部分的かつ曖昧なラベル付き構造データからのマルコフ条件付確率場の学習
Training Conditional Random Fields Using Partial and Ambiguous Structured Labels

坪井祐太*1*3　鹿島久嗣*1　森信介*2　小田裕樹　松本裕治*3

*1 日本アイ･ビー･エム株式会社
*2 京都大学
*3 奈良先端科学技術大学院大学

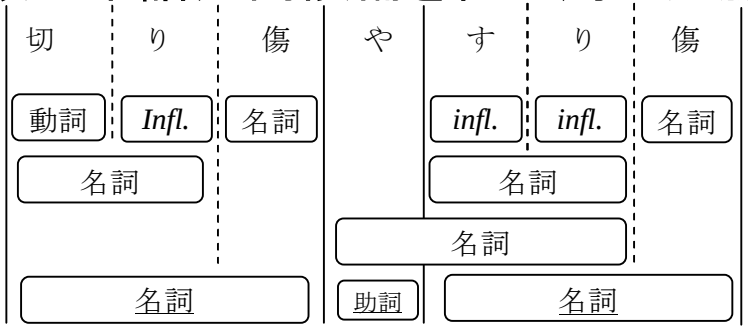
目次

研究の背景

単語分割タスク・品詞付与タスク

辞書方式では不十分→ルールや統計的手法により文脈を考慮する必要性

- **単語分割タスク：入力文を単語に分割**
  - 複数の単語分割候補を含み周辺文脈を考慮し決定する必要性



文字列「切り傷やすり傷」内で単語になりうる文字列
“｜”：正しい単語境界
“¦”：間違った単語境界
（infl.=活用語尾語尾）

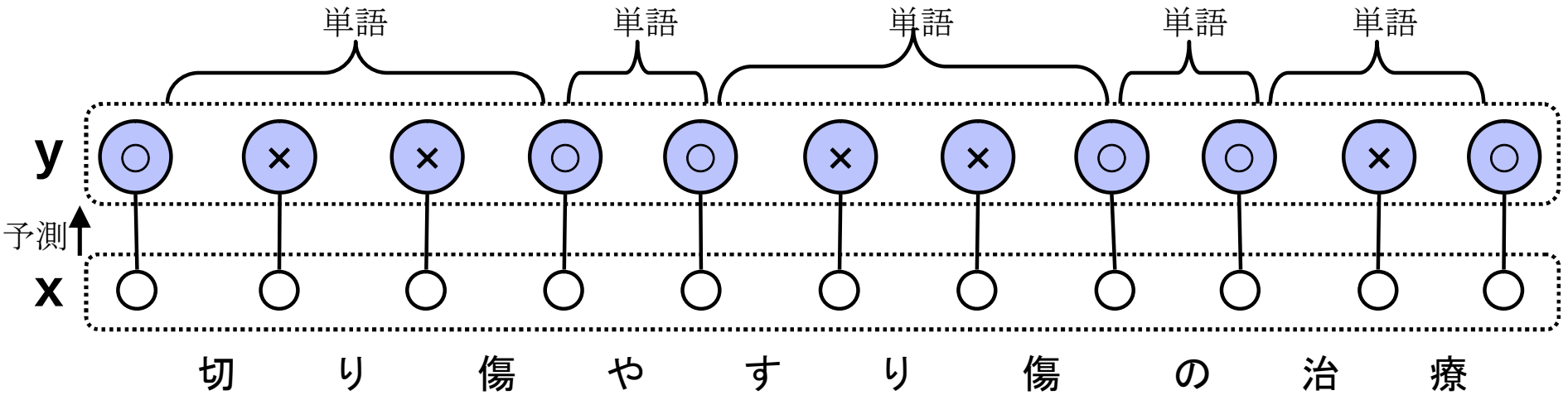
- **品詞付与タスク：単語に品詞を付与**
  - 同じ単語でも複数の品詞候補が存在し周辺文脈を考慮し決定する必要性
  - 英語：files → 飛ぶ（動詞） or ハエ（名詞）
  - 日本語 ： 高め→高め[た]（動詞） or 高め[の球]（名詞）

