リカレントニューラルネット言語モデルの識別学習*

○太刀岡勇気 (三菱電機・情報総研), 渡部晋治 (MERL)

1 はじめに

ニューラルネットワークによるモデリング技術は音声認識の分野で注目を集めている。最も成功した例が、音響モデリングにおける深層神経回路網 (Deep neural network; DNN) であろう。ニューラルネットワークは近年言語処理にも使われるようになってきており、中でも、リカレントニューラルネットワーク言語モデル (Recurrent neural network based language model; RNN-LM) が、その高い性能ゆえ、よく使われる [1, 2]。RNN は再帰的な入力を持つ 1 層以上の隠れ層を持つニューラルネットワーク (NN) である。計算コストはかかるものの、RNN-LM を利用することで音声認識の性能は大きく改善する。RNN-LM と従来の n グラムモデルの最も大きな違いは、利用できる単語コンテキストの長さである。言語モデルの役割は、その単語の前の単語のコンテキストに基づいて、当該単語の事後確率を正しく計算するところにある。長いコンテキストを利用することで、豊富な情報が得られるものの、従来の n グラムモデルでただ単純に長いコンテキストを利用するだけでは (すなわち 4-gram や 5-gram を使う)、データが疎であるという問題に直面してしまう。これらの問題に対処するため、RNN-LM は初めに高次元である認識したい単語の 1-of-N 表現を、隠れ層における低次元の連続空間に写像する。そして認識したい単語の事後確率を直接的に推定する。前のフレームからの隠れ層のユニットは、次のフレームでの入力ベクトルに結び付けられている。この再帰的な入力により、低次元の隠れ層のユニットに、単語の履歴が集積されることとなる。RNN-LM では暗に単語列の履歴を全部考慮することができるのに対して、よく使われている n グラム言語モデルでは、直前の $(n-1)$ 単語の履歴しか考慮することができない。RNN-LM による事後確率の計算にはフィードフォワード伝播が必要なので、テーブルルックアップで済む n グラム言語モデルに比べると、必要な計算量は非常に多くなる。この問題を扱った研究もいくつかあるものの [3, 4, 5]、通例、RNN-LM は N-best やラティスのリスコアリングのような後処理として使われる。

RNN-LM はこのように通常の n グラムモデルに比べて高い性能を発揮する。しかしながら、RNN-LM の学習基準は、正解単語と予測単語の間のクロスエントロピー (Cross entropy; CE) 基準に基づいている。そして CE 基準では音声認識の仮説と正解から計算される識別的な基準を陽に考慮することはできない。一方で、様々な音声認識のタスクにおいて GMM に基づく音響モデルや特徴量変換の学習において識別的基準が有効であることが広く示されてきた [6, 7, 8]。さらに、本来的に高いフレームレベルでの識別性を保つはずの DNN 音響モデルにおいても識別的基準による学習により音声認識の誤りを減らすことができる [9, 10, 11]。RNN-LM の CE 基準は、履歴が与えられたうえでの対象単語の事後確率を考慮できるという意味では識別的であるが、音声認識の仮説を考慮できる RNN-LM の識別基準があれば、音声認識の誤りをよりよく訂正することができる。本報では、RNN-LM のための新しい識別学習法を提案する。

これとは異なる N-best リスコアリングの枠組みでの識別モデルとしては、識別的言語モデリング (Discriminative language modeling; DLM) が広く知られている [12]。DLM は、正解単語列と学習データを認識した音声認識仮説から得られる n グラム数に基づいてそれを正すように学習する。これにより、特に短い単語コンテキストにおいて、デコーダーに固有の誤りを修正することができる。しかしながら、DLM のコンテキストは、上述の n グラム同様に、本質的に n グラム (通常は tri-gram) に限られる。更に、データが疎であることから、長いコンテキストを使ったとしても誤りを修正する能力がそれほど向上するわけではない。提案法は、RNN-LM の枠組みに基づくことで長いコンテキストの影響を考慮しつつ、識別的基準により音声認識の仮説傾向も取り入れた学習を行う。さらに、DLM と提案の識別的 RNN-LM を組み合わせることで DLM 単体よりも性能を向上させることができる。これにより、短いコンテキスト・長いコンテキスト双方にとって有効な識別的言語モデリングが行えると考えられる。

2 RNN-LM

図 1 には、今後の実験に用いる、1 つの隠れ層を持つ RNN-LM のトポロジーを示している。前のフレームの隠れ層のユニットは入力特徴量ベクトルに再帰的に連結される。重み行列 $\boldsymbol{U}$ と $\boldsymbol{V}$ ($\triangleq \Theta$) は、学習時に推定すべきモデルパラメータである。

*Discriminative Method for Recurrent Neural Network Language Models, by TACHIOKA, Yuuki (Mitsubishi Electric Corporation), WATANABE, Shinji (MERL).

