特集 6

ロボットとの会話 ―人工知能からのアプローチ―

田中 穂積
東京工業大学 大学院情報理工学研究科
tanaka@cl.cs.titech.ac.jp

徳永 健伸
東京工業大学 大学院情報理工学研究科
take@cl.cs.titech.ac.jp

自然言語理解研究の変遷

言語の利用は人間の知性を特徴付ける重要な要素である．知的なロボットを実現するためには，さまざまな 動作が自然に行えると同時に，自然言語を介して人間とコミュニケーションをする言語運用能力が不可欠である．

コンピュータで自然言語を処理しようとする研究は，コンピュータが発明されるとほぼ同時の 1940 年代に始まっている．1950 年代後半には，人工衛星の開発でソ連に遅れをとった米国がコンピュータを使ってロシア語の科学技術文献を英語に翻訳する可能性を摸索していた．当初は，対訳辞書を作り，それにより，文中の単語を置き換えることで比較的簡単に機械翻訳を実現できるという楽観的な見方がされていたが，現実は違っていた．1966 年に提出された ALPAC レポートによって，より基礎的な研究に研究資金が投じられるようになった．これが自然言語処理や計算言語学と呼ばれている研究分野の本格的な始まりである．

自然言語処理には大きく分けて，表層的な言語表現である文字列からその表現が表す意味，さらにはその表現を作り出した話者や著者の意図を抽出する言語解析と，反対に文やテキストの意味内容，あるいは話者や著者の意図から言語表現を作り出す言語生成の 2 つの方向の処理がある．

言語解析は，言語学が研究対象とする構造単位の分類に応じて，伝統的に以下の 4 つの処理に下位分類される．

- 形態素処理[☆1]
- 統語処理：文の統語構造の解析
- 意味処理：文の意味の解析
- 談話処理：複数の文の構造の解析

これらのうち，形態素処理と統語処理に関しては一般的に利用できるツールが多く存在し，その解析精度もアプリケーションによっては十分実用に耐えるほど向上している．特に大量の言語データと機械学習の手法を使う統計的自然言語処理の研究成果のおかげでその性能はこの 10 年で飛躍的に改善された．同様の手法を意味処理，談話処理についても適用しようとする試みはあるが，これらの処理については十分な精度のツールが手軽に使え

☆1 形態素とは意味を持つ最小の言語単位のことである．たとえば，英語の 1 語 “cats” は猫を表す “cat” とそれが複数であることを表す “-s” の 2 つの形態素からなる．

るような状況には至っていない．

これに対して，言語生成は，何を入力にするかがアプリケーションによって異なるので，その研究もアプリケーションに依存する傾向にあり，ツール化は解析に比べて容易ではない．言語解析が言語の構造単位に対応した処理に分けて考えられるのに対し，言語生成では，生成する内容を決定する処理とそれを表層表現へ変換する処理の2つを考えることが多い．入力が話者や著者の意図であることを考えると，言語解析では後送りできる談話の問題をまず最初に扱う必要がある．

ロボットとの自然言語によるインタラクションという観点からすれば，これらの言語解析，言語生成の両方を実現する必要がある．一般に自然言語によって人間とインタラクションするシステムは対話システムと呼ばれ，これまでにも多くの対話システムが研究されてきた．自然言語処理の研究の歴史の中で最も初期に作られ，まったく異なる方向性を持っている代表的なシステムにELIZA[3]とSHRDLU[4]がある．

ELIZAはWeizenbaumが作成したシステムで精神科医をシミュレートするプログラムである．以下はELIZAの対話の例である．

> 人間：Men are all alike.
> ELIZA: In what way?
> 人間：They're always bugging us about something or other.
> ELIZA: Can you think of a specific example?
> 人間：Well, my boyfriend made me come here.
> ELIZA: Your boyfriend made you come here.

