教師なし隠れマルコフモデルを利用した 最大エントロピータグ付けモデル

風間　淳一[†]　　宮尾　祐介[††]　　辻井　潤一[†††]

本論文では, 教師なし学習によって推定された隠れマルコフモデル (HMM) の隠れ状態を最大エントロピー (ME) モデルの素性として利用するタグ付けモデルを提案する. 教師なし学習された確率モデルを本手法に従って利用することにより, タグ付きコーパスが少ない状況でのタグ付け器作成コストを削減することが可能となる. 実験では, 英語品詞タグ付けと日本語の単語分割を対象として, 少量のタグ付きコーパスで学習する場合の精度が本手法により改善されることを示し, 提案手法がタグ付け器作成のコスト削減に寄与することを実証する. さらに, 英語品詞タグ付けでタグ付きコーパスを最大限利用できる場合には, 最高水準の精度 (96.84%) を達成し, 品詞タグ付けモデルとしても優れていることを示す.

キーワード:　タグ付け,　最大エントロピー法,　教師なし学習,　隠れマルコフモデル

A Maximum Entropy Tagging Model with Unsupervised Hidden Markov Models

Jun'ichi Kazama[†] , Yusuke Miyao[††] and Jun'ichi Tsujii[†††]

We describe a new tagging model where the states of a hidden Markov model (HMM) estimated by unsupervised learning are incorporated as the features in a maximum entropy model. Our method for exploiting unsupervised learning of a probabilistic model can reduce the cost of building taggers with a small annotated corpus. Experimental results on English POS tagging and Japanese word segmentation show that our method greatly improves the tagging accuracy when the model is trained with a small annotated corpus. Furthermore, our English POS tagger achieved a state-of-the-art POS tagging accuracy (96.84%) when a large annotated corpus is available.

KeyWords:　*Tagging, Maximum Entorpy Method, Unsupervised Learning, Hidden Markov Model*

1　はじめに

近年, 品詞タグ付け・チャンク同定・固有表現認識などの様々なタグ付けタスクに対して統計的機械学習手法が高い性能を示すことが明らかになっている (Brill 1994; Ratnaparkhi 1996;

† 北陸先端科学技術大学院大学　情報科学研究科, School of Computer Science, Japan Advanced Institute of Science and Technology
†† 東京大学大学院　情報理工学系研究科, Graduate School of Information Science and Technology, University of Tokyo
††† 東京大学大学院　情報理工学系研究科, Graduate School of Information Science and Technology, University of Tokyo　科学技術振興機構 CREST, CREST Japan Science and Technology Agency

Brants 2000; Kudoh and Matsumoto 2000; Borthwick 1999). これら既存研究の多くは学習のために大量のタグ付きコーパスや辞書の存在を仮定しているが, 実際に新しい分野や新しいタスクに対してタグ付け器を作成する場合, 大量のタグ付きコーパスやタスク専用の辞書（単語等の単位に対し可能なタグを列挙したもの）が存在しないという問題が頻繁に起こる. 従って, 大量のタグ付きコーパスや辞書が利用できる時に高い精度を達成する手法だけではなく, 大量のタグ付きコーパスや辞書を利用できない時にタグ付け器を作成するコストを削減するための技術が強く求められている.

本研究では, 新しい分野や新しいタスクに対して以下のような開発過程によりタグ付け器を開発することを想定し, 開発コストを削減することを目標とする.

1. その時点で利用できるタグ付きコーパスを用いてタグ付け器を教師あり学習する.
2. 学習されたタグ付け器を使用してタグなしテキストへタグを付与する.
3. 付与されたタグの誤りを人手により修正し, タグ付きコーパスへ追加する. 1 へ.

上のような過程においては, 利用できるタグ付きコーパスが少ない段階から高い精度を示すモデルがコストの削減に大きく寄与する. 同時に, 最終的に大量のタグ付きコーパスが得られた時には, 実用に十分な高精度を達成できるタグ付け器が得られることが望ましい.

本論文では, タグ付け器作成のコストを削減するために, 教師なし学習により推定された隠れマルコフモデル (HMM) の隠れ状態を最大エントロピー法の素性として利用する新しいタグ付けモデルを提案する. 隠れマルコフモデルの教師なし学習は, タグ付きコーパスなしで信頼性の高い言語モデルを与える. 一方, 最大エントロピー (ME) モデル (Berger,Della Pietra and Della Pietra 1996) はタグ付けタスクにおいて最高精度を達成するのに十分なモデル化能力をもつと考えられる (Ratnaparkhi 1996). これら二つのモデルを組み合わせることにより, （1）少量のタグ付きコーパスしか利用できない時に, 比較的高い精度を示す, （2）最終的に, 大量のタグ付きコーパスがあるときには, 知られている最高性能に匹敵する精度を示す, という性質をもつタグ付けモデルが得られる.

最大エントロピーモデルに教師なし HMM を組み合わせる我々のモデルの構成は, タグ付け器の学習には以下に挙げる**タスク学習**と**分野適応**という二つの側面があり, そのうち分野適応に関しては教師なし学習が利用できるという考えに基づいている.

タスク学習　どのような種類のタグ付け器か. (例: 品詞タグ付けか, 句構造チャンキングか)

分野適応　どのような種類のテキストを対象とするか. (例: 新聞記事か, ウェブテキストか. 英語か, 日本語か)

タスク学習に関しては, 観測される事象（単語列など）と観測し得ない事象（正しいタグの列）の関係の学習が必要であるから, タグ付きコーパスや辞書の利用は避けられない. 一方, 分野適応については, タグ付きコーパスを使わない教師なしの学習を利用できる可能性がある. 例えば, n-gram モデルや HMM などの確率モデルは, 分野のタグなしコーパスから特定の単語の出現や共起のパターンを捉えるように学習をすることができる. そのように教師なし学習で分野に適応した確率モデルを, タスク学習の際の基礎的な情報源として利用することで, タグ付きコーパ

スが少ないときの精度を改善することが可能になると考えられる. 別のいい方をすれば, タグ付け器の精度のいくらかはタスクとは関係のない分野適応の性能によっており, その部分はタグ付きコーパスなしに分野の生のテキストのみから学習できるという考え方である. 我々のモデルは, 分野適応のために隠れマルコフモデルの教師なし学習を利用し, タスク学習のために最大エントロピーモデルの教師あり学習を利用するモデルであり, 隠れマルコフモデルの状態遷移列を最大エントロピーモデルの素性とすることにより, 隠れマルコフモデルが上述の基礎的な情報源となることを意図したものである.

また, 少量のタグ付きコーパスで高精度を示すことは, データスパースネスに対して頑健であるということを意味するから, 我々のモデルにおける隠れマルコフモデルは, スムージングに相当し, より信頼性の高い素性を最大エントロピーモデルに与えているとも捉えることができる.

分野適応のためにどの確率モデルを用いるかは簡単な問題ではないが, 本研究では，少なくとも（1） 教師なし学習法が確立していること，（2）最大エントロピー法と組み合わせるために何らかの「状態」と解釈できる情報を持っていること, という条件を満たすものとして隠れマルコフモデル (HMM) を選択した. HMM 自体が様々なタグ付けタスクで利用され，有効性が示されていることも理由の一つである. HMM を用いたタグ付けモデルでは, 通常 HMM の隠れ状態をタグと同一視するが, そのことからも, 教師なし学習で得られるモデルの隠れ状態が, タスクのタグに近くなることも十分考えられる. そのような場合にはタスク適応は非常に簡単になり, 必要なタグ付きコーパスの量を大きく削減できる． 実際に, HMM ベースの英語品詞タグ付け器を教師なし学習により学習し成功した例もある (Cutting,Kupiec,Pedersen and Sibun 1992) が，HMM ベースのタグ付け器の教師なし学習は限定された状況でしか有効でないという報告もあり (Merialdo 1994; Elworthy 1994), 本研究の目的のために有効であるかどうかは実験によって検証する必要がある. 実験では，提案手法のように隠れ状態を最大エントロピーモデルの素性として用いることにより，本研究のように辞書の存在を仮定しない難しい状況でも，HMM の教師なし学習の効果を有効に発揮させることができることを示す.

