2T-06

対話相手への好感に基づく発話構成要素の選択とお見合い対話システムへの実装

田中 滉己　井上 昂治　中村 静　高梨 克也　河原 達也

京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

一問一答や短いやりとりを行う音声対話システムの実用化が進んでいる．我々は社会的かつ長いやりとりを通して人間らしい対話感を実現する音声対話システムの研究を進めている．人間らしい対話感を実現するための要素の一つとして，システムに感情などの内部状態を持たせることが挙げられる [1]．本研究では，システムの内部状態として対話相手への好感を扱う．好感が反映されるシステムのふるまいとして，発話の構成要素を用いる．ここでは，発話の構成要素として反応，エピソード，質問の３つを考える．反応は相手の発話に対する反応や回答，エピソードは自己開示，質問は話題を掘り下げたり転換したりする質問にそれぞれ相当する．図 1 に例を示す．好感が高い場合には反応に加えてエピソードあるいは質問を発話する．逆に，好感が低い場合には反応のみを発話するということが考えらえる．これを実現するための発話構成要素選択のモデルとして，システムの内部状態を中間層とする階層的ニューラルネットワークを提案する．入力はユーザのふるまい，出力はシステムの発話構成要素に対応する．内部状態を中間層に対応させるために，内部状態のラベルを用いて事前学習を行う．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ファインチューニングする．前回の発表 [2] では，上記のモデルを提案したが，本稿では，使用するコーパスでの傾向に基づき，発話構成要素を選択する複数のタスクについて再検討を行う．また，実応用として，現在実装を進めているお見合い対話システムについても述べる．

2. お見合い対話コーパス

本研究では，自律型アンドロイド ERICA[3] 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，被験者と別室のオペレータによって遠隔操作された ERICA との対話である．対話内容は ERICA によるお見合いの練習である．ERICA は女性の設定であるため被験者は男性である．各対話は約 10 分で，18 セッションを収録した．オペレータには好感と発話構成要素について事前に説明し，対話の自然さを損なわない範囲で，被験者への好感の度合いに応じて反応，エピソード，質問を選択して発話するよう指示した．対話終了後にアンケートを実施し，オペレータは，対話中の話題ごとに以下の項目を 7 段階で評価した．

1. ERICA（オペレータ）が被験者に抱いていた好感
2. 被験者が ERICA に抱いていた好感の予想
3. ERICA（オペレータ）がその話題に持っていた興味
4. 被験者がその話題に持っていた興味の予想

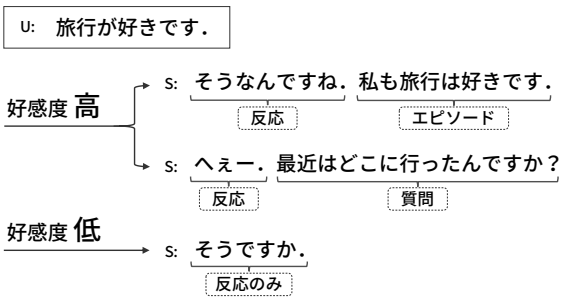

図 1: 好感に基づく発話構成要素の選択

3. 問題設定

本研究で扱うタスクについて述べる．ユーザのふるまいから得られる特徴量をもとに，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．

3.1 分類タスク

システムのターンに含まれる発話構成要素の選択は，２つのタスクに分割する．１つ目は，システムのターンが反応のみで構成されるか，反応以外の要素（エピソードまたは質問）を含むかを分類する．１つ目のタスクで後者に分類されたものに対して，２つ目のタスクを適用する．ここでは，エピソードと質問のどちらを含むかを分類する．これら２つのタスクの分類結果から，次のシステムのターンに含む要素を決定する．

3.2 特徴量

入力特徴ベクトル $\boldsymbol{o} = (\boldsymbol{o}_s, \boldsymbol{o}_l)$ として，ユーザの話し方と聞き方に関するものを用いる．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
- 直前のシステムのターンの終わりからの沈黙時間
- 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（平均およびレンジ）
- F0（平均およびレンジ）
- エピソードの長さ（長い発話単位 [4] の数）
- 笑いの頻度
- フィラーの頻度
- 発話構成要素の組合せ

ユーザの聞き方に関する特徴量 $\boldsymbol{o}_l$ は，直前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

3.3 提案モデル

本研究では，内部状態を中間層として持つ階層的ニューラルネットワークを提案する．ネットワークの構造を図

Selection of utterance constructional units based on favorable impression to a user and its implementation on a speed-dating dialogue system: *Koki Tanaka, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara (Kyoto Univ.)*

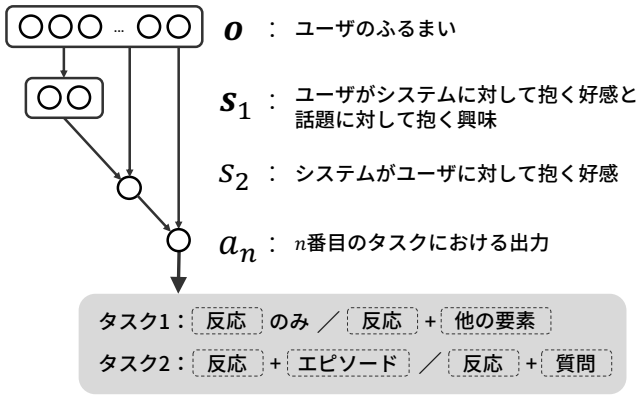

図 2: 提案モデル

2 に示す．まず，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 $\boldsymbol{s}_1$ を以下のように推定する．

