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 annotating 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 between three datasets from medical domain and SIGHAN2005[1] show that our model achieve persistent higher performance than single-task CWS and several transfer learning baselines, especially when domain-shift is large between source and target datasets.

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, probabilistic graphical models and recently, neural networks. They all relied heavily on human-annotated data, which is not only time-consuming but also expensive. Specially, for specific 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, 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 an 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, the popular universal Chinese CWS tool Jieba [2] mistakenly puts the characters "高" and "铁" together, which stands for the famous G-series high-speed train.

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.

The discussion above shows the dilemma of domain specific CWS task:

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

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

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

2016; Mou et al., 2016). These methods utilize transfer learning methods. For example, Peng and Dredze (2016) proposed a multi-task architecture, treating shared layers as *transferable* between different domains or tasks, considering domain-specific layers as *un-transferable*. However, the assumption that there is no domain-invariant knowledge that can be transferred between domain-specific layers, which in practice is not the case. For instance, for a multi-task training task for CWS and part-of-speech tagging, the task-specific layers can share some knowledge intuitively.

In this paper, we propose *Adaptive Multi-Task Transfer Learning* method for CWS. We analyze the effect of multi-task learning (Caruana, 1997) under three different share/specific settings for CWS. Inspired by the success of using maximum mean discrepancy (MMD) (Gretton et al., 2012) with domain adaptation in computer vision area (Saenko et al., 2010; Tzeng et al., 2014; Long and Wang, 2015b). We propose to use MMD to make domain-specific layers adapted to domain-invariant features, named *Adaptive Multi-Task Transfer Learning*. Finally, we open source 3 medical datasets from different medical departments and medical forum, and do extensive experiments over the datasets.

The contribution of this paper can be summarized as follows:

- To the best of our knowledge, an Adaptive Multi-Task Transfer Learning method is first introduced to CWS.
- We open source 3 medical CWS datasets with 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 or a word consists 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 out a mapping from X to $Y^* = \{y_1^*, \ldots, y_n^*\}$:

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

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

The general architecture of neural CWS often contains: (1) a character embedding layer; (2) an encoder automatically extracts feature and (3) a decoder inferences tag from the feature. The encoder can either be convolution neural network (Chen et al., 2017a) or recurrent neural network (Chen et al., 2015b; Cai et al., 2017).

In this paper, we utilize a basic model 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 methods based on neural network, an encoder is usually adopted to automatically extract feature instead of hand-crafted feature engineering work.

LSTM Recurrent neural networks (RNN) are capable of capturing contextual information over arbitrary length of sequence. However, it is unable to encode the long-term dependency due to the gradient vanishing problem (Bengio et al., 1994). Long short-term memory network is an popular variant, it introduces a memory cell to preserve previous states and gate mechanism to control the updates of hidden states and memory cell (Hochreiter and Schmidhuber, 1997). Mathematically, for a LSTM with parameters θ_a and sequence $X = \{x_1, \ldots, x_n\}$, the LSTM recurrently updates hidden states $h_t = LSTM(x_{t-1}, h_{t-1}, \theta_a)$ at timestep t.

BiLSTM In order to leverage information both side of the sequence, bi-directional LSTM was introduced with both forward and backward directions:

$$\begin{aligned} \overrightarrow{h_t} &= \text{LSTM}(x_t, \overrightarrow{h_{t-1}}, \theta_a) \\ \overleftarrow{h_t} &= \text{LSTM}(\overrightarrow{h_t}, \overleftarrow{h_{t-1}}, \theta_b) \\ h_t &= \overrightarrow{h_t} \oplus \overleftarrow{h_t} \end{aligned} \quad (2)$$

where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are the hidden states at timestep t for the forward and backward LSTM respectively; $\oplus$ is concatenation operation; θ_a and θ_b denotes the parameters of the LSTMs.

