2013 12th International Conference on Machine Learning and Applications

Empirical Co-occurrence Rate Networks For Sequence Labeling

Zhemin Zhu Djoerd Hiemstra Peter Apers Andreas Wombacher
CTIT Dabatase Group, Computer Science Department
University of Twente, The Netherlands
{z.zhu, d.hiemstra, p.m.g.apers, a.wombacher}@utwente.nl

***Abstract*—Structured prediction has wide applications in many areas. Powerful and popular models for structured prediction have been developed. Despite the successes, they suffer from some known problems: (i) Hidden Markov models are generative models which suffer from the mismatch problem. Also it is difficult to incorporate overlapping, non-independent features into a hidden Markov model explicitly. (ii) Conditional Markov models suffer from the label bias problem. (iii) Conditional Random Fields (CRFs) overcome the label bias problem by global normalization. But the global normalization of CRFs can be expensive which prevents CRFs from applying to big data.**

In this paper, we propose the Empirical Co-occurrence Rate Networks (ECRNs) for sequence labeling. ECRNs are discriminative models, so ECRNs overcome the problems of HMMs. ECRNs are also immune to the label bias problem even though they are locally normalized. To make the estimation of ECRNs as fast as possible, we simply use the empirical distributions as the estimation of parameters. Experiments on two real-world NLP tasks show that ECRNs reduce the training time radically while obtain competitive accuracy to the state-of-the-art models.

I. Introduction

Structured prediction has many important applications in natural language processing [1], computer vision [2], [3], bioinformatics [4], [5] and other areas. For example, in natural language processing, part-of-speech (POS) tagging [6] is a typical structured prediction task. The input of a POS tagger is a sentence which is treated as a sequence of words and the output is a sequence of POS tags assigned to each word in the sentence. Named entity recognition (NER) [7] is another important application in information extraction which transforms a sequence of words into a sequence of NER tags which identify people, organizations, locations or other named entities. In other applications, the structure of outputs can be more complex than sequences, e.g., for a syntactic parser, the output is a parse tree which is tree-structured.

Structured prediction attracts a lot of research interests and has been studied extensively in NLP and other areas. Many powerful models, such as Hidden Markov Models (HMMs) [8], Conditional Markov Models (CMMs) [9], [10], and Conditional Random Fields (CRFs) [11], have been proposed. They are widely applied to practical applications in different areas. Despite the great successes achieved by these models, they are not flawless. They suffer from some known problems. Hidden Markov Models are generative models whose objective function at training time is different from the objective function at decoding time. This is known as the mismatch problem [10]. At training time, HMMs optimize a joint probability, but at decoding time they try to find a sequence of tags which maximizes a conditional probability. Also it is difficult to incorporate overlapping, non-dependent features explicitly into a hidden Markov model. Conditional Markov models were proposed to overcome the drawbacks of hidden Markov models. Conditional Markov Models are discriminative models in which the objective functions are consistent with each other at training and decoding time. But conditional Markov models are affected by the label bias problem [11], [12]. Then Conditional Random Fields were proposed, which avoid the label bias problem. But the training of conditional random fields can be very expensive [13].

In this paper, we propose the Empirical Co-occurrence Rate Networks (ECRNs) for predicting structured outputs. ECRNs avoid the problems of the existing models. ECRNs are discriminative models. In a discriminative model, the objective functions are consistent at training and decoding time. And also it is easy to craft overlapping, non-independent features into ECRNs explicitly. We also show that ECRNs avoid the label bias problem naturally even though they are locally normalized. To make the training of ECRNs as fast as possible, we simply use the empirical distributions as the estimation of parameters. This results in very efficient training of ECRNs. Experiments on two real-world datasets show that ECRNs reduce the training time radically while obtain competitive results to the state-of-the-art models.

The rest of this paper is organized as follows. In Section II, we review the existing popular models, such as HMMs, CMMs and CRFs. We also illustrate their known problems. Section III is devoted to our model: Empirical Co-occurrence Rate Networks (ECRNs). In Section IV, we prove ECRNs do not suffer from the label bias problem by experiments on a simulated dataset. Also in this section we describe experiments on two real-world datasets. Conclusions and future work follow in the last two sections.

II. Related Work

In this section, we review some popular models (HMMs, CMMs, CRFs) for structured prediction. In Section III, we will introduce our model ECRNs. These models differ in some important characteristics, such as conditioning (generative or discriminative), graph structure (directed or undirected) and factorization. Table I summarizes the important characteristics of these models.