本研究では, 提案モデルの有効性を示すため, 英語の品詞タグ付けと日本語の単語分割という二つのタグ付けタスクにおいて実験を行った. 英語の品詞タグ付けの実験では, Penn Treebank の Wall Street Journal コーパスを用い, 少量のタグ付きコーパスしか利用できない時には, 提案モデルが通常の HMM や最大エントロピーモデルを用いた品詞タグ付けモデルよりも高い精度を示し, また, タグ付きコーパスを全て利用できる場合には, このコーパスに対する最高水準の 96.84%という精度を達成することが分かった. また, 提案モデルの使用により, 95%の精度のタグ付け器を開発するために必要な人手によるタグの修正を約 40%削減できることが分かった. 一方, 日本語の単語分割の実験では, 京大コーパスを用い, 英語の品詞タグ付けの場合と同様の結果を得た. これは, 提案手法が様々なタグ付けタスクに一般的に応用できる手法であることを示している.

まず, 第 2 節で, 本研究の基礎となる HMM を用いた確率的タグ付けモデルと最大エントロピーモデルを用いた確率的タグ付けモデルについて述べた後, 第 3 節で, HMM の状態を最大エ

ントロピーモデルを用いたタグ付けモデルの素性として用いる提案モデルについて述べる．そして, 第 4 節で, 英語品詞タグ付けと日本語単語分割の実験について報告する．最後に, 第 5 節で関連研究と今後の課題について議論する．

2　確率的タグ付けモデル

本論文では, タグ付けを与えられたシンボル列 $o_1 o_2 \cdots o_T$ の各位置 t のシンボル o_t にタグ τ_t を付与していく過程であるとする．例えば, 品詞タグ付けにおいては, シンボルは単語に相当し, タグは品詞タグに相当する．また, 日本語の文字ベースの単語分割においては, シンボルは文字に相当し, タグセットは単語の切れ目とそれ以外を示す $\{end, other\}$ のようなものになる．タグ付けの過程を確率モデルにより表し, 与えられたシンボル列に対して最も高い確率を持つタグ列を出力するのが確率的タグ付けである．本節では, 提案手法の基となる, 隠れマルコフモデル (HMM) によるタグ付けと, 最大エントロピーモデルを用いたタグ付けモデルについて説明する．以下, 記法として, 列 $x_k x_{k+1} \cdots x_{l-1} x_l$ を x_k^l と書く．例えば, 長さ T のシンボル列（文）は o_1^T と書く．

2.1　隠れマルコフモデルによるタグ付け

隠れマルコフモデル（HMM）は, 単純さと実際の性能の高さから様々なタグ付けタスクで利用されてきた確率モデルである (Cutting et al. 1992; Merialdo 1994; Bikel,Miller,Schwartz and Weischedel 1997)．HMM は, 状態遷移確率に従い状態遷移をし, 記号出力確率に従ってシンボルを出力する, 有限状態機械と考えられる．HMM の詳細については, (Rabiner 1989) 等を参照されたい．

HMM を用いた典型的なタグ付けモデルでは, HMM の状態がタグに対応する．タグ付けは, 与えられたシンボル列 o_1^T に対し, 最も大きい確率を持つ状態遷移列 $\hat{q}_1^T$ を求めること ($\hat{q}_1^T = \underset{q_1^T}{\operatorname{argmax}}\, p(q_1^T | o_1^T)$) に帰着される．この探索は,

$$
\begin{aligned}
\underset{q_1^T}{\operatorname{argmax}}\, p(q_1^T | o_1^T) &= \underset{q_1^T}{\operatorname{argmax}}\, \frac{p(o_1^T, q_1^T)}{p(o_1^T)} \quad (\text{ベイズの定理より}) \\
&= \underset{q_1^T}{\operatorname{argmax}}\, p(o_1^T, q_1^T) \quad (p(o_1^T) \text{ は探索中は定数であるから}) \\
&= \underset{q_1^T}{\operatorname{argmax}} \prod_{t=1}^{T} p(q_t | q_{t-1}) p(o_t | q_t)
\end{aligned}
$$

のように最終的には状態遷移確率 $p(q_t|q_{t-1})$ と記号出力確率 $p(o_t|q_t)$ の積を用いて表され, Viterbi アルゴリズム (Viterbi 1967) により効率的に行うことができる．

状態とタグを同一視する場合, タグ付きコーパスがあれば, HMM のパラメータ（状態遷移確率と記号出力確率）は, $C(q_j)$ をタグ付きコーパス中で状態 q_j を通過した回数, $C(q_j \to q_i)$ を

状態 q_j から状態 q_i へ遷移した回数, $C(q_i \uparrow o_k)$ を状態 q_i にいる時にシンボルが o_k であった回数として, 以下のように推定することができる.

$$p(q_i|q_j) = \frac{C(q_j \rightarrow q_i)}{C(q_j)}$$
$$p(o_k|q_i) = \frac{C(q_i \uparrow o_k)}{C(q_i)} \quad (1)$$

このような教師ありの学習に加えて，これらのパラメータを教師なし学習することも可能である．HMM の教師なし学習には, Baum-Welch アルゴリズムと呼ばれる学習法が知られている (Baum and Eagon 1967). Baum-Welch アルゴリズムでは, パラメータをある初期値から始めて, 与えられたシンボル列の尤度を必ず大きくするようなパラメータの更新を反復的に行うことによりパラメータの値を推定する.

(Cutting et al. 1992) は, HMM を Baum-Welch アルゴリズムにより学習し, 精度 96% の英語品詞タグ付け器を構築することに成功した. しかし, 彼らの研究では大量のタグ付きコーパスから構築した辞書を使用しており, それが教師なし学習を正しく働かせるために大きな役割を担っていたと考えられる. 前に述べた通り, 我々は辞書が存在しない新しい分野やタスクと対象としたタグ付け器の構築を想定しており, 同様の効果が得られるかは明らかではない.

加えて, 尤度の最大化が精度の向上に直接結びつく訳ではないことは一般的に知られており, HMM を用いたタグ付け器を Baum-Welch アルゴリズムで教師なし学習した場合に精度が低下してしまう場合も少なからずあることが報告されている (Merialdo 1994; Elworthy 1994). (Elworthy 1994) は Baum-Welch 学習に従ってタグ付けの精度の変化がとる類型を (1) 典型的（classical）: 反復を重ねるに従って精度が向上し続ける, (2) 初期最大（initial maximum）: 初期値が最高精度を示し, 反復に従って精度は低下する, (3) 早期最大（early maximum）: 数反復後に最高精度を示し, 後は低下する, の三つに分類し, 実験から, 初期値が大量のタグ付きコーパスから得られたような良い値である場合には初期最大や早期最大になり, Baum-Welch 学習の効果はほとんどないという報告をしている.

節 4 の英語品詞タグ付けの実験では, 我々の設定では, HMM ベースの品詞タグ付け器の精度は, 初期値を求めるのに使用したタグ付きコーパスの量に関係なく, Baum-Welch 学習により低下することを示す．この結果から, 我々の設定は, 一見, 前述の (Merialdo 1994; Elworthy 1994) における Baum-Welch 学習が精度を低下させてしまう場合のように思われる. しかし, 実験では, 提案手法により HMM の状態を最大エントロピーモデルの素性として利用する場合, Baum-Welch 学習によって精度が低下してしまった HMM の状態を使う方が学習をしていない HMM の状態を使うよりも精度を向上させる効果が高いことを示す. つまり, 我々の手法は, (Merialdo 1994; Elworthy 1994) の報告のように Baum-Welch 学習が逆効果に見える場合でも, それを有効に利用できる手法であるといえる.

