. (1) and infer the label sequence.

Department	# Train	# Dev	# Test
Cardiology	3,004	601	601
Respiratory	3,025	605	606
Neurology	932	187	187
Gastroenterology	1,517	303	304
Sum	8,478	1,696	1,698

Table 1: Sentence numbers for *CM-NER* corpus.

5 Experiments

In this section, we evaluate La-DTL[1] and other baseline methods on 12 cross-specialty NER problems based on real-world datasets. The experimental results show that La-DTL steadily outperforms other baseline models in all tasks significantly. We also conduct further ablation study and robustness study. We evaluate La-DTL on two more non-medical NER transfer tasks to validate its general efficacy over a wide range of applications.

5.1 Cross-Specialty NER

Datasets We collected a Chinese medical NER (*CM-NER*) corpus for our experiments. This corpus contains 1600 de-identified EHRs of our affiliated hospital from four different specialties in four departments: Cardiology (500), Respiratory (500), Neurology (300) and Gastroenterology (300), and the research had been reviewed and approved by the ethics committee. Named entities are annotated in the BIOES format (Begin, Inside, Outside, End and Single), with 30 types in total. The statistics of *CM-NER* is shown in Table 1.

Baselines The following methods are compared. For a fair comparison, we implement La-DTL and baselines with the same base model introduced in (Lample et al., 2016) but with different transfer techniques.

- **Non-transfer** uses the target domain labeled data only.
- **Domain mask** and **Linear projection** belong to the same framework proposed by Peng and Dredze (2017) but have different implementations at the projection layer, which aims to produce shared feature representations among different domains through a linear transformation.
- **Re-training** is proposed by Lee et al. (2017), where an artificial neural networks (ANNs) is first trained on the source domain and then re-trained on the target domain.
- **Joint-training** is a transfer learning method proposed by Yang et al. (2017) where different tasks are trained jointly.
- **CD-learning** is a cross-domain learning method proposed by He and Sun (2017), where each source domain training example's learning rate is re-weighted.

Experimental Settings We use 23,217 unlabeled clinical records to train the word embeddings (word2vec) at 128 dimensions using skip-gram model (Mikolov et al., 2013). The hidden state size is set to be 200 for word-level Bi-LSTM. We evaluate La-DTL for cross-specialty NER with *CM-NER* in 12 transfer tasks, results shown in Table 2. For each task, we take the whole source domain training set $\mathcal{D}_s$ and 10% sentences of the target domain training set $\mathcal{D}_t$ as training data. We use the development set in target domain to search hyper-parameters including training epochs. We then take the models to make the prediction in target domain test set and use F1-score as the evaluation metric. Statistical significance has been determined using a randomization version of the paired sample t-test (Cohen, 1995).

Results and Discussion From the results of 12 cross-specialty NER tasks shown in Table 2, we find that La-DTL outperforms all the strong baselines in all the 12 cross-specialty transfer learning tasks, with 2.62% to 6.70% F1-score lift over state-of-the-art baseline methods. Meanwhile, Linear projection and Domain mask (Peng and Dredze, 2017) do not perform as good as other three baselines, which may be because such linear transformation methods are likely to weaken the representations. While other three baseline methods all share the whole model between source/target domains but differ in the training schemes and performance.

To better understand the transferability of La-DTL, we also evaluate three variants of La-DTL: La-MMD, CRF-L2, and MMD-CRF-L2. La-MMD and CRF-L2 have the same networks and loss function as La-DTL but with different building blocks: La-MMD has $\beta = 0$, while CRF-L2 has $\alpha = 0$. In MMD-CRF-L2, we replace La-MMD loss $\mathcal{L}_{\text{La-MMD}}$ in La-DTL with a vanilla MMD loss:

