Vol.2015-NL-220 No.3
2015/1/19

隠れセミマルコフモデルに基づく品詞と単語の同時ベイズ学習

内海 慶[1,a)]　塚原 裕史[1,b)]　持橋大地[2,c)]

概要：本論文では，教師なし学習による品詞を含めた形態素解析手法を提案する．従来の教師なし形態素解析手法では分かち書きのみを対象としており，品詞の推定は扱われてこなかった．本稿では，品詞遷移確率と単語の生起確率の事前分布に階層 Pitman-Yor 過程を用いた隠れセミマルコフモデルに基づく形態素解析手法を提案し，分かち書きとその潜在的な品詞を同時に学習する．これにより，単語分割自体の精度も向上することを日本語，中国語，およびタイ語での実験により確認した．

キーワード：形態素解析，品詞推定，隠れマルコフモデル，言語モデル，ノンパラメトリックベイズ

1. はじめに

形態素解析は自然言語処理の基盤技術である．特に，日本語，中国語などのアジア語のように単語境界が与えられない言語では，文書検索の索引付けや名詞句，固有表現抽出，構文解析等の様々な自然言語処理手法を適用するための前処理として不可欠となっている．従来，形態素解析器には教師あり学習手法が用いられてきた．そのため，パラメータの学習には教師データが使用され，言語知識を持つ専門家によって整備された学習コーパスが用いられている．こうしたコーパスの多くは書き言葉，特に新聞データを対象として作られている．しかし，近年ではブログや交流サイト，ミニブログ等の Consumer Generated Media(CGM) が増加しており，こうした一般消費者が生成するメディアの処理の必要性は特別に高い．

CGM では書き言葉と話し言葉が混在されて用いられる．また，顔文字などを用いた感情表現など，これまでの書き言葉では現れなかった表現が作られており，未知語は常に新しく産まれ続けている．こうした近年の CGM に対して，従来のように人手で多量の正解データを作るのは現実的ではない．

こうした問題に対して，文字列の生データから，あるいは既存の教師データを援用しつつ分かち書きを行う教師なし学習や半教師あり学習の手法が提案されている [1][2][3][4] [5][6] [7]．しかしながら，これまで提案されてきた手法では，分かち書きは獲得できるものの，品詞情報の獲得は対象とされていなかった．品詞情報は固有表現抽出や係り受け解析等，形態素解析を前処理として用いる解析では重要な手がかりであり，本稿で示すように，分かち書き自体の精度にも貢献する文法的情報である．

そこで，本論文では，教師なし，半教師あり学習で分かち書きの学習を行うと同時に，品詞情報の獲得を行う手法を提案する．

以降，2 章では話し言葉を対象とした形態素解析に関連する選考研究について説明を行い，3 章で我々が基にした持橋らの NPYLM[1] について解説する．4 章で，我々の提案する形態素解析手法について述べる．5 章では，我々の手法について評価を行い，その効果を示す．6 章では，総論を行い，今後の課題を示す．

2. 関連研究

話し言葉の形態素解析においては，未知語が解析の問題となることが指摘されている [8]．内元らは少量のタグ付きコーパスから，最大エントロピーモデルを用いて文字列の単語と品詞の同時推定を行い，尤度の低い形態素を人手で修正することで効率よく未知語に対する情報を付与している．松本ら [9] は，既存の書き言葉を対象に作られた形態素解析が話し言葉では性能が出ないことを示し，これに対して少量のタグ付き話し言葉データを既存の書き言葉の教師データに加えることで，大きく性能を改善できること

[1] デンソーアイティーラボラトリ
DENSO IT LABORATORY, cross tower 28th Floor, 2-15-1 Shibuya Shibuya-ku Tokyo, 150–0002, Japan
[2] 統計数理研究所
The Institute of Statistical Mathematics
[a)] kuchiumi@d-itlab.co.jp
[b)] htsukahara@d-itlab.co.jp
[c)] daichi@ism.ac.jp

を示した．これらの手法は教師あり学習に基づく手法であり，形態素解析の学習をするためには教師データを必要とする．しかし，前述したように話し言葉は変化が早く，常に変化する表現にあわせて教師データを作成し続けるのは現実的ではない．

Creutz ら [10] は，英語及びフィンランド語の単語分割に，最小記述長 (MDL) に基づくグリーディアルゴリズムを用いる手法と，EM アルゴリズムで用いて入力データの対数尤度を最大化する手法の 2 つを提案し，MDL に基づく手法がより高い正解率となることを示した．Argamon ら [11] も同様に，MDL に基づくグリーディアルゴリズムによって単語分割を行っている．

