Conditional Random Fields Incorporating Incomplete Annotations

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

*1Tokyo Research Laboratory, IBM Research, IBM Japan.

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

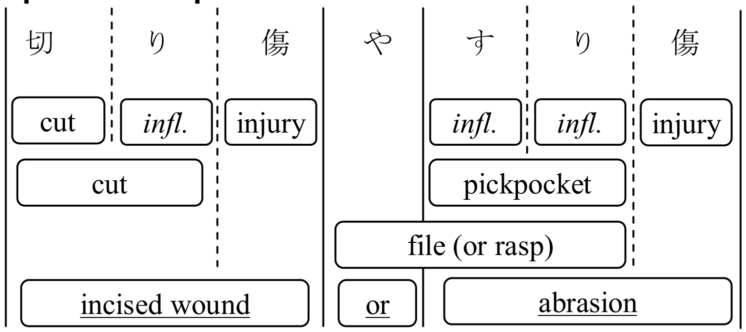
*3 Graduate School of Information Science, Nara Institute of Science and Technology

Contents

- Background
  - Word Segmentation Task & Part-of-speech Tagging Task as Structured Output Learning
- Incomplete annotations
  - Supervised learning using partial and ambiguous annotations
- Training Conditional Random Fields using Incomplete annotations
  - Conditional Random Fields: CRF
  - Marginal likelihood maximization
- Experiments
  - a domain adaptation task of Japanese word segmentation using partially annotated data created by word lists.
  - POS tagging task using ambiguous annotations which are contained in Penn treebank corpus.

Background
Word Segmentation Task & Part-of-speech Tagging Task
Those tasks have been solved by both rule-based or statistical approach using context information.

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

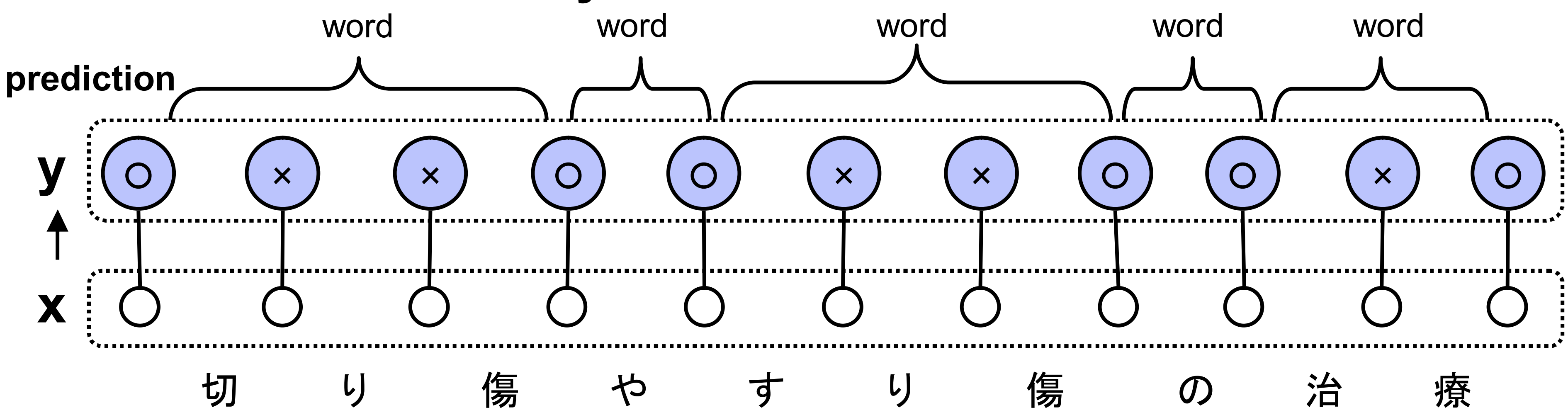


: word candidate
“ | ” : correct boundary
“ ¦ ” : incorrect boundary
infl. : inflectional suffix of verbs

- **Part-of-speech Tagging Task：identifying words as nouns, verbs, adjectives, adverbs, etc.**
  - Part-of-speech of words are depend on there context
    - English：flies → verb or noun?
    - Japanese：高め → 高め[た]（verb）or 高め[の球]（noun）？
- **Dictionary lookup is not enough for these tasks.**

Background
Word Segmentation as Structured Output Learning
Map: Character sequence → Boundary label sequence

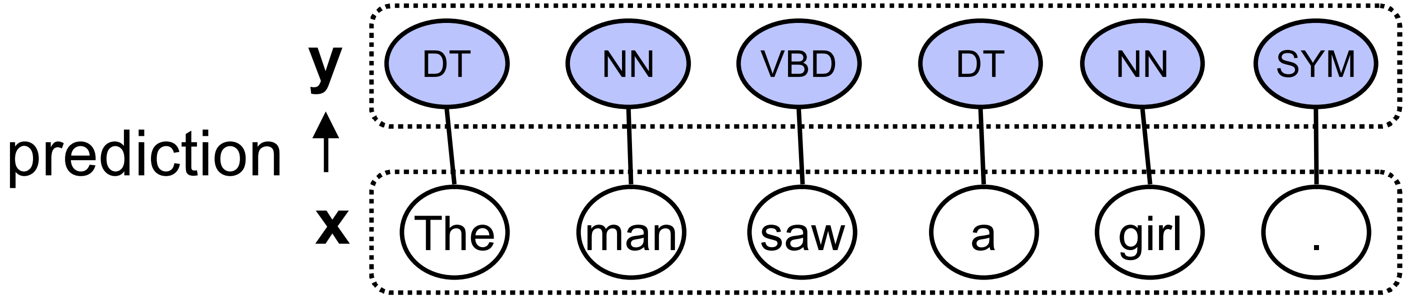
- **x: a given sequence of character boundaries**
- **y: a sequence of corresponding word boundary labels, which specify whether the current position is a word boundary or not.**



Label: × :non-word boundary ○:word boundary

Background
Part-of-speech Tagging as Structured Output Learning
Map: Word sequence → POS sequence

