大規模ラベルなしデータを利用した言語解析器の性能検証

鈴木 潤　　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,isozaki}@cslab.kecl.ntt.co.jp

概要

半教師あり学習法の枠組により，正解の付与されていないデータでも分類器の性能向上に貢献することが可能であることが最近いくつか報告されている．しかし，機械学習分野でも，ラベルありデータが比較的豊富にある状況下でのラベルなしデータ量に対する効果については，これまで明確に議論がなされていなかった．そこで本稿では，英語品詞タグ付け、チャンキング、固有表現抽出タスクを対象に，従来扱われてきたラベルなしデータの量を大幅に超えるギガワード (10 億単語) レベルの大規模データを用い，半教師あり学習法による性能の挙動を検証する．本稿の実験結果より，ラベルなしデータを増加することにより性能も大幅に向上することを示す。また，ラベルありデータではカバーされていない未知データに対する汎化性能の検証もおこない，ラベルありデータに出現しない単語を含む文に対しても，ラベルなしデータによりカバーすることで性能の劣化を抑えることができることも示す．

1　はじめに

近年，正解の付与された ‘ラベルありデータ’ と正解の付与されていない ‘ラベルなしデータ’ の両方を用いた学習，いわゆる半教師あり学習が機械学習分野のみならず自然言語処理や画像処理といった応用分野でも扱われるようになってきた．自然言語処理分野の多くのタスクでは，生テキストがラベルなしデータとなる．よって，ラベルなしデータを大量かつ容易に獲得することが可能である．つまり，自然言語処理分野の多くのタスクは半教師あり学習に適したタスクと考えられ，汎用的でかつ性能の良い半教師あり学習法の開発が望まれる．

一般的に，学習に用いるデータは多い方が良いと言われている．ラベルありデータに関しては，この仮定が成り立つことは明確であるが，ラベルなしデータに関しては厳密には良くわからない部分もある．特に，本稿で対象とする品詞タグ付け，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，フリーで使えるテストコレクションとして数万語レベル以上の大規模なラベルありデータが公開されている．機械学習の分野では，ラベルありデータがごくわずかしかない設定での半教師あり学習法の実験がほとんどで，本稿で扱うようなラベルありデータが比較的豊富にある状況でのラベルなしデータの効果はほとんど議論されてこなかった．また，現状では，これらの大規模テストコレクションに対しては，半教師あり学習を用いて教師あり学習で得られた最高性能を大幅に向上させたという報告はあまりみられない [1, 2]．これは，ラベルありデータが豊富に利用可能である状況では，ラベルありデータに汎化性能の良い分類器を学習するために有効な情報が豊富に含まれているため，ラベルなしデータから追加の有効な情報を得るのが困難であるためと考えられる．そこで，本稿では，これらラベルありデータが比較的豊富な大規模テストコレクションに対して，半教師あり学習でのラベルなしデータの効果について検証と議論をおこなう．

まずはじめに，大規模ラベルなしデータを扱うことができるように，ラベルなしデータに対する計算量が少ないスケーラブルな半教師あり学習法を提案する．次に，ラベルありデータの 1000 倍から 5000 倍になるギガワード (10 億単語) レベルの大規模ラベルなしデータを用意し，ラベルなしデータ量に対する性能の挙動について検証をおこなう．また同時に，ラベルありデータではカバーされていない未知データに対する汎化性能の検証もおこなう．

2　条件付確率モデルによる半教師あり学習

本稿では，従来の教師あり学習における条件付確率場 [3](CRF) に対して，半教師あり学習を可能とする自然な拡張をおこなうことで，あらたな半教師あり学習法を定義する．

2.1　条件付確率場 (CRF)