2.3 Decoder

After the encoder extracts feature from the sequence, the goal of CWS is to inference a sequence of labels Y from the feature. At this stage, we deploy a conditional random fields layer. Specifically, $p(Y|X)$ in Eq. (1) could be formulated as

$$p(Y|X) = \frac{\exp(\Phi(X, Y))}{\sum_{Y' \in \mathcal{L}^n} \exp(\Phi(X, Y'))} \quad (3)$$

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

$$\Phi(X, Y) = \sum_{j=1}^{n} \phi(X, i, y_i, y_{i-1}) \quad (4)$$

$$\phi(X, i, y_i, y_{i-1}) = s(X, i)_{y_i} + t_{y_i y_{i-1}} \quad (5)$$

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:

$$s(X, i) = \mathbf{W}^\top h_i + \mathbf{b} \quad (6)$$

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

As discussed in Section 1, CWS tools built on open source are not qualified for domain CWS tasks. The main reason is that supervised learning methods make the assumption that training data and test data are drawn from the same feature space, thus the performance drop is intuitively natural when there is a feature space drift.

However, in real world, it is common that we can learn a problem better and faster if we had learned another similar problem. For instance, learning French before may help to learn English. Similarly, learning to play the violin may help to accelerate learning the piano. Motivated by the intuition, people started to study the problem, named *transfer learning*, since a NIPS-95 workshop, focused on machine-learning methods that retain and reuse previously learned knowledge, as reported by Pan and Yang (2010).

In this paper, we utilize the framework of multi-task learning (Caruana, 1997), which is one of the method in the field of *transfer learning*, and further introduce three models which are variants of our proposed *Adaptive Multi-Task Transfer Learning*.

3.1 Notations and Definitions

In this section, we introduce some notations and definitions that are used in later discussion. We define the multi-task learning in the context of our work.

Multi-task learning is defined as a *dual-domain-mono-task* learning in this paper, or more precisely, cross-domain CWS learning. In this paper, the task contains two *Domains* $\mathcal{D}_S$ and $\mathcal{D}_T$, which can be interpreted as *source Domain* and *target Domain* respectively. Our purpose is to improve the performance of *target Domain* by co-training with *source Domain*.

Each domain $\mathcal{D}$ contains two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. For example, if we treat CWS as a sequence tagging problem, then $\mathcal{X}$ is the space of all characters in training data. And X is a sample sentence. If two domains are different, *e.g.* news and medical, they may have different feature space and different marginal probabilistic distribution.

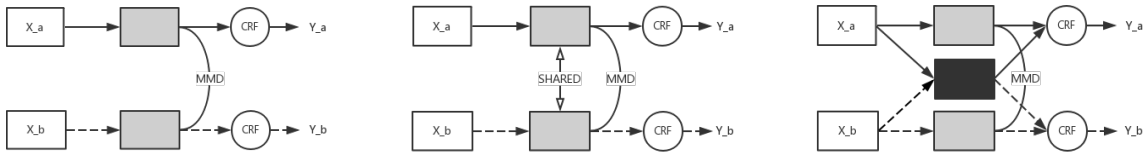
Given a single domain, $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a *task* contains two components: a label space $\mathcal{Y}$ and a predictive function $f(\cdot)$, which can be learned during the training phrase. Formally, $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$.

3.2 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. The test statistic is the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space $\mathcal{H}$ (RKHS). Given two distribution p and q, and a class of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$, MMD is defined as

$$\text{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} (\mathbb{E}_x[f(x)] - \mathbb{E}_y[f(y)]) \quad (7)$$

Consider $\mathcal{F}$ as the unit ball in reproducing kernel Hilbert space $\mathcal{H}$, it is proven that $\text{MMD}[\mathcal{F}, p, q] = 0$ *iff.* $p = q$. Given two sets of samples $X = \{x_i, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, the empirical estimate of MMD is defined as the distance



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

Figure 1: Three models of 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. And the "MMD" represents MMD measure for the hidden representation after corresponding layer. The solid arrow shows the flow of one task and the dotted arrow shows another.

between the empirical mean embedding of each distribution

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

We add MMD into our objective function for *Adaptive Multi-Task Transfer Learning*, which will be discussed in detail in later section.

