小特集
Special Feature

［身近になった対話システム］

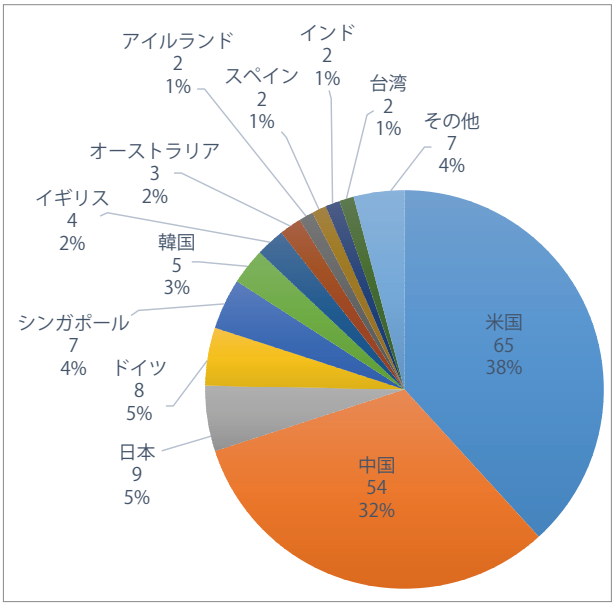
5 対話システムでは今何が問題になっているのか？

基 専

東中竜一郎　名古屋大学
光田　航　NTT

対話システム研究を俯瞰する

人間と会話を行うコンピュータのことを対話システムと呼ぶ．スマートフォン上の音声エージェントやスマートスピーカが一般に普及し，対話システムは，より身近なものとなってきた．メディアなどで対話システムの情報が流れる場面を見るにつけ，「最近の対話システムは随分進化した」と思われるかもしれない．もちろん，本稿でも述べるとおり，ディープラーニングの活用に伴い，その進化は著しい．しかし，一方でまだまだできないことも多い．



■図 -1　過去 1 年の国際会議における国別の対話関連論文数

本稿では，現在の対話システム研究を概観し，「対話システムでは今何が問題になっているのか」について述べたい．これによって，これから対話システムの研究を始めたいけれど何から始めればよいか分からないという方への問題設定の指針になればと考えている．

本稿は，過去 1 年に SIGDIAL/ACL/EMNLP[☆1] という国際会議で発表された対話システム関連の論文のサーベイに基づいている．SIGDIAL は対話・談話関連のトップ会議であり，ACL/EMNLP は言語処理関連のトップ会議で，対話システムのトラックがある．これらの会議が言語処理分野に偏っている点は否めないが，対話システムに関する主要な論文はこれらの会議で発表されることが多いため，今回はこれらに着目した．もちろん，IJCAI/AAAI といった人工知能関連の会議や，Interspeech/ICASSP といった音声・信号処理系の会議，ICMI/IVA といったマルチモーダル・バーチャルエージェント関連の会議でも対話関連論文が発表されているので，サーベイ範囲が限定されている点には注意されたい．

サーベイでは対話システムに関する論文（タイトルに dialog や conversation を含むもの）が全部で 170 編確認された．国別の論文数（第一著者の所属に基づく）を図 -1 に示す．米国と中国の二強であり，米国では主に Facebook, Microsoft, Sales-

☆1　具体的には，SIGDIAL 2020, ACL 2020, EMNLP 2020 を対象とした．

force といった企業が，中国では精華大学，北京大学，ハルビン工業大学といった大学が牽引している．なお，日本は国別 3 位に位置している．

さて，筆者らがこれらの論文の中で取り組まれている課題を吟味し分類した．その結果，対話によって所定のタスクを遂行するタスク指向型対話システムとタスクの遂行を主目的としない非タスク指向型対話システムは同程度の割合でアクティブに研究されていることが分かった．取り組まれている課題としては，ディープラーニングの適用の仕方や発話の質の改善といったものが中心である．また，文書や映像を交えた対話システムの研究が増加している．さらに，より複雑な対話にチャレンジするものや，対話システムが社会に普及しているからこそ現れる対話システムの倫理的な問題についての研究も見られる．本稿では，これらの課題について，それぞれかいつまんで紹介していく．なお，解説記事の特性上，引用は最小限にとどめているが，重要なデータやモデルについてはなるべく多く関連 URL を含めるようにした．必要に応じてリンク先を参照していただければと考えている．

