初対面対話における好感のモデリングと発話構成要素の選択

Modeling of Feeling and Selection of Utterance Constructional Units in First Encounter Dialogue

田中 滉己 *　井上 昂治　中村 静　高梨 克也　河原 達也
Koki Tanaka　Koji Inoue　Shizuka Nakamura　Katsuya Takanashi　Tatsuya Kawahara

京都大学 大学院情報学研究科
Graduate School of Informatics, Kyoto University

Abstract: Modeling internal mental states is important for a dialogue system that behaves like a human. In this paper, a model of feeling toward the dialogue partner is addressed. In the proposed model, feeling is modeled based on a hierarchical structure of logistic regression models. At first, user's feeling toward the system and user's interest in the current topic are predicted by user's multimodal dialogue behaviors. Then, system's feeling toward the user is determined by the predicted results. Finally, system's utterance content, that is, 'the utterance constitutional units' are selected based on the system's feeling. The utterance constitutional units include response, episode, and question parts. Each logistic regression model is individually pre-trained with a small amount of the annotated data of feeling. Afterward, the entire model is fine-tuned with a large amount of dialogue data. Experimental results show that the modeling of feeling contributes to improving accuracy of the utterance constructional units.

1　はじめに

音声言語処理や対話システムの研究の発展を受けて，スマートスピーカや会話ロボットなどが実用化されている．そこでのやりとりは，基本的には一問一答や少数のターンでのやりとりである．これに対して，我々は，人間どうしの対話のように，より深いやりとりを指向した対話システムの研究を進めている．人間らしいふるまいを対話システムで実現するための一要素として，感情などの内部状態をシステムに持たせることが挙げられる [1]．これまでに，対話相手のふるまいから，その相手の感情や興味などの内部状態を推定する研究が多く進められてきたが [2, 3, 4, 5]，そこからシステム自身の内部状態を生成し，さらにはシステムのふるまいまで反映させた例は少ない．

本研究では，システムの内部状態として，初対面対話における対話相手への好感を扱う．人間どうしの初対面対話では，相手への好感の状態が対話中のふるまいや態度に反映されることがある．本研究では，好感の状態が反映されるふるまいとして，システムの発話の構成要素を考える．発話の構成要素は，談話分析の研究 [6] に基づき，反応，エピソード，質問の 3 つとする．反応はユーザ発話への反応や回答，エピソードは自己開示などの情報提供，質問は話題の掘り下げなどの役割がそれぞれある．システムの発話を生成する際に，好感の状態に応じて，各構成要素の選択を判断する．図 1に提案システムの概要を示す．例えば，好感が高い場合，全ての要素が選択されてより多くの内容を発話する．一方，好感が低い場合，必要最小限である反応のみを発話する[1]．

本稿では，階層的なニューラルネットワークにより発話構成要素を選択する．ただし，システムの内部状態である好感は，ネットワークの中間層として表現する．はじめに，ユーザのふるまいから，ユーザがシステムに対して抱く好感と，話題に対する興味を推定する．次に，ユーザのふるまいと，さきほどの推定結果から，システムがユーザに対して抱く好感を決定する．最後に，ユーザのふるまいと，システムがユーザに抱く好感から，発話構成要素を選択する．本研究では，内部状態の学習ラベルを考慮した効率的なネットワークの学習方法を提案する．中間層に相当する好意や興味のデータは主観的であるため，アノテーションが難しくデータ量が限られてしまう．一方，入力と出力に相当するふるまいと発話構成要素は客観的であるため，より大量のデータを用意することができる．そこで，少

