オブジェクト同定のための語彙の視覚状況依存性の分析とモデル化

山肩 洋子 † 河原 達也 † 美濃 導彦 ‡
京都大学 情報学研究科 † 京都大学 学術情報メディアセンター ‡

1. はじめに

近年のロボット技術の発展により，ロボットによる自律行動が可能となるにつれ，汎用パーソナルロボットへの次なるステップとして，人間との円滑なコミュニケーションに重点を置かれた研究が広がりつつある．ロボットの利便性のためには，ロボットが家庭の中で何らかの明確な役割を持ち，ユーザの意図どおりのサービスを行うための意志伝達を達成できることが望ましい．また，ユーザに特別な装備や訓練を要求することなく，コミュニケーションをはかることも重要な要素となる．そこで我々は，人間対人間のコミュニケーションにおいて，論理的意志伝達にもっとも頻繁に用いられる音声による言語情報をユーザとロボットとのコミュニケーションメディアとし，ユーザが意図したオブジェクトをロボットに指示して持って来てもらうというタスクを対象にして，ユーザの自然発話を理解する機構の研究を行っている．

実世界において音声により指示されたオブジェクトを同定する際には，様々な曖昧性が生じる．本稿ではこれらの曖昧性のうち，言葉からオブジェクトに至る参照における，視覚的状況依存性に着目し，シミュレーション実験により，その特性を調べる．さらにその結果をふまえて，信念ネットワークを用いた言語理解モデルを提案する．

2. 音声対話によるオブジェクト同定タスク

本研究で扱うタスクは，参加者が命令者 (ユーザ側) と実行者 (ロボット側) の 2 名，両者のコミュニケーション手段は音声のみとする．両者は複数のオブジェクトが描かれた同一の画像を持っており，命令者はそのうち一つのオブジェクトをターゲットとして選択して，これがどのオブジェクトであるかを音声により実行者に伝達する．両者は音声対話により曖昧性を解消し，実行者が正しくターゲットを決定できればタスク達成とする．(図 1).

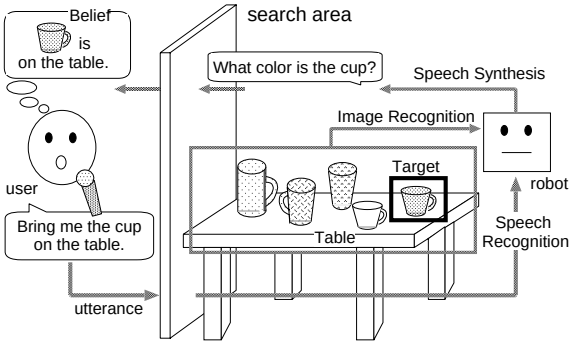

図 1: ユーザとロボットによる協調的なオブジェクト探索タスク

3. 単語によるオブジェクト参照モデル

ロボットはユーザの発話に対し，これが参照する視覚的特徴を推定して，これに合致するオブジェクトを探索するのであるが，ユーザが発した言葉が指し示す特徴は必ずしも定まらず，状況に大きく影響を受ける．たとえば [1] では，オブジェクトがどのように使われているかによって，同じオブジェクトでも名称が変化することを示した．また [2] では，ターゲットの用途

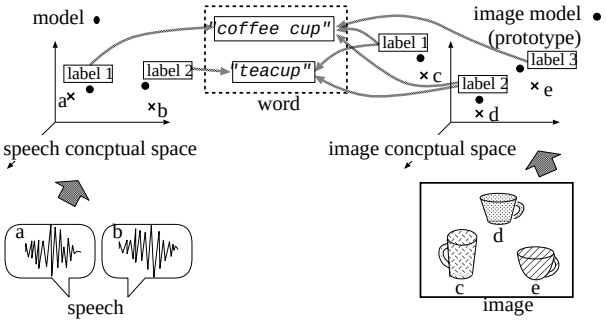

図 2: 音声から画像に至る参照の曖昧性

Analysis and Modeling of Visual Context Dependency in Object Reference by Natural Language

†Yoko YAMAKATA and Tatsuya KAWAHARA, School of Informatics, Kyoto University
‡Michihiko MINOH, Academic Center for Computing and Media Studies, Kyoto University