A. Hidden Markov Models

978-0-7695-5144-9/13 $31.00 © 2013 IEEE
DOI 10.1109/ICMLA.2013.23

TABLE I: Characteristics of Models

	HMM	CMM	CRF	ECRN
Conditioning[a]	Gen.	Dis.	Dis.	Dis.
Normalization[b]	Loc.	Loc.	Glo.	Loc.
Training[c]	Fast	Fast	Slow	Fast
Directionality[d]	Dir.	Dir.	Un.	Un.
LBP[e]	No	Yes	No	No

[a] generative or discriminative
[b] local or global
[c] whether training is fast or slow
[d] directed or undirected
[e] whether affected by the label bias problem

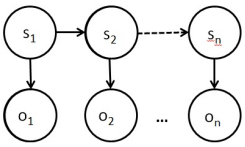

Fig. 1: Hidden Markov Models

1) Graph Structure: Figure 1 shows a first order HMM, where $S = [s_1, s_2, ..., s_n]$ is the state sequence (or tag sequence) and $O = [o_1, o_2, ..., o_n]$ is the observation sequence (or word sequence). For example, in part-of-speech tagging, S are the POS tags to be predicted and O are the words in a sentence. In graphical models, the graph structure encodes independence relations between nodes. Based on these independence relations, we can factorize a joint probability into small factors.

2) Factorization: The factorization of directed models is based on the mathematical concept of conditional probability. HMMs are generative models which factorize a joint probability as follows:

$$p(S, O) = p(s_1) \prod_{i=1}^{n} p(o_i|s_i) \prod_{j=1}^{n-1} p(s_{j+1}|s_j). \quad (1)$$

3) Known Issues: There are two known drawbacks of HMMs [13]. The first drawback is the mismatch problem which stems from their generative nature. At training time, HMMs optimize a joint probability $p(S, O)$, but at decoding time we want a tag sequence which maximizes the conditional probability $p(S|O)$. As $P(S, O) = p(S|O)p(O)$, where $p(O)$ is the distribution of observations, HMMs just pay unnecessary efforts to model $p(O)$. Klein et al. [14] show models with consistent objective functions at training and decoding time perform better than mismatch models. The second drawback is HMMs have difficulty to incorporate overlapping, non-independent features explicitly [10]. This can be observed from the factors $p(o_i|s_i)$ in Equation 1. o_i is assumed independent of other observations conditioned by s_i.

B. Conditional Markov Models

1) Graph Structure: Figure 2 shows a first order CMM. Maximum entropy markov models (MEMMs) [10] are typical CMMs which train the model using the maximum entropy framework.

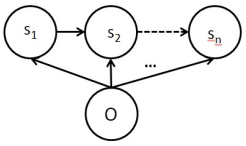

Fig. 2: Conditional Markov Models

2) Factorization: CMMs are discriminative models which factorize a conditional joint probability as follows:

$$p(S|O) = p(s_1|O) \prod_{i=1}^{n-1} p(s_{i+1}|s_i, O). \quad (2)$$

3) The Label Bias Problem: Due to their discriminative nature, CMMs do not suffer from the mismatch problem of HMMs and they can easily craft overlapping, non-independence features explicitly. But CMMs are affected by the Label Bias Problem [6], [11], [12], [15] which stems from the nature of their factorization. The factors $p(s_{i+1}|s_i, O)$ are *local conditional probabilities* with respect to s. These local conditional probabilities prefer the s_i with fewer outgoing transitions. The extreme case is when s_i has only one possible outgoing transition s_{i+1}, then $p(s_{i+1}|s_i, O)$ is always 1 no matter what o_{i+1} is. That is o_{i+1} is not used for predicting s_{i+1}. We use the following example to illustrate this problem. Suppose the training dataset consists of 21 training instances including 11 of $(\texttt{rib} : \texttt{XIB})$, 9 of $(\texttt{rob} : \texttt{YOB})$ and 1 of $(\texttt{rob} : \texttt{XIB})$, where $\{\texttt{r}, \texttt{o}, \texttt{i}, \texttt{b}\}$ are observations and $\{\texttt{X}, \texttt{Y}, \texttt{O}, \texttt{I}, \texttt{B}\}$ are tags. At test time, we want to predict the tags for the observation sequence $(\texttt{rob})$. Obviously, the correct tags for $(\texttt{rob})$ should be $(\texttt{YOB})$ rather than $(\texttt{XIB})$. Because there are 9 of $(\texttt{rob} : \texttt{YOB})$ and only 1 of $(\texttt{rob} : \texttt{XIB})$. But according to Equation 2, $p(\texttt{YOB}|\texttt{rob}) = p(\texttt{Y}|\texttt{r})p(\texttt{O}|\texttt{Y}, \texttt{o})p(\texttt{B}|\texttt{O}, \texttt{b}) = \frac{9}{21} \times 1 \times 1 = \frac{9}{21}$, which is smaller than $p(\texttt{XIB}|\texttt{rob}) = p(\texttt{X}|\texttt{r})p(\texttt{I}|\texttt{X}, \texttt{o})p(\texttt{B}|\texttt{I}, \texttt{b}) = \frac{12}{21} \times 1 \times 1 = \frac{12}{21}$. So CMMs will mislabel $(\texttt{rob})$ as $(\texttt{XIB})$. The reason is that $(\texttt{X})$ has only one outgoing transition $(\texttt{I})$. This constrains $p(\texttt{I}|\texttt{X}, \texttt{o})$ to be 1 even though $p(\texttt{I}|\texttt{o})$ is very small. That is $(\texttt{o})$ is not used for prediction its tag. The tag of $(\texttt{o})$ totally depends on the previous tag. From this example, we can see that the local conditional probabilities $p(s_{i+1}|s_i, O)$ cause the label bias problem.

