Training Conditional Random Fields Using Incomplete Annotations

Yuta Tsuboi *1*3**, Hisashi Kashima** *1**, Shinsuke Mori** *2**,**
Hiroki Oda, and **Yuji Matsumoto** *3.

*1 IBM Research, Tokyo Research Laboratory.

*2 Academic Center for Computing and Media Studies, Kyoto University

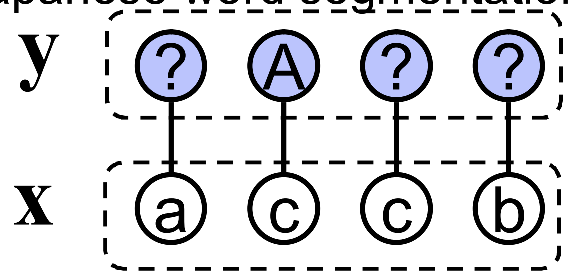
*3 Graduate School of Information Science, Nara Institute of Science and Technology

Contents

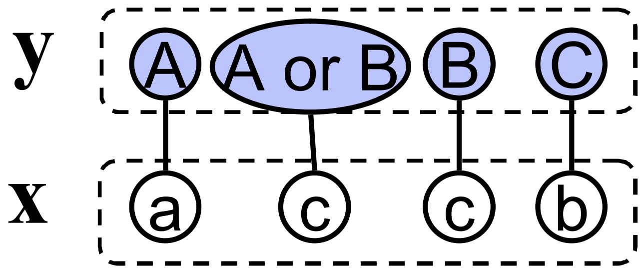
- Incomplete annotations in corpus building.
  - Partial annotations & Ambiguous annotations
  - Word segmentation & Part-of-speech tagging task
- Training CRFs using Incomplete annotations
  - Representation of incomplete annotations
  - Supervised learning setting
  - Marginal likelihood for CRFs
- Experiments
  - A domain adaptation task of Japanese word segmentation using **partial annotations** by domain-specific word lists
  - POS tagging task using **ambiguous annotations** which are contained in Penn treebank corpus.

Incomplete Annotations in Corpus Building

- Partial annotations
  - Some parts of a sentence are manually annotated.
  - e.g. the domain adaptation task of Japanese word segmentation

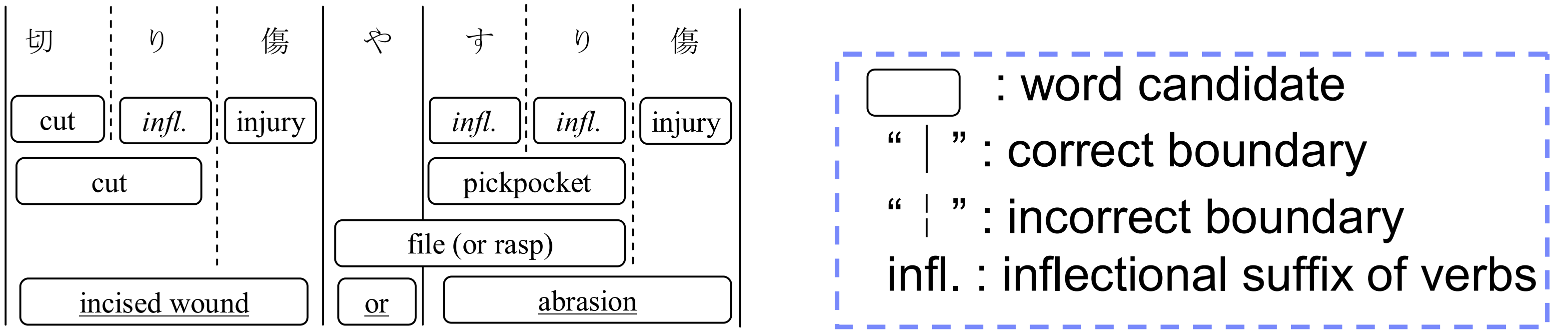


- Ambiguous annotations
  - Some parts of a sentence are annotated by a set of candidate labels instead of a single label.
  - e.g. POS tags in Penn treebank corpus.



Background
Word Segmentation & Part-of-speech Tagging Task

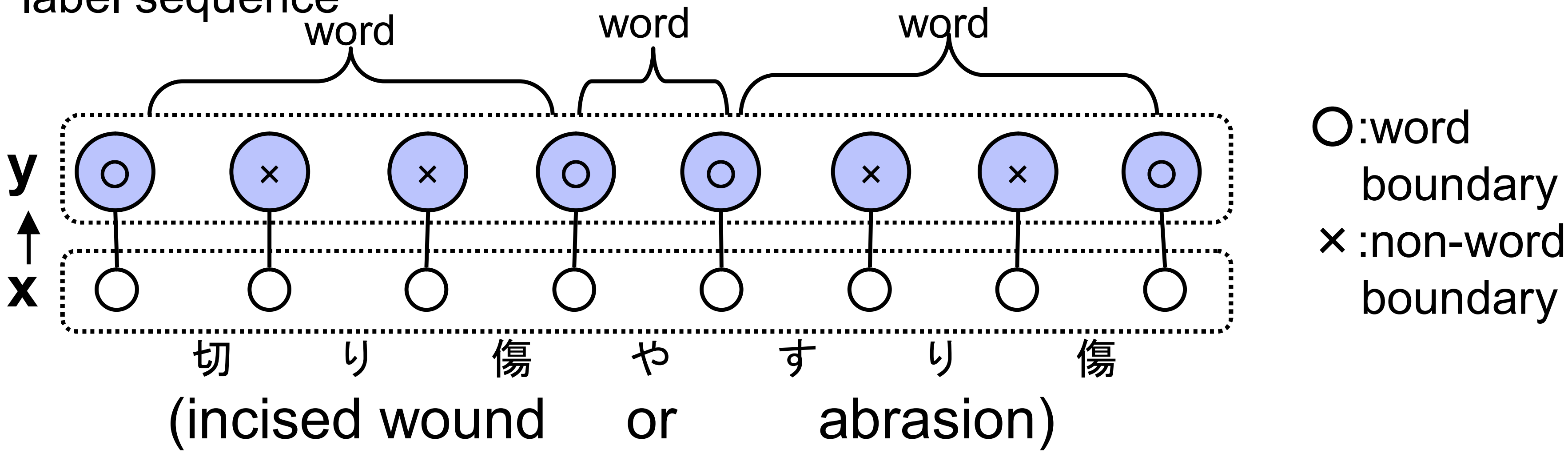
- **Word Segmentation Task：detecting word boundaries for non-segmented languages, such as Japanese, and Chinese.**
  - e.g. Japanese phrase ``切り傷やすり傷“ (incised wound or abrasion):



