画像とそれに対する発話を対象とした幼児の名詞獲得モデル

内田 ゆず　　荒木 健治
Yuzu Uchida　Kenji Araki
{yuzu, araki}@media.eng.hokudai.ac.jp
北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

１．はじめに

社会の情報化が急速に進む現在，ロボットが身近な存在になりつつあるが，我々が彼らと意思の疎通を図ることは大変困難である．現在でも，限定されたタスクの範囲内であれば，機械は人間と対話を行うことができる．しかし，それは，その機械の設計者が想定した状況を逸脱した瞬間にコミュニケーションがとれなくなるということを意味する．本当の意味で「一家に一台のロボット」，「ロボットが家族の一員に」という生活を実現するためには，より人間らしい対話能力が不可欠である．

既存の対話システムは，大人の言語処理能力をモデルとして構築されている．つまり，完璧に言語を使いこなせるモデルを一足飛びに作ろうとしている．しかし，人間の言語能力とは極めて複雑なものであるため，このようなアプローチでは汎用的な対話システムに到達することは非常に困難であると考えられる．そこで，我々は，人が言語能力を獲得する能力を模倣したシステムを作成することにより，最終的に人間と同等の対話能力の実現が可能なのではないかと考えている[1].

人間の幼児の言語獲得過程にはいくつかの発達段階が見られる．生後数ヶ月はまだ言葉にならない音（喃語）を発するだけであるが，生後１歳から１歳半くらいになると，意味のある一つの単語を発話するようになる．この時期は「一語期」と呼ばれる．さらに生後２歳くらいまでに「二語期」といわれる段階に達し，次に「あっち，お兄ちゃんいる」のような「電報文」と呼ばれる段階がくる．３歳くらいまでには語彙数も数百語になり，発音も大人の言葉に近いものになる[2].

本研究では，言語獲得能力を工学的に実現する第一歩として，幼児の「一語期」に数多く獲得される名詞語彙を，予め語彙および統語的知識を与えない状態でユーザの発話から獲得するシステムを構築した[3]．本稿では，作成したシステムの概要，及び評価実験，そして今後の展望について述べる．

２．名詞獲得システム

2.1 幼児の言語獲得モデル

幼児は，ある言葉が指し示すものを的確に捉えることで，効率的な言語獲得を成し遂げている．このように，無数にある概念を有効に制限する能力が，初期の言語獲得においては非常に重要である．この制限のひとつの方法が「制約の理論」[2][4]である．「制約の理論」のもとでは，いくつかの個別の「制約」が提案・検討されている．本システムは，それらの諸制約のうち，「なじみのない物体につけられた未知の名前は物体の部分，色，素材，特質などでなく物体全体の名前である．」という事物全体バイアス，「未知の物体に対してつけられた名前は，その物体を含むカテゴリに対する名前である．」という事物カテゴリバイアス，「幼児は獲得したラベルを形状が似ている他のものに拡張する」という形状類似バイアス，「異なる名詞が同じ対象を指示した場合，ふたつの語はまったく同一ではない．」という対比の原理，の 4 つの制約に基づいたモデルによって構築される．

2.2 タスク

幼児が言語を獲得するには，次の３つの過程が必要であるとされている[5].

・インプットを単語分割するため，母語の音声的な特徴を分析する．
・音声インプットから単語を切り出す.
・切り出した単語に意味を付与する．

このうち，音声を分析する部分，音声から単語を切り出す部分は本研究の対象外としている．したがって，テキストから単語を切り出し，その単語に意味を付与するという過程の実現が研究対象となる．

本研究のタスクは，画像を提示しながら発話された文から学習して，その画像に対応する名詞，つまりラベルを獲得するというものである．これを，前述の「制約の理論」をモデル化した言語獲得のための能力だけで実現する．予めシステムに与えられる言語的知識は無い．与えられるのは，名詞を獲得するためのいくつかのルールのみである．この点は，言語獲得過程のシミュレートを行うシステム Mlas (Multi-Language Acquisition System)[4]などとの大きな差異である．

2.3 処理過程

本システムの処理は大まかに分けて，入力・共通部分抽出・スコア計算・出力・ユーザの評価・名詞獲得・ラベル獲得ルール生成の７つで構成される．図１に処理の流れを示す.

