大域的な情報を用いた未知語の品詞推定

中川哲治[†1,*1] 松本裕治[†2]

本稿では，局所的な情報と大域的な情報を用いて未知語の品詞推定を行う手法を提案する．多くの従来手法において，未知語の品詞は局所的な情報（未知語の前後数単語内，あるいは未知語が含まれる文内の情報等）のみを用いて推定されるが，大域的な情報（同じ語形を持つ未知語が文書中の別の場所でどのような品詞として使われているかという情報）は未知語の品詞推定を行ううえでしばしば有用な手がかりとなる．局所的な情報だけではなく大域的な情報も利用して未知語の品詞を推定するために，文書中に出現する同じ語形を持つすべての未知語の品詞を同時に考慮した確率モデルを提案し，ギブスサンプリングを用いて解析を行う．また提案手法において，品詞情報が付与されていないようなラベルなしデータを利用する方法も検討する．複数のコーパスを使用して実験を行った結果，提案手法を用いることにより，特に中国語と日本語において高い精度で未知語の品詞を推定できることを確認した．

Guessing Parts-of-speech of Unknown Words Using Global Information

Tetsuji Nakagawa[†1,*1] and Yuji Matsumoto[†2]

In this paper, we present a method for guessing POS tags of unknown words using local and global information. Although many existing methods use only local information (i.e. limited window size or intra-sentential features), global information (such as consistency of POS tags of unknown words with the same lexical form) provides valuable clues for predicting POS tags of unknown words. We propose a probabilistic model, in which all the occurrences of the unknown words with the same lexical form in a document are taken into consideration at once, for guessing POS tags of unknown words using global information as well as local information, and predict POS tags of unknown words using Gibbs sampling. We also attempt to utilize unlabeled data which is not attached POS tags. We conduct experiments on multiple corpora, and show that the method improves accuracy of POS guessing of unknown words especially for Chinese and Japanese.

1. 背景と課題

単語の品詞を同定する品詞タグ付けは，基本的な言語解析タスクの１つである．一般に品詞タグ付けシステムは，人手により作成された辞書や，訓練データから自動的に学習されたパラメータを持っており，それらの情報を用いて入力された単語の品詞を推定する．しかしながら，実際に品詞タグ付けを行う際には，そのような辞書や訓練データの中には存在しない単語がしばしば出現する．このような単語は未知語と呼ばれる．未知語は，その情報が品詞タグ付けシステムの中に存在しないため，特別な処理によって扱われる．未知語の品詞を正確に推定することは，高精度な品詞タグ付けを行うために必要であり，また単語辞書を生テキストから自動的に作成するような場合にも重要である．

未知語の品詞推定に関して，これまでに様々な研究が行われている[5),13)–17)]．これらの既存手法の多くでは，未知語の品詞は局所的な情報，つまり未知語の前後に存在するいくつかの単語や未知語自身の情報（語尾や文字種等）のみを用いて推定されている．しかしながら，局所的な情報のみで品詞を推定するのは困難な場合が存在する．そのような場合に，同じ語形を持つ未知語が文書中の別の場所でどのような品詞として使われているかというような大域的な情報は非常に有用であると考えられる．そこで本稿では，局所的な情

†1 沖電気工業株式会社ユビキタスサービスプラットフォームカンパニー
Ubiquitous Service Platform Company, Oki Electric Industry Co., Ltd.
†2 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
*1 現在，情報通信研究機構知識創成コミュニケーション研究センター
Presently with Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology

報のみを手がかりとして未知語の品詞推定を行った場合に，品詞推定が困難である場合が存在するという従来手法の持つ課題に対して，大域的な情報も用いることにより高い精度で未知語の品詞推定を行う手法を提案する．以下本稿では特に断らない限り，同じ語形を持つ未知語が文書中の別の場所でどのような品詞として用いられているかという情報を，大域的情報と呼ぶことにする．

たとえば，名詞のように使われている未知語が存在した場合，それが普通名詞であるか固有名詞であるかを，局所的な情報のみを用いて判断するのは困難な場合がある．しかし，もしそのような曖昧な語と同じ語形を持つ未知語が，文書中の別の場所で，大きな手がかりとなる局所的な情報（たとえば,「〜先生」のような人名に付く敬称等）とともに出現していれば，そのような情報は曖昧な語の品詞を推定するうえで役に立つと思われる．別の例として，サ変名詞に関する問題があげられる．サ変名詞は普通名詞のように使うことができるが，単語の末尾に「する」を付けることにより動詞として使うこともできる．名詞のように使われている未知語が，サ変名詞と普通名詞のどちらであるかを判定することはしばしば困難である．しかし，サ変名詞なのか普通名詞なのか曖昧である未知語が存在した場合，もし同じ語形を持つ未知語が文書中の別の場所で「〜する」という形で出現していれば，その曖昧な未知語の品詞はサ変名詞である可能性が高いと判断することができる[⋆1]．これらの例のように，大域的な情報は未知語の品詞を推定するうえで有用であると思われる．

本稿で提案する手法では，未知語の個々の出現のみに注目してその品詞を推定するのではなく，同じ語形を持つすべての未知語の品詞の，文書[⋆2]全体における同時確率分布を定義して利用する．このような同時確率分布を用いることにより，同じ語形を持つすべての未知語の品詞間の依存関係（相互作用）を考慮することができる．そして，従来手法でしばしば用いられているように，単語や文の独立性を仮定して単語単位や文単位での出現確率を個別に最大化するのではなく，文書全体が与えられた条件での，同じ語形を持つすべての未知語の品詞の同時確率を考慮することにより，未知語の品詞推定を行う．

このような大域的な情報を利用した確率モデルは，局所的な情報のみを利用した確率モデルと比較した場合，確率変数間の複雑な依存関係を考慮する必要があるため多くの計算量を必要とする．そこで提案手法では，ギブスサンプリング[23)]を使用して効率的に確率モデルの計算を行う．ギブスサンプリングは，統計物理の分野で発展してきたマルコフ連鎖モンテカルロ（Markov Chain Monte Carlo; MCMC）法の一種であり，大規模な確率モデルを効率良く近似的に計算するために利用することができる一般性のある手法である．自然言語処理ではしばしば，単語の品詞は直前の数個の単語の影響のみを受けると仮定して品詞タグ付けを行ったり，文節の係り受け関係は互いに独立であると仮定して係り受け解析を行ったりするなど，確率変数間の依存関係を限定して計算しやすいモデルを構築することが多い．また，品詞タグ付けと構文解析を別々の処理に分けて行うなど，タスク間の独立性を仮定することが多い．しかしながら，大規模な確率モデルの計算を可能にするギブスサンプリングのような手法を用いることにより，多くの近似を行った単純なモデルを厳密に計算するのではなく，様々な情報を考慮した複雑なモデルを近似的に計算するアプローチを試みることができる．

提案手法の 1 つの特徴として，ラベルなしデータを取り入れることにより性能の改善を容易に行える可能性があることがあげられる．ここでいうラベルなしデータとは，品詞タグの付与されていないコーパスのことである．教師ありの機械学習を用いて品詞タグ付けを行うには，訓練データとして品詞タグの付与されたコーパスが必要となるが，品詞タグ付きコーパスを作成するには多くの人手による作業を要する．一方で，品詞タグが付与されていないラベルなしデータは容易に入手することができる．そのため，ラベルなしデータを利用することにより，未知語の品詞推定精度を改善することができれば非常に有用である．提案手法では，ラベルなしデータをテストデータに単純に加えてテストデータの分量を増やすことにより，確率モデル自体は変更せずに，テスト時に使用される大域的な情報を増やすことができると思われる．

提案手法では，文書全体が与えられた条件のもとでの確率分布を使用しているため，入力されたデータを 1 文ごとに逐次的に解析していくことはできず，解析を行う前に入力データの文書全体を読み込んでおいてバッチ的に処理を行う必要がある．しかしながら，生テキストから半自動的に辞書を作成するような状況を

⋆1 サ変名詞のような，複数の可能な用法を持つような品詞に関するこのような曖昧性の問題は,「可能性に基づく品詞の問題」として Asahara [1)] により指摘されている．

⋆2 本稿では，処理の対象であるデータ全体（訓練データやテストデータ全体等，複数の文から構成される集合）を表すために「文書」という言葉を使用することにする．

考えた場合，実時間で処理を行う必要性はないが，自動的に解析されたデータの修正に要する人手を少しでも減らすためには高い解析精度が必要とされるため，このような手法の用途として適していると思われる．

