言語処理における識別モデルの発展
-- HMMからCRFまで --

坪井祐太, **鹿島久嗣**（IBM Research）
工藤 拓 (Google)

今日話した内容

- HMMの話
- 復号(Viterbi)
- 内積で書けるよ
- 内積で書ける別のモデル CRF, HM Perceptron, HM SVM

- 違いはパラメータ推定法
- Generative Model
- 　　HMM
- 　　HMM課題

- Discriminative Modelのメリット：直接P(Y|X), 素性間の重なりを考慮できる

- Conditional Model
- 　　CRF
-
- Discriminative Model
- 　　HM Perceptron
- 　　メリット：
- Generative Model vs Discriminative のまとめた表

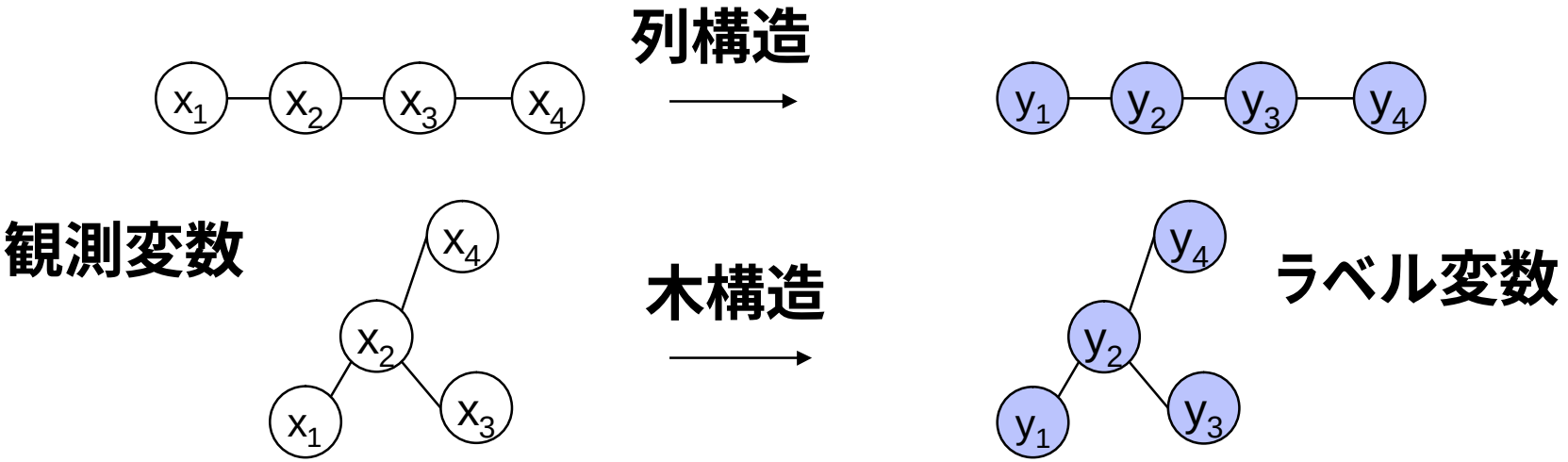
- 形態素解析の例

チュートリアルの流れ

- 構造のラベル付け問題の定義
- 2つのアプローチ：生成モデル vs. 識別モデル
  - 生成モデル：隠れマルコフモデル
  - 識別モデル：条件付確率場
- そのほかの識別モデル
- 計算機実験による性能比較

構造ラベル付与問題とは？

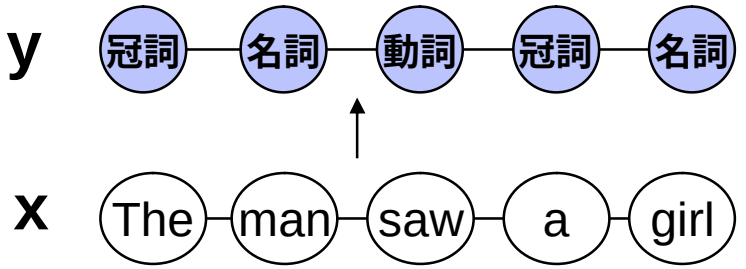
- 観測されたデータ構造xに対応するラベル構造yへの写像を学習する問題



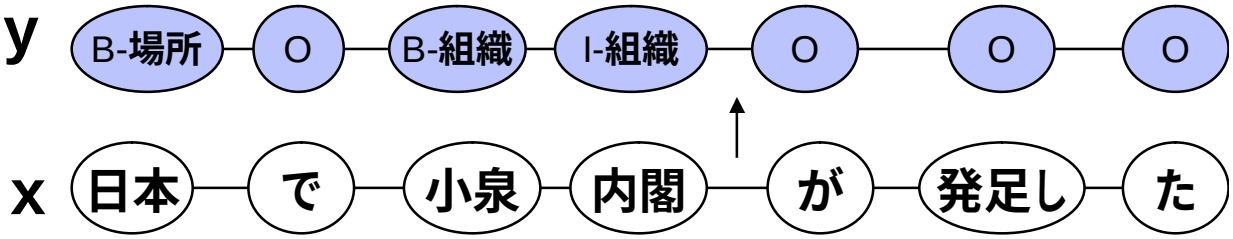
- 構造を持つ観測データの各ノードにラベルを付与する問題として捉えることが出来る応用が多く存在する。
  - 例：自然言語処理、バイオインフォマティックス

自然言語処理での構造ラベル付与問題の例 (列構造)

- **品詞タグ付与タスク**
  単語列に対して品詞ラベルを付与するタスク

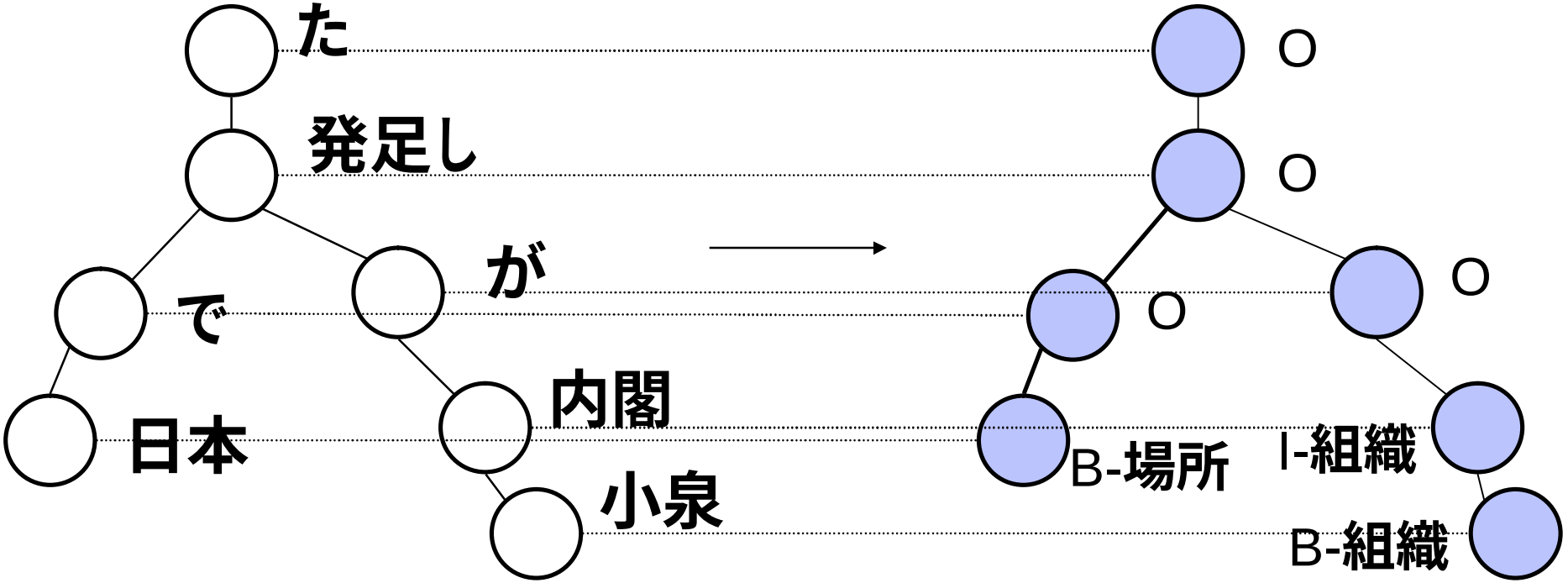


- **固有表現抽出タスク**
  人名・組織名等の固有表現をテキスト中から抽出
  単語列に対して固有表現の始まり(B-XXX) と続く固有表現(I-XXX) を示すラベルを付与するタスク (O は固有表現以外)