*連絡先：京都大学 大学院情報学研究科 知能情報学専攻
京都市左京区吉田本町
E-mail：tanaka@sap.ist.i.kyoto-u.ac.jp

[1]デモ動画 https://youtu.be/M3WLl4XcjMQ

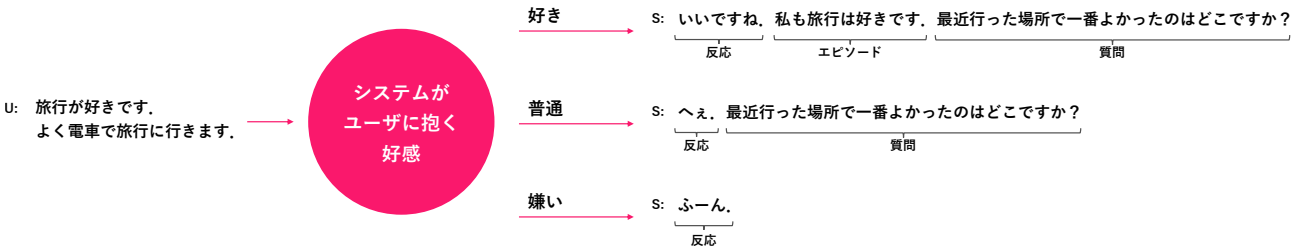

図 1: システム内部状態 (好感) に基づくシステムの発話の構成要素の選択

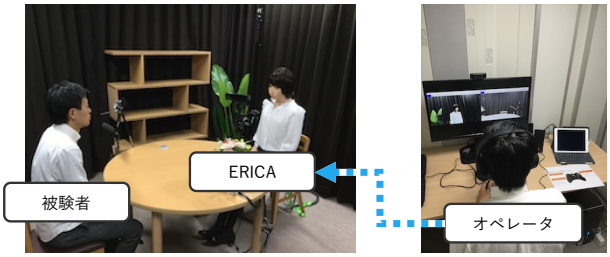

図 2: データ収録の様子

量の内部状態のラベルを用いて，中間層を事前学習する．その後，入力と出力のラベルを用いて，ネットワーク全体を End-to-end でファインチューニングする．以上のように，内部状態を効率的に学習することで，発話構成要素の選択の精度向上を目指す．

2　お見合い対話コーパス

本研究では，初対面対話のデータとして，アンドロイド ERICA[7] を用いて収録したお見合い対話コーパスを用いる．この対話は，被験者と ERICA による 1 対 1 のものであり，ERICA は別室のオペレータによって操作されている．データ収録の様子を図 2に示す．対話は，ERICA によるお見合いの練習であり，お互い初対面であるため，パーソナルな情報を交換することを目的とした．被験者には，趣味，職業，出身地など，お見合いで話す可能性の高い話題の一覧を事前に提示した．オペレータには，発話構成要素と好感の概念を事前に説明し，自然な対話を保ちつつ，好感に基づいて発話構成要素を選択するように指示した．また，被験者に対する好感が低い場合には，気を遣うことなく，それが発話構成要素に現れてもよいことを伝えた．

オペレータによる ERICA のふるまい制御は以下の通りである．ERICA の発話については，オペレータが話した音声を，ロボットに搭載したスピーカからそのまま再生した．また，ロボットの口の動きはその音声に合わせて自動で生成される [8]．視線，頷き，腕のジェスチャは，オペレータの手元のコントローラで操作した．

表 1: 発話構成要素の組合せの頻度

発話構成要素			頻度
反応	エピソード	質問	
✓	-	-	472
✓	✓	-	177
✓	-	✓	86
-	✓	-	69
-	-	✓	53
✓	✓	✓	8
その他			34
計			899

上記の設定で，18 セッションの対話を収録した．収録に用いたセンサは，ショットガンマイク，マイクロフォンアレイ，RGB カメラ，Kinect v2 である．収録したデータに対して，発話，相槌，笑い，フィラー，ターン，対話行為 [9]，長い発話単位 [10] を手動でアノテーションした．また，1 セッションあたりの平均時間は 10 分 55 秒であった．被験者として大学生および大学院生の男性 18 名，オペレータは 20 代から 30 代の女性 4 名が参加した．したがって，各オペレータは複数のセッションに参加したことになる．