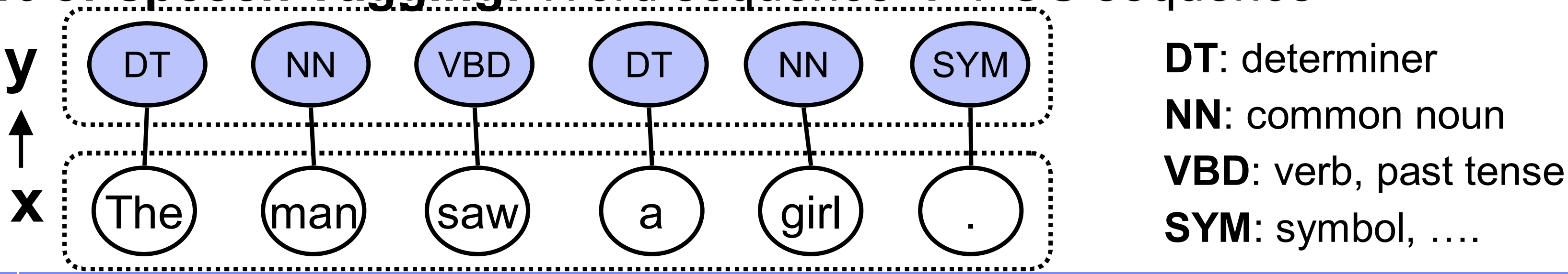
- **Part-of-speech Tagging Task：identifying words as nouns, verbs, adjectives, adverbs, etc.**
  - .English: flies → verb or noun?
- **Statistical methods are commonly used for these problems.**

Background
Word Segmentation & Part-of-speech Tagging as Structured Output Prediction

- **Word Segmentation :** Character boundary sequence → Word boundary label sequence

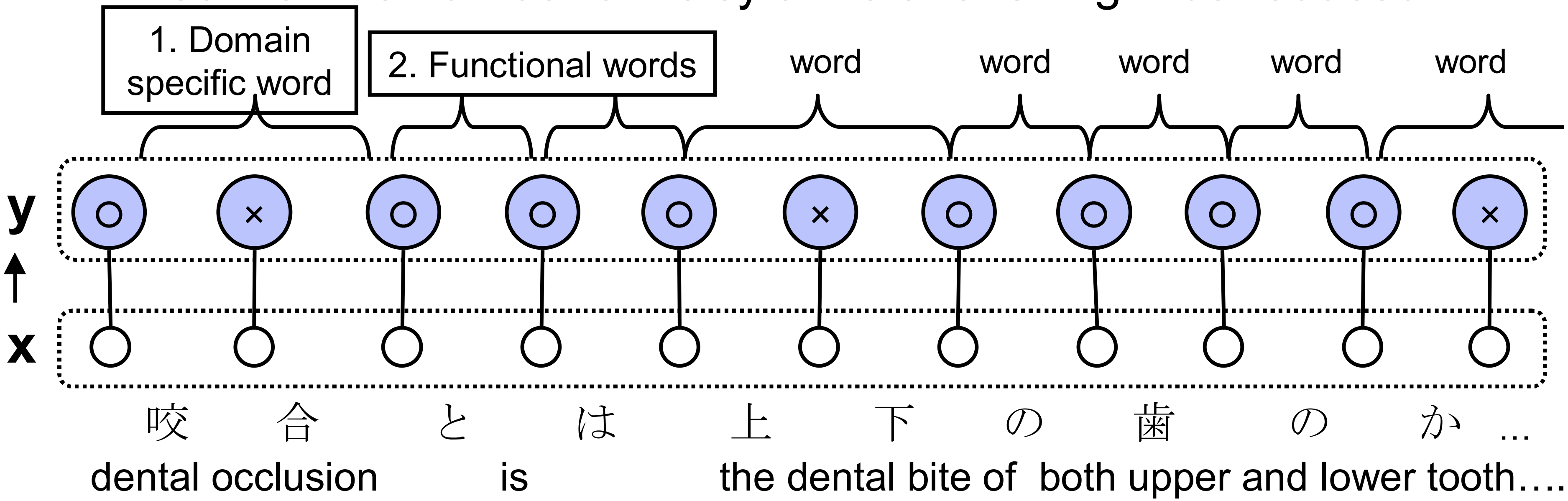


- **Part-of-speech Tagging:** Word sequence → POS sequence



An example of partial annotations
In the situation of domain adaptation, it is useful to allow partial annotations.

1. Annotators can concentrate on the important parts of sentences, which can be identified by domain-specific resources or active learning techniques.
2. Linguistically complicated parts can be left without annotation so that the number of noisy annotations might be reduced.



An example of ambiguous annotations Penn treebank English corpus includes more than 100 sentences containing POS ambiguities

- Frequent POS ambiguous words in Penn treebank corpus (Wall Street Journal).

frequency	word	POS tags
15	data	NN|NNS
10	more	JJR|RBR
7	pending	JJ|VBG
4	than	IN|RB

- Ambiguous annotations are more common in the tasks which deal with semantics, such as information extraction tasks.

