ラティス構造とニューラルネットワークに基づく系列ラベリング

佐藤 元紀　　進藤 裕之　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{sato.motoki.sa7, shindo, matsu}@is.naist.jp

1 はじめに

系列ラベリングにおいて, 入力文に対して深層学習の手法を適応し, 高い精度を報告している研究が多く存在する [5]. 本研究では，系列ラベリングにおいて系列タグを予測する Bidirectional LSTM(BiLSTM) を訓練し，さらに BiLSTM の出力層に基いてラティスを構築し，ラティス上の最適なパスを探索するニューラルネットワークモデルを提案する．評価実験を通して，ラティス構造を用いることで高い精度を得ることができることを示し，提案手法の有効性を示す．また提案手法が CoNLL 2003 コーパスの固有表現認識における最高精度，F 値 91.26 を達成した．

2 関連研究

系列ラベリングは，入力文 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ に対して，タグ系列 $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ を予測するタスクである．系列ラベリングの 1 つである固有表現認識 (Named Entity Recognition; NER) は，入力文に対して人名や地名などの固有表現に対して系列タグ (Begin; B, Inside; I, Other; O, End; E, Single; S) を予測するタスクとして定式化される．英語の NER で使われる CoNLL 2003[8] コーパスに対する最高精度の手法は，Ma ら [5] の研究である．Ma らは, BiLSTM に対して単語単位のベクトルと, 文字単位のベクトルを素性として与え，最終層で条件付確率場 (Conditional Random Fields; CRF)[4] を用いて最適なタグ系列を求めることで CoNLL 2003 における最高精度を達成した．

一方で，系列タグ単位での Linear CRF で最適化しているため，単語数が長い固有表現の場合,「BIIIIE」というタグを予測することになる．ラティスを用いることで固有表現を 1 つのノードとして扱うことができ, 文全体として最適な固有表現を同定することが期待される．Muller ら [6] の研究では，より高次の CRF を段階的に学習することでタギングのタスクで性能向上することが示されている．入力文に対して辞書を用いてラティスを構築し，ラティスの正解パスを探索する手法は, 日本語形態素解析 [3] で使われているが，固有表現の場合, 辞書の網羅率が低くなりやすいという問題がある．本研究では，まず系列タグを予測するニューラルネットワークを訓練し，ニューラルネットワークの出力層に基づき，ラティスを構築する．ラティスの構築については，3.2 章で詳しく述べる．

固有表現は複数単語から構成されるため，系列タグを単語単位で予測するモデルでは，辞書の素性をモデルに組み込むことが単純ではない．一方で固有表現をラティスのノードとして扱うことで，辞書素性が入れやすくなるという利点もある．

ラティスを用いる利点は以下の通りである．

- 単語数が長い固有表現を 1 つのノードとして扱うことができる．
- 固有表現を 1 つのノードとして扱うので辞書素性が入れやすくなる．

ラティス探索は，係り受け解析における高い性能を得ている手法を用いる [1]．Andor ら [1] は入力が単語単位であるが，ラティスの場合はノードの長さは可変長である．本研究では，Andor らの手法を可変長のノードに対して使えるように手法を拡張し，ラティス探索に応用した．

3 提案手法

本節では，提案手法について説明する．本研究で提案するモデルは 2 つのニューラルネットワークを用いる．提案手法の概要を 図 1 に示す．用いた 2 つのモデルは，系列タグを予測するリカレントニューラルネットワークと，ラティス探索をするためのモデルである．系列ラベリングのモデルは, 現在最高精度の先行研究 [5] のモデルを用いた．ニューラルネットワークの出

