Semi-Markov Conditional Random Fields のための 損失関数スムージング

Smoothing in Semi-Markov Conditional Random Fields

福岡 健太
Kenta Fukuoka

奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Japan
kenta.fukuoka@gmail.com

浅原 正幸
Masayuki Asahara

(同 上)
masayu-a@is.naist.jp

松本 裕治
Yuji Matsumoto

(同 上)
matsu@is.naist.jp, http://cl.naist.jp/

keywords: sequential labeling, conditional random fields, loss functions.

Summary

Linear-chain conditional random fields are a state-of-the-art machine learner for sequential labeling tasks. Altun investigated various loss functions for linear-chain conditional random fields. Tsuboi introduced smoothing method between point-wise loss function and sequential loss function. Sarawagi proposed semi-markov conditional random fields in which variable length of observed tokens are regarded as one node in lattice function. We propose a smoothing method among several loss functions for semi-markov conditional random fields. We draw a comparison among the loss functions and smoothing rate settings in base phrase chunking and named entity recognition tasks.

1. は じ め に

1.1 背 景

自然言語処理や情報抽出の分野では、与えられた系列にラベルを付与する問題（系列ラベリング問題）や与えられた系列を分割する問題（系列セグメンテーション問題）として定式化できる問題が多い。Linear-Chain Conditional Random Fields (Linear-Chain CRFs) [Lafferty 01] は Lafferty らによって開発された、系列ラベリング問題を解くための機械学習器である。従来のヒストリに基づく学習モデルの順次適用によるラベル付与手法がもつ、局所的なラベルにバイアスがかかるという問題を、大域的に正規化する手法を用いて解決し、種々の応用で採用されている。Linear-Chain CRFs は定義された損失関数を最小化することにより学習を行うが、原論文で採用されている負の対数尤度以外にも、種々の損失関数が提案されている [Kakade 02, Altun 03b, 坪井 05]。Semi-Markov Conditional Random Fields (Semi-CRFs) [Sarawagi 04] は、Linear-Chain CRFs を、系列セグメンテーション問題を解くために改良した機械学習器である。しかしながら、Semi-CRFs に関してはあまり損失関数の比較研究が行われていない。そこで、我々は、Semi-CRFs に適用可能な様々な損失関数の比較を行う。基本句構造同定と固

有表現抽出の2つの系列セグメンテーション問題について、少量データを用いて、提案する損失関数の性能の比較を行い、その結果を報告する。

1.2 系列ラベリング問題と系列セグメンテーション問題

系列ラベリング問題とは観測系列 $X = \langle x_1, \dots, x_T \rangle$ に対しラベル系列 $Y = \langle y_1, \dots, y_T \rangle$ を付与する問題のことをいう。品詞ラベル付与問題、固有表現抽出、基本句同定問題といった自然言語解析の問題は、与えられた単語系列に対しラベルを付与する系列ラベリング問題と定式化することができる。

X	Wolff	played	in	Real	Madrid
Y	NNP	VBD	IN	NNP	NNP

図 1 品詞ラベル付与問題

例えば、品詞ラベル付与問題は、 X を単語列とし、 Y を品詞ラベルとすることにより定式化できる。図 1 は、英語の品詞ラベル付与問題の例である。ここで、付与されているラベル系列 Y は、観測系列である単語系列 X の品詞を表す。NNP は固有名詞、VBD は動詞過去形、IN は前置詞を表す。

Y を範囲の境界を表すラベル (チャンクラベル) を用いることにより, 可変長の範囲 (チャンク) にラベルを付与する系列セグメンテーション問題も系列ラベリング問題の 1 つと見ることができる. 自然言語処理の分野では, 基本句同定や固有表現抽出などを, 系列セグメンテーション問題として解くことが多い. 様々なチャンクラベルが提案されており, 例えば, BIO ラベル集合 [Sang 99] の場合, B を同定する範囲の開始位置, I を同定する範囲の内部 (開始位置以外), O を同定する範囲外として用いる.

X	Wolff	played	in	Real	Madrid
Y	B-NP	B-VP	B-PP	B-NP	I-NP

図 2 基本句同定

基本句同定とは, テキスト中の再帰的に定義されない最小の句 (基本句) を同定する問題のことをいう. 図 2 は, 英語の基本句同定問題の例である. NP は名詞句, VP は動詞句, PP は前置詞句を表す. これら基本句を表すラベルと BIO ラベル集合の組み合わせにより, テキスト内の句である単語の範囲を表す. この例では “Wolff” 1 語が名詞句, “played” 1 語が動詞句, “in” 1 語が前置詞句, “Real Madrid” 2 語が名詞句を表す.

X	Wolff	played	in	Real	Madrid
Y	B-PER	O	O	B-ORG	I-ORG

図 3 固有表現抽出問題

固有表現抽出問題とは, テキスト中の固有表現の出現を同定する問題のことをいう. 図 3 は, 英語の固有表現抽出問題の例である. PER は人物を表すラベルとし, ORG は組織を表すラベルとする. これら固有表現を表すラベルと BIO ラベル集合の組み合わせにより, テキスト内の固有表現である単語の範囲を表す. 尚, O は固有表現以外の部分を表し, 固有表現以外の各単語に付与する. この例では “Wolff” 1 語が人物を表し, “Real Madrid” 2 語が組織を表す.

