2006－NL－172（4）
2006／3／27

大域的な情報を用いた未知語の品詞推定

中川 哲治[†,††]
nakagawa378@oki.com

松本 裕治[††]
matsu@is.naist.jp

† 沖電気工業株式会社 研究開発本部
†† 奈良先端科学技術大学院大学 情報科学研究科

本稿では，局所的な情報と大域的な情報を用いた未知語の品詞推定を行う．従来手法の多くでは，未知語の品詞は局所的な情報 (未知語の前後数単語内，あるいは未知語の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 (文書全体での未知語の使われ方等) は未知語の品詞推定を行う上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を利用して未知語の品詞を推定するために，同じ語形を持つ未知語の文書中での全ての出現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モデルのパラメータを推定する．また，このモデルの半教師あり学習への適用を検討し，複数のコーパスを使用して実験を行う．

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

Tetsuji Nakagawa[†,††]
nakagawa378@oki.com

Yuji Matsumoto[††]
matsu@is.naist.jp

†Corporate Research and Development Center, Oki Electric Industry Co., Ltd.
††Graduate School of Information Science, Nara Institute of Science and Technology

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 (extra-sentential features) 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 estimate its parameters using Gibbs sampling. We also attempt to apply the model to semi-supervised learning, and conduct experiments on multiple corpora.

1　背景

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一つである．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，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中に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．未知語に関する情報は品詞タグ付けシステム中に存在しないため，通常は特別な処理によって扱われる．できるだけ正確に未知語の品詞を推定することは，高精度な品詞タグ付けを行う上で必要であり，また単語辞書を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 [12, 11, 4, 13, 14]．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は品詞推定が困難な場合が存在する．例えば，名詞のように使われている未知語があった場合，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，文書中の別の箇所で，大きな手がかりとなる局所的な情報 (人名に付く敬称など) と共に出現していれば，そのような情報は曖昧な語の品詞を推定する上で役に立つ．

別の例として，サ変名詞に関する問題が挙げられる．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，サ変名詞か普通名詞か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．この問題は Asahara[1] によって，「可能性に基づく品詞の問題」として指摘されている．可能性に基づく品詞とは，その品詞が単語の個々の事例の性質を表すのではなく，

その単語が持つことが可能な全ての性質を表すような品詞である．例えばサ変名詞は，名詞として使われることも可能であり，「する」が末尾に付いて動詞として使われることも可能であるが，個々の事例はこれらの全ての性質を一度に有しているわけではない．このように，可能性に基づく品詞を持つ単語においては，個々の事例は全ての可能な性質の中の一部の性質を持っているだけなので，一つの事例の局所的な情報のみからその品詞を推定するのは難しい場合がある．しかしながら，例えば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箇所で「〜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．

以上のような問題に対処するために，本稿では局所的な情報だけではなく大域的な情報も利用した未知語の品詞推定手法を提案する．提案手法では，未知語の個々の出現のみに注目するのではなく，未知語が文書[1]全体でどのように出現しているのかを同時に考慮する．そして，従来手法のように単語や文の独立性を仮定して単語や文の出現確率を個別に最大化するのではなく，同じ語形を持つ全ての未知語間の相互作用を考慮して文書全体が与えられた場合の条件付き確率を最大化することにより，未知語の品詞推定を行う．大域的な情報の利用は，他の自然言語処理タスクにおいても役に立つことが知られており，特に固有表現抽出では，大域的な情報を利用した手法がいくつか提案されている [6, 7].