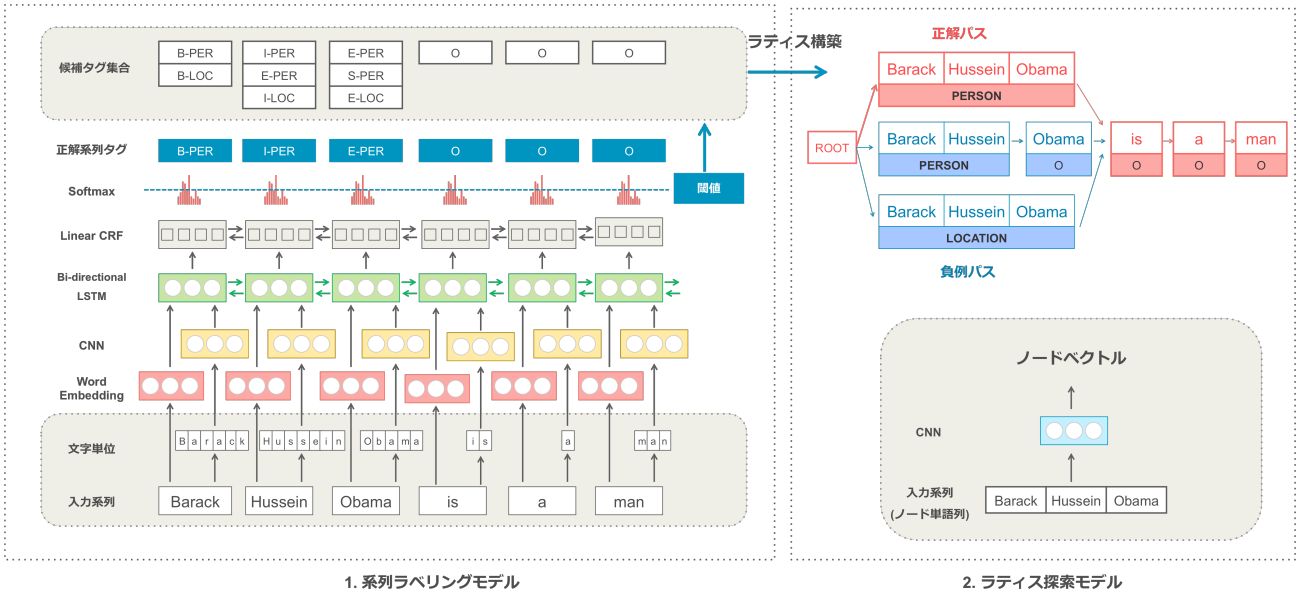

図 1: 提案手法の概要

力層に閾値を設定し，閾値以上の系列タグを系列タグ候補集合とし，ラティスを構築する．ラティス探索は，Andor ら [1] の手法と，ラティスに対して CRF で最適化する 2 つの手法について比較を行った．ただし，ラティス探索の場合，入力単位が可変長のノードとなるため，先行研究の手法を拡張した．ラティスの構築・探索，及びノードの扱いに関しては 3.2 章で詳しく述べる．

本研究の貢献を以下にまとめる．

- 系列タグを予測するニューラルネットワークの出力層に基いてラティスを構築し，ラティス上の最適なパスを探索する手法を提案した．
- CoNLL 2003 コーパスにおける最高精度 F 値 91.26 を達成した．

3.1 系列ラベリングモデル

まず単語列から系列タグ (IOBES タグ) を予測するモデルを訓練する．NER のタスクにおける最高精度を達成している Ma らの Bidirectional LSTM-CNN-CRF(BiLSTM-CNN-CRF)[5] モデルを採用した．

Bidirectional LSTM-CNN (BiLSTM-CNN) は，単語ベクトルと単語のベクトルを文字単位から畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) で計算したものを結合し，BiLSTM[2] に入力し，タグを予測するモデルである．

さらに出力層で Linear Chain CRF で予測する系列タグを最適化するのが BiLSTM-CNN-CRF である．本研究における系列ラベリングモデルは BiLSTM-CNN をベースラインとした．

BiLSTM は，前方向の系列の入力と後ろ方向の入力を LSTM に入力し，それぞれの入力を結合したものを用いる．そのため，各単語の前後の文脈の情報を用いることができる．BiLSTM の入力には，単語ベクトル r_{word} と文字レベルベクトル r_{char} を結合したものを入力とする．単語ベクトル r_{word} は，入力単語 x_i に対応する事前訓練済の単語ベクトルで，文字レベルベクトル r_{char} は，入力単語 x_i の文字列 $x_{char} = (c_1, c_2, \ldots, c_n)$ を CNN に入力として与えた固定次元のベクトルである．本研究で用いる CNN，BiLSTM については，Ma ら [5] の論文を参考にされたい．

$$r_{char} = CNN_{char}(x_{char}) \quad (1)$$

$$h = BiLSTM(r_{word}, r_{char}) \quad (2)$$

3.2 ラティス探索モデル

ラティスの構築

系列タグを予測するように訓練されたモデルの出力層における各系列タグの確率分布に対して，閾値 T 以上の系列タグ集合をラティス構築のための系列タグの候補タグ集合とする．閾値 T はモデルのハイパーパラ

