情報処理学会論文誌　Vol. 52　No. 11　3038–3051 (Nov. 2011)

大規模データを用いた半教師あり学習による高精度係り受け解析モデルの学習

鈴木　潤[†1]　磯崎秀樹[†1]　永田昌明[†1]

係り受け解析では，正解係り受け構造が付与されたデータを用いた教師あり学習により解析器を学習するのが現在最も一般的な方法であり，データ量が十分あれば非常に高い解析精度が得られることが実証されている．しかし，さらなる解析精度向上のため，正解データを増やし続けるのは作成に要する費用や時間の観点で現実的な方策ではない．そこで本論文では，正解係り受け構造が付与されていないデータも利用して解析精度を向上させる，いわゆる半教師あり学習に基づく係り受け解析モデルとその学習法を提案する．実験では，係り受け解析の標準評価データとして広く利用されている，係り受け構造が交差するチェコ語，交差しない英語の 2 言語の係り受け解析データを用いて，提案法の有効性を定性的，定量的に検証する，提案法は，従来の教師あり学習で得た係り受け解析器を大幅に上回る解析精度を達成することを示す．

Learning High-performance Dependency Parsing Models by Large-scale Semi-supervised Learning

Jun Suzuki,[†1] Hideki Isozaki[†1] and Masaaki Nagata[†1]

Intensive work have recently been undertaken to develop dependency parsing. Most of the recent developed dependency parsers are obtained by using supervised learning with labeled data. In contrast, this paper introduces a high-performance dependency parser trained by semi-supervised learning, which is able to effectively incorporate unlabeled data as an additional training data. We demonstrate the effectiveness of our proposed method on dependency parsing experiments using two widely used test collections: the Penn Treebank for English as a projective dependency parsing, and the Prague Dependency Treebank for Czech as a non-projective dependency parsing. Our results in the above datasets significantly outperform those obtained from conventional supervised learning approach.

1. はじめに

テキストの係り受け解析は，文内の語や節間の依存関係を解析する問題である．近年では，2006，2007 年の CoNLL shared task[1),2)] で取り上げられ，多数の言語で同一規格の係り受け解析データが作成される等，国際的にも自然言語解析技術の重要な基盤技術の 1 つとして研究が続けられている．

係り受け解析では，正解係り受けが付与されたデータから教師あり学習（近年では主に識別学習）により解析器の学習を行うのが現在最も一般的な方法である．教師あり学習による方法は，正解データが十分な量確保できれば，高い解析精度が得られることが数多くの文献で報告されている[3)–7)]．しかし，さらなる解析精度向上のため，正解データを増やし続けるのは作成に要する費用や時間の観点で現実的な方策とはいえない．よって次のステップとして，正解データ量の増加以外の方法でさらなる解析精度向上が可能な方法が模索されている．

係り受け解析では，一般的に 1 入力は 1 文である．正解の係り受け構造が付与されていないデータであれば，現在 web データや電子化文書等から比較的容易かつ大量に獲得することが可能である．よって，現実的な選択肢の 1 つとして，正解係り受け構造が分からないデータを利用して係り受け解析精度を向上させる取り組みが考えられる．このように，正解データに加え正解が不明なデータを用いて学習を行う枠組みは，一般的に半教師あり学習と呼ばれ，自然言語処理や機械学習の分野で近年さかんに研究されている．本論文では，係り受け解析タスクを対象として，半教師あり学習により係り受け解析モデルを学習し，解析精度を向上させる方法について議論を行う．

一般論として正解が不明なデータは，正解が不明であるがゆえに，教師あり学習のデータとして正解データと同じように利用することはできない．しかし，正解が不明なデータからでも教師あり学習時に利用する特徴は抽出可能であり，特徴空間に関して何らかの統計量を計算することができる．近年発展した半教師あり学習法の多くは，それぞれの方法に基づいて特徴空間に関する何かしらの統計量を正解が不明なデータから獲得し，教師あり学習時に利用する方法と見なすことができる[8)–11)]．特に係り受け解析の場合は，単語間の依存関係を学習するため，2 単語の組合せが学習時に利用する特徴の基本的な単位となる．つま

†1 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
NTT Communication Science Laboratories, NTT Corporation

り，高次元かつ疎な特徴空間を利用しないと問題全体を表現するのが難しい問題といえる．このような性質を持つ係り受け解析タスクでは，正解データに出現する特徴の種類は，タスク全体で出現可能な特徴の種類数に比べて，非常に少ないと考えられる．よって，大規模な正解が不明なデータから得られた特徴空間に関する統計量は，正解データだけでは未知の特徴空間の領域に対する何らかの有益な情報になると期待できる．そして，この統計量を教師あり学習でうまく利用することで，結果的に，解析精度の向上が期待できる．

本論文では，まず教師あり学習のモデルとして，係り受け解析の従来研究で用いられている条件付確率場[12] を係り受け解析に適用したモデル[13),14)] を取り上げる．この係り受け解析用の条件付確率場に，文献 9)，15) で提案された半教師あり学習の枠組みに基づいて，係り受け構造に対しそこに出現する各特徴の出現確率を正解が不明なデータから推定する目的で，生成モデルを組み合わせる．組み合わせた生成モデルで推定した各特徴の出現確率は，係り受け構造へのなりやすさの推定値と見なせるため，この情報が係り受け構造を決定するときに役立つと考えられる．最終的に，この組合せモデルに対して，正解データと正解が不明なデータの双方の情報を使って，各々の情報を補完しあいながら学習を行うような学習法を述べる．

本論文の主な貢献は，(1) 従来の教師あり学習による係り受け解析モデルを，半教師あり学習が可能なモデルへ拡張し，その学習法を考案した点，(2) 提案法では，正解データが少量の場合でも比較的大量にある場合でも，十分大きな効果が得られる方法であることを実証した点，(3) 提案法では，正解が不明なデータでも，データ量を増やすことによって係り受け解析精度をさらに向上させることができることを実証した点，等である．

以下，一般的な機械学習の用語に則って，本論文では正解データを ‘ラベルありデータ’，正解が不明なデータを ‘ラベルなしデータ’ と記述する．

2. 係り受け解析での一次依存条件付確率場と教師あり学習

本論文では，係り受け解析の基本モデルとして，系列ラベリング等でよく用いられる条件付確率場[12] を用いる[13),14)]．以下，条件付確率場による教師あり学習法について述べる．

2.1 一次依存条件付確率場の定義

係り受け解析の場合，入力は文であり，出力は係り受け構造である．ここでは，入力文を $\mathbf{x}$，出力係り受け構造を $\mathbf{y}$ と書く．h を係り受け構造の係り元の単語のインデックス，m を係り受け構造の係り先の単語のインデックス，r を係り受け関係を表すラベルのインデックスとする．ただし，係り受け関係を表すラベルは事前に定義されているものとし，全部で R 個とする．また，0 を文のルートを表すインデックスとする．このとき，n 単語[⋆1]からなる文 $\mathbf{x}$ が与えられたとき，ラベル付き係り受け構造 $\mathbf{y}$ は，n 個の 3 タプル (h, m, r) で表現できる．つまり，$\mathbf{y} = \{(h, m, r)_i\}_{i=1}^{n}$ である．また，$h \in \{0, ...n\}$，$m \in \{1, ...n\}$，$r \in \{1, ...R\}$ となる．

ある入力文 $\mathbf{x}$ と，出力係り受け構造 $\mathbf{y}$ 中の単一の係り受け構造 (h, m, r) から抽出される特徴ベクトルを $\mathbf{f}(\mathbf{x}, h, m, r)$ と表す．また，この特徴ベクトルに対するパラメータベクトルを $\mathbf{w}$ とする．このとき，$\mathbf{x}$ と $\mathbf{y}$ に対する線形判別関数 $g(\mathbf{x}, \mathbf{y})$ は以下のように定義できる．

$$g(\mathbf{x}, \mathbf{y}) = \sum_{(h,m,r)\in\mathbf{y}} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, h, m, r) \tag{1}$$

次に，可能なすべての入力文の集合を $\mathcal{X}$，出力係り受け構造の集合を $\mathcal{Y}$ と定義する．つまり，$\mathbf{x} \in \mathcal{X}$ および $\mathbf{y} \in \mathcal{Y}$ である．ある入力文 $\mathbf{x}$ に対して，得られる可能なすべての出力係り受け構造の集合を $\mathcal{Y}(\mathbf{x})$ とする．これは $\mathcal{Y}(\mathbf{x}) \subseteq \mathcal{Y}$ の関係が成り立つ．このとき，入力文 $\mathbf{x}$ が与えられたときの出力係り受け構造 $\mathbf{y}$ の条件付確率 $p(\mathbf{y}|\mathbf{x})$ は，以下のように定義できる．

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(g(\mathbf{x}, \mathbf{y})\right), \quad \text{ただし} \quad Z(\mathbf{x}) = \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \exp\left(g(\mathbf{x}, \mathbf{y})\right) \tag{2}$$

これは，条件付確率場を係り受け解析に適用したモデルと見なすことができる．

入力文 $\mathbf{x}$ と単一の係り受け構造 (h, m, r) から得られる特徴のみを利用した係り受け解析モデルを，一次依存係り受け解析モデルと呼ぶ[6]．そこで本論文では，式 (1) を式 (2) に代入した係り受け解析での条件付確率場を，一次依存条件付確率場と呼ぶ．

2.2 解析アルゴリズム

$\mathbf{x}$ が与えられたときの最尤出力 $\hat{\mathbf{y}}$ は，$\mathbf{x}$ が与えられた際の $\mathbf{y}$ の条件付確率 $p(\mathbf{y}|\mathbf{x})$ から求めることができる．ただし，$p(\mathbf{y}|\mathbf{x})$ は $g(\mathbf{x}, \mathbf{y})$ の単調増加関数なので，以下の式に帰着することができる．

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} g(\mathbf{x}, \mathbf{y}) \tag{3}$$

式 (3) は，$g(\mathbf{x}, \mathbf{y})$ が最も大きい $\mathbf{y}$ を $\mathcal{Y}(\mathbf{x})$ の中から探索する問題であり，自然言語処理分

⋆1 実際のデータは，Penn Treebank 形式の Tokenize を行うので，厳密には単語ではない．しかし，本論文では分かりやすさを重視して ‘トークン’ を単に ‘単語’ と記述することにする．

