半教師あり形態素解析 NPYCRF の修正

藤井 遼　　道本 龍　　持橋 大地
(株) 博報堂 研究開発局　　統計数理研究所
{ryo.b.fujii, ryo.domoto}@hakuhodo.co.jp　daichi@ism.ac.jp

1 はじめに

先に持橋らが提案した NPYCRF [5] は，識別モデルである CRF と生成モデルである NPYLM [6] を JESS-CM [4] の枠組で統合することで，両者の長所を組み合わせた半教師つき学習による単語分かち書きを行う手法である．

しかし我々は [5] に述べられているモデルの推定法のうち，Markov モデルである CRF と semi-Markov モデルである NPYLM の間で情報を相互変換するアルゴリズムに不備があることを発見した．本研究ではこの誤りを指摘し，代わりに新たなアルゴリズムを与えて NPYCRF の正しい推定法を提案する．

[5] 以降の主に中国語分かち書きにおける最新の研究成果を組み合わせることで，新たな推定法による実験を行い，見通しの良い定式化で一般的に良い精度が得られることを示す．

2 半教師あり形態素解析: NPYCRF

2.1 識別-生成統合モデル

単語分割の生成モデルであり教師なし学習を行うモデルである NPYLM と，精度の高い教師あり学習による識別モデルである CRF を統合することができれば，精度を保ちつつ未知語に適応できるような半教師あり学習を実現することができると考えられる．このために [5] では半教師あり学習法として，鈴木らの JESS-CM 法 [4] を用いている．JESS-CM では，入力 $\boldsymbol{x}$ に対するラベル $\boldsymbol{y}$ の確率 $P_{\mathrm{CONC}}(\boldsymbol{y}|\boldsymbol{x})$ を次式の形で表現する．

$$P_{\mathrm{CONC}}(\boldsymbol{y}|\boldsymbol{x}) \propto P_{\mathrm{DISC}}(\boldsymbol{y}|\boldsymbol{x};\Lambda)P_{\mathrm{GEN}}(\boldsymbol{y},\boldsymbol{x};\Theta)^{\lambda_0} \quad (1)$$

P_{DISC} は識別モデル，P_{GEN} は生成モデルであり，Λ, Θ はそれぞれのパラメータである．ここで，識別モデルに CRF のような対数線形モデルを用いることにすると，(1) 式は $\log P_{\mathrm{GEN}}(\boldsymbol{y},\boldsymbol{x})$ を（二値ではなく連続値を返す）一つの素性関数とみなせば，

$$\begin{aligned} P_{\mathrm{CONC}}(\boldsymbol{y}|\boldsymbol{x}) &\propto \exp(\lambda_0 \log P_{\mathrm{GEN}}(\boldsymbol{y},\boldsymbol{x}) + \sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{y},\boldsymbol{x})) \\ &= \exp(\Lambda^* \cdot F^*(\boldsymbol{y},\boldsymbol{x})) \end{aligned} \quad (2)$$

と，パラメータ $\Lambda^* = (\lambda_0, \lambda_1, \ldots, \lambda_K)$ を持つ対数線形モデルの形でも書くことができる．ここで $F^*(\boldsymbol{y},\boldsymbol{x}) = (\log P_{\mathrm{GEN}}(\boldsymbol{y},\boldsymbol{x}), f_1(\boldsymbol{y},\boldsymbol{x}), \ldots, f_K(\boldsymbol{y},\boldsymbol{x}))$ とおいた．式 (1) と (2) は等価であるから，JESS-CM ではこの 2 式を用いて，教師ありデータ (X_l, Y_l) および教師なしデータ X_u からなるデータについての目的関数

$$P(X_u, Y_l|X_l;\Lambda^*,\Theta) = P(Y_l|X_l)\cdot P(X_u) \quad (3)$$

の値を，

- Θ を固定し (X_l, Y_l) で識別モデルの重み Λ^* を最適化
- Λ を固定し X_u で生成モデル Θ を最適化

という 2 つのステップを交互に行って最大化する．

2.2 NPYCRF におけるモデル変換

JESS-CM では識別モデルと生成モデルが同じ構造を持つことを前提にしているが，NPYLM は semi-Markov モデルであり，CRF は精度面から Markov-CRF を採用したいという理由により，NPYCRF では異なる構造を持つモデルの情報を統合する方法が必要となる．[5] では NPYLM を学習する際に CRF の情報を semi-Markov モデルへ足し合わせる変換と，CRF を学習する際に NPYLM の情報を Markov モデルへ足し合わせる変換の二種類が提案されている．

