Tokyo Research Laboratory

構造データのラベル付け学習モデルの設計
Design of Discriminative Models for Labeling Structured Data

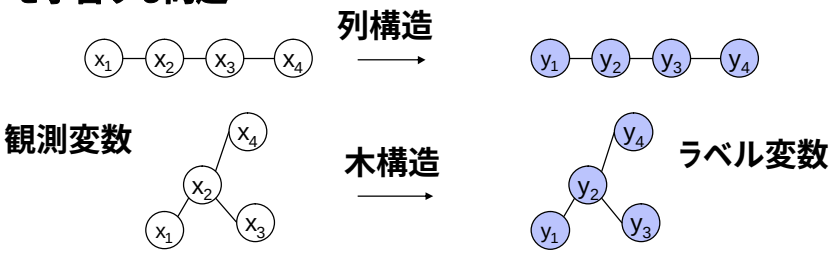
坪井祐太, 鹿島久嗣
2005-11-09

発表概要

- 構造ラベル付与学習とは？
- 構造ラベル付与学習モデル
  - 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)
- 構造ラベル付与学習の損失関数の設計
  - 全損失関数と点損失関数
  - 新しい損失関数の提案
    - λ混合損失関数とマルコフ損失関数
    - 自然言語処理での応用による評価結果

構造ラベル付与問題とは？

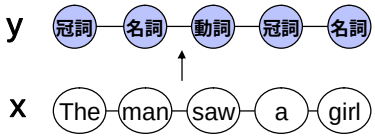
- 観測されたデータ構造xに対応するラベル構造yへの写像を学習する問題



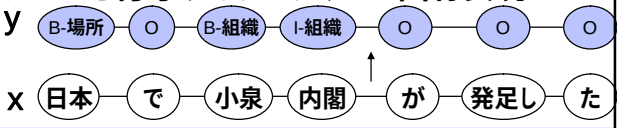
- 構造を持つ観測データの各ノードにラベルを付与する問題として捉えることが出来る応用が多く存在する。
  - 例：自然言語処理、バイオインフォマティックス

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

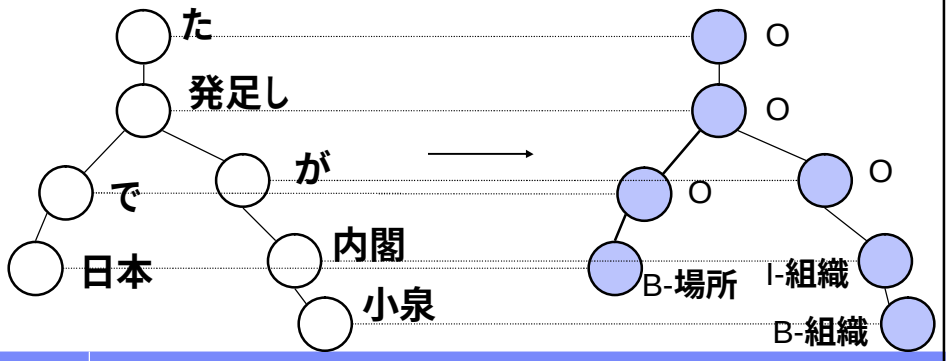


- 固有表現抽出タスク
  人名・組織名等の固有表現をテキスト中から抽出
  単語列に対して固有表現の始まり(B-XXX) と続く固有表現(I-XXX) を示すラベルを付与するタスク（O は固有表現以外）