C. Conditional Random Fields

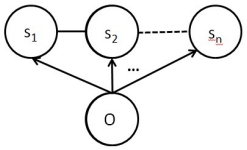

Fig. 3: Conditional Random Fields

1) Graph Structure: Figure 3 shows a linear-chain CRF. CRFs [11] are discriminative and undirected graphical models.

2) Factorization: The factorization for undirected models are based on the Hemmersley-Clifford Theorem. According to this theorem, a linear-chain CRF can be factorized as follows:

$$p(S|O) = \frac{1}{Z(O)} \prod_{i=1}^{n-1} \phi(s_i, s_{i+1}, O) \prod_{j=1}^{n} \psi(s_j, O),$$

where $Z(O)$ is the global normalization which ensures $\sum_S p(S|O) = 1$. ϕ and ψ are non-negative factors defined over pairwise and unary cliques, respectively. Unlike local models, such as HMM and CMM whose factors are probabilities, the factors of CRFs, ϕ and ψ, have no probabilistic interpretations[1]. So they cannot be locally normalized.

3) Known Issues: The global normalization makes CRFs unaffected by the Label Bias Problem. A known problem of CRFs is the training of CRFs for large-scale datasets can be slow. On linear-chain CRFs, the time complexity of the standard training method for CRFs is quadratic in the size of the tag set, linear in the number of features and almost quadratic in the size of the training sample [13], [16]. Approximate techniques [17]–[19] have been proposed for reducing the training time, but they are not guaranteed to converge or still take considerable time. Also advanced optimization techniques, such as stochastic gradient descent [20] and average perception [21], have been applied to accelerate convergence rate. Normally, they reduce the number of iterations, but they cannot reduce the time complexity of one iteration. Also sometimes they oscillate when getting close to the optimum and we need to pre-set the maximum number of iterations to stop the iterative process.

III. Empirical Co-occurrence Rate Networks

A. Graph Structure

ECRNs are discriminative, undirected models. A linear-chain ECRN has the same graph structure as Figure 3. But the factorization of ECRNs is different from that of CRFs.

B. Co-occurrence Rate Factorization

Co-occurrence Rate (CR) is the exponential function of Pointwise Mutual Information (PMI) [22]. PMI was first introduced into NLP community by Church and Hanks [23]. It instantiates Mutual Information [24] to specific events and was originally defined between two variables which can be extended to multiple variables [25], [26]. Copula is a similar concept in statistics [27]. To the best of our knowledge, we are the first to apply CR to factorize undirected graphs.

Definition 1 (CR and Conditional CR)**.**

$$\mathtt{CR}(X_1; ...; X_n) = \frac{p(X_1, ..., X_n)}{p(X_1)...p(X_n)},$$
$$\mathtt{CR}(X_1; ...; X_n \mid Y) = \frac{p(X_1, ..., X_n \mid Y)}{p(X_1 \mid Y)...p(X_n \mid Y)}.$$

According to this definition, if there is only one variable, then $\mathtt{CR}(X) = p(X)/p(X) = 1$. For convenience, $\mathtt{CR}(\emptyset) = 1$. CR is a proper quantity for measuring *compatibility*: (i) If $0 \leq \mathtt{CR} < 1$, events occur *repulsively*; (ii) If $\mathtt{CR} = 1$, events occur *independently*; (iii) If $\mathtt{CR} > 1$, events occur *attractively*. We distinguish the following two notations:

$$\mathtt{CR}(X_1; X_2; X_3) = \frac{p(X_1, X_2, X_3)}{p(X_1)p(X_2)p(X_3)},$$
$$\mathtt{CR}(X_1; X_2X_3) = \frac{p(X_1, X_2, X_3)}{p(X_1)p(X_2, X_3)}.$$

The first denotes CR between three random variables: X_1, X_2 and X_3. By contrast, the second denotes CR between two random variables: X_1 and a joint variable X_2X_3. We have the following two simple but very useful theorems.

