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

福岡 健太　　浅原 正幸　　松本 裕治

奈良先端科学技術大学院大学 情報科学研究科

Abstract

自然言語処理の中で系列ラベリングは品詞ラベルづけ、基礎句構造同定、固有表現抽出など種々のタスクに用いられている。系列ラベリングで用いられる 1 手法として、Linear-Chain Conditional Random Fields がある。この手法に関し Altun らは様々な損失関数を比較した。また、坪井らは損失関数のスムージングを提案し、精度比較を行った。ラティス中の 1 つのノードに対し 1 つの観測値が対応する Linear-Chain Conditional Random Fields に対し、Sarawagi らは 1 つのノードに対し 可変長の範囲の観測値を対応するようにラティスを構成する Semi-Markov Conditonal Random Fields を提案した。本稿では、Semi-Markov Conditional Random Fields における損失関数のスムージングを提案し、坪井らと同様の比較実験を行う。

1　背景

1.1　Linear-Chain CRFs と損失関数スムージング

Linear-Chain Conditional Random Fields (Linear-Chain CRFs) は、Lafferty ら [6] により提案された、系列ラベリングのために設計された教師付き学習器である。観測系列 $X = \langle x_1, \ldots, x_T \rangle$ に対し正しいラベル系列 $Y = \langle y_1, \ldots, y_T \rangle$ 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。自然言語処理の分野では、X を単語列とし、Y を品詞ラベルとする品詞ラベル付与や、Y をチャンクの境界を表すラベルとするチャンキングなどに用いられる。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：

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

ここで W は、モデルのパラメータ、$F(X,Y)$ 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” 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” 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。

モデルの学習時には、正しくラベルづけされた N 個の訓練事例 $\{\langle X^{(1)}, Y^{(1)} \rangle, \cdots, \langle X^{(N)}, Y^{(N)} \rangle\}$ 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 [6] では、損失関数を負の対数尤度の和により定義している：

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

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の性能が悪くなる可能性がある。これに対し、Kakade ら [3] 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：

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

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 1 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。Altun ら [1] 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。

坪井ら [8] 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)：

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

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。

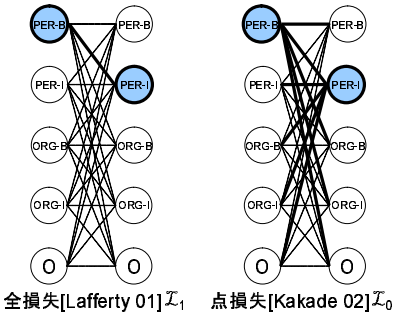

図 1: Linear-Chain CRFs における遷移素性の更新の違い

また、坪井らは損失関数の線形補間スムージングに関し、この λ 混合損失関数と k 次マルコフ損失関数の 2 つの見方を示し、相互のパラメータの関連についても示している。

1.2 Semi-Markov CRFs

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 x_t と対応づけられていた。Sarawagi ら [7] により提案された Semi-Markov Conditional Random Fields (Semi-Markov CRFs) では、観測系列 X 中の $l(l \geq 1)$ 要素 $x_t, \ldots, x_{t+l-1}$ が、セグメント系列 $S = \langle s_1, \ldots, s_n \rangle$ 中の 1 要素 s_k と対応づけるように拡張したモデルである。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 つ組で表現される。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 $Y = \langle y_1, \ldots, y_n \rangle$、各セグメントの長さを表わす長さ変数列を $L = \langle l_1, \ldots, l_n \rangle$ 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-Markov 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、$G(X,S)$ 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 $\{\langle X^{(1)}, S^{(1)} \rangle, \cdots, \langle X^{(N)}, S^{(N)} \rangle\}$ 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。

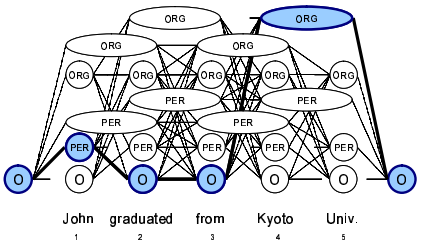

図 2: Semi-Markov CRFs

2 Semi-Markov CRFs のための損失関数スムージング

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

- 全損失 (Sequential loss):

