K-29　音声対話におけるオブジェクト同定のための曖昧性の解消
Disambiguation of Object Reference by Speech Dialogue for Object Search Task

山肩 洋子† 河原 達也† 奥乃 博† 美濃 導彦‡
Yoko Yamakata Tatsuya Kawahara Hiroshi G. Okuno Michihiko Minoh

1. はじめに

ユーザが音声発話により指示したオブジェクトを持って来るロボットの実現を目指す。ユーザがロボットに気軽に命令するためには、音声対話のみでロボットがユーザの意図を理解できることが望ましい。実世界を対象とした自然言語による指示の理解においては、曖昧性が大きな問題となるため、ユーザの発話が指し示しているオブジェクトを同定する機構について研究する。

まず、ユーザの発話に含まれる単語から、画像中のオブジェクトに至る参照において生じる種々の曖昧性を分類し、これらを信念ネットワークを用いて包括的に解消する機構について述べる。次に、この機構のうち、単語とオブジェクトの画像特徴における概念空間との参照関係に注目し、そこに一貫した個人差が存在することを実験により示す。また本研究で提案した機構の枠組において、従来研究で問題とされてきた、状況変化による参照関係への影響に関しても調査する。

2. 音声対話によるオブジェクト同定タスク

本研究では以下のようなタスクを対象に研究を行う。参加者は**命令者 (ユーザ側)** と**実行者 (ロボット側)** の2名、両者は互いに見えない位置にいて、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。

3. 音声によるオブジェクト参照における曖昧性

ユーザの発話から画像に至るオブジェクト参照を図1に示す。カメラにより撮影されたオブジェクトの画像特徴量は、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、ここでの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。本稿では、このようなイメージモデルがすでに得られているとし、これと単語との参照関係について考える。我々は、この参照関係が個人に強く依存していると考え、以降、これを**ユーザモデル**と呼ぶことにする。

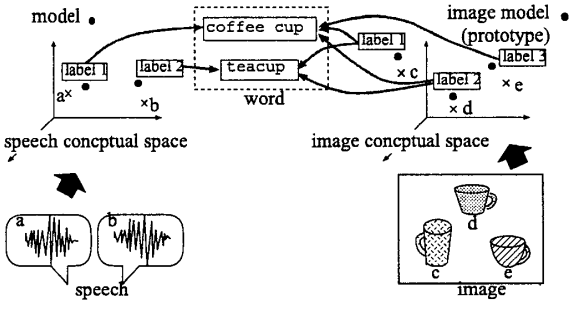

図 1: 音声から画像に至る参照の曖昧性

ユーザの発話に対する処理においても、画像と同様の曖昧性が発生するが、音声と単語の参照では音素を媒介するため、この曖昧性が比較的小さい。本研究では、ユーザの発話から単語に至る曖昧性は、音声認識部において、まとめて扱うことにする。

4. 信念ネットワークを用いた適応的言語理解

我々は、3章で述べた種々の曖昧性を、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研究している (図2)[1]。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

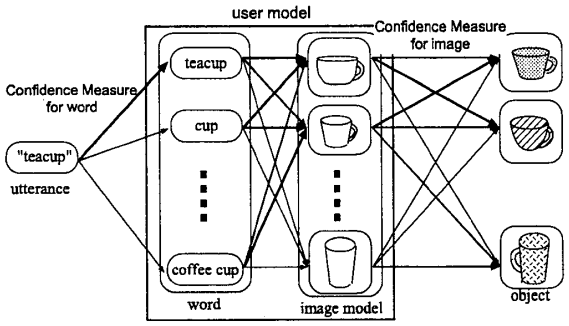

図 2: 信念ネットワークを用いた適応的言語理解

†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

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、初期値から始めて、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。

5. ユーザモデルの特性

ユーザが音声によりあるオブジェクトを指し示した際、たとえロボットがこの音声を正しく音声認識できたとしても、ユーザの意図したオブジェクトを一意に同定する事は容易ではない。同一のオブジェクトに対し、ある人は「ティーカップ」と呼ぶが、別の人は「コーヒーカップ」と呼ぶ場合もある。しかし、もしロボットがユーザモデルを予め獲得していれば、ユーザの発話に対し正しくターゲットの候補を選択し、さらにその発話がどの程度曖昧であるかを判定することが可能となる。例えば「コップを取ってほしい」と言われた際、「コップ」という言葉で常に特定のオブジェクトを指し示すユーザに対しては、すぐさまそのオブジェクトを取ることができるし、どんなタイプも「コップ」と呼ぶユーザに対しては、「コップ」という言葉だけではユーザの意図したオブジェクトを同定できない事を理解し、『色』や『柄』など、その他の属性情報について質問する判断を下すことができる。

