1I1-1

コミュニケーションの中に意味づけされる単語の学習

Learning Communicative Meanings of Utterances by Robots

田口 亮[*1]
Ryo Taguchi

岩橋 直人[*2*3]
Naoto Iwahashi

新田 恒雄[*1]
Tsuneo Nitta

[*1] 豊橋技術科学大学
Toyohashi University of Technology

[*2] (独)情報通信研究機構
National Institute of Information and Communications Technology

[*3] (株)国際電気通信基礎技術研究所
Advanced Telecommunications Research Institute International

This paper describes the computational mechanism that enables a robot to return suitable utterances or actions to a human by learning the meanings of interrogative words, such as "what" and "which". Previous works of language acquisition by robots proposed methods to learn the words, such as “blue” and “box,” that indicate objects or events in the real world. However the methods can’t learn and understand interrogative words because they don’t directly indicate objects or events. The meanings of them are grounded on communication itself and stimulate specific responses of a listener. We call the meanings communicative meanings. Our proposed method learns the relationship between human utterances and robot responses as communicative meanings on the basis of the graphical model of the human robot interaction.

1. はじめに

実世界でロボットが人とコミュニケーションするためには，多様な事物・事象に関する知識や，文法や慣習などの知識が必要になる．ロボットによる言語獲得の研究は，こうした知識をロボット自らがユーザとのインタラクションを通して獲得していくことを目的としている．先行研究では，「箱」や「青い」といった物の名前や特徴を表す単語や，「上げて」や「乗せて」といった動作を表す単語の学習が行われてきた[Roy02, Iwahashi07]．一方で，人の発話には，「なに」や「どれ」といった疑問詞や，「おはよう」や「バイバイ」といった挨拶など，実世界の事象を表しているわけではない単語が用いられる．これらの単語は，コミュニケーションにおいて直接相手の行為に変化を与える機能がある．図 1 に示すように，実世界の事象に接地した意味（**記述的意味**）を理解することで，ロボットは人が実世界状況の中の何に注意を向けているかを理解することができる．一方で，「なに？」と聞かれたら発話で答え，「どれ？」と聞かれたらオブジェクトを指し示すなど，コミュニケーション自体に接地した意味（**相互行為的意味**）を理解することで，ロボットは適切に応答することができる．通常，一つの発話には記述的意味と相互行為的意味の両方が含まれるが，従来研究では，「発話されたら動作を出力する」というように応答の種類が予め規定されていたため，記述的意味の理解だけを扱ってきた．しかし，より自然で複雑なコミュニケーションを実現するためには，それら両方の意味を理解する必要がある．本稿では，両者の意味を統合した発話応答のモデルを提案すると共に，発話と応答の対応関係を相互行為的意味として学習する手法を提案する．

連絡先：田口亮
〒466-8555 名古屋市昭和区御器所町
名古屋工業大学 19 号館 2 階 226 室
E-Mail: taguchi.ryo@nitech.ac.jp
TEL: 052-735-5552

図 1：記述的意味と相互行為的意味

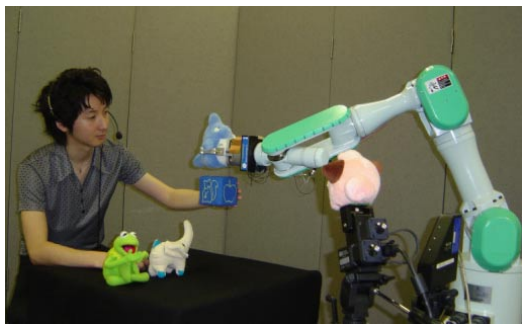

図 2：インタラクションの様子

2. 問題設定

ロボットは図 2 のように配置され，人とロボットが互いにテーブルの上のオブジェクト（ぬいぐるみや箱など）を用いながら言語によるインタラクションを行う．まず，ロボットは，オブジェクトやその動かし方についての概念や，概念に対応する単語，および文法などの信念を人との共有経験を基に学習する[Iwahashi07]．その後，相互行為的意味の学習実験を行う．