Baum-Welch 学習は与えられた状態数で尤度を最大限向上させることを目標に学習を行うため, 学習後の HMM の状態は, 必ずしも, タスクが求めるタグと一対一に対応するものになっていないと推測される（特にランダムな初期値から学習した場合）. しかし, その場合でも尤度は

最大化されるためより良い言語モデルになっていると考えられる．我々の手法は, Baum-Welch 学習によって得られた尤度を最大化する点で最適な状態と，タスクのタグの関係を, 最大エントロピーモデルによって学習していると考えられる.

また, 一旦, 状態とタグに一対一の対応がないことを認めれば, 例えば, タグ数よりも大きい状態数の HMM を用いた方が, 言語モデルとしてより詳細になり, かつ, 最大エントロピーモデルによる状態とタグの関係の学習も自由度が増すから, より高いタグ付け精度を達成するという可能性も考慮することが可能になる. もちろん, 最大エントロピーモデルの教師あり学習の際には, 状態数が大きいほど状態素性はスパースになるから, 利用できるタグ付きコーパスの量によって適切な状態数は変わると考えられるが, 実験では, 少なくとも英語品詞タグ付けにおいては, タグ数（=45）よりも大幅に多い状態数 (=320) の HMM と組み合わせた時に最高精度が得られることも示す.

2.2　最大エントロピーモデルによるタグ付け

最大エントロピーモデル (Della Pietra,Della Pietra and Lafferty 1997) を用いたタグ付け (Ratnaparkhi 1996) では, 付与されるタグは以下の条件付き確率により決定される.

$$p_{ME}(\tau|h_t) = \frac{1}{Z(h_t)} \prod_i \alpha_i^{f_i(\tau,h_t)},$$
$$\text{ただし,}\quad Z(h_t) = \sum_\tau \prod_i \alpha_i^{f_i(\tau,h_t)}, \tag{2}$$

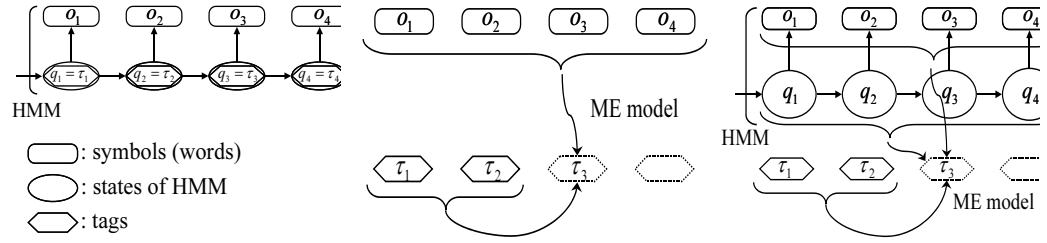
τ は位置 t に付与するタグであり, h_t は位置 t における文脈である. 文脈 h_t は, そこから現在のシンボル, 周囲のシンボル, 直前位置で付与したタグ等を保持する情報源と考えられる. 関数 $f_i(\tau, h_t)$ は素性関数と呼ばれ, 文脈中に出現した手がかりとその時のタグの組み合わせから決まる値を返す関数である. 素性関数同士が独立であるかに関係なくどのような素性も同時に利用できる（uni-gram 素性と bi-gram 素性を同時に使うなど）ことが, 最大エントロピー法の利点である. 素性関数は一般には非負の実数値を返す関数ならば何でもよいが, 本研究も含め自然言語処理では以下の形の二値の素性関数を用いることが多い.

$$f_i(\tau, h_t) = \begin{cases} 1 & \text{if } H(h_t) \wedge \tau = Y \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

これは, (Ratnaparkhi 1996) が英語の品詞タグ付けで用いた素性関数の定義と同様である. $H(h_t)$ の部分は h_t により真偽 (0,1) が決まる関数であり, 例えば,

$$o_{t-(n-1)}^{t} \text{は } X_1X_2\cdots X_n \text{ である } (\textit{symbol n-gram})$$
$$\tau_{t-n}^{t-1} \text{は } Y_1Y_2\cdots Y_n \text{ である } (\textit{previous tags}), \tag{4}$$

のようなものである. ただし, X_i や Y_i には実際の特定のシンボルやタグが入る.



(a) HMM ベースタガー　　(b) ME ベースタガー　　(c) 提案モデル

図 1: 既存モデルと提案モデル. HMM ベースタガー（a）では状態とタグは同一であるのに対して, 提案モデル（c）では必ずしも一致する必要はない

式 2 中の α_i はモデルのパラメータであり, 各素性関数 f_i の重要度を表していると解釈できる. 一般的に, 式 3 で書かれる素性関数が正しいタグとの組み合わせを示すときに 1 を返す場合, α_i は 1 より大きい値になる（例えば, $H(h_t)$ が「現在の単語は the である」で, Y が「定冠詞」のような場合). 逆に, 間違った判断をするような素性関数に対しては α_i は 1 より小さな値になる. この重み α_i は, 一般反復法 (GIS) (Darroch and Ratcliff 1972) や改良反復法 (IIS) (Della Pietra et al. 1997) などの最適化アルゴリズムにより求めることができる. その際, 学習データは, $\{(\tau_1, h_1), (\tau_2, h_2), \cdots, (\tau_N, h_N)\}$ という文脈と正しいタグの組の列として表され, このデータの尤度を最大化する最適化が行われる. この学習データはタグ付けコーパス (o_1^T と正しい τ_1^T の組の列) から構成される. 例えば, 式 4 で示した 2 つの $H(h_t)$ を用いる場合, h_t は o_1^T と τ_1^T から以下のように構成すればよい.

$$h_t = \langle o_{t-(n-1)}, \cdots, o_t, \tau_{t-n}, \cdots, \tau_{t-1} \rangle \tag{5}$$

タグ付けの際には, 与えられたシンボル列 o_1^T に対して最大確率をもつタグ列 $\hat{\tau}_1^T$ を求める:

$$\begin{aligned} \hat{\tau}_1^T &= \operatorname*{argmax}_{\tau_1^T} p(\tau_1^T | o_1^T) \\ &= \operatorname*{argmax}_{\tau_1^T} \prod_{t=1}^{T} p_{ME}(\tau_t | h_t). \end{aligned} \tag{6}$$

h_t が直前に付与したタグを含まない場合には, 各位置で最大の $p(\tau|h_t)$ を割り当てられるタグを選択し $\hat{\tau}_t$ とすればよい. h_t が直前に付与したタグを含む場合にも, Viterbi アルゴリズム (Viterbi 1967) と同様の探索法を用いてタグ列 $\hat{\tau}_1^T$ を求めることができる.

3　MEモデルを用いたタグ付けとHMMの統合

HMM と ME ベースのタグ付けモデルを組み合わせるための我々の手法は, HMM の隠れ状態を最大エントロピーモデルの素性として使用するという, きわめて単純なものである.

そのために, Viterbi アルゴリズムによって得られる HMM の最尤状態列 $\hat{q}_1^T$ を利用して値を返す以下の形の素性関数を定義し, 導入する.

$$f_i(\tau, h_t) = \begin{cases} 1 & \text{if } \hat{q}_{t+k-(n-1)}^{t+k} = Z_1 \cdots Z_n \wedge \tau = Y \\ 0 & \text{otherwise,} \end{cases}$$

ただし, k は現在位置 t からのオフセットであり, $Z_1 \cdots Z_n$ には実際の特定の状態の名前（番号）が入る. この素性関数を**状態素性**と呼ぶことにする. 図 1 は, 提案モデルと通常の HMM ベースのタグ付けモデルや ME ベースのタグ付けモデルの確率モデルとしての構造を比較して示している. 図中矢印は確率変数の依存関係が存在することを表している. 提案手法を用いた学習とタグ付けは以下のような手順で行われる.

