2005 年情報論的学習理論ワークショップ
2005 Workshop on Information-Based Induction Sciences (IBIS2005)
Tokyo, Japan, November 9 - 11, 2005.

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*
Yuta TSUBOI　　Hisashi KASHIMA

Abstract: We explain discriminative learning approaches for structured data labeling problems and propose a new loss function which is an intermediate loss function between sequential loss and pointwise loss. We show this loss function has “Markov property” to keep local consistencies and is useful for optimizing systems identifying structural segments, such as information extraction systems.

Keywords: Structured Output, Conditional Random Fields, Information Extraction

1 まえがき

自然言語処理やバイオインフォマティックスなどの分野では, 配列や木構造などの構造を持ったデータに, ラベルを付与する問題として定式化できる課題が多い. 構造ラベル付与問題の例としては, 自然言語処理では品詞付与や固有表現抽出, バイオインフォマティクスではタンパク質の二次構造予測や遺伝子発見などが挙げられる.

固有表現抽出を具体例として説明する. 固有表現抽出は文書中の人名・地名・組織名等を特定する問題である. 固有表現抽出は以下の様に単語に対して固有表現の始まり (B-XXX) と続く固有表現 (I-XXX) を示す目的ラベルを付与する問題として考えることができる[1]. 以下の例では「日本」を地名 (LOC), 「小泉内閣」を組織名 (ORG) として抽出する. ただし, O は固有表現以外を意味する.

単語列	日本	で	小泉	内閣	が	発足し	た	.
ラベル列	B-LOC	O	B-ORG	I-ORG	O	O	O	O

構造ラベル付与問題は, 構造データを入力とし対応する構造を持った目的ラベルを出力する分類問題として扱われることが多い. 目的ラベル構造を一つのクラスと考えると多クラス分類問題の一種と考えることができるが, 実際にはラベルの割り付け数は構造のサイズに対して指数的になるため容易ではない. そこで局所的な目的ラベル間だけの依存関係を仮定し動的計画法で効率的に計算する手法が一般的に使われる.

配列に対してのラベル付与問題では隠れマルコフモデル (HMM)[19, 9] が用いられ, 一定の成功を収めてきた. HMM は生成確率モデルであり, 観測変数と目的変数の同時確率 $P(\boldsymbol{x}, \boldsymbol{y})$ を使い, 次のように目的変数のラベル予測を行う.

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{y}} P(\boldsymbol{x}, \boldsymbol{y})$$

ただし, $\boldsymbol{x}$ は観測変数列, $\boldsymbol{y}$ は対応する目的変数列であるとする (図 1(a) 参照.). 同時確率を求めるためにはあらゆる観測変数列を並べあげる必要があるがその数は指数的になるため, 観測変数同士の独立性を仮定する方法が一般的である. その結果, 生成モデルでは相互に依存関係のある観測変数 (素性) を適切に扱えない. 現実の問題では独立性は成り立たないことが多いため, ラベル付与問題では $P(\boldsymbol{y}|\boldsymbol{x})$ を直接推定する識別モデルの方が, 生成モデルに比べて一般的に性能が良いといわれている.

そこで次節では構造ラベル付与問題に対する代表的な識別モデル手法である条件付確率場 (Conditional Random Field; CRF) を紹介する. 3 節では CRF の関連研究を紹介する. 第 4 節では CRF の新しい構造的な損失関数を提案し, 続く 5 節では情報抽出における実験結果を示す.

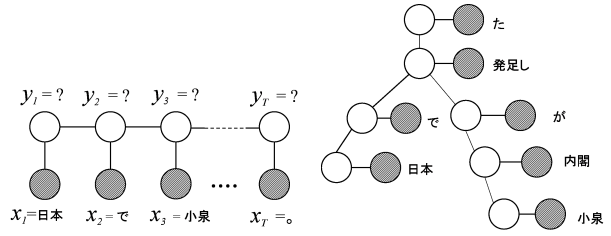
2 条件付確率場

$\boldsymbol{x} = (x_1, x_2, \ldots, x_T), x_t \in \Sigma_x$ を観測変数の集合とし, $\boldsymbol{y} = (y_1, y_2, \ldots, y_T), y_t \in \Sigma_y$ を目的変数の集合とする. ただし, Σ_x と Σ_y はそれぞれ観測変数と目的変数が取りうる値の集合とする. ラベルを付与する構造データは, 変数間の依存関係を示すグラフ構造で表現されているものと仮定する. グラフ構造の例を 1 節の固有表現抽出を使って示す. 図 1(a) は 1 節の単語列に対するラベル付与, 図 1(b) は文法的な単語間の係り受け関係を示した依存構造木と呼ばれる構造に対するラベル付与の例である.

