Hidden Dynamic Probabilistic Models for Labeling Sequence Data *

Xiaofeng YU **Wai LAM**

Information Systems Laboratory
Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{xfyu,wlam}@se.cuhk.edu.hk

Abstract

We propose a new discriminative framework, namely Hidden Dynamic Conditional Random Fields (HDCRFs), for building probabilistic models which can capture both internal and external class dynamics to label sequence data. We introduce a small number of hidden state variables to model the sub-structure of a observation sequence and learn dynamics between different class labels. An HDCRF offers several advantages over previous discriminative models and is attractive both, conceptually and computationally. We performed experiments on three well-established sequence labeling tasks in natural language, including part-of-speech tagging, noun phrase chunking, and named entity recognition. The results demonstrate the validity and competitiveness of our model. In addition, our model compares favorably with current state-of-the-art sequence labeling approach, Conditional Random Fields (CRFs), which can only model the external dynamics.

Introduction

The problem of annotating or labeling observation sequences arises in many applications across a variety of scientific disciplines, most prominently in natural language processing, speech recognition, information extraction, and bioinformatics. Recently, the predominant formalism for modeling and predicting label sequences has been based on discriminative models and variants. Conditional Random Fields (CRFs) (Lafferty, McCallum, & Pereira 2001) are perhaps the most commonly used technique for probabilistic sequence modeling.

More specifically, CRFs have been shown to achieve state-of-the-art performance in a variety of domains (Sutton & McCallum 2006). CRFs are undirected graphical models trained to maximize the conditional probability of the desired outputs given the corresponding inputs. The primary advantage of CRFs is their conditional nature, resulting in the relaxation of strong independence assumptions required by Hidden Markov Models (HMMs) in order to ensure tractable inference. Additionally, CRFs avoid the label bias problem of Maximum Entropy Markov Models (MEMMs) (McCallum, Freitag, & Pereira 2000) and other discriminative directed graphical based Markov models. And they have the great flexibility to encode a wide variety of arbitrary, non-independent features and to straightforwardly combine rich domain knowledge.

However, structured data are widely prevalent in the real world. Observation sequences tend to have distinct internal sub-structure and indicate predictable relationships between individual class labels, especially for natural language. For example, in the task of noun phrase chunking, a noun phrase begins with a noun or a pronoun, optionally accompanied by a set of modifiers. A noun phrase may contain one or more base noun phrases. In the named entity recognition task, many named entities have particular characteristics in their composition. A location name can optionally end with a location salient word, but cannot end with any organization salient word. A complex, nested organization name may be composed of a person name, a location name, or even another organization name. These complex and expressive structures can largely influence predictions. The efficiency of the CRF approach heavily depends on its first order Markov property - given the observation, the label of a token is assumed to depend only on the labels of its adjacent tokens. The CRF approach models the transitions between class labels to enjoy advantages of both generative and discriminative methods, thus capturing external dynamics, but unfortunately it lacks the ability to represent internal sub-structure.

A feasible solution to this problem is to directly model the internal sub-structure in sequence data. We incorporate a set of observed variables with additional latent, or hidden state variables to model relevant sub-structure in a given sequence, thus resulting a new discriminative framework, Hidden Dynamic Conditional Random Fields (HDCRFs), by modeling both external dependencies between different class labels and internal sub-structure in given sequences. Our proposed model learns the external dependencies by modeling a continuous stream of class labels, and it learns internal sub-structure by utilizing intermediate hid-

*The work described in this paper is substantially supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Nos: CUHK4193/04E and CUHK4128/07) and the Direct Grant of the Faculty of Engineering, CUHK (Project Codes: 2050363 and 2050391). This work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

den states. HDCRFs define a conditional distribution over the class labels and hidden state labels conditioned on the observations, with dependencies between the hidden variables expressed by an undirected graph. Similar to CRFs, this modeling is also able to deal with features that can be arbitrary functions of the observations. Efficient parameter estimation and inference can be carried out using standard graphical model algorithms such as belief propagation. In this paper, we first describe the HDCRF model, including the training and inference algorithms. We then describe experiments that demonstrate the effectiveness of HDCRFs to outperform the best previous discriminative fully-observable CRF model on three well-known natural language processing tasks.

Related Work

There is an extensive literature dedicated to sequence modeling. Graphical models are a natural formalism for modeling sequences. Traditionally, graphical models (e.g., HMMs and stochastic grammars) were generative, and they have been used to represent the joint probability to paired observation and label sequences. But modeling the joint distribution can lead to difficulties when using the rich local features that can occur in relational data, which can include complex dependencies. Modeling these dependencies among inputs can lead to intractable models, but ignoring them can lead to reduced performance.

