音声対話理解技術とソフトウェアロボットの行動 *

田中穂積　（東京工業大学・大学院情報理工学研究科）

1. はじめに

我々はこれまで、生活空間の大半を物理空間で過ごしてきたが、最近、情報空間で過ごす時間が急激に増大している。このような情報空間を使いやすく豊かな空間にするためには、物理空間と同様に情報空間でも自然言語を用いた対話が可能になることが望まれる。

このような対話システムの先駆として良く知られているものに、1960 年代後半から 1970 年代前半にかけて MIT の Winograd が開発した SHRDLU（ロボット）がある [16]。SHRDLU は端末から入力した英語の指令を理解し、仮想空間内の積木の世界でロボットに仕事を行なわせることができる。指令文に含まれる代名詞の指すものを同定したり、入力文の解釈に曖昧性が生じた場合には、積木の世界の様子を調べて曖昧性を解消すること、積木の世界の操作で障害となる積み木があればそれを除去して本来の仕事をおこなうという計画立案能力をもっている。当時のコンピュータ環境を考慮すれば、SHRDLU は画期的なシステムであったといえよう。しかし SHRDLU の動作は単純で、対話文も単純な言い回しに限られていた。対話は鍵盤入力を通じて行なうものであった。理想を言えば、音声による対話が可能なことが望ましい。

現在では、音声認識技術、自然言語処理技術ともに当時と比べて格段に進歩している。CG 技術にも目覚しい進歩があり、極めてリアルな 3 次元映像を作り出すことが可能になってきた。表情豊かで複雑な動作を行なう 3 次元ソフトウエアロボットを、仮想空間内に容易に作成することが可能になってきた [12]。しかも大量の計算パワーがパソコンのレベルで利用可能になってきた。

一方、ハードウェアの人間型の歩行ロボットの機械的な技術の進歩も著しい。しかし、現在のハードウェアロボットには、機械的な制約もある。単語のやりとり程度の会話は可能であっても、自然言語で会話したり、指令を理解して行動する能力は備えていない。このようなロボットと音声による対話が可能なことが望まれている。我々は SHRDLU を越えた言語理解システムを構築する時機が到来したと考えている。

以上のような背景から、学術創成研究「言語理解と行動制御」を 5 年間の予定で平成 13 年度から開始した。

2. 目的・研究課題

本学術創成研究では、SHRDLU と同様、ソフトウェアロボット（物理空間に存在するハードウェアのロボットではなく、ソフトウェアでできたロボット／エージェント）を研究対象とする。CG 技術によりコンピュータ内部の仮想空間に、極めてリアルで人間に近い姿をした 3 次元のソフトウェアロボット（Life-Like Robot）が容易に作成可能になったこと、ソフトウェアロボットは様々な動作が可能であること [5][8][4]、そのためロボットに様々な自然言語の動作指令を与えることが可能なったことが背景にあることをすでに述べた。

ソフトウェアロボットが音声・自然言語による対話を理解し、仮想空間内で動作させる研究を目的にする。それと並行してハードウェアロボットを物理空間（実空間）で動作させることも行う。ソフトウェアロボットの研究は、直接ハードウェアロボットの研究に応用可能であると考えたからである。

本学術創成研究で取り上げた研究課題を以下に列挙する。当初の研究計画には含まれず、その後重要であるとして加えられた課題もある。

1. 言語理解と対話

- **a.** 指示代名詞の指示するものの決定（Resolusion of anaphoric relation）
- **b.** 発話で省略されたものの推定 (Ellipsis handling)
- **c.** 指示物体の同定 (「それ」、「あれ」、「これ」など)
- **d.** 指示された場所の同定（「そこ」、「あそこ」、「ここ」など）
- **e.** 不明確性の処理 (Vagueness handling)
- **f.** 空間位置の言語表現とその理解 (Space understanding)
- **g.** 談話の管理 (Discourse management)
- **h.** 発話意図の理解（Intension understanding)
- **i.** 話し言葉の形態素解析
- **j.** 文章生成

2. 音声認識技術

- **a.** 雑音環境下での音声認識技術
- **b.** 画像情報（口唇情報など）の利用
- **c.** 韻律情報の利用
- **d.** 実時間音声認識
- **e.** 修復表現（言い直し、言い足し）の処理
- **f.** フィラーの処理
- **g.** 自然な音声合成技術