$$\mathcal{L}_{\text{MMD}} = \text{MMD}^2(\mathcal{R}^s, \mathcal{R}^t),$$

[1] https://github.com/felixwzh/La-DTL

Method	C→R	C→N	C→G	R→C	R→N	R→G	N→C	N→R	N→G	G→C	G→R	G→N	AVG
Non-transfer	67.20	54.51	49.01	65.63	54.51	49.01	65.63	67.20	49.01	65.63	67.20	54.51	59.09
Linear projection (Peng and Dredze, 2017)	69.01	67.02	57.40	69.79	65.87	57.71	67.70	68.77	51.33	68.00	69.65	61.12	64.45
Domain mask (Peng and Dredze, 2017)	70.76	63.97	58.62	70.18	64.27	58.16	67.93	69.89	56.18	68.87	69.89	63.49	65.18
CD-learning (He and Sun, 2017)	71.38	64.01	56.72	72.17	64.91	58.14	68.99	71.13	56.27	70.17	71.76	62.06	65.64
Re-training (Lee et al., 2017)	72.45	70.55	59.58	72.56	68.59	60.94	69.60	70.08	56.58	70.14	71.90	66.01	67.42
Joint-training (Yang et al., 2017)	69.82	70.49	63.52	71.45	67.03	67.71	70.96	71.43	60.54	69.68	71.55	68.15	68.53
La-MMD	73.08	69.48	59.86	72.53	70.28	60.16	71.31	73.04	57.94	69.80	73.99	67.19	68.22
CRF-L2	73.34	71.52	60.17	72.43	69.72	67.61	69.76	71.54	59.96	69.75	71.82	67.30	68.74
MMD-CRF-L2	73.05	72.35	60.80	72.65	69.87	66.82	70.25	71.75	58.98	70.48	73.98	67.43	69.03
La-DTL	**73.59**†	**72.91**†	**64.60**†	**73.88**†	**73.01**†	**70.17**†	**73.08**†	**73.11**†	**62.14**†	**71.61**†	**74.21**†	**71.49**†	**71.15**

Table 2: Results (F1-score %) of 12 cross-specialty medical NER tasks. C, R, N, G are short for the department of Cardiology, Respiratory, Neurology, and Gastroenterology, respectively. † indicates La-DTL outperforms the 6 baselines significantly ($p < 0.05$).

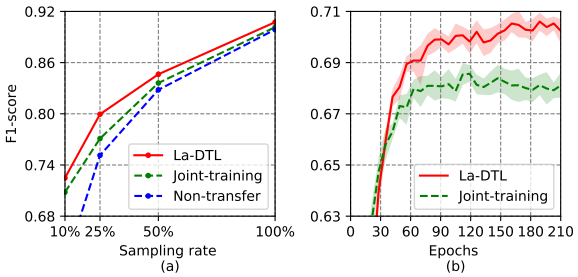

Figure 4: (a) F1-score of La-DTL, Joint-training and Non-transfer method in C→R task with different sampling rate. (b) The learning curve of La-DTL and Joint-training in C→R task.

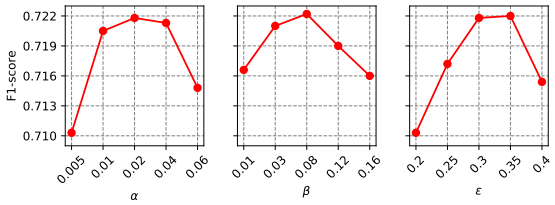

Figure 5: Hyperparameter study for α, β, and ε.

where $\mathcal{R}^s$ and $\mathcal{R}^t$ are sets of hidden representation from source and target domain. Results in Table 2 show that: (i) Using La-MMD alone does achieve satisfactory performance since it outperforms the best baseline Joint-training (Yang et al., 2017) in 7 of 12 tasks. And it has a significant improvement over Domain mask and Linear projection methods (Peng and Dredze, 2017), which indicates that using La-MMD to reduce the domain discrepancy of feature representations in sequence tagging tasks is promising. (ii) CRF-L2 is also a promising method when transferring between NER tasks, and it improves the La-MMD method significantly when these two methods are combined to form La-DTL. (iii) Label-aware characteristic is important in sequence labeling problems because there is an obvious performance drop when La-MMD is replaced with a vanilla MMD in La-DTL. But MMD-CRF-L2 still has very competitive performance compared to all the baseline methods. This shows positive empirical evidence that transferring knowledge at both Bi-LSTM feature representation level and CRF parameter level for NER tasks is better than transferring knowledge at only one of these two levels, as discussed in Section 4.1.

