自然言語処理と系列ラベリング技術

浅原　正幸

自然言語処理の分野は系列に対するラベル付与（系列ラベリング）問題として解かれるタスクが多くある．例えば，品詞ラベル付け問題は，入力を単語列とし，各単語に品詞を付与する系列ラベリング問題の1つである．このような背景から，教師あり学習による系列ラベリング技術が多く提案されてきた．本稿では，自然言語処理の分野でどのように系列ラベリング技術が利用されているかを概観するとともに，近年考案された系列全体において最適化を行う構造マッピング法に基づく系列ラベリング手法を紹介する．

キーワード：マルコフ過程，教師あり学習，系列ラベリング問題，構造マッピング

1．はじめに

系列ラベリング（sequential labeling）とは，ある観測された系列（sequence）$X=X_1, X_2, ..., X_n$ に対し，隠れ変数（hidden variable）列 $Y=Y_1, Y_2, ..., Y_n$ を付与する技術の総称である．確率モデルで解く場合，訓練時には $P(Y|X)$ をモデル化し，テスト時には新たに入力される X に対して $Y=arg\max_Y P(Y|X)$ を解くことによってラベル列を推定する．1990年代 $P(Y|X)$ のモデル化には，生成モデルである隠れマルコフモデルが用いられていた[10]が，2000年代に入り様々な識別モデルによる手法が提案されてきた．自然言語処理の分野では X_i が単語であり，X が単語列であることが多い．本稿では，自然言語処理の系列ラベリングの利用方法をいくつか紹介するとともに，近年提案されてきた識別モデルを用いた系列ラベリング手法を紹介する．

2．系列ラベリングと自然言語処理

系列ラベリング問題として解かれる代表的な問題として，品詞ラベリングがある．観測系列 X を単語列，ラベル系列 Y を品詞列とすることによって形式化することができる．また，系列中の連続部分系列にラベルを付与するようなタスクとして，基本名詞句（Base NP；Base Noun Phrase）同定や固有表現抽出がある．このような場合付与したいラベルとチャンクラベル（chunk label；範囲指定するようなラベル/Bを範囲の開始位置，Iを範囲の内側，Oを範囲の外側，Sを単独で範囲となる位置）を導入することにより，系列中の部分系列を同定することができる．基本名詞句同定の場合には名詞句を表す“NP”とチャンクラベルの対によるラベルを，固有表現抽出の場合には固有表現の種類を表すラベル（例えば，人名（PERSON）の場合は“PER”，場所名（LOCATION）の場合は“LOC”）とチャンクラベルの対によるラベルを付与する（図1：左）．

系列ラベリング問題は図2のようなトレリス状の図で説明することができる．“BOS”は系列の開始位置（Begining of Sequence），“EOS”は系列の終了位置（End of Sequence）を示す特別な記号である．図2では基本名詞句同定で用いられる4つのラベル{“$NP-B$”, “$NP-I$”, “$NP-S$”, “O”}を，系列上の対応する観測要素上に配置する．“BOS”から“EOS”への1つのパスが，観測列に対する1つのラベル割り当てを