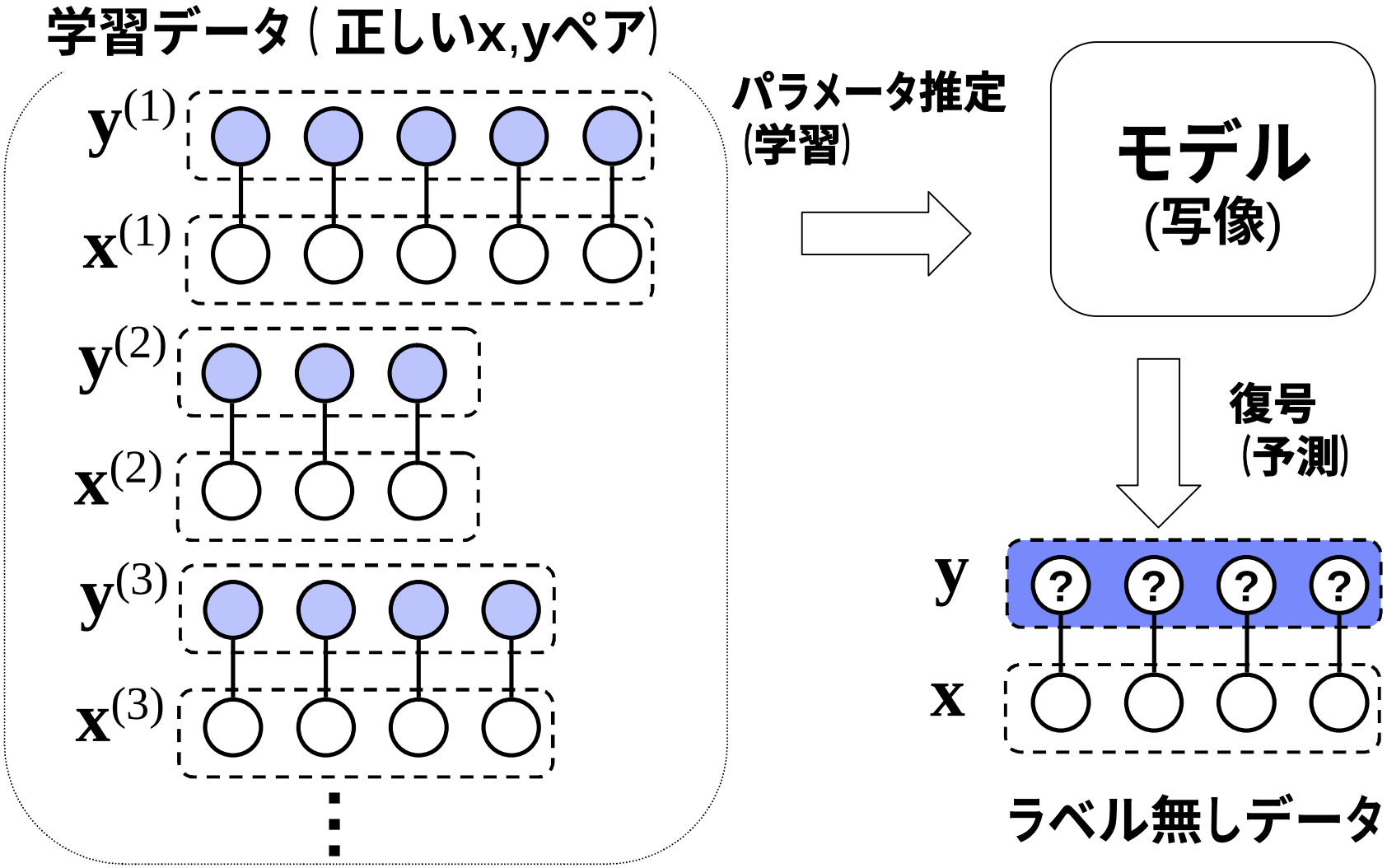


自然言語処理での構造ラベル付与問題の例 (木構造)

- 係り受け木に対する固有表現抽出タスク
  - 係り受け解析によって生成された単語間の関係を表す係り受け木に対して、ラベルを付与。
  - 主語と述語の関係など言語構造を考慮したラベル付与



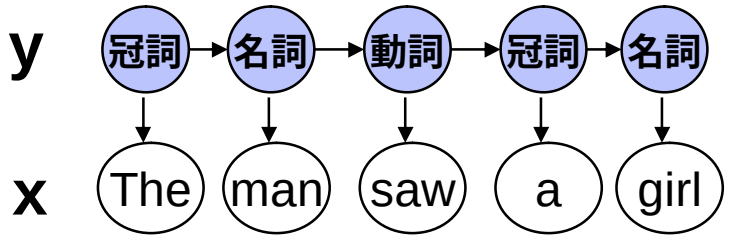
教師付き学習による構造ラベル付与問題



構造ラベル付与学習モデル

- **生成モデルに基づく手法**
  - **隠れマルコフモデル**(HMM)
- **識別モデルに基づく手法**
  - **条件付確率場**(CRF)

生成モデルによる構造ラベル付与学習
隠れマルコフモデル(HMM)

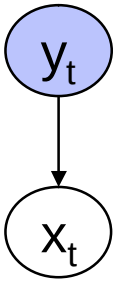


- xとyの同時分布に基づくモデル
- 生成確率を出力確率と遷移確率に分解してモデル化

$$P(\mathbf{x},\mathbf{y}) = P(\mathbf{x}\mid\mathbf{y})\,P(\mathbf{y})$$

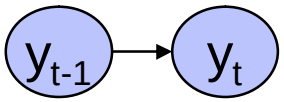
$$= \prod_{t=1}^{T} P(x_t|y_t)P(y_t \mid y_{t-1})$$

(Tは構造x, yのサイズ)



出力確率

(各ノード間の独立性仮定)



遷移確率

(マルコフ性仮定)

隠れマルコフモデル(HMM):
ラベル列の予測 (復号問題)

predict

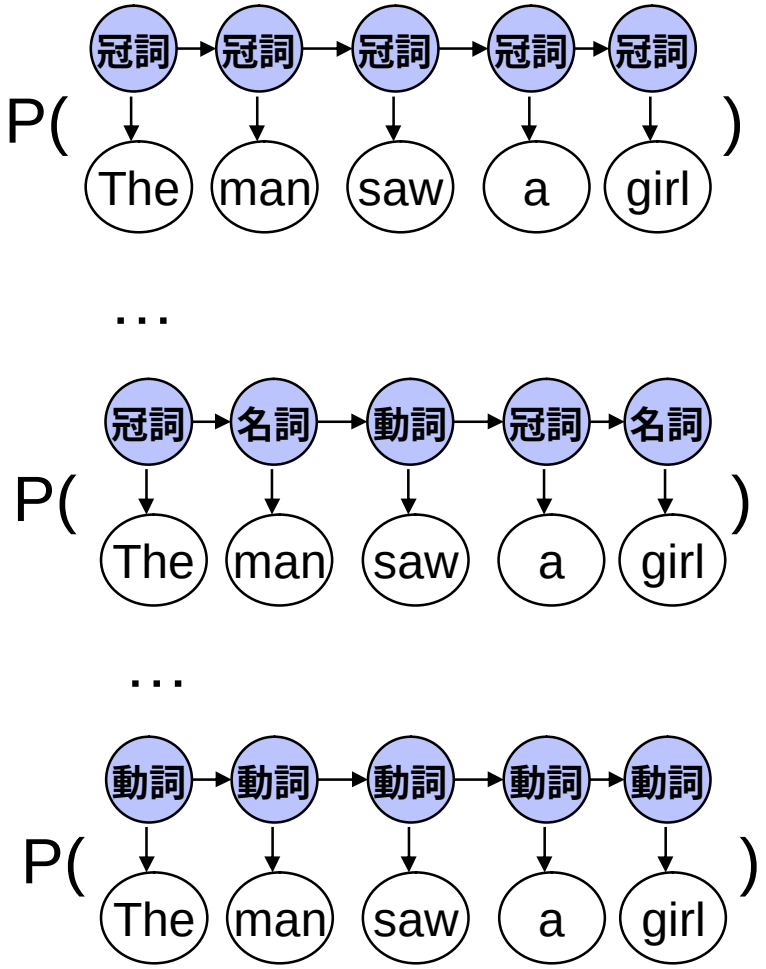
$$\mathbf{y} = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{x} \mid \mathbf{y}) = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{x}, \mathbf{y})$$

$$= \underset{\mathbf{y}}{\operatorname{argmax}} \prod_{t=1}^{T} P(x_t|y_t) P(y_t \mid y_{t-1})$$

あるxが与えられたときに、確率が最大になるラベル列yを見つけたい。

→ありうるラベル列全て(|Y|^{T}個)の列挙は非効率 (|Y| は目的ラベル集合のサイズ)