以下，2 章では大域的な情報を用いた未知語の品詞推定手法について説明する．3 章では実験結果を報告する．4 章では関連研究について議論し，5 章で結論を述べる．

2. 大域的な情報を用いた未知語の品詞推定

本稿では，未知語の品詞推定タスクを，品詞タグ付けの後処理として考える．つまり，単語分割はすでに済んでおり，既知語の品詞もすでに決定されていると仮定して，未知語に対する品詞の推定にのみ注目する[⋆1]．

以下この章では，大域的な情報を利用して未知語の品詞推定を行うための確率モデルについてまず説明する．次に，テストデータの解析方法と，訓練データからのモデルパラメータの推定方法について説明する．また，ラベルなしデータを利用する方法についても議論する．

2.1 大域的な情報を利用する確率モデル

提案手法では，同じ語形を持つ文書中のすべての未知語の品詞を同時に考慮した確率モデルを考える．そのような未知語の品詞は相互に影響を及ぼし，さらに各未知語の品詞は局所的な文脈の影響も受けると考える．これと似たような状況は，物理学でも扱われている．たとえば，ある系の中に多量の電子が存在しており，各電子がスピンを持っている場合を考える．このような電子のスピンは相互に影響を及ぼし，さらに各スピンは外部磁場の影響も受ける．物理学では，系の状態を $\mathbf{s}$ とし，系のエネルギーを $E(\mathbf{s})$ とした場合，$\mathbf{s}$ の確率分布は次のようなボルツマン分布により表現されることが知られている：

$$P(\mathbf{s}) = \frac{1}{Z}\exp\{-\beta E(\mathbf{s})\}, \tag{1}$$

ここで，β は逆温度と呼ばれる値であり，Z は次のように定義される正規化定数である：

$$Z = \sum_{\mathbf{s}} \exp\{-\beta E(\mathbf{s})\}. \tag{2}$$

高村ら[25] は，このモデルを単語の感情極性判定に応用したが，本研究ではこのモデルを未知語の品詞推定へ応用する．

⋆1 未知語の単語分割自体も難しい問題ではあるが，たとえば文献 26) のような既存手法により対処することができるため，本稿ではこの問題については議論しないことにする．

以下の説明では，文書中に同一の語形を持つ未知語が K 回出現するとする．未知語がとりうる品詞は N 種類あるとし，各品詞は 1 から N の整数で表現されることとする．k 番目に出現した未知語の品詞を t_k で表し，k 番目に出現した未知語の局所的文脈（未知語の前後の単語の語形や品詞等）を w_k で表すことにする．また $\mathbf{w}$ と $\mathbf{t}$ を，それぞれ w_k と t_k の集合とする：

$$\mathbf{w} = \{w_1, \cdots, w_K\},\ \ \mathbf{t} = \{t_1, \cdots, t_K\},$$
$$t_k \in \{1, \cdots, N\} \quad (k = 1, \cdots, K).$$

$\lambda_{i,j}$ は，品詞 i と品詞 j の間における相互作用の強さを表す重みとし，対称性を持つものとする（$\lambda_{i,j} = \lambda_{j,i}$）．そして，$\mathbf{w}$ が与えられた場合に未知語の品詞が $\mathbf{t}$ であるエネルギーを次のように定義する：

$$E(\mathbf{t}|\mathbf{w}) = -\left\{\frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{t_k,t_{k'}} + \sum_{k=1}^{K}\log p_0(t_k|w_k)\right\}. \tag{3}$$

ここで，$p_0(t|w)$ は局所的文脈 w のみを用いて計算される品詞 t の初期分布（局所的モデル）であり，最大エントロピーモデル等の任意の統計的モデルを用いて計算されるものとする．上記の式の右辺は，2 つの要素から構成されている．1 つは大域的な品詞間の相互作用を表す項であり，もう 1 つは局所的な文脈による影響を表す項である．この大域的な情報を表す項で，ある 2 つの品詞 i と j に対して $\lambda_{i,j}$ の値が大きければ（または小さければ），未知語がある場所で品詞 i として用いられていた場合，別の場所ではそれと同じ語形を持つ未知語は品詞 j として用いられやすい（または用いられにくい）という品詞間の相互作用が表現されることになる．

本研究では，逆温度 β の値は 1 に固定する．すると，式 (1)，(2)，(3) より，次のように $\mathbf{t}$ の条件付き同時確率分布が得られる：

$$P(\mathbf{t}|\mathbf{w}) = \frac{1}{Z(\mathbf{w})}p_0(\mathbf{t}|\mathbf{w})\exp\left\{\frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{t_k,t_{k'}}\right\}, \tag{4}$$

$$Z(\mathbf{w}) = \sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w})} p_0(\mathbf{t}|\mathbf{w})\exp\left\{\frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{t_k,t_{k'}}\right\}. \tag{5}$$

ただし，

$$p_0(\mathbf{t}|\mathbf{w}) \equiv \prod_{k=1}^{K} p_0(t_k|w_k). \tag{6}$$

ここで $\mathcal{T}(\mathbf{w})$ は，$\mathbf{w}$ が与えられた場合における，それらの未知語の品詞のあらゆる可能な候補を表す集合である．未知語の数は K 個であり，各未知語はそれぞれ N 個の品詞のうちのどれか 1 つをとるため，$\mathcal{T}(\mathbf{w})$ の要素の数は N^K 個である．上記の式は，次のように関数 $f_{i,j}(\mathbf{t})$ を定義して変形することができる：

$$f_{i,j}(\mathbf{t}) \equiv \frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\delta(t_k,i)\delta(t_{k'},j), \tag{7}$$

$$P(\mathbf{t}|\mathbf{w}) = \frac{1}{Z(\mathbf{w})}p_0(\mathbf{t}|\mathbf{w})\exp\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{i,j}f_{i,j}(\mathbf{t})\right\}, \tag{8}$$

$$Z(\mathbf{w}) = \sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w})}p_0(\mathbf{t}|\mathbf{w})\exp\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{i,j}f_{i,j}(\mathbf{t})\right\}. \tag{9}$$

ここで，$\delta(i,j)$ は次のように定義されるクロネッカーのデルタである：

$$\delta(i,j) = \begin{cases} 1 & (i=j), \\ 0 & (i\neq j). \end{cases} \tag{10}$$

上記の式において，$f_{i,j}(\mathbf{t})$ は文書中で品詞 i と品詞 j の組が出現した回数を表しており，式 (8) はそのような文書単位の大域的な素性関数を使用した最大エントロピーモデルの一種であると見なすこともできる．

以上のように，提案手法で用いる確率モデルでは，同じ語形を持つすべての未知語の品詞を同時に考慮し，それらの品詞間の依存関係を表現することができる．ただし本手法では，異なる語形を持つ未知語の品詞はそれぞれ独立であると仮定する．つまり，ある語形を持つ未知語の集合は，別の語形を持つ未知語の集合とは別に独立して処理される．

2.2 解析手法

テストデータ $\mathbf{w}$ と局所的モデル $p_0(t|w)$ とモデルのパラメータ $\Lambda = \{\lambda_{1,1},\cdots,\lambda_{N,N}\}$ が与えられた場合に，その未知語の品詞 $\mathbf{t}$ を求めることを考える．1 つの方法として，あらゆる可能な $\mathbf{t}$ の候補の中から，$P(\mathbf{t}|\mathbf{w})$ を最大化するものを解として選ぶことが考えられる．しかしながら，あらゆる可能な候補の数は N^K 個存在するため，このような計算を厳密に行うのは一般的に困難である．

品詞タグ付け等の系列タグ付け問題でしばしば用いられる HMM や MEMM，CRF 等のモデルでは動的計画法を用いて効率的に計算が行われる．また特別な構造を持った確率モデルに対しては，効率的なアルゴリズムが適用できる場合もある[21]．しかしながら，ここでは，同じ語形を持つすべての未知語の品詞間の相互作用（依存関係）を考慮しており，その同時確率分布を細かい要素に分解することができないため，動的計画法等の効率的な計算手法を利用することはできない．そこで，確率分布から生成された有限個のサンプルを使用して近似解を求めることにする．