構造ラベル付与問題では, 観測変数と目的変数の両方の値が付与された構造データが訓練データとして与えられているような, 教師付き学習問題であるとする. i 番目

*日本アイ・ビー・エム株式会社 東京基礎研究所, 神奈川県大和市下鶴間 1623–14, e-mail: {yutat,hkashima}@jp.ibm.com
IBM Tokyo Research Laboratory, Shimotsuruma 1623–14, Yamato-shi, Kanagawa, Japan

[1] この形式は IOB2 と呼ばれる固有表現のラベル方法で, その他様々な記述方法が提案されている [22].



(a) 単語列　　(b) 依存構造木

図 1: グラフ構造の例: 斜線のノードは観測変数 $\boldsymbol{x}$ を, 白抜きのノードが目的変数 $\boldsymbol{y}$ を示す.

の学習データを $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$ と書く, ただし $|\boldsymbol{x}^{(i)}| = |\boldsymbol{y}^{(i)}| = T^{(i)}$ を仮定する.

CRF は条件付確率を直接表現する多クラスのロジスティック回帰の形をとる.

$$f(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\langle\Theta, \Phi(\boldsymbol{x}, \boldsymbol{y})\rangle)}{\sum_{\tilde{\boldsymbol{y}}} \exp(\langle\Theta, \Phi(\boldsymbol{x}, \tilde{\boldsymbol{y}})\rangle)},$$

ただし, Θ はモデルのパラメータ, $\Phi(\boldsymbol{x}, \boldsymbol{y})$ は $(\boldsymbol{x}, \boldsymbol{y})$ に対する素性ベクトルで, $\Phi(\boldsymbol{x}, \boldsymbol{y})$ の要素 ϕ_i の値は i 番目の素性が $(\boldsymbol{x}, \boldsymbol{y})$ に現れた回数とする. 図 2 に示すように, 素性は通常, 連続する変数の組として定義される. 図 2(a) の形の素性は観測素性, 図 2(b) の形の素性は遷移素性と呼ばれる. ラベル付与は $\boldsymbol{x}$ に対する $\operatorname{argmax}_{\boldsymbol{y}} f(\boldsymbol{y}|\boldsymbol{x})$ による予測で行う.

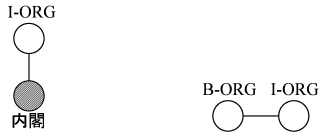
モデルの学習時には定義された損失関数を最小化する最適なパラメータを見つける. オリジナルの CRF [17] では負の対数尤度の和が損失関数として使われている.

定義 1 (全損失関数 (Sequential loss function) [4]).
全損失関数 L_1 は

$$L_1 \quad = \quad -\sum_i \log f(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)})$$

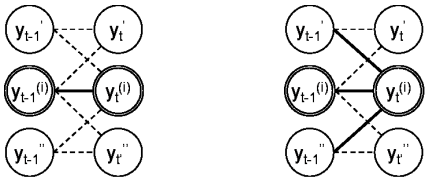
と定義する.
—
なお, 最適なパラメータは損失関数のパラメータによる偏微分を導くことで勾配法を用いて得られる.

全損失関数 L_1 は, 目的変数の集合 $\boldsymbol{y}^{(i)}$ の尤度を最大化しているため, 配列の要素全てをまとめて正しく予測するようなパラメータを学習していると解釈することができる. 一方, 問題によっては配列内のなるべく多くの目的変数を当てることが重要である場合もある. また, 学習データが少ない場合などの難しい問題に対しては, 構造全体を予測するという制約は厳しいため未知データに対する性能が悪くなる可能性がある. そこで Kakade ら [13] は各点での $y_t^{(i)}$ の周辺尤度 $\Pr(y_t = y_t^{(i)}|\boldsymbol{x}^{(i)})$ に基づいた別の損失関数 L_0 を提案した.



