14－03

自然言語を理解するアニメーテッドエージェントのための3次元仮想空間における位置の表現と処理

Expression and processing of 3-D spatial relations for animated agents acting on natural language instructions

新山 祐介　秋山 英久　鈴木 泰山　徳永 健伸　田中 穂積

{euske, haki, taizan, take, tanaka}@cs.titech.ac.jp

東京工業大学 情報理工学研究科 計算工学専攻

Department of Computer Science, Tokyo Institute of Technology

Abstract: We are developing animated agents which can understand natural language instructions and act upon them. This paper focuses on the representation of 3-D spatial expressions. The system accepts relative spatial expressions stated from the speaker's viewpoint. The same expression can refer to different positions depending on the speaker's viewpoint. We propose a spatial semantic representation using lambda abstraction, which can be processed without any consideration of the speaker's viewpoint and resolved to actual positions effectively. We also propose a translation method from natural language expressions into this representation.

1 はじめに

既存の自然言語処理技術の大部分は、機械翻訳や情報検索といった、おもに静的なテキストを扱うものが多かった。これらは言語に対して「フィルタ」のように動作する。これに対して、言語が直接行為そのものに結びつくような研究、とりわけ人間やロボットの物理的な行動を言語によって指示するような研究はSHRDLU [1] 以来ほとんど行われていない。我々は人間の行動と言語を結びつける研究のひとつとして、自然言語処理技術とCG技術とを組み合わせ、自然言語によってアニメーテッドエージェントを制御するシステムを開発している。

アニメーテッドエージェントとは仮想空間内に存在し、ユーザの自然言語による命令を理解し、それに従って動作するエージェントである。ユーザはエージェントの動作を画面上でアニメーションとして見ることができる。仮想空間上には、形状をもった「俳優」エージェントとともに仮想空間を撮影する「カメラ」エージェントも配置される。我々は最終的にはユーザが自然言語によって仮想空間上のすべてのエージェントを監督するようなシステムを目標としているが、現在その第一段階として、カメラエージェントを自然言語によって制御し、その撮影した画面を連続的に変化させるシステムを試作した。本論文ではこの際に生じる、自然言語による位置決定の問題を扱う。

本システムが受け付ける命令は、仮想空間上に置かれたカメラエージェントに対して同じ仮想空間に存在する物体を撮影させるものである。具体的には「物体A の前へ行ってください」「その後に回り込んでください」等の命令を想定している。これらの命令は内部的な意味表現に変換され、カメラエージェントに送られる。

本システムでは、ユーザはカメラエージェントの撮影する映像を画面上で見ながら命令を入力する。したがって、このユーザの発する命令はカメラエージェントの視点における発話行為とみなすことができる。一般に「～の前」「～の近く」などといった位置表現は、その表現を発した話者の視点を基準とした相対的なものであり、話者の視点を考慮しなければ実際の位

Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1, Ôokayama, Meguro-ku, Tokyo, 152, JAPAN

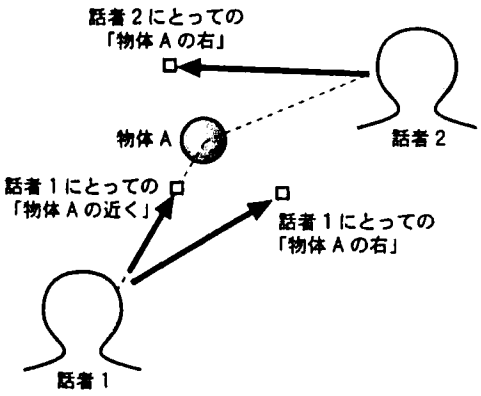

図 1: 話者の視点によって異なる位置表現

置は決定できない [3, 4]。また「前の近く」などのように、相対的な表現が重複して使用された場合、「前」「近く」などの語が表す意味をどのように表現し、組み合わせるかが問題となる。

