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

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/

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:

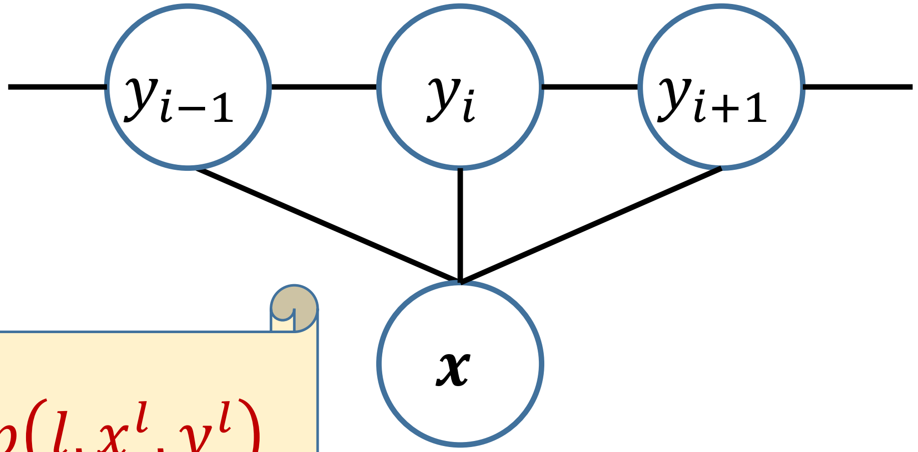
$$p_\theta(y^l|x^l) = \frac{1}{Z_\theta(x^l)} \exp(u_\theta(x^l, y^l)) \qquad Z_\theta(x^l) = \textstyle\sum_{y^l} \exp(u_\theta(x^l, y^l))$$

Potential function:

Node potential

Edge potential

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



- 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

Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

Content

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

Joint random field (JRF)

JRF

Define a joint distribution:

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

From JRF we have:

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

Which is a CRF

From JRF we have:

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

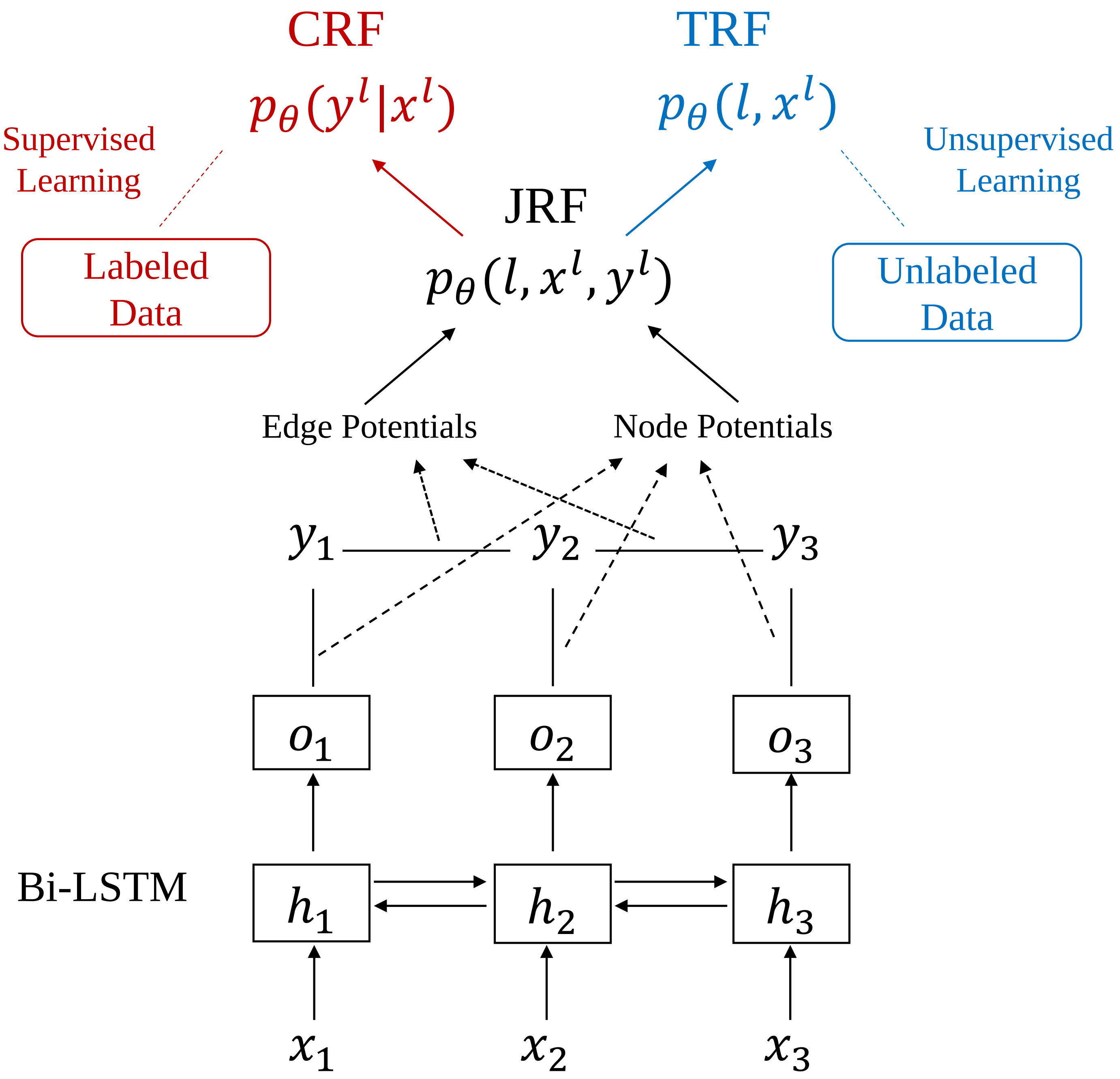
$$= \frac{\pi_l}{Z_\theta(l)} \exp\left(u_\theta(x^l)\right)$$

