Information Sciences 179 (2009) 169–179

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Minimum tag error for discriminative training of conditional random fields

Ying Xiong *, Jie Zhu, Hao Huang, Haihua Xu

*Department of Electronic Engineering, Shanghai Jiao Tong University, 1164# Shanghai Jiao Tong University, No. 800 Dong Chuan Road, Min Hang, Shanghai 200240, China*

ARTICLE INFO

*Article history:*
Received 26 May 2008
Received in revised form 17 September 2008
Accepted 18 September 2008

*Keywords:*
Natural language processing
Machine learning
Conditional random fields
Discriminative training
Chinese word segmentation

ABSTRACT

This paper proposes a new criterion called minimum tag error (MTE) for discriminative training of conditional random fields (CRFs). The new criterion, which is a smoothed approximation to the sentence labeling error, aims to maximize an average of transcription tagging accuracies of all possible sentences, weighted by their probabilities. Corpora from the second international Chinese word segmentation bakeoff (Bakeoff 2005) are used to test the effectiveness of this new training criterion. The experimental results have demonstrated that the proposed minimum tag error criterion can reliably improve the initial performance of supervised conditional random fields. In particular, the recall rate of out-of-vocabulary words ($R_{oov}$) is significantly improved compared with that obtained using standard conditional random fields. Furthermore, the new training method has the advantage of robustness to segmentation across all datasets.



## 1. Introduction

Conditional random fields (CRFs) [14] have recently become popular as models for sequence labeling tasks because they offer several advantages over traditional generative models such as hidden Markov models (HMM). Because CRFs are basically defined as conditional models of label sequences given observation sequences, they can make use of flexible overlapping features and overcome label bias problems. In recent years, CRFs have been successfully applied to many tasks, such as gene identification [11], spoken language understanding (SLU) [10], part-of-speech (POS) tagging [14], name entity recognition (NER) [3,17], and shallow parsing [23], especially Chinese word segmentation [6,21].

Unlike English and other Western languages, the Chinese language is based on characters rather than words. There are no blank spaces between words in Chinese sentences. Word segmentation is the first step in Chinese language processing tasks such as information retrieval (IR) [2], text mining (TM) [4,31], question answering (QA) [18], and machine translation (MT) [15,26]. The goal of Chinese word segmentation is to segment a Chinese sentence into a sequence of meaningful words. In early decades, methods for Chinese word segmentation mainly focused on dictionary-based approaches which matched input sentences against a given dictionary. However, the "word" in Chinese is actually not a well-defined concept, and no generally accepted lexicon exists. Furthermore, different tasks may have different granularities for defining Chinese word segmentation. In computer applications, "segmentation units" receive more attention than "words" [29]. For example, "并行计算机 (parallel computer)" may be segmented as two segmentation units "并行/计算机 (parallel/computer)" in an information retrieval task, but may be regarded as one unit in a key word extraction task.[1] Moreover, new words come into being all the time. Because of such factors, statistically based methods are the mainstream approach to Chinese word

* Corresponding author.
*E-mail addresses:* yingxiong@sjtu.edu.cn (Y. Xiong), zhujie@sjtu.edu.cn (J. Zhu), haohuang@sjtu.edu.cn (H. Huang), haihua_xu@sjtu.edu.cn (H. Xu).

[1] To facilitate the use of terminology, the term "words" will be used to mean "segmentation units" in the rest of this paper.

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.09.018

segmentation, especially supervised machine learning methods such as HMM, maximum entropy (ME) [30], and CRFs [14]. Unlike dictionary-based methods, machine learning methods rely on statistical models which are learned automatically from corpora, making them more adaptive and robust in processing different corpora. In the recent SIGHAN Bakeoff competitions [5,16], CRFs were widely used and outperformed other machine learning methods such as HMM, support vector machine (SVM) [7], and ME. However, little work has been done on training criteria for CRFs. In this paper, a new training criterion is presented which achieves further improvement in the performance of CRFs without adding to the commonly used unigram and bigram features.

CRF parameters were first estimated using the maximum log-likelihood (ML) criterion [14]. However, the ML criterion is prone to overfitting because CRFs are often trained with a very large number of overlapping features. The maximum a posteriori (MAP) criterion was then proposed in [23] to reduce overfitting. Large margin methods have also been applied to parameter optimization [1,25,27]. Furthermore, the minimum classification error (MCE) criterion, on which the speech and pattern recognition research communities often focused, was adapted to CRF parameter estimation [24]. Gross et al. [8] proposed a training procedure that maximized per-label predictive accuracy in [8]. The procedure was similar to MCE except that it was based on a pointwise loss function rather than a sequential loss function.