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

- ラベル損失 (Label-wise loss):

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

ここで $\sum_{\tilde{S}:\tilde{s}_t = s_t^{(i)} \& \tilde{y}_{t-1} = y_{t-1}^{(i)}}$ は、t 番目のセグメント変数が $s_t^{(i)}$ であり、かつ $t-1$ 番目のセグメントのラベルが $y_{t-1}^{(i)}$ であるような全てのセグメント系列に対する和を表現する。また $\#|S^{(i)}|$ は、各 i 番目の事例のセグメント系列候補の長さを表し、セグメント系列候補によって可変な値を示す。

- 長さ損失関数 (Length-wise loss):

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

ここで $\sum_{\tilde{S}:\tilde{s}_t = s_t^{(i)} \& \tilde{l}_{t-1} = l_{t-1}^{i}}$ は、t 番目のセグメント変数が $s_t^{(i)}$ であり、かつ $t-1$ 番目のセグメントの長さが $l_{t-1}^{(i)}$ であるような全てのセグメント系列に対する和を表現する。

- 点損失関数 (Point-wise loss):

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

ここで $\sum_{\tilde{S}:\tilde{s}_t = s_t^{(i)}}$ は、t 番目のセグメント変数が $S_t^{(i)}$ であるような全てのセグメント系列に対する和を表現する。

Sarawagi らの原論文 [7] 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。また、Linear-Chain CRFs のように前

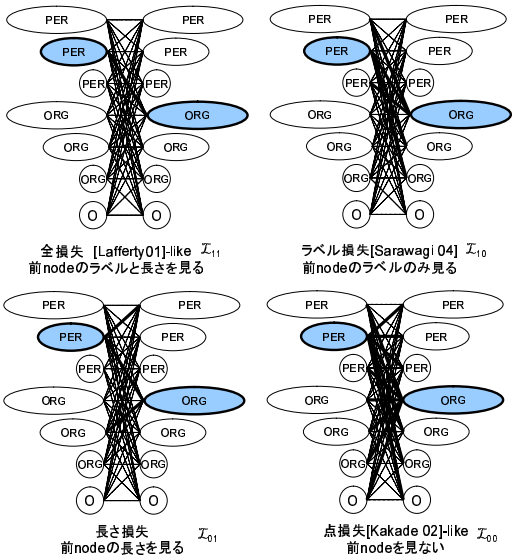

図 3: Semi-Markov CRFs における遷移素性の更新の違い

述の全損失関数、点損失関数の 2 つが定義可能である。さらに前状態の長さのみを見る損失関数が定義可能である。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 3 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、混合損失関数は次のようになる：

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

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連接することは少ない。固有表現以外のセグメントを 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情報が有効であろう。一方、一般的な基本句構造同定の場合、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 の英語の句構造同定データ [5] と CoNLL 2002 のスペイン語の固有表現抽出データ [4] により評価実験を行う。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-Markov 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。句構造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。坪井ら [8] 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 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。

Altun ら [1] 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 Window 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-Markov 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。最適化の手法として L-BFGS 法を用い、全ての実験において二次の正規化項 (Gaussian Prior) [2] を導入した。

訓練事例数を 100〜600 文の間で動かして精度の推移を見る。表 1 に基本句構造同定の結果を、表 2 に固有表現抽出の結果を示す。$\lambda(k)$ 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。詳細な値については、文献 [9] を参照されたい。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 $k=\inf$ もしくは $m=\inf$（つまり $\lambda=1$ もしくは $\mu=1$）を意味する。坪井らの結果と同様にスムージング率を変化させると、各オリジナルの損失関数よりも性能が良くなる場合があるという程度の結果であった。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 0 もしくは 1 の場合が多く、必ずしも 2 次元のスムージングが有効ではないことがわかる。各実験において F 値をプロットしてもきれいな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。固有表現抽出では、開発セットと評価セット両方について評価を行ったが、この 2 つのデータセットにおいて必ずしも最良のスムージング値が一致しないという結果になった。