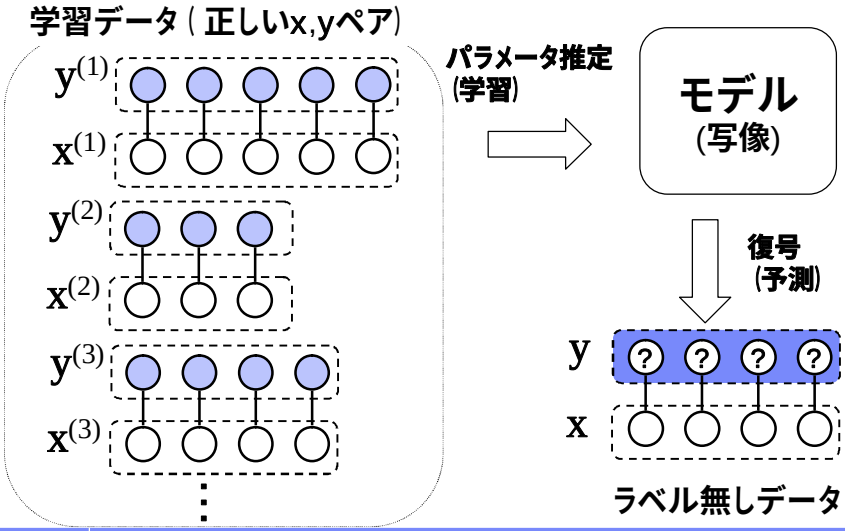


自然言語処理での構造ラベル付与問題の例（木構造）

- 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



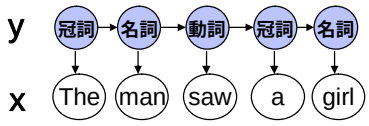
教師付き学習による構造ラベル付与問題



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
- **識別モデルに基づく手法**
  - –条件付確率場

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)

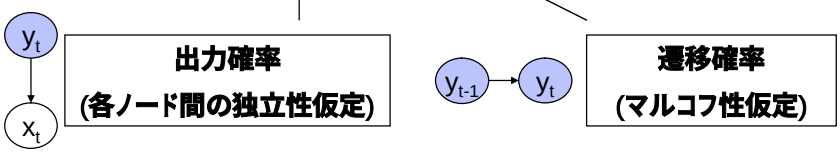


- **xとyの同時分布に基づくモデル**
- **生成確率を出力確率と遷移確率に分解してモデル化**

$$P(x,y) = P(x \mid y)\,P(y)$$

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

(Tは構造x, yのサイズ)



隠れマルコフモデル(HMM):
ラベル列の予測 (復号問題)

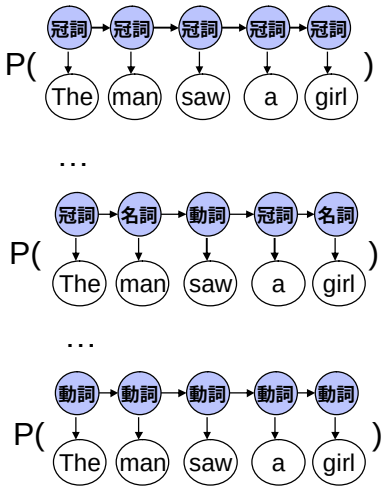
全部で$|Y|^T$個

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

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

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)



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**

$$\delta_1(y) = P(x_1 \mid y)$$

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

英語品詞タグ付けタスクでのδ

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

隠れマルコフモデル(HMM):
Viterbi 復号法による δ_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**

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_i P_\theta(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$$ （iは学習データの索引）

$$= \underset{\theta}{\operatorname{argmax}} \prod_i \prod_t^{T^{(i)}} P_\theta(x_t^{(i)} \mid y_t^{(i)}) P_\theta(y_t^{(i)} \mid y_{t-1}^{(i)})$$

- **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。**

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

- **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。**

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

→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(immediately|**副詞**), P(ly**で終わる単語**|**副詞**)
- **固有表現抽出**
  - P(New | B-**地名**), P(**最初が大文字の単語** |B-**地名**)
  - P(**小泉**|B-**組織**), P(**漢字だけからなる単語**| B-**組織**)

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**
  - **素性間の相互作用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**

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

(**Φ**はx,yの**素性**、**θ** は**素性に対する重み**)

識別モデルのパラメータ推定（学習）

- **最尤推定**

$$\hat{\theta} = \arg\max_{\theta} \prod_{i \in trainingdata} P_{\theta}(y^{(i)} \mid \mathbf{x}^{(i)})$$

