Joint Segmentation and Tagging with Coupled Sequences Labeling

Xipeng Qiu[1], *Feng Ji*[1,2], *Jiayi Zhao*[1] *and Xuanjing Huang*[1]
(1) School of Computer Science, Fudan University
(2) Suntec Software (Shanghai) Co. Ltd.
{xpqiu,fengji,11210240073,xjhuang}@fudan.edu.cn

Abstract
Segmentation and tagging task is the fundamental problem in natural language processing (NLP). Traditional methods solve this problem in either pipeline or joint cross-label ways, which suffer from error propagation and large number of labels respectively. In this paper, we present a novel joint model for segmentation and tagging, which integrates two dependent Markov chains. One chain is used for segmentation, and the other is for tagging. The model parameters can be estimated simultaneously. Besides, we can optimize the whole model by improving the single chain. The experiments show that our model could achieve higher performance over traditional models on both English shallow parsing and Chinese word segmentation and POS tagging tasks.

Title and Abstract in Chinese

基于双链序列标注的联合切分和标注模型

在自然语言处理中，序列标注模型是最常见的模型，也有着广泛地应用。针对常见的可分解为分段和标注两个子任务的复杂序列标注问题，我们提出了双链序列标注模型。该模型中存在着两条相互联系的马尔科夫链。为此我们提出了一个同时求解这两条链上最优序列的解码算法。同时利用这两条链，针对不同的实际应用场景可以组合出不同的标注模型，使用不同的解码算法完成实际的标注任务。为了能够适应不同的解码算法，我们还提出了一个能够利用异构语料训练模型的参数学习算法。在多个语料上的实验表明，我们提出的模型性能要优于其他模型，并能在同一个模型内完成多种标注任务。

Keywords: Coupled Sequences Labeling, Segmentation, Tagging.

Keywords in Chinese: 双链序列标注, 切分, 标注.

Proceedings of COLING 2012: Posters, pages 951–964,
COLING 2012, Mumbai, December 2012.

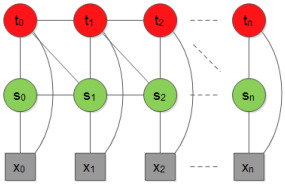


Figure 1: Coupled Sequences Labeling Model (双链序列标注模型)

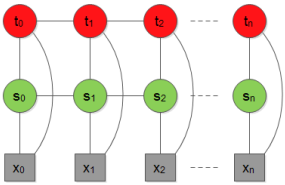
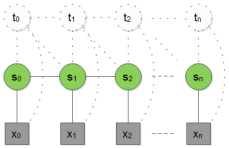
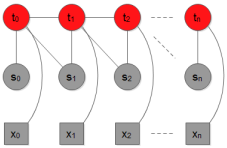


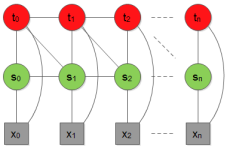
Figure 2: Factorial CRF Model (FCRF 模型)



(a) segment by only using the linear chain **s**



(b) label a segmented sentence



(c) simultaneously segment and label a sentence

Figure 3: Coupled Sequences Labeling Model for Different Tasks (处理不同任务时的双链序列标注模型变换), where gray nodes are observed nodes.

Table 1: Feature templates for shallow parsing (浅层句法分析特征模板)