メータである．閾値は, ラティス上に正解ノードが含まれる割合 (Oracle 値) を開発データで計測し, 決定する．閾値の決定については 4.3 章で詳しく述べる．

ラティス上のノードベクトルの計算

ラティス上のノード x_{node} は, 固有表現の単語列 $\boldsymbol{x_{node}} = (w_1, w_2, \ldots, w_n)$ である．w_n はノードに含まれる単語である．ラティス探索の素性としてニューラルネットワークを用いるため, 可変長である単語列を固定次元に写像する必要がある．本研究では, ノードベクトル r_{node} の計算に CNN を用いた．

$$r_{node} = CNN_{node}(x_{node}) \quad (3)$$

ラティスの探索

本節では，ラティスの探索手法について説明する．本研究では，ラティスの探索に 2 つの手法を用いた．それぞれ，Andor ら [1] の手法である Global Normalization と，ラティスレベルで Linear CRF で最適化する手法の 2 つを試した．

Global Normalization

$$h_1 = ReLU(W_1 r_{node} + b_1) \quad (4)$$
$$h_2 = ReLU(W_2 h_1 + b_2) \quad (5)$$
$$score(r_{node}) = W_3[r_{node}; h_1; h_2] + b_3 \quad (6)$$
$$p_G(d_{1:n}) = \frac{\sum_{j=1}^{n} score(d_{1:j-1}, d_j)}{Z_G} \quad (7)$$

$W_{1,2,3}$, $b_{1,2,3}$ はニューラルネットワークの重みパラメータ，活性化関数は ReLU を用いた． $d_{i:j}$ は可能な系列である．p_G は Global Normalization におけるスコア関数である．Z_G は全ての可能な系列の集合の総和である．

Andor ら [1] はビームサーチで最適化を行った．最適化する損失関数は以下の通りである．

$$L_{global-beam}(d^*_{1:j}) = -\sum_{i=1}^{j} p(d^*_{1:i-1}, d^*_i) \quad (8)$$
$$+ \ln \sum_{d_{1:j} \in B_j} \exp \sum_{i=1}^{j} p(d_{1:i-1}, d_i) \quad (9)$$

B_j はビームサーチ幅内にある系列集合である．d^* は正解系列を示す．

Lattice CRF　ラティスを探索する手法としてラティス単位で Linear CRF で最適化する方法を試した．本研究における実験では，Global Normalization に比べ Lattice CRF 方が高い精度を得ることができた．

4　評価実験

4.1　実験設定

本研究では，CoNLL 2003 のコーパスを用いて提案モデルの評価を行った．評価は，適合率と再現率の調和平均である F 値によって評価をした．

系列ラベリングモデルは BiLSTM-CNN をベースラインとした．BiLSTM-CNN の出力層に基いてラティスを構築し，ラティス探索をすることで精度向上が見られるか検証を行った．

コーパス

本研究では, 英語の固有表現認識で広く使われている CoNLL 2003 のコーパスを使用した．CoNLL 2003 コーパスでは, 人名, 地名, 組織名, その他 (PERSON, LOCATION, ORGANIZATION, MISC) の固有表現タイプを含んでいる．訓練・開発・評価データの文数はそれぞれ, 38,219， 5,527， 5,462 である．系列タグは，先行研究で高い性能が報告されている BIOES を利用した [5]．テキストデータに対する前処理は，単語ベクトルに対しては小文字化を行い，文字単位ベクトルに対しては何も前処理を行わなかった．これは先行研究同様に，人手による素性選択をする必要がないという利点がある．