3.3 Models

To exploit different settings of shared layer, domain-specific layer and maximum mean discrepancy loss, we propose three models for cross-domain CWS task as shown in Figure 1.

3.3.1 Model-I Specific LSTM

This model can be interrupted as two *parallel tasks* connected with MMD after specific Bi-LSTM layers of two tasks. It's not a traditional *multi-task learning* model which usually shares parameters. We design the model in order to see whether knowledge can be transfered through and only through MMD.

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

$$h_i^{(t)} = \text{Bi-LSTM}(X, \theta^{(t)}) \quad (9)$$

$$s(X,i)^{(t)} = \mathbf{W^{(t)}}^{\top} h_i^{(t)} + \mathbf{b}^{(t)} \quad (10)$$

where $h_i^{(t)} \in \mathbb{R}^{2d_h}$, $\mathbf{W^{(t)}} \in \mathbb{R}^{2d_h \times |\mathcal{L}|}$, and $\mathbf{b}^{(t)} \in \mathbb{R}^{|\mathcal{L}|}$. The MMD between two tasks, denoted by a and b, is formulated as:

$$\text{MMD}^2 := \left\| \frac{1}{m}\sum_{i=1}^{m}\phi(h_i^{(a)}) - \frac{1}{n}\sum_{i=1}^{n}\phi(h_i^{(b)}) \right\|_{\mathcal{H}} \quad (11)$$

3.3.2 Model-II Shared LSTM

Model-II are similar with the traditional *multi-task learning*. Unlike traditional ways of sharing low-level layers (Ruder, 2017), we adopt specific embedding layer for each task. The *heterology* of embedding space shouldn't be forced by sharing word embeddings. Then a shard Bi-LSTM layer is used to share semantic between different domains. And domain specific CRF layers are applied.

The formula of this model is the same with Eq. (9)(10)(11).

3.3.3 Model-III Shared & Specific LSTM

Model-III is a combination of Model-I and Model-II. We propose the model on considering that specific Bi-LSTM layers can be used to share knowledge between domains, while shared Bi-LSTM layer is naturally suitable for knowledge to be transferred. Together with MMD, the specific Bi-LSTM layers can also be forced to capture domain-invariant knowledge between two domains.

Similar with Model-I, the hidden representation and CRF score of *task* t at position i can be computed as:

$$h_i^{(t)} = \text{Bi-LSTM}(X, \theta^{(t)}) \oplus \text{Bi-LSTM}(X, \theta) \quad (12)$$

$$s(X,i)^{(t)} = \mathbf{W^{(t)}}^{\top} h_i^{(t)} + \mathbf{b}^{(t)} \quad (13)$$

where $h_i^{(t)} \in \mathbb{R}^{4d_h}$, $\mathbf{W^{(t)}} \in \mathbb{R}^{4d_h \times |\mathcal{L}|}$, and $\mathbf{b}^{(t)} \in \mathbb{R}^{|\mathcal{L}|}$. $\theta^{(t)}$ denotes the parameter of domain specific Bi-LSTM, and θ denotes the parameter of shared Bi-LSTM layer. MMD can be calculated the same as Eq. (11).

3.4 Formal Definition

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

Definition 3.1. Given a source domain $\mathcal{D}_S$, target domain $\mathcal{D}_T$, and corresponding tasks $\mathcal{T}_S$, $\mathcal{T}_T$, *Adaptive Multi-Task Transfer Learning* to help improve the learning of target predictive function $f_T(\cdot)$ by using *shared parameter* and *maximum mean discrepancy loss* to minimize the distance between $P(X_S)$ and $P(X_S)$, $P(Y_S|X_S)$ and $P(Y_T|X_T)$, where $\mathcal{D}_S \neq \mathcal{D}_T$, where $\mathcal{T}_S \neq \mathcal{T}_T$.