研究の背景

構造出力問題としての単語分割タスク

文字境界列→単語境界ラベル列 を予測

- 文字列境界の列を表す入力列X
- 単語境界・非境界の列を表す出力ラベルの列Y
- 構造出力問題：文字列境界列Xから単語境界・非境界の列Yを予測する問題



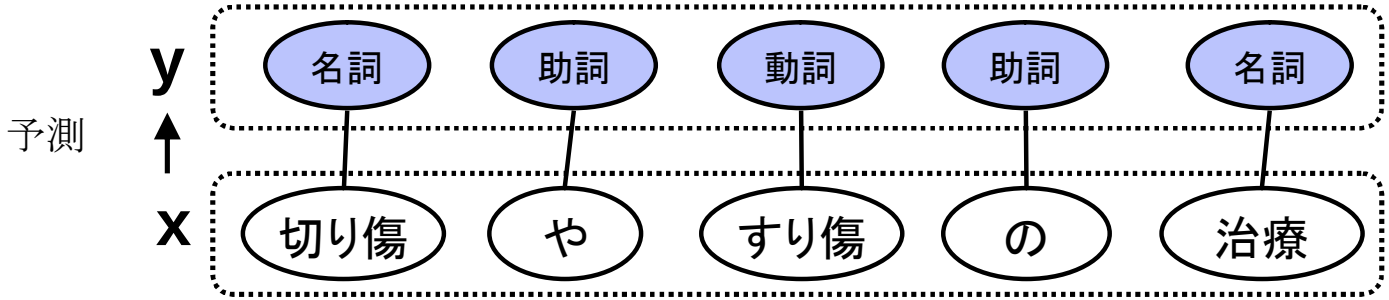
ラベル：　×:非単語境界　○:単語境界

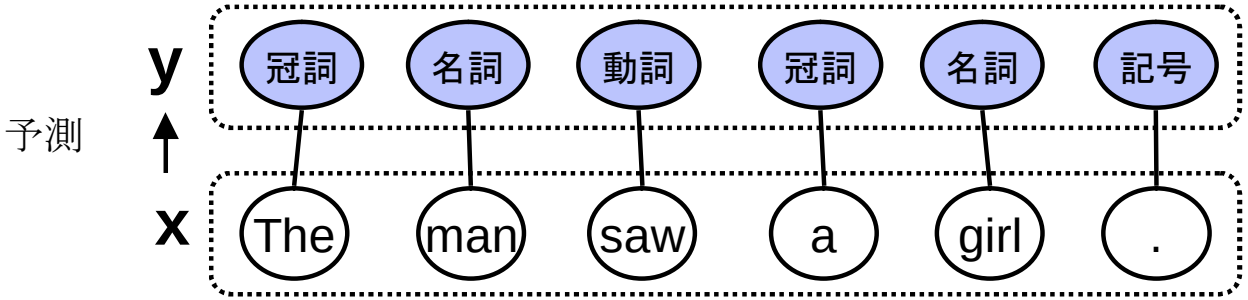
研究の背景

構造出力問題としての品詞付与タスク

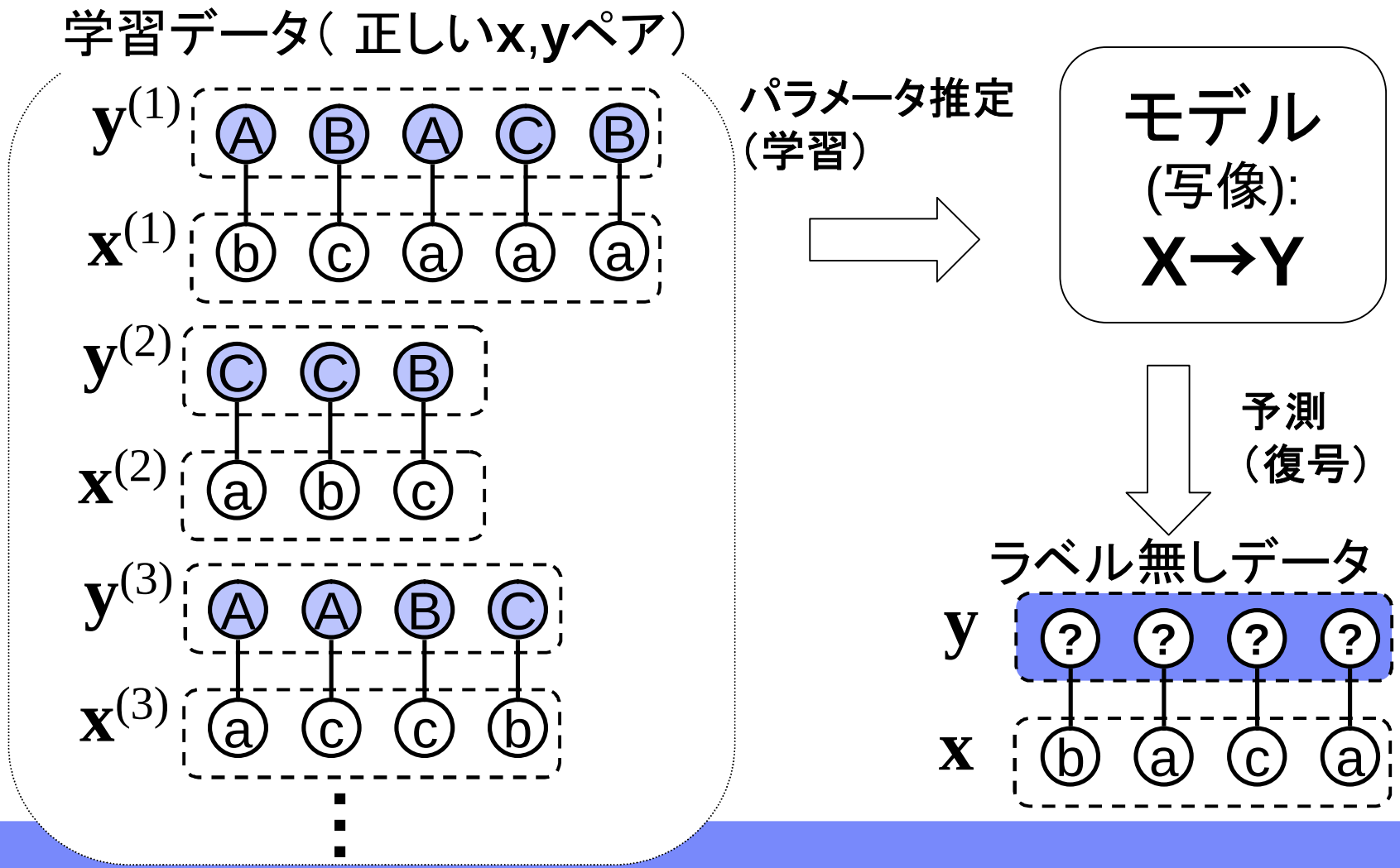
単語列→品詞列を予測

- 単語の列を表す入力列X
- 品詞の列を表す出力ラベルの列Y
- 構造出力問題：単語列Xから品詞列Yを予測する問題





研究の背景
教師付き学習による構造出力学習アプローチ
正しい入力列・ラベル列の対データを元に統計モデルを学習



目次

- 研究の背景
  - 構造出力問題としての単語分割・品詞付与タスク
- 部分的アノテーションと曖昧なアノテーション
  - 部分的かつ曖昧なアノテーションと教師付き学習の定式化
- 部分的かつ曖昧なアノテーションを用いた条件付確率場の学習
  - 条件付確率場(Conditional Random Fields: CRF)
  - 周辺尤度最大化による学習法の提案
- 実験
  - 単語リストによる部分的アノテーションを活用した日本語単語分割の分野適応実験
  - Penn Treebank コーパスの曖昧なアノテーションを用いた品詞付与実験
- まとめと今後の課題

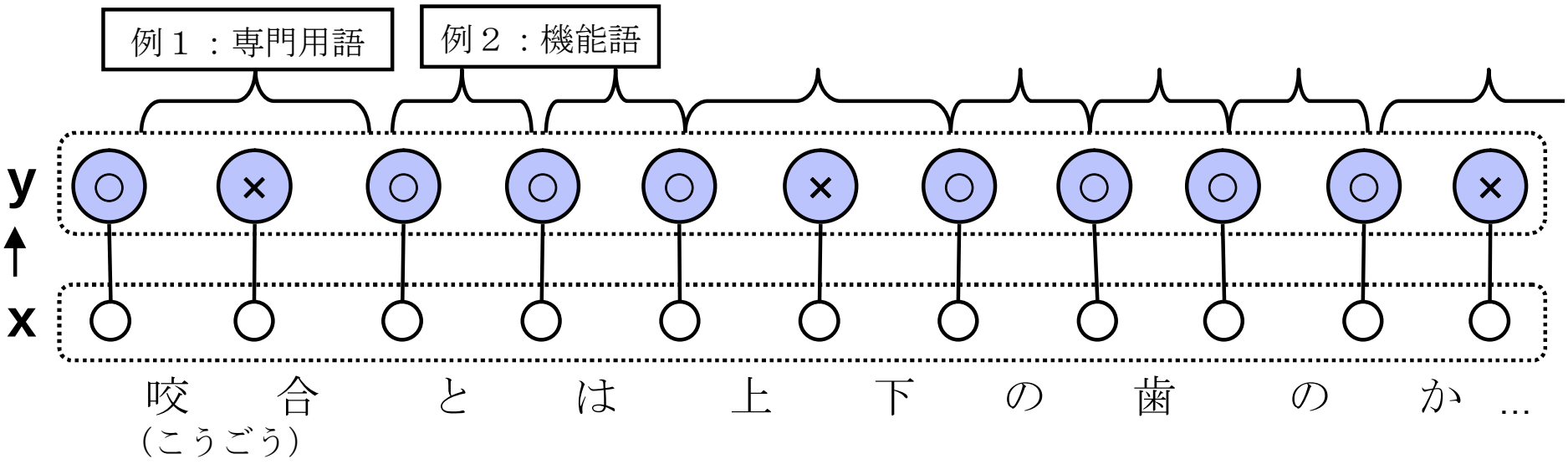
部分的アノテーションと曖昧なアノテーション
コーパスに存在する不完全なデータを扱う必要性