本論文ではλ抽象を用いた関数を意味表現として用いることにより、話者の視点が決定していない状態でもこれらの意味表現を処理できることを示す。λ抽象を用いた意味表現は従来からモンタギュー文法などで用いられてきたが、本手法により話者の視点に依存しない「前」「近く」などの意味が表現でき、それらを日本語の統語構造に類似した形で自然に組み合わせることも可能となる。本システムではこれらの意味表現を関数型言語 Scheme 用の関数とみなし、インタプリタに直接処理させることにより効率のよい実行を実現している。

2 λ抽象を用いた位置表現

ユーザが本システムに与える命令はおもにカメラエージェントの移動およびパンニングの方法を指定するものである。システム内では言語解析部分とカメラエージェントは分離されており、自然言語による命令は内部的な意味表現に変換されてカメラエージェントに渡される。カメラエージェントはその指示が達成されるまでそれ自身の状態を変化させ、この過程をシステムはアニメーションとして出力する。したがって想定されている自然言語の命令を効率よく表現できる意味表現を導入することが必要となる。

しかし人間が自然言語を用いて空間上の位置を表すとき、同じ表現でも話者の視点によって実際に指される位置は異なる。たとえば図 1 における話者 1 にとっての「物体 A の右」と、話者 2 にとっての「物体 A の右」は同じ表現でも実際には異なった位置を指す。そのためカメラエージェントは、同じ「物体 A の右へ行ってください」という命令でも、命令が発された時のカメラエージェントの視点によって異った動作をする必要がある。だがシステムにとってはその位置にかかわらず、同じ「物体 A の右」という表現は同じ意味表現で表されることが望ましい。

また、人間は「物体 A の近く」などといった表現も使用する。この場合、物体 A の「近く」とは一体どこなのかという問題が生じる。計算機が自然言語による命令を実際に実行する際には、たとえユーザの命令が曖昧さを含むものであっても目標となる処理は曖昧さなく実行しなければならない。本システムではこのような場合「近くの点」を話者と対象となる物体との間に引いた線分上に決定するが (図 1)、ここでも「近く」という意味は話者の位置によらない表現に変換されることが望まれる。

自然言語による空間的な位置表現の解決を扱った研究として、Kalita らによる英語の前置詞の研究がある [2]。ここでは空間上の複数の点あるいは領域に対して、それらがある特定の前置詞の関係を満たしているかどうかを決定する手法が提案されている。しかしここで挙げられている手法は話者の視点を含むようなものではなく、実時間で解決するのは困難であるため本システムのような要求には向かない。

本手法では位置を表す名詞「前」「右」「近く」「正面」などの意味表現を、λ抽象を用いた関数として定義する。これによって、話者の視点が未決定のままこれらの意味表現を組み合わせ、新たな意味表現を作成することができる。λ式そのものはオブジェクトであるから、システム中で記憶、代入、受け渡し等の操作が可能になる。この関数は話者の視点座標を与えると、その時点での話者にとっての「前」や「右」などの座標を決定し、その値を返す。たとえば本手法によって図 1 に示されている「物体 A の右」の意味表現 $S_{RightOfA}$ を表すと、次のようになる。まず最初に「物体 A の位置」を表す意味表現：

$$S_A \equiv \lambda p.P_A$$

を導入する。P_A は物体 A の座標を表す定数である。したがって、これは入力座標に関係なくつねに物体 A の固定した座標を返すような恒等関数となる。本

手法では意味表現の単位はすべて関数であるので、このような話者の位置に依存しない座標もすべて関数で表現する。

次に、ある意味表現 s が与えられたとき、「s の右」という意味表現を新たに返すような関数 $S_{RightOf}$ を導入する：

$$S_{RightOf} \equiv \lambda s.\lambda p.f_{RightOf}(p,(s)p)$$

ここで、$f_{RightOf}(p_1,p_2)$ は視点 p_1 から見た p_2 の右にある点の座標を返す関数である。「物体 A の右」という意味表現は、これら 2 つの意味表現を組み合せて次のように表すことができる：

$$\begin{aligned} S_{RightOfA} &\equiv (S_{RightOf})(S_A) \\ &\Rightarrow \lambda p_1.f_{RightOf}(p_1,(\lambda p_2.P_A)p_1) \end{aligned}$$