実験では，人はオブジェクトの操作を要求する発話（例えば「青い箱を赤い箱に乗せて」など）と，オブジェクトの名称や操作に対する質問をする．例えば，あるオブジェクトを指差しながら

「なに？」と聞いたり，オブジェクトを動かして「赤いぬいぐるみはどれを飛び越えた？」と聞いたりする．ただし，発話は助詞を含まず，活用変化のない単純なものとする（例えば，「青い 箱 赤い 箱 乗せて」など）．ロボットは人の発話と行動から，適切な応答を推論し，発話または行動（オブジェクト操作や指差し）を返す．ロボットの応答が間違っていた場合，人はロボットの手を軽く叩き，正しい応答の例を示す．これによりロボットは適切な応答を学習する．

3. 相互行為的意味の学習

3.1 発話応答のモデル

[Iwahashi07]は，発話と動作の関係をグラフィカルモデルで表現した．モデルには指差しなどのコンテキストも含まれており，状況に応じて適切に発話を理解・生成することができる．本稿では，相互行為的意味の学習・理解を実現するために，[Iwahashi07]のモデルを図 3 のように拡張する．従来のモデルは図中の発話 S，意味構造 Z，行動 B，話題 A から成っていたが，本モデルでは発話 S，意味構造 Z，行動 B に関して自分（ロボット）と相手（人）の2種類を設け，発話意図を介して統合する．これは人が発話をし，ロボットがそれに対して応答をするという発話応答のモデルとなっている．発話意図 I は人が期待するロボットの応答の種類をコンパクトにまとめたものである．これが応答に制約を与えることで，適切な発話応答が実現する．

以下，モデルの詳細を説明する．時刻 t における人の発話・行動を各々 S_{1t}, B_{1t}，ロボットの発話・行動を S_{2t}, B_{2t}とする．行動 B_{1t}, B_{2t} は，オブジェクトを動かしたり，指差したりしたことを表す．具体的には，行動内容 B_{C1t}, B_{C2t}（指差し/ 操作/ なし）と，その対象となるオブジェクト B_{O1t}, B_{O2t} で表現する（すなわち $B=(B_C,B_O)$）．両者の発話・行動 S_{1t}, B_{1t}, S_{2t}, B_{2t} は直接観測される．オブジェクトの動作はトラジェクタ・オブジェクト O_{Tt}（動かすもの），ランドマーク・オブジェクト O_{Lt}（動きの基準点となるもの），軌道 U_tから成る．トラジェクタ・オブジェクト O_{Tt}と軌道 U_tはどちらかがオブジェクトを操作した時に観測される．O_{Tt}, O_{Lt}, U_t をまとめて話題 A_tと呼ぶ．Z_{1t}と Z_{2t}は両者の発話の意味構造を表す．例えば，人が「なに 茶色 箱 乗せる」と聞いたときの正しい意味構造は次のようになる．

表 1：意味構造の例

トラジェクタ Z_T:	（なし）
ランドマーク Z_L:	茶色 箱
動作 Z_M:	乗せる
疑問詞 Z_I:	なに

話題 A_tは両者の発話の意味構造 Z_{1t}, Z_{2t}と行動 B_{1t}, B_{2t}, 前回の話題 A_{t-1} から推論される．これは，互いが共通の話題について発話・行動しあうことを意味する．また，前の話題 A_{t-1} は次の話題 A_tを決定する際のコンテキストとして利用される．

発話意図は人の意味構造 Z_{1t}・行動 B_{1t} と，ロボットの意味構造 Z_{2t}・行動 B_{2t}とを繋ぐための中間ノードの役割を果たし，応答に制約を与える．人の意味構造 Z_{1t}・行動 B_{1t}から発話意図 I_tへのマッピングを**相互行為的意味**と呼び，発話意図 I_tからロボットの意味構造 Z_{2t}・行動 B_{2t} へのマッピングを**応答制約**と呼ぶ．