タスク指向型対話システムにおける問題

近年はタスク指向型対話システムにおけるディープラーニング適用の研究が特に増加している．これは MultiWOZ と呼ばれる大規模なマルチドメインのタスク指向型対話のデータセットが公開されたことが大きい．MultiWOZ（ここでは MultiWOZ2.1[☆2] を想定している）では，レストラン検索やホテル検索といった複数のドメインにまたがる，ユーザ役とシステム役による人間同士の対話が 1 万程度収録されており，各発話に詳細なアノテーション（正解ラベルなどの付加情報）が準備されている．現在のタスク指向型対話システムの研究はほとんどこのデータを用いたものと言ってよい．

タスク指向型対話システムでは，数年前までは複数のモジュールによって構成することが一般的だった．具体的には，発話理解（Natural language understanding ; NLU），状態更新（Dialogue state tracking ; DST），行動選択 (Policy)，発話生成（Natural language generation ; NLG）から構成される．NLU は，ユーザ発話のテキストを対話行為と呼ばれる意味表現に変換する．具体的には，意図種別（inform/request/greet/bye/thank など）と付随する情報（属性値対で表現される）に変換される．たとえば，「西地区で安い価格帯のレストランを教えて」という発話は「domain=restaurant, intent=inform, area=west,price range=cheap」という対話行為となる．DST では，ユーザがこれまでの対話で伝えてきた内容（信念状態）を得る．たとえば，先ほどの発話のあとに，ユーザがさらに「イタリアンがいいです」と発話したとすると，信念状態は「domain=restaurant area=west, price range=chap, food=italian」となる．なお，ここでの属性値対，たとえば area=west において，area をスロット，west を値と呼ぶ．Policy では，信念状態をもとにデータベース（たとえばレストランのデータベース）を検索した上で，システムが次に行うべき対話行為を出力し，NLG ではその対話行為をテキストに変換して出力する．

近年では個別のモジュール，複数のモジュール，および，すべてのモジュールがニューラルネットワークによって実装されている．MultiWOZ のリーダーボード[☆3] を見ると，どのアルゴリズムがどの程度の性能が出ているかが一覧できる．なお，個々のモジュールの性能は改善されているが，全体としては問題が山積している．たとえば，ユーザシミュレータを用いた対話ではタスク達成率が 50% に満たない程度であることも報告されている．また，筆者の研究室においても，End-to-End のシステムを

☆2 https://github.com/budzianowski/multiwoz/tree/master/data

☆3 https://paperswithcode.com/sota/multi-domain-dialogue-state-tracking-on-1

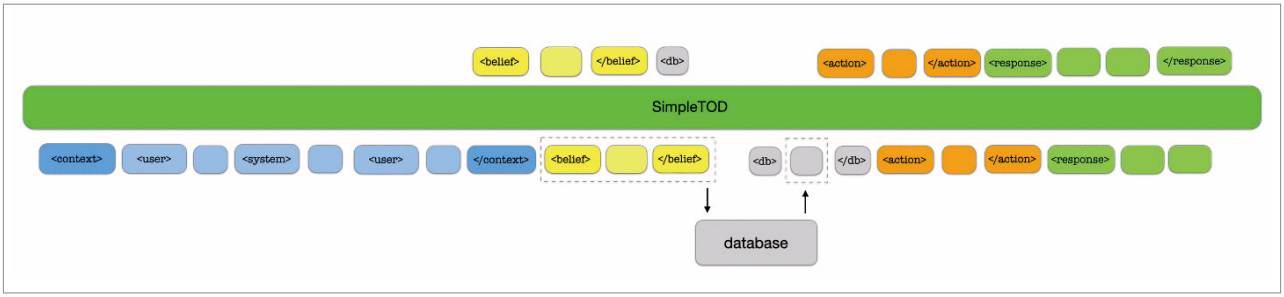
用いて被験者実験を行ったことがあるが，同様に精度が低かった．ここでは特にディープラーニングの適用に関する次の 3 つの問題を紹介する．

