暗黙の発話状況を考慮したニューラル対話モデル

佐藤 翔悦 [*1]　吉永 直樹 [*2]　豊田 正史 [*2]　喜連川 優 [*2*3]
[*1] 東京大学大学院 情報理工学系研究科
[*2] 東京大学 生産技術研究所　[*3] 国立情報学研究所
{shoetsu, ynaga, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

1　はじめに

我々は他人との会話の中で，時間・場所や話題，人間関係など多様な発話状況（ドメイン）を暗黙に了解して発言を行う．例えば，昼休みになればランチの話題をするだろうし，会話の途中で話題を変えたければ「ところで」と断るだろう．また家族と職場の人とでは話し方を変える．このように，会話においては発言の背後にある様々な要因によって適切な応答は変化する．

我々が日常行う雑談を対象とした対話システムでは，想定される発話状況が多様であることから，タスク指向型対話システムのように特定の発話状況を仮定し，パターンを作り込むアプローチが取りづらい．そのため，マイクロブログなど多様な発話状況を含む対話コーパスを利用した統計的対話モデルが有望であると考えられる [1][2]．しかし，これらの既存研究においても発話状況を陽に考慮したものは少なく，対話データが持つ発話状況を把握し，その違いを踏まえて応答を行うことは難しい．

これに対し我々は発話内容に基づくクラスタリングによって発話状況を自動的に認識した上で，発話状況依存の対話モデルを学習・運用する手法を提案した [3]．しかしこの手法では，発話状況を考慮することで，対話モデルの学習データの一貫性が向上する一方，そのサイズが小さくなるという致命的な問題が存在する．

そこで本研究では，汎的な対話モデルと発話状況に依存する対話モデルを統合し，これを同時に学習・運用するニューラル対話モデルを提案する．提案手法は，発話状況を自動付与した発話を入力として，発話状況を考慮して適切な応答を返すことが可能である．具体的な発話状況については，発話内容から得られる発話状況に加えて，季節のような外部環境に依存した発話状況の利用も検討する．

実験では，Twitter から収集した対話データを用いて提案手法の有効性を検証する．応答選択テストを用いた評価により，提案手法の有効性を確認した．

2　関連研究

本研究のように，発話状況（あるいはドメイン）が陽に与えられていない状況で汎用な発話状況を考慮して対話モデルの学習・テストを行う研究は我々の知る限り存在しないが，関連する研究としてドメイン適応の研究があるので，以下で紹介する．

Hasegawa らは発話が聞き手に対し喚起させる感情の種類に着目し，人手で作成した少数の規則によって Twitter から取得した大規模な対話データを怒り，喜び，悲しみなどといった 9 つのカテゴリに分類した感情タグ付き対話コーパスを構築している [4]．実験では，そのコーパスから対話モデルを学習することで特定の感情を喚起するような応答の生成を行った．

Li らは統計的手法により学習した対話システムが学習データ中に不特定多数の人間の発話・応答を含むために，例えば「出身はどこですか？」と尋ねられた時は「東京です」と答え,「どこ出身？」と尋ねられた場合は「大阪」と答えるといったように，一貫した応答を返さないことを問題とした．彼らの手法では応答者をいくつかのタイプに分類した上で user embedding として Decoder Recurrent Neural Network (RNN) に与えることで，応答者のタイプごとに一貫性のある応答を得ることに成功している [5]．

また，我々が採用する対話モデル [6] と同様の機構を用いたニューラル機械翻訳において，Johnson らは多国語の機械翻訳において翻訳先，あるいは翻訳元の言語を表す ID を入力の先頭に加える，というシンプルな方法で複数言語間での翻訳を単一のモデル内で可能にした [7]．

本稿では，同じニューラルネットワークに基づく Li らと Johnson らの手法と提案手法を比較し，提案手法の優位性を検証する．

3 提案手法

我々の提案する対話モデルは Sutskever らの SEQ2SEQ モデル [6] をベースに，発話に自動付与された発話状況 (ドメイン) を学習・テスト時に考慮する機構を導入する．以下で，SEQ2SEQ について簡単に導入した後，提案手法，及び考慮する発話状況について述べる．