1.3 Linear-Chain Conditional Random Fields と損失関数スムージング

Linear-Chain Conditional Random Fields (Linear-Chain CRFs) は, Lafferty ら [Lafferty 01] により提案された, 系列ラベリングのための教師付き学習器である. 観測系列 X に対し正しいラベル系列 Y を付与する識別モデルであり, その際正しいラベル系列とそれ以外の全てのラベル系列候補とを弁別するように学習を行う. Linear-Chain CRFs は, 条件付き確率 $P(Y|X)$ を次の

指数関数モデルにより表現する:

$$P(Y|X) = \frac{\exp(W \cdot F(X, Y))}{\sum_{\tilde{Y}} \exp(W \cdot F(X, \tilde{Y}))} \quad (1)$$

ここで W はモデルのパラメータ, $\tilde{Y}$ は可能な全てのラベル系列, $F(X, Y)$ は X, Y に関する素性ベクトルで $F(X, Y)$ の 1 要素 f_i は i 番目の素性が X, Y に現れた回数とする. 素性はラベル変数と任意の観測変数との 2 つ組 (観測素性), もしくは連続するラベル変数の 2 つ組 (遷移素性) 上に定義され, 例えば, 「ある観測変数が大文字で始まり, それに対応するラベル変数が “B-PER” である」「あるラベル変数が “O” で, その 1 つ前のラベル変数が “B-ORG” である」などといったものが導入される. ラベルづけの際には $P(Y|X)$ を最大化する Y を見つけることで予測を行う.

モデルの学習時は, 入力に対する出力の条件付き確率の尤度最大化により行われる. 正しくラベルづけされた N 個の訓練事例 $\{(X^{(1)}, Y^{(1)}), \dots, (X^{(N)}, Y^{(N)})\}$ を与えられたとして, 次に定義する損失関数を最小化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. Lafferty の原論文 [Lafferty 01] では, 損失関数を負の対数尤度の和により定義している:

$$\mathcal{L}_1^{\text{linear}} = - \sum_{1 \leq i \leq N} \log P(Y^{(i)}|X^{(i)})$$

これを全損失関数 (Sequential loss function) と呼ぶ. 全損失関数は, ラベル変数の集合 $Y^{(i)}$ の尤度を最大化し, 系列の要素全体をまとめて正しく予測すると解釈できる. しかし, 学習データが少ない場合など未知データが多く系列全体の予測が困難な場合がある. この場合, 系列全体の予測により未知データに対する性能が悪くなる可能性がある. これに対し, Kakade ら [Kakade 02] は, 前ノードのラベルを無視し, 系列の各点で正しくラベルを予測するパラメータを学習するような損失関数を提案した:

$$\mathcal{L}_0^{\text{linear}} = - \sum_{1 \leq i \leq N} \sum_{t=1}^{|Y^{(i)}|} \log \sum_{\tilde{Y}: \tilde{y}_t = y_t^{(i)}} P(\tilde{Y}|X^{(i)})$$

これを点損失関数 (Point-wise loss function) と呼ぶ. ここで $\sum_{\tilde{Y}: \tilde{y}_t = y_t^{(i)}}$ は, t 番目のラベル変数が $y_t^{(i)}$ であるような全てのラベル系列に対する和を表現する. 点損失関数は隣接するラベルとの整合性に関する情報を無視するため, 可能な限り多くの目的変数のラベルを正しく予測していると解釈できる. 訓練時のパラメータの違いは遷移素性に現われる. 図 4 に遷移素性の更新の違いを示す. 太い線の丸が学習事例にあるラベル変数である. 訓練時には太線の遷移素性に関して重みが増える. 全損失関数では観測された 2 つ組のラベル変数にのみ重みが増えるが, 点損失関数では 2 つ組のラベル変数のうち, どちらかが観測されていれば重みが増えるようにパラメータが更新される. Kakade らの FAQ からの質問と解答の切り出しの実験では, 全損失関数より点損失関数の方が良

い性能であったことが報告されている．Altun ら [Altun 03b] は、品詞ラベルづけと固有表現抽出の 2 つのタスクで、これらを含む種々の損失関数の比較を行った．彼女らの 300 文規模における実験では各損失関数間で精度の差異はあまり見られなかった．

図 4 Linear-Chain Conditional Random Fields における遷移素性の更新の違い

坪井ら [坪井 05] はこの 2 つの損失関数の線形補間スムージングを提案した．スムージング率を λ とすると新しい損失関数は以下のように定義される (λ 混合損失関数と呼ぶ)：

$$\mathcal{L}_\lambda^{linear} = \lambda \mathcal{L}_1^{linear} + (1 - \lambda) \mathcal{L}_0^{linear}$$

この λ の値を変更することにより、前ノードのラベルの情報をどのくらい反映させるかを制御できる．また坪井らは、 λ 混合損失関数のスムージング率 λ を決める指針として、次に示す k 次マルコフ損失関数との関連性について示している．全損失が系列全体に対する周辺尤度、点損失が 1 点を固定した周辺尤度であるのに対し、 k 次マルコフ損失関数は $k + 1$ 個の連続した点を固定した周辺尤度となるように定義される． k 次マルコフ損失関数を定義するために式 (1) を以下のように拡張する：

$$P(Y|X) = \frac{\exp(\sum_{\tau=-k+1}^{T+k+1} W \cdot F(X, Y_\tau^{\tau+1}))}{\sum_{\tilde{Y}} \exp(\sum_{\tau=-k+1}^{T+k+1} W \cdot F(X, \tilde{Y}_\tau^{\tau+1}))}$$