A significant amount of recent work has shown the power of discriminative models for sequence labeling tasks. In the speech and natural language processing communities, one of the first large-scale applications of CRFs was by Sha & Pereira (2003), who matched state-of-the-art performance on segmenting noun phrases in text. Since then, CRF models have been successfully used for tasks such as word recognition, part-of-speech tagging, text classification and information extraction. For example, Pinto *et al.* (2003) applied CRFs to extract tables from documents and showed improvements over an HMM approach. Peng & McCallum (2004) used a CRF model to extract information (e.g., titles, authors, and affiliations) from research papers with high performance.

More recently, several variations of CRF model have been investigated. Kristjansson *et al.* (2004) proposed a constrained CRF framework for interactive information extraction. A constrained Viterbi decoding was designed to find the optimal field assignments consistent with the fields explicitly specified or corrected by the user, and a mechanism was proposed for estimating the confidence of each extracted field to filter out low-confidence extractions. However, this framework needs manual assistance via a user interface, which is hardly available for sequence labeling. Zhu *et al.* proposed a two-dimensional Conditional Random Fields (2D CRFs) to better incorporate the neighborhood interactions for Web information extraction, based on the observation that the information on a Web page is two-dimensionally laid out. This is also different from our proposed model which handles one-dimensional sequence data that have internal sub-structure. Sutton, Rohanimanesh, & McCallum (2004) presented Dynamic Conditional Random Fields (DCRFs). As a particular case, a factorial CRF (FCRF) was used to jointly solve two sequence labeling tasks (noun phrase chunking and part-of-speech tagging) on the same observation sequence. Improved accuracy was achieved by modeling the interactions between the two tasks. Unfortunately, training a DCRF model with unobserved nodes (hidden variables) makes their approach difficult to optimize. Our HDCRF incorporates hidden state variables with an explicit partition, and inference can be efficiently computed using belief propagation during both training and testing.

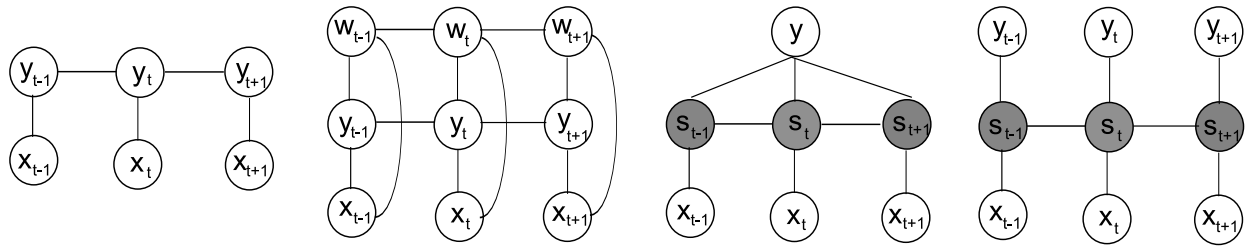
When observation data have distinct sub-structure, models that exploit hidden state are advantageous. A related model is presented by Gunawardana *et al.* (2005), who build a hidden-state CRF (HCRF) which can estimate a class label given a segmented sequence in a phone classification task. Since they are trained on sets of pre-segmented sequences, HCRF models do not capture the dynamics between class labels, only the internal structure. Gunawardana *et al.* (2005) applied HCRFs to segmented sequences, leaving segmentation as a pre-processing step. A similar model for natural language parsing is shown by Koo & Collins (2005).

Other related models are Hidden Markov Random Fields (HMRFs) (Kunsch, Geman, & Kehagias 1995) and Dynamic Bayesian Networks (DBNs) (Dean & Kanazawa 1989; Murphy 2002). HMRFs, DBNs and HDCRFs employ a layer of latent variables with an undirected graph specifying dependencies between those variables. However, there is an important difference that HMRFs and DBNs model a joint distribution over latent variables and observations, whereas HDCRFs are a discriminative model.

It is known that models which include latent or hidden-state structure may be more expressive than fully observable models, and can often find relevant sub-structure in a given domain. To the best of our knowledge, this is the first attempt at incorporating hidden-state structure into discriminative models, yielding a new framework that models both dependencies between external class labels and internal sub-structure in observation sequences. Experimental study shows that our model can be run directly on unsegmented sequence data, with improved recognition performance over conventional discriminative sequence methods.

