A Dual-layer CRFs Based Joint Decoding Method for Cascaded Segmentation and Labeling Tasks

Yanxin Shi
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
yanxins@cs.cmu.edu

Mengqiu Wang
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
mengqiu@cs.cmu.edu

Abstract

Many problems in NLP require solving a cascade of subtasks. Traditional pipeline approaches yield to error propagation and prohibit joint training/decoding between subtasks. Existing solutions to this problem do not guarantee non-violation of hard-constraints imposed by subtasks and thus give rise to inconsistent results, especially in cases where segmentation task precedes labeling task. We present a method that performs joint decoding of separately trained Conditional Random Field (CRF) models, while guarding against violations of hard-constraints. Evaluated on Chinese word segmentation and part-of-speech (POS) tagging tasks, our proposed method achieved state-of-the-art performance on both the Penn Chinese Treebank and First SIGHAN Bakeoff datasets. On both segmentation and POS tagging tasks, the proposed method consistently improves over baseline methods that do not perform joint decoding.

1 Introduction

There exists a class of problems which involves solving a cascade of segmentation and labeling subtasks in Natural Language Processing (NLP) and Computational Biology. For instance, a semantic role labeling (SRL) system relies heavily on syntactic parsing or noun-phrase chunking (NP-chunking) to segment (or group) words into constituents. Based on the constituent structure, it then identifies semantic arguments and assigns labels to them [Xue and Palmer, 2004; Pradhan *et al.*, 2004]. And for Asian languages such as Japanese and Chinese that do not delimit words by space, solving word segmentation problem is prerequisite for solving part-of-speech (POS) labeling problem. Another example in the Computational Biology field is the DNA coding region detection task followed by sequence similarity based gene function annotation [Burge and Karlin, 1997].

Most previous approaches treat cascaded tasks as processes chained in a pipeline. A common shortcoming of those approaches is that errors introduced in upstream tasks propagate through the pipeline and cannot be easily recovered. Moreover, the pipeline structure prohibits the use of predictions of tasks later in the chain to help making better prediction of earlier tasks [Sutton and McCallum, 2005a]. Several new techniques have been proposed recently to address these problems. Sutton *et al.* [2004] introduced the Dynamic Conditional Random Fields (DCRFs) to perform joint training/decoding of subtasks. One disadvantage of this model is that exact inference is generally intractable and can become prohibitively expensive for large datasets. Sutton and McCallum [2005a] presented an alternative model by decoupling the joint training and only performing joint decoding. Kudo *et al.* [2004] presented another Conditional Random Field (CRF) [Lafferty *et al.*, 2001] based model that performs Japanese word segmentation and POS tagging jointly. Another popular approach of joint decoding for cascaded tasks is to combine multiple predicting tasks into a single tagging or labeling task [Luo, 2003; Ng and Low, 2004; Yi and Palmer, 2005; Miller *et al.*, 2000].

However, all aforementioned approaches do not work well for cases where a segmentation task comes before a labeling task. That is because in such cases, the segmentation task imposes hard-constraints that cannot be violated in successive tasks. For example, if a Chinese POS tagger assigns different POS labels to characters within the same word, as defined by a word segmenter, the word will not get a single consistent POS labels. Similarly, the constituent constraints imposed by syntactic parsing and NP-chunking tasks disallow argument overlaps in semantic role labeling [Pradhan *et al.*, 2004]. From a graphical modeling's perspective, those models all assign nodes to the smallest units in the base segmentation task (e.g., in the case of Chinese word segmentation, the smallest unit is one Chinese character). As a result, those models cannot ensure consistency between the segmentation and labeling tasks. For instance, [Ng and Low, 2004] can only evaluate POS tagging results on a per-character basis, instead of a per-word basis. Hindered by the same problem, [Kudo *et al.*, 2004] only considered words predefined in a lexicon for constructing possible Japanese word segmentation paths, which puts limit on the generality of their model.

