RNNにより高次の依存を考慮したニューラル隠れマルコフモデル

平岡 達也[1], 高瀬 翔[1], 内海 慶[2], 欅 惇志[2], 岡崎 直観[1]

[1] 東京工業大学
[2] デンソーアイティーラボラトリ
[1]{tatsuya.hiraoka, sho.takase, okazaki}@nlp.c.titech.ac.jp
[2]{kuchiumi, akeyaki}@d-itlab.co.jp

1 はじめに

自然言語などの系列データを用いた問題を解く上で，品詞や話題遷移などの隠れ状態を的確に捉え活用することが重要である．アノテーションが大量にある場合は，隠れ状態そのものを教師信号として活用できる [5, 1, 2]．一方でデータが十分に無い場合は，生データから教師なし学習で隠れ状態を獲得する必要がある [7, 11, 9].

隠れ状態を推定する方法としては，隠れマルコフモデル (HMM) が有名である．近年では HMM の遷移確率と出力確率の計算にニューラルネットワークを用いることで，教師なし品詞タグ付けなどのタスクで性能が向上すると報告されている [10, 4].

これまでのニューラル HMM(NHMM) は単純マルコフ過程をベースとしており，遷移確率の計算に用いる隠れ状態は直前のもののみである．しかし，品詞などの自然言語の隠れ状態には高次の依存があると考えられる．こうした隠れ状態間の高次依存を考慮することで，より的確に隠れ状態の推定ができると期待される．

先行研究 [10] では，LSTM で観測系列をエンコードすることで高次の依存を間接的に考慮している．本研究ではこれに加え，可能なあらゆる隠れ状態の系列を RNN でエンコードし，隠れ状態間の高次依存を直接的に考慮する手法を提案する．

単純には全ての隠れ状態の系列に対応する RNN の計算結果を保持する必要があるが，一文に対する隠れ状態の組み合わせ数は膨大であり，現実的な手法ではない．これを解決するためには，より確からしい隠れ状態の遷移の情報に重きを置いた RNN のセルの計算を行えば良い．そこで提案手法では，遷移確率を計算するためのスコア行列を用いてセルに重み付けを行う．これにより，RNN のセルに隠れ状態間の高次依存関係が反映されると仮定する．

提案手法は教師なし品詞タグ付けで評価する．さらに，提案手法が隠れ状態の高次依存を適切に捉えられることを確認するため，人工データでの実験を行う．

2 ニューラル隠れマルコフモデル

本研究では既存の NHMM をベースとした拡張を行う．本章では Tran らにより提案されたニューラルネットワークを用いた HMM[10] について説明する．

長さ T の観測系列 $\mathbf{x} = x_1...x_t...x_T$ がそれぞれの時刻に対応する隠れ状態の系列 $\mathbf{z} = z_1...z_t...z_T$ から出力されるとする．HMM では，観測系列と隠れ状態の系列の同時確率を次のように求める．

$$p(\mathbf{x}, \mathbf{z}) = (\prod_{t=1}^{T} p(x_t|z_t))(\prod_{t=2}^{T} p(z_t|z_{t-1})) \quad (1)$$

ここで $p(x_t|z_t)$ は時刻 t に z_t から x_t が出力される出力確率，$p(z_t|z_{t-1})$ は z_{t-1} から z_t への遷移確率である．

NHMM では確率の計算にニューラルネットワークを用い，出力確率 $p(x_t|z_t)$ は次のように求める．

$$p(x_t|z_t) = (\mathrm{softmax}(f(\boldsymbol{v}_{z_t}^{\mathrm{emis}}))_{x_t} \quad (2)$$

$\boldsymbol{v}_{z_t}^{\mathrm{emis}}$ は出力確率計算のための z_t のベクトル，$f(\cdot)$ はベクトルを各単語に対応するスコアベクトルに変換するニューラルネットワークである．さらに $(\cdot)_{x_t}$ はベクトルから x_t に対応する値を抜き出す操作を表す．

遷移確率 $p(z_t|z_{t-1})$ は，クエリベクトル $\boldsymbol{q}_{t-1}^{\mathrm{obs}}$ から次のように計算した遷移スコア行列 $\boldsymbol{M}_{t-1}^{\mathrm{obs}} \in \mathbb{R}^{Z \times Z}$ に対して，行ごとの softmax 計算を適用することで求める．ここで Z は隠れ状態数である．

$$\boldsymbol{M}_{t-1}^{\mathrm{obs}} = \boldsymbol{W}^{\mathrm{obs}} \boldsymbol{q}_{t-1}^{\mathrm{obs}} + \boldsymbol{b}^{\mathrm{obs}} \quad (3)$$

$$p(z_t = i|z_{t-1} = j) = (\mathrm{softmax}(\boldsymbol{m}_{t-1,j}^{\mathrm{obs}}))_i \quad (4)$$