Joint Cross-Product Model	Coupled Sequence Labeling Model
$w_{i-2}y_i, w_{i-1}y_i, w_iy_i, w_{i+1}y_i, w_{i+2}y_i$	$w_{i-1}s_i, w_is_i, w_{i+1}s_i$
	$w_{i-2}t_i, w_{i-1}t_i, w_it_i, w_{i+1}t_i, w_{i+2}t_i$
$w_{i-1}w_iy_i, w_iw_{i+1}y_i$	$w_{i-1}w_is_i, w_iw_{i+1}s_i$
	$w_{i-1}w_it_i, w_iw_{i+1}t_i$
$p_{i-2}y_i, p_{i-1}y_i, p_iy_i, p_{i+1}y_i, p_{i+2}y_i$	$p_{i-1}s_i, p_is_i, w_{i+1}s_i$
	$p_{i-2}t_i, p_{i-1}t_i, p_it_i, p_{i+1}t_i, p_{i+2}t_i$
$p_{i-2}p_{i-1}y_i, p_{i-1}p_iy_i, p_ip_{i+1}y_i, p_{i+1}p_{i+2}y_i$	$p_{i-2}p_{i-1}s_i, p_{i-1}p_is_i, p_ip_{i+1}s_i, p_{i+1}p_{i+2}s_i$
	$p_{i-3}p_{i-2}t_i, p_{i-2}p_{i-1}t_i, p_{i-1}p_it_i, p_ip_{i+1}t_i,$ $p_{i+1}p_{i+2}t_i, p_{i+2}p_{i+3}t_i, p_{i-1}p_{i+1}t_i$
$p_{i-2}p_{i-1}p_iy_i, p_{i-1}p_ip_{i+1}y_i, p_ip_{i+1}p_{i+2}y_i$	$p_{i-2}p_{i-1}p_is_i, p_{i-1}p_ip_{i+1}s_i, p_ip_{i+1}p_{i+2}s_i$
$y_{i-1}y_i$	$w_is_it_i$
	$w_is_{i-1}s_i$
	$w_{i-1}t_{i-1}t_i, w_it_{i-1}t_i, p_{i-1}t_{i-1}t_i, p_it_{i-1}t_i$
	$s_{i-1}t_{i-1}s_i, t_{i-1}s_it_i$

Table 2: Feature templates for Chinese S&T (中文分词、词性标注特征模板)

Joint Cross-Label Model	Coupled Sequence Labeling Model
$c_{i-2}y_i, c_{i-1}y_i, c_iy_i, c_{i+1}y_i, c_{i+2}y_i$	$c_{i-2}s_i, c_{i-1}s_i, c_is_i, c_{i+1}s_i, c_{i+2}s_i$
	$c_{i-3}t_i, c_{i-2}t_i, c_{i-1}t_i, c_it_i, c_{i+1}t_i, c_{i+2}t_i, c_{i+3}t_i$
$c_{i-1}c_iy_i, c_ic_{i+1}y_i, c_{i-1}c_{i+1}y_i$	$c_{i-1}c_is_i, c_ic_{i+1}s_i, c_{i-1}c_{i+1}s_i$
	$c_{i-3}c_{i-2}t_i, c_{i-2}c_{i-1}t_i, c_{i-1}c_it_i, c_ic_{i+1}t_i,$ $c_{i+1}c_{i+2}t_i, c_{i+2}c_{i+3}t_i, c_{i-2}c_it_i, c_ic_{i+2}t_i$
$y_{i-1}y_i$	$c_is_it_i$
	$c_{i-1}t_{i-1}t_i,\ s_{i-1}s_i$
	$s_{i-1}t_{i-1}s_i, t_{i-1}s_it_i$

input : Taggging training dataset: $(\mathbf{x}_i, \mathbf{s}_i, \mathbf{t}_i), i = 1, \cdots, M, ;$
input : Segmentation training dataset (optional): $(\mathbf{x}_i, \mathbf{s}_i), i = M+1, \cdots, M+N, ;$
input : Parameters: $\mathcal{C}, K$.
output: $\mathbf{w}$

Initialize: $\mathbf{cw} \leftarrow 0, \mathbf{w} \leftarrow 0$;
for $k = 0 \cdots K-1$ **do**
  random select an integer number $l \in (1, \ldots, M+N)$ with no repeat;
  if $l \leq M$ **then**
    receive an example $(\mathbf{x}_l, \mathbf{s}_l, \mathbf{t}_l)$;
    predict (2nd Viterbi): $(\hat{\mathbf{s}}_l, \hat{\mathbf{t}}_l) = \arg\max_{\mathbf{s},\mathbf{t}} \langle \mathbf{w}, \Phi^{st}(\mathbf{x}_l, \mathbf{s}, \mathbf{t}) \rangle$;
    if $(\hat{\mathbf{s}}_l, \hat{\mathbf{t}}_l) \neq (\mathbf{s}_l, \mathbf{t}_l)$ **then**
      update with $\mathbf{w}$ with Eq. 10, where $(\cdot)$ is $(\mathbf{x}_l, \mathbf{s}_l, \mathbf{t}_l)$ and $(*)$ is $(\mathbf{x}_l, \hat{\mathbf{s}}_l, \hat{\mathbf{t}}_l)$;
    end
  else
    receive an example $(\mathbf{x}_l, \mathbf{s}_l)$;
    predict (1st Viterbi): $\hat{\mathbf{s}}_l = \arg\max_{\mathbf{s}} \langle \mathbf{w}, \Phi^{s}(\mathbf{x}_l, \mathbf{s}) \rangle$;
    if $\hat{\mathbf{s}}_l \neq \mathbf{s}_i$ **then**
      update with $\mathbf{w}$ with Eq. 10, where $(\cdot)$ is $(\mathbf{x}_l, \mathbf{s}_l)$ and $(*)$ is $(\mathbf{x}_l, \hat{\mathbf{s}}_l)$;
    end
  end
end
$\mathbf{w} = \mathbf{cw}/K$;

