Semi-Markov Conditional Random Fields における 文節情報を用いたラティス構造の制御

福岡 健太 † 浅原 正幸 † 績木 貴史 ‡ 松本 裕治 †

† 奈良先端科学技術大学院大学 情報科学研究科

‡ 松下電器産業株式会社

1 背景

1.1 日本語固有表現抽出と系列ラベリング

固有表現とは，組織の名称や人名 · 地名などの固有名詞や日時 · 時間 · 通貨などの数値表現を指す．固有表現は，次々と新しいものが生まれるため，全てを辞書に登録するのは不可能であり，テキストから自動的に抽出し，それがどの種類の固有表現に相当するのかを分類する技術が多く提案されている．本稿では，機械学習を用いた固有表現抽出を行う．機械学習を用いて固有表現抽出を行う場合，チャンク (かたまり) 同定問題として行われることが多く，系列ラベリング手法が用いられる．

系列ラベリングとは，ある観測系列 (入力系列) $X = \langle x_1, x_2, \ldots, x_n \rangle$ に対して，隠れ変数列 (出力系列) $Y = \langle y_1, y_2, \ldots, y_n \rangle$ を付与する技術を指す．$T = \{t^1, t^2, \ldots, t^m\}$ は隠れ変数 y_i に割り当てられる値の集合 (ラベル集合) とする．ラベルとしてチャンクの開始位置や終了位置を表すチャンクタグを導入することで，系列中に出現するチャンクを抽出することが可能である．固有表現抽出においては，x_i が単語 (文字) であり，X が単語列 (文字列) であることがほとんどである．x_i は単語 (文字) 情報以外に，単語 (文字) の品詞情報や文字種の情報などを持つ場合が多い．また，固有表現の種類とチャンクタグをハイフンで結んだものがラベルとして用いられる．

日本語の固有表現では，最大エントロピー法を用いる手法 [7] や，Support Vector Machines (SVMs) を用いる手法 [12, 10, 6]，などが提案されている．近年提案されている固有表現抽出手法の中に，Linear-Chain Conditional Random Fields (Linear-Chain CRFs) [2, 4] に基づく手法がある．Linear-Chain CRFs は，入力系列に対する正しい出力系列と他の全出力系列とを弁別するような学習をするため，系列全体から最適な出力系列を見つけだせる利点がある．本稿では，Semi-Markov Conditional Random Fields (Semi-Markov CRFs) [5] を用いた日本語固有表現抽出について述べる．Semi-Markov CRFs は，Linear-Chain CRFs を，1 つのノードに対し可変長の範囲の観測トークン列を対応するようにラティスにノードを展開するよう拡張したモデルである．そのため，より豊富な文脈情報を用いることができる．従来，Semi-Markov CRFs では，ある長さ以下の全ての部分観測トークン列に対応するノードをラティスに展開するため，計算量が増大する．この問題を解決するため，個々の文節長以下の部分トークン列のみを展開することでノード数を削減する手法を提案する．

1.2 Semi-Markov CRFs

Linear-Chain CRFs では，入力系列の 1 つの観測値に対し 1 つの隠れ変数を付与することで固有表現の範囲を同定する．一方，Semi-Markov CRFs は複数の観測値を 1 つのセグメントとして同定するモデルである．固有表現抽出においては，1 つの固有表現が 1 つのセグメントに相当する．Semi-Markov CRFs では，1 つの指数分布モデルによりセグメント系列 $S = \langle s_1, s_2, ..., s_p \rangle$ の入力系列 X に対する条件付き確率 $P(S|X)$ を表現する．ここで，$s_j = \langle t_j, u_j, y_j \rangle$ は，開始位置 t_j，終了位置 u_j，ラベル $y_j \in T$ の 3 つ組で表される．各セグメント s_j は，ラベル y_j が $t_j \leq i \leq u_j$ 間の全ての x_i に付与されることと同義とし，また s_j の長さは正とし，t_j と u_j は $1 \leq t_j \leq u_j \leq |S|$，$t_{j+1} = u_j + 1$ を満たす．Semi-Markov CRFs では各セグメント s_j の最大長をあらかじめ与える必要があり，これを L (ノードの最大長) とする．

