Fast Online Training with Frequency-Adaptive Learning Rates for Chinese Word Segmentation and New Word Detection

Xu SUN, Houfeng WANG, Wenjie LI

The Hong Kong Polytechnic University

Peking University

Outline

- **Introduction**
- **Method**
  - Joint modeling: word segmentation + new word detection
  - New features
- **A new online training method**
  - Feature-frequency adaptive online training
  - Finish the training in 10 passes
- **Experiments**
- **Conclusions**

Introduction

- 3 proposals in this work

Introduction

- **3 proposals in this work**

- **1) Joint modeling**
  - word segmentation + new word detection

- **2) New features**
  - high dimensional edge features on CRFs

Introduction

- 3 proposals in this work

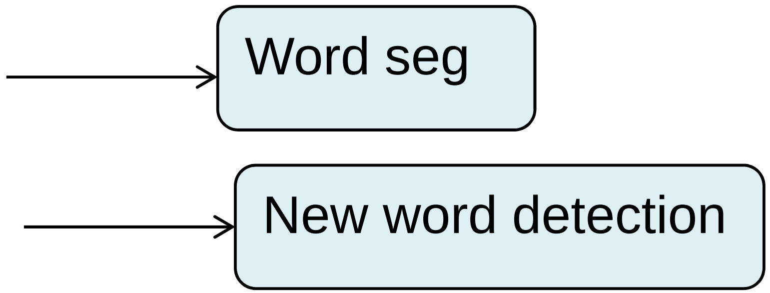
- 1) Joint modeling
  - word segmentation + new word detection

- 2) New features
  - high dimensional edge features on CRFs

- 3) Very Fast online training **major proposal**
  - finish the training in 10 passes

1) Joint modeling

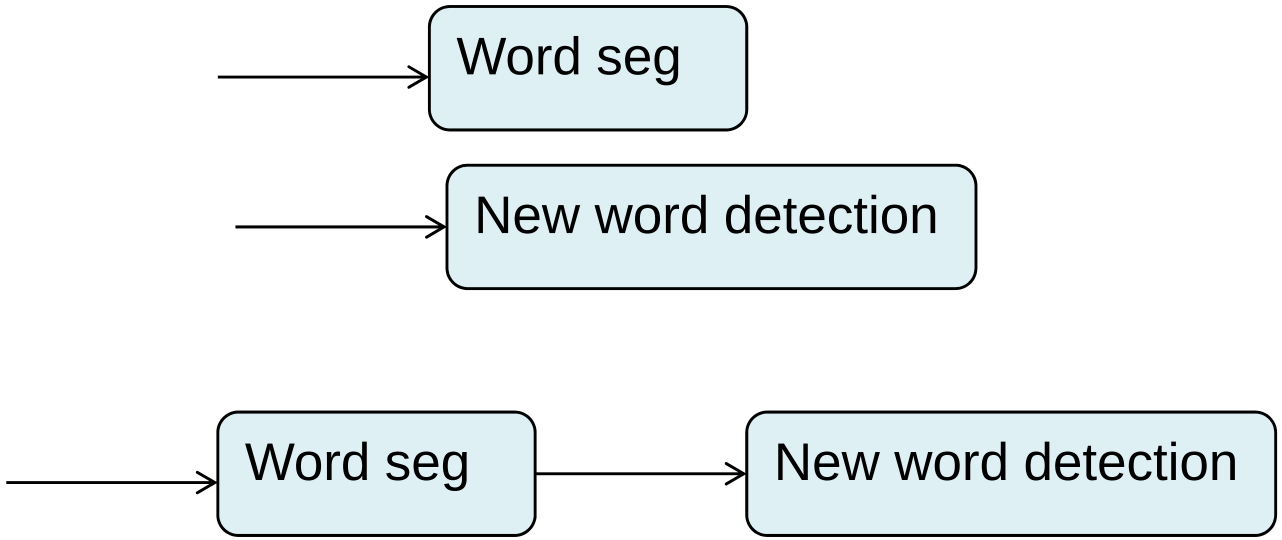
- **Prior work on word seg & new word detection**
  - Separate: two tasks are independent



1) Joint modeling

Prior work on word seg & new word detection

- Separate: two tasks are independent
- Pipeline: first word seg, then new word detection



1) Joint modeling

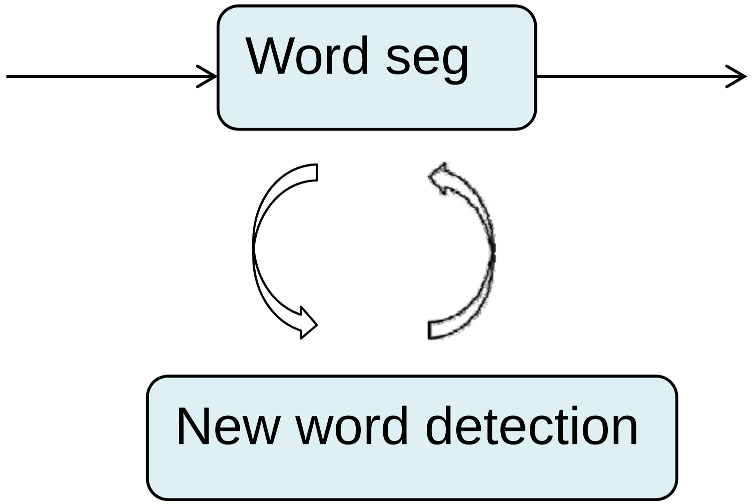
Our work

- Joint modeling: word seg + new word detection
- For simplicity, we use a single CRF model