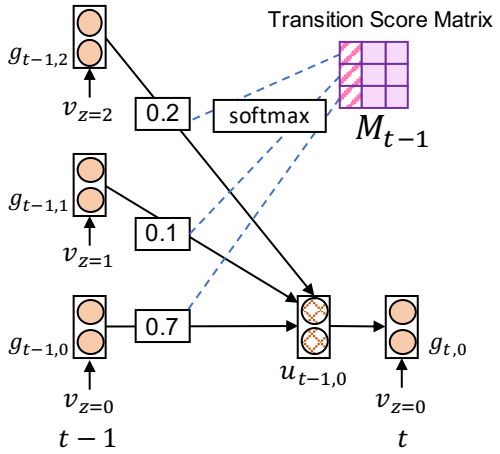

図 1: 提案手法で拡張した RNN の計算の概要．時刻 $t-1$ で入力された隠れ状態のベクトル $\boldsymbol{v}_z$ から計算されたセルの重み付き和が次の時刻の計算に使用される．

ここで $\boldsymbol{W}^{\text{obs}}$ はテンソル，$\boldsymbol{b}^{\text{obs}}$ はバイアス行列，$\boldsymbol{m}^{\text{obs}}_{t-1,j}$ は遷移スコア行列 $\boldsymbol{M}^{\text{obs}}_{t-1}$ の j 行目を示す．既存研究 [10] では $\boldsymbol{q}^{\text{obs}}_{t-1}$ として，観測系列を LSTM でエンコードしたものを用いる．すなわち，直前の隠れ状態が $z_{t-1}=j$ であるときの遷移スコアベクトル $\boldsymbol{m}^{\text{obs}}_{t-1,j}$ は，$\boldsymbol{M}^{\text{obs}}_{t-1}$ の j 行目として次のように求められる．

$$\boldsymbol{m}^{\text{obs}}_{t-1,j} = \boldsymbol{W}^{\text{obs}}_{j}\boldsymbol{q}^{\text{obs}}_{t-1} + \boldsymbol{b}^{\text{obs}}_{j} \tag{5}$$

ここで $\boldsymbol{W}^{\text{obs}}_{j}$ は $\boldsymbol{W}^{\text{obs}}$ の j 番目のスライス，$\boldsymbol{b}^{\text{obs}}_{j}$ は $\boldsymbol{b}^{\text{obs}}$ の j 行目である．

$p(x_t|z_t)$ と $p(z_t|z_{t-1})$ を用いて前向き・後ろ向きメッセージ α と β を計算することで事後確率 $p(z_t=i|\mathbf{x})$ を求める．また，隠れ状態の推論はビタビアルゴリズム [12] を用いて計算できる．

3 提案手法: RecNHMM

提案手法では遷移スコア行列の計算を以下のように変更し，隠れ状態の高次依存を捉えられるように拡張する．

$$\boldsymbol{M}_t = \boldsymbol{W}^{\text{score}}(\boldsymbol{M}^{\text{obs}}_t + \boldsymbol{M}^{\text{lat}}_t) + \boldsymbol{b}^{\text{score}} \tag{6}$$

ここで $\boldsymbol{W}^{\text{score}}$ と $\boldsymbol{b}^{\text{score}}$ はパラメータ，$\boldsymbol{M}^{\text{obs}}_t$ は式 (3) で計算されたスコア行列である．$\boldsymbol{M}^{\text{lat}}_t$ は各時刻までの隠れ状態の系列を RNN でエンコードして求めたスコア行列であり，本章で説明する．式 (3) と同様に，行ごとの softmax 計算を行うことで遷移確率を得る．

NHMM で隠れ状態の系列を利用する時，単純には RNN で系列をエンコードする方法が考えられる．正しい隠れ状態の系列が既知であれば，各時刻のセル $\boldsymbol{g}_t$ はエルマンネットワーク [3] で計算できる．

$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}^{\text{RNN}}(\boldsymbol{g}_{t-1} \oplus \boldsymbol{v}_{z_t}) + \boldsymbol{b}^{\text{RNN}}) \tag{7}$$

