Upgrading CRFs to JRFs and Its Benefits to Sequence Modeling and Labeling

Yunfu Song¹, Zhijian Ou¹, Zitao Liu², Songfan Yang²

¹Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University
²TAL AI Lab, Beijing, China

http://oa.ee.tsinghua.edu.cn/ouzhijian/

Presented at International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020
Content

1. Introduction
2. JRF
3. Experiments
4. Conclusions
Introduction

• **Sequence modeling**
  - For sequence of length $l$, $x^l \triangleq x_1, x_2, \ldots, x_l$, calculate $p(l, x^l)$
  - e.g. language modeling

• **Sequence labeling**
  - Given observation sequence $x^l$, predict the label sequence $y^l \triangleq y_1, y_2, \ldots, y_l$
  - e.g. part of speech (POS) tagging, named entity recognition (NER), and chunking.
Motivation

• **Sequence modeling**
  - Can be improved with additional relevant labels, e.g. incorporating POS tags for language modeling.
  - Labels usually not available in testing, use hypothesized labels in testing. 😞

• **Sequence labeling**
  - Mainly learn from **limited** labeled data. 😞

---

**Probabilistic generative modeling**

• Avoid need of labels in testing. 😃
• Leverage both labeled data and unlabeled, task-dependent semi-supervised learning. 😃
Conditional random field (CRF)

(Linear-chain) CRFs define a conditional distribution $y^l$ given $x^l$ of length $l$:

$$p_\theta(y^l|x^l) = \frac{1}{Z_\theta(x^l)} \exp(u_\theta(x^l, y^l))$$

$$Z_\theta(x^l) = \sum_{y^l} \exp(u_\theta(x^l, y^l))$$

Potential function:

$$u_\theta(x^l, y^l) = \sum_{i=1}^{l} \phi_i(y_i, x^l) + \sum_{i=1}^{l-1} \psi_i(y_{i-1}, y_i, x^l)$$

• Upgrade CRFs, a joint generative model of $x^l$ and $y^l$, $p(l, x^l, y^l)$
• Use $u(x^l, y^l)$ in the original CRF

Content

1. Introduction
2. JRF
3. Experiments
4. Conclusions
Joint random field (JRF)

Define a joint distribution:

\[ p_\theta(l, x^l, y^l) = \pi_l p_\theta(x^l, y^l; l) = \frac{\pi_l}{Z_\theta(l)} \exp\left( u_\theta(x^l, y^l) \right) \]

\[ Z_\theta(l) = \sum_{x^l, y^l} \exp\left( u_\theta(x^l, y^l) \right) \]

From JRF we have:

\[ p_\theta(y^l|x^l) = \frac{1}{\sum_{y^l} \exp(u_\theta(x^l, y^l))} \exp\left( u_\theta(x^l, y^l) \right) \]

Which is a CRF

From JRF we have:

\[ p_\theta(l, x^l) = \frac{\pi_l}{Z_\theta(l)} \sum_{y^l} \exp\left( u_\theta(x^l, y^l) \right) \]

\[ = \frac{\pi_l}{Z_\theta(l)} \exp\left( u_\theta(x^l) \right) \]

Where \( u_\theta(x^l) = \log \sum_{y^l} \exp\left( u_\theta(x^l, y^l) \right) \)

Which is a trans-dimensional random field (TRF)

[Bin Wang and Zhijian Ou, “Improved training of neural trans-dimensional random field language models with dynamic constraint”, in SLT 2018.]
JRF

CRF
\[ p_{\theta}(y^l|x^l) \]

Supervised Learning

Labeled Data

Unsupervised Learning

Unlabeled Data

TRF
\[ p_{\theta}(l, x^l) \]

Node Potentials

Edge Potentials

JRF
\[ p_{\theta}(l, x^l, y^l) \]

Bi-LSTM

\[ x_1 \rightarrow h_1 \rightarrow o_1 \]
\[ x_2 \rightarrow h_2 \rightarrow o_2 \]
\[ x_3 \rightarrow h_3 \rightarrow o_3 \]
JRF

**Supervised learning of JRFs:**
- Similar to training of CRFs
- Empirical distribution $p_L(x^l, y^l)$

$$\max_\theta L_S(\theta) = E_{(x^l, y^l) \sim p_L(x^l, y^l)} \left[ \log p_\theta(y^l|x^l) \right]$$