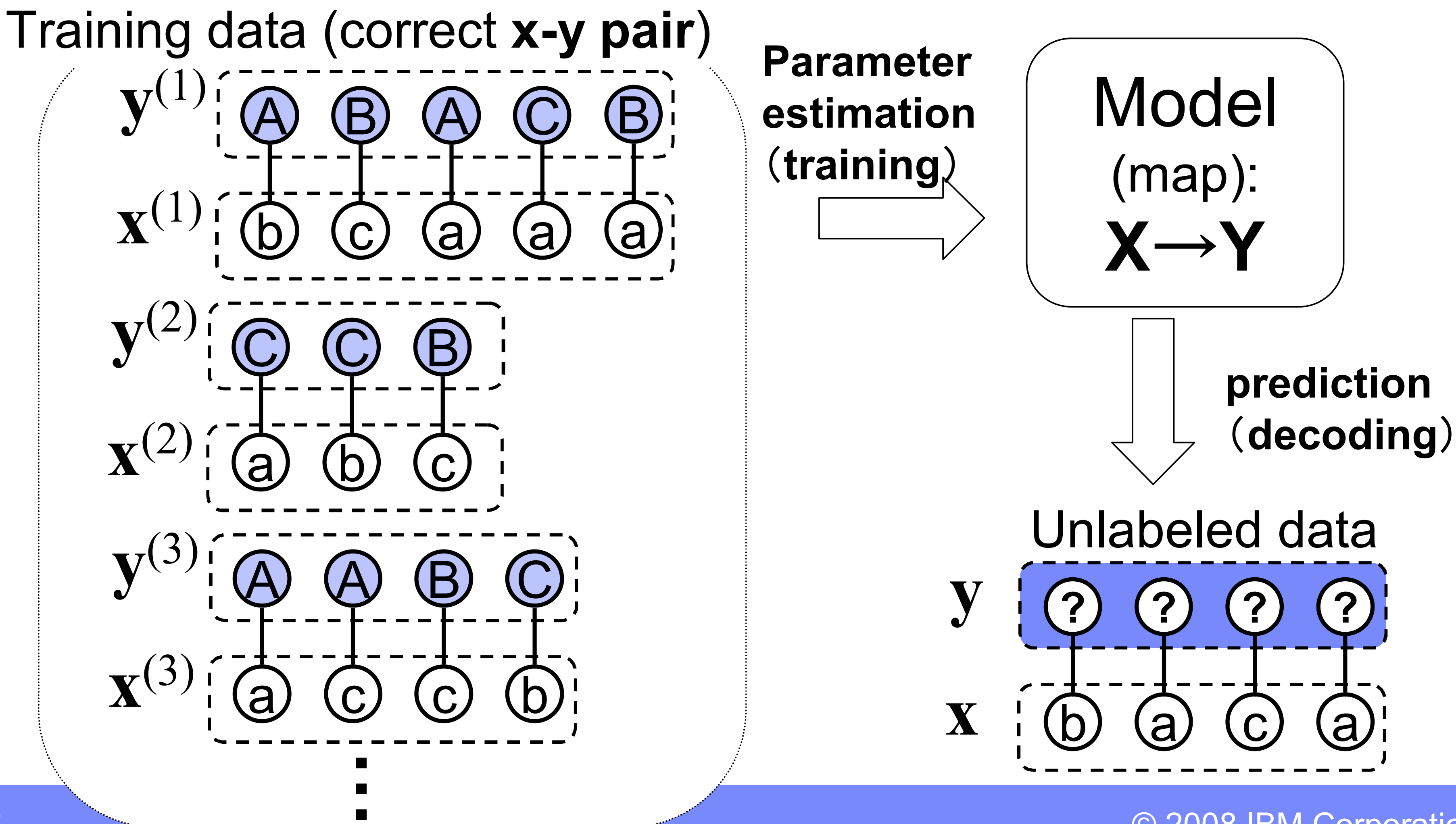
- **x: a word sequence**
- **y: a corresponding POS tag sequence**



Tags with corresponding part-of-speech

DT: determiner
NN: common noun
VBD: verb, past tense
SYM: symbol

Background
Supervised Structured Output Learning
Training a statistical models using correct pairs of an input and a label sequence

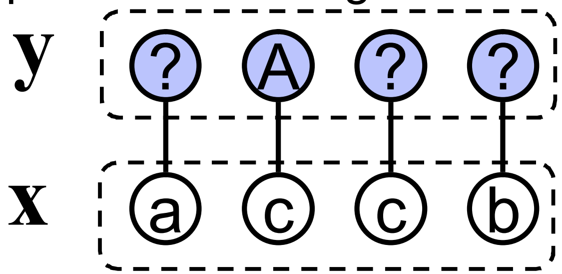


Contents

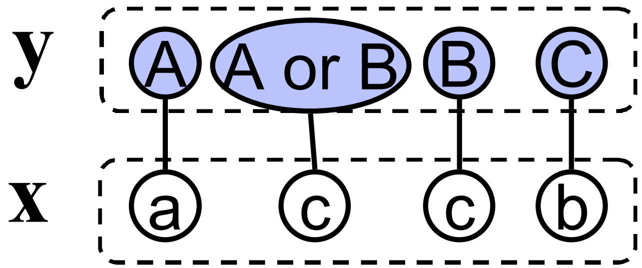
- Background
  - Word Segmentation Task & Part-of-speech Tagging Task as Structured Output Learning
- Incomplete annotations
  - Supervised learning using partial and ambiguous annotations
- Training Conditional Random Fields using Incomplete annotations
  - Conditional Random Fields: CRF
  - Marginal likelihood maximization
- Experiments
  - a domain adaptation task of Japanese word segmentation using partially annotated data by word lists.
  - POS tagging task using ambiguous annotations which are contained in Penn treebank corpus.

Partial annotations and ambiguous annotations
Incomplete annotations in corpus building phase