ある文において単語が位置 a で始まり位置 $b-1$ で終わり，次の単語が位置 b から始まる時，前者の単語を c_a^b と書くことにする．この時 Markov モデル上の 1 単語 c_a^b 分の経路に対応する，区間 $[a,b)$ $(a<b)$ の ポテンシャルを $\gamma(a,b)$ とおけば，これは例えば文頭を 1，それ以外を 0 の 2 値ラベルで表すならば，状態 1 で始まり 1 で終わる V 字型 ($b=a+1$ のときは直線) の区間となる（図 1 左）．

[5] では，semi-Markov モデルにおける単語 c_t^{u-1} から c_s^{t-1} への遷移確率 $P(c_t^{u-1}|c_s^{t-1})$ に対応するポテンシャルとして，

$$\begin{aligned} P(c_t^{u-1}|c_s^{t-1})) &\propto \exp(\gamma(s,t,u)) \\ &= \exp(\gamma(s,t)+\gamma(t,u)) \end{aligned} \quad (4)$$

を用い，NPYLM の学習を前向き確率

$$\begin{aligned} \alpha[t][k] = \sum_{j=0}^{t-k} &\exp\{\lambda_0 \log(P(c_{t-k}^{t-1}|c_{t-k-j}^{t-k-1})) \\ &+ \gamma(t-k-j, t-k, t)\} \cdot \alpha[t-k][j] \end{aligned} \quad (5)$$

を使って forward-filtering backward-sampling アルゴリズムで行っている．

ここで，(4) 式における semi-Markov モデルと Markov モデルの関係は等号 (=) ではなく，比例 ($\propto$) であることに注意する必要がある．比例関係であることは正しいが，semi-Markov モデルにおいては先行文脈は与えられており，右辺と左辺の間のスケールは一般にその文脈によって異なる．しかし (5) 式では，異なる先行文脈 $[s,t)$ 間において $\exp(\gamma(s,t,u))$ を直接足しあわせていることが分かる．これは確率の加算にはなっておらず，結果として正しいモデルの統合，ひいて

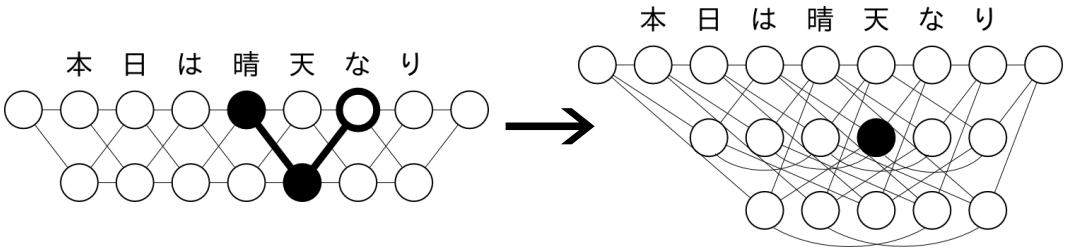

図 1　Markov モデルの semi-Markov モデルへの変換

はモデルの推定にはつながらない.