Hidden Dynamic Conditional Random Fields

Suppose X is a random variable over data sequences to be labeled, and Y is a random variable over corresponding label sequences. Our task is to learn a mapping between an observation sequence $X = (x_1, x_2, \ldots, x_T)$ and the corresponding label sequence $Y = (y_1, y_2, \ldots, y_T)$. Each Y is a number of a set $\mathcal{Y}$ of possible class labels. For each sequence, we also assume a vector of sub-structure variables $S = (s_1, s_2, \ldots, s_T)$, which are not observed in training examples, and will therefore form a set of hidden variables. Each s_j is a member of a finite set $\mathcal{S}_{y_j}$ of possible hidden states for the class label y_j. Suppose $\mathcal{S}$ is the set of all possible hidden states of all $\mathcal{S}_y$ sets. Intuitively, each s_j corresponds to a labeling of x_j with some member of $\mathcal{S}$, which may correspond to the sub-structure.



(a) CRFs (b) DCRFs (c) HCRFs (d) HDCRFs

Figure 1: Graphical representation and comparison of our HDCRFs with three previously published models. (a) linear-chain CRFs. $\{x_{t-1}, x_t, x_{t+1}\}$ is observation sequence and $\{y_{t-1}, y_t, y_{t+1}\}$ is label sequence. (b) Dynamic CRFs (DCRFs), where $\{y_{t-1}, y_t, y_{t+1}\}$ and $\{w_{t-1}, w_t, w_{t+1}\}$ are two different label sequences. DCRFs include links between different labels. (c) Hidden-state CRFs (HCRFs), where $\{s_{t-1}, s_t, s_{t+1}\}$ is a set of hidden variables assigned to observation sequence and only one class label y for an observation sequence. (d) our proposed Hidden Dynamic CRFs (HDCRFs). Hidden variables $\{s_{t-1}, s_t, s_{t+1}\}$ are used to model the internal sub-structure in observation sequence. Note that only the link with the current observation x_t is shown, but for all four models, long range dependencies are also possible.

Given the above definitions, we define a hidden dynamic probabilistic model as follows:

$$p(Y|X) = \sum_S p(Y|S, X) \cdot p(S|X). \quad (1)$$

By definition, sequences which have any $s_j \notin \mathcal{S}_{y_j}$ will obviously have $p(Y|S, X) = 0$, so we can rewrite our above model as:

$$p(Y|X) = \sum_{S:\forall s_j \in \mathcal{S}_{y_j}} p(S|X). \quad (2)$$

Similar to CRFs, the conditional probability distributions, $p(S|X)$, take the form

$$p(S|X) = \frac{1}{Z(X)} \exp\Big(\sum_k \lambda_k \cdot \sum_{j=1}^{T} f_k\big(s_{j-1}, s_j, X, j\big)\Big), \quad (3)$$

where $Z(X)$ is an instance-specific normalization function

$$Z(X) = \sum_S \exp\Big(\sum_k \lambda_k \cdot \sum_{j=1}^{T} f_k\big(s_{j-1}, s_j, X, j\big)\Big), \quad (4)$$

and $f_k(s_{j-1}, s_j, X, j)_{k=1}^K$ is a set of real-valued feature functions. $\Lambda = \{\lambda_k\} \in \Re^K$ is a parameter vector, reflecting the confidence of feature functions. Each feature function is either a transition function $t_k(s_{j-1}, s_j, X, j)$ over the entire observation sequence and the hidden variables at positions i and $i-1$, or a state function $s_k(s_j, X, j)$ depends on a single hidden variable at position i.

Note that our proposed model is different from HCRFs, which model the conditional probability of one class label y given the observation sequence X through

$$p(y|X) = \frac{1}{Z'(X)} \sum_{S \in y} \exp\Big(\lambda \cdot f\big(y, S, X\big)\Big) \quad (5)$$

where the partition function $Z'(X)$ is given by

$$Z'(X) = \sum_{y, S \in y} \exp\Big(\lambda \cdot f\big(y, S, X\big)\Big). \quad (6)$$

HCRFs do capture the internal sub-structure in the observation sequence by definition. However, they output only one class label y for an entire sequence X and they lack the ability to model the dependencies between class labels, which have been shown to be essentially important for labeling sequence data. Due to this disadvantage, HCRFs can only be used to label the segmented sequences. And without considering dependencies between class labels may lead to reduced performance. On the other hand, CRFs can model dependencies between class labels. But unfortunately they cannot represent internal sub-structure in sequence data. As we show, Our HDCRFs combine the strengths of CRFs and HCRFs by modeling both external dependencies between class labels and internal sub-structure. Yet, this modeling still allows practically efficient solutions to large-scale real world sequence labeling tasks.