提案手法では，未知語の品詞推定結果の解 $\hat{\mathbf{t}} = \{\hat{t}_1,\cdots,\hat{t}_K\}$ を，次のようにして求めることにする：

$$\hat{t}_k = \operatorname*{argmax}_t P_k(t|\mathbf{w}). \tag{11}$$

ここで，$P_k(t|\mathbf{w})$ は局所的な文脈の集合 $\mathbf{w}$ が与えられた場合における，k 番目の未知語の品詞の周辺確率であり，次のように $\delta(t_k,t)$ の $P(\mathbf{t}|\mathbf{w})$ に関する期待値として計算することができる：

$$P_k(t|\mathbf{w}) = \sum_{\substack{t_1,\cdots,t_{k-1},t_{k+1},\cdots,t_K\\ t_k=t}} P(\mathbf{t}|\mathbf{w}), = \sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w})}\delta(t_k,t)P(\mathbf{t}|\mathbf{w}). \tag{12}$$

このような，ある確率分布に対する期待値は，その確率分布から生成された多数のサンプルを用いて近似することができる[12]．たとえば，$A(\mathbf{x})$ を確率変数 $\mathbf{x}$ の関数とし，$P(\mathbf{x})$ を $\mathbf{x}$ の確率分布とし，$\{\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(M)}\}$ を $P(\mathbf{x})$ から生成された M 個のサンプルとする．この場合，$A(\mathbf{x})$ の $P(\mathbf{x})$ に関する期待値は以下のように近似することができる：

$$\sum_{\mathbf{x}}A(\mathbf{x})P(\mathbf{x}) \simeq \frac{1}{M}\sum_{m=1}^{M}A(\mathbf{x}^{(m)}). \tag{13}$$

よって，確率分布 $P(\mathbf{t}|\mathbf{w})$ から生成された M 個のサンプル $\{\mathbf{t}^{(1)},\cdots,\mathbf{t}^{(M)}\}$ を用いることにより，品詞の周辺分布は次のように近似することができる：

$$P_k(t|\mathbf{w}) \simeq \frac{1}{M}\sum_{m=1}^{M}\delta(t_k^{(m)},t). \tag{14}$$

次に，確率分布からのサンプルをどのようにして得るかについて説明する．ここでは，ギブスサンプリングを用いてサンプルの生成を行う．ギブスサンプリングを用いることにより，高次元の確率分布から効率的にサンプルを生成することができる[23]．そのアルゴリズムを**図 1** に示す．このアルゴリズムでは，まず最初に初期状態 $\mathbf{t}^{(1)}$ を決める．そして，ある 1 つの確率変数について，それ以外の確率変数をすべて固定した条件付き確率からのサンプリングを行い状態を更新する，という手続きを繰り返していく．ギブスサンプリングは実装するのが容易であり，また生成されるサンプルの分布は元の確率の定常分布へ収束することが知

図 1 ギブスサンプリング
Fig. 1 Gibbs sampling.

```
1 t^(1) を初期化する
2 for m := 2 to M
3   for k := 1 to K
4     t_k^(m) ~ P(t_k|w, t_1^(m), ..., t_{k-1}^(m), t_{k+1}^(m-1), ..., t_K^(m-1))
```

られており，大規模な確率モデルを用いた研究でしばしば使用されている[8),19)]．図 1 における条件付き確率 $P(t_k|\mathbf{w}, t_1, \cdots, t_{k-1}, t_{k+1}, \cdots, t_K)$ は次のようにして容易に計算できる：

$$
\begin{aligned}
&P(t_k|\mathbf{w}, t_1, \cdots, t_{k-1}, t_{k+1}, \cdots, t_K) \\
&= \frac{P(\mathbf{t}|\mathbf{w})}{P(t_1, \cdots, t_{k-1}, t_{k+1}, \cdots, t_K|\mathbf{w})}, \\
&= \frac{\frac{1}{Z(\mathbf{w})} p_0(\mathbf{t}|\mathbf{w}) \exp\{\frac{1}{2}\sum_{k'=1}^{K}\sum_{\substack{k''=1 \\ k''\neq k'}}^{K} \lambda_{t_{k'}, t_{k''}}\}}{\sum_{t_k^*=1}^{N} P(t_1, \cdots, t_{k-1}, t_k^*, t_{k+1}, \cdots, t_K|\mathbf{w})}, \\
&= \frac{p_0(t_k|w_k) \exp\{\sum_{\substack{k'=1 \\ k'\neq k}}^{K} \lambda_{t_{k'}, t_k}\}}{\sum_{t_k^*=1}^{N} p_0(t_k^*|w_k) \exp\{\sum_{\substack{k'=1 \\ k'\neq k}}^{K} \lambda_{t_{k'}, t_k^*}\}}, \quad (15)
\end{aligned}
$$

ここで，最後の式は次の関係から得られる：

$$
\begin{aligned}
\frac{1}{2}\sum_{k'=1}^{K}\sum_{\substack{k''=1 \\ k''\neq k'}}^{K} \lambda_{t_{k'}, t_{k''}} &= \frac{1}{2}\sum_{\substack{k'=1 \\ k'\neq k}}^{K}\sum_{\substack{k''=1 \\ k''\neq k, k''\neq k'}}^{K} \lambda_{t_{k'}, t_{k''}} \\
&\quad + \sum_{\substack{k'=1 \\ k'\neq k}}^{K} \lambda_{t_{k'}, t_k}.
\end{aligned}
$$

初期状態 $\mathbf{t}^{(1)}$ は，$p_0(\mathbf{t}|\mathbf{w})$ を最大化する品詞の集合に設定した．

式 (11) によって得られる解は，$\mathbf{w}$ が与えられた場合の各未知語の品詞の周辺確率を最大化するものであり，このようにして最適解を得るアプローチは最大事後周辺確率（Maximum Posterior Marginal; MPM）推定として知られている．Finkel ら[8)] は，確率変数間に複雑な依存関係が存在する確率分布に対して最適解を求める際に，焼きなまし法を使用した．焼きなまし法を使う場合と比較すると，上述の手法では冷却スケジュール（cooling schedule）を決める必要がなく，さらに最も尤度の高い解を得るだけではなく，2 番目，3 番目に尤度の高い解等も周辺分布 $P_k(t|\mathbf{w})$ を用いて得ることができる．このように優先度が付いた複数の解が得られることは，提案手法を形態素解析用辞書の半自動作成等に応用する際には，作業者にランク付けされた候補のリストを提示することができるため有用であると思われる．

2.3 パラメータ推定手法

L 個の事例からなる訓練データ $\{\langle\mathbf{w}^1, \mathbf{t}^1\rangle, \cdots, \langle\mathbf{w}^L, \mathbf{t}^L\rangle\}$[⋆1] と局所的モデル $p_0(t|w)$ が与えられた場合に，式 (8) のモデルのパラメータ $\Lambda=\{\lambda_{1,1}, \cdots, \lambda_{N,N}\}$ を推定することを考える．ここでは，次式のように定義される Gaussian prior[6)] をパラメータの事前分布として利用した（C は定数であり，ハイパーパラメータ σ の値は 0.5 とした）：

$$
\begin{aligned}
\log P(\Lambda) &= -\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\lambda_{i,j}^2}{2\sigma^2} + C, \\
\frac{\partial}{\partial\lambda_{i,j}} \log P(\Lambda) &= -\frac{\lambda_{i,j}}{\sigma^2},
\end{aligned}
$$

最大事後確率推定を行うため，目的関数 $\mathcal{L}_\Lambda$ を定義し，この値を最大化する Λ を求める（下付き文字 Λ は，Λ によりパラメータ化されていることを表している）：

$$
\begin{aligned}
\mathcal{L}_\Lambda &= \log\prod_{l=1}^{L} P_\Lambda(\mathbf{t}^l|\mathbf{w}^l) + \log P(\Lambda), \\
&= \log\prod_{l=1}^{L} \frac{1}{Z_\Lambda(\mathbf{w}^l)} p_0(\mathbf{t}^l|\mathbf{w}^l) \\
&\quad \cdot \exp\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{i,j} f_{i,j}(\mathbf{t}^l)\right\} + \log P(\Lambda), \\
&= \sum_{l=1}^{L}\Bigg[-\log Z_\Lambda(\mathbf{w}^l) + \log p_0(\mathbf{t}^l|\mathbf{w}^l) \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{i,j} f_{i,j}(\mathbf{t}^l)\Bigg] + \log P(\Lambda). \quad (16)
\end{aligned}
$$

これを偏微分すると次のようになる：

$$
\begin{aligned}
\frac{\partial\mathcal{L}_\Lambda}{\partial\lambda_{i,j}} &= \sum_{l=1}^{L}\left[f_{i,j}(\mathbf{t}^l) - \frac{\partial}{\partial\lambda_{i,j}} \log Z_\Lambda(\mathbf{w}^l)\right] \\
&\quad + \frac{\partial}{\partial\lambda_{i,j}} \log P(\Lambda), \\
&= \sum_{l=1}^{L}\left[f_{i,j}(\mathbf{t}^l) - \sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w}^l)} f_{i,j}(\mathbf{t}) P_\Lambda(\mathbf{t}|\mathbf{w}^l)\right] \\
&\quad + \frac{\partial}{\partial\lambda_{i,j}} \log P(\Lambda). \quad (17)
\end{aligned}
$$

上記の $\mathcal{L}_\Lambda$ と $\partial\mathcal{L}_\Lambda/\partial\lambda_{i,j}$ を計算すれば，準ニュー

⋆1 提案手法では，同じ語形を持つ未知語の集合を 1 つの訓練事例として扱う．そのため，事例の数が L 個であるということは，訓練データの文書中には L 種類の異なる未知語の語形が存在することを意味する．

表 1 使用したコーパス
Table 1 Statistical information of corpora.

