IBM Research, Tokyo Research Laboratory

A New Objective Function for Sequence Labeling

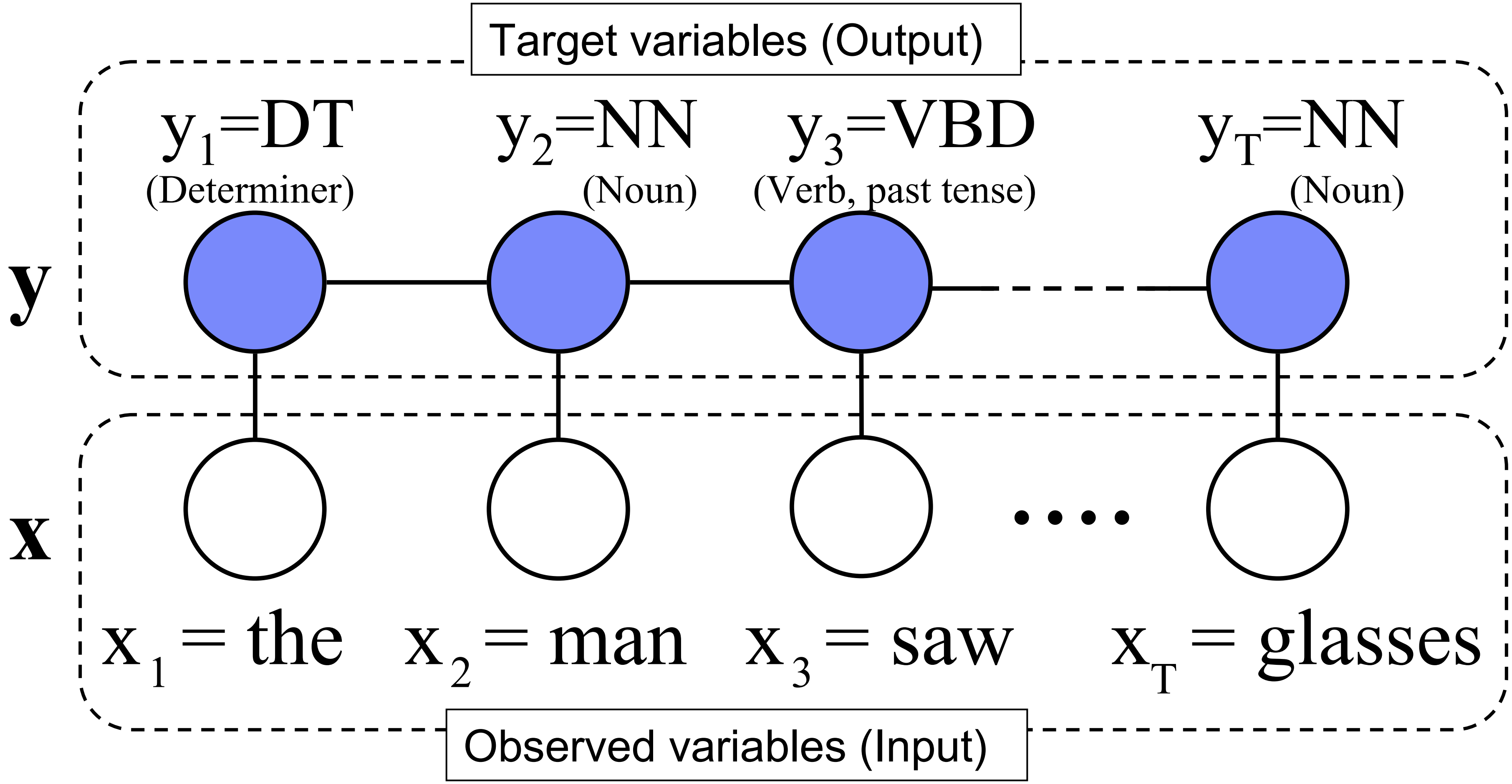
Yuta Tsuboi and Hisashi Kashima
IBM Research, Tokyo Research Laboratory

Outline

- Sequence labeling problem
  - An application in natural language processing
  - Supervised learning of sequence labeling
- Previous work
  - Conditional Random Fields (CRFs)
  - Two objective functions for sequence labeling: Sequential loss & Pointwise loss
- A new objective function with Markov property
  - A motivating application: an information extraction task
  - Mixed loss & Markov loss
- Experiment