An illustrative representation of HDCRFs and comparison with other three models is shown in Figure 1. In the graphical structure $G = (V, E)$, each transition function defines an edge feature, while each state function defines a node feature. All the features respect the structure of graph G, in that no feature depends on more than two hidden variables (s_i, s_j), and if a feature does depend on variables s_i and s_j, there must be an edge (i, j) in the graph G. The graph G can be encoded arbitrarily and it captures domain specific knowledge such as the internal sub-structure.

It is worth noticing that the weights λ_k associated with the transition function $t_k(s_{j-1}, s_j, X, j)$ model both the internal sub-structure and external dependencies between different class labels. Weights associated with a transition function for hidden states that are in the same subset $\mathcal{S}_{y_j}$ will model the sub-structure patterns, while weights associated with the transition functions for hidden states from different subsets will model the external dependencies between labels.

Parameter Estimation

Given some training data consist of n labeled sequences $\mathcal{D} = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, the parameters $\Lambda = \{\lambda_k\}$ are set to maximize the conditional log-likelihood. Following previous work on CRFs, we use the following ob-

jective function to estimate the parameters:

$$L(\Lambda) = \sum_{i=1}^{n} \log P_\Lambda(Y_i|X_i). \quad (7)$$

To avoid over-fitting, log-likelihood is usually penalized by some prior distribution over parameters that provides smoothing to help copy with sparsity in the training data. A commonly used prior is zero-mean (with variance σ^2) Gaussian (Lafferty, McCallum, & Pereira 2001; Peng & McCallum 2004). With a Gaussian prior, log-likelihood is penalized as follows:

$$L(\Lambda) = \sum_{i=1}^{n} \log P_\Lambda(Y_i|X_i) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2}. \quad (8)$$

We encode structural constraints with an undirected graph structure, where the hidden variables $\{s_1, s_2, \ldots, s_T\}$ correspond to some vertices in the graph. To keep training and inference tractable, we restrict our model to have disjoint sets of hidden states associated with each class label. We use a quasi-Newton gradient ascent method to search for the optimal parameter values, $\Lambda^* = \arg\max_\Lambda L(\Lambda)$, under this criterion.

$$\forall Y \in \mathcal{Y}, j \in 1 \ldots T, a \in \mathcal{S},$$
$$P(s_j = a|Y, X) = \sum_{S:s_j=a} P(S|Y, X) \quad (9)$$

$$\forall Y \in \mathcal{Y}, j, k \in 1 \ldots T, a, b \in \mathcal{S},$$
$$P(s_j = a, s_k = b|Y, X) = \sum_{S:s_j=a, s_k=b} P(S|Y, X) \quad (10)$$

where $P(s_j = a|Y, X)$ and $P(s_j = a, s_k = b|Y, X)$ are marginal distributions over individual variables s_j or pairs of variables $\{s_j, s_k\}$ corresponding to edges in the graph. The gradient of $L(\Lambda)$ can be defined in terms of these marginals and can therefore be calculated efficiently.

We first consider derivatives with respect to the parameters λ_k associated with a state function s_k. Taking derivatives gives

$$\begin{aligned}\frac{\partial L(\Lambda)}{\partial \lambda_k} &= \sum_S P(S|Y, X) \sum_{j=1}^{T} s_k(s_j, X, j) \\ &\quad - \sum_{Y', S} P(Y', S|X) \sum_{j=1}^{T} s_k(s_j, X, j) \\ &= \sum_{j,a} P(s_j = a|Y, X) s_k(j, a, X) \\ &\quad - \sum_{Y', j, a} P(s_j = a, Y'|X) s_k(j, a, X) \end{aligned} \quad (11)$$

It shows that $\frac{\partial L(\Lambda)}{\partial \lambda_k}$ can be expressed in terms of components $P(s_j = a|X)$ and $P(Y|X)$, which can be computable using belief propagation (Pearl 1988).

For derivatives with respect to the parameters λ_l corresponding to a transition function t_l, a similar calculation gives

$$\begin{aligned}\frac{\partial L(\Lambda)}{\partial \lambda_l} &= \sum_{j,k,a,b} P(s_j = a, s_k = b|Y, X) t_l(j, k, a, b, X) \\ &\quad - \sum_{Y', j, k, a, b} P(s_j = a, s_k = b, Y'|X) t_l(j, k, a, b, X), \end{aligned} \quad (12)$$

hence $\frac{\partial L(\Lambda)}{\partial \lambda_l}$ can also be expressed in terms of expressions (e.g., the marginal probabilities $P(s_j = a, s_k = b|Y, X)$) that can be computed efficiently using belief propagation. We perform gradient ascent with the limited-memory quasi-Newton BFGS optimization technique (Liu & Nocedal 1989).