- 部分的アノテーション
  - 文の一部にのみラベルが与えられたアノテーション
  - 日本語単語分割問題を例に説明
- 曖昧なアノテーション
  - 文の一部のラベルに複数の候補があるアノテーション
  - 英語品詞付与問題を例に説明

定式化は数ページ先までお待ち下さい

部分的アノテーション
文の一部にのみラベルが与えられたアノテーション
特に分野適応時に文全体にラベルを全て付与するのに比べて効率的

- 日本語単語分割の分野適応における部分的アノテーションの効果
  1. 学習効果の高い事例のラベル付けに集中でき効果的
     - 分野特有の用語の出現箇所を集中的にアノテーション
  2. 判断できない箇所のラベル付けが省略可能→アノテーションミス削減
     - 機能語などは単語分割基準を理解している必要があり専門家には困難。



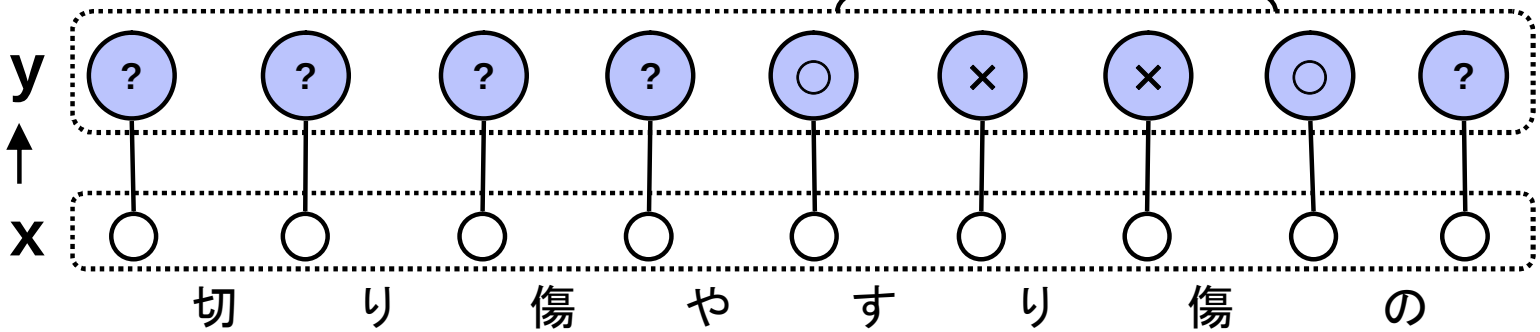
部分的アノテーション
単語リストを用いたKWIC形式インターフェース(森, 2006)
専門用語の出現箇所の文脈を表示し未知語にのみ集中的にアノテーションするGUI

- 適応先の単語リスト：専門用語辞書や商品名リストなど

部分的ラベル付与例：
対象分野の単語リスト内単語の出現箇所をKWIC表示し、単語として用いられている文に○をつける。

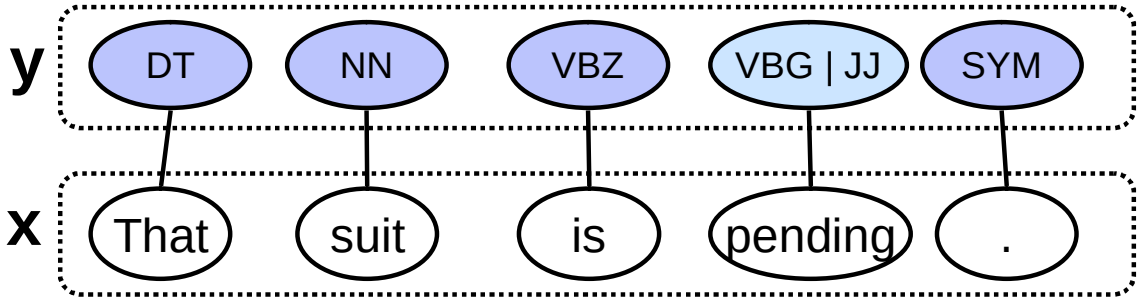
	が皮膚を強くこ	すり傷	ついてしまっ
	感染，角膜のこ	すり傷	，角膜潰瘍，
○	皮膚に切り傷や	すり傷	を負った場合
○	泥まみれの深い	すり傷	や，皮下深く

一単語分だけアノテーション付与



曖昧なアノテーション
文の一部のラベルに複数の候補があるアノテーション
作業者が付与するラベルを決定できない場合に、候補を列挙しておくことが可能になる。

- Penn Treebankコーパスにおける品詞が曖昧な単語の例
  - “pending” の品詞タグは”VBGまたはJJ”としてアノテーション



  - DT：限定詞 NN：名詞単数 VBZ：動詞3人称単数現在形
    VBG：動名詞または動詞現在分詞 JJ：形容詞 SYM: 記号
- Penn Treebank コーパスでは、タグの候補が複数ある場合の記述の順番は重要ではなくその順番には一貫性は無い(*Part-of-Speech Tagging Guidelines for the Penn Treebank Project, 1995*)。

曖昧なアノテーション
文の一部のラベルに複数の候補があるアノテーション
Penn Treebankコーパスには品詞に曖昧性が在る単語を含む文が100文以上存在

- 品詞に曖昧性のある例が頻度3以上の単語

頻度	単語	品詞
15	data	NN|NNS
10	more	JJR|RBR
7	pending	JJ|VBG
4	than	IN|RB
3	trading	NN|VBG
3	broker-dealer	JJ|NN

  - 候補品詞が3つある単語も存在：dividing/JJ|NN|VBG
- 意味を扱うタスクでは作業者間での一致率も低く、より曖昧性が残りやすい
  - 曖昧性が残らないタスク設計が基本だが、現実には困難

部分的アノテーションと曖昧なアノテーション
コーパスに存在する不完全なデータを扱う必要性

- 部分的アノテーション
  - 文の一部にのみラベルが与えられたアノテーション
- 曖昧なアノテーション
  - 文の一部のラベル候補が複数あるアノテーション
  - 注：複数のラベル(multi label)を付与する問題とは異なり、
    予測時はラベルは一意に決定

- 参考：それ以外の不完全なデータ：
  - 入力に誤りがあるデータ（欠損値） → 協調フィルタリング、素性削除学習
  - ラベルの間違いを含むデータ → 頑健推定、外れ値検出