学習:

1. HMM M_1 を Baum-Welch アルゴリズムにより（大量の）対象分野のタグなしコーパスから学習する.
2. タグ付きコーパス中のシンボル列 (文) o_1^T に対して, M_1 を用いた Viterbi アルゴリズムにより最尤の状態列 $\hat{q}_1^T$ を求める. 得られた状態列 $\hat{q}_1^T$, シンボル列 o_1^T, タグ付きコーパス中のタグ列 τ_1^T などから, 最大エントロピーモデルの学習のためのデータ $\{(\tau_1, h_1), \cdots, (\tau_T, h_T)\}$ を構成する.
3. 学習データを用い, 最大エントロピーモデル M_2 を推定する.

タグ付け:

1. 解析対象のシンボル列 o_1^T に対して, M_1 を用いた Viterbi アルゴリズムにより最尤状態列 $\hat{q}_1^T$ を求める.
2. $\hat{q}_1^T$ を含む h_1^T を参照しながら最大エントロピーモデル M_2 に基づくタグ付けを行う.

我々の手法はモデル構成の点からみると, 非常に柔軟なものになっている. HMM の教師なし学習は, タグ付与を行う最大エントロピーモデルからは分離されているので, タグセットに依存せず自由に HMM の状態数や初期パラメータ値などの設定を選ぶことができる. 例えば, HMM の状態数に関しては以下のような設定の可能性を考えることができる.

- 通常の HMM ベースタグ付けモデル同様, 状態とタグの一対一対応を仮定し, 初期パラメータ値は利用できる少量のタグ付きコーパスから得る.
- 状態とタグの一対一対応を仮定せず, 状態数はタグ数よりも大きく（タグ数が大きすぎる場合など, 場合によっては小さく）する. 初期パラメータ値は, 例えば, ランダムに与える.

前者を選択した場合には, 提案手法は既存の HMM ベースのタグ付け器を Baum-Welch アルゴリズムにより再学習し, そのタグ付けの出力を最大エントロピーモデルを用いるタグ付け器が利用するというカスケードされたタグ付け器と捉えることができる. この方法も十分あり得る選択であるが, 本研究では, 後者の方法に注目する. より大きな状態数を持つ HMM はモデルの容量が大きい. そのため, 対象分野の特徴をより良く捉えられる可能性があり, 全体としての精度を向上させる可能性がある. 実際, 英語品詞タグ付けの実験では, 後者の選択をした時に最高精度が達成されることを示す. 日本語の単語分割では, 実験で述べるように二種類のタグで単語の切れ目を示すモデルを用いる場合, 前者の方法では, HMM は二つの状態しかとらない非常に非力なモデルになり粗い分野適応になってしまう. しかし, 後者の方法で状態数を増やした HMM を用いればそのような問題はなく, 提案モデルが効果的に働くと期待される.

また, 本研究では HMM を分野適応のための確率モデルとして用いたが, HMM と同様に隠れ状態を持つ確率モデルならば, 本手法を応用して組み合わせることができる. この点で, 本手法は教師なし学習を組み合わせるための一般的なフレームワークになっている.

4　実験

4.1　英語品詞タグ付け

本節では, 提案手法を英語の品詞タグ付けに適用した実験について述べる. 実験は, Penn Treebank (Marcus,Santrini and Marcinkiewicz 1993) の Wall Street Journal (WSJ) コーパスを用いて行う. Wall Street Journal は約 50,000 の品詞タグ付き文を含む. 品詞タグは 45 種類である. 既存の研究では, この WSJ コーパスに対して 96.5%を超える精度が報告されている (Ratnaparkhi 1996; Brants 2000). この WSJ コーパスを, 学習用（section 00 から 22; 45,446 文; 1,084,229 単語）と評価用（section 23-24; 3,762 文; 89,537 単語）に分割して使用した. また, 学習用コーパスのうち最初の 32,000 文を, HMM の教師なし学習のためのテキストとして利用した.

実験 1

本実験では, 提案手法の有効性を示すため, 以下の三つのモデル（A）, （B）, （C）のタグ付け器の精度を, 使用できるタグ付きコーパスの量を 100 文から 40,000 文まで変化させて観察する.[1]

(A)　45 状態の HMM を用いたタガー（ベースライン）

(B)　標準的な最大エントロピーモデルベースのタガー

(C)　提案モデルによるタガー.

[1] 量を変化させる際には, コーパスの先頭の連続した部分を用いる.

表 1: 英語品詞タグ付けの素性セット. 表中 $(\{a,b\},c)$ のような表現は $(a,c),(b,c)$ を意味する.

category	$H(h_t)$
シンボル	$o_{t+k-(n-1)}^{t+k} = X_1 \cdots X_n$ $(k,n) = (\{-2,-1,0,1,2\},1)$
直前のタグ	$\tau_{t-n}^{t-1} = Y_1 \cdots Y_n$ $n = 1,2$
HMM 状態	$q_{t+k-(n-1)}^{t+k} = Z_1 \cdots Z_n$ $(k,n) = (\{-1,0,1\},1),(0,2),(0,3)$
数字	o_t が数字を含む
大文字	o_t が大文字を含む
ハイフン	o_t がハイフンを含む
接尾辞	o_t の接尾辞が W, $\|W\| \leq 4$
接頭辞	o_t の接頭辞が W, $\|W\| \leq 4$

(A) は, 学習用コーパスから相対頻度（式 1）で頻度の絶対ディスカウンティングを行った上で推定する（ディスカウント量は 0.5）. (B) は, (Ratnaparkhi 1996) で述べられているものとほとんど同一の最大エントロピーモデルに基づいたタグ付け器である. ただし, 細かい違いとして, (Ratnaparkhi 1996) では素性を抽出する時に単語が”rare” かどうかで抽出される素性が変わるようになっていたが, このモデルではそのような区別はしていない. また, (Ratnaparkhi 1996) では探索法としてビームサーチを用いていたが, このモデルでは Viterbi 探索を行う. (C) は, (B) で用いた素性に加えて前述の状態素性を用いた提案モデルによるタグ付け器である. 状態素性を取り出すための HMM は, 状態数 160 とタグ数よりも多くしてあり, 初期パラメータ値はランダムで与えた. HMM は Baum-Welch 学習で推定するが, 通常の Baum-Welch 学習では 160 状態の HMM の学習は計算量的に困難であったので, (Ghahramani and Jordan 1997) で述べられてる Gibbs サンプリングによる近似により計算量を減らした Baum-Welch アルゴリズムの変種を用いた. Baum-Welch 反復の回数は 100 とした. (B) と (C) の最大エントロピーモデルの推定には, IIS アルゴリズムを用い, 反復数は 250 とした. 表 1 に（C）で用いた素性を挙げる. これから状態素性を除いたものは（B）で用いた素性集合に一致する.[2] モデル (B),(C) の学習では (Ratnaparkhi 1996) と同様に, （量を変化させた後の）学習コーパスで 10 回以上出現した素性以外はあらかじめ取り除くことをした（cut-off; $threshold = 10$）. cut-off の影響をみるため, (C) で $threshold = 2$ とした場合を (C') として精度の変化をみた. また, 全てのモデルにおいて, コーパス中で最も頻繁に現れる単語の上位 10,000 単語のみを別々のシンボルとして扱い, それ以外の単語はすべて *UNK* という同一のシンボルとして扱った.

[2] 直前のタグは位置 $t-2$ まで使用しているが, 位置 $t-1$ と位置 $t-2$ のすべての組み合わせを考えて Viterbi 探索をするのではなく, 位置 $t-2$ に関しては Viterbi 探索中で保存されている最良の状態候補のみを考慮する.