野ではデコーディングと呼ぶ．係り受け解析の場合は $\mathcal{Y}(\mathbf{x})$ 内の出力候補数は文の長さ n に関して指数関数的に増加する．現実的な速度で解析を行うためには，多項式オーダの探索アルゴリズムが必要になる．

係り受け解析は大きく分けて，英語や日本語のように一般的に係り受け構造に交差がないと仮定する場合に使われる projective 係り受け解析と，チェコ語のように交差があると仮定する場合に使われる non-projective 係り受け解析がある．projective か non-projective かによって式 (3) を求める際の効率的な解析アルゴリズムが異なる．projective 係り受け解析の場合は，文献 4) で紹介されたように，CKY チャートパージングアルゴリズムを効率化した Eisner アルゴリズム[16] により，$O(n^3)$ の計算量で求めることができる．また，non-projective 係り受け解析の場合は，文献 5) で紹介されたように，式 (3) の推定を最大全域木の推定問題と見なすことで，Chu-Liu-Edmonds アルゴリズム[17),18)] により，$O(n^2)$ の計算量で求めることができる．

2.3　パラメータ推定：教師あり学習

ラベルあり学習データを $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ とし，N 個のサンプルで構成されているとする．条件付確率場では，式 (2) の条件付確率の対数尤度を最大化するパラメータを推定する方法が最も基本的な学習方法である．実用的には，ラベルあり学習データに過適応することを防ぐためにパラメータの事前分布 $p(\mathbf{w})$ を導入し，ラベルあり学習データが与えられたときのパラメータ $\mathbf{w}$ の事後確率 $p(\mathbf{w}|\mathcal{D}_L)$ を最大化するパラメータ $\mathbf{w}$ を推定する方法が主流である[13),14)]．ただし，

$$p(\mathbf{w}|\mathcal{D}_L) = \prod_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_L} p(\mathbf{w}|\mathbf{x},\mathbf{y}) = \prod_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_L} \frac{p(\mathbf{y}|\mathbf{x};\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})}$$

である．$p(\mathbf{w}|\mathcal{D}_L)$ を最大にするパラメータを $\hat{\mathbf{w}}$ とすると，$\hat{\mathbf{w}}$ は $p(\mathbf{w}|\mathcal{D}_L)$ の対数からパラメータ $\mathbf{w}$ に依存しない項 $p(\mathbf{y})$ を除いた式 $\mathcal{O}(\mathbf{w}|\mathcal{D}_L)$ の最大化問題の解と等価であり，以下の式で求まる[19)]．

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \mathcal{O}(\mathbf{w}|\mathcal{D}_L),\ \text{ただし}\ \mathcal{O}(\mathbf{w}|\mathcal{D}_L) = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_L} (\log p(\mathbf{y}|\mathbf{x};\mathbf{w}) + \log p(\mathbf{w})) \quad (4)$$

本論文では，事前分布 $p(\mathbf{w})$ にガウス分布 $p(\mathbf{w}) = \exp(-||\mathbf{w}||^2/2C)$ を用いる．C は人手により決定するチューニングパラメータである．一次依存条件付確率場の場合，式 (4) の最大化は凸最適化である．実際に $\hat{\mathbf{w}}$ を推定する際には，勾配法に属する数値最適化法を用いて反復計算により最適解を求める方法が最近の主流である．特に，準ニュートン法の一種である L-BFGS[20)] が収束性や省メモリ性の観点で近年よく利用されている．式 (4) の $\mathcal{O}(\mathbf{w}|\mathcal{D}_L)$ の $\mathbf{w}$ に関する勾配は以下のようになる．

$$\nabla\mathcal{O}(\mathbf{w}|\mathcal{D}_L) = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_L} \nabla g(\mathbf{x},\mathbf{y}) - \frac{1}{Z(\mathbf{x})} \sum_{\mathbf{y}'\in\mathcal{Y}(\mathbf{x})} \exp\left(g(\mathbf{x},\mathbf{y}')\right) \nabla g(\mathbf{x},\mathbf{y}') - \frac{\mathbf{w}}{C} \quad (5)$$

$$\nabla g(\mathbf{x},\mathbf{y}) = \sum_{(h,m,r)\in\mathbf{y}} \mathbf{f}(\mathbf{x},h,m,r)$$

式 (4)，(5) 中の $Z(\mathbf{x})$ と，式 (5) の右辺第 2 項にはすべての出力係り受け構造に対する計算 $\sum_{\mathbf{y}'\in\mathcal{Y}(\mathbf{x})}$ が必要である．この計算は，理論上は $\mathcal{Y}(\mathbf{x})$ 中の全候補を列挙すれば計算可能であるが，現実的な処理のためには効率的な計算アルゴリズムが必要になる．これらを多項式時間で計算するアルゴリズムは，最尤出力 $\hat{\mathbf{y}}$ を求めるアルゴリズムと同様に，係り受けの交差あり（non-projective），交差なし（projective）によって異なる．projective 係り受け解析の場合は，inside-outside アルゴリズム[21)] に Eisner のデータ構造[16)] を利用することで $O(n^3)$ の計算量で求めることができる[22)]．また，non-projective の場合は，Matrix-Tree Theorem[23)] による逆行列と行列の対角化の計算によって，同じく $O(n^3)$ の計算量で計算することができる[13),14),24)]．

3. 係り受け解析での一次依存条件付確率場と半教師あり学習

前章で述べた一次依存条件付確率場を半教師あり学習に対応するための拡張と，その学習法について述べる．提案法は，基本的に文献 15) で提案されている半教師あり学習法の枠組みを用いる．文献 15) の半教師あり学習の枠組みでは，まず生成モデル $p(\mathbf{x},\mathbf{y})$ を定義し，その生成モデルを組み込んだ識別モデル $p(\mathbf{y}|\mathbf{x})$ を定義する．生成モデルを利用するのは，識別モデルではラベルなしデータは出力 $\mathbf{y}$ が不明であるため学習に直接利用することは不可能であるが，生成モデルなら出力 $\mathbf{y}$ を欠損データと見なすことで，学習に利用することが可能なためである．つまり，ラベルなしデータの情報を取り込むために生成モデルの性質を利用する．また，ラベルありデータは，従来の教師あり学習と同じで，識別モデルの学習に利用する．ただし，識別モデルと組み込まれた生成モデルは，それぞれのデータから独立に学習が行われるのではなく，お互いの持つ情報を相補的に補完し合いながら学習を行う枠組みとなっている．

この枠組みを用い，本論文では，識別モデルとして前述した一次依存条件付確率場をベースにした係り受け解析に適した半教師あり学習法を提案する．ただし，係り受け解析タスク

の性質上，文献 15) の方法を単純に適用するだけでは不十分であるため，係り受け解析に適した拡張を行う．具体的には，任意の 2 単語間が係り受け関係になる尤度を，その 2 単語間の係り受けを含む係り受け構造が，出力として選択される分布に基づいて各特徴の出現確率を推定した生成モデル $u_{\boldsymbol{\theta}}$ と，出力として選択されない分布に基づいて各特徴の出現確率を推定した生成モデル $u_{\boldsymbol{\mu}}$ の対数尤度比で表現する．

この理由は，$u_{\boldsymbol{\theta}}$ において推定される係り受け構造が選出される分布に基づく各特徴の出現確率は，特徴の出現頻度によるバイアスを受けるが，このバイアスが係り受け解析タスクの性質と合わないためである．つまり，生成モデルの性質上，ラベルなしデータから各特徴の出現確率を推定する際に，高頻度語間の係り受けで得られる特徴の出現確率は相対的に高く，低頻度語間で得られる出現確率は相対的に低く見積もられることになる．しかし実際に，2 単語間が係り受け関係になるかならないかは単語の出現頻度とは関係なく決定される問題である．よって，この出現頻度バイアスにより，$u_{\boldsymbol{\theta}}$ 単体では，2 単語間の係り受けになりやすさを推定する統計量としては不十分であると考えられる．そこで提案法では，この出現頻度バイアスを補正するため，係り受けが選出されない分布に基づく各特徴の出現確率を推定した生成モデル $u_{\boldsymbol{\mu}}$ も導入し，これらの対数尤度比をとることで，出現頻度バイアスを打ち消すことができる．そして対数尤度比の値を，バイアスがない状態での各特徴の係り受け関係になりやすさを示す統計量として利用する．

参考までに，分類問題で用いるナイーブベイズ分類器や，系列ラベリングで用いる隠れマルコフモデル等では，このバイアス問題は起こらないことに注意されたい．係り受け解析が，木構造になるという制約下で 2 単語間の依存構造を予測する問題であるためにこのような問題が発生する．

最後に，ベイズ則 $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})/p(\mathbf{x})$ に従って，$\mathbf{x}$ が既知のときには，$u_{\boldsymbol{\theta}}$ は，係り受けになりやすさを表す尤度，$u_{\boldsymbol{\mu}}$ は，係り受けになりにくさを表す尤度とも見なすことができる．

3.1　ラベルなしデータ用の生成モデルの定義

ここでは，ラベルなしデータを扱うために導入する生成モデルを定義する．まず，各係り受け関係のラベルごとにそれぞれ特徴ベクトルを定義する．係り受け関係のラベルの総数を R とすると，独立した特徴ベクトルが R 個あることになる．ここでは，r 番目の係り受け関係のラベルに対する特徴ベクトルを $\mathbf{e}_r(\mathbf{x}, h, m)$ と書く．また，$\mathbf{e}_r(\mathbf{x}, h, m)$ は，d_r 次元の特徴ベクトルとし，$e_{r,a}(\mathbf{x}, h, m)$ をその a 番目の要素とする．