3. 非言語表現の理解

∗ Action Control for a Software Agent through Speech Dialogue.
By Hozumi Tanaka (Tokyo Institute of Technology)

a. うなずき、視線、身振り手振り、表情などによる無言の対話

b. あいづち

4. ロボットの行動制御

a. CG による Life-Like な 3 次元ソフトウェアロボットの構築

b. ソフトウェアロボットの多様な動作生成（非言語表現を含む）

c. 行動計画の立案

- 経路探索（最短経路探索アルゴリズム、衝突回避アルゴリズム）

e. 物体認識アルゴリズム（ハードウエアロボットの場合）

5. プロトタイプシステムの開発

a. 対話コーパスの作成と分析

b. 言語レベル（抽象レベル）の世界と動作レベル（具体レベル）の世界との結合

c. CG 技術、音声認識技術、自然言語処理技術、行動計画立案技術の統合

d. ソフトウェアロボットの動作の可視化･映像化

e. 自律動作可能なソフトウェアロボットの開発

f. 対話能力を持つハードウェアロボットの開発

g. マルチエージェントシステムの研究

- 一対多の対話（エージェント同士の対話）
- 協調動作

6. 言語理解と行動制御に関する基礎研究

a. 話し言葉の言語学的研究

b. 空間理解の認知科学的研究

c. 言語行為の研究

d. 認知科学から見た言語と行動に関する研究

3. 研究内容

音声対話理解技術とロボットの行動に関連し、興味深い研究課題を幾つか取り上げて説明する。

上記した研究項目 1 は、言語理解と対話に関する研究課題である。ロボットの行動は、動画として画面に表示される。そのため、より深い言語理解が求められる。ロボットは、状況に依存した発話を理解しなければならない。状況に依存して、ロボットは「どの対象物を、どこで、何をするのか」を判断しなければならない。話し言葉は指示代名詞や語の省略が多用される [11]。指示代名詞が指すものが何か、そしてどこでそれを対象に動作すべきかを決めたり、省略されている語を推測しなければならない。

別のタイプの指示物体の同定も問題になることがある。仮想空間に「赤い玉」が複数個存在している場合、「赤い玉を取れ」という指令では、どの「赤い玉」を意味しているかを同定しなくてはならない。これらが 1･a,b,c,d に述べた研究課題である。さらにマルチエージェントシステムの場合には (5e 参照)、どのロボットに対する指令であるかを決めなければならない。

1･e は、ロボットの行動を可視化するために解決しなければならない問題である。たとえば、「もうちょっと右へ行って」という指令は、「どの程度右か」を決めない限り、ロボットの行動を可視化できない。「コップを取れ」という指令では「取っ手のあるコップ」と「取っ手のないコップ」とでは、ロボットのコップの掴み方が異なるだろう。「コップを取れと」いう指令にはコップの掴み方に対する不明確性が含まれているのである。ロボットのミクロな動作レベルでは、取っ手の有無は、ロボットの指の曲げ方の違いを生むだろう。言語（指示）レベルでは、ロボットの行う動作の仔細を指示しない。いずれにしても、ロボットは動作に関する不明確性を解消しなければ動作できない。不明確性の問題はロボットの動作を可視化するために避けて通れない重要な問題でなのである。

「右」などという相対位置表現には、別の問題もある。「右」という語を解釈するためには、話者の位置、発話対象の位置、向き等を考慮しなければならない。発話状況を考慮してはじめて「右」という語の解釈が決まることにも注意しなければならない。これが 1･f に述べた研究である。

1･g は、現在のロボットの置かれている状況を記憶したり、対話の履歴を管理することである。1･h には、間接発話行為の解釈が関係してくる [1][6][2][3]。発話が文字通りの意味ではないことがあるからである。たとえば「右に曲がれ」という指令のあとで、「行き過ぎ」という発話がなされた場合、発話者の真の意図は「行き過ぎ」という文字通りの意味ではなく「右に曲がりすぎたので、すこし元に戻りなさい」という意味に解釈すべきである。ロボットは気を利かせなければならない。間接発話行為は、一般的な解決策を見出すことがが困難な研究課題である。

書き言葉については、形態素解析システムが既に幾つか開発され利用されている。「茶筅」は代表的なシステムである。書き言葉用の「茶筅」は、そのままでは話し言葉に応用できない。現在、我々の研究グループで話し言葉用の「茶筅」を開発中である（1･i)。これは次の音声認識の項で再び触れる。

項目 2 は音声認識技術に関する研究課題である。2･a,b は雑音環境下でのロバストな音声認識を行う手法を開発することである [9]。2･b,c は音声認識精度を向上させるための研究課題である。2･b には、動画像処理が含まれている。言語モデルとしてバイグラムやトライグラムを用いた統計的な音声認識法に、音響

情報の他に、口唇情報や韻律情報を補強して認識精度を向上させようとする研究である。ロボットとの対話では、音声認識システムに実時間性が求められる。この実時間性には、音声認識システム用の並列処理アルゴリズムを開発して高速化して対応しなければならないだろう。その実装方法が研究課題である (2・d)。

