階層型 RNN を用いた対話における応答の対話行為予測

大原康平 † 佐藤翔悦 † 吉永直樹 ‡ 豊田正史 ‡ 喜連川優 §‡
† 東京大学 ‡ 東京大学生産技術研究所 § 国立情報学研究所
{ohara, shoetsu, ynaga, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

1 はじめに

Siri のような知的対話エージェント，Amazon Echo や Google Home のようなスマートスピーカーの普及に伴い，身の回りの出来事についてデバイスと人間が雑談対話を行う機会が増えている．これらの雑談対話システムにおいて，ある発話にシステムが適切な応答を返すためには，その発話の対話行為を推定しそれに応じた応答を返すことが有効である．例えば，「質問」の発話であれば回答を行ったり，「自己開示」の発話に対しては「あいづち」を返すなどすることで，自然な会話を行うことが可能になると考えられる．

このような背景を受けて，与えられた発話の対話行為を推定する研究がこれまで広く行われている [1, 2, 3, 4]．これらの研究では，発話や話者交替，過去の対話履歴などを入力として発話の対話行為の推定を行っている．また，深層学習を利用した手法 [4] では，Switchboard コーパスで 80.1% の精度で対話行為の推定が可能であると報告されている．

これらの研究は，発話の対話行為に応じて適切な応答を返すことを目的としたものであるが，人間の間の会話では，対話行為が「質問」である発話に対して聞き返しを行ったり，「自己開示」である発話に対して「自己開示」で応答したりと，必ずしも発話の対話行為に対して一意に応答が決まるものではない．そのため，より直接的に，どのような対話行為の応答を返すべきか予測することができれば，それに応じた応答を返すことでより自然な会話を行うことが可能になると期待できる．

そこで，本研究では，与えられた発話に対する応答の対話行為の予測を行う問題を新たに設定し，階層型 Recurrent Neural Network (RNN) を用いて，これを解く手法を提案する．提案手法では，発話の系列（履歴）を入力とし，発話自体をモデル化する RNN と，発話の単語系列をモデル化する RNN を組み合わせて，（最後の発話者とは異なる話者による）応答の対話行為の予測を行う．また，このモデルを発話の対話行為の推定にも転用し，両タスクの差異についても考察を行う．

実験では JAIST タグ付き自由対話コーパス [2] を用いて提案手法の評価を行った．評価尺度としては，発話に対する応答の非決定性を考慮し，システムが予測した 3 ラベル中に正解が含まれていれば正解とみなす top-3 accuracy を用いて応答の対話行為の予測結果を評価する．また，人手による対話行為予測も行い混同行列を用いて結果を詳細に比較する．

2 関連研究

発話に対する応答の対話行為予測を行った研究は著者の知る限りこれまで存在しないが，発話の対話行為推定については広く研究が行われており，提案タスクと強く関連すると考えられるので以下で説明する．

発話の対話行為の推定は多クラス分類問題であり，機械学習を用いた手法が提案されている．Grau ら [1] は，Naive Bayes 分類器の特徴量として単語トライグラムを用い，Switchboard コーパスにおける対話行為の推定を行った．福岡ら [2] は，対話行為によって有効な素性は異なるとし推定毎に最適な特徴量の選択を行うことで，全ての特徴量を使う場合に比べて有意に高い性能が得られることを示した．

近年では汎化性能の高いニューラルネットワークを用いた対話行為推定モデルも考案されている．Kalchbrenner ら [3] は，発話の局所的な特徴を抽出する Convolutional Neural Network (CNN) と対話全体の特徴を捉える Recurrent Neural Network (RNN) を組み合わせた手法を提案し，Switchboard コーパスにおいて既存の機械学習ベースのモデルを上回る性能を挙げている．また Khanpour ら [4] は Deep Recurrent Neural Network を用いた手法を提案している．

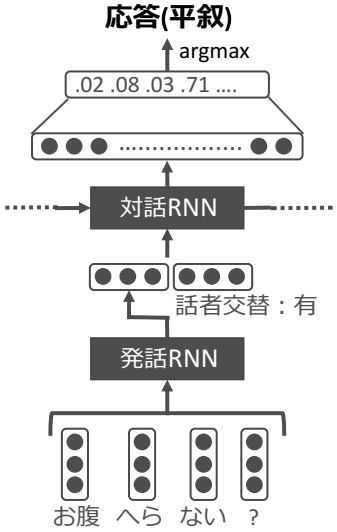

図1　提案手法：階層型 RNN

これらの研究で利用されている話者交替や過去の対話履歴は，応答の対話行為予測においても有効であると考えられるため，これらを考慮するモデルを構築し，発話に対する応答の対話行為予測タスクに取り組む．

