2007－NL－181 (4)
2007／9／25

高次元特徴空間に適した半教師あり条件付確率場の検証

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，半教師あり条件付確率場 (Semi-supervised conditional random fields) について議論をおこなう．自然言語処理の多くのタスクでは，効果的なモデル学習のために単語やその連接といった特徴を利用する必要があり，一般的に数万次元以上という高次元かつスパースな特徴空間を用いて学習をおこなう必要がある．よって，これらのタスクでは，半教師あり学習の枠組みにおいても，高次元スパース特徴空間に頑健な枠組が求められる．そこで，本稿では，文献 [1] の枠組をベースにし，高次元スパース特徴空間に対して頑健な半教師あり条件付確率場を新たに提案する．また，固有表現抽出およびチャンキングタスクを用いて半教師あり条件付確率場の性能と性質について検証をおこなった．提案法により，従来の教師あり条件付確率場 [2]，エントロピー正則化に基づく半教師あり条件付確率場 [3] と比較して大幅に良い結果が得られた．また，エントロピー正則化に基づく半教師あり条件付確率場は，理論的にも実験的にも，高次元スパース特徴空間を用いた学習では性能の向上が期待できないことを明らかにする．
キーワード: 半教師あり条件付確率場, 高次元スパース特徴空間，識別関数和最大化，エントロピー正則化

Semi-supervised Conditional Random Fields for Extremely Large and Sparse Feature Spaces

Jun Suzuki　　Akinori Fujino　　Hideki Isozaki
NTT Communication Science Laboratories, NTT Corp.
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
{jun,a.fujino,isozaki}@cslab.kecl.ntt.co.jp

Abstract

This paper proposes a novel semi-supervised conditional random field which provides good characteristics with respect to handling the large and sparse feature spaces. Experiments on two real NLP tasks with extremely large feature spaces, such as named entity recognition and syntactic chunking, show that our proposed method significantly improves the state-of-the-art performance obtained from supervised CRFs [2], and semi-supervised CRFs employing the entropy regularization approach [3]. Moreover, this paper reveals that, theoretically and experimentally, semi-supervised CRFs based on the entropy regularization approach [3] cannot work well for improving the performance of tasks with large and sparse feature spaces.
Keywords: Conditional Random Fields, Large and Sparse Feature Spaces, Maximum Marginal Discriminant Functions, Entropy Regularization

1 はじめに

自然言語処理分野のタスクは，インターネットや文書の電子化といった背景から，大量のラベルなしデータ (生テキスト) を確保するのが比較的容易であるという性質を持っている．しかし一方で，意味や文脈的な要素を特徴として取り扱う必要があり，全ての事象をカバーする高性能なモデルを学習するのは非常に難しいタスクが多い．一般論として，性能のよいモデルを学習するためには，対象とする問題空間を十分に被覆するデータ量が必要である．しかし，問題空間を被覆できる程のラベルありデータを作成するのはコスト的に困難である場合がほとんどである．従来，主に分類・回帰問題に属する自然言語処理タスクでは，このような状況でも教師あり学習により問題を解いてきたが，このような性質を持つタスクには半教師あり学習の枠組がより適していると言える．半教師あり学習は，獲得が比較的容易なラベルなしデータを (大量に) 用い，問題空間をより広く被覆することによって，ラベルありデータのみを用いた教師あり学習で得られる性能を向上させ

ることを目的としている．つまり，自然言語処理分野で取り扱われてきた様々なタスク(特に，分類や回帰に属するタスク)で性能を向上させる可能性を持った学習の枠組である．

また一方で，自然言語処理の主要なタスクとして，タギング，チャンキング，係り受け解析といった文書を文法・意味的に解析する問題がある．これら自然言語解析タスクは，出力間に依存関係が存在するという特徴をもっている．そこで近年では，条件付確率場 [2] に代表される出力間の依存関係を直接モデル化し大域的な最適化をおこなう学習法，いわゆる「構造学習」の枠組を適用してモデル学習をおこなう方法が一般的になってきた．構造学習法は，これらのタスクで，部分的な解析結果を組み上げる方法より概ね良い性能を示しており，自然言語処理分野では重要な学習の方法論の一つとして活用されている．

これらの自然言語解析タスクは，前後の文脈等を考慮して対象の学習をおこなう必要がある，或は，言語事象そのものをモデル化しようとしているというという観点で，扱う問題空間はとても広いものであることが容易に推測できる．しかも，これらの言語事象を全て網羅するような正解データを作成するのは不可能に近い．ただし，膨大な量のラベルなしデータは容易に獲得することができる，つまり，これら自然言語解析タスクは半教師あり学習に非常に適した性質をもつ問題であると考えられる．このような背景もあり，近年では，構造学習問題に対しても様々な半教師あり学習法が提案されるようになった．Co-training のようなマルチビューに基づく方法 [4]，グラフによるサンプル間類似度に基づく方法 [5]，特徴空間内のサンプルの密度に基づく方法 [3]，生成モデルに基づく方法 [1] 等が代表的な方法である．また，前者二つは，マージン最大化に基づく方法，後者二つは確率モデルに基づく方法と分類することもできる．

