1N2-1in1

Closed class に着目した教師なし品詞タグ推定性能向上の検討

On the Awareness of Closed Classes in Unsupervised POS Induction

柴田 尚樹 *1　　若林 啓 *2
Hisaki SHIBATA　　Kei WAKABAYASHI

*1筑波大学 図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba

*2筑波大学 図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba

Unsupervised part of speech (POS) induction plays an important role in the natural language processing, especially when the available linguistic resource in the target domain is insufficient. The principle of POS induction essentially relies on the assumption of the existence of a universal nature in the concept of word class shared throughout every languages. In this paper, we shed a light on the usefulness of the open/closed class, which is considered as one of the universal property of linguistic category of words. We set up the experimental method which explicitly integrates the concept of open/closed class with the training algorithm for hidden Markov models (HMMs) that is a major approach in the POS induction. The empirical result that compares with a standard HMM training supports the fact that the explicit awareness of closed class can improve the accuracy of POS induction.

1. はじめに

品詞タグ推定は，自然言語処理の基本的なタスクである．教師なし品詞タグ推定は，コーパスが十分に整備されていない言語への応用が考えられるタスクであり，しばしば，隠れマルコフモデル (Hidden Markov Model，HMM) が用いられる．

HMM を用いた場合，各単語に付与される状態を正解の品詞と照らし合わせることによって評価が行われる．

統語論における品詞の分類に，オープンクラス (Open class) とクローズドクラス (Closed class) という分類がある [田中 13]. オープンクラスには，新たな単語が増えやすい品詞 (名詞，動詞など) が属し，クローズドクラスには，増えにくい品詞 (前置詞，接続詞など) が属す．

この点に着目して，HMM による学習結果を観察すると，クローズドクラスに属す単語が大部分を占めている中に，オープンクラスに属す単語が少ない出現回数として多種類存在する状態が確認できる．単語に付与された状態を品詞とみなすため，このような状態からオープンクラスに属す単語を別の隠れ状態に移動させることで教師なし品詞タグ推定性能の向上が考えられる．

本研究では，明示的にクローズドクラスに着目することが教師なし品詞タグ推定に与える影響を確かめるべく，上述の点に着目した手法をベースとなる HMM に組み込む実験を行った．ベースとなる HMM のパラメータ学習手法は，Collapsed Variational Bayes(CVB) 法とした．

以下では，CVB 法を用いた HMM について述べた後，提案手法へと論旨を進める．

2. 隠れマルコフモデル

隠れマルコフモデルを表現するため，以下の記号を定義する．

$$\begin{aligned} \boldsymbol{A}_{ij} &= p(z_t = j | z_{t-1} = i) \\ \boldsymbol{B}_{iv} &= p(x_t = v | z_t = i) \end{aligned}$$

連絡先: 筑波大学 図書館情報メディア研究科 〒 305-8550 茨城県つくば市春日 1-2 E-mail: s1211505@klis.tsukuba.ac.jp

$$\begin{aligned} \boldsymbol{\pi}_i &= p(z_1 = i) \\ \boldsymbol{nA}_{ij} &: \text{状態 } i \text{ の次に状態 } j \text{ に遷移した回数} \\ \boldsymbol{nB}_{iv} &: \text{状態 } i \text{ から単語 } v \text{ が出力された回数} \\ \boldsymbol{n\pi}_i &: t = 1 \text{ で状態 } i \text{ であった回数} \\ x_t^{(d)} &: d \text{ 番目の系列データの時刻 } t \text{ における出力記号} \\ z_t^{(d)} &: d \text{ 番目の系列データの時刻 } t \text{ における状態} \\ \boldsymbol{X}^{(d)} &= [x_1^{(d)}, x_2^{(d)}, \ldots, x_t^{(d)}, \ldots, x_T^{(d)}] \\ \boldsymbol{Z}^{(d)} &= [z_1^{(d)}, z_2^{(d)}, \ldots, z_t^{(d)}, \ldots, z_T^{(d)}] \\ \boldsymbol{X} &= [\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(D)}] \\ \boldsymbol{Z} &= [\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \ldots, \boldsymbol{Z}^{(D)}] \end{aligned}$$