ここで $X = \langle x_{-k+1}, \dots, x_{T+k-1} \rangle$, $Y = \langle y_{-k+1}, \dots, y_{T+k-1} \rangle$, $Y_\tau^{\tau+k} = \langle y_\tau, \dots, y_{\tau+k} \rangle$ とする． $t < 1$ の変数 x_t, y_t には、ダミーの変数である “BOS” (Beginning of sequence) を、 $t > T^{(i)}$ の変数 x_t, y_t には、ダミーの変数である “EOS” (End of sequence) を便宜的に定義する．この仮定のもとで k 次マルコフ損失関数は以下のように定義される：

$$\begin{aligned} \mathcal{M}_k^{Linear} &= - \sum_i \sum_{t=-k+1}^{T^{(i)}} \log \sum_{\tilde{y}: \tilde{y}_t^{t+k} = y^{(i)t+k} } P(\tilde{Y}|X^{(i)}) \\ &= - \sum_i \sum_{t=-k+1}^{T^{(i)}} \log(\end{aligned}$$

$$\sum_{\tilde{Y}} \frac{\exp(\sum_{\tau=-k+1}^{T^{(i)}+k+1} W \cdot F(X^{(i)}, Y_\tau^{\tau+1}))}{\sum_{\tilde{Y}} \exp(\sum_{\tau=-k+1}^{T^{(i)}+k+1} W \cdot F(X^{(i)}, \tilde{Y}_\tau^{\tau+1}))}.$$

坪井らは、 λ 混合損失関数のスムージング率 λ と k 次マルコフ損失関数の k の間に、整数 $k \geq 1$ において、 $\lambda = \frac{k}{k+1}$ とすると $\frac{1}{1-\lambda} \mathcal{L}_\lambda^{Linear} = \mathcal{M}_k^{Linear}$ が成り立つことを示した．全損失関数ではラベル系列全体をまとめて正しく予測するのに対し、点推定では周辺のラベルの整合性を無視し 1 点 (もしくは 1 セグメント) のラベルを数多く正答する性質がある．Linear-Chain CRFs に関して λ 混合損失関数は同値な k 次マルコフ損失関数が定義される場合、 $k + 1$ 個の連続した部分系列を可能な限り多く正答する性質がある．この事実は、スムージング率を決定するうえでの指針 となる．彼らの実験では各損失関数間で精度の差異はあまり見られなかった．

1.4 Semi-Markov Conditional Random Fields

Linear-Chain CRFs では観測系列 X 中の 1 要素 x_t が、ラベル系列 Y 中の 1 要素 y_t と対応づけられていた．基本句同定や固有表現抽出の場合、観測系列の連続する複数の要素が 1 つのラベルと対応づけられる．Sarawagi ら [Sarawagi 04] は、この問題に対応するために Linear-Chain CRFs を拡張した Semi-Markov Conditional Random Fields (Semi-CRFs) を提案した．Semi-CRFs では、観測系列 X 中の $l (l \geq 1)$ 要素 $x_t, \dots, x_{t+l-1}$ が、セグメント系列 $S = \langle s_1, \dots, s_n \rangle$ 中の 1 要素 s_k と対応づける (図 5)．ここで、 s_k は各セグメント変数 $s_k = \langle t, t+l-1, y_k \rangle$ は、対応する観測系列の開始位置 t 、同終了位置 $t+l-1$ 、セグメントに対するラベル y_k の 3 組で表現される．式の記述の便宜のために、各セグメントのラベル変数列を $Y = \langle y_1, \dots, y_n \rangle$ 、各セグメントの長さを表す長さ変数列を $L = \langle l_1, \dots, l_n \rangle$ とする．尚、セグメント系列の長さ $|S| = n$ は観測系列の長さ $|X|$ 以下となる．Semi-CRFs を、条件付き確率 $P(S|X)$ を次のように指数関数モデルにより表現する：

$$P(S|X) = \frac{\exp(W \cdot G(X, S))}{\sum_{\tilde{S}} \exp(W \cdot G(X, \tilde{S}))}$$

ここで W は、モデルのパラメータ、 $\tilde{S}$ は可能な全てのセグメント系列、 $G(X, S)$ は、 X, S に関する素性ベクトルである．Linear-Chain CRFs で定義されていた素性に加えて、複数の観測変数に対応するセグメント全体に対する素性を加えることができる．モデルの学習時には、正しくラベルづけされた N 個の訓練事例 $\{\langle X^{(1)}, S^{(1)} \rangle, \dots, \langle X^{(N)}, S^{(N)} \rangle\}$ が与えられたとして、次節に示すような損失関数を最小化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．

図 5 Semi-Markov Conditional Random Fields

2. Semi-Markov Conditional Random Fields のための損失関数スムージング

本研究では Semi-CRFs に関して損失関数の混合モデルを新規に提案する．Semi-CRFs では，次の 4 つの損失関数が定義可能である：

- 全損失:

$$\mathcal{L}_{11}^{semi} = - \sum_{1 \leq i \leq N} \log P(S^{(i)} | X^{(i)})$$

- ラベル損失:

$$\mathcal{L}_{10}^{semi} = - \sum_{1 \leq i \leq N} \sum_{t=1}^{|S^{(i)}|} \sum_{\tilde{S}: \tilde{s}_t = s_t^{(i)} \& \tilde{y}_{t-1} = y_{t-1}^{(i)}} P(\tilde{S} | X^{(i)})$$