$$\boldsymbol{s}_1 = \sigma(A_1 \boldsymbol{o}^T + \boldsymbol{b}_1^T) \tag{1}$$

ここで，A_n と $\boldsymbol{b}_n$ はネットワークのパラメータ，$\sigma()$ はシグモイド関数で，$\cdot^T$ は転置を表す．次に，システムがユーザに対して抱く好感 s_2 を，前段の推定結果とユーザのふるまい（まとめて $\boldsymbol{s}_{1'} = (\boldsymbol{s}_1, \boldsymbol{o})$）から推定する．

$$s_2 = \sigma(A_2 \boldsymbol{s}_{1'}^T + b_2) \tag{2}$$

最後に，前段の推定結果とユーザのふるまい（まとめて $\boldsymbol{s}_{2'} = (\boldsymbol{s}_2, \boldsymbol{o})$）から，各タスクにおける発話構成要素の選択に対応する事後確率を出力する．

$$a_n = \sigma(A_3 \boldsymbol{s}_{2'}^T + b_3) \tag{3}$$

a_n は，n 番目のタスクの出力である．タスク 1 ではターンが反応のみである事後確率，タスク 2 ではターンがエピソードを含む事後確率にそれぞれ対応する．

上記のモデルにおいて，好感や話題への興味を中間層に対応させるための学習方法について述べる．そもそも好感や話題への興味は主観的であるため，学習ラベルを集めることが困難である．そこで，学習ラベルが少量であっても効率よく学習できるように，事前学習とファインチューニングを組み合わせた方法を提案する．はじめに，上記のニューラルネットワークの各層を個別に事前学習する．この時，前節で述べた対話コーパスの収集で得られたアンケート結果を用いる．ただし，アンケートは話題毎に行われたため，各話題内ではラベルの内容は同一であるとみなす．その後，入力 $\boldsymbol{o}$ とシステムの発話構成要素 $\boldsymbol{a}$ のラベルを用いてネットワーク全体をファインチューニングする．ファインチューニングには，好感や話題への興味のラベルを必要としないため，アンケート結果が得られていない他の対話データを追加することが可能である．ファインチューニングの際，事前学習の効果を保つために，事前学習で得られたパラメータ W_{pre} とファインチューニングで得られるパラメータ W との差のフロベニウスノルムを損失関数に追加する．

$$E'(W) = E(W) + ||W - W_{pre}||_F \tag{4}$$

4. 実験結果

既に述べたお見合い対話コーパスを用いて，5 分割交差検証により提案手法を評価した．比較手法は以下の 3 つである． 1 つ目は内部状態をモデル化しないもので，ユーザのふるまいから直接ロジスティック回帰によりシステムの発話構成要素を選択する（ベースライン）． 2 つ目と 3 つ目は提案モデルと同じネットワーク構造を有するが，2 つ目は事前学習，3 つ目はファインチューニングをそれぞれ行わない．表 1 と表 2 に結果を示す．提案手法と事前学習なしを比べると，提案手法による精度改善が見られた．このことは，中間状態のラベルを用いることの有効性を示している．また，事前学習のあとにファインチューニングを行うことで僅かではあるがさらなる精度改善がみられた．

表 1: タスク 1 の結果（反応のみ／他の要素も含む）

モデル	正解率
ベースライン	0.642
事前学習なし	0.606
ファインチューニングなし	0.650
提案手法	0.656

表 2: タスク 2 の結果（エピソード／質問）

モデル	正解率
ベースライン	0.567
事前学習なし	0.600
ファインチューニングなし	0.630
提案手法	0.632

5. お見合い対話システムへの実装

我々は，ERICA を用いたお見合い対話システムの実装を現在進めている‡．このシステムに，これまでに述べた発話構成要素を選択する機能の実装方法について述べる．入力として使用する特徴量は自動で抽出する必要がある．相槌，笑い，フィラーの検出はニューラルネットによる手法を用いる [5]．また，ユーザ発話の構成要素については対話行為の推定を行い，これで近似する．例えば，言明はエピソードに対応させる．また，お見合い対話は混合主導であるため，主導権およびターンテイキングの制御が必要である．さらに，発話構成要素を選択したのちに，これに基づく発話内容の生成も必要である．現状は人手で用意したものを用いているが，ニューラルネットによる学習ベースの手法を導入する予定である．

6. おわりに

本稿では，対話相手への好感に基づいてシステムの発話構成要素を選択する手法を述べた．階層的ニューラルネットワークにより対話相手への好感をその中間層で表現した．また，事前学習とファインチューニングの組合せにより予測精度が向上することを示した．最後に，発話構成要素を選択する機能をお見合い対話システムに実装するための方針について述べた．

謝辞 本研究は JST ERATO JPMJER1401 の支援を受けた．

参考文献

[1] Rosalind W. Picard. *Affective computing*. 1997.
[2] 田中滉己 *et al.* 初対面対話における好感の生成と発話構成要素の予測のモデル. 情処全国大会, 2018.
[3] Koji Inoue *et al.* Talking with ERICA, an autonomous android. *SIGdial*, 2016.
[4] Yasuharu Den *et al.* Two-level annotation of utterance-units in japanese dialogs: An empirically emerged scheme. *LREC*, 2010.
[5] Koji Inoue *et al.* Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. *APSIPA Trans. Signal & Information Processing*, 2018.

‡デモ動画 https://youtu.be/M3WLl4XcjMQ