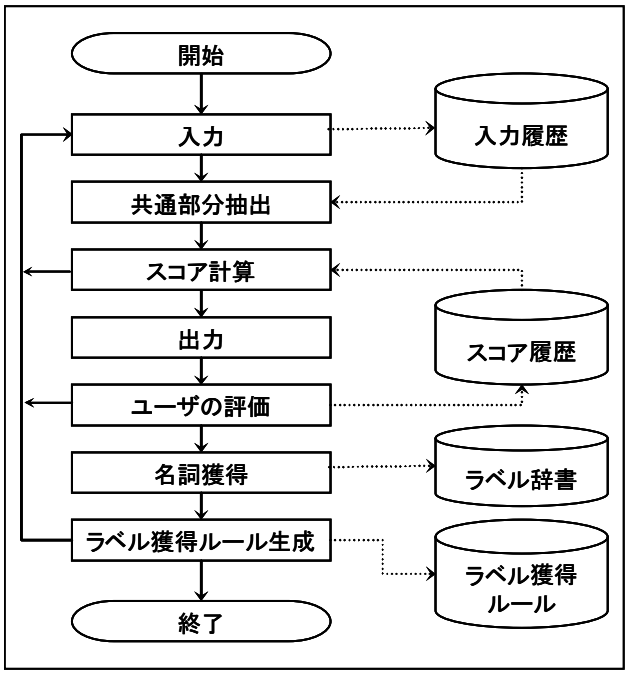

図 1 処理の流れ

2.3.1 入力

入力は単一の動物が写った写真 1 枚（以降画像 P と呼ぶ）と，その動物を見ながら大人が幼児に話かける発話 1 文（以降文 S と呼ぶ）の対である．

写真は 10 枚用意されており，ユーザはその中から任意の 1 枚を選択できるようになっている．

入力文は全てひらがなで表記され，入力文に形態素解析などの前処理は一切施されない．ひらがなで表記するのは，入力された文字列自体に意味が含まれてしまうことを避けるためであり，形態素解析などを行わないのは，幼児が正確な品詞分割などの能力を持っていないと考えるためである．

また，今回行った実験では音声入力を用いたが，その際に発生する音声認識誤りは修正しない．これは，大人はさまざまな背景知識や言語知識によって多少の音声認識誤りを修正して理解することが可能であるが，幼児にはそもそもそのような知識がないため，修正する能力もないという理由からである．

2.3.2 共通部分抽出

システムは入力を得ると，過去に画像 P とともに入力された文と文 S を比較して，字面が一致する文字列を切り出す．この切り出された文字列を共通部分と呼ぶ．これ以降の処理で共通部分は，画像 P に対応するラベルの候補として扱われる．

2.3.3 スコア計算

抽出された共通部分には基本スコアが付与される．基本スコアは，出現頻度が高く，文字数が多く，他の画像と共に出現することのない共通部分が高くなるように決定する.スコア計算式は式(1)のようになる．

ここで，α は共通部分が他の画像とともに出現している場合スコアを減少させるように働く係数，F は共通部分が同一画像と共に出現した頻度，PN は画像の出現回数，L は共通部分の文字数である．

$$SCORE = \alpha \times \frac{F}{PN} \times \sqrt{L} \quad \cdots\cdots\cdots (1)$$

本システムでは，画像の類似度によるラベル拡張という処理も行う[3]．これは，幼児が獲得したラベルを形状が似ている他のものに拡張する[2]という能力をモデル化したものである．このようなルールがあれば，全ての事物のラベルを 1 から学習する必要はなくなるので効率的な言語獲得が実現される．また，ラベルの拡張はラベルの対象の範囲を認識することにも役立つので，固有名詞獲得の過程でもこの処理は重要だと考えられる．

具体例を用いて手法を説明すると，ある犬の画像にラベル「わんわん」が獲得されたとき，その犬と形状が類似した動物について，「わんわん」という文字列がラベルになる可能性が高いと考え，スコアを増加させる，というものである．

画像の類似度の決定方法は，実用を考えると画像処理を用いることが望ましいが，実現には技術面でまだまだ問題が残されている．そこで，動物の写真を見てそれぞれの動物の形状がどの程度似ているかを 10 段階で評価する，というアンケートによって決定した．

2.3.4 出力

2.2.3 で述べたように求めたスコアが閾値を超えた共通部分は，画像 P のラベル候補となる．そして，テキストが表示されると同時に，音声合成で出力される．

2.3.5 ユーザの評価

システムの出力に対してユーザは次の 3 つのキーワードのうち，最も相応しいものを選び，音声で入力する．この場合も，音声認識誤りは修正しない．

- 「じょうず」
  ラベルとして適切である
- 「おしい」
  ラベルとしては適切でないが意味はわかる
- 「ちがうよ」
  意味がわからない

ユーザの反応によってその共通部分のスコアは再計算される．幼児がこれらのキーワードを理解するとは考えられないが，実際には，大人の表情や声の調子で感じ取ることのできる情報は多い．本手法ではそれらの代わりにキーワードを用いることとする．