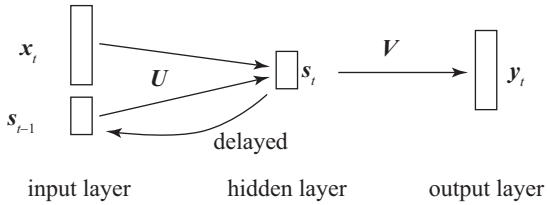

Fig. 1 Recurrent neural network language model (RNN-LM) topology. $|\mathcal{V}|$-dimensional input vector $\boldsymbol{x}_t$ is a 1-of-$|\mathcal{V}|$ representation of the t-th word of the utterance. Output vector $\boldsymbol{y}_t$ is an $|\mathcal{V}|$-dimensional posterior probability vector corresponding to input words conditioned on the previous context. The hidden layer has a low-dimensional vector $\boldsymbol{s}_t$. Hidden-layer units in the previous frame $\boldsymbol{s}_{t-1}$ are recursively concatenated to the input vector $\boldsymbol{x}_t$.

2.1 CE 学習

初めに RNN-LM を、評価関数 $\mathcal{F}^{\mathrm{CE}}$ を最小化するクロスエントロピー (CE) 基準に基づいて学習する。CE は、語彙セット $\mathcal{V}$ から予測された単語の事後確率 $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(n), \cdots, y_t(|\mathcal{V}|)]^\top$ と、正解ラベル系列 $C = \{c_t | t = 1, \cdots, T\}$ に基づいて計算される。

$$\mathcal{F}^{\mathrm{CE}}(C) = -\sum_{n=1}^{|\mathcal{V}|}\sum_{t=1}^{T} \delta(n, c_t) \log y_t(n) \qquad (1)$$

ここで c_t は、正解ラベルにおける t 番目の単語のインデックスである。$\delta(\cdot,\cdot)$ は、Kronecker のデルタ関数である。出力層はソフトマックス関数 y_t を持つ。

$$y_t(n) = \frac{\exp(a_t(n))}{\sum_{n'} \exp(a_t(n'))} \qquad (2)$$

ここで n は出力ソフトマックス層の要素の番号であり、a_t は n 番目の単語のアクティベーションである。

2.2 更新則

本節では、学習すべきパラメータ Θ の勾配法に基づく更新則について述べる。ニューラルネットの連鎖則に基づき (すなわち $\partial/\partial\Theta = \partial/\partial a_t(n) \cdot \partial a_t(n)/\partial\Theta$)、アクティベーション $a_t(n)$ についての評価関数 $\mathcal{F}^{\mathrm{CE}}$ の微分は

$$\frac{\partial \mathcal{F}^{\mathrm{CE}}}{\partial a_t(n)} = -[\delta(n, c_t) - y_t(n)] \triangleq \varepsilon_t(n) \qquad (3)$$

のようになる。$\partial/\partial a_t(n) \log y_t(n') = \delta(n, n') - y_t(n)$ であることを利用した。この等式は、正解単語と t 番目の位置における単語 n の誤りに相当する事後確率 $\varepsilon_t(n)$ の差異が、モデルパラメータ Θ の推定に伝播されることを示している。再帰的な連結があるため、これは時間方向に連結した誤差逆伝搬則 (back propagation through time) によりパラメータを最適化することとなる [1]。

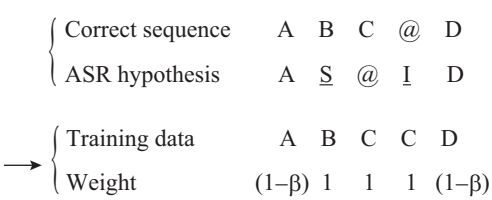

Fig. 2 Weight discount procedure of the proposed method. The weight of training data is discounted (i.e., $1-\beta$) for the correct data. A, B, C, and D are words, @ is a NULL token that follows the alignments of a correct word sequence and ASR hypothesis are fixed. S denotes a substitution and I denotes an insertion error. For insertion, repeated entry of the previous frame is used.