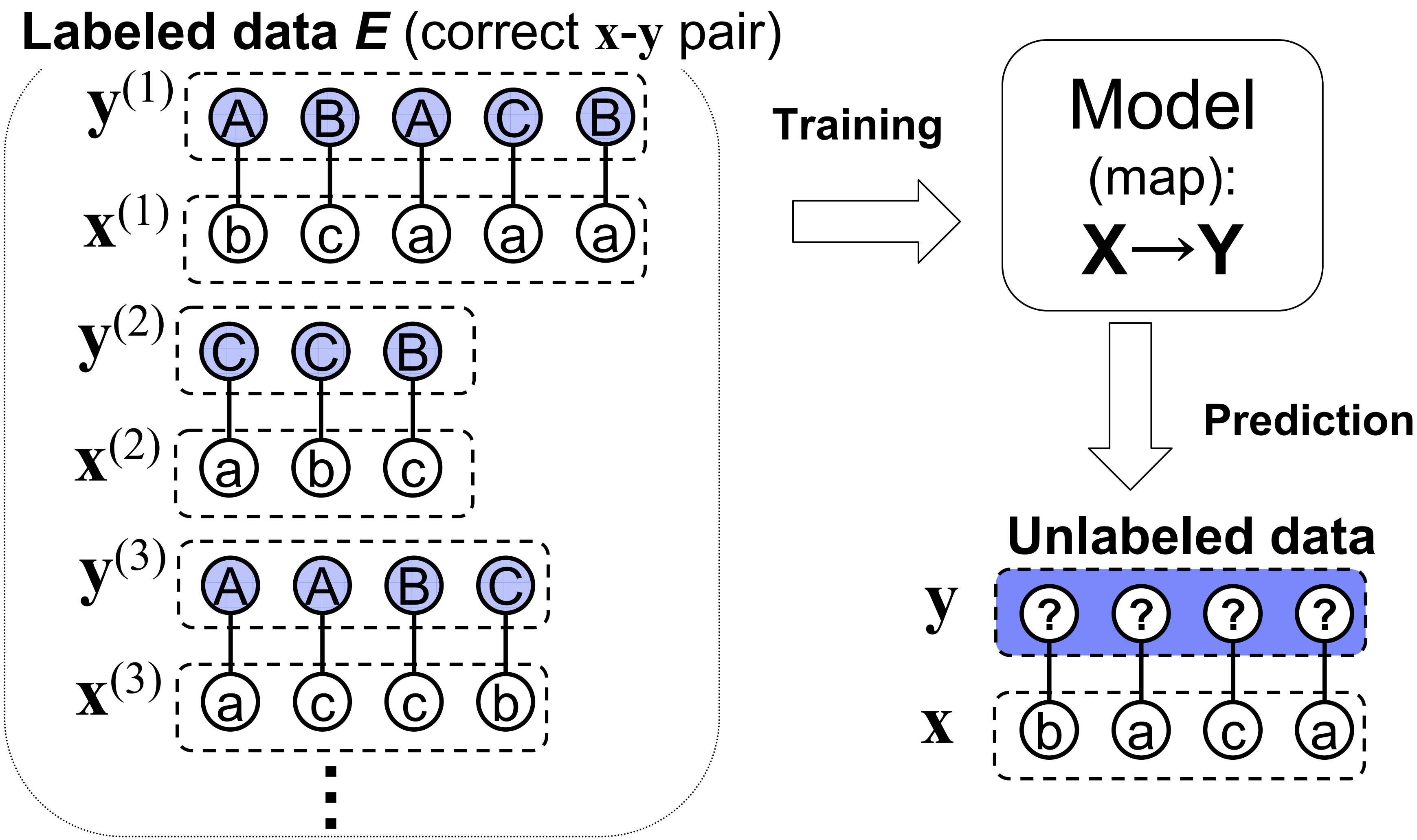
Applications of sequence labeling
Part-of-speech (POS) tagging task

- Predicting part-of-speech tags of words in a sentence.



Supervised learning of sequence labeling

Training a statistical model using correct pairs of an input x and a label sequence y.