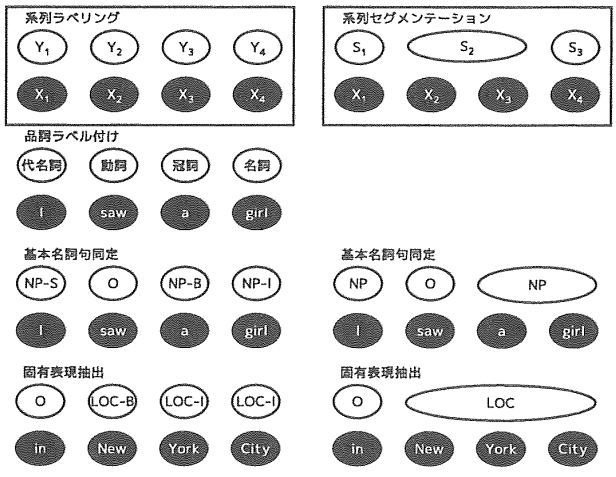

図1　自然言語処理の諸問題の定式化

あさはら　まさゆき
奈良先端科学技術大学院大学
〒630-0101 生駒市高山町8916-5

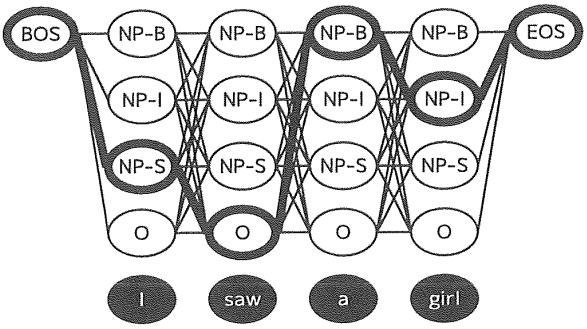

図2　系列ラベリングのトレリス

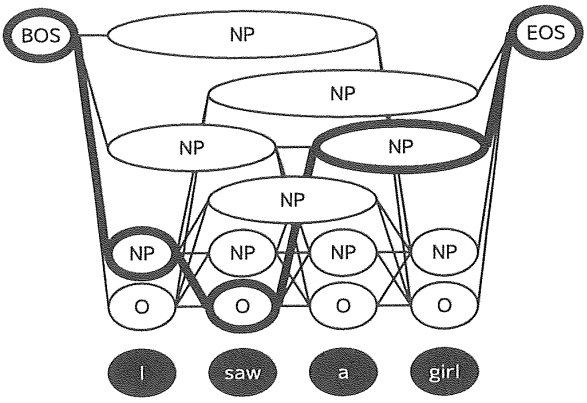

図3　系列セグメンテーションのトレリス

表している．ノードの組み合わせによりすべての可能なラベル列を考えることができる．このすべての可能なラベル列の中で，太線で囲まれたノードが正解ラベル列を表している．系列ラベリング手法の多くが，正解ラベル列とそれ以外のすべての可能なラベル列とを弁別するように学習を行うことによる．

系列ラベリング問題の拡張として，系列セグメンテーション問題がある．系列セグメンテーション問題は観測系列 X に対して，X の部分系列に対するラベル割り当て S_i の列（セグメント列）S を推定する問題である（図1：右）．本来基本名詞句同定や固有表現抽出は系列セグメンテーション問題として解くべき問題である．図3に系列セグメンテーションのトレリスを示す．系列セグメンテーション問題では観測系列上の準マルコフ過程を考えることによりモデル化を行う．解析時にセグメント全体に対する特徴を用いることができるため性能が良いが，空間計算量が大きくなるという問題がある．

日本語や中国語の形態素解析（単語わかち書き）は系列セグメンテーション問題とみなすことができる．文字列を観測系列として，あらかじめ用意した辞書に基づき，辞書に登録されている文字列と一致するすべての部分系列をトレリス上に展開する（図4）．

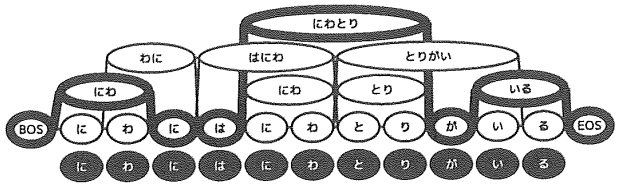

図4　日本語形態素解析のトレリス

3.　識別モデルによる系列ラベリング

系列ラベリングは前節で提示したトレリス上で，正解ラベル列とそれ以外のすべての可能なラベル列とを弁別するように学習を行う．本節では指数分布モデルを用いた手法である条件付確率場と，カーネル法を用いた手法である隠れマルコフパーセプトロンと隠れマルコフサポートベクトルマシンを紹介する．トレリスの形は異なるが系列セグメンテーション問題は，系列ラベリング問題と同様の手法でモデル化できる[11][7]．

3.1　条件付確率場

系列ラベリングの1解法として，Lafferty らによって提案された条件付確率場（CRF：Conditional Random Fields）[8]がある．条件付確率場は1つの指数分布モデルによって各出力系列 Y の入力文 X に対する条件付確率 $P(Y|X)$ を表現する．

$$P(Y|X)=\frac{1}{Z(X)}\exp\left(\sum_{i}^{n}\sum_{a}\lambda_a\phi_a(X, Y_i)\right)$$

$$Z(X)=\sum_{Y\in T^n}\exp\left(\sum_{i}^{n}\sum_{a}\lambda_a\phi_a(X, Y_i)\right)$$

ここで，ϕ_a は，観測系列と系列上のある位置のラベル割り当てによって決定される特徴量関数，$T=\{t^1, t^2, \ldots, t^m\}$ は，Y_i に割り当てられるラベルの集合を表し，T^n は長さ n の可能なすべてのラベルの組み合わせを表す．$\lambda_a(\in\Lambda=\{\lambda_1, \ldots, \lambda_A\}\in\mathcal{R}^A)$ は特徴量関数 ϕ_a に対する重みであり，正しいラベル系列を他の全出力候補と弁別するように選択される．$Z(X)$ は，全出力候補を考慮するための正規化項である．