によってオブジェクトを認識するといった「機能モデル」も提案されている．我々は，これらの『ターゲット自身に付随する状況』以外に,『ターゲットと競合するその他のオブジェクトの状況』にも影響を受けると考え，視覚的状況に依存した語彙のオブジェクト参照について研究する．

ユーザの発した『名称』から画像に至るオブジェクト参照のモデルを図 2 に示す．ロボットの視覚にあたるカメラにより撮影されたオブジェクトの画像特徴量は，オブジェクトの三次元構造の推定を経て，画像概念空間に投射される．この処理を画像認識部と位置付け，ここで生じる曖昧性を画像認識における曖昧性とする．次に，画像概念空間に投射されたオブジェクトはイメージモデルとマッチングされる．イメージモデルとは，画像概念空間において多数の画像データをクラスタリングしたものであり，オブジェクトのプロトタイプとしての意味合いを持つ．

我々は [3] で，このモデルに個人差がどう関わっているかを調べた．コップ類に属する『名称』15 単語と，線画で表したイメージモデル 14 種類との間の参照関係を調べるアンケート調査を行った結果,『名称』によるオブジェクト参照が個々のユーザに強く依存していることがわかった．これより，状況変化のオブジェクト参照への影響においても，オブジェクト参照に個人差があることを念頭に置いた上で解釈しなければならない．

4. 状況変化によるオブジェクト参照への影響

本章では，ターゲットと，これと競合するオブジェクトを配置させたときの，ユーザのターゲットに対する名称の変化を調べることにより，名称の選定における状況変化の影響を調べる．

本研究の目的に則したオブジェクトは，すなわち，ユーザにとってまったく別のオブジェクトとみなせるが，言語体系においては同一のカテゴリに属するようなオブジェクト集合である．[1] は，カップのような線画一つを基本オブジェクトとして選出し，これを縦・横それぞれの方向に引き伸ばしたとき，名称が変わり得ることを実験により示している．つまり，トポロジー的に同一の物体でも，そのいくつかのパラメータを変化させることにより，名称が変わる可能性があることを示している．我々はこの実験にならい，図 3 に示されるような 3D CG によるオブジェクトを基本形として選定し，これを縦・横にそれぞれ引き伸ばしたオブジェクトと合わせて，3 種類のオブジェクトを選出した．さらに [3] で行ったアンケートにより，名称が変化する上で,『取っ手』の有無が重要な要素となっていることがわかった．そこで，この特徴を合わせ，表 1 で示した 6 つのオブジェクト「NH(Normal with Hand)」「WH(Wide with Hand)」「HH(High with Hand)」「NN(Normal No-hand)」「WN(Wide No-hand)」「HN(High No-hand)」を以降の実験に用いることにする．被験者の数は 20 名である．

4.1 ユーザの個人差に基づくオブジェクトの選定

基本のオブジェクトを縦・横方向にそれぞれ引き伸ばして行く際，元のオブジェクトと『違う』とみなせる閾値は個々のユーザによって異ると考えられる．そこで，個々のユーザごとに次のような予備実験を行い，基本のオブジェクトを縦・横方向にそれぞれ引き伸ばした際,『コップ類』とみなせる限界のオブジェクトを選定することにした．

実験に用いたシステムは図 3 のようなものである．このシステムでは,「HIGH/LOW」ボタン，あるいは「WIDE/NARROW」ボタンを押すことにより，基本のオブジェクトの 3D CG データを縦・横方向の引き伸ばし率を調節すことができる．実験結果のブレを考慮して，実験は基本形から引き伸ばす方向と，明らかに『コップ類』とはみなせない程度に引き伸ばしたものを縮める方向の二種類を二回ずつ行った．被験者に与えた教示は,「『コップ類』とみなせる限界まで『WIDE(あるいは HIGH)』ボタンを押してください.」と「『コップ類』とみなせるようになるまで『NARROW(あるいは LOW)』ボタンを押してください.」である．実験の順序が影響を及ぼさないよう，実験の間には適宜妨害

図 3: 予備実験用システム

表 1: 実験結果

	 NH	 WH	 HH	 NN	 WN	 HN
オブジェクトが一つのときの名称の種類	4(種類)	8	7	5	15	4
オブジェクトが 2 個のとき 名称が変化した被験者数	7(名)	5	9	6	6	5