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



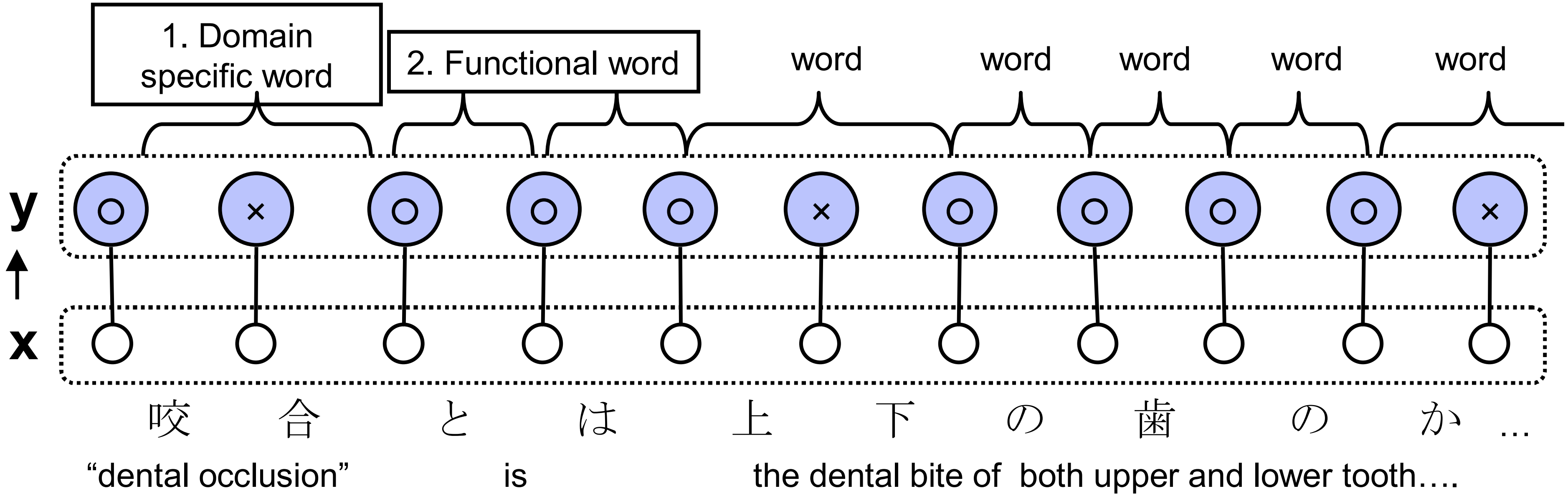
- Ambiguous annotations
  - A part of a structured instance are annotated by a set of candidate labels instead of a single label.
  - e.g. POS tags in Penn treebank corpus.



The notation for these annotations will be shown at some pages later.

Partial annotation
Partial annotations are effectively created in the situation of domain adaptation.

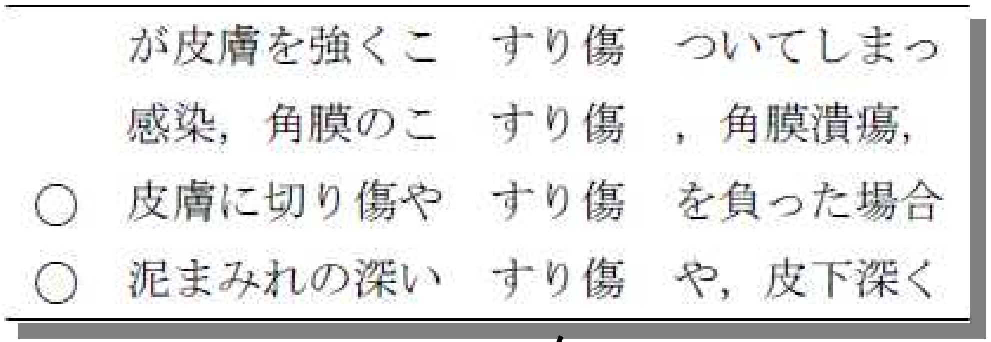
1. Annotators can concentrate on the higher learning effect instances
   – e.g. domain experts annotate only domain specific expressions.
2. Linguistically complicated parts can be left without annotation so that the number of noisy annotations might be reduced.
   – e.g. domain experts can leave functional words untouched.

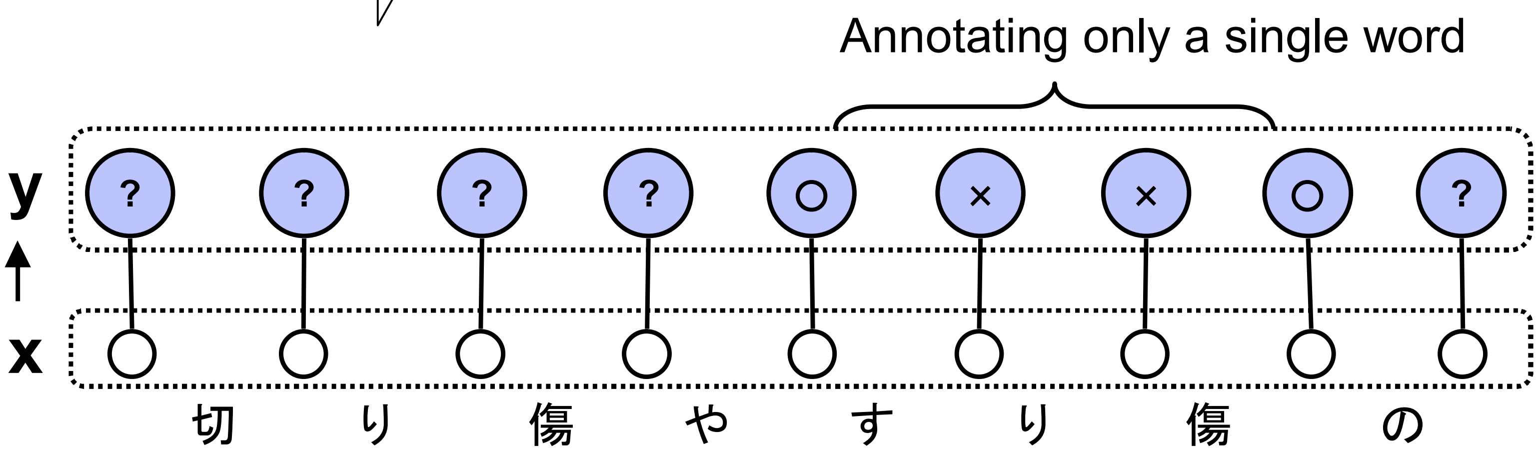


Partial annotations
KWIC (KeyWord In Context) style annotation UI (User Interface) using domain word lists (Mori, 2006)

- Domain word lists: product name list, technical term dictionary, …

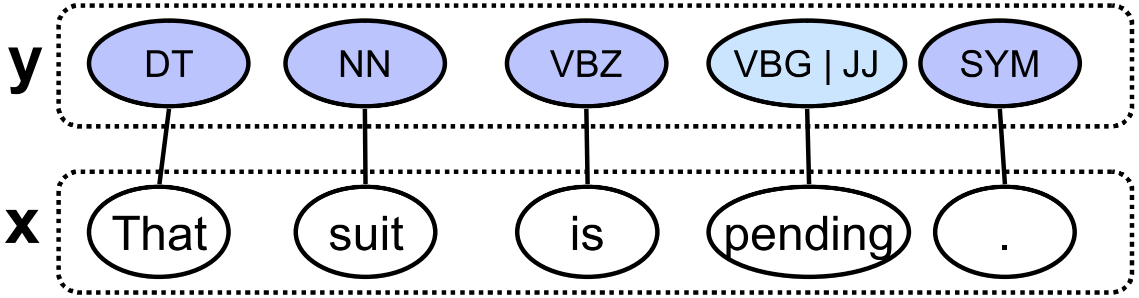
Example:
An annotator marks the occurrences only if the string ``すり傷'' (abrasion) of the domain word list is used as a real word in the given context.





