A New Objective Function for Sequence Labeling

Yuta Tsuboi Hisashi Kashima
IBM Research, Tokyo Research Laboratory
$\{$*yutat,hkashima*$\}$*@jp.ibm.com*

Abstract

We propose a new loss function for discriminative learning of Markov random fields, which is an intermediate loss function between the sequential loss and the pointwise loss. We show this loss function has "Markov property", that is, the importance of correct labeling for a particular position depends on the numbers of the correct labels around there. This property works to keep local consistencies among the assigned labels, and is useful for optimizing systems identifying structural segments, such as information extraction systems.

1 Introduction

The sequence labeling problem is an important generalization of the supervised classification problem, where the labels for a set of target variables are to be predicted when the labels for a set of observed variables are given. Many real-world tasks are formalized as sequence labeling problems in various fields such as natural language processing and bioinformatics. For example, information extraction is one of the most important applications of labeling problems, whose purpose is to identify semantic segments in sequences. The hidden Markov model (HMM) has been successful in the sequence labeling problem for years. Recently, conditional models such as the maximum entropy Markov model (MEMM) [5] and the conditional random field (CRF) [4] (Section 2) have been attracting considerable attentions because of their capabilities to allow overlapping features, and their performances overwhelming that of HMM. Especially, CRF is considered as one of the state-of-the-art labelers.

There are several works on designing and comparing various loss functions (i.e. objective functions) for labeling problems [1, 2, 3]. Two important classes of the loss functions are the sequential loss and the pointwise loss [3]. The sequential loss is the original objective function that maximizes the sum of log-likelihoods, and the pointwise loss maximizes the sum of marginal log-likelihoods with target variables fixed at each position. This indicates that the sequential loss aims to correctly predict the whole target variables in a sequence. On the other hand, the pointwise loss aims to correctly predict each of the target variables as many as possible. When applying the two loss functions to information extraction tasks, the sequential loss has a possibility of resulting in a bad performance in difficult problems with relatively small training data, on the other hand, the pointwise loss is not enough to represent the aim to extract segments adequately.

In Section 3, we propose the *mixed loss*, which is an intermediate loss function between the sequential loss and the pointwise loss defined as a linear combination of the sequential loss and the pointwise loss. We show that the mixed loss has a "Markov property", that is, the importance of correct labeling for a particular position depends on the numbers of the correct labels around there. Therefore, our new loss function is expected to be useful to predict clusters of correct labels.

Section 4 demonstrates that the proposed method is promising by preliminary experiments in natural language processing, and Section 5 concludes this paper.

2 Objective Function for Sequence Labeling

Let $\boldsymbol{x} = (x_1, x_2, \dots, x_T), x_t \in \Sigma_x$ be a set of *observed variables*, and $\boldsymbol{y} = (y_1, y_2, \dots, y_T), y_t \in \Sigma_y$ be a set of *target variables*, where Σ_x and Σ_y are the sets of labels for the observed and the target variables, respectively. Figure 1 is a graph representation of an example in a part-of-speech tagging task, where x_t indicates the t-th word, and y_t indicates the part-of-speech tag for the t-th word.

Given the labels for the observed variables in $\boldsymbol{x}$, we want to assign a correct label to each of the target variables in $\boldsymbol{y}$. For this goal, we may exploit training data, $E = (e^{(1)}, e^{(2)}, \dots, e^{(|E|)})$, whose i-th example is $e^{(i)} = (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$, and $|\boldsymbol{x}^{(i)}| = |\boldsymbol{y}^{(i)}| = T^{(i)}$.

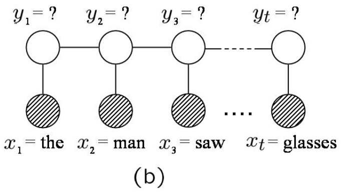

Figure 1. A graph representation of a sequence in part-of-speech tagging tasks. Given x as the sentence "the, man, saw ,$\cdots$, glasses.", y as the part-of-speech tags for the sequence, e.g. "DT, NN, VBD, $\ldots$, NNS", are to be predicted.