全部で|Y|^{T}個



隠れマルコフモデル(HMM):
Viterbi 復号法による最適ラベル列の求め方

- **位置tまでのラベル列の確率が最大になる値を記憶するテーブルδを使い再計算を避けることで、最適なラベル列を効率的に計算**

$$\delta_1(y) = \mathrm{P}(x_1 \mid y)$$

$$\delta_t(y) = \max_{\tilde{y} \in \mathbf{Y}} \delta_{t-1}(\tilde{y}) P(x_t \mid y) P(y \mid \tilde{y})$$

英語品詞タグ付けタスクでのδ

y \ x	the	man	saw	…
冠詞	$\mathrm{P}(\text{the} \mid \text{冠詞})$	$\max_{\tilde{y} \in \mathbf{Y}} \delta_1(\tilde{y}) \times \mathrm{P}(\text{man} \mid \text{冠詞})\, \mathrm{P}(\text{冠詞} \mid \tilde{y})$	$\max_{\tilde{y} \in \mathbf{Y}} \delta_2(\tilde{y}) \times \mathrm{P}(\text{saw} \mid \text{冠詞})\, \mathrm{P}(\text{冠詞} \mid \tilde{y})$	…
名詞	$\mathrm{P}(\text{the} \mid \text{名詞})$	$\max_{\tilde{y} \in \mathbf{Y}} \delta_1(\tilde{y}) \times \mathrm{P}(\text{man} \mid \text{名詞})\, \mathrm{P}(\text{名詞} \mid \tilde{y})$	$\max_{\tilde{y} \in \mathbf{Y}} \delta_2(\tilde{y}) \times \mathrm{P}(\text{saw} \mid \text{名詞})\, \mathrm{P}(\text{名詞} \mid \tilde{y})$	…
動詞	$\mathrm{P}(\text{the} \mid \text{動詞})$	$\max_{\tilde{y} \in \mathbf{Y}} \delta_1(\tilde{y}) \times \mathrm{P}(\text{man} \mid \text{動詞})\, \mathrm{P}(\text{動詞} \mid \tilde{y})$	$\max_{\tilde{y} \in \mathbf{Y}} \delta_2(\tilde{y}) \times \mathrm{P}(\text{saw} \mid \text{動詞})\, \mathrm{P}(\text{動詞} \mid \tilde{y})$	…

隠れマルコフモデル(HMM):
Viterbi 復号法による σ_t の計算例

- δ_t を最大にする y_{t-1} から y_t への遷移（矢印）が決まった時の δ の例（英語品詞タグ付け）

x / y	the	man	saw	…
冠詞	P(the | 冠詞)	P(the | 動詞) × P(man | 冠詞) P(冠詞 | 動詞)	P(the | 名詞) × P(man | 動詞) P(動詞 | 名詞) × P(saw | 冠詞) P(冠詞 | 動詞)	…
名詞	P(the | 名詞)	P(the | 冠詞) × P(man | 名詞) P(名詞 | 冠詞)	P(the | 名詞) × P(man | 動詞) P(動詞 | 名詞) × P(saw | 名詞) P(名詞 | 動詞)	…
動詞	P(the | 動詞)	P(the | 名詞) × P(man | 動詞) P(動詞 | 名詞)	P(the | 冠詞) × P(man | 名詞) P(名詞 | 冠詞) × P(saw | 動詞) P(動詞 | 名詞)	…

隠れマルコフモデル(HMM):
テーブルπを用いた最適ラベル列の求め方 (Viterbi 復号法)

- 位置tまでのラベル列の確率が最大になる位置t-1のラベルを記憶するテーブルπ も同時に計算
- 末端(T)において確率($\delta_T(y)$)が最大になるyの$\pi_T(y)$からバックトラックすることで、確率が最大になるラベル列を得ることができる。

$$\pi_t(y) = \underset{\tilde{y} \in \Sigma_y}{\operatorname{argmax}}\, \delta_{t-1}(\tilde{y}) P(x_t \mid y) P(y \mid \tilde{y})$$

y \ x	the	man	saw	…
冠詞		π_2(冠詞)=動詞	π_3(冠詞)=動詞	…
名詞		π_2(名詞)=冠詞	π_3(名詞)=動詞	…
動詞		π_2(動詞)=名詞	π_3(動詞)=名詞	…

隠れマルコフモデルのパラメタ推定（学習）

- **パラメータ θ**
  - **出力確率** $P(x_t|y_t)$
  - **遷移確率** $P(y_t|y_{t-1})$
- **最尤推定**

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_i \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$$

（iは学習データの索引）

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_i \prod_t^{T^{(i)}} \mathrm{P}_{\theta}(x_t^{(i)} \mid y_t^{(i)}) \, \mathrm{P}_{\theta}(y_t^{(i)} \mid y_{t-1}^{(i)})$$

- **最尤なパラメータは共起頻度のカウントで計算可能**

生成モデルの課題点

- 同時分布の推定をすることで、間接的に分類問題を解いている。(P(x)を余分に推定している)

$$\mathbf{y}^{predict} = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{y} \mid \mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{x}, \mathbf{y})$$

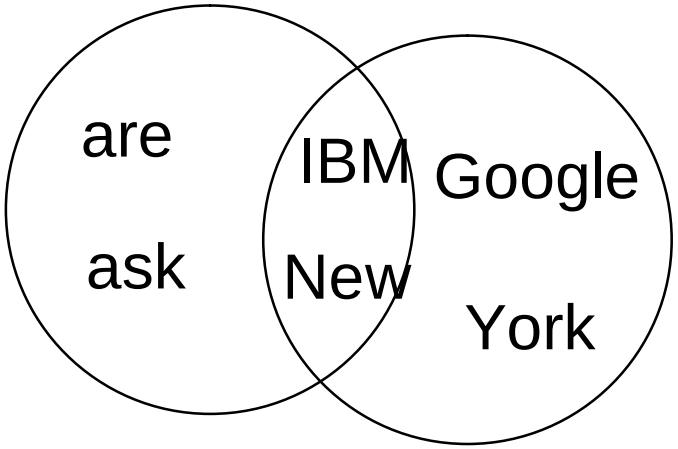
- 同時分布を推定する十分なデータを得るのは難しいため、変数間の独立性を仮定することが多い。

$$\mathbf{y}^{predict} = \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{t=1}^{T} P(x_t|y_t) P(y_t \mid y_{t-1}) \qquad \text{(HMM)}$$

→相互作用のある観測変数（素性）をうまく扱えない

相互作用のある素性の例

- **自然言語処理では単語それ自身以外に、単語の部分文字列等が素性に使われることが多い**
- **品詞タグ付け**
  - P(beautiful|**形容詞**), P(ful**で終わる単語**|**形容詞**)
  - P(be**で始まる単語**|**形容詞**)
  - P(immediately|**副詞**), P(ly**で終わる単語**|**副詞**)
- **固有表現抽出**
  - P(3**文字の単語** | B-**組織**), P(**最初が大文字の単語** |B-**組織**)
  - P(**小泉**|B-**人名**), P(**漢字だけからなる単語**| B-**人名**)



足して1にならないといけないのになってない

識別モデル

- xからyを直接推定するモデル
- 識別モデルの利点
  - 直接分類問題を解くことが出来る
  - 素性間の重なりを考慮して重みを学習
- 多クラスのロジスティック回帰モデル (最大エントロピーモデル)
  - 確率分布の形をした識別モデル
  - 条件付分布を直接モデル化

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

内積だよ

(Φはx,yの素性、θは素性に対する重み)

復号は一緒

識別モデルのパラメータ推定（学習）

- **最尤推定**

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \prod_{i \in \text{trainingdata}} \mathrm{P}_{\boldsymbol{\theta}}(y^{(i)} \mid \mathbf{x}^{(i)})$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i \in \text{trainingdata}} \frac{\exp(\left\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{x}^{(i)}, y^{(i)}) \right\rangle)}{\sum_{\tilde{y} \in \mathbf{Y}} \exp(\left\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{x}^{(i)}, \tilde{y}) \right\rangle)}$$

- **パラメータの計算**
  - 素性間の相互作用を考慮してパラメータ推定をする必要があり、計算手順が複雑
  - 損失関数の偏微分を用いて損失最小化問題を解く。

識別モデルによる構造ラベル付与学習：
条件付確率場 (Conditional Random Fields: CRF)