二値の出力をもつ特徴量関数 $\phi_a(X, Y_i)$ を定義する（便宜的に $a=\langle b_i, t^j\rangle$ とし，b_i を観測系列 X の i 番目の要素 X_i に対する二値の出力をもつ素性関数，t^j をあるラベルとする）：

$$\phi_a(X, Y_i)=\phi_{\langle b_i, t^j\rangle}(X, Y_i)$$
$$=\begin{cases}1 & \text{if } b_i(X) \text{ が真でかつ } Y_i=t^j\\ 0 & \text{otherwise}\end{cases}$$

$$E_{P(Y|X)}[\Phi_a(Y,X)] = \sum_{t^k \in T, t^j \in T, 0 \le i \le n} \frac{\alpha_{t^k}(i) \cdot \phi_a(X_{i+1}, Y_{i+1} = t^j) \cdot V(t^k, t^j) \cdot \beta_{t^j}(i+1)}{Z(X)}$$

ただし $V(t^k, t^j)$ は，スペースの関係上便宜的に入れた関数：

$$V(t^k, t^j) = \exp((\sum_a \lambda_a \phi_a(X_{i+1}, Y_{i+1} = t^j)))$$

ここで，α と β は forward-backward コストで以下のように再帰的に定義される：

$$\alpha_{BOS}(0) = 1$$

$$\alpha_{t^j}(i+1) = \sum_{t^k \in T} \alpha_{t^k}(i) \cdot \exp((\sum_a \lambda_a \phi_a(X_{i+1}, Y_i = t^k, Y_{i+1} = t^j)))$$

$$\beta_{EOS}(n+1) = 1$$

$$\beta_{t^k}(i) = \sum_{t^j \in T} \exp((\sum_a \lambda_a \phi_a(X_{i+1}, Y_i = t^k, Y_{i+1} = t^j))) \cdot \beta_{t^j}(i+1)$$

図5　特徴量の出現期待値と正規化項の導出

観測に対する素性関数 $b(X_i)$ の例として，「X_i が大文字で始まる」，「X_i が *ing* で終わる」などが考えられる．

記号の簡単化のために，大域特徴量ベクトル $\Phi(Y, X) = \{\Phi_1(Y, X), ..., \Phi_A(Y, X)\}$ を与える．$\Phi_A = \sum_{i=1}^{n} \phi_a(X_i, Y_i)$ とする．大域素性を用いると，$P(Y|X)$ は以下のように書ける：

$$P(Y|X) = \frac{1}{Z(x)} \exp(\Lambda \cdot \Phi(Y, X))$$

ここで，$\Lambda \cdot \Phi(Y, X)$ を系列 Y のコストと呼ぶ．入力 X に対するコストを最大化するラベル列 $\hat{Y}$ は，

$$\hat{Y} = arg \max_Y P(Y|X) = arg \max_Y \Lambda \cdot \Phi(Y, X)$$

となる．この最適ラベル列は隠れマルコフモデルと同様に Viterbi アルゴリズムを用いて効率良く探索できる．

一般的な最尤推定を用いてパラメータを選択する．あらかじめラベル付けされた訓練データ $D = \{\langle X^u, Y^u \rangle\}_{u=1}^{|D|}$ に対する対数尤度 $\mathcal{L}_\Lambda$ の最大化を行う．

$$\mathcal{L}_\Lambda = \sum_u \log(P(Y^u|X^u))$$

$$= \sum_u [\log(\sum_Y \exp(\Lambda \cdot [\Phi(Y^u, X^u) - \Phi(Y, X^u)]))]$$

$$\hat{\Lambda} = arg \max_{\Lambda \in \mathcal{R}^A} \mathcal{L}_\Lambda$$

対数尤度 $\mathcal{L}_\Lambda$ を大きくするためには，各訓練データ $\langle X^u, Y^u \rangle$ に対し，$\sum_Y \exp(\Lambda \cdot [\Phi(Y^u, X^u) - F(Y, X^u)])$ を大きくすればよい．これは，正解のパスコスト $\Lambda \cdot \Phi(Y^u, X^u)$ と，残りの全候補のコスト和 $\sum_{Y \neq Y^u} \Lambda \cdot \Phi(Y, X^u)$ との差をできるだけ大きくすることに相当する．これにより，ラティス中の全候補系列から正解の系列のみを弁別するような効果が生まれる．