指数分布モデルの形で表した Semi-Markov CRFs における条件付き確率 $P(S|X)$ を次に示す．

$$
\begin{aligned}
P(S|X) &= \frac{1}{Z(X)} \exp\Big(\sum_j^m \Big(\sum_v \lambda_v f_v(X, s_j) \\
&\quad + \sum_e \mu_e g_e(s_{j-1}, s_j)\Big)\Big) \\
Z(X) &= \sum_{S \in T^m} \exp\Big(\sum_j^m \Big(\sum_v \lambda_v f_v(X, s_j) \\
&\quad + \sum_e \mu_e g_e(s_{j-1}, s_j)\Big)\Big)
\end{aligned}
$$

ここで，$m = |S|$ とし，T^m は入力系列 X に対する可能なセグメント S 全ての集合を表す．f_v は観測素性，g_e は遷移素性をそれぞれ指す関数とし，λ_v は素性関数 f_v に対する重み，μ_e は素性関数 g_e に対する重みとなる．観測素性は，観測系列 X と次状態の s_i のペアで与えられる素性関数，遷移素性は，前状態 s_{i-1} と次状態 s_i のペアで与えられる素性関数と定義する．また，$Z(X)$ は他の全候補を考慮するための正規化項を表し，T^n は入力系列 X に対する可能なラベル系列 Y 全ての集合を表す．観測素性関数 f_v，遷移素性関数 g_e は，引数 s_{j-1}，s_j を展開することにより次のようにも表すことができる．

$$
\begin{aligned}
f_v(X, s_j) &= f'_v(X, y_j, t_j, u_j) \\
g_e(s_{j-1}, s_j) &= g'_e(y_{j-l}, t_{j-1}, u_{j-1}, y_j, t_j, u_j)
\end{aligned}
$$

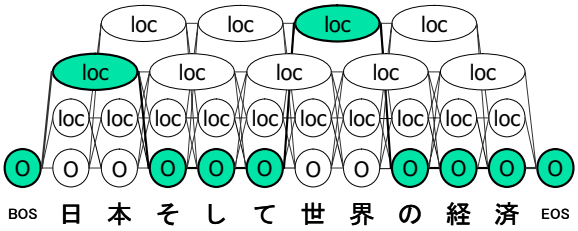

図 1: 通常の Semi-Markov CRFs

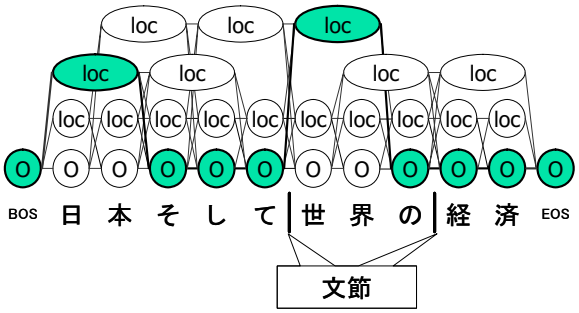

図 2: 文節情報を用いた Semi-Markov CRFs

本稿では，観測素性をノード内素性とノード外素性に分け次のように定義する．

ノード内素性: 推定するノード内に含まれる観測系列を表現するための観測素性

ノード外素性: 推定するノード外に含まれる観測系列を表現するための観測素性

さらに Semi-Markov CRFs では，素性としてセグメント長を考えることが可能である．

λ_v，μ_e の値の推定は，準ニュートン法 (Limited memory BFGS [3]) などのパラメータ推定アルゴリズムを用いて行うことができる．最適系列は，MEMMs や Linear-Chain CRFs 同様 Viterbi アルゴリズムを用いることで効率的に探索することができ，推論も Linear-Chain CRFs 同様 Forward-Backward アルゴリズムを用いることで効率的に行うことができる．