後者二手法は，前者二手法と比較して，大規模データへ適用する際に重要となる計算量という観点や，学習アルゴリズムや最適化法の複雑度という観点等，実タスクへ適用する際にいくつかの利点をもっている．また，後者二手法のベースとなっている条件付確率場が自然言語解析タスクで良好な性能を示しており，ベースの教師あり学習法として信頼性が高いという点も挙げられる．そこで，本稿では，後者二手法が属する確率モデルに基づく半教師あり構造学習法，具体的には，条件付確率場をベースにした半教師あり学習に焦点をあて議論をおこなう．特に，本稿では，自然言語解析タスクがもつ，高次元スパース特徴空間という性質に着目し，この条件下での半教師あり条件付確率場の性能や性質について検証する．

以下，第 2 節で対象とするタスクの性質を明らかにする．次に，第 3 節で提案法および比較手法の基礎となる条件付確率場について簡単に説明する．第 4 節で，提案法の対立手法となるエントロピー正則化に基づく条件付確率場について述べ，第 5 節で，提案する半教師あり条件付確率場を述べる．その後，第 6 節にて本稿での実験について述べる．ここでは，主に，高次元スパース特徴空間へ適用した際の性能の比較をおこなう．

表 1: 実験データ

	固有表現抽出 (Reuters corpus)	チャンキング (WSJ)
正解ラベル数 (w/ IOB タグ)	4 (9)	11 (23)
ラベルあり学習データ	14,987 文 203,621 語	8,936 文 211,727 語
ラベルなし学習データ	(Reuters Corpus) 1,029,122 文 17,003,926 語	
評価データ	3,684 文 46,435 語	2,012 文 47,377 語

2 対象タスク

本稿では，自然言語処理分野でよく用いられる固有表現抽出とチャンキングの二つのタスクを用いて半教師あり条件付確率場の性能の検証をおこなう．データには，それぞれ CoNLL-2003 と CoNLL-2000 の shared task で使用されたデータを用いる．各データのサイズを表 1 に示す．

これらのタスクでは，前後の文脈によりラベルが決定するため，前後の単語やその組合せを特徴として利用するのが一般的である．特に，半教師あり学習をおこなう際には，ラベルなしデータに含まれる異なり語も特徴としてして加わるため，非常に大きな数になる．

具体的な例として，表に示した固有表現抽出データの単語の異なり数は，ラベルあり学習データで 23,624 語，ラベルありとラベルなしデータを合わせたときのは 257,691 語である．同様に，チャンキングデータでは，ラベルありデータのみで 19,122 語，ラベルなしデータも含めると 257,392 語となる．ここで，一般的によく用いられる前後 2 単語に含まれる情報を特徴として利用することを考える．最も簡単に単語のみを特徴として用いると仮定しても，特徴空間の次元数は単語異なり数の 5 倍となるため，固有表現抽出データの例では，ラベルありデータのみを用いる場合でも十万次元を超え，半教師あり学習に至っては百万次元を軽く超える，さらに，条件付確率場を用いること場合には，出力ラベル数毎にパラメタが必要であり，単純計算で固有表現抽出タスクでさらに 9 倍，チャンキングタスクでは 23 倍の特徴数となる．実際には，さらに品詞やその他の情報，あるいは bi-gram といった情報を用いることもあり，扱う特徴数はさらに大きなものになる．

このように，本稿で対象とするような自然言語解析タスクにおいては，特に高次元特徴空間に対しても性能を十分に発揮できる学習の枠組が必要であることがわかる．

3 構造学習：条件付確率場 (CRF)

本稿では，確率モデルに基づく半教師あり構造学習法，主に条件付確率場をベースにした手法について議論をおこなう．そこで，はじめに教師あり学習での条件付確率場 [2] について簡単に述べる．

$\mathcal{X}$ および $\mathcal{Y}$ を，それぞれ全ての可能な入力，および，出力の集合とする．$\boldsymbol{x} = \{x_i\}_{i \in S} \in \mathcal{X}$ を (構造付きの) 入力，$\boldsymbol{y} = \{y_i\}_{i \in S} \in \mathcal{Y}$ を入力の各状態 (ステート) i に対応する (構造付きの) 出力とする．ただし，S を入力および出力間の依存関係を表すグラフ $\mathcal{G}(\boldsymbol{x}, \boldsymbol{y})$ 内の状態の集合とする．このとき，条件付確率場は，条件付確率 $p(\boldsymbol{y}|\boldsymbol{x})$ を $\mathcal{G}(\boldsymbol{x}, \boldsymbol{y})$ のクリーク $c \in \mathcal{C}$ 上のポテンシャル関数の対数線形の形式で定義している．つまり，$\boldsymbol{\lambda}$ をパラメタベクトル，Ψ_c をクリーク c 上のポテンシャル関数とすると，条件付確率場上での条件付確率 $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\lambda})$ は以下のように定義される：

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{x})} \prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\lambda}). \quad (1)$$

ただし，$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \prod_{c \in \mathcal{C}} \Psi_c(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\lambda})$ であり，パーティション関数を表す．