To tackle this problem, we propose a dual-layer CRFs based method that exploits joint decoding of cascaded sequence segmentation and labeling tasks that guards against violations of those hard-constraints imposed by segmentation task. In this method, we model the segmentation and labeling tasks by dual-layer CRFs. At decoding time, we first perform individual decoding in each layer. Upon these individual decodings, a probabilistic framework is constructed in order to find the most probable joint decodings for both subtasks. At training time, we trained a cascade of individual CRFs for the subtasks, for given our application's scale, joint training is much more expensive [Sutton and McCallum, 2005a]. Evaluated on Chinese word segmentation and part-of-speech (POS) tagging tasks, our proposed method achieved state-of-the-art performance on both the Penn Chinese Treebank [Xue *et al.*, 2002] and First SIGHAN Bakeoff datasets [Sproat and Emerson, 2003]. On both segmentation and POS tagging tasks, the proposed method consistently improves over baseline methods that do not perform joint decoding. In particular, we report the best published performance on the AS open track of the First SIGHAN Bakeoff dataset and also the best average performance on the four open tracks.

To facilitate our discussion, in later sections we will use Chinese segmentation and POS tagging as a working example to illustrate the proposed approach, though it should be clear that the model is applicable to any cascaded sequence labeling problem.

2 Joint Decoding for Cascaded Sequence Segmentation and Labeling Tasks

In this section, using Chinese sentence segmentation and POS tagging as an example, we present a joint decoding method applicable for cascaded segmentation and labeling tasks.

2.1 A Unified Framework to Combine Chinese Sentence Segmentation and POS Tagging

Let $\mathbf{C} = \{C_1, C_2, ..., C_n\}$ denote the observed Chinese sentence where C_i is the i^{th} Chinese character in the sentence, $\mathbf{S} = \{S_1, S_2, ..., S_n\}$ denote a segmentation sequence over $\mathbf{C}$ where $S_i \in \{B, I\}$ represents segmentation tags ($Begin$ and $Inside$ of a word), $\mathbf{T} = \{T_1, T_2, ..., T_m\}$ denote a POS tagging sequence where $m \leq n$ and $T_j \in \{the\ set\ of\ possible\ POS\ labels\}$. Our goal is to find a segmentation sequence and a POS tagging sequence that maximize the joint probability $P(\mathbf{S}, \mathbf{T}|\mathbf{C})$[1]. Let $\hat{\mathbf{S}}$ and $\hat{\mathbf{T}}$ denote the most likely segmentation and POS tagging sequences for a given Chinese sentence, respectively. By applying chain rule, $\hat{\mathbf{S}}$ and $\hat{\mathbf{T}}$ can be obtained as follows:

$$\begin{aligned} <\hat{\mathbf{S}}, \hat{\mathbf{T}}> &= \arg\max_{\mathbf{S},\mathbf{T}} P(\mathbf{S}, \mathbf{T}|\mathbf{C}) \\ &= \arg\max_{\mathbf{S},\mathbf{T}} P(\mathbf{T}|\mathbf{S}, \mathbf{C}) P(\mathbf{S}|\mathbf{C}) \quad (1) \\ &= \arg\max_{\mathbf{S},\mathbf{T}} P(\mathbf{T}|\mathbf{W(C,S)}) P(\mathbf{S}|\mathbf{C}) \quad (2) \end{aligned}$$

Equation 1 can be rewritten as Equation 2, since given a sequence of characters $\mathbf{C} = \{C_1, C_2, ..., C_n\}$ and a segmentation $\mathbf{S}$ over it, a sentence can be interpreted as a sequence of words $\mathbf{W(C,S)} = \{W_1, W_2, ..., W_m\}$.

Note that the joint probability $P(\mathbf{S}, \mathbf{T}|\mathbf{C})$ is factorized into two terms, $P(\mathbf{T}|\mathbf{W(C,S)})$ and $P(\mathbf{S}|\mathbf{C})$. The first term represents the probability of a POS tagging sequence $\mathbf{T}$ built upon the segmentation $\mathbf{S}$ over sentence $\mathbf{C}$, while the second term represents the probability of the segmentation sequence $\mathbf{S}$. Maximizing the product of these two terms can be viewed as a reranking process. For a particular sentence $\mathbf{C}$, we maintain a list of all possible segmentations over this sentence sorted by the their probability $P(\mathbf{S}|\mathbf{C})$. For each segmentation $\mathbf{S}$ in this list, we can find a POS tagging sequence $\mathbf{T}$ over $\mathbf{S}$ that maximizes the probability $P(\mathbf{T}|\mathbf{W(C,S)})$. Using the product of these two probabilities, we can then rerank the segmentation sequences. The segmentation sequence $\mathbf{S}$ that is reranked to be the top of the list of all possible segmentations is the final segmentation sequence output, and the most probable POS tagging sequence along with this segmentation is our final POS tagging output. Such a final pair always maximizes the joint probability $P(\mathbf{S}, \mathbf{T}|\mathbf{C})$.