表 2: 英語品詞タグ付けの精度

タグ付きコーパス（文）	(A)	(B)	(C)	(C')
100	72.29	70.74	**75.73**	**83.54**
200	78.25	80.03	**84.48**	**87.62**
500	83.89	86.03	**89.14**	**91.45**
1,000	87.19	89.76	**91.84**	**92.89**
2,000	89.98	92.58	**93.82**	**94.22**
4,000	91.63	94.16	**94.91**	**95.02**
8,000	92.86	95.25	**95.66**	95.54
16,000	93.61	96.11	**96.30**	96.13
32,000	93.86	96.54	**96.73**	96.52
40,000	93.94	96.74	**96.84**	96.63

表 2 に結果を示す．また, 図 2 ではこの結果をグラフで示す．提案手法が, 特に, 利用できるタグ付きコーパスの量が少ない時に他の比較手法に比べて高い精度を示すことが分かる．加えて, タグ付きコーパスを最大限に用いた時に, 提案手法 (C) が通常の最大エントロピー法を用いたタグ付け器よりもわずかではあるが高い精度を示していることが分かる．この時, 比較手法 (B) の精度は (Ratnaparkhi 1996) で報告されている精度 96.43%よりも高いため, 提案手法 (C) は最大エントロピーモデルを用いた既存手法の最高精度よりも優れた精度であると推測される.[3] これは, 40,000 文という比較的大量のタグ付きコーパスを利用できる場合でも, HMM 素性それ自体, 高精度を達成するのに有用な素性であることの現れだと考えられる（HMM に使用したタグなしコーパスは 32,000 文であり, 最終的に使用したタグ付きコーパス量 40,000 文よりも少ない）．また, (C') の結果から, 素性の cut-off の閾値を適切に選ぶ（小さめの閾値を用いる）ことにより, 少量のタグ付きコーパス時の精度をさらに向上させられることが分かる．ただし, タグ付きコーパスを最大限に用いた時の (C') の精度は (B) や (C) より低くなっている．これは, 利用できるコーパスの量に応じて適切な閾値が異なることを示唆しており, 状況に応じて閾値を適切に選択することができればさらなる精度向上を期待することができる．

次に, 本手法によりどの程度タグ付け器作成のコストが削減できるかを検証した．節 1 で述べたように, タグ付け器の学習・自動タグ付け・人手による修正, という過程により少しずつタグ付きコーパスの量を増やしていきながらタグ付け器を作成する場合に, ある精度に到達するま

[3] (Ratnaparkhi 1996) の実験とはデータが完全に同一ではないので断定することはできない．また，この時の (C) の精度の信頼水準 95%の信頼区間は約 0.11%であり, この結果だけから, (B) と (C) の差を統計的に有意とするのは難しいかもしれない．テストコーパスの量を増やしたり, クロスバリデーションを行うなどさらに調査が必要である．

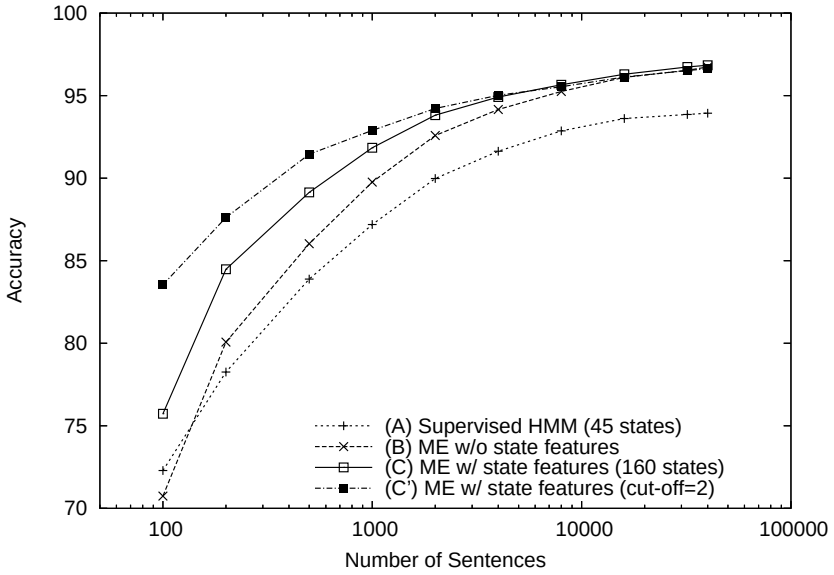


図 2: 英語品詞タグ付けの精度

でに必要な人手による修正の数がどのくらい変化するかをみる.[4] 人手による修正の量は,

$$\sum_{i=1}^{N-1}\{annot(i+1)-annot(i)\}\times\{1-acc(i)\}, \tag{7}$$

と表すことができる. $annot(i)$ は i 番目のループにおけるタグ付きコーパスの量であり, $acc(i)$ は i 番目のループでのタグ付け器の精度である. 表 3 は, 様々な目標精度に到達するまでにモデル (B) と (C) で必要な人手による修正の量をこの式で計算したものである. 開発過程としては, 各ループで前の実験結果の表 2 にある量までタグ付きコーパスを作成することを想定した.[5] 表には, 最終的に必要なタグ付きコーパスの量も載せた. 提案手法は, 最終的に必要なタグ付きコーパスの量だけでなく, その量のタグ付きコーパスを得るまでに必要な人手による修正の量を大幅に削減することが分かる. 例えば, 目標精度が 95%の時に, 提案手法は, 通常の最大エントロピーモデルで必要とする修正の 40%に相当する量の修正を削減することができる.

実験 2

次に, この実験では, Baum-Welch 学習の効果を検証する. 節 3 で述べた通り, HMM の状態とタグの一対一を仮定して, 組み合わせることも可能である. 本実験では, そのような HMM (45 状態の HMM) のパラメータ値をタグ付きコーパスから推定し (実験 1 におけるモデル

[4] 修正箇所を見つけ出すコストもあるが, ここでは無視する.
[5] 最終的に必要な量 $annot(N)$ は, 目標精度が表 2 中の精度の間にある場合には線形補完により推定した.

表 3: 英語タグ付け器開発に必要な人手による修正の量 (*corr*) と最終的に必要なタグ付きコーパスの量 (*total*). 量は単語（一文 ≈25 単語）.

	(B)		(C)			
目標精度					(B) からの削減量	
	corr.	*total*	*corr.*	*total*	*corr.*	*total*
90.00	3,983	23,355	2,101	13,259	47%	43%
94.00	9,526	89,391	5,502	54,218	42%	39%
95.00	14,171	167,600	8,532	105,162	40%	37%
96.00	23,503	359,025	17,275	292,191	26%	19%
96.50	38,186	730,802	27,713	558,571	27%	24%

（A）と同一), それを初期値に Baum-Welch 学習を行った場合の精度の変化をみる．これは, (Cutting et al. 1992; Merialdo 1994; Elworthy 1994) における実験に相当する. そして, これら Baum-Welch 学習前・学習後の HMM を提案手法に従って最大エントロピーモデルに組み合わせた場合の効果を調べる.

そのために, 以下のモデル (A), (A'), (D), (E) を用意する.

(A) 状態とタグの一対一対応を仮定した HMM（実験 1 のモデル (A)）

(A') (A) を Baum-Welch 学習で再学習した HMM

(D) 提案手法でモデル (A) を組み合わせた最大エントロピーモデル

(E) 提案手法でモデル (A') を組み合わせた最大エントロピーモデル

(A') については, 標準的な Baum-Welch アルゴリズムを用いて学習した（反復数は 20 回）. また, モデル (D), (E) については, 実験 1 と同様, 素性は表 1 の素性セットを用い, IIS の反復回数を 250 回, cut-off の閾値を 10 とした.