Robustness to Target Domain Data Sparsity We further study the sparsity problem (target domain) of La-DTL in C→R task comparing to Joint-training (Yang et al., 2017) and Non-transfer method. We evaluate La-DTL with different data volume (sampling rate: 10%, 25%, 50%, 100%) on the target domain training set. Results are shown in Figure 4(a). We observe that La-DTL outperforms Joint-training and Non-transfer results under all circumstances, and the improvement of La-DTL is more significant when the sampling rate is lower.

To show La-DTL's convergence and significant improvement over Joint-training, we repeat the 10% sampling rate experiment for 10 times with 10 random seeds. The F1-score on the target domain development set for two methods with a 95% confidence interval is shown in Figure 4(b) where La-DTL outperforms Joint-training method significantly.

Hyperparameter Study We study the influence of three key hyperparameters in La-DTL: α, β, and ε in C→R task with 10% target domain sampling rate. We first apply a rough grid search for the three hyperparameters, and the result is ($\alpha = 0.02, \beta = 0.03, \varepsilon = 0.3$). We then fix two hyperparameters and test the third one in a finer granularity. The results in Figure 5 indicate that setting $\alpha \in [0.01, 0.04]$ could better leverage La-MMD and further setting $\beta \in [0.03, 0.12]$ and $\varepsilon \in [0.3, 0.4]$ yields the best empirical perfor-

Corpus	# Train	# Dev	# Test
SighanNER	23,182	-	4,636
WeiboNER	1,350	270	270
CoNLL 2003	14,987	3,466	3,684
TwitterNER	1,900	240	254

Table 3: Sentence numbers for non-medical corpora.

Method	F1-score
Non-transfer	54.78
Linear projection (Peng and Dredze, 2017)*	56.40
Linear projection (Peng and Dredze, 2017)	56.99
Domain mask (Peng and Dredze, 2017)*	56.80
Domain mask (Peng and Dredze, 2017)	56.32
CD-learning (He and Sun, 2017)*	52.05
CD-learning (He and Sun, 2017)	56.46
Re-training (Lee et al., 2017)	55.36
Joint-training (Yang et al., 2017)	56.80
La-DTL	**57.74**

Table 4: Results (F1-score %) of *WeiboNER* transfer. $*$ indicates the result reported in the corresponding reference.

mance. This shows that we need to balance the learning objective of the source and target domains for better transferability.

5.2 NER Transfer Experiment on Non-medical Corpus

To show La-DTL could be applied in a wide range of NER transfer learning scenarios, we make experiments on two non-medical NER tasks. Corpora's details are shown in Table 3.

***WeiboNER* Transfer** Following He and Sun (2017); Peng and Dredze (2017), we transfer knowledge from *SighanNER* (MSR corpus of the sixth SIGHAN Workshop on Chinese language processing) to *WeiboNER* (a social media NER corpus) (Peng and Dredze, 2015). Results in Table 4 show that La-DTL outperforms all the baseline methods in Chinese social media domain.

***TwitterNER* Transfer** Following Yang et al. (2017) we transfer knowledge from CoNLL 2003 English NER (Tjong Kim Sang and De Meulder, 2003) to *TwitterNER* (Ritter et al., 2011). Since the entity types in these two corpora cannot be exactly matched, La-DTL and Joint-training (Yang et al., 2017) can be applied directly in this case while other baselines can not. Because the CRF parameter transfer of La-DTL is label-aware, and Joint-training simply leverages two independent CRF layers. The results are shown in Table 5, where La-DTL again outperforms Joint-training, indicating that La-DTL could be applied seamlessly to transfer learning scenarios with mismatched label sets and languages like English.

Method	F1-score
Non-transfer	34.65
Joint-training (Yang et al., 2017)*	43.24
La-DTL	**45.71**

Table 5: Results (F1-score %) of *TwitterNER* transfer. $*$ indicates the result reported in the corresponding reference.