2 文節情報を用いたラティス構造の制御

Semi-Markov CRFs では，ノードの最大長 (1 ノードに対応する観測値の最大数) L を設定する必要がある．固有表現抽出において，ノードの最大長は固有表現の最大トークン (形態素もしくは文字) 長に相当する．日本語固有表現抽出で文字単位の解析を行う場合，固有表現の最大文字長が非常に長く，ノードの最大長が増えることで計算量が膨大になるという問題が生じる．

日本語固有表現はほとんどが名詞から構成され，文節をまたぐ固有表現が少ない [10]．そこで，文節をまたぐノードをラティスに展開しないことでノード数を削減し，計算量を削減する手法を提案する．さらに，これによりノードの最大長も抑えることができる．通常の Semi-Markov CRFs のノードの最大長は，最大で 1 文に含まれる文字 (単語) 数となるが，文節をまたぐノードを展開しないことで最大でも 1 文節内に含まれる文字 (単語) 数に抑えることができる．なお，ここで用いる文節の定義は京都大学コーパス ver 3.0 [9] の基準に準ずる．Semi-Markov CRFs の通常のラティスを図 1 に，文節情報を用いたラティスを図 2 に示す (いずれの場合もノードの最大長を 2 とした)．ここでは，正解ノードを太線の枠で，正解系列を太線でそれぞれ表している．

また，文節をまたぐ固有表現も存在するため，これには各ノードにあらためて BEI のチャンクタグを用いることで対応する．なお，BEI タグの役割は SE 法のタグ付け手法に準ずるものとする．なお，1 つのノードで 1 つの固有表現に相当するものは，チャンクタグ S により表す．

タグ O: 長さ 1 のノードを全ての観測値に対し展開する (通常の Semi-Markov CRFs と同じ)．固有表現以外を表す．

タグ S: 文節内で，Semi-Markov CRFs で設定されるノードの最大長 L 以下の長さのノードを全て展開する．1 つのノードで 1 つの固有表現を表す．

タグ B: 文節をまたぐ固有表現の開始系列を表す．1 文の最後の文節以外の文節内にこのノードを展開する．1 つの文節内に展開するノードは，次の文節に隣接するノードだけでよい．

タグ I: 文節をまたぐ固有表現の先頭と終了以外の系列を表す．1 文の最初と最後の文節以外の文節内にこのノードを展開する．1 つの文節内に展開するノードは，文節長に等しいノードだけでよい．

タグ E: 文節をまたぐ固有表現の終了系列を表す．1 文の最初の文節以外の文節内にこのノードを展開する．1 つの文節内に展開するノードは，前の文節に隣接するノードだけでよい．

「医療と宗教を考える会を発足した」という文を例に，文節情報を用いてラティス構造を制御した例を図 3 に示す．この例では，固有表現 ⟨ ORG ⟩ のみを仮定したラティスを考え,「医療と宗教を考える会」が ⟨ ORG ⟩ であるとし，図中の縦線は文節の境界を表すものとする (この図では，表示スペースの関係でノードのみを表示し，⟨ ORG ⟩ は組織名を表す固有表現とする)．また，ノードの最大長は指定せず文節内に展開可能なノードを全て展開している．

3 評価実験

電子番組表 (EPG) データを対象に評価実験を行う．EPG データ 2 週間分に対し，我々が独自に定義した固有表現タグを人手で付与したものを実験に用いる．定義した固有表現全 9 種類を表 1 に示す．この中で，⟨ LOC ⟩ と ⟨ PRO ⟩ は，固有表現に入れ子構造が多く見られる．入れ子構造とは,「伏見稲荷神社」を例に説明すると,「伏