3.1 SEQ2SEQ 対話モデル

SEQ2SEQ 対話モデルは発話を入力とし，まず Encoder と呼ばれる RNN によって発話の単語列を順次読み込み，発話内容を表現する実数値ベクトルに変換する．続いて，得られた実数値ベクトルを Decoder と呼ばれる RNN の初期状態とし，RNN の内部状態に基づいて単語を一単語ずつ入出力する．本研究では RNN として Long-Short Term Memory (LSTM) [8] を採用し，応答選択テストの際は発話・応答ペアに対するソフトマックス損失を擬似的なスコアとして用いて応答候補の順位付けを行う．

3.2 発話状況を考慮した対話モデル

発話状況を考慮したニューラル対話モデルを訓練するにあたって，発話以外の情報をモデルで考慮する方法としては (1) 発話状況に応じて一部のパラメタを独立に訓練するか (2) 発話状況を特徴量として外部から与えるか [5, 7] の 2 通りが考えられるが，本研究では前者に基づく Local/Global SEQ2SEQ を提案する．

提案モデルでは Encoder, Decoder 共に 2 種類の RNN を同時に訓練する．1 つは発話状況によらず共通して用いる Global-RNN，もう 1 つは発話状況ごとに独立な Local-RNN であり，その 2 つの出力を平均することで全体の出力を得る (図 1)．また Local-RNN のみを用いたモデルでの予備実験も行ったが[1]，Global-RNN を加えた場合と比べいずれも低い性能しか得られなかったため，第 4 節では後者の結果のみを記す．

3.3 暗黙の発話状況

対話に付随する発話状況としては様々なものがあるが，本稿では発話内容に基づく内的な発話状況と，発話時間・場所や発話者間の人間関係などに基づく外的な発話状況に分けて，それぞれ用いることを検討する．提案モデルは発話状況が所与であると仮定するため，それぞれについて教師無しで（安価に）得られる発話状況として，以下を用いる．

content　内的な発話状況については既存研究 [3] に倣って発話内容を実数値ベクトルで表現し，それをクラスタリングすることで得る．具体的には，対話モデルの訓練データを用いて word2vec[2] で単語のベクトル表現を得た後，発話中の単語のベクトル表現を平均することで発話のベクトル表現を得る．得られた発話ベクトル表現を k-means クラスタリングを用いて分割・分類し，得られたクラスタを発話状況とみなす．実験では，発話状況の分割数 k は 10 に固定した．

season/month　外的な発話状況については，今回のデータセットにおいて自明に利用できるタイムスタンプを用いる．具体的には発話のタイムスタンプを元に 1-12 月の 12 種類（月ごと）への分割と，3-5 月, 6-8 月, 9-11 月, 12-2 月の 4 種類（春夏秋冬）への分割を行った．分割数ごとに季節単位で分けたものを season, 月単位で分けたものを month と呼ぶ．本稿では人間関係や場所のような，現在用いるマイクロブログの対話データでは陽に与えられず，推定が必要な発話状況の利用については今後の課題とする．

4 実験

提案手法の有効性を検証するため，応答選択タスクによって手法の評価を行う．

4.1 設定

学習・テストデータ　学習・評価のためのデータセットとしては，我々の研究室において 2011 年 3 月より継続的に収集している Twitter のデータセットを元に構築した[3]．具体的には，各投稿（ツイート）を発話とみなし，あるツイートとそれに対するリプライを 1 組の発話・応答ペアとした上で，学習データとして 2014 年から約 23,563,865 組，テストデータとして 2015 年の各月から 500 組ずつ合計 6,000 組を抽出し，それに加えてダミー応答候補として同年のツイートからランダムに 114,000 応答を抽出した．続いて MeCab[4] (辞

[1] embedding/softmax layer を共有しているという違いはあるが，[3] とほぼ同一のモデル

[2] https://code.google.com/p/word2vec/

[3] 2011 年 3 月に 30 名程度の著名な日本人ユーザを選択し，そのユーザのタイムラインを公式 API で継続的に収集するとともに，それらのユーザがメンション・リツイートを行ったユーザのタイムラインも収集対象として追加することで拡大していった．

