Joint Chinese Word Segmentation and POS Tagging on Heterogeneous Annotated Corpora with Multiple Task Learning

Xipeng Qiu, Jiayi Zhao, Xuanjing Huang
Fudan University, 825 Zhangheng Road, Shanghai, China
xpqiu@fudan.edu.cn, zjy.fudan@gmail.com, xjhuang@fudan.edu.cn

Abstract

Chinese word segmentation and part-of-speech tagging (S&T) are fundamental steps for more advanced Chinese language processing tasks. Recently, it has attracted more and more research interests to exploit heterogeneous annotation corpora for Chinese S&T. In this paper, we propose a unified model for Chinese S&T with heterogeneous annotation corpora. We first automatically construct a loose and uncertain mapping between two representative heterogeneous corpora, Penn Chinese Treebank (CTB) and PKU's People's Daily (PPD). Then we regard the Chinese S&T with heterogeneous corpora as two "related" tasks and train our model on two heterogeneous corpora simultaneously. Experiments show that our method can boost the performances of both of the heterogeneous corpora by using the shared information, and achieves significant improvements over the state-of-the-art methods.

1 Introduction

Currently, most of statistical natural language processing (NLP) systems rely heavily on manually annotated resources to train their statistical models. The more of the data scale, the better the performance will be. However, the costs are extremely expensive to build the large scale resources for some NLP tasks. Even worse, the existing resources are often incompatible even for a same task and the annotation guidelines are usually different for different projects, since there are many underlying linguistic theories which explain the same language with different perspectives. As a result, there often exist multiple heterogeneous annotated corpora for a same task with vastly different and incompatible annotation philosophies. These heterogeneous resources are waste on some level if we cannot fully exploit them.

However, though most of statistical NLP methods are not bound to specific annotation standards, almost all of them cannot deal simultaneously with the training data with different and incompatible annotation. The co-existence of heterogeneous annotation data therefore presents a new challenge to utilize these resources.

The problem of incompatible annotation standards is very serious for many tasks in NLP, especially for Chinese word segmentation and part-of-speech (POS) tagging (Chinese S&T). In Chinese S&T, the annotation standards are often incompatible for two main reasons. One is that there is no widely accepted segmentation standard due to the lack of a clear definition of Chinese words. Another is that there are no morphology for Chinese word so that there are many ambiguities to tag the parts-of-speech for Chinese word. For example, the two commonly-used corpora, PKU's People's Daily (PPD) (Yu et al., 2001) and Penn Chinese Treebank (CTB) (Xia, 2000), use very different segmentation and POS tagging standards.

For example, in Table 1, it is very different to annotate the sentence "刘翔进入中国区总决赛 (Liu Xiang reaches the national final in China)" with guidelines of CTB and PDD. PDD breaks some phrases, which are single words in

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 658–668,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

<table>
<tr><th></th><th>Liu</th><th>Xiang</th><th>reachs</th><th colspan="2">China</th><th colspan="2">final</th></tr>
<tr><td>CTB</td><td colspan="2">刘翔/NR</td><td>进入/VV</td><td colspan="2">中国区/NN</td><td colspan="2">总决赛/NN</td></tr>
<tr><td>PDD</td><td>刘/nrf</td><td>翔/nrg</td><td>进入/v</td><td>中国/ns</td><td>区/n</td><td>总/b</td><td>决赛/vn</td></tr>
</table>

Table 1: Incompatible word segmentation and POS tagging standards between CTB and PDD

CTB, into two words. The POS tagsets are also significantly different. For example, PDD gives diverse tags “n” and “vn” for the noun, while CTB just gives “NN”. For proper names, they may be tagged as “nr”, “ns”, etc in PDD, while they are just tagged as “NR” in CTB.

Recently, it has attracted more and more research interests to exploit heterogeneous annotation data for Chinese word segmentation and POS tagging. (Jiang et al., 2009) presented a preliminary study for the annotation adaptation topic. (Sun and Wan, 2012) proposed a structure-based stacking model to fully utilize heterogeneous word structures. They also reported that there is no one-to-one mapping between the heterogeneous word classification and the mapping between heterogeneous tags is very uncertain.

These methods usually have a two-step process. The first step is to train the preliminary taggers on heterogeneous annotations. The second step is to train the final taggers by using the outputs of the preliminary taggers as features. We call these methods as “**pipeline-based**” methods.

In this paper, we propose a method for joint Chinese word segmentation and POS tagging with heterogeneous annotation corpora. We regard the Chinese S&T with heterogeneous corpora as two “related” tasks which can improve the performance of each other. Since it is impossible to establish an exact mapping between two annotations, we first automatically construct a loose and uncertain mapping the heterogeneous tagsets of CTB and PPD. Thus we can tag a sentence in one style with the help of the “related” information in another heterogeneous style. The proposed method can improve the performances of joint Chinese S&T on both corpora by using the shared information of each other, which is proven effective by experiments.

