Adaptive Multi-Task Transfer Learning for Chinese Word Segmentation in Medical Text

Anonymous ACL submission

Abstract

Chinese word segmentation (CWS) based on open source corpus faces dramatic performance drop when dealing with domain text, especially for a domain with lots of terms and variant writing style, such as the medical domain. However, building domain-specific CWS requires extremely high annotation cost. In this paper, we propose Adaptive Multi-Task Transfer Learning for CWS by exploiting domain-invariant knowledge from high resource to low resource domains. Experiments on three datasets from medical domain and three open source datasets[1] show that our model achieves persistent higher performance than single-task CWS and several transfer learning baselines, especially when there is a large disparity between source and target domains.

1 Introduction

Chinese word segmentation (CWS) is a fundamental task for Chinese natural language processing (NLP). Most state-of-art methods are based on statistical supervised learning and neural networks. They all rely heavily on human-annotated data, which is a time-consuming and expensive work. Specially, for domain CWS, *e.g.* medical area, the annotation expense is even higher because only domain experts are qualified for the work.

Moreover, CWS tools based on open source datasets, *e.g.* SIGHAN2005[2], face a significance performance drop when dealing with domain text. The ambiguity caused by domain terms and writing style makes it extremely difficult to train a universal CWS tool. As shown in Table 1, given a medical term "高铁血红蛋白血症" (methemoglobinemia), Chinese medical experts tend to annotate it as "高/铁/血红蛋白/血症", which holds the correct definition, an anemia caused by hemoglobin with "high iron" (in Chinese, means iron with valence of three), corresponding to the morphology of "Methemoglobinemia". "PKU" stands for a model trained on PKU's People's Daily corpus, we can see that after segmentation the word "铁血" (jagged) is treated as a word, which is totally wrong semantically. Also, another popular Chinese CWS tool Jieba [3] mistakenly puts the characters "高" and "铁" together, which stands for the high-speed bullet train in China.

CWS tool	高铁血红蛋白血症			
PKU	高 high	铁血 jagged	红蛋白 albumen	血症 anemia
Jieba	高铁 train	血红蛋白 hemoglobin		血症 anemia
Medical	高 high	铁 iron	血红蛋白 hemoglobin	血症 anemia

Table 1: Medical CWS ambiguity with CWS tools. PKU stands for a model trained on PKU dataset. Jieba [4] is another popular CWS tool.

In summary, domain specific CWS task poses significant challenges because:

1. Tools built on open source annotated corpus works bad on domain specific CWS.

2. Domain annotated data is scarce and annotating domain specific data costs expensively.

3. Leaving open source annotated data behind is a waste of resources.

Recently, efforts have been made to exploit open source (high resource) data to improve the performance of domain specific (low resource)

[1] Datasets information is discussed in Sec. 4.1

[2] http://sighan.cs.uchicago.edu/bakeoff2005/

[3] https://github.com/fxsjy/jieba

tasks and decrease the amount of domain annotated data (Yang et al., 2017; Peng and Dredze, 2016; Mou et al., 2016).

In this paper, we further develop multi-task learning (Caruana, 1997; Peng and Dredze, 2016) and propose a novel framework, named *Adaptive Multi-Task Transfer Learning*. Inspired by the success of *Domain Adaptation* (Saenko et al., 2010; Tzeng et al., 2014; Long and Wang, 2015b), we propose to minimize distribution distance of hidden representation between source and target domain, thus make the hidden representations *adapt* to each other and obtain domain-invariant features. Finally, we annotated 3 medical datasets from different medical departments and medical forum, together with 3 open source datasets[1], and do extensive experiments.

The contribution of this paper can be summarized as follows:

- We propose a novel framework, *Adaptive Multi-Task Transfer Learning*, for Chinese word segmentation.
- To the best of our knowledge, we are the first to analyze the performance of transfer learning methods against the amount of disparity between target/source domains.
- Our framework outperforms strong baselines especially when there is substantial *disparity*.
- We open source 3 medical CWS datasets from different sources, which can be used for further study.

2 Single-Task Chinese word segmentation

In this section, we briefly formulate the Chinese word segmentation task and introduce our base model, Bi-LSTM-CRF (Huang et al., 2015).

2.1 Problem Formulation

Chinese word segmentation is often treated as a sequence tagging problem on character level. BIES tagging scheme is broadly accepted by annotators, each character in sentence is labeled as one of $\mathcal{L} = \{B, I, E, S\}$, indicating begin, inside, end of a word, and a word consisting of a single character.

Given a sequence with n characters $X = \{x_1, \ldots, x_n\}$, the aim of the CWS task is to find a mapping from X to $Y^* = \{y_1^*, \ldots, y_n^*\}$:

$$Y^* = \arg\max_{Y \in \mathcal{L}^n} p(Y|X) \tag{1}$$

where $\mathcal{L} = \{B, I, E, S\}$

The general architecture of neural CWS contains: (1) a character embedding layer; (2) an encoder automatically extracts feature and (3) a decoder inferences tag from the feature.