ELIZAで使われている技術は，あらかじめ用意したキーワードを手がかりにした単純なパターンマッチであり，人間の入力を理解しているわけではない．入力中に用意したキーワードに合致するものがなければ，この例の最後のELIZAの発話のように，人間の発話をおうむ返しにして人間がキーワードを含む別の発話をするのを待つ．ELIZAは，このように単純なメカニズムのプログラムではあったが，被験者を使った実験によると，Weizenbaumの意図に反して，人間が真剣にELIZAと対話をする現象が観察された．これは，精神科医のカウンセリングという，特に達成すべき明確な目的がない対象領域☆2を設定していることが主な理由であると考えられる．最近ではシーマンなどのビデオゲームにも同様の手法が使われているし，emacsエディタのdoctorコマンドもELIZAの派生である．

1960年代に開発されたWinogradのSHRDLUは，さまざまな色や形の積木の世界において，人間がロボットアームに自然言語で指令を出して動作させ，世界の状態を変えることができるシステムである（**図-1**）．人間は世界の状態や過去の履歴に関してシステムに質問をすることもできる．対話する人間に特に明確な目標がないELIZAに対し，SHRDLUでは，人間が意図する通りに世界の状態を変化させるという目的がある．SHRDLUは当時として考え得る自然言語処理のほとんどすべての処理を網羅し，多くの言語現象を扱うことができた最初の対話システムである．また，グラフィック・ディスプレイを備え，世界の変化を動画として確認できるという点でも画期的であった．

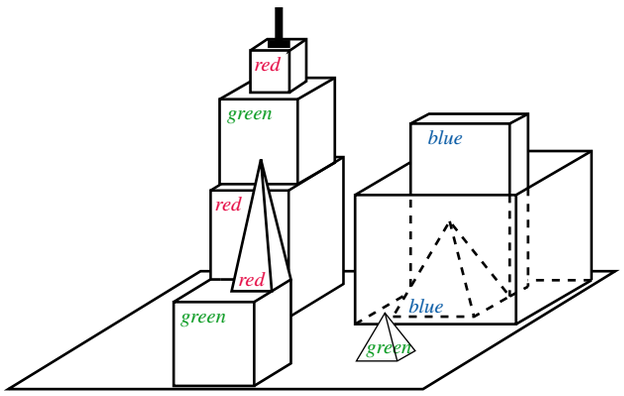

図-1　SHRDLUの積木の世界[4]

ロボットとの言語によるインタラクションという観点からみると，これらの2つのアプローチの用途が異なることは自明であろう．いわゆるAIBOなどに代表される「癒し」のためのロボットならELIZA的アプローチのチャットで十分かもしれないが，ある特定の目的を達成するためには，SHRDLUのように話者の意図を把握し，それに応じて適切な動作，あるいは発話をすることが必要である．

ELIZAやSHRDLU以降も多くの対話システムが研究されているが，残念ながらSHRDLUのような総合的なものは少なく，SHRDLUが扱っていなかった特定の言語現象に焦点を当てたものが多い．特にデータベース中の情報を対話を通じて検索したり，旅行計画の立案をしたりする，いわゆる情報探究型の対話システムが最近の対話システム研究の中心である．情報探究型の対話にも，システムから質問に対して質問で答える問い返しや，すでに決定した事項の変更に関する対話など，興味深い現象は多いが，SHRDLUのように人間とシステムが世界を共有できる類のものはほとんどない．

また，マルチモーダル対話システムと呼ばれるものも多く研究されているが，対話システムの中でグラフィカルな情報を言語表現と効率的に組み合わせることや，システムを擬人化して表情を持たせることによる心理的な

☆2　もちろん患者の精神的なトラブルを解消するという目的はあるが，後述するようないわゆる情報探究型の対話の目的とは異なる．

効果を狙うものなどが中心で，人間とシステムが世界を共有するためにグラフィカルな情報を利用しようとするシステムは非常に少ない．現在のコンピュータグラフィクスやロボティクス研究の進展を考慮すると，より人間に近いロボットを使い，人間と世界を共有する対話システムの研究を行う環境が整いつつあるといえる．

言語理解における身体性

例として部屋の家具の配置を相談している以下の会話を考えてみよう．

(1) 夫：このソファは君の右の壁の方がいいんじゃない？
(2) 妻：(振り返って) この出窓のあたり？
(3) 夫：(場所を指差しながら) いやその辺．
(4) 妻：ちょっとここには入らないんじゃないの．
(5) 夫：(子供に向かって) メジャーをお母さんに取ってあげて．
(6) 子：どこにあるの？
(7) 夫：机の引き出しにあるだろ．
(8) 子：あった．
(9) 夫：それをお母さんに渡して．