- ロジスティック回帰モデルを基に、ローカルな変数間の関係を素性(遷移素性)で表現したモデル

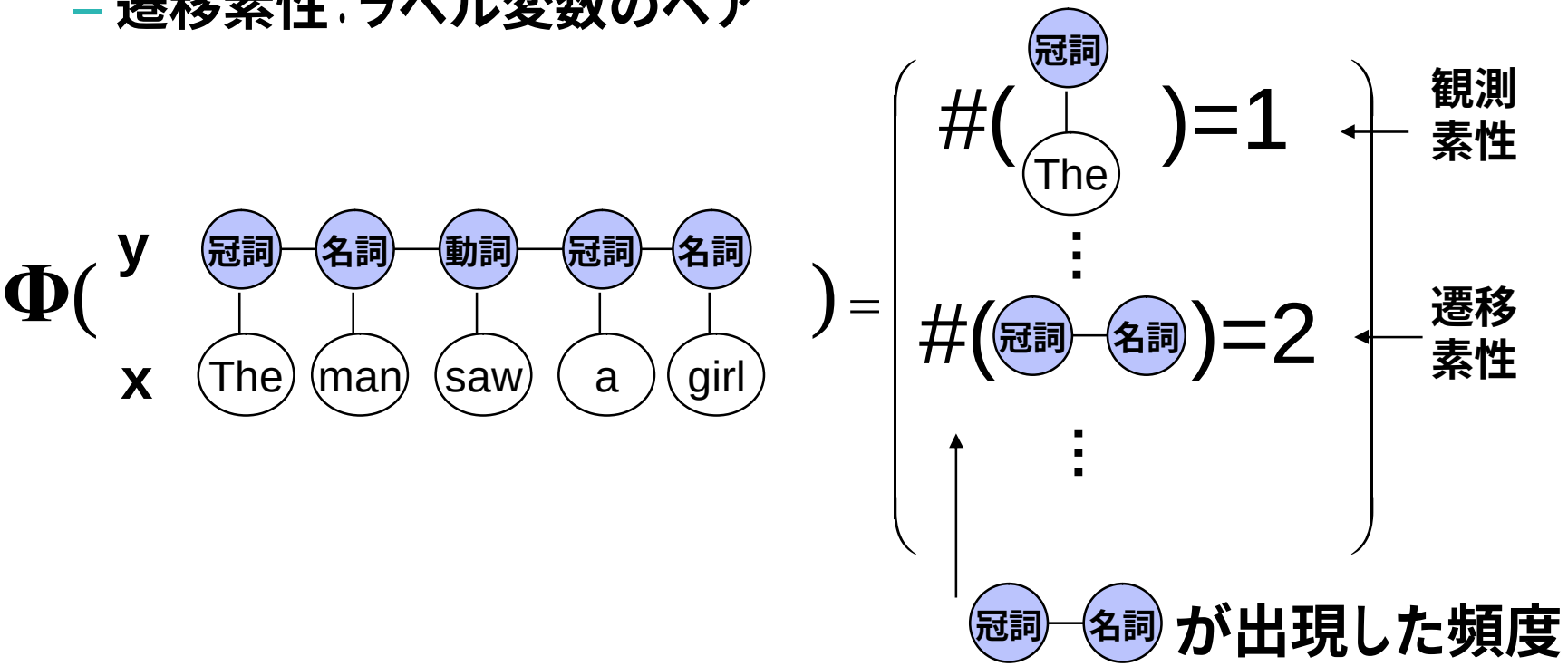
$$P(\mathbf{y} \mid \mathbf{x}) = \frac{\exp\left(\left\langle \boldsymbol{\theta}, \phi(\mathbf{x}, \mathbf{y}) \right\rangle\right)}{\sum_{\tilde{\mathbf{y}}} \exp\left(\left\langle \boldsymbol{\theta}, \phi(\mathbf{x}, \tilde{\mathbf{y}}) \right\rangle\right)}$$

$$= \frac{\exp\left(\sum_{\tau=1}^{T} \left\langle \theta, \phi(\mathbf{x}, \mathbf{y}_{\tau}^{\tau+1}) \right\rangle\right)}{\sum_{\tilde{\mathbf{y}}} \exp\left(\sum_{\tau=1}^{T} \left\langle \theta, \phi(\mathbf{x}, \tilde{\mathbf{y}}_{\tau}^{\tau+1}) \right\rangle\right)} \cdot \quad \mathbf{y}_{\tau}^{\tau+1} = (y_{\tau}, y_{\tau+1})$$

(Φはx,yの素性、θは素性に対する重み)

条件付確率場の素性

- 素性ベクトルの各要素は、素性が構造中に出現した頻度
  - 観測素性:観測変数とラベル変数のペア
  - 遷移素性:ラベル変数のペア



その他の識別モデル：分類手法に基づくアプローチ

- **CRFのパラメタ推定の別アプローチ**
  - **分類手法（パーセプトロン, SVM, …）にもとづいたアプローチ**
- **何がうれしいか？→ 通常のアプローチでは解けない問題に適用できる**

そのほかの識別モデル：隠れマルコフパーセプトロン

- CRFの推定は、訓練データの $y^{(i)}$ が出力される確率を「最大にする」ように学習される

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \prod_{i \in \text{trainingdata}} \mathrm{P}_{\boldsymbol{\theta}}(y^{(i)} \mid \mathbf{x}^{(i)})$$

- 別の考え方：訓練データの $y^{(i)}$ が出力される確率が、「ほかの $\tilde{y}^{(i)} \neq y^{(i)}$ の出力確率よりも大きければよい」

$$\log \mathrm{P}_{\boldsymbol{\theta}}(y^{(i)} \mid \mathbf{x}^{(i)}) > \log \mathrm{P}_{\boldsymbol{\theta}}(\tilde{y}^{(i)} \mid \mathbf{x}^{(i)})$$

$$\Leftrightarrow \left\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{x}^{(i)}, y^{(i)}) \right\rangle - \left\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{x}^{(i)}, \tilde{y}^{(i)}) \right\rangle > 0$$

隠れマルコフパーセプトロンのアルゴリズム

1. i 番目の訓練データに対して予測してみる

$$\mathbf{y}^{predict(i)} = \underset{\mathbf{y}}{\operatorname{argmax}}\, \mathrm{P}(\mathbf{y} \mid \mathbf{x}^{(i)})$$

2. 当たっていれば（ $\mathbf{y}^{predict(i)} \neq \mathbf{y}^{(i)}$ ）何もしない
3. 外れていたら、パラメータを修正

$$\boldsymbol{\theta}^{new} \leftarrow \boldsymbol{\theta}^{old} + \eta\left(\boldsymbol{\Phi}(\mathbf{x}^{(i)}, y^{(i)}) - \boldsymbol{\Phi}(\mathbf{x}^{(i)}, \tilde{y}^{(i)})\right)$$

4. 1.に戻って繰り返す

最尤推定と隠れマルコフパーセプトロンの違いは、argmax操作のみで実現できるところ

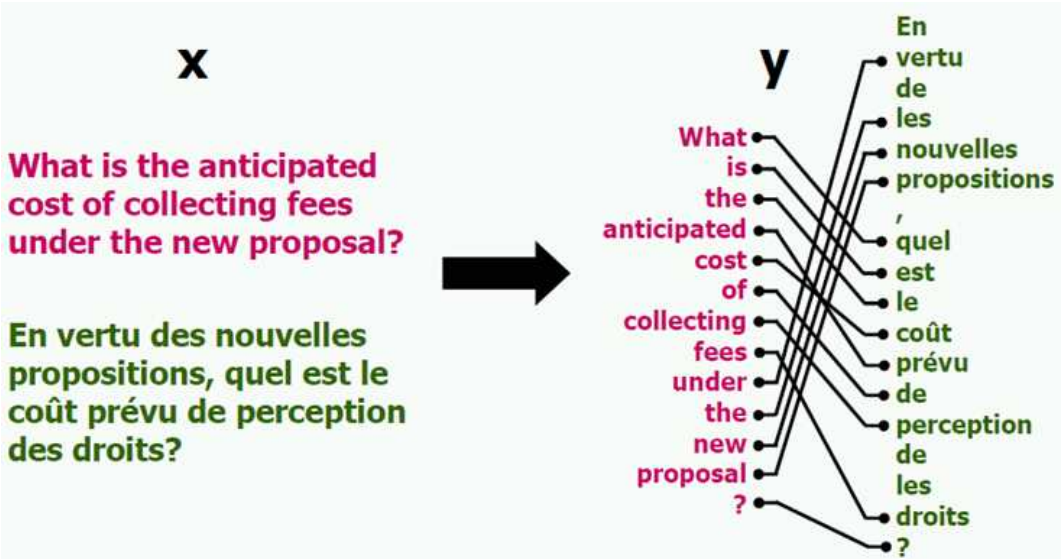
- 隠れマルコフパーセプトロンは訓練と予測が両方argmax操作

	CRF	隠れマルコフパーセプトロン
訓練	Σ Π	argmax
予測	argmax	argmax

- 通常、argmaxもΣ Πも動的計画法で実現できるのであまり変わりはない…しかし…

隠れマルコフパーセプトロンは、最尤推定では解けない問題が解ける？

- 問題によっては、argmax操作は、動的計画法以外の多項式時間アルゴリズムで実現できる
- 動的計画法で多項式時間で解けない問題は、最尤推定では解けない
  - たとえば、異なる言語の文章間での単語のマッチング
    - argmax操作が線形計画法（多項式時間）で解ける