3.5 Objective Function

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

$$\mathcal{J}(\theta^{(a)}, \theta^{(b)}) = \mathcal{J}_{seg} + \alpha \mathcal{J}_{\text{MMD}} + \beta \mathcal{J}_{L_2} \quad (14)$$

where α and β are hyper-parameter 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:

$$\begin{aligned} \mathcal{J}_{seg} = &-\frac{1}{n}\sum_{i=1}^{n} \log p(Y_i^{(a)}|X_i^{(a)}) \\ &-\frac{1}{m}\sum_{i=1}^{m} \log p(Y_i^{(b)}|X_i^{(b)}) \end{aligned} \quad (15)$$

$\mathcal{J}_{\text{MMD}}$ is the MMD loss used to capture domain-invariant knowledge between different domains. Given two sets of hidden representation sampled from two distribution, denoted as $\mathbf{h^{(a)}}$ and $\mathbf{h^{(b)}}$, $\mathcal{J}_{\text{MMD}}$ can be calculated as:

$$\begin{aligned} \mathcal{J}_{\text{MMD}}(\mathbf{h^{(a)}}, \mathbf{h^{(b)}}) = &\frac{1}{(N^a)^2}\sum_{i,j=1}^{N^a} k(h_i^a, h_j^a) \\ &+\frac{1}{(N^b)^2}\sum_{i,j=1}^{N^b} k(h_i^b, h_j^b) \\ &-\frac{2}{(N^a N^b)}\sum_{i,j=1}^{N^a, N^b} k(h_i^a, h_j^b) \end{aligned} \quad (16)$$

where N_a and N_b are the corresponding number of $\mathbf{h^{(a)}}$ and $\mathbf{h^{(b)}}$, $k(\cdot)$ is the reproducing kernel function, Eq. (16) is also proven in (Gretton et al., 2012).

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

$$\mathcal{J}_{L_2} = \left|\left|\theta^{(a)}\right|\right|_2^2 + \left|\left|\theta^{(b)}\right|\right|_2^2 \quad (17)$$

Type	#Train	#Dev	#Test
Cardiology(EMR)	?	?	?
Respiratory(EMR)	?	?	?
Forum	?	?	?
Sum	?	?	?

Table 2: Statistics of number of sentences for medical corpus.

Dataset	Single-task	(Cai and Zhao, 2016)
Cardiology	81.10	?
Respiratory	81.33	?
Forum	75.62	?
PKU	95.45	95.5

Table 3: Performance (F1-score) of Single-task model compared with state-of-art CWS.

4 Experiment

In this section, we evaluate our proposed models on real-word domain Chinese word segmentation tasks, especially in the medical domain where annotated data is scarce and domain-drift is significant with open source annotated data, *e.g.* SIGHAN2005, which is discussed in Section 1. We conduct 6 experiments between different medical corpus from different source, *e.g.* Electric Medical Record (EMR) which is totally professional and forum data which contains consumer language. Meanwhile, we conduct experiments between medical domain and news domain (PKU's People's Daily Corpus from SIGHAN2005). These experiments give thorough compare between the three models we proposed in Section 3 and strong baselines.

4.1 Datasets

Open-Source We utilize the open source CWS data provided by PKU, which is a part of SIGHAN2005.

Medical We collected three datasets of medical CWS data for our experiment and future research. The first two datasets are electric medical records from different departments. The third dataset is medical forum data from *Good Doctor Online*[3], which is a Chinese forum for medical consult. The information of the datasets is shown in Table 2.

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

Method	Cross-medical						News-Medical		
	R→C	F→C	C→R	F→R	C→F	R→F	P→C	P→R	P→F
Single-task	81.10	81.10	81.33	81.33	75.62	75.62	81.10	81.33	75.62
INIT	**90.62**	87.19	**88.88**	85.56	**79.81**	78.53	86.20	84.05	**82.54**
Our model without Adaptive									
Model-II w/o Adap.	86.71	85.27	85.34	83.40	77.62	78.34	85.63	84.17	78.60
Model-III w/o Adap.	84.39	83.59	83.80	83.27	77.18	77.38	84.43	84.38	77.61
Adaptive Multi-Task Transfer Learning									
Model-I	85.96	85.43	85.45	85.58	77.85	78.16	85.82	85.26	77.69
Model-II	87.55	**87.24**	86.27	**86.40**	78.31	**78.57**	**86.77**	**85.66**	79.04
Model-III	86.30	85.49	85.13	85.19	77.05	77.23	85.89	85.05	78.37

Table 4: The performance (F1-score) of 9 cross domain multi-task learning CWS tasks. R, C, F, P stand for *Respiratory*, *Cardiology*, *Forum*, *PKU* respectively. *Our model without Adaptive* are Multi-Task Learning with different setting according to our models.