コーパス (言語)	品詞の数 (オープンクラス)	単語数 (未知語の出現数) [コーパスの使用部位] 訓練データ	テストデータ	ラベルなしデータ
CTB (C)	34 (28)	84,937 [sec. 1–270]	7,980 (749) [sec. 271–300]	6,801 [sec. 301–325]
PFR (C)	42 (39)	304,125 [Jan. 1–Jan. 9]	370,627 (27,774) [Jan. 10–Jan. 19]	445,969 [Jan. 20–Jan. 31]
EDR (J)	15 (15)	1,018,561 [$id = 5n + 0$]	1,020,572 (30,765) [$id = 5n + 1$]	3,065,914 [$id = 5n + 2, 5n + 3, 5n + 4$]
KUC (J)	40 (36)	198,514 [Jan. 1–Jan. 8]	31,302 (2,477) [Jan. 9]	41,227 [Jan. 10]
RWC (J)	66 (55)	487,333 [1–10,000th sentences]	190,571 (11,177) [10,001–14,000th sentences]	210,096 [14,001–18,672th sentences]
GEN (E)	47 (36)	243,180 [1–10,000th sentences]	123,386 (7,775) [10,001–15,000th sentences]	134,380 [15,001–20,546th sentences]
SUS (E)	125 (90)	74,902 [sec. A01–08, G01–08, J01–08, N01–08]	37,931 (5,760) [sec. A09–12, G09–12, J09–17, N09–12]	37,593 [sec. A13–20, G13–22, J21–24, N13–18]
WSJ (E)	45 (33)	912,344 [sec. 0–18]	129,654 (4,253) [sec. 22–24]	131,768 [sec. 19–21]

トン法を用いて最適な Λ を求めることができる．本研究では，L-BFGS 法[11] を使用して最適解を求めた[⋆1]．ただし，式 (16) 中の $Z_\Lambda(\mathbf{w}^l)$ と（式 (9) 参照），式 (17) 中の項には，あらゆる品詞の組合せについて数え上げる計算が含まれており，計算量が非常に大きい．そこで，Rosenfeld ら[19] が行ったのと同様に，モンテカルロ法を用いてこれらの値を近似計算する．$Z_\Lambda(\mathbf{w}^l)$ は，$p_0(\mathbf{t}|\mathbf{w}^l)$ から生成された M' 個のサンプル $\{\mathbf{t}^{(1)}, \cdots, \mathbf{t}^{(M')}\}$ を使用して，次のように計算できる：

$$Z_\Lambda(\mathbf{w}^l) = \sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w}^l)} p_0(\mathbf{t}|\mathbf{w}^l) \exp\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{i,j} f_{i,j}(\mathbf{t})\right\}, \simeq \frac{1}{M'} \sum_{m=1}^{M'} \exp\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{i,j} f_{i,j}(\mathbf{t}^{(m)})\right\}. \quad (18)$$

式 (17) 中の項は，$P_\Lambda(\mathbf{t}|\mathbf{w}^l)$ からギブスサンプリングにより生成された M' 個のサンプル $\{\mathbf{t}^{(1)}, \cdots, \mathbf{t}^{(M')}\}$ を使用して，次のように計算できる：

$$\sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w}^l)} f_{i,j}(\mathbf{t}) P_\Lambda(\mathbf{t}|\mathbf{w}^l) \simeq \frac{1}{M'} \sum_{m=1}^{M'} f_{i,j}(\mathbf{t}^{(m)}). \quad (19)$$

実験では，サンプルの初期状態 $\mathbf{t}^{(1)}$ には，訓練データ中の正解ラベルを使用した．

⋆1 提案手法ではサンプリングを用いた近似を行うため，実験の際にはしばしば計算が完全には収束しなかった．そのような場合には，L-BFGS の反復計算を途中で中断し，その時点で得られている値を解として用いた．

2.4 ラベルなしデータの利用

提案したモデルでは，テストデータにラベルなしデータを単純に結合して，テスト時にその結合されたデータをまとめて解析することにより，容易にラベルなしデータを利用することができると思われる．つまり，テストデータの量を増やせば，有用な局所的な情報を持った事例の数も増加すると思われるが，そのような事例の品詞は容易に予測することができ，他の事例の品詞を予測する際に大域的な情報として利用できる可能性がある．たとえば，テストデータ中に普通名詞であるかサ変名詞であるかを判定するのが困難な未知語が存在した場合に，ラベルなしデータ中に同じ語形の単語が「〜する」という文脈で出現していれば，ラベルなしデータを使用することによりテストデータ中のそのような曖昧な単語に対する解析精度を向上させられる可能性がある．このように本手法でラベルなしデータを利用する場合，ラベルなしデータはテスト時にのみ利用し，訓練時はラベルなしデータを利用しない場合とまったく同じ手順でモデルのパラメータ推定を行う．

3. 実　　験

3.1 使用したデータ

実験には次の 8 つのコーパスを使用した（**表 1**）：Penn Chinese Treebank コーパス 2.0（CTB），PFR コーパス（PFR），EDR コーパス（EDR），京大コーパス version 2（KUC），RWCP コーパス（RWC），GENIA コーパス 3.02p（GEN），SUSANNE コーパス（SUS），Penn Treebank WSJ コーパス（WSJ）．これらはすべて品詞タグ付きコーパスであり，中国語（C），日本語（J），英語（E）のいずれかの言語のコー

パスである．これらのコーパスをそれぞれ，訓練データ，テストデータ，ラベルなしデータの 3 つの部分に分割した．ラベルなしデータ中の未知語の品詞はあらかじめ削除しておいた．

表 1 に，使用した各コーパスの，言語，品詞の数，オープンクラス（未知語がとることのできる品詞，説明は後述）の数，訓練データ・テストデータ・ラベルなしデータのサイズ，それらのデータの分割方法[⋆1]，を示す．以下の節で述べられるすべての実験結果は，各コーパスをこの表に示されるとおりに訓練データやテストデータに分割して実験することによって得られた結果である．

テストデータとラベルなしデータ中の単語に対しては，訓練データ中に存在しない単語をここでは未知語と定義する．テストデータに含まれる未知語の数を表 1 の括弧内に示す．本実験における未知語の品詞推定の精度は，ここに示された数だけ存在するテストデータ中の未知語のうち，正しく品詞を推定することができたものの割合として計算される．

この実験では，単語の区切りと既知語の品詞は，コーパス中の正しい情報を使用した．単語分割や品詞タグ付けの後処理として実際に未知語の品詞推定を行う場合には，単語区切りや既知語の品詞に誤りが含まれると思われる．しかしながら，そのような他の要因を除いて未知語の品詞推定精度のみを分析するためと，データへの加工は最小限にすることで同一のデータを用意して追試等を行うことが容易になるように，このような設定で実験を行った．

3.2 実験の手順

図 2 に，各コーパスを用いて実験を行う場合の手順を示す．まず，訓練データを前後 2 つに等分割し，片方は訓練データ A とし，もう片方は訓練データ B とする（図 2，*1）．次に，訓練データ A には出現するが訓練データ B には出現しない単語と，訓練データ B には出現するが訓練データ A には出現しない単語を特定する．そして，そのような単語を訓練データ中の擬似的な未知語として定義する．このような（2 分割）交差検定を用いることにより，訓練データ中における未知語を定義して，未知語に対する訓練事例を作成することができる[⋆2]．つまり交差検定を用いて，

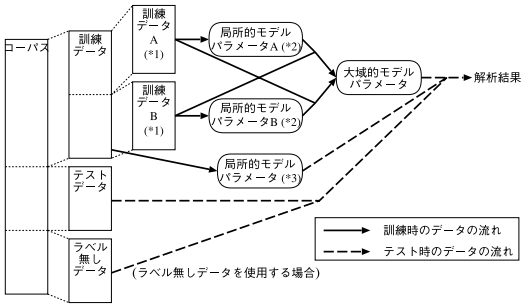

図 2 実験の手順
Fig. 2 Experimental procedure.