目的関数の凸性から，最適点では以下が成立する：

$$\frac{\delta \mathcal{L}_\Lambda}{\delta \lambda_a} = \sum_u (\Phi_a(Y^u, X^u) - E_{P(Y|X^u)}[\Phi_a(Y, X^u)])$$

$$= O_a - E_a = 0$$

ただし，$O_a = \sum_u \Phi_a(Y^u, X^u)$ は特徴量 a の学習データ D の出現頻度，$E_a = \sum_u E_{P(Y|X^u)}[\Phi_a(Y, X^u)]$ は特徴量 a のモデル分布の出現期待値である．この期待値の計算は，単純には出力系列の全候補を陽に枚挙することにより実現できるが，この候補数は入力文の長さに対して指数的に増えるために，事実上困難である．しかし，図5のように隠れマルコフモデルで用いられる forward-backward アルゴリズムの変種を用いることにより効率良く計算することができる．以上を用いると，正規化項は $Z(X) = \alpha_{EOS}(n+1) = \beta_{BOS}(0)$ となる．これらの値を用いて，目的関数に対する最適なパラメータを準ニュートン法などにより見つける．

3.2　隠れマルコフパーセプトロンと隠れマルコフサポートベクトルマシン

条件付確率場はロジスティック回帰により，最尤ラベル列を推定するモデルであった．他の手法として，カーネル法に基づく系列ラベリング手法，隠れマルコフパーセプトロン（HM-Perceptron：Hidden Markov Perceptron）と隠れマルコフサポートベクトルマシン（HMM-SVM：Hidden Markov Model-Support Vector Machines）を紹介する[1][13]．HM-Perceptron および HMM-SVM では，入力である観

測系列 $X\in\mathcal{X}$ と出力であるラベル列 $Y\in\mathcal{Y}$ を定義域として実数値を返す，重み w でパラメータ化された識別関数 $F: \mathcal{X}\times\mathcal{Y}\to R$ を考える．入力と出力の組に対して最大化するようにしてラベル列を推定する：

$$f(x)=arg\max_{y\in\mathcal{Y}} F(x, y\,;\, w)$$

F を特徴量関数 Φ との線形結合する場合を考える：

$$F(x, y\,;\, w)=\langle w, \Phi(x, y)\rangle$$

入力と可能な全出力に対して Φ を明示的に写像することを避けるために，入出力空間に対するカーネル関数 K を定義する：

$$K((x, y), (\bar{x}, \bar{y}))=\langle\Phi(x, y), \Phi(\bar{x}, \bar{y})\rangle$$

m 個の訓練データ $D=\{\langle X^u, Y^u\rangle\}_{u=1}^{|D|}$ が与えられた際の式は $\sum_{u=1}^{|D|}\alpha K((X^u, Y^u), (X, Y))$ という双対形に展開することができる．

3.2.1 隠れマルコフパーセプトロンの訓練

CRF では期待尤度最大化により訓練が行われた．隠れマルコフパーセプトロンでは，正しいラベル割り当て y^u に対する評価関数 $F(x^u, y^u)$ の値が，他の誤ったラベル割り当て $\bar{y}\neq y^u$ に対する評価関数 $F(x^u, \bar{y})$ の値より大きくなるように逐次更新することにより訓練が行われる．

観測系列 x^u に対する予測 $\tilde{y}^u$ を考える．

$$\tilde{y}^u=arg\max_{y\in\mathcal{Y}}\sum_{u=1}^{|D|}\sum_{\tilde{y}\in\mathcal{Y}}\alpha_u(\tilde{y})\langle\Phi(x^u, \tilde{y}), \Phi(x^u, y)\rangle$$

$$=arg\max_{y\in\mathcal{Y}}\sum_{u=1}^{|D|}\sum_{\tilde{y}\in\mathcal{Y}}\alpha_u(\tilde{y})K((x^u, \tilde{y}), (x^u, y))$$