4.2 Training

The training phrase aims to optimize the model parameters $\theta^{(a)}$ and $\theta^{(b)}$ by minimizing the objective function defined in Eq. (14). We use Adam (Kingma and Ba, 2014) with mini-batch. One mini-batch contains sentences from both domains. The hyper-parameter setting is discussed later.

4.3 Single-task Robustness

Before introducing our experiments on proposed *Adaptive Multi-Task Transfer Learning*, we first evaluate the robustness of the single-task model (Bi-LSTM-CRF), which is the base of our proposed models. We compare the model with the state-of-art on Chinese word segmentation, proposed by Cai and Zhao (2016). We run experiments on our datasets with their code released on github[4]. The results show that the performance of single-task model and state-of-art are close, as shown in Table 3.

4.4 Experiment Settings

The dimension of character embedding and the LSTM hidden state dimension are set to be 50. We evaluate our *Adaptive Multi-Task Transfer Learning* for 8 transfer learning tasks. For each task, we take all of source training data and 10% of target training data. Hyper-parameters are determined by development set.

4.5 Baselines

Several strong baselines 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 3 by removing $\mathcal{L}_{\text{MMD}}$, which is the *Adaptive* part in our proposal.

4.6 Hyper-parameter

In *Adaptive Multi-Task Transfer Learning*, we have two hyper-parameters α and β, which controls the weight of $\mathcal{J}_{\text{MMD}}$ and $\mathcal{J}_{L_2}$. Our experiments show that $\alpha \in [0.32, 0.48]$ and $\beta \in [0.2, 0.3]$ works best. The performance of development set according to α and β is shown in Figure TODO.

4.7 Result and Discuss

Table 4 show the performance of our 9 cross-domain CWS experiments with different methods. We'll discuss the result from several aspects:

(1) All transfer learning methods outperforms strong (discussed in Section 4.3) baseline of single-task method. Especially, our models outperforms from 2% to 6% than single-task baseline.

(2) The *Adaptive* part of our model, $\mathcal{J}_{\text{MMD}}$ is proven to be promising. First, Model-I, which is a

[4]https://github.com/jcyk/CWS

parallel training without sharing parameters, outperforms single-task baseline by 4% on average. It even outperforms Model-II w/o Adap on 5/9 experiments. Second, for Model-II and Model-III , the *Adaptive* version outperforms the *w/o Adap.* version on 16/18 experiments, the last two has close performance with gap 0.13% and 0.15% respectively. It shows that The *Adaptive* part adapts domain-invariant knowledge to each domain thus boosting the performance. Third, Model-III underperforms Model-II and outperforms Model-I all the way. We argue that the specific layers which may encode noise into the model account for the former experiment finding, and the use of shared parameters accounts for the latter finding.

(3) The experiment result of Model-II with *Adaption* and INIT is interesting. Model-II outperforms INIT on 5/9 tasks. We find that 5 tasks come from data with larger domain drift.For example, the domain drift between medical forum data and medical EMR data, *e.g.* F $\rightarrow$ C and F $\rightarrow$ R, is larger than the domain drift between medical EMR data, R $\rightarrow$ C and C $\rightarrow$ R. We argue that the INIT model is more likely to achieve better optimal for low domain shift transfer learning because it is trained sequentially and it is much easier for a single task training and provides a better global initialization for target domain. But for large domain shift task, the initialization can be a worse one. Meanwhile, Model-II is trained simultaneously for source and target domain, the model can learn domain-invariant knowledge through shared parameters and MMD. When domain-drift is small, the random initialization of Model-II may account for the loss. However, when domain-drift is large, Model-II performs best.

