The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

end-to-end 学習を用いたマルチモーダル多人数会話における対話ロボットの行動ターゲット生成

end-to-end training based action type generation for multi-party conversation robot

片山颯人 *1 藤江真也 *2 小林哲則 *1
Hayato Katayama Shinya Fujie Tetsunori Kobayashi

*1早稲田大学 Waseda University
*2千葉工業大学 Chiba Institute of Technology

Action generation method for multi-party conversation system using multi-modal information based on machine learning technique is proposed. Since data for conventional end-to-end training contains concrete system utterance contents as output, systems depend on tasks and domains. The study proposes end-to-end system that generates an action type, which is abstracted action. An action type is independent of conversation task, so the system has high versatility. Data collection is conducted under Wizard-of-Oz manner, where a wizard chooses appropriate action type and the utterance contents are generated from the existing modules automatically. The result of the preliminary experiment conducted using the data collected shows effectiveness of our framework.

1. はじめに

近年，スマートスピーカや会話ロボットなどが普及し，複数のユーザがそれらを囲んで会話を行う機会が増えた．このような環境で動作する多人数会話システムが注目されている [Matsuyama 15, Zhang 17, Shi 18b]．システムが多人数会話に参加するためには，参与構造理解や複数話者による談話構造理解など一対一の対話では生じない複雑な現象を扱う必要がある．さらに，そのような現象を扱いつつシステムの行動を適切に決定するためには，音声や言語にとどまらず，視覚を含むマルチモーダルな情報の利用が必須である．一方，扱う情報が増えることによって，行動決定に至るまでに様々な認識・理解モジュールを組み合わせることが避けられず，システムは大規模化し，手作り（Hand-crafted）のルールによる動作決定で適切な行動生成を実現することは極めて厳しい．

このような背景の下，近年では人同士や人と対話システムとの対話データを用いた機械学習により，入力から出力までを機械学習に任せる end-to-end なシステムが注目されている [Constantin 18, Shi 18a, Li 17, Vinyals 15, Serban 15]．end-to-end システムは，参与構造理解や談話構造理解など状況理解の枠組みを陽に扱う必要がなく，データ・ドリブンでシステムを構築できる．データ収集の際は，裏で人がシステムを操作することによってシステムの適切な出力のログを取得することのできる Wizard of Oz（WoZ）法が用いられる [Budzianowski 18, DeVault 15, Johansson 16] ．こうして収集されたデータを用いることで，具体的なルールを人手で決めることなく適切な行動生成ができることが期待される．

通常の end-to-end システムは，その学習に，入力としてユーザから得られる音声や画像情報，出力としてシステムの行動，すなわち具体的な発話内容や，身体をもつエージェントの場合は身振り手振りなどのコマンドを与える．WoZ 法で収集されるデータは，人である Wizard がシステムを操作した結果に基づく具体的な入出力ペアとなる．この際に利用される出力は，具体的であるがゆえ，会話のタスクおよびドメインに依存している．そのため，このデータで学習したシステムを汎用的に利用することは難しく，タスク毎にそのタスクに沿ったデータの収集を行いシステムの構築が必要になる．本研究では，機械学習を用いた多人数会話システムの行動決定手法において，タスクおよびドメイン依存性を低減した汎用的な枠組みを提案する．具体的には，システムの具体的な発話や振る舞いではなく，抽象化された行動タイプを出力とする方法である．質問応答システムが応答する際の「受動的な行動」や，雑談対話システムがユーザに対して自ら発話を行う「能動的な行動」のように，行動タイプはシステムの発話に伴って存在するものである．これらは抽象的であるために会話のタスクやドメインに共通して存在する．よって抽象的な行動タイプまでを end-to-end で決定する枠組みを構築することで，タスクやドメインへの依存性の低い，汎用性が高いシステムになることが期待される．

WoZ 法を用いたコーパス収集では，Wizard の操作がシステムの行動決定に直接つながる．そのため，Wizard がシステムの操作決定に迷いをもつことや，誤った操作をすることは避けるべきである．通常の end-to-end システムは発話内容まで Wizard が選択することから操作の遅延が発生し，発話のタイミング制御が困難になることもある [DeVault 15]．また，Wizard の操作を抽象的すぎる簡易的なコマンド操作に限定した場合，そのコマンドに対応したシステム発話文の生成が困難となる．そこで本研究では具体的な発話内容は別途自動的に計算しておき，Wizard は抽象化されたコマンドを選択することによって簡便に操作することが可能となるシステムを構築した．