In this paper, we utilize a widely-used model as the base if our framework, which consists of a bi-directional long short-term memory neural network (BiLSTM) as encoder and a conditional random fields (Lafferty et al., 2001) as decoder.

2.2 Encoder

In neural network models, an encoder is usually adopted to automatically extract feature instead of human-crafted feature engineering.

Bi-LSTM LSTM is a popular variant of RNN in order to alleviate the vanishing gradient problem (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997). In addition to considering *past* information from left, Bidirectional LSTM also captures *future* information from the right of the token.

2.3 Decoder

We deploy a conditional random fields layer as decoder. Specifically, $p(Y|X)$ in Eq. (1) could be formulated as

$$p(Y|X) = \frac{\exp(\Phi(X, Y))}{\sum_{Y' \in \mathcal{L}^n} \exp(\Phi(X, Y'))} \tag{2}$$

Here, $\Phi(\cdot)$ is a potential function, consider the situation that we only take the influence between two consecutive variables into account:

$$\Phi(X, Y) = \sum_{j=1}^{n} \phi(X, i, y_i, y_{i-1}) \tag{3}$$

$$\phi(X, i, y_i, y_{i-1}) = s(X, i)_{y_i} + t_{y_i y_{i-1}} \tag{4}$$

where $s(X, i) \in \mathbb{R}^{|\mathcal{L}|}$ is a function that measure the score of the i_{th} character for each label in $\mathcal{L} = \{B, I, E, S\}$, and $t \in \mathbb{R}^{|\mathcal{L}| \times |\mathcal{L}|}$ denotes the transition score between labels. More formally:

$$s(X, i) = \mathbf{W}^\top h_i + \mathbf{b} \tag{5}$$

where h_i is the hidden state of the i^{th} character after BiLSTM; $\mathbf{W} \in \mathbb{R}^{d_h \times |\mathcal{L}|}$ and $\mathbf{b} \in \mathbb{R}^{|\mathcal{L}|}$ are all parameters in the model.

3 Adaptive Multi-Task Transfer Learning

With the motivation to leverage domain-invariant knowledge from high resource domain, we utilize the framework of multi-task learning (Caruana, 1997), which is one of the methods in *transfer learning*, and further introduce three models which are variants of our proposed *Adaptive Multi-Task Transfer Learning*. We exploit three statistical distance measures as the *Adaptive* part to test the generality of our framework.

3.1 Notations and Definitions

In this paper, multi-task learning is defined as a *dual-task* learning, which contains two *Domains* $\mathcal{D}_S$ and $\mathcal{D}_T$. Our purpose is to improve the performance of *target Domain* by exploiting knowledge from *source Domain*.

Each domain $\mathcal{D}$ contains two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where X is a sample sentence, and $X = \{x_1, \dots, x_n\} \in \mathcal{X}$.

Given a single domain, $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a *task* contains two components: a label space $\mathcal{Y}$ and a predictive function $f(\cdot)$, which can be learned during the training phase. Formally, $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$.

3.2 Statistical Distance

In this section, we briefly introduce three statistical distance measures, Kullback–Leibler divergence, Maximum Mean Discrepancy, and Central Moment Discrepancy.

3.2.1 Kullback–Leibler divergence

Kullback–Leibler divergence (KL), is a non-symmetric measure of the divergence between two distributions. For two discrete probability distributions P and Q, the KL divergence from Q to P is defined as:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{6}$$

the Kullback–Leibler divergence between P and Q is defined as:

$$\mathrm{KL}(P, Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{7}$$

3.2.2 Maximum Mean Discrepancy

Proposed by Gretton et al. (2012), maximum mean discrepancy (MMD) is a nonparametric statistical test used to determine if two samples are drawn from different distribution. Given two sets of samples $X = \{x_i, \dots, x_m\}$ and $Y = \{y_1, \dots, y_n\}$, the empirical estimate of MMD is defined as the distance between the empirical mean embedding of each distribution:

$$\mathrm{MMD}^2[\mathcal{F}, p, q] := \left\| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) - \frac{1}{n} \sum_{i=1}^{n} \phi(y_i) \right\|_{\mathcal{H}} \tag{8}$$

where $\mathcal{F}$ is the unit ball in reproducing kernel Hilbert space $\mathcal{H}$.