提案手法の一つの利点として，ラベル無しデータを容易に利用できる可能性があることが挙げられる．つまり，ラベル無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，テスト時に使用される大域的な素性を増やすことができる．

このような文書全体を考慮した確率モデルは，局所的な素性のみを利用する確率モデルと比較して，多くの計算量を必要とする．また，入力されたデータを一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．そこで，提案手法では，ギブスサンプリングを使用して効率的に確率モデルの計算を行う．また，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考えた場合，実時間で処理を行う必要性は無いが，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，このようなモデルの用途として適していると思われる．

以下，2 節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 節では実験結果を報告する．4 節では，関連研究について議論し，5 節で結論を述べる．

[1]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など，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．

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，既知語の品詞は既に決定されており，文書中での未知語の出現位置も既に同定されていると仮定して，未知語に対する品詞の推定のみに注目する．

以下この節では，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始めに説明する．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推定方法について説明する．また，ラベル無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した確率モデル

提案手法では，同じ語形を持つ文書中の全ての未知語を同時に考慮してモデル化することを考える．そして，そのような未知語の品詞は相互に影響しあい，なおかつ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．例えば，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作用し，なおかつ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:

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

Takamura ら [17] 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ことを考える．

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:

$$\begin{aligned}\mathbf{w}&=\{w_1,\cdots,w_K\},\\ \mathbf{t}&=\{t_1,\cdots,t_K\},\\ t_k&\in\{1,\cdots,N\}\qquad(k=1,\cdots,K).\end{aligned}$$

$\lambda_{i,j}$ 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．一つは大域的な品詞間の相互作用を表す項であり，もう一つは局所的な文脈による影響を表す項である．

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:

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

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

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

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．文書中に出現する未知語の数は 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一つをとるため，$\mathcal{T}(\mathbf{w})$ 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:

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

以上のように提案手法で用いる確率モデルでは，同じ語形を持つ未知語の全ての出現を同時に考慮する．ただし本手法では，異なる語形の未知語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計算される．

2.2 解析手法

テストデータ $\mathbf{w}$ と初期分布 $p_0(t|w)$ 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．一つの方法として，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．また，全ての未知語の品詞間の相互作用 (依存関係) を考慮しているため，動的計画法のような手法を用いることはできない．そこで，マルコフ連鎖モンテカルロ法を用いることにより，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を考える．

未知語の品詞推定結果の解 $\hat{\mathbf{t}}=\{\hat{t}_1,\cdots,\hat{t}_K\}$ を，次

1 $\mathbf{t}^{(1)}$ を初期化する
2 **for** m **:=** 2 **to** M
3 **for** k **:=** 1 **to** K
4 $t_k^{(m)}\sim P(t_k|\mathbf{w},t_1^{(m)},\cdots,t_{k-1}^{(m)},t_{k+1}^{(m-1)},\cdots,t_K^{(m-1)})$

図 1: ギブスサンプリング

のようにして求めることにする:

$$\hat{t}_k=\underset{t}{\text{argmax}}\,P_k(t|\mathbf{w}), \tag{11}$$

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，次のように未知語の品詞の確率分布に対する期待値として計算することができる:

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

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 [10]．例えば，$A(\mathbf{x})$ 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，$\{\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(M)}\}$ は $P(\mathbf{x})$ 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はマルコフ連鎖モンテカルロ (Markov Chain Monte Carlo; MCMC) 法の一種であり，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 [20]．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全て固定した条件付き確率からのサンプリングを行い値を更新する，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られている．図 1 中の条件付き確率 $P(t_k|\mathbf{w},t_1,\cdots,t_{k-1},t_{k+1},\cdots,t_K)$ は次のようにして容易に計算できる:

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

ここで，最後の式は次の関係から得られる:

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

本研究では，サンプルの数 M は 100 とし，また初期状態 $\mathbf{t}^{(1)}$ は，$p_0(\mathbf{t}|\mathbf{w})$ を最大化する品詞に設定した．

式 (11) によって得られる解は，$\mathbf{w}$ が与えられた場合の各品詞の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 [7] は，同様の最適化を行う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した場合，上述の手法では冷却スケジュール (cooling schedule) を決める必要が無い．さらに上述の手法は，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なども周辺分布 $P_k(t|\mathbf{w})$ 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，提案手法を辞書の半自動的作成などに応用する際には，作業者にランク付けされた候補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[2]$\{\langle\mathbf{w}^1,\mathbf{t}^1\rangle,\cdots,\langle\mathbf{w}^L,\mathbf{t}^L\rangle\}$ と初期分布 $p_0(t|w)$ 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，次のような目的関数 $\mathcal{L}_\Lambda$ 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):