2・e の修復表現の処理は、言語処理と関係する。言語処理の分野では、2・f のフィラー表現とともに非文処理 (Ill-formed senentence analysis) の範疇で研究されているが、十分研究されてきたとはいえない。修復表現は音声対話に頻繁に現れる現象であるので重要である。修復表現は大別して言い直し表現と言い足し表現がある。フィラーは会話途中で挿入される「あのー」とか「えーと」などという語のことである。研究が進むにつれて、ロボットの行う行動と言い直し表現とが密接に関連することが明らかになってきた。

たとえば、「赤い玉を机の左に置きなさい」と言う指令に続けて「いや青い玉」という修復表現が発話されたとする。もし、ロボットが赤い玉を既に机の上に置いてしまった後なら、動作の再計画を行い、青い玉を掴みに行き、それを机の上に置かなければならない。このとき既に机の上に置いた赤い玉をどう処理すべきか、もとの位置に戻すべきだろうか。ロボットはこのような問題の解決を迫られることになる。修復表現は発話の修復だけでなく、ロボットの行った動作の修復も含まれていることに注意したい。

我々が想定しているロボットは自律したロボットとして存在し、会話能力を持っていることが望まれる。特に指令を出す人間との対話では、ロボット側が自然な音声による対話能力を持つことが望ましい。これが 2・g であり、これは後述する 5・f とも関係する。なお、1・i の話し言葉用の形態素解析システム「茶筅」は、発音記号を持った音声認識用の辞書を持つものを開発することになっている。

項目 3 は非言語表現によるコミュニケーション機能をエージェントにもたせるための研究課題である。非言語表現は、ロボットと人間とが円滑なコミュニケーションを行うために（必須ではないが）役立つとされている。

項目 4 は、主としてＣＧ技術を用いたロボットの行動制御に関する研究である [10]。4・b は、多様な動作を行うことができるソフトウェアロボットの開発である。それには、Newton 力学を用いる方法と用いない方法がある [4]。自然な動作の生成を行う場合、両者の方法には一長一短がある。Newton 力学の世界にどっぷり浸かっているハードウェアロボットの動作は好むと好まざるとに関わらず Newton 力学にしたがうので、このような問題は起きない。ソフトウェアロボットの場合、関節をもつソフトウェアロボットを作成して Newton 力学に従うやや複雑な動作を生成することもできる。モーションキャプチャで採取した定型的な動作をつなぎあわせて複雑な動作を合成することも考えられる。次の 4・c のロボットの動作計画立案は、ソフトウェア、ハードウェアを問わず重要な人工知能の研究課題である [14]。

項目 5 は、言語理解と行動制御の様々な問題を、プロトタイプシステムの試作を通じて発見したり、開発した手法、アルゴリズム、理論の有効性を検討検証するためのテストベッドとして役立つと思われる。プロトタイプシステムとして開発途上のシステムを以下に列挙する。

(1) 仮想空間内に存在する複数の物体を対象に、ロボットに物体移動を指令して動作させることを想定した K2 システム

(2) レシピから料理手順を動画として表示・教示するロボット [13]

(3) 手話を理解するロボット

(4) 案内タスク、コピータスク、お茶くみタスクなどオフィス業務を行うロボット

(5) ジェスチャの認識

(6) 首振り、視線を利用した対話ロボット

(7) 冷蔵庫内の物体を取り出すサービスロボット

ソフトウェアロボットは複数個のロボットを仮想空間内に作成することが容易であるので、マルチエージェントシステムの研究を行うのに好都合である [7][15] (5・f)。

項目 6 は、学際的な立場から言語と行動制御に関する基礎理論の構築を目指している。6・a では日本語の話し言葉に特有の言語現象を言語学的な立場か分析する。話し言葉には省略や助詞落ちなどが顕著であるが、これまで言語学者も十分な分析がなされていない。間接言語行為については、6・c で理論的な検討を行う。

4. 組織

工学者、哲学者、言語学者、の参加を得て以下の組織（研究代表者：田中穂積（東工大））で学際的な研究を進めている。

- 言語と行動に関する認知理論（項目 6)
  土屋俊（千葉大)、山田友幸（北大)、辻幸夫（慶応大)、山梨正明（京大)、楠見孝（京大)、丸山直子（東京女子大）
- 音声・言語理解 (項目 1,2,3)
  白井清昭（北陸先端大)、奥村学（東工大)、松本裕治（奈良先端大)、徳永健伸（東工大)、乾健太郎（奈良先端大)、牧野正三（東北大)、河原達也（京大)、古井貞熙（東工大)、鹿野清宏（奈良先端大)、田中穂積（東工大）