Inference

Given a new test sequence X, we would like to estimate the most probable labeling sequence Y^* that maximizes our conditional model:

$$Y^* = \arg\max_Y P(Y|X, \Lambda^*) \quad (13)$$

where the parameters Λ^* are learned via training process. Assuming each class label is associated with a disjoint set of hidden states, Equation 13 can be rewritten as:

$$Y^* = \arg\max_Y \sum_{S:\forall s_j \in \mathcal{S}_{y_j}} P(Y|X, \Lambda^*) \quad (14)$$

The marginal probabilities $P(s_j = a|X, \Lambda^*)$ are computed for all possible hidden states $a \in \mathcal{S}$ to estimate the label y_j^*. Then these marginal probabilities are summed according to the disjoint sets of hidden states $\mathcal{S}_{y_j}$ and the label associated with the optimal set is chosen. As discussed in the previous subsection, these marginal probabilities can also be computed efficiently using belief propagation. In our experiments we use the above maximal marginal probabilities approach to estimate the sequence of labels since it minimizes the error.

Experiments

We carried out three sets of experiments on part-of-speech (POS) tagging, noun phrase (NP) chunking, and named entity recognition (NER) to explore the performance of our HDCRF model against the CRF model, using three standard natural language datasets. We also investigated the number of hidden states per class labels and its impact on recognition performance and training time. To make fair and accurate comparison, we used the same set of features on each set of experiment for both HDCRFs and CRFs. All experiments were performed on the Linux platform, with a 3.2GHz Pentium 4 CPU and 4 GB of memory.

We trained our HDCRF model using the objective function specified by Equation 8. For training and validation, we varied the number of hidden states per class label from 1 to 3 and the regularization term with values $10^k, k \in [-2, 2]$

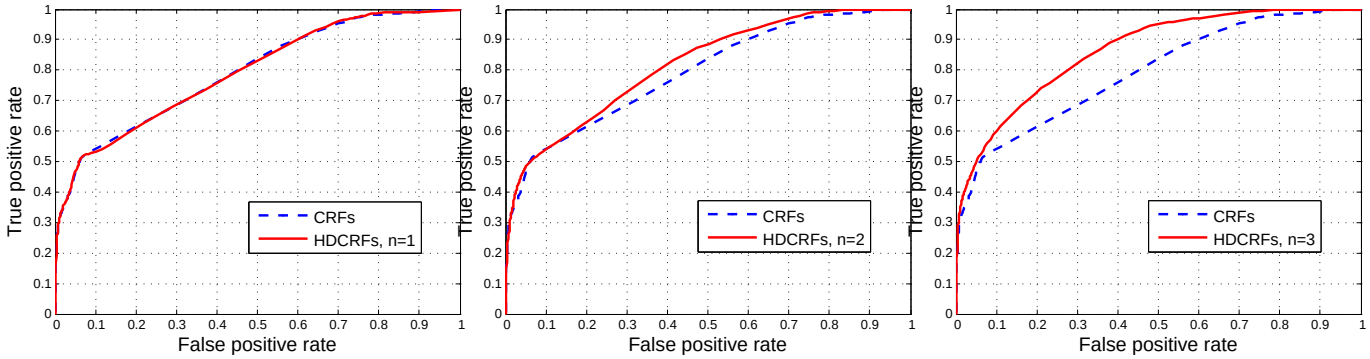


Figure 2: Performance comparison of HDCRFs and CRFs on POS tagging. n is the number of hidden states per class labels, $n \in [1, 3]$.