There are three main contributions of our model:

- First, we regard these two joint S&T tasks on different corpora as two related tasks which have interdependent and peer relationship.
- Second, different to the pipeline-based methods, our model can be trained simultaneously on the heterogeneous corpora. Thus, it can also produce two different styles of POS tags.
- Third, our model do not depend on the exactly correct mappings between the two heterogeneous tagsets. The correct mapping relations can be automatically built in training phase.

The rest of the paper is organized as follows: We first introduce the related works in section 2 and describe the background of character-based method for joint Chinese S&T in section 3. Section 4 presents an automatic method to build the loose mapping function. Then we propose our method on heterogeneous corpora in 5 and 6. The experimental results are given in section 7. Finally, we conclude our work in section 8.

2 Related Works

There are some works to exploit heterogeneous annotation data for Chinese S&T.

(Gao et al., 2004) described a transformation-based converter to transfer a certain annotation-style word segmentation result to another style. However, this converter need human designed transformation templates, and is hard to be generalized to POS tagging.

(Jiang et al., 2009) proposed an automatic adaptation method of heterogeneous annotation standards, which depicts a general pipeline to integrate the knowledge of corpora with different

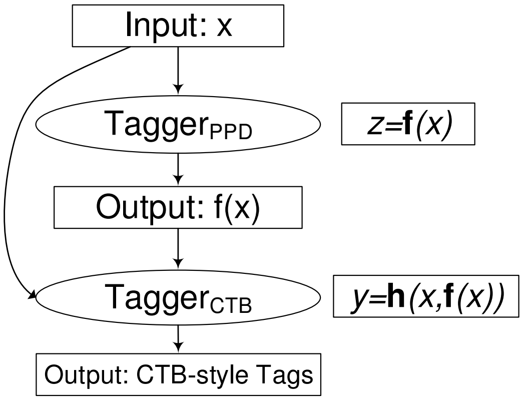

Figure 1: Traditional Pipeline-based Strategy for Heterogeneous POS Tagging

underling annotation guidelines. They further proposed two optimization strategies, iterative training and predict-self re-estimation, to further improve the accuracy of annotation guideline transformation (Jiang et al., 2012).

(Sun and Wan, 2012) proposed a structure-based stacking model to fully utilize heterogeneous word structures.

These methods regard one annotation as the main target and another annotation as the complementary/auxiliary purposes. For example, in their solution, an auxiliary tagger $\mathbf{Tagger_{PPD}}$ is trained on a complementary corpus PPD, to assist the target CTB-style $\mathbf{Tagger_{CTB}}$. To refine the character-based tagger, PPD-style character labels are directly incorporated as new features. The brief sketch of these methods is shown in Figure 1.

The related work in machine learning literature is multiple task learning (Ben-David and Schuller, 2003), which learns a problem together with other related problems at the same time, using a shared representation. This often leads to a better model for the main task, because it allows the learner to use the commonality among the tasks. Multiple task learning has been proven quite successful in practice and has been also applied to NLP (Ando and Zhang, 2005). We also preliminarily verified that multiple task learning can improve the performance on this problem in our previous work (Zhao et al., 2013), which is a simplified case of the work in this paper and has a relative low complexity.

Different with the multiple task learning, whose tasks are actually different labels in the same classification task, our model utilizes the shared information between the real different tasks and can produce the corresponding different styles of outputs.

3 Joint Chinese Word Segmentation and POS Tagging

Currently, the mainstream method of Chinese POS tagging is joint segmentation & tagging with character-based sequence labeling models(Lafferty et al., 2001), which can avoid the problem of segmentation error propagation and achieve higher performance on both subtasks(Ng and Low, 2004; Jiang et al., 2008; Sun, 2011; Qiu et al., 2012).

The label of each character is the cross-product of a segmentation label and a tagging label. If we employ the commonly used label set {B, I, E, S} for the segmentation part of cross-labels ({B, I, E} represent *Begin*, *Inside*, *End* of a multi-node segmentation respectively, and S represents a *Single* node segmentation), the label of character can be in the form of {B-T}(T represents POS tag). For example, *B-NN* indicates that the character is the begin of a noun.

4 Automatically Establishing the Loose Mapping Function for the Labels of Characters

To combine two human-annotated corpora, the relationship of their guidelines should be found. A **mapping function** should be established to represent the relationship between two different annotation guidelines. However, the exact mapping relations are hard to establish. As reported in (Sun and Wan, 2012), there is no one-to-one mapping between their heterogeneous word classification, and the mapping between heterogeneous tags is very uncertain.