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

これを偏微分すると次のようになる:

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

本研究では，$P(\Lambda)$ には Gaussian prior[5] を用いる:

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

ここで，C は定数であり，σ の値は 1 に設定した．

上記の $\mathcal{L}_\Lambda$ と $\frac{\partial\mathcal{L}_\Lambda}{\partial\lambda_{i,j}}$ を計算すれば，準ニュー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[9] を使用して最適解を求めた[3]．しかしながら，式 (16) 中の $Z_\Lambda(\mathbf{w}^l)$ と (式 (9) 参照)，式 (17) 中の項には，あらゆる品詞についての数え上げが含まれており，計算が困難である．そこで，Rosenfeld ら [16] が行ったのと同様に，ギブスサンプリングを用いてこれらの値を計算する．$Z_\Lambda(\mathbf{w}^l)$ 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:

$$\begin{aligned}Z_\Lambda(\mathbf{w}^l)&=\sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w}^l)}p_0(\mathbf{t}|\mathbf{w}^l)\exp\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{i,j}f_{i,j}(\mathbf{t})\right\},\\&\simeq\frac{1}{M}\sum_{m=1}^{M}\exp\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{i,j}f_{i,j}(\mathbf{t}^{(m)})\right\}.\end{aligned}\tag{18}$$

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:

$$\sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w}^l)}f_{i,j}(\mathbf{t})P_\Lambda(\mathbf{t}|\mathbf{w}^l)\simeq\frac{1}{M}\sum_{m=1}^{M}f_{i,j}(\mathbf{t}^{(m)}).\tag{19}$$

実験では，ギブスサンプリングの初期分布 $\mathbf{t}^{(1)}$ には，訓練データ中の正解ラベルを使用した．

2.4 ラベル無しデータの利用

提案したモデルでは，テストデータに単純にラベル無しデータを結合し，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，容易にラベル無しデータを利用することができると思われる．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．例えば，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，ラベル無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，ラベル無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単語に対するの解析精度を向上できる可能性がある．このように，本手法における半教師あり学習では，ラベル無しデータはテスト時のみに利用し，訓練時はラベル無しデータを利用しない場合と全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 使用したデータと実験の手順

実験には次の 8 つのコーパスを使用した: 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) (表 1 参照)．これらは全て品詞タグ付きコーパスであり，中国語，日本語，英語

[2]つまり，訓練データとなる文書中に含まれる未知語の語形は L 種類存在し，訓練データ中の各事例は同じ語形を持つ未知語の集合である．

[3]提案手法ではギブスサンプリングを用いた近似を行っているため，実験の際にはしばしば L-BFGS が完全には収束しなかった．そのような場合には，L-BFGS の繰り返しを途中で中断した．

コーパス (言語)	品詞の数 (オープンクラス)	単語数 (未知語の数) [コーパス中の分割位置]		
		訓練データ	テストデータ	ラベル無しデータ
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)	2,550,532 $[id = 4n + 0, id = 4n + 1]$	1,280,057 (24,178) $[id = 4n + 2]$	1,274,458 $[id = 4n + 3]$
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]

表 1: 使用したコーパス

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，ラベル無しデータの 3 つの部分に分割した．ラベル無しデータは半教師あり学習の実験に使用するためのものであり，ラベル無しデータ中の未知語の品詞はあらかじめ削除した．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・ラベル無しデータのサイズ，それらのデータの分割手法，を示す．テストデータとラベル無しデータ中の単語に対して，訓練データ中に存在しない単語を未知語と定義する．テストデータ中の未知語の数を表 1 の括弧内に示す．未知語の品詞推定の精度は，ここに示された数の未知語のうち，どれだけを正しく品詞推定することができたかによって計算される．