次に，$\mathbf{e}_r$ に対するパラメータベクトルを $\boldsymbol{\theta}_r = (\theta_{r,1}, \ldots, \theta_{r,d_r})$ とする．ただし，パラメータ $\boldsymbol{\theta}_r$ の各要素は，$\theta_{r,a} \geq 0$ と $\sum_{a=1}^{d_r} \theta_{r,a} = 1$ の 2 つの制約を満たす値をとると仮定する．また，$\boldsymbol{\theta}_r$ と同様に，$\mathbf{e}_r$ に対する別のパラメータベクトルを $\boldsymbol{\mu}_r = (\mu_{r,1}, \ldots, \mu_{r,d_r})$ とし，$\mu_{r,a} \geq 0$ と $\sum_{a=1}^{d_r} \mu_{r,a} = 1$ を満たすこととする．ここで，パラメータが多項分布となる特徴ベクトルによる以下の 2 つの生成モデル $u_{\boldsymbol{\theta}}$ と $u_{\boldsymbol{\mu}}$ を導入する．

$$u_{\boldsymbol{\theta}}(\mathbf{x}, h, m, r) = p(r) \prod_{a=1}^{d_r} (\theta_{r,a})^{e_{r,a}(\mathbf{x},h,m)}\,,\ u_{\boldsymbol{\mu}}(\mathbf{x}, h, m, r) = p(r) \prod_{a=1}^{d_r} (\mu_{r,a})^{e_{r,a}(\mathbf{x},h,m)}$$

パラメータ $\boldsymbol{\theta}_r$ と $\boldsymbol{\mu}_r$ の推定方法の詳細は 3.3 節で述べる．直感的な説明としては，$\theta_{r,a}$ は，h，m の係り受け関係が係り受け構造 $\mathbf{y}$ の一部に含まれるとき，その $\mathbf{y}$ が出力として選択される分布に基づいて特徴 $e_{r,a}(\mathbf{x}, h, m)$ が出現する確率をラベルなしデータから推定した値である．一方，$\mu_{r,a}$ は，h，m の係り受け関係がある係り受け構造 $\mathbf{y}$ の一部に含まれるとき，その $\mathbf{y}$ が出力として選択されない分布に基づいて特徴 $e_{r,a}(\mathbf{x}, h, m)$ が出現する確率をラベルなしデータから推定した値である．

生成モデル用の特徴 $\mathbf{e}_r(\mathbf{x}, h, m)$ は，理論的には人手により自由に定義すればよい．しかし，ここでは計算コストの増大を防ぐ効果を期待し，教師あり学習で用いる特徴ベクトル $\mathbf{f}(\mathbf{x}, h, m, r)$ を再利用する．係り受け解析では，係り先と係り元の品詞の組合せ，係り先と係り元の単語の組合せ，係り先と係り元の品詞とその間に出現する品詞の組合せといった具合に，数十種類の特徴テンプレートを用いて特徴ベクトル $\mathbf{f}(\mathbf{x}, h, m, r)$ を作成するのが一般的である[4),5)]．そこで提案法では，文献 15) に従って，特徴テンプレートごとに 1 つの生成モデルを定義し，複数の生成モデルを利用する方法を用いる．つまり，一次依存条件付確率場に 10 種類の特徴テンプレートを用いる場合は，10 個の生成モデルを使う構成となる．

3.2　判別関数の拡張

まず，前述の $u_{\boldsymbol{\theta}}$ と $u_{\boldsymbol{\mu}}$ の対数尤度比を用いて q を以下のように定義する．

$$q(\mathbf{x}, h, m, r) = \log \frac{u_{\boldsymbol{\theta}}(\mathbf{x}, h, m, r)}{u_{\boldsymbol{\mu}}(\mathbf{x}, h, m, r)} = \sum_{a=1}^{d_r} e_{r,a}(\mathbf{x}, h, m)(\log \theta_{r,a} - \log \mu_{r,a}) \qquad (6)$$

次に，前節で述べたように J 種類の特徴テンプレートから得られた J 個の生成モデルの対数尤度比があるとし，j 番目の対数尤度比を q_j を表す．また，$v_1, \ldots, v_J$ を q_j に対応する $\mathbf{w}$ と同様な J 個のパラメータとする．このとき，q_j を用いて，半教師あり学習用の一次依存条件付確率場の判別関数 $g(\mathbf{x}, \mathbf{y})$ を，以下のように定義する．

$$g(\mathbf{x},\mathbf{y}) = \sum_{(h,m,r)\in\mathbf{y}} \mathbf{w}\cdot\mathbf{f}(\mathbf{x},h,m,r) + \sum_{(h,m,r)\in\mathbf{y}}\sum_{j=1}^{J} v_j q_j(\mathbf{x},h,m,r) \tag{7}$$

直感的な解釈としては，v_j は q_j の信頼性をラベルありデータ上で評価した値を反映するパラメータといえる．表記を簡略化するため $\mathbf{v}=(v_1,\ldots,v_J)$ と $\mathbf{q}=(q_1,\ldots,q_J)$ を導入する．

3.3　パラメータ推定：ラベルありとラベルなし学習データからの半教師あり学習

提案法では，3 ステップのパラメータ推定を用いて半教師あり学習を行う．まず，ラベルあり学習データを $\mathcal{D}_L=\{\mathbf{x}_i,\mathbf{y}_i\}_{i=1}^N$，ラベルなし学習データを $\mathcal{D}_U=\{\mathbf{x}'_i\}_{i=1}^M$ とする．全体の学習アルゴリズムへの入力は，教師あり学習データ $\mathcal{D}_L$，教師なし学習データ $\mathcal{D}_U$，事前に人手により定義した特徴ベクトル $\mathbf{f}$ である．前述したように，生成モデル用の特徴ベクトル $\mathbf{e}_j(\forall j)$ の定義は $\mathbf{f}$ の定義から自動的に生成される．

出力は，パラメータベクトル $\mathbf{w}$ と $\mathbf{v}$ と生成モデルの対数尤度比の集合 $\mathbf{q}$ となる．ただし，パラメータベクトル $\mathbf{w}$ と $\mathbf{v}$ は，ラベルあり学習データから教師あり学習により推定され，$\mathbf{q}$ はラベルなし学習データから推定される．

以下に，一次依存条件付確率場による半教師あり学習の手続きを述べる．

第 0 ステップ：初期化

生成モデルの出力する確率が一様分布に従うように生成モデルを初期化する．ここでは，すべての j，r，a の組に対して $\theta_{j,r,a}=\mu_{j,r,a}=1/d_{j,r}$ を初期値として得られる値を $\mathbf{q}^0$ とする．つまり，$\mathbf{q}^0$ は，式 (6) からゼロベクトルとなる．

第 1 ステップ：教師あり学習

ラベルあり学習データ $\mathcal{D}_L$ と初期化した $\mathbf{q}^0$ を用いて一次依存構造確率モデルのパラメータ推定を行う．ここで，式 (2) の判別関数 g に式 (7) を代入して得られる条件付確率を $p(\mathbf{y}|\mathbf{x};\mathbf{w},\mathbf{v},\mathbf{q})$ とする．初めに生成モデル $\mathbf{q}^0$ を固定した状態で，ラベルあり学習データ $\mathcal{D}_L$ に対してパラメータ $\mathbf{w}$，$\mathbf{v}$ の対数事後確率が最大になるパラメータ $\mathbf{w}^0$，$\mathbf{v}^0$ を求める．

$$(\mathbf{w}^0,\mathbf{v}^0) = \arg\max_{\mathbf{w},\mathbf{v}} \mathcal{O}_L(\mathbf{w},\mathbf{v}|\mathcal{D}_L,\mathbf{q}^0)$$

$$\mathcal{O}_L(\mathbf{w},\mathbf{v}|\mathcal{D}_L,\mathbf{q}^0) = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_L} \log p(\mathbf{y}|\mathbf{x};\mathbf{w},\mathbf{v},\mathbf{q}^0) + \log p(\mathbf{w},\mathbf{v})$$

$$\log p(\mathbf{w},\mathbf{v}) = -\frac{||\mathbf{w}||^2+||\mathbf{v}||^2}{2C} \tag{8}$$

ただし $\log p(\mathbf{w},\mathbf{v})$ は正則化項に相当し，上式に示したように，式 (4) と同様に $p(\mathbf{w},\mathbf{v})$ にはガウス分布を用いる．C も同様に人手により決定するチューニングパラメータである．

この最適化は，$\mathbf{q}^0$ を教師あり学習の特徴の一種と思えば，式 (4) の最適化と等価である．つまり，この最適化は凸最適化問題であり，勾配に基づく反復計算法で最適解を得ることができる[⋆1]．よって，この第 1 ステップの $\mathbf{w}^0$，$\mathbf{v}^0$ の推定は，2.3 節で述べた従来の一次依存条件付確率場の教師あり学習と同じ方法で求めることができる．

第 1 ステップで得た $\mathbf{w}^0$，$\mathbf{v}^0$，$\mathbf{q}^0$ による一次依存条件付確率場を $p^0(\mathbf{y}|\mathbf{x})$ と書く．

第 2 ステップ：生成モデルの推定

第 2 ステップでは，第 1 ステップで得た $p^0(\mathbf{y}|\mathbf{x})$ と，ラベルなし学習データ $\mathcal{D}_U$ を用いて $\mathbf{q}^1$ を推定する．実際には，生成モデル $u_{\boldsymbol{\theta}}$ と $u_{\boldsymbol{\mu}}$ の推定を行い，その後 q を計算する．ここでの処理は，直感的には，ラベルなし学習データでは正解係り受け構造が不明なので，第 1 ステップで得た一次依存条件付確率場 $p^0(\mathbf{y}|\mathbf{x})$ の推定値を用いて係り受け構造を補完し，その補完した係り受け構造の情報を用いて，ラベルなしデータが与えられたときの生成モデルのパラメータの事後確率を最大にするパラメータを求める処理となっている．

この処理は，すべての q_j に対して共通な式を用いるので，ここではある j に対する式のみを示す．また，式を簡潔にするために，本節では j の添え字を省略する．ただし実際には，すべての j に対して同様な計算を行う．