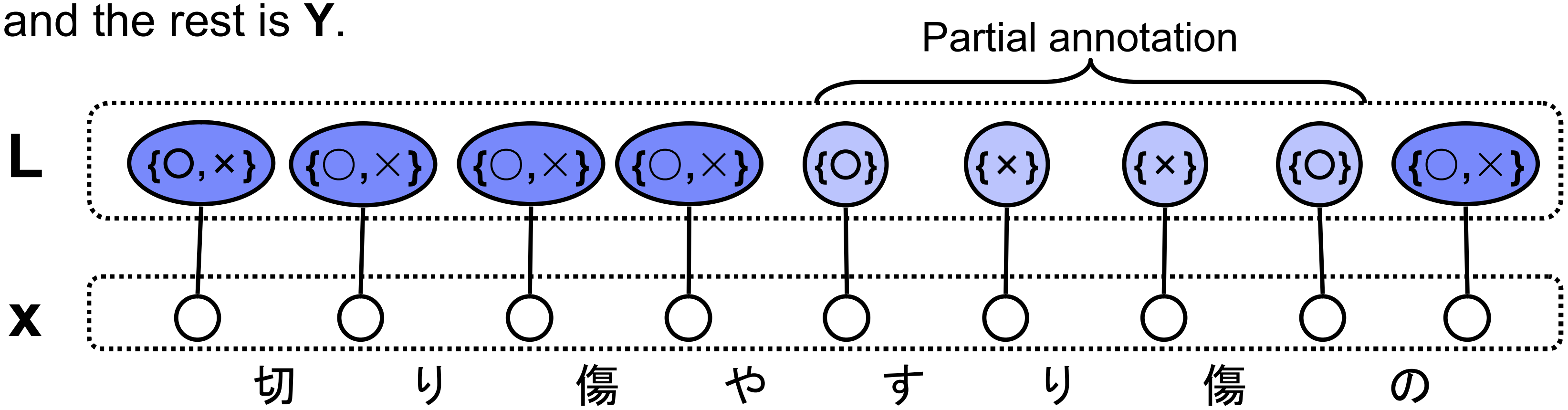
Contents

- Incomplete annotations in corpus building.
  - Partial annotations & Ambiguous annotations
  - Word segmentation & Part-of-speech tagging task
- Training CRFs using Incomplete annotations
  - Representation of incomplete annotations
  - Supervised learning setting
  - Marginal likelihood for CRFs
- Experiments
  - A domain adaptation task of Japanese word segmentation using **partial annotations** by domain-specific word lists
  - POS tagging task using **ambiguous annotations** which are contained in Penn treebank corpus.

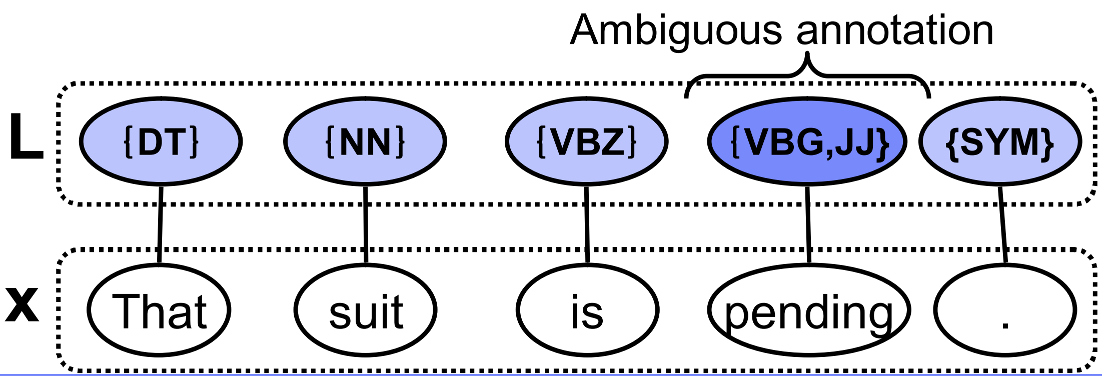
Representation for partial and ambiguous annotations
a sequence of the possible value set:

$$\mathrm{L} = \left(L_t \subseteq Y \text{ for } t = 1 \cdots T\right)$$

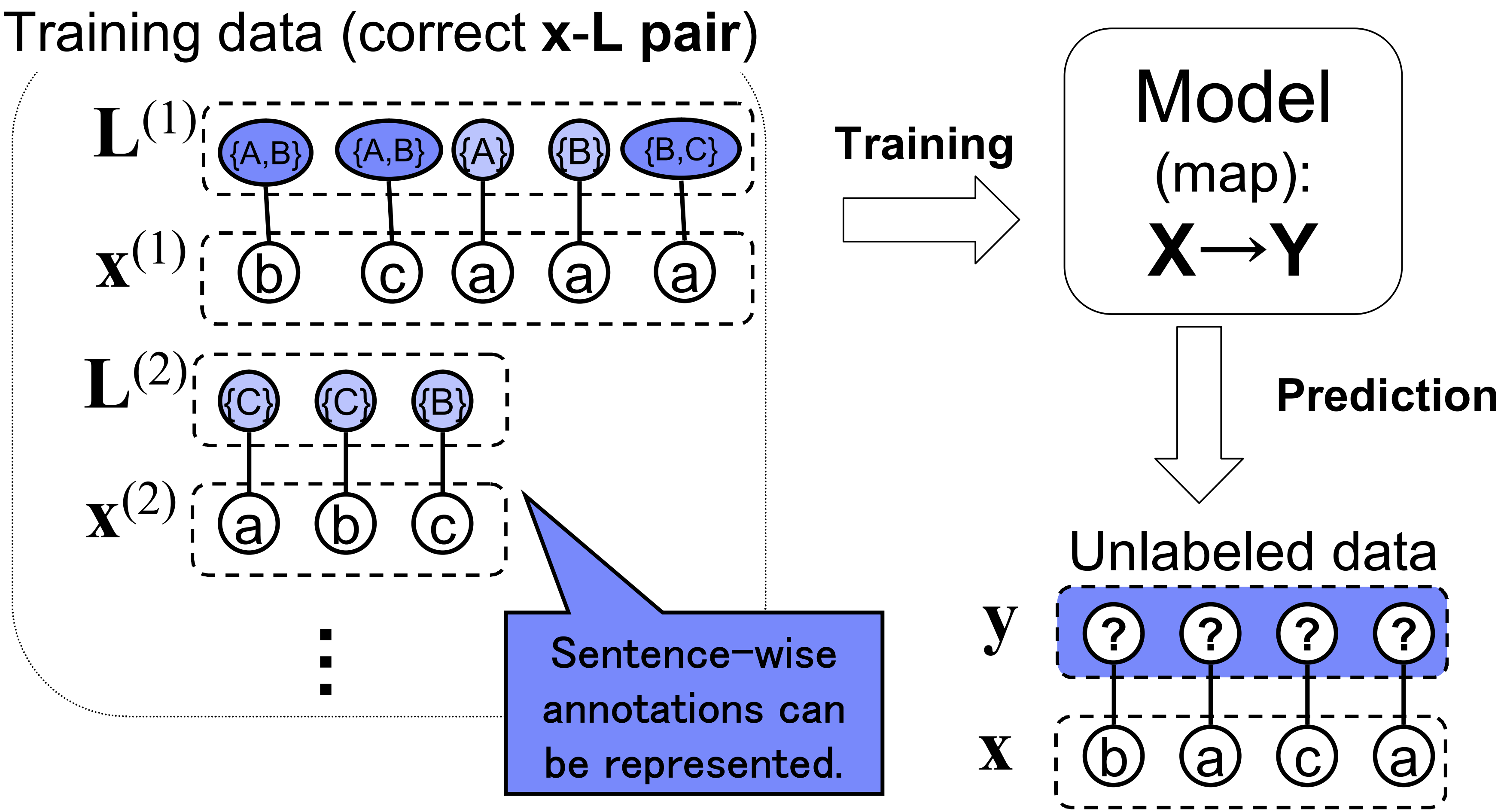
- **Partial Annotations**
  - The partial annotation at position t is a case where the set L_t is a singleton and the rest is **Y**.



- **Ambiguous Annotations**
  - L_t represents a set of candidate labels at the position t.



Supervised learning using incomplete annotations Training data is pairs of input x and label set sequence $\mathrm{L} = \left(L_t \subseteq Y \text{ for } t = 1 \cdots T\right)$.



Conditional Random Fields (CRFs)
State of the art model for structured output prediction

CRFs model the conditional probability $\mathrm{P}_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$ of a label sequence **y** given an observed sequence **x**.

$$\mathrm{P}_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) = \frac{\overbrace{\exp\left(\left\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{x}, \mathbf{y}) \right\rangle\right)}^{\text{The score of y}}}{\underbrace{\sum_{\tilde{\mathbf{y}} \in \mathbf{Y}} \exp\left(\left\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{x}, \tilde{\mathbf{y}}) \right\rangle\right)}_{\text{The summed score of all the possible ys. (Efficient computation algorithm is known)}}}$$

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

$\theta \in \mathbf{R}^d$: the vector of model parameters.

Prediction: $\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathbf{Y}} \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$

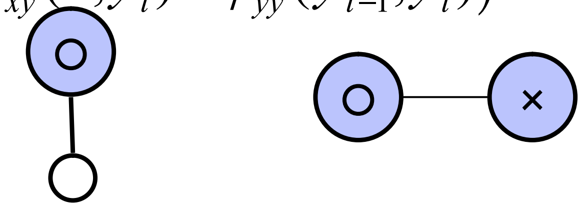
The advantage of CRFs in NLP
Supporting over-wrapping features and label correlations

- Advantage of discriminative model
  - Using freely correlated features, such as both unigram and bigram, or substrings and string itself

  ←→ In generative model, it is hard to estimate the joint probability $p(\mathbf{x}, \mathbf{y})$ of these features from limited samples

- Advantage of structured output learning
  - Representing correlations between elements in the output structure (e.g. y_{t-1} and y_t) as feature ϕ_{yy}.

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

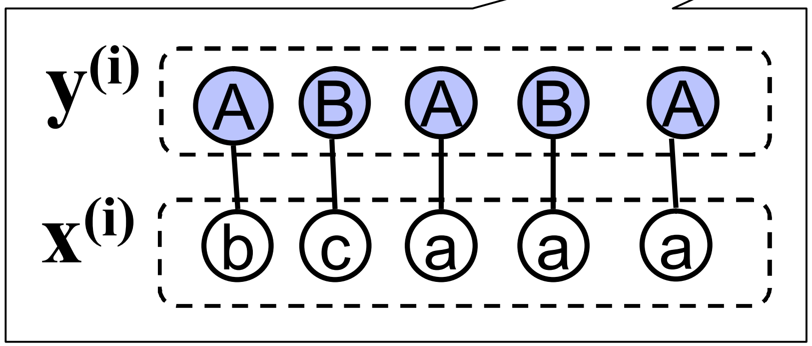


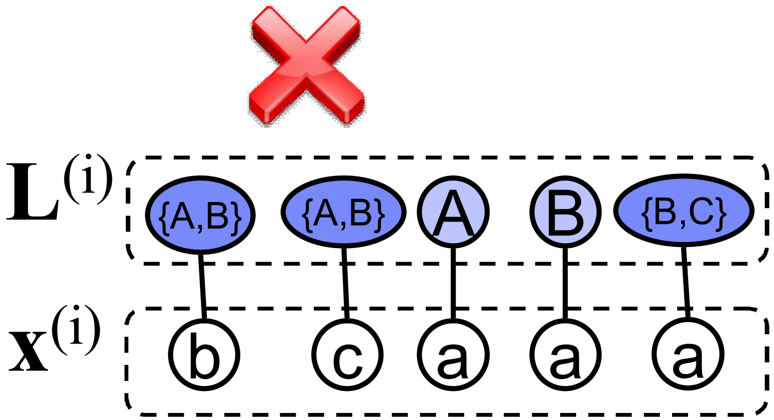
The original CRF learning algorithm requires completely annotated sequence (x, y)