人の発話・行動 S_{1t}, B_{1t} が与えられると，ロボットは次に示す決定関数Ψを最大化するように，話題 A_tと発話意図 I_tを推論し，応答 S_{2t}, B_{2t}を決定する．

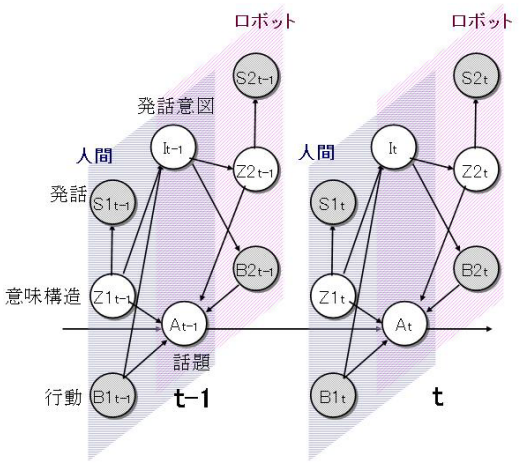

図 3：発話応答のモデル（直接観測可能なノードを灰色で示す）

$$
\begin{aligned}
&\psi(S_{1t},B_{1t},A_t,S_{2t},B_{2t}) = \\
&\max_{A_t,S2t,B2t}\{\gamma_1 \log p(S_{1t}\mid Z_{1t};L,G) && \text{【音声】}\\
&\quad+\gamma_2 \log p(O_{Tt}\mid Z_{Tlt};L)+\gamma_2 \log p(O_{Lt}\mid Z_{Llt};L) \\
&\qquad && \text{【オブジェクト】}\\
&\quad+\gamma_3 \log p(U_t\mid Z_{M1t},O_T,O_L;L) && \text{【動作】}\\
&\quad+\gamma_4 \log p(O_{Tt},O_{Lt}\mid Z_{M1t};R) && \text{【関係】}\\
&\quad+\gamma_5 \log p(O_{Tt},O_{Lt}\mid B_{1t},A_t;H) && \text{【コンテキスト】}\\
&\quad+\gamma_6 \log p(I_t\mid Z_{1t},B_{1t};CM) && \text{【相互行為的意味】}\\
&\quad+\gamma_7 \log p(Z_{2t},B_{2t}\mid I_t;RC) && \text{【応答制約】}\\
&\quad+\gamma_1 \log p(S_{2t}\mid Z_{2t};L,G) && \text{【音声】}\\
&\quad+\gamma_2 \log p(O_{Tt}\mid Z_{T2t};L)+\gamma_2 \log p(O_{Lt}\mid Z_{L2t};L) \\
&\qquad && \text{【オブジェクト】}\\
&\quad+\gamma_3 \log p(U_t\mid Z_{M2t},O_T,O_L;L) && \text{【動作】}\\
&\quad+\gamma_4 \log p(O_{Tt},O_{Lt}\mid Z_{M2t};R) && \text{【関係】}\\
&\quad+\gamma_5 \log p(O_{Tt},O_{Lt}\mid B_{2t},A_t;H)\} && \text{【コンテキスト】}
\end{aligned}
$$

ただし， L, G, H, R, CM,RC は語彙，文法，コンテキスト，オブジェクトと動作の関係，相互行為的意味，応答制約の各信念モデルを表す．またγは共有確信度を表す[Iwahashi07]．図 3 のグラフィカルモデルはこの決定関数Ψを図示したものであり，人の行為（発話，行動）と，ロボットの行為（発話，行動），および話題の対応の適切さを表す．語彙 L，文法 G，コンテキスト H，オブジェクトと動作の関係 R および共有確信度γは[Iwahashi07]の手法で学習される．相互行為的意味の信念 CM の学習については次節で説明する．