また，ある特定の単語分割 $\boldsymbol{y}$ に対して (1) 式の右辺を展開すると

$$P_{\mathrm{DISC}}(\boldsymbol{y}|\boldsymbol{x};\Lambda)P_{\mathrm{GEN}}(\boldsymbol{y},\boldsymbol{x};\Theta)^{\lambda_0} = \prod_{\{s,t\}\in\boldsymbol{y}} \exp(\gamma(s,t)) \cdot \prod_{\{s,t,u\}\in\boldsymbol{y}} P(c_t^{u-1}|c_s^{t-1}))^{\lambda_0} \quad (6)$$

と，Markov モデルおよび semi-Markov モデルそれぞれで特定のパス上のポテンシャルをそれぞれ一度ずつ足しあわせた形になっているはずであるが，(5) 式のように足しあわせて行くと，それぞれ $\gamma(s,t)$ が二回ずつ現れることが分かる．この分解を見てもモデルの統合が正しくないことが直感的に理解できる．

同様に，semi-Markov モデルの Markov モデルへの変換においても，(4) 式における等号と比例の混同による誤りがある．[5] では，Markov モデル上で位置 t でのラベル bigram (z_t, z_{t+1}) について場合分けを行い，それぞれについて semi-Markov モデル上で計算しやすい形での周辺化を行うことで変換を試みている．一例を上げると，$\cdots$ の部分の変数は全て 0 だとして

$$p(z_t=0, z_{t+1}=1) = \sum_k \sum_l p(z_k=1,\cdots,z_t=0,z_{t+1}=1,\cdots,z_l=1) \quad (7)$$

などのように周辺化している．一方同じ記法を使うと，(4) 式で定義した $\gamma(s,t,u)$ は

$$\exp(\gamma(s,t,u)) = p(z_s=1,\cdots,z_t=1,\cdots,z_u=1) \quad (8)$$

であるから, 範囲内の各 z_i の値をたどることで (4) 式の左辺である semi-Markov モデル上の確率から Markov モデル上のポテンシャルへの変換が得られた，としている．しかし同様に (7) 式内で異なる文脈間での足し算が行われてしまうので，この変換も正しくない．

3　提案手法

本研究では [1] と同じように Markov モデルの情報を semi-Markov モデルに変換し，semi-Markov モデル上で Markov-CRF の学習を行うことでモデルを推定する．以下, この変換と学習アルゴリズムについて述べる．

3.1　semi-Markov モデル上での統合

[5] におけるポテンシャルの代わりに，semi-Markov モデル上で先行文脈に依らず，現在の単語にのみ依存するポテンシャル（semi-Markov モデル上のパスではなく, ノードで表すこととする）を仮定することで，無理なくモデルを変換することができる．再び図 1 を考えると, semi-Markov モデル上で（任意の先行単語）→「晴天」に当たるポテンシャルは, Markov モデル上の太線で示した経路に沿った重みを足し合わせたものになる．一般に Markov モデル上で確率 $P(c_t^{u-1}|\cdot)$ に対応するポテンシャルは，再び $\gamma(\cdot,\cdot)$ を用いて

$$P(c_t^{u-1}|\cdot)) = \frac{\exp(\gamma(t,u))}{Z(\boldsymbol{x})} \quad (9)$$

と書くことができる．ここで $Z(\boldsymbol{x})$ は CRF の規格化定数である．$\exp(\gamma(t,u))$ と $P(c_t^{u-1}|\cdot)$ が比例関係であることは [5] と変わらないが，その係数が（ある文に対して）全ての P で等しいことから，係数をくくりだした上で直接加算することには問題ないことが分かる．

3.2　統合モデル上での NPYLM のパラメタ更新

(9) 式を用いて, CRF の情報を取り入れた NPYLM の前向き確率は, (5) 式の代わりに

$$\alpha[t][k] = \sum_{j=0}^{t-k} \exp\{\lambda_0 \log(P(c_{t-k}^{t-1}|c_{t-k-j}^{t-k-1})) + \gamma(t-k,t)\} \cdot \alpha[t-k][j] \quad (10)$$

と計算することができ, 後向きサンプリングも同様に行うことができる．

また，(10) 式を終端まで計算すると $\gamma(s,t)$ および $P(c_{t-k}^{t-1}|c_{t-k-j}^{t-k-1})$ が各系列の各単語についてちょうど 1 回ずつ出現するので，特定の系列に注目すれば (6) 式の展開と一致していることが分かり，モデル統合として辻褄が合っていることが分かる．