訓練データをさらに小さな別の新しい訓練データとテストデータに分けることを 2 回繰り返すことにより，その新しい訓練データ中には存在しないがその新しいテストデータ中に存在する単語を擬似的な未知語として定義する．これらの擬似的な未知語の持つ品詞を，オープンクラスの品詞として定義し，オープンクラスの品詞のみを未知語に対する品詞の候補として考えることにする．よって，未知語のとりうる品詞の数として定義した N の値は，オープンクラスの品詞の数と等しい．

訓練時には，2 種類のモデルのパラメータを推定する必要がある．1 つは，$p_0(t|w)$ の値の計算に必要な局所的モデルのパラメータであり，もう 1 つは大域的モデルのパラメータ（$\lambda_{i,j}$ の値）である．大域的モデルパラメータを推定する際には，局所的モデルパラメータと訓練データが必要となるが，局所的モデルパラメータを推定するために使用した訓練データを用いて大域的モデルパラメータを推定すると，局所的モデルパラメータが訓練データに過適応した場合に適切な大域的モデルパラメータを推定できない恐れがある．そこで，大域的モデルパラメータを推定する際にはまず，訓練データ A と訓練データ B を使用して，局所的モデルパラメータ A と局所的モデルパラメータ B をそれぞれ学習させる（図 2，*2）．大域的モデルパラ

⋆1 CTB，KUC，WSJ コーパスの訓練データとテストデータへの分割方法については，それぞれ文献 4)，20)，24) のような既存研究で用いられている標準的な分割方法が存在するため，それらの分割方法に従った．

⋆2 このような擬似的な未知語を生成するためによく用いられる方法として，コーパス中に 1 回しか出現しない単語を未知語と見なして使用する方法がある[15]．このような単語は，hapax legomena と呼ばれ，実際の未知語に近い性質を持つことが知られている[2]．このような単語は，次のように，(交差検定法の特殊な場合である）leave-one-out 法によって集められたものと解釈することができる．まず，コーパスから単語を 1 つ取り出し，コーパスの残りの部分を訓練データと考えることにする．もし取り出された単語がその訓練データ中に存在しなければ，その語を未知語と見なす．その後，取り出された単語をコーパスに戻し，再び別の単語を取り出して同様の処理を行う．これを，コーパス中のすべての単語に対して繰り返すと，結局コーパス中に 1 回しか出現しない単語が未知語と見なされることになる．しかしながらこの方法で定義される未知語は，文書中に 1 度しか出現せず大域的な情報を利用することができないため，大域的なモデルの学習には利用することができない．そのため，我々の実験では 2 分割交差検定を使用した．

メータを推定する際には，訓練データ A に含まれる未知語に対しては局所的モデルパラメータ B を，訓練データ B に含まれる未知語に対しては局所的モデルパラメータ A を使用して，$p_0(t|w)$ の値を計算する．テスト時には，テストデータ中の未知語に対する $p_0(t|w)$ の値は訓練データ全体から学習された局所的モデルパラメータ（図 2，*3）を用いて計算する．そして，この局所的モデルパラメータを大域的モデルパラメータとともに利用することにより，未知語の品詞推定を行い，解析結果を得る．

訓練データやテストデータ中に 1 度しか出現しない未知語に対しては大域的な情報が利用できないため，2 回以上出現した未知語に対してのみ提案手法を用いた処理を行う．テストデータ中に 1 回しか出現しない未知語に対しては，$p_0(t|w)$ の値を最大化するような品詞を選ぶことにより，局所的な情報のみを使用して品詞推定を行う．

ラベルなしデータを利用した実験では，テストデータとラベルなしデータは単純に結合され，その結合されたデータ全体が与えられた場合の確率を最大化する品詞の集合が解として選択される．

解析時にギブスサンプリングにより生成するサンプルの数 M や，パラメータ推定時にモンテカルロ法で使用するサンプルの数 M' は，どちらの値も 100 とした．

3.3 局所的モデル

提案手法では，局所的文脈 w で条件付けられた品詞 t の確率を与える局所的モデル $p_0(t|w)$ が使用される（式 (8) 参照）．この実験では，局所的モデルの計算には最大エントロピー（Maximum Entropy; ME）モデルを使用することにした．$p_0(t|w)$ は ME モデルを使用して次のように計算できる[3]：

$$p_0(t|w) = \frac{1}{Y(w)} \exp\left\{\sum_{h=1}^{H} \alpha_h g_h(w,t)\right\}, \quad (20)$$

$$Y(w) = \sum_{t=1}^{N} \exp\left\{\sum_{h=1}^{H} \alpha_h g_h(w,t)\right\}. \quad (21)$$

ここで，$g_h(w,t)$ は 2 値の素性関数である．局所的文脈 w は，未知語に関する次のような情報を持っているとする：

- 未知語の前後 2 つの単語の品詞：τ_{-2}，τ_{-1}，τ_{+1}，τ_{+2}[⋆1]．
- 未知語自身と未知語の前後 2 つの単語の出現形：ω_{-2}，ω_{-1}，ω_0，ω_{+1}，ω_{+2}．
- 未知語を構成する文字種：$y_1, \cdots, y_{|\omega_0|}$．文字種としては次の 6 つを使用した：アルファベット，数字，記号，漢字，ひらがな，カタカナ．

素性関数 $g_h(w,t)$ は，次の例のように，w と t がある条件を満たす場合にのみ 1 を，それ以外では 0 を返す関数である：

$$g_{123}(w,t) = \begin{cases} 1 & (\omega_{+1} = \text{“先生”} \ \&\ \tau_{+1} = \text{“名詞”} \ \&\ t = 5), \\ 0 & (\text{otherwise}). \end{cases}$$

使用した基本素性を**表 2** に示す．これらの基本素性は，Ratnaparkhi[18] や内元ら[24] によって使用された素性を参考に決めたものである．ME モデルでは，素性の組合せは明示的に別の素性として与えなければならない．そこでこれらの基本素性に加えて，すべての異なる 2 つの基本素性を対とした組合せ素性も，素性として使用する．

式 (20) 中のパラメータ α_h の値は，訓練データ中に存在するオープンクラスの品詞を持つすべての単語を用いて推定した．

表 2 局所的モデルに使用した基本素性
Table 2 Basic features used for initial distribution.

言語	素性
英語	ω_0 の 4 文字までの語頭
	ω_0 の 4 文字までの語尾
	ω_0 が数字を含むか
	ω_0 が大文字を含むか
	ω_0 がハイフンを含むか
中国語 日本語	ω_0 の 2 文字までの語頭
	ω_0 の 2 文字までの語尾
	y_1
	$y_{\|\omega_0\|}$
	y_1 & $y_{\|\omega_0\|}$
	$\bigcup_{i=1}^{\|\omega_0\|}\{y_i\}$ (文字種の集合)
英語 中国語 日本語	$\|\omega_0\|$ (ω_0 の長さ)
	τ_{-1}
	τ_{+1}
	τ_{-2} & τ_{-1}
	τ_{+1} & τ_{+2}
	τ_{-1} & τ_{+1}
	ω_{-1} & τ_{-1}
	ω_{+1} & τ_{+1}
	ω_{-2} & τ_{-2} & ω_{-1} & τ_{-1}
	ω_{+1} & τ_{+1} & ω_{+2} & τ_{+2}
	ω_{-1} & τ_{-1} & ω_{+1} & τ_{+1}

3.4 実験結果

3.4.1 大域的な情報の利用による解析精度の変化

実験結果を**表 3** に示す．この表において，**局所**，**局所＋大域**，**局所＋大域＋ラベルなしデータ**はそれぞ

⋆1 訓練時およびテスト時には，既知語に対する品詞はコーパスから正しい品詞が与えられるものとする．もし，これらの局所的文脈が別の未知語を含んでいた場合，その品詞は *Unk* という特殊な品詞で表現することとする．

表 3 未知語の品詞推定の実験結果
Table 3 Results of POS guessing of unknown words.

