McGill

Department of Electrical and Computer Engineering

FULL-RANK LINEAR-CHAIN NEUROCRF FOR SEQUENCE LABELING

Marc-Antoine Rondeau marc-antoine.rondeaubeauchamp@mail.mcgill.ca

Yi Su yi.su@nuance.com

Introduction

Goal: Improve sequence labelling performance by directly modelling label to label transitions with a neural network

The successful combination of deep neural network (DNN) and hidden Markov model (HMM) in acoustic modeling inspired the combination of NN and conditional random fields (CRF). Those "NeuroCRFs" used a HMM-like output layer:

- DNN generated emission scores
- Constant transition matrix

We propose to use a NN to generate *transition* scores directly.

HMM-like output layer: Low-Rank NeuroCRF

NN used to model label emissions

CRFs are similar to softmax applied to sequences:

$$P(\mathbf{y}|\mathbf{x}) = \frac{\exp F(\mathbf{y})}{\sum_{\mathbf{y}'} \exp F(\mathbf{y}')}$$

$$F(\mathbf{y}) = \sum_t G_{y_t}(\mathbf{x}_t) + A_{y_{t-1},y_t},$$

The neural network output score all possible labels for a given word. This score is combined to a transition matrix.

Full-Rank NeuroCRF

NN used to model label to label transitions

$F(\mathbf{y})$ is replaced by

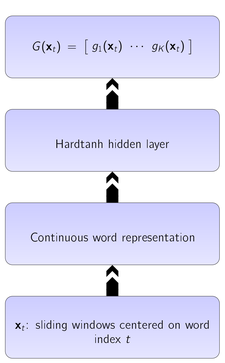
$$F^{(f)}(\mathbf{y}) = \sum_t G_{y_{t,1},y_t}(\mathbf{x}_t)$$

- NN learns to detect transitions rather than emissions
- Model emission as dependent on input *and* previous label
- Can adapt transition scores to input
- Full-rank can learn parameters equivalent to low-rank NeuroCRF

Overview

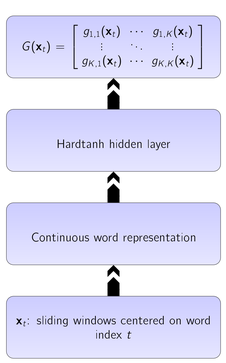
Low-rank

$F(\mathbf{y}, \mathbf{x}) = \sum_t g_{y_t}(\mathbf{x}_t) + A_{y_{t-1},y_t}$



Full-rank

$F(\mathbf{y}, \mathbf{x}) = \sum_t g_{y_{t-1},y_t}(\mathbf{x}_t)$



Tasks

We applied low and full rank NeuroCRFs to two segment labelling tasks:

- Syntactic chunking (CoNLL-2000): segments defined by syntactic role
- Named entity recognition (NER, CoNLL-2003): segments are named entities

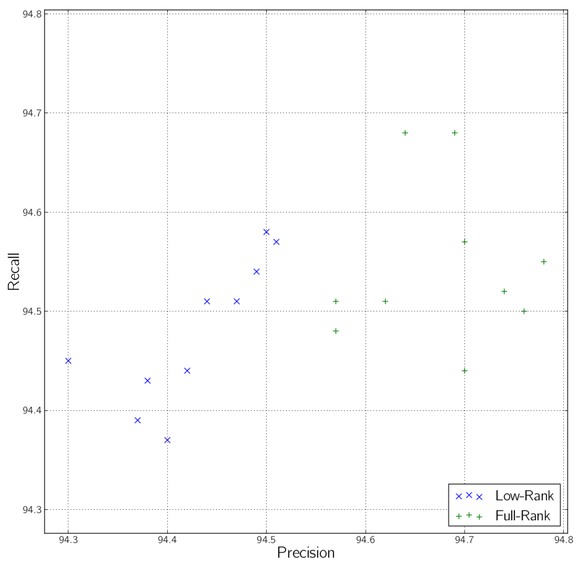
Table : Training sets' details.

	Chunking	NER
# Classes	11	4
# Labels	45	17
# Words	188,112	203,621
# Words inside segment	**163,700**	**34,600**
Entropy (labels)	3.36	1.24
Conditional entropy	1.52	0.87
Mutual information	**1.84**	**0.37**

Performance measured by $F_1 = 2pr/(p+r)$, averaged for 10 random initializations

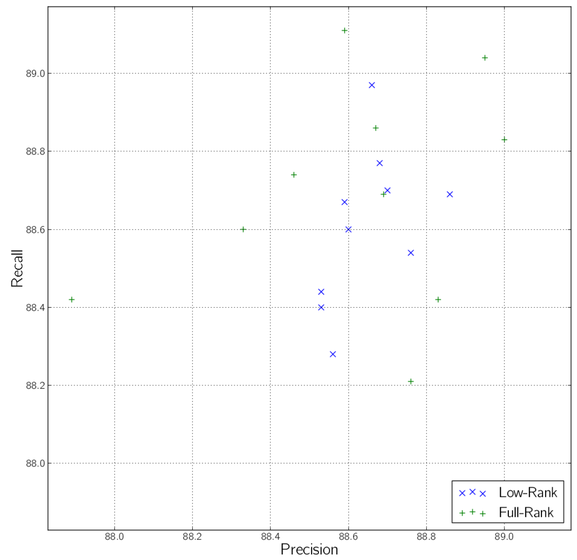
- Precision p: # correctly labelled segments divided by # decoded segments
- Recall r: # correctly labelled segments divided by # segments in test

Task 1: Chunking (CoNLL-2000)



- Obtained significant improvements
- Precision-Recall graph confirms difference
- Improved precision
- High mutual information between successive labels
- Full-rank NeuroCRF helpful: label emission depends on previous label
- Label to label transitions *not* well modelled by constant transition matrix

Task 2: Named entity recognition (CoNLL-2003)



- Added parameters cause overfiting; corrected by dropout
  - Without dropout: 87.92 from 88.53
  - With dropout: 88.65 from 88.63
- Precision-Recall graph confirms similarity
- Low mutual information between successive labels: emission scores equivalent to transition scores
- Label to label transitions well modelled by constant transition matrix
- Good regularization prevent degradation

Experimental results for 10 random initializations

	Chunking		NER	
	Low-Rank	Full-Rank	Low-Rank	Full-Rank
Average	94.45	**94.61**	88.63	**88.65**
Minimum	94.37	**94.52**	**88.42**	88.15
Maximum	94.54	**94.68**	88.81	**88.99**
Std. Dev	0.0664	0.0561	0.1344	0.2482

Conclusions

- Full-rank improved performance on task with significant dependencies between labels
- Full-rank model was equivalent to low-rank on task without significant dependencies between labels
- Regularization prevented overfitting and enabled full-rank NeuroCRF to learn parameters equivalent to low-rank