3　提案手法

本研究では，発話の対話行為推定において高い性能を挙げたニューラルネットワークベースのモデルを参考とし，階層型 RNN を用いた応答の対話行為予測モデルを設計した．

提案モデルを図 1に示す．このモデルは予測対象の発話だけでなく，その直前の発話も考慮して応答の対話行為予測を行う．発話 RNN はそれぞれの発話に含まれる単語の系列から発話全体を表すベクトル (発話ベクトル) を出力する RNN であり，対話 RNN はそれぞれの発話ベクトルを入力として一連の対話を表現するベクトル (対話ベクトル) を出力する RNN である．ここで，対話において話者交替が発生したかどうかの情報を考慮するため，発話ベクトルに話者交替の有無を表す埋め込みを結合した上で対話 RNN への入力とする．[*1]

出力層では対話 RNN から出力される対話ベクトルを元に応答の対話行為の確率分布を推定する．図 1の例では一つ目の発話「お腹へらない?」に対する応答「へったね」の対話行為である「応答 (平叙)」の確率を最大化するように学習が行われる．最適化においては，各タイムステップの発話それぞれから後続する発話の対話行為の予測を行い，その交差エントロピー誤差の平均を最小化する．

提案手法では応答以前の発話をどれだけ遡って使うかによって対話行為予測の性能は変わると考えられるが，本研究では現在の発話に加え過去の 3 つの発話，計 4 つの発話系列を用いて対話行為を予測する．

また1節で述べたように，発話に対する応答の対話行為予測タスクと，発話の対話行為推定タスクとの比較を行うため，同モデルを発話の対話行為推定タスクにも転用する．差異は出力のみであり，例えば図 1の一つ目の発話「お腹へらない?」の対話行為「質問 (YesNo)」を推定するように学習を行う．

表1　JAIST タグ付き自由対話コーパスの対話行為の分布．割合は%．応答数は話者交替が発生した回数．

対話行為	全発話数 (割合 (%))		応答数 (割合)	
自己開示	53,701	(58.4)	22,870	(44.6)
質問 (YesNo)	6,430	(7.0)	4,103	(8.0)
質問 (What)	3,950	(4.3)	2,632	(5.1)
応答 (YesNo)	2,130	(2.3)	2,086	(4.1)
応答 (平叙)	7,508	(8.2)	6,932	(13.5)
あいづち	9,216	(10.0)	7,788	(15.2)
フィラー	4,405	(4.8)	1,718	(3.4)
確認	3,940	(4.3)	2,727	(5.3)
要求	751	(0.8)	424	(0.8)

4　実験

4.1　データセット

学習・評価用データセットとして福岡ら [2] の研究で用いられた，名大対話コーパス [5] の一部の対話に対して対話行為ラベルを付与した JAIST タグ付き自由対話コーパス[*2]を用いる．本研究では，コーパスから固定長[*3]の発話系列を抜粋したものを実験に使用する．対話数は 97，発話数は 92,020 であり，各発話に対して対話行為タグが人手により付与されている．対話行為とその分布を表 1に示す．

実験のための学習・開発・テストデータは福岡ら [2] と同様[*4]にデータセットを 80%，10%，10% に分割した．各分割でそれぞれどの対話を用いたかについては，

[*1] 本研究では発話者とは異なる話者による応答の対話行為の予測を行うため，入力の会話末尾の発話に対しては常にその後の発話で話者交替が行われるが，それ以前の発話については同じ話者による連続した発話も含まれうることに注意されたい．

[*2] http://www.gsk.or.jp/catalog/gsk2017-b/

[*3] 本研究では対話の系列長を応答を含めて 5 に設定した．

[*4] 階層型 RNN のために対話の系列長を 5 に設定し，その内側は二者になるように切り出しているため，正確には同じデータセットにはなっていない．

表2　モデルのハイパーパラメタ．

RNN の層数	1
RNN の隠れ層	500 次元
単語/話者交替の埋め込み	500 次元
語彙サイズ	11,613
dropout rate	0.5
バッチサイズ	300
学習率 (Adam)	0.0001
epsilon (Adam)	1e-07