1) Joint modeling

Our work

- Joint modeling: word seg + new word detection
- For simplicity, we use a single CRF model



2) New features

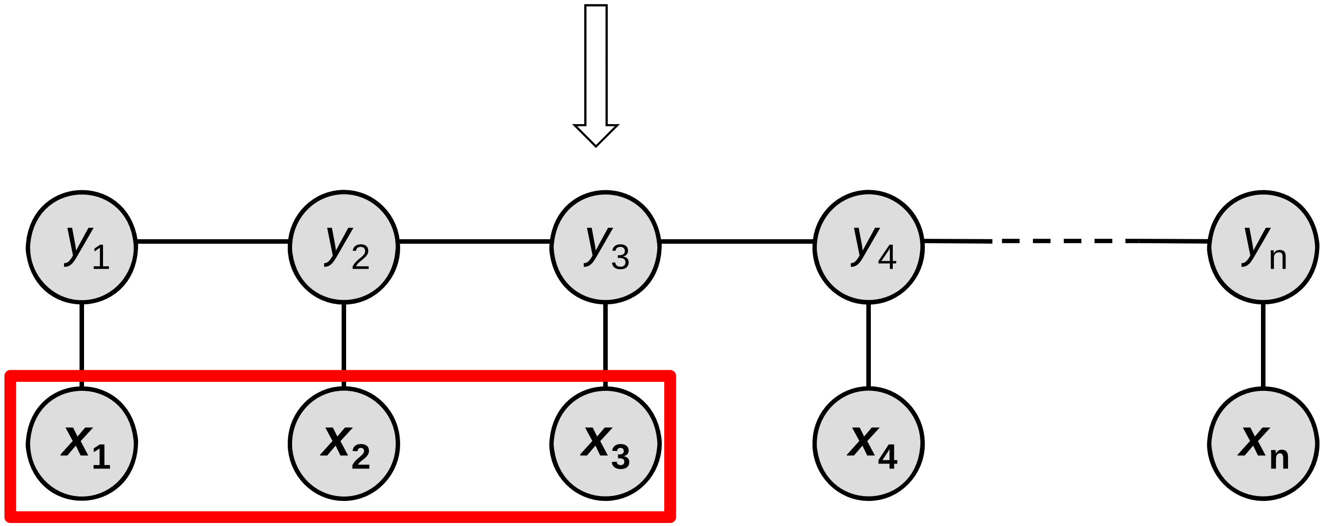
- Traditional features are character ngrams

2) New features

- **Traditional features are character ngrams**
- **New features: word-based ngrams**
  - Word lexicon is collected from training data
  - Word unigram features
  - Word bigram features

2) New features

- Traditional features are character ngrams
- New features: word-based ngrams
  - Word lexicon is collected from training data
  - Word unigram features
  - Word bigram features

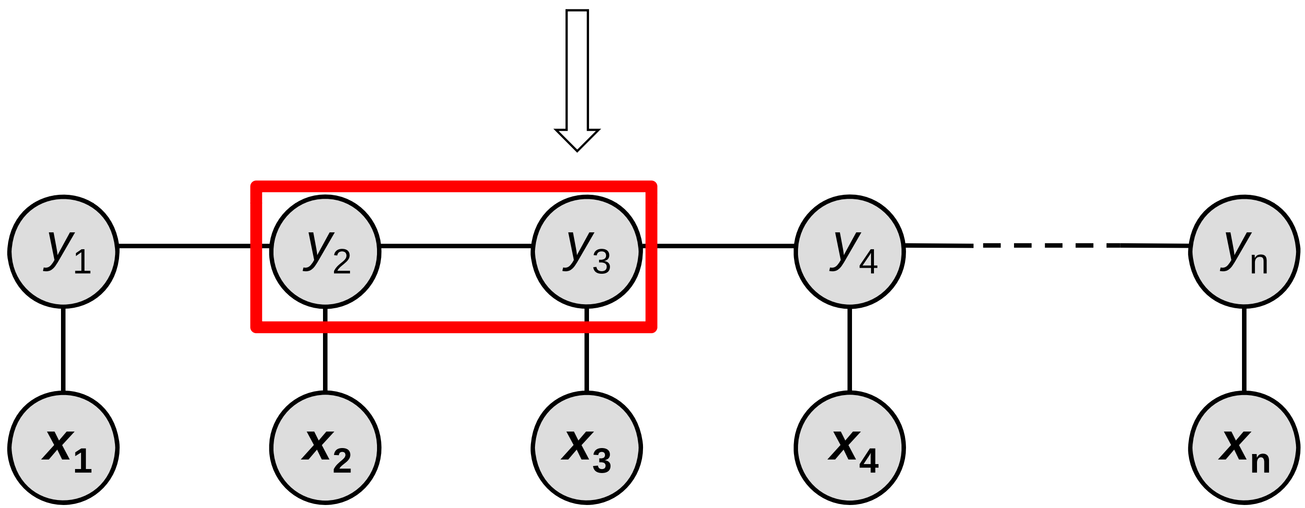


2) New edge (transition) features

- **Traditional edge features avoid observations**
  - to avoid feature explosion

