データの分布特性を利用した半教師あり系列構造学習: 言語解析への適用

鈴木 潤　　藤野昭典　　磯崎 秀樹
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
{jun,a.fujino,isozaki}@cslab.kecl.ntt.co.jp

概要

本稿では，生成/識別ハイブリッドアプローチによる半教師あり系列構造学習法を提案する．提案法では，識別モデルと生成モデルを log-linear 形式で識別的に統合する形で目的関数を定義する．また，ラベルなしデータは生成モデル内で，全ての出力に対する識別関数 (discriminant function) の値の総和を増加させる目的でのみ利用する．英語固有表現抽出 (CoNLL-2003) と英語チャンキング (CoNLL-2000) データを用いた実験において，生成/識別ハイブリッドアプローチによる半教師あり系列構造学習法は，条件付き確率場 (CRF) といった教師あり学習の設定での性能を大幅に上回る性能を示した．

1　はじめに

チャンキング，固有表現抽出といった自然言語解析タスクでは，近年，条件付き確率場 [1] に代表される大域的最適化に基づく構造学習法 (以下，略して構造学習法) が用いられるようになり，従来の局所的な最適解を組み合わせる方法より良い性能を示している．これらのタスクは，(単語) 系列にラベルを付与する問題とみなすことができるため，系列ラベリング問題と総称される．また，出力間に相互依存性があるため，学習時の特徴空間が非常に大きなものになるという特徴を持っている．つまり，一般論として，性能のよいモデルを学習するためには，この特徴空間を十分に覆う程の膨大なデータ量が必要となる．しかし，従来の構造学習法は，主に教師あり学習の設定で構築されており，限られた量のラベルありデータのみしか利用することができなかった．また，相互依存性により，系列ラベリング問題のラベルありデータを作成するコストが非常に高いということもあげられる．

これらの状況から，構造学習法では，獲得が比較的簡単なラベルなしデータを取り込んだ半教師あり学習の枠組が求められており，事実，近年いくつかの研究成果が報告されるようになった [2, 3, 4, 5, 6].

生成アプローチでは，EM アルゴリズム [7] をもちいることで，自然かつ簡単にラベルなしデータを取り入れることができる．しかし，本稿で対象とするチャンキングや固有表現抽出といったよく扱われる系列ラベリング問題では，一般的に識別モデルを学習するのに十分な量のラベルありデータが存在するため，生成アプローチでは，ラベルなしデータを取り込んでも，教師あり学習の設定での識別モデルに性能が及ばない場合がほとんどである．

そこで，識別アプローチで半教師あり構造学習法を考えたいが，識別アプローチではラベルなしデータをどのように取り込むかという方法は，生成モデルのように自明ではない．例えば，古典的な条件付き確率により識別モデルを設計することを考えると，ラベルなしデータの項は式から消去されてしまう．そこで，ラベルなしデータを識別モデルに取り込むためには新たな仮定が必要になる．その方法の一つとして，サンプル間の類似度に基づく方法が提案されている [2, 4, 6]．しかし，これはテスト時に全てのサンプルとの類似度を計算する必要があり，本稿の対象とするチャンキングや固有表現抽出といったタスクではデータ量が膨大であるため，実用的とはいえない．また，識別アプローチでラベルなしデータを取り込む別の方法として，エントロピー正則化項 (entropy regularizer)[8] を用いる方法があり，構造学習法としては，最小エントロピー正則化に基づく条件付き確率場が提案されている [5]．これは，(条件付き) エントロピー基準に下で，ラベルなしデータをなるべく識別するように学習が行われる．

これらに対して，本稿では，生成/識別ハイブリッドアプローチによる構造学習法を提案する．生成/識別ハイブリッドアプローチは文献 [9] によりはじめて提案された枠組であり，文献 [10] によって半教師あり学習法への適用法が提案された．本稿では，この方法を構造学習法へ拡張をおこなった．また同時に，目的関数を再定義し，従来の枠組では考慮されてこなかった，教師ありデータから識別的に学習されたモデルを導入可能とした．これらの改良により，提案ハイブリッドモデルでは，条件付き確率場といった教師あり学習の設定での構造学習法を上回る性能を得ることが可能となった．

実験では，英語固有表現抽出 (CoNLL-2003) と英語チャンキング (CoNLL-2000) データを用い，ラベルありデータが比較的大量にあり，教師あり学習で良好な性能が得られる状況でも，提案法による半教師あり学習を用いることでさらに大幅に性能を向上させることができることを示す．