Linear-Chain CRFs と Semi-Markov CRFs を比較した際に Semi-Markov CRFs の性能が悪いのは、今回の実験で Linear-Chain CRFs に合わせた素性を導入したためである。Semi-Markov CRFs に導入することが可能なセグメント単位の素性を入れていない。セグ

表 1: 実験結果:基本句構造同定

	Linear-Chain CRFs			Semi-Markov CRFs				
	点	全	λ(k) 最良	点	ラベル	長さ	全	$\lambda\mu$ (k,m) 最良
100	81.35	84.01	**84.49**(5)	79.79	80.18	79.93	79.91	**80.40**(4,0)
200	84.78	**87.10**	87.07(5)	84.25	84.49	84.25	84.39	**84.61**(3,0)
300	85.64	87.94	**87.96**(4)	86.75	86.80	86.89	86.86	**87.09**(4,2)
600	87.05	**89.75**	89.40(5)	88.85	88.81	88.93	89.00	**89.03**(5,2)

表 2: 実験結果:固有表現抽出

		Linear-Chain CRFs			Semi-Markov CRFs					
		点	全	λ(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)	39.18	39.29	39.36	39.38	**39.76**(0,2)	39.70(0,1)
100	評価	41.99	**50.59**	49.58(5)	46.29	46.44	46.28	46.11	46.58(0,2)	**46.91**(0,1)
200	開発	40.85	**47.24**	45.35(5)	44.91	**46.02**	45.46	45.98	45.99(-,1)	45.50(1,0)
200	評価	48.10	**55.81**	54.51(5)	53.39	53.54	53.32	53.56	53.56(-,1)	**53.67**(1,0)
300	開発	48.12	**54.47**	52.78(5)	51.24	51.88	51.61	51.89	**52.03**(0,1)	51.77(-,1)
									52.03(1,0)	
300	評価	53.24	**61.00**	60.43(5)	58.94	59.36	59.13	59.47	58.95(0,1)	**59.48**(-,1)
									59.05(1,0)	
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)	

メント単位の素性を入れることにより、Linear-Chain CRFs の性能を越すことは可能[1]だと考えている。

4 まとめ

本稿では、Semi-Markov CRFs に対するスムージング手法を提案し、複数のタスクで比較実験を行った。残念ながらスムージングによる性能改善の可能性はあるが、スムージング率の決定など、扱いが困難であることがわかった。今後の改善の可能性として、tri-gram への拡張がある。坪井らの手法も我々の提案手法も基本的に uni-gram（点損失）と bi-gram（全損失）間のスムージングである。Linear-Chain CRFs で 全損失が良いということは、それより長い文脈を見た tri-gram の情報も有効である可能性があり、uni-gram 〜 tri-gram 間でのスムージングもしくは、選択的に tri-gram を入れる手法が考えられる。Semi-Markov CRFs でも同様のスムージングが定義でき、さらに各点でラベル、長さのどの情報を利用すべきかなどの素性エンジニアリングが可能であろう。

参考文献

[1] Y. Altun, M. Johnson, and T. Hofmann. Investigating Loss Functions and Optimization Methods for Discriminative Learning of Label Sequences. In *EMNLP-2003*, 2003.

[2] S. F. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University, 1999.

[3] S. Kakade, Y. W. Teh, and S. Roweis. An alternative objective function for Markovian fields. In *ICML-2002*, 2002.

[4] E. F. T. Kim. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *CoNLL-2002*, pp. 155–158, 2002.

[5] E. F. T. Kim and S. Buchholz. Introduction to the CoNLL-2000 Shared Task: Chunking. In *CoNLL-2000*, pp. 127–132, 2000.

[6] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML-2001*, 2001.

[7] S. Sarawagi and W. W. Cohen. Semi-markov conditional random fields for information extraction. In *NIPS-2004*, 2004.

[8] 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.

[9] 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.

[1]実際、現在位置のノードの長さ素性の用い方で精度が変動する。発表では、これについても言及したい。