(a)　　(b)

図 2: CRF で使用される素性



(a) 全損失　　(b) 点損失

図 3: 遷移素性の更新の違い: 二重丸が学習サンプル中に観測された目的変数ラベル, 一重丸が観測されない目的変数ラベルを示す. 太字のエッジはこの学習サンプルによって重みが増える素性を, 点線のエッジは重みが減る素性を示す.

定義 2 (点損失関数 (Point-wise loss function) [13]).
点損失関数 L_0 は

$$L_0 \quad = \quad -\sum_i \sum_{t=1}^{T^{(i)}} \log \sum_{\tilde{\boldsymbol{y}}:\tilde{y}_t = y_t^{(i)}} f(\tilde{\boldsymbol{y}}|\boldsymbol{x}^{(i)}), \qquad (1)$$

と定義する. ただし, $\sum_{\boldsymbol{y}:y_t = y_t^{(i)}}$ は t 番目の目的変数が $y_t^{(i)} \in \Sigma_y$ であるような全ての目的変数の値に対する和を表す.
—

点損失関数 L_0 は周辺のラベルとの整合性を無視し, 可能な限り多くの点の目的変数のラベルを正しく予測するパラメータを学習していると解釈できる. 点損失関数は全損失関数と同程度の性能を持つことが実験的に示されている [1, 13].

2 つの損失関数に基づいて得られるパラメータの違いは遷移素性の重みに特徴的に現れる. 全損失 (図 3(a)) では観察されなかった遷移素性に対しては大きな負の重みが学習されるのに対し, 点損失 (図 3(b)) では遷移素性のどちらかのラベルが観測されれば遷移素性に正の重みが付与される.

3　条件付確率場の関連研究

CRF の提案以降, 様々な CRF の拡張が提案されている. 近年注目を集めている最大マージン基準による CRF の拡張として, サポートベクタマシンに基づくアルゴリズム [5, 21] やブースティングに基づくアルゴリズムが提案されている [2].

他方, CRF やその拡張に対してカーネル法を適用することで, より高次の素性の利用をする手法が提案されて

いる [5, 24, 18, 3, 21]. さらに, 木カーネル [7, 14], グラフカーネル [11, 16] など任意サイズの部分構造を素性として利用可能な畳み込みカーネル [12, 10] を利用できる手法として, 学習と予測を, 候補生成と分類の 2 段階に分けて行う手法も提案されている [8, 15].

CRF やその拡張では入力構造と出力構造の組み合わせに対して特徴空間を定義し, その上での条件付分布の推定問題あるいはランキング問題として捉えられているが, 一方, 入力構造と出力構造それぞれに対して特徴空間を定義し, これらの間の写像関係を推定する方法も提案されている. 例えば, 入力空間から出力空間の主成分への回帰問題としてとらえる手法 [25] や, 出力空間でのマージン最大化を基準とする手法 [26] などがある。

4 情報抽出に適した損失関数

2 節で紹介した 2 つの損失関数はそれぞれ適した場面があると考えられる. 本節では全損失 L_1 と点損失 L_0 の中間的な性質を持つ新しい損失関数を考え, その損失関数が局所的な目的変数の塊の整合性を重視する性質を持つことを示す. 固有表現抽出やタンパク質二次構造予測では, 予測対象の名詞句や α ヘリックス領域や β シート領域などは目的変数の塊によって表現されるため, これら構造中の局所的断片を特定する目的に向いた損失関数だと予想される.

まず, 新しい損失関数 L_λ を次のように定義する.

定義 3 (λ 混合損失関数). 定数 $\lambda (0 \leq \lambda \leq 1)$ で,

$$\begin{aligned} L_\lambda &:= \lambda L_1 + (1-\lambda) L_0 && (2)\\ &= -\sum_i \Big(\lambda \log f(\boldsymbol{y}^{(i)} | \boldsymbol{x}^{(i)}) && (3)\\ &\quad + (1-\lambda) \sum_{t=1}^{T^{(i)}} \log \sum_{\tilde{\boldsymbol{y}}: \tilde{y}_t = y_t^{(i)}} f(\tilde{\boldsymbol{y}} | \boldsymbol{x}^{(i)}) \Big) \end{aligned}$$

を λ-混合損失関数と定義する.