3.2 相互行為的意味の学習

(1) 発話意図のモデル

人の発話・行動と発話意図のマッピングを相互行為的意味と呼ぶ．発話意図 It は人が期待するロボットの応答として，

「I_{At}を I_{Ot}で示して欲しい」

という簡単な形でモデル化する．I_{At} は応答すべき事柄（トラジェクタ or ランドマーク or 軌道）を表し，I_{Ot} はその事柄を何で示すのか，すなわち出力モダリティ（発話 or 行動 or 応答なし）を表

す. 例えば, ぬいぐるみが箱に乗るというシーンを共有している時に, 人が「ぬいぐるみ なに 乗せる」と発話した場合, その発話意図は「ランドマークを(I_{At}), 発話(I_{Ot})で示して欲しい」であると考える.

発話意図のモデルを図 4 に示す. 発話意図は意味構造 Z_{1t} と行動 B_{1t} から導出される. ここで, 少ないインタラクションでの学習を実現するためにモデルの複雑さを必要最低限に抑えるよう, 意味構造 Z_{1t} をそのまま使うのではなく, Z_{1t} に含まれる各単語の有無を 4bit で表した Y_{1t}を利用する. 例えば表 1 の例の場合には, Z_T に単語がなく, Z_L, Z_M, Z_I に単語が含まれるため, Y_{1t} は=0111 となる. このように Y_{1t} は, 発話 S_{1t} の句構造を近似したものとなっていることから, 句構造と呼ぶ. また, 行動 B_{1t} に含まれる対象オブジェクト B_{O1t} はシーンに依存するため, 同様の理由から利用しないこととした. 発話意図 I_t は句構造 Y_{1t}, 疑問詞 Z_{I1t}, 行動内容 B_{C1t}から導出する.

(2) 応答制約

応答制約とは発話意図 I_t からロボットの意味構造 Z_{2t}・行動 B_{2t} へのマッピングである. ロボットは必ず推定した発話意図に沿った応答を行うこととし, 応答制約(発話意図 I_t から Z_{2t}, B_{2t} へのマッピング)はトップダウンで与えた. 例えば, 上のような発話意図に対しては, 応答の仮説の中でも, ランドマークの語を含まない意味構造, または, ランドマーク以外の語を含む意味構造の確率 $p(Z_{2t},B_{2t}|I_t;RC)$を 0 とすることで, 発話意図に沿わない応答を除外する. また, I_{Ot} が行動である時は, I_{At} に応じてオブジェクト操作および指差しの確率を決定する.

(3) 相互行為的意味の学習

相互行為的意味とは人の意味構造 Z_{1t}・行動 B_{1t} から発話意図 I_t へのマッピングである. 人が発話に対するロボットの応答が正しくない場合, 人はロボットの手を叩き, 正しい応答の例を示す. ロボットはその正例をロボットの発話・行動 S_{2t}, B_{2t} に入力し, $p(I_{At}|Y_{1t}, Z_{I1t}, B_{C1t})$および $p(I_{Ot}|Y_{1t}, Z_{I1t}, B_{C1t})$を学習する. しかし, このままでは Y_{1t}, Z_{I1t}, B_{C1t} の組み合わせが確率の条件となるため, 異なる句構造間や疑問詞間での学習結果の汎化は期待できない. そこで, それぞれを以下のように確率の重み付け和として近似し, それらの汎化を可能にする.

$$p(I_{At}|Y_{1t},W_{I1t},B_{C1t}) \approx$$
$$w_{a1}p(I_{At}|Y_{1t})+w_{a2}p(I_{At}|W_{I1t})+w_{a3}p(I_{At}|B_{C1t})$$
$$p(I_{Ot}|Y_{1t},W_{I1t},B_{C1t}) \approx$$
$$w_{O1}p(I_{Ot}|Y_{1t})+w_{O2}p(I_{Ot}|W_{I1t})+w_{O3}p(I_{Ot}|B_{C1t})$$

それぞれの確率と重みは EM アルゴリズムによって推定する. $p(I_{At}|Y_{1t}, Z_{I1t}, B_{C1t})$の初期分布は等確率, $p(I_{Ot}|Y_{1t}, Z_{I1t}, B_{C1t})$の初期分布は「応答なし」を出力するように与えた.