DST の精度改善　ディープラーニング時代の DST とは，NLU を挟まず，対話履歴（ユーザ発話とシステム発話の系列）を入力として，直接信念状態を出力する処理を指すことが多い．NLU を経ることで情報が落ちてしまったり，エラーが伝播してしまったりするからである．信念状態を得る方法にはいくつかの方法論が試されてきている．黎明期では，ドメインごとのスロットと値のすべての可能性を列挙し，そのどれが正しいかを分類問題として解くという方法論がとられていた．しかし，ドメイン，スロット，値の組合せが多くなってくるとそれでは処理が困難である．そこで，最近では生成ベースの方法がとられるようになっている．生成ベースの方法とは，スロットの値を埋める文字列を対話履歴などから生成することを指す．ただ，一から文字列を作り出さなくても，文脈に値が含まれているかもしれないので，文脈からのコピー（スパン抽出とも呼ばれる）も併用される．これにより，すべての組合せを事前に考慮する必要はなく，また，未知のドメインであっても対応することが可能となる．ただ，婉曲的な言い方など，コピーなどですべての値が取得できるとは限らないため，生成ベースと分類ベースの併用も検討されている．ディープラーニングの利用においては，大量の学習データが必要となる．そこで，大規模なタスク指向型対話システムのデータで事前学習されたモデル（たとえば TOD-BERT[☆4]）を用いたり，メタ学習といった学習を効率化する方法論が用いられている．

NLG の精度改善　NLG では，対話行為から発話文字列を生成するが，現在の大きな課題は発話の信頼性である．入力に存在しない内容を生成してしまう問題を hallucination と呼ぶが，システムが発話すべき内容と異なる内容を話してしまうことはタスク指向型対話システムにとって致命的である．また，システムの扱うドメインが増えてくるにしたがって，そのドメイン向けの発話生成のデータを準備する必要があり，そのコストも問題になっている．現状の解決策としては，発話に含めるべき単語を指定して生成するなど，制御性を高めるアプローチがとられている．

End-to-End のモデル化　対話履歴を入力として次のシステム発話を出力する End-to-End 対話システムのモデル化は 2016 年ごろから研究が進められており，最近は MultiWOZ をベースに研究が進められてきた．MultiWOZ のみではデータ量が十分ではないため，事前学習済みのモデルの利用が中心的な課題となっている．たとえば，SimpleTOD[☆5]（図 -2）は現在有力なモデルの 1 つであるが，これはユーザ発話から信念状態を出力し，その信念

☆4 https://github.com/jasonwu0731/ToD-BERT
☆5 https://github.com/salesforce/simpletod



■図 -2　SimpleTOD の入力（下部）と出力（上部）の例[1]．対話履歴（青色）から信念状態（黄色）を出力し，その信念状態から得られるデータベースの検索結果（灰色）を入力することで，システムの対話行為（橙色）を出力し，その対話行為をさらに入力することで，システム発話（緑色）を出力する．

