隠れセミマルコフモデルに基づく教師なし完全形態素解析

内海 慶　　塚原 裕史　　持橋 大地
デンソーアイティーラボラトリ　統計数理研究所
{kuchiumi, htsukahara}@d-itlab.co.jp　daichi@ism.ac.jp

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学習を行う．

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コストが大きい．そのため，教師なし学習手法が望まれる．

これまでにも文字列の生データ，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 [9][12] [3][6] [5][7] [11]．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 章では教師なし形態素解析の先行研究について説明を行い，3 章で，我々の提案する形態素解析手法について述べる．4 章では，我々の手法について評価を行い，その効果を示す．5 章では，総論を行い，今後の課題を示す．

2　関連研究

教師なし学習に基づく形態素解析手法では，最小記述長 (MDL) に基づく手法がこれまで多く提案されている [2] [1][11] [7]．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．他にも，ベイズ学習に基づく手法 [9] や，識別モデルとベイズモデルの協調学習を行う手法 [12] も提案されている．

これまで提案されてきた教師なし学習に基づく形態素解析手法では，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．すなわち，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　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．

3.1　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．

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

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

$P(w_i, z_i|h_{i-1})$ を，(2) 式のように変形する．

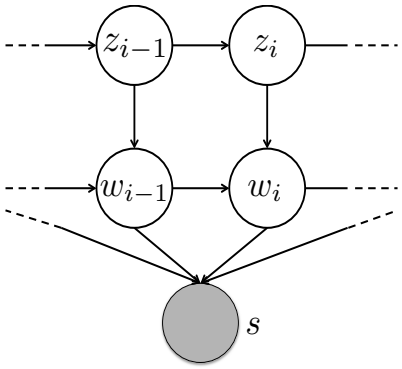

図 1　提案手法の生成モデル

$$P(w_i, z_i|h_{i-1}) = P(w_i|z_i, h_{i-1})P(z_i|h_{i-1}) \quad (2)$$

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

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

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．

3.2　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 [10] 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 [8] 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，[10] に従って行う．

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

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

3.3　学習

我々の学習アルゴリズムも持橋ら [9] 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．

3.3.1　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　評価

我々は日本語，中国語，2 つの言語でアルゴリズムの動作を検証した．日本語の定量的な評価には京都大学テキストコーパス[*1]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[*2]を用いる．

4.1　評価データ

4.1.1　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．

4.1.2　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．

[*1] http://nlp.ist.i.kyoto-u.ac.jp
[*2] http://www.sighan.org/bakeoff2005/

4.2 実験条件

表 1 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 [9] 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．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 30 とした．

表 1　評価データのサイズ

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	40K	10K	1K
SIGHAN MSR	90K	10K	4K
SIGHAN CITYU	50K	10K	1K

4.3 実験結果

4.3.1 単語分割

単語分割の評価には F 値を用いた．

表 2　教師なし単語分割の評価

-	PYHMM	NPYLM	Zhikov 2010
京大コーパス	0.700	0.680	**0.713**[*9]
MSR	**0.806**	0.802[*8]	0.782[*10]
CITYU	0.767	**0.824**[*8]	0.787[*10]

表 2 に，教師なし学習での分かち書きの評価結果を表す．表 3 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，教師なし学習の実験と訓練データ，テストデータのサイズを揃えてはいるが，別途サンプリングされたものであるため直接的に比較できる数値ではないことに注意されたい．

ベイズ学習手法の比較として [9] の数値を，MDL に基づく手法との比較として [11] の数値を記載した．中国語については MSR では NPYLM を F 値で上回った．使用したデータ数が他の手法と比較して 1/9 であることを考慮するとこれはかなり高い数値と言える．一方，CITYU のデータでは他の手法と比較して低い数値となった．データ数が他の手法と比較して 1/5 ということもあるが，他にも CITYU と MSR では正解のアノテーションに違いがあるということが挙げられる．提案手法では，[9] と同様にポアソン分布による補正を入れているが，ポアソン分布のパラメータ λ をサンプリングする際のガンマ分布のハイパーパラメータ a_0, b_0 は今回の実験では共に 1 に固定し，特に調整を行わなかった．コーパスの単語の長さに合わせて数値の調整を行うことで，精度の改善が行えると考えられる．

表 3　半教師あり単語分割の評価

-	PYHMM	Mochihashi 2009
京大コーパス	**0.930**	0.913[*8]

4.3.2 品詞推定

品詞推定は京大コーパスのみで行った．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[4] 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 4 に表す．半教師あり学習での品詞推定の評価結果を表 5 に表す． 表 2，表 4 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分かる．

表 4　教師なし品詞推定の精度

-	PYHMM	NPYLM+BayesianHMM
京大コーパス	**0.598**	0.508

表 5　半教師あり品詞推定の精度

-	PYHMM
京大コーパス	0.894

コーパス作成が困難なデータの例として，方言の解析を行う．方言のデータは，別途 Twitter [*3]から収集した，三河弁の文章を用いた．三河弁のツイートのデータは約 1000 件である．解析結果を図 2 に示す．定量的な評価は正解が付与されていないため，ここでは行わない．各潜在クラスに対応する獲得された単語のリストを表 6 に示す．潜在クラスの 2 には助詞が，3，13 には三河弁特有の文末表現が集まっている．また，36 には形容詞が多く表れている．他にも，41 には地名がまとまっているなど，意味的に近いものがまとまる傾