- 長さ損失:

$$\mathcal{L}_{01}^{semi} = - \sum_{1 \leq i \leq N} \sum_{t=1}^{|S^{(i)}|} \sum_{\tilde{S}: \tilde{s}_t = s_t^{(i)} \& \tilde{l}_{t-1} = l_{t-1}^{(i)}} P(\tilde{S} | X^{(i)})$$

- 点損失:

$$\mathcal{L}_{00}^{semi} = - \sum_{1 \leq i \leq N} \sum_{t=1}^{|S^{(i)}|} \sum_{\tilde{S}: \tilde{s}_t = s_t^{(i)}} P(\tilde{S} | X^{(i)})$$

Sarawagi らの原論文 [Sarawagi 04] の損失関数は前状態のラベルのみを見るラベル損失関数 (Label-wise loss function) である*1．また，Linear-Chain CRFs のように前述の全損失関数 (Sequential loss function)，点損失関数 (Point-wise loss function) の 2 つが定義可能である．さらに前状態の長さのみを見る長さ損失関数 (Length-wise loss function) が定義可能である．Linear-Chain CRFs のときと同様に，訓練時のパラメータの違いは遷移素性に現われる．図 6 に遷移素性の更新の違いを示す．太い線の丸が学習事例にあるラベル変数である．訓練時には太線の遷移素性に関して重みが増える．

この 4 つの損失関数間で，1 つ前のセグメントのラベル変数を見るか否か，もしくは，1 つ前のセグメントの

図 6 Semi-Markov Conditional Random Fields における遷移素性の更新の違い（現在位置の長さを考慮する場合）

長さ変数を見るか否かの 2 つの観点から混合損失関数が定義可能である．1 つ前のセグメントのラベル変数を見るか否か方向のスムージング率を λ ，1 つ前のセグメントの長さ変数を見るか否か方向のスムージング率を μ とすると，提案する $\lambda\mu$ 混合損失関数は次のようになる：

$$\mathcal{L}_{\lambda\mu}^{semi} = \mu(\lambda\mathcal{L}_{00}^{semi} + (1-\lambda)\mathcal{L}_{10}^{semi}) + (1-\mu)(\lambda\mathcal{L}_{01}^{semi} + (1-\lambda)\mathcal{L}_{11}^{semi})$$

Linear-Chain CRFs における議論と同様に，ラベル k 次長さ m 次マルコフ損失関数が次のように定義できる：

$$\mathcal{M}_{km}^{Semi} = - \sum_i \sum_{t=-\omega+1}^{T^{(i)}} \log \left(\sum_{\tilde{y}: \tilde{y}_{t+\omega-k} = y_{t+\omega-k}^{(i)} \& \tilde{l}_{t+\omega-m} = l_{t+\omega-m}^{(i)}} P(\tilde{S} | X^{(i)}) \right)$$

但し， $\omega = \max(k, m)$ とする．この損失関数は， $k+1$ 個の連続した部分系列のラベルおよび $m+1$ 個の連続した部分系列の長さを可能な限り多く正答する性質がある． $\lambda\mu$ 混合損失関数のスムージング率 λ, μ とラベル k 次長さ m 次マルコフ損失 k, m の間に，整数 $k \geq 1$ および整数 $m \geq 1$ において， $\lambda = \frac{k}{k+1}, \mu = \frac{m}{m+1}$ とすると $\frac{1}{1-\lambda} \frac{1}{1-\mu} \mathcal{L}_{\lambda\mu}^{Semi} = \mathcal{M}_{km}^{Semi}$ が成り立つ．この事実は，Linear-Chain CRFs のときと同様に，スムージング率を決定する上での指針となる．

上記損失関数は 1 つ前のセグメントのどの情報を考慮するかによる．現在位置のセグメントに対しては，ラベルの情報さえ定義できれば，長さの情報がなくても，訓練，予測ともに可能である．実際，訓練データ量が少ない場合，現在位置のセグメントの長さの情報にバイアスがかかり，未知の長さのセグメントをうまく抽出できない問題が起きる．以下に現在位置のセグメントの長さの情報を考えない場合の損失関数を定義する．

*1 原論文 2.1. 節の segment feature function g に前状態の長さの情報は取り入れられていない．

図 7 Semi-Markov Conditional Random Fields における遷移素性の更新の違い（現在位置の長さを考慮しない場合）

- 全損失（現在位置のセグメント長を考慮しない）：

$$\mathcal{L}_{11}^{semi2} = - \sum_{1 \leq i \leq N} \sum_{t=1}^{|S^{(i)}|} \sum_{\tilde{S}: \tilde{y}_t = y_t^{(i)} \& \tilde{s}_{t-1} = s_{t-1}^{(i)}} P(\tilde{S} | X^{(i)})$$

- ラベル損失（現在位置のセグメント長を考慮しない）：

$$\mathcal{L}_{10}^{semi2} = - \sum_{1 \leq i \leq N} \sum_{t=1}^{|S^{(i)}|} \sum_{\tilde{S}: \tilde{y}_t = y_t^{(i)} \& \tilde{y}_{t-1} = y_{t-1}^{(i)}} P(\tilde{S} | X^{(i)})$$

- 長さ損失（現在位置のセグメント長を考慮しない）：

$$\mathcal{L}_{01}^{semi2} = - \sum_{1 \leq i \leq N} \sum_{t=1}^{|S^{(i)}|} \sum_{\tilde{S}: \tilde{y}_t = y_t^{(i)} \& \tilde{l}_{t-1} = l_{t-1}^{(i)}} P(\tilde{S} | X^{(i)})$$