このような会話をロボットによって行おうとしたら，ロボットには少なくとも以下のような能力がなければならない．

- 相手の位置やオブジェクトの位置を把握し，前後左右の空間的な位置関係を正しく理解できること．
  この例のような空間的な関係を理解するためには適切な参照枠の設定が必要となる．たとえば，2人が向い会って「君の右」といった場合，話者から見て「君の右」なのか，聴者から見て「右」なのか，参照枠の設定によってはまったく逆の場所を指すことになる．
- 言語情報とパラ言語情報[☆3]を統一的に扱えること．
  単に言語表現だけではなく，言語表現と視線の移動や指差し動作との同期を考慮しないと，(2)や(3)のような発話は正しく理解できない．
- 漠然性が扱えること．
  ソファを置く正確な場所がどこであるのかは，言語表現の上では漠然としているが，実際にソファを置く際には厳密な位置を決定する必要がある．
- 協調作業ができること
  この例では対話を通して複数の人間が協調的に計画を立てている．また，実際にソファを移動しようとすれば複数の人間が協調して動作をしなければならない．

☆3 ここでは，音響的な情報だけでなく，発話にともなうジェスチャ，表情などの非言語的情報も含む．

	ハードウェアロボット	ソフトウェアロボット
動作の多様性	制約あり	比較的自由
環境のセンシング	必要	不要
マルチエージェント化	高コスト	低コストで可能
物理シミュレーション	不要	必要

表-1 ハードウェアロボットとソフトウェアロボットの比較

この例から分かるように，ロボットとの対話を実現するためには，言語を理解するロボットがその世界の状況に根ざしていなければならないことが分かる．これまでのように言語を単に記号として扱い，言語処理が記号処理にとどまっていたのでは，このような対話は扱えない．このように記号に実世界に根ざした意味を与えることは記号のグラウンディングと呼ばれている．また，これは，近年，認知科学や人工知能の分野で，「身体性」というキーワードでもって語られる概念と関係がある．身体性を重要視する研究者は，身体を持たない知能はあり得ず，知能とは世界とのインタラクションによって初めてもたらされると主張している．

ここで注意すべきは，「身体性」といったときにそれは必ずしもハードウェアロボットのように実世界における物理的な身体を意味しないという点である．「身体」は計算機内に仮想的にシミュレートしたものでも十分な場合もある．ハードウェアロボットとシミュレーションによって実現するソフトウェアロボットの違いについてまとめたものを**表-1**に示す．

HONDAのASIMOに代表されるように，最近の二足歩行ロボットの技術の進歩はめざましい．しかし，ハードウェアロボットの動作にはいまだに制限がある．ハードウェアロボットにより複雑な動作を求めれば，それだけ精密な機械とならざるを得ないため，保守のコストも無視できなくなる．これに対してソフトウェアロボットは，動作を作り込むことによって複雑な動作も比較的容易に実現することができる．特に表情の生成などは，コンピュータグラフィクスによる画像のほうが柔軟で多様なものが容易に実現できる．

ハードウェアロボットではまともに動作するために，外界のさまざまな情報をセンサによって計測し，それをロボットの動作に利用するための計算が必要となる．また，センサの誤差に関しても考慮しなければならない．これに対してソフトウェアロボットでは，世界が計算機の中に構築されているため，センシングの処理を回避できる．どのようなオブジェクトがどの位置に存在するかを確実に把握することができる．

また，ソフトウェアロボットではマルチエージェント環境を容易に構築できるという利点がある．ハードウェアロボットは，高価なので，これを複数台用意して，協

調動作などの研究を気軽に行うことは難しい．一方，ソフトウェアロボットでは，各ロボットの個性は別として，別の個体を容易に複製することができるため，マルチエージェント環境を安価に実現できる．

以上は，ソフトウェアロボットの利点であるが，ソフトウェアロボットにも問題はある．ソフトウェアロボットにおいて，現実に近い，より自然なロボットの動作や世界の変化を実現しようとすれば，実世界を完全に計算機内にシミュレートする必要がある．これはニュートンの力学系をシミュレートすることになり，膨大な計算量を必要とする．たとえば，実世界ではロボットが物体にぶつかれば，それ以上は進めないのは当然であるが，ソフトウェアロボットの場合，物体との接触を検出するようにプログラムした世界を用意してやらないと，ロボットは物体をすり抜けてしまう．このように，実世界では，特に考慮しなくても自然に実現される物理的な制約も仮想世界では，すべてプログラムして作り込まなければならないという問題がある．