図引用：

Dan Klein and Ben Tasker, “Max-Margin Methods for NLP: Estimation, Structure, and Applications”,

ACL 2005 Tutorial より

End of the Presentation

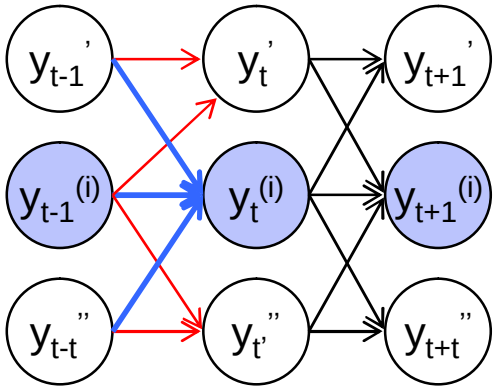
Reference

- **Dan Klein and Ben Tasker. Max-Margin Methods for NLP: Estimation, Structure, and Applications. ACL-05 Tutorial.**
- **Dan Klein and Chris Manning. Maxent Models, Conditional Estimation, and Optimization, without the Magic. ACL-03 Tutorial.**

Gradient of Point-wise Log-Loss

$$\frac{L}{\partial \boldsymbol{\theta}} = -\sum_{i \in \text{trainingdata}} \left(\sum_{t}^{T^{(i)}} \sum_{\hat{\mathbf{y}} \in \hat{y}_t = y_t^{(i)}} \mathrm{P}(\hat{\mathbf{y}} \mid \mathbf{x}^{(i)}) \boldsymbol{\Phi}(\mathbf{x}^{(i)}, \hat{\mathbf{y}}) - \sum_{\breve{\mathbf{y}} \in \mathbf{Y}} \mathrm{P}(\breve{\mathbf{y}} \mid \mathbf{x}^{(i)}) \boldsymbol{\Phi}(\mathbf{x}, \breve{\mathbf{y}}) \right)$$

$$\sum_{\hat{\mathbf{y}} \in \hat{y}_t = y_t^{(i)}} \mathrm{P}(\hat{\mathbf{y}} \mid \mathbf{x}^{(i)}) \boldsymbol{\Phi}(\mathbf{x}^{(i)}, \hat{\mathbf{y}}) = \frac{\sum_{\hat{\mathbf{y}} \in \hat{y}_t = y_t^{(i)}} \mathrm{P}(\hat{\mathbf{y}} \mid \mathbf{x}^{(i)})}{\sum_{\breve{\mathbf{y}}} \mathrm{P}(\breve{\mathbf{y}} \mid \mathbf{x}^{(i)})} \boldsymbol{\Phi}(\mathbf{x}^{(i)}, \hat{\mathbf{y}})$$



Classification of Structured Output (e.g. HMMs, CRFs)

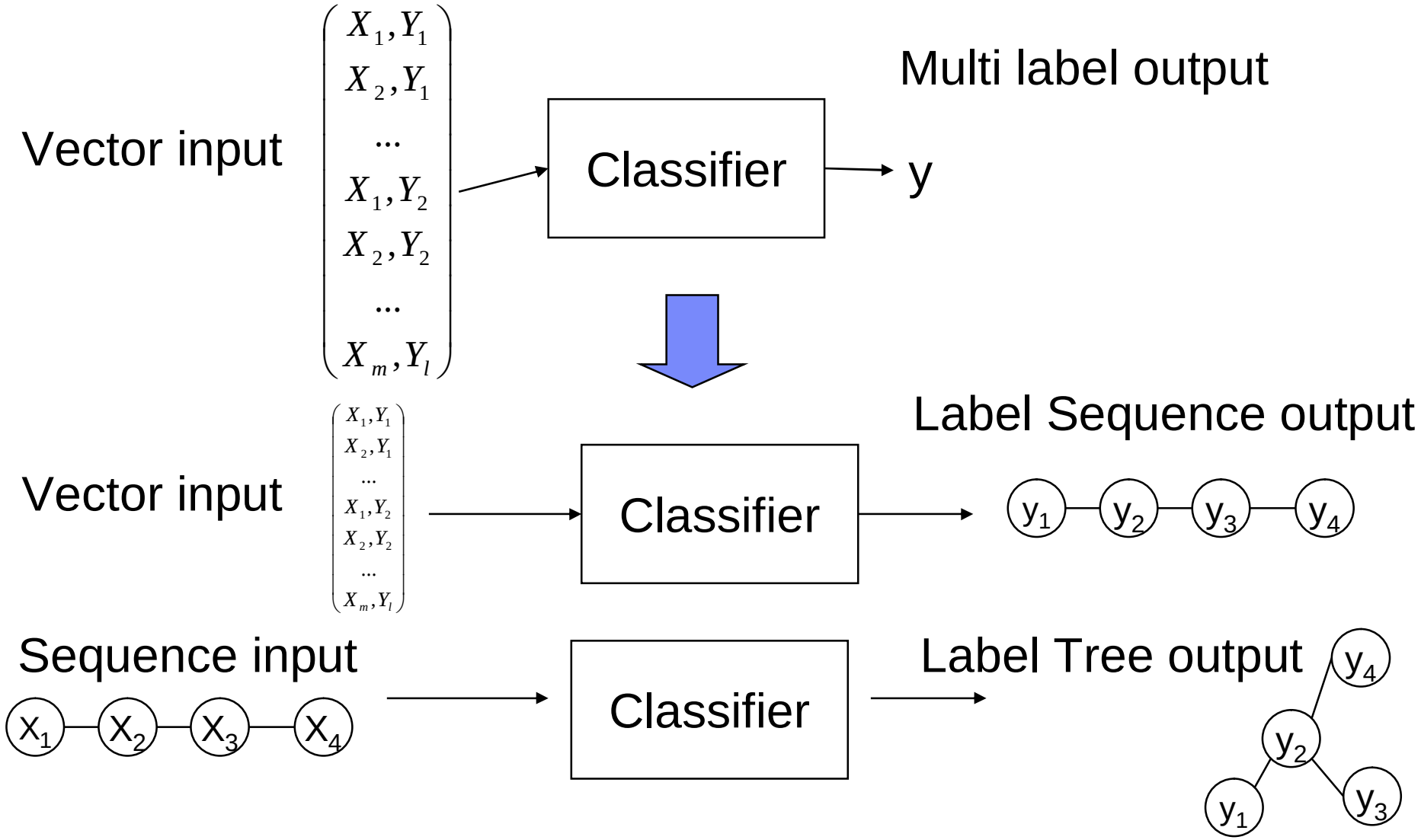
- **Viterbi Decoding**
- **Finding the most probable label sequence.**
- **Avoids re-calculation**

Parameter Estimation of Structured Output (CRFs)

$$-\sum_i \log P(Y \mid \mathbf{X}) = -\sum_i \log \frac{\exp(\langle \boldsymbol{\theta}, \varphi(\mathbf{X}, Y) \rangle)}{\sum_{\tilde{Y} \in \mathbf{Y}} \exp(\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\mathbf{X}, \tilde{Y}) \rangle)}$$