Ambiguous annotations
A set of candidate labels annotated in a part of a structured instance.

- Ambiguous POS tags in Penn treebank corpus
  - The proper POS tag of ``pending'' is represented by disjunctive POS tag (``VBG and JJ'') which is separated by a vertical bar.



DT: determiner, NN: common noun

VBZ: present tense and 3rd person singular verb

VBG: gerund or present participle verb

JJ: adjective SYM: symbols

- Note: the order in which the candidate tags appear has not been standardized in Penn Treebank corpus (*Part-of-Speech Tagging Guidelines for the Penn Treebank Project, 1995*).

Ambiguous annotations
Penn treebank English corpus, whose annotation procedure is relatively well-defined, includes more than 100 sentences containing POS ambiguities

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

freqency	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

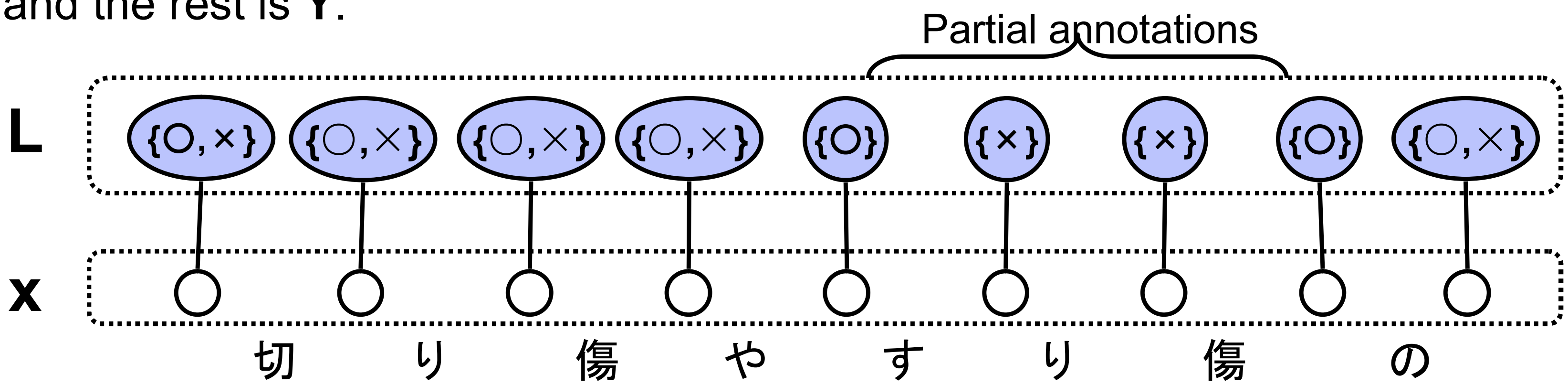
- Background
  - Word Segmentation Task & Part-of-speech Tagging Task as Structured Output Learning
- Incomplete annotations
  - Supervised learning using partial and ambiguous annotations
- Training Conditional Random Fields using Incomplete annotations
  - Conditional Random Fields: CRF
  - Marginal likelihood maximization
- Experiments
  - a domain adaptation task of Japanese word segmentation using partially annotated data by 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 L:

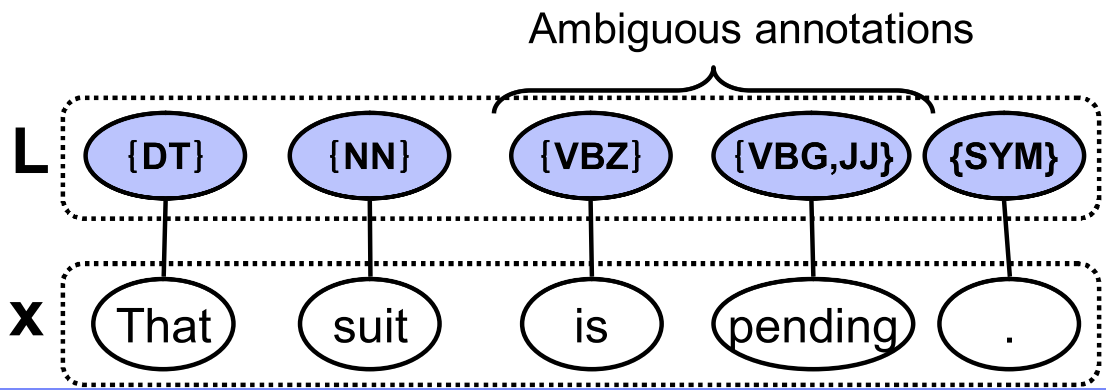
$$\mathbf{L} = \left(L_t \subseteq Y \text{ for } t = 1 \cdots T\right)$$