—

この損失関数は $\lambda = 0$ の時 L_0 と等しく, $\lambda = 1$ の時 L_1 と等しい関数になっていることに注意する.

次に, 局所的な整合性を重視し目的変数の塊を可能な限り正しく予測するようにパラメータを学習する損失関数を考える. 配列ラベル付与において, 次のような CRF を考える.

$$f(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\sum_{\tau=-k+1}^{T+k-1} \langle \Theta, \Phi(x_\tau, \boldsymbol{y}_\tau^{\tau+1}) \rangle)}{\sum_{\tilde{\boldsymbol{y}}} \exp(\sum_{\tau=-k+1}^{T+k-1} \langle \Theta, \Phi(x_\tau, \tilde{\boldsymbol{y}}_\tau^{\tau+1}) \rangle)}, \qquad (4)$$

ただし $\boldsymbol{x} = (x_{-k+1}, \ldots, x_{T+k-1})$, $\boldsymbol{y} = (y_{-k+1}, \ldots, y_{T+k})$, また $\boldsymbol{y}_\tau^{\tau+1} = (y_\tau, y_{\tau+1})$ である. $t < 1$ または $t > T^{(i)}$ の変数は特別な値 σ_0 を取るダミー変数である. ダミー変数は以下の説明のために便宜的に使用するものであり, ダミー変数抜きのモデルと式 (4) は等しい. この仮定の元で次のようなマルコフ損失関数と呼ぶ関数を定義する.

定義 4 (k 次マルコフ損失関数). 整数 $k > 0$ で,

$$\begin{aligned} M_k &:= -\sum_i \sum_{t=-k+1}^{T^{(i)}} \log \sum_{\tilde{\boldsymbol{y}}: \tilde{\boldsymbol{y}}_t^{t+k} = \boldsymbol{y}^{(i)}{}_t^{t+k}} f(\tilde{\boldsymbol{y}} | \boldsymbol{x}^{(i)}) \qquad (5)\\ &= -\sum_i \sum_{t=-k+1}^{T^{(i)}} \log \sum_{\tilde{\boldsymbol{y}}: \tilde{\boldsymbol{y}}_t^{t+k} = \boldsymbol{y}^{(i)}{}_t^{t+k}} \\ &\qquad \cdot \frac{\exp(\sum_{\tau=-k+1}^{T^{(i)}+k-1} \langle \Theta, \Phi(x_\tau^{(i)}, \tilde{\boldsymbol{y}}_\tau^{\tau+1}) \rangle)}{\sum_{\tilde{\boldsymbol{y}}} \exp(\sum_{\tau=-k+1}^{T^{(i)}+k-1} \langle \Theta, \Phi(x_\tau^{(i)}, \tilde{\boldsymbol{y}}_\tau^{\tau+1}) \rangle)}, \end{aligned}$$

を k 次のマルコフ損失関数と定義する.

—

点損失 (式 (1)) が一点のみを固定した周辺尤度であるのに対して k 次のマルコフ損失関数は $k+1$ 個の連続した目的変数を固定した周辺尤度となっている. 言い換えると, M_k は点 t から次の k 変数の条件付分布にのみに依存した各点の損失の和で, 長さ $k+1$ の塊を可能な限り多く正答することを目的とした損失関数となっている. 本稿では, この様に各点の損失が周辺の変数にのみ依存する性質を損失のマルコフ性と呼ぶ.

最後に次の定理により混合損失関数とマルコフ損失関数の同一性を示す.

定理 1. 整数 $k \geq 0$ において,

$$\lambda = \frac{k}{k+1} \qquad (6)$$

と置くと,

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

—

証明. 証明は簡単な式変形によって得られる. 詳細は付録 *1* 参照. □

—

この定理によって, どんな正の整数 $k > 0$ でも, 等式 (6) を充たす λ を選ぶことで $\frac{1}{1-\lambda} L_\lambda$ の最小化と M_k の最小化が等しいことが示される. よって, 混合損失関数を使うことで局所的な整合性を保ったラベル構造を正しく予測するパラメータを学習できる.