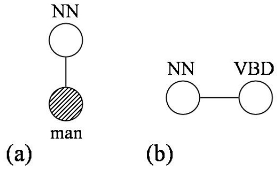

Figure 2. Each feature is defined as a pair of two consecutive variables such as (a) a pair of an observed variable and a target variable, or (b) a pair of two target variables.

The model of CRF is an extension of multi-class logistic regression with multiple target variables,

$$f(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\langle \Theta, \Phi(\boldsymbol{x}, \boldsymbol{y}) \rangle)}{\sum_{\tilde{\boldsymbol{y}}} \exp(\langle \Theta, \Phi(\boldsymbol{x}, \tilde{\boldsymbol{y}}) \rangle)},$$

where Θ is the vector of the model parameters, and $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is the feature vector for $(\boldsymbol{x}, \boldsymbol{y})$. Each element ϕ_i of $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is the number of times the i-th feature appears in $(\boldsymbol{x}, \boldsymbol{y})$. Usually, each feature is defined to be a pair of consecutive two variables such as in Figure 2. One type of such features is a pair of an observed variable and a target variable (Figure 2(a)), and the other is a pair of two target variables (Figure 2(b)). Given the labels for $\boldsymbol{x}$, the labels for $\boldsymbol{y}$ are predicted by $\mathrm{argmax}_{\boldsymbol{y}}\, f(\boldsymbol{y}|\boldsymbol{x})$.

The model is trained by finding the optimal parameters that minimize a loss function. In the original CRF model [4], the sum of negative log-likelihoods is used as the loss function.

Definition 1 (Sequential loss function [2]). *The sequential loss function L_1 is defined as*

$$L_1 \quad = \quad -\sum_i \log f(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}).$$

Let us consider the implication of the sequential loss function L_1. This loss function tries to learn the parameters that predict the labels for the whole target variables in a sequence simultaneously, since the likelihood of the set of target variable $\boldsymbol{y}^{(i)}$ for each example is maximized in this loss. However, there is a possibility of resulting in a bad performance in difficult problems with relatively small training data since a large negative weight is given to the features whose transitions was never observed in the training set. In addition, there are some tasks, e.g. part-of-speech tagging, where it is enough to correctly predict target variables as many as possible.

Based on those ideas, Kakade et al. [3] proposed another loss function L_0, which is based on the marginal likelihood $\Pr(y_t = y_t^{(i)}|\boldsymbol{x}^{(i)})$ of the label $y_t^{(i)}$ at each position t.

Definition 2 (Pointwise loss function [3]). *The pointwise loss function L_0 is defined as*

$$L_0 \quad = \quad -\sum_i \sum_{t=1}^{T^{(i)}} \log \sum_{\tilde{\boldsymbol{y}}:\tilde{y}_t = y_t^{(i)}} f(\tilde{\boldsymbol{y}}|\boldsymbol{x}^{(i)}), \qquad (1)$$

where $\sum_{\boldsymbol{y}:y_t = y_t^{(i)}}$ indicates summation over all possible label assignments for the target variables with the t-th target variable fixed as $y_t^{(i)} \in \Sigma_y$.

The pointwise loss function L_0 aims to correctly predict each of the target variables as many as possible. and does not care the consistencies among consecutive labels. The pointwise loss function is experimentally shown to be competitive to the sequential loss [1, 3].

The above loss functions and their derivatives can be calculated efficiently by using the dynamic programming technique.

3 A New Objective Function with Markov Property

Although each of the objective functions reviewed in the previous section makes sense in each context, we might imagine an intermediate situation where it is desired to correctly predict clusters of variables. For example, in information extraction tasks such as named entity recognition and protein secondary structure prediction, we want to find local segments that indicate named entities, alpha helices or beta sheets regions, and they are represented as clusters of labels.

Therefore, a suitable loss function for information extraction is the one with the characteristics of both L_1 and L_0. In other words, we want a loss function with "Markov property", that is, the importance of correct labeling for a particular position depends on the numbers of the correct labels around there. We define the following new loss function L_λ for this purpose.

Definition 3 (**λ-mixed loss function**). *For a given constant* $0 \leq \lambda \leq 1$, *we call*

$$L_\lambda \quad := \quad \lambda L_1 + (1-\lambda) L_0 \tag{2}$$

the λ-mixed loss function. We can see this loss function lies between the sequential loss and the pointwise loss, since L_λ is identical to L_0 when $\lambda = 0$, and identical to L_1 when $\lambda = 1$.