ただし，D は独立な系列データの数を，t は系列データの時刻を表している．T は系列データごとに定義される系列データの長さである．英語コーパスを対象としたことにより，以下では，系列データ中の一つの出力記号を単語，単語についての系列データ一つを文章，HMM に用いる文章を全てまとめたものを文書とも表記する．

2.1 ハイパーパラメータと事前分布

多項分布のパラメータである $\boldsymbol{A}$, $\boldsymbol{B}$,$\boldsymbol{\pi}$ の事前分布にディリクレ分布を導入する．ディリクレ分布に関するパラメータ $\boldsymbol{\alpha}$ および $\boldsymbol{\beta}$ を用いて，これらを定義すると以下のようになる．ただし，$\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$，$N$ を状態数，V を語彙数とする．

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{A})p(\boldsymbol{B}) \quad (1)$$

$$p(\boldsymbol{\pi}) = \frac{\Gamma(\sum_{i=1}^{N} \alpha_i)}{\prod_{i=1}^{N} \Gamma(\alpha_i)} \prod_{i=1}^{N} \pi_i^{\alpha_i - 1} \quad (2)$$

$$p(\boldsymbol{A}) = \prod_{i=1}^{N} p(\boldsymbol{A}_i) = \prod_{i=1}^{N} \frac{\Gamma(\sum_{j=1}^{N+1} \alpha_j)}{\prod_{j=1}^{N+1} \Gamma(\alpha_j)} \prod_{j=1}^{N+1} A_{ij}^{\alpha_j - 1} \quad (3)$$

$$p(\boldsymbol{B}) = \prod_{i=1}^{N} p(\boldsymbol{B}_i) = \prod_{i=1}^{N} \frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} B_{iv}^{\beta_v - 1} \quad (4)$$

ディリクレ分布のパラメータ $\boldsymbol{\alpha}$ および $\boldsymbol{\beta}$ はパラメータ $\boldsymbol{\pi}$, $\boldsymbol{A}$, $\boldsymbol{B}$ の分布を決めるハイパーパラメータである．

2.2 Collapsed Variational Bayes 法を用いた HMM によるパラメータ推定

HMM を用いた学習では，任意の所与の $\boldsymbol{\theta}' = \{\boldsymbol{\pi}', \boldsymbol{A}', \boldsymbol{B}'\}$ に対して，式 (5) が成立する $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$ を求める．

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) - \log p(\boldsymbol{X}|\boldsymbol{\theta}') \geq 0 \tag{5}$$

$\boldsymbol{\theta}$ を新たな $\boldsymbol{\theta}'$ とし，式 (5) による計算を繰り返すことで，反復的に対数尤度 ($\log p(\boldsymbol{X}|\boldsymbol{\theta})$) を更新する．

以下では，反復的なパラメータ推定手法に Collapsed Variational Bayes 法を用いた HMM(CVBHMM) のパラメータ推定手法を述べる．

対数尤度を以下のように変形する．

$$\begin{aligned}
\log p(\boldsymbol{X}) &= \log \sum_{\boldsymbol{Z}} \int p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}) d\boldsymbol{\theta} && (6)\\
&= \log \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \frac{\int p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{q(\boldsymbol{Z})} && (7)\\
&\geq \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \frac{\int p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{q(\boldsymbol{Z})} && (8)
\end{aligned}$$

ここで，以下の平均場近似を仮定する．

$$q(\boldsymbol{Z}) = \prod_{d=1}^{D} q_d(\boldsymbol{Z}^{(d)}) \tag{9}$$

この平均場近似を適用した分布の最大化を目指す．すなわち式 (9) を式 (8) に適用した式

$$F(q_1, ..., q_D) = \sum_{\boldsymbol{Z}} \left\{ \prod_{d=1}^{D} q_d(\boldsymbol{Z}_d) \times \log \frac{\int p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\prod_{d=1}^{D} q_d(\boldsymbol{Z})^{(d)}} \right\} \tag{10}$$