3.2.3 Central Moment Discrepancy

Proposed by (Zellinger et al., 2017), Central Moment Discrepancy (CMD) is a new distance function on probability distributions on compact intervals. Let X and Y be bounded random samples with respective probability distributions p and q on the interval $[a, b]^N$. The central moment discrepancy CMD_K is defined as an empirical estimate of the CMD metric:

$$\begin{aligned} \mathrm{CMD}_K(X, Y) = & \frac{1}{|b-a|} ||\mathbf{E}(X) - \mathbf{E}(Y)||_2 \\ & + \sum_{k=2}^{K} \frac{1}{|b-a|^k} ||C_k(X) - C_k(Y)||_2 \end{aligned} \tag{9}$$

where $\mathbf{E}(X) = \frac{1}{|X|} \sum_{x \in X} x$ is the empirical expectation vector computed on the sample X and $C_k(X) = \mathbf{E}((x - \mathbf{E}(X))^k)$ is the vector of all k_{th} order sample central moments of the coordinates of X. In experiment, we set K to 5, following (Zellinger et al., 2017).

3.3 Formal Definition

We now give the definition of *Adaptive Multi-Task Transfer Learning*.

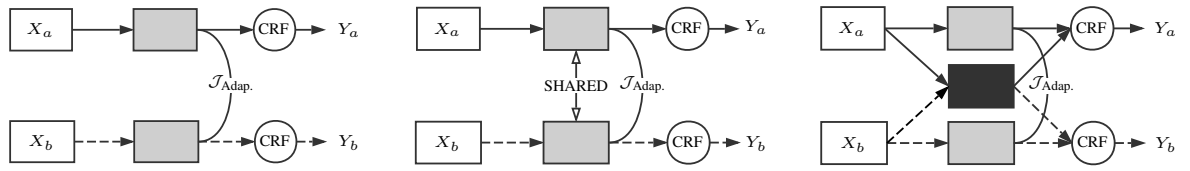
Definition 3.1. Given two domains $\mathcal{D}_S$ and $\mathcal{D}_T$, and corresponding tasks $\mathcal{T}_S$, $\mathcal{T}_T$, *Adaptive Multi-Task Transfer Learning* aims to improve the learning of target predictive function $f_T(\cdot)$ by using *shared parameter* and *minimizing the distance* between $P(X_S)$ and $P(X_S)$, $P(Y_S|X_S)$ and $P(Y_T|X_T)$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$.

3.4 Objective Function

The objective function of our proposed *Adaptive Multi-Task Transfer Learning* can be formulated as follows:

$$\mathcal{J}(\theta^{(a)}, \theta^{(b)}) = \mathcal{J}_{seg} + \alpha \mathcal{J}_{Adap.} + \beta \mathcal{J}_{L_2} \tag{10}$$

where $\theta^{(a)}$ and $\theta^{(b)}$ are model parameters for task a and b, α and β are hyper-parameters to be chosen.



(a) Model-I (b) Model-II (c) Model-III

Figure 1: Three models with different settings. The white block represents Embedding lookup layer, while the gray and black block represents Bi-LSTM layer. The "SHARED" in Figure 1b stands for shared Bi-LSTM for both tasks. The "$\mathcal{J}_{\text{Adap.}}$" represents *Adaptive* loss for the hidden representation after corresponding layer, which is formally discussed in Sec 3.4. The solid arrow and dotted arrow show the flow of task a and task b respectively.

$\mathcal{J}_{seg}$ stands for the negative log likelihood for source domain and target domain. At each training step, we minimize the mean negative log likelihood:

$$\begin{aligned}\mathcal{J}_{seg} = &-\frac{1}{n}\sum_{i=1}^{n}\log p(Y_i^{(a)}|X_i^{(a)}) \\ &-\frac{1}{m}\sum_{i=1}^{m}\log p(Y_i^{(b)}|X_i^{(b)})\end{aligned} \tag{11}$$

$\mathcal{J}_{Adap.}$ is the *Adaptive* loss used to capture domain-invariant knowledge between different domains, which forces the hidden representations between two domains to *adapt* to each other. Given two sets of hidden representation, denoted as $\mathbf{h^{(a)}}$ and $\mathbf{h^{(b)}}$, and a statistic distance function $g(\cdot)$, $\mathcal{J}_{\text{Adap.}}$ can be calculated as:

$$\mathcal{J}_{\text{Adap.}} = g(\mathbf{h^{(a)}}, \mathbf{h^{(b)}}) \tag{12}$$

where $g(\cdot) \in \{\text{KL}(\cdot), \text{MMD}(\cdot), \text{CMD}(\cdot)\}$ in our paper, but not limited in practical use; $\mathbf{h^{(a)}}$ and $\mathbf{h^{(b)}}$ are different for different model setting, which will be defined in Sec 3.5.

$\mathcal{J}_{L_2}$ is the L_2 regularization which is used to control overfitting problem:

$$\mathcal{J}_{L_2} = \left|\left|\theta^{(a)}\right|\right|_2^2 + \left|\left|\theta^{(b)}\right|\right|_2^2 \tag{13}$$

3.5 Models

In this section, we present the design of three variants of our framework in detail. The architectures are represented in Figure 1.