- The incompletely annotated data (x, L) is not directly applicable to CRFs.
  - Conventional objective function for CRFs (log-likelihood):

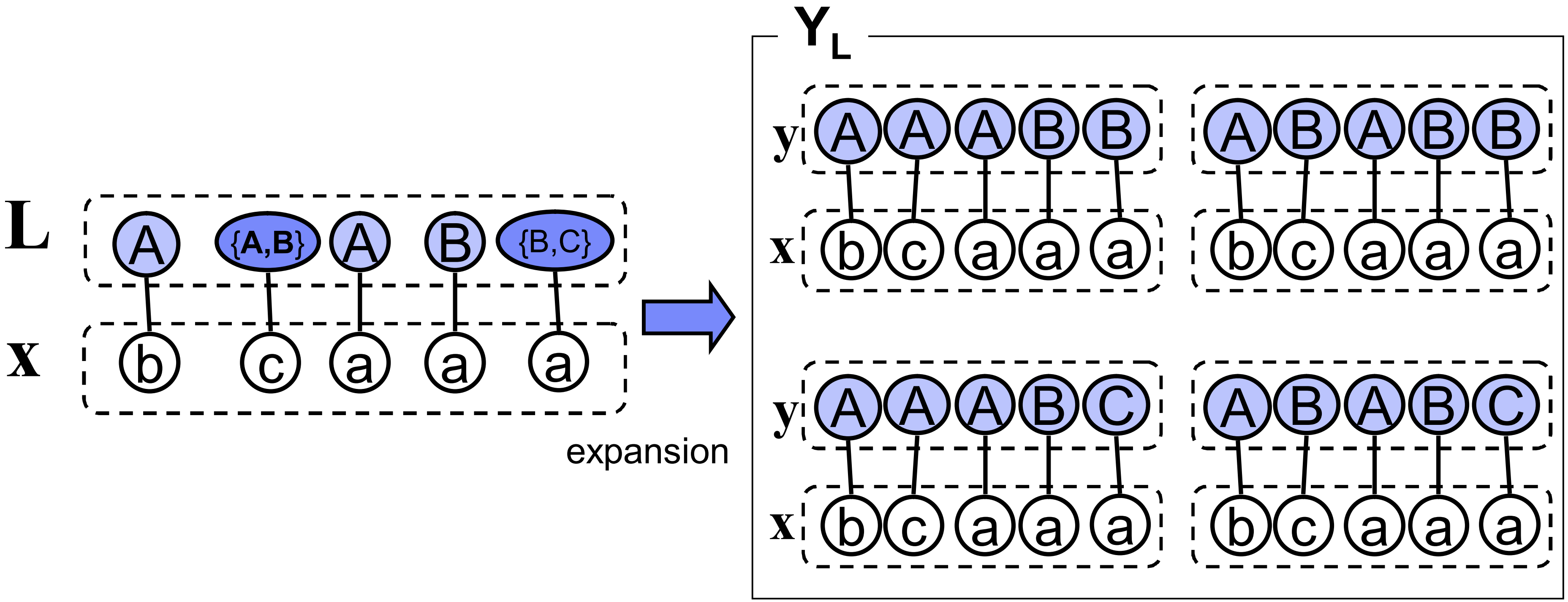
$$O(\boldsymbol{\theta}) = \sum_{i \in \mathrm{data}} \log P_{\boldsymbol{\theta}}(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)})$$

completely annotated sequence





Training CRFs using Y_L as the training data where Y_L denote all of the possible label sequence consistent with L. (Marginalized Likelihood)



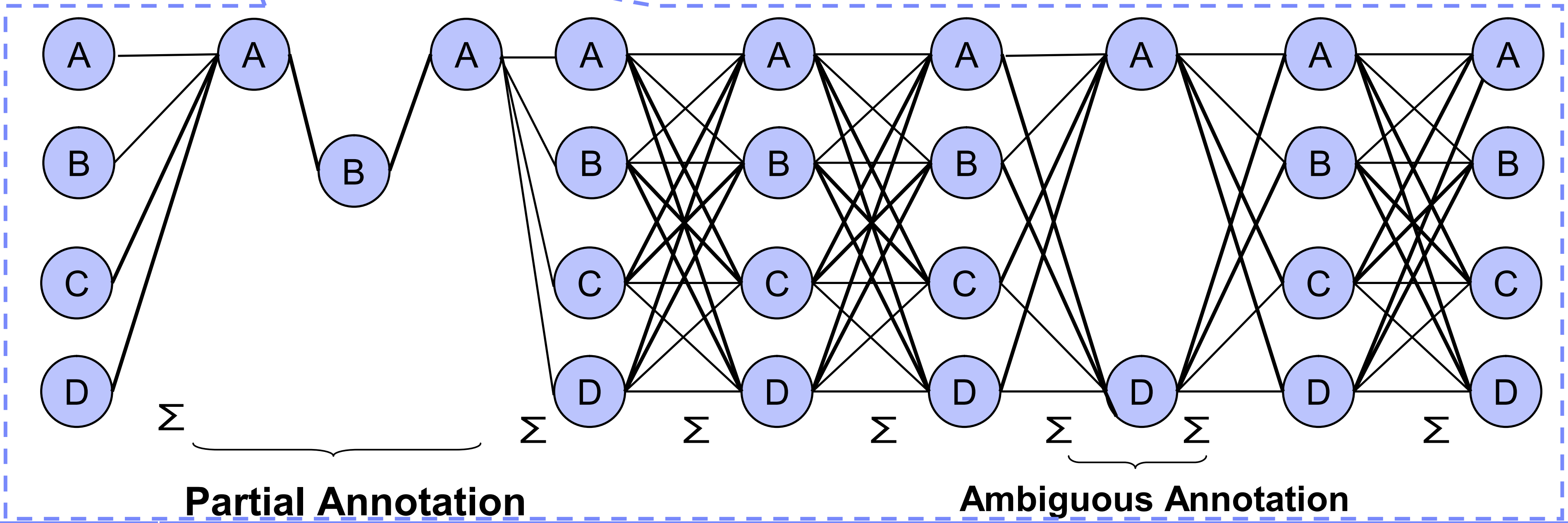
- **The CRF objective function for incomplete annotations**