$\mathcal{X}$ および $\mathcal{Y}$ を，それぞれ全ての可能な入力サンプル，および，出力ラベルの集合とする．$\boldsymbol{x} \in \mathcal{X}$ を (構造付きの) 入力サンプル，$\boldsymbol{y} \in \mathcal{Y}$ を (構造付きの) 出力ラベルとする．$\mathcal{C}$ を無向グラフ (グラフィカルモデル)$\mathcal{G}(\boldsymbol{x}, \boldsymbol{y})$ 中のクリークの集合とする．$\boldsymbol{y}_c$ を対応するクリーク c から出力されるラベルとする．また，各 $c \in \mathcal{C}$ には，**ポテンシャル関数** Ψ_c が定義されるとする．このとき，CRF は条件付確率 $p(\boldsymbol{y}|\boldsymbol{x})$ を Ψ_c の対数線形の形式で定義する．ここで，$\boldsymbol{f} = (f_1, \ldots, f_I)$ を特徴ベクトル，$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_I)$ をパラメタベクトルとすると，CRF 上の条件付確率 $p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\lambda})$ は以下のように定義される．

$$p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{x})} \prod_{c \in C} \Psi_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}) \qquad (1)$$

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．

$\boldsymbol{f}_c(\boldsymbol{y}_c, \boldsymbol{x})$ を $\mathcal{G}(\boldsymbol{x}, \boldsymbol{y})$ 中の対応するクリーク c から得られる特徴ベクトルとする．ポテンシャル関数の構成要件として，非負関数である必要がある．ゆえに，現在では，指数関数，つまり，

$$\Psi_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}) = \exp(\boldsymbol{\lambda} \cdot \boldsymbol{f}_c(\boldsymbol{y}_c, \boldsymbol{x}))$$

が一般的に広く利用されている．

2.2　CRF の半教師あり学習拡張

まず，J 種の確率モデルを仮定する．j 番目の (同時) 確率モデルを $p_j(\boldsymbol{x}_j, \boldsymbol{y}; \boldsymbol{\theta}_j)$ とする．ただし，$\boldsymbol{\theta}_j$ はモデルパラメタを表し，$\boldsymbol{x}_j = \mathcal{T}_j(\boldsymbol{x})$ は，$\boldsymbol{x}$ から事前に定義された関数 $\mathcal{T}_j$ により抽出された入力サンプルとする．このとき，x_j は x と同じ (グラフ) 構造を持つこととする．つまり，$p_j(\boldsymbol{x}_j, \boldsymbol{y})$ は，$\mathcal{G}(\boldsymbol{x}, \boldsymbol{y})$ 中のクリーク c によって x 同様に分割できることを意味し，$p_j(\boldsymbol{x}_j, \boldsymbol{y}; \boldsymbol{\theta}_j) = \prod_c p_j(\boldsymbol{x}_{jc}, \boldsymbol{y}_c; \boldsymbol{\theta}_j)$ と書くことができると仮定する．

Input: training data $\mathcal{D} = \{\mathcal{D}_l, \mathcal{D}_u\}$
where labeled data $\mathcal{D}_l = \{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$,
and unlabeled data $\mathcal{D}_u = \{\boldsymbol{x}^m\}_{m=1}^M$
Initialize: $\boldsymbol{\Theta}^{(0)} \leftarrow$ uniform distribution, $t \leftarrow 1$
do
1. (Re)estimate $\boldsymbol{\lambda}'$:
maximize $\mathcal{L}^1(\boldsymbol{\lambda}'|\boldsymbol{\Theta})$ with fixed $\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta}^{(t-1)}$ using $\mathcal{D}_l$.
2. Estimate $\boldsymbol{\Theta}^{(t)}$: (Initial values = $\boldsymbol{\Theta}^{(t-1)}$)
maximize $\mathcal{L}^2(\boldsymbol{\Theta}|\boldsymbol{\lambda}')$ with fixed $\boldsymbol{\lambda}'$ using $\mathcal{D}_u$.

do_until $\frac{|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^{(t-1)}|}{|\boldsymbol{\Theta}^{(t-1)}|} < \epsilon$.
Reestimate $\boldsymbol{\lambda}'$: perform the same procedure as 1.
Output: a JESS-CM, $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\lambda}', \boldsymbol{\Theta}^{(t)})$.