Zhikov ら [7] らは，MDL と branching entropy[12] の 2 つを用いた手法を提案し，branching entropy によって単語境界の候補を絞り込むことで高速な単語分割を提案した．Magistry ら [6] も同様に MDL と branching entropy を用いた中国語の単語分割を行っている．Magistry らは同時に，記述長が小さいことが良い分割に対応するわけではないことを指摘している．MDL に基づく手法は，単語境界を決定する際にはそこまでの単語列等の文脈が考慮されておらず，文脈に応じて単語分割を変化させることは難しい．

持橋ら [1] は，文字・単語ベイズ n グラム言語モデルのベイズ学習を用い，言語に依存しない単語分割の提案を行った．彼らの手法では，最適なスムージングを行うベイズ n グラム言語モデルを用いることで，学習データに対する言語モデルの過学習の問題を解決している．しかし，言語モデルの性能を最適化しているため，人間の分割基準とは異なる [*1] 場合があった．この問題に対処するため，持橋らは，CRF と NPYLM の協調学習を行う半教師有り学習手法による単語分割も提案している [2]．

これまで提案された教師あり学習に基づく形態素解析手法では，話し言葉を扱う際にも教師データを必要としており，また教師なし学習による手法では分かち書きは扱っているものの，単語の潜在的な意味クラスを考慮していないため品詞推定は行えなかった．

我々の提案する手法では，単語の潜在的な意味クラスを考慮し，単語分割と同時にその推定を行う．これまでの手法では単語分割のみが対象であったが，提案手法では単語の意味クラスと，意味クラスの間の依存関係をデータから獲得することで，単語の意味，すなわち品詞と，品詞間の依存関係，すなわち文法を獲得する．

3. Nested Pitman-Yor Language Model

我々の提案する手法は持橋らの教師なし形態素解析手法を基にしている．ここでは最初に，持橋らの NPYLM について説明する．

[*1] 「見/る」のように活用語尾が分割されたり，複合語が単語として切り出されることが多い．

Algorithm 1 NPYLM の学習アルゴリズム

Input: $s \in S$

```
Add w_0(S) to Θ
for j = 1 ··· J do
    for s in randperm (S) do
        Remove customers of w(s) from Θ
        Draw w(s) according to p(w|s, Θ)
        Add customers of w(s) to Θ
    end for
    Sample hyperparameters of Θ
end for
```

持橋らの手法では，教師なし形態素解析を最も基本的な単語分割問題として扱っている．NPYLM では，動的計画法と MCMC を組み合わせた学習を行うことで文字-単語の階層 n グラム言語モデルと単語分割を直接最適化する．Algorithm 1 に，NPYLM の学習アルゴリズムを示す．$\mathbf{w}_0(S)$ は，与えられた入力文集合 $s \in S$ の各文 s それぞれに対する初期状態の単語分割を表す．初期状態では s の単語分割は与えられていないため，$\mathbf{w}_0(S)$ は入力文全てをそれぞれ 1 つの単語と 1 見なし，階層 n グラム言語モデルのパラメータを初期化する．各文 s の単語分割の影響を言語モデルのパラメータから除去し，各文 s に対する単語分割確率 $P(\mathbf{w}|s)$ に従って新しい単語分割 $\mathbf{w}(s)$ をサンプリングし，言語モデルのパラメータの更新を行う．単語分割のサンプリングは，Forward filtering-Backward sampling 法によって効率的に行われる．

Remove customers 及び，Add customers の処理は，基本的には [13] と同様であるが，単語 n グラム言語モデルの基底測度 G_0 には，可変長 n グラム言語モデル [14] が用いられている．そのため，単語 n グラム言語モデルの文脈木の根では，与えられた単語に含まれる文字列を用いて確率的に可変長文字 n グラム言語モデルの更新を行う．

4. 提案手法

品詞推定と分かち書きを同時に行うために，我々は次のように問題の定式化を行う．

4.1 教師なし形態素解析

形態素解析の問題を，文字列 $s = c_1 c_2 \dots c_N$ が与えられた際に，s を分割して得られる単語列及び単語列に対応した品詞列の確率 $P(\mathbf{w}|s)$ を最大化する問題と考える．ここで，$\mathbf{w} = \{w_1 w_2 w \dots w_M, z_1 z_2 \dots z_M\}$ であり，w_n，z_n はそれぞれ単語と品詞を表す．$P(\mathbf{w}|s)$ はそのままでは計算が難しいため，(1)　式とおくことで部分問題に分割する．