まず，式の簡略化のため $\bar{p}=1-p$，$\boldsymbol{\theta}=\{\boldsymbol{\theta}_r\}_{r=1}^R$，$\boldsymbol{\mu}=\{\boldsymbol{\mu}_r\}_{r=1}^R$ を導入する．第 2 ステップでは以下の最大化問題により，係り受け構造として選択される分布に基づいた特徴の出現確率 $\boldsymbol{\theta}$ と選択されない分布に基づいた特徴の出現確率 $\boldsymbol{\mu}$ の推定値 $\hat{\boldsymbol{\theta}}$ と $\hat{\boldsymbol{\mu}}$ を得る．

$$(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\mu}}) = \arg\max_{\boldsymbol{\theta},\boldsymbol{\mu}} \mathcal{O}_U(u_{\boldsymbol{\theta}},u_{\boldsymbol{\mu}}|\mathcal{D}_U,p^0)$$

$$\mathcal{O}_U(u_{\boldsymbol{\theta}},u_{\boldsymbol{\mu}}|\mathcal{D}_U,p^0) = \sum_{\mathbf{x}'\in\mathcal{D}_U}\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}')}\Bigg[p^0(\mathbf{y}|\mathbf{x}')\sum_{(h,m,r)\in\mathbf{y}}\log\big(u_{\boldsymbol{\theta}}(\mathbf{x}',h,m,r)\big) + \bar{p}^0(\mathbf{y}|\mathbf{x}')\sum_{(h,m,r)\in\mathbf{y}}\log\big(u_{\boldsymbol{\mu}}(\mathbf{x}',h,m,r)\big)\Bigg] + \sum_r\Big[\log p(\boldsymbol{\theta}_r)+\log p(\boldsymbol{\mu}_r)\Big]$$

$$\text{s.t.}\quad \forall r\quad \sum_{a=1}^{d_r}\theta_{r,a}=1, \sum_{a=1}^{d_r}\mu_{r,a}=1,\quad \forall r,a\quad \theta_{r,a}\geq 0,\mu_{r,a}\geq 0 \tag{9}$$

ただし，$\log p(\boldsymbol{\theta}_r)=(\eta-1)\sum_{a=1}^{d_r}\log\theta_{r,a}$，$\log p(\boldsymbol{\mu}_r)=(\eta-1)\sum_{a=1}^{d_r}\log\mu_{r,a}$

⋆1 $\mathbf{q}^0$ は，負の値となることもありうるが，凸関数であることには変わらない．

$p(\boldsymbol{\theta}_{j,r})$ および $p(\boldsymbol{\mu}_{j,r})$ は，多項分布モデルでよく用いられるディリクレ事前分布に相当し，η は，人手により決定するチューニングパラメータである．ただし，$\eta > 1$ である．

上式は制約付き最適化問題なので，ラグランジュ未定乗数 $\{\alpha_r\}_{r=1}^{R}$，$\{\beta_r\}_{r=1}^{R}$ を用いて以下の最適化問題を解く．

$$\begin{aligned}\mathcal{O}_U(u_{\boldsymbol{\theta}}, u_{\boldsymbol{\mu}}|\mathcal{D}_U, p^0) = &\sum_{\mathbf{x}'\in\mathcal{D}_U}\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}')}\Bigg[p^0(\mathbf{y}|\mathbf{x}')\sum_{(h,m,r)\in\mathbf{y}}\log\big(u_{\boldsymbol{\theta}}(\mathbf{x}', h, m, r)\big)\\ &+\bar{p}^0(\mathbf{y}|\mathbf{x}')\sum_{(h,m,r)\in\mathbf{y}}\log\big(u_{\boldsymbol{\mu}}(\mathbf{x}', h, m, r)\big)\Bigg]\\ &+\sum_r\Bigg[\log p(\boldsymbol{\theta}_r)+\log p(\boldsymbol{\mu}_r)-\alpha_r\Big(\sum_{a=1}^{d_r}\theta_{r,a}-1\Big)-\beta_r\Big(\sum_{a=1}^{d_r}\mu_{r,a}-1\Big)\Bigg]\end{aligned}$$

このとき，$\mathcal{O}_U$ のパラメータ $\theta_{r,a}$ に対する偏微分は，以下の式となる．

$$\frac{\partial\mathcal{O}_U(u_{\boldsymbol{\theta}}, u_{\boldsymbol{\mu}}|\mathcal{D}_U, p^0)}{\partial\theta_{r,a}} = \frac{1}{\theta_{r,a}}\sum_{\mathbf{x}'\in\mathcal{D}_U}\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}')}p^0(\mathbf{y}|\mathbf{x}')\sum_{(h,m,r)\in\mathbf{y}}e_{r,a}(\mathbf{x}', h, m)+\frac{(\eta-1)}{\theta_{r,a}}-\alpha_r$$

式 (9) を最大化するパラメータはすべての $\theta_{r,a}$ で上の偏微分が 0 になる点である．よって

$$\hat{\theta}_{r,a} = \frac{\hat{\theta}'_{r,a}}{\alpha_r},\quad \hat{\theta}'_{r,a} = \sum_{\mathbf{x}'\in\mathcal{D}_U}\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}')}p^0(\mathbf{y}|\mathbf{x}')\sum_{(h,m,r)\in\mathbf{y}}e_{r,a}(\mathbf{x}', h, m)+(\eta-1)\quad(10)$$

となる．ここで，$\sum_{a=1}^{d_r}\hat{\theta}_{r,a}=1$ の関係を用いると，$\alpha_r=\sum_{a=1}^{d_r}\hat{\theta}'_{r,a}$ が求まる．よって，最終的に式 (9) を最大化するパラメータ $\hat{\theta}_{r,a}$ は以下で求めることができる．

$$\hat{\theta}_{r,a} = \frac{\hat{\theta}'_{r,a}}{\sum_{a=1}^{d_r}\hat{\theta}'_{r,a}}\quad(11)$$

また，$\hat{\mu}_{r,a}$ も同様の導出で以下のようになる．

$$\hat{\mu}_{r,a}=\frac{\hat{\mu}'_{r,a}}{\sum_{a=1}^{d_r}\hat{\mu}'_{r,a}},\ \hat{\mu}'_{r,a}=\sum_{\mathbf{x}'\in\mathcal{D}_U}\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}')}\bar{p}^0(\mathbf{y}|\mathbf{x}')\sum_{(h,m,r)\in\mathbf{y}}e_{r,a}(\mathbf{x}', h, m)+(\eta-1)\quad(12)$$

実際の計算には，全出力候補に対する総和 $\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}')}$ の計算が必要である．これは，2.3 節で述べた一次依存条件付確率場の学習と同じ inside-outside アルゴリズム，または，Matrix-Tree Theorem による逆行列計算により計算できる．

注意点として，用いる特徴の種類 j によらず p^0 は同じものを用いるので，式 (9) で得られる J 個の独立な最適化は一括して推定することができる．つまり，inside-outside アルゴリズム，または，Matrix-Tree Theorem による逆行列計算は，J に依存せず 1 入力あたりたかだか 1 回で処理することができる．また，式 (11) および式 (12) から分かるように，式 (9) の解は解析的に求まるので，一次依存条件付確率場のように反復計算によるパラメータ推定は必要なく，データを 1 度処理するだけで解が求まる．このように提案法では，ラベルなし学習データが大規模化しても効率的な処理が可能なように考慮して構築されている．

最終的に，求めた $\hat{\boldsymbol{\theta}}$，$\hat{\boldsymbol{\mu}}$ から式 (6) に従って q を計算する．第 2 ステップで推定した J 個の q を合わせて $\mathbf{q}^1$ とする．

第 3 ステップ：教師あり学習による再推定

第 3 ステップでは基本的に第 1 ステップと同じ処理をする．ただし，第 1 ステップの目的関数式 (8) の $\mathcal{O}_L(\mathbf{w}, \mathbf{v}|\mathcal{D}_L, \mathbf{q}^0)$ を $\mathbf{q}^0$ から $\mathbf{q}^1$ に置き換えた関数 $\mathcal{O}_L(\mathbf{w}, \mathbf{v}|\mathcal{D}_L, \mathbf{q}^1)$ を用いて $\mathbf{w}^1$ と $\mathbf{v}^1$ を推定する．

最終的に第 2，3 ステップで得られた $(\mathbf{w}^1, \mathbf{v}^1, \mathbf{q}^1)$ が出力される．**図 1** にこれら 3 ステップによる提案法の学習アルゴリズムをまとめる．提案法のアルゴリズム上は，ステップ 3 終了後，ステップ 2，3 を複数回繰り返して学習を行い，さらなる解析精度向上を狙うことも可能である．この場合は，ステップ 2 へは，$p^1 = p(\mathbf{y}|\mathbf{x}; \mathbf{w}^1, \mathbf{v}^1, \mathbf{q}^1)$ を代入する．また，学習の前に事前に繰返し数を決定しておく．本論文では，特に記述がない限り繰返し数は 1（つまり 1 度目のステップ 3 終了後に学習終了）とする．

Input: 特徴ベクトル $\mathbf{f}$ の定義，学習データ $\mathcal{D} = \{\mathcal{D}_L, \mathcal{D}_U\}$，
　ただし，ラベルあり学習セット $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$，ラベルなし学習セット $\mathcal{D}_U = \{\boldsymbol{x}'_i\}_{i=1}^{M}$
Initialize: $\mathbf{q}^0 : \theta_{j,r,a} = \mu_{j,r,a} = 1/d_{j,r}$(uniform distribution)　$\forall j, r, a$
Step1: $\mathbf{q}^0$ と $\mathcal{D}_L$ を用いて $\mathbf{w}^0$ と $\mathbf{v}^0$ の推定：$(\mathbf{w}^0, \mathbf{v}^0) = \arg\max_{\mathbf{w},\mathbf{v}}\mathcal{O}_L(\mathbf{w}, \mathbf{v}|\mathcal{D}_L, \mathbf{q}^0)$
　$p^0 = p(\mathbf{y}|\mathbf{x}; \mathbf{w}^0, \mathbf{v}^0, \mathbf{q}^0)$
Step2: p^0 と $\mathcal{D}_U$ を用いて $\mathbf{q}^1$ を推定：$\mathbf{q}^1 = (q_1^1, \ldots, q_J^1),\ \forall j\ q_j^1 = \log\frac{u_{\hat{\boldsymbol{\theta}}_j}(\mathbf{x},h,m,l)}{u_{\hat{\boldsymbol{\mu}}_j}(\mathbf{x},h,m,l)}$
　$(\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\mu}}_j) = \arg\max_{\boldsymbol{\theta}_j,\boldsymbol{\mu}_j}\mathcal{O}_U(u_{\boldsymbol{\theta}_j}, u_{\boldsymbol{\mu}_j}|\mathcal{D}_U, p^0)$
Step3: $\mathbf{q}^1$ と $\mathcal{D}_L$ を用いて $\mathbf{w}^1$ と $\mathbf{v}^1$ の推定：$(\mathbf{w}^1, \mathbf{v}^1) = \arg\max_{\mathbf{w},\mathbf{v}}\mathcal{O}_L(\mathbf{w}, \mathbf{v}|\mathcal{D}_L, \mathbf{q}^1)$
Output: $(\mathbf{w}^1, \mathbf{v}^1, \mathbf{q}^1)$

図 1　一次依存条件付確率場の半教師あり学習の処理ステップ
Fig. 1　Entire parameter estimation algorithm for semi-supervised DepCRFs.