3.3　統合モデル上での CRF のパラメタ更新

semi-Markov モデルを Markov モデルへ変換する代わりに，Markov-CRF の勾配計算を直接 semi-Markov モデル上で行うことが出来ることが，[1] で触れられている．一般に CRF の勾配計算は各素性に対する周辺確率が得られれば十分であることから，必要な周辺確率の計算方法について述べる．

semi-Markov モデル素性に対する勾配は通常の semi-Markov モデルにおける Forward-Backward アルゴリズムと同様の方法で求められる．求める周辺確率は $P(c_{t-k-j}^{t-k-1}, c_{t-k}^{t-1}|\boldsymbol{x})$ の形であるから,

$$P_{\mathrm{CONC}}(c_{t-k-j}^{t-k-1}, c_{t-k}^{t-1}|\boldsymbol{x}) = \alpha[t-k][j] \cdot \beta[t][k] \cdot \exp\{\lambda_0 \log(P(c_{t-k}^{t-1}|c_{t-k-j}^{t-k-1})) + \gamma(t-k,t)\}/Z(\boldsymbol{x})^* \quad (11)$$

という形となる．$Z(\boldsymbol{x})^*$ は統合されたモデル上での規格化定数である．ここで $\alpha[t][k]$ は (10) 式で計算される

前向き確率であり, $\beta[t][k]$ は

$$\beta[t][k] = \sum_{j=1}^{N} \exp\{\lambda_0 \log(P(c_{t+1}^{t+j}|c_{t-k+1}^{t})) + \gamma(t+1, t+j+1)\} \cdot \beta[t+j][j] \quad (12)$$

で計算される後ろ向き確率である.

一方 Markov-CRF の重み λ_i に対応する周辺確率 $p(z_t, z_{t+1})$ の計算は自明ではない. だが, (7) 式のような周辺化を考えれば, Markov モデル上の各素性が semi-Markov モデル上のいくつかのノードに分散化されていると見なすことが出来る. まず, semi-Markov モデルのノードに対する周辺確率は, 単語 c_{t-k}^{t} に対して

$$P_{\mathrm{CONC}}(c_{t-k}^{t}|\boldsymbol{x}) = \frac{\alpha[t][k] \cdot \beta[t][k]}{Z(\boldsymbol{x})^*} \quad (13)$$

である. そこで, 例えば (7) 式の代わりに

$$\begin{aligned} & p(z_t = 0, z_{t+1} = 1) \\ =& \sum_k p(z_k = 1, \cdots, z_t = 0, z_{t+1} = 1) \\ =& \sum_k P_{\mathrm{CONC}}(c_k^{t+1}|\boldsymbol{x}) \end{aligned} \quad (14)$$

のように周辺化された確率を考えれば, ありうるすべてのパターンについて周辺化された確率 $p(z_t, z_{t+1})$ を計算することができる.

また, あるノードでどの素性が現れているかは

$$\frac{\exp(\gamma(t-k, t))}{Z(\boldsymbol{x})} = p(z_{t-k} = 1, \cdots, z_t = 1) \quad (15)$$

であることに注意すれば, $[t-k, t]$ の区間について位置 i と (bigram 素性であれば) 隣り合った $\{z_i, z_{i+1}\}$ の値を調べることで知ることが出来る. これら素性毎に (13) 式を合計することで最終的な結果は (14) のような周辺化と等しくなり, (11) 式と同時に Forward-Backward アルゴリズムで全ての必要な周辺確率を得るアルゴリズムを構成できる.

3.4 デコーディング

[5] では,「解が不安定になる」という理由でデコードを行う際に教師なしデータに対する最尤分割を教師データとみなして CRF を再学習する必要があった. 正しいモデル変換によるデコーディングでは, (10) 式で与えた前向き確率を用いて Viterbi アルゴリズムを使用でき, 特に不安定になることもなく満足行く品質の解が得られる.

4 実験

4.1 学習データ

実験は中国語と日本語の二種類の言語で行った.