図 1: JESS-CM による半教師あり学習のパラメタ推定法

ここで，特徴ベクトル $\boldsymbol{f}$ と p_j の対数尤度をつなげたベクトル $\boldsymbol{h}=(f_1, \ldots, f_I, \log p_1, \ldots, \log p_J)$ と，それに対応するパラメタベクトル $\boldsymbol{\lambda}'=(\lambda_1, \ldots, \lambda_I, \lambda_{I+1}, \ldots, \lambda_{I+J})$ を導入する．このとき，同時確率モデルを埋め込む形式で新たなポテンシャル関数を定義できる．

$$\begin{aligned}\Psi'_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) &= \exp(\boldsymbol{\lambda} \cdot \boldsymbol{f}_c(\boldsymbol{y}_c, \boldsymbol{x})) \cdot \prod\nolimits_j p_j(\boldsymbol{x}_{jc}, \boldsymbol{y}_c; \boldsymbol{\theta}_j)^{\lambda_{I+j}} \\ &= \exp(\boldsymbol{\lambda}' \cdot \boldsymbol{h}(\boldsymbol{y}_c, \boldsymbol{x}))\end{aligned}$$

ただし，$\boldsymbol{\Theta}=\{\boldsymbol{\theta}_j\}_{j=1}^J$ とする．

各 $p_j(\boldsymbol{x}_{jc}, \boldsymbol{y}_c)$ の値域は $[0, 1]$ であり，非負関数なので，Ψ'_c もまたポテンシャル関数としての条件を満たしている．ゆえに，半教師あり学習用の条件付確率モデルは以下のように定義できる．

$$P(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) = \frac{1}{Z'(\boldsymbol{x})} \prod\nolimits_{c \in C} \Psi'_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) \qquad (2)$$

ただし，$Z'(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \prod_{c \in \mathcal{C}} \Psi'_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta})$ である．本稿では，式 (2) で定義される半教師あり学習用の条件付確率モデルを‘*Joint probability model Embedding style Semi-Supervised Conditional Model*’ (**JESS-CM**) と呼ぶ．

ラベルありデータ $\mathcal{D}_l=\{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$ が与えられたとき，$\boldsymbol{\Theta}$ を固定した条件下での $\boldsymbol{\lambda}'$ の MAP 推定は以下の式で書ける．

$$\mathcal{L}^1(\boldsymbol{\lambda}'|\boldsymbol{\Theta}) = \sum_n \log P(\boldsymbol{y}^n|\boldsymbol{x}^n; \boldsymbol{\lambda}', \boldsymbol{\Theta}) + \log p(\boldsymbol{\lambda}'),$$

ただし，$p(\boldsymbol{\lambda}')$ は $\boldsymbol{\lambda}'$ の事前確率分布を表す．式 (2) と式 (1) の比較から，二つの式は利用しているポテンシャル関数以外は全く同じ型であることがわかる．よって，従来の教師あり学習の CRF で用いられる勾配ベースの最適化法とその勾配を求めるために用いられる forward-backward アルゴリズムを用いてパラメタ $\boldsymbol{\lambda}'$ の推定をおこなうことが可能である．また，$\boldsymbol{\Theta}$ を固定した条件下で $\boldsymbol{\lambda}'$ の大域的最適解が得られることも同様に保障される．

文献 [2] によりラベルなしデータを効果的に取り込むことが可能なパラメタ推定法が提案されている．本稿では，その手法を‘*Maximum Discriminant Functions sum*’ (**MDF**) 推定法として参照する．MDF 推定法は，識別関数 g の総和を最大化するようにパラメタ推定をおこなう方法である．よって，ラベルなしデータ $\mathcal{D}_u=\{\boldsymbol{x}^m\}_{m=1}^M$ が与えられた時に，JESS-CM のパラメタ $\boldsymbol{\Theta}$ の推定に MDF 推定を適用すると，パラメタ $\boldsymbol{\lambda}'$ を固定した条件下で以下の目的関数を最大化することになる．

$$\mathcal{L}^2(\boldsymbol{\Theta}|\boldsymbol{\lambda}') = \sum_m \log \sum_{\boldsymbol{y} \in \mathcal{Y}} g(\boldsymbol{x}^m, \boldsymbol{y}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) + \log p(\boldsymbol{\Theta})$$

data	学習	開発	評価
品詞タグ付け: (PTB III データ中の WSJ)			
ラベルの種類数			45
総文数	38,219	5,527	5,462
総単語数	912,344	131,768	129,654
(WSJ セクション ID)	0–18	19–21	22–24
チャンキング: (PTB III データ中の WSJ: CoNLL'00 データ)			
ラベルの種類数			23 (w/ IOB-tagging)
総文数	8,936	N/A	2,012
総単語数	211,727	N/A	47,377
(WSJ セクション ID)	15–18	N/A	20
固有表現抽出: (Reuters Corpus2: CoNLL'03 データ)			
ラベルの種類数			9 (w/ IOB-tagging)
総文数	14,987	3,466	3,684
総単語数	203,621	51,362	46,435
(記事の期間)	22–30/08/96	30–31/08/96	06–07/12/96