2　生成/識別ハイブリッドアプローチによる半教師あり系列構造学習

ここでは，提案法である生成/識別ハイブリッドアプローチによる半教師あり系列構造学習法について，その定式化の方法とパラメタ推定法について述べる．

ここでは，ラベルありデータを $\mathcal{D}_l = \{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$ とし，ラベルなしデータを $\mathcal{D}_u = \{\boldsymbol{x}^m\}_{m=1}^M$ と表す．

2.1　識別アプローチによるモデルの統合

入力系列 $\boldsymbol{x}$ と出力系列 $\boldsymbol{y}$ の同時確率分布 $p(\boldsymbol{x}, \boldsymbol{y})$ とする．また，$\boldsymbol{\lambda}$ を識別モデルのパラメタベクトル，$\boldsymbol{\theta}$ を生成モデルのパラメタベクトルとする．

本稿で提案するハイブリッドモデルを，識別モデルから推定される $p^D(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda})$ と，生成モデルから推定される

$p^G(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$ を用いて識別的に統合する形で定式化する．つまり，提案ハイブリッドモデルでの入力系列 $\boldsymbol{x}$ が与えられたときの出力系列 $\boldsymbol{y}$ の事後確率を，log-linear モデルの形式を用いて以下のように定義する．

$$\begin{aligned} & R(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma}) \\ &= \frac{\prod_i p_i^D(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\lambda}_i)^{\gamma_i}\prod_j p_j^G(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}_j)^{\gamma_j}}{\sum_{\boldsymbol{y}}\prod_i p_i^D(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\lambda}_i)^{\gamma_i}\prod_j p_j^G(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}_j)^{\gamma_j}} \\ &= \frac{\prod_i p_i^D(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\lambda}_i)^{\gamma_i}\prod_j p_j^G(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}_j)^{\gamma_j}p_j^G(\boldsymbol{y};\boldsymbol{\theta}_j)^{\gamma_j}}{\sum_{\boldsymbol{y}}\prod_i p_i^D(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\lambda}_i)^{\gamma_i}\prod_j p_j^G(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}_j)^{\gamma_j}p_j^G(\boldsymbol{y};\boldsymbol{\theta}_j)^{\gamma_j}} \end{aligned} \quad (1)$$

ここで，$\boldsymbol{\Gamma}$=$\{\{\gamma_i\}_{i=1}^I,\{\gamma_j\}_{j=1}^J\}(\gamma_i,\gamma_j\in[0,1])$, $\boldsymbol{\Lambda}$=$\{\boldsymbol{\lambda}_i\}_{i=1}^I$, $\boldsymbol{\Theta}=\{\boldsymbol{\theta}_j\}_{j=1}^J$ とする．式 (1) の 3 行目は，2 行目から $p^D(\boldsymbol{x},\boldsymbol{y})=p^D(\boldsymbol{y}|\boldsymbol{x})p^D(\boldsymbol{x})$ と $p^G(\boldsymbol{x},\boldsymbol{y})=p^G(\boldsymbol{x}|\boldsymbol{y})p^G(\boldsymbol{y})$ を用い，かつ，全ての i に対する $p_i^D(\boldsymbol{x};\boldsymbol{\lambda}_i)^{\gamma_i}$ が，分母分子で相殺されて式から削除されることで得られる．ゆえに，提案ハイブリッドモデルは，識別モデルと生成モデルの log-linear モデルとして定式化される．

ここで，識別モデル (CRF) を log-linear 形式で統合する枠組は LOP-CRF とよばれ，文献 [11] で提案されている．提案法で，全ての j に対して $\gamma_j=0$ の場合は，LOP-CRF と一致する．つまり，提案法は，LOP-CRF にラベルなしデータを取り込むための生成モデルを導入可能にする拡張をおこなった方法とも捉えることができる．

ここで，ラベルありデータを用いて学習された (I 個の) 識別モデル $p_i^D(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\lambda}_i)$ が存在するとする．提案ハイブリッドモデルでは．任意の $\boldsymbol{\Theta}$ に対して，以下の目的関数を最大化することで $\boldsymbol{\Gamma}$ のパラメタ推定をおこなう．

$$\mathcal{L}^{\text{SS-Hyb.}}(\boldsymbol{\Gamma}|\boldsymbol{\Theta})=\sum_n \log R(\boldsymbol{y}^n|\boldsymbol{x}^n;\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma})+\log p(\boldsymbol{\Gamma}). \quad (2)$$