中国語の教師データおよびテストデータとして SIGHAN Bakeoff 2005 の MSR セットを用い, Chinese Gigaword Corpus から 50000 文ランダムに抽出し教師なしデータとして用いた. SIGHAN Bakeoff のテストスクリプトを使って精度を検証した.

日本語の教師データとして京大コーパスを用い, Twitter のクロールデータ 100 万ツイートと「しょこたんブログ」のクロールデータ 4 万文の二種類を教師なしデータとして用いた. それぞれの教師なしデータの単語分割を観察し, モデルの振る舞いの直感的な評価のために使った.

表 1　SIGHAN Bakeoff MSR での結果

モデル	CRF	NPYLM	NPYCRF
Token F 値	0.962	0.958	0.973
OOV 再現率	0.742	0.316	0.654
IV 再現率	0.964	0.982	0.981

4.2 実験設定

CRF の素性は二言語とも先行研究 [3] と同じものを用い, それに加えて Unicode 文字種の unigram および bigram を素性として加えた.

SIGHAN Bakeoff 2005 MSR の State-of-the-art である [2] においては, CRF の表現力を高めるためにもっともよく使われる {B, I} で文頭とそれ以外を表す二次元ラベリングではなく, {B, E, I} でそれぞれ文頭, 文末, それ以外を表す三次元のラベリングを使って高い精度を達成している. またこれにより素性が高次元になったことで学習が難しくなったことに対処するため, 最小化問題を解く際のアルゴリズムとして一般的によく使われる L-BFGS ではなく, 確率的勾配法の一種である ADF と呼ばれる手法を提案している.

本研究でも CRF の表現力の高さが重要であることと, モデル間重み λ_0 とそれ以外の素性 λ_i との関係が複雑であることなどから, [2] の枠組を利用し, ラベリングの次元数および最適化手法での違いを考慮した実験を行った. しかし CRF 単体では上手く働くこの枠組みは, 本研究で提案したアルゴリズムと三次元のラベリングの組み合わせは上手く働かず, 実験では λ_0 の値が極端に大きくなるなどモデル推定が不安定になった. その原因究明は今後の課題ではあるが, まずは二次元ラベリングと L-BFGS による結果を示すことにする.

また, semi-Markov モデルの状態数を枝刈りするために, 負の二項分布による一般線形モデル (NBGLM) で文の各位置で最大単語長の予測を行い, それを上回る状態については考慮しないことにした.

正則化を行うため, λ_0 の prior として $N(1, \sigma^2)$ なる平均をずらした正規分布を仮定した. 平均が 1 であるのはそれぞれのモデルの「信頼度」が対等である, という仮定を置くことに相当する. 他の素性は一般的な L2 正則化を行うため, 分布は正規分布を仮定した.

4.3 結果

中国語での集計結果を表 1, 具体例を図 2 に示す. また, 日本語での結果として, しょこたんブログの結果を図 3, Twitter での結果を図 4 に示す.

中国語でのテスト結果は, F 値は CRF 単独, および NPYLM 単独よりも向上しており, 両者の長所を組み合わせて更なる精度向上を目指すというモデルの目的は達成できていると言える. ただ, [2] における State-of-the-art の結果には及ばない結果となったため, [2] の手法との統合を上手く行うことで, 今後精度向上を目指したい.

各データセットでの結果を見ると、教師つきデータには含まれておらず, 文字単位の素性での分割は難しいであろう固有名詞を NPYLM が発見し, うまく分

CRF	NPYLM	NPYCRF	正解
有些	有些	有些	有些
大学生	大学生	大学生	大学生
眼	眼高手低	眼	眼高手低
高手		高手	
低		低	
,	,	,	,
不屑	不屑于	不屑于	不屑于
于			
做	做	做	做
小	小	小	小
事情	事情	事情	事情
。	。	。	。
王思斌	王	王思斌	王思斌
	思		
	斌		
,	,	,	,
男	男	男	男
,	,	,	,
１９４９年１０月	１９４９年	１９４９年１０月	１９４９年１０月
	１０月		
生	生	生	生
。	。	。	。

図 2　SIGHAN Bakeoff MSR での単語分割例.