- 点損失（現在位置のセグメント長を考慮しない）：

$$\mathcal{L}_{00}^{semi2} = - \sum_{1 \leq i \leq N} \sum_{t=1}^{|S^{(i)}|} \sum_{\tilde{S}: \tilde{y}_t = y_t^{(i)}} P(\tilde{S} | X^{(i)})$$

図 7 に現在位置の長さを考慮しない場合の遷移素性の更新の違いを示す．太い線の丸が学習事例にあるラベル変数である．訓練時には太線の遷移素性に関して重みが増える．同様に，混合損失関数が以下のように定義できる*2：

$$\mathcal{L}_{\lambda\mu}^{semi2} = \mu(\lambda\mathcal{L}_{00}^{semi2} + (1-\lambda)\mathcal{L}_{10}^{semi2})$$

*2 尚，この損失関数に対応するマルコフ損失関数は以下になる：

$$\mathcal{M}_{km}^{Semi2} = - \sum_i \sum_{t=-\omega+1}^{T^{(i)}} \log \left(\sum_{\tilde{S}: \tilde{y}_{t+\omega-k}^{t+\omega} = y_{t+\omega-k}^{t+\omega} \& \tilde{l}_{t+\omega-m}^{t+\omega-1} = l_{t+\omega-m}^{t+\omega-1}} P(\tilde{S} | X^{(i)}) \right)$$

元の損失関数とは異なり，長さに関して $m+1$ 個ではなく m 個の連続した部分系列を可能な限り多く正答する性質がある．

$$+(1-\mu)(\lambda\mathcal{L}_{01}^{semi2} + (1-\lambda)\mathcal{L}_{11}^{semi2})$$

我々はこのスムージング手法をタスクにおける前のセグメント情報の重要性をモデルに反映するために提案する．例えば，固有表現抽出の場合，固有表現を示すセグメントが接続することは少ない．固有表現以外について 1 トークンを 1 セグメントとして扱うとすると，殆どのセグメントに関して，前セグメントの長さの情報は意味をなさず，固有表現でないというラベルの情報のみが有効であろう．一方，一般的な基本句同定の場合，基本句を示す様々な長さのセグメントが隣接する．このような場合には前セグメントの長さ，ラベルともに有効であろうと考える．

3. 評価実験

CoNLL 2000 の英語の基本句同定データ [Sang 00] と CoNLL 2002 のスペイン語の固有表現抽出データ [Sang 02] により評価実験を行う．CoNLL 2002 のデータでは訓練データ，開発データ，評価データの 3 つが用意されている．CoNLL 2000 のデータでは訓練データと評価データの 2 つが公開されており，開発データが公開されていない．公開されている訓練データのうち 600 文を“本実験における訓練データ”，残りの部分を“本実験における開発データ”として用いる．各タスクにおいて，訓練データを用いて訓練を行い，開発データ，評価データにより性能の比較を行う．

Linear-Chain CRFs の各損失関数とスムージング率を変更した λ 混合損失関数，Semi-CRFs の各損失関数とスムージング率を変更した $\lambda\mu$ 混合損失関数などの比較を行う．坪井ら [坪井 05] の実験にならい，スムージング係数は $\lambda = \frac{k}{k+1}$, $\mu = \frac{m}{m+1}$ とした場合の k, m を 0, 1~5, ∞ の範囲で動かして，各タスクの F 値の変化を検証する．また，Altun ら [Altun 03b] の実験にならい，Linear-Chain CRFs の実験については，前後 1 単語ずつを含む 3 単語のウィンドウ幅の次の事象を用いた：単語，品詞，1~3 文字の接辞，最初の文字が大文字か小文字か数字かそれ以外か，単語中の全ての単語が大文字もしくはハイフンか，単語中の全ての単語が大文字か否か，単語にハイフンが含まれるか否か．また付与するラベルとして，BIO タグ (B がチャンクの開始位置，I がチャンクの内部で開始位置以外，O がチャンクの外部) を用いる．Semi-CRFs の実験については，固有表現の前後 1 単語ずつと最初と最後の単語に関する上記事象を素性として用い，展開するノードの最大長を訓練データ中の事例の最大長に設定した．また，Semi-CRFs において，現在位置の長さを考慮しないモデルの方が性能が良かったため，その結果を示す．最適化の手法として L-BFGS 法を用い，全ての損失関数において二次の正規化項 (Gaussian Prior) [Chen 99] $-\frac{\|W\|}{2\sigma}$ を導入した．二次の正規化項の

図 8 スムージング率の変化による精度の変化：CoNLL-2000 基本句同定 開発データ
(300 文，現在位置の長さを考慮しない，正規化項あり ($\sigma = 1.0$))

図 9 スムージング率の変化による精度の変化：CoNLL-2000 基本句同定 評価データ
(300 文，現在位置の長さを考慮しない，正規化項あり ($\sigma = 1.0$))

図 10 スムージング率の変化による精度の変化：CoNLL-2002 固有表現抽出 開発データ
(300 文，現在位置の長さを考慮しない，正規化項あり ($\sigma = 1.0$))

図 11 スムージング率の変化による精度の変化：CoNLL-2002 固有表現抽出 評価データ
(300 文，現在位置の長さを考慮しない，正規化項あり ($\sigma = 1.0$))