$$O(\boldsymbol{\theta}) = \sum_{i \in \text{data}} \log P_{\boldsymbol{\theta}}(\mathbf{Y}_{\mathbf{L}^{(i)}} \mid \mathbf{x}^{(i)}) = \sum_{i \in \text{data}} \log \sum_{\mathbf{y} \in \mathbf{Y}_{\mathbf{L}^{(i)}}} P_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}^{(i)})$$

The score for Y_L (all of the possible label sequence consistent with L) is efficiently computable using a dynamic programming technique under the Markov assumption.

$$\sum_{\mathbf{y}\in\mathbf{Y}_L} \exp\left(\left\langle \boldsymbol{\theta}, \Phi(\mathbf{x},\mathbf{y}) \right\rangle\right)$$

Reuse the previous (t-1)-th computation which is consistent with $L_1, L_2, \ldots L_{t-1}$



Summary of the proposed method

- The proposed problem definition can deal with partial annotations, ambiguous annotations, and complete annotations in the same manner.
  - Non-concave (→ local maxima) objective function for CRF learning

$$\mathrm{O}(\boldsymbol{\theta}) = \sum_{i \in \mathrm{data}} \log \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{Y}_{\mathbf{L}^{(i)}} \mid \mathbf{x}^{(i)}) = \sum_{i \in \mathrm{data}} \log \sum_{\mathbf{y} \in \mathbf{Y}_{\mathbf{L}^{(i)}}} \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}^{(i)})$$

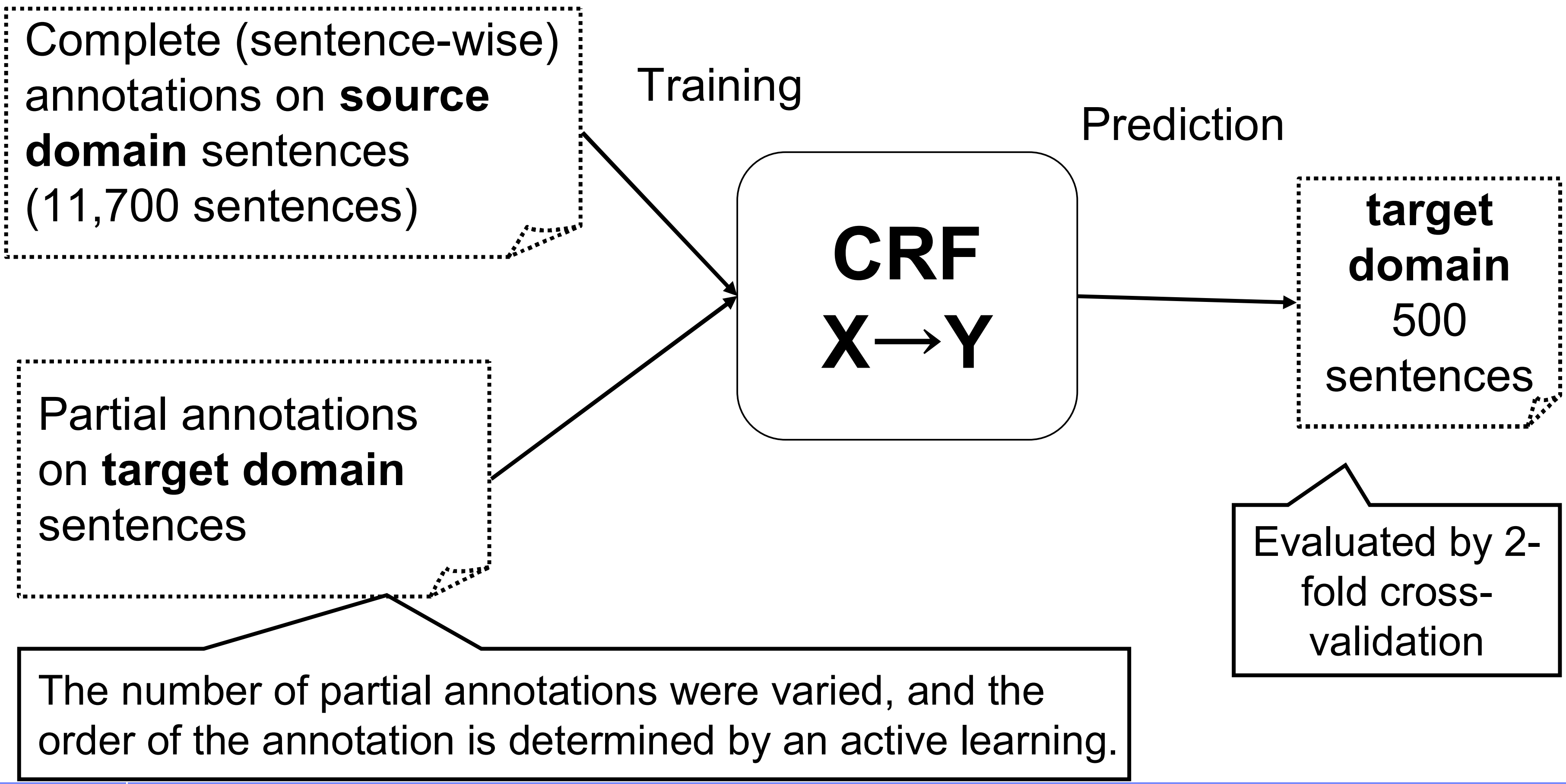
  - To optimize the function, we can use variants of gradient-based methods, such as conjugate gradient, L-BFGS, ….