Fortunately, there is a loose mapping can be found in CTB annotation guideline[1] (Xia, 2000). Table 2 shows some

[1] Available at http://www.cis.upenn.edu/ ~chi-

	CTB’s Tag	PDD’ Tag[1]
Total tags	33	26
verbal noun	NN	v[+nom]
proper noun	NR	n
是 (shi4)	VC	v
有 (you3)	VE, VV	v
conjunctions	CC, CS	c
other verb	VV, VA	v, a, z
number	CD, OD	m

[1] The tag set of PDD just includes the 26 broad categories in the mapping table. The whole tag set of PDD has 103 sub categories.

Table 2: Examples of mapping between CTB and PDD’s tagset

mapping relations in CTB annotation guideline. These loose mapping relations are many-to-many mapping. For example, the mapping may be “NN/CTB↔{n,nt,nz}/PDD”, “NR/CTB↔{nr,ns}/PDD”, “v/PDD↔{VV, VA}/CTB” and so on.

We define $\mathcal{T}_1$ and $\mathcal{T}_2$ as the tag sets for two different annotations, and $t_1 \in \mathcal{T}_1$ and $t_2 \in \mathcal{T}_2$ are the corresponding tags in two tag sets respectively.

We first establish a loose mapping function $\mathbf{m} : \mathcal{T}_1 \times \mathcal{T}_2 \rightarrow \{0, 1\}$ between the tags of CTB and PDD.

$$\mathbf{m}(t_1, t_2) = \begin{cases} 1 & \text{if } t_1 \text{ and } t_2 \text{ have mapping relation} \\ 0 & \text{else} \end{cases} \tag{1}$$

The mapping relations are automatically build from the CTB guideline (Xia, 2000). Due to the fact that the tag set of PPD used in the CTB guideline is just broad categories, we expand the mapping relations to include the sub categories. If a PPD’s tag is involved in the mapping, all its sub categories should be involved. For example, for the mapping “NR/CTB↔nr/PDD”, the relation of *NR* and *nrf/nrg* should be added in the mapping relations too (*nrf/nrg* belong to *nr*).

Since we use the character-based joint S&T model, we also need to find the mapping function between the labels of characters.

In this paper, we employ the commonly used label set {B, I, E, S} for the segmentation part of cross-labels and the label of character can be in the form of {B-T}(T represents POS tag). Thus, each mapping relation $t_1 \leftrightarrow t_2$ can be automatically transformed to four forms: B-t_1 ↔B-t_2, I-t_1 ↔I-t_2, E-t_1 ↔E-t_2 and S-t_1 ↔S-t_2. (“B-NR/CTB↔{B-nr,B-ns}/PPD” for example).

Beside the above transformation, we also give a slight modification to adapt the different segmentation guidelines. For instance, the person name “莫言 (Mo Yan)” is tagged as “B-NR, E-NR” in CTB but “S-nrf, S-nrg” in PPD. So, some special mappings may need to be added like “B-NR/CTB↔S-nrf/PPD”, “E-NR/CTB↔{S-nrg, E-nrg}/PPD”, “M-NR/CTB↔{B-nrg, M-nrg}/PPD” and so on. Although these special mappings are also established automatically with an exhaustive solution. In fact, we give segmentation alignment only to proper names due to the limitation of computing ability.

Thus, we can easily build the **loose bidirectional mapping function $\tilde{\mathbf{m}}$** for the labels of characters. An illustration of our construction flowchart is shown in Figure 2.

Finally, total 524 mappings relationships are established.

5 Joint Chinese S&T with Heterogeneous Data with Multiple Task Learning

Inspired by the multiple task learning (Ben-David and Schuller, 2003), we can regard the joint Chinese S&T with heterogeneous data as two “related” tasks, which can improve the performance of each other simultaneously with shared information.

5.1 Sequence Labeling Model

We first introduce the commonly used sequence labeling model in character-based joint Chinese S&T.

Sequence labeling is the task of assigning labels $\mathbf{y} = y_1, \ldots, y_n (y_i \in \mathcal{Y})$ to an input sequence $\mathbf{x} = x_1, \ldots, x_n$. $\mathcal{Y}$ is the set of labels.

nese/posguide.3rd.ch.pdf

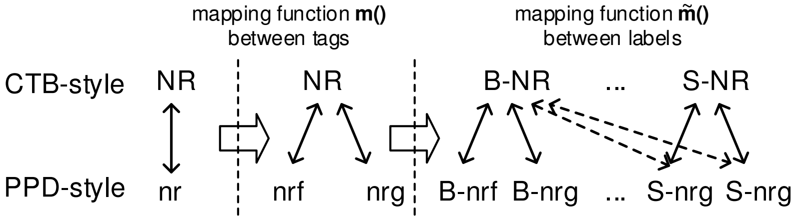

Figure 2: An Illustration of Automatically Establishing the Loose Mapping Function