図 2 に実験の手順を示す．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) 未知語として扱う．このような (2 分割) 交差検定を用いることにより，訓練データ中での未知語を定義して，未知語に対する訓練事例を作成することができる[4]．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慮することにする (つまり，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 種類のモデルパラメータを推定する必要がある．一つは，$p_0(t|w)$ の計算に必要な局所的モデル (パラメータ) であり，もう一つは大域的モデル (パラメータ)($\lambda_{i,j}$ の値) である．局所的モデルパラメータは，全ての訓練データを用いて推定される (図 2，*2)．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の推定に使用したのと同一の訓練データを使用して大域的モデルパラメータを推定することはできない．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 A と局所的モデル B をそれぞれ学習させる (図 2，*3)．大域的モデルパラメータを推定する際には，訓練データ A 中の未知語に対しては局所的モデル B を，訓練データ B 中の未知語に対しては局所的モデル 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 $p_0(t|w)$ の値は訓練データ全体から学習された局所的モデルを用いて計算する．そして，この局所的モデルを大域的モデルと共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一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．テストデータ中の 1 回しか出現しない未知語に対しては，$p_0(t|w)$ 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．

半教師あり学習の実験では，ラベル無しデータも利用する．その場合，テストデータとラベル無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が解として選択される．

3.2 初期分布

提案手法では，局所的文脈 w が与えられた場合の品詞 t の確率を与える初期分布 $p_0(t|w)$ が使用される (式 (8))．この実験では，初期分布の計算には最大エントロピー (Maximum Entropy; ME) モデルを使用した．$p_0(t|w)$ は ME モデルを使用して次のように計算

[4] 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みなして使用する方法がある [13]．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，その語を未知語とみなす．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，コーパス中の全ての単語に対して繰り返す．しかしながらこの方法で定義される未知語は，文書中に一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．

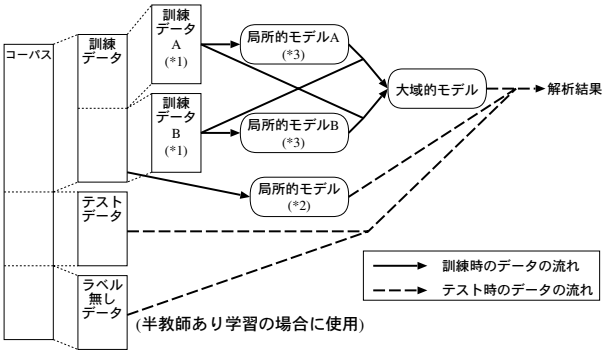

図 2: 実験の手順

できる [3]:

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

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

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こととする:

- 未知語の前後 2 つの品詞: $\tau_{-2},\tau_{-1},\tau_{+1},\tau_{+2}$.[5]
- 未知語自身と未知語の前後 2 つの単語: $\omega_{-2},\omega_{-1},\omega_0,\omega_{+1},\omega_{+2}$.
- 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:

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

使用した素性を表 2 に示す．これらの素性は，Ratnaparkhi[15] や Uchimoto ら [18] によって使用された素性を基に決めたものである．

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全ての単語を用いて推定した．

3.3 実験結果

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＋ラベル無しデータ**は，それぞれ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無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，次のようにして初期分布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:

$$\hat{t}_k=\underset{t}{\text{argmax}}\, p_0(t|w_k). \tag{22}$$

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に対して解析精度のマクネマー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．

[5]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．

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)
(共通)	$\|\omega_0\|$ (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}