4. 関 連 研 究

半教師あり学習の研究は，機械学習の研究分野で近年さかんに研究されている研究課題の1つである．しかし，本論文が対象としている係り受け解析タスクといった自然言語処理タスクでは，高次元かつ疎な特徴空間を用いて学習を行うことや，出力が構造を持っており，離散最適化に属する解析アルゴリズムを用いて最尤出力を求める必要があるといった比較的難しいタスク設定となっている．そのため，機械学習分野で発展した半教師あり学習をそのまま適用しても，必ずしも良い効果が得られるとは限らない．そこで，自然言語処理タスクの性質を考慮した半教師あり学習法を考案し，教師あり学習の精度を大幅に上回る結果を実証した方法がいくつか報告されている．

1つ目の方法として，文献 8) で提案された補助問題を利用する方法がある．単純なテキスト分類タスクや，系列ラベリング問題で大きな効果が得られることが報告されている．しかし，この方法の利点でもあり，また難点でもある点として，補助問題は，対象タスクに合わせて人手により設計する必要があるため，効果はその定義に大きく依存することがあげられる．また，本論文の対象である係り受け解析タスクに対してどのような補助問題が適切であるかといった議論はこれまでになされておらず，今後の研究課題の1つと考えられる．

次に，文献 11)，25) で使われている単語クラスタリングを利用する方法がある．この方法は比較的簡単であり，ラベルなしデータから解きたい問題とは独立に単語クラスタリングを行い，そのクラスタリング結果を教師あり学習の特徴として利用する方法である．この手法の利点は，解きたい問題と独立に単語クラスタリングを作成するので，比較的汎用性が高いという点である．この方法に関しては，文献 25) において，係り受け解析タスクでの実証実験がすでになされており，大きな効果が得られることが示されている．本論文では，この手法を提案法の比較手法の1つとして取り上げる．

提案法は，文献 9) で提案された識別モデルと生成モデルを組み合わせたモデルを用いる方法を，文献 15) によって構造学習タスクに適用するために拡張した方法をベースとしている．この枠組みの利点は，理論的背景がしっかりしている点，人手により必要な設定が他の手法に比べて相対的に少ない点，他の自然言語処理タスクで大きな効果が得られたという実証がある点等があげられる．

5. 実　　験

2章で述べた，一次依存条件付確率場での教師あり学習による係り受け解析器と，3章で述べた，提案法による係り受け解析器の性能比較を行い，提案法の有効性を検証する．ここでは，従来法である一次依存条件付確率場での教師あり学習による係り受け解析器を‘**supervised DepCRF**’と表す．また，提案法を‘**SS-DepCRF**’と略記する．

本実験では，projective 係り受け解析の代表的な言語として英語，また non-projective 係り受け解析の代表的な言語としてチェコ語を用いて評価を行った．

5.1 デ　ー　タ

学習や評価に用いたデータには，従来研究との解析精度比較を行うため，係り受け解析の従来研究[4)–6),25)] と同じ方法で同一のデータを準備した．

英語の係り受け構造が付与されたデータは，Penn Treebank（PTB）III[26)] 中の Wall Street Journal（WSJ）セクションから獲得した．ただし，head-selection ルール[3)] に基づいて PTB III のフレーズ構造から係り受け構造に変換したデータである．変換後のデータを，セクション 02–21 を学習セット，セクション 22 を開発セット，セクション 23 を評価セットとして分割した．ラベルなし学習セット用のデータには，Brown Laboratory for Linguistic Information Processing（BLLIP）コーパスを用いた．ただし，BLLIP コーパスは PTB III WSJ セクションを含むデータであるが，PTB III WSJ セクションは，ラベルなし学習セットからは除外した．

チェコ語のデータは，Prague Dependency Treebank（PDT）1.0[27)] から獲得した．学習/開発/評価セットの分割は，コーパスに事前に割り振られているラベルに従って分割した．ラベルなし学習セットに関しては，PDT 1.0 コーパス中の正解係り受け構造が付与されていないセクションを用いた．**表 1** に実験で用いたデータセットの詳細を示す．

5.2 解析モデル

本実験では，データに即して英語は係り受け間に交差がない場合に用いる projective 係り受け解析，チェコ語は交差がある場合に用いる non-projective 係り受け解析を用いた．

表 1　実験で用いたデータセットの詳細
Table 1　Details of data sets used in our experiments.

(a) 英語係り受け解析

データセット	総文数	総単語数
ラベルあり学習セット	39,832	950,028
開発セット	1,700	40,117
評価セット	2,012	47,377
ラベルなし学習セット	1,796,379	43,380,315

(b) チェコ語係り受け解析

データセット	総文数	総単語数
ラベルあり学習セット	73,088	1,255,590
開発セット	7,507	126,030
評価セット	7,319	125,713
ラベルなし学習セット	2,349,224	39,336,570

5.3 特徴および特徴テンプレート

学習に用いた特徴は，従来研究[4),25)] に従って，品詞と単語の組合せで特徴を作成した．**表 2** に実際に用いた特徴テンプレートを示す．最終的に，30 種類の特徴を用いて学習を行った．

品詞に関しては，コーパス中の正解ではなく，実際に係り受け解析を行う状況に即して，一般的な品詞タガーを用いて，本実験で利用したすべてのデータに対して品詞を再付与した．英語では，具体的な品詞タガーに，本実験での学習セットを使って学習した MXPOST [28)] を利用した．チェコ語では，コーパスに同梱されている 'feature-based tagger' を利用した．ただし，チェコ語の品詞は種類数が多くほとんど現れない品詞が複数あるので，文献 29) の方法を用いて簡単化した品詞を用いた．

5.4 チューニングパラメータ

3 章で示したとおり，SS-DepCRF は 2 種類のチューニングパラメータ C と η を持っている．C は，第 1，3 ステップの教師あり学習時の正規化項の重みを決定する値，η は，第 2 ステップの生成モデルの推定時の正規化項の重みを決定する値である．本実験では，η に関しては，$\eta = 2$ に固定してすべての実験を行った．この直感的な解釈は，生成モデルのパラメータ推定時に各特徴が仮想的に 1 回余計にデータ中に出現したと仮定して，推定することを意味している[⋆1]．次に，C の値に関しては，開発セットを用いて最適な値を選択した．

表 2　実験に用いた特徴の種類（特徴テンプレート）
Table 2　Feature templates used in our experiments.

$[\mathrm{w}_h, \mathrm{w}_m]$, $[\mathrm{t}_h, \mathrm{t}_m]$, $[\mathrm{c}_h, \mathrm{c}_m]$, $[\mathrm{t}_h, \mathrm{t}_m, \mathrm{w}_m]$, $[\mathrm{t}_h, \mathrm{w}_h, \mathrm{t}_m]$, $[\mathrm{t}_h, \mathrm{w}_h, \mathrm{t}_m, \mathrm{w}_m]$
$[\mathrm{t}_h, \mathrm{t}_m, \mathrm{t}_{m-1}]$, $[\mathrm{t}_h, \mathrm{t}_m, \mathrm{t}_{m+1}]$, $[\mathrm{t}_h, \mathrm{t}_{h-1}, \mathrm{t}_m]$, $[\mathrm{t}_h, \mathrm{t}_{h+1}, \mathrm{t}_m]$,
$[\mathrm{t}_h, \mathrm{t}_{h-1}, \mathrm{t}_m, \mathrm{t}_{m-1}]$, $[\mathrm{t}_h, \mathrm{t}_{h-1}, \mathrm{t}_m, \mathrm{t}_{m+1}]$, $[\mathrm{t}_h, \mathrm{t}_{h+1}, \mathrm{t}_m, \mathrm{t}_{m-1}]$, $[\mathrm{t}_h, \mathrm{t}_{h+1}, \mathrm{t}_m, \mathrm{t}_{m+1}]$
$[\mathrm{c}_h, \mathrm{c}_m, \mathrm{c}_{m-1}]$, $[\mathrm{c}_h, \mathrm{c}_m, \mathrm{c}_{m+1}]$, $[\mathrm{c}_h, \mathrm{c}_{h-1}, \mathrm{c}_m]$, $[\mathrm{c}_h, \mathrm{c}_{h+1}, \mathrm{c}_m]$,
$[\mathrm{c}_h, \mathrm{c}_{h-1}, \mathrm{c}_m, \mathrm{c}_{m-1}]$, $[\mathrm{c}_h, \mathrm{c}_{h-1}, \mathrm{c}_m, \mathrm{c}_{m+1}]$, $[\mathrm{c}_h, \mathrm{c}_{h+1}, \mathrm{c}_m, \mathrm{c}_{m-1}]$, $[\mathrm{c}_h, \mathrm{c}_{h+1}, \mathrm{c}_m, \mathrm{c}_{m+1}]$
$[\mathrm{w}_h, \mathrm{w}_{m-1}]$, $[\mathrm{w}_h, \mathrm{w}_{m+1}]$, $[\mathrm{w}_{h-1}, \mathrm{w}_m]$, $[\mathrm{w}_{h+1}, \mathrm{w}_m]$
$[\mathrm{c}_h, \mathrm{c}_m, \mathrm{Betc}(h, m)]$, $[\mathrm{d}]$, $[\mathrm{I(d)}]$, $[\mathrm{c}_h, \mathrm{c}_m, \mathrm{I(d)}]$

w：単語　t：品詞　c：簡易品詞（品詞タグの前 2 文字の prefix）
$\mathrm{Betc}(h, m)$：h と m の間に出現する簡易品詞
d：係り元 h と係り先 m の距離（$\mathrm{d} = |h - m|$）
I(d)：量子化した距離（1，2〜4，5〜9，10〜19，20〜29，30〜39，40〜）