$\boldsymbol{f}_c(\boldsymbol{y}, \boldsymbol{x})$ をクリーク c から得られる (ローカルな) 特徴ベクトルとする．ポテンシャル関数の条件は，非負の値を出力する関数であり，指数関数，つまり $\Psi_c(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\lambda}) = \exp(\boldsymbol{\lambda} \cdot \boldsymbol{f}_c(\boldsymbol{y}, \boldsymbol{x}))$ が広く用いられている．

条件付確率場のパラメタ推定 (学習) には，事後確率最大化に基づくパラメタ推定法が近年最もよく用いられている．つまりラベルあり学習データを $\mathcal{D}_l = \{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$ とすると，$\log p(\boldsymbol{\lambda}|\mathcal{D}_l)$ を最大化する学習基準である．このとき，条件付確率場の事後確率最大化学習の目的関数は以下のようになる：

$$\mathcal{L}^0(\boldsymbol{\lambda}) = \sum_n \log p(\boldsymbol{y}^n|\boldsymbol{x}^n;\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}). \quad (2)$$

このときの，右辺第二項は事前分布を表し，ガウシアン事前分布が最もよく用いられている．

以下，本稿では，教師あり学習での条件付確率場を CRF と省略表記する．

4 エントロピー正則化に基づく半教師あり条件付確率場 (SSCRF-ER)

エントロピー正則化に基づく半教師あり学習は，Grandvalet ら [6] により提案された．その後，これを構造学習問題へ適用する方法として，エントロピー正則化に基づく半教師あり条件付確率場が，Jiao ら [3] によって提案された．また，Lee らによって同手法を画像処理に適用した結果も報告されている [7]．本稿では，以降，「エントロピー正則化に基づく半教師あり条件付確率場」を SSCRF-ER と省略表記する．

SSCRF-ER では，条件付確率 $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\lambda})$ の定義は式 (1) で示した教師あり学習と同じ定義で与えられる．一方，SSCRF-ER の学習時の目的関数はラベルなしデータの項を追加して以下のように定義される．

$$\mathcal{L}^1(\boldsymbol{\lambda}) = \sum_n \log p(\boldsymbol{y}^n|\boldsymbol{x}^n;\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}) + \gamma \sum_m \sum_{\boldsymbol{y}} p(\boldsymbol{y}^m|\boldsymbol{x}^m;\boldsymbol{\lambda}) \log p(\boldsymbol{y}^m|\boldsymbol{x}^m;\boldsymbol{\lambda}) \quad (3)$$

式 (3) 右辺第一項および第二項は，式 (2) の教師あり学習時の条件付確率場と同じ事後確率最大化を表している．つまり第一項はラベルありデータの対数尤度，第二項は事前分布を表す．第三項は，ラベルなしデータから計算されるエントロピーの総和を表す項である．また，γ は，第一項のラベルありデータから計算される対数尤度の項と，第三項のラベルなしデータから計算されるエントロピー項の比率を調整することを目的としたトレードオフパラメタである．

実際の学習では，目的関数の勾配を求め，最急降下法や準ニュートン法といった勾配に基づく反復計算法を用いてパラメタ推定をおこなう．目的関数の勾配や効率的な計算法については，元文献 [3] を参考にされたい．また，Mann らによって，エントロピー項およびその勾配の効率的な計算アルゴリズムも紹介されている [8]．

直観的な意味付けとしては，エントロピー正則化による半教師あり学習 (条件付確率場を含む) では，ラベルありデータの事後確率最大化しつつエントロピーという文脈でラベルなしデータを最も良く分離するように最適化をおこなうことを意味する．

5 識別関数和最大化に基づく半教師あり条件付確率場 (SSCRF-MD)

Suzuki らによって，SSCRF-ER とは別の定式化を用い，CRF をベースとした半教師あり学習法が提案されている [1]．SSCRF-ER では，エントロピー項を用いてラベルなしデータを導入しているが，この手法では，ラベルなしデータに対して生成モデルを仮定することでラベルなしデータを導入している．つまりこれは，生成モデルに基づく半教師あり学習法に分類される方法である．特に，生成モデルの周辺対数尤度を最大化することでラベルなしデータを効率的に利用する方法 [9] と同じアナロジーを用い，周辺対数尤度の代わりに識別関数和を最大化することにより，ラベルなしデータから情報を獲得すると

いう特徴をもつ手法である．

文献 [1] では，実際には生成/識別ハイブリッドアプローチの文脈で定式化されている．本稿では，生成モデルを利用してラベルなしデータの導入する点，識別関数和を最大化することでラベルなしデータを効率的に利用する点，CRF がベースである点は同じであるが，教師あり CRF の自然な拡張という形式での新たに定式化をおこなう．ここで，本稿で提案する半教師あり条件付確率場を省略して SSCRF-MD と記述する．