2) New edge (transition) features

- **Traditional edge features avoid observations**
  - to avoid feature explosion

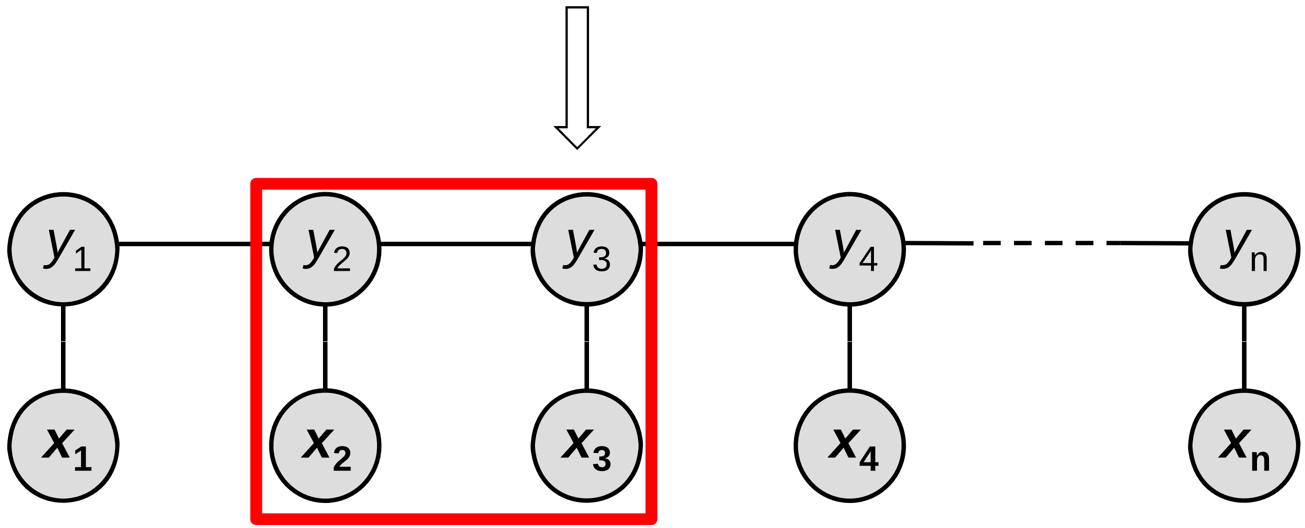


2) New edge (transition) features

- **Traditional edge features avoid observations**
  - to avoid feature explosion
- **Proposal: New edge features involving observations**
  - Tractable feature explosion: #feature increased 9 times
  - →10 millions features

2) New edge (transition) features

- **Traditional edge features avoid observations**
  - to avoid feature explosion
- **Proposal: New edge features involving observations**
  - Tractable feature explosion: #feature increased 9 times
  - → 10 millions features



3) Very fast online learning ***

- **Large-scale word segmentation is already slow**

3) Very fast online learning ***

- Large-scale word segmentation is already slow
- With high dimensional new features (10 millions)

→Even more slower training speed than before

3) Very fast online learning ***

- Large-scale word segmentation is already slow

- With high dimensional new features (10 millions)

→Even more slower training speed than before

→Need a fast training method, even with 10 millions of features

3) Very fast online learning ***

Batch training

- Limited memory BFGS (LBFGS)
- Too slow, need 300 passes for training

3) Very fast online learning ***

- **Batch training**
  - Limited memory BFGS (LBFGS)
  - Too slow, need 300 passes for training
- **Existing online training methods**
  - Stochastic gradient descent (SGD)
  - Moderately fast, need 50 passes for training

3) Very fast online learning ***

- **Batch training**
  - Limited memory BFGS (LBFGS)
  - Too slow, need 300 passes for training
- **Existing online training methods**
  - Stochastic gradient descent (SGD)
  - Moderately fast, need 50 passes for training

- **Main proposal: a very fast online training method**
  - Finish the training in 10 passes

3) Very fast online learning ***

- **Potential problem of existing online training**
  - Over-simplified setting
  - Learning rate: use only 1 scalar for all weights

3) Very fast online learning ***

Potential problem of existing online training

- Over-simplified setting
- Learning rate: use only 1 scalar for all weights

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \boxed{\gamma_t} \nabla_{\boldsymbol{w}_t} \mathcal{L}_{stoch}(\boldsymbol{z}_i, \boldsymbol{w}_t)$$

3) Very fast online learning ***

- better to use different learning rates for different features?