Contents

- Incomplete annotations in corpus building.
  - Partial annotations & Ambiguous annotations
  - Word segmentation & Part-of-speech tagging task
- Training CRFs using Incomplete annotations
  - Representation of incomplete annotations
  - Supervised learning setting
  - Marginal likelihood for CRFs
- Experiments
  - A domain adaptation task of Japanese word segmentation using **partial annotations** by domain-specific word lists
  - POS tagging task using **ambiguous annotations** which are contained in Penn treebank corpus.

Domain adaptation experiments for the Japanese word segmentation task
Partial annotations were given the occurrences of words in the domain specific word list.

Domain adaptation task: daily conversation → medical reference manual



Domain adaptation task of Japanese word segmentation
Features and Performance Measure

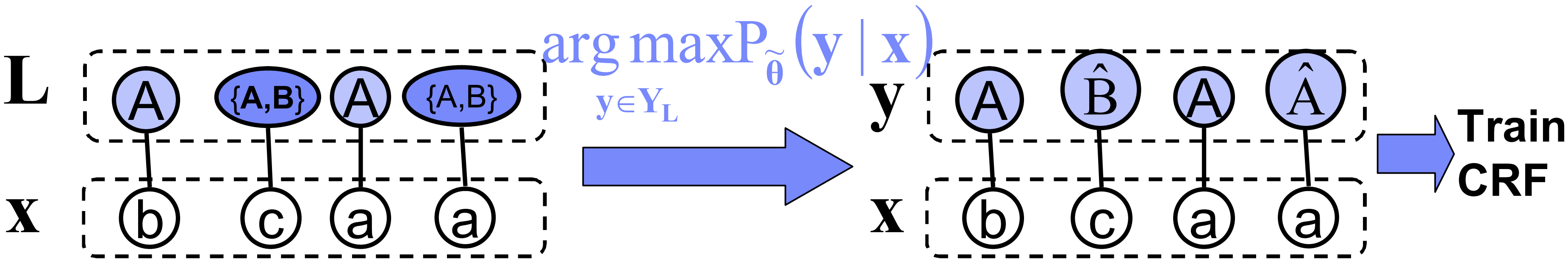
- As the features for observed variables, we use the **n-gram (n=1,2,3) characters and character types** around the current character boundary.
- We also used **lexical features consulting a dictionary**.
  - General domain dictionary and Target domain dictionary:
- Implementing **the first order Markov CRFs** and using L_2 regularizer
- The performance measure in the experiments is **F measure score** F =2PR/(R+P)

$$R = \frac{\text{\# of correct words}}{\text{\# of words in test data}} \times 100$$

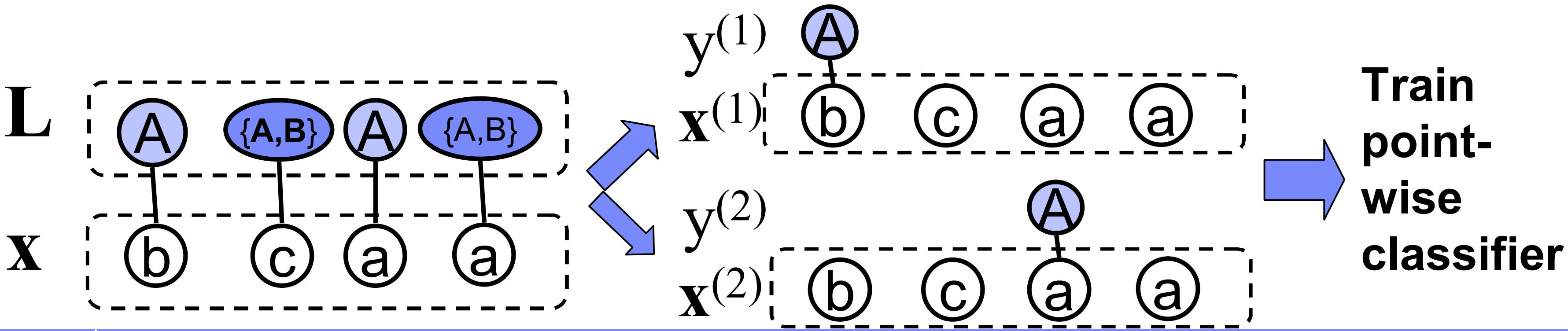
$$P = \frac{\text{\# of correct words}}{\text{\# of words in system output}} \times 100.$$