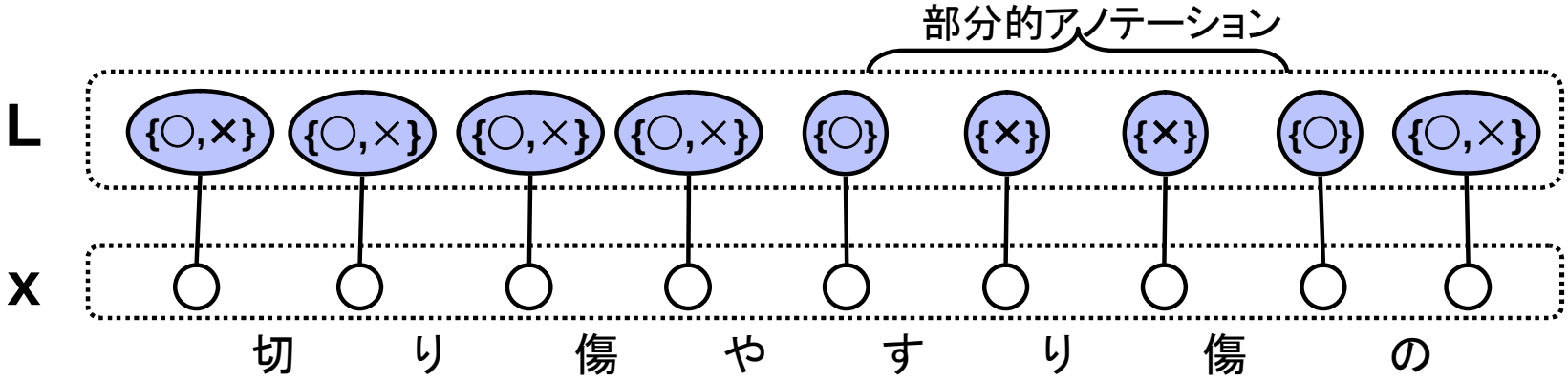
目次

- 研究の背景
  - 構造出力問題としての単語分割・品詞付与タスク
- 部分的アノテーションと曖昧なアノテーション
  - 部分的かつ曖昧なアノテーションと教師付き学習の定式化
- 部分的かつ曖昧なアノテーションを用いた条件付確率場の学習
  - 条件付確率場(Conditional Random Fields: CRF)
  - 周辺尤度最大化による学習法の提案
- 実験
  - 単語リストによる部分的アノテーションを活用した日本語単語分割の分野適応実験
  - Penn Treebank コーパスの曖昧なアノテーションを用いた品詞付与実験
- まとめと今後の課題

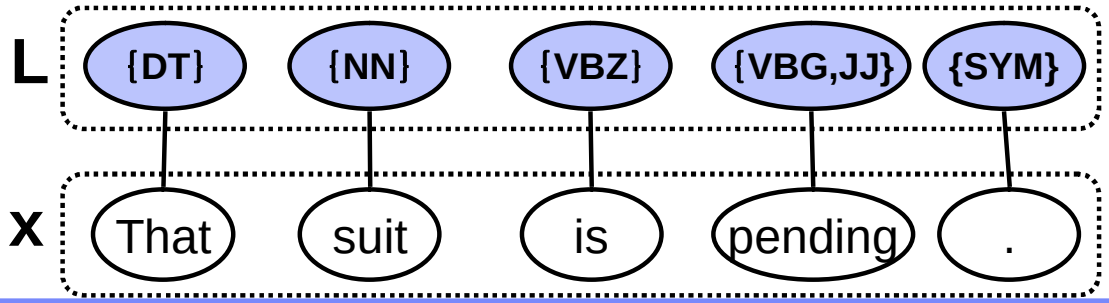
部分的アノテーションと曖昧なアノテーションの定式化
ラベル集合列Lで表現。要素L_tは点tで取りうるラベル候補集合。

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

- ラベルが付与された箇所のL_tの要素は一つ
- **部分的アノテーション**
  - ラベルが付与されていない箇所(?)のL_tは全てのラベルの集合



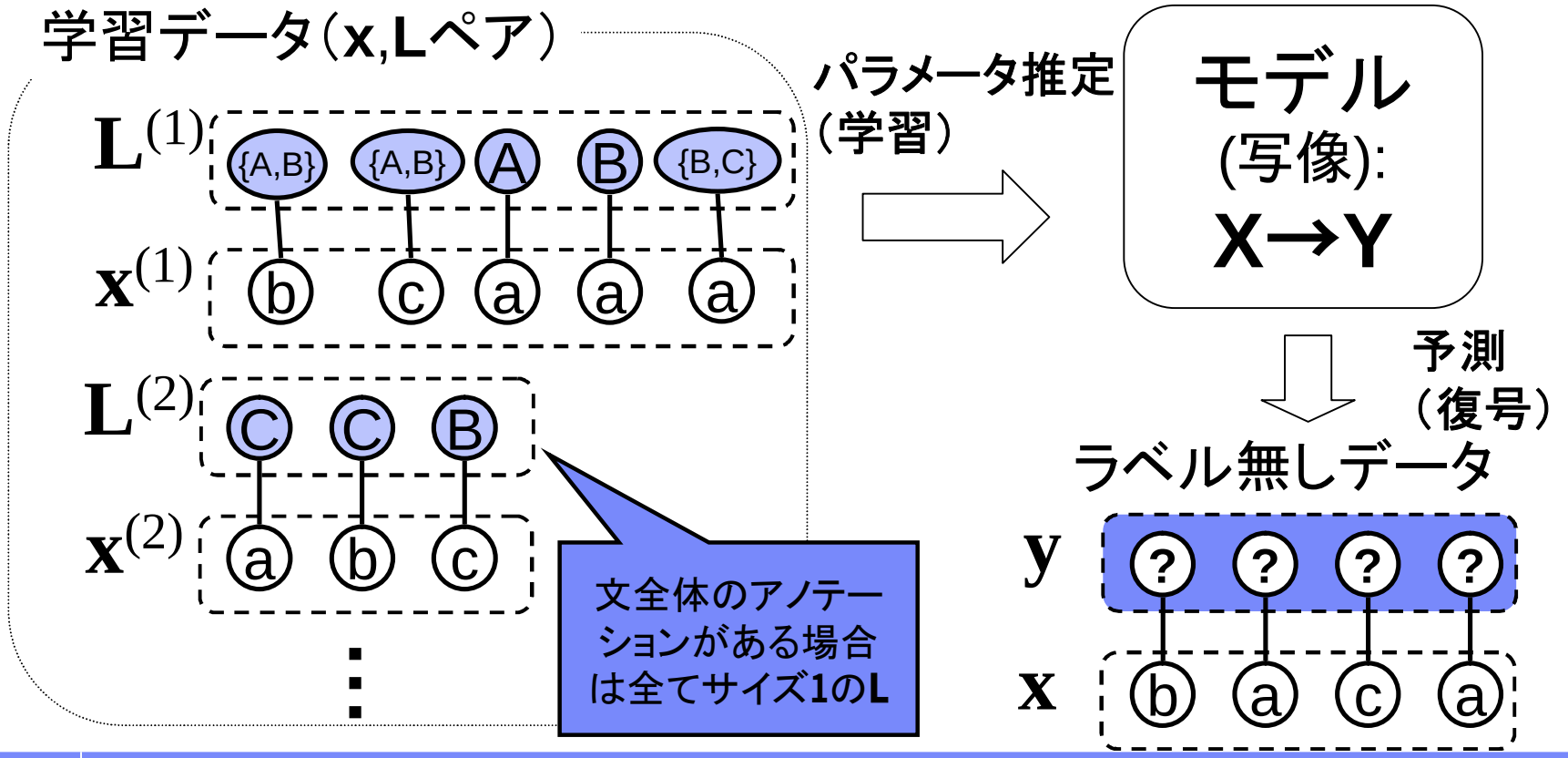
- **曖昧なアノテーション**
  - 点tのラベル候補をL_tで表現



部分的かつ曖昧な構造アノテーションからの教師付き学習

入力列xとラベル集合列Lの対データを用いて統計モデルを学習

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



目次

- 研究の背景
  - 構造出力問題としての単語分割・品詞付与タスク
- 部分的アノテーションと曖昧なアノテーション
  - 部分的かつ曖昧なアノテーションと教師付き学習の定式化
- 部分的かつ曖昧なアノテーションを用いた条件付確率場の学習
  - 条件付確率場(Conditional Random Fields: CRF)
  - 周辺尤度最大化による学習法の提案
- 実験
  - 単語リストによる部分的アノテーションを活用した日本語単語分割の分野適応実験
  - Penn Treebank コーパスの曖昧なアノテーションを用いた品詞付与実験
- まとめと今後の課題