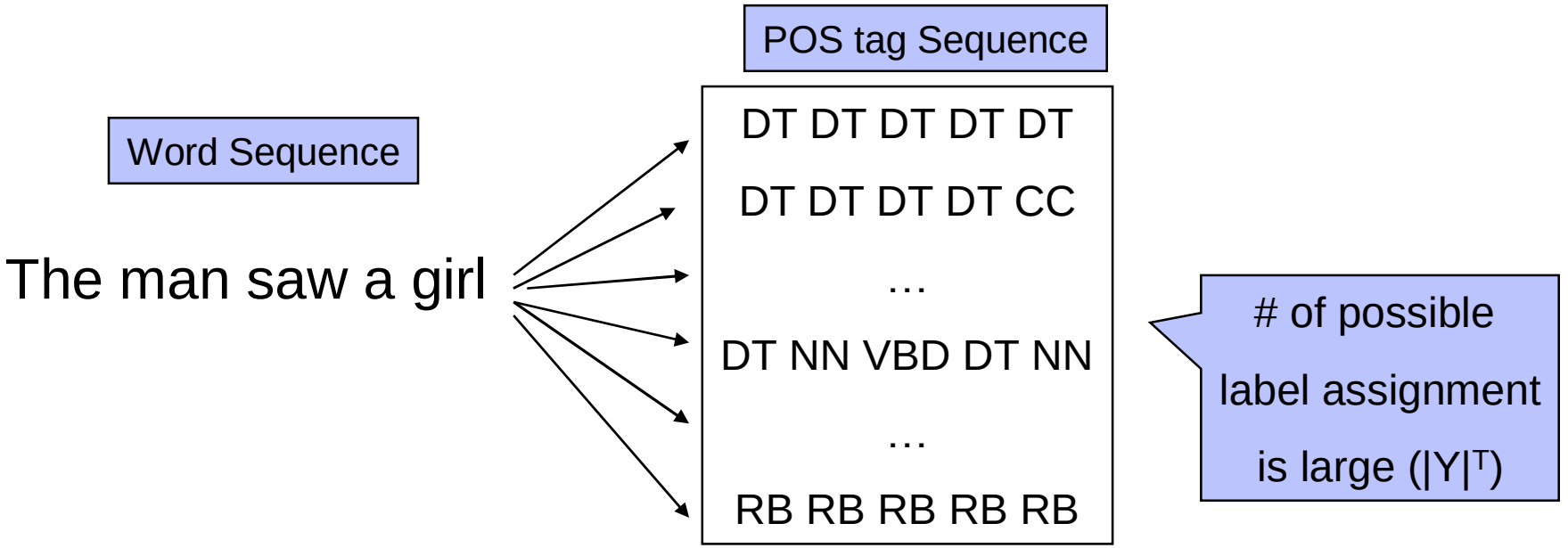
$$= -\sum_i \left(\langle \boldsymbol{\theta}, \varphi(\mathbf{X}, Y) \rangle - \log \sum_{\tilde{Y} \in \mathbf{Y}} \exp(\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\mathbf{X}, \tilde{Y}) \rangle) \right)$$

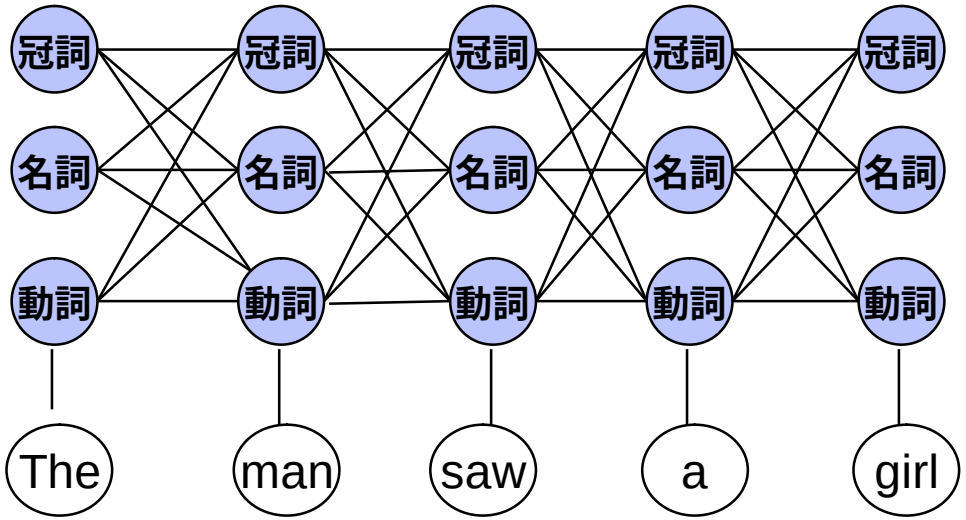
Structured Output Learning (e.g. POS tagging, NER, Parsing)



Structured Output Learning

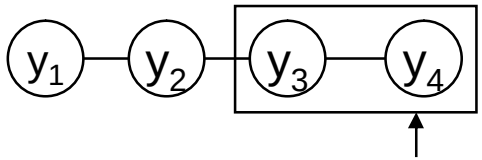
- **In general, structured output problem is one of multi label problems.**
- **In practice, all the labels are not enumerable and sparse in training data.**




冠詞
名詞
動詞
The
man
saw
a
girl

条件付確率場(CRF)：
識別モデルによる構造ラベル付与学習

- **Dynamic Programming (Belief Propagation) algorithm**
  - solves graphical inference problems via a series of local message-passing operations.
  - a label depends only its neighborhood labels (Markov assumption)
  - reuses previous calculation to avoid re-calculation



When we decide y_4, we only see the decision of y_3.

- **Transition features**
  - represent structural label consistency

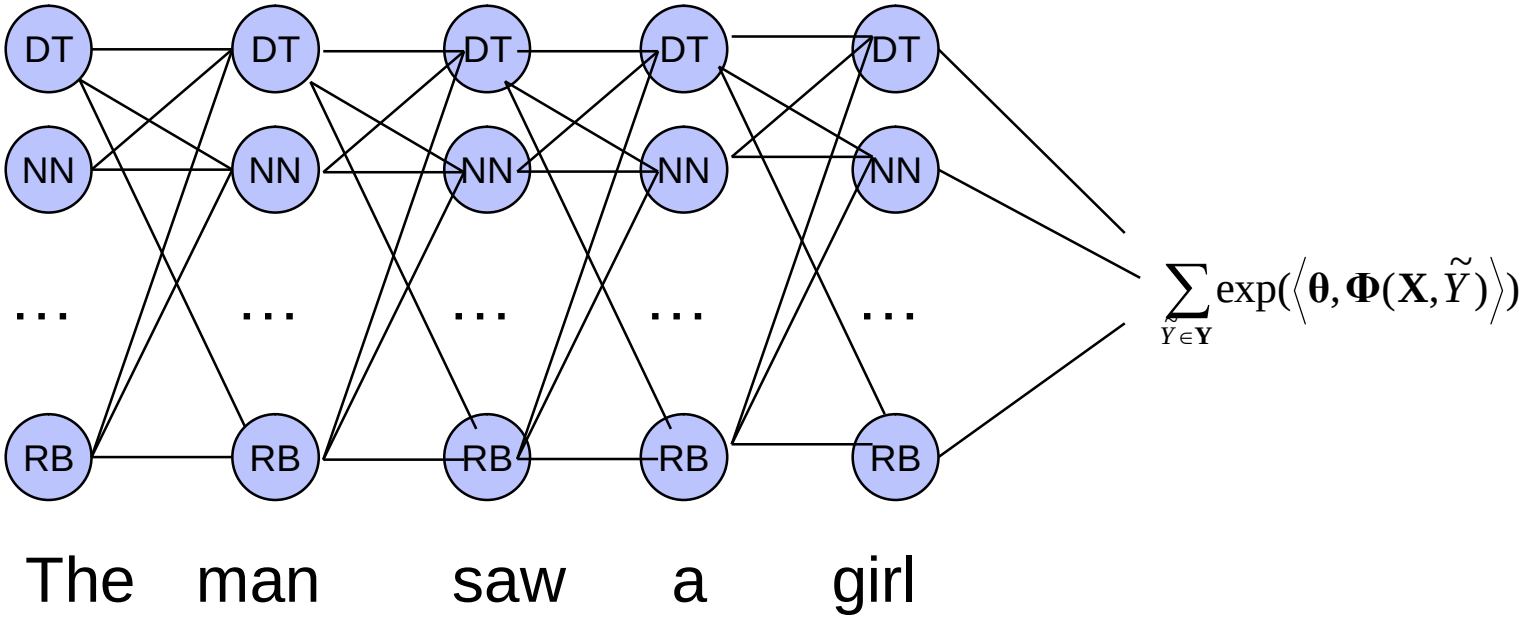
Machine Learning on Structured data

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

Dynamic Programming

$$P(Y \mid \mathbf{X}) = \frac{\exp(\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{X}, Y) \rangle)}{\sum_{\tilde{Y} \in \mathbf{Y}} \exp(\langle \boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{X}, \tilde{Y}) \rangle)}$$

- **reuses previous calculation to avoid re-calculation**
- **計算したい値**
  - **存在しうる全てのLabel Pathのスコアの和**
  - **あるポジションtにある素性が出てくるLabel Pathのスコアの和**



分類問題で使われる生成モデル (例：Naïve Bayes)

- **ベイズ分類器**

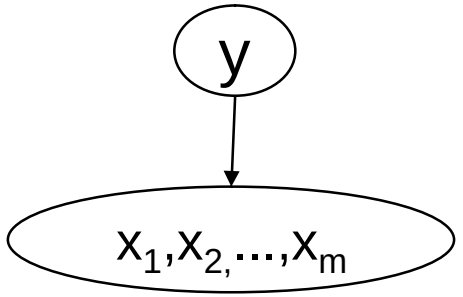
ベイズの法則

$$\text{predict} \quad y = \underset{y}{\operatorname{argmax}} \mathrm{P}(y \mid \mathbf{x}) = \underset{y}{\operatorname{argmax}} \mathrm{P}(\mathbf{x} \mid y)\,\mathrm{P}(y)$$