ただし，$p(\boldsymbol{\Gamma})$ を $\boldsymbol{\Gamma}$ の事前確率分布とする．

任意の固定された $\boldsymbol{\Theta}$ 上で，$\mathcal{L}^{\text{SS-Hyb.}}(\boldsymbol{\Gamma}|\boldsymbol{\Theta})$ はパラメタ $\boldsymbol{\Gamma}$ に対して凸関数となるので，この最適化は大域的最適解が保証される．よって L-BFGS [12] といった勾配を用いる最適化アルゴリズムを適用して容易に解を得ることができる．

2.2 ラベルなしデータの導入

式 (1) の $R(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma})$ に基づき，入力系列 $\boldsymbol{x}$ を与えた際の出力系列 $\boldsymbol{y}$ の識別関数 (discriminant function)g を考える．式 (1) 右辺の分母は正規化項なので $\boldsymbol{y}$ の決定には寄予しないため，識別関数 g は，式 (1) 右辺の分子のみを用いて以下のように定義できる．

$$g(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma})=\prod_i p_i^D(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\lambda}_i)^{\gamma_i}\prod_j p_j^G(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}_j)^{\gamma_j}$$

ここで，任意の (未知) 入力系列 $\boldsymbol{x}$ を与えた時，全ての出力系列に対して識別関数 g が与える値を観測したと仮定する．このとき，識別関数 g が全ての出力系列に対して非常に小さい値を与えたとする．このような状況では，全ての出力系列間での識別関数 g の値の差が非常に小さく，ほぼ同じ値となっていることと等価となるため，識別の信頼性は低いと考えることができる．つまり理想的には，あらゆる入力系列 $\boldsymbol{x}$ に対して，出力系列間で識別関数が与える値の差が大きい状況が望ましい．そこで，未知入力に対する出力系列間の識別関数が与える値の差が大きくなること期待して，ラベルなしデータを用いて，全ての出力系列に対する識別関数 g の出力値の総和を最大化することを考える．全ての出力 $\boldsymbol{y}$ に対する識別関数 g の出力値の合計の最大化には以下の式を用いる．

$$\mathcal{G}(\boldsymbol{\Theta}|\boldsymbol{\Gamma})=\sum_{m=1}^M \log\sum_{\boldsymbol{y}} g(\boldsymbol{x}^m,\boldsymbol{y};\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma})+\log p(\boldsymbol{\Theta}), \quad (3)$$

ここで，$p(\boldsymbol{\Theta})$ は $\boldsymbol{\Theta}$ に対する事前確率分布を表している．

$\boldsymbol{\Gamma}$ が既知のとき，$\mathcal{G}(\boldsymbol{\Theta}|\boldsymbol{\Gamma})$ を初期値近傍での最大化する $\boldsymbol{\Theta}$ を，EM アルゴリズムのような反復計算によって推定することができる．

2.3 パラメタ推定法

提案法である生成/識別ハイブリッドアプローチによる系列構造学習法でのパラメタ推定法について述べる．パラメタ $\boldsymbol{\Gamma}$ の推定には，任意の固定した $\boldsymbol{\Theta}$ の下で式 (2) を用いて推定をおこなう．一方，パラメタ $\boldsymbol{\Theta}$ の推定には，任意の固定したパラメタ $\boldsymbol{\Gamma}$ の下で式 (3) を用いて推定をおこなう．つまり，$\boldsymbol{\Theta}$ と $\boldsymbol{\Gamma}$ のパラメタ学習には相互に依存関係がある．そこで，$\boldsymbol{\Theta}$ と $\boldsymbol{\Gamma}$ を反復的に交互に学習する方法を用いる．

2.4 効率的なパラメタ推定アルゴリズム

提案ハイブリッドモデルの式 (1) を，識別モデル p_i^D には条件付き確率場 (CRF)，生成モデル p_j^G には隠れマルコフモデル (HMM) により構成する場合について，そのパラメタ推定法を述べる．