表 1 実験結果:CoNLL-2000 基本句同定 (文数変化)
(現在位置の長さを考慮しない, 正規化項あり $\sigma = 1.0$)

		Linear-Chain CRFs				Semi-CRFs					
		点	全	$\lambda(k)$		点	ラベル	長さ	全	$\lambda\mu(k,m)$	
				開発最良	評価最良					開発最良	評価最良
100	開発	80.82	83.49	83.86 (5)		85.26	85.02	85.14	85.04	85.39 (4,1)	85.33(2,0)
100	評価	81.35	84.01	84.49 (5)		85.46	85.16	85.24	85.15	85.48(4,1)	85.52 (2,0)
200	開発	83.94	86.46	86.36(5)		87.52	87.35	87.44	87.41	87.57 (1,4)	87.56(0,3)
200	評価	84.78	87.10	87.07(5)		88.04	87.98	88.01	87.85	88.12(1,4)	88.16 (0,3)
300	開発	84.82	87.29	87.26(4)		88.60	88.45	88.51	88.38	88.64 (5,2)	88.64(1,1)
300	評価	85.64	87.94	87.96 (4)		89.11	89.05	88.99	88.85	89.11(5,2)	89.12 (1,1)
600	開発	86.42	89.10	88.81(4)	88.79(5)	89.64	89.53	89.59	89.55	89.69 (5,3)	89.65(∞ ,1)
600	評価	87.05	89.75	89.39(4)	89.40(5)	90.05	90.01	90.03	90.01	90.06(5,3)	90.14 (∞ ,1)

表 2 実験結果:CoNLL-2002 固有表現抽出 (文数変化)
(現在位置の長さを考慮しない, 正規化項あり $\sigma = 1.0$)

		Linear-Chain CRFs				Semi-CRFs					
		点	全	$\lambda(k)$		点	ラベル	長さ	全	$\lambda\mu(k,m)$	
				開発最良	評価最良					開発最良	評価最良
100	開発	35.54	42.95	41.21(5)		42.17	41.66	41.36	41.91	42.17 (2,4)	41.99(4, ∞)
100	評価	41.99	50.59	49.58(5)		48.34	48.02	48.08	49.03	48.70(2,4)	49.21 (4, ∞)
200	開発	40.85	47.24	45.35(5)		44.95	44.62	44.60	47.40	46.04(∞ ,5)	45.85(4, ∞)
200	評価	48.10	55.81	54.51(5)		52.24	51.96	52.23	54.84	53.80(∞ ,5)	53.98(4, ∞)
300	開発	48.12	54.47	52.78(5)		54.85	54.36	54.46	54.29	55.08 (3, ∞)	54.92(∞ ,2)
300	評価	53.24	61.00	60.43(5)		58.72	58.48	58.60	60.56	59.31(3, ∞)	59.37(∞ ,2)
600	開発	52.22	57.90	57.35(5)		53.83	55.75	55.41	55.54	55.78 (∞ ,1)	
600	評価	55.56	64.34	63.50(5)		60.19	62.11	61.66	61.72	62.05(∞ ,1)	

ためのパラメータ σ は, $\sigma = 0.001 \sim 10$ の値で, 各タスクにおいて訓練データ 300 文により比較実験を行い, 点損失, ラベル損失, 長さ損失, 全損失の 4 つの損失関数で F 値が最良であった $\sigma = 1.0$ を用いる. 尚, 4 つの損失関数で F 値が最良となる σ の値に齟齬はなかった.

訓練事例数を 100 ~ 600 文の間で動かした場合の性能の推移について, 基本句同定の評価データの実験結果を表 1 に, 固有表現抽出の開発データと評価データの実験結果を表 2 に示す. $\lambda(k)$ と $\lambda\mu(k,m)$ は, スムージング率を変えて最も良かった値のみを掲載する. 太字は, 各学習器で F 値最良を示す. 括弧内の値は, 最良値が得られたスムージング率で, 0 は $k = 0$ もしくは $m = 0$ (つまり $\lambda = 0$ もしくは $\mu = 0$), ∞ は $k = \infty$ もしくは $m = \infty$ (つまり $\lambda = 1$ もしくは $\mu = 1$) を意味する.

基本句同定の実験結果に関しては, Linear-Chain CRFs と Semi-CRFs の結果を比較すると Semi-CRFs の方が性能が良いことがわかった. Semi-CRFs については, スムージングを用いた損失関数が他の損失関数と同等もしくはそれ以上の性能を持つ可能性があることがわかった.

固有表現抽出の実験結果に関しては, Linear-Chain CRFs と Semi-CRFs の結果を比較すると, データサイズによってどちらの方が性能がよいか様々であった. Linear-Chain CRFs については, 全てのデータサイズで全損失関数の性能が良かった. これに対し, Semi-CRFs については, データサイズ毎に良いモデルが異なった. また, スムージングをした際の最良の点も λ, μ のどちらかのスムージング係数が 1 の場合が多く, 必ずしも 2 次元方向のスムージングが有効ではないことがわかる. さらに必ずしも開発データと評価データとで最良の性能をも

つスムージング率は異なった.