Given a sample $\mathbf{x}$, we define the feature $\Phi(\mathbf{x}, \mathbf{y})$. Thus, we can label $\mathbf{x}$ with a score function,

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} S(\mathbf{w}, \Phi(\mathbf{x}, \mathbf{y})), \tag{2}$$

where $\mathbf{w}$ is the parameter of score function $S(\cdot)$. The feature vector $\Phi(\mathbf{x}, \mathbf{y})$ consists of lots of overlapping features, which is the chief benefit of discriminative model. Different algorithms vary in the definition of $S(\cdot)$ and the corresponding objective function. $S(\cdot)$ is usually defined as linear or exponential family function.

For first-order sequence labeling, the feature can be denoted as $\phi_k(\mathbf{x}, y_{i-1:i})$, where i stands for the position in the sequence and k stands for the number of feature templates. For the linear classifier, the score function can be rewritten in detail as

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \sum_{i=1}^{L} \left(\langle \mathbf{u}, \mathbf{f}(\mathbf{x}, y_i) \rangle + \langle \mathbf{v}, \mathbf{g}(\mathbf{x}, y_{i-1:i}) \rangle \right), \tag{3}$$

where $y_{i:j}$ denotes label subsequence $y_i y_{i+1} \cdots y_j$; $\mathbf{f}$ and $\mathbf{g}$ denote the state and transition feature vectors respectively, $\mathbf{u}$ and $\mathbf{v}$ are their corresponding weight vectors; L is the length of $\mathbf{x}$.

5.2 The Proposed Model

Different to the single task learning, the heterogeneous data have two sets of labels $\mathcal{Y}$ and $\mathcal{Z}$.

The heterogeneous datasets $\mathbb{D}_s$ and $\mathbb{D}_s$ consist of $\{\mathbf{x}_i, \mathbf{y}_i\}(i = 0, \cdots, m)$ and $\{\mathbf{x}_i, \mathbf{z}_i\}(i = 0, \cdots, n)$ respectively.

For a sequence $\mathbf{x} = x_1, \ldots, x_L$ with length L. , there may have two output sequence labels $\mathbf{y} = y_1, \ldots, y_L$ and $\mathbf{z} = z_1, \ldots, z_L$, where $y_i \in \mathcal{Y}$ and $z_i \in \mathcal{Z}$.

We rewrite the loose mapping function $\tilde{\mathbf{m}}$ between two label sets into the following forms,

$$\varphi(y) = \{z | \tilde{\mathbf{m}}(y, z) = 1\}, \tag{4}$$

$$\varphi(z) = \{y | \tilde{\mathbf{m}}(y, z) = 1\}, \tag{5}$$

where $\varphi(z) \subset \mathcal{Y}$ and $\varphi(y) \subset \mathcal{Z}$ are the subsets of $\mathcal{Y}$ and $\mathcal{Z}$. Give a label y(or z) in an annotation, the loose mapping function φ returns the corresponding mapping label set in another heterogeneous annotation.

Our model for heterogeneous sequence labeling can be write as

$$\begin{aligned} \hat{\mathbf{y}} = \arg\max_{\mathbf{y}, y_i \in \mathcal{Y}} \sum_{i=1}^{L} \Big(& \langle \mathbf{u}, \mathbf{f}(\mathbf{x}, y_i) \rangle \\ & + \langle \mathbf{s}, \sum_{z \in \varphi(y_i)} \mathbf{h}(\mathbf{x}, z) \rangle \\ & + \langle \mathbf{v}_1, \mathbf{g}_1(\mathbf{x}, y_{i-1:i}) \rangle \\ & + \langle \mathbf{v}_2, \sum_{\substack{z_{i-1} \in \varphi(y_{i-1}) \\ z_i \in \varphi(y_i)}} \mathbf{g}_2(\mathbf{x}, z_{i-1:i}) \rangle \Big), \end{aligned} \tag{6}$$

and

$$\begin{aligned} \hat{\mathbf{z}} = \arg\max_{\mathbf{z}, z_i \in \mathcal{Z}} \sum_{i=1}^{L} \Big(& \langle \mathbf{u}, \sum_{y \in \varphi(z_i)} \mathbf{f}(\mathbf{x}, y) \rangle + \\ & \langle \mathbf{s}, \mathbf{h}(\mathbf{x}, z_i) \rangle \\ & + \langle \mathbf{v}_1, \sum_{\substack{y_{i-1} \in \varphi(z_{i-1}) \\ y_i \in \varphi(z_i)}} \mathbf{g}_1(\mathbf{x}, y_{i-1:i}) \rangle \\ & + \langle \mathbf{v}_2, \mathbf{g}_2(\mathbf{x}, z_{i-1:i}) \rangle \Big), \end{aligned} \tag{7}$$

where $\mathbf{f}$ and $\mathbf{h}$ represent the state feature vectors on two label sets $\mathcal{Y}$ and $\mathcal{Z}$ respectively.