**Unsupervised learning of JRFs:**
- Similar to training of TRFs
- Use dynamic noise-contrastive estimation (DNCE)
- Introduce a noise distribution $p_\phi(l, x^l)$ (generally a LSTM language model)
- Empirical distribution $p_U(l, x^l)$

**Semi-supervised learning of JRFs:**
- Combine supervised and unsupervised training

$$\begin{cases} 
\max_\theta L(\theta) = L_S(\theta) + \alpha L_U(\theta) \\
\min_\phi KL(p_U(l, x^l) || p_\phi(l, x^l)) 
\end{cases}$$

$$\max_\theta \mathbb{E}_{(l, x^l) \sim p_U(l, x^l) + p_\phi(l, x^l)} \left[ \log \frac{p_\theta(l, x^l)}{p_\theta(l, x^l) + p_\phi(l, x^l)} \right] +$$

$$\mathbb{E}_{(l, x^l) \sim p_\phi(l, x^l)} \left[ \log \frac{p_\phi(l, x^l)}{p_\theta(l, x^l) + p_\phi(l, x^l)} \right] \equiv L_u(\theta)$$

$$\min_\phi KL(p_U(l, x^l) || p_\phi(l, x^l))$$
Content

1. Introduction
2. JRF
3. Experiments
4. Conclusions
Experiments (Sequence modeling)

- Dataset: WSJ portion PTB
- Rescore the 1000-best list from WSJ’92 test set
- Evaluate the word error rate (WER).

- KN5 (n-gram), LSTM, TRF language models are trained without POS tags
- JRF is trained with POS tags, and avoids the need of POS tags during testing

<table>
<thead>
<tr>
<th>Method</th>
<th>KN5</th>
<th>LSTM</th>
<th>TRF</th>
<th>JRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER (%)</td>
<td>8.78</td>
<td>7.36</td>
<td>6.99</td>
<td>6.77</td>
</tr>
</tbody>
</table>

8% ↓

3% ↓
Experiments (Sequence labeling)

- POS tagging (PTB), NER (CoNLL-2003) and chunking (CoNLL-2000)
- **Accuracy** for POS tagging, **F1 score** for NER and chunking (BIOES)

<table>
<thead>
<tr>
<th>Method</th>
<th>POS (10%)</th>
<th>POS (100%)</th>
<th>NER (10%)</th>
<th>NER (100%)</th>
<th>Chunking (10%)</th>
<th>Chunking (100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>96.83</td>
<td>97.45</td>
<td>86.85</td>
<td>90.87</td>
<td>89.98</td>
<td>94.76</td>
</tr>
<tr>
<td>Self-training</td>
<td>96.91</td>
<td>97.46</td>
<td>86.92</td>
<td>90.88</td>
<td>90.64</td>
<td>94.84</td>
</tr>
<tr>
<td>JRF</td>
<td>96.96</td>
<td>97.47</td>
<td>86.99</td>
<td>90.90</td>
<td><strong>91.12</strong></td>
<td><strong>95.10</strong></td>
</tr>
</tbody>
</table>

- CRF performs purely supervised learning
- Self-training and JRF perform semi-supervised learning

*Note: The table shows accuracy percentages.*
Content

1. Introduction
2. JRF
3. Experiments
4. Conclusions
Conclusions

- We propose to upgrade CRFs to JRFs, obtained as a joint generative model of observation and label sequences.

- This development from CRFs to JRFs enables semi-supervised learning and benefits both sequence modeling and labeling tasks.
  - In language modeling rescoring task, the JRF model outperforms traditional language models and avoids the need of POS tags during testing.
  - For sequence labeling, JRFs achieve consistent improvements over the CRF baseline and self-training on POS tagging, NER and chunking tasks.

- Going to release the codes for reproducing this work.
Thanks for your attention!

Yunfu Song$^1$, Zhijian Ou$^1$, Zitao Liu$^2$, Songfan Yang$^2$

$^1$Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University

$^2$TAL AI Lab, Beijing, China

http://oa.ee.tsinghua.edu.cn/ouzhijian/