2.3.6 名詞獲得

「入力」から「ユーザの評価」の処理を繰り返した結果，再計算されたスコアが閾値を超え，さらに「じょうず」という評価を得たことがある共通部分は画像 P のラベルとして獲得される．

2.3.7 ラベル獲得ルールの生成[7]

ラベル獲得ルールとは，再帰的な名詞獲得を行うためのルールである．人間は過去に得た知識を活用し，より効率的に学習を進めていく．本手法ではそのような再帰的な学習を次のように行う．

ある事物に対するラベル L を獲得すると，その事物に関する過去の入力文のラベル L を含む文から，ラベル獲得ルールを生成する．ラベル獲得ルールとは，図 1 のようにラベルの部分を変数とすることで，入力文を抽象化したものである．次に，生成したラベル獲得ルールに合致する入力文があった場合，変数部@1 に相当する部分を切り出し，スコアを上昇させる．

獲得したラベル　:わんちゃん
過去の入力　:あっちにわんちゃんがいるよ
↓
ラベル獲得ルール:あっちに@1がいるよ

図 2 ラベル獲得ルールの生成

これは，人間は様々な表現を聞いているうちに，どのような表現がラベルを示すものなのかを学習して，より効率的に学習を進めていると考え，その様子をモデル化したものである．

2.4 ユーザインタフェース

本システムのユーザインタフェースを図 3 に示す．なお，Microsoft Visual Studio .NET 2003 を用いてこのインタフェースの構築を行った．

図 3 ユーザインタフェース

3．実験と考察

3.1 実験方法

被験者に実際に本システムを使用してもらい，実験を行った．被験者に与えられた情報は，システムの操作方法，このシステムには提示した画像にラベルをつける能力があること，システムが何か出力した場合はそれに対して評価をすること，である．また，全ての被験者には音声を用いて入力を行ってもらった．使用した音声認識ソフトは IBM ViaVoice VoiceCenter である．実験の際には，10 枚の動物の画像から 1 枚を選び，その動物を見せながら赤ちゃんに話しかけるように発話することを被験者に依頼した．音声入力に慣れていない被験者にとって，この実験は多大な負担があるので，入力を 100 回程度行った段階で，結果に関わらず実験を終了することとした．被験者は 20 代女性 2 人（被験者 A・C），40 代女性 1 人（被験者 B），20 代男性 1 人(被験者 D)の 4 人である．

3.2 実験結果

表 1 に実験結果をまとめる．

表 1 実験結果

被験者	A	B	C	D	テキスト
入力回数	119	95	101	137	94
ラベル獲得数	10	5	7	10	7
平均入力回数	8.1	11.6	9.8	10.2	8.3

表中の入力回数とは，該当する被験者が行った入力の回数，ラベル獲得数とは，実験開始から終了までの間にシステムが獲得したラベルの数，平均入力回数とは，一つの画像に対応するラベルを獲得するまでに，その画像とともに行った入力の回数を表している．例えば，被験者 A はシステムに一つのラベルを獲得させるまでに平均 8.1 回の入力を行ったということを意味する．表の右端のテキストという欄は，入力が音声で行われる場合とテキストで行われる場合を比較し，結果にどのような差が現れるかを検証するために，第一著者がテキスト入力で同様の実験をした結果である．

3.3 考察

まず，本手法の根本的な目的である「名詞獲得」の観点から実験結果を考察する．被験者ＡからＤまでの結果を見ると，個人差はあるものの，半数以上の画像に対して正しくラベルを獲得できている．このことから，本手法によりユーザの入力から画像に対応するラベルを獲得できることがわかった．

次に，音声入力の場合とテキスト入力の場合の比較を行う．音声入力の場合の平均入力回数（被験者 A~D の平均）は 9.9 回であるのに対し，テキスト入力は 8.3 回となっている．音声入力の方が平均入力回数が多くなる原因としては，入力文に音声認識誤りが含まれることがある．しかし，誤認識率が約 26.1%であることを考慮すると，この差は大きなものだとはいえない．ここで，なぜこのように多くの音声認識誤りが含まれる入力から適切にラベルを獲得することができたのかを分析する．

図 4 に示した例 1 では，ほとんどの部分が正確に認識できていない．しかし，写真に対応するラベルを示した「しかさん」や「はむすたー」などの部分は正しく認識されている．逆に，例 2 では，文全体での誤認識率は低いが，ラベルを示す「うしさん」が「うすいさん」、「わんちゃん」が「おにいちゃん」と誤認識されている．本システムでは，入力文同士を比較して字面が一致した部分を抽出することでラベル候補を生