はじめに，$1 \leq j \leq J$ とし，任意の特徴抽出関数 $\mathcal{T}_j$ により生成された $\boldsymbol{x}$ と同じ構造をもつ入力オブジェクトを $\boldsymbol{x}_j$ を定義する：$\boldsymbol{x}_j = \mathcal{T}_j(\boldsymbol{x})$．次に，$J$ 個の生成モデル $p_j(\boldsymbol{x}_j, \boldsymbol{y})$ を仮定する．ただし，各 $p_j(\boldsymbol{x}_j, \boldsymbol{y})$ は，式 (1) で定義される CRF と同じクリーク c により分割することが可能という条件を満たすとする：$p_j(\boldsymbol{x}_j, \boldsymbol{y}) = \prod_c p_{jc}(\boldsymbol{x}_j, \boldsymbol{y})$．また，生成モデルのパラメタベクトルの集合 $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_j\}_{j=1}^J$ とし，パラメタ $\boldsymbol{\lambda}$ の次元数が I のとき，$\boldsymbol{\lambda}' = (\boldsymbol{\lambda}, \lambda_{I+1}, \ldots, \lambda_{I+J})$ と定義する．

SSCRF-MD では，式 (1) で定義される教師あり CRF とは違う新たなポテンシャル関数を以下のように定義する：

$$\Phi'_c(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) = \exp(\boldsymbol{\lambda} \cdot \boldsymbol{f}_c(\boldsymbol{x}, \boldsymbol{y})) \prod_j p_{jc}(\boldsymbol{x}_j, \boldsymbol{y}; \boldsymbol{\theta}_j)^{\lambda_{I+j}}.$$

このとき，$p_j(\boldsymbol{x}_{jc}, \boldsymbol{y}_c)$ の値域は $[0, 1]$ なので，Ψ'_c はポテンシャル関数の条件を満たしている．

このポテンシャル関数を用いた条件付確率場での条件付確率は以下のように定義される：

$$p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) = \frac{1}{Z'(\boldsymbol{x})} \prod_{c \in C} \Phi'_c(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta}). \quad (4)$$

よって，SSCRF-MD の学習時の目的関数は，教師あり CRF と同様の形式で以下のようになる：

$$\mathcal{L}^2(\boldsymbol{\lambda}'|\boldsymbol{\Theta}) = \sum_n p(\boldsymbol{y}^n|\boldsymbol{x}^n; \boldsymbol{\lambda}', \boldsymbol{\Theta}) + \log p(\boldsymbol{\lambda}'). \quad (5)$$

ただし，式 (5) は，パラメタ $\boldsymbol{\Theta}$ を固定した下で，$\boldsymbol{\lambda}'$ の推定を行うための目的関数である．

次に，文献 [1] にしたがってラベルなしデータを用いてパラメタタ $\boldsymbol{\Theta}$ を推定する目的関数を導出する．文献 [1] では，識別関数和を最大化することで $\boldsymbol{\Theta}$ を推定している．通常，CRF の識別関数は，条件付確率の分母にあたるパーティション関数は出力の決定には寄与しないため，分子にあたるポテンシャル関数の積で定義される：

$$g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) = \prod_c \Phi'_c(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta}).$$

1. Given training set: $\mathcal{D} = \{\mathcal{D}_l, \mathcal{D}_u\}$ where $\mathcal{D}_l = \{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$ and $\mathcal{D}_u = \{\boldsymbol{x}^m\}_{m=1}^M$
2. Initialize $\boldsymbol{\Theta}^{(0)} \leftarrow$ uniform distribution and $t \leftarrow 0$.
3. Estimate $\boldsymbol{\lambda}'$ by maximizing Equation (5) with a fixed $\boldsymbol{\Theta}^{(0)}$ using $\mathcal{D}_l$.
4. Perform the following until $t = T$, where T is the pre-defined maximum number of iterations.
   4.1. Estimate $\boldsymbol{\Theta}^{(t+1)}$ to maximize Equation (6) with a fixed $\boldsymbol{\lambda}'$ using $\mathcal{D}_u$.
   4.2. (Re)estimate $\boldsymbol{\lambda}'$ to maximize Equation (5) with a fixed $\boldsymbol{\Theta}^{(t+1)}$ using $\mathcal{D}_l$.
   4.3. go to 5 if a convergence condition, $\frac{|\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{\Theta}^{(t)}|}{|\boldsymbol{\Theta}^{(t)}|} < \epsilon$, is met.
   4.4. $t \leftarrow t + 1$.
5. Output $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\lambda}', \boldsymbol{\Theta}^{(t)})$.

図 1: SSCRF-MD の学習アルゴリズム

よって，$\boldsymbol{\lambda}'$ を固定した下で $\boldsymbol{\Theta}$ の推定するための目的関数は，以下のように定義される：

$$\mathcal{L}^3(\boldsymbol{\Theta}|\boldsymbol{\lambda}') = \sum_m \log \sum_{\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) + \log p(\boldsymbol{\Theta}). \quad (6)$$

実際の学習時には，文献 [1] と同様に式 (6) と式 (5) を反復して最大化する学習によりパラメタ推定をおこなう．図 1 に，SSCRF-MD の学習アルゴリズムの概略を示す．