表 2: 初期分布に使用した素性

CTB，PFR，KUC，RWC，WSJ コーパスにおいて，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 ($p<0.05$ で統計的に有意)．しかしながら，英語のコーパス (GEN，SUS コーパス) では，精度の向上は小さかった．表 4 は，PFR，RWC，SUS コーパスにおいて，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との区別を行うのは容易であることが多いが，中国語と日本語では，そのような習慣は無いため，局所的な情報のみを利用してそれらの区別を行うことは，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低いことが考えられる．可能性に基づく品詞の問題に関しては，RWC コーパス中のサ変名詞 (名詞-サ変接続) の正解はそれほど増えてはいないが，PFR コーパス中の名詞的動詞の正解数は提案手法により大きく増加させることができた．

ラベル無しデータを使用した場合，テストデータに 2 回以上出現した (提案手法での処理対象となる) 未知語の数は増加している．ラベル無しデータを使用しない場合と比べて，いくつかのコーパスでは精度が向上したが，CTB，KUC，WSJ コーパスでは精度が低下した．

提案手法では，訓練時にもテスト時にもギブスサンプリングを使用するため，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，異なる疑似乱数列を使用して 10 回の試行を行い，精度のばらつきの度合いを測った．さらに，Finkel ら [7] 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 $\beta\approx\infty$ まで等差的に減少さ

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+ラベル無しデータ
CTB (C)	0.7423 (193)	0.7717 (171) [0.0000] ⟨344⟩	0.7704 (172) [0.0001] ⟨361⟩
PFR (C)	0.6499 (9723)	0.6690 (9193) [0.0000] ⟨16019⟩	0.6785 (8930) [0.0000] ⟨18861⟩
EDR (J)	0.9639 (874)	0.9643 (863) [0.1775] ⟨4903⟩	0.9651 (844) [0.0034] ⟨7770⟩
KUC (J)	0.7501 (619)	0.7634 (586) [0.0000] ⟨788⟩	0.7562 (604) [0.0872] ⟨936⟩
RWC (J)	0.7699 (2572)	0.7785 (2476) [0.0000] ⟨5044⟩	0.7787 (2474) [0.0000] ⟨5878⟩
GEN (E)	0.8836 (905)	0.8837 (904) [1.0000] ⟨4094⟩	0.8863 (884) [0.0244] ⟨4515⟩
SUS (E)	0.7934 (1190)	0.7957 (1177) [0.1878] ⟨3210⟩	0.7979 (1164) [0.0116] ⟨3583⟩
WSJ (E)	0.8345 (704)	0.8368 (694) [0.0162] ⟨1412⟩	0.8352 (701) [0.7103] ⟨1627⟩

表 3: 未知語の品詞推定の実験結果

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)	0.7696±0.0021	0.7682±0.0028
PFR (C)	0.6707±0.0010	0.6712±0.0014
EDR (J)	0.9644±0.0001	0.9645±0.0001
KUC (J)	0.7595±0.0031	0.7612±0.0018
RWC (J)	0.7777±0.0017	0.7772±0.0020
GEN (E)	0.8841±0.0009	0.8840±0.0007
SUS (E)	0.7997±0.0038	0.7995±0.0034
WSJ (E)	0.8366±0.0013	0.8360±0.0021

表 5: 精度のばらつきと焼きなまし法との比較

せた．実験の結果を表 5 に示す．この表は，10 回実行した試行における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．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．

4 関連研究

大域的な情報を用いた自然言語処理の研究は，従来から様々なものが行われている．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 [6] 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 [7] 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，conditional random field 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 つのモデルを掛け合わせ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 [8] 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によって求めたり人手によって決定されていたが，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一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，このようなアプローチは他のタスクでも用いられている．Takamura ら [17] 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において，各電子は上向きか下向きの 2 つのうちのどちらか一方の状態 (スピン) をとり，その状態の確率分布が与えられる．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，平均場近似を用いて系の状態を計算している．

Yarowsky[19] は，ラベル無しデータを用いた語義の曖昧性解消の手法を提案した．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