コーパス (言語)	未知語の解析精度 (%) (誤りの数) [p 値] ⟨2 回以上出現した未知語の数 ⟩		
	局所	局所+大域	局所+大域+ラベルなしデータ
CTB (C)	73.97 (195)	76.77 (174) [0.0403] ⟨344⟩	76.77 (174) [0.0422] ⟨361⟩
PFR (C)	68.25 (8,819)	70.09 (8,306) [0.0000] ⟨16,019⟩	70.54 (8,181) [0.0000] ⟨18,861⟩
EDR (J)	95.92 (1,254)	95.92 (1,254) [1.0000] ⟨9,838⟩	95.89 (1,263) [0.1282] ⟨18,064⟩
KUC (J)	76.46 (583)	77.55 (556) [0.0010] ⟨788⟩	77.71 (552) [0.0013] ⟨936⟩
RWC (J)	77.58 (2,506)	78.64 (2,387) [0.0000] ⟨5,044⟩	78.65 (2,386) [0.0000] ⟨5,878⟩
GEN (E)	88.91 (862)	88.91 (862) [1.0000] ⟨4,094⟩	88.80 (871) [0.5277] ⟨4,515⟩
SUS (E)	78.68 (1,228)	79.76 (1,166) [0.0034] ⟨3,210⟩	79.58 (1,176) [0.0311] ⟨3,583⟩
WSJ (E)	83.56 (699)	83.71 (693) [0.2569] ⟨1,412⟩	83.14 (717) [0.3952] ⟨1,627⟩

れ，従来手法と同じアプローチである局所的な情報のみを利用した場合，提案手法により局所的な情報と大域的な情報を利用した場合，提案手法により局所的な情報と大域的な情報を利用してさらにラベルなしデータも利用した場合，の未知語の品詞推定結果を表す．局所的な情報のみを利用した場合の解は，次のようにして局所的モデルの値を最大化するような品詞 $\hat{\mathbf{t}} = \{\hat{t}_1, \cdots, \hat{t}_K\}$ として求めた：

$$\hat{t}_k = \operatorname*{argmax}_t p_0(t|w_k). \qquad (22)$$

この表では，精度，誤りの数，局所的な情報のみを使用した場合と比較して統計的検定を行った場合の p 値，テストデータ中に 2 回以上出現した未知語の数，を示している．統計的検定は，同じ語形を持つ未知語の集合を単位として，対応のある t 検定を行った[9]．具体的には，未知語の各語形ごとに品詞が誤って推定された未知語の数を集計し，比較を行う 2 つのシステム間で，未知語の各語形ごとにそれらの値の差をとる．2 つのシステムの解析精度に差は存在しないのであれば，この差の値の平均値は 0 であるはずなので，差の平均値が 0 であることを帰無仮説として検定を行う[⋆1]．

実験の結果，表 3 に示されるように，CTB，PFR，KUC，RWC，SUS，WSJ コーパス（表 1 に示されるように訓練・テストデータの分割を行ったもの）に対して，大域的な情報も用いることにより，局所的な情報しか利用しない場合に比べて精度を向上させることができた．そのうちの，CTB，PFR，KUC，RWC，SUS コーパスに関しては，2 つの方式間の精度には $p < 0.05$ で統計的に有意差が見られた．EDR，GEN，WSJ コーパスでは，精度はほとんど変化しなかった．

ラベルなしデータを使用した場合，提案手法において処理対象となる，テストデータに 2 回以上出現した未知語の数は増加している．ラベルなしデータを使用しない場合と比べて，PFR，KUC 等のコーパスでは精度が向上したが，SUS，WSJ 等のコーパスでは逆に精度が低下しており，単純にラベルなしデータをテストデータに加えて解析を行うだけでは精度が向上しない場合が見られた．

表 4 2 回以上出現した未知語の解析精度
Table 4 Accuracy for non-unique unknown words.

コーパス (言語)	2 回以上出現した未知語の解析精度 (%)	
	局所	局所+大域 (増減)
CTB (C)	80.23	86.34 (+6.11)
PFR (C)	65.95	69.15 (+3.20)
EDR (J)	94.72	94.72 (0.00)
KUC (J)	74.24	77.66 (+3.42)
RWC (J)	74.21	76.57 (+2.36)
GEN (E)	89.35	89.35 (0.00)
SUS (E)	80.37	82.31 (+1.94)
WSJ (E)	87.82	88.24 (+0.42)

表 4 は，提案手法において処理対象となる，テストデータに 2 回以上出現した未知語のみに対する解析精度を示している．多くのコーパス（表 1 に示されるよ

⋆1 提案手法では，同じ語形を持つ未知語の品詞間の依存性を考慮しており，個々の未知語のテスト結果に対する独立性を仮定できないため，各未知語を単位として検定を行うことはできない．そのため，同じ語形を持つ未知語をグループ化して，同じ語形を持つ未知語の集合を単位として検定を行った．この検定方法は，Matched Pairs Sentence Segment Word Error (MAPSSWE) Test と呼ばれているものと同等のものである (http://www.nist.gov/speech/tests/sigtests/mapsswe.htm)．

表 5 正しく品詞が推定された未知語の品詞別での数の変化
Table 5 Ordered list of increased/decreased number of correctly tagged words.

PFR (C)		RWC (J)		SUS (E)	
+143	nz（その他の固有名詞）	+35	名詞-固有名詞-組織	+36	NP（固有名詞）
+111	vn（名詞的動詞）	+30	名詞-固有名詞-地域	+9	NNU（単位名）
+105	nr（人名）	+28	名詞-固有名詞-人名-姓	+7	NN（普通名詞）
+58	j（略語）	+22	名詞-固有名詞-人名-名	+6	JJ（形容詞）
+52	ns（地名）	+8	名詞-固有名詞	+2	VV（本動詞の原形）
+37	d（副詞）	+6	名詞-サ変接続	+2	VVN（本動詞の過去分詞形）
...	...	...	...	...	...
−33	v（動詞）	−1	形容詞	−1	CS（従位接続詞）
−59	m（数）	−21	名詞	−1	VVG（本動詞の現在分詞形）

うに訓練・テストデータの分割を行ったもの）において，大域的な情報を用いることにより，これらの未知語に対する解析精度が改善されていることが分かる．

3.4.2 品詞別に見た解析精度の変化

表 5 は，PFR，RWC，SUS コーパス（表 1 に示されるように訓練・テストデータの分割を行ったもの）において，大域的な情報を利用することにより，正しく品詞を推定できるようになった未知語数の増減を品詞別で示している．中国語の PFR コーパスと日本語の RWC コーパスでは，大域的な情報を使うことによって多くの固有名詞が正しく解析できるようになっている．

英語では，固有名詞が大文字で書き始められるため，固有名詞と普通名詞の区別を行うのは容易であることが多い．しかしながら中国語と日本語では，そのような習慣はないため，局所的な情報のみを利用して区別を行うことはしばしば難しい．大域的な情報を利用しても，英語のコーパス（GEN，WSJ コーパス）ではあまり大きな精度の向上が得られなかった原因として，英語ではこのような固有名詞に関する曖昧性が小さく，大域的な情報の必要性が少ないことが考えられる．また日本語のコーパスである EDR コーパスでも精度は向上しなかったが，このコーパスで用いられている品詞は種類が少なく，普通名詞と固有名詞が区別されていないため曖昧性が少なかったことがその原因として考えられる．

3.4.3 未知語の出現頻度と解析精度との関係

大域的な情報を利用して未知語の品詞推定を行う本手法では，テストデータ中に同じ語形を持つ未知語がより多く出現していれば，それだけたくさんの情報を考慮できる可能性がある．そこで，テストデータ中での未知語の出現頻度と，その解析精度の関係を調べた．未知語の数が一番多い PFR コーパスについての結果をまとめたものを**表 6** に示す．この表において，**頻度**はコーパス中に同じ語形を持つ未知語が何回出現したかを，**トークン数**はそのような未知語の延べ数を，**タイプ数**はそのような未知語の種類数を表す．頻度が 1 回の場合は，大域的な情報は使用されないため，解析精度の増減はない．頻度が 8〜10 回の場合には解析精度が大きく増加しているが，頻度と，解析精度やその増減との間には単純な比例関係は見られなかった．つまり，出現頻度が多くなれば解析精度が増加するとは，必ずしもいえない．

表 6 PFR コーパスにおける未知語の出現頻度と精度の関係
Table 6 Frequency and accuracy of unknown words in PFR corpus.