Algorithm 1: Online Learning Algorithm for Coupled Sequences Labeling Model. (双链序列标注在线学习算法)

1 Introduction

In the fields of natural language processing (NLP), joint segmentation and tagging (S&T) task is an important research topic. Many NLP problems can be transformed to joint S&T task, such as shallow parsing(Sha and Pereira, 2003), named entity recognition(Zhou and Su, 2002), Chinese part-of-speech (POS) tagging(Ratnaparkhi, 1996) and so on. For example, there are no explicitly boundaries between words in Chinese sentence. Therefore, sentence must be segmented into sequence of words, in which each word would be assigned with a POS tag.

Recently many research works focused on joint S&T tasks, which can be categorized into two ways: pipeline and cross-label.

The pipeline approaches are to solve two subtasks in order, segmentation and tagging. However, the obvious disadvantage of these approaches is error propagation, which significantly affects the whole performance.

The cross-label approaches can avoid the problem of error propagation and achieve more higher performance on both subtasks (Ng and Low, 2004). However, due to the large number of labels, two problems arise: (1) The amount of parameters increases rapidly and would be apt to overfit to the training corpus; (2) The decoding efficiency by dynamic programming would decrease.

In addition, joint cross-label approaches cannot segment or tag sentences separately. For example, in Chinese POS tagging task, the joint model cannot segment sentences individually without tagging the sentences. Moreover, if the sentences are already segmented, the joint model can not tag individually with the existing segmentation information.

In this paper, we present a novel joint model for S&T task with coupled sequences labeling. The proposed model integrates two linear Markov chains with a two dimensional structure. One chain is used for segmentation, and the other is for tagging. These two chains are labeled simultaneously, so our method does not suffer from error propagation. Unlike cross-label model, the number of labels in our model is much smaller. Experiments on two tasks, shallow parsing and Chinese POS tagging, demonstrate the effectiveness of our model.

The contributions of our methods are as follows:

1. Instead of cross-product labels, two types of nodes in our model make us represent features more flexibly.
2. Exact decoding algorithm can be employed to find the best S&T sequences simultaneously.
3. Our method not only can do joint S&T task, it can also segment or tag sentences separately.
4. Our model can be trained simultaneously with the heterogeneous data sources.

It is very important in practice that to utilize the heterogeneous data sources. For example in Chinese POS tagging, we can use two datasets (segmentation dataset and POS tagging dataset) for training parameters. This character is especially useful since the segmentation dataset is more easily annotated than POS tagging dataset.

The rest of the paper is organized as follows: In section 2, we describe the general sequence labeling method. In section 3 we present our novel model with coupled sequences labeling, then we analysis its complexity and discuss its applications. The experimental results are shown in section 4. In section 5, we introduce the related works. Finally, we conclude our work in section 6.

2 Joint Sequences Labeling Model

In this section, we first introduce and analyze joint S&T task with common sequence labeling model. Then, we present the joint cross-label approach and analyze its complexity.

2.1 Sequence Labeling Model

Sequence labeling is the task of assigning labels $\mathbf{y} = y_1, \ldots, y_L$ to an input sequence $\mathbf{x} = x_1, \ldots, x_L$.

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

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} F(\mathbf{w}, \Phi(\mathbf{x}, \mathbf{y})), \tag{1}$$

where $\mathbf{w}$ is the parameter of function $F(\cdot)$. The feature vector $\Phi(\mathbf{x}, \mathbf{y})$ consists of lots of overlapping features, which is the chief benefit of discriminative model.