が最大化を目指す下限である．

この方法では，D 個の分布を交互に最適化する．q_d を最適化する際には，q_d を除く全ての分布を固定する．q_d は以下の近似式で与えられる．

$$q_d(\boldsymbol{Z}^{(d)}) \approx \tilde{\pi}_{z_1^{(d)}} \prod_{t=1}^{T} \tilde{A}_{z_t^{(d)} z_{t+1}^{(d)}} \tilde{B}_{z_t^{(d)} x_t^{(d)}} \tag{11}$$

$\tilde{\pi}_i$，$\tilde{A}_{ij}$，$\tilde{B}_{iv}$ は以下の通りである．

$$\begin{aligned}
\tilde{\pi}_i &= \frac{\tilde{n}_{\pi i}^{\backslash d} + \alpha_i}{\sum_{i'=1}^{N} (\tilde{n}_{\pi i'}^{\backslash d} + \alpha_i')} && (12)\\
\tilde{A}_{ij} &= \frac{\tilde{n}_{Aij}^{\backslash d} + \alpha_j}{\sum_{j'=1}^{N+1} (\tilde{n}_{Aij'}^{\backslash d} + \alpha_{j'})} && (13)\\
\tilde{B}_{iv} &= \frac{\tilde{n}_{Biv}^{\backslash d} + \beta_v}{\sum_{v'=1}^{V} (\tilde{n}_{Biv'}^{\backslash d} + \beta_{v'})} && (14)
\end{aligned}$$

ただし $\boldsymbol{Z}_{\backslash d}$ は,$\boldsymbol{Z}$ から $\boldsymbol{Z}^{(d)}$ を除いた系列集合を，$\tilde{\boldsymbol{Z}}_{\backslash d}$ は,$\boldsymbol{Z}_{\backslash d}$ からサンプルされた文章 d についての状態系列を表している．$\tilde{n}_{\pi i}^{\backslash d}$，$\tilde{n}_{Aij}^{\backslash d}$，$\tilde{n}_{Biv}^{\backslash d}$ はそれぞれサンプル $\tilde{\boldsymbol{Z}}_{\backslash d}$ においてカウントした初期状態，状態遷移，単語出力の回数である．

Algorithm 1 CVBHMM におけるパラメータの学習

```
α, β, π, A, B を初期化
ite ⇐ 0
while ite < Iteration do
  for d = 1 to D do
    if ite ≥ 1 then
      REMOVE Z̃^(d)
    end if
    UPDATE π
    for i = 1 to N do
      UPDATE A_i
      UPDATE B_i
    end for
    X^(d) に対応する Z̃^(d) を p(Z^(d)|Z_\d, X^(d)) に従う分布からサンプリング．...※
    X^(d) と Z̃^(d) を用いることで得られる各状態からの状態遷移数，各状態からの単語の出現数を nA, nB, nπ の該当する部分に加える．
    p(X^(d)) を尤度に加える．
  end for
  ite ⇐ ite + 1
  UPDATE Hyper Parameter
end while
```

以上のようにして，尤度の最大化を目指す方法は Collapsed Variational Bayes(CVB) 法と呼ばれる．Algorithm(1) は上記の CVBHMM によるパラメータ更新アルゴリズムである．

上記アルゴリズムについて補足する．

$REMOVE\ \tilde{\boldsymbol{Z}}^{(d)}$ では，前回の反復でサンプリングした $\tilde{\boldsymbol{Z}}^{(d)}$ について，各状態からの状態遷移数，各状態からの単語の出現数を $\boldsymbol{nA}$, $\boldsymbol{nB}$, $\boldsymbol{n\pi}$ の該当する部分から差し引いている．