(**x**は観測変数ベクトル、yはラベル変数)

  - **結合分布**

$$\mathrm{P}(\mathbf{x} \mid y) = \mathrm{P}(x_1, x_2, \ldots, x_m \mid y)$$

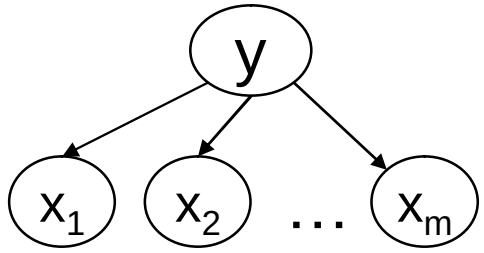


  - **ナイーブ分布**

(結合条件付分布を緩め、観測変数間の独立性を仮定)

$$\mathrm{P}(\mathbf{x} \mid y) = \prod_{j}^{|\mathbf{x}|} \mathrm{P}(x_j \mid y)$$

(|x|はベクトルのサイズ)



テーブルπを用いた最適ラベル列の求め方 (Viterbi 復号法)

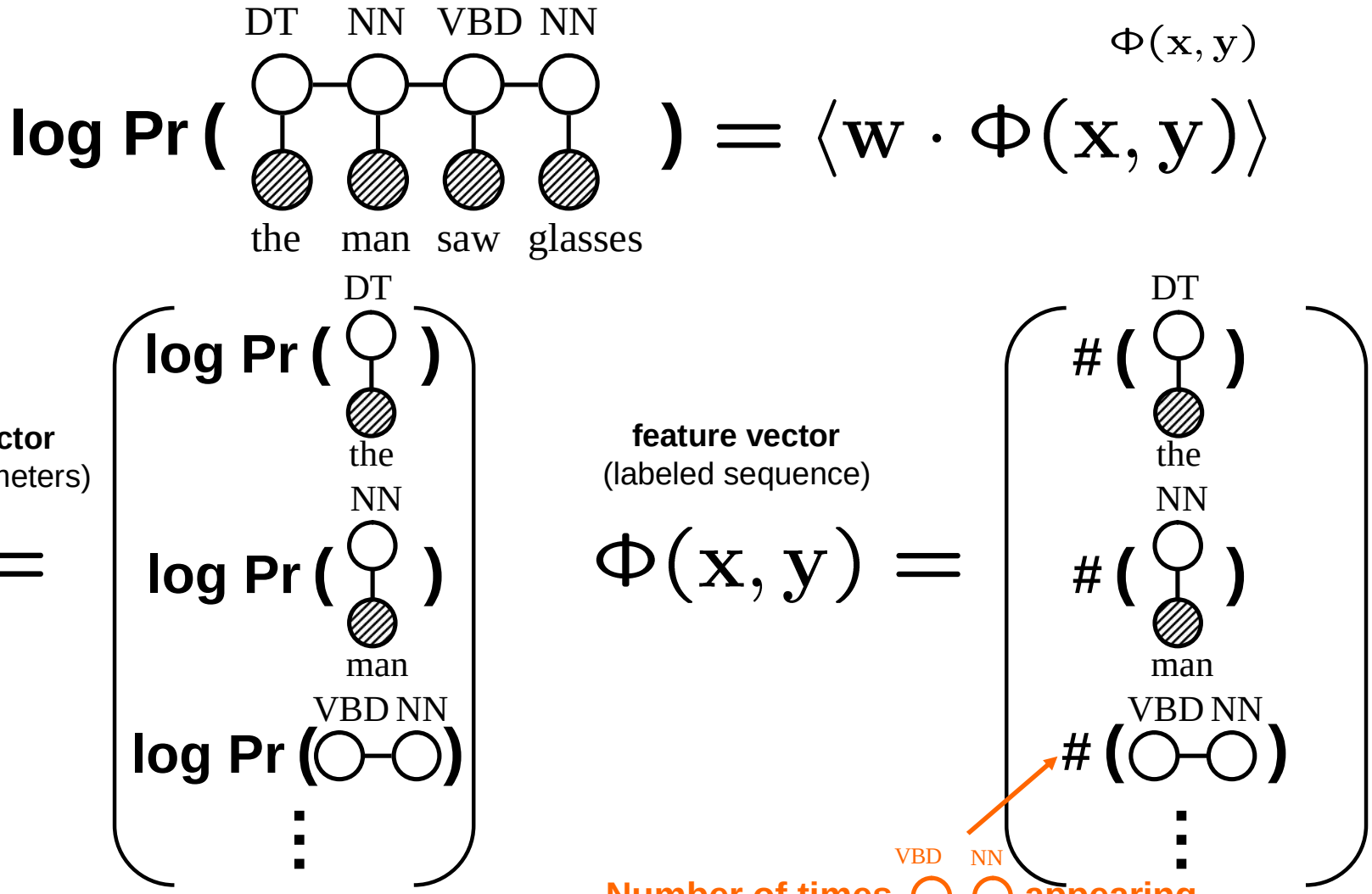
- 位置tまでのラベル列の確率が最大になる位置t-1のラベルを記憶するテーブルπ も同時に計算
- 末端(T)において確率($\sigma_T(y)$)が最大になるyの$\pi_T(y)$からバックトラックすることで、確率が最大になるラベル列を得ることができる。

$$\pi_t(y) = \underset{\tilde{y} \in \Sigma_y}{\operatorname{argmax}} \delta_{t-1}(\tilde{y}) P(x_t \mid y) P(y \mid \tilde{y})$$

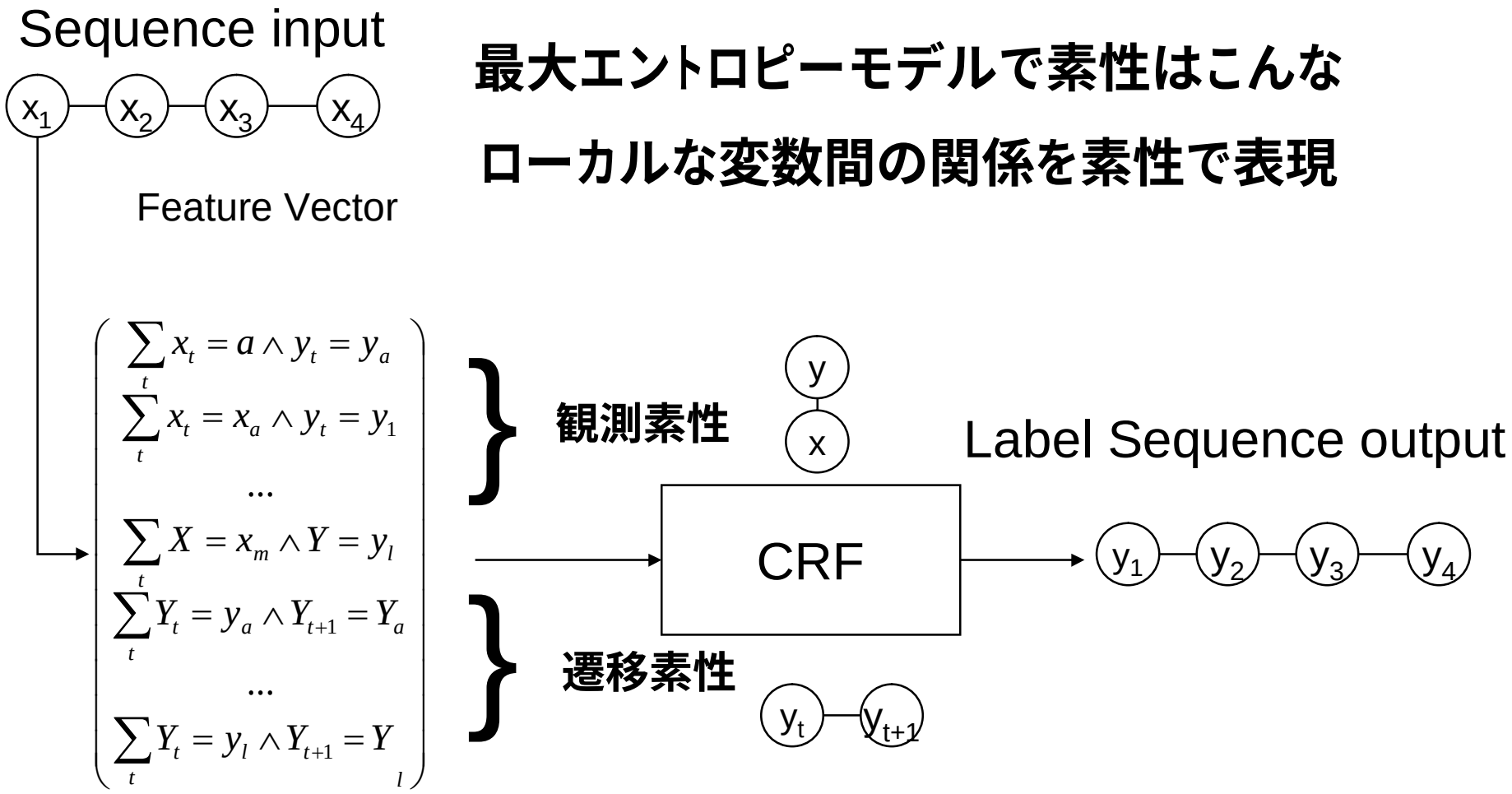
y \ x	the	man	saw	…
冠詞		π_2(冠詞)=動詞	π_3(冠詞)=動詞	…
名詞		π_2(名詞)=冠詞	π_3(名詞)=動詞	…
動詞		π_2(動詞)=名詞	π_3(動詞)=名詞	…