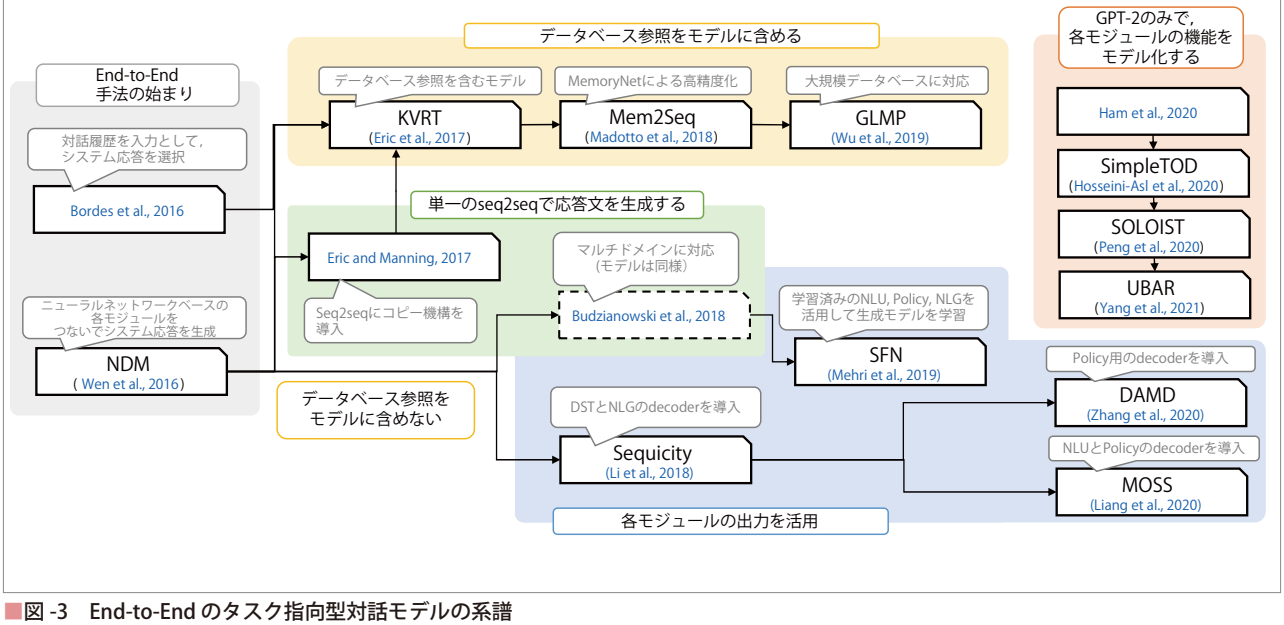
状態をもとに得られるデータベースの検索結果をさらに入力することで，システムの対話行為を出力し，その対話行為をさらに入力することで，発話を出力するという一連の処理を，事前学習済みの大規模言語モデルである GPT-2 上で行っている．図 -3 は，End-to-End の対話モデルの系譜（名古屋大学東中研究室にて，大橋厚元氏の協力により作成）である．文脈からシステム発話を生成，もしくは，選択するものから始まり，データベース参照をモデルに含めるタイプと含めないタイプに大きく分かれる．後者の場合は，推定された信念状態に基づいてあらかじめ設定されたルールなどによりデータベースが参照され，その結果が後段の処理に入力される．図 -3 下部の流れでは，当初は Seq2Seq のモデルですべての処理を賄っていたが，高精度化のために，NLU，DST，Policy，NLG などの個々のモジュールの出力を活用するようになった．一方で，2020 年になり，大規模言語モデル（主に GPT-2）の流れが入り込み，すべてを言語モデルによって解決する方法論が脚光を浴びている．

このようなタスク指向型対話システムの進展は MultiWOZ が牽引したが，中国語でも CrossWOZ[☆6] や RiSAWOZ[☆7] といった大規模なマルチドメインタスク指向型対話のデータが構築されており，今後の進展が期待される．

非タスク指向型対話システムにおける問題

非タスク指向型対話システムは雑談対話システムとも呼ばれる．主目的はタスクの達成ではなく，社会的な関係維持である．タスク指向型対話システムとは異なり，古くから，Twitter や Weibo などの SNS から大規模な対話データが取得され，学習データとして存在していたことから，2015 年ごろから