$V_{i,s}^D$ を，系列中の位置 s に対する i 番目の CRF のポテンシャル関数の値を表すとし，$V_{j,s}^G$ を，系列中の位置 s に対する j 番目の HMM が出力する確率値を表すとする．ここで，$\boldsymbol{\lambda}$ を条件付き確率場のパラメタベクトルとし，$\boldsymbol{f}_s$ を入力系列 $\boldsymbol{x}$ が与えられた際の系列中の位置 s から得られる局所的な素性ベクトルとすると，指数関数を用いたポテンシャル関数は $\exp(\boldsymbol{\lambda}\cdot\boldsymbol{f}_s)$ と表せる．また，θ_{y_{s-1},y_s} と θ_{y_s,x_s} を，系列中の位置 $s-1$ から s の状態 y_{s-1} と y_s 間の遷移確率と s でのシンボル出力確率を表すとすると，$V_{i,s}^D=\exp(\boldsymbol{\lambda}\cdot\boldsymbol{f}_s)$ と，$V_{j,s}^G=\theta_{y_{s-1},y_s}\theta_{y_s,x_s}$ となる．

式 (2) の最適化は，γ_i と γ_j の偏微分を計算することができれば，一般的な数値最適化法を適用することができる．まず，識別モデル (ここでは CRF) のパラメタである γ_i に関する偏微分は以下のように書き表すことができる．

$$\begin{aligned} \frac{\partial\mathcal{L}^{\text{SS-Hyb.}}(\boldsymbol{\Gamma}|\boldsymbol{\Theta})}{\partial\boldsymbol{\gamma_i}} &= \sum_n \log p_i^D(\boldsymbol{y}^n|\boldsymbol{x}^n)+\sum_n \log Z_i^D(\boldsymbol{x}^n) \\ &\quad -\sum_n E_{R(\mathcal{Y}|\boldsymbol{x}^n;\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma})}\Big[\sum_s \log V_{i,s}^D\Big] \end{aligned} \quad (4)$$

右辺第一項と第二項は最適化処理中は定数でとなるため，事前に一度計算しておけばよい．

また，同様に，生成モデルのパラメタ γ_j に対する偏微分は以下のように書き表せる．

$$\begin{aligned} & \frac{\partial\mathcal{L}^{\text{SS-Hyb.}}(\boldsymbol{\Gamma}|\boldsymbol{\Theta})}{\partial\boldsymbol{\gamma_j}} \\ &= \sum_n \log p_j^G(\boldsymbol{x}^n,\boldsymbol{y}^n)-\sum_n E_{R(\mathcal{Y}|\boldsymbol{x}^n;\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma})}\Big[\sum_s \log V_{j,s}^G\Big] \end{aligned} \quad (5)$$

右辺第一項は，式 (4) の第一，二項と同様に最適化処理中は定数でとなるため，事前に一度計算しておけばよい．

ここで，式 (4) の右辺第三項と式 (5) の右辺第二項の計算法について考える．$\mathcal{N}_R(\boldsymbol{x})$ を式 (1) 右辺の分母を表す

とすると，式 (1) は以下のように書き表すことができる．

$$R(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma}) = \frac{\prod_s \prod_i \left[V^D_{i,s}\right]^{\gamma_i} \prod_j \left[V^G_{j,s}\right]^{\gamma_j}}{\mathcal{N}_R(\boldsymbol{x}) \prod_i [Z_i(\boldsymbol{x})]^{\gamma_i}}, \quad (6)$$

式 (1) から式 (6) の導出の詳細は付録に示す．この式から，提案ハイブリッドモデルでの系列中の各位置 s のコストは，識別モデルと生成モデルの各位置 s に対応する値の総乗で求められ，条件付き確率 $R(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma})$ は，そのコストの全ての位置での総乗と全体の比率で表される．式 (4) の右辺第三項と式 (5) の右辺第二項は，各モデルの出力値に対する期待値であるため式 (6) から，forward-backward アルゴリズムを用いて効率的に計算できることがわかる．つまり，$\boldsymbol{\Gamma}$ の推定は，従来，条件付き確率場で用いられていたのと全く同じ forward-backward アルゴリズムを用いて効率的に計算することができる．

また，$\boldsymbol{\Theta}$ の推定にも式 (3) の形式から，通常の隠れマルコフモデルと同様に forward-backward アルゴリズムが適用できることがわかる．唯一の違いは，隠れマルコフモデルでは，$p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$ を用いて周辺確率を計算し，パラメタ推定を行うが，提案ハイブリッドモデルでは，R を用いて周辺確率を計算するところである．