PFR (C)		RWC (J)		SUS (E)	
+162	vn (名詞的動詞)	+33	名詞-固有名詞-人名-姓	+13	NP (固有名詞)
+150	ns (地名)	+32	名詞-固有名詞-地域	+6	JJ (形容詞)
+86	nz (その他の固有名詞)	+28	名詞-固有名詞-組織	+2	VVD (本動詞の過去形)
+85	j (略語)	+17	名詞-固有名詞-人名-名	+2	NNL (場所名詞)
+61	nr (人名)	+6	名詞-固有名詞	+2	NNJ (組織名)
+27	l (慣用語)	+4	名詞-サ変接続	...	...
...	...	...	...	−2	VVN (本動詞の過去分詞)
−26	m (数)	−2	名詞-固有名詞-地域-国	−3	NN (普通名詞)
−100	v (動詞)	−29	名詞	−6	NNU (単位名)

表 4: 正しく品詞が推定された未知語の数の変化

報の両者を用いて教師無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 [16] は，whole-sentence exponential language model を提案した．この手法では，ある文 s の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 $f_{i,j}(\mathbf{t})$(式 7) を素性関数とみなすと，式 (8) は本質的に上の式と等しいといえる．彼らのモデルは，shallow parser によって得られる統語的な素性などの，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 MCMC 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直接応用できる．その場合，本稿で検討したようにラベル無しデータを容易に利用することができる．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，同じ語形をもった未知語の品詞間の相互作用を考慮して，文書全体をモデル化する．モデルのパラメータは，ギブスサンプリング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．本手法を半教師あり学習へ応用することも試みたが，ラベル無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．ラベル無しデータを利用する方法を改善することは，今後の課題である．

謝辞

本研究は，情報通信研究機構平成 14 年度民間基盤技術研究促進制度に係る研究開発課題「多言語標準文書処理システムの研究開発」の一環として行われたものである．

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] Chen, C., Bai, M. and Chen, K.: Category Guessing for Chinese Unknown Words, *Proceedings of NLPRS '97*, pp. 35–40 (1997).

[5] Chen, S. and Rosenfeld, R.: A Gaussian Prior for Smoothing Maximum Entropy Models (1999). Technical Report CMUCS-99-108, Carnegie Mellon University.

[6] Chieu, H. and Ng, H.: Named Entity Recognition: A Maximum Entropy Approach Using Global Information, *Proceedings of COLING 2002*, pp. 190–196 (2002).

[7] Finkel, J., Grenager, T. and Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, *Proceedings of ACL 2005*, pp. 363–370 (2005).

[8] Klakow, D.: Log-linear interpolation of language models, *Proceedings of ICSLP '98*, pp. 1695–1699 (1998).

[9] 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).

[10] MacKay, D. J. C.: *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press (2003).

[11] Mikheev, A.: Automatic Rule Induction for Unknown-Word Guessing, *Computational Linguistics*, Vol. 23, No. 3, pp. 405–423 (1997).

[12] Mori, S. and Nagao, M.: Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis, *Proceedings of COLING '96*, pp. 1119–1122 (1996).

[13] Nagata, M.: A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context, *Proceedings of ACL '99*, pp. 277–284 (1999).

[14] Orphanos, G. S. and Christodoulakis, D. N.: POS Disambiguation and Unknown Word Guessing with Decision Trees, *Proceedings of EACL '99*, pp. 134–141 (1999).

[15] Ratnaparkhi, A.: A Maximum Entropy Model for Part-of-Speech Tagging, *Proceedings of EMNLP '96*, pp. 133–142 (1996).

[16] 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).

[17] Takamura, H., Inui, T. and Okumura, M.: Extracting Semantic Orientations of Words using Spin Model, *Proceedings of ACL 2005*, pp. 133–140 (2005).

[18] Uchimoto, K., Sekine, S. and Isahara, H.: The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary, *Proceedings of EMNLP 2001*, pp. 91–99 (2001).

[19] Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proceedings of ACL '95*, pp. 189–196 (1995).

[20] 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).