Another interpretation of HMM

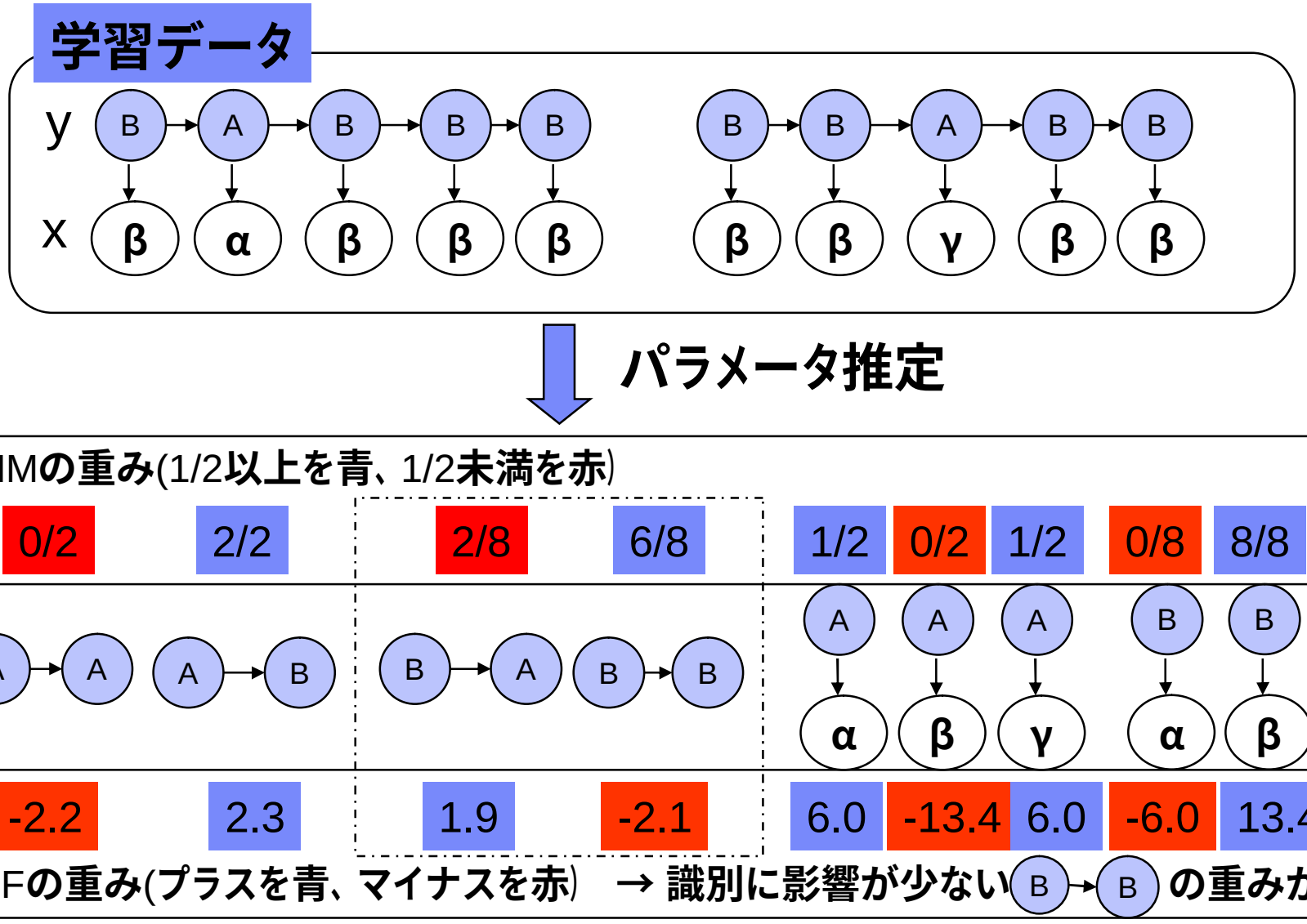
– Log likelihood can be written as
the inner product of weight vector **w** & feature vector



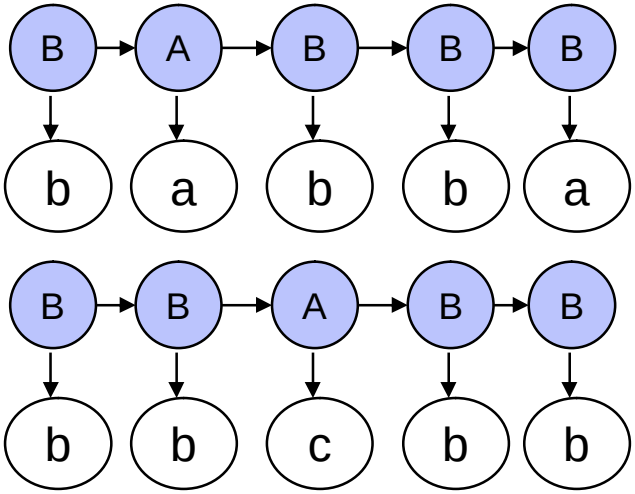
識別モデルによる構造ラベル付与学習：
条件付確率場 (Conditional Random Fields: CRF)

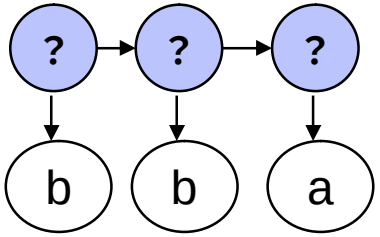


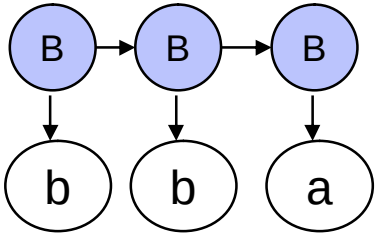
生成モデルと識別モデルの素性の重みの違い



生成モデルと識別モデル



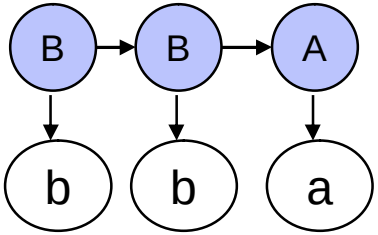




9/10*6/8*9/10*6/8*1/10

=2916/64000

=0.4556

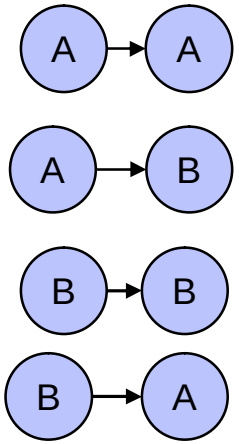


9/10*6/8*9/10*2/8*1/2

=972/12800

=0.0759

- **HMM**

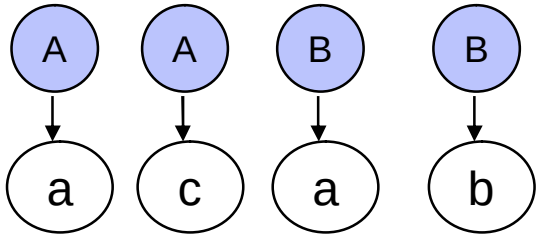


0/2

2/2

6/8

2/8



1/2 1/2 1/10 9/10

- # weights for observation
- A-a:5.96 A-b:-13.41 A-c:5.96
- B-a:-5.96 B-b:13.41 B-c:-5.96
- # weights for transition
- A-A:-2.14 A-B:2.28
- B-A:1.91 B-B:-2.05

- **Y-X**
  - A-a:6.09 A-b:-20.67
  - B-a:-6.09 B-b:20.67

- **Y-Y**
  - A-A:-3.69 A-B:10.08
  - B-A:-2.45 B-B:-3.93