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

Junjie Xing **Kenny Q. Zhu**
Shanghai Jiao Tong University
{jjxing@,kzhu@cs.}sjtu.edu.cn

Shaodian Zhang
Synyi AI Inc.
shaodian@synyi.cn

Abstract

Chinese word segmentation (CWS) trained from open source corpus faces dramatic performance drop when dealing with domain text, especially for a domain with lots of special terms and diverse writing styles, such as the biomedical domain. However, building domain-specific CWS requires extremely high annotation cost. In this paper, we propose an approach by exploiting domain-invariant knowledge from high resource to low resource domains. Extensive experiments show that our model achieves consistently higher accuracy than the single-task CWS and other 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 field , the annotation expense is even higher because only domain experts are qualified for the work.

Moreover, CWS tools trained from open source datasets, e.g., SIGHAN2005[1], 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 would annotate it as "高/铁/血红蛋白/血症", which means anemia caused by hemoglobin with "high iron" (in Chinese, means iron with valence of 3), 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 one word, which is wrong semantically. Also, another popular Chinese CWS tool Jieba [2] mistakenly puts the characters "高" and "铁" together, which stands for the high-speed bullet train in China.

<table>
<tr><th>CWS tool</th><th colspan="4">高铁血红蛋白血症</th></tr>
<tr><td>PKU</td><td>高
high</td><td>铁血
jagged</td><td>红蛋白
albumen</td><td>血症
anemia</td></tr>
<tr><td>Jieba</td><td>高铁
train</td><td colspan="2">血红蛋白
hemoglobin</td><td>血症
anemia</td></tr>
<tr><td>Medical</td><td>高
high</td><td>铁
iron</td><td>血红蛋白
hemoglobin</td><td>血症
anemia</td></tr>
</table>

Table 1: Medical CWS ambiguity with CWS tools. PKU stands for a model trained on PKU dataset.

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

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

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

1. Tools built on open source annotated corpus works badly on domain specific CWS.
2. Annotated domain data is scarce due to high cost.
3. How to leverage open source annotated data despite their generality is an open question.

Recently, efforts have been made to exploit open source (high resource) data to improve the performance of domain specific (low resource) 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 this line of work by developing a multi-task learning (Caruana, 1997; Peng and Dredze, 2016) 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 the 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[??]. The contribution of this paper can be summarized as follows:

- We propose a novel framework for Chinese word segmentation in the medical domain.
- 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 Related Work

2.1 Chinese word segmentation

Statistical Chinese word segmentation has been studied for decades. Xue and others (2003) was the first to treat it as a sequence tagging problem, using a maximum entropy model. Peng et al. (2004) achieved better results by using a conditional random field model (Lafferty et al., 2001). This method has been followed by many other works (Zhao et al., 2006; **?**).

Recently, neural network models have been applied on CWS. These methods use automatically derived features from neural network instead of hand-crafted discrete features. Zheng et al. (2013) first adopted neural network architecture to CWS. Chen et al. (2015b) used Long short-term memory(LSTM) to capture long term dependency. Chen et al. (2015a) proposed a gated recursive neural network (GRNN) to incorporate context information. In this paper, we adopt Bidirectional LSTM-CRF Models (Huang et al., 2015) as our base model.

2.2 Transfer Learning

Transfer learning distills knowledge from source domain and helps target domain to 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; Argyriou et al., 2007), parameter transfer(Lawrence and Platt, 2004; Bonilla et al., 2007) and relation knowledge transfer(Mihalkova et al., 2007; Mihalkova and et al., 2009).

Recently, the transferability of neural networks is also studied. For example, (Mou et al., 2016) studied two methods (INIT, MULT) on NLP applications. Peng and Dredze (2016) proposed to use domain mask and linear projection upon multi-task learning (MTL) (Long and Wang, 2015a). In this paper, we follow MTL and extend the framework with a novel loss function.

3 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).

