エントレインメント分析に基づく応答文選択モデルの評価

水上雅博 †　吉野幸一郎 †§　Graham Neubig †‡　中村哲 †
†奈良先端科学技術大学院大学 情報科学研究科
§国立研究開発法人 科学技術振興機構
‡カーネギーメロン大学
{masahiro-mi, koichiro, s-nakamura}@is.naist.jp, gneubig@cs.cmu.edu

1　はじめに

エントレインメント（Entrainment）は，同調傾向やシンクロニー（Synchrony）とも呼ばれ，対話中の話者間において，話し方や声の調子などの振る舞いが同調，類似する現象を指す．この現象は，語彙 [2, 10]，統語構造 [15, 16]，文体 [13, 5]，韻律 [11, 4, 16, 7]，発音 [14]，ターンテイク [3, 1]，対話行為 [10] など，対話中の多様な要素で観測されることが知られている [8]．

エントレインメントは対話のタスク成功率や自然性，対話意欲（Engagement）と相関することが報告されており [12]，エントレインメントの分析を通して対話システムの性能や対話の質を測る試みが行われている．一方で，エントレインメントを対話中に考慮し，対話行為選択や応答文選択に利用する取り組みは未だなされていない．

そこで本研究では，エントレインメントを考慮した用例ベース対話システムの応答文選択手法を提案，評価する．具体的には，対話行為によって語彙のエントレインメントの度合いが変化する現象に着目し，ユーザの現在の対話行為とそれに対する応答におけるシステムの対話行為から次の発話のエントレインメントの度合いを推定し，適切なエントレインメントが生じている応答を選択する．対話中の各発話の対話行為に合わせてエントレインメントの度合いを変化させることで，ユーザと対話システム間でより高い自然性と対話意欲を持った対話の実現する．

2　関連研究

対話のエントレインメントを分析する研究として，聞き役対話における聞き役側の相槌（Backchannel）のエントレインメント傾向の分析がある．具体的には，ピッチ（F0）や声量（Power）の相関関係を示す研究 [7] や，発話のタイミングのみでなく，対話のターン管理や主導権についてエントレインメントが起きることを示す研究 [9] である．

著者らは，対話において重要な要素として，発話の意味や役割を示すに対話行為に着目した分析 [10] を行っている．この研究によって，対話行為の選択傾向においてもエントレインメントが生じ，また，語彙のエントレインメントが生じる度合いには発話の対話行為が非常に強く影響していることが分かっている．これらのことから，対話におけるエントレインメントでは対話行為を考慮してエントレインメントを行うかどうかを決定する必要があることがわかっている．

また，エントレインメントの分析を通して得られた知見をモデルや戦略に組み込む研究もなされている．例えば，エントレインメントを考慮することで発話タイミングを高精度に予測する研究 [3] や，エントレインメントを誘発することで，より音声認識の成功率を高める対話戦略を栄養する研究 [6] がある．さらに，これらを統合し，エントレインメントを利用して音声認識とターン管理の性能を向上する研究 [8] もある．

このように，ユーザのエントレインメントのモデル化や予測が行われる一方で，対話システムがユーザに対してエントレインメントを行うようなシステムは提案されてこなかった．

3　エントレインメントを考慮した応答文選択モデル

本論文では，対話行為と語彙のエントレインメント分析の結果に着目し，非タスク指向の対話システムがユーザに対して自然なエントレインメントを行うための応答文選択モデルを構築する．まず，応答文選択モデルの全体像を図 1 に示す．この応答文選択モデルでは，用例ベース対話モデル [17] の枠組みに沿って応答

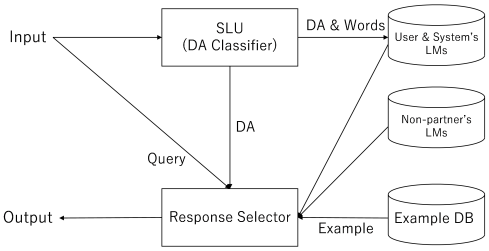

図 1: 提案法の全体図

文選択を行うが，応答文選択の際に入力文との類似度のみでなく，エントレインメントの度合いを考慮して応答文を選択する．応答文選択はユーザから与えられたユーザ発話 q' に対して，適当なシステム応答 r_j を持つ用例 $\langle q_j, r_j \rangle$ を用例データベース $\mathbf{e}$ から探索する処理である．この際，提案手法ではユーザ発話 q' とクエリ発話 q_j とのコサイン類似度 $\mathrm{Cos}(q', q)$ と，その応答文のユーザに対するエントレインメント度合いの妥当性 $\mathrm{Entr}(r_j|d)$ の二つに対してパラメータ λ_d による重み付けを行い，最大化を行うことで応答文を決定する．

$$\hat{r} = \underset{\langle q_j, r_j \rangle \in \mathbf{e}}{\mathbf{argmax}} \lambda_{da} \mathrm{Cos}(q', q_j) + (1 - \lambda_d) \mathrm{Entr}(r_j|d). \quad (1)$$