各対話セッションの終了後に，オペレータにアンケートに答えてもらった．はじめに，対話中に現れた話題を挙げてもらった．その後，各話題について話していたときを振り返ってもらい，以下の要素について 7 段階で評価してもらった．

a. ERICA(オペレータ) が被験者に抱いていた好感
b. 被験者が ERICA に抱いていた好感の予想
c. ERICA(オペレータ) が話題に持っていた興味
d. 被験者が話題に持っていた興味の予想

収録データおよびアンケート結果の分析について述べる．はじめに，オペレータのターン数は全体で 899 であった．また，各ターンに対して発話構成要素の組合せをアノテーションしたところ，表 1に示す分布となった．最も多い組合せは「反応のみ」の 472 ターンであった．また，オペレータに対して行ったアンケートの各項目の分布を図 3から図 6に示す．オペレータに

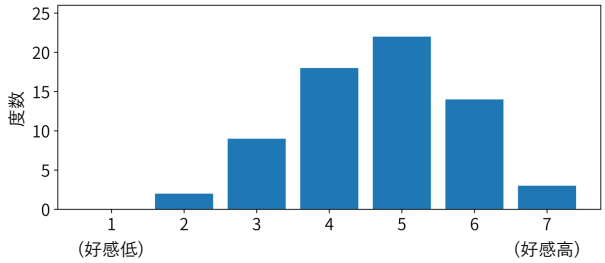

図 3: オペレータが被験者に抱く好感の分布

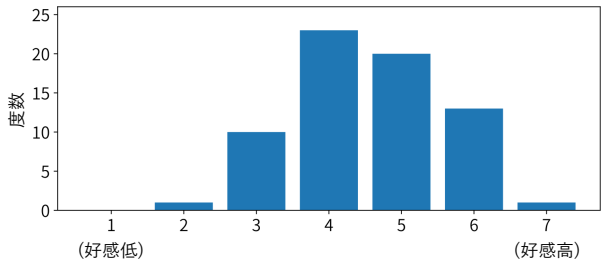

図 4: 被験者がオペレータに抱く好感（オペレータによる推定）の分布

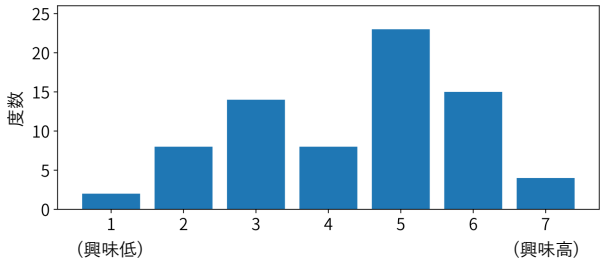

図 5: オペレータが話題に持つ興味の分布

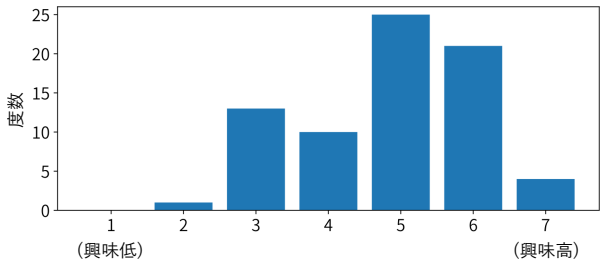

図 6: 被験者が話題に持つ興味（オペレータによる推定）の分布

よって挙げられたトピックの数は全体で 74 であった．興味の分布（図 5と図 6）の方が好感の分布（図 3と図 4）よりも広く分散していることから，興味の度合いの方が話題による影響が大きいと考えられる．一方で，好感は興味よりも対話を通して比較的安定しており，ゆるやかに変化するものと考えられる．