☆6 https://github.com/thu-coai/CrossWOZ
☆7 https://github.com/terryqj0107/RiSAWOZ



■図 -3　End-to-End のタスク指向型対話モデルの系譜

さまざまな手法が取り組まれてきた．大きくは，複数の発話候補から発話を選択する発話選択と一単語ずつ発話を出力することで発話を生成する発話生成の２つの課題がある．ただ，BERT [2] に代表される大規模言語モデルが一般的となり，現在では発話生成が中心課題である．ここでは特に発話生成に関する３つの問題を紹介する．

Dull response　システムの応答に情報量がなく，「そうですね」や「分かりません」などのありきたりな応答になってしまうという問題を指す．Dull response は，雑談の発話生成において問題であり続けてきた．雑談において，ありきたりな発話はデータ中に頻度が高く，生成されやすい．これに対処するために，近年では外部知識を用いる手法，対話の全体的な流れ（将来的な話の流れも含む）を考慮する手法，高品質なデータから学習する手法などが取り組まれている．大規模な対話データに基づく高性能なチャットボットとして Facebook の BlenderBot[☆8] が有名であるが，最近 BlenderBot2[☆9] が発表された．このシステムは Web 検索を応答生成に活用している．外部知識を取り入れることで，より情報のある発話が実現されている．なお，国内で開催されている対話システムライブコンペティション [3] において，昨年優勝したシステムは BlenderBot の日本語版のシステム Hobbyist[☆10] であった．Hobbyist の対話については，本特集の中野の解説記事 [4] を参照されたい．Dull response の問題が残るとは言え，現状の非タスク指向型対話システムの精度の高まりが感じられる．

個性の一貫性　あるときには「犬が好き」と言い，あるときには「猫が好き」と言うように，個性（ペルソナ）が一貫しないという問題を指す．Persona-Chat[☆11] と呼ばれる，対話と話者のプロフィール文がセットになった大規模なデータが Facebook から公開されており，このデータを中心に多数の研究がされて数年が経つが，いまだに個性の一貫性は問題となっている．最近は，やりとりがプロフィール文と矛盾しているかどうかの判定を組み込んだり，常識的知識を用いて，プロフィール文から想像できる内容についても矛盾しているかどうかを判定するといった工夫が行われている．

論理の一貫性　自身の発話した内容と矛盾した発話や事実と異なる内容を話してしまう問題が近年取り組まれ始めている．以前は発話そのものが文法的に誤っていたり意味が通らなかったりしたが，近年は大規模言語モデルによってその点は解消された．しかし，その反面，文脈を通して論理が一貫していないといった問題が生じるようになってきた．これは個性の一貫性にも共通の問題である．たとえば，DialogueNLI[☆12] と呼ばれる大規模なデータセットが 2019 年に公開された．これは，発話が文脈と矛盾しているか，文脈に含意されているかなどがアノテーションされたデータである．このデータの公開を皮切りに，論理の一貫性に関する研究が加速しており，今後の研究の重要なテーマになっていくと考えられる．