本稿では，2. で提案システムの詳細を説明し，3. では WoZ 法を用いた多人数会話収集システムの設計と構築について説明し，実際に収集したデータについて述べる．4. では収集したデータを用いて対話システムの顔向きを決定するモデルの実験とその結果を報告する．最後に 5. にて本稿の結論を述べる．

2. end-to-end 学習を利用した行動タイプ決定システム

提案する行動決定システムの概観を図 1 に示す．モジュールは大きく分けて，行動タイプ決定部と発話内容決定部からなる．それぞれの詳細について述べる．

連絡先: 片山颯人，早稲田大学，東京都新宿区早稲田町 27，katayma@pcl.cs.waseda.ac.jp

2.1 行動タイプ決定部

行動タイプ決定部では，ユーザの音声や画像の系列から，システムが出力する具体的な行動を抽象化して分類した行動タイプのうち，その時点でシステムが取るべきものを決定し，その宛先（行動を行うターゲット）とともに出力する．本研究では，この部分を機械学習モデルで実現することにより，end-to-end 学習を部分的に利用した行動決定システムを実現する．本研究で設定した行動タイプを，表 1 に示す．基本的に行動タイプは従来研究における Dialogue-Act に対応するものである．後述する WoZ 法におけるデータ収集では，Wizard はこの行動タイプを選択することにより システムを操作する．

2.2 発話内容決定部

発話内容決定部では，行動タイプ決定部で決定された行動タイプに応じて，その時点で適切なシステムの発話内容を出力する．この構成は，ドメインに非依存の行動決定部とドメインに依存する発話内容決定部をそれぞれ独立に機能させることができる．言い換えると，既存の言語処理モジュール（発話内容決定部）を使ってシステムを構築することも可能であり，会話のドメインが変わっても行動決定部は再学習する必要がない．

本研究ではユーザが見に行く映画を決定するタスクを対象とする．ここでは，従来研究 [Matsuyama 15] で用いたものと同様の発話内容生成システムを利用する．すなわち，ユーザの音声から音声認識によって得られる言語情報からキーワード抽出を行い，行動タイプごとに用意されたテンプレートを用いて具体的な発話内容を生成する．

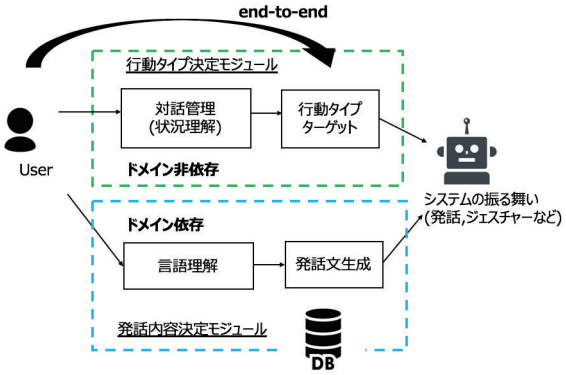

図 1: 提案システム

表 1: 対話ロボットの行動タイプ

行動タイプ	概要	例
能動的行動	自ら発話を行う	推薦，質問
受動的行動	話者からの発話に返答する	応答
非言語的行動	話者の発話に頷く	頷き
なし	何もしない	-

3. WoZ 法を用いた会話コーパス収集

3.1 WoZ System

2 で述べた行動タイプ決定部を機械学習で実現するために，入出力ペアのデータが相当数必要となる．本研究ではこれを，人である Wizard がシステムを操作する Wizard of Oz 法によって得られるログとして収集する．これにより，データ収集と同時に理想的な入出力ペアが獲得できるため，データ収集後にアノテーションを行う必要がない．通常の WoZ システムでは，Wizard は環境情報（ここではユーザの発話内容や振る舞いの様子）を得ながら，会話システムの具体的な行動や発話内容を生成する．一方，本研究では抽象化された行動タイプを選択する．具体的な発話内容は，発話内容決定部により自動的に生成されたものを利用する．収集の様子を図 2 に，Wizard が利用するコントローラのインターフェースを図 3 に示す．

