対話システムのための対話状況を制約とした応答変換の検討

千葉祐弥 [1]　東中竜一郎 [2]
[1]NTT コミュニケーション科学基礎研究所　[2] 名古屋大学 大学院情報学研究科
yuuya.chiba.ax@hco.ntt.co.jp　higashinaka@i.nagoya-u.ac.jp

概要

我々はこれまで，対話状況に応じて対話戦略を適応する対話システムの実現を目標として，日常会話を対象とした対話状況の推定と状況ごとの対話の分析を行ってきた．本研究では，日常会話のように多様に変化する状況の中で，状況に合わせて応答を切り替えるための応答変換手法を検討する．本稿では，検討手法による応答の変換例を示すとともに，応答変換モデルのエンコーダ出力とデコーダ出力を可視化することで，ネットワークがある程度適切に対話相手との関係を捉えて応答生成できていることを示す．最後に人間による応答の評価を行い，検討手法の性能と改善点について議論する．

1　はじめに

スマートスピーカやパーソナルアシスタントのような，日常的に使われる対話システムが一般的になるつつある．このようなシステムが人間どうしの日常会話に自然に参与できるようになるためには，単に自然な応答ができるだけではなく，対話の目的やシステム自身を含む対話参与者間の関係を正確に理解し，状況に合わせて応答を調整できる必要がある．

我々はこれまで，日本語日常会話コーパス (CEJC) [1] に含まれる対話を分析し，日常会話が対話の目的や対話のやり方に相当する 7 つの因子の組み合わせによって説明でき，かつ状況ごとにそれらの因子の重みが異なることを示した [2]．図 1 は CEJC に含まれる対話場面の例である．分析では，特に対話形式や対話相手との関係が対話に大きな影響を与えることが示唆された．例えば，家族との雑談においては物語的な会話が多く行われ，同僚や上司との会議においては丁寧さの度合いが高く問題解決を指向した対話が行われる傾向がある．したがって，対話システムは，これらの対話状況を考慮した応答生成を行うことで，より自然にユーザの会話に参与できるようになると考えられる．

対話形式: 雑談
場所: 自宅
活動: 食事
対話参与者の関係: 家族

対話形式: 用談相談
場所: 施設_その他
活動: 社会参加
対話参与者の関係: 社会関係

図 1　日本語日常会話コーパスにおける対話場面の例．写真は公開のためマスキングしたものである．

状況に合わせた対話制御に関しては，これまでも様々に検討されてきた．例えば，対話システムライブコンペティションでは 2019 年よりシチュエーショントラックが開催され，設定された状況の中で適切な対話を行う対話システムが様々に構築されている [3]．Utami and Bickmore [4] は，恋人同士のような親密な間柄に特有の会話現象を考慮した会話が可能なカウンセリング対話システムを構築している．また，対話相手との物理的距離を認識し，適切な距離を保つように調整するシステム [5] や，周囲の環境について対話を行う車載機器用音声対話システム [6]，与えられたシーンに関する受け答えが可能な対話システム [7, 8, 9] なども検討されている．しかしながら，これらの研究はあらかじめ指定された特定の状況下でシステムが動作することを想定しており，日常会話のように多様に変化する状況の中で，適切に応答を調整する対話システムに関してはほとんど検討されてこなかった．

本稿では，何らかの方法で決定されたシステム発話文を，指定された対話の状況に適した発話文へと変換する応答変換手法を検討する．この手法により，発話生成のたびに応答変換を行うことで，逐次変化する対話の状況に合わせた応答が可能となる．近年，文献 [10, 11, 12] などを筆頭に，大規模対話モデルの応答の流暢性と自然性が飛躍的に向上している．したがって，大規模に学習された対話モデルを

対話状況を表すラベルを付与した対話データで適応することで，流暢性と自然性が高く，かつ状況に依存した応答が生成できるようになると期待される．

2 日本語日常会話コーパス (CEJC)

実験データとして，日本語日常会話コーパス (CEJC) [1] を実験に用いる．CEJC は，日常生活の中における様々な活動とともに起こる自然な会話を収録したコーパスである．