For example, in first-order Markov sequence labeling model, the feature can be denoted as $\phi_k(y_{i-1}, y_i, \mathbf{x}, i)$, where i is the position in the sequence. Then the score function can be rewritten as

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} F(\sum_{i=1}^{L} \sum_{k} w_k \phi_k(y_{i-1}, y_i, \mathbf{x}, i)), \tag{2}$$

where L is length of $\mathbf{x}$.

2.2 Joint S&T with Cross-Label Sequence Labeling Model

In the traditional approach for joint S&T, each label y_i is the cross-product of segmentation label s_i and tagging label t_i, usually with the form of s_i-t_i. Therefore, the state space of cross-labels is $|\mathcal{Y}| = |\mathcal{S}| \times |\mathcal{T}|$, where $|\mathcal{S}|$, $|\mathcal{T}|$ is the number of segmentation labels and tagging labels, respectively.

In real applications, $|\mathcal{S}|$ is always small, while $|\mathcal{T}|$ will be very large. In segmentation task, there are several commonly used label sets such as {B, I}, {B, I, O}, {B, I, E, S}, etc. For example, {B, I, E, S} represent *Begin*, *Inside*, *End* of a multi-node segmentation, and *Single* node segmentation respectively. In tagging task, the label set depends on the detail definition of the task, such as {PER, LOC, ORG, MISC} in classic name entity recognition task, and {NNS, NNPS, NNP, ...} in Part-of-Speech tagging task.

Although joint learning with cross-label can avoid error propagation, which usually occurs in pipeline frameworks, the complexity of decoding algorithm would be increased rapidly due to the increased state space. Suppose we use first order Viterbi algorithm for decoding in linear chain model, the complexity is $(|\mathcal{S}||\mathcal{T}|)^2 L$ in such joint labeling frameworks, while $(|\mathcal{S}|^2 + |\mathcal{T}|^2)L$ in pipeline frameworks.

3 Coupled Sequences Labeling Model

In this section, we will describe the coupled sequences labeling model in detail, and propose an exact inference algorithm for finding two best sequences simultaneously. Then we apply this model to the problems mentioned in the beginning of this paper. Finally an online training algorithm is proposed to learn the parameters of our model by optimizing two difference inference algorithms.

3.1 Model Description

Different from the cross-label model, we define two sequences $\mathbf{s} = s_1, \ldots, s_L$ and $\mathbf{t} = t_1, \ldots, t_L$ for an input sequence $\mathbf{x} = x_1, \ldots, x_L$. $\mathbf{s}$ and $\mathbf{t}$ represent the segmentation and tagging labels respectively.

Then we employ a hybrid model by integrating these two linear chains. While keeping relative independence and completeness of these two chains, we also consider the interactions between them in order to cope with error propagation. The graphic structure of our model is shown in Figure 1.

Besides the original undirected edges (hereinafter to be referred as edges) existed in two linear chains, corresponding to $e(s_{i-1}, s_i)$ and $e(t_{i-1}, t_i)$, we also append two kinds of edges between different chains. $e(s_i, t_i)$ is equivalent to the representation of "Cross-Label" mentioned in section 2.2. Meanwhile, we also add an edge $e(t_{i-1}, s_i)$ into the model. This change brings about two new different cliques, respectively associated with variables $C_1 = \{s_{i-1}, t_{i-1}, s_i\}$ and $C_2 = \{t_{i-1}, s_i, t_i\}$,which essentially gives rise to the increment of the complexity of our model.

The reason for this change is to avoid the "label bias" problem citeLafferty:2001 in factorial CRF (FCRF) (Sutton et al., 2004). FCRF model has a similar graphic structure to our model, shown in Figure 2. As in the case of Chinese POS tagging, if given a context s_{i-1} = "B", t_{i-1} = "NN" and s_i = "E", t_i would be assigned to "JJ" instead of "NN" with a higher probability, since the transition from "NN" to "JJ" defeats against the transition to "NN" while s_i = "E" has no effect. However, s_i = "M" provides a strong clue, which implies that word w_{i-1} and w_i are in the same segmentation and would be assigned to the same label.