Partial Annotations

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

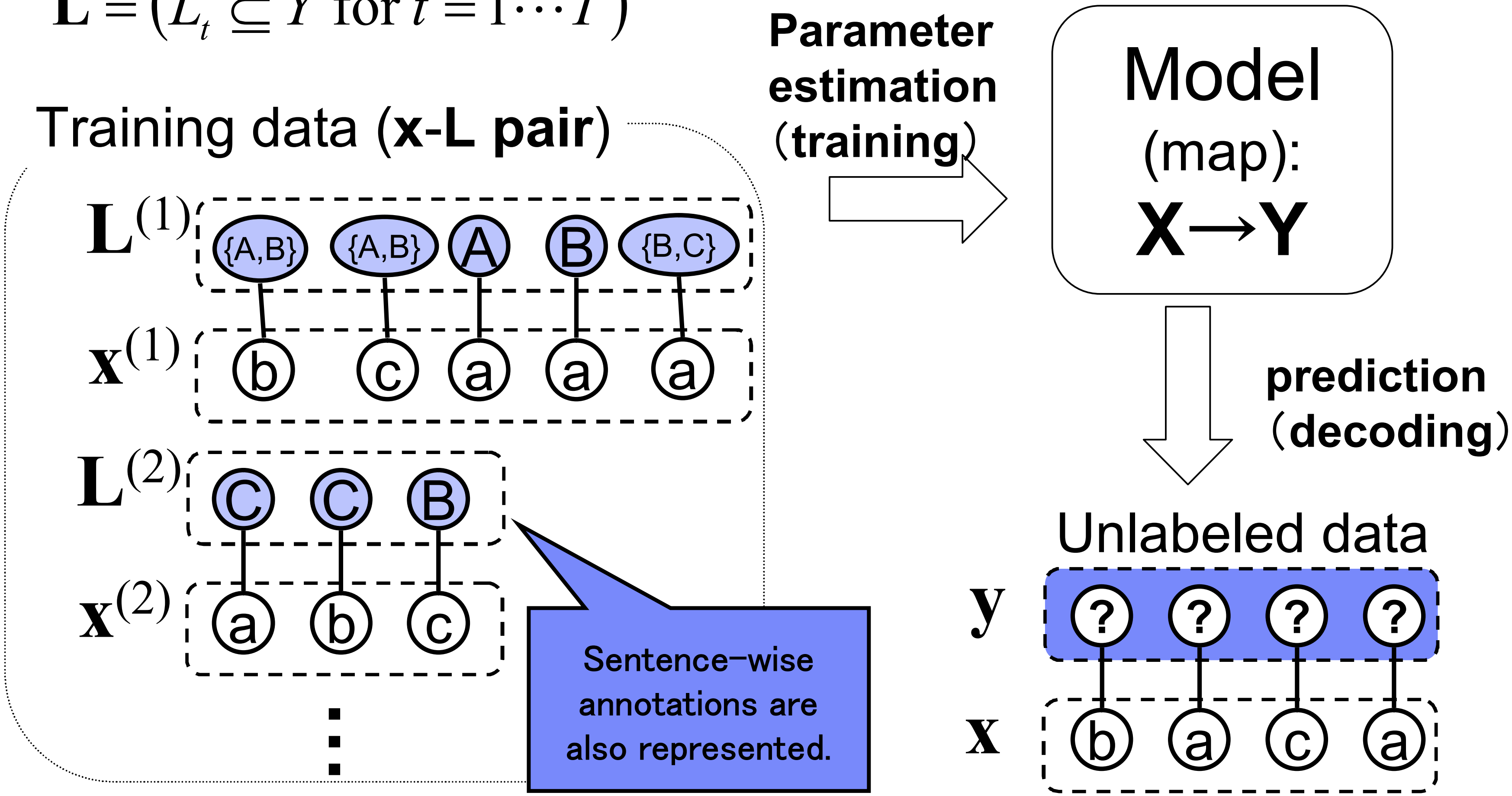


- **Ambiguous Annotations**
  – L_t represents a set of candidate labels at the position t.



Supervised learning using incomplete annotations Training data is pairs of input x and label set sequence L.

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



Contents

- Background
  - Word Segmentation Task & Part-of-speech Tagging Task as Structured Output Learning
- Incomplete annotations
  - Supervised learning using partial and ambiguous annotations
- Training Conditional Random Fields using Incomplete annotations
  - Conditional Random Fields: CRF
  - Marginal likelihood maximization
- Experiments
  - a domain adaptation task of Japanese word segmentation using partially annotated data by word lists.
  - POS tagging task using ambiguous annotations which are contained in Penn treebank corpus.

Conditional Random Fields (CRFs)
CRFs model conditional probability of a label sequence y given an observed sequence x.

A discriminative model for structured output

$$P_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) = \frac{\exp\left(\left\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{x}, \mathbf{y}) \right\rangle\right)}{\sum_{\tilde{\mathbf{y}} \in \mathbf{Y}} \exp\left(\left\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{x}, \tilde{\mathbf{y}}) \right\rangle\right)}$$

The score of y

The summation of all the possible label sequences' score

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

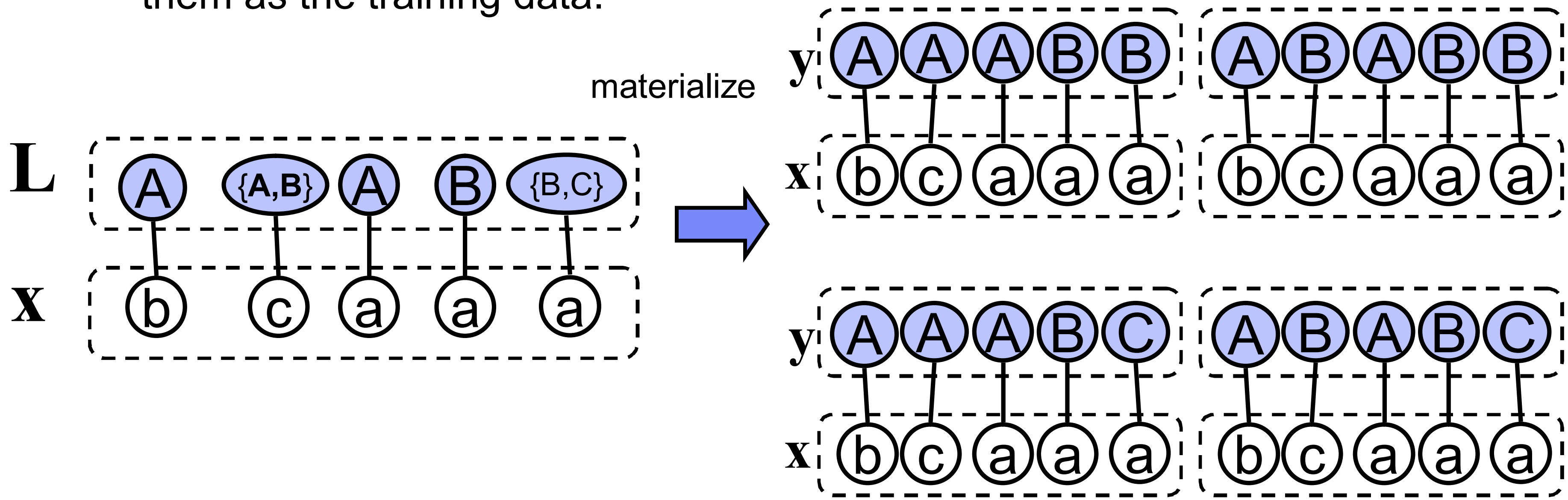
$\theta \in \mathbf{R}^d$ denote the vector of model parameters.

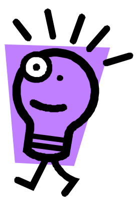
Once θ is estimated, the label sequence can be predicted by

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathbf{Y}} P_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$$

Training CRFs incorporating Incomplete annotations. Since original CRFs require completely labeled sequence y for learning, the incompletely annotated data (x, L) is not directly applicable to CRFs.