$UPDATE\ \boldsymbol{\pi}$ では式 (12) を，$UPDATE\ \boldsymbol{A}_i$ では式 (13) を，$UPDATE\ \boldsymbol{B}_i$ では式 (14) を求めている．サンプリングには FB 法 [Scott 02] を用いた．$UPDATE\ Hyper\ Parameter$ については 2.3 節に記述した．変数 $Iteration$ はパラメータであり，反復学習回数をあらかじめ設定する．

2.3 ハイパーパラメータの学習

$\boldsymbol{A}$ と $\boldsymbol{\pi}$ に関係するハイパーパラメータ $\boldsymbol{\alpha}$ のみ，学習を行う．

多項分布の $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$ の事前分布として導入するディリクレ分布のパラメータ $\boldsymbol{\alpha}$ の下で系列データ $\boldsymbol{X}$ が得られる確率 $p(\boldsymbol{X}|\boldsymbol{\alpha})$ は以下のように変形される．

$$\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{\alpha}) &= \prod_{d=1}^{D} \int p(\boldsymbol{X}_d|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} && (15)\\
&= \prod_{d=1}^{D} \frac{\Gamma(\sum_{i=1}^{N} \alpha_i)}{\Gamma(n_d + \sum_{i=1}^{N} \alpha_i)} \prod_{i=1}^{N} \frac{\Gamma(n_{di} + \alpha_i)}{\Gamma(\alpha_i)} && (16)
\end{aligned}$$

ただし，n_{di} は $\boldsymbol{X}^{(d)}$ において i の値をとるものの数，$n_d = \sum_{i=1}^{N} n_{di}$ である．式 (16) は $\boldsymbol{\alpha}$ について凸関数であるが，解析的に求めることはできない．今，$\boldsymbol{X}$ が観測されているとした場合，ハイパーパラメータ $\boldsymbol{\alpha}$ は $p(\boldsymbol{X}|\boldsymbol{\alpha})$ を大きくするように更新することができる．Algorithm(1) の $UPDATE\ Hyper\ Parameter$ では，Wallach[Wallach 08] による Method 1 によってハイパーパラメータを更新している．

Method 1 は，Minka の 2000 年のテクニカルレポート

[Minka 00] で提案された fixed-point iteration 法を効率的に計算するアルゴリズムである．

Minka の fixed-point iteration では，定数 $\boldsymbol{\alpha}'$ が与えられたとき，式 (16) の対数について以下の不等式が成り立つことを利用して，反復最適化を行う．

$$\log p(\boldsymbol{X}|\boldsymbol{\alpha}) \geq \sum_{d=1}^{D}\left[-\left\{\Psi(n_d+\sum_{i=1}^{N}\alpha_i')-\Psi(\sum_{i=1}^{N}\alpha_i')\right\}\sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}\left\{\alpha_i'(\Psi(n_{di}+\alpha_i')-\Psi(\alpha_i'))\log\alpha_i\right\}\right]+C$$

右辺は解析的に溶ける凸関数である．微分して 0 と置くことで，右辺を最大化する $\hat{\boldsymbol{\alpha}}$ が以下のように得られる．

$$\hat{\alpha}_i = \alpha_i' \frac{\sum_{d=1}^{D}\Psi(n_{di}+\alpha_i')-\Psi(\alpha_i')}{\sum_{d=1}^{D}\Psi(n_d+\sum_{i'=1}^{N}\alpha_{i'}')-\Psi(\sum_{i'=1}^{N}\alpha_{i'}')} \tag{17}$$

式 (17) は以下のように変形される．$C_i(n)$ は，$n_{di}=n$ であるような $\boldsymbol{X}^{(d)}$ の数であり，$C_{\cdot}(n)$ は含まれる変数の数が n であるような $\boldsymbol{X}^{(d)}$ の数である．

$$\hat{\alpha}_i = \alpha_i' \frac{\sum_{d=1}^{D}\sum_{f=1}^{n_{di}}\frac{1}{f-1+\alpha_i'}}{\sum_{d=1}^{D}\sum_{f=1}^{n_d}\frac{1}{f-1+\sum_{i'=1}^{N}\alpha_{i'}'}} \tag{18}$$

$$= \alpha_i' \frac{\sum_{n=1}^{\max_d n_{di}} C_i(n)\sum_{f=1}^{n}\frac{1}{f-1+\alpha_i'}}{\sum_{n=1}^{\max_d n_d} C_{\cdot}(n)\sum_{f=1}^{n}\frac{1}{f-1+\sum_{i'=1}^{N}\alpha_{i'}'}} \tag{19}$$

