# Conditional Random Fields for Punctuation Prediction Learning Algorithms, Project 2

Mohsen Malmir, Erfan Sayyari

February 11, 2014

## 1 Abstract

## 2 Introduction

In this paper, we provide the results and details of implementation of a Conditional Random Field (CRF) model for predicting English language text punctuations. CRFs are a variant of undirected graphical models that are well suited for predicting structured labels. We train our model by maximizing log conditional likelihood (LCL) of the training data. During our experiments, we found that the model could predict individual tags by more than 94% accuracy. However, as we performed more and more experiments, we discovered the over fitting to some prevalent tags in the training set. Using several experiments and methods that will be described later in this paper, we could overcome the over-fitting problem to some degrees.

We choose two techniques to train the proposed model: Collin's Perceptron and Contrastive Divergence. Both of these methods are approximations to the general gradient following method. We choose these two methods because of their two characteristics: simplicity and elegance. Both Collin's Perceptron and Contrastive Divergence provides simplified updates rule for the parameters, which are less expensive compared to the general gradient following for maximizing LCL. Despite simplicity, both methods have very clear and intuitive explanation behind them: they maximize the LCL of the data by throwing spurious samples at the model. However, as we explain our experiments in more details, we do not stick to the basic algorithms and tweak them in several ways to improve their performance.

One of the most important parts of the CRF model for predicting text punctuation is the choice of feature functions as it effects both the performance and accuracy of the model. The feature functions we develop here use Part-of-Speech (POS) tags for the input sentence. The main characteristic for the proposed feature functions is their efficiency in computation time: training of our method only takes a few minutes on a single core machine. We show that using these feature functions, the model can predict the punctuation tags with high accuracy.

For the given training and test data, danger of over-fitting is high because of the imbalanced distribution of punctuation tags. We deal with this problem with different techniques. We show how bounding the weight parameters and early stopping can improve the accuracy for the Collin's Perceptron. By introducing random sampling and guided sampling, we achieve similar improvements in Contrastive Divergence. Then we introduce the Turn-taking train procedure: for each tag we train a different predictor and then combine them using .... We show this last technique is much more effective and achieves the highest accuracy for punctuation prediction for the given data set.

# 3 Design and Analysis of Algorithm

## 3.1 The general log-linear model

## 3.2 Feature functions

### 3.2.1 Part of Speech tagging

## 3.3 Conditional random fields

## 3.4 The Collins perceptron

## 3.5 Contrastive divergence

### 3.5.1 Gibbs sampling

# 4 Design of Experiments

## 4.1 Initialization

## 4.2 Preprocessing

## 4.3 Performance Measure

# 5 Results of Experiments

## 5.1 Collins perceptron

## 5.2 Gibbs sampling

# 6 Findings and Lessons Learned

## 6.1 Numerical Issues and Preprocessing

## 6.2 Overfitting

## 6.3 Model Selection

## 6.4 Future works