本研究では会話システムのプラットフォームとして，市販されているロボット SOTA[*1] を用いた．また，ロボットの前にいる 2 名のユーザが，お互いに相談をしながら見に行く映画を決める日常的に起こりうるタスクを想定した．ロボットは，ユーザからの映画に関する質問に答えたり，必要に応じて自ら映画を推薦したりすることを想定している．

Wizard は図 3 に示したコントローラに表示されているボタンを押すことで，行動タイプ決定部の出力を模擬する．各ボタンは行動タイプの種類を表しており，非言語的な行動として「Look」「Nod」を，能動的な行動として「推薦」「詳細」「質問」，受動的な行動として「応答」「Yes」「No」「Unknown」を設定している．推薦や質問は少なからずタスク依存性が存在するが，多くのタスクに共通して存在しているものであるため，ドメイン依存性の影響がないと判断した．Wizard がこのボタンを選択することで，そのボタンに応じた発話内容をデータベースから参照することでシステム発話を行う．図に示したボタンは 2 名のユーザそれぞれについて別のものが画面上に表示されており，行動のターゲットはどちらのボタンを押したかによって指定されるものとする．Wizard がボタンを押したタイミングを，その行動タイプの行動を出力すべきタイミングとして記録するため，Wizard の判断が遅れることは好ましくない．Wizard の操作はボタンを 1 度押すのみにとどめており，遅延は最小限に抑えられるようにした．

3.2 データ収集

実際の収集は，親しい間柄のユーザ 2 名（A，B と呼ぶ），ロボット 1 体で対話を行った様子を収録した（図 2）．各話者が装着したピンマイクから音声を取得し，対話ロボットの頭部に設置した web カメラから映像を取得する．Wizard は web カメラによる画像情報と壁越しに聞こえる音声を基に対話ロボットの行動選択を図 3 の中から行う．実際の対話の一例を

図 2: コーパス収集の様子

∗1 https://sota.vstone.co.jp/home/

図 3: Controller of Wizard

表 2: 対話の一例

発話者	宛先	発話内容	Wizard の選択
A	S	それの監督は誰なの	-
S	A	監督は是枝裕和だよ	応答
A	-	聞いたことある	-
B	A	有名だよね	Look（B）

表 2 に示す．システムの適切な行動タイプ決定モデルを学習するために，Wizard による各行動タイプの生成は表 3 に沿って決定している．ユーザには，ロボットが公開中の映画情報を持っているので自由なことを聞いて見たい映画を決めるように指示をした．会話ので使用した言語は日本語であり，実験参加者は研究室に所属する 20 代の男女学生 16 人で， 40 対話分（総時間 150 分）のデータを収集した．

4. 実験

4.1 行動ターゲット推定

3.2 で収集したデータを用いて，対話ロボットの行動決定がどの程度できるかを確認するために，行動のターゲットが二人のユーザのうちどちらにあるかを識別する実験を行った．ネットワークの概観図を図 4 に示す．音声情報を元に顔向きを決定する場合，話者が誰も話をしていない場合や 2 人以上話をしている場合に顔向きを決定することができない（声の大きさは話者により異なるため，それのみでは判断できない）．そこで画像情報の中でも，現在の speaker の特定や発話先の推定として有効な視線の情報を含めたマルチモーダルな情報を用いる．さらに，行動ターゲットを end-to-end で推定するために，各モーダル情報に対してサブタスクを設定しニューラルネットワークによって各推定に有効な特徴量が抽出できることを期待する．x_t^s ， x_t^{img} はそれぞれ音声，画像の入力を表し，音声に対しては時系列情報を考慮した LSTM を用い（ h_t^s は LSTM 層の出力 ），画像に対しては畳み込みニューラルネットワーク

表 3: Wizard の行動タイプ選択基準

行動タイプ	行動基準
能動的行動 （推薦，質問，詳細）	全てのユーザが 発話をしていない時
受動的行動 （応答,Yes,No,Unknouuwn）	ユーザからシステムに 対して質問した時
非言語的行動 （Look）	他のユーザが 発話ターンを取得した時
非言語的行動（Nod）	ユーザが話している時