(被験者; 20 名)

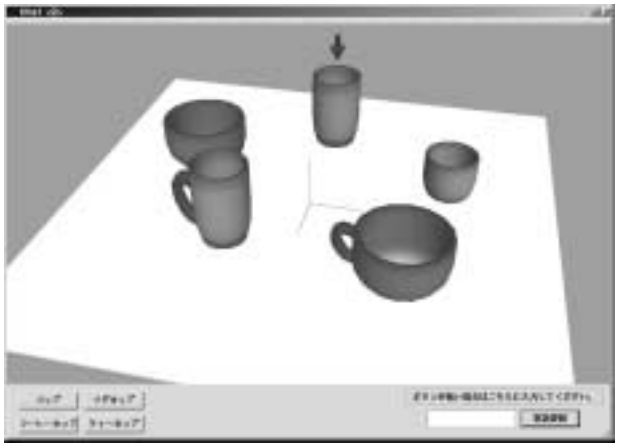

図 4: 実験用システム

課題を加えた．最後に，縦・横それぞれに限界の引き伸ばし率の平均値を算出し，実験に用いるオブジェクト集合を決定した．

以上の予備実験により，被験者ごとに選出された 6 種類のオブジェクトに対して，名称を調べるシステムを図 4 に示す．このシステムでは，画面中に一つ以上のオブジェクト (すべて同一色) が，テーブルの上に置かれている状態で映し出される．相対表現をなるべく避けるよう，視点を常に動かしている．矢印がそのうちの一つのオブジェクトを指しており，これがターゲットである．被験者は，画面の状況において，ターゲットを指し示す呼び名を，キーボードにより入力するよう求められる．被験者の入力負荷を軽減するため，一度入力された名称は，それ以降ボタンとなって画面下部に現れ，二度目以降はボタンを押すだけで入力できるようになっている．被験者に与える教示は以下の通りである．

「あなたはある部屋に入り，画面のような光景を見たあと出てきました．その後，その部屋に入った人に対し，矢印が指しているオブジェクトを携帯電話で指示する場合，あたなが用いる名称を答えてください.」(一部省略)

本研究では『名称』に注目するため,「大きいほう」「背の高いほう」といった相対的な表現は使わないよう，被験者に注意を行った．

4.2 オブジェクトが一つのときの『名称』による参照

まずターゲット一つしか配置されていない場合の名称を調査した．各オブジェクトに対して，被験者の用いた名称の種類の数を表 1 に示す．これによると，たとえ同じ形状のオブジェクトであっても，少なくとも 4 種類以上の名称が用いられることがわかり，ユーザの個人差を学習することが，ユーザの発話を理解する上で重要であることがわかる．すべてのオブジェクトに別々の名称を付けた被験者は 6 名しかおらず，平均 4.9 単語 (最小 3 単語) の名称を用いている．これより，名称によるオブジェクト指定は，少なくとも一対多となる可能性が高いといえる．同じ名称を付けられた場合が多いオブジェクトの組み合わせは，[NH,HH](7 名)，[NH,WH](6 名)，[HH,HN](5 名) であった．

4.3 複数オブジェクトを配置させたときの『名称』による参照

次にターゲット以外のオブジェクトを配置したときの名称の変化について調べる．ターゲット以外のオブジェクトを一つ加えたとき，名称を変えた被験者の数を同じく表 1 に示す．これによると，すべてのオブジェクトについて，名称が変わる可能性があることがわかる．このうち，基本のオブジェクトである [NH] に注目すると，名称が変わった被験者のうち 5 名が,「コップ」や「カップ」といった抽象的な言葉から「コーヒーカップ」「マグカップ」のような抽象度の低い言葉に

変化したものであった．それ以外のオブジェクトについても，抽象度が低から高への変化はほとんど見られず，競合するオブジェクトが増えると，その分曖昧性の少ない名称を選ぶ傾向があった．しかし,「マグカップ」から「コーヒーカップ」といった，抽象度が同程度の名称への変化も頻繁に見られ，名称からオブジェクトへの参照関係を表現するためには，一対多ではなく，多対多の関係を扱う必要があることが確認された．