6 実験による性能・性質の検証

本稿では，SSCRF-ER と SSCRF-MD の性能や性質の比較を，固有表現抽出やチャンキングといった実タスクを用いて検証する．また，ベースラインとして，教師あり学習での条件付確率場 (CRF) と，もっとも単純な半教師あり学習法の一つである Self-training を条件付確率場に適用した方法 (SSCRF-self) を用いる．

本稿で用いる比較手法は全て CRF をベースにした手法である．よって，用いる特徴や事前分布によるスムージングといった比較条件は全て同じになる．本稿の実験では，全ての比較手法で，事後確率最大化学習で使用する事前分布にガウシアン事前分布を用いる．また，分散に相当するハイパーパラメタは，それぞれの手法毎に事前に準備した開発データで最も良い性能が得られた値を使用した．

6.1 特徴空間の大きさの違いによる性能比較

まずはじめに，比較的小さい特徴空間を使用して性能の比較をおこなった．ここでは，前後 1 単語に含まれる特徴タイプのみを特徴に用いた．特徴タイプには，固有表現抽出では，単語，品詞，チャンクタグ，単語のタイプを表す特徴タイプ (大文字や数字を含むかというような特徴) の 4 種類を用い，チャン

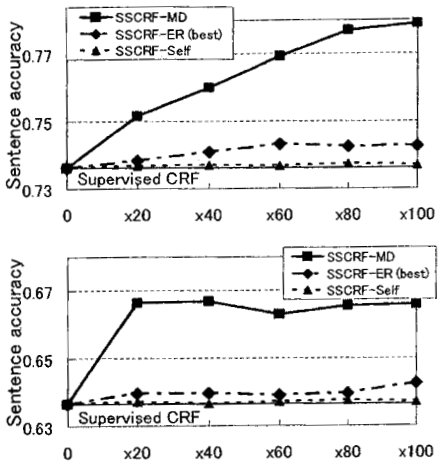

図 2: 比較的小さな特徴セットを用いた際の学習曲線 (上: 固有表現抽出, 下: チャンキング)

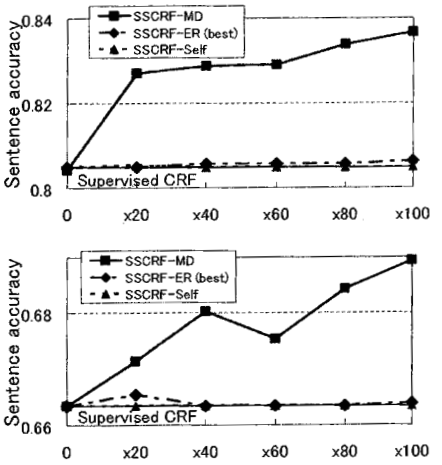

図 3: 一般的な特徴セットを用いた際の学習曲線 (上: 固有表現抽出, 下: チャンキング)

キングでは，単語，品詞の 2 種類の特徴タイプを用いた．また，ここでの実験では，問題を簡単にし議論をしやすくする目的で，固有表現抽出タスクでは正解出力ラベルを固有表現かそれ以外，チャンキングでは NP チャンクかそれ以外という 2 種類のみとして実験をおこなった．

図 2 に，前後 1 単語の範囲に含まれる特徴タイプのみを使用した際の結果を示す．ここでは，ラベルありデータを先頭から 2,000 文，ラベルなしデータをその $\{0, 20, 40, 80, 100\}$ 倍の量を用いて学習した際の性能が示されている．また，図中の縦軸は文正解率による性能評価の値を示し，横軸はラベルありデータに対するラベルなしデータ量の比率を示している[1]．ただし，これらの図に示された SSCRF-ER の結果は，「評価データ」で最も良い性能が得られたトレードオフパラメタの値を用いた際の性能である．SSCRF-ER のトレードオフパラメタは，性能に対して非常にセンシティブであり，開発データで得られた値を用いても良い結果が得られないことがしばしば起こるためである．これに関しては 6.3 節の実験で詳しく述べる．この図からは，SSCRF-MD と SSCRF-ER の双方でラベルなしデータの量にしたがって性能が向上していることが見てとれる．ただし，SSCRF-MD の方がラベルなしデータを効率良く用いて学習していることがわかる．

次に，一般的によく用いられる特徴セットである前後 2 単語の範囲に含まれる特徴タイプを用いて実験をおこなった．その結果を図 3 に示す．結果の傾向は図 2 と同じであるが，SSCRF-ER では教師あり CRF と比較して性能の向上がほぼ無い状態になっている．

表 2: 通常の設定での実験結果

固有表現抽出

methods	$F_{\beta=1}$	(gain)	Seq. acc.	(gain)
supervised CRF	84.36	-	77.66	-
SSCRF-MD	**86.69**	(+2.33)	**80.37**	(+2.71)

チャンキング

methods	$F_{\beta=1}$	(gain)	Seq. acc.	(gain)
supervised CRF	93.79	-	59.15	-
SSCRF-MD	**94.45**	(+0.66)	**61.38**	(+2.23)