表 1: ラベルあり学習，開発，評価データ

data	abbr.	記事の期間	総文数	総単語数
Reuters	reu	09/96–08/97*	14,362,393	225,017,761
Corpus2		*(excluding 06–07/12/96)		
English	afp	05/94–12/96	5,510,730	135,041,450
Gigaword	apw	11/94–12/96	7,207,790	154,024,679
	ltw	04/94–04/97	3,484,709	82,096,565
	nyt	07/94–12/96	15,977,991	357,952,297
	xin	01/95–04/97	2,046,725	47,280,370
合計	all		48,590,338	1,001,413,122

表 2: ラベルなしデータ

ただし，$p(\boldsymbol{\Theta})$ は $\boldsymbol{\Theta}$ の事前確率分布を表す．JESS-CM の識別関数は，$g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) = \prod_{c \in \mathcal{C}} \Psi'_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta})$ と定義できる．

最終的に，実際の学習時には，$\mathcal{L}^1(\boldsymbol{\lambda}'|\boldsymbol{\Theta})$ と $\mathcal{L}^2(\boldsymbol{\Theta}|\boldsymbol{\lambda}')$ 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，学習アルゴリズムの概略を示す．

3 実験の設定

本稿の実験では，英語品詞タグ付け，チャンキング，固有表現抽出タスクを用いて性能評価をおこなう．

3.1 データセット

品詞タグ付けには，Penn TreeBank III に含まれる Wall Street Journal データを文献 [4] と同じデータ分割で用いて実験をおこなった．また，チャンキングと固有表現抽出実験には，それぞれ CoNLL'00 と'03 の共有タスクで用いられたデータを利用した．表 1 に，学習，開発，評価データの詳細を示す．

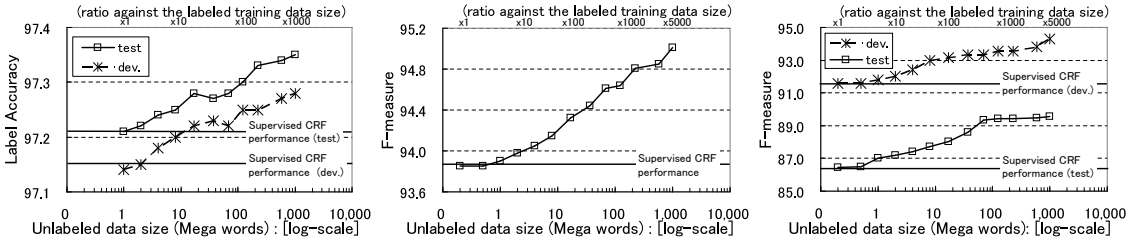
ラベルなしデータには，Reuters Corpus2 と English Gigaword, third edition (LDC2007T07) 中の新聞記事を利用した．ラベルなしデータの詳細を表 2 に示す．注目すべき点として，ラベルなしデータは約 5000 万文，10 億単語もの大規模データであり，ラベルあり学習データの 1000 倍から 5000 倍にものぼる点である．

3.2 JESS-CM の設計

本稿の実験では，JESS-CM の設計として linear chain CRF と同じグラフ構造を用いる．用いた特徴集合としては，利用するテストコレクションで従来よく用いられている特徴のみを利用した．ただし，品詞タグ付けと固有表現抽出では，文献 [5] で利用されている正規表現以外の単語タイプを追加情報として用いた．しかし，固有表現抽出で