Two other possible methods dealing with partial annotations

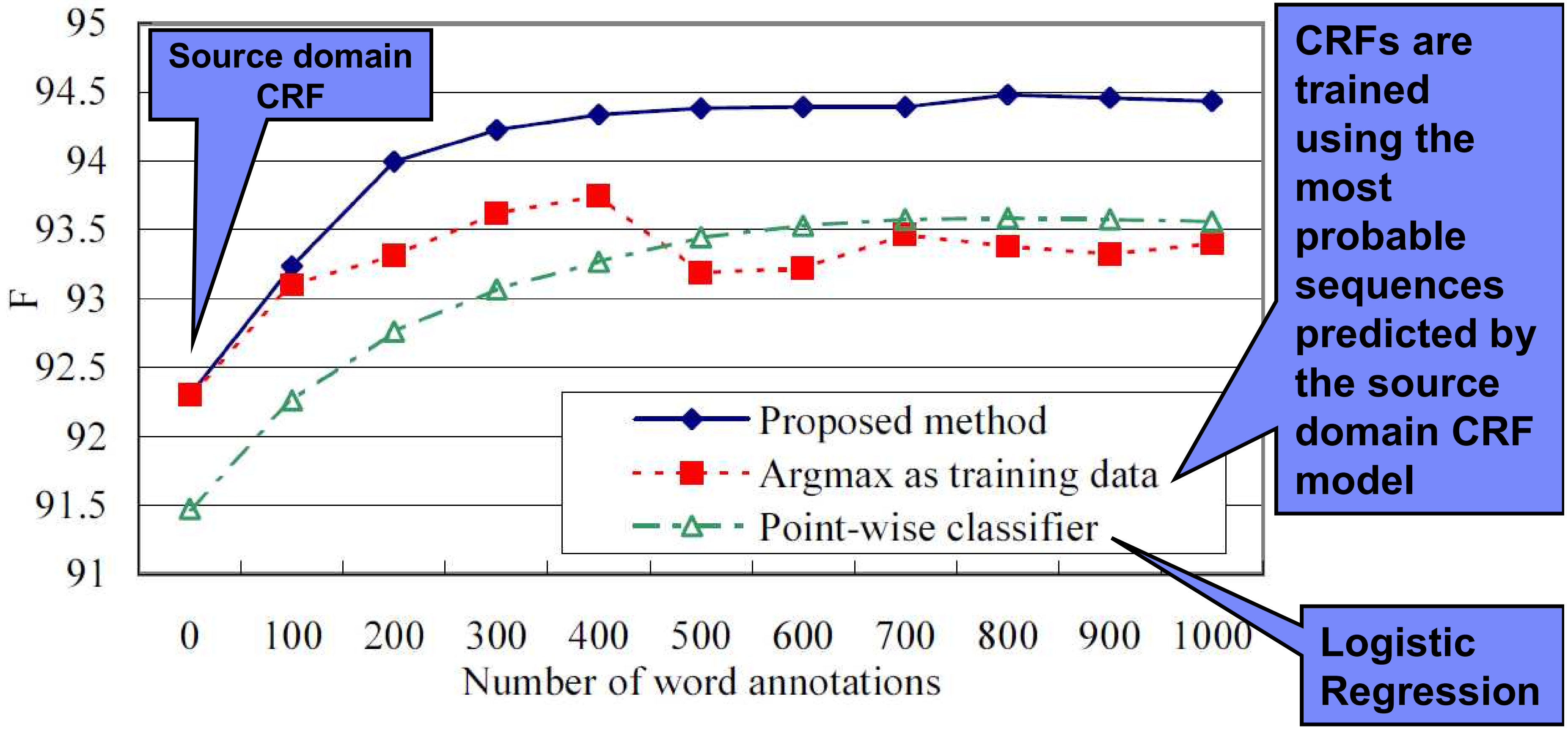
1. Filling unlabeled parts by prediction which is consistent with partial annotations (argmax as training data).



2. Training point-wise classifier which exclude label correlations.



This experimental result suggests that the proposed method maintains CRFs' advantage over the *point-wise classifier* and properly incorporates partial annotations.



Contents

- Incomplete annotations in corpus building.
  – Partial annotations & Ambiguous annotations
  – Word segmentation & Part-of-speech tagging task
- Training CRFs using Incomplete annotations
  – Representation of incomplete annotations
  – Supervised learning setting
  – Marginal likelihood for CRFs
- Experiments
  – A domain adaptation task of Japanese word segmentation using **partial annotations** by domain-specific word lists
  – POS tagging task using **ambiguous annotations** which are contained in Penn treebank corpus.

POS tagging task using ambiguous annotations which are contained in Penn treebank corpus. Experiment Settings

Training data

That/DT suit/NN is/VBZ **pending/VBG|JJ** ./SYM

… calls/VBZ for/IN MCI/NNP to/TO provide/VB **data/NN|NNS** service/NN ,/SYM…

⋮

"POS ambiguous sentences" (118)

… on/IN the/DT **pending/VBG** spinoff/NN disclosed/VBD that/IN….

⋮

POS unique sentences (1/10 or 2/10)

Test data

…. than/IN the/DT **pending/JJ** deal/NN suggests/VBZ ./SYM

⋮

POS unique sentences (8/10)

⇒ 5 trials for different data sets

For the comparison with the proposed method, we employed heuristic rules which disambiguate annotated candidate POS tags in the POS ambiguous sentences.

- **Heuristic POS disambiguation rules**
  That/DT suit/NN is/VBZ **pending/VBG|JJ** ./SYM →
  That/DT suit/NN is/VBZ **pending/VBG** ./SYM

1. **rand**: random selection
   pending/VBG|JJ → **pending/JJ**

2. **first**: selecting the first tag of the description order
   pending/VBG|JJ → **pending/VBG**

3. **frequent**: selecting the most frequent tag in the corpus
   pending/VBG|JJ → **pending/VBG** （where #VBG ＞ #JJ.）

4. **discarded**: the POS ambiguous sentences are ignored in training data.

The proposed method always outperformed other heuristic POS disambiguation

- Evaluation measures:

$$\mathrm{P} = \frac{\text{\# of correctly tagged word}}{\text{\# of all the word occurrences}} \times 100,$$

$$\mathrm{APA} = \frac{1}{|\boldsymbol{A}|} \sum_{w \in \boldsymbol{A}} \frac{\text{\# of the correctly tagged } w}{\text{\# of all the occurrences of } w} \times 100,$$

A: a word set and is composed of the word one of whose occurrences is ambiguously annotated

- Results

		mrg (proposed)	random	first	frequent	discarded
Ex.1	P	**94.274**	**94.274**	94.262	**94.274**	94.198
	APA	**73.272**	71.582	72.658	71.68	71.91
Ex.2	P	**94.982**	94.98	94.974	94.976	94.98
	APA	**76.242**	74.276	75.28	74.326	75.16

Table 5: The average POS tagging performance over 5 trials.

Conclusions

- We introduced
  - supervised learning setting incorporating partial annotations, ambiguous annotations, and complete annotations.
  - a parameter estimation method for CRFs using incomplete annotations under Markov assumption.