なお本報告では, オブジェクトや動作を表す単語は, 予め学習しておき, 今回の実験で新たに使われた未知の単語は全て疑問詞であるとみなし学習する. 単語がオブジェクト, 動作, 疑問詞のどのクラスに属するかの判定は今後の課題とする.

4. 実験

4.1 実験条件

人の発話に対して, ロボットが応答する実験を行う. 1回の発話・応答を1ターンと呼ぶ. 用意したオブジェクトは7種類で, そのうち3つをテーブルに配置し, 1ターン毎にその配置を変更する. また, 動作は「乗せる」「飛び越える」など計7種類とした.

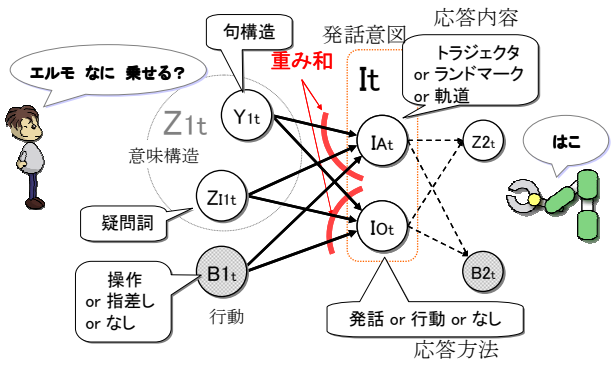

図 4:発話意図のモデル

人の発話として, ①操作の要求, ②「なに」を利用した質問, ③「どれ」を利用した質問, の3種類を想定する. ①操作の要求は, 発話された内容に合わせオブジェクトを操作させることを目的とし, 「エルモ 乗せる」や「エルモ 箱 乗せる」といった4種類の句構造(「T:L:M」「T:M」「L:M」「M」, ただし T,L,M は各々トラジェクタ, ランドマーク, 動作を表す)で発話する. ②「なに」を利用した質問では, オブジェクトを指差しながら「なに」と聞く「I」(I は疑問詞)と, オブジェクトを動かしながら「エルモ なに 乗せる」などと聞く「I:L:M」, 「I:M」「T:L:M」の計4種類の句構造を想定する. 「I」は指差したオブジェクト, 「I:L:M」「I:M」はトラジェクタ, 「T:I:M」はランドマークを発話することを要求する. ③「どれ」を利用した質問では, 「エルモ どれ」と聞く「T:I」と, オブジェクトを動かしながら聞く「I:L:M」, 「I:M」「T:L:M」の計4種類の句構造を想定する. 「T:I」は発話されたオブジェクト, 「I:L:M」「I:M」はトラジェクタ, 「T:I:M」はランドマークを要求する.

提案する学習法の汎化性能をわかりやすく示すため, 本実験では, ①操作の要求, ②「なに」を利用した質問, ③「どれ」を利用した質問を各20ターン(各句構造5ターン)ずつ順番に行った.

4.2 実験結果

実験でのロボットの応答を図 5 に示す. 横軸はターン数で, 各ターンに使用した句構造も合わせて載せる. 縦軸は応答を表す. 灰色の線は想定する正解の応答, 青丸が実際のロボットの応答を示す. 正解の応答とロボットの応答が重なる部分は全て, 質問した話題に沿った応答がなされていた.

(1) 操作の要求に対する応答

$p(I_{Ot}|Y_{1t}, Z_{I1t}, B_{C1t})$ の初期分布を「応答なし」としているため, 初めロボットは応答を返さない(図中「なし」). その時, 人が実際にオブジェクトを操作して見せることで, 次のターンからロボットは人の発話通りにオブジェクトを操作するようになる. 操作の要求には4種類の句構造を用いたが, 句構造の違いには関係なく正しい応答を返した.