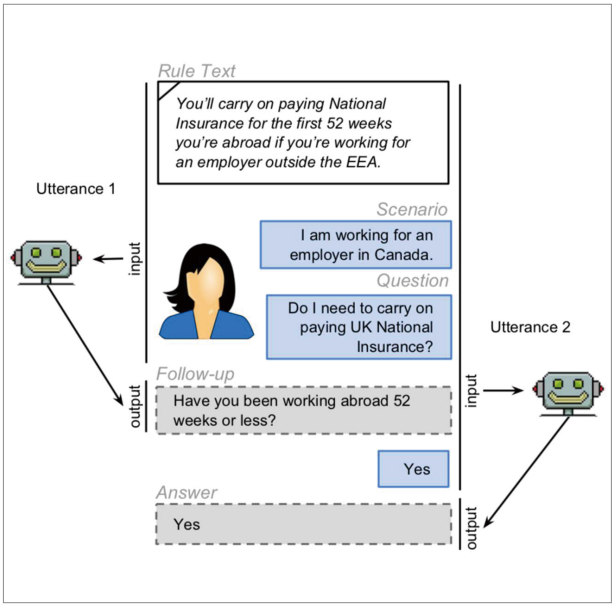
発話生成以外の重要な課題として自動評価がある．雑談対話システムの評価の難しさは長年にわたって指摘されてきた．初めのころは，自動評価尺度が用いられていた．入力に対するシステム応答を，参照発話（正解）との類似度によって計算する BLEU（機械翻訳で用いられる単語のオーバラップに基づく評価尺度）などの尺度が利用されていたが，対話システムにおいて，入力に対する応答は 1 つとは限ら

☆8 https://ai.facebook.com/blog/state-of-the-art-open-source-chatbot/
☆9 https://ai.facebook.com/blog/blender-bot-2-an-open-source-chatbot-that-builds-long-term-memory-and-searches-the-internet/
☆10 https://dialog-system-live-competition.github.io/dslc3/index.html
☆11 https://github.com/facebookresearch/ParlAI/tree/master/projects/personachat
☆12 https://wellecks.github.io/dialogue_nli/

ないことから，あらかじめ用意された正解との比較は妥当ではない．そこで，対話システムの論文では長らく人間による主観評価がなされている．よくある方法は，対話の満足度を多段階で評価するといったものである．クラウドソーシングを活用し，多数のワーカーに主観評価をしてもらう．ただ，人間の絶対評価は主観性が高く結果が安定しないという問題があった．そこで，個々のシステムから得られる対話ログを見比べてどちらがよいかを相対評価してもらうという評価法（ACUTE-EVAL[☆13]）が近年用いられている．

そうは言っても，人手の評価はコストが高いという問題は絶えず問題となってきており，現在はどうやって自動評価を行うかという課題が中心となっている．最近の取り組みとして顕著なものは，大規模言語モデルを用いた評価であろう．たとえば，FED[☆14] と呼ばれる手法では，大規模対話モデルの DialoGPT[☆15] を用いて，システム応答の次の発話として，「それは面白いですね！」のような発話がどのくらい生成されるかで評価する．この方法では，人間の評価値との比較的高い相関が報告されている．自動評価法が確立されれば，対話システムの改善が容易になるため，多くの研究者が本課題に取り組んでいる．

☆13 https://github.com/facebookresearch/ParlAI/tree/master/parlai/crowdsourcing/tasks/acute_eval
☆14 https://github.com/Shikib/fed
☆15 https://github.com/microsoft/DialoGPT



■図 -4　ShARC[5] の対話例．システムはルールを参照しながらユーザからの質問に的確に応答する．必要に応じて不足情報を補うフォローアップの質問を行う．

その他の問題

対話システムにおける課題は，タスク指向型対話システムと非タスク指向型対話システムについてのものばかりではない．ここでは特に近年取り組まれている問題を 4 つ取り上げたい．

読解との融合　現状のタスク指向型対話システムが扱っている対話は現実的な対話とは言えない．正直，レストラン検索やホテル予約などであれば，音声で行うよりも Web フォームで行った方が簡便であるケースは多い．対話がより効果的な場面は，話すことで状況が変わっていくような場面であろう．たとえば，ShARC[☆16] と呼ばれるデータセット（図 -4）がある．これは，ある規則（たとえばどういった条件で年金が受け取れるかなど）が書かれた文書があって，それに基づいた対話が含まれたデータセットである．ユーザは自身の状況を述べ，システムは規則に照らし，必要であれば，追加で質問を行う．これを繰り返すことで，最終的にユーザに回答する．このような対話を実現するには，規則を理解（読解）した上で，対話を行う必要がある．