3.5.1 Model-I Specific LSTM

This model can be interrupted as two *parallel tasks* connected with $\mathcal{J}_{\text{Adap.}}$ after specific Bi-LSTM layers of two tasks. We design the model in order to see whether knowledge can actually be transfered through the *Adaptive* loss alone.

The hidden representation and CRF score of *task* t at position i can be computed as:

$$h_i^{(t)} = \text{Bi-LSTM}(X^{(t)}, \theta^{(t)}) \tag{14}$$

$$s(X,i)^{(t)} = \mathbf{W^{(t)}}^{\top} h_i^{(t)} + \mathbf{b}^{(t)} \tag{15}$$

where $h_i^{(t)} \in \mathbb{R}^{2d_h}$, $\mathbf{W^{(t)}} \in \mathbb{R}^{2d_h \times |\mathcal{L}|}$, $\mathbf{b}^{(t)} \in \mathbb{R}^{|\mathcal{L}|}$, $\theta^{(t)}$ denotes parameters of domain specific Bi-LSTM. The $\mathcal{J}_{\text{Adap.}}$ between two tasks, denoted by a and b, is formulated as:

$$\mathcal{J}_{\text{Adap.}} = g(\mathbf{h}^{(a)}, \mathbf{h}^{(b)}) \tag{16}$$

where $\mathbf{h}^{(t)} = \{h_i^{(t)} | X^{(t)} \in \mathcal{X}^{(t)}\}$, $\mathcal{X}^{(t)}$ is a batch of input sequences.

3.5.2 Model-II Shared LSTM

Model-II is designed to adopt domain specific embedding layers, shard Bi-LSTM layer and domain specific CRF layers. Note that traditional *multi-task learning* uses shared embedding (Ruder, 2017).

The hidden representation of *task* t at position i can be computed as:

$$h_i^{(t)} = \text{Bi-LSTM}(X^{(t)}, \theta) \tag{17}$$

where two tasks share Bi-LSTM parameter θ, which is the only difference with Model-I. CRF score and $\mathcal{J}_{\text{Adap.}}$ is the same with (15)(16).

3.5.3 Model-III Shared & Specific LSTM

Model-III is a combination of Model-I and Model-II, with both domain specific and shared Bi-LSTM layers.

The hidden representation and CRF score of *task* t at position i can be computed as:

$$\begin{aligned}h_i^{(t)} &= \text{Bi-LSTM}(X, \theta^{(t)}) \oplus \text{Bi-LSTM}(X, \theta) \\ &= h_{i(specific)}^{(t)} \oplus h_{i(shared)}^{(t)}\end{aligned} \tag{18}$$

$$s(X,i)^{(t)} = \mathbf{W^{(t)}}^{\top} h_i^{(t)} + \mathbf{b}^{(t)} \quad (19)$$

where $h_i^{(t)} \in \mathbb{R}^{4d_h}$, $\mathbf{W^{(t)}} \in \mathbb{R}^{4d_h \times |\mathcal{L}|}$, and $\mathbf{b}^{(t)} \in \mathbb{R}^{|\mathcal{L}|}$. $\theta^{(t)}$ and θ denote the parameter of domain specific and shared Bi-LSTM. $\mathcal{J}_{\text{Adap.}}$ can be calculated as :

$$\mathcal{J}_{\text{Adap.}} = g(\mathbf{h}^{(a)}, \mathbf{h}^{(b)}) \quad (20)$$

where $\mathbf{h}^{(t)} = \{h^{(t)}_{i(specific)} | X^{(t)} \in \mathcal{X}^{(t)}\}$, $\mathcal{X}^{(t)}$ is a batch of input sequences.

4 Experiment

In this section, we evaluate our proposed models on real-world medical Chinese word segmentation tasks, where annotated data is scarce and domain-drift is significant with open source annotated data. We conduct extensive experiments and discuss the result in detail. We also conduct an Ablation test.

4.1 Datasets

Table 2: Statistics of number of sentences for open source corpus.

Type	#Train	#Dev	#Test
PKU	70498	8369	1945
MSR	173850	19453	3985
WEIBO	38086	3834	16673

Table 3: Statistics of number of sentences for medical corpus.

Type	#Train	#Dev	#Test
Cardiology(EMR)	5636	1658	1658
Respiratory(EMR)	5191	1661	1549
Forum	4863	1412	1474
Sum	15690	4731	4691

Open-Source We utilize three open source CWS datasets, respectively are PKU and MSR from SIGHAN2005 Bakeoff[5] and WEIBO from (Qiu et al., 2016). The information of the datasets is shown in Table 2.