$$P(\mathbf{w}|s) = \prod_{i=1}^{M} P(w_i, z_i|h_{i-1}) \tag{1}$$

$$h_i = \{w_1, w_2, \dots, w_i, z_1, z_2, \dots, z_i\}$$

$P(w_i, z_i|h_{i-1})$ を，ベイズ則を用いて (2)　式のように変形する．

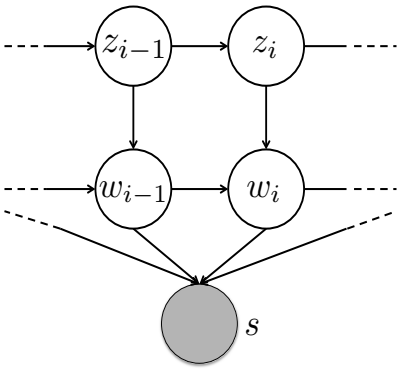

図 1　提案手法の生成モデル
Fig. 1　Our proposed generative model

$$P(w_i, z_i|h_{i-1}) = P(w_i|z_i, h_{i-1})P(z_i|h_{i-1}) \quad (2)$$
$$P(w_i|z_i, h_{i-1}) = P(w_i|w_{i-N+1}^{i-1}, z_i) \quad (3)$$
$$P(z_i|h_{i-1}) = P(z_i|z_{i-N+1}^{i-1}) \quad (4)$$

ここで，i 番目の単語は $N-1$ 個前までの単語列と i 番目の品詞のみに，i 番目の品詞は $N-1$ 個前までの品詞列のみに依存すると仮定した．

図 1 に我々の提案する生成モデルを表す．単語境界は与えられていないため，$\mathbf{w}$ についても観測はできない．観測できるのは文字列 s のみである．我々の手法は，s の部分文字列からなるセグメントを単語候補とし，単語候補の隠れ変数として品詞が加わった隠れセミマルコフモデルとなっている．

4.2　n グラム確率モデル

$P(w_i|w_{i-N+1}^{i-1}, z_i)$ は，品詞毎の単語 n グラム確率，$P(z_i|z_{i-N+1}^{i-1})$ は品詞 n グラム確率を表す．品詞毎の単語 n グラム確率には，持橋らと同様 Pitman-Yor 過程による n グラムモデル [13] を，単語ユニグラムの事前確率にも同様に可変長文字 n グラム言語モデル [14] を用いる．品詞 n グラム確率についても同様に，事前分布に Pitman-Yor 過程を用いる．Pitman-Yor 過程を事前分布に用いた単語 n グラム確率を (5)　式，品詞 n グラム確率を (6)　式に表す．$t_{|h|}$ は，文脈 h において，親の文脈から単語 w_i が生成されたと見なされた回数を表し，文脈毎の Chinese Restaurant Process によってデータから最適化される．$d_{|h|}$ と $\theta_{|h|}$ は単語 n グラムの Pitman-Yor 過程のハイパーパラメータを $e_{|h|}$ と $\eta_{|h|}$ は品詞 n グラムの Pitman-Yor 過程のハイパーパラメータを表す．ハイパーパラメータの推定は，[13] に従って行う．

$$P(w_i|w_{i-N+1}^{i}, z_i) = \frac{c(w_i|w_{i-N+1}^{i-1}, z_i) - d_{|h|}t_{hw_i}}{\theta_{|h|} + c(w_{i-N+1}^{i-1})} \quad (5)$$
$$+ \frac{\theta_{|h|} + d_{|h|}t_{h\cdot}}{\theta_{|h|} + c(w_{i-N+1}^{i-1}, z_i)} P(w_i|w_{i-N+2}^{i-1}, z_i)$$

$$P(z_i|z_{i-N+1}^{i-1}) = \frac{c(z_i|z_{i-N+1}^{i-1}) - \eta_{|h|}t_{hz_i}}{e_{|h|} + c(z_{i-N+1}^{i-1})} \quad (6)$$
$$+ \frac{e_{|h|} + \eta_{|h|}t_{h\cdot}}{\eta_{|h|} + c(z_{i-N+1}^{i-1})} P(z_i|z_{i-N+2}^{i-1})$$

4.3　学習

我々の学習アルゴリズムも持橋らと同様，動的計画法と MCMC を組み合わせた手法で行う．我々と持橋らの手法の違いは，単語分割と品詞列の両方を隠れ変数とみなし，同時にサンプリングを行う点である．

4.3.1　単語分割と品詞列のサンプリング

単語分割と品詞列を同時にサンプリングするためには，単語の品詞の同時確率を求める必要がある．そのため，提案手法の Forward-filtering における前向き確率は，品詞を考慮して (7)　式の再帰式のようになる（$N=2$）．ここで，$\alpha[t][k][z]$ は位置 $t-k$ から t までの長さ k の文字列 c_{t-k}^{t} が品詞 z の単語として生成される確率を表す．Z は品詞クラス数を表す．

$$\alpha[t][k][z] = \sum_{j=1}^{t-k}\sum_{r=0}^{Z} P(c_{t-k}^{t}|c_{t-k-j+1}^{t-k}, z)P(z|r)\alpha[t-k][j][r] \quad (7)$$

$\alpha[t][k][z]$ が求まると，文末から単語分割と品詞を同時にサンプリングすることができる．$\alpha[t][k][z]$ は，c_{t-k}^{t} が品詞 z の単語となる確率であり，文末を示す特別な単語への遷移確率は $P(E_w|c_{t-k}^{t}, E_p)P(E_p|z)\alpha[t][k][z]$ となる．この確率に従って文末から繰り返し，文頭に至るまで単語と品詞のサンプリングを行う．