部分的かつ曖昧なアノテーションを用いた条件付確率場の学習
条件付確率場 (Conditional Random Fields; CRF)

- 入力xが与えられた時の出力ラベル構造yの条件付確率をモデル化（識別モデル）

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

あるラベル構造yのスコア

あらゆるラベル構造yのスコアの合計（マルコフ性を仮定すると動的計画法で効率的に計算可能）

$\Phi : X \times Y \rightarrow R^d$ はx, yのペアの特徴を現す任意の素性ベクトル

$\theta \in R^d$は素性に対する重みを表すモデルパラメータ

θ が決まった下では、ラベル構造予測は $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathbf{Y}} P_\theta(\mathbf{y}\mid\mathbf{x})$

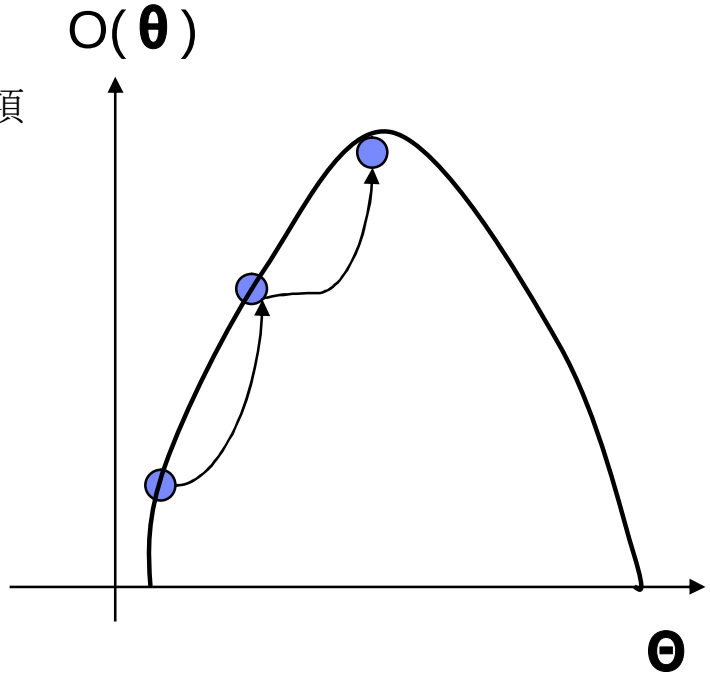
- 柔軟な素性情報を用いることが可能
- 構造全体の整合性を評価してモデルを学習

部分的かつ曖昧なアノテーションを用いた条件付確率場の学習
CRFのパラメータ推定（学習）の概要
勾配法による目的関数の最大値探索問題

1. 現在のパラメータの目的関数と偏微分(傾き)の計算(MAP推定)

$$O(\boldsymbol{\theta}) = \underbrace{\sum_{i \in \text{data}} \log P_{\boldsymbol{\theta}}(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)})}_{\text{対数尤度}} + \underbrace{\log P(\boldsymbol{\theta})}_{\substack{\text{パラメータの} \\ \text{事前分布=正則化項}}}$$

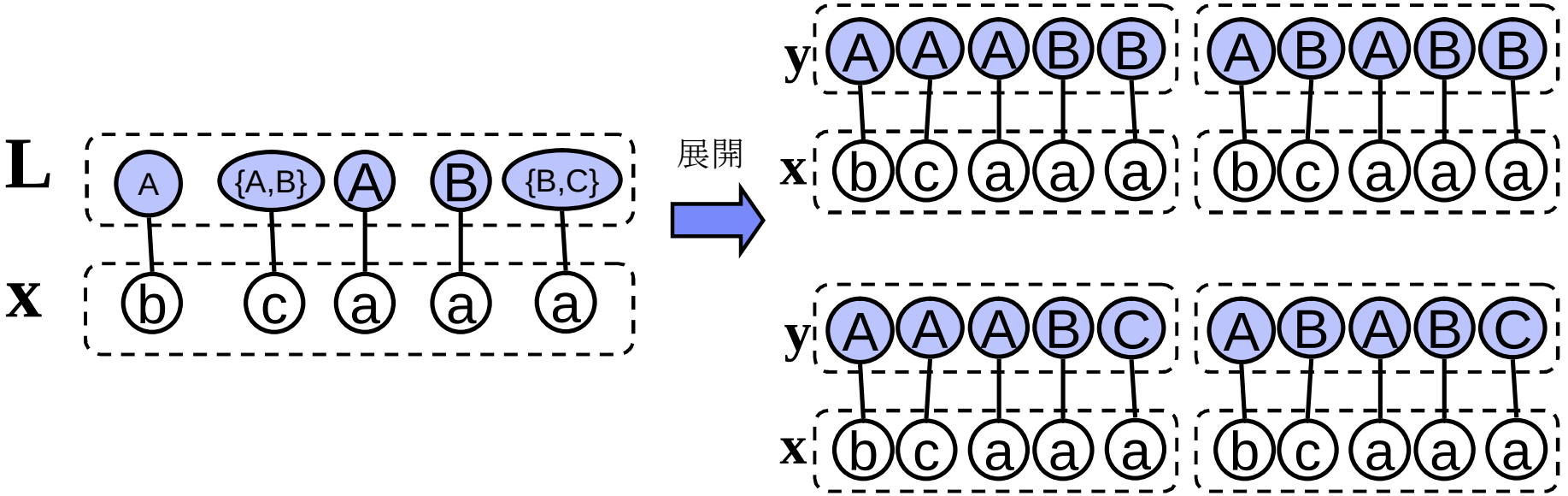
$$\frac{O(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\sum_{i \in \text{data}} \log P_{\boldsymbol{\theta}}(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}) + \log P(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$