ここで $\boldsymbol{W}^{\text{RNN}}$ と $\boldsymbol{b}^{\text{RNN}}$ はパラメータ，$\boldsymbol{v}_{z_t}$ は隠れ状態 z_t のベクトル，$\oplus$ はベクトル結合を表す．また σ は活性化関数を表し，本研究では tanh 関数を用いる．

実際の HMM の計算時には，正しい隠れ状態の系列は不明である．そのため，あらゆる隠れ状態の系列を考慮してエンコードできるように RNN を修正する必要がある．素朴には全ての系列を RNN に与える方法が考えられるが，組合せ爆発のために現実的ではない．

そこで提案手法では，各隠れ状態の系列を RNN のセルに保持する方法を提案する．本研究で拡張した RNN(RecNHMM) は，各時刻に入力された隠れ状態のベクトルの数だけセルを持つ．各セルは遷移スコア行列 $\boldsymbol{M}_t$ を元に計算された値で重み付けをして次の時刻の計算に利用する．

具体的には，時刻 t の隠れ状態 $z_t=i$ に対応するセル $\boldsymbol{g}_{t-1,i}$ は次のように計算する．

$$\boldsymbol{g}_{t,i} = \sigma(\boldsymbol{W}^{\text{RNN}}(\boldsymbol{u}_{t-1,i} \oplus \boldsymbol{v}_{z_t=i}) + \boldsymbol{b}^{\text{RNN}}) \tag{8}$$

$$\boldsymbol{u}_{t-1,i} = \sum_j a^{t-1}_{j,i}\boldsymbol{g}_{t-1,j} \tag{9}$$

ここで $\boldsymbol{W}^{\text{RNN}}$ と $\boldsymbol{b}^{\text{RNN}}$ はパラメータである．また $a^{t-1}_{j,i}$ は遷移スコア行列から計算されたセルの重みである．

式 (9) では直前の時刻までのセルの重み付き和を計算する．本研究では，これにより各セルが再帰的に隠れ状態の系列を保存していると仮定する．

安定した学習のために，セルへの重みは $\sum_j a^{t-1}_{j,i} = 1$ を満たすように設定する．提案手法では $\boldsymbol{M}_{t-1}$ の各列に対して softmax 計算を行うことで，これを満たす重みを得る[1]．図 1 は時刻 t に隠れ状態 $z_t=0$ が入力されたときの式 (8，9) の計算を示している．

RNN の計算結果を用いて，各時刻のスコア行列 $\boldsymbol{M}^{\text{lat}}_t$ の j 行目は次のように求める．

$$\boldsymbol{m}^{\text{lat}}_{t,j} = \boldsymbol{W}^{\text{lat}}\boldsymbol{g}_{t,j} + \boldsymbol{b}^{\text{lat}}, \tag{10}$$

ここで $\boldsymbol{W}^{\text{lat}}$ と $\boldsymbol{b}^{\text{lat}}$ はパラメータである．

観測系列から計算したスコア行列 $\boldsymbol{M}^{\text{obs}}_t$ は各時刻の単語ベクトルを使うため，系列によって時刻ごとの

[1] $\boldsymbol{M}_{t-1}$ の各行への softmax 計算では遷移確率を計算している．

Method		de	en	es	fr	id	it	ja	ko	pt-br	sv
論文報告値	HMM [9]	45.5	59.8	60.6	60.1	49.6	51.5	59.5	51.7	59.5	42.4
	ANCHOR [9]	61.1	66.1	69.0	68.2	63.7	60.4	65.3	53.8	64.9	51.1
	+FEAT [9]	63.4	71.4	74.3	71.9	67.3	60.2	69.4	61.8	65.8	61.0
	Stratos (2019)[8]	75.4	73.1	73.1	70.4	73.6	67.4	77.9	**65.6**	70.7	67.1
著者実装	Tran et al. (2016)[10]*	71.4	64.1	69.9	73.1	72.8	67.4	81.0	58.0	68.5	62.9
	He et al. (2018) [4]	52.0	52.8	58.7	55.9	64.6	51.8	66.3	49.4	59.9	46.4
再実装	NHMM+conv (1)	71.7	66.5	71.2	68.5	61.5	59.9	80.6	59.2	67.1	58.6
	(1) + LSTM* (2)	79.8	73.8	75.1	73.3	71.2	71.2	**82.7**	63.2	**75.4**	66.3
提案手法	(2) + RecNHMM	**79.8**	**74.0**	75.4	75.5	72.4	**72.6**	81.9	63.2	75.1	66.5
	(2) + RecNHMM+emb	79.2	72.8	**76.1**	**75.6**	**73.8**	72.6	81.8	62.8	74.8	67.2
	(1) + RecNHMM+emb	78.6	71.5	75.5	74.9	73.1	72.2	81.8	62.6	74.7	**68.0**