Medical We collected three datasets of medical CWS data for our experiment and future research. The first two datasets are electric medical records (EMR) from different departments. The third dataset is medical forum data from *Good Doctor Online*[6], which is a Chinese forum for medical consult. The information of the datasets is shown in Table 3.

[5]http://sighan.cs.uchicago.edu/bakeoff2005/

[6]http://www.haodf.com

4.2 Disparity Study

Transfer Learning aim to improve the performance of low-resource domain task by exploiting the annotated data form high-resource domain, thus the *Disparity* between different tasks is a leading factor to influence the *transferability* between different domains with different methods.

In this paper, we used $\mathcal{X}^2$ test (Kilgarriff and Rose, 1998) to quantify the *Disparity* between three medical corpus. If the size of corpus 1 and corpus 2 are N_1, N_2 and word w has observed frequencies $o_{w,1}$, $o_{w,2}$, then expected value $e_{w,1} = \frac{N_1 \times (o_{w,1}+o_{w,2})}{N_1+N_2}$, and likewise for $e_{w,2}$, then

$$\mathcal{X}^2 = \sum \frac{(o-e)^2}{e} \quad (21)$$

$\mathcal{X}^2$ test shows that *Disparity* between forum dataset and two EMR datasets are close, both far larger than the *Disparity* between two EMR datasets, as shown in Table 4.

Due to the fact that $\mathcal{X}^2$ test doesn't permit comparison between corpus of different sizes (Kilgarriff and Rose, 1998), we propose a simple *agreement* test, using the size of the intersection between the most common n tokens (bi-gram) to quantify the *disparity* between medical corpus and open source corpus. We set n to 500.

Table 4: Result of $\mathcal{X}^2$ test between medical datasets, the larger the higher disparity.

Dataset	Cardiology	Respiratory	Forum
Cardiology	0	0.069	0.126
Respiratory	0.069	0	0.122
Forum	0.126	0.122	0

Table 5: Result of *agreement* test between medical datasets and open source datasets, the smaller the higher disparity.

Dataset	Cardiology	Respiratory	Forum
PKU	25	27	76
MSR	23	25	80
WEIBO	54	50	135

Agreement test shows that the *Disparity* between PKU/MSR and two EMR datasets are close, both far larger than the *Disparity* between PKU/MSR and forum dataset. WEIBO dataset is more similar with medical datasets than PKU and MSR.

Table 6: Performance (F1-score) of Single-task model compared with state-of-art CWS.

Models	Cardiology	Respiratory	Forum
Single-task	81.10	81.33	75.62
(Cai and Zhao, 2016)	80.1	81.5	73.0
(Zhang et al., 2016)	82.46	81.74	77.14

4.3 Training

The training phrase aims to optimize the model parameters $\theta^{(a)}$ and $\theta^{(b)}$ by minimizing the objective function defined in Eq. (10). We use Adam (Kingma and Ba, 2014) with mini-batch. Each batch contains sentences from both domains. The hyper-parameter setting is discussed later.

4.4 Single-task Performance

Before introducing our experiments on proposed *Adaptive Multi-Task Transfer Learning*, we first evaluate the effectiveness of the single-task model (Bi-LSTM-CRF), which is our base model. We compare the model with the two state-of-art on Chinese word segmentation, proposed by Cai and Zhao (2016) and Zhang et al. (2016) respectively. We run experiments on our datasets with their code released on github[7,8]. The results show that the performance of single-task model and state-of-art are close, as shown in Table 6, which indicates the single-task model is a strong baseline for our advanced models.

4.5 Experiment Settings

The dimension of character embedding and the LSTM hidden state dimension are 50. The batch size is 30. We evaluate our *Adaptive Multi-Task Transfer Learning* for totally 15 transfer learning tasks. For each task, we take all of source training data and 10% of target training data. Hyper-parameters are determined by tuning against the development set.

4.6 Baselines

Several baseline methods are compared.

- **Single-task** uses target domain data only, as discussed in Section 2.
- **INIT** loads parameters of model trained on source domain data and then fine-tune the model on target domain data.
- **Multi-Task** shares parameter for both source and target domain, the model is trained simultaneously.

Our implementation of **INIT** follows Mou et al. (2016), and the implementation of **Multi-Task** follows the models we proposed in Sec. 3 by removing $\mathcal{J}_{\text{Adap.}}$, annotating *Model w/o* $\mathcal{J}_{\text{Adap.}}$ in Table 7 and 8.

4.7 Hyper-parameter

In *Adaptive Multi-Task Transfer Learning*, we have two hyper-parameters α and β, which controls the weight of $\mathcal{J}_{\text{Adap.}}$ and $\mathcal{J}_{L_2}$. Our experiments show that $\alpha \in [0.3, 0.7]$ and $\beta \in [0.2, 0.3]$ works best.