3.2 Inference Algorithm

According to the theory of probabilistic graphical models (Koller and Friedman, 2009), we can define a score function $F(\cdot)$ as the logarithmic potential function:

$$F(\mathbf{w}, \Phi(\mathbf{x}, \mathbf{s}, \mathbf{t})) = \sum_{i}^{L} \{\mathbf{w}^T \Phi_{C_1}(s_{i-1}, t_{i-1}, s_i, \mathbf{x}, i) + \mathbf{w}^T \Phi_{C_2}(t_{i-1}, s_i, t_i, \mathbf{x}, i)\}, \tag{3}$$

Given an observed sequence $\mathbf{x}$, the aim of inference algorithm is to find two best label sequences simultaneously with the highest score. In order to adapt to our model with two kinds of 3-variable cliques, we make some modifications of a second order Viterbi algorithm (Thede and Harper, 1999). We define two functions for recording the score of the best partial path from the beginning of the sequence to the position i:

$$\begin{aligned} \delta_i(t_{i-1}, s_i) \triangleq F(\mathbf{w}, \Phi(\mathbf{x}, s_{0:i}, t_{0:i-1})) &= \arg\max_{\mathbf{s_{i-1}}} \{\eta_{i-1}(s_{i-1}, t_{i-1}) + \mathbf{w}^T \Phi_{C_1}(s_{i-1}, t_{i-1}, s_i, \mathbf{x}, i)\}, \\ \eta_i(s_i, t_i) \triangleq F(\mathbf{w}, \Phi(\mathbf{x}, s_{0:i}, t_{0:i})) &= \arg\max_{\mathbf{t_{i-1}}} \{\delta_i(t_{i-1}, s_i) + \mathbf{w}^T \Phi_{C_2}(t_{i-1}, s_i, t_i, \mathbf{x}, i)\}, \end{aligned} \tag{4}$$

Initially, only features associated with variables s_0 and t_0 are hired. Without loss of generality, we set s_{-1} = "BoS" [1], t_{-1} = "BoT" [2] and $\eta_{-1}(s_{-1}, t_{-1}) = 0$. Then iteratively calculate these two

[1] denotes "Beginning of Segmentation"
[2] denotes "Beginning of Tagging"

score functions for any possible partial path. At last, the final score of two label sequences is

$$F(\mathbf{w}, \Phi(\mathbf{x}, \mathbf{s}, \mathbf{t})) = \eta_L(s_L, t_L), \tag{5}$$

Compared with the complexity of other joint models, the complexity of our model is $O((|\mathcal{S}|^2|\mathcal{T}| + |\mathcal{T}|^2|\mathcal{S}|)L)$, which is lower than cross-label model but higher than pipeline model. Although the asymptotic complexity is higher than pipeline model, the advantage is that our model would not suffer from error propagation and could make use of label information more efficiently.

3.3 Discussion of Coupled Sequences Labeling Model

As shown in Figure 1, our model can label two sequences simultaneously. However, we hope our model can be applied to solve the "inconsistent" problem (Section 1). Although two linear chains **s** and **t** are modeled in a hybrid framework, they still retain its complete structure. This means that we can independently use the linear chain **s** to segment a sentence (Figure 3a), while use the linear chain **t** to label a segmented sentence (Figure 3b) or use the whole structure to label two sequences together (Figure 3c). Therefore, we need two inference algorithms, respectively a first order Viterbi algorithm for segmenting a sentence when only using the linear chain **s**, and a second order Viterbi algorithm (see Section 3.2) for other two applications. The main idea behind this method is the fact that there are many overlapping features used in both segmentation and tagging tasks since most of the features are extracted from a local context.

Another reason to employ different inference algorithms is to maintain the decoding speed for different applications. If only used in the segmentation task, the complexity of our method is $O(|\mathcal{S}|^2 L)$. If applied to tag a segmented sentence, the complexity is $O((|\mathcal{T}| + |\mathcal{T}|^2)L) = O(|\mathcal{T}|^2 L)$. If used in the joint labeling task, its complexity is still $O((|\mathcal{S}|^2|\mathcal{T}| + |\mathcal{T}|^2|\mathcal{S}|)L)$.

However, in the coupled sequences labeling model, two linear chains are highly dependent due to the edge $e(t_{i-1}, s_i)$. It implies that if we train a model by using the whole structure, we cannot directly use the segmentation features, which are only related to the segmentation chain **s**. Therefore, we need to optimize the whole structure together with the segmentation chain.