表 1: 教師なし品詞タグ付けでの Many-to-One 指標の比較．上段は先行研究による報告値，中段は NHMM を用いた先行研究の著者実装と再現実装による結果，下段は提案手法の結果を示す．*は同じ実験設定であることを表す．

スコアが変化する．一方で提案手法によるスコア行列 $\boldsymbol{M}_t^{\mathrm{lat}}$ は観測系列を使わないため，$\boldsymbol{M}_t^{\mathrm{obs}}$ を用いない場合はスコア行列が全ての系列で同じになってしまう．$\boldsymbol{M}_t^{\mathrm{obs}}$ を使わない場合は，式 (8) の計算に観測された単語ベクトル $\boldsymbol{v}_{x_t}$ を利用することで系列ごとに異なるスコアを計算できる．

$$\boldsymbol{g}_{t,j} = \sigma(\boldsymbol{W}^{\mathrm{RNN}'}(\boldsymbol{u}_{t-1,j} \oplus \boldsymbol{e}_t) + \boldsymbol{b}^{\mathrm{RNN}}), \quad (11)$$

$$\boldsymbol{e}_t = \boldsymbol{v}_{z_t=j} \oplus \boldsymbol{v}_{x_t}, \quad (12)$$

$\boldsymbol{v}_{x_t}$ は式 (5) での計算に用いたものと共有している．

提案手法は既存手法 [10] と同様に，前向き・後ろ向きメッセージを用いた Baum-Welch アルゴリズムをベースとした学習を行う．

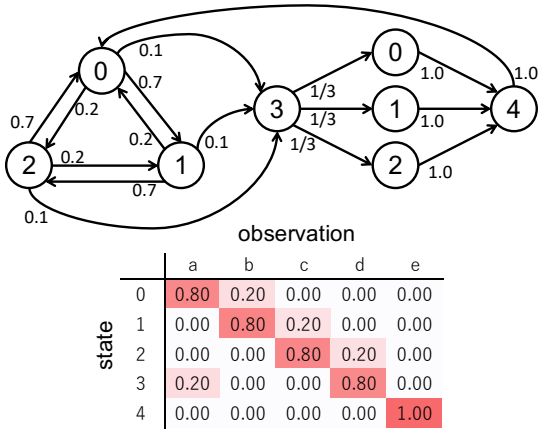

図 2: 人工データの作成に使用した遷移確率（上）と出力確率（下）．

4 実験と考察

4.1 教師なし品詞タグ付け

HMM の検証に広く使われる教師なし品詞タグ付けタスクで提案手法を評価する．実験では 12 の品詞タグが付与された Universal Treebank Version 2.0[6] を用いる．先行研究 [10] と同様に全ての数字を 0 に置き換える前処理を行い，全てのデータで学習と評価を行った．

表 1 に実験結果を示した．NHMM+conv は式 (2) の $f(\cdot)$ として，文字レベルの畳み込みニューラルネットワークを用いる手法である．さらに LSTM は観測系列をエンコードする手法を追加したモデルであり，先行研究 [10] のモデルと一致する．RecNHMM は本論文で拡張した RNN を用いたもの，embed は式 (11) で説明した RNN に単語分散表現を入力するモデルである．NHMM は初期値の差が性能に大きな影響を与えるため，シンプルな手法である NHMM+conv の学習結果を初期値として再実装と提案手法の学習を行った．

実験結果より，多くの言語で提案手法の性能が既存手法を上回り，品詞タグ付けタスクでの有効性が確認された．また，いくつかの言語では拡張した RNN に単語分散表現を入力することでさらなる性能の向上が見られた．一方で先行研究で提案された，観測系列をエンコードする LSTM を用いないモデルの性能は高くなく，LSTM の必要性が示された．

4.2 人工データを用いた実験

本研究では隠れ状態間の高次依存を捉えられるモデルの提案を目的としている．しかし，教師なし品詞タグ付けタスクだけでは実際に提案手法が高次依存を捉えられているかの確認が難しい．そこで図 2 に示したルールで人工データを作成し，提案手法が隠れ状態のトライグラム関係を捉えられているかを確認した．1 文は 20 トークンからなり，訓練，検証，評価のため