	(a) 品詞タグ付け				(b) チャンキング		(c) 固有表現抽出			
評価尺度	ラベル正解率		文正解率		$F_{\beta=1}$ 値	文正解率	$F_{\beta=1}$ 値		文正解率	
data	開発	評価	開発	評価	評価	評価	開発	評価	開発	評価
JESS-CM	97.28	97.35	55.53	56.81	95.01	64.46	94.31	89.58	91.00	84.66
(教師あり CRF からの性能変化)	(+0.13)	(+0.14)	(+1.68)	(+1.74)	(+1.13)	(+4.42)	(+2.74)	(+3.14)	(+4.16)	(+3.47)

表 3: 品詞タグ付け (PTB III データ)，チャンキング (CoNLL'00 データ)，固有表現抽出 (CoNLL'03 データ) における JESS-CM の性能評価および教師あり学習 (CRF) との性能比較



(a) 品詞タグ付け　(b) チャンキング　(c) 固有表現抽出

図 2: ラベルなしデータ量に対する JESS-CM の学習曲線

しばしば用いられるような人名，地名辞書のような外部資源は利用していない．本稿で用いた特徴は全て与えられたラベルあり学習データから自動的に抽出できるものだけである．

次に，同時確率モデルとしては，linear chain CRF と同じグラフ構造を持つ一次隠れマルコフモデル (HMM) を用いた．また，HMM の特徴集合には，JESS-CM で用いた特徴と全く同じものを用いた．ただし，一つの HMM に対して一種類の特徴 (単語 or 品詞など) を割り振り，用いた特徴の種類に応じた数の HMM を導入した．

4 実験結果および考察

4.1 性能に対するラベルなしデータ量の効果

表 3 は，10 億単語のラベルなしデータを用いた際の JESS-CM の性能評価と，教師あり学習での CRF との性能比較の結果である．教師あり学習 (CRF) に対する提案法の大幅な性能の向上は，ラベルなしデータを利用したことのみで得られたものである．つまり，提案法では，単純かつ良く知られた CRF と HMM の組合せのようなモデルで，辞書等の外部資源，複雑なモデルの導入，人手による特徴の選別，タスク依存の人間の知識といったコストを一切かけずに，大幅な性能向上を得ることができたことを示している．

図 2 にラベルなしデータ量に対する JESS-CM の学習曲線を示す．ただし，x 軸は 100 万単語単位 (Mega-word) でかつ log スケールで示されている．また，上の x 軸はラベルあり学習データに対するラベルなしデータの比率を表している．図から，ラベルなしデータ量を多くすることによって，性能が向上することが読み取れる．また，その上昇は log スケールに対しておおよそ線形に増加しており，さらにラベルなしデータを増加させることによって性能の向上が見込めることを示している．

4.2 未知データに対する汎化性能の検証

ここでは，ラベルあり学習データではカバーされていない未知データに対する汎化性能について検証をおこなう．まず．開発および評価データを二つの文集合に分割する．一つ目の文集合は，文中に出現する単語全てがラベルあり学習データに出現した単語で構成されている文の集合とする (**L.cov**)．もう一方は，それ以外の文であり，文中の単語の少なくとも一つがラベルあり学習データに出現しなかった単語を含む文の集合とする (**L.¬cov**)．つまり，ラベルあり学習データからみて未知の単語を含む文の集合ということもできる．

			文正解率			
data		文の比率	CRF	JESS-CM (変化)		U.cov
品詞タグ付け	L.¬cov	**46.1%**	45.91	48.43	**(+2.51)**	81.0%
(開発)	L.cov	53.9%	60.62	61.40	(+0.78)	95.9%
品詞タグ付け	L.¬cov	**40.4%**	48.57	50.07	**(+1.49)**	80.3%
(評価)	L.cov	59.6%	59.48	61.02	(+1.53)	96.3%
チャンキング	L.¬cov	**70.7%**	56.71	61.56	**(+4.85)**	82.5%
(評価)	L.cov	29.3%	68.08	71.14	(+3.06)	93.2%
固有表現抽出	L.¬cov	**54.3%**	78.32	85.65	**(+7.33)**	93.9%
(開発)	L.cov	45.7%	96.97	97.35	(+0.38)	99.6%
固有表現抽出	L.¬cov	**64.3%**	75.39	80.12	**(+4.73)**	93.4%
(評価)	L.cov	35.7%	91.63	92.85	(+1.22)	100%