3 問題設定

本研究でのタスクは，ユーザのふるまいから得られた特徴量をもとに，次のシステムのターンに含まれる発話構成要素を選択することである．問題設定を図 7に示す．入力特徴量はユーザの話し方と聞き方から得られる．ユーザの話し方に関する特徴量 $\boldsymbol{o}_s$ は，先行するユーザのターンでのふるまいから抽出する．ユーザの聞き方に関する特徴量 $\boldsymbol{o}_l$ は，最後のシステムのターン中のユーザのふるまいから抽出する．話し方に関する特徴量と聞き方に関する特徴量を結合して以下のように表す．

$$\boldsymbol{o} := (\boldsymbol{o}_s \ , \ \boldsymbol{o}_l) \tag{1}$$

特徴量についての詳細は 4節で述べる．出力は，反応・エピソード・質問の 3 つの発話構成要素の組合せのパターン $\boldsymbol{a}$ である．また，システムの好感などを内部状態として扱い，ベクトル $\boldsymbol{s}$ で示す．したがって，本研究で扱う問題は，観測 $\boldsymbol{o}$ をもとに内部状態 $\boldsymbol{s}$ を考慮しながら，次のシステムの行動 $\boldsymbol{a}$ を予測することである．対話システムに関する従来研究と比較すると，内部状態はスロットなどの対話状態に対応する．タスク指向対話では，対話状態が客観的に定義されるため，確率統計モデルを学習するためのラベルデータを集めることが比較的容易であった．しかし，本研究で扱うお見合い対話では，内部状態は好感という主観的なものであるため，十分なデータを集めることが困難である．

発話構成要素の組合せを推定するタスクについて述べる．表 1の結果より，発話構成要素の分布には偏りがあるため，発話構成要素の組合せを直接予測するのではなく，図 8に示す 3 つのサブタスクに分けて行う．1 つ目のタスクは，システムのターンが反応のみか，他の要素（エピソードまたは質問）も含むかを予測するタスクである．他の要素も含むと決定された場合には，以下の 2 つのタスクをそれぞれ実行する．2 つ目のタスクは，システムのターンにエピソードが含まれるか否かを予測する．同様に，3 つ目のタスクは，システムのターンに質問が含まれるか否かを予測する．ここでは，各タスクの予測は独立に行い，各予測を組合せて最終的な発話構成要素を決定する．表 2に，発話構成要素の各組合せが，各タスクにおいて正例であるか負例であるかを示す．

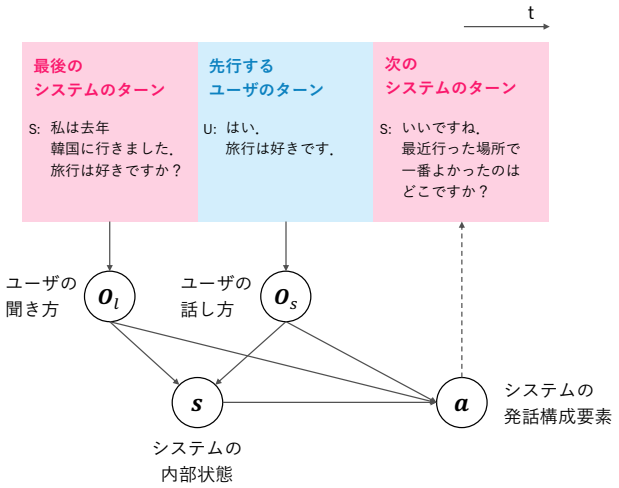

図 7: 本研究における問題設定

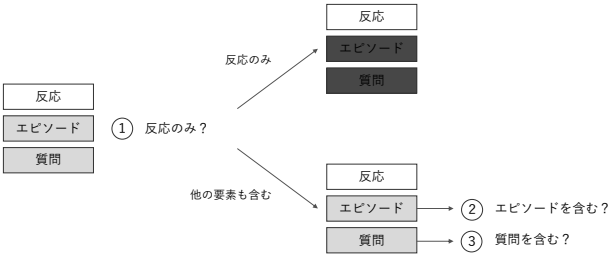

図 8: 発話構成要素の選択のための 3 つの分類タスク

4 提案モデル

本研究では，次のシステムの発話構成要素を選択するために好感という内部状態を考える．しかし，内部状態のラベルは主観的であるため，そのデータを大規模に収集することは難しい．これは内部状態をモデル化する上では普遍的な問題であるといえる．一方で，観測 $\boldsymbol{o}$ と行動 $\boldsymbol{a}$ のラベルは客観的なものであるから，収録した対話データから容易に得ることができる．そこで，内部状態のラベルが少量の場合でも効率的に学習を行う階層的なニューラルネットワークを提案する．はじめに，内部状態に対応するネットワークの中間層を個別に事前学習する．その後，End-to-End でモデル全体をファインチューニングする．