Outline

- Sequence labeling problem
  - An application in natural language processing
  - Supervised learning of sequence labeling
- Previous work
  - Conditional Random Fields (CRFs)
  - Two objective functions for sequence labeling: Sequential loss & Pointwise loss
- A new objective function with Markov property
  - A motivating application: an information extraction task
  - Mixed loss & Markov loss
- Experiment

State of the art sequence labeler
Conditional Random Fields: CRFs

- Modeling conditional probability $\Pr(\mathbf{y}|\mathbf{x})$ of an entire label sequence $\mathbf{y}$ over a given input $\mathbf{x}$.

$$f_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) = \frac{\exp\left(\langle \boldsymbol{\theta}, \phi(\mathbf{x}, \mathbf{y}) \rangle\right)}{\sum_{\tilde{\mathbf{y}}} \exp\left(\langle \boldsymbol{\theta}, \phi(\mathbf{x}, \tilde{\mathbf{y}}) \rangle\right)}.$$

$\phi : \mathbf{X} \times \mathbf{Y} \rightarrow \Re^d$ **:** a map from a pair of $\mathbf{x}$ and $\mathbf{y}$ to a feature vector

$\boldsymbol{\theta} \in \Re^d$: the vector of model parameters (weight vector).

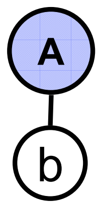
CRFs are the generalization of multinomial logistic regressions.

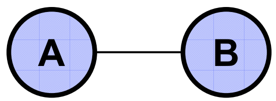
The advantage of CRFs for sequence labeling

- We can represent the consistency of a target variable sequence by features ϕ_{yy} (consecutive target variables y_{t-1} and y_t).

Feature vector of a whole sequence of length T

$$\phi(\mathbf{x},\mathbf{y}) = \sum_{t=1}^{T}\left(\phi_{xy}(\mathbf{x},y_t) + \phi_{yy}(y_{t-1},y_t)\right)$$





Observation feature

Transition feature

Previous work: Two objective functions of training CRFs

- Sequential loss function (Lafferty et al., 2001)
  - Maximizes the likelihood of whole label sequence of each example.

$$L_1 = -\sum_{i=1}^{|E|} \log f_{\boldsymbol{\theta}}(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)})$$

⇨ Maximizing the number of correctly predicted *sequences*

- Pointwise loss function (Kakade et al., 2002)
  - Maximizes the likelihood of each labels in the sequences.

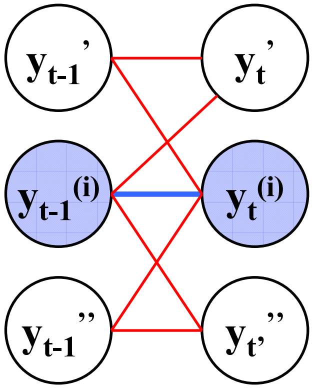
$$L_0 = -\sum_{i=1}^{|E|} \sum_{t}^{T^{(i)}} \log \sum_{\tilde{\mathbf{y}}:\tilde{y}_t = y_t^{(i)}} f_{\boldsymbol{\theta}}(\tilde{\mathbf{y}} \mid \mathbf{x}^{(i)})$$

Marginalize all the possible label assignments with fixed label $y_t^{(i)}$ at the t-th position

⇨ Maximizing the number of correctly predicted *variables*

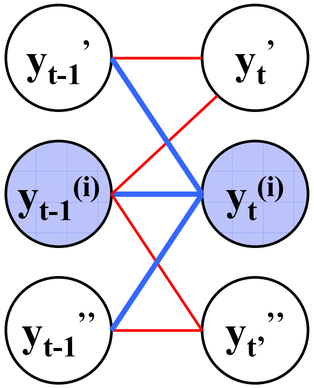
Weight updates of transition features under sequential loss and pointwise loss.

Sequential loss



Transitions occurred in the training samples.

Pointwise loss



The blue edges indicate rewarded transitions, and the red edges indicate punished transitions in the maximum likelihood estimation.