3.1 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 of our framework, which consists of a bi-directional long short-term memory neural network (BiLSTM) as encoder and conditional random fields (CRF) (Lafferty et al., 2001) as decoder.

3.2 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.

3.3 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.

4 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 under the proposed *Adaptive Multi-Task Transfer Learning* framework (**AMTTL**). We exploit three statistical distance measures as the *Adaptive* part to test the generality of our framework.

4.1 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 $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$.

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, $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$.

4.2 Formal Definition

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

Definition 4.1. 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 $\mathcal{T}_S \neq \mathcal{T}_T$.

4.3 Objective Function

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

$$\mathcal{J}(\theta^{(a)}, \theta^{(b)}) = \mathcal{J}_{seg} + \alpha \mathcal{J}_{Adap.} + \beta \mathcal{J}_{L_2} \tag{6}$$

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

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

$$\begin{aligned} \mathcal{J}_{seg} = &-\frac{1}{n}\sum_{i=1}^{n} \log p(Y_i^{(a)}|X_i^{(a)}) \\ &-\frac{1}{m}\sum_{i=1}^{m} \log p(Y_i^{(b)}|X_i^{(b)}) \end{aligned} \tag{7}$$

$\mathcal{J}_{Adap.}$ 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:

$$\mathcal{J}_{\text{Adap.}} = g(\mathbf{h^{(a)}}, \mathbf{h^{(b)}}) \tag{8}$$

where $g(\cdot)$ can be, but is not limited to, KL divergence, maximum mean discrepancy (MMD) (Gretton et al., 2012) or central moment discrepancy (CMD) (Zellinger et al., 2017); $\mathbf{h^{(a)}}$ and $\mathbf{h^{(b)}}$ are different for different model setting, which will be defined in Sec 4.4.

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

$$\mathcal{J}_{L_2} = \left|\left|\theta^{(a)}\right|\right|_2^2 + \left|\left|\theta^{(b)}\right|\right|_2^2 \tag{9}$$

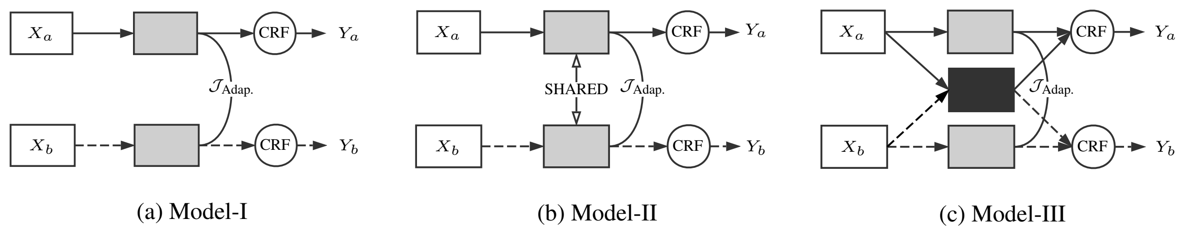
4.4 Models

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

4.4.1 Model-I Specific LSTM

This model can be interpreted 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:



(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 4.3. The solid arrow and dotted arrow show the flow of task a and task b respectively.

$$h_i^{(t)} = \text{Bi-LSTM}(X^{(t)}, \theta^{(t)}) \tag{10}$$

$$s(X,i)^{(t)} = \mathbf{W^{(t)}}^\top h_i^{(t)} + \mathbf{b}^{(t)} \tag{11}$$

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:

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

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

4.4.2 Model-II Shared LSTM

Model-II is designed to adopt domain specific embedding layers, shared Bi-LSTM layer and domain specific CRF layers. Note that traditional *multi-task learning* uses shared embedding (Ruder, 2017). Shared embedding means that source and target domain share the same set of embedding parameters while domain-specific embedding means that the two domains maintain their own sets.

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

$$h_i^{(t)} = \text{Bi-LSTM}(X^{(t)}, \theta) \tag{13}$$

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 as (11)(12).