4.1 特徴量

はじめに，使用する特徴量について述べる．ここでは，観測ベクトル $\boldsymbol{o} = (\boldsymbol{o}_s, \boldsymbol{o}_l)$ として，ユーザの話し方と聞き方に関する以下の特徴量を用いる．ユーザの話し方に関する特徴量 $\boldsymbol{o}_s$ は，先行するユーザのターンから以下を抽出する．

- ターンの継続長

表 2: 各タスクにおけるラベルの設定 (p: 正例，n: 負例，-: 不使用)

発話構成要素			タスク		
反応	エピソード	質問	1	2	3
✓	-	-	p	-	-
✓	✓	-	n	p	n
✓	-	✓	n	n	p
-	✓	-	n	p	n
-	-	✓	n	n	p
✓	✓	✓	n	p	p

- 直前のシステムのターンの終わりからのポーズ長
- ターン中の発話区間の割合
- 対話の開始からの発話区間の割合
- 発話速度
- パワー（平均とレンジ）
- F0（平均とレンジ）
- エピソードの長さ（エピソードがない場合は 0）
- 笑いの頻度
- フィラーの頻度
- 発話構成要素の組合せ

パワーと F0 の抽出には Praat[11] を用いた．エピソードの長さには，長い発話単位（LUU）の数を用いた．LUU は意味的なまとまりを表し，エピソードの意味的な量を捉えることができると考えられる．発話構成要素の組合せは 3 次元の二値ベクトルで表され，各次元は発話構成要素それぞれの有無を表す．以上より，ベクトル $\boldsymbol{o}_s$ の次元は 18 である．

ユーザの聞き方に関する特徴量 $\boldsymbol{o}_l$ は，最後のシステムのターンにおけるユーザのふるまいから以下を抽出する．

- 相槌の頻度
- 笑いの頻度

ベクトル $\boldsymbol{o}_l$ の次元は 2 である．将来的には，視線やうなずきといったふるまいも用いる予定である．

4.2 ネットワーク構成

提案する階層的なニューラルネットワークの構成を図 9に示す．このネットワークは，3節で定義した 3 つの分類タスクにおいて，モデルの学習および分類を別々に行う．最終的には，3 つのタスクそれぞれの出力を統合して，システムの発話構成要素 $\boldsymbol{a}$ を決める．以下では，ある 1 つのタスクに対するネットワークの構成として説明する．はじめに，ユーザのふるまいから D_o 次元（ここでは 20 次元）の入力特徴量のベクトル $\boldsymbol{o}$ を得る．観測から直接システムがユーザに抱く好感を推

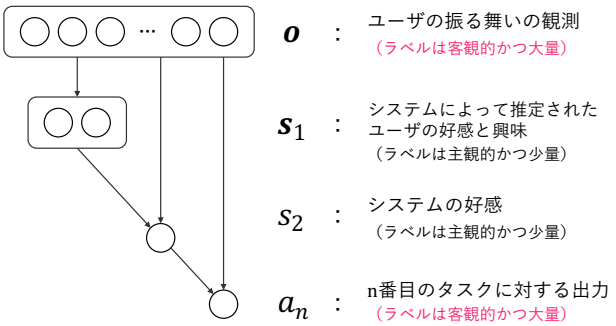

図 9: 提案モデル

定することも可能ではあるが，まずは，ユーザがシステムに対して抱く好感と話題に対して持つ興味を以下のように推定する．

$$\boldsymbol{s}_1 = \sigma(A_1 \boldsymbol{o}^T + \boldsymbol{b}_1^T) \tag{2}$$

ここで，$\boldsymbol{s}_1$ はユーザの好感と興味に対応する 2 次元のベクトルである．A_1 と $\boldsymbol{b}_1$ はネットワークパラメータであり，それぞれの大きさが $2 \times D_o$ と 2 である．$\sigma()$ はシグモイド関数で，T は転置を表す．次に，システムがユーザに対して抱く好感を，ユーザのふるまいと，ユーザが抱く好感と興味（まとめて $\boldsymbol{s}_{1'} = (\boldsymbol{s}_1, \boldsymbol{o})$ と書く）を元に推定する．