ここまでは k が整数であるとしたが以下 k が整数でない場合, $\lfloor k \rfloor < k < \lceil k \rceil$, を考える. 直感的には L_λ は $M_{\lfloor k \rfloor}$ と $M_{\lceil k \rceil}$ の中間となる損失関数であると思われる. 実際, 次の命題が定理 1 より導かれる.

命題 1. 実数 $k \geq 0$ で $\lambda = k/(k+1)$ とした時, 以下が成り立つ.

$$\frac{1}{1-\lambda} L_\lambda = (\lceil k \rceil - k) M_{\lfloor k \rfloor} + (k - \lfloor k \rfloor) M_{\lceil k \rceil}.$$

—

証明. 省略. □

—

$0 \leq \lceil k \rceil - k, k - \lfloor k \rfloor \leq 1$ かつ $(\lceil k \rceil - k) + (k - \lfloor k \rfloor) = 1$ であるから, L_λ は $M_{\lfloor k \rfloor}$ と $M_{\lceil k \rceil}$ の内分点となっていると言える.

別の視点として, 指数的に減衰する重みによるマルコフ損失の加重平均として考えることも可能である.

命題 2. 実数 $0 < \lambda < 1$ において以下が成り立つ.

$$\frac{1}{1-\lambda} L_\lambda = (1-\lambda) \sum_{\kappa=0}^{\infty} \lambda^\kappa M_\kappa.$$

—

証明. 省略. □

—

この加重平均では κ が小さい時に M_κ に大きな重みを与え, 逆に κ が大きくなるに従い指数的に重みが減少する. そして λ は重みの減衰速度をコントロールするパラメータとしての役割となっており, λ が小さいほど早く減衰する.

命題 2 により, 混合損失はある特定の大きさの目的変数の塊を重視しているというだけでなく, あらゆる大きさの領域をその大きさごとに決まる重みで重み付けしているという別の解釈が与えられる.

提案手法と関連する手法に準マルコフ条件付確率場 (semi-CRF) [20] がある. semi-CRF は配列データの分割を目的とした手法で, 領域を表現する目的変数に対応する観測変数の塊を作り, 全損失関数でパラメータを学習する. 一方で提案手法は目的変数の塊を予測することを目的とする. 提案手法と semi-CRF の動機は類似しているが, これらは直交した概念であるため, 組み合わせることでお互いに補完しうる関係と言える.

また, 配列ラベル付与だけでなく, 木構造に対してのラベル付与に対しても L_λ を考えることは可能である. 定理 1 は図 1(b) のような根付木構造に対しても適用できる (付録 2 を参照).

5 損失関数の比較実験

4 節で提案した損失関数の性能を固有表現抽出タスクにおいて全損失, 点損失との比較実験を行った.

5.1 実験設定

本実験では固有表現抽出を配列ラベル付与問題として扱った (図 1(a) 参照). 実験には CoNLL 2002 の共有タスクで提供されているスペイン語のデータを用いた [23]. データは学習・開発・評価セットに分かれており、それぞれ 8322 文・1914 文・ 1516 文からなる. 抽出対象は人名・地名・組織名・その他の 4 種類で, 1 節の例のように固有表現の最初とそれ以降を示すラベルと固有表現以外を示すラベルがあり, 総ラベル数 $|\Sigma_y| = 9$ である. 固有表現に含まれる平均単語数は 1.74 である. 素性は Altun ら [4] と同じ遷移素性と観測素性を用いた (図 2). 観測変数は単語それ自身以外に, “単語にピリオドが含まれているか” などの部分文字列の情報を用い, さらに前後 1 単語の同じ情報も観測変数として含めた. また, 損失関数の最小化には共役勾配法を用いた.

比較は二つの点から評価した. 一つ目は共有タスクでの標準的な評価法で, 学習セットでモデルを学習し, 開発セットでパラメータのチューニングを行い, 開発セットで得られた最良のパラメータを使用したモデルで評価セットの性能を比較した. 開発セットでチューニングしたパラメータは正規化項 [6] の分散値である. 二つ目の実験は学習データサイズを変化させた時の性能比較である.　損失関数の違いに焦点を絞るためにこの実験では正規化項を導入しなかった. 性能評価は開発・評価セットで行った.