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

- **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**

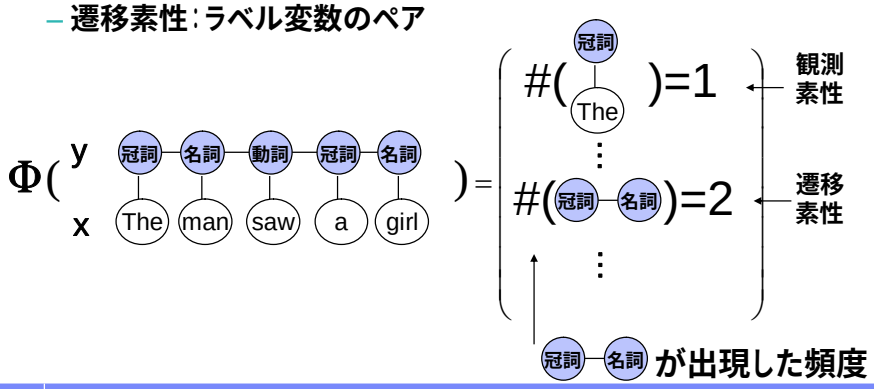
$$P(\mathbf{y} \mid \mathbf{x}) = \frac{\exp(\langle \theta, \phi(\mathbf{x}, \mathbf{y}) \rangle)}{\sum_{\tilde{\mathbf{y}}} \exp(\langle \theta, \phi(\mathbf{x}, \tilde{\mathbf{y}}) \rangle)}$$

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

(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



条件付確率場のパラメータ推定

最尤最大化

$$\arg\max_{\theta} \text{Likelihood}(\theta) = \arg\max_{\theta} \left(\prod_{i \in \text{trainingdata}} P_\theta(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}) \right)$$

負の対数尤度＝損失関数

損失最小化

$$\arg\min_{\theta} \text{Loss}(\theta) = \arg\min_{\theta} \left(-\sum_{i \in \text{trainingdata}} \log P_\theta(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}) \right)$$

構造ラベル付与学習における損失関数の設計

- **全損失 (Sequential Loss) 関数** (Lafferty, 2001)

$$L_1 = -\sum_{i \in \text{training data}} \log P(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)})$$

$$\frac{L_1}{\partial \boldsymbol{\theta}} = -\sum_{i \in \text{trainingdata}} \left(\boldsymbol{\Phi}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) - \sum_{\tilde{\mathbf{y}} \in \mathbf{Y}} P(\tilde{\mathbf{y}} \mid \mathbf{x}^{(i)}) \boldsymbol{\Phi}(\mathbf{x}^{(i)}, \tilde{\mathbf{y}}) \right)$$

ラベル構造全体を正しく正答するように学習

- **点損失(Point-wise Loss) 関数** (Kakade, 2002)

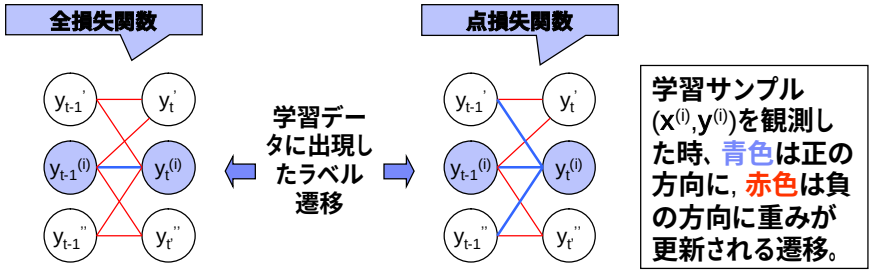
$$L_0 = -\sum_{i \in \text{training data}} \sum_{t}^{T^{(i)}} \log \sum_{\tilde{\mathbf{y}}: \tilde{y}_t = y_t^{(i)}} P(\tilde{\mathbf{y}} \mid \mathbf{x}^{(i)})$$