表 4 は, モデル (A), (A'), (D), (E) の精度を, 実験 1 と同様に, 使用できるタグ付きコーパスの量を変化させて示したものである. (A') の精度から明らかなように, 我々の設定では, 利用できるタグ付きコーパスの量に関係なく（初期値の良さに関係なく）, Baum-Welch 学習がタグ付けの精度を低下させることが分かる. つまり, 全てのタグ付きコーパス量で, (Elworthy 1994) で述べられている初期最大や早期最大のケースになっていると考えられる. (Merialdo 1994; Elworthy 1994) によれば, 初期値がかなり悪い時には Baum-Welch 学習により精度が改善されることが予想されるのであるが, 我々の設定でそうならないのは, 我々の設定では辞書的な情報を一切用いていないためと考えられる.

しかしながら, (D), (E) の精度をみると, タグ付きコーパス量が 100 文から 8,000 文の範囲では, Baum-Welch 学習後の (A') を組み合わせた (E) のほうが精度が高い. これは, 一見逆効果に見える Baum-Welch 学習が, 提案手法に従って最大エントロピーモデルと組み合わされた場合, 少量のタグ付きコーパスしか利用できないときの精度を改善するのに確かに有用であることを示している.

この実験結果からは, 次のようなことも分かる. 第一には, Baum-Welch 学習による再推定を

表 4: 45 状態 HMM の Baum-Welch 学習による精度の変化と, 提案手法で組み合わせた場合の精度

タグ付きコーパス（文）	(A)	(A')	(D)	(E)	参考 (B)	参考 (C)
100	72.29	65.46	74.58	**79.13**	70.74	75.73
200	78.25	65.38	82.88	**85.01**	80.03	84.48
500	83.89	64.98	88.45	**89.18**	86.03	89.14
1,000	87.19	64.81	91.52	**91.58**	89.76	91.84
2,000	89.98	65.97	93.21	**93.74**	92.58	93.82
4,000	91.63	67.63	94.73	**94.80**	94.16	94.91
8,000	92.86	68.75	95.46	**95.62**	95.25	95.66
16,000	93.61	70.31	**96.27**	96.21	96.11	96.30
32,000	93.86	71.31	**96.71**	96.67	96.54	96.73

していない HMM を組み合わせた提案モデル (D) でも, HMM を組み合わせない通常の最大エントロピーモデル (B) より精度が高いということである. これは, HMM を素性として組み合わせる手法自体の有用性を示している. HMM は生成モデルであり, 条件付きモデルである最大エントロピーモデルが捉えない情報を捉えられているからかもしれない. この点については, さらなる分析が必要である. 第二には, タグ付きコーパス量が 1,000 文以上の領域では, 160 状態の HMM をランダムな初期値から始めて Baum-Welch 学習し最大エントロピーモデルと組み合わせたモデル (C) のほうが, タグと一対一対応した 45 状態を持つ HMM を組み合わせるモデル (D) や (E) よりも精度が高いということである. これは, HMM の状態に関して状態とタグの一対一対応を仮定しないという方法が有効であることを示している. 前に述べた通り, 最大エントロピーモデルの教師あり学習の時には状態数が大きいほど状態素性はスパースになるから, 利用できるタグ付きコーパスの量によって適切な状態数は変わる. このことは, 1,000 文より少ない領域では 160 状態の HMM と組み合わせる (C) よりも 45 状態の HMM と組み合わせる (E) が精度が良いことからも分かる. つまり, タグ付きコーパスの量に従って適切な状態数を見つけることが必要になる. これは, 例えば, ヘルドアウトデータを用いたテストやクロスバリデーションにより決定すれば良いと考えられる. しかし, ここで強調したいのは, 標準的な WSJ コーパスを最大限利用できるような状況では, タグ数よりも大きい状態数の HMM と組み合わせた場合に最高精度を達成したということである. また, タグと状態が一対一に対応しない場合には初期値をタグ付きコーパスから簡単に推定することはできないため, ランダムな初期化を行っていることにも注目されたい. ランダムな初期化が有利に働いているか不利に働いているかについては調査が必要であり, タグと状態が一対一対応しない場合にタグ付きコーパスの情報を初期値に反映する方法についても今後研究する必要がある.

表 5: 日本語単語分割の素性セット

category	$H(h_t)$
シンボル	$o_{t+k-(n-1)}^{t+k} = X_1 \cdots X_n$ $(k,n) = (\{-1,0,1\},1),(\{-1,0,1\},2),(\{-1,0,1\},3)$
HMM 状態	$q_{t+k-(n-1)}^{t+k} = Z_1 \cdots Z_n$ $(k,n) = (\{-1,0,1,2,3\},1),(\{0,1,2\},2),(0,3)$

4.2 日本語単語分割

次に, 我々の手法が様々なタグ付けタスクに一般的に適用可能であることを示すため, 日本語の単語分割を用いて実験を行う. 日本語や中国語などの言語においては, 英語のように単語がスペースで区切られていないため, 様々な NLP の前処理としてどこが一つの単語（形態素）なのかを認識しなければならない. このタスクを**単語分割**と呼ぶ.[6] 本実験では, 文字をベースに単語分割を行うモデルを想定して実験を行う. 文字をベースにした手法は, 辞書を必要としないため未知語問題に対して頑健であると考えられ, 既存研究では 95–96%の分割精度を実現している (Nagamatsu and Tanaka 1997; Oda and Kita 1999; 小田, 森, 北 1999). 本実験では, これらの既存研究を参考にし, 単語分割を単語の最後の文字に *end*, それ以外の文字には *other* というタグを付与するタグ付けタスクと捉える.

実験のためのコーパスとしては, 京大コーパス (ver. 3.0)(黒橋, 長尾 1997) を用いた. 京大コーパスでは, 約 40,000 文の新聞記事（毎日新聞の 20,000 文の通常記事と 20,000 文の社説）がアノテーションされている. このコーパスの通常記事 20,000 文に含まれる形態素の情報から分割情報のみを取り出し, 学習用の 17,694 文と評価用の 1,966 文に分割し使用した. 教師なし学習のためには, 1994 年の毎日新聞の記事約 110,000 文を用いた.[7]

単語分割の精度は文字単位のタグ付け精度, 分割の precision・recall によって測ることができる. Std を評価コーパス中の単語数, Sys をシステムが認識した単語数, Cor をシステムが正しく認識した単語数とすると, precision は Cor/Sys, recall は Cor/Std と表すことができる. ここでは, 紙面を節約するため, 以下で定義される, precision と recall の調和平均である F-score を用いて精度を示す.

$$F\text{-}value = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

本実験では, 以下の三つのモデル (A), (B), (C) を用意し, 英語品詞タグ付けの実験と同様に使用できるタグ付きコーパスの量を変化させて, 精度の変化を比較した.

(A) (Stolcke and Shriberg 1996) で述べられているモデルに基づいた文字トライグラムモデ

[6] もちろん, 形態素解析がこの処理を同時に行うこともある.

[7] CD-毎日新聞'94, 日外アソシエイツ. 京大コーパスは 1995 年の記事を対象としているため, これらの文と京大コーパスの文に共通部分はない.

表 6: 日本語単語分割の精度 (F-score)

タグ付きコーパス（文）	(A)	(B)	(C)
50	40.80	66.57	**76.97**
100	49.43	71.38	**82.46**
200	62.88	74.57	**85.80**
500	77.30	80.55	**89.31**
1,000	84.36	85.09	**91.32**
2,000	88.01	87.90	**92.54**
4,000	91.35	90.75	**93.97**
8,000	93.75	92.92	**94.84**
10,000	94.21	93.53	**95.12**
12,000	94.61	93.93	**95.32**
16,000	95.18	94.76	**95.68**

ル.[8]

(B) 状態素性を用いない最大エントロピーモデル. 使用した素性は表 5 のうち, シンボル素性のみである.