この予測 $\tilde{y}^{(i)}$ が正しいラベル割り当て $y^{(i)}$ と異なる場合に以下のようにパラメータ α を更新する：

$$\alpha_i(y^i)\leftarrow\alpha_i(y^i)+1$$

$$\alpha_i(\tilde{y}^i)\leftarrow\alpha_i(\tilde{y}^i)-1$$

3.2.2 隠れマルコフサポートベクトルマシンの訓練

隠れマルコフサポートベクトルマシンでは，正しいラベル割り当て y^u に対する評価関数 $F(x^u, y^u)$ の値と，他の誤ったラベル割り当て $\bar{y}\neq y^u$ に対する評価関数 $F(x^u, \bar{y})$ の値との差（マージン）γ_u を考える：

$$\gamma_u=F(X^u, Y^u)-\max_{Y\neq Y^u} F(X^u, Y)$$

マージン最大化と呼ばれるこの手法では，マージン $\min_u\gamma_u$ を最大化する w を発見することにより訓練が行われる．

目的関数は

$$\min\frac{1}{2}\|w\|^2$$

$$\text{s.t.}\quad F(X^u, Y^u)-\max_{Y\neq Y^u} F(X^u, Y)\geq 1, \forall i$$

となる．この式の制約は非線形なものである．最適化の手法を導入するために同値な線形制約に置き換える：

$$F(X^u, Y^u)-F(X^u, Y)\geq 1, \forall i \text{ and } \forall Y\neq Y^u$$

新たにパラメータ θ_u を各事例に導入することにより以下のように書き換えられる：

$$z_u(Y)(F(X^u, Y)+\theta_u)\geq\frac{1}{2}$$

$$z_u(Y)=\begin{cases} 1 \text{ if } Y=Y^u \\ -1 \text{ if } Y\neq Y^u\end{cases}$$

この二次計画問題の双対問題をラグランジュ法を用いて導出すると以下のようになる．

$$\max W(\alpha)$$

$$=-\frac{1}{2}\sum_{u, Y}\sum_{v, \bar{Y}}\alpha_u(Y)\alpha_v(\bar{Y})z_u(Y)z_v(\bar{Y})k_{u, v}(Y, \bar{Y})+\sum_{u, Y}\alpha_u(Y)$$

$$\text{s.t. } \alpha(Y)\geq 0, \forall i=1, \ldots, n, \forall Y\in\mathcal{Y}$$

$$\sum_{Y\in\mathcal{Y}} z_i(Y)\alpha_i(Y)=0, \forall i=1, \ldots, n$$

ここで $k_{u, v}(Y, \bar{Y})=\langle\Phi(X^u), Y), \Phi(X^{(v)}, \bar{Y})\rangle$ である．この最適なパラメータを得るためには，通常のサポートベクトルマシンと同様に二次計画法が用いられる．最適化法の詳細はサポートベクトルマシンの教科書[3]を参照してほしい．

4. 関連情報

自然言語処理の分野では前節に示したモデル作成を行う学習パッケージが数多く公開されている（表1）．それぞれ実装されている言語や利用できるプラットホームが異なるため，自分の環境に合ったパッケージを選んで利用するとよい．

このように多くの学習パッケージが公開されている背景には，評価型ワークショップの影響が考えられる．評価型ワークショップでは，期間を決めて複数の参加者が様々な手法を同じデータを用いて評価をする．データが共通化されているために，研究者は機械学習モデルの改善や，与える特徴量関数の設計などに集中して開発することができる．表2に系列ラベリング関連の評価型ワークショップを示す．

関連する技術に関するチュートリアルについて示す．自然言語処理における最も大きな国際会議 ACL（the Association for Computational Linguistics）では，2003年に条件付確率場に関連するチュートリアル[5][1]が，2005年にマージン最大化による学習器に関連するチュートリアル[6][2]が開催された．また，日

表1　公開されている学習パッケージ

条件付確率場の学習パッケージ	
CRF++	http://crfpp.sourceforge.net/
FlexCRFs	http://flexcrfs.sourceforge.net/
MALLET	http://mallet.cs.umass.edu/
MinorThird	http://minorthird.sourceforge.net/
隠れマルコフサポートベクトルマシンの学習パッケージ	
SVM^{HMM}	http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html

表2　評価型ワークショップ

会議名	タスク	URL
SIGHAN Bakeoff	中国語わかち書きほか	http://www.sighan.org/
CoNLL-2003	固有表現抽出	http://www.cnts.ua.ac.be/conll2003/ner/
CoNLL-2002	固有表現抽出	http://www.cnts.ua.ac.be/conll2002/ner/
CoNLL-2001	節の同定	http://www.cnts.ua.ac.be/conll2001/clauses/
CoNLL-2000	句の同定	http://www.cnts.ua.ac.be/conll2000/chunking/
IREX-NE	固有表現抽出	http://nlp.cs.nyu.edu/irex/NE/
CoNLL-1999	名詞句の同定	http://www.cnts.ua.ac.be/conll99/npb/
MUC-6	固有表現抽出ほか	http://cs.nyu.edu/cs/faculty/grishman/muc6.html