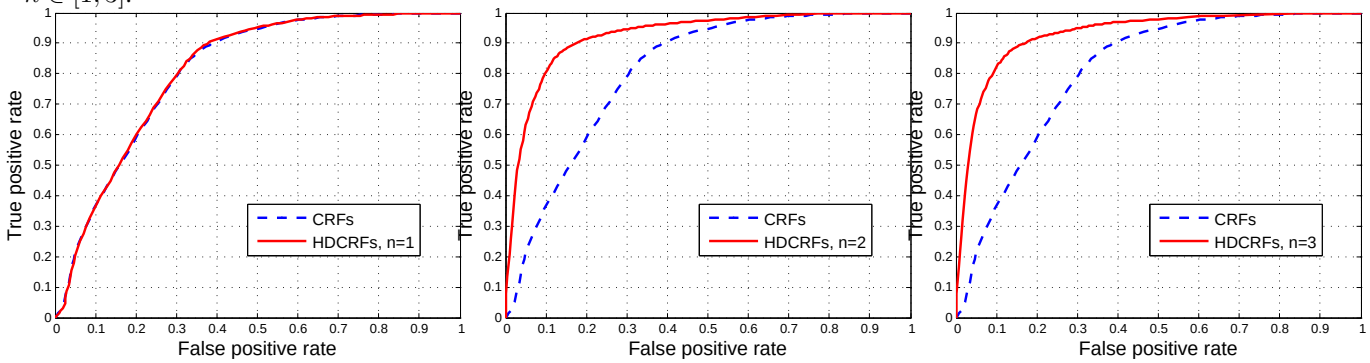


Figure 3: Performance comparison of HDCRFs and CRFs on NP chunking. n is the number of hidden states per class labels, $n \in [1, 3]$.

to choose the optimal validation parameters. For testing, marginal probabilities were computed for each class label and each token of the sequence using belief propagation. The class label with the highest likelihood is assigned to the token in the sequence.

We also trained a CRF model as a baseline. The objective function of this model contains a regularization term similar to the regularization term shown in Equation 8 for our HDCRF model. For training and validation, this regularization term was also validated with values $10^k, k \in [-2, 2]$. For testing, marginal probabilities were computed for each class label and each token of the sequence using the Viterbi algorithm. The optimal class label for a specific token is selected to be the label with the highest marginal probability.

We then describe the above three well-established natural language processing tasks and corresponding datasets used in our experiments, and the well-engineered feature set used for each task in the following three subsections.

Part-of-Speech Tagging

As a well-established task in natural language processing, POS tagging is the process of marking up the tokens (words) in a sequence (sentence) as corresponding to a particular class label (part-of-speech such as noun, verb, adjective, adverb, etc), based on both its definition, as well as its context. We used Penn Treebank POS tagging dataset for our experiment. This dataset consists of sentences from the Wall Street Journal, and each word in a given input sentence must be labeled with one of 45 syntactic tags. We randomly selected 15,000 sentences from this dataset, and conducted the experiment with a rich feature set using a 80%-20% train-test split. Besides contextual features (e.g., several words before and after the current word), we used a set of morphological features: whether the word begins with a capitalized letter or all letters are capitalized, whether it contains a hyphen or digits, and whether it ends in one of the following suffixes: `-ance`, `-ence`, `-eer`, `-ist`, `-ician`, `-ion`, `-ity`, `-ies`, `-ify`, `-ize`, `-yse`, `-able`, `-ible`, `-ical`, `-ish`, `-ous`.

Noun Phrase Chunking

NP chunking can be viewed as a sequence labeling task by assigning each word as either BEGIN-PHRASE, INSIDE-PHRASE, or OTHER. Our data comes from the CoNLL 2000 shared task (Sang & Buchholz 2000). We consider each sentence to be a training instance, with single words as tokens. The data are divided into a standard training set of 8,936 sentences and a test set of 2,012 sentences. We used the POS features provided in the official CoNLL 2000 dataset. The POS features were generated by the Brill tagger (Brill 1995), which was trained on over 40,000 sentences of Wall Street Journal. The labeling accuracy of this tagger is around 95-97%. In addition, we also used some contextual and morphological features, similar to the features used for our POS tagging experiment.

Named Entity Recognition

NER is the task of identifying and classifying phrases that denote certain types of named entities (NEs), such as person names (PERS), locations (LOCS) and organizations (ORGS)

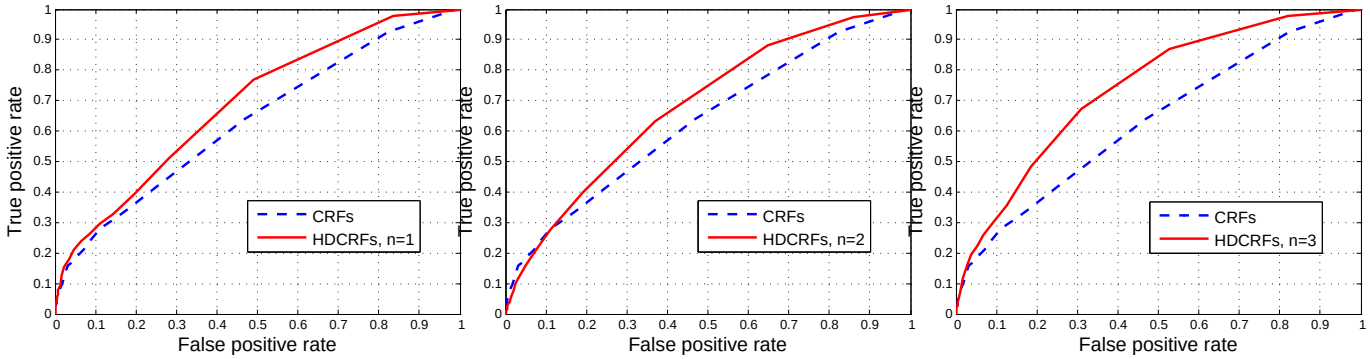


Figure 4: Performance comparison of HDCRFs and CRFs on NER. n is the number of hidden states per class labels, $n \in [1, 3]$.