これらの特徴を踏まえると，研究の目的によってはソフトウェアロボットによる「身体性」を利用することが可能な場合も多い．実世界のシミュレーションを近似し，動作の自然さをある程度犠牲にすれば，身体性を持ったロボットの言語能力や行動計画などの人工知能の問題を研究するにはソフトウェアロボットで十分である場合も多い．

また，ソフトウェアロボットで研究した成果が，すべてそのままではないにしろ，ハードウェアロボットにも適用できる可能性は大きい．次章では，状況に依存した言語理解のための研究課題について述べる．

状況に依存した言語理解のための研究課題

前章では，状況に根ざした対話をロボットが行うために必要な能力について述べた．ここでは，それをさらに詳細化し，ロボットに状況に依存した言語理解を行わせるための研究課題について述べる．

空間的関係の把握

前後左右や上下など物体間の空間的な関係とその言語表現の間の関係については哲学や認知科学などの分野で数多くの研究がある．前章でも述べたように互いに向い合っている状況で「君の右」と言った場合，話者の視点に立つのか，聴者の視点に立つのかで，解釈が逆になってしまう．これは単に視点だけの問題ではなく，参照物体（上の例では「君」）自身に方向性があるかないかなど，さまざまな 要因が言語理解に関係する☆4．

このように空間的な関係の解釈を絞り込むために制約を設定することを参照枠を設定するという．参照枠を決めるモデルは認知科学の立場からいくつか提案されている．たとえば，Levelt は座標系と参照物がそれぞれ話者であるか話者以外であるかによって，参照枠を 3 種類に分類しているし，Retz-Schmidt は，参照物自身が方向性を持つかどうかという要因と視点からやはり 3 種類に分類している．

このように認知科学の研究は参照枠を分類することに主な興味があるが，その具体的な手続きについては教えてくれない．そのほかにも，Herskovits は参照枠の分類よりも，それを決定付ける要因を中心にこの問題を整理し，座標系の原点（常に参照物），軸の順序（前後左右の正順とその鏡像の逆順），軸の方向（「前」の方向）の 3 つの要因によって参照枠を決定する枠組みを提案している．さらに，最初の 2 つの要因については決定方法を述べているが，肝心の軸の方向の決定については明確な答えを出していない．実世界あるいは仮想世界の上で実験システムを構築し，これらの要因について実証的に明らかにしてゆくことは工学の役割であろう．

パラ言語情報

人間同士の対話の中では，暗黙のうちに多くの情報が言語表現以外の手段によって伝わっている．たとえば，表情や手の動き，視線，あるいは声の調子などの音響的な情報は言語で表現された情報を補完する役割を担っている．Cassell ら [1] は話しをする人間の動作をビデオに撮影・分析し，ソフトウェアロボットに実装する試みを行っている．また，自然な視線の動きの実現，コンピュータアニメーションにおいて発話と口唇の動きを同期させる Lip Sync と呼ばれる技術，表情の生成などは，人工知能における仮想エージェントの研究分野では活発に研究されているテーマである．たとえば，長尾らは実際に音声対話システムに表情生成を組み込み，音声認識に失敗したことを，顔をしかめることによってユーザに伝える実験を行っている．その結果，音声認識の失敗を言語表現で伝えるより，ユーザの主観的な評価は改善されたと報告している．これらの研究は，コンピュータグラフィクスや音声認識の技術の進歩を前提としており，最近のこれらの研究分野の進展によって初めて可能になったものである．

曖昧性と漠然性

言語解析においては曖昧性はさまざまな解析の段階で問題になる．たとえば，前節の親子の対話例において，発話(9)で使われている代名詞の「それ」が「メジャー」を指しているということは，人間ならばすぐに分か

☆4 「君」を「ボール」に置き換えるとこのような曖昧性は生じない．

るが，計算機でこれを同定するのはそれほど容易ではない．先行文脈にはメジャーのほかにも机や引き出しやソファなど「それ」で指せそうなものがいくつもある．これは照応の曖昧性と呼ばれ，これを解決する処理は照応の解消と呼ばれている．