構造の各点をなるべく多く正答するように学習

各tでのラベル$y_t^{(i)}$の周辺尤度

$$\frac{L_0}{\partial \boldsymbol{\theta}} = -\sum_{i \in \text{trainingdata}} \left(\sum_{t}^{T^{(i)}} \sum_{\tilde{\mathbf{y}} \in \tilde{y}_t = y_t^{(i)}} P(\tilde{\mathbf{y}} \mid \mathbf{x}^{(i)}) \boldsymbol{\Phi}(\mathbf{x}^{(i)}, \tilde{\mathbf{y}}) - \sum_{\tilde{\mathbf{y}} \in \mathbf{Y}} P(\tilde{\mathbf{y}} \mid \mathbf{x}^{(i)}) \boldsymbol{\Phi}(\mathbf{x}, \tilde{\mathbf{y}}) \right)$$

損失関数による遷移素性のパラメータ推定の違い



学習サンプル $(x^{(i)}, y^{(i)})$ を観測した時、青色は正の方向に、赤色は負の方向に重みが更新される遷移。

- **全損失関数** (L_1)
  - データに出てこない遷移素性の重みは負の無限大となる
  - 学習データが少ないとき、遷移素性を過学習する可能性
- **点損失関数** (L_0)
  - データに出現しない遷移素性もy_tがデータ中に出現すれば正の重みを持つ
  - ラベル間の一貫性 (遷移素性)を重視しない

マルコフ性を持った新しい損失関数の提案

- λ混合損失関数
- k次マルコフ損失関数
- λ混合損失関数と k次マルコフ損失関数の比例関係
- 固有表現抽出タスクでの評価結果

λ混合損失関数

- 全損失と点損失 の線形和

$$L_\lambda = \lambda L_1 + (1-\lambda) L_0$$
$$= -\sum_i \left(\lambda \log P(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}) + (1-\lambda) \sum_t^{T^{(i)}} \log \sum_{\tilde{\mathbf{y}}: \tilde{y}_t = y_t^{(i)}} P(\tilde{\mathbf{y}} \mid \mathbf{x}^{(i)}) \right)$$
$$(0 \le \lambda \le 1)$$

構造全体の整合性(L_1)と局所的な精度(L_0)のバランスを図った損失関数

k次のマルコフ損失関数

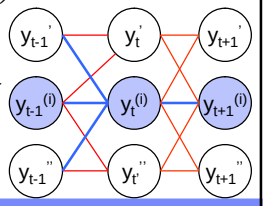
長さk+1の領域をなるべく多く正答するように学習

- 点tにおける損失が長さk+1のラベル列(t〜t+k)にのみ依存する性質を「損失のマルコフ性」と呼び、この損失をk次のマルコフ損失関数 と定義する。

$$M_k = -\sum_i \sum_{t=-k+1}^{T^{(i)}} \log \frac{\sum_{\tilde{\mathbf{y}}:\tilde{\mathbf{y}}_t^{t+k}=\mathbf{y}^{(i)}{}_t^{t+k}} \exp\left(\sum_{\tau=-k+1}^{T^{(i)}+k-1} \left\langle \theta, \phi(\mathbf{x}^{(i)}, \tilde{\mathbf{y}}_\tau^{\tau+1}) \right\rangle\right)}{\sum_{\tilde{\mathbf{y}}} \exp\left(\sum_{\tau=-k+1}^{T^{(i)}+k-1} \left\langle \theta, \phi(\mathbf{x}^{(i)}, \tilde{\mathbf{y}}_\tau^{\tau+1}) \right\rangle\right)}$$