(2) 「なに」を利用した質問応答

「なに」とだけ質問した場合(図中「なに」I(+P))も同様に, 一度の教示で正しく応答できるようになる. その後, 「なに 箱 乗せる」(「なに」I:L:M)や「なに 乗せる」(「なに」I:M)と句構造を変えても, 間違った応答は見られなかった. これは, 「なに」と言われたらトラジェクタを表す単語を返す(図中の発話 T)ように応答を汎化しているためである. この汎化によって, 「エルモ なに 乗せる」(「なに」T:I:M)という質問に対しても, 「エルモ」とトラジ

ェクタを発話する例が見られた．この誤った汎化は正しい応答を2回示すことで，修正された．

(3) 「どれ」を利用した質問応答

「エルモ　どれ」(図中「どれ」T:I)の場合も一度の教示で，正しい応答を学習した．しかし，「なに」の時とは異なり，句構造「I:L:M」の質問で誤った応答を返す．これは，句構造が同じ「なに　箱　乗せる」の知識を利用し，トラジェクタを発話したためである．それに対し，トラジェクタを指差すことを教えると，次のターンからは句構造が変わっても適切な応答ができるようになる．

(4) 発話意図モデルの重みの変化

この汎化の振る舞いは，学習した重みから説明できる．図 6 に $p(I_{At}|Y_{1t}, Z_{I1t}, B_{C1t})$および $p(I_{Ot}|Y_{1t}, Z_{I1t}, B_{C1t})$で用いた重みの変化を示す．操作の要求から「なに」の中盤までは，句構造 Y_{1t}，疑問詞 Z_{I1t}，行動内容 B_{C1t}を単独で用いても正しい応答ができるため重みに変化は見られない．しかし，ロボットが2度の汎化誤りをした「エルモ　なに　乗せる」を学習すると，$p(I_{At}|Y_{1t}, Z_{I1t}, B_{C1t})$の計算に用いる重みが変化し，$p(I_{At}|Y_{1t})$の重みが最も高くなる．これは，「応答すべき事柄($I_{At}$)は，句構造 Y_{1t} で規定される」ことを学習したことを意味する．

また，$p(I_{Ot}|Y_{1t}, Z_{I1t}, B_{C1t})$の計算に用いる重みが変化するのは，「どれ　箱　乗せる」を学習した時である．前述した通り，ここでロボットは「なに」と「どれ」を区別せずに応答した．しかし，正しい応答例を示すことで，$p(I_{Ot}|Z_{I1t})$の重みが最も高くなる．これは，「出力モダリティ(I_{Ot})は，疑問詞 Z_{I1t} で規定される」ことを学習したことを意味する．すなわち，「なに」なら発話し，「どれ」なら指差しすることを学習した．

このように，応答すべき事柄か(I_{At})を句構造 Y_{1t}から求め，疑問詞 Z_{I1t} から出力モダリティ(I_{Ot})を求めるようになると，「どれ　乗せる」や，「エルモ　どれ　乗せる」といった未知の質問に対しても，適切に応答ができるようになる．

(5) 相互行為的意味の正解率

最後に $p(I_{At}|Y_{1t}, Z_{I1t}, B_{C1t})$および $p(I_{Ot}|Y_{1t}, Z_{I1t}, B_{C1t})$の正解率を図 7 に示す．各時点のモデルに対し，想定する発話の全パターン(Y_{1t} t, Z_{I1t}, B_{C1t} の組み合わせ16種類)を入力し，確率が最大となる発話意図 It が，想定する応答と一致すれば正解とした．比較のため，$p(I_{At}|Y_{1t}, Z_{I1t}, B_{C1t})$および $p(I_{Ot}|Y_{1t}, Z_{I1t}, B_{C1t})$を重み和ではなく，3 変数の組み合わせとして計算した場合の結果(図中 AND モデル)を載せる．この結果から，提案モデルの汎化性能の高さがわかる．また，全てを学習してもモデルが破綻することなく，各質問に対する適切な応答が保存されることが示されている．

5. まとめ

本報告では，対話者との事物に対する共同注意と共有信念に基づいた発話応答の対話構造を表すダイナミカルグラフィカルモデルと，これを学習する計算機構を示した．これにより，ロボットが，人の発話と行動の意図を推定することにより，状況に応じて適切な応答を生成する能力を，人とのインタラクションを通して学習できることを実験により検証した．また，発話意図と単語の関係を学習することで，疑問詞(「なに」と「どれ」)の意味を，事物にではなくコミュニケーション機能に接地するものとして学習することができた．

本実験で獲得した知識は人とロボットの役割が反転した場合，すなわちロボットが発話し，人が応答する場合にも再利用ができる．今後は，曖昧な発話があった場合などに，ロボットが人に質問して曖昧性を解消していくような，より自然なコミュニケーションの実現を目指す．

謝辞

本研究は，国立情報学研究所共同研究「能動的ハンドインタラクションによる実世界言語コミュニケーションの学習に関する研究」による研究助成を受け実施したものである．

参考文献

[Roy02] Roy, D.: Learning visually-grounded words and syntax for a scene description task, Computer Speech and Language, Vol.16, No.3, pp353-386 (2002).

[Iwahashi07] Iwahashi, N. “Robots That Learn Language: A Developmental Approach to Situated Human-Robot Conversations,” In Sankar, N. ed. Human-Robot Interaction, pp.95-118, I-Tech Education and Publishing (2007).

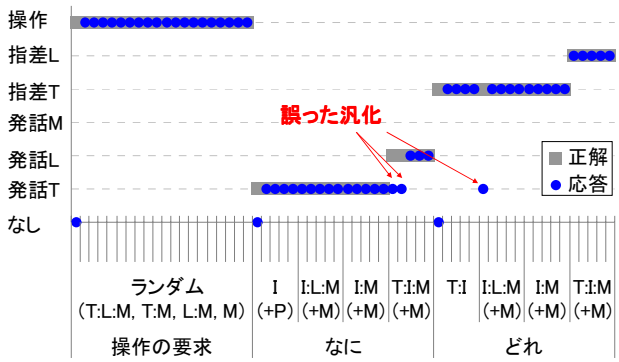

図 5：ロボットの応答

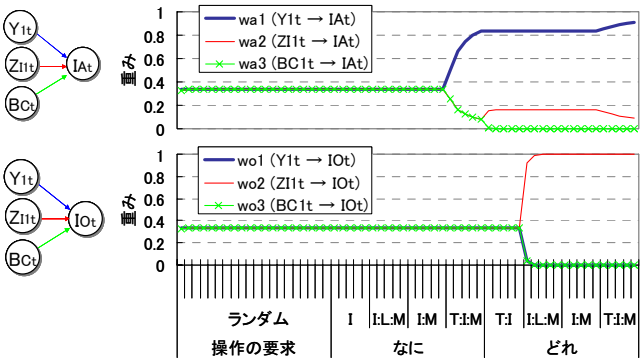

図 6：発話意図モデルの重みの変化(w_a：上，w_o：下)

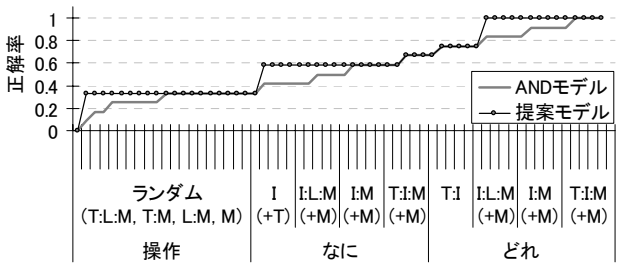

図 7：相互行為的意味の正解率