画像処理との融合　読解では，文書をベースにした対話を行うが，文書ではなく静止画や動画を対象とした対話システムの研究も多く研究されている．Visual Dialog[☆17] を皮切りに，画像に対する質問応答や動画に対する質問応答が活発に取り組まれ

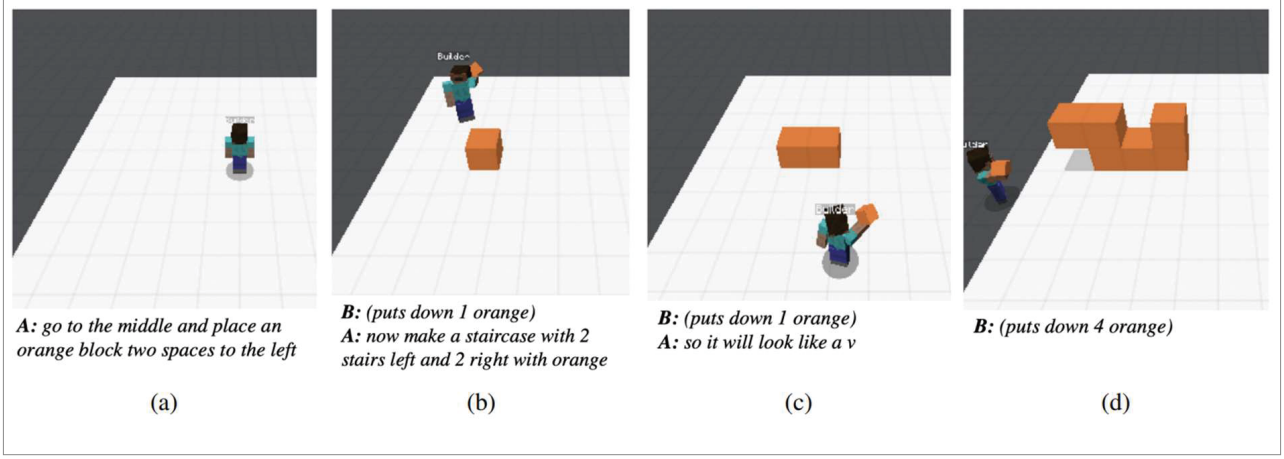
☆16 https://sharc-data.github.io/
☆17 https://visualdialog.org/

ている．映像と対話が紐づいたデータは多くない．そこで，大規模言語モデルを活用した研究や，映像とテキストを対応づけるマルチモーダル版の BERT を学習する研究（VisDial-BERT☆18，VD-BERT☆19）などが取り組まれている．

より複雑な対話　人間同士のよりリアルな対話を収集し，それを実現しようという営みも行われている．たとえば，Minecraft（ブロックを積み上げることでさまざまな構造物を作ることができるゲーム）におけるユーザ同士の対話を収集したデータ（図 -5）がある．一方がもう一方に指示を与える形のデータであるが，ブロックが積み上げられていくごとに，ユーザが置かれている状況が変わっていく．そのような中で相手の発話をどのように理解し，どのように行動するかといったことが扱われている．また，WAY☆20 というデータでは，3D 空間内の家にいる話者とその家の俯瞰的な地図を有している話者との対話が収集されている．対話によって家のどこに話者がいるかを特定するのだが「そこから何が見えますか？」「横の部屋に行ってみてください」といった人間同士のようなリアルな対話が含まれている．そのほか，Dungeons and Dragons と呼ばれる，会話で進行していくロールプレイングゲームの対話ログを集積したデータ☆21 もある．対話システムが実社会で用いられるようになると，対話の状況が刻々と変わっていくことになる．これらの研究は，ゲームを題材としているが，対話の本質的な問題に着目しており，今後重要なデータになってくると思われる．