In Eq.(6) and (7), the score of the label of every character is decided by the weights of the corresponding mapping labels and itself.

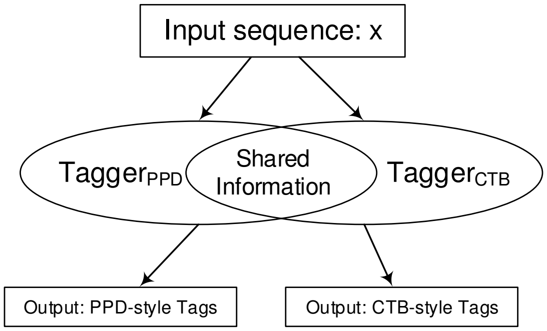

Figure 3: Our model for Heterogeneous POS Tagging

The main challenge of our model is the efficiency of decoding algorithm, which is similar to structured learning with latent variables(Liang et al., 2006) (Yu and Joachims, 2009). Most methods for structured learning with latent variables have not expand all possible mappings. In this paper, we also only expand the mapping that with highest according to the current model.

Our model is shown in Figure 3 and the flowchart is shown in Algorithm 1. If given the output type of label T, we only consider the labels in T to initialize the Viterbi matrix, and the score of each node is determined by all the involved heterogeneous labels according to the loose mapping function.

```
input  : character sequence $x_{1:L}$
         loose mapping function $\varphi$
         output type: $T (T \in \{T_y, T_z\})$
output: label sequence ls
if $T == T_y$ then
    calculate ls using Eq. (6);
else if $T == T_z$ then
    calculate ls using Eq. (7) ;
else
    return null;
end
return ls
```

Algorithm 1: Flowchart of the Tagging process of the proposed model

6 Training

We use online Passive-Aggressive (PA) algorithm (Crammer and Singer, 2003; Crammer et al., 2006) to train the model parameters. Following (Collins, 2002), the average strategy is used to avoid the overfitting problem.

For the sake of simplicity, we merge the Eq.(6) and (7) into a unified formula.

Given a sequence $\mathbf{x}$ and the expect type of tags T, the merged model is

$$\hat{\mathbf{y}} = \arg\max_{\substack{\mathbf{y}\\ t(\mathbf{y})=T}} \langle \mathbf{w}, \sum_{\mathbf{z}\in\psi(\mathbf{y})} \Phi(\mathbf{x},\mathbf{z})\rangle, \tag{8}$$

where $t(\mathbf{y})$ is a function to judge the type of output tags; $\psi(\mathbf{y})$ represents the set $\{\varphi(y_1) \otimes \varphi(y_2) \otimes \cdots \otimes \varphi(y_L)\} \cup \{\mathbf{y}\}$, where $\otimes$ means Cartesian product; $\mathbf{w} = (\mathbf{u}^T, \mathbf{s}^T, \mathbf{v}_1^T, \mathbf{v}_2^T)^T$ and $\Phi = (\mathbf{f}^T, \mathbf{h}^T, \mathbf{g}_1^T, \mathbf{g}_2^T)^T$.

We redefine the score function as

$$S(\mathbf{w},\mathbf{x},\mathbf{y}) = \langle \mathbf{w}, \sum_{\mathbf{z}\in\psi(\mathbf{y})} \Phi(\mathbf{x},\mathbf{z})\rangle. \tag{9}$$

Thus, we rewrite the model into a unified formula

$$\hat{\mathbf{y}} = \arg\max_{\substack{\mathbf{y}\\ t(\mathbf{y})=T}} S(\mathbf{w},\mathbf{x},\mathbf{y}). \tag{10}$$

Given an example $(\mathbf{x},\mathbf{y})$, $\hat{\mathbf{y}}$ is denoted as the incorrect label sequence with the highest score

$$\hat{\mathbf{y}} = \arg\max_{\substack{\bar{\mathbf{y}}\neq\mathbf{y}\\ t(\bar{\mathbf{y}})=t(\mathbf{y})}} S(\mathbf{w},\mathbf{x},\bar{\mathbf{y}}). \tag{11}$$

The **margin** $\gamma(\mathbf{w}; (\mathbf{x},\mathbf{y}))$ is defined as

$$\gamma(\mathbf{w}; (\mathbf{x},\mathbf{y})) = S(\mathbf{w},\mathbf{x},\mathbf{y}) - S(\mathbf{w},\mathbf{x},\hat{\mathbf{y}}). \tag{12}$$