Method		M-1	V-M
Baseline	NHMM (1)	77.83	57.41
	(1) + LSTM (2)	82.95	64.93
Proposed	(2) + RecNHMM	83.18	64.82
	(2) + RecNHMM+emb	84.36	66.86
	(1) + RecNHMM+emb	**84.94**	**67.33**

表 2: 人工データにおける実験結果．評価には Many-to-One 指標 (M-1) と V-Measure(V-M) の値を用いた．

にそれぞれ 2,000，500，500 文のデータを作成した．

表 2 に示した実験結果より，提案手法はベースラインに比べて適切に隠れ状態の遷移を捉えられていることがわかる．さらに図 3 は，学習済みの既存手法と提案手法が実際に予測した遷移確率を表す．図 2 のように，隠れ状態 3 の 2 ステップ後に隠れ状態 4 に行き着く．すなわち，ここには高次の依存関係があり，図 3 ではこれに関わる遷移確率を示している．図より，提案手法は隠れ状態 3 から生成された観測値 a を手がかりとして，適切に 2 ステップ先の隠れ状態 4 を予測できていることが確認できる．

5 まとめ

本研究では，隠れ状態間の高次依存を捉えるための NHMM を提案した．実験より，提案手法は教師なし品詞タグ付けタスクの性能向上に寄与することが確認された．さらに人工データを用いた実験より，従来の手法では捉えられない隠れ状態間のトライグラム関係が学習できることを確認した．

謝辞 本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究「多言語音声翻訳高度化のためのディープラーニング技術の研究開発」により得られたものです．

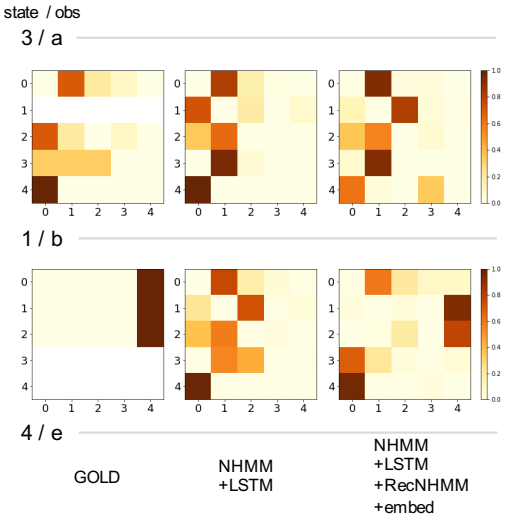

図 3: 観測された系列 “a, b, c” と，対応する隠れ状態の系列 “3, 1, 4” に対して既存手法と提案手法が予測した遷移確率の比較．

参考文献

[1] James Allan and Hema Raghavan. Using part-of-speech patterns to reduce query ambiguity. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 307–314, 2002.

[2] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, 2008.

[3] Jeffrey L Elman. Finding structure in time. *Cognitive science*, Vol. 14, No. 2, pp. 179–211, 1990.

[4] Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. Unsupervised learning of syntactic structure with invertible neural projections. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1292–1302, 2018.

[5] Eliyahu Kiperwasser and Yoav Goldberg. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, Vol. 4, No. 1, pp. 313–327, 2016.

[6] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, T Oscar, et al. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 92–97, 2013.

[7] Xi Shao, Changsheng Xu, and Mohan S Kankanhalli. Unsupervised classification of music genre using hidden markov model. In *2004 IEEE International Conference on Multimedia and Expo*, Vol. 3, pp. 2023–2026, 2004.

[8] Karl Stratos. Mutual information maximization for simple and accurate part-of-speech induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1095–1104, 2019.

[9] Karl Stratos, Michael Collins, and Daniel Hsu. Unsupervised part-of-speech tagging with anchor hidden markov models. *Transactions of the Association for Computational Linguistics*, Vol. 4, No. 1, pp. 245–257, 2016.

[10] Ke M Tran, Yonatan Bisk, Ashish Vaswani, Daniel Marcu, and Kevin Knight. Unsupervised neural hidden markov models. In *Proceedings of the Workshop on Structured Prediction for NLP*, pp. 63–71, 2016.

[11] Jurgen Van Gael, Andreas Vlachos, and Zoubin Ghahramani. The infinite hmm for unsupervised pos tagging. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pp. 678–687, 2009.

[12] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, Vol. 13, No. 2, pp. 260–269, 1967.