# Assignment 2: Designing a CRF Model from Scratch
# ELL884

Animesh Lohar - 2024EET2368

**Abstract**

This report explores the implementation of Conditional Random Fields (CRFs) in Named Entity Recognition (NER). It discusses the theoretical foundations, mathematical derivations, and practical implementation approaches, alongside the challenges faced during implementation and the results derived from model evaluation.

## 1 Introduction

Named Entity Recognition (NER) is a vital subfield in Natural Language Processing (NLP), focusing on the identification and classification of entities in text into predefined categories such as persons, organizations, locations, and dates. Its significance is underscored in various applications, including information retrieval, automated summarization, and sentiment analysis.

Conditional Random Fields (CRFs) are a class of discriminative models that predict sequences of labels for structured prediction tasks. Unlike generative models, CRFs model the conditional probability of the output label sequence based on the observed input features. This characteristic is particularly beneficial for NER tasks, where interdependencies among labels are crucial.

The objective of this report is to detail the theoretical and practical implementations of CRFs for NER, illustrating both the mathematical foundations and the performance of the developed model.

## 2 Mathematical Derivations

CRFs model the conditional probability of a label sequence $Y$ given an observed input sequence $X$:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_k \sum_t \psi_k(y_t, y_{t-1}, X, t) w_k\right) \tag{1}$$

Where:

- $Z(X)$ is the partition function, which normalizes the probability distribution:

$$Z(X) = \sum_{Y'} \exp\left(\sum_k \sum_t \psi_k(y'_t, y'_{t-1}, X, t) w_k\right) \tag{2}$$

- $\psi_k(y_t, y_{t-1}, X, t)$ are feature functions representing the relationship of the current label $y_t$ with its previous label $y_{t-1}$ and the observed sequence $X$.
- $w_k$ are the weights associated with the feature functions, which are learned during training.

The aim is to optimize the weights $w$ using Maximum Likelihood Estimation, maximizing the log-likelihood of observed sequences:

$$\hat{w} = \arg\max_w \sum_i \log P(Y_i|X_i; w) \tag{3}$$

The gradient of the log-likelihood with respect to the weights is computed as:

$$\nabla_w L(w) = \sum_i \left( \sum_t \psi(y_t, y_{t-1}, X_i, t) - \mathbb{E}_P[\psi(y_t, y_{t-1}, X_i, t)] \right) \tag{4}$$

Where $\mathbb{E}_P$ denotes the expected value under the current model distribution.

# 3 Implementation Approach

The implementation of the CRF model for NER followed several key steps:

## 3.1 Data Preparation

We used a dataset consisting of text annotated with named entities. Data preprocessing included tokenization and alignment of input and output sequences to ensure each token's corresponding label was accurately represented.

## 3.2 Feature Engineering

A variety of feature functions were designed to improve the model's predictive capacity:

- **Word-level Features**: Current word, prefixes, and suffixes.
- **Contextual Features**: Previous and next tag features to capture label dependencies.
- **Syntactic Features**: Word shape features, such as capitalization and numeric presence.

These features were chosen to exploit linguistic patterns and enhance entity recognition.

## 3.3 Model Training

We employed stochastic gradient ascent for training the model's weights. Cross-validation techniques were utilized to monitor performance and fine-tune hyperparameters, aiming to improve the model's ability to generalize to unseen data.

## 3.4 Challenges and Solutions

Several challenges arose during the implementation:

- **Data Imbalance**: Certain entity types were underrepresented. We addressed this by using data augmentation techniques to balance the dataset.
- **Computational Load**: The retraining of the model on large datasets proved intensive. Batch processing and efficient data handling techniques were explored to improve processing times.
- **Overfitting Risks**: Overfitting during training necessitated the use of regularization techniques to lap performance improvements on unseen data.

# 4 Results and Evaluation

The CRF model was evaluated on a separate test set using precision, recall, and F1-score metrics, which are crucial for NER tasks. The results are summarized in Table 1.

The analysis of results showed the model's effectiveness at identifying commonly occurring entities, while challenges remained with less frequent entity types. These findings indicate areas for future improvement, particularly in feature representation and data collection strategies.

| Entity Type | Precision | Recall | F1-Score |
|---|---|---|---|
| PER | 0.88 | 0.85 | 0.86 |
| LOC | 0.92 | 0.90 | 0.91 |
| ORG | 0.85 | 0.83 | 0.84 |

Table 1: Performance Metrics for the CRF Model

# 5 Conclusion

In summary, this report highlights the application of Conditional Random Fields for Named Entity Recognition. We explored the theoretical underpinnings, implemented the model, and evaluated its performance. The CRF model demonstrated promising results, although certain challenges persist, indicating further research avenues, including deeper exploration into neural network architectures for improved performance.