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 (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 handcrafted features. Recently, neural networks have been for NLP tasks. Zheng et al. (2013) first introduced the neural network architecture to CWS task. Pei et al. (2014) further exploited the interactions between tags and context characters to improve performance. Chen et al. (2015b) adopted the long short-term memory(LSTM) to keep long dependency and avoided the limit of window size of local context. Chen et al. (2015a) proposed to employ a gated recursive neural network to incorporate the complicated combinations of the context characters. Cai and Zhao (2016) employed a factory to produce word representation given governed characters and proposed sentence-level likelihood evaluation system for CWS. (Cai et al., 2017) proposed a greedy neural word segmenter with balanced word and character embedding. 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* method and three models with different settings. Model-I is a novel multi-task learning model which doesn't share parameters between different tasks by exploiting *maximum mean discrepancy* (MMD), to the best of our knowledge. While Model-II/III utilizes MMD as well as sharing parameters. Our 6 experiments between medical text from different source and 3 experiments between medical text and news text show that: (1) *Adaptive Multi-Task Transfer Learning* outperforms multi-task learning all the way; (2) Model-2 of *Adaptive Multi-Task Transfer Learning* outperforms baselines and other set-

tings when the domain shift is large (between medical EMR data and medical forum data or news data), but underperforms INIT method when domain shift is smaller (between medical MER data). We also conduct an Ablation Study on our model. 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 Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *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. http://papers.nips.cc/paper/3143-multi-task-feature-learning.

Andreas Argyriou, Charles A. Micchelli, Massimiliano Pontil, and Yiming Ying. 2007. A spectral regularization framework for multi-task structure learning. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *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. https://doi.org/10.1109/72.279181.

Edwin V. Bonilla, Kian Ming Adam Chai, and Christopher K. I. Williams. 2007. Multi-task gaussian process prediction. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *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. http://arxiv.org/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. https://doi.org/10.18653/v1/P17-2096.

Rich Caruana. 1997. Multitask learning. *Machine Learning* 28(1):41–75. https://doi.org/10.1023/A:1007379606734.

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. http://arxiv.org/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. https://doi.org/10.24963/ijcai.2017/553.

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. http://aclweb.org/anthology/P/P15/P15-1168.pdf.

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. http://aclweb.org/anthology/D/D15/D15-1141.pdf.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017b. Adversarial multi-criteria learning for chinese word segmentation. In Regina Barzilay and Min-Yen Kan, editors, *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. https://doi.org/10.18653/v1/P17-1110.

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. http://dl.acm.org/citation.cfm?id=2188385.2188410.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

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

CoRR abs/1508.01991. http://arxiv.org/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. http://aclweb.org/anthology/P07-1034.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/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*. https://doi.org/10.1145/1015330.1015382.

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. https://doi.org/10.1145/1102351.1102415.

Mingsheng Long and Jianmin Wang. 2015a. Learning multiple tasks with deep relationship networks. *CoRR* abs/1506.02117. http://arxiv.org/abs/1506.02117.

Mingsheng Long and Jianmin Wang. 2015b. Learning transferable features with deep adaptation networks. *CoRR* abs/1502.02791. http://arxiv.org/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. http://www.aaai.org/Library/AAAI/2007/aaai07-096.php.

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. http://arxiv.org/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. https://doi.org/10.1109/TKDE.2009.191.

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. http://aclweb.org/anthology/P/P14/P14-1028.pdf.

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. https://doi.org/10.3115/1220355.1220436.

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

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. http://aclweb.org/anthology/D/D13/D13-1062.pdf.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR* abs/1706.05098. http://arxiv.org/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. http://dl.acm.org/citation.cfm?id=1888089.1888106.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *CoRR* abs/1412.3474. http://arxiv.org/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. http://arxiv.org/abs/1703.06345.

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. http://aclweb.org/anthology/Y/Y06/Y06-1012.pdf.

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. http://aclweb.org/anthology/D/D13/D13-1061.pdf.