Intuitively, given a segmentation over a sentence, if the maximum of probabilities of all POS tagging sequences built upon this segmentation is very small, it can be a signal that tells us there is high chance that the segmentation is incorrect. In this case, we may be able to find another segmentation that does not have a probability as high as the first one, but the best POS tagging sequence built upon this segmentation has a much more reasonable probability, so that the joint probability $P(\mathbf{S}, \mathbf{T}|\mathbf{C})$ is increased.

2.2 N-Best List Approximation for Decoding

To find the most probable segmentation and POS tagging sequence pair, exact inference by enumerating over all possible segmentation is generally intractable, since the number of possible segmentations over a sentence is exponential to the number of characters in the sentence.

To overcome this problem, we propose a N-best list approximation method. Instead of exhaustively computing the list of all possible segmentations, we restrict our reranking targets to the N-best list $\mathcal{S} = \{\mathbf{S_1}, \mathbf{S_2}, ..., \mathbf{S_N}\}$, where $\{\mathbf{S_1}, \mathbf{S_2}, ..., \mathbf{S_N}\}$ is ranked by the probability $P(\mathbf{S}|\mathbf{C})$. Then, the approximated solution that maximizes the joint probability $P(\mathbf{S}, \mathbf{T}|\mathbf{C})$ can be formally described as:

$$\begin{aligned} <\hat{\mathbf{S}}, \hat{\mathbf{T}}> &= \arg\max_{\mathbf{S}\in\mathcal{S},\mathbf{T}} P(\mathbf{S}, \mathbf{T}|\mathbf{C}) \\ &= \arg\max_{\mathbf{S}\in\mathcal{S},\mathbf{T}} P(\mathbf{T}|\mathbf{W(C,S)}) P(\mathbf{S}|\mathbf{C}) \quad (3) \end{aligned}$$

Comparing to other similar work that uses N-best lists and SVM for reranking [Daume and Marcu, 2004; Asahara *et al.*, 2003], or perform rule-based post-processing for error correction [Xue and Shen, 2003; Gao *et al.*, 2004; Ng and Low, 2004], our method has a unique advantage that it outputs not just the best segmentation and POS sequence but also a joint probability estimate. This probability estimate allows more natural integration with higher level NLP applications that are also based on probabilistic models, and even reserves room for further joint inference.

3 Dual-layer Conditional Random Fields

In Section 2, we have already factorized the joint probability into two terms $P(\mathbf{S}|\mathbf{C})$ and $P(\mathbf{T}|\mathbf{W(C,S)})$. Notice that

[1] Note that a segmentation-POS tagging sequence pair is meaningful only when the POS tagging sequence $\mathbf{T}$ is labeled on the basis of the segmentation result $\mathbf{S}$, or the pair of $\mathbf{S}$ and $\mathbf{T}$ is consistent. In our proposed method, the joint probabilities $P(\mathbf{S}, \mathbf{T}|\mathbf{C})$ of inconsistent pairs of $\mathbf{S}$ and $\mathbf{T}$ are defined to be 0.

both terms are probabilities of a whole label sequence given some observed sequences. Thus, we use Conditional Random Fields (CRFs) [Lafferty *et al.*, 2001] to define these two probability terms.

CRFs define conditional probability, $P(\mathbf{Z}|\mathbf{X})$, by Markov random fields. In the case of Chinese segmentation and POS tagging, the Markov random fields in CRFs are chain structure, where **X** is the sequence of characters or words, and **Z** is the segmentation tags for characters (B or I, used to indicate word boundaries) or the POS labels for words (NN, VV, JJ, etc.). The conditional probability is defined as:

$$P(\mathbf{Z}|\mathbf{X}) = \frac{1}{N(\mathbf{X})}\exp\,(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(\mathbf{Z},\mathbf{X},t)) \qquad (4)$$

where $N(\mathbf{X})$ is a normalization term to guarantee that the summation of the probability of all label sequences is one. $f_k(\mathbf{Z},\mathbf{X},t)$ is the k^{th} $localfeaturefunction$ at sequence position t. It maps a pair of **X** and **Z** and an index t to {0,1}. $(\lambda_1, ..., \lambda_K)$ is a weight vector to be learned from training set.