On first sight, the new objective function does not seem to enhance local consistencies of labels, but we can show that it really does in sequence labeling. Now let us consider another loss function defined as follows.

Definition 4 (**k-th order Markov loss function**). *For a given integer $k > 0$, we call*

$$M_k := -\sum_i \sum_{t=-k+1}^{T^{(i)}} \log \sum_{\tilde{\boldsymbol{y}}:\tilde{\boldsymbol{y}}_t^{t+k}=\boldsymbol{y}^{(i)}{}_t^{t+k}} f(\tilde{\boldsymbol{y}}|\boldsymbol{x}^{(i)}) \tag{3}$$

k-th order Markov loss function. y_t for $t < 1$ and $t > T^{(i)}$ are dummy variables which always take a special label σ_0, i.e. $x_t = y_t = \sigma_0$.

In contrast with the marginal likelihood (1) fixing only one target variable at a time, The k-th order Markov loss function M_k fixes $k+1$ consecutive target variables at a time. Therefore, this loss function tries to correctly predict as many chunks of length $k+1$ as possible.

Now, we obtain the following main theorem that claims equivalence of the mixed loss and the Markov loss.

Theorem 1. *For any integer $k \geq 0$, let*

$$\lambda = \frac{k}{k+1}. \tag{4}$$

Then,

$$\frac{1}{1-\lambda} L_\lambda = M_k.$$

Proof. *By simple algebraic substitutions.* □

This theorem indicates that, for any positive integer $k > 0$, the minimization of $\frac{1}{1-\lambda} L_\lambda$ is equivalent to the minimization of M_k by choosing λ that satisfies (4). Therefore, our new loss function works for correct prediction of labels while keeping local consistencies among them and we can see sequential loss and pointwise loss are the special cases of the Markov loss when $k = \infty$ and $k = 0$, respectively.

In above case, since we inherently assumed that k is integer, corresponding λ can take only discrete values. Then, what if k is not integer, i.e. $\lfloor k \rfloor < k < \lceil k \rceil$? Intuitively, L_λ is just an intermediate loss between $M_{\lfloor k \rfloor}$ and $M_{\lceil k \rceil}$. The following two corollaries are easily derived from Theorem 1. The first one just justifies this intuition, and the other gives another interpretation.

Corollary 1. *For any $k \geq 0$, let $\lambda = k/(k+1)$. Then,*

$$\frac{1}{1-\lambda} L_\lambda = (\lceil k \rceil - k) M_{\lfloor k \rfloor} + (k - \lfloor k \rfloor) M_{\lceil k \rceil}. \tag{5}$$

Note that, since $0 \leq \lceil k \rceil - k, k - \lfloor k \rfloor \leq 1$ and $(\lceil k \rceil - k) + (k - \lfloor k \rfloor) = 1$, L_λ is just an internally dividing point between $M_{\lfloor k \rfloor}$ and $M_{\lceil k \rceil}$.

From another perspective, this can be understood as a weighted sum of Markov losses with exponentially decaying weights.

Corollary 2. *For any $0 < \lambda < 1$,*

$$\frac{1}{1-\lambda} L_\lambda = (1-\lambda) \sum_{\kappa=0}^{\infty} \lambda^\kappa M_\kappa. \tag{6}$$

This weighted sum gives large weights to M_κ with small κ, and the weight decays exponentially as κ becomes larger. λ is the parameter controlling the speed of the decay, and small λ means fast decay. This corollary related a weighted sum of Markov losses to L_λ with a particular $0 < \lambda < 1$, and we can interpret that the mixed loss cares not only for a particular length of chunks, but also all length of chunks by weighting them depending on their lengths.

4 Experiment

We compared the performances of the three objective functions, the sequential loss, the pointwise loss, and the mixed loss for CRF on a Named Entity Recognition task (NER). The NER task is a subtask of information extraction which deals with identifying phrases that contain the names of persons, organizations, locations, times and quantities in sentences. We used the Spanish corpus provided for the shared task of CoNLL2002 on NER [6]. The corpus is composed of a training set, a development set, and a test set, which contain 8,322 (264,680), 1,914 (52,849), and 1,516 (51,487) sentences (tokens), respectively. Each of the tokens in the corpus is annotated with one of the 9 kinds of target labels, i.e. $|\Sigma_y| = 9$. The average phrase length of the named entities is 1.74.