3 RNN-LM の識別学習

3.1 RNN-LM の識別的基準

RNN-LM に識別学習を導入するために、単語単位の尤度比の評価関数 $\mathcal{F}^{\mathrm{LR}}$ から始めることとする [1]:

$$\mathcal{F}^{\mathrm{LR}}(C, H) = -\sum_{t} \log \frac{y_t(c_t)}{y_t(h_t)^\beta} \qquad (4)$$

ここで h_t は、 正解単語系列 C と 1 位の音声認識の仮説系列 $H = \{h_t | t = 1, \cdots, T\}$ で整列した 1 位の音声認識仮説の t 番目の単語番号である。β はスケーリング係数であり、この係数の意味に関しては後に議論する。この対数尤度比は、識別的基準の性質を持つことに注意されたい。なんとなれば、$\mathcal{F}^{\mathrm{LR}}(C, H)$ を最小化することは、 誤認識した h_t を正解単語 c_t に正すことに相当するからである。

式 (4) は以下のように書き換えることもできる。

$$\begin{aligned}\mathcal{F}^{\mathrm{LR}}(C, H) &= -\sum_{n}\sum_{t}\Big[\delta(n, c_t)\log y_t(n) \\ &\qquad - \beta\delta(n, h_t)\log y_t(n)\Big] \\ &= \mathcal{F}^{\mathrm{CE}}(C) - \beta\mathcal{F}^{\mathrm{CE}}(H).\end{aligned} \qquad (5)$$

それゆえ、式 (4) は、正解ラベルと音声認識仮説の CE の重み付き誤差であると解釈することもできる。

3.2 更新則

提案法では、式 (3) に対応する更新則は、同じく式 (5) を微分することで、

$$\frac{\partial \mathcal{F}^{\mathrm{LR}}(C, H)}{\partial a_t(n)} = -[\delta(n, c_t) - \beta\delta(n, h_t) - (1-\beta)y_t(n)] \qquad (6)$$

[1] これは系列の識別学習ではなく、正解単語と音声認識仮説の間の整列結果に基づく単語単位の識別学習である。

のように得られる。我々の実装では簡単のため、$(1-\beta)y_t(n)$ を $y_t(n)$ で近似することで、更新則は

$$\frac{\partial \mathcal{F}^{\mathrm{LR}}(C,H)}{\partial a_t(n)} \approx -[\delta(n,c_t) - \beta\delta(n,h_t) - y_t(c_t)] \quad (7)$$

のように得られる。図 2 には、提案法の重みを割り引く具体例を示す。第 1 に、動的計画法により、正解単語系列と音声認識の単語系列を整列させる。第 2 に、正解ラベルに対する重みを割り引き (すなわち $1-\beta$)、割り引かれた重みに対してモデルを再学習する。対象の単語に対する重みが負になることをさけるため、$\delta(n,c_t) - \beta\delta(n,h_t)$ が負となる際には $\delta(n,c_t) - \beta\delta(n,h_t) = 0$ と仮定していることに注意されたい。

3.3 元の CE モデルとの平滑化

最後に、RNN-LM は提案の識別的手法によって求められたモデル $\boldsymbol{U}^{\mathrm{LR}}, \boldsymbol{V}^{\mathrm{LR}}$ と元の CE モデル $\boldsymbol{U}^{\mathrm{CE}}, \boldsymbol{V}^{\mathrm{CE}}$ のパラメータを平滑化して、以下のように求められる。

$$\{\boldsymbol{U},\boldsymbol{V}\} \leftarrow \tau\{\boldsymbol{U}^{\mathrm{CE}},\boldsymbol{V}^{\mathrm{CE}}\} + (1-\tau)\{\boldsymbol{U}^{\mathrm{LR}},\boldsymbol{V}^{\mathrm{LR}}\} \quad (8)$$

ここで、τ は平滑化係数である。

4 実験

4.1 実験の設定

日本語話し言葉コーパス (Corpus of Spontaneous Japanese; CSJ) [13] を用いて、提案法の有効性の検証を行った。語彙サイズは 70k である。3 種類のテストセットがあり、それぞれは 10 話者による講演や模擬講演からなる。テストセット E1、E2 と E3 はそれぞれ、22,682、23,226 と 14,896 単語からなる。