4.4　ガンマー般化線形回帰モデルによる計算の効率化

提案手法の計算量は，文字列長を N，考慮する最大のセグメント長を L，品詞数を K とした時に $O(K^2L^2N)$ となる．最大のセグメント長を適切な値に決めることで，言語に依存せず本手法は適用可能であるが，日本語のカタカナからなる名詞やタイ語等の言語では一部の固有表現の長さが大きくなる．そこで，我々は実験の高速化のために予め文字列の各位置毎で考慮すべき最大のセグメント長を予測し，ラティスの圧縮を行うことで計算の効率化を行った．

(8)　式に示すように，単語の長さ x はガンマ分布に従うと仮定し，与えられた単語列からガンマ分布のパラメータ a，b の回帰を行った．

$$\mathrm{Ga}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1}e^{-bx} \quad (8)$$

a, b を (9)　式でモデル化した．$\mathbf{w}$, $\mathbf{v}$ はそれぞれ a, b を回帰する D 次元のベクトルである．$\mathbf{f} = (f_0, f_1, \cdots, f_D)$ は与えられた文字列のもつ D 次元の特徴ベクトルを表し，ここでバイアス項に対応する特徴 f_0 は常に 1 であるとする．

$$\begin{cases} a = \exp(\mathbf{w}^T \mathbf{f}) \\ b = \exp(\mathbf{v}^T \mathbf{f}) \end{cases} \quad (9)$$

特徴ベクトルは，時刻 n までの文字列 $s_n = \{c_1 \dots c_n\}$ から抽出される特徴量からなるとし，$\mathbf{f_n} = \phi(s_n)$ と表す．

パラメータ $\mathbf{w}$，$\mathbf{v}$ の学習には MCMC を用いる．

4.4.1 特徴量

最大セグメント長の予測に使用した特徴量を表 1 に示す．特徴は，入力文字列 s_n の各位置 t 毎に抽出される．

表 1 セグメント長予測のための素性

Table 1 Features for segment size prediction

特徴	概要
w_i	位置 $t-i$ の文字 $(0 \le i \le 1)$
t_i	位置 $t-i$ の文字種 $(0 \le i \le 4)$
$type$	位置 t の文字種が直前で何文字続いているか
c	位置 t から最大 8 文字前までで文字種の変化した回数

5. 評価

提案手法の評価を行うため，我々は日本語，中国語，タイ語の 3 つの言語でアルゴリズムの動作を検証した．日本語の評価には，書き言葉のデータとして京都大学テキストコーパス [*2] を，口語体を含む文章の例として，BCCWJ コーパス [*3] を用いる．中国語には，SIGHAN bakeoff 2005 の中国語単語分割で用いられたデータセット [*4] 及び The Chinese Treebank 8.0[*5] を用いる．タイ語には InterBEST 2009 [*6] のタイ語単語分割で用いられたデータセットを用いる．

5.1 評価データ

5.1.1 京都大学テキストコーパス

京都大学テキストコーパスは，毎日新聞の 1995 年 1 月 1 日から 17 日までの全記事約 2 万文，1 月から 12 月までの社説記事約 2 万文の計約 4 万記事が含まれており，人手による正解の分かち書き及び各形態素の品詞情報が付与されている．

5.1.2 BCCWJ コーパス

BCCWJ コーパスは現代日本語の書き言葉の全体像を把握するために国立国語研究所によって作成されたコーパスで，書籍全般，新聞，白書，ブログ，ネット掲示板，教科書，法律などのジャンルにまたがった 1 億 430 万語からなる．我々はこのうち，新聞データ (PN) と知恵袋 (OC) を評価データとして用いた．

5.1.3 SIGHAN bakeoff

SIGHAN bakeoff 2005 には，分かち書きの基準が異なる 4 つのコーパスが含まれている．我々はこれらのうち，従来手法と比較可能な 2 つのデータ (MSR，CITYU) で中国語の単語分割の評価を行った．

5.1.4 The Chinese Treebank 8.0

Chinese Treebank 8.0 には，中国語のオンラインニュース，政府文書，雑誌記事，テレビ放送，ブログデータ等 71,369 文，1,620,561 単語が含まれており，分かち書きと品詞情報が付与されている．

5.1.5 InterBEST 2009

InterBEST 2009 は，新聞，事典，科学論文，ノベルの複数ジャンルからなるタイ語の単語境界付きデータセットである．評価には，ノベルデータを用いた．

5.2 実験条件

表 2 に，実験に用いたデータのサイズを表す．単位はそれぞれ文である．訓練データとテストデータは，重複の無いようデータからランダムでサンプルした．訓練データに含まれる記号と数値表現の一部については，計算の効率化のためセグメント長を 1 に固定した．今回の実験で使用した訓練データのサイズは，10,000 文で統一した．これは比較に用いた従来手法の訓練データサイズと比べて小さい．例えば [1] では，SIGHAN bakeoff の評価で 50,000 文を訓練データとして用いている． 評価は，単語分割と品詞推定で行う．