3) Very fast online learning ***

- better to use different learning rates for different features?
- →YES. We use a vector of learning rates

3) Very fast online learning ***

- better to use different learning rates for different features?
- →YES. We use a vector of learning rates

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \boxed{\boldsymbol{\gamma}_t} \cdot \boldsymbol{g}_t$$

$$\boldsymbol{\gamma}_t \quad \in \quad \mathbb{R}_+^f$$

$$\boldsymbol{g}_t = \nabla_{\boldsymbol{w}_t} \mathcal{L}_{stoch}(\boldsymbol{z}_i, \boldsymbol{w}_t) = \nabla_{\boldsymbol{w}_t} \left\{ \ell(\boldsymbol{z}_i, \boldsymbol{w}_t) - \frac{||\boldsymbol{w}_t||^2}{2n\sigma^2} \right\}$$

3) Very fast online learning ***

- better to use different learning rates for different features?
- →YES. We use a vector of learning rates

3) Very fast online learning ***

- better to use different learning rates for different features?
- →YES. We use a vector of learning rates
- →Learning of Learning rates: How to learn the values of the learning rates?

3) Very fast online learning ***

- better to use different learning rates for different features?
- → YES. We use a vector of learning rates
- → Learning of Learning rates: How to learn the values of the learning rates?

*Learning rates are learned from feature frequency information

*Higher frequency feature → lower learning rate

3) Very fast online learning ***

ADF learning algorithm

procedure $\text{ADF}(q, c, \alpha, \beta)$
  $\boldsymbol{w} \leftarrow 0, t \leftarrow 0, \boldsymbol{v} \leftarrow 0, \boldsymbol{\gamma} \leftarrow c$
  repeat until convergence
    Draw a sample $\boldsymbol{z}_i$ at random
    $\boxed{\boldsymbol{v} \leftarrow \text{UPDATE}(\boldsymbol{v}, \boldsymbol{z}_i)}$
    if $t > 0$ and $t \bmod q = 0$
      $\boldsymbol{\gamma} \leftarrow \text{UPDATE}(\boldsymbol{\gamma}, \boldsymbol{v})$
      $\boldsymbol{v} \leftarrow 0$
    $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{w}} \mathcal{L}_{stoch}(\boldsymbol{z}_i, \boldsymbol{w})$
    $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{\gamma} \cdot \boldsymbol{g}$
    $t \leftarrow t + 1$
return $\boldsymbol{w}$

3) Very fast online learning ***

ADF learning algorithm

procedure ADF(q, c, α, β)

  $\boldsymbol{w} \leftarrow 0$, $t \leftarrow 0$, $\boldsymbol{v} \leftarrow 0$, $\boldsymbol{\gamma} \leftarrow c$

  repeat until convergence

    Draw a sample $\boldsymbol{z}_i$ at random

    $\boldsymbol{v} \leftarrow \textsc{Update}(\boldsymbol{v}, \boldsymbol{z}_i)$

    if $t > 0$ and $t \bmod q = 0$

      $\boxed{\boldsymbol{\gamma} \leftarrow \textsc{Update}(\boldsymbol{\gamma}, \boldsymbol{v})}$

      $\boldsymbol{v} \leftarrow 0$

    $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{w}} \mathcal{L}_{stoch}(\boldsymbol{z}_i, \boldsymbol{w})$

    $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{\gamma} \cdot \boldsymbol{g}$

    $t \leftarrow t + 1$

  return $\boldsymbol{w}$

3) Very fast online learning ***

ADF learning algorithm

procedure ADF(q, c, α, β)
  $\boldsymbol{w} \leftarrow 0, t \leftarrow 0, \boldsymbol{v} \leftarrow 0, \boldsymbol{\gamma} \leftarrow c$
  repeat until convergence
  . Draw a sample $\boldsymbol{z}_i$ at random
  . $\boldsymbol{v} \leftarrow \text{UPDATE}(\boldsymbol{v}, \boldsymbol{z}_i)$
  . **if** $t > 0$ and $t \bmod q = 0$
  . . $\boldsymbol{\gamma} \leftarrow \text{UPDATE}(\boldsymbol{\gamma}, \boldsymbol{v})$
  . . $\boldsymbol{v} \leftarrow 0$
  . $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{w}} \mathcal{L}_{stoch}(\boldsymbol{z}_i, \boldsymbol{w})$
  . $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boxed{\boldsymbol{\gamma}} \cdot \boldsymbol{g}$
  . $t \leftarrow t + 1$
  return $\boldsymbol{w}$

Convergence Analysis

- **Good convergence properties of the proposed method**

The ADF training is convergent

Theorem 1 *Assume ϕ is the largest eigenvalue of the function $\boldsymbol{C}_t = \prod_{m=1}^{t}(\boldsymbol{I} - \boldsymbol{\gamma}_0\beta^m \boldsymbol{H}(\boldsymbol{w}^*))$. For the proposed ADF training, its convergence rate is bounded by ϕ, and we have*

$$\phi \leq \exp\left\{\frac{\boldsymbol{\gamma}_0\lambda\beta}{\beta - 1}\right\},$$

where λ is the minimum eigenvalue of $\boldsymbol{H}(\boldsymbol{w}^*)$.