音声データは個別の話者が装着した IC レコーダ (Sony ICD-SX734) と場面中央に置かれた IC レコーダ (Sony ICD-SX1000) によって収録されている．また，屋外や移動シーンに関しては Panasonic HX-A500 によって撮影された映像，その他のシーンに関しては全天球型カメラ (Kodak PIXPRO SP360 4K) と 2 台の GoPro Hero3+で撮影された映像が含まれている．全ての対話データには書き起こし文が付属しており，形態素情報や話者ラベル，発話時間などが詳細にアノテーションされている．加えて，各対話データには，対話形式，対話が行われた場所，対話とともに行われている活動，対話参与者の関係といった対話状況を表すラベルも付与されている．本稿の応答変換では，書き起こし文と，それに付随する対話状況ラベル，会話や話者の属性を記したメタ情報を用いる．

コーパスには 152 時間の対話データが収録されており，全体で 1,462 名の話者による 427 対話が含まれる．発話とトークンの総数は約 40 万発話，約 184 万トークンである．

3 状況を制約とした応答変換

3.1 問題設定

本研究の目的は，所与のシステム応答発話を，指定した状況に適した応答発話へと変換するモデルを学習することである．そこで，本稿では対話状況ラベルと応答発話文から抽出された内容語を入力として，応答発話文を出力するように応答変換モデルを学習する．これによって，入力発話文の内容を極力維持したうえで，状況によって異なる部分のみを変化させるように学習されたモデルが構築されることを期待する．また，指定する対話状況は特に対話への影響が大きい対話形式と対話参与者間の関係である．対話形式に関しては，対話に対して付与された 3 種類のラベルを用いる．すなわち，会議会合，用談相談，雑談である．対話参与者間の関係には，話者からみた受け手との関係 (e.g.,「親」や「子」など) を用いる．

表 1 話者関係行例の例 (会話 ID: T009_020)

	IC01	IC02	IC03	IC04	IC05	IC06
IC01	本人	親	親	祖父母	祖父母	祖父母
IC02	子	本人	夫婦	親	義父母	義父母
IC03	子	夫婦	本人	義父母	親	親
IC04	孫	子	婿嫁	本人	親族	親族
IC05	孫	婿嫁	子	親族	本人	夫婦
IC06	孫	婿嫁	子	親族	夫婦	本人

応答変換モデルへの入力は下記の通りである．

$$[\text{状況}]s_i[\text{SEP}]r_i[\text{応答}]c_1^i[\text{SEP}]c_2^i[\text{SEP}]\cdots[\text{SEP}]c_n^i$$

ここで，s_i は i 番目の発話に付与された対話形式であり，r_i は i 番目の発話における，話者から見た受け手との関係である．s_i と r_i は自然言語によって表現される．$\boldsymbol{c}^i = (c_1^i, c_2^i, \cdots, c_n^i)$ は，発話から抽出された内容語の系列である．[状況] や [応答]，[SEP] は，状況や内容語の区切りを示す特殊トークンである．例えば，友人同士の用談相談の「ちょっと寒くなってきた」という発話からは，“[状況] 用談相談 [SEP] 友人知人 [応答] ちょっと [SEP] 寒い [SEP] なる [SEP] くる” という入力が構築される．出力は元の発話と同一である．学習時には，CEJC から抽出された自然発話を用いる．

3.2 話者間の関係の認定

対話参与者の関係に関しては，CEJC には情報提供者を中心にした各話者への関係だけが与えられている．そこで，会話・話者のメタ情報のうち，会話の説明や，年齢，職業に基づいて対話参与者間の関係を認定した．この方法で，一部の同僚間の上下関係などを除き，全ての対話参与者間の関係を概ね一意に定めることができる．表 1 に会話 ID T009_020 の対話における対話参与者間の関係の行列を示す．

また，CEJC に収録された対話は多人数対話が多く含まれるため，必ずしも発話の受け手が明確ではない．そこで，本稿では当該発話の次に発話権を取得した話者を発話の受け手とみなした．表 2 に対話例と話者間の関係を示す．表中太字で示したように一部の発話で認定された関係に誤りがあるが，概ね正しい受け手が認定できていることがわかる．