表 2 評価データのサイズ

Table 2 Dataset size for evaluation

データセット	全体サイズ	訓練データ	テスト
京大コーパス	38400	10000	1000
BCCWJ PN	78607	10000	1000
BCCWJ OC	678475	10000	1000
SIGHAN MSR	90909(86924+3985)	10000	3985
SIGHAN CITYU	54511(53019+1492)	10000	1492
CTB8.0	20412	10000	937
InterBEST Novel	50139	10000	1000

(1) 教師なし学習
訓練データに付与されている分かち書きを削除し，文字列のみとした上で教師なし学習を行う．教師無し学習の評価では，潜在クラスの数は 15 とした．

(2) 半教師あり学習
訓練データとテストデータに含まれないデータからランダムに抽出した 10K 文を教師データとして用いる．

5.2.1 最大セグメント長の予測精度

最大セグメント長の予測には，半教師あり学習用に生成したラベル付きデータを用いてガンマ分布のフィッティングを行い，訓練データに含まれないデータから 1000 文字を用いて最大セグメントの予測精度を評価した [*7]．予測は，ガンマ分布の累積密度関数の値が 0.99 を超えるセグメ

[*2] http://nlp.ist.i.kyoto-u.ac.jp
[*3] http://www.ninjal.ac.jp/corpus_center/bccwj/
[*4] http://www.sighan.org/bakeoff2005/
[*5] https://catalog.ldc.upenn.edu/LDC2013T21
[*6] http://thailang.nectec.or.th/interbest/
[*7] 京大コーパスと CTB は，更に少量の 50000 文字で行った

表 3　ガンマ分布による最大セグメント長の予測精度
Table 3　Precision of max segment size by gamma distribution

-	京大コーパス	BCCWJ PN	BCCWJ OC	MSR	CITYU	CTB8.0	BEST
精度	0.995	0.996	0.989	0.995	0.995	0.996	0.981
長さ 5 以上の単語についての精度	0.849	0.842	0.452	0.366	0.440	0.588	0.839
評価データ中の最大セグメント長	12	10	55	11	13	8	88

ント長を求め，実際のセグメント長がこれに含まれる場合には正解として評価した．ガンマ分布のフィッティングには，gamglm[*8] を用いた．

表 3 に，最大セグメントの予測精度を表す．短い単語についてはどのデータについても予測した最大セグメント長の中に含めることができたが，長い単語についてはラティス中に正しい単語を含められていない場合があることが分かる．しかし，どの言語についても長い単語の出現頻度は多くないため，全体の精度で見ると殆どの単語は学習・予測時のラティスの中に含めることができる．

5.3　実験結果

5.3.1　単語分割

単語分割の評価には F 値を用いた．(10)　式に我々が用いた評価尺度を表す．

$$F_w = \frac{2 \times R_w \times P_w}{R_w + P_w} \tag{10}$$

$$P_w = \frac{\text{正しく分割できた単語数}}{\text{正解データ中の単語数}}$$

$$R_w = \frac{\text{正しく分割できた単語数}}{\text{デコーダの出力した単語数}}$$

表 4　教師なし単語分割の評価
Table 4　evaluation in unsupervised word segmentation

-	PYHMM	Mochihashi 2009	Zhikov 2010
京大コーパス	**0.714**	0.631	0.713[*9]
CTB8.0	**0.743**	0.693	-
BCCWJ PN	**0.716**	0.656	-
BCCWJ OC	**0.787**	0.595	-
MSR	0.787	**0.802**[*8]	0.782[*10]
CITYU	0.795	**0.824**[*8]	0.787[*10]
BEST	0.777	**0.821**[*11]	0.733[*9]

表 4 に，教師なし学習での分かち書きの評価結果を，表 5 に半教師あり学習での分かち書きの評価結果を表す．半教師ありの実験は，品詞情報の与えられているデータのみで行った．

ベイズ学習手法の比較として [1] の数値を，MDL に基づく手法との比較として [7] の数値を記載した．中国語については NPYLM を F 値で 0.02 ∼ 0.03 ほど下回ったが，実験で用いた訓練データが比較手法の 1/5 である点を考慮すると，これは十分に高い数値と言える．タイ語でも同様に NPYLM を下回った．タイ語では最大セグメント長の予測を行わずに固定値を用いている．タイ語は長い単語が日本語や中国語に比べて多いため，ガンマ一般線形モデルによる最大セグメント長の予測を用いると長い単語について誤りやすく，実験ではこの影響が出ていると考えられる．