式 (19) により得られる $\hat{\boldsymbol{\alpha}}$ を新たに $\boldsymbol{\alpha}'$ とし反復的に適用することで，式 (16) の最適化を行う．これが Wallach の Method 1 である．

3. 提案手法

本章では，クローズドクラスを考慮した手法の有効性を検討するため，閉じた隠れ状態という言葉を定義した後，ベースとなる CVBHMM に手を加えた手法を述べる．

3.1 閉じた隠れ状態

本研究では，その状態に属す単語の出現回数の一覧から，クローズドクラスに属す単語とオープンクラスに属す単語の推定できると考えられる状態を “閉じた隠れ状態” と定義する．

具体的には，その状態に属す単語のうち，出現回数が多い単語はクローズドクラスに属し，出現回数が少ない単語はオープンクラスに属す，と考えられる状態のことである．

どれほどの出現回数を多いとみなすか，少ないとみなすかについては実験ごとに定める．また，閉じた隠れ状態に属さない状態は “開いた隠れ状態” と呼ぶこととする．

3.2 手法

実験ごとに，どのような状態を閉じた隠れ状態とみなすか，どのような出現回数の単語を移動対象とするかを定め，CVBHMM における $\tilde{\boldsymbol{Z}}^{(d)}$ の推定ごとに，

1. $\tilde{z}_t^{(d)}$ が，閉じた隠れ状態であるか判別し，
2. $\tilde{z}_t^{(d)}$ が閉じた隠れ状態であり，かつ $\tilde{z}_t^{(d)}$ において $x_t^{(d)}$ が出現回数が少ない単語であると判断される場合，$\tilde{z}_t^{(d)}$ を開いた隠れ状態で置き換える

という操作を追加した．これは，Algorighm(1) における操作 (※) の後に，上記の操作を追加するということである．

このアルゴリズムのステップ 1，ステップ 2 の具体的な方法として，本研究では 2 つの基準を用い，それぞれ方法 1，方法 2 と呼ぶ．方法 1 は，状態 i が閉じた隠れ状態であるかどうかを判別するとき，出力回数 nB_{iv} において出現回数 1 位の単語 $\tilde{v}=\arg\max_v nB_{iv}$ のみの情報を用いる．ステップ 1 では，状態 i において単語 $\tilde{v}$ の出力回数が 5 割を占めていた場合，すなわち $nB_{i\tilde{v}} \geq 0.5\times\sum_v nB_{iv}$ のとき，状態 i を閉じた隠れ状態と判定する．ステップ 2 では，状態 $\tilde{z}_t^{(d)}$ において，$v=x_t^{(d)}$ が「出現回数の少ない単語」であるかどうかを判定するが，この基準として方法 1–1，方法 1–2 の 2 種類を実験する．方法 1–1 では，当該の単語の出現回数 nB_{iv} が 1 の場合のみを「出現回数の少ない単語」とする．方法 1–2 では，出現回数の割合 $\frac{nB_{iv}}{\sum_{v'} nB_{iv'}}$ が 0.001 未満の場合に「出現回数の少ない単語」とする．いずれの場合においても，閉じた隠れ状態 i において出現回数が少ないと判断された単語に状態 i が割り当てられている場合には適切でないと考え，別の状態に置き換える対象とする．