- Sequential loss function (L_1): A large negative weight will be given to features not observed in the training set.
- Pointwise loss function (L_0): does not care consistencies among consecutive labels

Motivation:

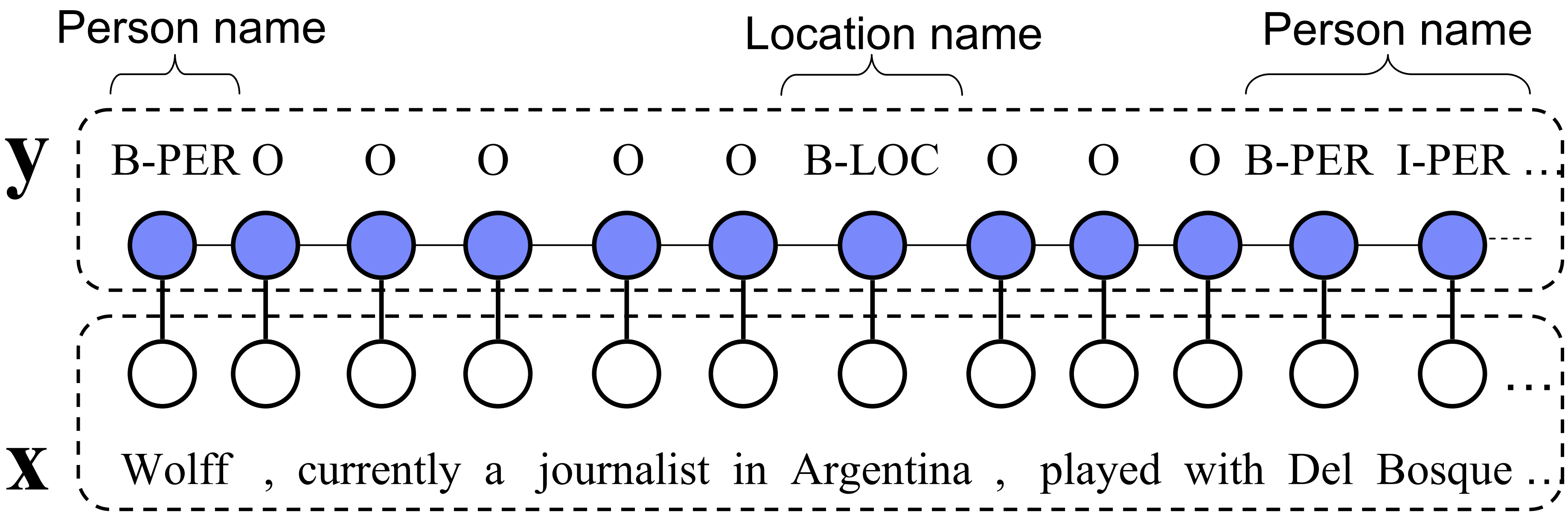
Needs of loss functions to predict each segment correctly

Outline

- Sequence labeling problem
  - An application in natural language processing
  - Supervised learning of sequence labeling
- Previous work
  - Conditional Random Fields (CRFs)
  - Two objective functions for sequence labeling: Sequential loss & Pointwise loss
- A new objective function with Markov property
  - A motivating application: an information extraction task
  - Mixed loss & Markov loss
- Experiment

Motivating applications of sequence labeling
Named Entity Recognition (NER) task

- An information extraction task to extract phrases containing names of *persons* (PER), *organizations*, *locations* (LOC), *times* and *quantities* in texts.
- Labeling each word by either of "*beginning* (B-x)", "*continuation* (I-x)" and "*non-named entities* (O)".



Outline of the proposed loss function

- We propose two equivalent forms of a new loss function which is suitable for information extraction tasks.
- λ-mixed loss function: Intermediate between sequential loss and pointwise loss.

equivalent

- k-th order Markov loss function: The loss at position t depends only on the labels of the next k positions.

λ-mixed loss function

- Linear combination of sequential loss (L_1) and pointwise loss (L_0) with mixing parameter λ

Sequential loss **Pointwise loss**

$$\mathrm{L}_\lambda := \lambda\, \mathrm{L}_1 + (1-\lambda)\, \mathrm{L}_0$$

$$= -\sum_{i=1}^{|E|}\left(\lambda \log \mathrm{f}_{\boldsymbol{\theta}}\left(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}\right) + (1-\lambda)\sum_{t}^{T^{(i)}} \log \sum_{\tilde{\mathbf{y}}:\tilde{y}_t = y_t^{(i)}} \mathrm{f}_{\boldsymbol{\theta}}\left(\tilde{\mathbf{y}} \mid \mathbf{x}^{(i)}\right)\right),$$