Theorem 1 (Partition Theorem)**.**

$$\mathtt{CR}(X_1; ..; X_k; X_{k+1}; ..; X_n)$$
$$= \mathtt{CR}(X_1; ..; X_k)\mathtt{CR}(X_{k+1}; ..; X_n)\mathtt{CR}(X_1..X_k; X_{k+1}..X_n).$$

Proof.

$$\begin{aligned}
&\mathtt{CR}(X_1; ..; X_k)\mathtt{CR}(X_{k+1}; ..; X_n)\mathtt{CR}(X_1..X_k; X_{k+1}..X_n) \\
&= \frac{p(X_1, .., X_k)}{\prod_{i=1}^{k} p(X_i)} \frac{p(X_{k+1}, .., X_n)}{\prod_{j=k+1}^{n} p(X_j)} \\
&\quad \frac{p(X_1, .., X_k, X_{k+1}, .., X_n)}{p(X_1, .., X_k)p(X_{k+1}, .., X_n)} \\
&= \frac{p(X_1, .., X_k, X_{k+1}, .., X_n)}{\prod_{i=1}^{n} p(X_i)} \\
&= \mathtt{CR}(X_1; ..; X_k; X_{k+1}; ..; X_n). \qquad \square
\end{aligned}$$

Theorem 2 (Conditional Independence Theorem)**.** *If $X \perp Y \mid Z$, then we have* $\mathtt{CR}(X; YZ) = \mathtt{CR}(X; Z)$.

$X \perp Y \mid Z$ means X is independent of Y conditioned by Z.

Proof: $X \perp Y \mid Z \Rightarrow p(X, Y|Z) = p(X|Z)p(Y|Z)$.

$$\begin{aligned}
p(X, Y, Z) &= p(Z)p(X, Y|Z) = p(Z)p(X|Z)p(Y|Z) \\
&= p(Z)\frac{p(X, Z)}{p(Z)}\frac{p(Y, Z)}{p(Z)} = p(X, Z)p(Y, Z)/p(Z).
\end{aligned}$$

So $\mathtt{CR}(X; YZ) = \frac{p(X,Y,Z)}{p(X)p(Y,Z)} = \mathtt{CR}(X; Z)$. ■

It is easy to prove that Theorem 1 and Theorem 2 also apply to Conditional CR. There are more nice theorems of CR [28]. Even we can obtain the factors of the Hemmersley-Clifford Theorem and Junction Tree by CR (Section 4 and 5 in [28]).

With Theorem 1 and Theorem 2, the linear-chain undirected graph in Figure 3 can be factorized as:

$$p(s_1, .., s_n|O) = \prod_{j=2}^{n} \mathtt{CR}(s_{j-1}; s_j|O) \prod_{i=1}^{n} p(s_i|O). \qquad (3)$$

This factorization can be obtained by CR as follows:

[1]Sometimes they are intuitively explained as the compatibility between nodes in cliques. But the notion compatibility has no formal definition.

$$p(s_1,..,s_n|O) = \texttt{CR}(s_1;..;s_n|O)\prod_{i=1}^{n} p(s_i|O) \quad (4)$$

$$= \texttt{CR}(s_1|O)\texttt{CR}(s_1;s_2..s_n|O)\texttt{CR}(s_2;..;s_n|O)\prod_{i=1}^{n} p(s_i|O) \quad (5)$$

$$= \texttt{CR}(s_1;s_2|O)\texttt{CR}(s_2;..;s_n|O)\prod_{i=1}^{n} p(s_i|O) \quad (6)$$

...

$$= \prod_{j=2}^{n} \texttt{CR}(s_{j-1};s_j|O)\prod_{i=1}^{n} p(s_i|O).$$

Equation 4 is obtained by Definition 1. Equation 5 is based on Theorem 1. We obtain Equation 6 because $\texttt{CR}(s_1|O) = 1$ and $s_1 \perp s_2..s_n \mid s_2$. Following Theorem 2, we have $\texttt{CR}(s_1;s_2..s_n|O) = \texttt{CR}(s_1;s_2|O)$. By repeating this process we can obtain the final result.