このエントレインメント度合いの妥当性 $\mathrm{Entr}(r_j|d)$ は，対話行為 d に応じた理想のエントレインメント指標比率と，対話中の話者とのエントレインメント指標比率の差から計算する．具体的には，ある応答候補 r_j ごとに計算されたエントレインメント指標比率 $R(V|d, P_{user}, P_{system+r_j})$ と，事前に計算された理想とするエントレインメント指標比率 $R_{ideal}(V|d)$ との差を最大が 1 になるように正規化し，その絶対値を 1 から引くことで，エントレインメントの妥当性 $\mathrm{Entr}(r_j|d)$ とする．

$$\mathrm{Entr}(r_j|d) = 1 - \left| \frac{\mathrm{R}(V|d, P_{user}, P_{system+r_j}) - \mathrm{R_{ideal}}(V|d)}{\max\{1 - \mathrm{R_{ideal}}(V|d), \mathrm{R_{ideal}}(V|d)\}} \right| \quad (2)$$

この指標は，理想とするエントレインメント指標比率に近い応答ほど 1 に近くなり，逆に過剰にエントレインメントを起こした場合やエントレインメントしなかった場合に 0 に近くなる．

エントレインメントの妥当性を計算するためには，応答候補ごとのエントレインメント指標比率 $R(V|d, P_{user}, P_{system+r_j})$ を計算する必要がある．まず，応答文選択を行う際，応答候補群の発話 r_j を対話システムの言語モデルにそれぞれ追加し，応答候補ごとの条件付き言語モデル確率 $P_{system+r_j}(w|d)$ を計算する．そして，応答候補ごとの条件付き言語モデル確率 $P_{system+r_j}(w|d)$ と，対話中のユーザの条件付き言語モデル確率 $P_{user}(w|d)$ からユーザとのエントレインメント指標 $\mathrm{En_{participants}}(V|d)$ を求める．なお，指標計算に利用する語彙 V は先行研究と同様に 25MFC(コーパス中で出現頻度の高かった 25 単語) とした．

$$\mathrm{En_{participants}}(V|d) = -\sum_{w \in V} \left| \mathrm{P_{user}}(w|d) - \mathrm{P_{system+r_j}}(w|d) \right| \quad (3)$$

同様に，応答候補ごとの条件付き言語モデル確率 $P_{system+r_j}(w|d)$ と，事前に用意された対話および無関係な話者の発話から学習された N 個の条件付き言語モデル確率群 $P_{non-partner_i}(w|d)$ $(i \in N)$ から非ユーザとのエントレインメント指標 $\mathrm{En_{non-partner_i}}(V|d)$ を計算する．

$$\mathrm{En_{non-partner_i}}(V|d) = -\sum_{w \in V} \left| \mathrm{P_{non-partner_i}}(w|d) - \mathrm{P_{system+r_j}}(w|d) \right| \quad (4)$$

最後に，ユーザおよび非ユーザのエントレインメント指標を比較することでエントレインメント指標比率 $R(V|d, P_{user}, P_{system+r_j})$ を求める．

$$\mathrm{R}(V|d, P_{user}, P_{system+r_j}) = \frac{1}{N} \sum_{i \in N} \begin{cases} 1 & \mathrm{En_{participants}}(V|d) > \mathrm{En_{non-partner_i}}(V|d) \\ 0.5 & \mathrm{En_{participants}}(V|d) = \mathrm{En_{non-partner_i}}(V|d) \\ 0 & \mathrm{En_{participants}}(V|d) < \mathrm{En_{non-partner_i}}(V|d) \end{cases} \quad (5)$$

これにより，理想的なエントレインメントの度合いに近いかどうかを判断することが可能になる．

4 実験

4.1 モデルの学習とコーパス

実験では，分析に用いたものと同じ Switchboard Dialogue Act Corpus を用いる．これには各発話に DASML タグセットに従って DA タグが与えられている．このコーパスをテスト用データ 115 対話と学習用データ 1045 対話の 2 つのデータに分割する．

まず，学習用データから，応答文選択モデルを構築する．はじめに，理想的なエントレインメント度合いを示す $\mathrm{R_{ideal}}(V|d)$ を，対話行為と語彙のエントレインメント分析の手法に基づいて計算する．次に，対話システムの用例データベース $\mathbf{e}$ を学習し，応答文選択モデルで必要なパラメータ λ_d を算出する．具体的には，学習用データの発話と応答のペアとして，クエリ