あと 後ろ に いる 霊 が すごく 強い から って 。 親族 で 身近な ひと らしい 。 まさか パパ じゃぁ ない よ なぁ …
五 番 くらい まで つくった ら ファンクラブ イベント で うたい たい ねぇ
だめだ 、 ギザマミタス が ある と 何 回 みて も 見つめて しまう マミタス ギザカワユス ギザカワユス
今日 は モンゴル か ！ やっぱり 翔子 は 3 時 以降 に 脳みそ 覚醒 タイム が くる なあ
wwwww ピンク の スカシカシパン カワユス wwwww ウレシス なぁあ の すぐ 燃え 尽きた くらい 燃え盛った あの 一瞬 あの 瞬間

図 3　しょこたんブログの単語分割例.

今月 は あと 1 g 通信 料 使える から つべ 行って 動画 漁ろう (^o^)
その 発言 が ひど いわ (;o;)(;o;) 笑
眠い けど 風呂 は いる と 目ぇ 覚めん や で
それ が この きのこ に なった 写真 は 風呂 上がり だっ
コメダ なう だ けど 、 やべーぞ これ は 。
エイプリルフール 企画 ある んだろう から さっさと ほかろ …
なんか line の 友達 も 減って る … マヂ むり … リスカ しょ …
爆笑 ！ さっこ 失恋 じゃ ない 失恋 なう やからな ！ わら

図 4　Twitter の単語分割例.

割している様子や，教師なしデータ特有の言い回しを NPYLM が捉えて分割に反映されている様子が分かる．一方で，特にひらがなの連続に対してくだけた言い回しとフォーマルな日本語における助詞助動詞との区別がつかないような局面では，言語モデルを信用しすぎた結果，正しくない分かち書きを選択している場合や，逆に言語モデルを信用しきれず，間違った CRF の解を採用している様子が観察できる．主観的には上手く機能していると考えているが，このように改善した部分と悪化した部分が混在しているため，今後はこれらの日本語 SNS データについて，正解データを作成して Token F 値などの客観的な指標を計算し比較したい．

5　結論

semi-Markov モデルの言語生成モデルと Markov モデルの識別モデルを JESS-CM 法の枠組みで統合するために先行研究の誤りを訂正し，Markov CRF のパラメタを semi-Markov モデルに変換しその上でパラメタ推定をする方法を提案した．提案法を日本語および中国語の分かち書きに適応することで CRF 単体および NPYLM 単体よりも精度が向上すること，さらに半教師あり学習を行うことでさらに高精度なモデルを得られることを示した．

今後の発展としては，モデル間の重みを常に一定ではなく，文脈に応じて変更することで片方のモデルが不得意な部分をもう片方で補う，というようなより高度な統合が考えられる．また今回は単語分かち書きを対象としたが，この枠組みを一般化して品詞推定を含めたものに拡張することなども考えられる．

参考文献

[1] Galen Andrew. A hybrid markov/semi-markov conditional random field for sequence segmentation. In *EMNLP*, pp. 465–472, 2006.

[2] Xu Sun, Wenjie Li, Houfeng Wang, and Qin Lu. Feature-frequency–adaptive on-line training for fast and accurate natural language processing. In *ACL-14: HLT*, pp. 563–586. Association for Computational Linguistics, 2014.

[3] Xu Sun, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, and Jun'ichi Tsujii. A discriminative latent variable chinese segmenter with hybrid word/character information. In *ACL-09: HLT*, pp. 56–64. Association for Computational Linguistics, 2009.

[4] Jun Suzuki and Hideki Isozaki. Semi-supervised sequential labeling and segmentation using gigaword scale unlabeled data. In *ACL-08: HLT*, pp. 665–673. Association for Computational Linguistics, 2008.

[5] 持橋大地, 鈴木潤, 藤野昭典. 条件付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 回年次大会発表論文集, pp. 1071–1074, 2011.

[6] 持橋大地, 山田武士, 上田修功. ベイズ階層言語モデルによる教師なし形態素解析. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2009, No. 36, p. 49, 20090318.