Besides training a model with a corpus annotated segmentation and tagging labels, we can also use heterogeneous corpora because two inference algorithms are jointly optimized. It is very meaningful in real applications. As we all known, difficulties of annotating corpus for different tasks are different. In our setting, segmentation corpus are easy to annotate while tagging corpus are difficulty. As a result, we can easily obtain a large segmentation corpus while a small tagging corpus. Because we aim to optimize two chains simultaneously, it is possible for us to training a unified model with these two different scale corpora. We can learn parameters from a small tagging corpus for two chains, and learn parameters from a large segmentation corpus for the segmentation chain.

3.4 Learning Parameters with Passive-Aggressive Algorithm

In the training stage, we use passive-aggressive algorithm to learn the model parameters. Passive-aggressive (PA) algorithm (Crammer and Singer, 2003; Crammer et al., 2006) was proposed for normal multi-class classification and can be easily extended to structure learning (Crammer et al., 2005). Like perceptron, PA is an online learning algorithm.

Because two inference algorithms are needed to optimize in our framework, without loss of generality, we use $(\cdot)$ to represent the gold answer while $(*)$ to the response of an inference algorithm

with the highest score. In the segmentation task, $(\cdot)$ equals to $(\mathbf{x}, \mathbf{s})$ and $(*)$ is $(\mathbf{x}, \hat{\mathbf{s}})$. In the joint task, $(\cdot)$ denotes $(\mathbf{x}, \mathbf{s}, \mathbf{t})$ and $(*)$ is $(\mathbf{x}, \hat{\mathbf{s}}, \hat{\mathbf{t}})$. Here $\hat{\mathbf{s}}, \hat{\mathbf{t}}$ are the incorrect labels with the highest scores.

We can define the **margin** $\gamma(\mathbf{w}; (\cdot))$ as

$$\gamma(\mathbf{w}; (\cdot)) = F(\mathbf{w}, \Phi(\cdot)) - F(\mathbf{w}, \Phi(*)), \tag{6}$$

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

$$\ell_w = \begin{cases} 0, & \gamma(\mathbf{w}; (\cdot)) > 1 \\ 1 - \gamma(\mathbf{w}; (\cdot)), & \text{otherwise} \end{cases} \tag{7}$$

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

$$\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w} - \mathbf{w}_k||^2 + \mathcal{C} \cdot \xi, \tag{8}$$

$$\textbf{s.t. } \ell(\mathbf{w}; (\cdot)) <= \xi \textbf{ and } \xi >= 0 \tag{9}$$

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,

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \tau_k(\Phi(\cdot) - \Phi(*)), \tag{10}$$

where

$$\tau_k = \min(\mathcal{C}, \frac{\ell_k(\mathbf{w}; (\cdot))}{\|\Phi(\cdot) - \Phi(*)\|^2}) \tag{11}$$

Our training algorithm is based on PA algorithm and shown in Algorithm 1. In our algorithm, the input examples are randomly selected in each round k. According to the source of the selected example, we obtain the best response by using the proper inference algorithm, and finally update the parameters $\mathbf{w}$. Following (Collins, 2002), the average strategy is also adopted to avoid overfitting problem.

4 Experiments

We employ two joint sequence labeling tasks to show the performance of our model. In the following section, we would report our experiment settings and discuss the experiment results.

We compare our method with cross-label model and factorial model with PA algorithm. The factorial model is similar with Factorial CRF(Sutton et al., 2007), but its parameters are learning with PA algorithm.

We use the standard evaluation metrics $F1$ score, 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).

4.1 Datasets

In order to demonstrate the performance of our proposed model, we employ two joint segmentation and tagging tasks, respectively English shallow parsing and Chinese word segmentation and POS tagging (Chinese S&T).

In English shallow parsing, the corpus from CoNLL 2000 shared task is commonly used, which contains 8936 sentences for training and 2012 sentences for testing. We employ the commonly used label set $\{B, M, E, S\}$ in the segmentation task. 12 tagging labels, such as noun phrase (NP), verb phrase (VP),…and others (O), are used in the sequence tagging task.

In Chinese S&T, we employ the Chinese Treebank (CTB) corpus, obtained from the Fourth International SIGHAN Bakeoff datasets (Jin and Chen, 2008). The label set $\{B, M, E, S\}$ is also used for segmentation task.