4.8 Result and Discussion

Table 7 show the performance of 6 cross medical CWS experiments, Table 8 show the performance of 9 experiments between open source datasets and medical datasets. **Bold** indicates scores that outperforms all baselines. Underline indicates the highest score for each task. We first discuss the result from several general aspects:

(1) All transfer learning methods outperforms strong baseline of single-task method (discussed in Section 4.4). Especially, our models outperforms from 2% to 6% than single-task baseline.

(2) The *Adaptive* part of our model, $\mathcal{J}_{\text{Adap.}}$, is proven to be promising. First, Model-I, which is a parallel training without sharing parameters and leveraging pretrained optimized initialization, outperforms single-task baseline by 4% on average. Second, $\mathcal{J}_{\text{Adap.}}$ improves the performance by 1% on average for both Model-II and Model-III. It shows that the $\mathcal{J}_{\text{Adap.}}$ do capture domain-invariant knowledge apart from the shared parameters.

(3) Within the three models we proposed, Model-II performs best, outperforms other two on 40/45 experiment instances. Model-I and Model-III are equal in match. We argue that it is because the missing of shared parameter of Model-I and the possible noise encoded by the specific layer of Model-III.

(4) For the three statistic distance measures we test in experiment, the overall performance is close. Compared with MMD and CMD, KL gains a more stable improvement on all experiments. However, CMD performs better to hit more best scores than KL and MMD.

[7] https://github.com/jcyk/CWS
[8] https://github.com/SUTDNLP/NNTransitionSegmentor

Table 7: F1-score of 6 cross domain multi-task learning CWS tasks. R, C, F, P stand for *Respiratory*, *Cardiology*, *Forum*, *PKU* respectively. *Model without Adaptive* are Multi-Task Learning with different setting according to our models.

Method	Cross Medical					
	R→C	F→C	C→R	F→R	C→F	R→F
Baselines						
Single-task	81.10	81.10	81.33	81.33	75.62	75.62
INIT	90.62	87.19	88.88	85.56	79.41	78.53
Model-II w/o $\mathcal{J}_{\text{Adap.}}$	86.71	85.27	85.34	83.40	77.62	78.34
Model-III w/o $\mathcal{J}_{\text{Adap.}}$.	84.39	83.59	83.80	83.27	77.18	77.38
Adaptive Multi-Task Transfer Learning-KL						
Model-I	86.94	86.70	85.64	**85.57**	78.35	78.46
Model-II	87.73	87.05	86.65	**86.51**	**79.44**	**78.92**
Model-III	86.66	86.53	85.86	85.39	78.67	**78.72**
Adaptive Multi-Task Transfer Learning-MMD						
Model-I	85.96	85.43	85.45	**85.58**	77.85	78.16
Model-II	87.55	**87.24**	86.27	**86.40**	**79.45**	**78.57**
Model-III	86.30	85.49	85.13	85.19	77.05	77.23
Adaptive Multi-Task Transfer Learning-CMD						
Model-I	86.17	86.03	85.58	**85.83**	78.61	78.39
Model-II	87.49	86.95	86.79	**86.29**	**79.52**	**79.08**
Model-III	86.54	86.36	85.68	**86.05**	78.23	**78.63**

Table 8: F1-score of 9 multi-task learning CWS tasks between open source datasets and medical datasets. R, C, F, P, M, W stand for *Respiratory*, *Cardiology*, *Forum*, *PKU*, *MSR*, *WEIBO* respectively. *Model without Adaptive* are Multi-Task Learning with different setting according to our models.

Method	Open Source - Medical								
	P→C	M→C	W→C	P→R	M→R	W→R	P→F	M→F	W→F
Baselines									
Single-task	81.10	81.10	81.10	81.33	81.33	81.33	75.62	75.62	75.62
INIT	86.20	84.32	87.72	84.05	82.83	86.56	82.54	81.78	84.37
Model-II w/o $\mathcal{J}_{\text{Adap.}}$	85.63	85.84	86.14	84.17	85.42	86.09	78.60	78.80	78.32
Model-III w/o $\mathcal{J}_{\text{Adap.}}$	84.43	86.19	85.61	84.38	85.02	85.79	77.61	77.87	78.38
Adaptive Multi-Task Transfer Learning-KL									
Model-I	**86.30**	**86.60**	86.64	**85.66**	**85.44**	85.69	78.55	78.21	78.11
Model-II	**87.01**	**86.20**	86.94	**85.88**	**85.61**	85.96	78.82	78.69	79.37
Model-III	**86.56**	**86.25**	87.29	**85.30**	**85.60**	85.52	78.20	77.45	78.56
Adaptive Multi-Task Transfer Learning-MMD									
Model-I	85.82	**86.62**	86.47	**85.26**	**85.48**	85.87	77.69	78.26	79.01
Model-II	**86.77**	**86.34**	86.82	**85.98**	**86.17**	85.86	79.04	79.21	78.80
Model-III	85.89	85.68	86.59	**85.05**	85.27	85.64	78.37	78.30	78.39
Adaptive Multi-Task Transfer Learning-CMD									
Model-I	**86.52**	85.93	86.39	**85.71**	85.36	85.97	78.66	78.29	78.49
Model-II	**87.21**	**86.92**	86.83	**85.83**	**85.82**	86.24	78.82	79.01	78.90
Model-III	**86.54**	85.99	86.64	**86.12**	**85.66**	85.63	78.73	78.15	78.71