各tでのラベル列$y^{(i)}{}_t^{t+t}$の周辺尤度

1次の マルコフ損失関数での重みの更新の例



λ混合損失関数と k次マルコフ損失関数の比例関係

- 定理

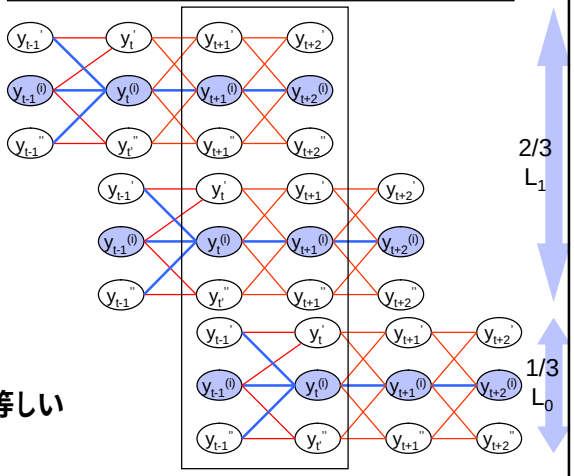
$$\lambda = \frac{k}{k+1},$$

と置くと,

$$M_k = \frac{1}{1-\lambda} L_\lambda$$

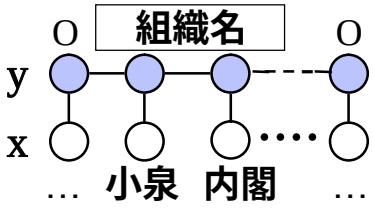
→λ=k/k+1としたとき、
L_λを最小化することは、
M_k を最小化することに等しい

2次のマルコフ損失関数による直感的な例証



提案損失関数の特徴

- **λ混合損失関数**
  - 全損失と点損失 の線形和を取ることで局所的な精度と構造全体の整合性のバランス
- **k次マルコフ損失関数**
  - λ=k/k+1の混合損失は長さk+1のセグメントの正答率を重視したマルコフ損失関数と等しく、λを決定する指標となる。
- **仮説**
  - ある長さを持った部分領域の抽出に向いてるのではないか。
    →固有表現抽出での評価実験



固有表現抽出タスクによる提案損失関数の性能検証

- **固有表現抽出タスク用データ**
  - CoNLL2002固有表現抽出タスクで提供されたスペイン語のデータ
  - 人名、地名、組織名、その他固有表現を示す計9つのラベル
- **CRFの観測素性x** (S2 feature in [Altun et al. 2003])
  - 単語および部分文字列素性
  - 前後に隣接する単語および部分文字列
- **マルコフ損失の解釈からkを1-5まで変化させた性能 (F値) を、全損失および点損失と比較。**
  - 評価手法 (1) 固有表現抽出タスクにおける標準的な評価法
  - 評価手法 (2) 少数の学習データ数での性能評価

(F値は精度と再現率の調和平均)

結果(1) CoNLL2002固有表現抽出タスクでの標準的評価法

- **学習用データで学習 (8322 文)**
- **開発用データ(1914文)で正則化パラメータをチューニング、評価用データ(1516文)で評価**
- **k=3=(固有表現の長さ＋2(境界))-1で最良の性能を得た**
  - **固有表現の平均長** = 1.74

	損失関数						
	点	k=1	k=2	k=3	k=4	k=5	全
精度	77.91	77.96	77.95	**78.10**	78.03	77.91	**78.10**
再現率	76.71	76.85	76.88	**76.96**	76.85	76.82	76.85
F1 値	77.30	77.40	77.41	**77.53**	77.43	77.36	77.47

結果 (2) 少数の学習データでの性能評価

- **学習データサイズを100-800まで変化**
- **1914文と1516文の評価 (F1値) の平均値**
- **正則化パラメータは未使用**
- **全損失と比べて提案手法は良い性能を示したが、kの値による性能は安定していない。**

データサイズ	損失関数						
	点	k=1	k=2	k=3	k=4	k=5	全
100	46.62	46.65	**47.43**	47.33	46.66	45.60	46.51
200	51.58	51.65	51.69	52.01	**52.07**	51.02	51.68
300	54.56	54.13	54.55	54.72	**54.79**	54.43	54.39
400	**55.55**	55.49	55.04	55.32	55.24	55.5	55.08
600	58.25	58.25	**58.35**	57.97	57.70	57.74	56.8
800	59.56	**59.72**	59.71	59.65	59.35	59.49	58.31

まとめ

- 構造ラベル付与学習モデルの隠れマルコフモデルと条件付確率場による手法を紹介。
- 構造を考慮した二つの損失関数（全損失と点損失関数）を紹介。
- 全損失関数と点損失関数の線形和を取ることで、構造全体の整合性と局所的な精度のバランスを図ったλ混合損失関数を設計した。
  - 長さk+1の領域のみに損失が依存するマルコフ損失関数がλ混合損失関数と等しいことを示した。
  - 提案損失関数の有効性を検証中

End of the Presentation