(C) 状態素性を用いた最大エントロピーモデル. 使用した素性は表 5 に挙げたものである. cut-off の閾値は 1 とし, HMM は, 320 状態の HMM を前述の Gibbs サンプリングを用いた Baum-Welch アルゴリズムでランダムな初期値から始めて学習し, 使用した.

最も頻繁に出現した 2,000 文字をシンボルとして扱い, このシンボルを次ぎに来る文字のタイプが変わるかどうかを示す二値の素性 $ct \in \{+, -\}$ により拡張する. つまり, シンボル o_i に対して $\langle o_i, ct \rangle$ という拡張されたシンボルを生成し, 用いる（これにより, シンボルの数は実質的に約 4,000 となる）. これは, 日本語には, 漢字・ひらがな・カタカナといった文字種があり, 各文字種の連続が単語（形態素）になることが多いため, その変化が単語を認識するための強力な手がかりとなるからである. ただし, モデル (A) に関してはこの拡張が精度にほとんど影響しなかったため, 元の 2,000 シンボルをそのまま使用した.

タグ付きコーパスの量を変えた時の, これらのモデルの精度を表 6 に示すとともに, 図 3 でグラフとして示す. 英語品詞タグ付けの時と同様, 提案手法によって, タグ付きコーパスが少量の時の精度が大きく改善されることが分かる. また, タグ付きコーパスを最大限に利用できる時には提案手法が最高精度を達成することも分かる.

表 7 は, モデル (A) と (C) が目標精度 (F-score) に達するまでに必要な人手による修正の量

[8] このモデルの推定には SRILM (The SRI Language Modeling Toolkit)(SRI 2000) を用いた. このモデルは, 文字クラスタリングによる文字クラスを用いていないという点を除けば, (小田他 1999) が述べているモデルと構造は同じである.

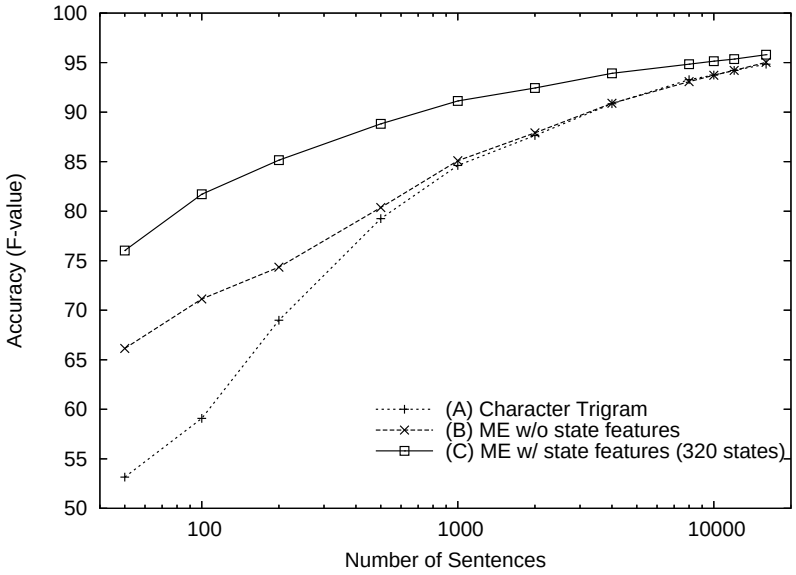


図 3: 日本語単語分割の精度 (F-score)

表 7: 日本語単語分割器開発に必要な人手による修正の量 ($corr$) と最終的に必要なタグ付きコーパスの量 ($total$). 単位は文字（1 文 ≈46 文字）.

	(A)		(C)			
目標精度					(A) からの削減量	
	corr.	*total*	*corr.*	*total*	*corr.*	*total*
94.0	22,167	411,491	7,177	185,412	68%	55%
95.0	28,399	673,520	12,824	414,486	55%	38%

を, 表 7 から推定したものである.[9] 提案手法は, 95%の F-score を達成するために必要な修正を 55%削減することがわかる.

5　関連研究と今後の課題

提案手法に関連した手法としては，単語や文字のクラスタリングによるデータスパースネス問題の回避がある (Brown,Pietra,deSouza,Lai and Mercer 1992; 小田他 1999). 例えば, 320 状態の HMM を考えると, 入力の各シンボルに対して Viterbi アルゴリズムにより求まる最尤の状態は, 320 クラスの中でそのシンボルに最もふさわしいクラスと考えることができる. ただし, 提案手法は, 既存のクラスタリングを用いた手法とは, クラスが Viterbi アルゴリズムによ

[9] 文字毎の精度を式 7 中の $acc(i)$ として用いた.

表 8: HMM 状態素性と単語クラスタ素性の比較

タグ付きコーパス（文）	ME+HMM	ME+Brown	ME+HMM+Brown
100	75.73	73.81	77.73
200	84.48	84.29	86.41
500	89.14	89.85	90.90
1,000	91.84	92.57	93.10
2,000	93.82	94.10	94.51
4,000	94.91	94.96	95.29
8,000	95.66	95.77	95.87
16,000	96.30	96.38	96.42
32,000	96.73	96.76	96.76
40,000	96.84	96.86	96.85

りコンテキストに従って動的に決まる点で異なる. 我々はタグ付けにはこのような動的なクラスタリングが適していると考えている. 試みに, 英語の品詞タグ付けで, Brown らによる相互情報量を用いた単語クラスタリング (Brown et al. 1992) で求めた単語クラスを素性として用いる場合と提案手法で HMM 状態を素性として用いる場合の精度の比較をした（状態数, クラス数とも 160). 表 8 が結果である. 単独の効果としてはタグ付きコーパスの量がかなり小さい時には, HMM 状態素性の方が効果が高く, それ以外では単語クラスタの方が効果が高いことが分かった. ただし, HMM の状態素性と Brown のクラスタ素性とを同時に使った場合（表中の ME+HMM+Brown）には, 両者が単独で用いられる場合よりも高い精度を示すことも分かる. これは, HMM の状態と単語クラスタが異なる性質を持ち, それぞれがタグ付けに有用なためと考えられる.

本研究では, Viterbi アルゴリズムにより求まる最尤の状態列のみを用いたが, それ以外の状態を利用することも可能である. 例えば, ある位置である状態を通過する確率を返す実数値の素性を組み込むことなどが考えられる（各位置で状態数だけの素性が確率値に応じて同時に発火する）. そのような場合, 単語が特定のクラスに属するのではなく各クラスへの所属の度合いが数値で表されるという, ソフトクラスタリングと類似の手法になり, データスパースネス問題をさらに解消できる可能性がある. ただし, その場合, 最大エントロピーモデル推定の計算量が膨大になるので, 学習の高速化や素性選択の工夫が必要になる.

また, HMM などの生成確率モデルを分類器や条件付き確率モデルで利用するという点で関連する研究としては, (Jaakkola and Haussler 1998; Tsuda,Kin and Asai 2002) などの研究がある. (Jaakkola and Haussler 1998) では例（シンボル列）X の対数尤度のモデルパラメータに対する敏感度（Fisher スコア; $U_X = \nabla_\theta \log P(X|\theta)$）を例 X から抽出される素性ベクトルとして利用する. ただし, (Jaakkola and Haussler 1998) の方法そのままでは, タグ付けで重要と考え

られる文字列中の位置の情報を取り扱うことができない．（例えば, $U_{X,t} = \nabla_\theta \log P(o_1^t|\theta)$ のような量を考える必要がある). また, (Jaakkola and Haussler 1998) では, 文の長さ T と関係なくその文の生成に関係するパラメータ数だけの素性が発火する. HMM の場合, 状態数を N, 文中の単語種数を W として $O(N + N^2 + WN)$ の素性が発火することになる. 前述のように位置情報を取り扱おうとすれば, 一文あたり $O(T(N + N^2 + WN))$ 程度の素性が発火することになる. 一方, HMM 状態素性では, 一文あたり $O(T)$ の素性が発火するだけである. 上で述べたソフトクラスタリング的な素性にしたとしても, 状態の組などを考えなければ, $O(NT)$ 程度であり, この点で, 提案手法はより計算コストの小さい方法であると言える. ただし, HMM 状態素性のような生成モデルの利用の仕方と, (Jaakkola and Haussler 1998; Tsuda et al. 2002) のような手法の精度に対する効果の比較は必要であり, 今後の研究の一方向になると考える. また, 前述した通り, 提案手法も (Jaakkola and Haussler 1998; Tsuda et al. 2002) も HMM のみに限定される手法ではなく, 他の確率モデルとの組み合わせの可能性の探索も今後の研究の一つと考えられる.