この時点では、まだ値 p_1 は定まっていないため、ここで評価を止めておく。これがカメラエージェントに渡される「物体 A の右」意味表現となる。ここで、実際のカメラエージェントの視点座標 P_{Agent1} をこの関数に適用すると、この視点における話者が意図した「物体 A の右」の座標 $P_{RightOfA}$ が得られる：

$$\begin{aligned} (S_{RightOfA})P_{Agent1} &\Rightarrow f_{RightOf}(P_{Agent1},P_A) \\ &\Rightarrow P_{RightOfA} \end{aligned}$$

このような意味表現は何段階にもわたって組み合わせることができる。カメラエージェントはこの意味表現をシステムの意味解析モジュールから受けとり、実行時に自分自身の座標をこの関数に適用しながらその時点でのゴール座標を決定する。関数への適用はアニメーションの各フレームごとに行われる。これにより、ユーザの話者としての視点を適切に考慮した動作が実行される。

本手法では位置はすべてその座標を返す関数として表現するため、物体が空間上で固定した座標をもたず、つねに移動しているような場合でも「物体 A のその時刻における位置」といった意味を表現できる。これによって動いている物体をつねに追いかけるような命令も表現が可能になる。

3 自然言語から意味表現への変換

本システムでは、ユーザの入力した命令は音声認識モジュール、構文解析モジュールを経て、その構文解析結果が意味解析モジュールへと渡される (図 2)。

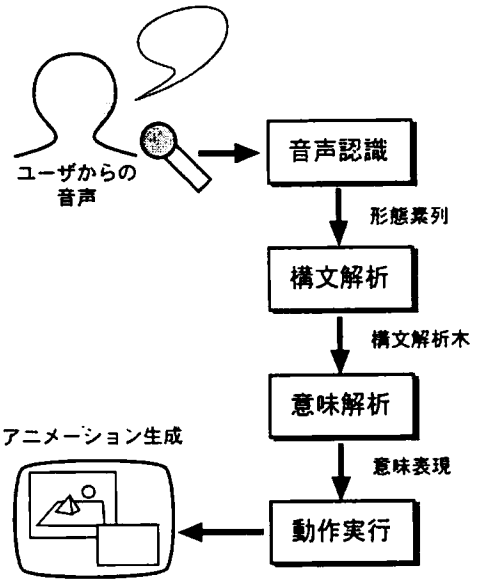

図 2: システムの構成

2. 節で示した意味表現を用いることによって、位置を表す名詞句からの意味表現の生成は次のような手順で行うことができる。例として、以下のような文法によって解析された名詞節 (NP) の解析木が与えられた場合を考える：

NP → NP pp n | n
pp → の
n → 物体A | 前 | 後 | 左 | 右 | 近く

ここで、与えられた名詞節からλ抽象を含んだ意味表現を返す関数 f_{sem} を定義する。この関数は、λ抽象を含んだ関数を生成するような高階関数であるとみなすことができる。すると、名詞 n については、上の文法の単語と意味表現とをすべて対応づけることができ、言語表現から意味表現への変換が効率よく行える。たとえば f_{sem}(物体 A) は 2. 節における意味表現 S_A を返し、f_{sem}(右) は意味表現 $S_{RightOf}$ を返す。さらに、f_{sem} を次のように定義することで、この関数によって名詞節の意味表現も生成できるようになる：

$$f_{sem}(\mathrm{NP\ pp\ n}) \equiv (f_{sem}(\mathrm{n}))(f_{sem}(\mathrm{NP}))$$

ところで、単なる「物体 A の右」という表現でもその「右」という位置は実際に対象となる物体からどれくらい離れた地点のことを指すのか判然としない。このため、本システムでは副詞の処理も行っている。