音響モデルは DNN-HMM であり、CE 学習により構築した。音響特徴量には、23 次元のメルフィルタバンク特徴量とその 1 次 2 次の動的特徴量を用いた。コンテキスト依存 HMM の状態数は 3,500 であり、DNN は 7 層の隠れ層を持ちそれぞれの層は 2,048 ノードからなる。この設定は、文献 [14] での設定を参考にした。初期学習率は 0.01 であり、学習の終盤にむけて 0.001 まで低減させた。CE 学習された DNN 音響モデルが得られた後、DNN のための boosted MMI 識別学習 [11] を行った。DNN 学習のためには、Kaldi ツールキット [15] の Povey による実装を使った。

元の言語モデルのサイズは 70k であるものの、RNN-LM の言語モデルのサイズは頻出 10k 語に絞った。このサイズが RNN-LM の入力層の次元数に一致する (すなわち $|\mathcal{V}|$)。隠れ層のユニット数は 30 である。RNN-LM の学習率 η は、0.1 もしくは 0.05 とした。RNN-LM は、RNN-LM ツールキット [16] により構築した。言語モデルスコアは RNN-LM のスコアと元の tri-gram 言語モデルのスコアを線形に内挿して求めた。内挿の重みは 0.5 とし、各発話につき上位 100 位までの仮説をリスコアリングに使った。RNN-LM と提案の識別的 RNN-LM を、DLM と組み合わせる実験も行った。

Table 1 WER [%] on CSJ using a DNN acoustic model with a conventional n-gram and discriminative language model (DLM).

	E1	E2	E3	Avg.
baseline	12.81	10.64	11.13	11.53
+DLM	12.60	10.52	10.82	11.31

Table 2 WER [%] on CSJ using a DNN acoustic model with RNN-LM-based and DLM-based rescoring.

	E1	E2	E3	Avg.
+RNN-LM	11.97	10.18	10.51	10.89
+RNN-LM+DLM	11.74	9.98	10.03	10.58

Table 3 WER [%] on CSJ with the proposed discriminative RNN-LM (d-RNN-LM).

β	τ	η	E1	E2	E3	Avg.
0.05	0.85	0.1	11.99	10.19	10.50	10.89
		0.05	**11.84**	10.07	10.61	10.84
	0.9	0.1	11.91	**10.02**	10.51	10.81
		0.05	**11.84**	10.03	10.49	10.79
0.10	0.85	0.1	12.20	10.45	10.69	11.11
		0.05	11.86	10.09	10.47	10.81
	0.9	0.1	11.93	10.19	10.41	10.84
		0.05	11.90	10.04	**10.39**	**10.78**
0.15	0.85	0.1	12.06	10.38	10.49	10.98
		0.05	11.93	10.09	10.40	10.81
	0.9	0.1	11.98	10.17	**10.39**	10.85
		0.05	11.98	10.03	**10.39**	10.80

4.2 ベースラインの結果

表 1 には、 識別学習を行った DNN 音響モデルによるベースラインの結果を示す。これは CSJ コーパスに対して、よい性能であるということができる [14]。DLM によるリスコアリングにより、単語誤り率 (Word error rate; WER) は平均で 0.22%改善した。

この高いベースラインに対して、RNN-LM によるリスコアリングは顕著に WER を改善し、表 2 に示す通り、平均で 0.64%の改善が見られた。RNN-LM に加えて、この結果からわかるように、DLM は依然として有効であり、ここから識別的モデルの効果を知ることができる。

Table 4 WER [%] on CSJ with the proposed discriminative RNN-LM (d-RNN-LM) and DLM rescoring.

β	τ	η	E1	E2	E3	Avg.
0.05	0.85	0.1	12.00	10.20	10.51	10.90
		0.05	11.68	9.98	10.04	10.57
	0.9	0.1	11.72	10.01	10.04	10.59
		0.05	11.63	**9.90**	10.05	10.53
0.10	0.85	0.1	12.07	10.19	10.70	10.99
		0.05	11.75	10.03	10.28	10.69
	0.9	0.1	11.77	10.03	10.12	10.64
		0.05	11.64	9.94	10.08	10.55
0.15	0.85	0.1	11.81	10.07	10.26	10.71
		0.05	11.63	10.00	10.14	10.59
	0.9	0.1	11.61	9.95	10.01	10.52
		0.05	**11.60**	9.95	**9.99**	**10.51**