5.2 実験結果

表 1 は共有タスクの標準的な評価法による比較結果である. 列の点, *k*=*l*, 全 はそれぞれ点損失, 混合損失 (*l* 次のマルコフ損失), 全損失の結果である. 提案手法はマルコフ損失としての解釈に基づきパラメータ k を 1 から 5 まで変化させた結果を示す. 抽出性能は精度, 再現率, F1 値で評価した. 精度は正しく予測した固有表現の割合, 再現率はテストセット中の抽出できた固有表現の割合であり, F1 値は精度と再現率の調和平均である.

結果より, 提案損失関数が他の損失関数と同等の性能を示すことがわかった. 特に $k = 3$ 次のマルコフ損失が若干よい性能を示した. これは固有表現とその両隣を正しく認識することが固有表現表現の抽出に重要になるため, 長さ $k+1 \approx 4$ の塊の予測を重視した損失関数が有効であった為と考えられる.

表 2 は学習データ量を変えた時の結果である. 混合損失はパラメータ k を 1 から 4 まで変化させた. 性能は開発・評価セットでの F1 値の平均で比較している.

全般的には点損失や混合損失の性能が全損失の性能を上回ったが、点損失と混合損失の性能差はデータサイズにより安定しなかった. データサイズ 100 と 200 では混合損失 $k = 3$ と $k = 2$ の性能がよく、それより大きなデータサイズでは点損失や混合損失 $k = 1$ がよい性能を示した. 結果よりデータサイズが小さい場合にも混合損失は有効であることが示された.

6 まとめ

本稿では構造ラベル付与問題に対して近年注目を集めている識別モデルである CRF とその拡張について紹介した. さらに, CRF の新しい目的関数として混合損失関数を提案し, それを用いた最適化が構造ラベルの部分的整合性を保つ学習と等しいことを示した.

表 1: CoNLL-2002 の固有表現抽出共有タスクの標準的な評価法での性能

	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

表 2: 学習データ量変化時の固有表現抽出性能比較

学習データ量	損失関数					
	点	k=1	k=2	k=3	k=4	全
100	45.36	46.12	**46.96**	46.94	42.72	43.96
200	47.76	47.39	47.44	**47.77**	47.30	47.16
300	**53.37**	52.91	52.68	52.92	52.86	52.40
600	**59.32**	58.68	58.25	58.11	57.34	56.00
1000	61.26	**61.91**	61.38	61.33	61.34	61.05

配列ラベル付与と木構造ラベル付与において混合損失関数と対応する部分的整合性を保つ損失関数が考えられるが, 一般的なグラフ構造に対しては明確な対応関数は得られていない. しかし, 同様に混合損失関数は大域的な整合性と局所的な精度のバランスを取る性質の損失関数となると考えられる.

参考文献

[1] Y. Altun and T. Hofmann. Large margin methods for label sequence learning. In *Proceedings of EuroSpeech*, 2003.

[2] Y. Altun, T. Hofmann, and M. Johnson. Discriminative learning for label sequences via boosting. In *Advances in Neural Information Processing Systems*, 2003.

[3] Y. Altun, T. Hofmann, and A. J. Smola. Gaussian process classification for segmenting and annotating sequences. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[4] Y. Altun, M. Johnson, and T. Hofmann. Investigating loss functions and optimization methods for discriminative learning of label sequences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2003.

[5] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov support vector machines. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

[6] S. F. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University, 1999.

[7] M. Collins and N. Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems*, 2002.

[8] M. Collins and N. Duffy. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[9] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[10] T. Gärtner. A survey of kernels for structured data. *SIGKDD Explorations*, 5(1):S268–S275, 2003.

[11] T. Gärtner, P. Flach, and S. Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Proceedings of the 16th Annual Conference on Computational Learning Theory*, 2003.

[12] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, University of California in Santa Cruz, 1999.

[13] S. Kakade, Y. W. Teh, and S. Roweis. An alternative objective function for Markovian fields. In *Proceedings of the 19th International Conference on Machine Learning*, 2002.