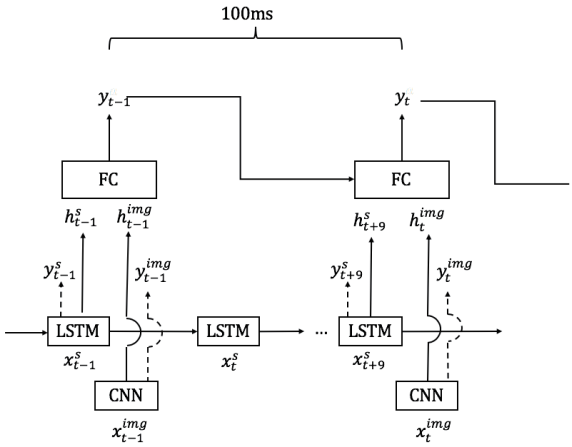

図 4: ネットワーク概観図

（ h_t^{img} は畳み込み後の出力）を用いることで各モーダル情報の統合を行う．ロボットの行動のターゲットを決めるために有効な情報として，ユーザが発話をしているかどうか（ y^s は発話有無の出力 ），またその発話が誰に対して（他方のユーザに向けられているか，ロボットに向けられているか）行われているか（ y^{img} は視線有無の出力）が 挙げられる．提案するネットワークは，これらの推定を内部に埋め込み，マルチタスク学習によって行動タ ーゲットと同時に学習する．発話をしているかどうかは，入力音声に対して Voice Activity Detection を行うことでわかる．また，発話がどちらに向けられているかは，ユーザの視線方向で推定することが可能である．

4.2 特徴量抽出

入力は 3.2 で収集したユーザ音声とロボットのカメラから得られた画像 である．音声の特徴量は，10ms ごとに算出される energy,F0（基本周波数）,MFCC と，それらの差分を用いた．差分の抽出に は過去 50 フレーム分の回帰係数を用いた．画像からは視線の情報を抽出するために，dlib[*2] を用いて，顔の検出を行う（図 5）．抽出された顔形状をもとに上 1 / 3 部分を切り出し目元の画像を入力とした．画像情報は 100 ms 毎に抽出されるため，音声とフレームレートが異なる．そのため，音声が 10 フレーム得られるたびに画像情報と統合し，行動ターゲットの推定を行うことで解決した （図 4）．行動ターゲットは，Wizard の操作に基づき，ロボットの顔方向をユー ザのどちらに向けるべきかを決めたものである．発話状態は，WoZ のシステム動作時に利用した，我々が独自に開発した音声認識システムによる発話区間検出の結果に基づき発話中，非発話中いずれかの状態にしたものである．視線方向は，本システムとは別途開発した視線方向認識器の結果を，カメラ方向を見ている，見ていないで決定している．なお，画像情報はロボット頭部に搭載したカメラから得られるものであり，一度に取得できるのは片方のユーザの顔画像だけである．したがって，補助情報としてロボットがどちらの方向を見ているか（前フレームの行動ターゲットの出力結果がなんであったか）入力として利用している．

4.3 実験結果

収集したデータを用いて学習，識別の実験を行った．提案手法に加え，2 つの手法を用いて比較を行った．1 つ目は発話状態推定と視線方向推定を事前に学習して，それぞれの推定結果

*2 http://dlib.net/

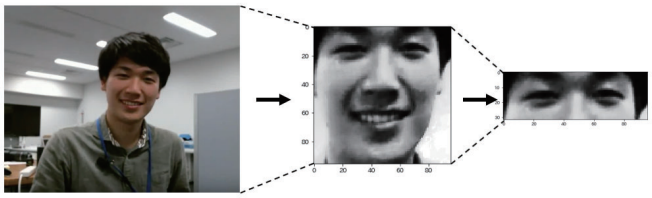

図 5: 目部分の抽出方法

表 4: 推定結果

入力	Accuracy[%]
事後確率	62.07
個別学習	92.08
全体学習	92.34

（それぞれの出力の事後確率）を入力として行動ターゲットを決定する手法，2 つ目は，個別に学習した各推定器の中間出力を抽出して入力とし，行動ターゲットを推定するモデル（すなわちマルチタスク学習でないもの）である．評価は， 40 対話のデータを 10 分割交差検証で行った. 結果を表 4 に示す.