We model separately the two probability terms defined in our model ($P(\mathbf{S}|\mathbf{C})$ and $P(\mathbf{T}|\mathbf{W(C,S)})$) using the dual-layer CRFs (Figure 1). The probability $P(\mathbf{S},\mathbf{T}|\mathbf{C})$ that we want to maximize can be written as:

$$\begin{aligned} P(\mathbf{S},\mathbf{T}|\mathbf{C}) &= P(\mathbf{T}|\mathbf{W(C,S)})P(\mathbf{S}|\mathbf{C}) \qquad (5)\\ &= \frac{1}{N_T(\mathbf{W(C,S)})}\times\frac{1}{N_S(\mathbf{C})}\\ &\quad\times\exp\,(\sum_{j=1}^{m}\sum_{k=1}^{K_T}\lambda_k f_k(\mathbf{T},\mathbf{W(C,S)},j))\\ &\quad\times\exp\,(\sum_{i=1}^{n}\sum_{k=1}^{K_S}\mu_k g_k(\mathbf{S},\mathbf{C},i)) \qquad (6)\end{aligned}$$

where m and n are the number of words and characters in the sentence, respectively, $N_T(\mathbf{W(C,S)})$ and $N_S(\mathbf{C})$ are the normalizing terms to ensure the sum of the probabilities over all possible **S** and **T** is one. $\{\lambda_k\}$ and $\{\mu_k\}$ are the parameters for the first and the second layer CRFs, respectively, f_k and g_k are the $localfeaturefunctions$ for the first and the second layer CRFs, respectively. Their properties and functions are the same as common CRFs described before.

The N-best list of segmentation sequences $\mathcal{S}$ and the value of their corresponding probabilities $P(\mathbf{S}|\mathbf{C})$ ($\mathbf{S} \in \mathcal{S}$) can be obtained using modified Viterbi algorithm and A* search [Schwartz and Chow, 1990] in the first layer CRF. Given a particular sentence segmentation **S**, the most probable POS tagging sequence **T** and its probability $P(\mathbf{T}|\mathbf{W(C,S)})$ can be inferred by the Viterbi algorithm [Lafferty *et al.*, 2001] in the second layer CRF. Having the N-best list of segmentation sequences and their corresponding most probable POS tagging sequences, we can use the joint decoding method proposed in Section 2 to find the optimal pair of segmentation and its POS tagging defined by Equation 3.

We divide the learning process into two: one for learning the first layer segmentation CRF, and one for learning the second layer POS tagging CRF. First we learn the parameters $\boldsymbol{\mu}$ = $\{\mu_1, ..., \mu_{K_S}\}$ using algorithm based on improved iterative scaling algorithm (IIS) [Della Pietra *et al.*, 1997] to maximize the log-likelihood of the first layer CRF. Then we learn the parameters $\boldsymbol{\lambda} = \{\lambda_1, ..., \lambda_{K_T}\}$ also using IIS, to maximize the log-likelihood of the second layer CRF. A detailed derivation of this learning algorithm for each learning step can be found in [Lafferty *et al.*, 2001].

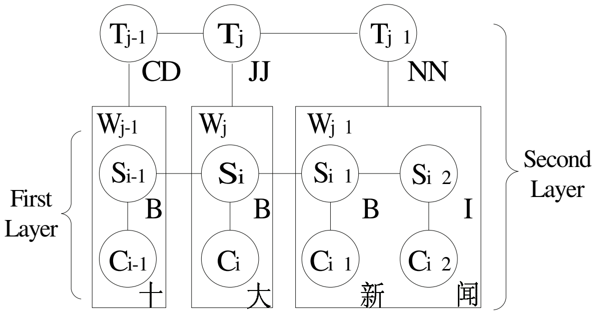

Figure 1: Dual-layer CRFs. $P(\mathbf{S},\mathbf{T}|\mathbf{C})$, the joint probability of a segmentation sequence **S** and a POS tagging sequence **T** given sentence **C** is modeled by the dual-layer CRFs. In the first layer CRF, the observed nodes are the characters in the sentence and the hidden nodes are segmentation tags for these characters. In the second layer CRF, given the segmentation results from the first layer, characters combine to form ”supernodes” (words). These words are the observed variables, and POS tagging labels for them are the hidden variables.