[14] H. Kashima and T. Koyanagi. Kernels for semi-structured date. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 291–298, 2002.

[15] H. Kashima and Y. Tsuboi. Kernel-based discriminative learning algorithms for labeling sequences, trees, and graphs. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[16] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[17] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

[18] J. Lafferty, X. Zhu, and Y. Liu. Kernel conditional random fields: Representation and clique selection. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[19] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.

[20] S. Sarawagi and W. W. Cohen. Semi-Markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems*, 2005.

[21] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems*, 2004.

[22] E. F. Tjong and K. Sang. Text chunking by system combination. In *Proceedings of Conference on Computational Natural Language Learning*, 2000.

[23] E. F. Tjong and K. Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of Conference on Computational Natural Language Learning*, pp. 155–158, 2002.

[24] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[25] J. Weston, O. Chapelle, A. Elisseeff, B. Schölkopf, and V. Vapnik. Kernel depndency estimation. In *Advances in Neural Information Processing Systems*, 2003.

[26] 賀沢秀人, 鈴木潤, 前田英作. マージン最大化に基づく写像近似法 SVMAP. 第 6 回情報論的学習理論ワークショップ予稿集, pp. 205–210, 2003.

付録 1: 配列での定理 1 の証明

証明に重要な役割を果たす次の分割を準備する.

$$
\begin{aligned}
&\sum_{\tilde{\boldsymbol{y}}:\tilde{\boldsymbol{y}}_t^{t+k}=\boldsymbol{y}^{(i)}{}_t^{t+k}} \exp\Big(\sum_{\tau=-k+1}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big) \\
&= \Bigg(\sum_{\tilde{\boldsymbol{y}}_{-k+1}^{t}:\tilde{y}_t=y_t^{(i)}} \exp\Big(\sum_{\tau=-k+1}^{t-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big)\Bigg) \\
&\quad\cdot\exp\Big(\sum_{\tau=t}^{t+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\boldsymbol{y}^{(i)}{}_\tau^{\tau+1})\rangle\Big) \\
&\quad\cdot\Bigg(\sum_{\tilde{\boldsymbol{y}}_{t+k}^{T^{(i)}+k}:\tilde{y}_{t+k}=y_{t+k}^{(i)}} \exp\Big(\sum_{\tau=t}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big)\Bigg) \\
&= F_t^{(i)}(y_t^{(i)})\cdot\exp\Big(\sum_{\tau=t}^{t+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\boldsymbol{y}^{(i)}{}_\tau^{\tau+1})\rangle\Big) \\
&\quad\cdot B_{t+k}^{(i)}(y_{t+k}^{(i)})
\end{aligned}
\tag{A.1}
$$

ただし、

$$
\begin{aligned}
F_t^{(i)}(y_t^{(i)}) &:= \sum_{\tilde{\boldsymbol{y}}_{-k+1}^{t}:\tilde{y}_t=y_t^{(i)}} \exp\Big(\sum_{\tau=-k+1}^{t-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big), \\
B_t^{(i)}(y_t^{(i)}) &:= \sum_{\tilde{\boldsymbol{y}}_t^{T^{(i)}+k}:\tilde{y}_t=y_t^{(i)}} \exp\Big(\sum_{\tau=t}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big).
\end{aligned}
$$

次に, 式 (3) と (4) より,

$$
\begin{aligned}
\frac{1}{1-\lambda}L_\lambda &= -\sum_i\Big(\frac{\lambda}{1-\lambda}\sum_{\tau=-k+1}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\boldsymbol{y}^{(i)}{}_\tau^{\tau+1})\rangle \\
&+\sum_{t=-k+1}^{T^{(i)}}\log\sum_{\tilde{\boldsymbol{y}}:\tilde{y}_t=y_t^{(i)}}\exp\Big(\sum_{\tau=-k+1}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big) \\
&-\Big(\frac{\lambda}{1-\lambda}+T^{(i)}\Big)\log\sum_{\tilde{\boldsymbol{y}}}\exp\Big(\sum_{\tau=-k+1}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big)\Big)
\end{aligned}
$$