最後に，CoNLL-2003 や CoNLL-2000 での通常のタスク設定で固有表現抽出，チャンキングをおこなった際の結果を表 2 に示す．$F_{\beta=1}$ は F 値を，'seq. acc.' は文正解率を表す．ここで，SSCRF-ER と SSCRF-self では，教師あり CRF の性能を向上させることができなかったので，表中には結果をのせていない．この表からわかることは，SSCRF-MD は，実際の固有表現抽出やチャンキングといった超高次元スパース特徴空間を利用するタスクであっても効果的な学習が可能であることを示している．参考情報として，この実験では，概ね十億次元程度の特徴空間を用いて学習がおこなわれている．

6.2 エントロピー正則化の直観的な解釈

まずはじめに，SSCRF-ER の元文献 [3] の実験結果を注意深く見てみると，現実のタスク評価という

[1] 通常これらのタスクは F 値により評価をおこなう．しかし，CRF は文正解率を最大化する学習の枠組なので，手法間の性能比較という観点では文正解率の方がより正確な評価指標となるため，本稿では文正解率で性能評価をしている．

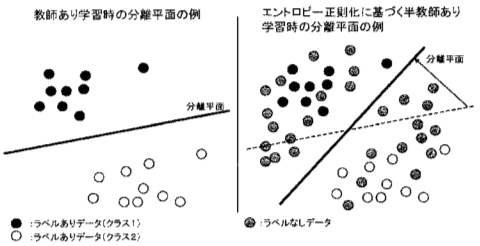

図 4: エントロピー正則化のイメージ

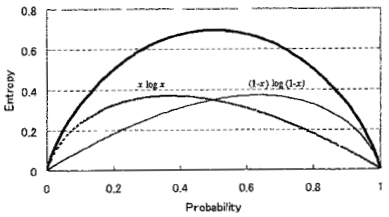

図 5: 二クラス分類時のエントロピーの値

観点では性能が向上したという実験結果は全く示されていないことがわかる.

機械学習の分野では，一般的にエントロピー正則化に基づく半教師あり学習は，高次元特徴空間を用いる場合には不向きであることが知られている．以下にその理由を述べる．エントロピー正則化に基づく半教師あり学習は,「特徴空間内でサンプルが密に分布している部分は同一の出力ラベルをもち，サンプルが疎な空間に分離平面がある」という仮定により構成されている．つまり，ラベルなしデータのエントロピーを最小化する (負のエントロピーを最大にする) という意味は，ラベルなしデータが疎な部分に積極的に分離平面を与えるようにパラメタを補正することと等価である.

ここで，直観的な説明のために，二クラス分類問題を考えてみる．図 4 にエントロピー正則化のイメージ図を示す.

次に，図 5 に，二クラス分類問題時の式 (3) のエントロピー項 (右辺第三項) がとる値を示す．この図からも明らかなように，エントロピーを最小化するように学習するという意味は，全ての (ラベルなし) サンプル $\boldsymbol{x}$ の全ての可能な出力 $\boldsymbol{y}$ での確率 $p(\boldsymbol{y}|\boldsymbol{x})$ ができるだけ確率 0 または 1 をとる方向にパラメタを変化させることに相当する．識別学習的な観点で見ると，CRF 上の条件付確率は，特徴空間内の分離平面からの距離に比例すると言える．つまり，直観的な説明としては，エントロピー正則化は，全ての (ラベルなし) サンプルからの距離がなるべく離れた空間に分離平面を決定しようとする働きをもっていると言える.

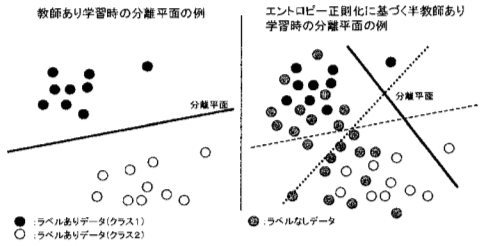

図 6: エントロピー正則化が失敗する例のイメージ

このことから，仮に，特徴空間上で異なったクラスに属するサンプルが重なるように密に分布している場合，つまり，前述したエントロピー正則化が利用してるサンプルの分布の仮定が合わない状況では，性能の向上は期待できないことは自明である．図 6 に，エントロピー正則化が失敗する端的な例を示す．また，エントロピー正則化項は，全ての (ラベルなし) データが単一のクラスに属するように，つまり，全ての x で単一の出力 y に対して確率 1 をとり，それ以外の出力に対しては 0 をとるようにする解がエントロピーを最小できるという性質があることも既知の問題として知られている.

次に，高次元特徴空間のもつ特徴を考えてみる．高次元特徴空間では大量のデータ用いてもサンプルが非常に疎に分布する．そのため，結果として特徴空間内のどの部分に分離平面を決定する場合でも，エントロピー項の値を同じように小さく (負のエントロピーを大きく) することができる．つまり，高次元かつスパースな特徴空間上では，エントロピー正則化により分離平面を決定するのに有効な情報が得られないことを意味する．その結果，最終的に全く無意味なところに分離平面を決定する可能性が非常に高い.