- ロボット (項目 3,4,5,2・g)
  中嶋正之（東工大)、白井良明（阪大)、小林哲則（早大)、佐藤誠（東工大)、北橋忠宏（関西学院大)、原島博（東大)、広瀬啓吉（東大)、小林隆夫（東工大）

- プロトタイプ（項目 3,4,5)
  徳永健伸（東工大)、中嶋正之（東工大)、奥村学（東工大)、牧野正三（東北大)、白井清昭（北陸先端大)、白井良明（阪大)、小林哲則（早大）

- 研究顧問
  長尾真（ＮＩＣＴ理事長)、辻三郎（阪大名誉教授)、白井克彦（早大総長)、野家啓一（東北大)、井出祥子（日本女子大）

5. おわりに

本学術創成研究「言語理解と行動制御」は、本格的なプロジェクトとしては世界的にも過去に例を見ない。プロトタイプシステムの作成過程では、言語理解とロボットの行動と言う立場から、これまで無視されてきた問題が実際には重要であることが明らかになってきた。ロボット動作の視覚化・映像化では、指令に含まれる不明確性（Vaguness) の解決が重要であること、修復表現では、指令は言語的な修復だけでなく、ロボットの動作の修復も必要になること、対話では、状況に依存した深い言語理解が必要になること、言語レベルの高次の指令を、具体的なロボットへの動作指令にどう結びつけるかなど、新しい問題を明らかにしてきた。本学術創成研究から得られたさまざまな知見を今後どう一般化し、理論化し解決するか、興味ある未解決の問題が山積している。哲学、認知科学、認知心理学、言語学の観点からも面白い問題が山積している。本学術創成研究により、わが国で「言語理解と行動」に関する研究分野に関心を寄せる研究者が増え、この分野の学術がさらに発展することを期待したい。

前章でもプロトタイプシステムとして応用の一端を示したが、最後に本研究の応用を幾つか列挙して結びとする。

(1) ゲームなどの Entertainment

(2) 介護ロボットシステム

(3) 手話ロボット

(4) サービスロボット

(5) ナビゲーションシステム

(6) 情報家電

参考文献を以下に挙げるが、本学術創成研究の個々の成果の詳細は次の URL を参照してほしい。

“http://www.cl.cs.titech.ac.jp/sinpro/Report2002e.pdf”
“http://www.cl.cs.titech.ac.jp/sinpro/Report2003e.pdf”

参考文献

[1] J. Allen. *Natural Language Understanding*. Benjamin/Cummings Publishing Company, Inc., 1995.

[2] J. Allen and C. R. Perrault. Analyzing intention in utterances. *Artificial Intelligence*, pages 143–178, 1980.

[3] J.L. Austin. *How to Do Things with Words*. Oxford University Press, New York, 1962.

[4] N. I. Badler, C. B. Phillips, and B. L. Webber. *Simulating Humans - Computer Graphics Animation and Control*. Oxford University Press, 1993.

[5] J. Cassell, T. Bickmore, L. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan. Embodiment in conversational interfaces: Rea. In *Proceedings of CHI'99 Conference*, pages 520–527, 1999.

[6] P. R. Cohen, J. Morgan, and M. E. Pollack, editors. *Intentious in Communication*. The MIT Press, 1990.

[7] J. Febler. *Multi-Agent Systems - An Introduction to Distributed Artificial Intelligence*. Addison-Wesley Longman, 1999.

[8] M. N. Huhns and M.P. Singh, editors. *Readings in AGENTS*. Morgan Kaufmann, 1998.

[9] J.-C. Junque and G. van Noord, editors. *Robustness in Language and Speech Technology*. Kluwer Academic Publishers, 2001.

[10] D.J. Litman and J.F. Allen. Discourse processing and commonsense plans. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions and Communications*, chapter 17, pages 365–388. The MIT Press, 1990.

[11] B.J. oGrosz, A.K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226, 1995.

[12] H. Predinger and Y. Ishizuka, editors. *life-Like Characters*. Springer, 2004.

[13] J. Rickel, Ruth Aylett, and Daniel Ballin, editors. *Intelligent Virtual Agents for Education and Training: Opportunities and Challenges*. Springer, 2001.

[14] S. Russell and P. Norvig. *Artificial Intelligence*. Prentice-Hall, 2nd edition edition, 1995.

[15] G. Weiss, editor. *Multiagent Systems*. The MIT Press, 1999.

[16] T. Winograd, editor. *Understanding Natural Language*. Academic Press, 1972.