本でも2006年の言語処理学会年次大会において，条件付確率場に関連するチュートリアル[14][3]が開催された．

5. おわりに

本稿では自然言語処理における系列ラベリングの利用と，系列ラベリングに関する様々な手法を概観した．この系列ラベリング技術自体は，バイオインフォマティクスをはじめとして，他分野にも適用できる一般的なものであり，実際利用されている．

最後に最近の系列ラベリング技術に関する話題を二つ挙げる．一つ目は，系列ではなく，木構造やグラフ構造に対するラベル付け手法に関する研究である．木構造[2]や有向非循環グラフに関しては，効率的に周辺確率などを計算する手法が提案されている．そうでない一般のグラフの場合には近似手法を用いてモデルを作成する手法が提案されている[12]．

二つ目は系列ラベリング学習器の半教師あり学習手法に関する研究である．実際に言語解析器を作成する際，初めから大規模なラベル付きデータを準備することは困難であり，小規模なラベル付きデータと大規模なラベルなしデータが存在する場合が多い．このような場合に，系列ラベリングモデルを作成する手法が提案されており[4]，高速化もなされている[9]．

[1] http://www.cs.berkeley.edu/~klein/papers/maxent-tutorial-slides.pdf

[2] http://www.cs.berkeley.edu/~klein/papers/max-margin-tutorial.pdf

[3] http://www.geocities.co.jp/Technopolis/5893/publication/NLP2006slide.pdf

参考文献

[1] Y. Altun, I. Tsochantaridis and T. Hofmann. Hidden Markov Support Vector Machines. In *Proc. of the Twentieth International Conference on Machine Learning* (*ICML-2003*), 2003.

[2] T. Cohn and P. Blunson. Semantic Role Labelling with Tree Conditional Random Fields. In *Proc. of the 9 th Conference on Computational Natural Language Learning* (*CoNLL-2005*), 2005.

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[4] F. Jiao, S. Wang, C.-H. Lee, R. Greiner and D. Schuurmans. Semi-Supervised Conditional Random Fields for Improved Sequence Segmentation and Labeling. In *Proc. of the 21 st International Conference on Computational Linguistics and 44 th Annual Meeting of the Association for Computational Linguistics* (*COLING-ACL-2006*), 2006.

[5] D. Klein and C. D. Manning. Maxent Models, Conditional Estimation, and Optimization, without the

Magic. In *Proc. of 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), Tutorial*, 2003.

[6] D. Klein and B. Taskar. SVM's and Structured Max-Margin Methods. In *Proc. of 43 rd Annual Meeting of the Association for Computational Linguistics (ACL-2005), Tutorial*, 2005.

[7] T. Kudo, K. Yamamoto and Y. Matsumoto. Appliying Conditional Random Fields to Japanese Morphological Analysis. In *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, 2004.

[8] J. Lafferty, A. McCallum and F. Pereira. Conditional Random Fields: Probabilisitic Models for Segmenting and Labeling Sequence Data. In *Proc. of the Eighteenth International Conference on Machine Learning (ICML-2001)*, 2001.

[9] G. S. Mann and A. McCallum. Efficient computation of entropy gradient for semi-supervised conditional random fields. In *Proc. of Joint Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-2007)*, 2007.

[10] C. D. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*, chapter 9. Markov Models. The MIT Press, 1999.

[11] S. Sarawagi and W. W. Cohen. Semi-Markov Conditional Random Fields for Information Extraction. In *Proc. of Eighteenth Annual Conference on Neural Information Processing Systems (NIPS-2004)*, 2004.

[12] C. Sutton, K. Rohanimanesh and A. McCallum. Dynamic Conditional Random Fields: Factorized Probabilisitic Models for Labeling and Segmenting Sequence. In *Proc. of the Twenty-First International Conference on Machine Learning (ICML-2004)*, 2004.

[13] B. Taskar, C. Guestrin and D. Koller. Max-Margin Markov Networks. In *Proc. of Seventeenth Annual Conference on Neural Information Processing Systems (NIPS-2003)*, 2003.

[14] 坪井, 鹿島, 工藤. 言語処理における識別モデルの発展—HMM から CRF まで—. 言語処理学会第 12 回年次大会, チュートリアル, 2006.