6 Conclusions

In this paper, we propose La-DTL, a label-aware double transfer learning framework, to conduct both Bi-LSTM feature representation transfer and CRF parameter transfer with label-aware constraints for cross-specialty medical NER tasks. To our best knowledge, this is the first work on transfer learning for medical NER in cross-specialty scenario. Experiments on 12 cross-specialty NER tasks show that La-DTL provides consistent performance improvement over strong baselines. We further perform a set of experiments on different target domain data size, hyperparameter study and other non-medical NER tasks, where La-DTL shows great robustness and wide efficacy. For future work, we plan to jointly perform NER and entity linking for better cross-specialty media structural information extraction.

Acknowledgments

The work done by SJTU is sponsored by Synyi-SJTU Innovation Program, National Natural Science Foundation of China (61632017, 61702327, 61772333) and Shanghai Sailing Program (17YF1428200).

References

Minmin Chen, Kilian Q Weinberger, and John Blitzer. 2011. Co-training for domain adaptation. In *Advances in Neural Information Processing Systems 24*, pages 2456–2464. Curran Associates, Inc.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1012, Cambridge, MA. Association for Computational Linguistics.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. 2013. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522.

Paul R Cohen. 1995. *Empirical methods for artificial intelligence*, volume 139. MIT press Cambridge, MA.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773.

Ben Hachey, Will Radford, and James R. Curran. 2011. Graph-based named entity linking with wikipedia. In *Proceedings of the 12th International Conference on Web Information System Engineering*, WISE'11, pages 213–226, Berlin, Heidelberg. Springer-Verlag.

Hangfeng He and Xu Sun. 2017. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In *AAAI*, pages 3216–3222.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Vivek Kulkarni, Yashar Mehdad, and Troy Chevalier. 2016. Domain adaptation for named entity recognition in online media with word embeddings. *arXiv preprint arXiv:1612.00148*.

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

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

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017. Transfer learning for named-entity recognition with neural networks. *arXiv preprint arXiv:1705.06273*.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France. PMLR.

Wei Lu, Hai Leong Chieu, and Jonathan Löfgren. 2016. A general regularization framework for domain adaptation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 950–954, Austin, Texas. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Mnica Marrero, Julin Urbano, Sonia Snchez-Cuadrado, Jorge Morato, and Juan Miguel Gmez-Berbs. 2013. Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482 – 489.

Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 591–598, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

V Murthy, Mitesh Khapra, Pushpak Bhattacharyya, et al. 2016. Sharing network parameters for crosslingual named entity recognition. *arXiv preprint arXiv:1607.00198*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 149–155, Berlin, Germany. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2017. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100, Vancouver, Canada. Association for Computational Linguistics.

Claudia Perlich, Brian Dalessandro, Troy Raeder, Ori Stitelman, and Foster Provost. 2014. Machine learning for targeted display advertising: Transfer learning in action. *Mach. Learn.*, 95(1):103–127.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. 2016. Beyond sharing weights for deep domain adaptation. *arXiv preprint arXiv:1603.06432*.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2017. Wasserstein distance guided representation learning for domain adaptation. *arXiv preprint arXiv:1707.01217*.

Nitish Srivastava and Ruslan R Salakhutdinov. 2013. Discriminative transfer learning with tree-based priors. In *Advances in Neural Information Processing Systems 26*, pages 2094–2102. Curran Associates, Inc.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Yonghui Wu, Min Jiang, Jianbo Lei, and Hua Xu. 2015. Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216:624.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *ICLR*.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3320–3328, Cambridge, MA, USA. MIT Press.

Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098.

Weinan Zhang, Ulrich Paquet, and Katja Hofmann. 2016. Collective noise contrastive estimation for policy transfer learning. In *AAAI*, pages 1408–1414.

Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. 2015. Supervised representation learning: Transfer learning with deep autoencoders. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 4119–4125. AAAI Press.

A Appendix

A.1 Detailed Proof

Recall the bound as in Eq. (6):

Lemma A.1. *$c_1(\|\mathbf{W}^s - \mathbf{W}^t\|_2^2 + \|\mathbf{A}^s - \mathbf{A}^t\|_2^2)$ is the upper bound of $(s^s(\mathbf{H}, \mathbf{y}) - s^t(\mathbf{H}, \mathbf{y}))^2$.*