CR, e.g. $\texttt{CR}(x;y) = \frac{p(x,y)}{p(x)p(y)}$, is a symmetric concept which fits the symmetric nature of undirected graphs. By contrast, as we know the factorization of directed graphs (Bayesian networks) are based on the conditional probability. Conditional probability, e.g. $p(x|y) = \frac{p(x,y)}{p(y)}$, is an asymmetric concept which fits the asymmetric nature of directed graphs.

C. Unaffected By The Label Bias Problem

ECRNs avoid the label bias problem due to the nature of their factorization. In Section II-B3, we can see that the label bias problem stems from the *local conditional probabilities* $p(s|s',O)$, where s' is the previous tag and s is the current tag. But in the factorization of ECRNs (Equation 3), there is no such local conditional probabilities. The factors in Equation 3 are *local joint probabilities* $\texttt{CR}(s';s|O) = \frac{p(s',s|O)}{p(s'|O)p(s|O)}$ and unary probabilities $p(s|O)$. In a local joint probability $p(s',s|O)$, both o' and o can be used for predicting s. If $p(s|o)$ is very small, $p(s',s|O)$ is also very small. That is s' cannot dominate the prediction of s. The current observation o also matters. So ECRns avoid the label bias problem naturally. We further check this by the example given in Section II-B3. According to Equation 3, $p(\texttt{YOB}|\texttt{rob}) = \texttt{CR}(\texttt{Y};\texttt{O}|\texttt{ro})\texttt{CR}(\texttt{O};\texttt{B}|\texttt{ob})p(\texttt{Y}|\texttt{r})p(\texttt{O}|\texttt{o})p(\texttt{B}|\texttt{b}) = \frac{9/10}{9/21\times 9/10} \times \frac{9/10}{9/10\times 1} \times \frac{9}{21} \times \frac{9}{10} \times 1 = \frac{9}{10}$, which is bigger than $p(\texttt{XIB}|\texttt{rob}) = \texttt{CR}(\texttt{X};\texttt{I}|\texttt{ro})\texttt{CR}(\texttt{I};\texttt{B}|\texttt{ob})p(\texttt{X}|\texttt{r})p(\texttt{I}|\texttt{o})p(\texttt{B}|\texttt{b}) = \frac{1/10}{12/21\times 1/10} \times \frac{1/10}{1/10\times 1} \times \frac{12}{21} \times \frac{1}{10} \times 1 = \frac{1}{10}$. So ECRNs will make the correct prediction YOB for rob. This is confirmed by experiments in Section IV-A.

D. Parameter Estimation

Traditionally, we can use the Maximum Entropy framework (MaxEnt) [10] to train the parameters of a graphical model. In MaxEnt, the probability is given by an exponential function of features. The constraints of MaxEnt require the expected value of each feature in the estimated distribution be equal to the empirical values in the training dataset. This is equivalent to maximizing some log likelihood function. The maximization can be done by optimization algorithms, such as Limited-memory BFGS.

To make the parameter estimation as fast as possible, also there are some very interesting challenges of applying MaxEnt to CRNs (Section III-D1), we leave MaxEnt training of CRNs as future work. Instead, we simply use the empirical distributions to estimate the parameters. From the factorization of ECRNs (Equation 3), we can see that we need to estimate two kinds of parameters: the unary probability $p(s_i|O)$ and the $\texttt{CR}(s_{j-1};s_j|O)$. Since $\texttt{CR}(s_{j-1};s_j|O)$ can be calculated through the unary and pairwise probabilities as $\texttt{CR}(s_{j-1};s_j|O) = \frac{p(s_{j-1},s_j|O)}{p(s_{j-1}|O)p(s_{j-1}|O)}$, we can estimate the unary $p(s_i|O)$ and pairwise probabilities $p(s_{j-1},s_j|O)$ instead of $\texttt{CR}(s_{j-1};s_j|O)$. We simply estimate these probabilities directly by the frequencies of patterns in the training dataset as follows:

$$\tilde{p}(s_i|O) = \frac{\#(s_i,o_i)}{\sum_{s_i}\#(s_i,o_i)},$$

$$\tilde{p}(s_i,s_{i+1}|O) = \frac{\#(s_i,s_{i+1},o_i,o_{i+1})}{\sum_{s_is_{i+1}}\#(s_i,s_{i+1},o_i,o_{i+1})},$$

where $\#(s_i,o_i)$ is the times (or frequency) of the pattern (s_i,o_i) appears in the training dataset.