表 5　半教師あり単語分割の評価
Table 5　evaluation in semi-supervised word segmentation

-	PYHMM	Mochihashi 2009
京大コーパス	**0.930**	0.913[*8]
CTB8.0	0.934	-
BCCWJ PN	0.947	-
BCCWJ OC	0.926	-

5.3.2　品詞推定

品詞推定は，正解の品詞が付与されている京大コーパス及び BCCWJ，CTB8.0 のみで行った．また，従来手法では教師なし学習で単語分割と品詞の同時推定を行う手法がなかったため，ここでは NPYLM と BayesianHMM[15] をカスケードした手法と比較を行った．BayesianHMM の学習には，学習済みの NPYLM を用いて訓練データを単語分割した結果を用いた．

評価尺度には，単語分割が正しく行えた単語について，付与された品詞クラスの精度を用いた．提案手法では潜在クラスを品詞クラスと見なすが，潜在クラスと品詞クラスの対応は自明ではない．ここでは，タグ付けされた潜在クラス毎に，最も多く共起した品詞クラスを対応するクラスと見なして精度を評価した．

教師なし学習での品詞推定の評価結果を表 6 に表す．半教師あり学習での品詞推定の評価結果を表 7 に表す．表 4，表 6 より，事前に単語分割を行った結果に対し品詞推定を行う手法と比較して，単語分割と品詞推定を同時に行うことで単語分割，品詞推定ともに良い結果になることが

表 6　教師なし品詞推定の精度
Table 6　evaluation in unsupervised POS tagging

-	PYHMM	NPYLM+BayesianHMM
京大コーパス	**0.590**	0.508
CTB8.0	**0.489**	0.416
BCCWJ PN	**0.559**	0.455
BCCWJ OC	**0.549**	0.450

[*8] http://chasen.org/˜daiti-m/dist/gamglm/
[*8] [1] より引用した
[*9] [7] より引用した
[*10] [6] より引用した
[*11] ガンマ一般線形回帰モデルによる最大セグメント長の予測を用いずにセグメント長の最大値を 7 に固定して実験を行った．

表 7　半教師あり品詞推定の精度
Table 7　evaluation in semi-supervised POS tagging

-	PYHMM
京大コーパス	0.894
CTB8.0	0.916
BCCWJ PN	0.906
BCCWJ OC	0.866

分かる．

5.4　エラー分析

表 8 に，京大コーパスのテストデータで高頻度に現れた誤りの例を示す．誤りの多くは活用形を分割していることに起因している．教師なし学習では，これらの活用形には助詞と同じ潜在クラスが割当てられていた．助詞と活用形のクラスを分離できるのであれば，後処理でチャンキングを行うことでこれらの誤りは対処が可能となると考えられる．

助詞と活用形がまとまった理由の 1 つに，潜在クラスの数を 15 としたことが挙げられる．この数は品詞の細分類と比べて少ない．そこで，潜在クラスの数を 50 にして再度京大コーパスについて評価を行った．表 9 に，潜在クラスの数を 50 とした時の F 値を示す．潜在クラスの数を大きくすることで F 値の向上が確認できた．

正解の品詞と潜在クラスの対応図を $K = 15$ の場合を図 2，$K = 50$ の場合を図 3 に示す．図 2 では，全体として，潜在クラスの 7 番が出やすい傾向があり，正解の品詞の多くはこのクラスに対応付けられるという結果になった．7 番のクラスで見ると，特に品詞の“特殊”と“助詞”と多く共起した．特殊という品詞が与えられているのは句読点などである．京大コーパス中で句読点は，名詞の後に出てくることが多く，そのために助詞との分離ができなかったものと考えられる．潜在クラスの数を 50 とした時は，$K = 15$ と比べて品詞と潜在クラスの対応がよりスパースになっていることが分かる．

より高精度な品詞推定を行うためには，潜在クラスが分離するように品詞 n グラム確率のハイパーパラメータの推定に用いるガンマ分布のハイパーパラメータを調整するなどが考えられるが，今回の実験ガンマ分布のハイパーパラメータは，a，b ともに 1 として行い，特に調整は行わなかった．