2. 勾配が0ならば終了
そうでなければ、パラメータ θ を
（たとえば、最大勾配方向に）更新して、1に戻る

部分的かつ曖昧なアノテーションを用いた条件付確率場の学習
CRFが入力列xとラベル列yの組を学習データとして要求する
→xとラベル列集合Lから学習でない。

- 素朴な解決方法→ラベル集合Lを展開して、複数のx, y組を生成する



- 問題1：完全にアノテーションがある文の4倍の学習データ！？
- 問題2：サイズ2以上のラベル集合L_tが増えると指数的に事例数が増える

→適切な重み付けと動的計画法により解決

部分的かつ曖昧なアノテーションを用いた条件付確率場の学習

周辺尤度最大化によるCRFの学習

Y_L（Lを満たす全てのラベル列"集合"）の尤度 $P_\theta(\mathbf{Y}_L \mid \mathbf{x})$ を最大化

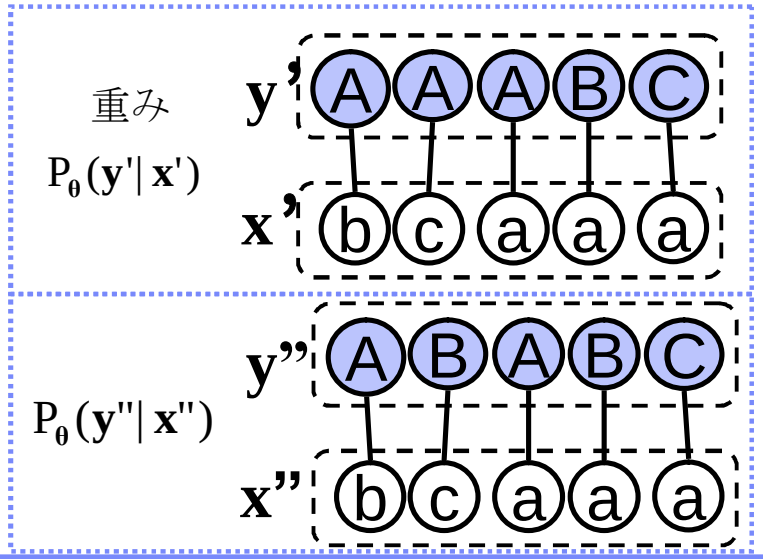
- **CRFの目的関数拡張→ L**を満たすラベル列の尤度合計（周辺化）

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

凸関数

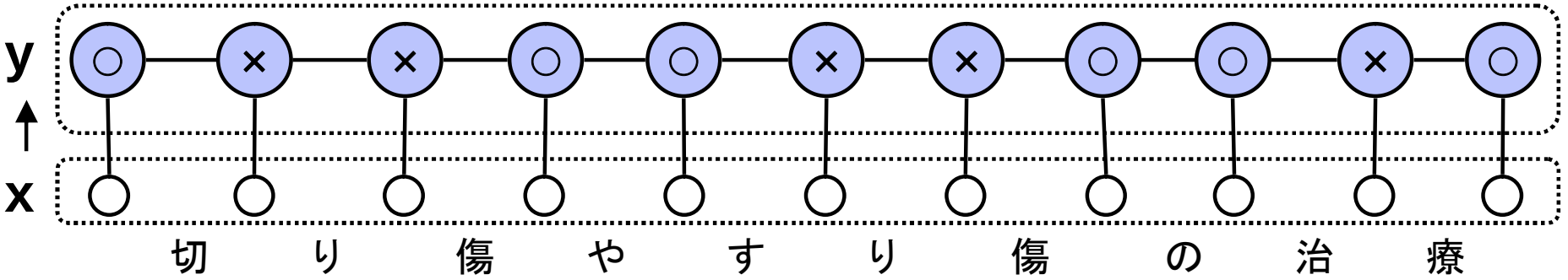
各点tで $y_t \in L_t^{(i)}$ となるラベル列の"集合"の条件付確率

- **周辺尤度最大化の直感的な解釈**
  - 学習中のモデルP_θで展開した**x, y**組それぞれに重み付け
  - P_θは他の学習データの尤度を最大化するように決まるため、他のアノテーションからの学習の邪魔をしにくい



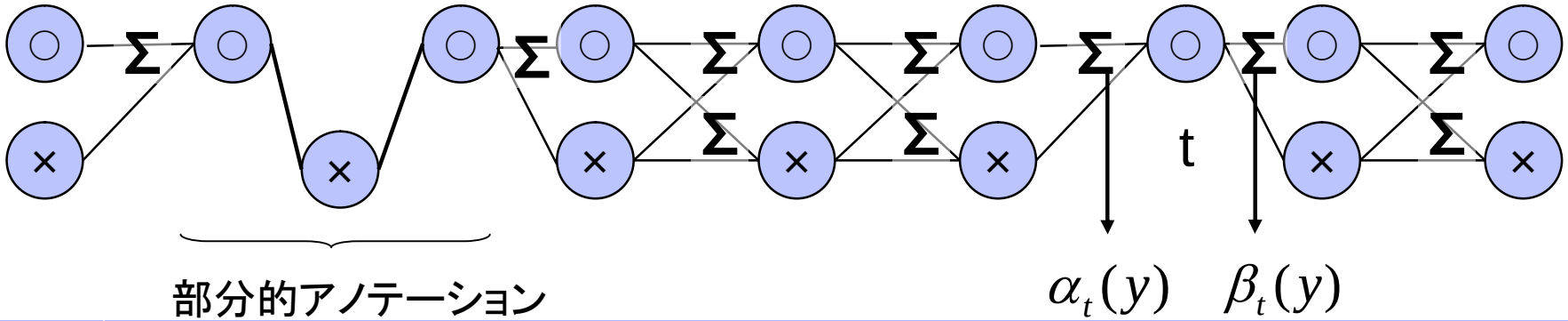
Lを満たす全てのラベル列の尤度合計の計算
ラベル間の依存関係にマルコフ性を仮定すると動的計画法により効率的に計算可能(詳細略)

例：1次のマルコフ性を仮定（各点のラベルは前後のラベルにのみ依存）



各点tで $y_t \in L_t$ となる全てのラベル列の合計スコアは、
前後ラベルまでの合計スコアで再帰的に計算可能

重複計算を避けるために
配列に計算結果を格納



目次

- 研究の背景
  - 構造出力問題としての単語分割・品詞付与タスク
- 部分的アノテーションと曖昧なアノテーション
  - 部分的かつ曖昧なアノテーションと教師付き学習の定式化
- 部分的かつ曖昧なアノテーションを用いた条件付確率場の学習
  - 条件付確率場(Conditional Random Fields: CRF)
  - 周辺尤度最大化による学習法の提案
- 実験
  - 単語リストによる部分的アノテーションを活用した日本語単語分割の分野適応実験
  - Penn Treebank コーパスの曖昧なアノテーションを用いた品詞付与実験
- まとめと今後の課題

日本語単語分割の分野適応実験
分野適応実験データ（会話辞典の例文から医療マニュアルへの分野適応）

- 適応元データ(source domain) : 会話辞典の例文
  - 例文 : こんな 失敗 は ご 愛敬 だ よ .
- 適応先データ(target domain) : 医療マニュアル
  - 例文 : 細胞膜には受容体があり、これによって細胞を識別することができます。

	分野	文数	単語数	分割済	目的
A	会話例文(source)	11,700	145,925	○	学習・素性選択
B	会話例文(source)	1,300	16,348	○	ハイパーパラメータのチューニング
C	医療マニュアル(target)	1,000	29,216	○	アノテーション・学習・テスト
D	医療マニュアル(target)	53,834	N/A	×	素性選択