表 4: L.cov and L.¬cov 文集合に対する性能比較

表 4 に，これら二つの文集合に対する教師あり学習の CRF と 10 億単語のラベルなしデータを用いた JESS-CM の文正解率を示す．表中の (**U.cov**) は，(**L.cov**) と同様に，ラベルなしデータに出現した単語のみで構成された文の比率を示す．表 4 より，教師あり学習時では，L.¬cov に対する文正解率は L.cov と比べて明らかに低いことが見て取れる．これは従来から知られているように，教師あり学習では，ラベルあり学習データに出現しない情報 (未知単語) に対する汎化性能は既知情報に対する性能と比較して相対的に良くないことを示している．一方，半教師あり学習の設定では，ラベルなしデータを導入することで L.¬cov の性能を大幅に向上させることができたことを示している．これは，ラベルあり学習データに対しては未知単語である単語も，ラベルなしデータでカバーすることによって正解する可能性が向上したと捉えることができる．

自然言語解析をおこなうタガーやチャンカーの実用上の設定を考えると，全ての入力文がラベルあり学習データに含まれる単語だけで構成されるとは考えづらい．むしろ，ラベルあり学習データがカバーする単語や文脈は，実用から考えると限りなく少なく，実際の利用時にはテストコレクション中の評価データで得られる性能程のよい結果は得られていないと考えられる．つまり，教師あり学習の設定では，汎用的なタガーやチャンカーという観点では，不十

system	開発	評価	外部リソース
JESS-CM	**97.28**	**97.35**	1G-word unlab. data
(Shen+, 2007)[4]	**97.28**	97.33	–
(Toutanova+, 2003)[6]	97.15	97.24	crude company name det.
[教師あり CRF (baseline)]	97.15	97.21	–

表 5: PTBIII データでの品詞タグ付け実験のトップシステムとの性能比較 (ラベル正解率)

system	評価	外部リソース
JESS-CM	**95.01**	1G-word unlabeled data
	94.67	15M-word unlabeled data
(Ando+, 2005)[1]	94.39	15M-word unlabeled data
(Suzuki+, 2007)[2]	94.36	17M-word unlabeled data
(Zhang+, 2002)[7]	94.17	full parser output
(Kudo+, 2001)[8]	93.91	–
[教師あり CRF (baseline)]	93.88	–

表 6: CoNLL'00 データでのチャンキング実験のトップシステムとの性能比較 (F 値)

分な性能しか得られないと言える．一方，半教師あり学習の設定では，ラベルなしデータによりラベルあり学習データに未出現の単語をカバーすることによって，汎化性能の向上が得られることが今回の実験結果により示された．この結果は，一般的な議論として，半教師あり学習のほうが教師あり学習よりも，より汎用的なシステム構築が可能であることを示している．つまり，ラベルありデータを大量に獲得することが困難な通常の状況でより汎用的なシステムの構築を目的とする場合は，大規模ラベルなしデータを効果的に取り込んだ半教師あり学習を利用した方が，より高い汎化性能が得られると考えられる．

4.3 既存のトップシステムとの性能比較

表 5 に示したように，PTB III データでの品詞タグ付け実験では，これまでに文献 [4] のシステムが最高の成績を示していた．表からもわかるように，我々のシステムでも文献 [4] と同等か若干良い結果が得られた．文献 [4] のシステムは，ラベルありデータのみの教師あり学習で学習されているが，系列データに対してデコード順序とラベル付与を同時に学習する非常に複雑なモデルとなっている．一方，JESS-CM のモデルは単純な一次マルコフモデルである．基本的にこのモデル構造の差が教師あり学習の設定での性能の差となっている．ゆえに，逆の見方をすると，JESS-CM は単純なモデルで最新の研究成果として得られた複雑なモデルと同等の性能を，ラベルなしデータを活用することで得ることができたと言うこともできる．

次に，表 6 に，CoNLL'00 データによるチャンキング実験，表 7 に，CoNLL'03 データによる固有表現抽出実験のトップシステムの成績をそれぞれ示す．これらの実験では，これまで文献 [1] のシステムが最も良い成績を示していた．文献 [1] のシステムは，本稿のシステムと同様に半教師あり学習を用いたシステムである．ただし，文献 [1] では，補助問題を定義し，その補助問題から対象タスクの学習に適した特徴を得るという半教師あり学習法を提案し用いている．よって，ラベルなしデータは，その補助問題の学習する際に利用される．