- Let $\mathbf{Y}_L$ denote all of the possible label sequence consistent with **L**, a naive approach can be explicitly materialize all the entry of $\mathbf{Y}_L$ and use them as the training data.





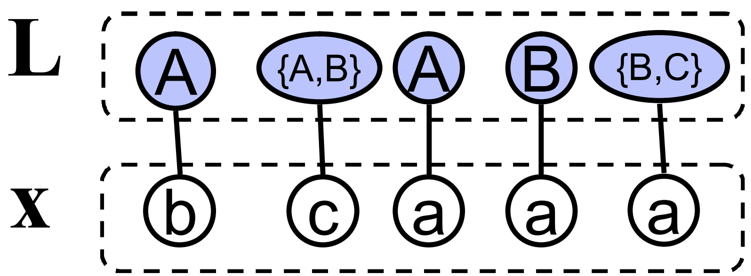
- The number of annotated sentences are quadruplicate which is exponential on the number of positions t with $|L_t| > 1$.
- →**Solving by appropriate weighting and dynamic programming**

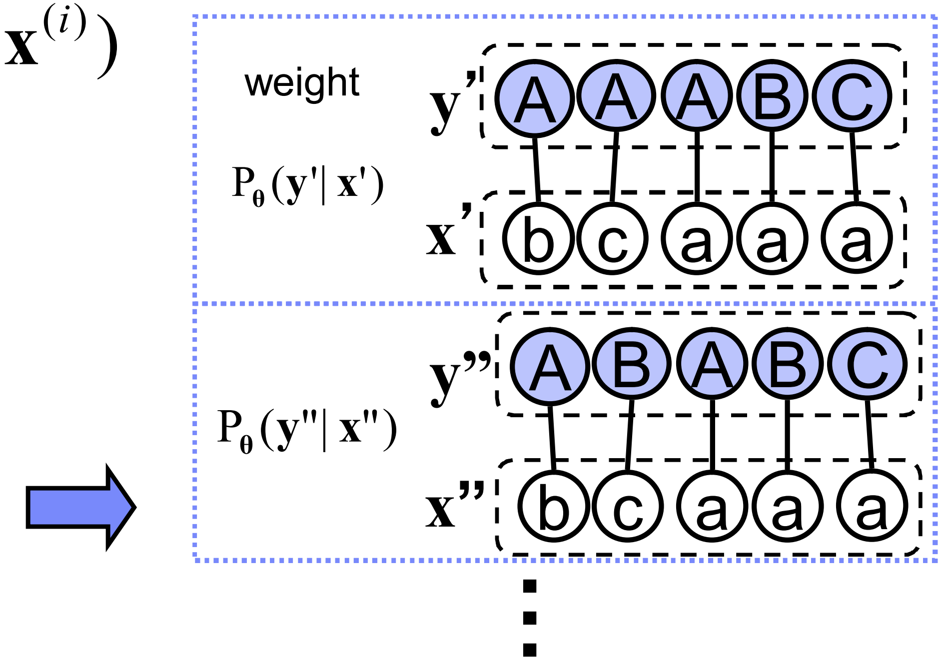
Training CRFs incorporating Incomplete annotations.
Maximum Marginalized Likelihood for CRFs
Maximizing the likelihood of a set Y_L

- **The proposed objective function** **Convex function**

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

  - Implicitly weighting **x-y** pairs by the current model P_θ





Summation for all of the possible label sequence consistent with L is efficiently computable using the dynamic programming technique under the Markov assumption.

Objective function

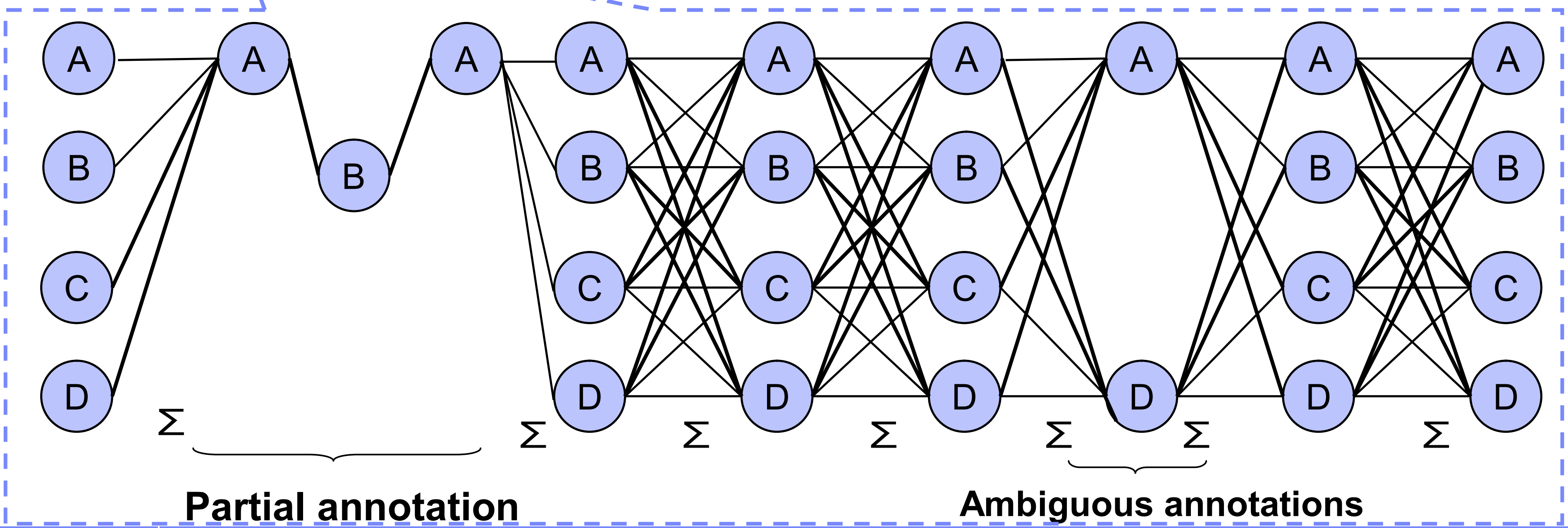
$$O(\boldsymbol{\theta}) = \sum_{i \in \text{data}} \log Z_{\boldsymbol{\theta},\mathbf{x}^{(i)},\mathbf{Y}_{\mathbf{L}^{(i)}}} - \log Z_{\boldsymbol{\theta},\mathbf{x}^{(i)},\mathbf{Y}}$$

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

Partial derivative function

$$\frac{\partial O(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i \in \text{data}} F\left(\mathbf{Y}_{\mathbf{L}^{(i)}}\right) - F(\mathbf{Y})$$

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



Conditional Random Fields Incorporating Incomplete Annotations

Contents

- Background
  – Word Segmentation Task & Part-of-speech Tagging Task as Structured Output Learning
- Incomplete annotations
  – Supervised learning using partial and ambiguous annotations
- Training Conditional Random Fields using Incomplete annotations
  – Conditional Random Fields: CRF
  – Marginal likelihood maximization
- Experiments
  – A domain adaptation task of Japanese word segmentation using partially annotated data by word lists.
  – POS tagging task using ambiguous annotations which are contained in Penn treebank corpus.