$$s_2 = \sigma(A_2 \boldsymbol{s}_{1'}^T + b_2) \tag{3}$$

ここで，s_2 はシステムがユーザに抱く好感に対応するスカラー値である．A_2 と b_2 はネットワークパラメータであり，それぞれの大きさは $1 \times (2 + D_o)$ と 1 である．最後に，各タスクにおける発話構成要素の選択に対応する確率値が予測される．

$$a_n = \sigma(A_3 \boldsymbol{s}_{2'}^T + b_3) \tag{4}$$

ここで，$\boldsymbol{s}_{2'}$ はシステムの好感とユーザのふるまいを結合したベクトル $\boldsymbol{s}_{2'} = (s_2, \boldsymbol{o})$ である．A_3 と b_3 はネットワークパラメータであり，それぞれのサイズは $1 \times (1 + D_o)$ と 1 である．a_n は，n 番目のタスクにおける出力である．例えば，タスク 1 （反応のみか，他を含むか）の場合は，反応以外の要素を含む確率に対応する．

4.3 モデルの学習

モデルの学習は事前学習とファインチューニングの 2 つの段階からなる．事前学習では，内部状態に対応するネットワークの各層を入力層から順番に学習する．はじめに，$\boldsymbol{s}_1$ を出力する層について，入力 $\boldsymbol{o}$ とアンケート結果から得た $\boldsymbol{s}_1$ のラベルを用いて学習する．次に，

表 3: タスク 1（反応のみ/反応以外も含む）の予測結果

モデル	適合率	再現率	F 値
ベースライン	0.672	0.622	0.646
事前学習なし	0.648	0.643	0.646
fine-tuning なし	0.714	0.568	0.632
提案手法	0.687	0.654	0.670

表 4: タスク 2（エピソードあり/エピソードなし）に対する予測結果

モデル	適合率	再現率	F 値
ベースライン	0.642	0.798	0.712
事前学習なし	0.655	0.757	0.702
fine-tuning なし	0.671	0.814	0.735
提案手法	0.672	0.817	0.738

表 5: タスク 3（質問あり/質問なし）に対する予測結果

モデル	適合率	再現率	F 値
ベースライン	0.387	0.659	0.488
事前学習なし	0.377	0.768	0.506
fine-tuning なし	0.427	0.674	0.522
提案手法	0.386	0.848	0.531

s_2 を出力する層について，入力 $\boldsymbol{o}$ とアンケート結果から得た $\boldsymbol{s}_1$ および s_2 のラベルを用いて学習する．最後に，a_n を出力する層について，入力 $\boldsymbol{o}$ とアンケート結果から得た s_2 のラベル，および発話構成要素を選択する各タスクにおける a_n の正解ラベルを用いて学習する．ファインチューニングでは，入力 $\boldsymbol{o}$ と出力 a_n の正解ラベルのみを用いて，誤差逆伝播法によりネットワーク全体のパラメータを調整する．事前学習の効果を保つために，ファインチューニングの際には，事前学習されたパラメータとファインチューニング後のパラメータが大きく異ならないような制約を誤差関数 $E'(W)$ に加える．

$$E'(W) = E(W) + ||W - W_{pre}||_F \tag{5}$$

ここで，$E(W)$ はネットワークの出力層に対する誤差関数，W_{pre} は事前学習で得たパラメータ，W はファインチューニング後のパラメータ，$||\cdot||_F$ は行列のフロベニウスノルムをそれぞれ表す．