4.2 Performance of Coupled Sequences Labeling Model

In the first experiment, we aim to compare the performances of our coupled sequences labeling model with other traditional joint models. Feature templates used in this experiment are summarized in Table 1 for English shallow parsing and in Table 2 for Chinese word segmenation and POS tagging, in which w_i denotes i^{th} word, p_i denotes i^{th} POS tag, c_i denotes i^{th} Chinese character.

We compare the total performance between traditional joint cross-label model, factorial model and our model. To learn the parameters of these models, we employ PA algorithm with an average parameter strategy to avoid the overfitting problem. The maximum amount of iterations is fixed to be 50.

The experiment results are shown in Table 3 for English shallow parsing, and in Table 4 for Chinese S&T. We also provides the performances of other methods reported in papers.

Table 3: Performances in English Shallow Parsing

Method	F1
Cross-Label CRFs	93.88
Voted Perceptrons (Carreras and Marquez, 2004)	93.74
Cross-Label model	93.47
Factorial model	93.11
Our model	93.94

Table 4: Performances in Chinese S&T

Method	$F1$
Pipeline	89.04
100-best reranking	89.23
Cross-label model	89.18
Factorial model	88.64
Our model	89.32

In English shallow parsing, our coupled sequences labeling model achieves the best performance than other two methods. We are surprised to find that the performance of the factorial model is lower than the joint cross-label model because of the "label bias" problem. However, a cross edge $e(t_{i-1}, si)$ is added into our coupled sequences model and shows its ability to avoid this problem. Experimental results in Chinese S&T show the similar conclusions.

4.3 Performance on Heterogeneous Corpora

The second experiment is to jointly train a unified model with heterogeneous corpora. In this experiment, we are expected to find out whether additional resources could increase the performance simultaneously on two different tasks.

We randomly divide the training corpus into two equal parts. One part is used as the joint S&T training corpus while another part is used for just segmentation training corpus.

For joint sequence labeling task, we employ our second order decoding algorithm (see 3.2) and the same feature templates (listed in Table 1 and Table 2) to extract features. For segmentation task, we

use the first order Viterbi algorithm to find the most possible segmentation. Only feature templates (listed in Table 1) irrelevant to tagging task are used for the segmentation task. Therefore template such as $w_{i-1}t_i$ would not be used to extract the segmentation features. Moreover this means two different tasks would share many common features in the training stage. The final test performances are shown in Table 5.

We experiment three real scenarios, respectively only segmenting a sentence, tagging a segmented sentence and jointly labeling a sentence. Notice that these different tasks use the same model in this experiment. Joint cross-label approach is chosen as the baseline, but this method cannot segment a sentence individually or tag a segmented sentence. However, our coupled sequences labeling model can handle these tasks in a unified model.

The experimental results are shown in Table 5 for English shallow parsing, and in Table 6 for Chinese S&T.

Table 5: Performances on English shallow parsing

	Corpus 1	Corpus 2	Segments	Segments(Joint)	Joint	Tagging
Cross-label	used	-	-	94.97	92.67	-
Cross-label	used	used	-	95.56	93.47	-
Our model	used	-	94.24	95.14	93.02	95.20
Our model	used	used	94.89	95.65	93.94	96.02
Our model	used	used as Seg	94.68	95.56	93.61	95.73

Note 1: "used" means that the corpus is used joint S&T task with both segmentation and tagging labels, "used as Seg" means that the corpus is just used with segmentation labels.

Note 2: "Segments" means the performance of segmentation which just used the segmentation chain, "Segments(joint)" means the performance of segmentation in joint labeling, "Tagging" is the performance of tagging when given gold segmentation labels.

Table 6: Performances on Chinese S&T

	Corpus1	Corpus2	Segments	Segments(Joint)	Joint	Tagging
Cross-label	used	-	-	93.53	87.05	-
Cross-label	used	used	-	94.85	89.18	-
Our model	used	-	92.56	93.70	87.39	89.28
Our model	used	used	94.37	94.71	89.32	91.87
Our model	used	used as Seg	94.15	95.03	89.21	91.30

In Chinese S&T, we can find that our coupled model on joint labeling task can outperform the cross-label model in both the experiments of using half of the corpus and full corpus. With using the second corpus, all models have better performances. This result demonstrates a common sense in machine learning community "more data, more performance". However, after adding the second corpus, the performance of the joint task is promoted, but still lower than using the full annotations. It is reasonable because the second corpus is only used as a segmentation corpus. This means we only employ half of the POS annotations to train our model. Experimental results on segmentation task show that after adding the second corpus, our model improves the performance and slightly behind the model trained with the full annotations. With the performance increases on segmentation, the performance of tagging a segmented sentence is increased as well.