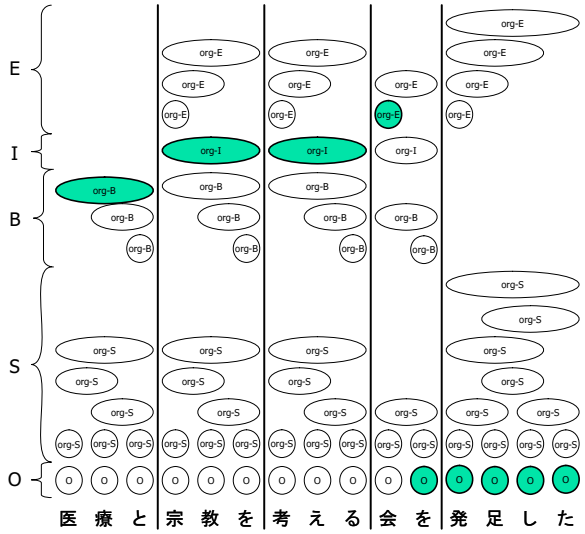

図 3: 文節情報を用いた Semi-Markov CRFs

見稲荷神社」も「伏見」もどちらも ⟨ LOC ⟩ ということになる．このように，ある固有表現何に同じ種類の違う固有表現が含まれている場合，入れ子構造があるという．実験では，固有表現毎に抽出モデルを作成することにする．また，入れ子構造が多い ⟨ LOC ⟩ と ⟨ PRO ⟩ に関しては，固有表現の最も外側を取るモデル ⟨ LOC-L ⟩，⟨ PRO-L ⟩ と，最も内側を取るモデル ⟨ LOC-S ⟩，⟨ PRO-S ⟩ を考える．なお，入れ子構造になっていない ⟨ LOC ⟩ と ⟨ PRO ⟩ に関しては，モデル ⟨ LOC-S ⟩，⟨ PRO-S ⟩ で抽出することにする．⟨ LOC ⟩ と ⟨ PRO ⟩ 以外の種類の固有表現に関しては，入れ子構造となっている場合は最も外側の固有表現だけを抽出対象とし，抽出対象のみを評価対象とした．評価は EPG データ 1 週間分を訓練データとし，別の 1 週間分の 3 日分を開発用データ，残りの 4 日分をテストデータとする．評価尺度としては，F1 ($\beta = 1$) を用いモデルの比較を行った．実験では，Linear-Chain CRFs と Semi-Markov CRFs の精度を比較する．また，文字単位の解析と形態素単位の解析での精度比較も行う．Linear-Chain CRFs で推定するタグは，IOB2 手法[1]を用いて付与した．Gaussian Prior [1] を用いてパラメータの正規化を行い，最適化手法には準ニュートン法を用いた．観測素性は出現回数 10 回以上のもののみを用いた．Semi-Markov CRFs で用いる文節情報は CaboCha[8] を用いて付与し，ノードの最大長 L は訓練データ中に含まれる各固有表現以上の数に設定した．

表 3 に実験結果を示す．太文字の数字は最も精度の高いものを表している．各モデルの素性は次の通りである．lienar 1 及び semi 1，semi 2，semi 3 は形態素単位の解析を行うモデルとし，それ以外のモデルは文字単位の解析を行うものとした．linear 1，linear A は Linear-Chain CRFs を用いるモデルであり，semi 1,，semi 2，semi A は文節情報を用いた Semi-Markov CRFs を用いるモデルである．また，semi 3 は文節情報を用いてノードを削減していない Sarawagi らの Semi-Markov CRFs [5] を用いるモデルである．

[1]タグ I，O，B を用いる．B は固有表現の開始位置，I は固有表現の開始位置以外，O は固有表現以外を表す．

表 1: EPG データで定義されている固有表現の種類及び例

固有表現の種類		例
ADJ	形容表現	スペシャル，ノスタルジック
GEN	ジャンル	ニュース，ドキュメンタリー
JOB	職業	アナウンサー，監督
LOC	国名 · 場所	アメリカ，伏見稲荷神社
ORG	組織名	新選組，石原軍団
PER	人名	小泉純一郎，KONISHIKI
PRO	番組名	とくダネ!，ワイド!スクランブル
SPO	スポーツ名	ゴルフ，テニス男子シングルス
TEA	チーム名	阪神，デトロイト · ピストンズ

linear 1: 観測素性として推定する位置とその前後各 2 形態素 · 品詞．遷移素性として前状態のラベル．

semi 1: ノード内素性として先頭と末尾から各 2 形態素 · 品詞，ノード外素性として，前後各 2 形態素 · 品詞．遷移素性として次状態ののラベル · 長さ，前状態のラベル · 長さ．観測素性は全て長さも考慮．