この基本句同定と固有表現抽出の実験結果の差異はデータの性質の差異によるものだと考える. 基本句同定データは切り出されるチャンクが連続して出現するのに対し, 固有表現抽出データは切り出されるチャンクがあまり連続して出現しない. 基本句同定問題では, 様々な長さやラベルのチャンク (句構造) が接続するため, Semi-CRFs の方がよりモデルにこの性質を反映することができ, 性能が良くなったと考える. しかしながら固有表現抽出は, チャンク (固有表現) 間に, チャンク以外のトークン (固有表現以外, いわゆる O タグ) が挟まれることが多い. チャンク以外のトークンは長さが 1 であるので, 今回の実験設定では Linear-Chain CRFs と Semi-CRFs とでは, ほとんどそのトークンに対する素性が変わらない. このため, 両者で性能の差異が見られなかったと考える. このことは Semi-CRFs の優位性を否定しているわけではない. Sarawagi ら [Sarawagi 04] のようにセグメント単位の素性を入れることで Linear-Chain CRFs より性能の良いモデルを作成できると考える.

次に二次の正規化項を用いない場合との比較を行う. 実験の設定は, 300 文のデータを用い, 二次の正規化項を用いる場合の σ の値は, 先ほどと同じ値 1.0 を用いる. この実験設定でも現在位置セグメント長素性を用いていない. 結果を表 3, 表 4 に示す. 二次の正規化項を用いない場合には, 用いる場合と比べて, 今回提案するスムージングによる性能の向上幅が大きいことがわかった. また, 二次の正規化項のみを用いる場合と今回提案するスムージングとを用いる場合と比べた場合には, 二次の正規化項の方が性能の向上が顕著であった.

表 3 実験結果:CoNLL-2000 基本句同定 (正規化項の有無)
(訓練データ 300 文, 現在位置の長さを考慮しない)

正規化項 /データ	Semi-CRFs				
	点	ラベル	長さ	全	$\lambda\mu(k,m)$ 最良
なし/開発	85.62	86.03	85.19	85.23	86.71 (3,1)
なし/評価	86.53	86.37	85.92	85.80	87.78 (3,4)
$\sigma=1.0$ /開発	88.60	88.45	88.51	88.38	88.64 (5,2)
$\sigma=1.0$ /評価	89.11	89.05	88.99	88.85	89.12 (1,1)

表 4 実験結果:CoNLL-2002 固有表現抽出 (正規化項の有無)
(訓練データ 300 文, 現在位置の長さを考慮しない)

正規化項 /データ	Semi-CRFs				
	点	ラベル	長さ	全	$\lambda\mu(k,m)$ 最良
なし/開発	49.26	48.48	44.75	48.89	53.47 ($\infty,5$)
なし/評価	55.98	53.74	52.33	55.09	58.08 (5, ∞)
$\sigma=1.0$ /開発	54.85	54.36	54.46	54.29	55.08 (3, ∞)
$\sigma=1.0$ /評価	58.72	58.48	58.60	60.56	59.37($\infty,2$)

次に現在位置の長さを考慮する場合としない場合との比較を行う。実験の設定は, 300 文のデータを用い, 二次の正規化項の σ の値は, 先ほどと同じ値 1.0 を用いる。結果を表 5, 表 6 に示す。有意な差ではないが, 現在位置のセグメント長素性がない場合の方が性能が良いことが多い。これは, データ量が少ない場合に, 現在位置セグメント長素性のバイアスがかかり, 現在位置のセグメント長が全く同じであるようなセグメントが少ないため精度を悪化させているのであろうと推察される。

最後に, 基本句解析と固有表現抽出の両タスクにおいて, スムージング率 λ, μ を $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ と動かして F 値, 再現率, 精度がどのように変化するかを調査した。基本句解析の実験において F 値, 再現率, 精度をプロットしたグラフを図 8 (開発データ) 図 9 (評価データ) に示す。どの評価指標でも凸グラフにはならず, 最適なスムージング率を推定することが困難である。また固有表現抽出の実験において F 値, 再現率, 精度をプロットしたグラフを図 10 (開発データ) 図 11 (評価データ) に示す。グラフから, 必ずしも最良のスムージング値が一致しないことがわかった。

表 5 実験結果:CoNLL-2000 基本句同定 (現在位置セグメント長素性の有無)

(訓練データ 300 文, 正規化項あり ($\sigma=1.0$))					
現在位置 セグメント 長素性 /データ	Semi-CRFs				
	点	ラベル	長さ	全	$\lambda\mu(k,m)$ 最良
なし/開発	88.60	88.45	88.51	88.38	88.64 (5,2)
なし/評価	89.11	89.05	88.99	88.85	89.12 (1,1)
あり/開発	88.42	88.34	88.41	88.20	88.65 (2,1)
あり/評価	88.84	88.88	88.98	88.80	89.09 (3,1)

表 6 実験結果:CoNLL-2002 固有表現抽出 (現在位置セグメント長素性の有無)

(訓練データ 300 文, 正規化項あり ($\sigma=1.0$))					
現在位置 セグメント 長素性 /データ	Semi-CRFs				
	点	ラベル	長さ	全	$\lambda\mu(k,m)$ 最良
なし/開発	54.85	54.36	54.46	54.29	55.08 (3, ∞)
なし/評価	58.72	58.48	58.60	60.56	59.37($\infty,2$)
あり/開発	54.64	53.90	53.97	53.92	54.89 (1,2)
あり/評価	60.39	59.89	59.71	60.40	60.22(1,0)