  - 適応先の単語リストは、会話例文Aに出てこない医療マニュアルCの単語（未知語）から作成

日本語単語分割の分野適応実験

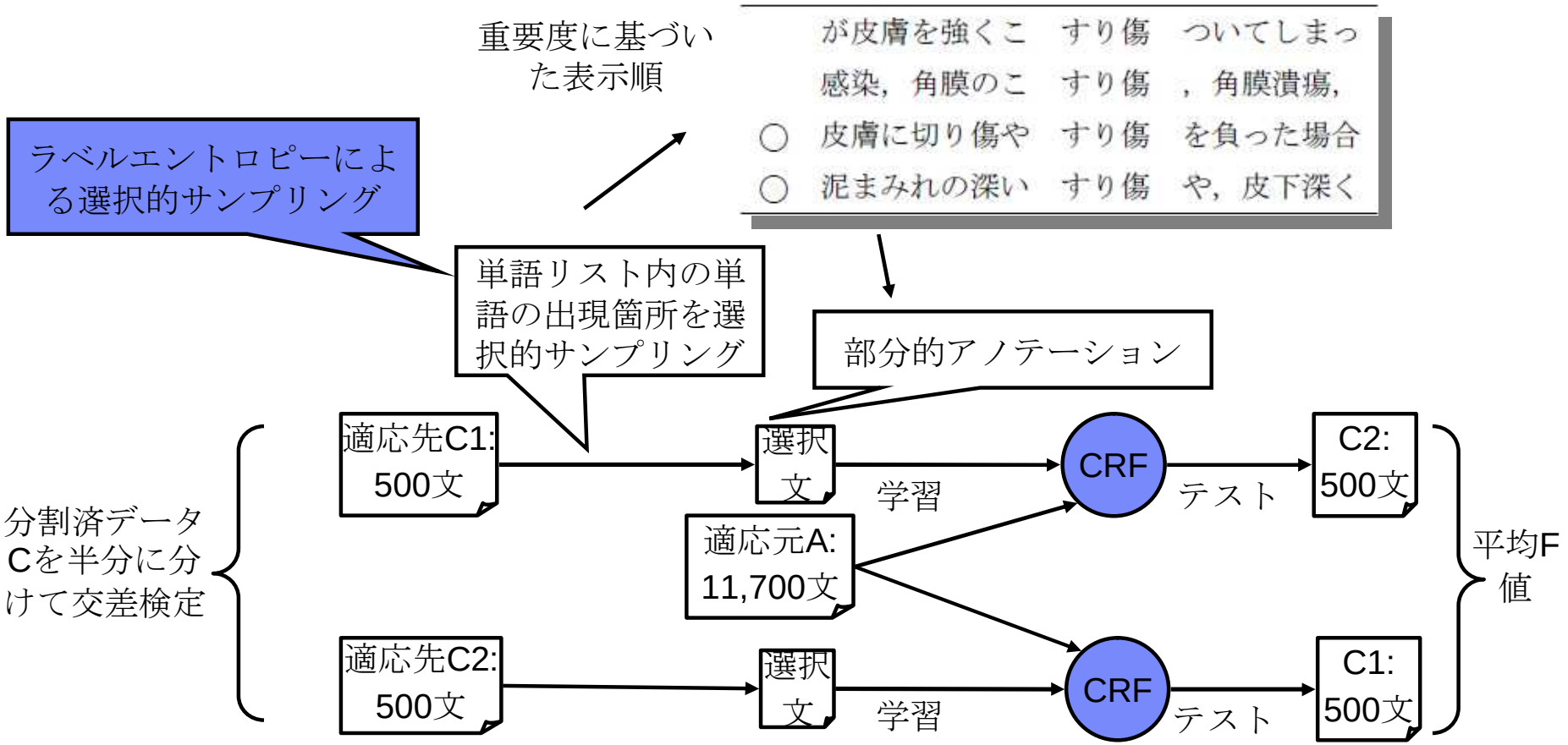
分野適応の必要性

- **学習データとは分野が異なるデータでは性能低下**
  - 「会話例文」で学習した単語分割器で「医療マニュアル」文書を処理
    →5%性能低下
    - 会話例文：こんな 失敗はご愛敬だよ .
    - 医療マニュアル：細胞膜には受容体があり、これによって細胞を識別することができます。
  - 特に、学習データには出現しない単語（未知語）がエラーの主要因

日本語単語分割の分野適応実験

実験シナリオ：単語リストによるアノテーション結果から学習

単語リストを使用したKWIC形式のUIによる部分的アノテーションを想定



日本語単語分割の分野適応実験

実験の設定と評価方法

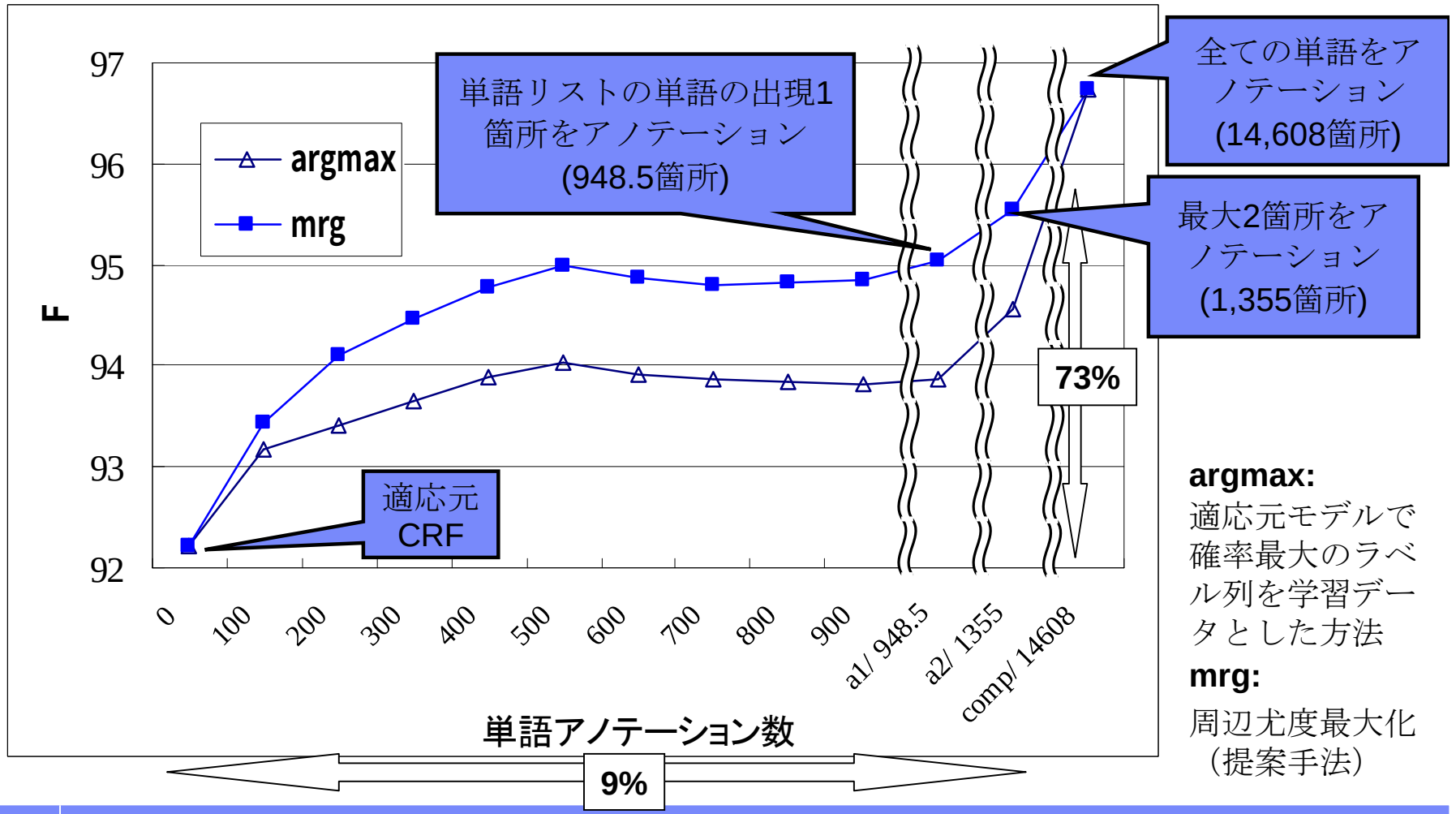
- 入力x_tの素性：各文字境界の前後2文字(4文字)の範囲から抽出できる文字N-gram (n=1,2,3)と字種N-gram (n=1,2,3)すべて。
  - 字種は、ひらがな(H)、カタカナ、漢字(C)、アルファベット、アラビア数字、記号
  - 例：「やすり傷」の中央文字境界を表す素性→
    {す|, | り, やす|, す|り, | り傷, やす| り, す| り傷, H|, |H, HH|, H|H, |HC, HH|H, H|HC} (“|” は注目する文字境界との相対位置を示す補助記号)
  - パラメータ数削減のため高頻度の素性のみを選択
  - 素性数は298, 363
- CRFは1次のマルコフモデル
- 評価指標：F値
  - 精度と再現率の調和平均 F=2PR/(R+P)