著者のホームページ[*5]にて公開している．

4.2　実験設定

RNN としては LSTM，最適化には Adam [6] を用い，tensorflow[*6]により提案モデルを実装した．表 2に開発データを用いて調整したハイパーパラメタを示す．なお，単語埋め込みの初期化には Wikipedia コーパスで学習した Gensim の word2vec[*7]を用い，話者埋め込みの初期化には Xavier initialization [7] を用いた．word2vec のモデルには CBOW を用いた．

また応答のバリエーションは応答者の性格や状況などによって異なることを考慮し，評価尺度として，システムが予測した 3 ラベル中に正解が含まれていれば正解とする top-3 accuracy を用いる．性能比較のため以下の 3 つのモデルで実験を行った．

weighted random　表 1の応答の割合に基づきランダムに対話行為を予測するベースライン．

w/o history　提案手法において，過去の発話履歴（及び話者交替）を使わないモデル (直前の発話に対する図 1の発話 RNN の出力を直接予測に用いる)．

proposed　提案手法 (3節)．

4.3　実験結果

各モデルの対話行為別の予測精度を表 3に示す．提案手法と weighted random の結果を比較すると，比較的低頻度のラベルでは提案手法の性能が下回る一方で，「自己開示」や「応答」などのデータセット中で高頻度の対話行為では上回っており，全体の精度でも 5.1% 優れた結果となっている．これは学習データのラベルの偏りを学習し低頻度のラベルはほとんど予測しないようなモデルになっているためであると考えられる．

[*5] http://www.tkl.iis.u-tokyo.ac.jp/~ohara/nlp-17
[*6] https://www.tensorflow.org/
[*7] https://radimrehurek.com/gensim/models/word2vec.html

表3　応答の対話行為予測結果：top-3 accuracy の比較．太字は各ラベルに対する最高値．

	weighted random	w/o history	proposed
自己開示	.881	.860	**.950**
質問 (YesNo)	.311	**.380**	.205
質問 (What)	.167	**.589**	.525
応答 (YesNo)	.155	.708	**.734**
応答 (平叙)	.516	**.709**	.662
あいづち	.532	.375	**.785**
フィラー	**.159**	.013	.013
確認	.194	**.622**	.119
要求	**.038**	.000	.000
全体 (マイクロ平均)	.597	.654	**.705**

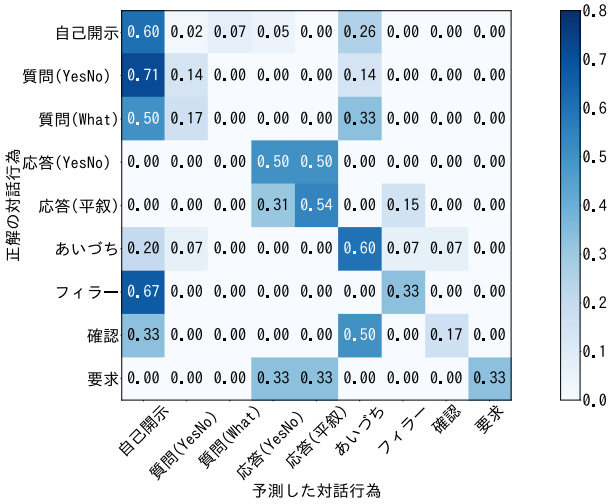

図2　混同行列: 人による応答の対話行為予測結果

提案手法と w/o history のモデルとの差に注目すると，「応答」と「質問 (What)」ではその差は数% 程度である一方で，「あいづち」で 40% 上回る結果であった．この理由としては，一般に人間は対話において聞き役に徹する際，「あいづち」を連続で続ける場合が多く，これに対し提案手法における過去の発話系列の考慮が有効に働いたのだと考えられる．

次に著者 2 人により，テストデータの中からそれぞれ異なる対話，計 100 件について応答の対話行為の予測を行い，提案手法との比較を行った．提案手法と人手評価による top-1 の予測結果についての混同行列を図 2，図 3に示す．図 3より，提案手法による予測結果は正解が「応答」の時以外はほとんど「自己開示」へと偏っている．これはモデルがデータセット中で高頻度の対話行為の分布を学習してしまっていることなどが原因と言える．　一方で人手による予測結果 (図 2) に注目すると，自己開示への偏りは提案手法より少ないものの依然みられ，特に「質問」のラベルについては大きく「自