頻度	未知語の解析精度 (%)			トークン数	タイプ数
	局所	局所+大域	増減		
1	71.38	71.38	0.00	11,755	11,755
2	68.67	70.21	+ 1.54	5,186	2,593
3	65.19	67.73	+ 2.54	2,904	968
4	66.23	69.44	+ 3.21	1,996	499
5	66.55	68.00	+ 1.45	1,100	220
6	63.68	67.59	+ 3.91	972	162
7	64.06	67.74	+ 3.68	651	93
8	62.90	71.98	+ 9.08	496	62
9	59.18	68.60	+ 9.42	414	46
10	70.37	82.96	+12.59	270	27
≥11	62.71	67.59	+ 4.88	2,030	121

提案手法では，同一の語形を持つ未知語が文書中の別の場所に有用な局所的な情報とともに出現していた場合に，その情報を間接的に利用することができる．しかしながら，文書中の別の場所に大きな曖昧性を持って出現していた場合，その情報は品詞推定の役には立たず，それらが出現していない場合と比べて利用できる情報は変わらないと思われる．さらに，文書中の別の場所に紛らわしい局所的情報とともに出現していた場合，その情報を利用することによって誤りを起こすことも考えられる．ラベルなしデータを使用した実験において，それを使用しなかった場合と比較して精度が低下した事例を調べたところ，ラベルなしデータの中に，誤った品詞推定結果を与えるような局所的な情報とともに出現している未知語がしばしば見られ

表 7 精度のばらつきと焼きなまし法との比較
Table 7 Results of multiple trials and comparison to simulated annealing.

コーパス (言語)	平均値 (%)± 標準偏差 (%)	
	周辺確率最大化	焼きなまし法
CTB (C)	76.49±0.21	76.45±0.27
PFR (C)	70.33±0.15	70.30±0.11
EDR (J)	95.94±0.03	95.94±0.02
KUC (J)	77.61±0.27	77.63±0.18
RWC (J)	78.56±0.10	78.62±0.08
GEN (E)	88.92±0.05	88.91±0.07
SUS (E)	79.70±0.24	79.65±0.20
WSJ (E)	83.78±0.06	83.82±0.10

た．ラベルなしデータを追加しても必ずしも精度が向上しなかった原因の 1 つとして，このような大域的な情報を考慮した場合の副作用の影響が考えられる．

3.4.4 乱数の影響と焼きなまし法との比較

提案手法では，訓練時にもテスト時にも乱数により生成されたサンプルを使用して近似計算を行うため，実験結果はサンプリングで使用する乱数列の影響を受ける．その影響について調べるため，異なる擬似乱数列を使用して 10 回の実験を行い，精度のばらつきの度合いを測った．さらに，Finkel ら[8] が行ったように，焼きなまし法を用いて解析する方法も試みた．その場合，式 (1) の逆温度 β を，$\beta = 1$ から $\beta \approx \infty$ まで変化させた．

実験の結果を**表 7** に示す．この表は，10 回の実験を行って得られた精度の平均値と標準偏差を示しており，**周辺確率最大化**と**焼きなまし法**はそれぞれ式 (11) を使用して解析を行った場合と焼きなまし法を使用して解析を行った場合を表している．この実験結果からは，乱数列を原因とする精度のばらつきや，解析方法の違いによる差はあまり見られなかった．

3.4.5 SVM を用いた既存手法との比較

提案手法を，既存の別の未知語品詞推定手法と比較するために実験を行った．ここでは，局所的な情報のみを利用する，サポートベクターマシン（Support Vector Machine; SVM）に基づく未知語の品詞推定手法[16] との比較を行った．この手法は，SVM に one-versus-rest 法を適用して多値分類を行うことで，未知語の品詞を推定する．SVM のパラメータ等は文献 16) に合わせて，カーネルは 2 次の多項式カーネルを使用して C の値は 1 とした．素性は，表 2 に示された基本素性を用いた．SVM では，多項式カーネルを利用することにより素性の組合せを自動的に考慮することができるため，組合せ素性は使用しなかった．

実験の結果を**表 8** に示す．この表において，**局所 (SVM)**，**局所 (ME)**，**局所+大域**は，それぞれ SVM を用いた既存手法を使用した場合，提案手法で用いている ME モデルに基づく局所的モデルを使用した場合，提案手法により局所的な情報と大域的な情報を利用した場合，の未知語の解析精度を表す．また，SVM を用いた既存手法と比較して統計的検定を行った場合の p 値も示してある．CTB，PFR，KUC，RWC，GEN，WSJ の 6 つのコーパスにおいて，大域的情報を用いた提案手法は，SVM を用いた手法よりも高い精度を得た．その中で CTB，PFR，GEN，WSJ コーパスにおいては，$p < 0.05$ で解析精度に有意差があった．PFR，KUC，RWC コーパスを用いた実験においては，ME モデルに基づく局所的モデルのみを使用した場合は SVM を用いた手法よりも精度は低かったが，大域的な情報も用いた場合は SVM を用いた手法よりも精度が高かった．

表 8 既存手法との比較
Table 8 Comparison to an existing method.

コーパス (言語)	未知語の解析精度 (%) [p 値]		
	局所 (SVM)	局所 (ME)	局所+大域
CTB (C)	73.43	73.97 [0.5721]	76.77 [0.0155]
PFR (C)	68.57	68.25 [0.0758]	70.09 [0.0000]
EDR (J)	96.02	95.92 [0.0459]	95.92 [0.0532]
KUC (J)	76.83	76.46 [0.4095]	77.55 [0.1522]
RWC (J)	78.40	77.58 [0.0007]	78.64 [0.4496]
GEN (E)	88.48	88.91 [0.0288]	88.91 [0.0225]
SUS (E)	79.97	78.68 [0.0001]	79.76 [0.6089]
WSJ (E)	83.05	83.56 [0.0482]	83.71 [0.0133]

局所 (SVM)，局所 (ME)，局所＋大域の各手法を用いて，8 つのすべてのコーパスで実験を行うのに要した学習時間/テスト時間は，それぞれ，367 時間/10,547 秒，5 時間/54 秒，15 時間/84 秒であり（実験は，Opteron 250 プロセッサの計算機で行った），大域的な情報を利用する提案手法は SVM を用いた手法よりも学習およびテストに必要な時間が少なかった．

4. 関 連 研 究

広い意味での大域的な情報を用いた自然言語処理の研究は，従来から様々な方法が試みられている．特に，未知語の品詞推定と若干似たタスクである固有表現抽出において，大域的な情報を利用したいくつかの手法が提案されている．Chieu ら[7] は，局所的な素性だけではなく大域的な素性も用いた，ME モデルに基づく

固有表現抽出手法を提案している．彼らの手法では，「この単語が文頭以外の場所に最初に出現したとき，大文字で書き始められていたかどうか」というような，大域的な素性をいくつか利用する．このような大域的な素性は，解析中に変化しない静的なものであるため，局所的な素性と同じように扱うことができる．そのため，解析には Viterbi アルゴリズムを使用している．この手法は効率的であるが，ラベル間の相互作用は考慮しない．

Finkel ら[8] は，大域的な情報を利用した情報抽出手法を提案した．彼らの手法では，同じ語形を持つ固有表現は同じラベルを持つ傾向があるという，「ラベルの一貫性」の性質を利用している．この手法では，CRF に基づく局所的なモデルと，対数線形モデルに基づく大域的なモデルの 2 つを定義している．そして，これらの 2 つのモデルの確率分布の積をとることにより最終的なモデルを得ているが，これは 2 つのモデルを等しく重み付けして対数線形補間[10] を行ったモデルと解釈することができる．この方法では，文書全体でのラベル間の相互作用を考慮し，ギブスサンプリングと焼きなまし法を用いて解析を行っている．我々の提案手法は，彼らの手法に非常に近いといえる．しかしながら彼らの手法では，大域的なモデルのパラメータはラベルの相対頻度から求めたり人手によって設定したりしていたが，提案手法では，大域的モデルのパラメータは目的関数を最大化する解として訓練データから得られる．