以上のことから，従来からよく用いられている forward-backward アルゴリズムのみで，提案ハイブリッドモデルのパラメタ推定は効率的に実行することができることがわかる．注意点として，式 (1) で定義しているように，提案ハイブリッドモデルでは，複数の生成/識別モデルを利用することができる枠組となっているため，一見するとモデルの数だけ forward-backward アルゴリズムを実行しなくてはいけないように感じるかもしれないが，パラメタ推定で必要な forward-backward アルゴリズムの実行回数は，これらのモデルの数とは独立であり，$\boldsymbol{\Gamma}$ と $\boldsymbol{\Theta}$ の推定での繰り返し計算で 1 サンプルにつき 1 回実行すればよい．よって，1 サンプルあたりの計算量のオーダーは通常の CRF や HMM と同じである．

パラメタ推定後，パラメタ $\boldsymbol{\Lambda}$，$\boldsymbol{\Theta}$，$\boldsymbol{\Gamma}$ は，CRF で得られるような一つのパラメタベクトルに簡単に統合することができる．ゆえに，テスト時には，一般的な CRF と同様に Viterbi アルゴリズムを用いて効率的に最尤出力を得ることができる．

3 実験: 系列ラベリングタスク

本稿の実験では，ラベルありデータが比較的大量にあり，教師あり学習で良好な性能が得られる状況でも，提案法による半教師あり学習を用いることでさらに性能を向上させることができることを示す．

ゆえに，CoNLL-2000 [13]，CoNLL-2003 [14] の shared task で使用された英語チャンキングと英語固有表現抽出のデータを用いて実験をおこなった．また，比較法としては，条件付き確率場 (CRF)，LOP-CRF[11] 用いて，提案ハイブリッドモデル (SS-Hyb.) との性能比較をおこなった．ただし，ベースラインとして，SVM の順次適用モデルを用いた汎用チャンカー yamcha[*1]の性能評価もおこなった．

素性には，固有表現・チャンキングタスク共通で，yamcha および CRF の学習には，推定する出力ラベルの位置の前後 2 単語に含まれる範囲から抽出した素性を用いた．また，LOP-CRF と SS-Hyb. に関しては，4 つの識別モデルを用い，それぞれ，(1) 前 2 単語から抽出される素性，(2) 対象位置の単語から抽出される素性，(3) 後 2 単語から抽出される素性，(4) 前後 2 単語の範囲に含まれる素性 (全素性) を用いて学習をおこなった．また，SS-Hyb. の生成モデルには HMM を用い，素性には，識別モデルで用いたものと同じ素性集合から，一つの素性タイプを一つの HMM に割り当てて用いた．これは，複数の HMM 全体として識別モデルと同じ素性集合を扱うように設定したためである．

[*1] http://chasen.org/~taku/software/yamcha/

3.1 英語固有表現抽出実験 (CoNLL-2003 データ)

CoNLL-2003 で用いられた固有表現抽出データは，学習データ 14,987 文 203,621 単語，評価データ 3,684 文 46,435 単語，開発データ 3,466 文 51,362 単語に 4 種類の固有表現 (PER,ORG,LOC,MISC) と固有表現以外の O タグの計 5 種類のセグメント情報が与えられている．また，CoNLL-2003 では，1,029,122 文，17,003,926 単語からなるラベルなしデータも提供されており，本稿では，このデータを用いて実験をおこなった．

固有表現の実験では，CoNLL-2003 用の配布データに，単語の 1 文字から 4 文字までの prefix と suffix，また単語のタイプを表す素性タイプを追加した．

3.2 英語チャンキング実験 (CoNLL-200 データ)

チャンキングデータは，学習データ 8,936 文 211,727 単語，評価データ 2,012 文 47,377 単語，11 種類のチャンク情報 (NP，VP など) とチャンク以外を表す O タグの計 12 種類のセグメント情報が与えられている．

チャンキングデータには，開発データとラベルなしデータは定義されていないため，学習データの 1/5 を開発データとし，ラベルなしデータは英語固有表現抽出と同じデータを用いた．

英語チャンキング実験の素性は，配布データに含まれる単語，品詞のみを用いた．

4 実験結果および考察

$F_{\beta=1}$ 値と文正解率を用いて性能評価をおこなった．表 1 と 2 に英語固有表現抽出実験と英語チャンキング実験の結果を示す．表中の hyper-params は，それぞれの手法で用いられるハイパーパラメタを表し，CRF では，ガウシアン事前確率分布の分散 δ^2，LOP-CRF ではディリクレ事前確率分布の ξ，SS-Hyb. では，式 (2) のディリクレ事前確率分布の ξ と，式 (3) のディリクレ事前確率分布の η を用いるとする．