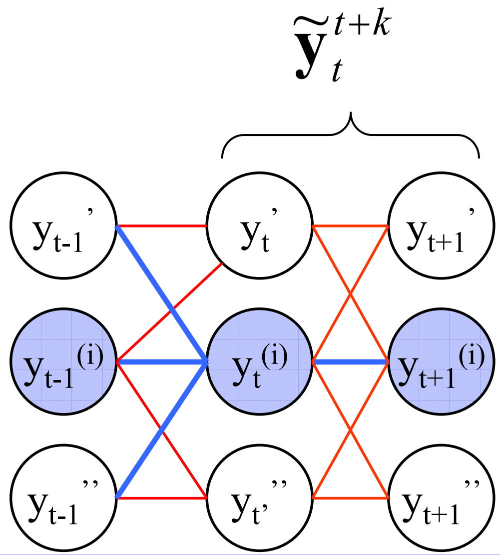
$$(0 \le \lambda \le 1).$$

k-th order Markov loss function

- The summation of marginalized negative log-likelihood of all the possible label assignments with fixed target segment $y^{(i)}{}_t^{t+k}$ of length $k+1$ at the t-th position.

$$\mathrm{M}_k := -\sum_i \sum_{t=-k+1}^{T^{(i)}} \log \sum_{\tilde{\mathbf{y}}:\tilde{\mathbf{y}}_t^{t+k}=\mathbf{y}^{(i)}{}_t^{t+k}} \mathrm{f}_{\boldsymbol{\theta}}\left(\tilde{\mathbf{y}} \mid \mathbf{x}^{(i)}\right).$$

Weight updates of transition features under k=1st order Markov loss



maximization

Tries to correctly predict as many segments $\mathbf{y}_t^{t+k}$ as possible.

Markov property of λ-mixed loss function (1)

- For a integer $k > 0$, the minimization of λ-mixed loss function is equivalent to the minimization of k-th order Markov loss function.
- Theorem 1:

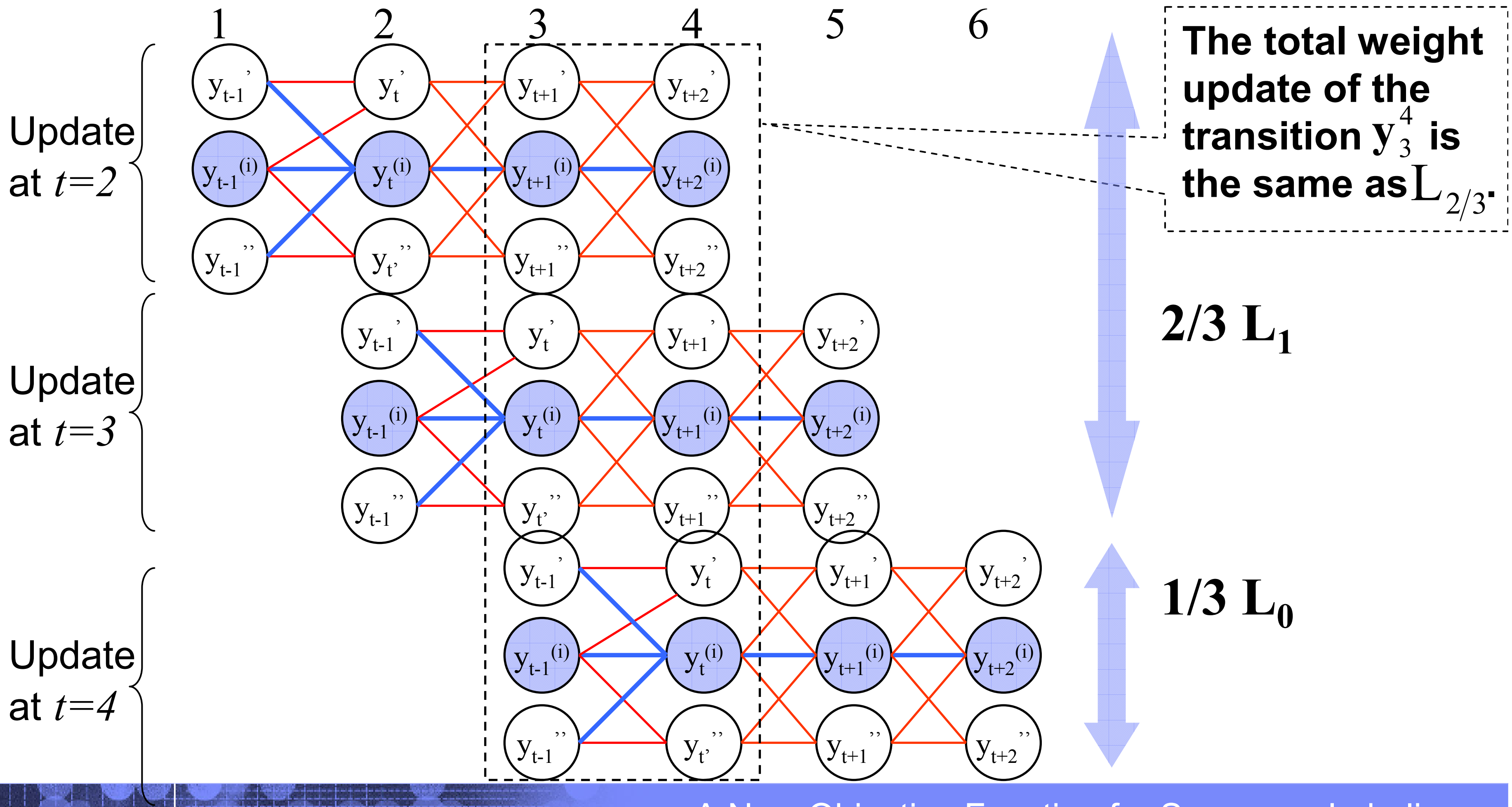
$$\lambda = \frac{k}{k+1},$$

Then,

$$\underbrace{\mathrm{M}_{\mathrm{k}}}_{k\text{-th order Markov loss}} = \frac{1}{1-\lambda} \underbrace{\mathrm{L}_{\lambda}}_{\lambda\text{-mixed loss}}.$$

Intuitive explanation of the relationship between Markov loss and λ-mixed loss $\left(L_\lambda := \lambda L_1 + (1-\lambda) L_0\right)$.

- An example of weight updates of transition features under a $k=2^{nd}$ order Markov loss $\left(\lambda = k/(k+1) = 2/3\right)$.



Markov property of λ-mixed loss function (2)

- λ-mixed loss function is equivalent to a weighted sum of Markov losses with exponentially decaying weights.
  - Corollary. For any $0 < \lambda < 1$,

λ-mixed loss

Infinite sum of Markov loss

$$\frac{1}{1-\lambda} \underbrace{L_{\lambda}}_{\lambda\text{-mixed loss}} = (1-\lambda)\underbrace{\sum_{\kappa=0}^{\infty} \lambda^{\kappa} M_{\kappa}}_{\text{Infinite sum of Markov loss}}.$$

➡ λ-mixed loss function is intended for all the lengths of segments while giving them weights depending on their lengths.

Summary of the proposed loss function

λ-mixed loss function: L_λ

k-th order Markov loss function: M_k

- Interpretation 1

$$\mathrm{M}_\mathrm{k} = \frac{1}{1-\lambda}\mathrm{L}_\lambda\left(\lambda = \frac{k}{k+1}\right).$$

- Interpretation 2

$$\frac{1}{1-\lambda}\mathrm{L}_\lambda = (1-\lambda)\sum_{\kappa=0}^{\infty}\lambda^\kappa \mathrm{M}_\kappa.$$

- The minimization of λ-mixed loss function tries to predict each segment correctly.
  - Suitable for information extraction tasks such as:
    - named entity recognition: finds local segments indicating named entities
    - protein secondary structure prediction: finds local segments indicating alpha helices and beta sheets regions.

Outline

- Sequence labeling problem
  - An application in natural language processing
  - Supervised learning of sequence labeling
- Previous work
  - Conditional Random Fields (CRFs)
  - Two objective functions for sequence labeling: Sequential loss & Pointwise loss
- A new objective function with Markov property
  - A motivating application: an information extraction task
  - Mixed loss & Markov loss
- Experiment

Experimental Setup

- Named entity extraction (NER) task
  - CoNLL2002 shared task on NER
  - 9 labels to indicate *person name*, *organization name*, *place*, and *names of miscellaneous entities*.
  - Using word and spelling features (S2 feature in [Altun et al. 2003]) for observation.
  - Sentences (tokens) of standard experiment settings
    - Training set: 8,322 (264,680)
    - Development set: 1,914 (52,849)
    - Test set: 1,516 (51,487)

Results: Evaluation with gold-standard settings

- Trained CRFs using training data based on λ-mixed loss function.
- Tuned regularization parameter (σ) using development data (model selection).
- Evaluated by F1 measure on test data.