Proof of Lemma A.1. $\otimes$ refers to convolutional product, $\mathbf{H}^W, \mathbf{H}^A$ are mask matrices corresponding to the given hidden vectors $\mathbf{H}$, and c_1 is a constant. We have:

$$
\begin{aligned}
&(s^s(\mathbf{H}, \mathbf{y}) - s^t(\mathbf{H}, \mathbf{y}))^2 \\
=&(\sum_{i=1}^{n} \mathbf{E}^s_{i,y_i} + \sum_{i=1}^{n-1} \mathbf{A}^s_{y_i,y_{i+1}} - \sum_{i=1}^{n} \mathbf{E}^t_{i,y_i} - \sum_{i=1}^{n-1} \mathbf{A}^t_{y_i,y_{i+1}})^2 \\
=&(\mathbf{W}^s \otimes \mathbf{H}^W + \mathbf{A}^s \otimes \mathbf{H}^A - \mathbf{W}^t \otimes \mathbf{H}^W - \mathbf{A}^t \otimes \mathbf{H}^A)^2 \\
=&((\mathbf{W}^s - \mathbf{W}^t) \otimes \mathbf{H}^W + (\mathbf{A}^s - \mathbf{A}^t) \otimes \mathbf{H}^A)^2 \\
\leq&2((\mathbf{W}^s - \mathbf{W}^t) \otimes \mathbf{H}^W)^2 + 2((\mathbf{A}^s - \mathbf{A}^t) \otimes \mathbf{H}^A)^2 \\
=&2(\sum_{i,j} (\mathbf{W}^s - \mathbf{W}^t)_{i,j} \cdot \mathbf{H}^W_{i,j})^2 + 2(\sum_{p,q} (\mathbf{A}^s - \mathbf{A}^t)_{p,q} \cdot \mathbf{H}^A_{p,q})^2 \\
\leq&2(\sum_{i,j} (\mathbf{W}^s - \mathbf{W}^t)^2_{i,j} \cdot \sum_{i,j} (\mathbf{H}^W_{i,j})^2) + 2(\sum_{p,q} (\mathbf{A}^s - \mathbf{A}^t)^2_{p,q} \cdot \sum_{p,q} (\mathbf{H}^A_{p,q})^2) \\
=&2(\|\mathbf{W}^s - \mathbf{W}^t\|_2^2 \cdot \|\mathbf{H}^W\|_2^2) + 2(\|\mathbf{A}^s - \mathbf{A}^t\|_2^2 \cdot \|\mathbf{H}^A\|_2^2) \\
\leq&c_1(\|\mathbf{W}^s - \mathbf{W}^t\|_2^2 + \|\mathbf{A}^s - \mathbf{A}^t\|_2^2).
\end{aligned}
$$

□

Lemma A.2. *$c(\|\mathbf{W}^s - \mathbf{W}^t\|_2^2 + \|\mathbf{A}^s - \mathbf{A}^t\|_2^2)^{\frac{1}{2}}$ is the upper bound of $D_{KL}(p^s(\mathbf{y}|\mathbf{H})||p^t(\mathbf{y}|\mathbf{H}))$.*

Proof of Lemma A.2. With Lemma. (A.1), we set $\varepsilon = (c_1(\|\mathbf{W}^s - \mathbf{W}^t\|_2^2 + \|\mathbf{A}^s - \mathbf{A}^t\|_2^2))^{\frac{1}{2}} \geq 0$ and $c = 2c_1^{\frac{1}{2}}$, and we have:

$$
s^s(\mathbf{H}, \mathbf{y}) - \varepsilon \leq s^t(\mathbf{H}, \mathbf{y}) \leq s^s(\mathbf{H}, \mathbf{y}) + \varepsilon, \tag{7}
$$

$$
\log\{\sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{H})} \exp[s^s(\mathbf{H}, \mathbf{y}')]\} - \varepsilon \leq \log\{\sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{H})} \exp[s^t(\mathbf{H}, \mathbf{y}')]\} \leq \log\{\sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{H})} \exp[s^s(\mathbf{H}, \mathbf{y}')]\} + \varepsilon. \tag{8}
$$