同じラベルなしデータ量の比較では，チャンキングでは JESS-CM の結果が大幅に良いのに対して固有表現抽出では若干下回る結果となった．この理由のひとつとして，文献 [1] では，補助問題を構成する際に，'固有表現は主に名詞と形容詞で形成される' という人間の知識を利用するこ

system	開発	評価	外部リソース
JESS-CM	**94.31**	**89.58**	1G-word unlab. data
	93.32	**89.33**	62M-word unlab. data
(Ando+, 2005)[1]	93.15	89.31	27M-word unlab. data
(Florian+, 2003)[9]	93.87	88.76	own large gazetteers, 2M-word labeled data
(Suzuki+, 2007)[2]	N/A	88.41	27M-word unlab. data
[教師あり CRF (baseline)]	91.54	86.44	–

表 7: CoNLL'03 データでの固有表現抽出実験のトップシステムとの性能比較 (F 値)

とで性能の向上を得ているという点が挙げられる．一方，提案法ではこのような知識を用いる必要はない．

計算量の観点でも，提案法は，ラベルなしデータ量に対して線形オーダーの手法であり，大規模ラベルなしデータでも容易に扱うことが可能である．一方，文献 [1] では，ラベルなしデータを用いて多数の補助問題を学習する必要があるため計算量は相対的に大きくなる．表 7 に示したように，固有表現抽出実験では，およそ 2.3 倍のラベルなしデータ量で文献 [1] と同等の成績が得られた．しかし，ラベルなしデータが大量に得られる状況では，ラベルなしデータ量が同じ時の性能を比較するより，同じ計算量に対して性能を比較する方が学習法の適切な比較になると考えることもできる．最終的に，10 億単語のラベルなしデータを利用した JESS-CM は，従来の最高の成績を大幅に上回る成績を示した．

5 まとめ

本稿では，大規模ラベルなしデータでも扱うことが可能なスケーラブルな半教師あり学習法を提案した．提案法は，大規模ラベルなしデータを用いることで，機械学習法の性能評価に広く利用されている大規模テストコレクションを用いた英語品詞タグ付け，チャンキング，固有表現抽出タスク全てにおいて，既存のシステムとして最高の成績を示した．本稿の実験結果より，これらのタスクでも，ラベルなしデータ量の増加が半教師あり学習の性能向上につながることを実証した．また，ラベルあり学習データではカバーされていない未知データに対する汎化性能も，ラベルなしデータによりカバーすることにより向上することを示した．

参考文献

[1] Ando, R. and Zhang, T.: A High-Performance Semi-Supervised Learning Method for Text Chunking, *Proc. of ACL-2005*, pp. 1–9 (2005).

[2] Suzuki, J., Fujino, A. and Isozaki, H.: Semi-Supervised Structured Output Learning Based on a Hybrid Generative and Discriminative Approach, *Proc. of EMNLP-CoNLL*, pp. 791–800 (2007).

[3] 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).

[4] Shen, L., Satta, G. and Joshi, A.: Guided Learning for Bidirectional Sequence Classification, *Proc. of ACL-2007*, pp. 760–767 (2007).

[5] Sutton, C., Sindelar, M. and McCallum, A.: Reducing Weight Undertraining in Structured Discriminative Learning, *Proc. of HTL-NAACL 2006*, pp. 89–95 (2006).

[6] Toutanova, K., Klein, D., Manning, C. and Yoram Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network, *Proc. of HLT-NAACL-2003*, pp. 252–259 (2003).

[7] Zhang, T., Damerau, F. and Johnson, D.: Text Chunking based on a Generalization of Winnow, *Machine Learning Research*, Vol. 2, pp. 615–637 (2002).

[8] Kudo, T. and Matsumoto, Y.: Chunking with Support Vector Machines, *Proc. of NAACL 2001*, pp. 192–199 (2001).

[9] Florian, R., Ittycheriah, A., Jing, H. and Zhang, T.: Named Entity Recognition through Classifier Combination, *Proc. of CoNLL-2003*, pp. 168–171 (2003).