semi 2: semi 1 の素性で，観測素性は長さを考慮しない．また，遷移素性も長さを考慮しない．

semi 3: semi 2 と同じ素性で，文節情報を用いてノードを削減しない (Sarawagi ら [5] と同じ設定)．

linear A: 観測素性として推定する位置とその前後各 2 文字 · 品詞 · 形態素．遷移素性として前状態のラベル．

semi A : ノード内素性として先頭と末尾から各 2 文字 · 品詞 · 形態素，ノード外素性として．前後各 2 文字 · 品詞 · 形態素，遷移素性として次状態の長さ，前状態のラベル長さ．観測素性は長さを考慮しない．

全体的に形態素単位で解析を行うより文字単位で解析を行う方が精度が良いことがわかる．EPG データでは未知語となる語が多く，未知語の間に固有表現の境界が存在することが多いということが 1 つの原因として挙げられる．⟨ PRO-L ⟩，⟨ PRO-S ⟩ といった固有表現長の平均長が長いものは形態素単位で解析を行う方が良い精度を示している．Linear-Chain CRFs と Semi-Markov CRFs を比較すると，Semi-Markov CRFs の方が全体的に精度がよいことがわかる．Semi-Markov CRFs においては，観測素性と遷移素性にセグメント長を考慮したもの (semi 1) より考慮しないもの (semi 2) の方が全体の精度は高かった．また同じ素性を用いた実験で，文節情報を用いてラティス構造を制御した手法 (semi 2) の方が，制御しない Sarawagi らの手法 (semi 3) に比べて固有表現の平均長が長い ⟨ PRO-L ⟩，⟨ PRO-S ⟩ の精度は高かった．しかし，他のほとんどの固有表現においては提案手法の精度の方が優れており，全体の精度も Sarawagi らの手法より優れていた．⟨ LOC ⟩ と ⟨ PRO ⟩ 以外の固有表現は入れ子構造になっている場合，最も外側の固有表現を抽出対象としている．このため，例えば ⟨ JOB ⟩ で，「日本画家」が正解であるにも関わらず「画家」と抽出してしまっているような例が多く見られた．このような

表 2: 固有表現の頻度と平均文字長

		訓練	開発	評価
ADJ	頻度	4981	2260	287
	平均長	2.76	2.74	2.71
GEN	頻度	8637	3766	4896
	平均長	3.28	3.30	3.22
JOB	頻度	7363	3072	4375
	平均長	3.14	3.16	3.18
LOC-L	頻度	647	246	491
	平均長	8.02	9.16	8.76
LOC-S	頻度	5937	2524	3780
	平均長	3.10	3.27	3.24
ORG	頻度	3334	1358	2005
	平均長	5.67	5.50	5.53
PER	頻度	13889	5850	8107
	平均長	4.94	5.00	4.95
PRO-L	頻度	2493	1123	1460
	平均長	20.18	20.51	21.09
PRO-S	頻度	7153	3150	4102
	平均長	7.99	8.01	7.97
SPO	頻度	685	172	297
	平均長	4.54	4.16	4.29
TEA	頻度	238	65	148
	平均長	4.62	5.54	5.13
TOTAL	頻度	55357	32537	23586
	平均長	5.20	5.29	5.25

