McGill Department of Electrical and Computer Engineering

NUANCE Nuance Communications Inc.

# RECENT IMPROVEMENTS TO NEUROCRFS FOR NAMED ENTITY RECOGNITION

**Marc-Antoine Rondeau** marcantoine.rondeau@gmail.com

**Yi Su** yi.su@nuance.com

## Introduction

**Goal:** Improve NeuroCRFs' sequence labelling performance with feature engineering, large margin training and ensemble learning.

- The similarities between labels can be exploited to add parameters shared by groups of similar transitions.
- A modified CRF partition function $Z(\mathbf{x})$ increases the margin between correct and incorrect hypotheses.
- The non-convexity of NNs is exploited to combine models with different random initializations into a single ensemble model.

By combining those approaches, we obtain $F_1 = 88.50$, a significant improvement over the 87.49 baseline on a named entities recognition task.

## NeuroCRF

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left(G(\mathbf{x}_t)F(y_{t-1}, y_t) + A_{y_{t-1},y_t}\right)$$

**Low Rank: NN used to model label emissions**

$$F(y_{t-1}, y_t) = F(y_t) = [f_1(y_t), \cdots, f_N(y_t)]^\top$$
$$f_i(y_t) = \begin{cases} 1, & i = y_t \\ 0, & i \neq y_t \end{cases}$$

**Full Rank: NN used to model label to label transitions**

$$F(y_{t-1}, y_t) = [f_1(y_{t-1}, y_t), \cdots, f_{N^2}(y_{t-1}, y_t)]^\top$$
$$f_i(y_{t-1}, y_t) = \begin{cases} 1, & i = Ny_{t-1} + y_t \\ 0, & i \neq Ny_{t-1} + y_t \end{cases}$$

## Shared Parameters

Labels are assigned to groups such as:

- $B(\text{O}) = \{\text{O, OUT}\}$
- $B(\text{B-LOC}) = \{\text{B-LOC, B, LOC, ENT}\}$

Labels to labels transitions are assigned to group set:

$$D(y_{t-1}, y_t) = (B(y_{t-1}) \times B(y_t)) \cup B(y_t)$$

**Shared parameters: NN used to model group transition and emission**

$$F(y_{t-1}, y_t) = \frac{1}{|D(y_{t-1}, y_t)|} \left[f_1(y_{t-1}, y_t), \cdots f_{|\mathcal{D}|}(y_{t-1}, y_t)\right]^\top$$
$$f_i(y_{t-1}, y_t) = \begin{cases} 1, & S(i) \in D(y_{t-1}, y_t) \\ 0, & \text{otherwise} \end{cases}$$

- $S(i)$ is the element of $\mathcal{D} = \bigcup D(y_{t-1}, y_t)$ assigned to NN output $G_i(\mathbf{x}_t)$
- An extension of full rank NeuroCRFs
- Parameters are shared by group of related emission and transition
- Requires feature engineering to define "related"

## Large Margin Training

**Large margin training: weight hypotheses in $Z$ to increase margin**

$$Z(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{y}' \in \text{gen}(\mathbf{x})} C(\mathbf{y}, \mathbf{y}') \exp\left(\sum_{t=1}^{T} G(\mathbf{x}_t)F(y'_{t-1}, y'_t)\right)$$
$$C(\mathbf{y}, \mathbf{y}') = \exp \sum_{t=1}^{T} \begin{cases} -1, & y_t = y'_t \\ 0, & y_t \neq y'_t \end{cases}$$

- $Z(\cdot)$ minimized during training: correct hypothesis is penalized
- Modified $Z(\cdot)$ to reduce penalty of good hypotheses during training
- Reduction proportional to similarity with $\mathbf{y}$

## Ensemble Learning

**Ensemble Learning: Exploit non-convexity to combine models**

$$\hat{G}(\mathbf{x}_t) = \frac{1}{M} \sum_{m=1}^{M} G^{(m)}(\mathbf{x}_t)$$

- NNs' non-convexity: training find local minimum
- Combine trained models to exploit complementarity of local-minimums
- Averaging output, not parameters
- Ensemble model concatenates hidden layers parameters and averages output layer parameters
- Resulting model much larger than original trained models