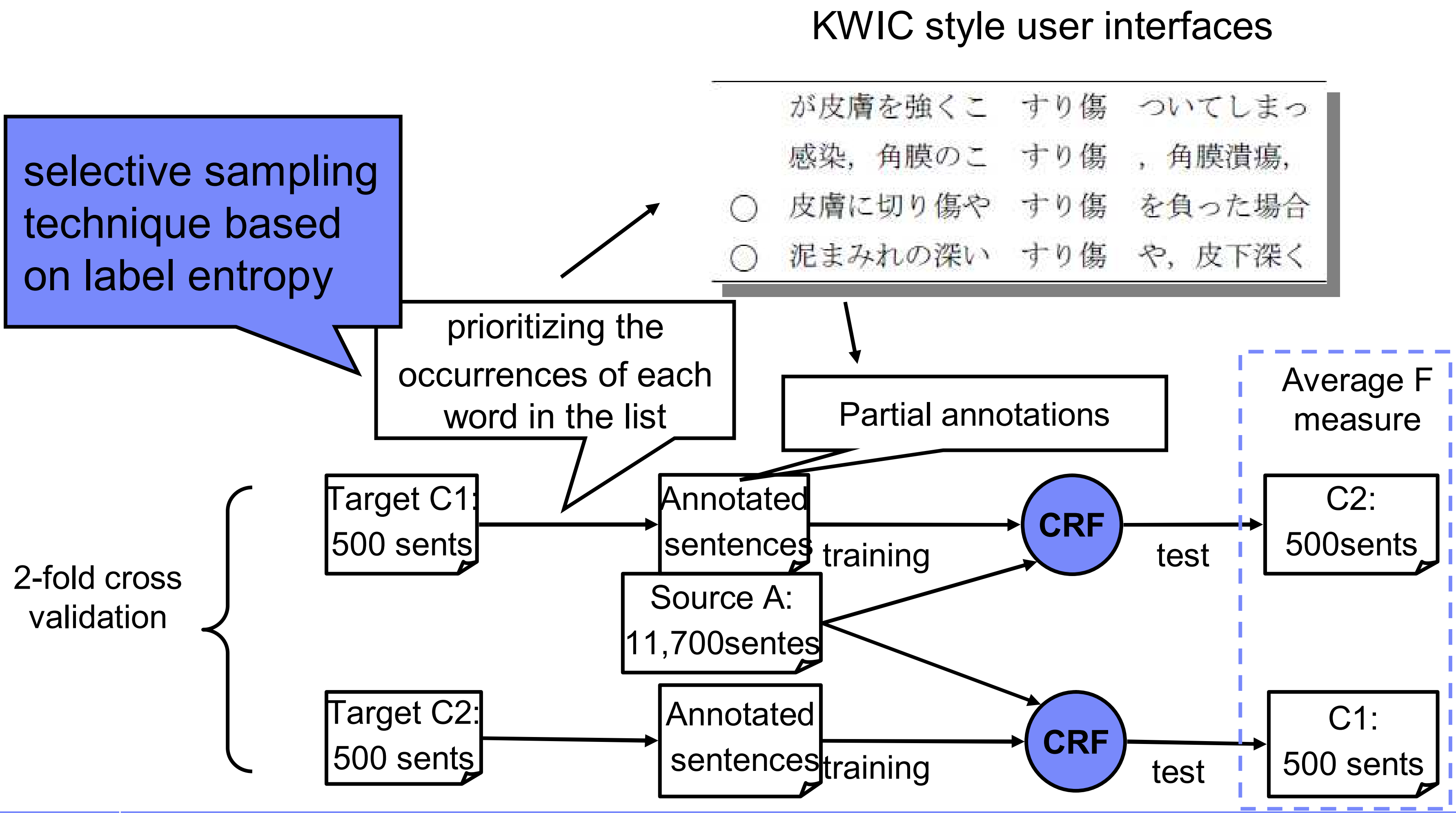
Domain adaptation experiments for the Japanese word segmentation task from daily conversation to medical reference manual

- Source domain data : example sentences in a dictionary of daily conversation
- Target domain data : medical reference manual

	domain	#sentences	#words
(A)	conversation	11,700	145,925
(B)	conversation	1,300	16,348
(C)	medical manual	1,000	29,216

  – We create the word list from the target domain words which do not appeared in the source domain data (A). The averaged number of distinct new words in the data (C1) is 948.5, which equals to the size of the word list.

Experiment scenario: a user selects the occurrences of words in the word list using the KWIC interface.



A domain adaptation task of Japanese word segmentation
Features and Performance Measure