Thus, we calculate the **hinge loss** $\ell(\mathbf{w}; (\mathbf{x},\mathbf{y}))$, (abbreviated as ℓ_w) by

$$\ell_w = \begin{cases} 0, & \gamma(\mathbf{w}; (\mathbf{x},\mathbf{y})) > 1 \\ 1 - \gamma(\mathbf{w}; (\mathbf{x},\mathbf{y})), & \text{otherwise} \end{cases} \tag{13}$$

In round k, the new weight vector $\mathbf{w}_{k+1}$ is calculated by

$$\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w} - \mathbf{w}_k||^2 + \mathcal{C} \cdot \xi,$$
$$\textbf{s.t. } \ell(\mathbf{w}; (\mathbf{x}_k, \mathbf{y}_k)) <= \xi \textbf{ and } \xi >= 0 \tag{14}$$

where ξ is a non-negative slack variable, and $\mathcal{C}$ is a positive parameter which controls the influence of the slack term on the objective function.

Following the derivation in PA (Crammer et al., 2006), we can get the update rule,

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \tau_k \mathbf{e}_k, \tag{15}$$

where

$$\begin{aligned} \mathbf{e}_k &= \sum_{\mathbf{z} \in \psi(\mathbf{y}_k)} \Phi(\mathbf{x}_k, \mathbf{z}) - \sum_{\mathbf{z} \in \psi(\hat{\mathbf{y}}_k)} \Phi(\mathbf{x}_k, \mathbf{z}), \\ \tau_k &= \min(\mathcal{C}, \frac{\ell_{w_k}}{\|\mathbf{e}_k\|^2}). \end{aligned}$$

As we can see from the Eq. (15), when we update the weight vector, the update information includes not only the features extracted from current input, but also that extracted from the loose mapping sequence of input. For each feature, the weights of its corresponding related features derived from the loose mapping function will be updated with the same magnitude as well as itself.

Our method regards two annotations to be interdependence and peer relationship. Therefore, the two heterogeneous annotated corpora can be simultaneously used as the input of our training algorithm. Because of the tagging and training algorithm, the weights and tags of two corpora can be used separately with the only dependent part built by the loose mapping function.

Our training algorithm based on PA is shown in Algorithm 2.

```
input  : mixed heterogeneous datasets:
         $(\mathbf{x}_i, \mathbf{y}_i), i = 1, \cdots, N$;
         parameters: $\mathcal{C}, K$;
         loose mapping function: $\varphi$ ;
output: $\mathbf{w}_K$
Initialize: $\mathbf{wTemp} \leftarrow 0, \mathbf{w} \leftarrow 0$;
for $k = 0 \cdots K-1$ do
    for $i = 1 \cdots N$ do
        receive an example $(\mathbf{x}_i, \mathbf{y}_i)$;
        predict: $\hat{\mathbf{y}}_i$ with Eq.(11);
        if hinge loss $\ell_w > 0$ then
            update $\mathbf{w}$ with Eq. (15);
        end
    end
    $\mathbf{wTemp} = \mathbf{wTemp} + \mathbf{w}$ ;
end
$\mathbf{w}_K = \mathbf{wTemp}/K$ ;
```

Algorithm 2: Training Algorithm

6.1 Analysis

Although our mapping function between two heterogeneous annotations is loose and uncertain, our online training method can automatically increase the relative weights of features from the beneficial mapping relations and decrease the relative weights of features from the unprofitable mapping relations.

Consider an illustrative loose mapping relation "NN/CTB↔n,nt,nz/PDD". For an input sequence **x** and PDD-style output is expected. If the algorithm tagging a character as "n/PDD"(with help of the weight of "NN/CTB") and the right tag isn't one of "n,nt,nz/PDD", the weight of "NN/CTB" will also be decreased, which is reasonable since it is beneficial to distinguish the right tag. And if the right tag is one of "n,nt,nz/PDD" but not "n/PDD" (for example, "nt/PDD"), which means it is a "NN/CTB", the weight of "NN/CTB" will remain unchanged according to the algorithm (updating "n/PDD" changes the "NN/CTB", but updating "nt/PDD" changes it back).

Therefore, after multiple iterations, useful features derived from the mapping function are typically receive more updates, which take relatively more responsibility for correct prediction. The final model has good parameter estimates for the shared information.

We implement our method based on FudanNLP(Qiu et al., 2013).

7 Experiments

7.1 Datasets

We use the two representative corpora mentioned above, Penn Chinese Treebank (CTB) and PKU' s People' s Daily (PPD) in our experiments.