[4] http://taku910.github.io/mecab/

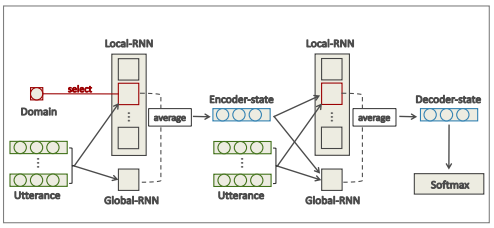

図 1: Local/Global SEQ2SEQ

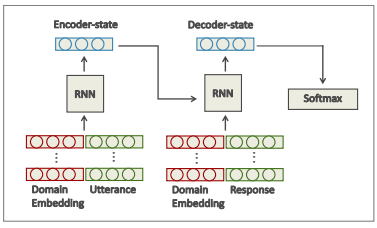

図 2: Domain-Embedding

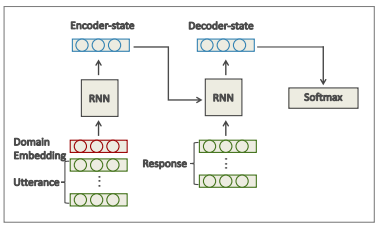

図 3: Domain-Header

表 1: モデルと注目した発話状況ごとの 1 in t P@k．モデル全体で最も高い精度のものを太字，同モデル内で最も高い精度のものをイタリックで各評価尺度ごとに記す．

Model	Domain	1 in 2P@1	1 in 5P@1	1 in 5P@2	1 in 20P@3
Baseline	-	66.7%	37.8%	60.1%	32.7%
Local/Global SEQ2SEQ	season	**70.2%**	**39.9%**	**64.2%**	**35.5%**
	month	68.4%	38.9%	62.9%	34.1%
	content	69.0%	39.5%	62.5%	35.4%
Domain-Embedding	season	*69.1%*	39.3%	*63.3%*	33.9%
	month	68.0%	38.0%	61.8%	32.8%
	content	*69.1%*	**39.9%**	62.3%	*35.0%*
Domain-Header	season	*69.1%*	*39.3%*	*62.5%*	*35.2%*
	month	65.7%	36.6%	58.8%	32.3%
	content	68.2%	38.6%	61.4%	33.2%

書には mecab-ipadic-neologd[5]を用いた) を用いて各発話・応答を形態素に分割した．テストに用いたデータはすべて 20 単語以下のものに制限している．また発話は他のツイートに対するリプライとなっていないものに限定することで，リプライの連鎖の中に存在する文脈の不足によって回答不能となるケースを可能な限り減らしている．

比較モデル　実験では以下の 4 つのモデルを比較する．

Baseline(SEQ2SEQ) [6] 第 3.1 節参照.

Local/Global SEQ2SEQ (提案手法) 第 3.2 節参照.

Domain-Embedding [7] 入力に加えて発話状況を表現するベクトルを RNN の各時刻の入力と結合させる．Li らと異なる点として，本研究では発話状況に対応したベクトルを Encoder と Decoder の両方で用いている (図 2).

Domain-Header [5] 発話を構成する単語列の先頭にその発話の発話状況を表す ID を加える (図 3).

以上のモデルを TensorFlow[6]を用いて実装した．RNN は 3 層の多層 LSTM を採用し，最適化は Adam [9] によって行った．主要なハイパーパラメータは語彙: 100,000 word，隠れ層: 200 次元，dropout 率: 0.25，初期学習率:1e-5 に設定した．

[5]https://github.com/neologd/mecab-ipadic-neologd
[6]https://www.tensorflow.org/

評価手順　前述した学習データを用いて訓練した各モデルについて応答選択テストを行う．具体的にはあるツイートに対し，実際に行われたリプライとダミー応答候補から最大 19 応答，合計最大 20 応答候補から応答としての妥当さを順位付けする．評価尺度としては Wu ら [10] の **1 in t P@k** を用いる．これは，t 個の応答候補の中から k 応答選択し，その中に実際に行われたリプライが含まれている割合を意味する．

4.2　結果

第 3.2 節で述べた 2 種類の発話状況に対し，第 4.1 節で述べた各種法を用いて実験を行った (表 1).