5 評価

2節で述べたお見合い対話コーパスを用いて提案手法を評価した． 5 分割交差検定により，発話構成要素を選択する各タスクについて，適合率，再現率，F 値を

計算した．モデルの実装には TensorFlow 1.7.0[2]を用いた．最適化手法には Adam[12] を用い，学習率はタスク 1 とタスク 3 では 10^{-2}，タスク 2 では 10^{-6} とした．

比較手法は以下の 3 つとした．1 つ目はユーザのふるまいの観測から直接ロジスティック回帰により発話構成要素を予測するものである．これは，提案モデルにおいて中間層がないネットワークであり，内部状態のモデル化がなされていない場合に対応する．このモデルをベースラインとする．2 つ目と 3 つ目は提案モデルと同じネットワーク構成をしているが，2 つ目では事前学習を行わず，3 つ目ではファインチューニングを行わない．各タスクにおける正例の割合（チャンスレベル）はそれぞれ 0.527，0.619，0.325 であった．また，タスク 3 はラベルが偏っているため，誤差関数をサンプル数の比で重み付けしている．

表 3から表 5に各タスクの予測結果を示す．まず，ベースラインモデルと提案モデルを比べると，全体的に提案モデルによる精度向上がみられた．したがって，内部状態をモデル化することによる効果が示されたといえる．次に，タスク毎に，事前学習なしおよびファインチューニングなしのモデルと提案モデルを比較する．タスク 1 とタスク 3 では，提案モデルはいずれの比較手法よりも高い精度を示した．したがって，これらのタスクでは，事前学習とファインチューニングの組合せが有効であるといえる．タスク 2 では，ファインチューニングによる効果はみられなかった．タスク 2 の対象であるエピソードは，好感の正解ラベルとの関係がより強く，事前学習のみで十分な学習が行われた可能性があると考えられる．

6 おわりに

本稿では，ユーザのふるまいの観測をもとに，好感という内部状態を考慮してシステムの発話の構成要素を選択する階層的なニューラルネットワークを提案した．内部状態のラベルデータを大量に得るのは困難なため，事前学習とファインチューニングの組合せによって効率的にネットワークの学習を行った．収録した初対面対話コーパスを用いた評価の結果，提案手法は比較手法に比べて高い精度で発話構成要素を予測できることを示した．今後の課題として，提案モデルを実際の対話システムに組み込むことのほか，ユーザの発話以外のマルチモーダルなふるまいを利用することが挙げられる．

[2] https://www.tensorflow.org/

謝辞

本研究は，JST ERATO 石黒共生ヒューマンロボットインタラクションプロジェクト JPMJER1401 の支援を受けて実施した．

参考文献

[1] Rosalind W. Picard. *Affective computing*, Vol. 252. MIT press Cambridge, 1997.

[2] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing*, Vol. 3, pp. 1–18, 2014.

[3] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review*, Vol. 43, No. 2, pp. 155–177, 2015.

[4] Björn Schuller, Niels Köhler, Ronald Müller, and Gerhard Rigoll. Recognition of interest in human conversational speech. In *Interspeech*, pp. 793–796, 2006.

[5] William Yang Wang, Fadi Biadsy, Andrew Rosenberg, and Julia Hirschberg. Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification. *Computer Speech & Language*, Vol. 27, No. 1, pp. 168–189, 2013.

[6] John McHardy Sinclair and Malcolm Coulthard. *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford University Press, 1975.

[7] Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. Talking with ERICA, an autonomous android. In *SIGDIAL*, pp. 212–215, 2016.

[8] Kurima Sakai, Carlos T. Ishi, Takashi Minato, and Hiroshi Ishiguro. Online speech-driven head motion generating system and evaluation on a tele-operated robot. In *ROMAN*, pp. 529–534, 2015.

[9] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. Towards an ISO standard for dialogue act annotation.

[10] Yasuharu Den, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida. Two-level annotation of utterance-units in japanese dialogs: An empirically emerged scheme. In *LREC*, pp. 1483–1486, 2010.

[11] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, Vol. 5, No. 9, pp. 341–345, 2001.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.