in text documents. We used People's Daily corpus (January-Jun, 1998) in our experiments, which contains approximately 357K sentences, with tagged entities of 156K PERS, 219K LOCS and 87K ORGS, respectively. To reduce the training time, we selected 10% of the data from People's Daily to create a smaller training dataset containing 15,000 sentences and a testing dataset containing 3,000 sentences. We used features that have been shown to be very effective for NER, namely the current character and its POS tag, several characters surrounding the current character and their POS tags, current word and several words surrounding the current word, and some clue word features which can capture non-local dependencies. This gives us a rich feature set using both local and non-local information.

Results

We plot ROC curves to present the performance of our HDCRF model. For all the ROC curves, the true positive rate is computed by dividing the number of recognized labels by the total number of truth labels. Similarly, the false positive rate is computed by dividing the number of falsely recognized labels by the total number of class labels.

Figure 2, Figure 3 and Figure 4 compare the performance of HDCRFs and CRFs on POS tagging, NP chunking and NER, respectively. As shown in these figures, our HDCRF model significantly outperforms the baseline CRF methods. The difference is statistically significant ($p < 0.05$ with a 95% confidence interval) according to McNemar's paired tests on all the experiments.

Effect of Number of Hidden States on Performance: It is particularly interesting to know that the performance boosted on all the experiments when increasing the number of hidden states per class labels n from 1 to 3. For POS tagging and NP chunking, HDCRFs perform almost the same as CRFs when n is set to 1. For NER, HDCRFs significantly outperform CRFs when $n = 1$. This can be explained by the fact that there are much more sub-structures existing in named entities than in part-of-speech or noun phrases. These results also show that HDCRFs can model the internal substructure well with a small number ($1 \sim 3$) of hidden states. Our HDCRFs perform the best on NP chunking and worst on NER, illustrated by the AUC (Area Under the Curve) in the figures. This finding shows the truth that in the three tasks, NER is the most difficult, NP chunking is the easiest, and POS tagging is the one in between.

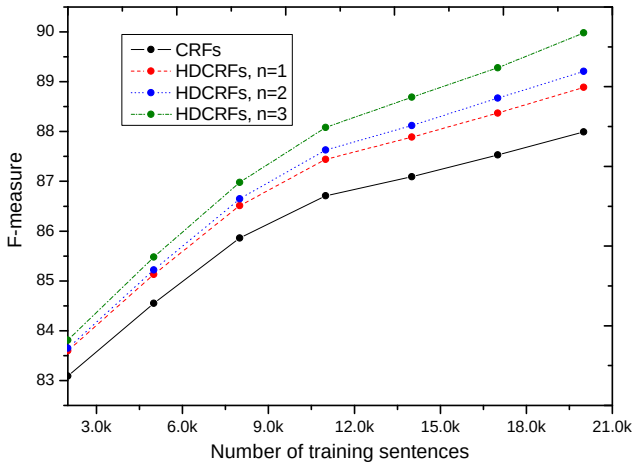


Figure 5: Effect of training set size on performance

Effect of Training Set Size on Performance: Figure 5 compares the F_1 [1] performance of both HDCRFs and CRFs on the NER task [2] with the increase of training set size. We increased the training set size from 2,000 sentences to 20,000 sentences, with an incremental step of 3,000. We found that increasing the training set size improves the performance for both HDCRFs and CRFs. As illustrated by the curves, HDCRFs consistently outperform CRFs. Moreover, with the same training set size, larger n leads to better performance for HDCRF models. When the training size only contains 2,000 sentences, for example, the F-measure for CRF model is 83.09%, and the F-measures for HDCRF models (n from 1 to 3) are 83.60%, 83.65%, and 83.81% respectively. This effect is prominent for larger training datasets. Using 20,000 sentences for training, HDCRFs obtain a maximum F-measure of 89.98% when n is set to 3, resulting in a 16.57% relative error reduction (RER) compared to the CRF model.