自然言語処理において大域的な情報を利用する 1 つのアプローチとして，前述したラベルの一貫性を利用する方法があるが，これと似たアプローチは他のタスクでもこれまでに利用されている．高村ら[25] は，物理学で用いられるイジングスピンモデルを単語の感情極性判定に応用した．イジングスピンモデルは，物質中の個々の電子が上向きか下向きの 2 つのうちのどちらか一方の状態（スピン）を持つとして，その状態の確率分布を与える．各電子の状態は相互作用し，隣り合った電子は同じスピンを持ちやすいという傾向がある．彼らの手法では，単語の感情極性（「望ましい」か「望ましくない」か）を電子のスピンと同じように見なすことにより，辞書の見出し語とその語釈文にある単語とは同じ感情極性を持ちやすいという性質をモデル化した．彼らの手法では，平均場近似と最大事後周辺確率推定を用いて系の状態を計算している．

Yarowsky[22] は，ラベルなしデータを用いた語義の曖昧性解消のための方法を提案した．彼の方法では確率的なモデルを明示的に考えているわけではないが，“one sense per discourse” と呼ばれるラベルの一貫性と，“one sense per collocation” と呼ばれる局所的な情報の両者を用いて教師なし学習を行っている．

大域的な情報を用いるアプローチとして，ラベルの一貫性を利用する以外の方法も試みられている．Rosenfeld ら[19] は，whole-sentence exponential language model を提案した．この手法では，与えられた文 s の生成確率を次のように計算する：

$$P(s) = \frac{1}{Z} p_0(s) \exp\left\{ \sum_i \lambda_i f_i(s) \right\}.$$

ここで，$p_0(s)$ は s の初期分布であり，trigram 等の任意の言語モデルを使用することができる．$f_i(s)$ は素性関数であり，文単位の素性を扱うことができる．我々の手法における式 (7) で定義される $f_{i,j}(\mathbf{t})$ を素性関数と見なすと，式 (8) は上の式と本質的には等しいといえる．彼らのモデルは，shallow parser によって得られる統語的な素性等の，任意の文単位の素性を利用することができる．モデルの計算にはギブスサンプリングや他のサンプリング手法が利用され，モデルのパラメータは generalized iterative scaling 法により訓練データから推定された．彼らは文全体のモデル化を対象としたが，この手法は文書全体のモデル化にもそのまま応用できる．その場合，本稿で検討したようにラベルなしデータを取り入れることができる．このような，広範囲の事象全体を対数線形モデルを用いてモデル化し，MCMC 法を利用して計算するアプローチは，他のタスクへの応用も可能な柔軟な枠組みであると思われる．

5. 結　　論

本稿では，局所的な情報だけではなく大域的な情報も用いて未知語の品詞推定を行う手法を提案した．この方法は，同じ語形を持つすべての未知語の品詞間の相互作用を考慮して，文書全体に対するモデル化を行う．モデルのパラメータは，サンプリング手法を用いて訓練データより推定した．

実験により局所的な情報のみを使用する場合と比較した結果，提案手法を用いることにより，特に中国語と日本語において，高い精度で未知語の品詞を推定できることを確認した．

本手法でラベルなしデータを利用することも試みたが，ラベルなしデータを使わない場合と比べて精度が低下する場合が見られた．ラベルなしデータを利用する方法を改善することは，今後の課題の 1 つである．また本稿では未知語の品詞推定のみを対象としたが，

未知語の単語分割を正確に行うことも重要な問題であり，大域的な情報を用いて未知語の単語分割を行うことも残された課題の 1 つである．

参 考 文 献

1) Asahara, M.: Corpus-based Japanese morphological analysis, Nara Institute of Science and Technology, Doctor's Thesis (2003).
2) Baayen, H. and Sproat, R.: Estimating Lexical Priors for Low-Frequency Morphologically Ambiguous Forms, *Computational Linguistics*, Vol.22, No.2, pp.155–166 (1996).
3) Berger, A.L., Pietra, S.A.D. and Pietra, V.J.D.: A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, Vol.22, No.1, pp.39–71 (1996).
4) Bikel, D.M. and Chiang, D.: Two Statistical Parsing Models Aplied to the Chinese Treebank, *Proc. 2nd Chinese Language Processing Workshop*, pp.1–6 (2000).
5) Chen, C., Bai, M. and Chen, K.: Category Guessing for Chinese Unknown Words, *Proc. NLPRS '97*, pp.35–40 (1997).
6) Chen, S. and Rosenfeld, R.: A Gaussian Prior for Smoothing Maximum Entropy Models, Technical Report CMUCS-99-108, Carnegie Mellon University (1999).
7) Chieu, H. and Ng, H.: Named Entity Recognition: A Maximum Entropy Approach Using Global Information, *Proc. COLING 2002*, pp.190–196 (2002).
8) Finkel, J., Grenager, T. and Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, *Proc. ACL 2005*, pp.363–370 (2005).
9) Gillick, L. and Cox, S.: Some Statistical Issues in the Comparison of Speech Recognition Algorithms, *Proc. ICASSP 1989*, pp.532–535 (1989).
10) Klakow, D.: Log-linear interpolation of language models, *Proc. ICSLP '98*, pp.1695–1699 (1998).
11) Liu, D.C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, Vol.45, No.3, pp.503–528 (1989).
12) MacKay, D.J.C.: *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press (2003).
13) Mikheev, A.: Automatic Rule Induction for Unknown-Word Guessing, *Computational Linguistics*, Vol.23, No.3, pp.405–423 (1997).
14) Mori, S. and Nagao, M.: Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis, *Proc. COLING '96*, pp.1119–1122 (1996).
15) Nagata, M.: A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context, *Proc. ACL '99*, pp.277–284 (1999).
16) Nakagawa, T., Kudoh, T. and Matsumoto, Y.: Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines, *Proc. NLPRS 2001* (2001).
17) Orphanos, G.S. and Christodoulakis, D.N.: POS Disambiguation and Unknown Word Guessing with Decision Trees, *Proc. EACL '99*, pp.134–141 (1999).
18) Ratnaparkhi, A.: A Maximum Entropy Model for Part-of-Speech Tagging, *Proc. EMNLP '96*, pp.133–142 (1996).
19) Rosenfeld, R., Chen, S.F. and Zhu, X.: Whole-Sentence Exponential Language Models: A Vehicle For Linguistic-Statistical Integration, *Computers Speech and Language*, Vol.15, No.1, pp.55–73 (2001).
20) Toutanova, K., Klein, D., Manning, C.D. and Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, *Proc. HLT-NAACL 2003*, pp.173–180 (2003).
21) Wang, S., Wang, S., Greiner, R., Schuurmans, D. and Cheng, L.: Exploiting Syntactic, Semantic and Lexical Regularities in Language Modeling via Directed Markov Random Fields, *Proc. ICML 2005*, pp.948–955 (2005).
22) Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proc. ACL '95*, pp.189–196 (1995).
23) 伊庭幸人，種村正美，大森裕浩，和合 肇，佐藤整尚，高橋明彦：統計科学のフロンティア 12 計算統計 II マルコフ連鎖モンテカルロ法とその周辺，岩波書店 (2005).
24) 内元清貴，関根 聡，井佐原均：最大エントロピーモデルに基づく形態素解析—未知語の問題の解決策，自然言語処理，Vol.8, No.1, pp.127–142 (2001).
25) 高村大也，乾 孝司，奥村 学：スピンモデルによる単語の感情極性抽出，情報処理学会論文誌，Vol.47, No.2, pp.627–637 (2006).
26) 中川哲治，松本裕治：単語レベルと文字レベルの情報を用いた中国語・日本語単語分割，情報処理学会論文誌，Vol.46, No.11, pp.2714–2727 (2005).

(平成 18 年 10 月 30 日受付)
(平成 19 年 12 月 4 日採録)

中川　哲治（正会員）

2000 年筑波大学第三学群情報学類卒業．2002 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年沖電気工業（株）入社．2006 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．2007 年より情報通信研究機構専攻研究員．博士（工学）．統計的自然言語処理および機械学習に興味を持つ．

松本　裕治（正会員）

1977 年京都大学工学部情報工学科卒業．1979 年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984〜1985 年英国インペリアルカレッジ客員研究員．1985〜1987 年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993 年より奈良先端科学技術大学院大学教授，現在に至る．工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM 各会員．