4.4.3 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:

$$\begin{aligned} h_i^{(t)} &= \text{Bi-LSTM}(X, \theta^{(t)}) \oplus \text{Bi-LSTM}(X, \theta) \\ &= h_{i(specific)}^{(t)} \oplus h_{i(shared)}^{(t)} \end{aligned} \tag{14}$$

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

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 :

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

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

Table 2: Statistics of number of sentences for corpus.

(a) Open Source

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

(b) Medical

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

5 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.

5.1 Datasets

We use three open source CWS datasets, namely PKU and MSR from SIGHAN2005 Bakeoff[3] and WEIBO from (Qiu et al., 2016). The information of the datasets is shown in Table 2a.

We annotated three medical datasets for our experiment and future research. The first two datasets are electronic medical records (EMR) from different departments. The third dataset is medical forum data from *Good Doctor Online*[4], which is a Chinese forum for medical consult. The information of the datasets is shown in Table 2b.

The electronic medical records are collected from our partner hospital, the data only permits non-commercial/academical use. The annotation was done by several doctors. It was carried out following a Chinese word segmentation criteria [5] created by Institute of Computational Linguistics at Peking University. For quality control, annotators were trained until they achieve about 80% inter-annotator agreement on previously annotated materials. Then we conducted double-blind annotation, with resolution of disagreements by a senior annotator. The entire annotation process follows Cohen et al. (2017).

On medicolegal issues, we removed all names (of patients and hospitals), addresses, hospital numbers and EMR IDs in the original data to anonymize and de-identify it.

5.2 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

$$\mathcal{X}^2 = \sum \frac{(o-e)^2}{e} \tag{17}$$

$\mathcal{X}^2$ test shows that *Disparity* between forum dataset and two EMR datasets are similar, but both are much larger than the *Disparity* between the two EMR datasets, as shown in Table 3.

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.

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

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

[5]`http://sighan.cs.uchicago.edu/bakeoff2005/data/pku_spec.pdf`

Table 3: 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 4: 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

Table 5: 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

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.

5.3 Single-task Performance

Before introducing our experiments on proposed framework, 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[6,7]. The results show that the performance of single-task model and state-of-art are close, as shown in Table 5, which indicates the single-task model is a strong baseline for our advanced models.

5.4 Training

The training phrase aims to optimize the model parameters $\theta^{(a)}$ and $\theta^{(b)}$ by minimizing the objective function defined in Eq. (6). We use Adam (Kingma and Ba, 2014) with mini-batch. Each batch contains sentences from both domains. The hyper-parameter setting is discussed later.

5.5 Experiment Settings

The dimension of character embedding and the LSTM hidden state dimension are 50. The batch size is 30. We evaluate our framework for a total of 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.

5.6 Baselines

Several baseline methods are compared.
Single-task uses target domain data only, as discussed in Section 3.
INIT fine-tunes the model trained on source domain using target domain data.
Multi-Task shares parameter for both source and target domain, the model is trained simultaneously.
Linear Projection shares encoder and projects hidden representations into specific feature space.
Domain Mask shares encoder and select different part of hidden representation for source and target domains.

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

Table 6: F1-score of 6 cross domain multi-task learning CWS tasks. R, C, F stand for *Respiratory*, *Cardiology*, *Forum* 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
Linear Projection	85.57	82.95	84.95	84.54	78.25	77.65
Domain Mask	85.01	85.03	85.08	84.74	77.24	78.07
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**	86.17	**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 7: 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
Linear Projection	86.21	86.08	85.35	85.18	84.58	85.27	77.62	77.15	77.54
Domain Mask	85.60	86.17	84.99	84.83	84.16	84.65	77.50	77.46	77.14
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

Our implementation of **INIT** follows Mou et al. (2016), and the implementation of **Multi-Task** follows the models we proposed in Sec. 4 by removing $\mathcal{J}_{\text{Adap.}}$, annotating *Model w/o* $\mathcal{J}_{\text{Adap.}}$ in Table 6 and 7. **Linear Projection** and **Domain Mask** both come from (Peng and Dredze, 2016).