Next, we analyze the result from a special aspect, the *Disparity* between source and target datasets:

(1) In Table 7, INIT outperforms all other baselines and our approaches in task R → C and C → R, but downperforms our approaches in the others. We argue that the effectiveness of INIT on task between domain R and C result from the low *Disparity* between the two domains. As shown in Table 4.

(2) We first refer to Table 5. We can simply categorize the *Disparity* of 9 combinations into 4 levels. P → C, P → R, M → C and M → R indicate high *Disparity*, W → C, W → R indicate low *Disparity*, P → F, M → F indicate low *similarity*, W → F indicates high *similarity*. Then we can find that, in 4 tasks of high *Disparity*, our approach outperforms all baselines. When *Disparity* goes down to the second level, our approach underperforms INIT but only with gap of 0.4%. However, when *Disparity* continuously goes down to the third and forth level, INIT outperforms our

approach by 3-4%.

4.9 Ablation Test

To investigate the effectiveness of different components in our *Adaptive Multi-Task Transfer Learning* framework, we do ablation test based on Model-II on task (P → R) with $\mathcal{J}_{\text{Adap.}}$ calculated by MMD. Results are reported in Table 4.9. *Model*-II *w/o shared Bi-LSTM* uses domain-specific Bi-LSTM, while *Model*-II *w/o specific embedding* uses shared embedding for both domains.

Results show that the choice of statistic distance measure weights least in the components, since the performance of different measures are close. The test verifies our choice of *shared Bi-LSTM* and *specific embedding*, since their significance is clear.

Table 9: Comparisons of different settings of our method.

Settings	F1-score	δ
Model-II + $\mathcal{J}_{\text{Adap.}}$-MMD	85.98	0
Model-II + $\mathcal{J}_{\text{Adap.}}$-KL	85.88	-0.10
Model-II + $\mathcal{J}_{\text{Adap.}}$-CMD	85.83	-0.15
Model-II w/o $\mathcal{J}_{\text{Adap.}}$	84.17	-1.49
Model-II w/o shared Bi-LSTM	85.26	-0.40
Model-II w/o specific embedding	82.09	-3.57

5 Related Work

Chinese word segmentation CWS is a preliminary step for Chinese natural language processing. It has long been treated as a sequence tagging problem since (Xue et al., 2003). Supervised learning methods are used, including maximum entropy (Low et al., 2005), conditional random fields (Lafferty et al., 2001; Peng et al., 2004; Zhao et al., 2006). These methods depend heavily on hand-crafted features. Recently, neural networks have been for CWS tasks. Zheng et al. (2013) first introduced the neural network architecture to CWS task. Later, different variants of RNN and score functions are developed to improve the performance (Pei et al., 2014; Chen et al., 2015b,a; Cai and Zhao, 2016; Cai et al., 2017). Besides, joint CWS with part-of-speech tagging was proven to improve both tasks (Chen et al., 2016, 2017a). Also, the heterogeneous annotating problem was discussed (Qiu et al., 2013; Chen et al., 2017b).

Transfer Learning Transfer learning distills knowledge from source domain to help target domain achieve a higher performance (Pan and Yang, 2010). In feature-based models, many transfer approached have been studied, including instance transfer (Jiang and Zhai, 2007; Liao et al., 2005), feature representation transfer (Argyriou et al., 2006, 2007), parameter transfer(Lawrence and Platt, 2004; Bonilla et al., 2007) and relation knowledge transfer(Mihalkova et al., 2007; Mihalkova and et al., 2009). However, there's little study on transfer learning for neural networks. (Mou et al., 2016) used intuitive methods (INIT, MULT) to study the transferability of neural networks on NLP applications. Peng and Dredze (2016) proposed to use domain mask and linear projection upon multi-task learning (Long and Wang, 2015a).

6 Conclusion

In this paper, we propose *Adaptive Multi-Task Transfer Learning* framework and three model instances with different settings. 15 experiments between medical datasets and open source datasets show that: (1) *Adaptive Multi-Task Transfer Learning* outperforms multi-task learning all the way; (2) *Adaptive Multi-Task Transfer Learning* outperforms all baselines when the *Disparity* between target and source dataset is high. For future work, we plan to study the transferability between different tasks for Chinese NLP and cross-lingual NLP tasks.