表 3: 実験結果

解析単位		形態素				文字	
モデル		linear 1	semi 1	semi 2	semi 3	linear A	semi A
ADJ	開発	92.15	92.86	**92.93**	92.81	96.10	96.09
	評価	91.32	92.41	92.49	92.43	95.72	**95.77**
GEN	開発	93.07	93.84	93.96	93.89	97.76	**97.85**
	評価	90.92	92.08	92.33	92.32	96.49	**96.51**
JOB	開発	94.42	95.26	95.64	95.56	**96.54**	96.53
	評価	91.78	92.40	92.82	92.76	95.41	**95.69**
LOC-L	開発	73.79	63.46	71.26	74.09	**75.47**	74.77
	評価	63.66	60.00	64.15	**65.38**	63.72	63.34
LOC-S	開発	88.28	88.87	88.70	87.42	93.51	**93.65**
	評価	84.29	85.00	84.58	83.28	90.28	**90.27**
ORG	開発	88.33	90.04	89.13	89.15	**90.14**	89.86
	評価	84.70	86.04	85.81	85.48	86.83	**86.93**
PER	開発	94.58	95.00	95.09	94.94	95.81	**96.13**
	評価	92.73	93.52	93.28	93.30	94.33	**94.38**
PRO-L	開発	94.24	93.52	93.47	**94.46**	93.12	93.68
	評価	91.57	88.91	91.44	**91.84**	89.36	90.99
PRO-S	開発	91.38	91.80	**91.84**	91.64	91.25	91.64
	評価	87.72	88.40	**88.43**	88.09	86.97	87.80
SPO	開発	70.68	74.64	74.64	74.64	80.14	**81.63**
	評価	79.11	82.14	83.43	82.97	87.10	**87.52**
TEA	開発	77.36	79.63	79.63	78.50	78.50	**88.14**
	評価	78.69	79.67	**80.16**	78.19	73.03	77.24
TOTAL	開発	92.26	92.77	92.88	92.69	94.62	**94.82**
	評価	89.57	90.32	90.39	90.13	92.36	**92.62**

間違いがかなり多くみられたため，入れ子構造の固有表現をうまく抽出できるモデルを考える必要がある．

4 まとめ

本稿では，Semi-Markov CRFs を用いた日本語固有表現抽出手法に対し，文節情報を用いることでラティス構造を制御し，計算量を削減する手法を提案した．実験では，提案手法が精度の観点で，Linear-Chain CRFs より優れていることを示した．また形態素単位の解析において，平均長が極端に長くない固有表現以外では従来の Semi-Markov CRFs より精度が良いことを示した．今後の課題として，セグメント単位の素性を用いること，セグメント長の素性を単独で用いることなどが考えられる．

謝辞

本研究は，奈良先端科学技術大学院大学と松下電器産業株式会社との共同研究により行われた．ここに記して謝意を表します．

参考文献

[1] S. Chen and R. Rosenfeld. A Gaussian Prior for Smoothing Maximum Entropy Models. 1999.

[2] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML-2001*, pp. 282–289, 2001.

[3] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. Mathematical Programming, 45. pp. 503–528, 1989.

[4] A. McCallum and W. Li. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In *CoNLL-2003*, pp. 188–191, 2003.

[5] Sunita Sarawagi and William W. Cohen. Semi-Markov Conditional Random Fields for Information Extraction. In *NIPS-2004*, 2004.

[6] 浅原正幸, 松本裕治. 日本語固有表現抽出におけるわかち書き問題の解決. 情報処理学会論文誌 Vol.45 No.5, pp. 1442–1450, 2004.

[7] 内元清貴, 馬青, 村田真樹, 小作浩美, 内山将夫, 井佐原均. 最大エントロピーモデルと書き換え規則に基づく固有表現抽出. 自然言語処理 Vol.7 No.2, pp. 63–90, 2000.

[8] 工藤拓, 松本裕治. Support Vector Machine による日本語係り受け解析. 情報処理学会研究報告 (自然言語処理研究会), 2000-NL-138, pp. 79–86, 2000.

[9] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会 第 3 回年次大会 発表論文集, pp. 115–118, 1997.

[10] 中野桂吾, 平井有三. 日本語固有表現抽出における文節情報の利用. 情報処理学会論文誌 Vol.45 No.3, pp. 934–941, 2004.

[11] 福岡健太. Semi-Markov Conditional Random Fields を用いた固有表現抽出に関する研究. Master's thesis, 奈良先端科学技術大学院大学 修士論文, 2006.

[12] 山田寛康, 工藤拓, 松本裕治. Support Vector Machine を用いた日本語固有表現抽出. 情報処理学会論文誌 Vol.43 No.1, pp. 44–53, 2002.