発話 q'_i とそれに対する応答 r'_i を $\mathbf{e}$ から取得し，それに対して以下の数式を解くことでパラメータ λ_d を対話行為ごとに算出する．

$$\lambda_d = \mathop{\mathbf{argmin}}_{\lambda} \sum_{\langle q'_i, r'_i \rangle \in \mathbf{e_{test}}} \sum_{\langle q_j, r_j \rangle \in \mathbf{c_i}} \left\{ \Big(\lambda \mathrm{Cos}(q'_i, q_j) + (1-\lambda)\mathrm{Entr}(r_j|d) \Big) - \mathrm{Cos}(r'_i, r_j) \right\}^2 \tag{6}$$

ここで，式中の応答候補 $\mathbf{c_i}$ は，クエリ発話 q'_i に対する類似度 $\mathrm{Cos}(q'_i, q_j)$ が高い順に 20 個を用例データベース $\mathbf{e}$ の中から選択する．ただし，選択の際は応答 r_j と実際の応答 r' が同じ対話行為であること，対象としているペア $\langle q'_i, r'_i \rangle$ は選択しないことを条件として追加した．実験は，この学習用データから構築された応答文選択モデルを用いて行う．

4.2 評価指標

本実験では，テスト用データの中のあるクエリ q' が与えらえた際に，応答文選択モデルが選択した応答 r_j とテスト用データの実際の応答 r' を比較し，客観評価指標を計算する．また，応答文選択の客観評価として以下の二つの指標を利用する．

一つ目は，応答の選択関数 $\big(\lambda_d \mathrm{Cos}(q', q_j) + (1-\lambda_d)\mathrm{Entr}(r_j|d)\big)$ と，選択された応答 r_j と実際の応答 r' との類似度 $\mathrm{Cos}(r', r_j)$ の二乗誤差であり，以下の式で表される．

$$\mathrm{EVAL_{MSE}} = \sum_{\langle q_j, r_j \rangle \in \mathbf{c_i}} \left\{ \Big(\lambda \mathrm{Cos}(q', q_j) + (1-\lambda)\mathrm{Entr}(r_j|d) \Big) - \mathrm{Cos}(r', r_j) \right\}^2 \tag{7}$$

この評価は，選択関数がどれだけ実際の応答と類似しているかを示す指標となる．

二つ目は，提案法によって選択された応答と実際の応答との類似度である．提案法によって選択される応答 $\hat{r}$ は，以下の式に従って選択関数が最大となる応答を探すことで得られる．

$$\hat{r} = \mathop{\mathbf{argmax}}_{\langle q_j, r_j \rangle \in \mathbf{c_i}} \Big(\lambda_d \mathrm{Cos}(q', q_j) + (1-\lambda_d)\mathrm{Entr}(r_j|d) \Big) \tag{8}$$

また，評価指標は選択された応答 $\hat{r}$ と実際の応答 r' のコサイン類似度を計算することで得られる．

$$\mathrm{EVAL_{COS}} = \mathrm{Cos}(r', \hat{r}) \tag{9}$$

この評価は，提案法によって選択された応答がどれだけ実際の応答に近いかを示す指標となる．

最後に，選択された応答のエントレインメント指標比率を示す．

$$\mathrm{EVAL_{ENTR}} = \mathrm{Entr}(\hat{r}|d) \tag{10}$$

これは応答文の評価となる指標ではないが，選択された応答がどの程度エントレインメントしているかを示す．

4.3 評価実験

提案法による有効性を検証するため，テスト用データの各対話に対して，順にクエリに対する応答を推定し，各評価指標の平均を計算した．また，提案法であるエントレインメントの適切さを考慮した応答文選択に加え，比較のためのベースラインとして，応答文選択モデルのパラメータ λ_d を常に 1（クエリの類似度のみを考慮する場合に相当）と，λ_d を常に 0（エントレインメントの適切度のみを考慮する場合に相当）に設定した場合の評価尺度を示す．結果を表 4.3 に示す．

評価実験の結果，相槌，ノンバーバル発話，yes/no 返答，ヘッジなどのエントレインメント分析で高いエントレインメント指標比率となった対話行為において，λ_d が 0 に近くなり，エントレインメントの適切度を強く考慮することで精度が向上することがわかった．特に，ヘッジにおいては，エントレインメントの適切度を考慮することで，実際の応答との類似度 $\mathrm{EVAL_{COS}}$ が 0.25 となり，クエリ発話の類似度のみを考慮した場合の 0.20 に比べて 25%向上した．

一方で，謝罪，軽視，第三者に対する言及などの対話行為を持った応答を行う際には λ_d が 1 に近くなり，エントレインメントの適切度よりも，用例とクエリ発話との類似度 $\mathrm{Cos}(q'_i, q_j$ が精度に強く影響を与えることがわかった．