成するため，上記の例 1 のような誤り方はラベル獲得に悪影響を与えない．つまり，誤認識率よりも，どの部分に認識誤りが存在するか，が重要になる．誤認識は文末表現や助詞などに多く現れるため，今回の実験にはそれほど大きな影響を与えなかったのだと考えられる．

一般的な対話システムでは，入力文に形態素解析や構文解析の処理を施すため，認識誤りが含まれていると後の処理にまで大きく影響を及ぼしてしまうが，本システムではそのような処理を行わないため，多少の音声認識誤りは許容できる．

＜音声認識誤りの例＞
・例1
認識結果:まるわしかさんだいを
正解文　:あれはしかさんだよ
認識結果:ほんぶごろうははむすたーまでを
正解文　:はむたろうははむすたーなんだよ

・例2
認識結果:うすいさんっていうんだよ
正解文　:うしさんっていうんだよ
認識結果:あのおにいちゃんはおおきいね
正解文　:あのわんちゃんはおおきいね

図 4 音声認識誤りの例

さらに，たとえ誤認識を含んだ文字列をラベル候補として出力したとしても，その出力に対してユーザが「じょうず」という評価を行なわない限り，システムは正しいラベルとして獲得しない．これは，幼児が大人の発話を聞き間違い，間違った言葉を話すが，それを大人が正していくうちに正しい言葉を覚える過程によく似ている．

４．まとめ

幼児の言語獲得過程を参考にして，ユーザの入力から画像に対応するラベルを獲得するシステムを構築した．また，音声認識誤りが含まれる入力からも適切にラベルを獲得できることを示した．

現在は入力される画像は予め用意されている写真であるが，今後は，カメラからリアルタイムに得た画像を対象とした，同様のシステムを構築する予定である．また，今回の実験で被験者から指摘された，音声合成が幼児の発話としては不自然であるという点なども改善していきたい．

ロボットが様々な環境において，より柔軟な対話を行うためには，言葉を辞書に記された定義のように客観的に理解するだけではなく，実際にそのロボットが置かれた状況での主観的な理解が必要であると考えている．本システムの機能を，当研究室で開発された，雑談対応の学習型音声対話システムである GA-ILSD(Spoken Dialogue Processing Method Using Inductive Learning with Genetic Algorithm)[8]に組み込むことで，主観的な理解を行いながら雑談を行うシステムを実現することも視野に入れている．

謝辞

本研究の一部は大川情報通信基金研究助成の援助を受けて行われた．

参考文献

[1] 荒木健治：自然言語処理ことはじめ―言葉を覚え会話のできるコンピュータ，森北出版，2004.

[2] 今井むつみ：ことばの学習のパラドックス，共立出版，1997.

[3] 内田ゆず・荒木健治：言語獲得システムにおける類似度に基づくラベル拡張手法の提案，平成 17 年 電気･情報関係学会北海道支部連合大会講演論文集，180，2005.

[4] 小林郁夫・古川康一・今井むつみ・尾崎知伸：帰納論理プログラムによる幼児の名詞語彙獲得のモデル化，電子情報通信学会技術研究報告 言語理解とコミュニケーション研究会(NLC)， Vol.99， No.387，pp.29-36，1999.

[5] 今井むつみ・野島久雄：人が学ぶということ―認知学習論からの視点，北樹出版，2003.

[6] 須賀哲夫・久野雅樹：ヴァーチャルインファントー言語獲得の謎を解く，北大路書房，2000.

[7] 内田ゆず・荒木健治：幼児の普通名詞および固有名詞獲得モデルに基づく帰納的学習を用いた再帰的獲得手法の提案，言語獲得と理解研究会(LAU)，Vol.1，No.1，pp.21-27，2005.

[8] 木村泰知，荒木健治，桃内佳雄，栃内香次：遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法，電子情報通信学会論文誌，D-II，Vol.J84-D-II， No.9， pp.2079-2091，2001

[9] Paul C. Quinn：Category representation in young infants，Current Directions in Psychological Science，Vol.11，No.2，pp.17-22，2002.

[10] 但馬香里：幼児における一語発話の獲得について―1 歳 10 ヶ月から 2 歳 0 ヶ月児の 3 人の幼児による観察報告，東京工芸大学工学部紀要，Vol.27，No.2，pp.59-64，2004.

[11] 須藤珠水・茂木健一郎：言語獲得期における語意学習とカテゴリー認知のメカニズム，電子情報通信学会 スマートインフォメディアシステム研究会 信学技報，SIS2004-4，pp.17-22，2004.