At decoding time, we may encounter o_i or o_{i+1} which is a unknown word[2]. The formulas above cannot be applied to unknown words, because the denominator is equal to 0 due to $\#(s,o_i)$ is 0 for any unknown word. In this case, we simply use the feature $f(o_i)$ to replace the o_i itself in the patterns $(s_i,f(o_i))$ as follows.

$$\tilde{p}(s_i|O) = \frac{\#(s_i,f(o_i))}{\sum_{s_i}\#(s_i,f(o_i))},$$

$$\tilde{p}(s_i,s_{i+1}|O) = \frac{\#(s_i,s_{i+1},f(o_i),f(o_{i+1}))}{\sum_{s_is_{i+1}}\#(s_i,s_{i+1},f(o_i),f(o_{i+1}))},$$

where o_i and o_{i+1} are unknown words. Since the pattern $(s_i,f(o_i))$ has been seen in the training data, the denominator cannot be equal to 0. It is important that when we calculate CR using pairwise and unary probabilities, the same features should be used to estimate these three probabilities. That is in $\texttt{CR}(s';s|O) = \frac{p(s',s|O)}{p(s'|O)p(s|O)}$, the three factors on right hand side are conditioned by the same features. In other words, CR should be treated as a whole. Otherwise, the accuracy decreases.

1) Challenges of MaxEnt Training of CRNs: In Equation 3, the unary probabilities can be normalized locally, but unfortunately the CR factors cannot[3]. The normalization conditions (or log-partition function in log form) play critical role in estimation [29]. They are the constraints to compute the moments of the distribution. Without local normalizations, CR factors can not be directly estimated. A promising way to obtain CR estimations is to train the pairwise and unary probabilities in a CR factors separately using MaxEnt at training time, and at decoding time CR can be calculated by pairwise and unary probabilities.

Another option may be to transform Equation 3 into the following form and maximize the joint probability:

$$p(s_1,..,s_n|O) = \frac{\prod_{j=2}^{n} p(s_{j-1},s_j|O)}{\prod_{i=2}^{n-1} p(s_i|O)}.$$

[2]Unknown words are the words which do not appear in the training dataset.

[3]Not really cannot. It is very interesting to use the empirical $\sum_{x,y}\texttt{CR}(x;y|O)$ as the normalization of $\texttt{CR}(x;y|O)$ in future.

Unfortunately, even the unary and pairwise factors can be locally normalized, in our preliminary experiment, this method did not work. We guess the reason is $p(s', s|O)$ and $p(s'|O)$ or $p(s|O)$ are not independent factors (this is obvious) and the objective function seems not convex. If we maximize the joint probability, $p(s', s|O)$ is maximized but $p(s'|O)$ or $p(s|O)$ is minimized. Then the estimated moments of these distributions deviate far from the empirical moments and this method failed. The failure of this method tells us to treat the CR factor as a whole is important because CR factors are independent of unary probabilities. This will be explored further in future.

The Decoder of ECRNs can be efficiently implemented using the traditional Viterbi algorithm.

IV. Experiments

We adopt MALLET version 0.4 [30] as the implementation of MEMM, CRF++ version 0.57 [31] as the implementation of standard training of CRFs. ECRNs were implemented by us in Java. For piecewise (**PW**) [19] training of CRFs, we adopt the MALLET version 2.0 as the implementation.

A. The Label Bias Problem

We test LBP on simulated data following [11]. The simulated data were generate as follows. There are five types of tags: $\{R1, R2, I, O, B\}$ and four types of observations: $\{r, i, o, b\}$. The designated observation for both $R1$ and $R2$ is r, for I it is i, for O it is o and for B it is b. We generate the paired sequences from two tag sequences: $[R1, I, B]$ and $[R2, O, B]$. Each tag emits the designated observation with probability of $29/32$ and each of other three observations with probability $1/32$. For training, we generate 1000 pairs for each tag sequence, so totally the size of training dataset is 2000. For testing, we generate 250 pairs for each tag sequence, so totally the size of testing dataset is 500. We run the experiment for 10 rounds and report the average per-token accuracy ($\frac{\#CorrectTags}{\#AllTags}$) in Table II.

TABLE II: Accuracy For Label Bias Problem

ECRN	CRF++	PW	MEMMs
95.8	95.9	96.0	66.6

The experimental results show that ECRN, piecewise training (PW) and the standard training (CRF++) of CRFs are all unaffected by the label bias problem. But MEMMs suffer from this problem because its accuracy is significantly lower than other models. This experiments confirm our discussions in Section III-C and Section II-B3.