表 2 実験に用いる対話の例 (会話 ID: T019_005b)．太字は認定された受け手が誤っている発話を示す．

話者関係	発話文
部下 → 上司	こないだのその中古の話しで中古車に結構流れてるってったじゃないですか
部下 → 上司	だから新車は売れないんじゃないすか
部下 → 上司	そこの自分たちで首絞めてるってゆうか
上司 → 部下	ちょっと頭のいいお客さんだとえー中古登録済み未使用車
部下 → 上司	ン全然そっちのほうがお得な感じしますもんね
上司 → 部下	三桁違うらしいからね
部下 → 上司	すごいすね
上司 → 部下	だって輸入車の登録台数は過去んー最高とかって
部下 → 上司	その裏ではこうやってまディーラーのにこう無理にこう台数を目標達成だとか言って
上司 → 部下	**裏がありまくりですよね**

4 実験条件

4.1 実験データ

実験データは，文献 [2, 13] と同様の基準で学習セット，開発セット，テストセットに分割した．すなわち，状況のラベルの分布がそれぞれのセットで可能な限り近くなるように分割した．対話数は，それぞれ 299 対話，84 対話，44 対話であった．それぞれの対話から発話文を抽出し，学習に用いた．各発話は UniDic[1] を辞書とした MeCab[2] を用いて分割し，名詞，代名詞，形状詞，副詞，感動詞，動詞，形容詞と解析された単語の原形を内容語とした．

内容語が 3 つ未満の発話と，トークン長が 50 を越える発話は除外した．受け手との関係に関しては，学習データ中の出現頻度が N 回未満の関係と，「本人」となる発話を除外した．ここで，$N = 1,000$ とした．そのうえで，各関係の出現頻度が概ね 1 万発話となるようにアップサンプリングした．

4.2 大規模対話モデルの学習

本稿では NTT が公開する大規模対話モデル [10] のうち，事前学習モデルをベースのモデルとして用いた．学習時のバッチサイズは 64 であった．損失関数は Cross Entropy 損失とし，最適化手法は Adafactor とした．学習率は 1e-04，warmup steps は 500 とした．また，最大エポック数は 1,000 と設定し，開発セットに対する損失が最少となるモデルを選択した．モデルのファインチューニングには fairseq [14] を用いた．生成時には $k = 50$ とした top-k サンプリングを行い，ビーム幅は 80 とした．

1) https://clrd.ninjal.ac.jp/unidic/
2) https://taku910.github.io/mecab/

表 3 関係のラベルによる変換結果の例 (入力発話文:「たぶん世界で一番安い」)

話者関係	変換発話文
部下 → 上司	でもたぶん世界一安いんですよ
同僚 → 同僚	たぶん世界一安いですよ
上司 → 部下	でたぶん世界一安いからね
子 → 親	たぶん世界一安いじゃん
親 → 子	たぶん世界一安いじゃん
サービス受領者 → 提供者	たぶん世界一安いからね
サービス提供者 → 受領者	たぶん世界一安いからね

4.3 人手評価

生成したサンプルに関して人間による評価を行った．テストセットのうち，内容語の個数が 3 以上 8 未満となるものからランダムに 100 文選択し，対話相手との関係のみ変化させて応答を変換した．変換には，出現頻度上位 10 位までの関係を用いた．この時，言い直しなどで同じ内容語が 2 回以上出現するサンプルや，内容語の長さが 1 のものは除外した．

実験では，5 人のワーカが評価を行った．5 人のワーカは，「適切性: システムが関係を反映した応答ができているか」と「自然性: 日本語としての自然性」に関して，5 段階 (1: 全くそう思わない，5: とてもそう思う) で評価した．

5 実験結果

5.1 生成例

学習されたモデルによる応答変換の例を表 3 に示す．入力発話文は「たぶん世界で一番安い」であり，抽出された内容語は「たぶん」，「世界」，「一番」，「安い」であった．また，対話形式については用談相談を指定した．表より，この例では，応答変換モデルは入力発話文の内容を維持しつつ受け手との関係によって異なる応答を生成していることがわかる．特に受け手が上司や部下，親子の場合は語尾に違いが見られ，それぞれ適切な丁寧さの発話が生成できている．一方で，店員と客の関係を表す，サービス提供者と受領者の間では同じ発話が生成されている．これは，データの中に床屋など，店員もカジュアルな発話を行う対話が含まれている影響であると考えられる．