結果，我々が提案する Local/Global SEQ2SEQ がベースラインの結果を上回り，平均的に最も良い応答性能を示した．その理由としては Domain-Embedding, Domain-Header は発話状況を表現するベクトルを入力するモデルであるため，Embedding が入力系列もしくは RNN 内部で入力とともに繰り返し行列演算が行われるにつれ，その影響が低くなってしまうからであると考えられる．特に，Domain-Header モデルでは発話状況に関する情報は入力の先頭でしか与えられないため，多くの場合 Domain-Embedding モデルが Domain-Header を上回っていた．

表 2 に各システムの応答例を示す．発話状況については season，評価尺度は 1 in 5P@1 の結果から引用し

表 2: season, 1 in 5P@1 における応答例 (太字が正解応答)

Domain	Utterance	Baseline	Local/Global SEQ2SEQ	Domain-Embedding	Domain-Header
春	寒いと思ったら雪降ってた	おつかれ！	**まじ?この時期にやめてほしいサクラサケ**	多分なにも考えてない。こいつはそうだ	おつかれ！
夏	アイス界に行ってくる	**寒そう**	マチキンもあっただろ!	なんでこんなにも起きるのが遅いんですか?	呼ばれてないけど春ですよ!!
秋	秋の花粉きてるなあ	ご苦労だった、ゆうこ。	**もうかなり前から鼻ヤバイです…**	ご苦労だった、ゆうこ。	関西以西だと思いますけど、最近は標準語化してるかもですねー
冬	冬らしからぬ紅い髪で除雪するエレノア	あいこん	**冬は白髪になろう**	**冬は白髪になろう**	**冬は白髪になろう**

た．Li らが述べるように，マイクロブログなどの雑談対話コーパスを用いて訓練した一般的なニューラル対話システムでは典型的な応答（例：おつかれ，おはよう）が非常によく見られる [11]．この現象は訓練データ中で典型応答の頻度自体が高いこと，また機械翻訳のような入出力間で概ね 1 対 1 の対応が取れている場合と異なり，雑談対話においては発話状況のような制約無しには発話に対する応答の自由度が高すぎるため汎用的な応答を選びがちであるということに起因すると考えられる．あるテストケースに季節性が存在するか否かについては主観的なものになるため季節性がどこまで結果に反映されているかについての定量的な評価は難しいが，テスト結果を俯瞰した結果全体としては季節性があると考えられる例についての応答性能は向上していた．一方で，発話状況が「夏」の例では 正解の応答は「寒そう」という夏にはそぐわないと考えられる表現である．このように，発話状況を考慮することで逆に応答に失敗するような例も存在した．

5 おわりに

本研究では多様な発話状況を持つ対話データに対し，教師無しで得られる発話状況に注目し，発話状況を考慮した対話モデルを提案した．実験ではマイクロブログから抽出した対話データにおいて，発話内容や時節に基づく発話状況を考慮した提案モデルが，既存手法に比べて高い応答性能を発揮することを確認した．

本研究で対象とした発話状況以外にも，会話相手との関係性など応答に影響を与える発話状況は多数存在するため，これらを同時に考慮した対話モデルの評価を行いたい．

謝辞 本研究の一部は JSPS 科研費 16H02905, 16K16109 の助成を受けたものです．

参考文献

[1] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *NAACL-HLT*, pp. 172–180, 2010.

[2] O. Vinyals and Q. V. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[3] N. Yoshinaga M. Toyoda S. Sato, S. Ishiwatari and M. Kitsuregawa. UT dialogue system at NTCIR-12 STC. In *NTCIR-12*, pp. 518–522, 2016.

[4] T. Hasegawa, N. Kaji, N. Yoshinaga, and M. Toyoda. Predicting and eliciting addressee's emotion in online dialogue. In *ACL*, pp. 964–972, 2013.

[5] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.

[6] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pp. 3104–3112, 2014.

[7] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan. A persona-based neural conversation model. In *ACL*, pp. 994–1003, 2016.

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[9] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] B. Wu, B. Wang, and H. Xue. Ranking responses oriented to conversational relevance in chat-bots. In *COLING*, pp. 652–662, 2016.

[11] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, pp. 110–119, 2016.