4.2　モデルパラメータの学習

モデルパラメータに関しては，先行研究 [5] のパラメータを利用した．単語ベクトルは，事前学習済 [1] の 100 次元の GloVe[7] を初期値に設定した．文字単位の CNN における文字ベクトルは，各ユニット次元 dim に対して，$[-\sqrt{\frac{3}{dim}}, \sqrt{\frac{3}{dim}}]$ を範囲とする一様分布からの乱数に従って初期化を行った．CNN，BiLSTM の重みパラメータは，$[-0.08, 0.08]$ を範囲とする一様分布からの乱数に従って初期化を行 バイアス項は 0 とした．文字単位の CNN は，窓枠は k=3，フィルターサイズは 30 に設定した．BiLSTM の隠れ次元は 200 とし，入力ベクトルと出力層に対して，Dropout(rate=0.5)

[1] http://nlp.stanford.edu/projects/glove/

閾値	Oracle 値 訓練データ	開発データ	評価データ
T=0.05	0.9992	0.9968	0.9909
T=0.01	0.9996	0.9980	0.9938
T=0.001	0.9998	0.9993	0.9964
T=0.0001	0.9999	0.9997	0.9981
T=0.00001	0.9999	0.9998	0.9993

表 1: CoNLL 2003 における閾値 T と Oracle 値

を適応し訓練を行った．各パラメータの学習は，確率的勾配法 (学習率=0.015, momentum=0.9) を用い，ミニバッチサイズは 10 とした．勾配ノルムは 5 でクリップした．

4.3 ラティス構築のための閾値の選択

本研究における系列ラベリングモデルは BiLSTM-CNN をベースラインとした．BiLSTM-CNN の Softmax 層の各系列タグの確率分布に対して，閾値 T 以上の系列タグ集合をラティス構築のための系列タグの候補タグ集合とする．閾値は，大きい値を取ることでラティス上のノード数を減らすことができるため，探索空間を減らすことができるが，正解ノードがラティス上から消えてしまう可能性が高まる．また閾値に小さい値を取ることで，正解ノードを保持することができるが，探索空間が広くなってしまう．

そこで閾値の決定のために，閾値と正解ノードが含まれる割合 (Oracle 値) について表 1 に示す．本研究における実験では閾値 $T = 0.00001$ を用いた．

5 結果・考察

実験の結果を表 2 に示す．先行研究の BiLSTM-CNN をベースラインとする．BiLSTM-CNN の F 値 89.72 に対して，Global Normalization は 0.3 ポイント F 値が下回った．Global+Pretrain は，事前学習 [1] を行った後に Global Normalization で最適化したモデルある．事前学習をすることで，ベースラインに比べ，F 値が 0.59 ポイント上回った．ラティスに対して CRF で最適化するモデルである Lattice CRF が，最も高い F 値 91.26 を得ることができた．現在の最高精度として報告されている BiLSTM-CNN-CRF[5] の F 値 91.21 に対して 0.05 ポイント上回る F 値を得ることができた．

	適合率	再現率	F 値
BiLSTM-CNN	89.04	90.40	89.72
Global	88.74	90.19	89.46
Global+Pretrain	90.24	90.38	90.31
Lattice CRF	90.98	91.53	**91.26**
BiLSTM-CNN-CRF [5]	91.35	91.06	**91.21**

表 2: CoNLL 2003 NER に対する実験結果

6 おわりに

本研究では，系列タグを予測するニューラルネットワークの出力層に基いてラティスを構築し，ラティス上の最適なパスを探索するモデルを提案した．評価実験を通して，提案手法が，CoNLL 2003 コーパスに対する最高精度 91.26 を達成した．今後はラティス上のノードに対して辞書素性を加え，さらなる精度向上を目指す予定である．

参考文献

[1] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins. Globally normalized transition-based neural networks. *ACL*, 2016.

[2] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith. Transition-based dependency parsing with stack long short-term memory. In *Proc. ACL*, 2015.

[3] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to japanese morphological analysis. In *EMNLP*, volume 4, pages 230–237, 2004.

[4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

[5] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th ACL*, pages 1064–1074, Berlin, Germany, August 2016. ACL.

[6] T. Müller, H. Schmid, and H. Schütze. Efficient higher-order crfs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, 2013.

[7] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[8] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. ACL, 2003.