表 4　関係ごとの評価値の平均スコア (5 段階)．太字は 4 以上，斜体は 3 以下のスコアを示す．

	友人知人	同級生	夫婦	親	子	同僚	仕事関係者	部下	上司	生徒	平均
適切性	3.97	**4.18**	**4.18**	**4.10**	**4.15**	3.91	3.06	3.80	*2.35*	3.33	3.70
自然性	*2.93*	3.28	3.30	3.35	3.35	3.25	3.23	3.10	*2.93*	3.13	3.19

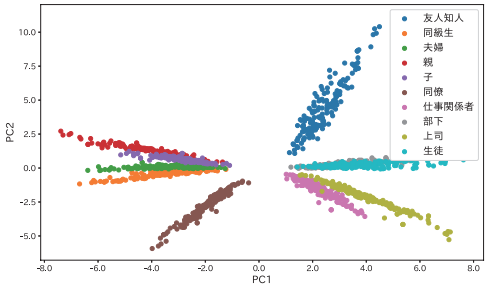

図 2　主成分分析によるエンコーダ出力の可視化結果．凡例は話者から見た受け手との関係を示す．

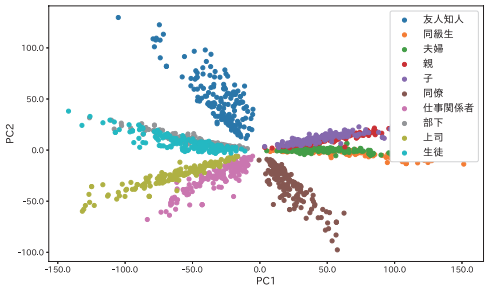

図 3　主成分分析によるデコーダ出力の可視化結果．凡例は話者から見た受け手との関係を示す．

5.2　エンコーダ・デコーダ出力の分析

エンコーダとデコーダが関係の違いをどのように捉えているかを調査した．テストセットからランダムに選択された 500 文に対して，頻度上位 10 位までの関係を用いて応答を変換した．各発話から，エンコーダとデコーダの隠れ層の出力を取り出し，可視化した．エンコーダに関しては，最終層の出力を系列に対して平均した．ベースモデルの隠れ層の次元数は 1,920 次元であるため，抽出される表現ベクトルも 1,920 次元である．一方で，デコーダに関しては，最終層の出力についてまずビームに対して平均をとり，さらに系列に対して平均をとることで 1,920 次元の表現ベクトルを抽出した．また，内容語が共通の発話間で平均をとり，それを減算することで発話内容の影響を除外した．その後，それぞれの表現ベクトル集合に対して主成分分析を行い，2 次元に圧縮した．

エンコーダ・デコーダの出力の可視化結果を図 2, 3 に示す．それぞれの図において，関係によって分布が分かれていることが見て取れる．特に，家族と仕事関係者・上司はほぼ同軸上の反対の領域に分布しており，親しみの違いがある程度捉えられていることが示唆される．一方で，図 2 と図 3 を比較すると，エンコーダ出力では差のあった親や子，夫婦や同級生などはデコーダ出力では重なっており，これらの関係においては発話の表層的な表現にあまり差がないことが示唆される．また，友人知人と同僚はその他の関係と直行しており，家族や上下関係のある間柄とは異なった発話表現が得られている可能性が示唆される．

5.3　人手評価の結果

表 4 に評価の結果を示す．表中の数値は関係ごとの評価値の平均である．5 人のワーカの全体の平均値は，適切性が 3.70，自然性が 3.19 であった．

適切性に関しては，仕事関係者と上司が特に低かった．これらの関係は図 3 上で親しみを表すと考えられる軸上に存在する．そのため，モデルは受け手との親しさの違いを捉えてはいるものの，丁寧さや敬語の表出は十分にできていないことが示唆される．自然性に関しては友人知人や上司において特に低かった．友人知人への発話ではカジュアルな発話に現れやすいフィラーの断片が不自然な箇所で出現しやすく，上司への発話では「です」や「っすか」といった語尾の接続が不自然になりやすかった．