方法 2 では，単語の出現回数の分布がなだらかさに基づいた判別を行う．状態 i において出現回数が最も多い単語を ν_{i1}，2 番目に多い単語を ν_{i2}，... と並べた配列を $\nu_i=\{\nu_{i1},...,\nu_{iV}\}$ とする．方法 1–2 は，$t=1$ から $V-1$ について $\nu_{it}\geq\eta\nu_{it+1}$ を満たしているかどうかを判定し，1 つでも条件を満たす t がある場合には状態 i を閉じた隠れ状態と判定する．η が 0.2 と 0.25 について実験を行い，これをそれぞれ方法 2–1，方法 2–2 と呼ぶ．閉じた隠れ状態と判定したら，$\nu_{it}\geq\eta\nu_{it+1}$ を満たす最小の t を t' としたとき，順位 t' 以降の全ての単語 $\{\nu_{it'+1},...,\nu_{iV}\}$ を出現回数が少ない単語と判断する．これらの単語に状態 i が割り当てられている場合には適切でないと考え，別の状態に置き換える対象とする．

上記の手順によって置き換える対象となった状態 $\tilde{z}_t^{(d)}$ が割り当てられている単語を $x_t^{(d)}$ とすると，状態 $\tilde{z}_t^{(d)}$ の置き換えは，単語 $x_t^{(d)}$ の割り当て先を変更することに対応する．本稿では，状態の置き換えを単語の割当先の変更という観点で見るとき，「単語 $x_t^{(d)}$ の移動」という用語を用い，割当先のことを「単語の移動先」と呼ぶ．

単語の移動先を決定する際には，サンプリングによって推定した状態系列 $\tilde{\boldsymbol{Z}}^{(d)}$ について，単語を移動させる対象となる時刻の状態のみを別の状態で書き換える必要がある．それ以外の状態は変更する必要がない．このため，単語を別状態に書き換える必要がない時刻の状態を固定したビタービアルゴリズムを用いて，新たに状態系列を得る．ただし，移動先の状態の候補は開いた隠れ状態のみとする．

4. 実験

提案手法による教師なし品詞タグ推定性能の向上の有無を検討するため，CVBHMM，方法 1–1，方法 1–2，方法 2–1，方法 2–2 の 5 つの実験条件で評価指標の比較を行う．CVBHMM は，Algorighm(1) で示した CVB 法を用いた HMM によるパラメータ推定であり，クローズドクラスを考慮していない方法である．

正解の品詞タグが付与された Wall Steet Journal(WSJ) コーパスを実験に用いる．学習の反復回数 $Iteration$ は 100 とし，状態数は全ての実験条件で 50 とする．$\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ の各値の初期値は，0.001 とする．

学習終了後，学習時の実験条件に従い再度全文章について

サンプリングを行い，各単語に状態を付与したものを用いて，推定結果とする．

上記 5 つの実験条件について，実験条件毎に 5 回独立な実験 (合計 25 実験) を行う．その後，学習終了後のサンプリングで得られた状態系列集合を 4.1 節で述べる評価に用いる．

4.1 評価方法

Many-to-One と V-Measure を用いる [Christodoulopoulos 10]．加え，独自定義の Vocabulary Based Many-to-One を用いる．

4.1.1 Vocabulary Based Many-to-One

Vocabulary Based Many-to-One(VBMO) は，本研究で独自に定義した評価手法である．以下は，VBMO の算出方法である．

1. Many-to-One と同様にして，各状態を品詞に当てはめる
2. 単語種類ごとに，正解率を算出
3. 正解率の平均値を，VBMO とする

VBMO は，一般に正しく品詞が推定された単語が増えるほど向上する．加え，正しく品詞推定される単語一つについて，出現回数の少ない単語による上がり幅が出現回数の多い単語による上がり幅より大きい，という性質を持つ．提案手法では，出現回数が少ない単語について正解率向上が見られると考えられることから VBMO を用いた．