Experiments

- **Data**
- **Sighan bakeoff 2004**
  - Microsoft Research data (MSR)
  - Peking University data (PKU)
  - City University of Hongkong data (CU)

	#W.T.	#Word	#C.T.	#Char
MSR	8.8×10^4	2.4×10^6	5×10^3	4.1×10^6
CU	6.9×10^4	1.5×10^6	5×10^3	2.4×10^6
PKU	5.5×10^4	1.1×10^6	5×10^3	1.8×10^6

Experiments

Results

- Baseline: CRF with SGD training

Data	Method	Passes	Train-Time (sec)	NWD Rec	Pre	Rec	CWS F-score
MSR	Baseline	50	4.7e3	72.6	96.3	95.9	96.1
	+ New features	50	1.2e4	75.3	97.2	97.0	97.1
	+ New word detection	50	1.2e4	78.2	97.5	96.9	97.2
	+ ADF training	10	2.3e3	77.5	97.6	97.2	**97.4**
CU	Baseline	50	2.9e3	68.5	94.0	93.9	93.9
	+ New features	50	7.5e3	68.0	94.4	94.5	94.4
	+ New word detection	50	7.5e3	68.8	94.8	94.5	94.7
	+ ADF training	10	1.5e3	68.8	94.8	94.7	**94.8**
PKU	Baseline	50	2.2e3	77.2	95.0	94.0	94.5
	+ New features	50	5.2e3	78.4	95.5	94.9	95.2
	+ New word detection	50	5.2e3	79.1	95.8	94.9	95.3
	+ ADF training	10	1.2e3	78.4	95.8	94.9	**95.4**

Experiments

Results

- New feature (word feature + high dimensional edge features) helps

Data	Method	Passes	Train-Time (sec)	NWD Rec	Pre	Rec	CWS F-score
MSR	Baseline	50	4.7e3	72.6	96.3	95.9	96.1
	+ New features	50	1.2e4	75.3	97.2	97.0	97.1
	+ New word detection	50	1.2e4	78.2	97.5	96.9	97.2
	+ ADF training	10	2.3e3	77.5	97.6	97.2	**97.4**
CU	Baseline	50	2.9e3	68.5	94.0	93.9	93.9
	+ New features	50	7.5e3	68.0	94.4	94.5	94.4
	+ New word detection	50	7.5e3	68.8	94.8	94.5	94.7
	+ ADF training	10	1.5e3	68.8	94.8	94.7	**94.8**
PKU	Baseline	50	2.2e3	77.2	95.0	94.0	94.5
	+ New features	50	5.2e3	78.4	95.5	94.9	95.2
	+ New word detection	50	5.2e3	79.1	95.8	94.9	95.3
	+ ADF training	10	1.2e3	78.4	95.8	94.9	**95.4**

Experiments

Results

- Joint modeling (word seg + new word detection) helps slightly on word segmentation, but largely on NWD rec.

Data	Method	Passes	Train-Time (sec)	NWD Rec	Pre	Rec	CWS F-score
MSR	Baseline	50	4.7e3	72.6	96.3	95.9	96.1
	+ New features	50	1.2e4	75.3	97.2	97.0	97.1
	+ New word detection	50	1.2e4	78.2	97.5	96.9	97.2
	+ ADF training	10	2.3e3	77.5	97.6	97.2	**97.4**
CU	Baseline	50	2.9e3	68.5	94.0	93.9	93.9
	+ New features	50	7.5e3	68.0	94.4	94.5	94.4
	+ New word detection	50	7.5e3	68.8	94.8	94.5	94.7
	+ ADF training	10	1.5e3	68.8	94.8	94.7	**94.8**
PKU	Baseline	50	2.2e3	77.2	95.0	94.0	94.5
	+ New features	50	5.2e3	78.4	95.5	94.9	95.2
	+ New word detection	50	5.2e3	79.1	95.8	94.9	95.3
	+ ADF training	10	1.2e3	78.4	95.8	94.9	**95.4**

Experiments

Results

- ADF training greatly reduced training time, yet with even higher F-score!