これらの結果から，提案ハイブリッド法は，教師あり学習の設定での識別アプローチによる学習法の性能を大幅に上回ることができることがわかる．一つ注意点として，英語チャンキング実験において LOP-CRF が CRF よりも若干性能が悪いのは，3.2 節で述べたように，LOP-CRF の識別モデルの学習に 4/5 のデータしか用いていないためである．

ラベルなしデータによる効果は，LOP-CRF の性能と比較することでみることができる．つまり，英語固有表現抽出で F 値 2.17，文正解率 2.87 向上し，英語チャンキングで F 値 0.34，文正解率 1.89 向上した．

提案ハイブリッド法において，ラベルなしデータが性能の向上に貢献した主な理由として 2 点考えられる．一つ目は，ラベルなしデータは識別関数の全ての出力に対する総

表 1: 英語固有表現抽出実験の結果 (CoNLL-2003)

methods	(hyper-params)	$F_{\beta=1}$	(gain)	Sent	(gain)
yamcha(SVM)	(C=1)	83.09	-	76.79	-
CRF	(δ^2=1)	84.61	-	78.04	-
LOP-CRF	(ξ=2)	85.06	(+0.45)	78.56	(+0.52)
SS-Hyb.	(ξ=2,η=1.0001)	87.23	(+2.68)	81.43	(+3.39)

表 2: 英語チャンキング実験の結果 (CoNLL-2000)

methods	(hyper-params)	$F_{\beta=1}$	(gain)	Sent	(gain)
yamcha(SVM)	(C=1)	93.61	-	58.00	-
CRF	(δ^2=1)	93.73	-	59.49	-
LOP-CRF	(ξ=2)	93.72	(-0.01)	59.29	(-0.20)
SS-Hyb.	(ξ=2,η=1.0001)	94.06	(+0.33)	61.18	(+1.69)

和を最大化 (増加) させるためにのみ利用しているという点である．逆に言うと，ラベルなしデータは最終的な出力系列の識別器の最適化には直接用いられない．これは，ラベルなしデータは正解出力が不明であるため，出力系列に対する識別器の最適化には貢献できない．ゆえに，正解出力系列の情報が不要で，かつ，識別器の性能を向上に貢献する情報となる識別関数の総和の増加という補助的な関数の最大化にのみ利用している点である．二点目として，複数の生成モデルを用いることにより，扱える素性の自由度をあげた点が考えられる．つまり，複数の生成モデルを組み合わせた全体として，識別モデルと同等の素性集合を扱えるようにした点である．

5 関連研究

半教師あり構造学習法としては，最小エントロピー正則化に基づく半教師あり条件付き確率場 [5] が現在一番の対立手法としてあげられる．提案ハイブリッドモデルでは，パラメタ推定時のラベルなしデータに対する計算アルゴリズムに通常の forward-backward アルゴリズムで利用できるが，この手法では，'nested' forward-backward アルゴリズムという通常の forward-backward アルゴリズムより計算オーダーが高い計算アルゴリズムが必要となる．一般的に，半教師あり学習の設定では，ラベルなしデータの量はラベルありデータの数百倍，数千倍以上になる．また，今後，扱うラベルなしデータの量は増加する一方であることが容易に推測できる．よって，提案ハイブリッドモデルは，最小エントロピー正則化に基づく条件付き確率場と比較して，ラベルなしデータの量に対するスケーラビリティが高いという大きな利点があるといえる．また，最小エントロピー正則化に基づく条件付き確率場の目的関数には，ラベルなしデータの影響をコントロールするセンシティブなハイパーパラメタが存在する．一方，提案ハイブリッドモデルには，MAP 推定で広く用いられている事前確率分布に関するハイパーパラメタのみである．このように，提案ハイブリッド法は，最小エントロピー正則化に基づく条件付き確率場と比較して，非常に良い特性を持っているということができる．

6 まとめ

本稿では，生成/識別ハイブリッドアプローチによる半教師あり系列構造学習法を提案した．英語固有表現と英語チャンキング実験により，ラベルなしデータを提案ハイブリッド法により取り込むことで，条件付き確率場といった教師あり学習の設定で得られる結果を大幅に上回る性能が得られることを示した．

参考文献

[1] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. of ICML-2001*, pp. 282–289 (2001).