これより，対話の状況を推定したのちにロボットの顔向きを決定するモデルよりも，同時に学習したモデルの方が性能が良いと言える. 図 6 に結果の一部を可視化したものを示す. 縦軸がユーザ B がターゲットである尤度を確率値で表したものである. 本システムではターゲットは常に A か B のどちらかに向けられるため，この値が低ければ A がターゲットらしいということになる．この図からマルチタスク学習による手法が，顔向きの切り替わりを最も柔軟に捉えられていることが確認できる．

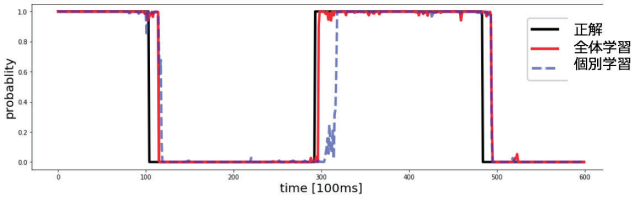

図 6: 結果の可視化

5. まとめ

マルチモーダル多人数会話におけるロボットの行動を生成するための手法としてユーザ入力からシステムの抽象的な行動タイプまでを end-to-end 学習する行動生成手法を提案した. WoZ 法を用いて収集されるデータを用いた機械学習を適用する手法として，従来の end-to-end システムの最大の欠点であるタスク・ドメイン依存性の問題を， 出力を行動タイプという形に抽象化することによって解決した．また，多人数会話システムの顔向き決定を行う行動ターゲット推定を通じて本システムの有効性を確認した．今後の予定として，実際の行動タイプを出力するシステムを構築するため，音声認識結果を利用した行動生成タイミングの決定モデルの作成と，その評価が挙げられる．

参考文献

[Budzianowski 18] Budzianowski, P., Wen, T., Tseng, B., Casanueva, I., Ultes, S., Ramadan, O., and Gasic, M.: MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling, *CoRR*, Vol. abs/1810.00278, (2018)

[Constantin 18] Constantin, S., Niehues, J., and Waibel, A.: An End-to-End Goal-Oriented Dialog System with a Generative Natural Language Response Generation, *CoRR*, Vol. abs/1803.02279, (2018)

[DeVault 15] DeVault, D., Mell, J., and Gratch, J.: Toward Natural Turn-Taking in a Virtual Human Negotiation Agent, in *AAAI Spring Symposia* (2015)

[Johansson 16] Johansson, M., Hori, T., Skantze, G., Hothker, A., and Gustafson, J.: Making Turn-Taking Decisions for an Active Listening Robot for Memory Training, in *SOCIAL ROBOTICS, (ICSR 2016) :*, No. 9979 in Lecture Notes in Artificial Intelligence, pp. 940–949 (2016), QC 20170125

[Li 17] Li, X., Chen, Y., Li, L., and Gao, J.: End-to-End Task-Completion Neural Dialogue Systems, *CoRR*, Vol. abs/1703.01008, (2017)

[Matsuyama 15] Matsuyama, Y., Akiba, I., Fujie, S., and Kobayashi, T.: Four-Participant Group Conversation: A Facilitation Robot Controlling Engagement Density as the Fourth Participant, *Computer Speech Language, 33(1):1-24.* (2015)

[Serban 15] Serban, I. V., Sordoni, A., Bengio, Y., Courville, A. C., and Pineau, J.: Hierarchical Neural Network Generative Models for Movie Dialogues, *CoRR*, Vol. abs/1507.04808, (2015)

[Shi 18a] Shi, W. and Yu, Z.: Sentiment Adaptive End-to-End Dialog Systems, *CoRR*, Vol. abs/1804.10731, (2018)

[Shi 18b] Shi, Z. and Huang, M.: A Deep Sequential Model for Discourse Parsing on Multi-Party Dialogues, *CoRR*, Vol. abs/1812.00176, (2018)

[Vinyals 15] Vinyals, O. and Le, Q. V.: A Neural Conversational Model, *CoRR*, Vol. abs/1506.05869, (2015)

[Zhang 17] Zhang, R., Lee, H., Polymenakos, L., and Radev, D.: Addressee and Response Selection in Multi-Party Conversations with Speaker Interaction RNNs, *arXiv e-prints* (2017)