With Eq. (7) and Eq. (8), we can derive

$$
\begin{aligned}
& -\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H})\log p^t(\mathbf{y}|\mathbf{H}) \\
=& -\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H})\log\frac{\exp[s^t(\mathbf{H},\mathbf{y})]}{\sum_{\mathbf{y}'\in\mathcal{Y}(\mathbf{H})}\exp[s^t(\mathbf{H},\mathbf{y}')]} \\
=& -\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H})\big\{s^t(\mathbf{H},\mathbf{y})-\log\{\sum_{\mathbf{y}'\in\mathcal{Y}(\mathbf{H})}\exp[s^t(\mathbf{H},\mathbf{y}')]\}\big\} \\
\leq& -\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H})\big\{s^s(\mathbf{H},\mathbf{y})-\varepsilon-\log\{\sum_{\mathbf{y}'\in\mathcal{Y}(\mathbf{H})}\exp[s^s(\mathbf{H},\mathbf{y}')]\}-\varepsilon\big\} \\
=& -\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H})\big\{\log\frac{\exp[s^s(\mathbf{H},\mathbf{y})]}{\sum_{\mathbf{y}'\in\mathcal{Y}(\mathbf{H})}\exp[s^s(\mathbf{H},\mathbf{y}')]}-2\varepsilon\big\} \\
=& -\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H})\big\{\log p^s(\mathbf{y}|\mathbf{H})-2\varepsilon\big\} \\
=& H(p^s(\mathbf{y}|\mathbf{H}))+2\varepsilon.
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
& D_{\mathrm{KL}}(p^s(\mathbf{y}|\mathbf{H})||p^t(\mathbf{y}|\mathbf{H})) \\
=& \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H})\log(\frac{p^s(\mathbf{y}|\mathbf{H})}{p^t(\mathbf{y}|\mathbf{H})}) \\
=& -H(p^s(\mathbf{y}|\mathbf{H}))-\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H})\log p^t(\mathbf{y}|\mathbf{H}) \\
\leq& -H(p^s(\mathbf{y}|\mathbf{H}))+H(p^s(\mathbf{y}|\mathbf{H}))+2\varepsilon \\
=& c(\|\mathbf{W}^s-\mathbf{W}^t\|_2^2+\|\mathbf{A}^s-\mathbf{A}^t\|_2^2)^{\frac{1}{2}}.
\end{aligned}
$$

□

A.2 Case Analysis

In clinical practice, patients with specific diseases would be assigned to different departments, and specialist doctors in their department may pay more attention to the specific disease. When writing a medical chart, these specific diseases and related clinical findings would have a more detailed description. Therefore, some medical terms would have enriched meanings in different departments accordingly. For example, patients with rheumatic heart disease are often treated in the department of Cardiology. The term, "rheumatic", a modifier, describes and limits the type of "heart disease". In English, "rheumatic" is an adjective modifying "heart disease". However, in Chinese, "rheumatic heart disease" can be regarded as two diseases, "rheumatism" and "heart disease". In the department of Cardiology, "rheumatic heart disease" is usually mentioned as a single term. While in other departments, "rheumatism" and "heart disease" are mostly two independent named entities in annotated datasets. As such, it is difficult to train an NER model to capture the relationship between "rheumatism" and "heart disease", and band them as a whole. In the training set of our study, the diagnostic term "rheumatic heart disease" (including synonym) is mentioned for 17 times in Dept. Cardiology, 16 times in Dept. Respiratory, none in Dept. Neurology and 3 times in Dept. Gastroenterology. We use the data from the first 3 departments as source domain training set respectively, and the data from Dept. Gastroenterology as the target domain training set. We test our models on the test set from Dept. Gastroenterology, where "rheumatic heart disease" is mentioned 3 times, and compare the results across models

Disease	Transfer Task	# disease term in source domain training set	# disease term in target domain training set	# disease term in target domain test set	# accurate labeling without transfer	# accurate labeling with transfer
rheumatic heart disease	C→G N→G R→G	17 0 16	0	3	0	3 0 3
pulmonary heart disease	C→G N→G R→G	4 0 24	0	2	0	2 0 2
coronary atherosclerotic heart disease	G→N C→N R→N	5 136 23	0	15	10	3 15 11

Table 6: Case analysis for cross-specialty medical NER tasks. C, R, N, G are short for department of Cardiology, Respiratory, Neurology, and Gastroenterology, respectively.

with/without transfer learning. As expected, models with source training data from Dept. Cardiovascular and Respiration correctly predict all these entities, but the model using source data from Dept. Neurology fails and so does a model without transfer learning.