この例は指示対象を先行文脈中，すなわち対話を書き起したテキストの中に見つけることができるが，状況を考慮しないと解消できない照応もある．このような照応は一般に外界照応と呼ばれる．たとえば，八百屋に行っていきなり野菜を指さして「これ，ください」という場合，先行文脈がないので「これ」の指示対象を先行文脈中に見つけることはできない．この場合，指示対象はその場の状況に存在するが，その状況は対話を書き起したテキストには現れない．この例でも分かるように，外界照応は言語を記号の中に閉じた系として考えていたのでは扱えない．言語と状況との関連を考えて初めて浮き掘りとなる問題である．

言語処理の研究において言語の曖昧性は中心的な課題であり，多くの研究が行われてきたのに対して，漠然性に関する研究は非常に少ない．上述した照応の曖昧性のように，一般に言語処理における曖昧性の解消は，多くの候補の中から正しいものを選択する離散的な過程としてとらえることができる．これに対して漠然性は数え上げることができない候補からもっともらしい答を見つける連続的な過程であるといえる．たとえば，前節の対話例の発話(3)においてある場所を「その辺」という表現で指示しているが，この場合，指示されている場所を数え上げてその中から正解を選ぶという処理は適切ではない．もちろん最終的に「その辺」に物を置く場合には，正確な位置は一意に決まることになるが，対話の中でやりとりされる「その辺」という表現が指示する場所にはかなりの幅が許容されている．

ここで興味深いのは，「その辺」という言語表現が記号的なものであるのに対して，それが指示する場所は連続的な広がりがあるという点である．このようなミスマッチは，言語を記号の中に閉じた系として考えている限り現われてこない問題である．この例からも分かるように，言語を使用するロボットを実／仮想世界に置いて，そのロボットとの間で言語や行動を通じてインタラクションしようとすると，このような漠然性を扱うことが不可欠となる．

ロボットの行動計画は古典的な人工知能の分野では記号処理を基礎としたプランニングによって行われてきた．「その辺」という表現を単に記号として扱っている限り，古典的な手法は使えるかもしれないが，この例のように特定の状況の中で，言語表現から具体的な空間的な位置関係を計算しようとすると連続的な座標系の計算も不可欠となってくる．このためには従来の古典的な記号処理と空間座標のような連続量のギャップを埋める枠組が必要となる．

協調作業

複数の人間が協力して作業を行うためには，お互いが共通の基盤を共有する必要がある．前章の対話例において夫が子供に向かって「それ（メジャー）をお母さんに渡して」と言っているが，子供がメジャーを母親に渡すためには，子供が「渡す」動作をするだけではなく，母親の方も同時に「受け取る」動作をしなければならない．この例にはもう1つ興味深い点がある．発話(9)は表面的には子供に向けられたものであるが，実際にはその場にいる母親にも聞こえていて，母親に対するメッセージも込められている．多くの対話システムでは1対1の対話を扱うものが多いが，このように複数の人間が同じ場を共有するような例では，1人の発話が必然的に複数の人間に聞こえることになる．このような場合，特定の1人に対するメッセージの場合もあれば，複数，あるいは全員に対するメッセージの場合もある．誰に対するメッセージなのかは状況に依存し，これを判断することが必要となる．

次世代自然言語理解システムへの展望

身体性を持ち，言語を通して人間とインタラクションできるソフトウェアロボットの研究が近年注目を集めている．これらは「身体を持つ会話エージェント(embodied conversational agents)」と呼ばれている．これらの研究で重要な点は，単にコンピュータグラフィクスによって精緻なアニメーションを生成するだけではなく，その多くが言語能力を重要視していることである．

このような研究分野は必然的に学際的なものとなる．すぐに思い付く関連分野として，コンピュータグラフィクス，音声言語処理，計算言語学，認知科学，哲学，言語学などが挙げられる．著者らのグループでもこれらの関連分野の研究者を組織し，2001年度から5年間の予定で「言語理解と行動制御」という研究題目で研究を行っている（文科省科学研究費補助金 学術創成研究 13NP0301）．本章では，一例として我々のプロジェクトを取り上げ，具体的な研究の取り組みについて述べる☆5．