4 Features for CRFs

Features for Word Segmentation

The features we used for word segmentation are listed in the top half of Table 1. Feature (1.1)-(1.5) are the basic segmentation features we adopted from [Ng and Low, 2004]. In (1.6), $L_{Begin}(C_0)$, $L_{End}(C_0)$ and $L_{Mid}(C_0)$ represent the maximum length of words found in a lexicon that contain the current character as either the first, last or middle character, respectively. In (1.7), $Single(C_0)$ indicates whether the current character can be found as a single word in the lexicon.

Besides the adopted basic features mentioned above, we also experimented with additional semantic features (Table 1, (1.8)-(1.9)). In these features, $Sem(C_0)$ refers to the semantic class of current character, and $Sem(C_{-1})$, $Sem(C_1)$ represent the semantic class of characters one position to the left and right of the current character, respectively. We obtained a character's semantic class from HowNet [Dong and Dong, 2006]. Since many characters have multiple semantic classes defined by HowNet, it is a non-trivial task to choose among the different semantic classes. We performed contextual disambiguation of characters' semantic classes by calculating semantic class similarities. For example, let us assume the current character is (*look,read*) in a word context of (*read newspaper*). The character (*look*) has two semantic classes in HowNet, i.e. (*read*) and (*doctor*). To determine which class is more appropriate, we check the example words illustrating the meanings of the two semantic classes given by HowNet. For (*read*), the example word is (*read book*); for (*doctor*), the example word is (*see a doctor*). We then calculated the semantic class

Segmentation features
(1.1) $C_n, n \in [-2, 2]$
(1.2) $C_nC_{n+1}, n \in [-2, 1]$
(1.3) $C_{-1}C_1$
(1.4) $Pu(C_0)$
(1.5) $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$
(1.6) $L_{Begin}(C_0)$, $L_{End}(C_0)$
(1.7) $Single(C_0)$
(1.8) $Sem(C_0)$
(1.9) $Sem(C_n)Sem(C_{n+1}), n \in -1, 0$
POS tagging features
(2.1) $W_n, n \in [-2, 2]$
(2.2) $W_nW_{n+1}, n \in [-2, 1]$
(2.3) $W_{-1}W_1$
(2.4) $W_{n-1}W_nW_{n+1}, n \in [-1, 1]$
(2.5) $C_n(W_0), n \in [-2, 2]$
(2.6) $Len(W_0)$
(2.7) Other morphological features

Table 1: Feature templates list

similarity scores between (*newspaper*) and (*book*), and (*newspaper*) and (*illness*), using HowNet's built-in similarity measure function. Since (*newspaper*) and (*book*) both have semantic class (*document*), their maximum similarity score is 0.95, where the maximum similarity score between (*newspaper*) and (*illness*) is 0.03. Therefore, $Sem(C_0)Sem(C_1)$ = (*read*)(*document*). Similarly, we can figure out $Sem(C_{-1})Sem(C_0)$. For $Sem(C_0)$, we simply picked the top four frequent semantic classes ranked by HowNet, and used "NONE" for absent values.

Features for POS Tagging

The bottom half of Table 1 summarizes the feature templates we employed for POS tagging. W_0 denotes the current word. W_{-n} and W_n refer to the words *n* positions to the left and right of the current word, respectively. $C_n(W_0)$ is the n^{th} character in current word. If the number of characters in the word is less than 5, we use "NONE" for absent characters. $Len(W_0)$ is the number of characters in the current word. We also used a group of binary features for each word, which are used to represent the morphological properties of current word, e.g. whether the current word is punctuation, number, foreign name, etc.

5 An Illustrating Example of the Joint Decoding Method

In this section, we use an illustrating example to motivate our proposed method. This example found in the real output of our system gives suggestive evidences that POS tagging helps predicting the right segmentation, and the right segmentation is more likely to get a better POS tagging sequence. We are only showing a snippet of the full sentence due to space limit:

The production and sales situation of foreign owned companies is relatively good.