が得られる. ここで $k=\frac{\lambda}{1-\lambda}$ と置き換え第 2 項を分割し,

$$
\begin{aligned}
\frac{1}{1-\lambda}L_\lambda = -\sum_i\Bigg(&k\sum_{\tau=-k+1}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\boldsymbol{y}^{(i)}{}_\tau^{\tau+1})\rangle \\
&+\sum_{t=-k+1}^{T^{(i)}}\log F_t^{(i)}(y_t^{(i)})+\sum_{t=-k+1}^{T^{(i)}}\log B_t^{(i)}(y_t^{(i)}) \\
&-(k+T^{(i)})\log\sum_{\tilde{\boldsymbol{y}}}\exp\Big(\sum_{\tau=-k+1}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big)\Bigg)
\end{aligned}
$$

を得る. さらに, 最初の 3 項をまとめ,

$$
\begin{aligned}
&\frac{1}{1-\lambda}L_\lambda \\
&= -\sum_i\Bigg(\sum_{t=-k+1}^{T^{(i)}}\log\Big(F_t^{(i)}(y_t^{(i)}) \\
&\qquad\cdot\exp\Big(\sum_{\tau=t}^{t+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\boldsymbol{y}^{(i)}{}_\tau^{\tau+1})\rangle\Big)\cdot B_{t+k}^{(i)}(y_{t+k}^{(i)})\Big) \\
&\quad-(k+T^{(i)})\log\sum_{\tilde{\boldsymbol{y}}}\exp\Big(\sum_{\tau=-k+1}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big)\Bigg).
\end{aligned}
$$

最後に, (A.1) を適用し終了.

$$
\begin{aligned}
&\frac{1}{1-\lambda}L_\lambda \\
&= -\sum_i\Bigg(\sum_{t=-k+1}^{T^{(i)}}\log\Big(\\
&\qquad\sum_{\tilde{\boldsymbol{y}}:\tilde{\boldsymbol{y}}_t^{t+k}=\boldsymbol{y}^{(i)}{}_t^{t+k}}\exp\Big(\sum_{\tau=-k+1}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big)\Big) \\
&\quad-(k+T^{(i)})\log\sum_{\tilde{\boldsymbol{y}}}\exp\Big(\sum_{\tau=-k+1}^{T^{(i)}+k-1}\langle\Theta,\Phi(x_\tau^{(i)},\tilde{\boldsymbol{y}}_\tau^{\tau+1})\rangle\Big)\Big)\Bigg) \\
&= M_k
\end{aligned}
$$

□

付録 2: 根付木での定理 1 の略証

配列ラベル付与と同様, 特別な定数ラベル σ_0 を常に取る適当なダミー変数を木構造の根と葉に想定する. 根から葉までの経路の長さが全て $k+1$ である部分グラフを M_k における目的変数の塊とするならば, 式 (A.1) と同様に以下の分割により L_λ と M_k の同一性が言える.

$$
\begin{aligned}
&\sum_{\boldsymbol{y}:y_t=y_t^{(i)}}\exp\Big(\sum_{\tau\in V}\Theta\Phi(x_\tau^{(i)},\boldsymbol{y}_\tau^{\pi(\tau)})\Big) \\
&= \Bigg(\sum_{\boldsymbol{y}_{IN(t)}\cup\{y_t\}:y_t=y_t^{(i)}}\exp\Big(\sum_{\tau\in IN(t)}^{t-1}\Theta\Phi(x_\tau^{(i)},\boldsymbol{y}_\tau^{\pi(\tau)})\Big)\Bigg) \\
&\quad\cdot\Bigg(\sum_{\boldsymbol{y}_{OUT(t)}\cup\{y_t\}:y_t=y_t^{(i)}}\exp\Big(\sum_{\tau\in OUT(t)\cup\{t\}}\Theta\Phi(x_\tau^{(i)},\boldsymbol{y}_\tau^{\pi(\tau)})\Big)\Bigg)
\end{aligned}
\tag{A.2}
$$

ただし, V は木の各ノードに対応する番号の集合であり, $\pi(t)$ は t 番ノードの親ノード番号を示す. また, $OUT(t)$ は t 番ノードの “外側”に属するノード番号の集合、$IN(t)$ は t 番ノード以下の “内側” に属するノード番号の集合である.