画像メディアを対象とした言語理解における曖昧性の研究としては、オブジェクトの名付けに対する文脈の影響を調べた [2] や、言語による空間描写からの 3 次元情報の復元を目指した『SPRINT』[3] などが挙げられるが、これらでは個人差による曖昧性は扱われなかった。

5.1 単語とイメージモデルとの参照関係における個人差

我々は、イメージモデルと単語の参照関係において、(1) 個人差があり、(2) その個人差は (高々一ヶ月間は) 一貫しているという仮説を立て、同一被験者に対し一ヶ月おいて二度、図 3 に示すようなアンケートを行った。これは、EDR 電子化辞書において「コップ類」に分類された 15 単語と、線画で表されたコップのイメージモデル 14 種類との参照関係を、3 段階で評価するものである。

被験者 12 名の回答を、ピアソンの積率相関係数により評価した。一ヶ月後の本人との相関係数の平均は 0.77 と、かなり強い相関があり、一ヶ月では個々人が持つ参照関係のモデルはそれほど変化しないと言える。また、他者との相関係数の平均は 0.56 であり、他者についても、ある程度共通した参照関係のモデルが期待できると言える。つまり、大雑把な一般化ユーザモデルが仮定できるが、個々のユーザに適応した方が良い結果が得られることがわかった。我々はこの結果に基づき、このアンケート回答の平均値を、4 章におけるシステムのユーザモデルの初期値とする。

5.2 状況変化に対するユーザモデルの影響

従来研究では、文脈や状況の変化により、同一オブジェクトに対する言語表現が影響を受けることが報告されている [2]。そこで、本システムにおけるユーザモデルの部分が、実際には文脈や状況の変化によりどのような変化を示すか調査した。本研究で最も大きな状況変化である、画像中のオブジェクトの数や組み合わせが変わったときの、単語とイメージモデルの参照への影響を調べるため、次のようなアンケートを行った。まず、5.1 節のアンケートで用いたイメージモデルのうち 6 種類について、(A)1 種類ずつの画像 6 枚と、(B)3 種類を組み合わせた画像 5 枚を用意した。被験者に (A) と (B) の画像を一枚ずつ交互に提示し、画像中の各イメージモデルに対し、その形状特徴をもつオブジェクトをだれかに取って欲しいとき、どのような名称で呼ぶかを書いてもらった。回答数は被験者一名につき (A)6 項目、(B)15 項目で、参加した被験者数は 13 名である。

image model / word				
glass	1	3	2	3
tumbler	3	1	3	3
teacup	1	3	1	1

図 3: 単語とイメージモデルの関係についてのアンケート回答例

結果、同じイメージモデルにも関わらず (B) で (A) と違う名称を選んだ回答数は、全回答数 195 項目中 68 個と、65.1%に上った。代わりとして選ばれた名称の 48.5%は「カップ」や「コップ」であり、状況に応じて、より抽象度の高いメタレベルの単語を、ターゲットの名称に選ぶ傾向があることが分かった。また、17.6%は、もともと (A) で「カップ」や「コップ」を選んでいたものが、そのサブクラスである抽象度の低い単語に変化したものであった。つまり、状況変化はイメージモデルに対し選ぶ単語の抽象度に影響を与えるが、単語とイメージモデルの参照の強さが反転するといった、互いに素となるような変化ではないことが推察される。

6. まとめ

本稿ではまず、音声によるオブジェクト参照において生じる曖昧性を分類した。次に、ユーザの個人差をユーザモデルで扱い、音声・言語・画像レベルの情報を信念ネットワークによって統合することにより、曖昧性を包括的に解消する機構を提案した。最後にアンケートにより、単語とイメージモデルの参照関係に一貫した個人差があり、状況に応じてオブジェクトに対し選ぶ単語の抽象度が変わる現象が見られることを示した。被験者数を増やし、調査結果の信頼性を上げることが今後の課題である。

参考文献

[1] 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.

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

[3] 山田篤, 網谷勝俊, 星野泰一, 西田豊明, 堂下修司. 自然言語における空間描写の解析と情景の再構成. 情報処理, No. 5 in Vol. 31, pp. 660–672, 1990.