The segmentation that has the highest probability (0.52) is:

(*foreign owned*) (*company*) (*production*)
(*situation*) (*situation*) (*relatively good*)

The second best segmentation which has probability 0.36 is:

(*foreign owned*) (*company*) (*production*)
(*situation*) (*situation*) (*relatively*) (*good*)

The only difference from the first sequence is that (*relatively good*) was segmented into two words (*relatively*) (*good*). Despite the lower probability, the second segmentation is more appropriate, since the two characters that compose the word (*relatively good*) carry their own meanings as individual words.

The traditional methods would have stopped here and use the first segmentation as the final output, though it is actually incorrect according to the gold-standard. Our joint decoding method further performs POS tagging based on each of the segmentation sequences. The POS tagging sequence with the highest probability (0.23) assigned to the first segmentation is:

(NN) (NN) (NN) (NN) (NN)

(VV) (PU)

where NN represents noun, VV represents other verb, and PU represents punctuation. The second segmentation was assigned the following POS label sequence with the highest probability 0.45:

(NN) (NN) (NN) (NN) (NN)

(AD) (VA) (PU)

where AD represents adverb, VA represents predicative adjective.

The best POS sequence arising from the second segmentation is more discriminative than the best sequence based on the first segmentation, which indicates the second segmentation is more informative for POS tagging. The joint probability of the second segmentation and POS tagging sequence (0.16) is higher than the joint probability of the first one (0.12), and therefore our method reranks the second one as the best output. According to the gold-standard, the second segmentation and POS tagging sequences are indeed the correct sequences.

6 Results

We evaluate our model using the Penn Chinese Treebank (CTB) [Xue *et al.*, 2002] and open tracks from the First International SIGHAN Chinese Word Segmentation Bakeoff datasets [Sproat and Emerson, 2003].

Using a linear-cascade of CRFs with the same set of features listed in Table 1 as baseline, we compared the performance of our proposed method. The accuracies of both word segmentation and POS tagging are measured by recall (R), precision (P), and F-measure which equals to $\frac{2RP}{R+P}$. For segmentation, recall is the percentage of correct words that were produced by the segmenter, and precision is the percentage of automatically segmented words that are correct. For POS tagging, recall is the percentage of all gold-standard words that are correctly segmented and labeled by our system, and precision is percentage of words returned by our system that are correctly segmented and labeled. We chose a N value of 20 for using in the N-best list, based on cross-validation results.

6.1 Results of Segmentation

For segmentation, we first evaluated our joint decoding method on the CTB corpus. 10-fold cross-validation was performed on this corpus. Results are summarized in Table 2.

	1	2	3	4	5	6	7	8	9	10	average
Baseline	97.3%	97.2%	95.4%	96.7%	96.2%	93.1%	95.9%	94.8%	95.7%	96.2 %	95.85%
Joint decoding	97.4%	97.3%	95.7%	96.9%	96.4%	93.4%	96.0%	95.2%	95.9%	96.3%	96.05%

Table 2: Comparison of 10-fold cross-validation segmentation results on CTB corpus. Each column represents one out of the 10-fold cross-validation results. The last column is the average result over the 10 folds.

As can been seen in Table 2, the results in all of the 10-fold tests improved with our joint decoding method. We conducted pairwise t-test and our joint decoding method was found to be statistically significantly better than the baseline method under confidence level 5.0^{-4} (p-values).

We also evaluated our proposed method on the open tracks of the SIGHAN Bakeoff datasets. These datasets are designed only for evaluation of segmentation results, and no POS tagging information were provided in the training corpus. However, since the learning of the POS tagging model and the segmentation model is decoupled, we can use a separate training corpus to learn the second layer POS tagging CRF model, and still be able to take advantage of the proposed joint decoding method. The results comparing to the baseline method are summarized in Table 3.