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Best performance at $k=3$

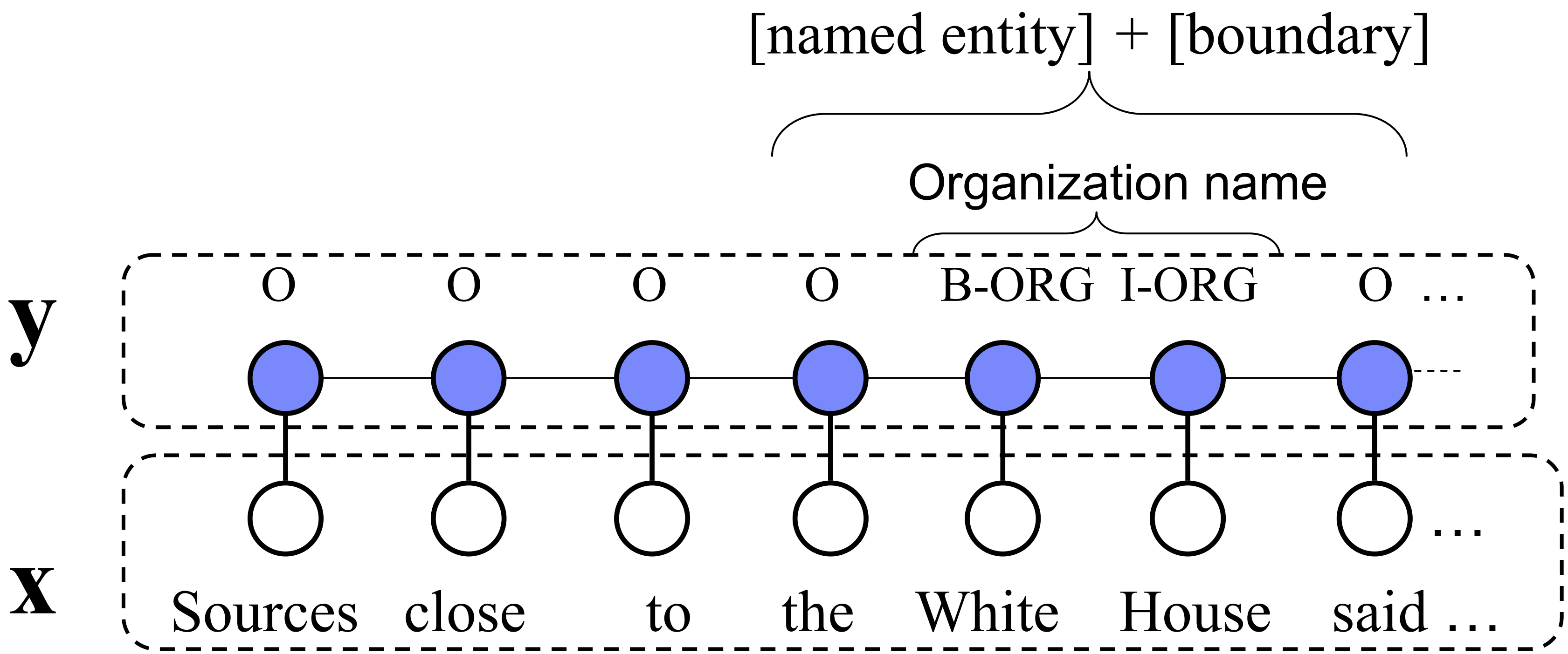
	Pointwise loss	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	Sequence loss
best σ	1.8	1.8	1.6	1.6	1.4	1.6	1.6
Precision	77.91	77.96	77.95	**78.10**	78.03	77.91	**78.10**
Recall	76.71	76.85	76.88	**76.96**	76.85	76.82	76.85
F1	77.30	77.40	77.41	**77.53**	77.43	77.36	77.47

Markov loss interpretation of the proposed loss $\lambda = \frac{k}{k+1}$.