[*8] [9] より引用した
[*9] [11] より引用した
[*10] [7] より引用した
[*3] http://twitter.com

向が見られる．

```
ウェーブスタジアム/34 刈谷/28 に/2 ＦＣ/1 刈谷/28 の/2 試合/31 を/2 観/35 に/2 行/27 って/40 み/35 りん/3
フォロバ/17 ありがと/19 ございます/19 ！ /2 よろしく/19 頼む/35 のん/3 。/10
これ/20 ぎし/37 しか/37 ない/12 だ/12 かん/3 ？ /10
あけおめ/19 だ/12 ぞん/3 ！！ /10 今年/18 も/2 よろしく/19 頼む/35 ぞん/3 ！！ /10
おま/13 ー/5 の/2 頭/25 、/2 ちんじゅう/35 だのん/3 ！！ /8 w/8
ぐろ/36 と/24 も/2 言/15 う/12 のん/3 ！！ /8
とちんこで/35 結んで/19 まったもん/12 で/12 、/2 と/37 れ/12 や/45 せん/13 に/13 ー/5 (^_^;)/10
のんほい/12 は/2 若い/24 人/20 は/2 あんまし/30 使/15 わん/12 ぞん/3 ！/10 じいさん/24 、/2 ばあさん/37 世代/25 の/2 言葉/27 だ/12 のん/3 ！ /8  /6
```

図 2　三河弁ツイートの解析例

表 6　潜在クラスに対応する獲得した単語

2	の 、 は に が で と も を 「
3	ぞん かん ね のん だに だん りん か ん だのん
9	(*^^*) ！ (^-^; (^_^;) (^^;; ！ (^^;; (^-^;
10	。 ！ ！！ ？ 」 (≧∇≦) ！！」「
11	楽 入 ど寒 大丈夫 会 受 停電 良 美味 台風が
13	に ら わ な よ ね だら じゃんね え ぁ
18	今年 最近 豊川 地元 誰 豊田 今度 次 豊川高校
19	さん ん め 食べ って よろしく ありがとう じゃん
20	これ 知 人 それ どこ まあ みんな 東京 いや 方
24	三河弁 この よ お 何 そ ほい 今日 また ほ
25	他 一緒 5 大変 頭 春 参加 指 世代 地域
26	マジ 豊橋カレー コレ トキワ コーヒー プロ ファン
27	行 」 方言 & 言葉 普通 夜店 始 確認
29	(！ (; (´・ ！！ (*｀ ？ (´・ (*^_^*)
30	気 うち 店 ほう ここ こっち 先生 友人 いろいろ こういう
31	女子 無理 決 近い 安心 標準語 感動 蒲郡 試合
32	((*\(^ ＼ (^ (^ ！ *\(^ 〜 (^_^ (*^
34	ヤマサ マーラ オレ ハイジ イメージ クッピーラムネ ツボ
35	な ー そう 好き こと らん なん ら み 意味
36	いい ど うまい 杏果 ぐろ めっちゃ かわい はよ
41	豊橋 名古屋 三河 西三河 名古屋弁 名古屋人 大阪

5　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．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 CI TYU のデータで NPYLM を下回ったが，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．

参考文献

[1] Shlomo Argamon, Navot Akiva, Amihood Amir, and Oren Kapah. Efficient unsupervised recursive word segmentation using minimum description length. In *Proceedings of the 20th international conference on Computational Linguistics*, p. 1058. Association for Computational Linguistics, 2004.

[2] Marco Baroni, Johannes Matiasek, and Harald Trost. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, pp. 48–57. Association for Computational Linguistics, 2002.

[3] Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, Vol. 112, No. 1, pp. 21–54, 2009.

[4] Sharon Goldwater and Tom Griffiths. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Vol. 45, pp. 744–751. Citeseer, 2007.

[5] Daniel Hewlett and Paul Cohen. Fully unsupervised word segmentation with bve and mdl. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 540–545. Association for Computational Linguistics, 2011.

[6] Daniel Hewlett and Paul R Cohen. Bootstrap voting experts. In *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 1071–1076, 2009.

[7] Pierre Magistry, Benoît Sagot, et al. Can mdl improve unsupervised chinese word segmentation? In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pp. 2–10, 2013.

[8] Daichi Mochihashi and Eiichiro Sumita. The infinite markov model. In *Advances in neural information processing systems*, pp. 1017–1024, 2007.

[9] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *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*, pp. 100–108. Association for Computational Linguistics, 2009.

[10] Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 985–992. Association for Computational Linguistics, 2006.

[11] Valentin Zhikov, Hiroya Takamura, and Manabu Okumura. An efficient algorithm for unsupervised word segmentation with branching entropy and mdl. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 832–842. Association for Computational Linguistics, 2010.

[12] 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.