全体としてエントレインメントの適切度を考慮して応答を選択することで，平均 $\mathrm{EVAL_{MSE}}$ は 0.079 となり，ベースラインの 0.10 に比べて改善した．平均 $\mathrm{EVAL_{COS}}$ はベースラインの 0.66 から，提案法を用いることで 0.67 となり，平均 $\mathrm{EVAL_{ENTR}}$ はベースラインの 0.56 から，提案法を用いることで 0.59 となった．これらのことから，提案法であるエントレインメントの適切度を考慮した応答文選択モデルを用いることによって，特にエントレインメント指標比率が高い対話行為において，ベースラインに比べて高い精度で応答を選択することが可能となったと言える．

	$\mathrm{EVAL}_{\mathrm{MSE}}$	$\mathrm{EVAL}_{\mathrm{COS}}$	$\mathrm{EVAL}_{\mathrm{ENTR}}$
類似度のみ考慮（ベースライン，$\lambda = 1$）	0.10	0.66	0.56
エントレインメントのみ考慮（$\lambda = 0$）	0.23	**0.67**	**0.59**
提案法（λ_d を対話行為 d ごとに設定）	**0.079**	**0.67**	**0.59**

表 1: 応答文選択モデルの評価

5 まとめ

本稿では，対話におけるエントレインメント分析の結果に基づいて，対話中のユーザに適切にエントレインメントすることが可能な応答文選択モデルを提案した．

評価実験を通して，エントレインメントの適切さを考慮した応答文選択を行うことで，選択関数と応答文の類似度間の事情誤差が減少し，選択された応答と実際の応答の類似度が向上した．これらの結果は，既存の応答文選択モデルに比べて提案法がより実際の応答に近い出力が可能であることを示している．また，提案法によって選択された応答は，エントレインメント分析の結果と同様に，ユーザに適切に同調した応答となった．

本研究の今後の課題として，客観評価尺度のみでなく，実際の対話と主観評価を通して提案法の有効性を検証する．

謝辞

本研究は，JST，CREST の支援を受けたものである．

参考文献

[1] Štefan Beňuš, Agustín Gravano, Rivka Levitan, Sarah Ita Levitan, Laura Willson, and Julia Hirschberg. Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, Vol. 71, pp. 3–14, 2014.

[2] Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 22, No. 6, p. 1482, 1996.

[3] Nick Campbell and Stefan Scherer. Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In *INTERSPEECH*, pp. 2546–2549, 2010.

[4] Rachel Coulston, Sharon Oviatt, and Courtney Darves. Amplitude convergence in children's conversational speech with animated personas. In *Proc. ICSLP*, Vol. 4, pp. 2689–2692, 2002.

[5] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pp. 745–754. ACM, 2011.

[6] Andrew Fandrianto and Maxine Eskenazi. Prosodic entrainment in an information-driven dialog system. In *INTERSPEECH*, pp. 342–345, 2012.

[7] Tatsuya Kawahara, Takashi Yamaguchi, Miki Uesato, Koichiro Yoshino, and Katsuya Takanashi. Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening. In *APSIPA*, pp. 392–395. IEEE, 2015.

[8] Rivka Levitan. Entrainment in spoken dialogue systems: Adopting, predicting and influencing user behavior. In *HLT-NAACL*, pp. 84–90, 2013.

[9] Rivka Levitan, Stefan Benus, Agust￥inGravano, Julia Hirschberg. Entrainment and turn-taking in human-human dialogue. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*, 2015.

[10] Masahiro Mizukami, Koichiro Yoshino, Graham Neubig, David Traum, and Satoshi Nakamura. Analyzing the effect of entrainment on dialogue acts. In *Proc. SIGDIAL*, 2016.

[11] Michael Natale. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, Vol. 32, No. 5, p. 790, 1975.

[12] Ani Nenkova, Agustín Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In *Proc. ACL*, pp. 169–172. Association for Computational Linguistics, 2008.

[13] Kate G Niederhoffer and James W Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, Vol. 21, No. 4, pp. 337–360, 2002.

[14] Jennifer S Pardo. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, Vol. 119, No. 4, pp. 2382–2393, 2006.

[15] David Reitter and Johanna D Moore. Predicting success in dialogue. 2007.

[16] Arthur Ward and Diane Litman. Measuring convergence and priming in tutorial dialog. In *University of Pittsburgh*, 2007.

[17] 水上雅博, Lasguido Nio, 木付英士, 野村敏男, Graham Neubig, 吉野幸一郎, Sakriani Sakti, 戸田智基, 中村哲. 快適度推定に基づく用例ベース対話システム．人工知能学会論文誌, 2016.