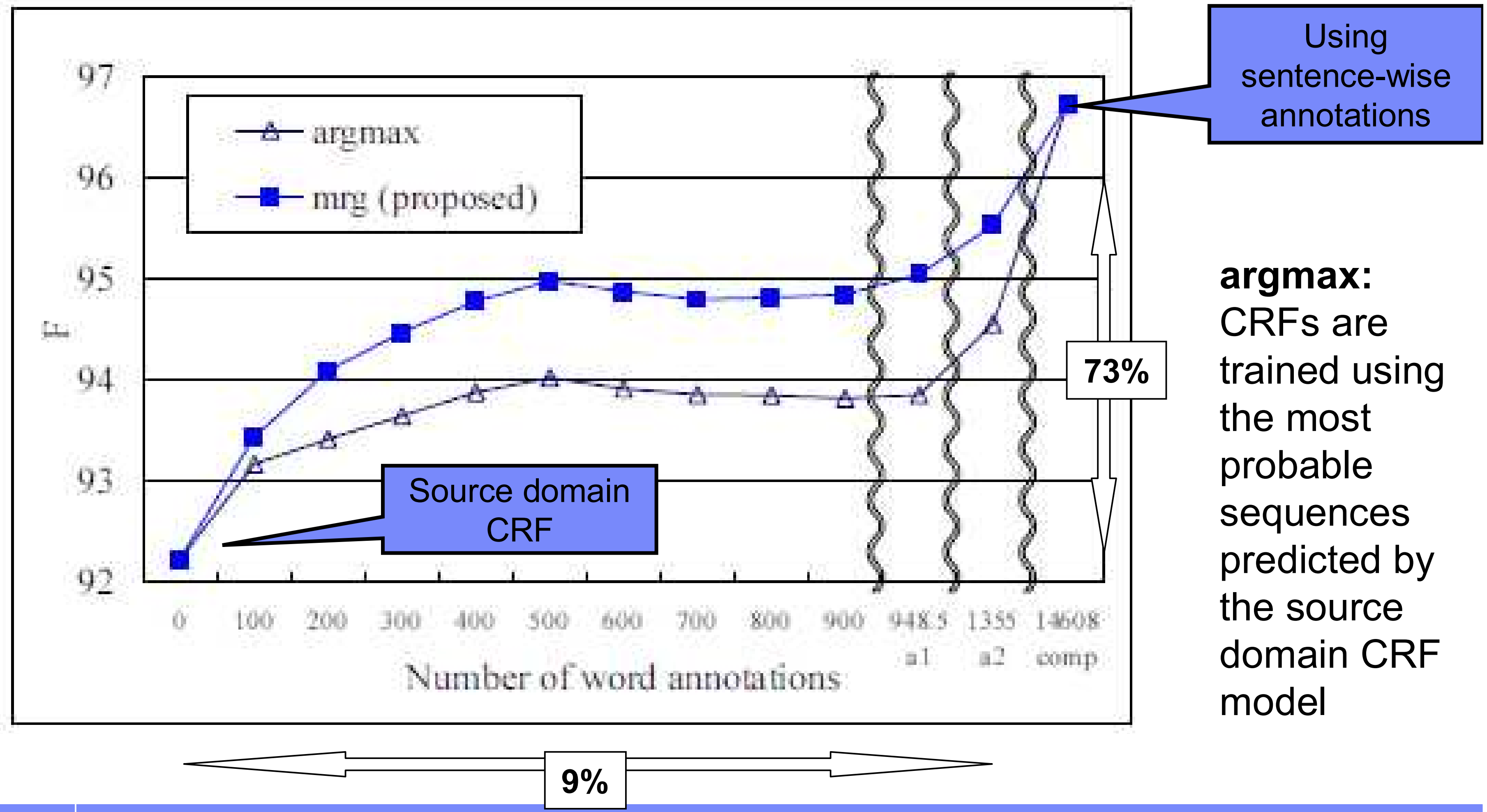
- As the features for observed variables, we use the **n-gram (n=1,2,3) characters and character types** including or adjoining the current character boundary.
  - The character type set is composed of Hiragana, Katakana, Kanji, alphabet, Arabic numerals, and symbols.
  - The total number of distinct features 298, 363
- Implementing **the first order Markov CRFs** and using L_2 regularizer

- The performance measure in the experiments
  - the standard **F measure score** F=2PR/(R+P)

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

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

The combination of both the proposed method and the selective sampling method achieved 73% of the performance gain by only 9.3% (a2) of the number of word occurrences for sentence-wise annotation.



argmax: CRFs are trained using the most probable sequences predicted by the source domain CRF model

Contents

- Background
  - Word Segmentation Task & Part-of-speech Tagging Task as Structured Output Learning
- Incomplete annotations
  - Supervised learning using partial and ambiguous annotations
- Training Conditional Random Fields using Incomplete annotations
  - Conditional Random Fields: CRF
  - Marginal likelihood maximization
- Experiments
  - a domain adaptation task of Japanese word segmentation using partially annotated data by 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**: both POS ambiguous and unique sentences
- **Test data**: POS unique sentences (11, 840 sentences)

5 trials for different data sets

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 (1,480 or 2,960)

Test data

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

⋮

POS unique sentences (11,840)

POS tagging task using ambiguous annotations which are contained in Penn treebank corpus. Features (mostly adapted from Altun et al. 2003.)

- The feature sets for each word are **case-insensitive spelling**, **orthographic features** of the current word, and **sentence last word**.
  - The orthographic features are whether a spelling begins with a number, upper case letter; whether it begins upper case letter and contains period(``."); whether it is all upper case letter, all lower case letter; whether it contains a punctuation symbol, a hyphen; and the last one, two, and three letters of the word.
  - The sentence last word corresponds to a punctuation mark (e.g. ``.", ``?", ``!")
  - the total number of resulting distinct features is 14,391.
- Implementing **the first order Markov CRFs** using L_2 regularizer

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

- **Disambiguation**
  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

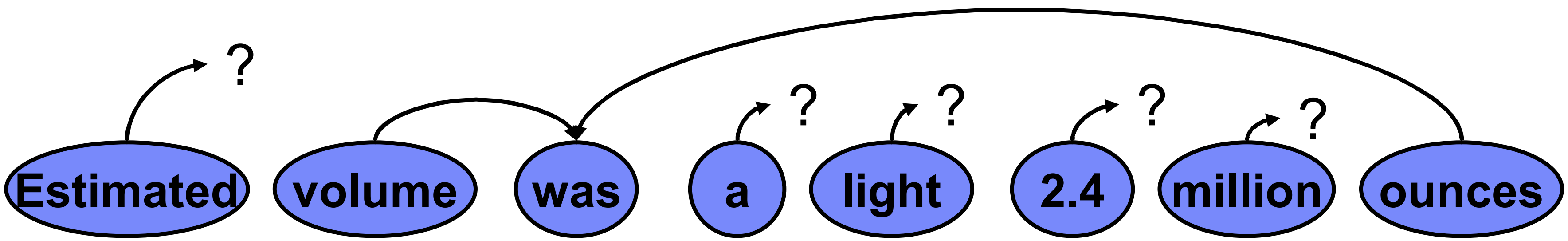
		mrg (proposed)	random	first	frequent	discarded
Ex.1	P	**94.274**	**94.274**	94.262	**94.274**	94.198
	APA	**73.272**	71.582	72.658	71.68	71.91
Ex.2	P	**94.982**	94.98	94.974	94.976	94.98
	APA	**76.242**	74.276	75.28	74.326	75.16

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

Conclusions and Future Work

- We propose a parameter estimation method for CRFs incorporating partial and ambiguous annotations of structured data.
- Future work: We believe partial annotations might also effectively reduce annotation work for dependency parsing.

Partially annotated dependency tree



Thank you for your attention!