[2] Zhu, X., Ghahramani, Z. and Lafferty, J.: Semi-Supervised Learning using Gaussian Fields and Harmonic Functions, *Proc.of ICML-2003*, pp. 912–919 (2003).

[3] Li, W. and McCallum, A.: Semi-Supervised Sequence Modeling with Syntactic Topic Models, *Proc. of AAAI-2005*, pp. 813–818 (2005).

[4] Altun, Y., McAllester, D. and Belkin, M.: Maximum Margin Semi-Supervised Learning for Structured Variables, *Proc. of NIPS*2005* (2005).

[5] Jiao, F., Wang, S., Lee, C.-H., Greiner, R. and Schuurmans, D.: Semi-Supervised Conditional Random Fields for Improved Sequence Segmentation and Labeling, *Proc. of COLING/ACL-2006*, pp. 209–216 (2006).

[6] Brefeld, U. and Scheffer, T.: Semi-Supervised Learning for Structured Output Variables, *Proc. of ICML-2006* (2006).

[7] Dempster, A. P., Laird, N. M. and Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, Vol. 39, pp. 1–38 (1977).

[8] Grandvalet, Y. and Bengio, Y.: Semi-Supervised Learning by Entropy Minimization, *Proc. of NIPS*2004*, pp. 529–536 (2004).

[9] Raina, R., Shen, Y., Ng, A. Y. and McCallum, A.: Classification with Hybrid Generative/Discriminative Models, *Proc. of NIPS*2003* (2003).

[10] Fujino, A., Ueda, N. and Saito, K.: Semi-Supervised Learning for Multi-Component Data Classification, *Proc. of IJCAI-2007*, pp. 2754–2759 (2007).

[11] Smith, A., Cohn, T. and Osborne, M.: Logarithmic Opinion Pools for Conditional Random Fields, *Proc. of ACL-2005*, pp. 10–17 (2005).

[12] Liu, D. C. and Nocedal, J.: On the Limited Memory BFGS Method for Large Scale Optimization, *Math. Programming, Ser. B*, Vol. 45, No. 3, pp. 503–528 (1989).

[13] Tjong Kim Sang, E. F. and Buchholz, S.: Introduction to the CoNLL-2000 Shared Task: Chunking, *Proc. of CoNLL-2000 and LLL-2000*, pp. 127–132 (2000).

[14] Tjong Kim Sang, E. T. and Meulder, F. D.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, *Proc. of CoNLL-2003*, pp. 142–147 (2003).

付録

$V^D_{i,s} = \exp(\boldsymbol{\lambda} \cdot \boldsymbol{f}_s)$，および，$V^G_{j,s} = \theta_{y_{s-1},y_s}\theta_{y_s,x_s}$ とする．式 (6) は，式 (1) から以下の式変形により導出することができる．

$$
\begin{aligned}
& R(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{\Gamma}) \\
&= \frac{\prod_i p^D_i(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\lambda}_i)^{\gamma_i}\prod_j p^G_j(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_j)^{\gamma_j} p^G_j(\boldsymbol{y}|\boldsymbol{\theta}_j)^{\gamma_j}}{\sum_{\boldsymbol{y}}\prod_i p^D_i(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\lambda}_i)^{\gamma_i}\prod_j p^G_j(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_j)^{\gamma_j} p^G_j(\boldsymbol{y}|\boldsymbol{\theta}_j)^{\gamma_j}} \\
&= \frac{1}{\mathcal{N}_R(\boldsymbol{x})}\prod_i\Big[\frac{\prod_s V^D_{i,s}}{Z_i(\boldsymbol{x})}\Big]^{\gamma_i}\prod_j\big[V^G_{j,s}\big]^{\gamma_j} \\
&= \frac{1}{\mathcal{N}_R(\boldsymbol{x})\prod_i[Z_i(\boldsymbol{x})]^{\gamma_i}}\prod_i\Big[\prod_s V^D_{i,s}\Big]^{\gamma_i}\prod_j\Big[\prod_s V^G_{j,s}\Big]^{\gamma_j} \\
&= \frac{1}{\mathcal{N}_R(\boldsymbol{x})\prod_i[Z_i(\boldsymbol{x})]^{\gamma_i}}\prod_s\prod_i\big[V^D_{i,s}\big]^{\gamma_i}\prod_j\big[V^G_{j,s}\big]^{\gamma_j}.
\end{aligned}
$$