評価者からは，自然性と適切性を切り分けて評価するのは難しいといった指摘があり，自然性の改善が適切性の評価の向上にも貢献すると考えられる．

6　おわりに

本研究では，対話状況に合わせた応答生成が可能な対話システムの構築を目指し，特に対話形式と対話相手との関係を制約とした応答変換モデルを学習した．実験では，モデルの変換例を示すとともに，エンコーダ出力とデコーダ出力を可視化することで，モデルが指定した状況の違いを捉えていることを示した．人手による評価実験では，ある程度状況に適した応答ができていることが示唆されたが，自然性については課題が残った．

今後は，発話スタイル変換 [15, 16] などの技術を駆使することで，応答の自然性の向上を目指す．

謝辞

本研究は科研費（JP19H05692）の助成を受けたものである．また，日本語日常会話コーパスの利用を承諾いただいた国立国語研究所に感謝する．

参考文献

[1] Hanae Koiso, Yasuharu Den, Yuriko Iseki, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, and Yasuyuki Usuda. Construction of the corpus of everyday Japanese conversation: An interim report. In **Proc. LREC**, pp. 4259–4264, 2018.

[2] Yuya Chiba and Ryuichiro Higashinaka. Analyzing variations of everyday Japanese conversations based on semantic labels of functional expressions. **ACM Transactions on Asian and Low-Resource Language Information Processing**, 2022.

[3] 吉川克正, 川本稔己, 山崎天, 水本智也, 小林滉河, 大萩雅也, 佐藤敏紀. シチュエーションに合わせたシナリオ誘導と HyperCLOVA を利用した応答生成によるハイブリッド対話システム. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 (第 96 回), pp. 124–129, 2022.

[4] Dina Utami and Timothy Bickmore. Collaborative user responses in multiparty interaction with a couples counselor robot. In **Proc. HRI**, pp. 294–303, 2019.

[5] Dan Bohus, Sean Andrist, and Eric Horvitz. A study in scene shaping: Adjusting F-formations in the wild. In **Proc. AAAI Fall Symposium**, pp. 1–7, 2017.

[6] Teruhisa Misu. Situated reference resolution using visual saliency and crowdsourcing-based priors for a spoken dialog system within vehicles. **Computer Speech & Language**, Vol. 48, pp. 1–14, 2018.

[7] Huda Alamri, Vincent Cartillier, Abhishek Das, et al. Audio visual scene-aware dialog. In **Proc. CVPR**, pp. 7558–7567, 2019.

[8] Shachi Kumar, Eda Okur, Saurav Sahay, Jonathan Huang, and Lama Nachman. Leveraging topics and audio features with multimodal attention for audio visual scene-aware dialog. **arXiv preprint arXiv:1912.10131**, 2019.

[9] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2Text: End-to-end learning of video-based text generation from multimodal inputs. **arXiv preprint arXiv:2101.12059**, pp. 1–14, 2021.

[10] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of Transformer-based Japanese chit-chat systems. In **Proc. SLT**, 4-1-17-TLP, 2023.

[11] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. **arXiv preprint arXiv:2208.03188**, 2022.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Proc. NeurIPS**, Vol. 33, pp. 1877–1901, 2020.

[13] Yuya Chiba and Ryuichiro Higashinaka. Dialogue situation recognition for everyday conversation using multimodal information. In **Proc. INTERSPEECH**, pp. 241–245, 2021.

[14] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. Fairseq: A fast, extensible toolkit for sequence modeling. **arXiv preprint arXiv:1904.01038**, 2019.

[15] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. **Computational Linguistics**, Vol. 48, No. 1, pp. 155–205, 2022.

[16] Atsumoto Ohashi and Ryuichiro Higashinaka. Adaptive natural language generation for task-oriented dialogue via reinforcement learning. In **Proc. COLING**, pp. 242–252, 2022.