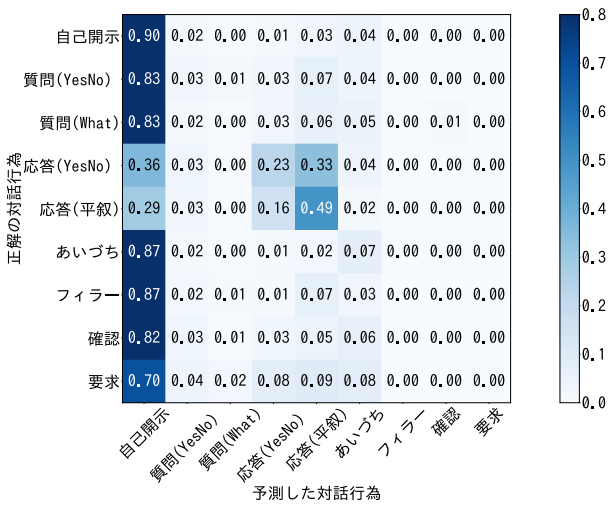

図3　混同行列: 提案手法による応答の対話行為予測結果

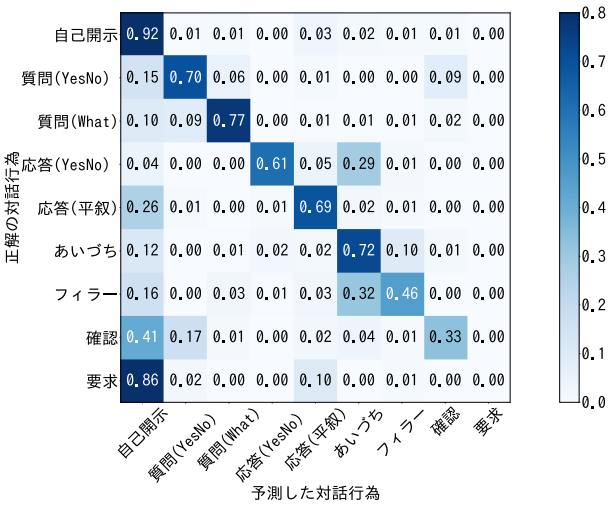

図4　混同行列: 提案手法による発話の対話行為推定結果

己開示」へ偏っている．このことから発話に対して質問をするか否かを予測するためには，今回利用した予測を行う発話およびその直前の 3 つの発話だけでは不十分であることが分かる．[*8]

最後に，同様のモデルで発話の対話行為推定をした場合の結果の混同行列を図 4に示す．図より「確認」と「要求」以外のラベルについては比較的高い精度で予測できていることが分かる．ここで興味深いのは，発話の対話行為推定の学習データもラベルの割合に偏りがあるにも関わらず，応答の対話行為予測よりも明らかに高い性能を挙げていることである．これより，応答の対話行為予測は発話の対話行為推定よりも解くのが難しいタスクだと言える．

5　おわりに

本論文は与えられた発話に対する応答の対話行為予測を行う萌芽的なタスクに取り組むため，予測対象となる発話とそれ以前の対話の情報，話者交替の情報を考慮する階層型 RNN の手法を提案した．実験では JAIST タグ付き自由対話コーパスをデータセットとして用い，同一の発話に対しても人や状況によって応答が異なることを考慮し top-3 accuracy を用いて評価を行った．その結果，提案手法はベースラインの性能を上回ることを確認し，人手評価などとの比較により応答の対話行為予測が難しいタスクであるという知見も得た．

今後の課題としては，応答生成において非言語の発話状況を考慮した手法 [8] を参考に，応答の対話行為を限定する様々な外的要因を考慮し，より高精度で応答の対話行為を推定する手法を開発することが挙げられる．

[*8] 応答する人間の知識や対話をしている二者間で共有している文脈など，対話の表層には表れない情報もあるため難しいタスクであると言える．

謝辞

本研究の一部は JSPS 科研費 17J06394，16K16109 の助成を受けたものです．

参考文献

[1] Maria Jose Castro Sergio Grau, Emilio Sanchis and David Vilar. Dialogue act classification using a bayesian approach. In *Proc. of SPECOM*, 2004.

[2] 福岡知隆, 白井清昭. 対話行為に固有の特徴を考慮した自由対話システムにおける対話行為推定. 自然言語処理, Vol. 24, No. 4, 2017.

[3] Nal Kalchbrenner and Phil Blunsom. Recurrent convolutional neural networks for discourse compositionality. In *Proc. of CVSC*, 2013.

[4] Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proc. of COLING*, 2016.

[5] 国立国語研究所. 名大会話コーパス. 科学研究費基盤研究 (B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」(平成 13 年度～ 15 年度).

[6] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICML*, 2015.

[7] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of AISTATS*, 2010.

[8] Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. Modeling situations in neural chat bots. In *Proc. of ACL-SRW*, 2017.