4.3 提案法

表 3 には提案の識別的 RNN-LM (d-RNN-LM) の性能を示す。提案法には 3 つのパラメータが存在するため、パラメトリックスタディを行った。ほとんどすべての場合で、提案法の平均の WER は表 2 に示す RNN-LM の WER を上回っている。ここから、パラメータの調整はそれほど難しくないことがわかる。表 4 には提案法と DLM を併用した場合、DLM が有効であることがわかる。これは n-gram モデルによる短いコンテキストの明示的な利用は提案法による短いコンテキストの間接的な利用に比べて、効果が大きいためと考えられる。

全体的には、実験における提案法の性能改善はそれほど大きくないが、この設定に対するベースラインが相当高性能であることも影響している。より誤りが多くなるタスクにおいて、提案法により RNN-LM のモデル推定がより頑健なものになると考えられる。

5 まとめと今後の課題

本報では、RNN-LM の識別学習法を提案した。正解ラベルに対する CE 学習に比べて、音声認識仮説に対する識別学習を提案した。提案の識別学習は、音響モデルの識別学習と同様に正解ラベルと音声認識結果の事後確率の差分統計量を用いる部分が CE 学習と異なっている。実験の結果、大語彙音声認識タスクにおいて提案法により性能が改善することが示された。短いコンテキストと長いコンテキストを暗に扱うことのできる提案の識別的 RNN-LM と、短いコンテキストだけを明示的に扱うことのできる DLM を組わせることでさらに性能が向上した。これはこれら 2 つのモデルの性質が異なることで、それぞれのモデルを補い合ったためと考えられる。今後の課題は、系列の識別学習の検討と N ベスト仮説を学習に取り入れることである。

参考文献

[1] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, “Recurrent neural network based language model,” Proceedings of INTERSPEECH, pp.1045–1048 (2010).

[2] M. Sundermeyer, I. Oparin, J.-L. Gauvain, B. Freiberg, R. Schlüter, and H. Ney, “Comparison of feedforward and recurrent neural network language models,” Proceedings of ICASSP, pp.8430–8434 (2013).

[3] Y. Shi, W.-Q. Zhang, M. Cai, and J. Liu, “Efficient one-pass decoding with NNLM for speech recognition,” IEEE Signal Processing Letters, **21**, 377–381 (2014).

[4] T. Hori, Y. Kubo, and A. Nakamura, “Real-time one-pass decoding with recurrent neural network language model for speech recognition,” Proceedings of ICASSP, pp.6414–6418 (2014).

[5] Z. Huang, G. Zweig, and B. Dumoulin, “Cache based recurrent neural network language model inference for first pass speech recognition,” Proceedings of ICASSP, pp.6404–6407 (2014).

[6] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, “Boosted MMI for model and feature-space discriminative training,” Proceedings of ICASSP, pp.4057–4060 (2008).

[7] M.J.F. Gales, S. Watanabe, and E. Fosler-Lussier, “Structured discriminative models for speech recognition: An overview,” IEEE Signal Processing Magazine, **29**, 70–81, Nov. 2012.

[8] Y. Tachioka, S. Watanabe, and J. Hershey, “Effectiveness of discriminative training and feature transformation for reverberated and noisy speech,” Proceedings of ICASSP, pp.6935–6939, May 2013.

[9] G. Wang and K. Sim, “Sequential classification criteria for NNs in automatic speech recognition,” Proceedings of INTERSPEECH, pp.441–444, Aug. 2011.

[10] B. Kingsbury, T. Sainath, and H. Soltau, “Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization,” Proceedings of INTERSPEECH, pp.485–488, Sept. 2012.

[11] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” Proceedings of INTERSPEECH, Aug. 2013.

[12] B. Roark, M. Saraçlar, M. Collins, and M. Johnson, “Discriminative language modeling with conditional random fields and the perceptron algorithm,” Proceedings of ACL, pp.47–54 (2004).

[13] S. Furui, K. Maekawa, and H. Isahara, “A Japanese national project on spontaneous speech corpus and processing technology,” Proceedings of ASR, pp.244–248 (2000).

[14] N. Kanda, R. Takeda, and Y. Obuchi, “Elastic spectral distortion for lowresource speech recognition with deep neural networks,” Proceedings of ASRU, pp.309–314, Dec. 2013.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, “The Kaldi speech recognition toolkit,” Proceedings of ASRU, pp.1–4 (2011).

[16] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, J. Černocký, and S. Khudanpur, “RNNLM–recurrent neural network language modeling toolkit,” Proceedings of ASRU, pp.1–4 (2011).