References

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2006. Multi-task feature learning. In Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006. MIT Press, pages 41–48.

Andreas Argyriou, Charles A. Micchelli, Massimiliano Pontil, and Yiming Ying. 2007. A spectral regularization framework for multi-task structure learning. In Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007. Curran Associates, Inc., pages 25–32.

Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. Trans. Neur. Netw. 5(2):157–166.

Edwin V. Bonilla, Kian Ming Adam Chai, and Christopher K. I. Williams. 2007. Multi-task gaussian process prediction. In Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007. Curran Associates, Inc., pages 153–160.

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for chinese. CoRR abs/1606.04300.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for chinese. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers. pages 608–615.

Rich Caruana. 1997. Multitask learning. Machine Learning 28(1):41–75.

Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. A long dependency aware deep architecture for joint chinese word segmentation and POS tagging. CoRR abs/1611.05384.

Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2017a. A feature-enriched neural model for joint chinese word segmentation and part-of-speech tagging. In Carles Sierra, editor, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. ijcai.org, pages 3960–3966.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, and Xuanjing Huang. 2015a. Gated recursive neural network for chinese word segmentation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. The Association for Computer Linguistics, pages 1744–1753.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015b. Long short-term memory neural networks for chinese word segmentation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. pages 1197–1206.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017b. Adversarial multi-criteria learning for chinese word segmentation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. Association for Computational Linguistics, pages 1193–1203.

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.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. CoRR abs/1508.01991.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In John A. Carroll, Antal van den Bosch, and Annie Zaenen, editors, ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic. The Association for Computational Linguistics.

Adam Kilgarriff and Tony Rose. 1998. Measures for corpus similarity and homogeneity. In Nancy Ide and Atro Voutilainen, editors, Proceedings of the Third Conference on Empirical Methods for Natural Language Processing, Palacio de Exposiciones y Congresos, Granada, Spain, June 2, 1998.. ACL, pages 46–52.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. CoRR abs/1412.6980.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data .

Neil D. Lawrence and John C. Platt. 2004. Learning to learn with the informative vector machine. In Carla E. Brodley, editor, Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004. ACM, volume 69 of ACM International Conference Proceeding Series.

Xuejun Liao, Ya Xue, and Lawrence Carin. 2005. Logistic regression with an auxiliary data source. In Luc De Raedt and Stefan Wrobel, editors, Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005. ACM, volume 119 of ACM International Conference Proceeding Series, pages 505–512.

Mingsheng Long and Jianmin Wang. 2015a. Learning multiple tasks with deep relationship networks. CoRR abs/1506.02117.

Mingsheng Long and Jianmin Wang. 2015b. Learning transferable features with deep adaptation networks. CoRR abs/1502.02791.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to chinese word segmentation. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing. volume 1612164, pages 448–455.

Lilyana Mihalkova and et al. 2009. Transfer learning from minimal target data by mapping across relational domains.

Lilyana Mihalkova, Tuyen N. Huynh, and Raymond J. Mooney. 2007. Mapping and revising markov logic networks for transfer learning. In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada. AAAI Press, pages 608–614.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? CoRR abs/1603.06111.

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

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for chinese word segmentation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers. The Association for Computer Linguistics, pages 293–303.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '04.

Nanyun Peng and Mark Dredze. 2016. Multi-task multi-domain representation learning for sequence tagging. CoRR abs/1608.02689.

Xipeng Qiu, Peng Qian, and Zhan Shi. 2016. Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word segmentation for micro-blog texts. In Proceedings of The Fifth Conference on Natural Language Processing and Chinese Computing & The Twenty Fourth International Conference on Computer Processing of Oriental Languages.

Xipeng Qiu, Jiayi Zhao, and Xuanjing Huang. 2013. Joint chinese word segmentation and POS tagging on heterogeneous annotated corpora with multiple task learning. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, pages 658–668.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. CoRR abs/1706.05098.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In Proceedings of the 11th European Conference on Computer Vision: Part IV. Springer-Verlag, Berlin, Heidelberg, ECCV'10, pages 213–226.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. CoRR abs/1412.3474.

Nianwen Xue et al. 2003. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing 8(1):29–48.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. CoRR abs/1703.06345.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (CMD) for domain-invariant representation learning. CoRR abs/1702.08811.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Transition-based neural word segmentation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics.

Hai Zhao, Changning Huang, Mu Li, and Bao-Liang Lu. 2006. Effective tag set selection in chinese word segmentation via conditional random field modeling. In Proceedings of the 20st Pacific Asia Conference on Language, Information and Computation, PACLIC 20, Huazhong Normal University, Wuhan, China, November 1-3, 2006. ACL.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. 2013. Deep learning for chinese word segmentation and POS tagging. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, pages 647–657.