5.7 Hyper-parameter

In our framework, 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.

5.8 Result and Discussion

Table 6 and Table 7 respectively shows the performance of 6 cross medical CWS experiments and 9 experiments between open source datasets and medical datasets. **Bold** indicates scores that outperforms all baselines. Underline indicates the highest score for each task. In general, we learn that

1. All transfer learning methods outperforms strong baseline of single-task method (discussed in Section 5.3). 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 $\mathcal{J}_{\text{Adap.}}$ does capture domain-invariant knowledge apart from the shared parameters.

3. Within the three models we proposed, Model-II performs best, outperforming 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.

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

1. In Table 6, INIT outperforms all other baselines and our approaches in task R $\rightarrow$ C and C $\rightarrow$ 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 3. We speculate that the INIT approach works so well between domains with low disparity because: (a) well trained model in the source domain provides a good start point for training in the target domain, which is very similar to the source; (b) the final model is fine-tuned against the target domain only. Our method is disadvantaged in this scenario because: (a) our model parameters are randomly initialized and are independent between two domains (except for the shared parameters), thus it cannot inherit so much information from the source domain as INIT does; (b) the final model is fine-tuned against both the source and the target domain at the same time; thus noise from the source domain may be introduced into the target domain. This is a research problem we want to tackle in the future.

2. We first refer to Table 4. We can simply categorize the 9 types of domain transfers into 4 levels. P $\rightarrow$ C, P $\rightarrow$ R, M $\rightarrow$ C and M $\rightarrow$ R indicate high disparity, W $\rightarrow$ C, W $\rightarrow$ R indicate low disparity, P $\rightarrow$ F, M $\rightarrow$ F indicate low similarity, W $\rightarrow$ 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%.

At last, we'd like to discuss the effect of transferring from a general-domain dataset (which has the advantage of larger quantity) against that of transferring from a medical dataset (which is better at quality). After comparing the tasks with the same target domain, we conclude that quality weighs more than quantity. Taking Cardiology as an example, the size of source training set used in cross-medical (high quality) tasks is only 1 percent of that used in general-to-medical (high quantity) tasks, but the cross-medical results still outperform the latter.

5.9 Ablation Test

To investigate the effectiveness of different components in our framework, we do ablation test based on Model-II on task (P $\rightarrow$ R) with $\mathcal{J}_{\text{Adap.}}$ calculated by MMD. Results are reported in Table 8. *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, since the performance of different measures are close. The test verifies our choice of *shared Bi-LSTM* and *specific embedding*.

Table 8: 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

6 Conclusion

In this paper, we propose an adaptive multi-task transfer learning framework and three model instances with different settings. 15 experiments between medical datasets and open source datasets show that: *AMTTL*(1) outperforms multi-task learning all the way; (2) 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.

Acknowledgments

This work was partially supported by Synyi-SJTU joint research funding. We would like to thank Kevin Bretonnel Cohen for his suggestions through the COLING Writing Mentoring Program.

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, pages 41–48. MIT Press.

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, pages 25–32. Curran Associates, Inc.

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

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, pages 153–160. Curran Associates, Inc.

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

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

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, pages 1744–1753. The Association for Computer Linguistics.

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.

K Bretonnel Cohen, Karin Verspoor, Karën Fort, Christopher Funk, Michael Bada, Martha Palmer, and Lawrence E Hunter. 2017. The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain. In Handbook of Linguistic Annotation, pages 1379–1394. Springer, Dordrecht, Dordrecht.

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, March.

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

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., pages 46–52. ACL.

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, volume 69 of ACM International Conference Proceeding Series. ACM.

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, volume 119 of ACM International Conference Proceeding Series, pages 505–512. ACM.

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.

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, pages 608–614. AAAI Press.

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, October.

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

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, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

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, ECCV'10, pages 213–226, Berlin, Heidelberg. Springer-Verlag.

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, pages 647–657. ACL.