5.5　潜在クラスを用いたチャンキング

潜在クラスを 50 とした時の結果に対し，潜在クラス間の遷移のルールを用いて，表 8 にある活用形についてチャンキングを行った数値も併記した．表 10 に，チャンキングを行った後の高頻度の単語分割誤りを示す．潜在クラスを用いたルールを用いることで，単語分割誤りが改善し，京大コーパスの形態素の基準に合わせたまとめあげが行えていることが分かる．こうした品詞を用いたまとめあげは，従来の単語分割のみを対象とした教師なし形態素解析手法では困難であり，仮に品詞を使わずに行うのであれば，表層文字列の文脈を用いて複雑なルールを記述することになる．　表 10 で残っている誤りのうち，最も高頻度の「に」は「積極的に」や「一気に」，「急速に」などの形容詞が分割され，名詞と助詞に対応する潜在クラスが割り当てられていた．これについては，ルールを用いたチャンキングを用いると精度が低下するため，修正は難しい．「ている」は単語分割の誤りが原因である．「増えている」，「表れている」，「している」などのタ形連用テ形の動詞と接尾辞が正しく分割できていないために誤りとなっており，これも潜在クラスを用いたルールでは対処ができなかった．「の」についても，「に」の時と同様に「他の」や「どの」などの連体詞，指示詞が分割され，「名詞」と「助詞」に対応する潜在クラスが割り当てられていた．「し」は動詞の「果たし」，「した」などが分割されたことによる．京大コーパスの中では「崩壊し」，「記録し」のようなサ変名詞と動詞「し」の連接が多く見られるため，「し」で終わる動詞とサ

表 8　単語分割誤りの例
Table 8　examples of failed word segmentations

誤りと見なされた単語	頻度	例
て	412	求め/て, し/て, 抑え/て
た	411	生まれ/た, 過ぎ/た, し/た
る	245	上回/る, され/る, 付け/る
し	221	し/て, し/た, し/てい
っ	203	っ/て, っ/た
に	101	わずか/に, 堅調/に, 敏感/に
の	81	予想/外/の, 他/の, 新鋭/の, そ/の
り	80	返/り/咲き, 上が/り, 移/り
な	80	貴重/な, 大き/な, よう/な
され	73	され/る, され/た, され/て

表 9　京大コーパスの単語分割 (K=50)
Table 9　word segmentation in Kyoto corpus(K=50)

品詞精度	単語分割の F 値	単語分割の F 値（チャンキング後）
0.603	0.716	0.755

表 10　チャンキング後の誤り
Table 10　segmentation errors after chunking

誤った単語分割	頻度
に	88
ている	81
の	75
し	62
で	47
た	44
大	41
出	40
新	39
として	37

変名詞と動詞「し」の連接を区別できなかったためと考えられる．文脈の中では動詞としての役割を持つため，両者の区別は本手法では難しい．

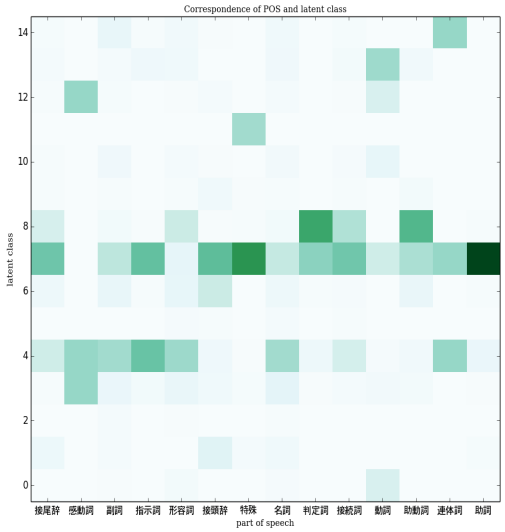

図 2　品詞と潜在クラスの対応図 (K=15)
Fig. 2　Correspondence of POS and latent class(K=15)

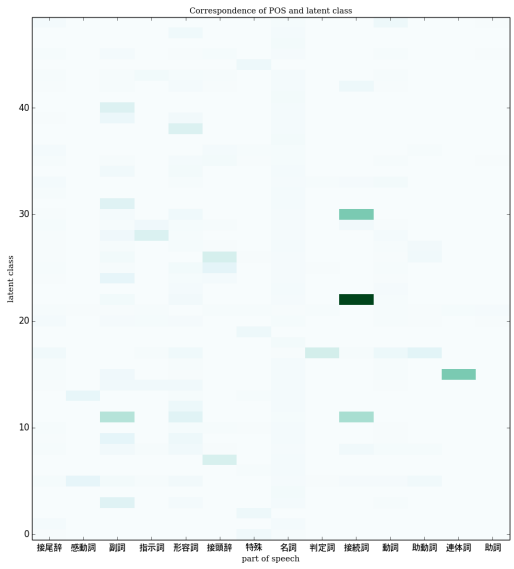

図 3　品詞と潜在クラスの対応図 (K=50)
Fig. 3　Correspondence of POS and latent class(K=50)

6. まとめ

本論文では，品詞推定と単語分割の同時推定手法の提案を行い，複数の言語においてその効果を検証した．日本語，中国語，タイ語について提案手法と従来の教師なし単語分割手法との比較を行い，品詞を考慮して推定を行うことで日本語については単語分割の性能が向上することを示した．中国語については NPYLM を F 値で 0.02 ∼ 0.03 ほど下回ったが，実験で用いた訓練データが比較手法の 1/5 である点を考慮すると，これは十分に高い数値と言える．タイ語でも同様に NPYLM を下回った．提案手法ではガンマ一般線形モデルによって最大セグメント長を事前に予測して学習時のラティスを構築している．NPYLM でも同様に最大セグメント長の予測結果を用いると F 値が低下したことから，これは最大セグメント長の予測精度が原因と考えられる．最大セグメント長の予測は，ラティスの圧縮による計算の高速化を目的としているため，提案手法の高速化を行うことでこのような問題は解決可能と考えられる．