Dataset	Partition	Sections	Words
CTB-5	Training	1−270 400−931 1001−1151	0.47M
	Develop	301−325	6.66K
	Test	271−300	7.82K
CTB-S	Training		0.64M
	Test	-	59.96K
PPD	Training	-	1.11M
	Test	-	0.16M

Table 3: Data partitioning for CTB and PD

7.1.1 CTB Dataset

To better comparison with the previous works, we use two commonly used criterions to partition CTB dataset into the train and test sets.

- One is the partition criterion used in (Jin and Chen, 2008; Jiang et al., 2009; Sun and Wan, 2012) for CTB 5.0.

- Another is the CTB dataset from the POS tagging task of the Fourth International Chinese Language Processing Bakeoff (SIGHAN Bakeoff 2008)(Jin and Chen, 2008).

7.1.2 PPD Dataset

For the PPD dataset, we use the PKU dataset from SIGHAN Bakeoff 2008.

The details of all datasets are shown in Table 3. Our experiment on these datasets may lead to a fair comparison of our system and the related works.

7.2 Setting

We conduct two experiments on **CTB-5 + PPD** and **CTB-S + PPD** respectively.

The form of feature templates we used is shown in Table 7.2, where C represents a Chinese character, and T represents the character-based tag. The subscript i indicates its position related to the current character.

Our method can be easily combined with some other complicated models, but we only use the simple one for the purpose of observing the sole influence of our unified model. The parameter C is tested on develop dataset, and we found that it just impact the speed of convergence and have no effect on the accuracy. Moreover, since we use the averaged strategy, we wish more iterations to avoid overfitting and set a small value 0.01 to it. The maximum number of iterations K is 50.

$C_i, T_0 (i = -2, -1, 0, 1, 2)$
$C_i, C_{i+1}, T_0 (i = -1, 0)$
T_{-1}, T_0

Table 4: Feature Templates

The $F1$ score is used for evaluation, which is the harmonic mean of precision P (percentage of predict phrases that exactly match the reference phrases) and recall R (percentage of reference phrases that returned by system).

7.3 Evaluation on CTB-5 + PPD

The experiment results on the heterogeneous corpora CTB-5 + PPD are shown in Table 5. Our method obtains an error reductions of 24.08% and 90.8% over the baseline on CTB-5 and PDD respectively.

Our method also gives better performance than the pipeline-based methods on heterogeneous corpora, such as (Jiang et al., 2009) and (Sun and Wan, 2012).

The reason is that our model can utilize the information of both corpora effectively, which can boost the performance of each other.

Although the loose mapping function are bidirectional between two annotation tagsets, we may also use unidirectional mapping. Therefore, we also evaluate the performance when we use unidirectional mapping. We just use the mapping function $\psi_{\mathbf{PDD} \to \mathbf{CTB}}$, which means we obtain the PDD-style output without the information from CTB in tagging stage. Thus, in training stage, there are no updates for the weights of CTB-features for the instances from PDD corpus, while instances from CTB corpus can result to updates for PDD-features.

Surprisedly, we find that the one-way mapping can also improve the performances of both corpora. The results are shown in Table 7. The

Method	Training Dataset	Test Dataset	P	R	F1
(Jiang et al., 2009)	CTB-5, PDD	CTB-5	-	-	94.02
(Sun and Wan, 2012)	CTB-5, PDD	CTB-5	94.42	94.93	94.68
Our Model	CTB-5	CTB-5	93.28	93.35	93.31
Our Model	PDD	PDD	89.41	88.58	88.99
Our Model	CTB-5, PDD	CTB-5	**94.74**	**95.11**	**94.92**
Our Model	CTB-5, PDD	PDD	**90.25**	**89.73**	**89.99**

Table 5: Performances of different systems on CTB-5 and PPD.

Method	Training Dataset	Test Dataset	P	R	F1
Our Model	CTB-S	CTB-S	89.11	89.16	89.13
Our Model	PDD	PDD	89.41	88.58	88.99
Our Model	CTB-S, PDD	CTB-S	**89.86**	**90.02**	**89.94**
Our Model	CTB-S, PDD	PDD	**90.5**	**89.82**	**90.16**

Table 6: Performances of different systems on CTB-S and PPD.

model$_{PPD\to CTB}$ obtains an error reductions of 14.63% and 6.12% over the baseline on CTB-5 and PDD respectively.

Method	P	R	F1
Model$_S$ on CTB-5	93.86	94.73	94.29
Model$_S$ on PDD	90.05	89.28	89.66

"Model$_S$" is the model which is trained on both CTB-5 and PDD training datasets with just just using the unidirectional mapping function $\psi_{\mathbf{PDD\to CTB}}$.