B. Part-of-speech Tagging Experiment

We use the Brown Corpus [32] for part-of-speech tagging experiments. The raw data were preprocessed by excluding the incomplete sentences which are not ending with a punctuation. This results in 34,623 sentences. The size of the tag space is 252. Following [11], we introduce features and parameters for each tag-word pair and tag-tag pair. We also use the same spelling features as those used by [11]:

1) (f_1) Whether a token begins with a number or upper case letter.
2) (f_2) Whether a token contains a hyphen.
3) (f_3) Whether a token ends in one of the following suffixes: -ing, -ogy, -ed, -s, -ly, -ion, -tion, -ity, -ies.

TABLE III: Accuracy On POS Tagging

	Overall	Known	Unknown	Time (Sec.)
CRF++	95.4	96.1	**71.7**	4,571,807
ECRN	**95.6**	**96.9**	70.5	**3.9**

Table III lists the per-token accuracy obtained by CRF++ and ECRN on the part-of-speech tagging dataset. The overall accuracy gives the per-token accuracy obtained by models covering all words including known and unknown words. Known and unknown accuracy show the per-token accuracy only considering known or unknown words. From the experimental results, ECRN is much faster than the traditional training (CRF++) of CRFs, and achieves better results on known words. Results also show that CRF++ outperforms ECRN on unknown words by 1.2 percent. On the overall accuracy, ECRN and CRF++ perform almost the same well. Because we simply use the empirical distribution to estimate the parameters. The training of ECRN is just by counting the frequencies in the training data without iterative optimization.

C. Named Entity Recognition

We use the the Dutch part of CoNLL-2002 NER Corpus[4] as our experimental dataset. There are three files in this corpus: ned.train (13,221) for training, ned.testa (2,305) for development and ned.testb (4,211) for testing.

The size of the tag space is 9. We use the same features as those described in the part-of-speech tagging experiment. The results are listed in Table IV.

TABLE IV: Accuracy On NER

	Overall	Known	Unknown	Time (Sec.)
CRF++	96.13	98.2	**77.4**	794
ECRN	**96.23**	**98.8**	73.7	**1.3**

On this dataset, ECRN obtains better results on overall and known word accuracy. But on unknown words, CRF++ perfomes significantly better than ECRN. This is the trade-off between speed and accuracy. The results also show ECRN is much faster than CRF++. The experimental results on NER dataset are consistent with the result on POS tagging dataset.

V. Conclusion

The existing models for structured prediction have their own drawbacks. We proposed the empirical Co-occurrence Rate Networks (ECRNs) for predicting structured outputs. ECRNs avoid the problems with the existing models. ECRNs are discriminative, local models which are unaffected by the label bias problem. ECRNs can be trained very fast by simply using the empirical distribution to estimate the parameters. Experiments on two real-world NLP datasets show that ECRNs speeds up the training radically and obtains competitive results to state-of-the-art models. ECRNs can be very useful for practitioners on big data.

[4] http://www.cnts.ua.ac.be/conll2002/ner/

VI. Future Work

As discussed in Section III-D1, MaxEnt training of CRNs is very interesting. Also in this paper, we did not try global features for training ECRN. Even Equation 3 shows CRNs can be conditioned by global features (the big O), we still need experimental evidence to support this. More comprehensive comparisons between models will be done including statistical tests. Moreover, we focus on linear-chain graphs in this paper. CR factorization can also be applied to tree-structured and cyclic graphs [28]. We will explore this direction. We will also study other important models for structured prediction, such as structured SVMs [33], [34] which minimize large-margin risks and apply factorization to kernel representations (kernel decomposition), exemplar-based methods [35], constrained conditional models [36] and so on. It is very interesting to apply CR factorization to kernel representations and minimize large-margin risks.

Acknowledgment

We thank the four reviewers of ICMLA2013 for their inspiring and helpful comments. This work has been supported by the Dutch national program COMMIT/.

References

[1] N. A. Smith, *Linguistic Structure Prediction*, G. Hirst, Ed. Morgan & Claypool Synthesis Lectures on Human Language Technologies, 2011, vol. 13.

[2] S. Kumar and M. Hebert, "Discriminative fields for modeling spatial dependencies in natural images," in *NIPS'04*, S. Thrun, L. Saul, and B. Schölkopf, Eds.

[3] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *NIPS*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 1097–1104.