Patients with pulmonary heart disease were often referred to Dept. Respiratory and Dept. Cardiology. In our training set, "pulmonary heart disease" (including synonym) is labeled for 24 times in Dept. Respiratory and 4 times in Dept. Cardiology. In English, "pulmonary" modified "heart disease". In Chinese, "pulmonary heart disease" contains body structure "lung" and disease name "heart disease". The model trained with the source set from both from department of respiratory and cardiology could correctly recognize the relation between lung and heart disease and predict the entity in the test set from Dept. Gastroenterology.

Similarly, "coronary atherosclerotic heart disease" contains two disease names, "coronary atherosclerosis" and "heart disease". Training model using source set from a department where the terms are enriched could improve the performance of recognizing the whole entity.

A.3 Medical Experiments Details

The 30 entity types for medical domain are: Symptom, Disease, Examination, Treatment, Laboratory index, Products, Body structure, Frequency, Negative word, Value, Trend, Modification, Temporal word, Noun of locality, Degree modifier, Probability, Object, Organism, Location, Person, Pronoun, Privacy information, Accident, Action, Header, Instrument and material, Non-physiological structure, Dosage, Scale, and Preposition.

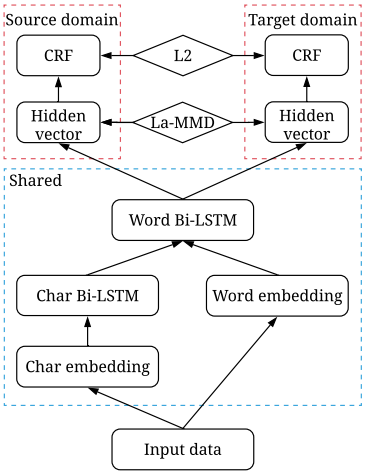

Figure 6: La-DTL framework for language like English.

A.4 Non-medical Experiments Details

WeiboNER Transfer

Both *SighanNER* and *WeiboNER* are annotated in the BIO format (Begin, Inside and Outside), but there is one more entity type (geo-political) in *WeiboNER*. For a fair comparison, we follow Peng and Dredze (2017); He and Sun (2017) to merge geo-political entities and locations in *WeiboNER*, to match different labeling schemes between *WeiboNER* and *SighanNER*. We use the inconsistencies fixed second version of *WeiboNER* data and word embeddings provided by *WeiboNER*'s developers (Peng and Dredze, 2015)[2] in this experiment.

TwitterNER Transfer

To show that La-DTL could be applied in transfer learning for NER scenario with mismatched

[2] https://github.com/hltcoe/golden-horse

named entity types and languages like English, we conduct this experiment transfer from CoNLL 2003 English NER to *TwitterNER*. The four entity types in CoNLL 2003 English NER are LOC, PER, ORG, and MISC. The ten entity types in *TwitterNER* are company, facility, geo-loc, movie, musicartist, other, person, product, sportsteam, and tvshow.

The Joint-training method (Yang et al., 2017) separates the CRF layers for each domain to bypass the label mismatch problem. Since our La-DTL is label-aware, we match four pairs of named entities between two CoNLL 2003 English NER and *TwitterNER*: LOC with geo-loc, PER with person, ORG with company and MISC with other to compute $\mathcal{L}_{\text{La-MMD}}$ and $\mathcal{L}_p$, and leave six named entities unmatched. Following Yang et al. (2017), We leverage char-level Bi-LSTM to generate better word representations, concatenate it with pre-trained word embeddings and feed concatenated embeddings to the word-level Bi-LSTM. The framework used for language like English is illustrated in Figure 6.

We also convert all characters to lowercase and use the same word embeddings provided by Yang et al. (2017)[3]. Also, we concatenate the training set and the development set for both domains and sample the same 10% from *TwitterNER* as (Yang et al., 2017) to be target domain training data. Since Yang et al. (2017) merge training and development set into training data, both Yang et al. (2017) and we report the best performance in the target domain test set.

[3] https://github.com/kimiyoung/transfer