## Experimental Study

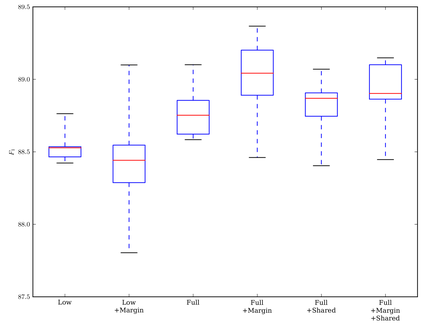
- Experiment on named entities recognition (NER)
- Two tasks:
  - CoNLL-2003: Manually annotated newswire
  - WikiNER: Semi-automatically annotated Wikipedia
- Small size of CoNLL-2003 limit training: improvement masked by overfitting
- Use similar task WikiNER to compensate
- WikiNER derived from links added by editors:
  - Manual segmentation
  - Automatic classification of segment
- Effective annotation directives different from CoNLL-2003

| Corpus Sizes | | | | | | |
|---|---|---|---|---|---|---|
| | CoNLL-2003 | | | WikiNER | | |
| | Train | Val. | Test | Train | Val. | Test |
| #Sentence | 14,987 | 3,466 | 3,684 | 113,812 | 14,178 | 14,163 |
| #Words | 203,621 | 51,362 | 46,435 | 2,798,532 | 351,322 | 349,752 |
| Entities | | | | | | |
| #LOC | 7,140 | 1,837 | 1,668 | 68,737 | 8,718 | 8,580 |
| #MISC | 3,438 | 922 | 702 | 58,826 | 7,322 | 7,462 |
| #ORG | 6,321 | 1,341 | 1,661 | 39,795 | 4,912 | 4,891 |
| #PER | 6,600 | 1,842 | 1,617 | 77,010 | 9,594 | 9,613 |
| All | 23,499 | 5,942 | 5,648 | 244,368 | 30,546 | 30,546 |

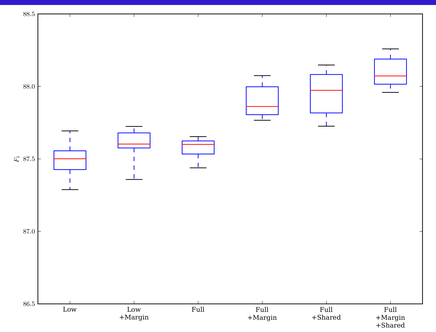
## Results

| System | ConLL-2003 | | | WikiNER | | |
|---|---|---|---|---|---|---|
| | Mean $F_1$ | Max $F_1$ | Ens. $F_1$ | Mean $F_1$ | Max $F_1$ | Ens. $F_1$ |
| Low Rank | 88.54 | 88.76 | 88.88 | 87.49 | 87.69 | 88.02 |
| +Margin | 88.34 | 89.10 | 88.77 | 87.60 | 87.72 | 87.79 |
| Full Rank | 88.77 | 89.10 | 89.14 | 87.58 | 87.65 | 88.03 |
| +Margin | **88.97** | **89.37** | 89.23 | 87.90 | 88.07 | 88.29 |
| +Shared | 88.81 | 89.07 | 89.37 | 87.95 | 88.15 | 88.40 |
| +Margin+Shared | 88.92 | 89.15 | **89.62** | **88.10** | **88.26** | **88.50** |

## Task 1: Named entity recognition (CoNLL-2003)




## Task 2: Named entity recognition (WikiNER)




## Conclusions

- CoNLL-2003 improved by large margin and ensemble learning
- Overfitting prevent improvement with shared parameters
- WikiNER large enough to support added parameters
- Combination of large margin training and ensemble learning improved further
- Future work:
- Better feature engineering of shared parameters
- Feature learning of shared parameters
- Improved regularization of larger model to replicate ensemble learning