Interpretation of this empirical result

- The best performance at $k=3$ agrees with our intuitions since [named entity length] + [boundary length=2] - 1 represents proper local consistency to recognize the boundaries of segment.

←Average phrase length of named entities in the data set ≒2



Conclusion

- We show the “Markov property” of the mixed loss between sequential and pointwise loss, that is the importance of correct labeling for a particular position depends on the numbers of the correct labels around there in sequence labeling.

λ-mixed loss function: L_λ

k-th order Markov loss function: M_k

- Interpretation 1

$$\mathrm{M}_\mathrm{k} = \frac{1}{1-\lambda}\mathrm{L}_\lambda\left(\lambda = \frac{k}{k+1}\right).$$

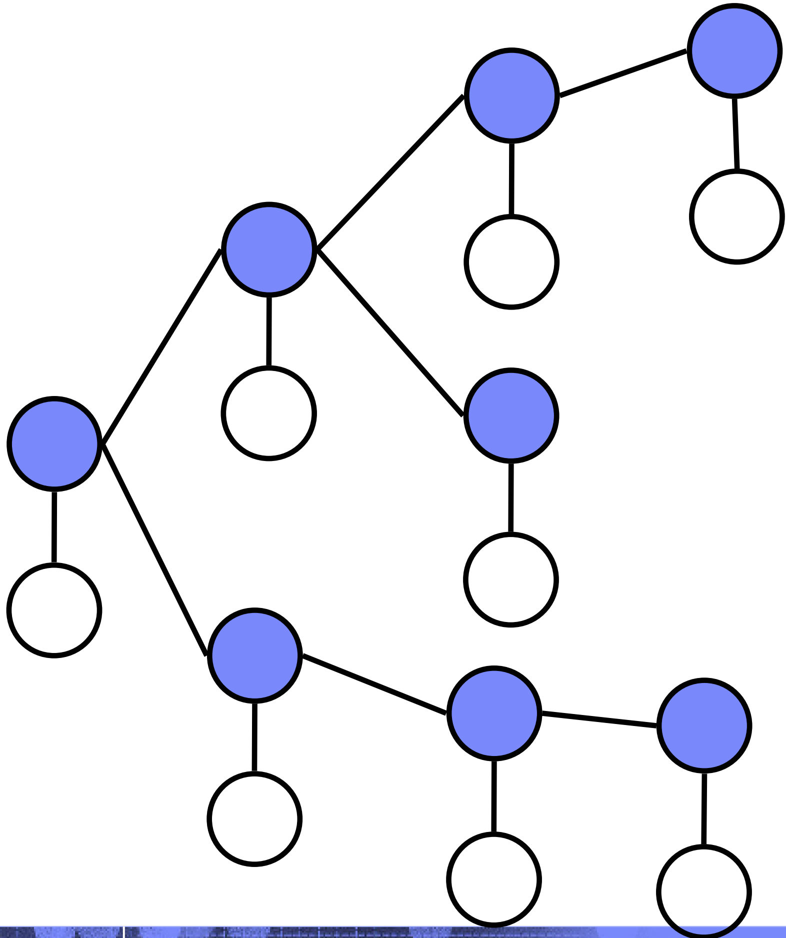
- Interpretation 2

$$\frac{1}{1-\lambda}\mathrm{L}_\lambda = (1-\lambda)\sum_{\kappa=0}^{\infty}\lambda^\kappa \mathrm{M}_\kappa.$$

End of presentation

More complex structures?

- Theorem 1 holds for data with rooted tree structures.



- However, for more general graph-structured data, we have no clear correspondence between L_λ and other objective functions so far.

Results (2) Emphasis for contrast

- Training CRFs without Gaussian prior
- Average of F1 measure of dev. & test data set.

The proposed works well for a relatively small data set.

Training Set Size	Loss Function point	k=1	k=2	k=3	k=4	seq
100	45.36	46.12	**46.96**	46.94	42.72	43.96
200	47.76	47.39	47.44	**47.77**	47.30	47.16
300	**53.37**	52.91	52.68	52.92	52.86	52.40
600	**59.32**	58.68	58.25	58.11	57.34	56.00
1000	61.26	**61.91**	61.38	61.33	61.34	61.05