Similar conclusions can be found in the experiments of English shallow parsing.

The results also indicate that two different tasks efficiently help each other via shared features.

Therefore, we believe that additional resources could be introduced into our model more flexible and be helpful to the final performance.

4.4 Decoding Speed

At last, we also list the decoding speed for different tasks in Table 7.

Table 7: Decoding speed on English Shallow Parsing and Chinese S&T task. (sentences/second)

	English Shallow Parsing				Chinese S&T			
	Seg	Seg(Joint)	Joint	Tagging	Seg	Seg(Joint)	Joint	Tagging
Cross-label	-	1503	1453	-	-	117	113	-
Our model	22995	1467	1435	1601	17572	130	124	153

Compared to the joint cross-label approach in both corpora, our model has the equivalent decoding speed on the joint task. While on the task of tagging a segmented sentence, our model provides a slight decoding speedup. The reason is that the states of segmentation are much less than tagging. However, on the only segmentation task, our model provides a decoding speedup over 10 times, since we can use the segmentation chain independently in our model.

5 Related Works

Several methods have been proposed to cope with the problems of joint S&T task.

Sutton et al. (2004, 2007) proposed Dynamic Conditional Random Fields (DCRF) to jointly represent the different tasks in a single graphical model. However, the exact training and inference for DCRF are time-consuming.

Duh (2005) proposed a model for jointly labeling multiple sequences. The model is based on the Factorial Hidden Markov Model (FHMM). Since FHMM is directed graphical model, FHMM requires considerably less computation than DCRFs and exact inference is easily achievable. However, the FHMM's generative framework cannot take full advantage of context features, so its performance is lower than DCRF.

Different with our model applied in joint S&T task, both the DCRF and FHMM are used in POS tagging and NP Chunking tasks. These two tasks are not strongly dependent on each other. Therefore, their models are relatively simplified for joint S&T task.

6 Conclusion

In this paper, we propose a novel joint S&T model by integrating two linear chains into a coupled sequence labeling model. Our approach does not suffer from the problem of error propagation, which usually occurs in pipeline models. Meanwhile, our proposed model would not result in the rapid increase of states as cross-label models. Our model also takes the advantage of more flexible feature representation, a uniform model with a flexible combination of labeling tasks, etc.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work was funded by NSFC (No.61003091 and No.61073069), 863 Program (No.2011AA010604) and 973 Program (No.2010CB327900).

References

Carreras, X. and Marquez, L. (2004). Phrase recognition by filtering and ranking with perceptrons. *Recent advances in natural language processing III: selected papers from RANLP 2003*, 260:205.

Collins, M. (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*.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Crammer, K., McDonald, R., and Pereira, F. (2005). Scalable large-margin online learning for structured classification. In *NIPS Workshop on Learning With Structured Outputs*. Citeseer.

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

Duh, K. (2005). Jointly labeling multiple sequences: A factorial HMM approach. In *Proceedings of the ACL Student Research Workshop*, pages 19–24, Ann Arbor, Michigan. Association for Computational Linguistics.

Jin, C. and Chen, X. (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.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Ng, H. and Low, J. (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.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In Brill, E. and Church, K., editors, *Proceedings of the Empirical Methods in Natural Language Processing*, pages 133–142.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics.

Sutton, C., McCallum, A., and Rohanimanesh, K. (2007). Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *J. Mach. Learn. Res.*, 8:693–723.

Sutton, C., Rohanimanesh, K., and McCallum, A. (2004). Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *Proceedings of the twenty-first international conference on Machine learning*, page 99. ACM.

Thede, S. M. and Harper, M. P. (1999). A second-order hidden markov model for part-of-speech tagging. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 175–182. Association for Computational Linguistics.

Zhou, G. and Su, J. (2002). Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics.