自然言語を理解するソフトウエアロボットにおける照応と省略の解決

新山 祐介, 徳永 健伸, 田中 穂積
{euske,take,tanaka}@cl.cs.titech.ac.jp
東京工業大学 大学院情報理工学研究科

1　はじめに

我々は自然言語を理解する仮想世界上のロボットを、ユーザの音声によって対話的に動作させるシステム 傀儡 (かいらい) を開発している。言語によってロボットに世界を操作させる対話システムとしては、Winograd による SHRDLU が先駆的である [7]。SHRDLU では、ユーザは英語でシステムに積み木を動かすよう指示する。システムはユーザの入力した文を解析し、積み木を動かすための手順を自動的にプランニングし、曖昧な点があればそれをユーザに問い返す。我々はコンピュータグラフィックス、音声認識、および自然言語処理を密接に関連づけることによって、SHRDLU よりも複雑な動作が可能で、扱える言語表現の複雑さも増したシステムの開発を目指している。

傀儡は、自然な話し言葉の理解に重点を置いている。ユーザは実世界に近い仮想世界を目にして発話するため、その発話には従来の言語理解システムでは扱われてこなかった様々な現象が現れる。たとえば本システムではユーザの視点が変化するため、ユーザの発話はその視点を考慮した解釈を行う必要がある。言語表現から3次元空間を構成する試みとしては [8, 2, 6] などがあるが、これらはいずれもユーザによる視点の変化や漠然性の問題を考慮しておらず、扱っている言葉もおもに書き言葉が中心である。

傀儡では、ユーザは仮想世界におけるソフトウエアロボットと共同で物体の配置をおこなう。ロボットが可能な動作は現在「行く (移動する)」「向く」「何かを押す」の3つである。たとえば「馬はその球を押して」「もうすこし」「ニワトリは右の赤い球の後に行って」などの指示をユーザが音声によって与え、ロボットの動作はアニメーションとしてユーザに提示される。仮想世界上にはロボットのほかにカメラが置かれており、ユーザはこのカメラを通して仮想世界を観察する。

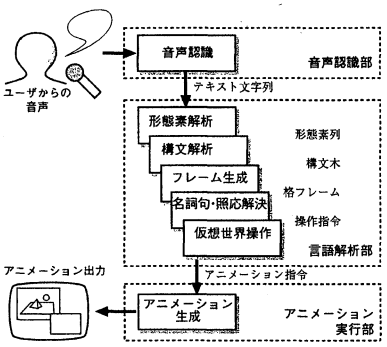

図 1: 本システムの構成

このような状況では、対話中に照応や省略がひんぱんに現れる。そのためシステムは、ユーザの発話履歴と仮想世界の状態を考慮して照応を解決する必要がある。本論文ではこのシステムにおける照応および省略解決の手法を述べる。まずシステムの構成を説明し、その処理の流れを簡単に解説する。つぎに照応・省略解決のためにユーザの焦点を推測する手法を示す。システムは照応解決の際、ユーザがその発話を発した視点を考慮することにより、直示的な表現も扱うことができる。最後にまとめと今後の課題を述べる。

2　システムの構成

本システムの構成を図1に示す。図中の点線による枠はおおまかなコンポーネントを表しており、上からそれぞれ順に音声認識部、言語解析部、そしてアニメーション実行部となっている。内部の長方形はさらに細かなモジュールを表している。ユーザの発話はまず音声認識部によって文字列に変換され、言語解析部によってアニメーション実行指令に変換される。これがアニメーション実行部に送られ、ユーザは結果を得る。

言語解析部における処理の流れを説明する。ユーザが入力した文はまず形態素解析モジュールに送られる。形態素解析モジュールは単語辞書を用いてこの文を形態素列に変換し、構文解析モジュールに送る。構文解析モジュールは与えられた文法にもとづき、この形態素列に対して構文解析を行なう。そしてこの結果生成された構文木をフレーム生成モジュールに送る。フレーム生成モジュールはこれをさらに格フレームに変換する。フレームとは、いくつかのスロットをもつデータ構造である。

各スロットは値をもち、スロットの中にさらに別のフレームを入れ子状に格納することができる。本システムではこの構造を、自然言語をアニメーション指令に変換する中間言語として用いる。フレームのスロットの内容として文の格情報を格納したものをとくに格フレームとよぶ [1]。たとえば「馬はその赤い球をすこし押して」は、以下のような格フレームに変換される:

```
Frame: "馬はその赤い球をすこし押して"
  agent:  Frame: "馬"
    class: HORSE
  object: Frame: "その赤い球"
    head: Frame "赤い球" (物体)
      head: Frame "球" (物体)
        class: SPHERE
      modifier: RED
    modifier: THE
  amount: 1
  verb:   PUSH
```

「馬」「その赤い球」などの名詞句もまたフレームによって表現される。「その赤い球」のように、係り受け関係をもつ名詞句は入れ子になったフレームとして表現される。**head** スロットはその名詞句のヘッドを表し、修飾語は **modifier** や **position-to** などのスロットによって表される。

この格フレームはつぎに意味・照応解決モジュールに送られる。このような格フレームから実際のアニメーション指令を生成するには、格フレーム中の「馬」「その赤い球」などの名詞句を、実際の仮想世界上の物体に対応づける必要がある。このため意味・照応解決モジュールは格スロットの名詞句に合致するオブジェクトを仮想世界から探索する。たとえば名詞句「その赤い球」には連体詞「その」が含まれているため、システムはこれを照応表現と解釈し、ユーザの発話履歴および仮想空間の状況の両方を考慮して対象となる物体を決定する。名詞句から仮想空間上のオブジェクトを得る処理を名詞句の解決と呼ぶ。最後に仮想世界操作モジュールが決定されたオブジェクトの情報をもとに、アニメーション生成手続きを作成する。

ここで名詞句を解決する具体的な手法について説明する。名詞句を解決するためには、まずその指令がどの視点から発されたのかを決定する必要がある。システムは名詞句が表現されたフレームを受けとると、まずユーザがどの視点にたってその指令を発話しているか、可能性のある視点を列挙する。つぎに、それぞれの視点からユーザがその指令を発したと仮定し、それぞれの解釈をスコア付けしたうえでもっとも妥当な解釈と思われるものを選ぶ。視点が特定できたら、システムは与えられた名詞句フレームをその視点から見たものとして解決する。この方法は、対象となる名詞句の種類によって異なる。以下にその手法を簡単に説明する。

物体を表す名詞句の場合、システムは仮想世界上のすべての物体を探索し、その解釈にあてはまる物体が実際に仮想世界上に存在するかどうかによって解釈の妥当性を判断する。これは、ある名詞に対して「仮想世界上のあるオブジェクトが、その名詞句の表しているオブジェクトである適合度」を返すような手続きを仮想世界上の各物体に適用することにより行う。「赤い」、「右にある」などの表現は、原始的な判定手続きをあらわす λ 式として辞書に格納されている。この手続きはオブジェクトの他に、それが解釈される際の視点を受けとるようになっており、その視点から見た適合度を返す。入れ子状になっているフレームは、もっとも内側のフレームから再帰的に解決される。システムは付与された適合度のもっとも高いものを選び、最終的に一意のオブジェクトを得る。

いっぽう位置を表す名詞句の場合、システムはその名詞句の特徴から仮想世界上のある一点を直接算出する。本システムでは、位置表現は最終的に仮想世界上のひとつの点とみなされる。しかし仮想世界上の点は無限にあるため、物体を特定する場合のように仮想世界上の候補すべてを探索するわけにはいかない。そこでシステムは「右」「前」などの名詞から、そのような点を計算する手続きを呼び出し、探索を行わずに位置を決定する。システムは位置関係を表すそれぞれの語に対し、それを算出する手続きを対応づけている。この手続きは、それが解釈される視点および位置の基点となる物体 (例えば「球

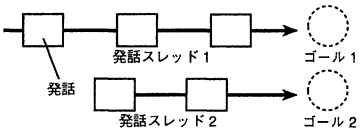

図 2: 発話スレッド

の右」という表現であれば「球」) を引数として受けとることで妥当な点を計算する。

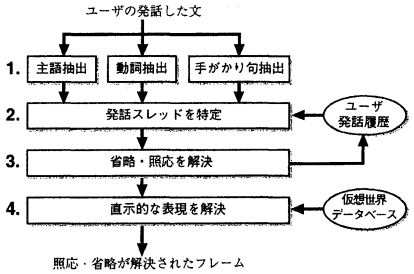

図 3: 照応・省略解決のアルゴリズム

3 照応・省略の解決

本節では照応および省略の解決手法を述べる。名詞句のなかには「それ」や「その球」などといった、代名詞や連体詞が含まれるものがある。このような語が名詞句中に現れると意味・照応解決モジュールはこれを照応表現であるとみなし、照応解決のための処理を実行する。

Grosz らによれば、ユーザが照応表現を用いるのは現在の発話の焦点となる名詞句を表すためである [5, 3]。本システムでは、ユーザはあることをソフトウエアロボットに行わせるために、そのロボットに自分が望む仮想世界の状態、すなわち「ゴール」を伝えている、とみなすことができる。このような状況では、ユーザの焦点はそのユーザがこれから達成しようとしているゴールによっても変化する [4]。一般的に、ユーザの望むゴールは一回の発話ですべて表現できるわけではない。そのためユーザは複数回の発話によってひとつのゴールを表現するが、このようなときに照応表現が用いられることがある。そこで本システムではユーザのゴールを保持することでユーザが用いる照応表現の指示対象を決定する。

実際には、システムはゴールそのものではなく、同一のゴールを表現する一連の発話列を保持する。この発話列を発話スレッドと呼ぶ。本システムでは対話中のある瞬間に、複数の発話スレッドが同時に存在しうる状況を想定している (図 2)。本システムは発話スレッドを発話履歴データベース内に保持しており、ある発話がなされたときにそれが既存のスレッドを受けたものであるのか、あるいは新たなスレッドの生成を示すものであるのかを判定する。これはその発話が既存のスレッドのどれかと主語や動詞が一致しているか、あるいは手がかり句が存在しているか、といった条件をもとに判定される。

ユーザが照応表現を用いる場合、その指示対象はユーザが表現しているゴールに属する発話スレッド中にすでに現れているはずである。そのため、ユーザの発話に照応や省略が含まれていたり、その内容に前回の続きを示唆するような表現が含まれている場合、システムはその内容に一致するスレッドを探索し、そのスレッドにあるこれまでの発話のフレームを使って照応および省略が解決できる。ユーザの発話に照応や省略が含まれていなかったり、妥当なスレッドが見つからない場合、システムはそれを新規のスレッド生成と解釈する。ユーザが代名詞を用いているにもかかわらず、それ以前の発話に適当な先行詞が見つからないときには、システムはユーザが直示を行っていると解釈する。先の処理によってユーザの視点が特定できているため、システムはそこから見た仮想世界の状況を考慮して直示的な表現を解釈することが可能となる。

本システムにおける照応・省略解決の手順を図 3 に示す:

1. ユーザの発話から、まず主語や動詞、および手がかり句を探索する。手がかり句とは「そのまま」「もうすこし」などの副詞句で、これはユーザが同一ゴールを指定するときの目印になることが多い。
2. ユーザが発話とこれまでの発話スレッドの発話とを比較し、ユーザのゴールを表現しているとみられるスレッドを探索する。この探索はもっとも新しい発話をもつスレッドから順に行われる。主語および動詞の両方が一致している発話がスレッド中にあれば、そのスレッドが選ばれる。
3. 発話スレッドが特定されると、システムは同一スレッド上にある過去の発話を取り出す。シス