	AS			CTB			PK			HK		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Baseline	96.7%	96.8%	96.7%	88.5%	88.3%	88.4%	94.9%	94.9%	94.9%	94.9%	95.5%	95.2%
Joint Decoding	96.9%	96.7%	96.8%	89.4%	88.7%	89.1%	95.3%	95.0%	95.2%	95.0%	95.4%	95.2%

Table 3: Overall results on First SIGHAN Bakeoff open tracks. *P* stands for precision, *R* stands for recall, *F1* stands for the F1 measure.

For comparison of our results against previous work and other systems, we summarize the results on the four open tracks in Table 4. We adopted the table used in [Peng *et al.*, 2004] for consistency and ease of comparison. There were 12 participating teams (sites) in the official runs of the First International SIGHAN Bakeoff, here we only show the 8 teams that participated in at least one of the four open tracks. Each row represents a site, and each cell gives the F1 score of a site on one of the open tracks. The second to fifth columns contain results on the 4 open tracks, where a bold entry indicates the best performance on that track. Column S-Avg contains the average performance of the site over the tracks it participated, where a bold entry indicates that this site on average performs better than our system; column O-Avg is the average of our system over the same runs, where a bolded entry indicates our system performs better on average than the other site. Our results are shown in the last row of the table.

In the official runs, no team achieved best results on more than one open track. We achieved the best runs on AS open (ASo) track with a F1 score of 96.8%, 1.1% higher than the second best system [Peng *et al.*, 2004]. Comparing to Peng *et al.* [2004], whose CRFs based Chinese segmenter were also evaluated on all four open tracks, we achieved higher performance on three out of the four tracks. Our average F1 score over all four tracks is 94.1%, 0.5% higher than that of Peng *et al.*'s system. Comparing with other sites using the average measures in the right-most two columns, we outperformed seven out of the nine sites. And the two sites that have higher average performance than us both did significantly better on the CTB open (CTBo) track. The official results showed that almost all systems obtained the worst performance on the CTBo track, due to inconsistent segmentation style in the training and testing sets [Sproat and Emerson, 2003].

	ASo	CTBo	HKo	PKo	S-Avg	O-Avg
S01		88.1%		**95.3%**	91.7%	**92.2%**
S02		**91.2%**			**91.2%**	89.1%
S03	87.2%	82.9%	88.6%	92.5%	87.8%	**94.1%**
S04				93.7%	93.7%	**95.2%**
S07				94.0%	94.0%	**95.2%**
S08			**95.6%**	93.8%	94.7%	**95.2%**
S10		90.1%		**95.9%**	**93.0%**	92.2%
S11	90.4%	88.4%	87.9%	88.6%	88.8%	**94.1%**
Peng *et al.* '04	95.7%	89.4%	94.6%	94.6%	93.6%	**94.1%**
Our System	**96.8%**	89.1%	95.2%	95.2%		94.1%

Table 4: Comparisons against other systems (this table is adopted from Peng *et al.* 2004)

6.2 Results of POS Tagging

Since the Bakeoff competition does not provide gold-standard POS tagging outputs, we only used CTB corpus to compare the POS tagging results of our joint decoding method with the baseline method. We performed 10-fold cross-validation on the CTB corpus. The results are summarized in Table 5.

	1	2	3	4	5	6	7	8	9	10	average
Baseline	93.8%	93.7%	90.2%	92.0%	93.3%	87.2%	92.2%	90.8%	91.5%	92.0 %	91.67%
Joint Decoding	94.0%	93.9%	90.4%	92.2%	93.4%	87.5%	92.4%	91.0%	91.7%	92.1%	91.86%

Table 5: Comparison of 10-fold cross-validation POS tagging results on CTB corpus. Each column represents one out of the 10-fold cross-validation results. The last column is the average result over the 10 folds.

From Table 5 we can see that our joint decoding method has higher accuracy in each one of the 10 fold tests than the baseline method. Pairwise t-test showed that our method was found to be significantly better than the baseline method under the significance level of 3.3^{-5} (p-value). This improvement on POS tagging accuracy can be understood as the result of the improved segmentation accuracy through joint decoding (as shown in Table 2). Therefore, these results showed that our joint decoding method not only helps to improve segmentation results, it also benefits POS tagging results.