倫理　言語処理において倫理的な側面が議論されるようになって久しいが，対話システムの分野においても議論が活発化している．今回のサーベイにおいて見られたものの 1 つがジェンダーにかかわるものである．大規模言語モデルに基づく対話システムではどうしても人間のバイアスが含まれてしまい，男性・女性に関するステレオタイプの発言を行ってしまう．そこで，そのような単語が出力されないように制約を与えつつ，対話システムとしての発言の多様性を担保する方法論が模索されている．その他，対話システムとの会話による個人情報の漏洩の検出や相手を不快にする発言の抑制といった課題も取り組まれている．

☆18 https://github.com/vmurahari3/visdial-bert
☆19 https://github.com/salesforce/VD-BERT
☆20 https://github.com/batra-mlp-lab/WAY
☆21 https://github.com/RevanthRameshkumar/CRD3



■図 -5　Minecraft Dialogue Corpus[6] の例（元図から一部抜粋）．プレイヤ A（Architect）が B（Builder）に指示を与えることで，目的の形（この例ではアルファベットの v）を作成する共同作業を行う．

実社会で活用できる対話システムに向けて

本稿では，ここ 1 年の対話システムに関する論文のサーベイに基づき，現状の課題について述べた．全体として，ディープラーニングの時代にあって，データセットが研究を牽引している様子が顕著である．しばらくは MultiWOZ や Persona-Chat に基づく研究が続きつつも，扱われるデータセットはよりリアルな対話状況を扱うものに変わっていくと思われる．倫理的な問題は対話システムが社会との接点を持ち始めたからこそ現れた．筆者らも早く対話システムと実社会で会話ができるようにしていきたいと考えているし，本稿を読んでより多くの方が対話システムの課題に取り組むことを期待している．

参考文献

1）Hosseini-Asl, E., McCann, B., Wu, C.-S., Yavuz, S. and Socher, R. : A Simple Language Model for Task-oriented Dialogue, arXiv preprint arXiv:2005.00796 (2020).
2）Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. : BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding, In Proc. of ACL, pp.417-4186 (2019).
3）東中竜一郎，船越孝太郎，稲葉通将，角森唯子，高橋哲朗，赤間怜奈，宇佐美まゆみ，川端良子，水上雅博：対話システムライブコンペティションから何が得られたか，人工知能，35(3):333–343 (2020).
4）中野幹生：対話システムを知ろう―自然言語による機械と人間とのコミュニケーション―，特集「身近になった対話システム」，情報処理，Vol.62, No.10, pp.e1-e6 (Oct. 2021).
5）Saeidi, M., Bartolo, M., Lewis, P., Singh, S., Rocktäschel, T., Sheldon, M., Bouchard, G. and Riedel, S. : Interpretation of Natural Language Rules in Conversational Machine Reading, In Proc. of EMNLP, pp.2087-2097 (2018).
6）Jayannavar, P., Narayan-Chen, A. and Hockenmaier, J. : Learning to Execute Instructions in a Minecraft Dialogue, In Proc. of ACL, pp.2589-2602 (2020).

（2021 年 7 月 26 日受付）

■東中竜一郎（正会員） higashinaka@i.nagoya-u.ac.jp

2001 年慶應義塾大学大学院政策・メディア研究科修士課程，2008 年博士課程修了．2001 年日本電信電話（株）入社．2020 年より，名古屋大学大学院情報学研究科教授．NTT 客員上席特別研究員．慶應義塾大学環境情報学部特別招聘教授．対話システムの研究に従事．著書に「Python でつくる対話システム」（オーム社），「AI の雑談力」（KADOKAWA）など．博士（学術）．

■光田 航 koh.mitsuda.td@hco.ntt.co.jp

2013 年東京工業大学情報工学科卒業．2015 年同大学大学院情報理工学研究科修士課程修了．2021 年筑波大学大学院システム情報工学研究科博士課程修了．2015 年日本電信電話（株）入社．自然言語処理，対話システムの研究開発に従事．博士（工学）．