最後に，本論文では, HMM の状態数や cut-off の閾値は限られたものを用いたが, 実験で述べた通り, 最適な値はタグ付けコーパスの量などに依存して変わると考えられる. ヘルドアウトデータなどを用いて最適な値を求められるかどうかも, 今後の実験で確かめる必要がある.

6　結論

本論文では, タグ付け器作成のコストを削減することを目的に, 教師なし学習された HMM の状態を最大エントロピーモデルの素性として用いるタグ付けモデルを提案した. 実験では, 本手法が少量のタグ付きコーパスしか利用できないときの精度を大幅に改善することを示し, コストの削減に寄与することを実証した. さらに, タグ付きコーパスを最大限利用できる時には, 提案モデルが最高水準の精度を達成することを示した. また, 英語の品詞タグ付けと日本語の単語分割という二種類のタグ付けタスクで有効性を示すことにより, 本手法が様々なタグ付けタスクに一般的に適用可能な方法であることを示した.

参考文献

Baum, L. E. and Eagon, J. A. (1967). “An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of ecology..” *Bull. Amer. Math. Soc.*, **73**, 360–363.

Berger, A. L.,Della Pietra, S. A.,and Della Pietra, V. J. (1996). “A Maximum Entropy Approach to Natural Language Processing.” *Computational Linguistics*, **22** (1), 39–71.

Bikel, D.,Miller, S.,Schwartz, R.,and Weischedel, R. (1997). “Nymble: a high-performance learning name-finder.” In *Proceedings of the Fifth Conference on Applied Natural Lan-*

guage Processing, pp. 194–201.

Borthwick, A. (1999). "A Maximum Entropy Approach to Named Entity Recognition." Ph.D. Thesis. New York University.

Brants, T. (2000). "TnT – A Statistical Part-of-Speech Tagger." In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*.

Brill, E. (1994). "Some Advances in Transformation-Based Part of Speech Tagging." In *Proceedings of the Twelefth National Conference on Artificial Intelligence (AAAI-94)*.

Brown, P. F.,Pietra, V. J. D.,deSouza, P. V.,Lai, J. C.,and Mercer, R. L. (1992). "Class-based n-gram models of natural language." *Computational Linguistics*, **18** (4), 467–479.

Cutting, D.,Kupiec, J.,Pedersen, J.,and Sibun, P. (1992). "A Practical Part-of-Speech Tagger." In *Proceedings of the Third Conference on Applied Language Processing*, pp. 133–140.

Darroch, J. N. and Ratcliff, D. (1972). "Generalized Iterative Scaling for log-linear models." *The Annals of Mathematical Statistics*, **43**, 1470–1480.

Della Pietra, S.,Della Pietra, V.,and Lafferty, J. (1997). "Inducing features of random fields." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19** (4), 380–393.

Elworthy, D. (1994). "Does Baum-Welch Re-estimation Help Taggers?." In *Proceedings of the 4th ACL Conference on Applied Natural Language Procecssing*, pp. 53–58.

Ghahramani, Z. and Jordan, M. I. (1997). "Factorial Hidden Markov Models." *Machine Learning*, **29**, 245–273.

Jaakkola, T. S. and Haussler, D. (1998). "Exploiting generative models in discriminative classifiers." In *Proceedings of the 10th Advances in Neural Information Processing Systems (NIPS)*.

Kazama, J.,Miyao, Y.,and Tsujii, J. (2001). "A Maximum Entropy Tagger with Unsupervised Hidden Markov Models." In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS)*, pp. 333–340.

Kudoh, T. and Matsumoto, Y. (2000). "Use of Support Vector Learning for Chunk Identification." In *Proceedings of CoNLL-2000 and LLL-2000*.

Marcus, M.,Santrini, B.,and Marcinkiewicz, M. A. (1993). "Building a large annotated corpus of English: Penn Treebank." *Computational Linguistics*, **19** (2), 313–330.

Merialdo, B. (1994). "Tagging English Text with a Probabilistic Model." *Computational Linguistics*, **20** (2), 155–171.

Nagamatsu, K. and Tanaka, H. (1997). "A Stochastic Morphological Analysis for Japanese employing Character n-Gram and k-NN Method." In *NLPRS'97*, pp. 23–28.

Oda, H. and Kita, K. (1999). "A character-based Japanese word segmenter using a PPM*-based language model." In *Proceedings of the 18th International Conference on Computer Processing of Oriental Languages (ICCPOL'99)*, pp. 527–532.

Rabiner, L. R. (1989). “A tutorial on hidden Markov models and selected applications in speech recognition.” *Proceedings of the IEEE*, **77** (2), 257–285.

Ratnaparkhi, A. (1996). “A Maximum Entropy Model for Part-of-Speech Tagging.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 133–142.

SRI (2000). “SRILM – The SRI Language Modeling Toolkit Ver. 1.0.”. available via `http://www.speech.sri.com/projects/srilm/`.

Stolcke, A. and Shriberg, E. (1996). “Automatic Linguistic Segmentation of Conversational Speech.” In *Proc. ICSLP '96*, Vol. 2, pp. 1005–1008 Philadelphia, PA.

Tsuda, K.,Kin, T.,and Asai, K. (2002). “Marginalized kernels for biological sequences.” *Bioinformatics*, **18**, S268–S275.

Viterbi, A. J. (1967). “Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm.” *IEEE Transactions on Information Theory*, **IT-13**, 260–267.

黒橋禎夫, 長尾真 (1997). “京大テキストコーパス・プロジェクト.” 言語処理学会 第 3 回年次大会, pp. 115–118.

小田裕樹, 森信介, 北研二 (1999). “文字クラスモデルに基づく日本語単語分割.” 自然言語処理, **6** (7), 93–108.

略歴

風間　淳一: １９９９年東京大学理学部情報科学科卒業．２００４年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程修了．博士（情報理工学）．同年４月より，北陸先端科学技術大学院大学情報科学研究科知識工学講座助手．

宮尾　祐介: １９９８年東京大学理学部情報科学科卒業，２０００年東京大学大学院修士課程修了．２００１年９月より，東京大学大学院情報理工学系研究科コンピュータ科学専攻助手．

辻井　潤一: 京都大学大学院工学博士．１９７１年京都大学工学部電気工学科卒業，１９７３年同大学大学院修士課程修了．同年４月より，同大学電気工学第２教室助手，助教授を経て，１９８８年から英国 UMIST (University of Manchester Institute of Science and Technology) 教授．同大学計算言語学センター所長などを経て，１９９５年より東京大学大学院理学系研究科情報科学専攻教授．組織変更により，現在は同大学院情報理工学系研究科・コンピュータ科学専攻教授．また，１９８１〜１９８２，フランス CNRS (グルノーブル) の招聘研究員．言語処理学会，情報処理学会，ACL 各会員．

(1995 年 5 月 6 日 受付)
(1995 年 7 月 8 日 再受付)
(1995 年 9 月 10 日 採録)