⋆1 この値が解析精度的に最適という意味ではない．現実には試行錯誤的に良い値を見つければ，本実験で示す結果より良い性能が得られる可能性がある．しかし，ここでは，チューニングパラメータ選択コストを従来法と公平にするため，このような設定を用いてる．

supervised DepCRF でも同様に式 (4) の C は開発セットを用いて最適な値を選択した．このように，supervised DepCRF と SS-DepCRF の間でチューニングパラメータ選択コストは同じに設定して実験を行った．

5.5 SS-DepCRF で用いる生成モデル

3.1 節で述べたように，提案法では，実際に用いる特徴テンプレートに基づいて機械的に生成モデルを定義した．5.3 節で示したように，本実験では 30 種類の特徴を用いて学習を行った．よって，SS-DepCRF で用いる生成モデルの数 J は 30 である．各生成モデルには，それぞれ 1 つの特徴テンプレートを割り当てる．1 つの生成モデルは，割り当てられた特徴テンプレートから生成される特徴のみで構成される．

5.6 評 価 指 標

すべての実験は，親子間の係り受けの正解率で評価した．ただし，英語係り受け解析では，句読点に関する評価を含めない方法が従来研究の主流[4),25)] なので，それに従ってここでも評価対象外とした．

本実験で示す解析精度は，開発セット，評価セットともに，開発セットで最も良い解析精度が得られたチューニングパラメータ C の値を用いて得られた結果である．

6. 実験結果および考察

表 3 に一次依存条件付確率場での教師あり学習による係り受け解析器（supervised DepCRF）と，提案法による半教師あり学習による係り受け解析器（SS-DepCRF）の実験結果を示す．なお，表 3 中の第 3 行（最下行）目は，式 (6) で定義した q を係り受けになる確率分布に基づいて推定された生成モデル $u_{\boldsymbol{\theta}}$ のみで置き換えた場合の解析精度を参考として示したものである．

評価セットの結果に対し，supervised DepCRF と SS-DepCRF 間で，文単位の paired Wilcoxon signed rank test を行ったところ，すべての組合せで p 値が 0.01 を下回った．こ

表 3　係り受け解析精度：括弧内は supervised DepCRF との差分
Table 3　Parent-prediction accuracies using the best setting in terms of development data performance.

	英語		チェコ語	
	開発セット	評価セット	開発セット	評価セット
supervised DepCRF	91.21	90.97	84.43	84.40
SS-DepCRF	91.81 (+0.60)	91.43 (+0.46)	85.00 (+0.57)	84.93 (+0.53)
（$q = \log u_{\boldsymbol{\theta}}$ とした場合）	91.66 (+0.45)	91.29 (+0.32)	84.64 (+0.21)	84.60 (+0.20)

表 4　開発，評価セットの文内単語共起集合に対するラベルあり，なし学習セットの被覆率
Table 4　Coverage rates of word co-occurrences in the development and test sets by those in labeled and unlabeled data.

	英語		チェコ語	
	開発セット	評価セット	開発セット	評価セット
ラベルあり学習セット	56.0%	53.8%	31.2%	31.7%
ラベルなし学習セット	85.6%	85.5%	64.0%	65.1%
ラベルあり+なし学習セット	86.1%	86.0%	64.4%	65.5%

の結果から SS-DepCRF は supervised DepCRF より有意水準 0.01 において統計的に有意に解析精度に差があるといえる．このように，係り受け解析の国際的な標準評価セットにおいて SS-DepCRF は supervised DepCRF より良い結果が得られることが分かった．

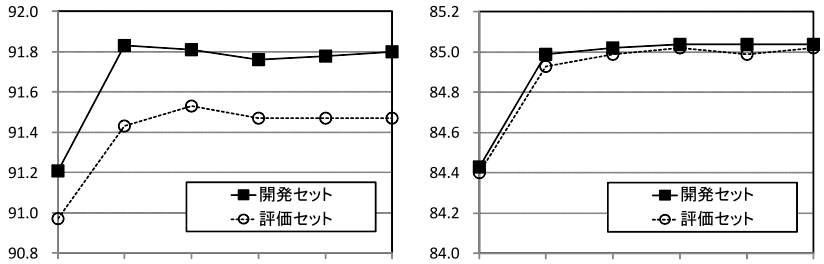
また，生成モデルを識別モデルに組み込む方法として，$u_{\boldsymbol{\theta}}$ 単体を用いる場合でも，ある程度の精度向上が得られることが分かった．しかし，3 章で述べたように，提案法のように $u_{\boldsymbol{\theta}}$ と $u_{\boldsymbol{\mu}}$ 間の対数尤度比を用いた方がさらに解析精度が向上することも確認できた．この結果は，係り受け解析では，通常の分類問題のような一般的な生成モデルにより推定した尤度よりも，出現頻度によるバイアスを消去した尤度比の方がより適していることを示す結果といえる．

次に，精度向上の裏付けとなる統計量として，開発セットまたは評価セットでの同一文中に共起する単語の集合 S，ラベルあり学習セット，ラベルなし学習セット，または，ラベルあり + なし学習セットでの同一文中に共起する単語の集合 T としたとき，S に対する T の被覆率 $|S \cap T|/|S| \times 100\%$を，**表 4** に示す．

この表から，評価セットに出現する 2 単語の依存関係の約 7 割（チェコ語），または，4 割強（英語）は，ラベルあり学習セットには出現していないことが分かる．つまり，学習後のこれらの 2 単語間に関する複数の特徴に対する重みは 0 となっており，解析時の手がかりとなる情報が不足していることを示している．これに対し，ラベルなし学習データを利用することで，被覆率を大幅に改善できていることが分かる．この改善の分だけ，従来重みが 0 であった特徴に何かしらの重みが与えられていることになる．その結果，解析時の手がかりとなる情報が増え，解析精度が向上したと考えることができる．もちろん，この被覆率の改善だけが本手法の精度向上の理由ではない．ここでは，直感的かつ統計的に示しやすい精度向上の理由の一例として示した．

6.1　Step2，3 の繰返し学習による効果

図 1 に示した提案法の学習ステップにおいて，Step2 と 3 を繰り返し学習することによる効果を**図 2** に示す．図中の横軸は繰返し数を表し，繰返しによる解析精度の変化を示している．0 は Step1 のみ，つまり教師あり学習の解析精度を表す．



(a) 英語係り受け解析　(b) チェコ語係り受け解析

図 2　提案法での学習ステップ 2，3 の繰返し学習による効果
Fig. 2　Impact of the iteration of Step 2. and 3. in our method.

この結果から，本実験では，1 回目の時点でほぼ解析精度が収束し，2 回目以降の繰返しの効果は得られなかったことが分かる．これは，1 回目と 2 回目以降で推定した生成モデルの分布がそれほど変化しなかったためだと思われる．生成モデルの分布が変化しなかった場合，提案法の教師あり学習（Step3）は凸最適化なので，必ず同じ解に収束する性質があるからである．

6.2　ラベルあり学習セットのデータ量に対する効果

従来研究で標準的に用いられている実験設定は，ラベルあり学習セットのデータ量が比較的多い状況での評価となっている．そこで，ラベルあり学習セットのデータ量が少量の場合に，SS-DepCRF の解析精度がどのように影響を受けるか検証する．

まず，実験に用いたラベルあり学習セットの部分集合を用いて，新たに小規模のラベルあり学習セット用意した．**表 5** に，作成した小規模ラベルあり学習セットの抽出基準とデータ量を示す．表中の “割合” は，元のラベルあり学習セットのデータ量に対するデータ量の割合を示している．

次に，**表 6** に，それぞれの小規模ラベルあり学習セットを用いたときの解析精度を示す．すべての小規模ラベルあり学習セットで，SS-DepCRF は，supervised DepCRF に対して解析精度を向上させることができた，このことから，ラベルあり学習セットのデータ量によらず，SS-DepCRF は，解析精度の向上が期待できる方法であることが実証できた．

また，ラベルあり学習セットのデータ量がより少ないときほど，より解析精度が向上する

表 5　学習曲線計測用のラベルあり学習セットの量と抽出方法
Table 5　Amount and extraction methods of the data sets for evaluating learning curve.

(a) 英語係り受け解析

学習セット名	総文数	総単語数	(割合)	抽出方法
ET1,671	1,671	40,039	4.2%	WSJ セクション 21
ET2,000	2,000	48,577	5.1%	ランダム
ET8,000	8,000	190,958	20.1%	ランダム
ET8,936	8,936	211,727	22.2%	WSJ セクション 15–18

(b) チェコ語係り受け解析

学習セット名	総文数	総単語数	(割合)	抽出方法
CT2,000	2,000	34,722	2.8%	ランダム
CT3,526	3,526	53,982	4.3%	c[0-9]*セクション
CT8,000	8,000	140,423	11.2%	ランダム
CT14,891	14,891	261,545	20.8%	l[a-i]*セクション

表 6　ラベルあり学習セットのデータ量の違いによる解析精度の比較
Table 6　Dependency parsing results for the SS-DepCRF and supervised DepCRF with different amounts of labeled training data.