これらの直観的な解釈や前節の実験結果からわかるように，高次元特徴空間でエントロピー正則化に基づく半教師あり学習をおこなうと，教師あり学習時の性能からまったく性能が向上しないか，逆に非常に低下させる結果しか得られないことがわかる．よって，自然言語処理分野の多くのタスクで，エントロピー正則化に基づく半教師あり学習法は不向きであることがわかる．特に，本稿で対象としているタギング・チャンキング問題に対して，現状の定式化による SSCRF-ER では性能の向上は期待できない.

しかしこれは，SSCRF-ER の方法論自体が悪いということを意味するわけではない．逆に，特徴空間の次元数が比較的小さく，サンプルが密に配置されるようなタスクの場合は，サンプルが比較的疎な空間に分離平面があるというエントロピー正則化の仮定が当てはまりやすい状況であるため，エントロピー

正則化による半教師あり学習での性能向上が大いに期待できる．実際，エントロピー正則化による半教師あり学習法が提案された Grandvalet らの文献 [6] では，人工データや画像認識 (顔認識) タスクに用いて実験をおこなっており，特徴空間は非常に小さいものである．また，SSCRF-ER をもちいた Lee らの文献 [7] でも，一般的に特徴数が数百程度の画像処理タスクに適用しており，性能向上が示されている．

結論としては，SSCRF-ER を用いる際には，用いる (ラベルなし) データのサンプル数と対象とするタスクの特徴空間の次元数の関係を十分考慮し，サンプルが特徴空間上で十分密に分布している場合にのみ用いるべきであると言える．

6.3 ハイパーパラメタの影響

SSCRF-ER と SSCRF-MD のハイパーパラメタの値の違いによる性能の変化を検証する．ただし前述のように，事後確率最大化学習で用いられる事前分布のハイパーパラメタは比較手法間で同じ条件で扱われているので性能比較はおこなわない．ここでは，それぞれの手法特有のハイパーパラメタの振舞いについて議論をおこなう．

式 (3) で定義される SSCRF-ER は，γ で表されるラベルなしデータ (或はラベルありデータ) の項をどの程度信頼して学習をおこなうかを示すトレードオフパラメタを持っている．一方，SSCRF-MD では，式 (5) や式 (6) から分かるように，明示的に示されてるハイパーパラメタは存在しない．しかし，式 (6) を最大化する際に利用する事前分布にはしばしばハイパーパラメタが存在する．特に本稿では，この事前分布にディリクレ事前分布を用いることとするため，ディリクレ事前分布のハイパーパラメタ α が存在する．よって，ここでの実験では SSCRF-ER のトレードオフパラメタ γ と SSCRF-MD で利用されるディリクレ事前分布のハイパーパラメタ α を変化させたときの性能の変化を比較検証する．

図 7 に結果を示す．注意点として，SSCRF-ER と SSCRF-MD で比較されているハイパーパラメタの性質は同じではないため，横軸のスケールにはあまり意味がないことである．ただし，表の左端は，SSCRF-ER の場合は $\log\gamma = -\infty$ なので，$\gamma = 0$，つまりラベルなしデータの項を用いないこととなり，教師あり学習と同じ結果となる．一方，SSCRF-MD の左端の意味するものは，式 (6) の (ディリクレ) 事前分布の項を用いないで学習するとことを意味する．次に，SSCRF-ER の右端は，γ が大きい値，つまりラベルなしデータの項をラベルありデータよりも重く見積もって学習を行うことを意味する．一方，SSCRF-MD の右端は，式 (6) で事前分布を一様分布に近い分布を仮定して生成モデル学習することに等しく，ほぼ教師あり学習で学習しているのと同じこととなる．

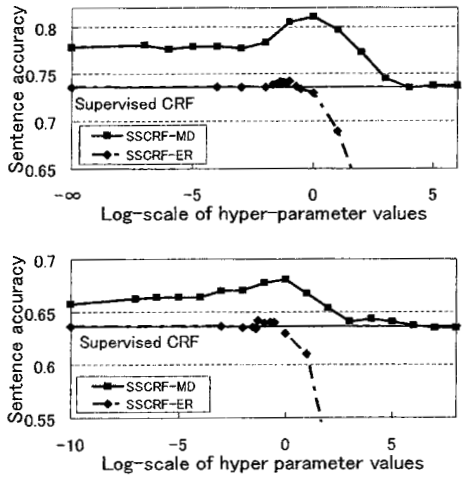

図 7: ハイパーパラメタに対する性能の変化 (上:固有表現抽出, 下:チャンキング)

よって，SSCRF-ER の左端と SSCRF-MD の右端は教師あり学習の結果と同じ値をとることが示されている．

図から明らかなように，ほぼ全値域で SSCRF-MD の性能が SSCRF-ER の最大性能を上回っている．特に，SSCRF-ER の致命的な欠点として，γ を大きくする，つまりラベルなしデータから計算されるエントロピーの項を重く評価して学習をおこなうと性能が急激に低下することが見て取れる．また，性能が最大になる値から性能が急速に低下する値がそれほど離れていないという観点からも，性能がハイパーパラメタの値に大きく影響を受けるということがわかる．実際に，これらの図を描画するためには，ピークの性能が得られる値を見付けるためにピーク周辺で細かく値を変化させて捜し出す必要があった．