Table 7: Performances of unidirectional PPD→CTB mapping on CTB-5 and PPD.

7.4 Evaluation on CTB-S + PPD

Table 6 shows the experiment results on the heterogeneous corpora CTB-S + PPD. Our method obtains an error reductions of 7.41% and 10.59% over the baseline on CTB-S and PDD respectively.

7.5 Analysis

As we can see from the above experiments, our proposed unified model can improve the performances of the two heterogeneous corpora with unidirectional or bidirectional loose mapping functions. Different to the pipeline-based methods, our model can use the shared information between two heterogeneous POS taggers. Although the mapping function is loose and uncertain, it is still can boost the performances. The features derived from the wrong mapping function take relatively less responsibility for prediction after multiple updates of their weights in training stage. The final model has good parameter estimates for the shared information.

Another phenomenon is that the performance of one corpus can gains when the data size of another corpus increases. In our two experiments, the training set's size of CTB-S is larger than CTB-5, so the performance of PDD is higher in latter experiment.

8 Conclusion

We proposed a method for joint Chinese word segmentation and POS tagging with heterogeneous annotation data. Different to the previous pipeline-based works, our model is learned on heterogeneous annotation data simultaneously. Our method also does not require the exact corresponding relation between the standards of heterogeneous annotations. The experimental results show our method leads to a significant improvement with heterogeneous annotations over the best performance for this task. Although our work is for a specific task on joint Chinese word segmentation and POS, the key idea to leverage heterogeneous annotations is very general and applicable to other NLP tasks.

In the future, we will continue to refine the proposed model in two ways: (1) We wish to use the unsupervised method to extract the loose mapping relation between the different annotation standards, which is useful to the corpora without loose mapping guideline. (2) We will analyze the shared information (weights of the features derived from the tags which have the mapping relation) in detail and propose a more effective model. Besides, we would also like to investigate for other NLP tasks which have different annotation-style corpora.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work was funded by NSFC (No.61003091), Key Projects in the National Science & Technology Pillar Program (2012BAH18B01), Shanghai Municipal Science and Technology Commission (12511504500), Shanghai Leading Academic Discipline Project (B114) and 973 Program (No.2010CB327900).

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853, December.

S. Ben-David and R. Schuller. 2003. Exploiting task relatedness for multiple task learning. *Learning Theory and Kernel Machines*, pages 567–580.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

K. Crammer and Y. Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

J. Gao, A. Wu, M. Li, C.N. Huang, H. Li, X. Xia, and H. Qin. 2004. Adaptive chinese word segmentation. In *Proceedings of ACL-2004*.

W. Jiang, L. Huang, Q. Liu, and Y. Lu. 2008. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. In *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Citeseer.

W. Jiang, L. Huang, and Q. Liu. 2009. Automatic adaptation of annotation standards: Chinese word segmentation and POS tagging: a case study. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 522–530.

Wenbin Jiang, Fandong Meng, Qun Liu, and Yajuan Lü. 2012. Iterative annotation transformation with predict-self reestimation for Chinese word segmentation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 412–420, Jeju Island, Korea, July. Association for Computational Linguistics.

C. Jin and X. Chen. 2008. The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese pos tagging. In *Sixth SIGHAN Workshop on Chinese Language Processing*, page 69.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 761–768. Association for Computational Linguistics.

H.T. Ng and J.K. Low. 2004. Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based. In *Proceedings of EMNLP*, volume 4.

Xipeng Qiu, Feng Ji, Jiayi Zhao, and Xuanjing Huang. 2012. Joint segmentation and tagging with coupled sequences labeling. In *Proceedings of COLING 2012*, pages 951–964, Mumbai, India, December. The COLING 2012 Organizing Committee.

Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. FudanNLP: A toolkit for Chinese natural language processing. In *Proceedings of ACL*.

Weiwei Sun and Xiaojun Wan. 2012. Reducing approximation and estimation errors for Chinese lexical processing with heterogeneous annotations. In *Proceedings of the 50th Annual Meeting of the*

Association for Computational Linguistics, pages 232–241.

W. Sun. 2011. A stacked sub-word model for joint Chinese word segmentation and part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1385–1394.

F. Xia, 2000. *The part-of-speech tagging guidelines for the penn Chinese treebank (3.0)*.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1169–1176. ACM.

S. Yu, J. Lu, X. Zhu, H. Duan, S. Kang, H. Sun, H. Wang, Q. Zhao, and W. Zhan. 2001. Processing norms of modern Chinese corpus. Technical report, Technical report.

Jiayi Zhao, Xipeng Qiu, and Xuanjing Huang. 2013. A unified model for joint chinese word segmentation and pos tagging with heterogeneous annotation corpora. In *International Conference on Asian Language Processing, IALP*.