我々のプロジェクトでは，これまで記号の世界に閉じて行われてきた言語理解の研究を，実／仮想世界とのインタラクションを導入することによって，より状況に根ざした言語理解に発展させることを主な目的としている．特に言語理解の結果として生じるロボットの行動を

☆5 http://www.cl.cs.titech.ac.jp/sinpro

重要視している．ただし，これは単に言語を解析した結果を視覚化するというだけの意味ではない．

AustinやSearlらの言語行為論に見られるように，言語の使用（発話）も行為の一種であると考えることができる．逆に発話に対して，物理的な動作や音調などのパラ言語的な手段によって応対することもできることを考えると，ある種の行動は言語の使用と同類であるともいえる．このように人間の活動において言語と行動は密接な関係にあるにもかかわらず，これまで言語処理は言語を閉じた記号系として扱い，ロボティクスでは行動を単なる制御系の問題として扱ってきた．知的なロボットを実現するためには，言語と行動を統一的に扱う必要があると我々は考えている．

この目的を達成するために，我々は仮想世界中に存在する複数のソフトウェアロボットと音声対話によってインタラクションできるプロトタイプシステムを作成し，これをテストベッドとしていくつかの研究テーマに取り組んでいる．

図-2 はプロトタイプシステムのスクリーンショットである．この図では，仮想空間中に黄色と黒色の2体のロボットと色のついた机とボールが置かれている．人間は音声入力によってロボットに指令を出し，ロボットはそれに従って世界の状態を変化させる．ロボットの行動と世界の変化の様子はアニメーションによって人間に提示される．

現在，このプロトタイプシステムを使って以下の項目について研究を行っている．

(1) 音声入力における言い直しや言い誤りを扱うための言語処理
(2) 世界の状況の情報を利用した照応の解消
(3) 空間的な漠然性の表現とそれを利用した行動計画
(4) 構成的な動作の辞書の構成

この中で，特にロボットの行動に関係が深い(3)と(4)について補足する．すでに述べたように古典的な人工知能の行動計画では，空間の位置を表現するのに記号が使われてきた．しかし，位置を指示する言語表現は漠然としており，それによって指示される位置もある程度の広がりが許される．このような位置の漠然性を古典的な行動計画の手法で扱うために，我々のシステムでは記号表現と位置のもっともらしさを表すポテンシャル関数を組み合わせたオブジェクトを使っている．これによって，ある程度の空間的表現に対応できることを明らかにしている．

アニメーションを生成するためには，ロボットの動作を定義した辞書が必要となる．しかし，すべての動作を定義して，その他の動作は基本動作から構成的に作り出すような機構が必要である．基本動作をどのように定義するか，あるいはそもそも基本動作なるものが定義できるのかについては哲学の分野でも長い論争がある．我々は対象領域を決めて，その対象領域のコーパスに含まれる動詞に関するボトムアップな情報と既存の動詞辞書の統語・意味属性などのトップダウンな情報を利用して基本動作を選択するアプローチを採用している[2]．現在は人手によって基本動作の動きを記述しているが，将来的にはモーションキャプチャなどによって基本動作の収集を行えるシステムを開発する予定である．

図-2　プロトタイプシステムのスクリーンショット

本稿では，人工知能，特に言語理解という観点から次世代のロボットに求められる機能について概観してきた．言語を理解し，適切に行動できるロボットが実用になれば，手話通訳や介護サービスなどに応用できるだろう．また，最近のビデオゲームの一部には音声入力をインタフェースとして，ゲーム中のキャラクタを制御するものも出始めているが，キーワードのみを認識して反応する非常に初歩的なものにすぎない．言語を理解するロボットはビデオゲームなどのエンタテイメントにも応用できよう．

参考文献

1）Cassell, J., Sullivan, J., Prevost, S. and Churchill, E.: Embodied Conversational Agents, The MIT Press (2000).
2）Tokunaga, T., Okumura, M, Saitô, S. and Tanaka, H.: Constructing a Lexicon of Action, the 3rd International Conference on Language Resources and Evaluation (LREC), pp.172-175 (2002).
3）Weizenbaum, J.: ELIZA-A Computer Program For the Study of Natural Language Communication Between Man and Machine, Communications of the ACM, Vol.9, No.1, pp.36-45 (1996).
4）Winograd, T.: Understanding Natural Language, Academic Press (1972).

（平成15年10月29日受付）