「右」という意味表現に程度を表す副詞を組み込むには、先の意味表現を拡張し、さらにいくつかの変数をλ抽象として追加すればよいが、基本的な枠組みは変わらない。また本システムで扱う話者の視点は、実際にはその座標だけでなく、視点の向きを表すベクトルも含まれているため、ユーザの指定する座標と向きをひとつの意味表現で表すことができる。これによって本システムはカメラエージェントの移動だけでなく、パンニングや移動とパンニングの組み合せなどが柔軟に行える。

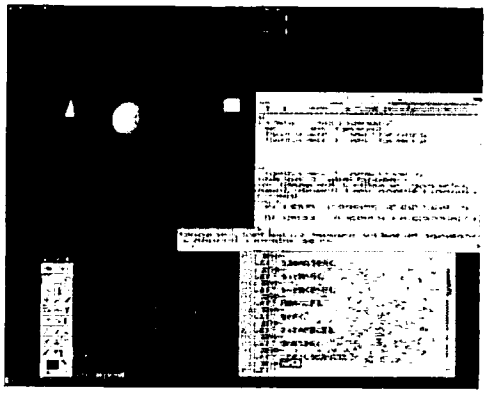

図 3: システムの動作画面

4 試作したシステム

我々は以上のような意味表現の生成および処理ルーチンを含んだシステムを試作した。λ抽象を含んだ関数を効率よく扱うためには、言語仕様そのものにこれらの型が効率よく処理できるような機能が含まれていることが望ましい。我々はそのような言語のひとつとして、関数型言語 Scheme を用いた。これによりλ抽象を含んだ意味表現の処理が容易に実現できるだけでなく、その実行効率も向上することが期待できる。

本システムでは「移動する」や「パンする」といった指令のほかに、移動とパンを組み合せた「回り込む」といった動作も可能になっている。その他に対話的な機能として省略の補完や履歴の記憶などの機能も持たせるようにした。

ユーザが発した指令はまず最初に音声認識モジュールによって解析される。構文解析モジュールでは現在のところ ATN を用いて文を解析する。意味解析モジュールは 3. 節で提案した手法を用いて構文木を意味表現に変換し、カメラエージェントに渡す。文脈にともなう「これ」「さっきの位置」などの代名詞の扱いや、命令の省略の補完も意味解析モジュールで行っている。最後にこの意味表現はカメラエージェントに伝えられ、カメラエージェントは各フレームごとにこの意味表現に自分の座標を適用しながら移動をくりかえし、画面のアニメーションを出力する (図 3)。

2. 節で述べたように、本システムは言語解析部分とカメラエージェント制御部分を独立に設計している。そのため、仮想空間上にカメラエージェントを複数配置することも容易に実現できる。

5 おわりに

本論文ではアニメーテッドエージェント作成の第一段階として、カメラエージェントを自然言語によって制御するための意味表現を提案した。これによって話者の視点によって異なる位置の表現に対し同一の意味表現を用いることができる。また、本手法を実現するシステムを試作しその有効性を確認した。

今後、形状をもったエージェントも自然言語で制御できるよう本システムを拡張していく予定であるが、その際にはさらに困難な曖昧性が現れることが予想される。また、位置表現や動作表現に関連した副詞の処理や、言語による並列動作の表現、対話における漸進的なゴールの設定などの課題にも取り組む必要がある。

参考文献

[1] Winograd, T., "Understanding Natural Language," Academic Press, Ph. D. Thesis, 1972.

[2] Kalita, J. K., Badler, N. I., "Interpreting prepositions physically," AAAI-91 Proceedings Ninth National Conference on Artificial Intelligence, Volume One, pp. 105-110, July 14-19, 1991.

[3] 片桐 恭弘, 談話の世界, 自然言語理解, 田中 穂積, 辻井 潤一 共編, pp. 159-190, ISBN4-274-07398-X, オーム社, 1988

[4] Herskovits, A., 空間認知と言語理解, 堂下 修司, 西田 豊明, 山田 篤 共訳, ISBN4-274-07676-8, オーム社, 1991