These training criteria achieved excellent performances on various tasks. For the task of sequence labeling, ideally a CRF model is desirable because it can provide high accuracy when labeling new sequences. However, it is difficult to find parameters which provide the best possible accuracy on training data. In particular, to maximize sequence tagging accuracy, which is measured by the number of correct labels, gradient-based optimization methods cannot be used directly. Therefore, other optimization methods such as those mentioned above are used.

This paper presents a new discriminative training criterion called minimum tag error (MTE) which can be seen as being in the same spirit as MCE, but with a different objective function that is more naturally applicable to the sequence labeling task. The MTE criterion is a smoothed approximation to the tag accuracy measured on the output of a sequence labeling system given the training data, which can be directly optimized by the gradient-based method without providing a smoothing function as with the MCE criterion. The effectiveness of this new criterion is tested on the Chinese word segmentation task because Chinese word segmentation is a prerequisite step for Chinese information processing. The experimental results presented here show that the proposed new criterion can reliably enhance the initial results yielded by the MAP trained model. Furthermore, the new approach has the advantage of being able to recognize out-of-vocabulary (OOV) words (i.e., the set of words in the test corpus which do not occur in the training corpus) better than the standard MAP training method.

The remainder of this paper is structured as follows: Section 2 reviews standard conditional random fields. The main focus is in Section 3 which introduces the MTE training method. Section 4 describes the experiments, Section 5 presents a discussion of the results, and Section 6 states conclusions. Finally, acknowledgements are expressed.

## 2. Conditional random fields

Let $\underline{X} = \langle X_1, X_2, \ldots, X_R \rangle$ be observation input data sequences to be labeled, and let $\underline{Y} = \langle Y_1, Y_2, \ldots, Y_R \rangle$ be a set of corresponding label sequences, where $R$ is the number of data sequences. All components of $Y_i$ ($i = 1, 2, \ldots, R$) are assumed to range over a finite tag set $T$. For example, $\underline{X}$ might consist of unsegmented Chinese sentences, and $\underline{Y}$ might range over the boundary tags of these sentences, with $T$ a set of boundary tags such as the commonly used BIO ("B" means the beginning of a word, "I" indicates a character other than the beginning of a word and "O" represents a single-character word). CRF models define the conditional probability of a particular label sequence $Y = \langle y_1, y_2, \ldots, y_n \rangle$ given the observation sequence $X = \langle x_1, x_2, \ldots, x_n \rangle$ as

$$p(Y|X) = \frac{1}{Z_X} \exp\left( \sum_{i=1}^{n} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, X, i) \right), \tag{1}$$

$$Z_X = \sum_{Y} \exp\left( \sum_{i=1}^{n} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, X, i) \right), \tag{2}$$

where $Z_X$ is a normalization factor over all state sequences, $f_k(y_{i-1}, y_i, X, i)$ is an arbitrary feature function which can be defined as a transitional function for the state pair $(y_{i-1}, y_i)$ or a state function for the state–observation pair $(y_i, X)$, $\lambda_k$ is a learned parameter associated with feature $f_k$, and $n$ is the length of the sequence.

For simplicity of presentation, the expressions defined in [23] are used as follows:

The global feature vector for $X$ and $Y$ is given as

$$\boldsymbol{F}(Y, X) = \sum_{i} \boldsymbol{f}(y, X, i), \tag{3}$$

where $i$ ranges over input positions. The conditional distribution of CRF can be rewritten as

$$p_\lambda(Y|X) = \frac{\exp(\boldsymbol{\lambda} \cdot \boldsymbol{F}(Y, X))}{\sum_Y \exp(\boldsymbol{\lambda} \cdot \boldsymbol{F}(Y, X))} = \frac{\exp(\boldsymbol{\lambda} \cdot \boldsymbol{F}(Y, X))}{Z_\lambda(X)}. \tag{4}$$

The parameter vector $\boldsymbol{\lambda}$ can be estimated by the maximum log-likelihood method. Sha and Pereira have proven in [23] that the L-BFGS algorithm can converge much faster when training CRF models. To reduce overfitting, the log-likelihood function is often penalized by a Gaussian prior distribution over the parameters which can be written as