(a) 英語係り受け解析

学習セット	ET1,671		ET2,000		ET8,000		ET8,936	
評価対象	開発	評価	開発	評価	開発	評価	開発	評価
supervised DepCRF	85.63	85.87	87.09	86.87	89.22	89.01	89.42	89.06
SS-DepCRF	87.19	87.16	88.00	87.87	90.19	89.71	90.31	89.88
差分	+1.56	+1.29	+0.91	+1.00	+0.97	+0.70	+0.89	+0.82

(b) チェコ語係り受け解析

学習セット	ET2,000		ET3,526		ET8,000		ET14,891	
評価対象	開発	評価	開発	評価	開発	評価	開発	評価
supervised DepCRF	75.67	75.07	76.88	76.70	80.61	80.39	81.94	81.76
SS-DepCRF	76.47	75.83	77.61	77.35	81.34	81.04	82.79	82.45
差分	+0.80	+0.76	+0.73	+0.65	+0.73	+0.65	+0.85	+0.69

傾向がみられた．ただし，この “ラベルあり学習セットのデータ量が少ないときほど，より性能が向上する現象” は，SS-DepCRF 特有の性質というよりは，半教師あり学習全般でよくみられる傾向である．これは，ラベルあり学習セットのデータがより少ない場合は，ラベルあり学習セットが持つ情報量がより少ないことを意味するので，ラベルなし学習セットが不足している情報をより補完できる可能性が高いからであると考えることができる．

ただし，逆に必ずしもすべての半教師あり学習法がこの性質を持つわけではない．SS-DepCRF は，教師あり学習の学習結果を再利用するような形で定義されているため，直感的には，ラベルあり学習セットのデータ量が少なく，教師あり学習で良い性能が得られないときには，性能の向上が限定される可能性が想像できた．よって，ラベルあり学習セットのデータ量が少ないときにも良好に機能することが確認できたことは意味があるといえる．このことから，SS-DepCRF は，ラベルあり学習セットのデータ量によらず十分に活用できることが分かった．

表 7　学習曲線計測用のラベルなし学習セットのデータ量
Table 7　Details of the maximum size of unlabeled data set used in English dependency parsing.

コーパス名	書誌名	期間 (mm/yy)	総文数	総単語数
BLLIP	wsj	00/87–00/89	1,796,379	43,380,315
Tipster	wsj	04/90–03/92	1,550,026	36,583,547
North	wsj	07/94–12/96	2,748,803	62,937,557
American	reu	04/94–07/96	4,773,701	110,001,109
Reuters	reu	09/96–08/97	12,969,056	214,708,766
English	afp	05/94–12/06	21,231,470	513,139,928
Gigaword	apw	11/94–12/06	46,978,725	960,733,303
	ltw	04/94–12/06	10,524,545	230,370,454
	nyt	07/94–12/06	60,752,363	1,266,531,274
	xin	01/95–12/06	12,624,835	283,579,330
合計			175,949,903	3,721,965,583

6.3　ラベルなしセットのデータ量に対する効果

次に，ラベルなし学習セットのデータ量が，SS-DepCRF の解析精度に与える影響を検証する．ただし，チェコ語に関しては，実験に用いたデータ以外に入手することができなかったため，この実験は英語係り受け解析のみで検証を行う．本実験用に準備したラベルなし学習セットの元コーパスとデータ量の詳細を**表 7** に示す．ただし，本データは，コーパス中で 128 単語以上の文を削除している．最終的に，約 1 億 8 千万文，約 37 億単語のラベルなし学習セットとなった．これは，ラベルあり学習セットのおよそ 4,000 倍のデータ量である．

図 3 に，ラベルなし学習セットのデータ量を変化させた際の SS-DepCRF の解析精度を表す．各ラベルなし学習セットは表 7 に示したデータから抽出して作成したものである．注意点として，横軸はラベルなし学習セットのデータ量の対数スケールを表している．また，一番左の点は，ラベルなし学習セットをまったく用いない場合の解析精度を示している．

興味深いことに，ラベルなし学習セットのデータ量の対数スケールに対してほぼ線形に解析精度が上昇していることが分かる．このことから，ラベルなし学習セットのデータ量は多

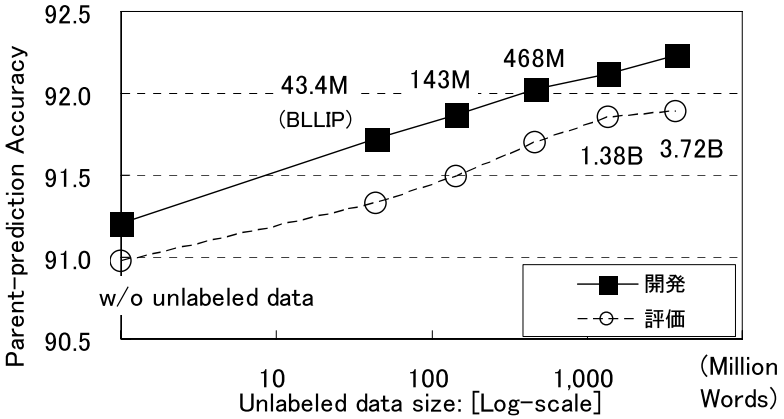


図 3　英語係り受け解析におけるラベル学習セットのデータ量に対する効果
Fig. 3　Impact of unlabeled data size for the SS-DepCRF of English dependency parsing.

ければ多いほど解析精度向上につながることが実証された．1 章で述べたように，係り受け解析タスクでは，正解データの作成コストは高いが，正解係り受けが付与されていない単なる文章は比較的容易かつ大量に獲得できる．よって，正解が分からないラベルなし学習セットでもデータ量を増やすことによって解析精度を向上させられるということは，非常に良い性質であり，SS-DepCRF の利点の 1 つといえる．

次に，図 3 から見て取れるように英語係り受け解析では，今回用いたラベルなし学習セットの最大データ量付近でも，まだデータ量の対数スケールに対し線形に解析精度が向上している．つまり，今回の実験で用いたラベルなし学習セットの最大データ量でも，まだ解析精度向上の限界点に達しておらず，ラベルなしデータをさらに増やすことによってさらに解析精度が向上する可能性がうかがえる結果となった．ただし，解析精度の伸びはデータの対数スケールに対して線形であることから，非常に大規模なデータを準備することが条件となる．

6.4　学習時の計算コスト

前節の実験で，SS-DepCRF を用いた学習は，ラベルなしデータを増加させることでさらに解析精度が向上可能であることが分かった．しかし，ラベルなしデータを増加させるということは，それだけ計算コストが増大することを意味する．SS-DepCRF の場合は，図 1 中の第 2 ステップにあたる生成モデルのパラメータ推定の計算コストが増大することを意味する（第 1，第 3 ステップはラベルなしデータを用いていないので計算コストは変わらない）．本実験で用いた実装では，2.93 GHz Xeon プロセッサ上で，4,300 万単語の BLLIP コーパスでの生成モデルの推定におよそ 5 時間かかった．SS-DepCRF の生成モデルの計算量のオーダはデータ量に対して線形なので，37 億単語のデータでは単純計算でおよそ 18 日かかることになる．

ここでは分散並列処理を利用することで，実際にかかる計算時間を削減する方法をとった．式 (11) および式 (12) の分子の計算，つまり $\hat{\theta}'_{r,a}$，$\hat{\mu}'_{r,a}$ の計算は，各データで完全に独立に処理可能である．参考までに，分母の計算は最後に 1 度だけすべてのパラメータの総和を計算すればよいので，相対的に計算時間はほとんどかからない．また提案法の必要メモリ量も，データ量に依存せず，学習に用いるパラメータベクトルと 1 入力分のメモリが確保できれば $\hat{\theta}'_{r,a}$，$\hat{\mu}'_{r,a}$ の計算が可能である．実際には，300 個のプロセッサを用いて分散並列処理を行うことで，数時間程度（実行環境により実測値は異なる）で生成モデルの推定が可能であった．

6.5　解析時の計算コスト

学習後はパラメータが固定される．また，SS-DepCRF では，一次依存条件付確率場と生成モデルで同じ特徴ベクトルを用いている．このため，式 (7) に示した判別関数から，単純な計算ですべてのパラメータを 1 つのベクトル表現に変換できることが分かる．変換されたパラメータベクトルは，式 (1) に示した supervised DepCRF で学習した際の線形判別関数と同じ形にできる．よって，解析時の計算コストは本質的に supervised DepCRF で学習したモデルと完全に一致する．前節で述べたように，SS-DepCRF は，学習時にはラベルなし学習セットのデータ量に比例して計算コストは増大する．しかし，解析時は学習時に用いたデータ量によらず一定であり，従来の教師あり識別学習で得られたモデルと同じ計算コストとなる．係り受け解析の現実的な利用状況を考えると，学習は 1 度行えばよいので，ある程度時間がかかることは許容できるが，解析速度は学習時の計算時間に比べて非常に重要な要素となるといえる．このことから，SS-DepCRF の持つ解析速度が従来法と同じという性質は，実用上非常に優れている性質といえる．

6.6　従来研究との比較

ここでは，従来研究との性能比較を行う．**表 8** に代表的な従来研究の解析精度を示す．

係り受け解析の従来研究では，MIRA や Averaged Perceptron（AP）といった，オンライン学習が用いられることが多い．これは，係り受け解析自体の計算コストが比較的大きいため，学習の速度面を重視してこのような傾向となったと考えられる．しかし精度面で

表 8　従来研究との解析精度の比較
Table 8　Comparisons with the previous systems.

(a) 英語係り受け解析

係り受け解析器	開発	評価	モデル+学習法	ラベルなしセット量
(McDonald, 2005)[4]	–	90.9	1 次依存 MIRA	–
supervised DepCRF	91.21	90.97	1 次依存 DepCRF	–
(McDonald, 2006)[6]	–	91.5	2 次依存 MIRA	–
SS-DepCRF	91.81	91.43	1 次依存 SS-DepCRF	43M 単語
	92.19	91.86	1 次依存 SS-DepCRF	3.7B 単語
(Koo, 2008)[25]	92.33	92.23	1 次依存 AP	43M 単語
SS-DepCRF +(Koo, 2008)[25]	93.01	92.70	1 次依存 SS-DepCRF	43M 単語

(b) チェコ語係り受け解析

係り受け解析器	dev	test	モデル+学習法	ラベルなしセット量
(McDonald, 2005)[5]	–	84.4	1 次依存 MIRA	–
supervised DepCRF	84.43	84.40	1 次依存 DepCRF	–
(McDonald, 2006)[6]	–	85.2	2 次依存 MIRA	–
SS-DepCRF	85.00	84.93	1 次依存 SS-DepCRF	39M 単語
(Koo, 2008)[25]	86.09	86.07	1 次依存 AP	39M 単語
SS-DepCRF+(Koo, 2008)[25]	87.03	87.14	1 次依存 SS-DepCRF	39M 単語