4. 関 連 研 究

一般的な CRFs の学習は尤度最大化もしくは事後確率最大化 [Sha 03] により行う。鈴木ら [鈴木 06] は CRFs の学習手法に識別誤り最小化による手法を提案している。彼らの手法では系列セグメンテーション問題の 1 評価手法である F 値を 1 から引いたものを用いる。学習時には Semi-CRFs に用いられている手法がそのまま用いることができる。彼らの実験結果では, 固有表現抽出のような切り出されるチャンクがあまり連続して出現しない問題で, より顕著な改善が見られた。また, 彼らは, 識別誤り最小化は尤度最大化もしくは事後確率最大化による学習手法の一般化であることも示している。その他の CRFs の学習の枠組みとして, Altun [Altun 03c] らおよび Taskar [Taskar 04] らはマージン最大化による手法を Altun [Altun 03a] らはブースティングによる手法をそれぞれ提案している。

5. お わ り に

本稿では, Semi-CRFs に対するスムージング手法を提案し, 基本句同定と固有表現抽出の 2 つの系列セグメンテーション問題で比較実験を行った。基本句同定のような長さ 2 以上のチャンクが連続するような系列データの場合には提案手法が有効であった。一方, 固有表現抽出のような長さ 2 以上のチャンクがあまり連続しないような系列データの場合には有効性が見られなかった。またスムージング率の設定に関しても困難であることがわかった。Linear-Chain CRFs における k 次マルコフ損失関数は $k+1$ 個の部分系列を, 正当するように定義されている。同様に Semi-CRFs におけるラベル k 次長さ m 次マルコフ損失関数はラベルに関して $k+1$ 個の部分系列を, 長さに関して $m+1$ 個の部分系列を正当するように定義されている。実際には正規化項による影響の方が大きく上記事実を裏付ける実験結果にならなかった。スムージング率と F 値の相関についても凸グラフにならず, 固有表現抽出における交差検定においても開発セットと評価セットとで最良となるスムージング率が一致しなかった。今回は小規模なデータを用いた実験であったが, 大規模なデータを用いる場合交差検定による最良のスムージング率の推定は, 膨大な計算時間を必要とする。

今後の課題として、交差検定以外の手法でスムージング率を決定する枠組を考える必要がある。

◇ 参 考 文 献 ◇

- [Altun 03a] Altun, Y., Hofmann, T., and Johnson, M.: Discriminative Learning for Label Sequences via Boosting, in *NIPS-2003* (2003)
- [Altun 03b] Altun, Y., Johnson, M., and Hofmann, T.: Investigating Loss Functions and Optimization Methods for Discriminative Learning of Label Sequences, in *EMNLP-2003* (2003)
- [Altun 03c] Altun, Y., Tsochantaridis, I., and Hofmann, T.: Hidden Markov Support Vector Machines, in *ICML-2003* (2003)
- [Chen 99] Chen, S. F. and Rosenfeld, R.: A gaussian prior for smoothing maximum entropy models, Technical report, Carnegie Mellon University (1999)
- [Kakade 02] Kakade, S., Teh, Y. W., and Roweis, S.: An alternative objective function for Markovian fields, in *ICML-2002* (2002)
- [Lafferty 01] Lafferty, J., McCallum, A., and Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in *ICML-2001* (2001)
- [Sang 99] Sang, E. F. T. K. and Meulder, F. D.: Representing Text Chunks, in *EACL-1999*, pp. 173–179 (1999)
- [Sang 00] Sang, E. F. T. K. and Buchholz, S.: Introduction to the CoNLL-2000 Shared Task: Chunking, in *CoNLL-2000*, pp. 127–132 (2000)
- [Sang 02] Sang, E. F. T. K.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition, in *CoNLL-2002*, pp. 155–158 (2002)
- [Sarawagi 04] Sarawagi, S. and Cohen, W. W.: Semi-markov conditional random fields for information extraction, in *NIPS-2004* (2004)
- [Sha 03] Sha, F. and Pereira, F.: Shallow Parsing with Conditional Random Fields, in *HLT/NAACL-2003* (2003)
- [Taskar 04] Taskar, B., Guestrin, C., and Koller, D.: Max-Margin Markov Networks, in *NIPS-2004* (2004)
- [鈴木 06] 鈴木 潤, 磯崎 秀樹: 学習誤り最小化に基づく条件付き確率場の学習: 言語解析への適用, 言語処理学会第 12 回年次大会発表論文集, pp. 548–551 (2006)
- [坪井 05] 坪井 祐太, 鹿島 久嗣: 構造データのラベル付け学習モデルの設計, in *IBIS-2005*, pp. 15–20 (2005)

〔担当委員: 奥村 学〕

2006 年 5 月 2 日 受理

—— 著 者 紹 介 ——

福岡 健太

2004 年名古屋工業大学工学部知能情報システム学科卒業。
2006 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了。同年日本電気株式会社入社、現在に至る。

浅原 正幸

1998 年京都大学総合人間学部基礎科学科卒業。2001 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了。同年より日本学術振興会特別研究員。2003 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。2004 年同大学助手、現在に至る。博士(工学)。自然言語処理の研究に従事。情報処理学会, 言語処理学会会員。

松本 裕治(正会員)

1977 年京都大学工学部情報工学科卒。1979 年同大学大学院工学研究科修士課程情報工学専攻修了。同年電子技術総合研究所入所。1984~85 年英国インペリアルカレッジ客員研究員。1985~87 年(財)新世代コンピュータ技術開発機構に出向。京都大学助教授を経て、1993 年より奈良先端科学技術大学院大学教授、現在に至る。工学博士。専門は自然言語処理。情報処理学会, 日本ソフトウェア科学会, 言語処理学会, 認知科学会, AAAI, ACL, ACM

各会員。