4.2 結果と考察

実験の結果を評価手法ごとに表 1，表 2，表 3 にまとめた．

表 1: Many-to-One についての評価

実験方法	1 回目	2 回目	3 回目	4 回目	5 回目	平均
CVBHMM	0.5418	0.5269	0.5567	0.5405	0.5436	0.5419
方法 **1–1**	0.5543	0.5611	0.5417	0.5453	0.5482	0.5501
方法 **1–2**	0.5326	0.5622	0.5238	0.5517	0.5439	0.5428
方法 **2–1**	0.5663	0.5612	0.5587	0.5598	0.5534	0.5598
方法 **2–2**	0.5258	0.5149	0.5588	0.5606	0.5572	0.5434

表 2: V-Measure についての評価

実験方法	1 回目	2 回目	3 回目	4 回目	5 回目	平均
CVBHMM	0.4595	0.4492	0.4720	0.4660	0.4788	0.4651
方法 **1–1**	0.4777	0.4829	0.4689	0.4690	0.4734	0.4743
方法 **1–2**	0.4629	0.4925	0.4620	0.4779	0.4738	0.4738
方法 **2–1**	0.4934	0.4921	0.4952	0.4951	0.4782	0.4908
方法 **2–2**	0.4746	0.4442	0.5015	0.4950	0.5004	0.4831

表 3: Vocabulary Based Many-to-One についての評価

実験方法	1 回目	2 回目	3 回目	4 回目	5 回目	平均
CVBHMM	0.3248	0.2756	0.2853	0.2993	0.3216	0.3013
方法 **1–1**	0.2724	0.3270	0.3404	0.3200	0.3042	0.3128
方法 **1–2**	0.3277	0.3036	0.2754	0.3234	0.3082	0.3076
方法 **2–1**	0.3218	0.3121	0.3039	0.3013	0.3404	0.3159
方法 **2–2**	0.3185	0.3036	0.2874	0.2979	0.3227	0.3060

すべての評価手法において若干の向上が見られる．

これより，クローズドクラスを考慮に入れることは，教師なし品詞タグ推定性能の向上につながると考えられる．

先行研究では，Many-to-One にて 0.8023，V-Measure にて 0.7207 という結果が得られている [Yatbaz 12]．HMM を用いた研究では，Many-to-One にて 0.775，V-Measure にて 0.698 という結果が得られている [Blunsom 11]．

先行研究においても，クローズドクラスを考慮した教師なし品詞タグ推定性能の向上を検討する必要がある．

5. 結論

本論文では，教師なし品詞タグ推定におけるクローズドクラスの考慮の必要性について検討した．

クローズドクラスに着目した HMM では，Many-to-One，V-Measure に加え，独自に定義した Vocabulary Based Many-to-One において若干の精度向上を確認できた．これより，クローズドクラスは教師なし品詞タグ推定において考慮すべき要因であると言える．

今後の展望としては，明示的にクローズドクラスを考慮する手法の先行研究への適用が挙げられる．

謝辞

本研究の一部は，JSPS 科研費（課題番号 25280110,25540159）および筑波大学図書館情報メディア系プロジェクト研究（Research Projects of Facalty of Library, Information and Media Science）の助成によって行われた．

参考文献

[Blunsom 11] Blunsom, P. and Cohn, T.: A Hierarchical Pitman-Yor Process HMM for Unsupervised Part of Speech Induction, in *Proc. ACL* (2011)

[Christodoulopoulos 10] Christodoulopoulos, C., Goldwater, S., and Steedman, M.: Two Decades of Unsupervised POS Induction: How Far Have We Come?, in *Proc. EMNLP* (2010)

[Minka 00] Minka, T. P.: Estimating a Dirichlet distribution, Technical report, MIT (2000)

[Scott 02] Scott, S. L.: Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century, *Journal of the American Statistical Association* (2002)

[Wallach 08] Wallach, H. M.: *Structured topic models for language*, PhD thesis, University of Cambridge (2008)

[Yatbaz 12] Yatbaz, M. A., Sert, E., and Yuret, D.: Learning Syntactic Categories Using Paradigmatic Representations of Word Context, in *Proc. EMNLP-CoNLL* (2012)

[田中 13] 田中 智之：統語論, 朝倉書店 (2013)