7 Discussion on Reranking

Reranking technique has been successfully applied in many NLP applications before, such as speech recognition [Stolcke *et al.*, 1997], NP-bracketing [Daume and Marcu, 2004] and semantic role labeling (SRL) [Toutanova *et al.*, 2005]. However, It is worth pointing out that the contexts in which they applied reranking method differ from ours in that we use reranking as an approximation technique for joint decoding. One similar work which also used reranking as approximation to joint decoding is [Sutton and McCallum, 2005b]. Nevertheless, their experiments showed negative results when reranking was applied to the task of joint parsing and SRL. One possible explanation is that the maximum entropy classifier they used is based on a local log-linear model, while CRFs employed by our method model the joint probability of the entire sequence, and therefore are more natural for our proposed joint decoding method.

8 Conclusion

We introduced a unified framework to integrate cascaded segmentation and labeling tasks by joint decoding based on dual-layer CRFs. We applied our method to Chinese segmentation and POS tagging tasks, and demonstrated the effectiveness of our method. Our proposed method not only enhances both segmentation and POS tagging accuracy, but it also offers an insight to improving performance of a task by learning from its related tasks.

References

[Asahara *et al.*, 2003] M. Asahara, C. Goh, X. Wang, and Y. Matsumoto. Combining segmenter and chunker for Chinese word segmentation. In *Proceedings of ACL SIGHAN Workshop*, 2003.

[Burge and Karlin, 1997] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 1997.

[Daume and Marcu, 2004] H. Daume and D. Marcu. NP bracketing by maximum entropy tagging and SVM reranking. In *Proceedings of EMNLP*, 2004.

[Della Pietra *et al.*, 1997] S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE TPAMI*, 1997.

[Dong and Dong, 2006] Z. Dong and Q. Dong. *HowNet And The Computation Of Meaning*. World Scientific, 2006.

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

[Kudo *et al.*, 2004] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of EMNLP*, 2004.

[Lafferty *et al.*, 2001] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, 2001.

[Luo, 2003] X. Luo. A maximum entropy Chinese character-based parser. In *Proceedings of EMNLP*, 2003.

[Miller *et al.*, 2000] S. Miller, H. Fox, L. Ramshaw, and R. Weischedel. A novel use of statistical parsing to extract information from text. In *Proceedings of ANLP*, 2000.

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

[Peng *et al.*, 2004] F. Peng, F. Feng, and A. McCallum. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of COLING*, 2004.

[Pradhan *et al.*, 2004] S. Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky. Shallow semantic parsing using support vector machines. In *Proceedings of HLT*, 2004.

[Schwartz and Chow, 1990] R. Schwartz and Y. Chow. The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses. In *Proceedings of ICASSP*, 1990.

[Sproat and Emerson, 2003] R. Sproat and T. Emerson. The first international Chinese word segmentation bakeoff. In *Proceedings of ACL SIGHAN Workshop*, 2003.

[Stolcke *et al.*, 1997] A. Stolcke, Y. Konig, and M. Weintraub. Explicit word error minimization in N-best list rescoring. In *Proceedings of Eurospeech*, 1997.

[Sutton and McCallum, 2005a] C. Sutton and A. McCallum. Composition of conditional random fields for transfer learning. In *Proceedings of HLT/EMNLP*, 2005.

[Sutton and McCallum, 2005b] C. Sutton and A. McCallum. Joint parsing and semantic role labeling. In *Proceedings of CoNLL*, 2005.

[Sutton *et al.*, 2004] C. Sutton, K. Rohanimanesh, and A. McCallum. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *Proceedings of ICML*, 2004.

[Toutanova *et al.*, 2005] K. Toutanova, A. Haghighi, and C. Manning. Joint learning improves semantic role labeling. In *Proceedings of ACL*, 2005.

[Xue and Palmer, 2004] N. Xue and M. Palmer. Calibrating features for semantic role labeling. In *Proceedings of EMNLP*, 2004.

[Xue and Shen, 2003] N. Xue and L. Shen. Chinese word segmentation as LMR tagging. In *Proceedings of ACL SIGHAN Workshop*, 2003.

[Xue *et al.*, 2002] N. Xue, F. Chiou, and M. Palmer. Building a large-scale annotated Chinese corpus. In *Proceedings of COLING*, 2002.

[Yi and Palmer, 2005] S. Yi and M. Palmer. The integration of syntactic parsing and semantic role labeling. In *Proceedings of CoNLL*, 2005.