テムはこの発話から照応表現の参照先を決定し、新しい発話を追加して発話スレッドを最新の状態に更新する。

4. 照応表現が含まれているにもかかわらず発話スレッドが見つからない場合、システムはユーザが直示的な表現を使っていると解釈し、ユーザ(カメラ)の視界とソフトウエアロボットの視界を考慮して仮想世界上のオブジェクトを決定する。この場合、システムはユーザの視点からみてもっとも近い場所にあるオブジェクトを直示の適合度が高いとして選び出す。

また、本システムでは文中の省略も解決することができる。ユーザは同一ゴールを修正する際には、前回の発話で指定したものについては省略する可能性がある。しかしひとたびユーザのゴールが特定できれば、省略されている名詞句に関しても照応表現と同じように前回の発話から自動的に補うことができる。本システムでは現在これを格スロット単位で行っており、システムは不足している格を補完することによって指令を実行する。本手法では、省略された名詞句はすでに仮想空間中に存在しているか、あるいは対話中に最低一度は現れていなければならない。しかし「後に下がれ」などの文では、「後」という名詞句は自明であり省略できる。本システムは、このような省略をアドホックなルールによって処理している。

4 おわりに

本論文では我々が開発しているシステム傀儡における照応および省略の解決手法について述べた。本システムでは、ユーザはカメラを通して仮想世界の映像を見ながら発話するため、その表現の解釈はユーザの視点によって変わる。また、ユーザは直示を用いる場合もある。このような状況で仮想世界上の位置や物体を表現する名詞句を解決し、仮想世界上のオブジェクトを一意に決定するための手法を提案した。ユーザはある動作を指令するときに、自分の持っているゴールを伝達しようと試みる。本システムは発話スレッドを用いることでユーザのゴールをデータベース内に保持する。これによってユーザが現在どの物体あるいは位置に焦点を置いているかが推定できる。システムはユーザの視界や仮想世界の状況を考慮してユーザの焦点を推測する。これによって照応表現や省略、および直示が含まれる発話を適切に解釈することができる。

現在のところ、本システムが扱える文は命令文だけである。また発話スレッドによる焦点の推測は、文の比較的表層的な面しか考慮に入れていない。ときにユーザは暗黙の焦点の移動を行う場合がある。たとえばある動作を行ったあとは、次にくる動作が予想できる場合などである。このような場合、本システムでは発話スレッドの識別に失敗し、照応や省略を正しく解決できない。また、本質的に曖昧な指令を受けた場合、本来システムはユーザに問い返すべきであるが、現在のシステムではどちらか一方の解釈に決めている。今後、仮想世界の構造を複雑化し、ユーザの指示できる範囲を広げる予定である。このような拡張をより自然に行えるようなアーキテクチャを提案することも重要な課題となる。

参考文献

[1] Charles J. Fillmore. 格文法の原理. 三省堂, 1975. 田中 春美, 船城 道雄 訳, ISBN 4-385-30085-2.

[2] C. W. Geib, L. Levison, and M. B. Moore. Sodajack: An architecture for agents that search for and manipulate objects. Technical Report MS-CIS-94-16/LINC LAB 265, 1994.

[3] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstrin. Providing a univied account of definite noun phrases in discourse. In *ACL Proceedings*, pp. 44–49, 1983.

[4] Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, Vol. 12, No. 3, pp. 175–204, July-September 1986.

[5] Candace L. Sidner. Focusing in the comprehension of definite anaphora. In Michael Brady and Robert C. Berwick, editors, *Computational Models of Discourse*. MIT Press, 1983.

[6] Steve Strassmann. Semi-autonomous animated actors. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 128–134, 1994.

[7] Terry Winograd. *Understanding Natural Language*. Academic Press, 1972.

[8] 佐藤泰介, 田中穂積, 渕一博. Visualizer - 自然言語理解システムの立場からみた機械による空間の把握. 電子通信学会誌, November 1976.