は，本実験が示すように，一次依存条件付確率場と一次依存 MIRA の解析精度はほぼ同等である．

オンライン学習は，学習が高速にできるといった利点がある一方，学習の終了条件の判定基準が明確ではない場合が多い点や，学習データをどの順番で用いるかにより結果が大きく異なる場合があるといった欠点があげられる．また一次依存条件付確率場は，確率モデルを用いることで厳密な条件付確率を計算できることや，最小ベイズリスクデコーディング[13]といった，確率モデルの研究で発達した技術を容易に取り入れることができるといった利点がある．このように，それぞれ一長一短あり，解析精度的にも機能的にもどちらがより優れているといえる状況ではないことが分かる．よって，使用者の要件に合わせて学習器を選択すればよいといえる．

SS-DepCRF と教師あり学習の 2 次依存モデルの比較を行うと，大規模ラベルなしセットを用いた場合には，2 次依存モデルをも上回る結果が得られている．一方，チェコ語のようにラベルなし学習セットがあまり十分に準備できなかった場合は，2 次依存モデルの解析精度に及ばない結果となった．ただし，2 次依存モデルは解析時の計算量の次数が上がり解析速度は極端に遅くなる．よって，データ量さえ準備することができれば一次依存モデルの解析速度で，2 次依存モデルに匹敵，あるいはそれを超える解析精度が得られるという SS-DepCRF の性質は，非常に意味のあることだといえる．

最後に，同じ半教師あり学習法である Koo らの方法[25] と比較すると，彼等の方法の方がより大きな解析精度向上が得られていることが分かる．彼等の方法は，簡単にいうと，ラベルなし学習セットの文章中の単語をクラスタリングし，そこで得られた単語クラスを教師あり学習の特徴として導入する方法である．係り受け解析においては，品詞的な情報が非常に有効であることが分かっており，単語クラスがちょうど品詞的な役割を担うことができることから，より解析精度向上に効果的な特徴を導入できたためと考えられる．

ただし，Koo らの方法[25] は，半教師あり学習法というよりは，単に，よりタスクに適した特徴を教師あり学習に導入した方法ととらえることができる．その意味で，新しく導入された特徴は，SS-DepCRF でも利用することが可能であり，それによってさらなる解析精度向上を見込むことができる．よって，たとえば，MIRA と DepCRF のように手法としてどちらか一方を選択しなくてはならないような排他的な方法を比較している状況ではないため，Koo らの方法[25] と SS-DepCRF の間で，解析精度での優劣をつける意味は薄いといえる．表 8 の最後の行に単純に組み合わせた際の解析精度を示す．このように，学習時間等を無視して，解析精度のみを重視する必要がある場合は，どちらか一方を選択するのではなく，これらの手法を組み合わせて利用することが可能である．

最後に，Koo らの手法に対して，SS-DepCRF の性質上の利点について述べる．まず，解析速度がより高速になる点があげられる．Koo らの実験では，単語クラスの組合せを新たな特徴として用いたため，用いた特徴の種類数が，比較対象となる教師あり学習で用いた特徴の種類数の 2 倍以上となっていた．係り受け解析では文内の単語数の 2〜3 乗オーダの解析アルゴリズムを用いているため，特徴の種類が 2 倍になるとおよそ 4〜8 倍解析速度が遅くなる．次に，人手の試行錯誤による新しい特徴の選別作業が不要である点がある．Koo らの方法では，単語クラスの組合せの選定は，基本的に人手により試行錯誤的に決定する必要がある．一方，SS-DepCRF では，従来教師あり学習で用いていた特徴をそのまま再利用しているので，追加で人手により試行錯誤する手間は不要である．

7. ま と め

本論文では，教師あり学習による一次依存条件付確率場を拡張し，半教師あり学習による係り受け解析器の学習法を提案した．提案法は，3 ステップによる学習を行う方法であり，ラベルあり学習データとラベルなし学習データを交互に情報を補完しあうような形で学習

を行う．提案法の特徴としては，ラベルなし学習データから情報を得るために，2 単語間の係り受けに対して係り受け関係になる確率とならない確率の双方をモデル化して利用している点である．また，その生成モデルの対数尤度比を判別関数，および，教師あり学習時の特徴として利用している点も特徴としてあげられる．実験では，国際的な標準評価セットにおいて，従来の教師あり学習による解析精度を大幅に上回る結果が得られた．また，提案法では，ラベルあり学習データが少ないときでも多いときでも良好に機能すること，ラベルなし学習データ量を増やすことでより精度向上がみられるといった性質があることを実験により示した．

参 考 文 献

1) Buchholz, S. and Marsi, E.: CoNLL-X Shared Task on Multilingual Dependency Parsing, *Proc. CoNLL-X*, pp.149–164 (2006).
2) Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S. and Yuret, D.: The CoNLL 2007 Shared Task on Dependency Parsing, *Proc. EMNLP-CoNLL*, pp.915–932 (2007).
3) Yamada, H. and Matsumoto, Y.: Statistical Dependency Analysis with Support Vector Machines, *Proc. IWPT* (2003).
4) McDonald, R., Crammer, K. and Pereira, F.: Online Large-margin Training of Dependency Parsers, *Proc. ACL*, pp.91–98 (2005).
5) McDonald, R., Pereira, F., Ribarov, K. and Hajič, J.: Non-projective Dependency Parsing using Spanning Tree Algorithms, *Proc. HLT-EMNLP*, pp.523–530 (2005).
6) McDonald, R. and Pereira, F.: Online Learning of Approximate Dependency Parsing Algorithms, *Proc. EACL*, pp.81–88 (2006).
7) Carreras, X.: Experiments with a Higher-Order Projective Dependency Parser, *Proc. EMNLP-CoNLL*, pp.957–961 (2007).
8) Ando, R.K. and Zhang, T.: A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data, *Journal of Machine Learning Research*, Vol.6, pp.1817–1853 (2005).
9) Fujino, A., Ueda, N. and Saito, K.: Semisupervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle, *IEEE Trans. Pattern Analysis and Machine Intelligence* (*TPAMI*), Vol.30, pp.424–437 (2008).
10) Mann, G.S. and McCallum, A.: Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data, *Journal of Machine Learning Research*, Vol.11, pp.955–984 (2010).
11) Turian, J., Ratinov, L.-A. and Bengio, Y.: Word Representations: A Simple and General Method for Semi-Supervised Learning, *Proc. ACL-2010*, pp.384–394 (2010).
12) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. ICML-2001*, pp.282–289 (2001).
13) Smith, D.A. and Smith, N.A.: Probabilistic Models of Nonprojective Dependency Trees, *Proc. EMNLP-CoNLL*, pp.132–140 (2007).
14) Koo, T., Globerson, A., Carreras, X. and Collins, M.: Structured Prediction Models via the Matrix-Tree Theorem, *Proc. EMNLP-CoNLL*, pp.141–150 (2007).
15) Suzuki, J. and Isozaki, H.: Semi-supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data, *Proc. ACL-08: HLT*, pp.665–673 (2008).
16) Eisner, J.: Three New Probabilistic Models for Dependency Parsing: An Exploration, *Proc. COLING-96*, pp.340–345 (1996).
17) Chu, Y. and Liu, T.: On the Shortest Arborescence of a Directed Graph, *Science Sinica*, Vol.14, pp.1396–1400 (1965).
18) Edmonds, J.: Optimum Branchings, *Journal of Research of the National Bureau of Standards*, Vol.71B, pp.233–240 (1967).
19) Sha, F. and Pereira, F.: Shallow Parsing with Conditional Random Fields, *Proc. HLT/NAACL-2003*, pp.213–220 (2003).
20) 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).
21) Baker, J.K.: Trainable Grammars for Speech Recognition, *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, pp.547–550 (1979).
22) Paskin, M.A.: Cubic-time Parsing and Learning Algorithms for Grammatical Bigram, Technical Report, University of California at Berkeley, Berkeley, CA, USA (2001).
23) Tutte, W.T.: *Graph Theory*, Addison-Wesley (1984).
24) McDonald, R. and Satta, G.: On the Complexity of Non-Projective Data-Driven Dependency Parsing, *Proc. IWPT*, pp.121–132 (2007).
25) Koo, T., Carreras, X. and Collins, M.: Simple Semi-supervised Dependency Parsing, *Proc. ACL-08: HLT*, pp.595–603 (2008).
26) Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A.: Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, Vol.19, No.2, pp.313–330 (1994).
27) Hajič, J.: Building a Syntactically Annotated Corpus: The Prague Dependency Treebank, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pp.12–19, Prague Karolinum, Charles University Press (1998).
28) Ratnaparkhi, A.: A Maximum Entropy Model for Part-of-Speech Tagging, *Proc.*

EMNLP, pp.133–142 (1996).
29) Collins, M., Hajic, J., Ramshaw, L. and Tillmann, C.: A Statistical Parser for Czech, *Proc. ACL*, pp.505–512 (1999).

(平成 22 年 12 月 30 日受付)
(平成 23 年 9 月 12 日採録)

鈴木　潤（正会員）

1999 年慶應義塾大学理工学部数理科学科卒業．2001 年同大学院理工学研究科計算機科学専攻修士課程修了．同年日本電信電話株式会社入社．2005 年奈良先端大学院大学博士後期課程修了．2008〜2009 年 MIT CSAIL 客員研究員．現在，NTT コミュニケーション科学基礎研究所に所属．博士（工学）．主として自然言語処理，機械学習に関する研究に従事．ACL，言語処理学会各会員．

磯崎　秀樹（正会員）

1983 年東京大学工学部計数工学科卒業．1986 年同工学系大学院修士課程修了，同年日本電信電話（株）入社．1990〜1991 年スタンフォード大学ロボティクス研究所客員研究員．2011 年退社．現在，岡山県立大学情報工学部教授．博士（工学）．平成 15 年度情報処理学会論文賞・山下記念研究賞受賞．自然言語処理の研究に従事．電子情報通信学会，人工知能学会，言語処理学会，ACL 各会員．

永田　昌明（正会員）

1987 年京都大学大学院工学研究科修士課程修了．工学博士．同年 NTT 入社．1989 年から 4 年間 ATR 自動翻訳電話研究所へ出向．1999 年から 1 年間 AT&T 研究所客員研究員．統計的自然言語処理の研究に従事．現在，NTT コミュニケーション科学基礎研究所主幹研究員．情報処理学会奨励賞（1991 年），情報処理学会論文賞（1995 年），人工知能学会研究奨励賞（1995 年）等受賞．電子情報通信学会，人工知能学会，言語処理学会，ACL 各会員．