$$R = \frac{\text{正解単語数}}{\text{全単語数}} \times 100$$

$$P = \frac{\text{正解単語数}}{\text{システムの出力単語数}} \times 100.$$

単語のアノテーション箇所数を変化させたときの単語分割性能向上（交差検定）

提案手法は文全体をアノテーションするのに比べて、9%のアノテーション箇所で性能向上の73%を達成 (F値は精度と再現率の調和平均)



argmax:
適応元モデルで確率最大のラベル列を学習データとした方法

mrg:
周辺尤度最大化（提案手法）

目次

- 研究の背景
  - 構造出力問題としての単語分割・品詞付与タスク
- 部分的アノテーションと曖昧なアノテーション
  - 部分的かつ曖昧なアノテーションと教師付き学習の定式化
- 部分的かつ曖昧なアノテーションを用いた条件付確率場の学習
  - 条件付確率場(Conditional Random Fields: CRF)
  - 周辺尤度最大化による学習法の提案
- 実験
  - 単語リストによる部分的アノテーションを活用した日本語単語分割の分野適応実験
  - Penn Treebank コーパスの曖昧なアノテーションを用いた品詞付与実験
- まとめと今後の課題

曖昧なアノテーションを用いた品詞付与実験
実験データ

- 学習データ：品詞が曖昧な単語と曖昧でない単語を含む文
- 評価用データ：品詞が曖昧でない単語を含む文(11, 840文)

→ 異なるデータで5回試行

Penn Treebank コーパス品詞タグ付け学習データ

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…

⋮

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

⋮

品詞が曖昧な単語を含む文 (118文)

全ての単語の品詞が確定した文 (1,480文 or 2,960)

Penn Treebank コーパス品詞タグ付けテストデータ

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

⋮

全ての単語の品詞が確定した文 (11,840文)

曖昧なアノテーションを用いた品詞付与実験
実験の設定

- 入力x_tの素性：二値素性
  - 単語自身の文字列
  - 単語末尾1,2,3 文字列
  - 先頭が大文字
  - 先頭が数字
  - 先頭が大文字かつドット(．)を含む
  - 全て大文字
  - 全て小文字
  - ハイフン(-)を含む
  - 句読点を含む，文の最後が“,”，“?”，“!” で終わる
- CRFは1次のマルコフモデル

曖昧なアノテーションを用いた品詞付与実験

提案手法との比較手法

- 品詞の曖昧性をルールによって解消し学習データとする
  That/DT suit/NN is/VBZ **pending/VBG|JJ** ./SYM →
  （品詞を一つ選択）That/DT suit/NN is/VBZ **pending/VBG** ./SYM

1. ランダム：乱数によって曖昧な品詞集合から1つ品詞を選択
   pending/VBG|JJ → **pending/JJ**

2. 記述順：最初に記述された品詞を選択
   pending/VBG|JJ → pending/VBG

3. 頻度順：高頻度の品詞を選択
   pending/VBG|JJ → pending/VBG
   （VBGの頻度 ＞ JJの頻度）

曖昧な品詞タグからの学習
品詞タグ付けタスクの性能（試行5回の平均）
曖昧性をルールで解消に比べて安定した性能（データ全体の性能を保持しつつ、曖昧な品詞の単語の性能も満たす。）

- **評価指標**:

$$\text{全体正解率} = \frac{\text{全単語品詞タグ正解数}}{\text{全単語出現数}} \times 100$$

$$\text{曖昧語正解率} = \frac{1}{|A|}\sum_{w \in A}\frac{w\text{の品詞タグ正解数}}{w\text{の出現数}} \times 100$$

A:曖昧な品詞アノテーションが存在する単語**82** 種の集合

- **実験結果**

品詞曖昧文（118文） ＋

	正解率	提案法	ランダム	記述順	頻度順
品詞確定文 (1,480文)	全体	**94.274**	**94.274**	94.262	**94.274**
	曖昧語	**73.272**	71.582	72.658	71.68
品詞確定文 (2,960文)	全体	**94.982**	94.98	94.974	94.976
	曖昧語	**76.242**	74.276	75.28	74.326

部分的かつ曖昧なラベル付き構造データからのマルコフ条件付確率場の学習 発表の概要

- 研究の背景
  - 構造出力問題としての単語分割・品詞付与タスク
- 部分的アノテーションと曖昧なアノテーション
  - 部分的かつ曖昧なアノテーションと教師付き学習の定式化
- 部分的かつ曖昧なアノテーションを用いた条件付確率場の学習
  - 条件付確率場(Conditional Random Fields: CRF)
  - 周辺尤度最大化による学習法の提案
- 実験
  - 単語リストによる部分的アノテーションを活用した日本語単語分割の分野適応実験
  - Penn Treebank コーパスの曖昧なアノテーションを用いた品詞付与実験
- まとめと今後の課題

まとめと今後の課題

- **曖昧かつ部分的アノテーションを用いた条件付確率場の学習方法提案**
  - 日本語単語分割の分野適応実験において、部分的アノテーションによって効率的な適応先性能向上を実現
  - Penn Treebankコーパスに存在する曖昧な品詞タグからの学習において、ルールによる解決に比べて安定的な性能を示した
- **今後の課題**
  - **部分的アノテーションの追加に関して必ずしも性能が単調増加しない**
    アノテーション数500-800の区間で短く切るべき単語を誤って長めに区切るエラーが増加。ラベルエントロピーによる選択的サンプリングでは、長めの単語が500-700に選択されていた。
    →単語非境界ラベル（×）が学習データに多く出現
    部分的アノテーションによる分布の歪みの補正が必要