We conducted two types of experiments, an experiment following the standard procedure of the shared task of CoNLL 2002, and an experiment evaluating the performances for various sizes of the training sets. In the first experiment, CRF with each loss function was trained by using the training set. The development data was used for tuning the regularization parameter. In the second experiment, we used the first 100, 200, 300, 600, and 1000 sentences of the training set. Both the development set and the test set were used in evaluation phase. In this experiment, we did not use the regularization term in the objective functions to concentrate on the comparison of the performances of loss functions. In both experiments, we defined the features by following the S3 definition in [2]. The parameters were estimated by using the conjugate gradient descent method.

Table 1 shows the results according to the standard procedure of the shared task of CoNLL 2002. The column "point", "k=l", and "seq" represent the results of pointwise loss function, mixed (l-th Markov) loss function, and sequential loss function, respectively. Under the "Markov loss" interpretation of the proposed loss function, we investigated the performances varying the parameter k from 1 to 4. The performances were evaluated according to P(recision), R(ecall), and F1 measure on the test set. Precision is the percentage of named entities found that are correct. Recall is the percentage of found named entities present in the corpus. F1 measure is the harmonic mean of the precision and recall. The results indicate that the performances of the proposed loss function are competitive with the existing loss functions. Especially, the $k = 3$ Markov loss slightly performs better than others. Since the average length of named entities is 1.74, this results agree with our intuitions since the size of segment plus two represents proper local consistency to recognize the edges of segment.

Table 2 shows the results of the NER task varying training data sizes. We compared the performances of sequential loss, pointwise loss, and mixed losses varying the parameter k from 1 to 4. The performances were evaluated according to the average F1 measure of development set and test set.

In total, the pointwise loss and the mixed loss show higher performance than the sequential loss, though the difference between the pointwise loss and the mixed loss was not stable when the size of training data was varied. The CRFs trained on mixed loss with $k = 3$ and 2 perform better than the others with the smaller training data sets of 100 and 200. With the larger training data sets, the performances of pointwise loss and mixed loss with $k = 1$ show higher performance than the others. This empirical result suggests that the proposed loss function works well for relatively small data sets.

Table 1. Precision, Recall, and F1 measure according to the standard evaluation procedure of CoNLL-2002 NER shared task.

	point	k=1	k=2	k=3	k=4	seq
P	77.91	77.96	77.95	**78.10**	78.03	**78.10**
R	76.71	76.85	76.88	**76.96**	76.85	76.85
F1	77.30	77.40	77.41	**77.53**	77.43	77.47

Table 2. The average F1 measure of NER varying the size of training data set.

Size	point	k=1	k=2	k=3	k=4	seq
100	45.36	46.12	**46.96**	46.94	42.72	43.96
200	47.76	47.39	47.44	**47.77**	47.30	47.16
300	**53.37**	52.91	52.68	52.92	52.86	52.40
600	**59.32**	58.68	58.25	58.11	57.34	56.00
1000	61.26	**61.91**	61.38	61.33	61.34	61.05

5 Conclusion

We proposed a new loss function called the *mixed loss* for information extraction, which is an intermediate loss function between the two loss functions, sequential loss and pointwise loss. We showed its "Markov property", that is, the importance of correct labeling for a particular position depends on the numbers of the correct labels around there.

References

[1] Y. Altun and T. Hofmann. Large margin methods for label sequence learning. In *Proc. EuroSpeech*, 2003.

[2] Y. Altun, M. Johnson, and T. Hofmann. Investigating loss functions and optimization methods for discriminative learning of label sequences. In *Proc. EMNLP*, 2003.

[3] S. Kakade, Y. W. Teh, and S. Roweis. An alternative objective function for Markovian fields. In *Proc. 19th ICML*, 2002.

[4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th ICML*, 2001.

[5] A. McCallum, D.Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *Proc. 17th ICML*, 2000.

[6] E. F. Tjong and K. Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proc. CoNLL*, pages 155–158, 2002.