Data	Method	Passes	Train-Time (sec)	NWD Rec	Pre	Rec	CWS F-score
MSR	Baseline	50	4.7e3	72.6	96.3	95.9	96.1
	+ New features	50	1.2e4	75.3	97.2	97.0	97.1
	+ New word detection	50	1.2e4	78.2	97.5	96.9	97.2
	+ ADF training	10	2.3e3	77.5	97.6	97.2	**97.4**
CU	Baseline	50	2.9e3	68.5	94.0	93.9	93.9
	+ New features	50	7.5e3	68.0	94.4	94.5	94.4
	+ New word detection	50	7.5e3	68.8	94.8	94.5	94.7
	+ ADF training	10	1.5e3	68.8	94.8	94.7	**94.8**
PKU	Baseline	50	2.2e3	77.2	95.0	94.0	94.5
	+ New features	50	5.2e3	78.4	95.5	94.9	95.2
	+ New word detection	50	5.2e3	79.1	95.8	94.9	95.3
	+ ADF training	10	1.2e3	78.4	95.8	94.9	**95.4**

Experiments

Results

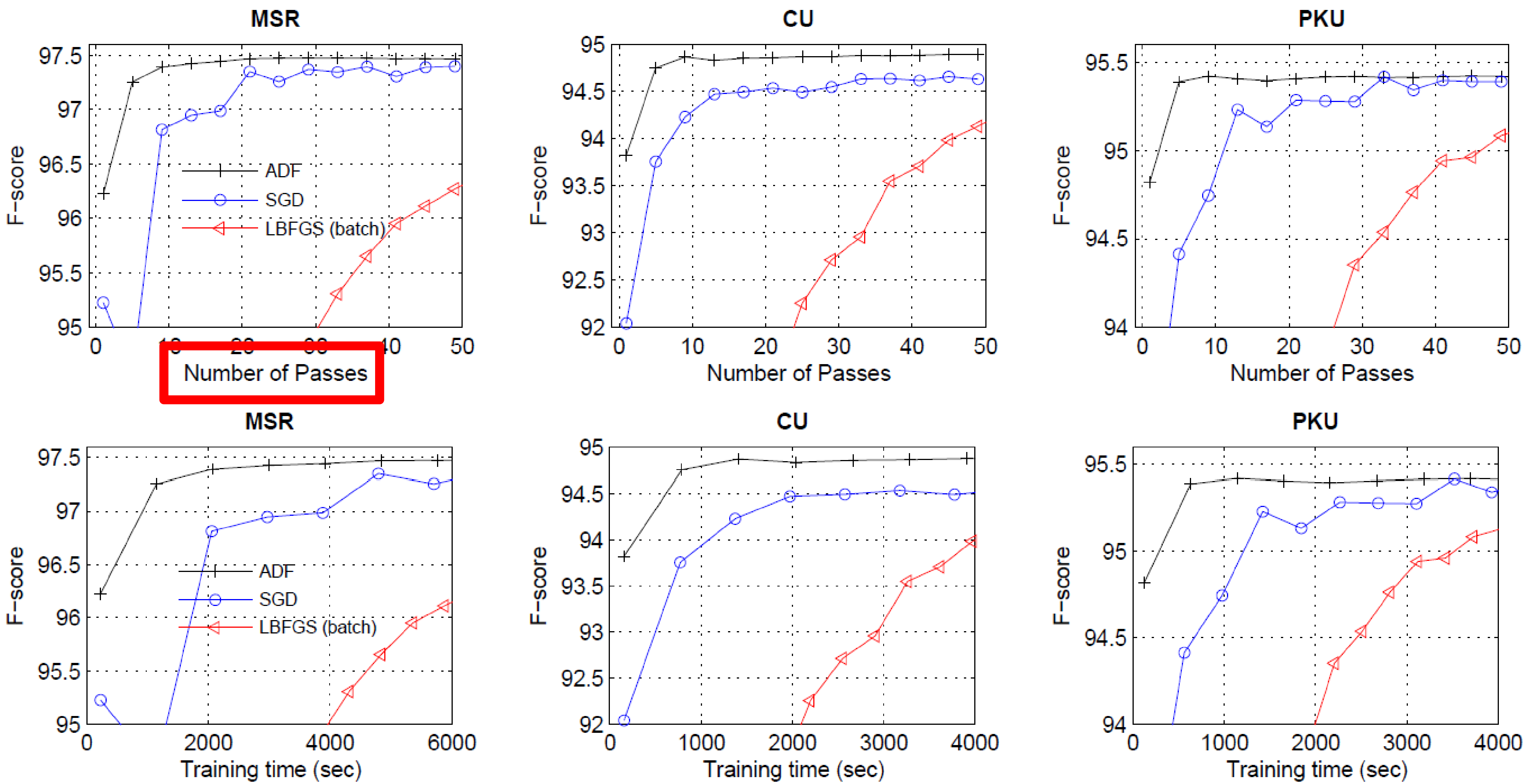
- ADF training greatly reduced training time, yet with even higher F-score!