さらに，ターゲット以外のオブジェクトの数を二つ以上に増やした場合，抽象度が低くなる方への名称の変遷は見られるが，同程度の抽象度の名称への変化は少なくなってゆくことがわかった．これは，候補となるオブジェクトが増えて，ターゲット同定における状況の曖昧性が増加したためであると考えられる．

以上の実験により，名称とオブジェクトとの参照関係は多対多で表現するのが適当で，ユーザモデルを学習した場合でも，1 つのオブジェクトに絞るのは無理があり，適合する可能性のあるオブジェクトをすべて考慮に入れる必要があることがわかった．

5. 信念ネットワークを用いた適応的言語理解

4 章の実験結果をふまえて，我々は、音声・言語・画像レベルの情報を信念ネットワークを用いて統合することにより，包括的に解消する機構を提案した [3]．信念ネットワークは,『名称』や『色』など，オブジェクトの属性ごとに独立に構築する．例えば『名称』に関する信念ネットワークでは，まず,『名称』に関する音声と，その認識候補である単語が，音声認識の信頼度によって結びつけられる．次に，単語とイメージモデルとは,『名称』における関連性を示すユーザモデルで結びつけられ，ユーザに適応的な言語理解を実現する．さらに，イメージモデルとオブジェクトは，画像概念空間における類似度で結びつけられる．この信念ネットワークを音声レベルから画像レベルまでたどり，確信度を伝搬することにより,『名称』に関する音声が画像中の各オブジェクトを指示する確信度が算出される．最後に全ての属性における確信度を統合し，最終的な確信度とする．ここで，確信度の最も高いオブジェクトをターゲットと仮定，その曖昧度を確信度のエントロピーで評価することにより，ターゲットを確定するか，あるいはより詳細な情報を求めてユーザに質問するといった対話戦略を決定する．ユーザモデルは，ターゲットの同定結果により各ユーザに適応するよう学習する．

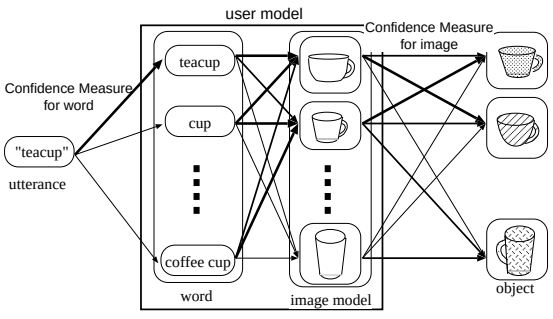

図 5: 信念ネットワークを用いた適応的言語理解

以上のシステムをソフトウェアロボットとして実装した．音声認識には Julian 3.3 を，音声合成には CHATR version 9.4 をそれぞれ用い，対話管理部は Perl，GUI は C++と OpenGL により作成している．4 章の実験と同様のタスクで実験を行ったところ，ユーザモデルの学習と，音声・言語・画像情報を合わせた包括的な曖昧性の解消について，動作が確認された．

6. おわりに

本稿では，ユーザがロボットに音声によりターゲットオブジェクトを指示する際，音声からオブジェクトに至る参照の曖昧性を解消することを目標として研究を行った．そこで，名称によるオブジェクト参照の状況依存性に注目し，コップ類のオブジェクトが一個，あるいは複数個配置されたときの，名称との参照関係を調べる実験を行った．その結果，名称とオブジェクトの参照関係が多対多で表され，複数の参照関係を維持しながら，最尤の候補を探索する必要があることが示された．これにより，信念ネットワークを用いたオブジェクト同定モデルを提案し、ソフトウェアロボットとして実装して，ユーザモデルの学習と，複数情報源による曖昧性の解消について動作を確認した．

謝辞

実験デザインおよびデータ収集に協力して下さった，京都大学教育学研究科 小島隆次氏に深く感謝する．

参考文献

[1] William Labov. The boundaries of words and their meanings. In *New Ways of Analyzing Variation in English*, 1973.

[2] 服部洋一, 黄瀬浩一, 北橋忠宏, 福永邦雄. 動的機能のモデルに基づく物体の機能認識. 情報処理学会論文誌, Vol. 36, No. 10, 1995.

[3] 山肩洋子, 河原達也, 奥乃博. ロボットとの音声対話のための信念ネットワークを用いた適応的言語理解. 人工知能学会研究会資料, SIG-SLUD-A201-3, 2002.