品詞精度は，従来手法と比較すると高い数値を示してはいるが，人手で付けた品詞クラスとの対応を見ると，現状では複数の品詞が一つの潜在クラスにまとまってしまう等の問題がある．潜在クラスの数を大きくしたり，品詞 n グラムのハイパーパラメータの推定に用いるガンマ分布のハイパーパラメータの調整等を行い，潜在クラスが分離するよう工夫を行う必要があるが，本稿では時間の都合から調整等は行っていない．

品詞推定精度の改善と手法の高速化については，今後の課題である．

参考文献

[1] Mochihashi, D., Yamada, T. and Ueda, N.: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, Association for Computational Linguistics, pp. 100–108 (2009).
[2] 持橋大地，　鈴木潤，藤野昭典：条件付き確率場とベイズ階層言語モデルの統合による半教師あり形態素解析，言語処理学会第 17 回年次大会 (NLP2011) (2011).
[3] Goldwater, S., Griffiths, T. L. and Johnson, M.: A Bayesian framework for word segmentation: Exploring the effects of context, *Cognition*, Vol. 112, No. 1, pp. 21–54 (2009).
[4] Hewlett, D. and Cohen, P. R.: Bootstrap Voting Experts., *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 1071–1076 (2009).
[5] Hewlett, D. and Cohen, P.: Fully unsupervised word segmentation with bve and mdl, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, pp. 540–545 (2011).
[6] Magistry, P., Sagot, B. et al.: Can MDL Improve Unsupervised Chinese Word Segmentation?, *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pp. 2–10 (2013).
[7] Zhikov, V., Takamura, H. and Okumura, M.: An efficient algorithm for unsupervised word segmentation with branching entropy and MDL, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 832–842 (2010).
[8] 内元清貴，　野畑周，　山田篤，　関根聡，井佐原均：日本語話し言葉コーパスの形態素解析，言語処理学会第 9 回年次大会 (NLP2003) (2003).

[9] 松本裕治, 伝康晴：話し言葉の形態素解析，情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2001, No. 54, pp. 40–54 (2001).
[10] Baroni, M., Matiasek, J. and Trost, H.: Unsupervised discovery of morphologically related words based on orthographic and semantic similarity, *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, Association for Computational Linguistics, pp. 48–57 (2002).
[11] Argamon, S., Akiva, N., Amir, A. and Kapah, O.: Efficient unsupervised recursive word segmentation using minimum description length, *Proceedings of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics, p. 1058 (2004).
[12] Jin, Z. and Tanaka-Ishii, K.: Unsupervised segmentation of Chinese text by use of branching entropy, *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics on Main conference poster sessions*, Association for Computational Linguistics, pp. 428–435 (2006).
[13] Teh, Y. W.: A hierarchical Bayesian language model based on Pitman-Yor processes, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 985–992 (2006).
[14] Mochihashi, D. and Sumita, E.: The infinite Markov model, *Advances in neural information processing systems*, pp. 1017–1024 (2007).
[15] Goldwater, S. and Griffiths, T.: A fully Bayesian approach to unsupervised part-of-speech tagging, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Vol. 45, No. 1, Citeseer, pp. 744–751 (2007).

3 ページ

< 誤 >

$$P(w_i|w_{i-N+1}^{i}, z_i) = \frac{c(w_i|w_{i-N+1}^{i-1}, z_i) - d_{|h|}t_{hw_i}}{\theta_{|h|} + c(w_{i-N+1}^{i-1})} + \frac{\theta_{|h|} + d_{|h|}t_{h\cdot}}{\theta_{|h|} + c(w_{i-N+1}^{i-1}, z_i)P(w_i|w_{i-N+2}^{i-1}, z_i)} \tag{5}$$

< 正 >

$$P(w_i|w_{i-N+1}^{i}, z_i) = \frac{c(w_i|w_{i-N+1}^{i-1}, z_i) - d_{|h|}t_{hw_i}}{\theta_{|h|} + c(w_{i-N+1}^{i-1}, z_i)} + \frac{\theta_{|h|} + d_{|h|}t_{h\cdot}}{\theta_{|h|} + c(w_{i-N+1}^{i-1}, z_i)}P(w_i|w_{i-N+2}^{i-1}, z_i) \tag{5}$$

5 ページ 5.2.1 節 5 行目

< 誤 >

累積密度関数

< 正 >

累積分布関数