[4] K. Sato and Y. Sakakibara, "Rna secondary structural alignment with conditional random fields." *Bioinformatics*, vol. 21:ii, pp. 237–242, 2005.

[5] Y. Liu, J. Carbonell, P. Weigele, and V. Gopalakrishnan, "Protein fold recognition using segmentation conditional random fields (scrfs)," *Journal of Computational Biology*, vol. 13(2), pp. 394–406, 2006.

[6] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *NAACL '03*, 2003, pp. 173–180.

[7] R. Grishman and B. Sundheim, "Message understanding conference-6: a brief history," in *Proceedings of the 16th conference on Computational linguistics - Volume 1*, ser. COLING '96. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996, pp. 466–471. [Online]. Available: http://dx.doi.org/10.3115/992628.992709

[8] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, pp. 257–286.

[9] K. C. Eric Brill, "A maximum entropy model for part-of-speech tagging," in *EMNLP96*, 1996, pp. 133–142.

[10] A. McCallum, D. Freitag, and F. C. N. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *ICML '00*, 2000, pp. 591–598.

[11] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[12] L. Bottou, "Une approche theorique de lapprentissage connexionniste: Applications 'a la reconnaissance de la parole." Ph.D. dissertation, Universite de Paris XI., 1991.

[13] T. A. Cohn, "Scaling conditional random fields for natural language processing," Ph.D. dissertation, 2007.

[14] D. Klein and C. D. Manning, "Conditional structure versus conditional estimation in nlp models," in *EMNLP '02*, 2002, pp. 9–16.

[15] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*, 1st ed. Springer Publishing Company, Incorporated, 1999.

[16] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. 4(4), pp. 267–373, 2012.

[17] M. J. Wainwright, T. Jaakkola, and A. S. Willsky, "Tree-reweighted belief propagation and approximate ML estimation by pseudo-moment matching," in *9th Workshop on Artificial Intelligence and Statistics*, January 2003.

[18] C. Sutton and A. McCallum, "Piecewise pseudolikelihood for efficient training of conditional random fields," in *Proceedings of the 24th international conference on Machine learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 863–870. [Online]. Available: http://doi.acm.org/10.1145/1273496.1273605

[19] C. A. Sutton and A. McCallum, "Piecewise training for undirected models," in *UAI 05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, July 26-29 2005, Edinburgh, Scotland*. AUAI Press, 2005, pp. 568–575.

[20] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *ICML '06*, 2006, pp. 969–976.

[21] M. Collins, "Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms," in *EMNLP '02*, 2002, pp. 1–8.

[22] R. Fano, *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA: The MIT Press, 1961.

[23] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguist.*, vol. 16, no. 1, pp. 22–29, Mar. 1990. [Online]. Available: http://dl.acm.org/citation.cfm?id=89086.89095

[24] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, pp. 379–423, 1948.

[25] Z. Zhu, "Factorizing probabilistic graphical models using cooccurrence rate," in *arXiv:1008.1566v1*, August 2010.

[26] T. Van de Cruys, "Two multivariate generalizations of pointwise mutual information," in *Proceedings of the Workshop on Distributional Semantics and Compositionality*, ser. DiSCo '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 16–20. [Online]. Available: http://dl.acm.org/citation.cfm?id=2043121.2043124

[27] G. Elidan, "Copula bayesian networks," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010, pp. 559–567.

[28] Z. Zhu, "Factorizing probabilistic graphical models using co-occurrence rate," Centre for Telematics and Information Technology, University of Twente, Enschede, Technical Report TR-CTIT-12-30, May 2011. [Online]. Available: http://eprints.eemcs.utwente.nl/22603/

[29] S. L. La, *Graphical Models*. Oxford University, 1996.

[30] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://www.cs.umass.edu/ mccallum/mallet.

[31] T. Kudo, "Crf++: Yet another crf toolkit," March 2012.

[32] W. N. Francis and H. Kucera, "Brown corpus manual," Tech. Rep., 1979.

[33] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *International Conference on Machine Learning (ICML)*, 2004, pp. 104–112.

[34] B. Taskar, C. Guestrin, and D. Koller, "Max-margin markov networks," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.

[35] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch, "Timbl: Tilburg memory-based learner version 6.3 reference guide," Tilburg University, Tech. Rep. ILK Technical Report ILK 10-01, May 2010.

[36] M. Chang, L. Ratinov, and D. Roth, "Structured learning with constrained conditional models," *Machine Learning*, vol. 88, no. 3, pp. 399–431, 6 2012. [Online]. Available: http://cogcomp.cs.illinois.edu/papers/ChangRaRo12.pdf