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

JRF



CRF
$p_\theta(y^l|x^l)$
TRF
$p_\theta(l, x^l)$
Supervised Learning
Unsupervised Learning
Labeled Data
Unlabeled Data
JRF
$p_\theta(l, x^l, y^l)$
Edge Potentials
Node Potentials
y_1
y_2
y_3
o_1
o_2
o_3
Bi-LSTM
h_1
h_2
h_3
x_1
x_2
x_3

JRF

Supervised learning of JRFs:

- Similar to training of CRFs
- Empirical distribution $p_L(x^l, y^l)$

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

Semi-supervised learning of JRFs:

- Combine supervised and unsupervised training

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

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 $p_U(l, x^l)$

$$\begin{cases} \max_\theta E_{(l,x^l)\sim\frac{p_U(l,x^l)+p_\phi(l,x^l)}{2}}[\log \frac{p_\theta(l,x^l)}{p_\theta(l,x^l)+p_\phi(l,x^l)}] + \\ \qquad E_{(l,x^l)\sim p_\phi(l,x^l)}[\log \frac{p_\phi(l,x^l)}{p_\theta(l,x^l)+p_\phi(l,x^l)}] \triangleq L_u(\theta) \\ \min_\phi KL(p_U(l,x^l)||p_\phi(l,x^l)) \end{cases}$$

Content

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

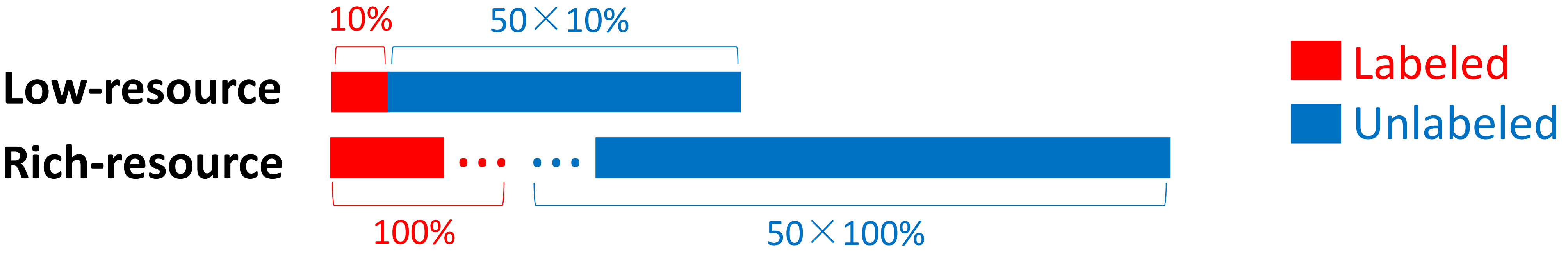
↓8%

Method	KN5	LSTM	TRF	JRF
WER (%)	8.78	7.36	6.99	**6.77**

↓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)



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

Method	POS (10%)	POS (100%)	NER (10%)	NER (100%)	Chunking (10%)	Chunking (100%)
CRF	96.83	97.45	86.85	90.87	89.98	94.76
Self-training	96.91	97.46	86.92	90.88	90.64	94.84
JRF	**96.96**	**97.47**	**86.99**	**90.90**	**91.12**	**95.10**

11%

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/