一方，SSCRF-MD は，ハイパーパラメタの値を大きくし過ぎても (右端)，少なくとも教師あり学習と同じ程度の性能は得られることが保障され，また，ハイパーパラメタを用いなくても (左端)，そこそこ良い性能が得られるという利点がある．このように，SSCRF-MD は，人手あるいは開発セット等によって決定されるハイパーパラメタの値に対して得られる性能の信頼性が比較的高いと言える．

ただしこれは，SSCRF-ER が尤度とエントロピー項の足し算モデルにより定式化されていることに起因するため，SSCRF-ER も SSCRF-MD のように，掛け算モデルで定式化するとった方法が開発されれば，トレードオフパラメタによる性能の影響を受けないように改良することが可能であると考えられる．

6.4 計算時間

半教師あり学習では，通常ラベルありデータの数十倍以上のラベルなしデータを利用して学習をおこなう．第2節で示した本稿の実験で用いたデータの例では，約80倍のラベルなしデータを用いている．また，将来的にラベルなしデータは増える一方だが，ラベルありデータが増えることはまれであるという事実がある．つまり，半教師あり学習の計算時間に関しては，ラベルなしデータに対する計算アルゴリズムの効率が非常に重要な要素となる．

SSCRF-ERでは，ラベルなしデータに対するエントロピー項の勾配計算が学習時間に最も支配的な要素となる．文献[3]では，Lを正解ラベル数，Sをサンプルの状態数とすると，1サンプルあたり計算オーダー$O(L^3S^2)$のアルゴリズムが提案されている．しかし，文献[8]により$O(L^2S)$の計算アルゴリズムが開発された．

一方，SSCRF-MDでは，識別関数和最大化のときのQ関数の計算が計算時間の支配的要素である．この計算は，文献[1]にあるようにHMMのパラメタ推定と同じforward-backward(Baum-Welch)アルゴリズムにより構成され，計算オーダーは$O(L^2S)$である．つまり，SSCRF-ERとSSCRF-MDでのラベルなしデータ1サンプルに対する計算量は同じである．

ただし，実際の計算時間という点では，SSCRF-ERとSSCRF-MDともに反復計算で収束するまで学習が行われるため，収束するまでの反復回数によって計算時間は異なってくる．SSCRF-ERでは，エントロピーの勾配計算に1サンプルあたり$O(L^2S)$の計算を3回おこなう必要があることや，経験的には，SSCRF-MDの方が収束までの反復計算回数が少ないため，実際の計算時間はSSCRF-MDの方が短い場合がほとんどである．

参考として，評価時は，SSCRF-ERもSSCRF-MDも通常のCRFと全く同じViterbiアルゴリズムで計算されるため，計算時間は同じである．

7 まとめ

本稿では，自然言語解析タスクに対して，エントロピー正則化，および，識別関数和最大化に基づく半教師あり条件付確率場の性能と性質を実験的に検証した．

本稿で提案した識別関数和最大化に基づく半教師あり条件付確率場は，固有表現抽出とチャンキングタスクにおいて，教師あり学習時の条件付確率場の性能を大幅に向上できることを示した．同時に，自然言語解析タスクのように高次元スパース特徴空間を用いる問題に対しても効率的に半教師あり学習がおこなえることを示した．

一方，エントロピー正則化に基づく半教師あり条件付確率場は，高次元スパース特徴空間を用いる問題では性能の向上が得られないことを示した．一般的な議論として，半教師あり学習法では，サンプルの分布等，なにかしらの仮定を用いることでラベルなしデータから情報を得て性能の向上を実現している．よって，その仮定が合わない状況では性能の向上は難しい．高次元スパース空間では全てのサンプルが疎に分布しているため，エントロピー正則化の仮定に問題が適さない状況であったことが性能が向上が得られなかった原因である．

現在，汎用的な半教師あり学習法が盛んに開発されているが，実タスクへ適用する際には，利用する半教師あり学習法の仮定と，解きたい問題の性質が適合しているかを十分検証する必要がある．

参考文献

[1] J. Suzuki, A. Fujino, and H. Isozoki. Semi-Supervised Structured Output Learning based on a Hybrid Generative and Discriminative Approach. In *Proc. of EMNLP-CoNLL*, pages 791–800, 2007.

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

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

[4] U. Brefeld and T. Scheffer. Semi-Supervised Learning for Structured Output Variables. In *Proc. of ICML-2006*, 2006.

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

[6] Y. Grandvalet and Y. Bengio. Semi-Supervised Learning by Entropy Minimization. In *Proc. of NIPS*2004*, pages 529–536, 2004.

[7] C.-H. Lee, S. Wang, F. Jiao, D. Schuurmans, and R. Greiner. Learning to Model Spatial Dependency: Semi-Supervised Discriminative Random Fields. In *Proc. of NIPS*2006*, 2006.

[8] G. S. Mann and A. McCallum. Efficient Computation of Entropy Gradient for Semi-Supervised Conditional Random Fields. In *Proc. of NAACL-HLT-2007 (short paper)*, 2007.

[9] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39:103–134, 2000.