Data	Method	Passes	Train-Time (sec)	NWD Rec	Pre	Rec	CWS F-score
MSR	Baseline	50	4.7e3	72.6	96.3	95.9	96.1
	+ New features	50	1.2e4	75.3	97.2	97.0	97.1
	+ New word detection	50	1.2e4	78.2	97.5	96.9	97.2
	+ ADF training	10	2.3e3	77.5	97.6	97.2	**97.4**
CU	Baseline	50	2.9e3	68.5	94.0	93.9	93.9
	+ New features	50	7.5e3	68.0	94.4	94.5	94.4
	+ New word detection	50	7.5e3	68.8	94.8	94.5	94.7
	+ ADF training	10	1.5e3	68.8	94.8	94.7	**94.8**
PKU	Baseline	50	2.2e3	77.2	95.0	94.0	94.5
	+ New features	50	5.2e3	78.4	95.5	94.9	95.2
	+ New word det						
	+ ADF training						

We achieved best accuracy reports on the MSR & PKU datasets

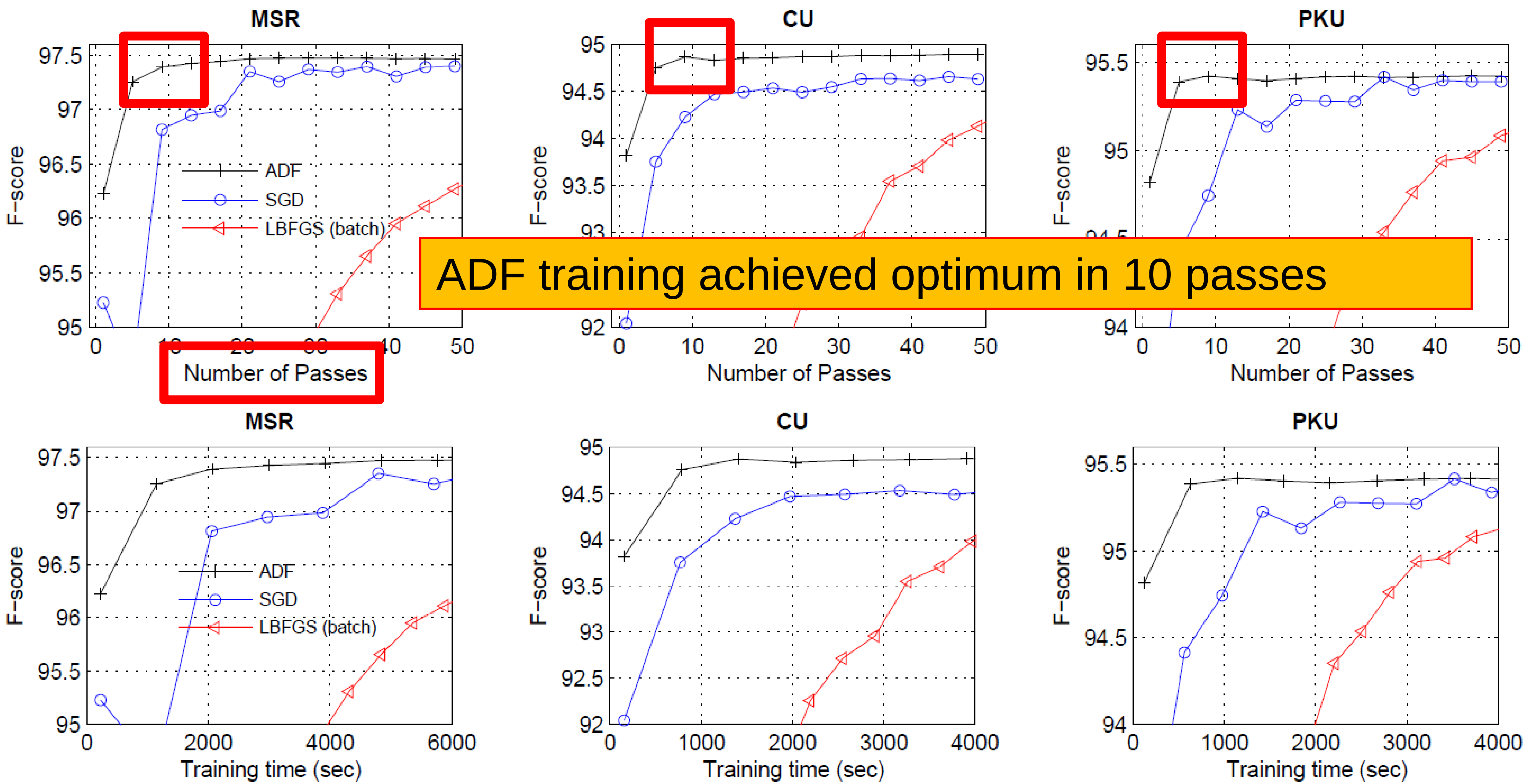
Experiments

◆New training ADF vs. SGD vs. LBFGS



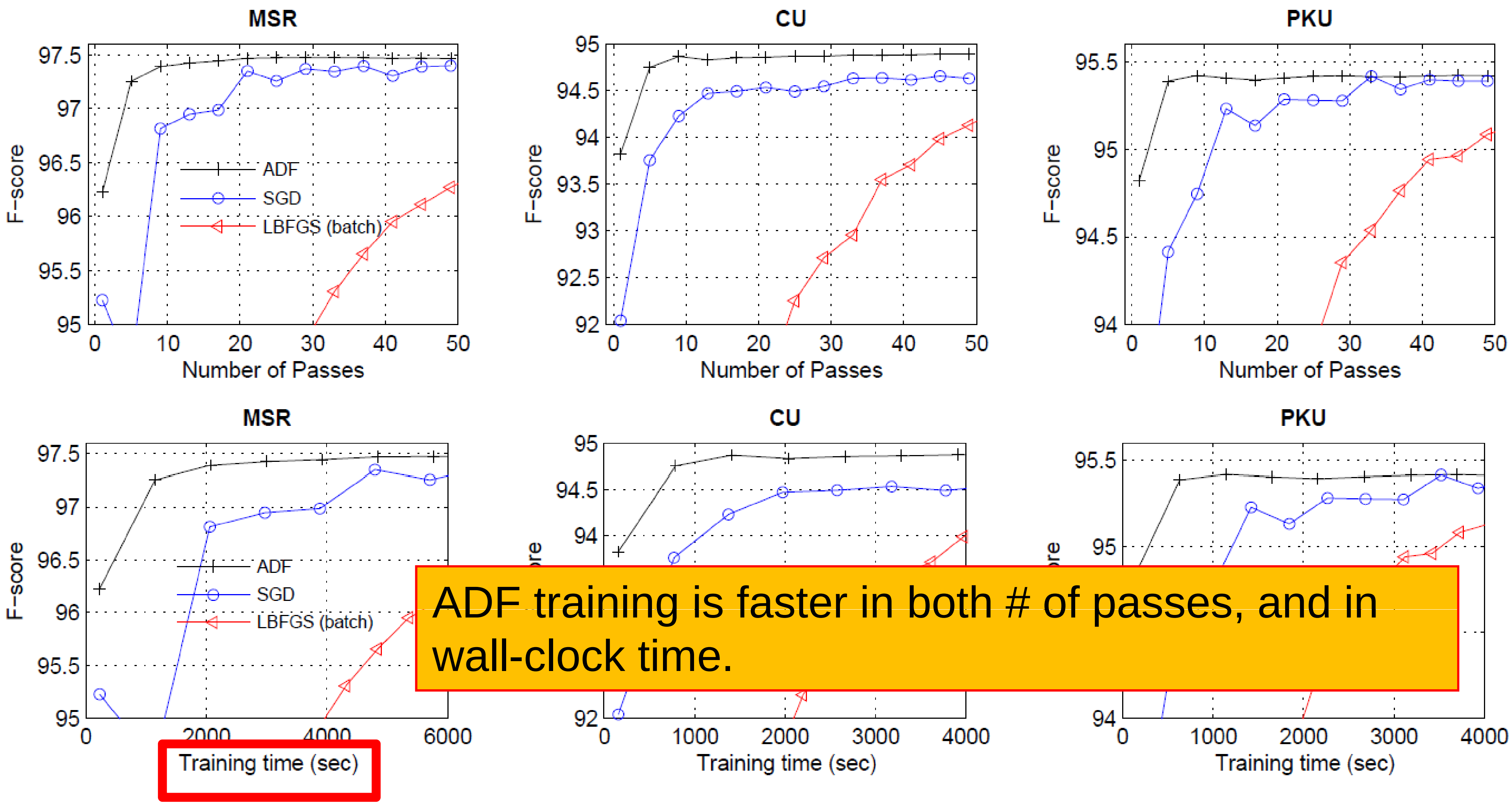
Experiments

◆New training ADF vs. SGD vs. LBFGS



Experiments

◆New training ADF vs. SGD vs. LBFGS



ADF training is faster in both # of passes, and in wall-clock time.

Conclusions

- **3 proposals**
  - Joint modeling: segmentation + new word detection
  - New features: word features + high dimensional edge features
  - A very fast online training method (main proposal)

Conclusions

- **3 proposals**
  - Joint modeling: segmentation + new word detection
  - New features: word features + high dimensional edge features
  - A very fast online training method (main proposal)
- **Experiments**
  - Joint modeling helps

Conclusions

- **3 proposals**
  - Joint modeling: segmentation + new word detection
  - New features: word features + high dimensional edge features
  - A very fast online training method (main proposal)
- **Experiments**
  - Joint modeling helps
  - New features helps a lot on model accuracy

Conclusions

- **3 proposals**
  - Joint modeling: segmentation + new word detection
  - New features: word features + high dimensional edge features
  - A very fast online training method (main proposal)
- **Experiments**
  - Joint modeling helps
  - New features helps a lot on model accuracy
  - The new training method finishes the training in 10 passes

Conclusions

- 3 proposals
  - Joint modeling: segmentation + new word detection
  - New features: word features + high dimensional edge features
  - A very fast online training method (main proposal)
- Experiments
  - Joint modeling helps
  - New features helps a lot on model accuracy
  - The new training method finishes the training in 10 passes
  - Final results beat the existing best reports on the datasets
    - More accurate & faster!

- **Thanks!**
- **Any question?**