Effect of Number of Hidden States on Training Time: In Figure 6 we show the impact of increasing the number of

[1] F-measure is the harmonic mean of Precision (P) and Recall (R). There exists a deep relationship between ROC and PR spaces and Davis & Goadrich (2006) proved that a curve dominates in ROC space if and only if it dominates in PR space.

[2] For POS tagging and NP chunking, the curves are very similar, we omitted for space.

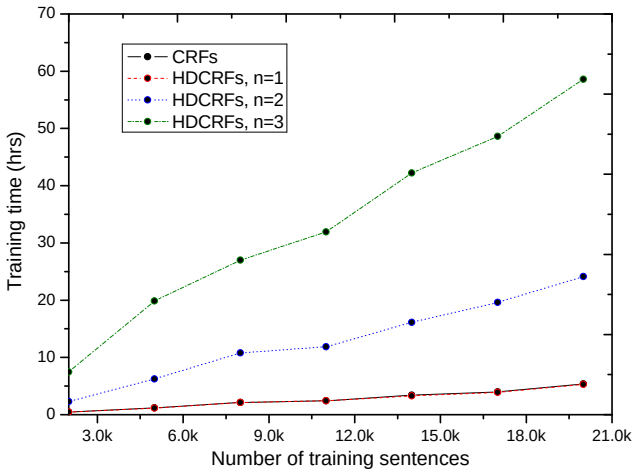


Figure 6: Effect of number of hidden states on training time

hidden states n (also from 1 to 3) on the training time for the NER task. Increasing n leads to an increase in the computational complexity of training procedure in such models. When $n = 1$, the training time of HDCRFs is almost the same as that of CRFs. Often when the training size is small, n does not effect the training time much, but when the training size increases, the gap between the curves broadens.

Conclusion

We have presented a discriminative framework which can model both the internal sub-structure and the external dependencies between different class labels for labeling sequence data. We incorporated a small number of hidden state variables to model the internal sub-structure for observation sequences. We found that only a small number of hidden variables are needed to efficiently model and capture the sub-structure in the observation sequence. We also found that assuming each class label is associated with a disjoint set of hidden states largely simplifies the model, thus leading to efficient training and inference. Our experiments on three important sequence labeling problems: part-of-speech tagging, noun phrase chunking, and named entity recognition showed that HDCRF model outperforms the state-of-the-art discriminative CRF model, demonstrating the effectiveness of our model.

References

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 21(4):543–565.

Davis, J., and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of ICML-06*, 233–240.

Dean, T., and Kanazawa, K. 1989. A model for reasoning about persistence and causation. *Computational Intelligence* 5(3):142–150.

Gunawardana, A.; Mahajan, M.; Acero, A.; and Platt, J. C. 2005. Hidden conditional random fields for phone classification. In *Proceedings of INTERSPEECH-05*, 1117–1120.

Koo, T., and Collins, M. 2005. Hidden-variable models for discriminative reranking. In *Proceedings of HLT/EMNLP-05*, 507–514.

Kristjansson, T. T.; Culotta, A.; Viola, P.; and McCallum, A. 2004. Interactive information extraction with constrained conditional random fields. In *Proceedings of AAAI-04*, 412–418.

Kunsch, H.; Geman, S.; and Kehagias, A. 1995. Hidden Markov random fields. *Annals of Appied Probability* 5(3):577–602.

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01*, 282–289.

Liu, D. C., and Nocedal, J. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45:503–528.

McCallum, A.; Freitag, D.; and Pereira, F. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of ICML-2000*, 591–598.

Murphy, K. P. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. Dissertation, University of California at Berkley.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Peng, F., and McCallum, A. 2004. Accurate information extraction from research papers using conditional random fields. In *Proceedings of HLT-NAACL 2004*, 329–336.

Pinto, D.; McCallum, A.; Wei, X.; and Croft, W. B. 2003. Table extraction using conditional random fields. In *Proceedings of ACM SIGIR-03*, 235–242.

Sang, E. T. K., and Buchholz, S. 2000. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of CoNLL-2000*, 127–132.

Sha, F., and Pereira, F. 2003. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL 2003*, 213–220.

Sutton, C., and McCallum, A. 2006. An introduction to conditional random fields for relational learning. In Getoor, L., and Taskar, B., eds., *Introduction to Statistical Relational Learning*. MIT Press.

Sutton, C.; Rohanimanesh, K.; and McCallum, A. 2004. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *Proceedings of ICML-04*.

Zhu, J.; Nie, Z.; Wen, J.-R.; Zhang, B.; and Ma, W.-Y. 2005. 2D conditional random fields for Web information extraction. In *Proceedings of ICML-05*, 1044–1051.