画像に対する発話からの名詞概念獲得システムにおける音素認識の導入について

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

対話などのエンターテイメント性を主眼に置いたロボットが次々と開発され，ロボットの活躍の場が工場から人間の生活空間に移ってきている．こうしたロボットはコミュニケーションロボットと呼ばれることが多いが，ユーザを飽きさせないほどの豊かなコミュニケーション能力を持ち合わせているものは今のところ存在しない．これは，シナリオベースの対話システム[1]を用いていることに原因のひとつがある．しかし，特に音声対話システムには音声認識誤りの問題があるため，シナリオベースの対話システムを用いなければ，ユーザ満足度の高いシステムを実現することは難しいのが現状である．

しかし，長いスパンでユーザを楽しませるためには，新しい知識を次々に学習し，予め決められていない対話を行うことができるような機能をロボットに持たせることが必要だと考えられる[2]．我々は，幼児のように知識のない状態から，ユーザによる画像に関する内容の発話（書き起こしテキスト）を手がかりに名詞概念（画像に対するラベル）を獲得していくシステムの構築を行い，性能評価を行ってきた[3]．これまでの性能評価実験では対象を書き起こしテキストに制限してきた．これは，音声入力に伴う音声認識誤りなどを排除した状態での評価を行うためである．しかし，ロボットにこのようなシステムを搭載することを考えたとき，入力方法としては音声を用いたものが最適かつ現実的であると考えられる．したがって，音声入力を用いた場合の本システムの性能評価を行う必要がある．一般的に，音声認識を行う場合には認識精度の向上のため，言語モデルを使用するが，「幼児のように何も知らない状態から」言語獲得を行うという本研究の目的と言語モデルの使用は矛盾している．そこで，本システムに音素認識を導入し，そこから名詞概念を獲得するシステムを考案した．

本稿では，システムの入力としてユーザ発話の音素認識結果（言語モデルなし）を用いた場合と，ユーザ発話を書き起こしたテキスト，音声認識結果（言語モデルあり）を用いた場合のそれぞれを比較した実験について述べ，その結果を報告する．

2. 名詞概念獲得システムの概要

本研究で用いる名詞概念獲得システムの処理の流れを図 1 に示す．また，本章では，それぞれの処理の詳細について述べる．

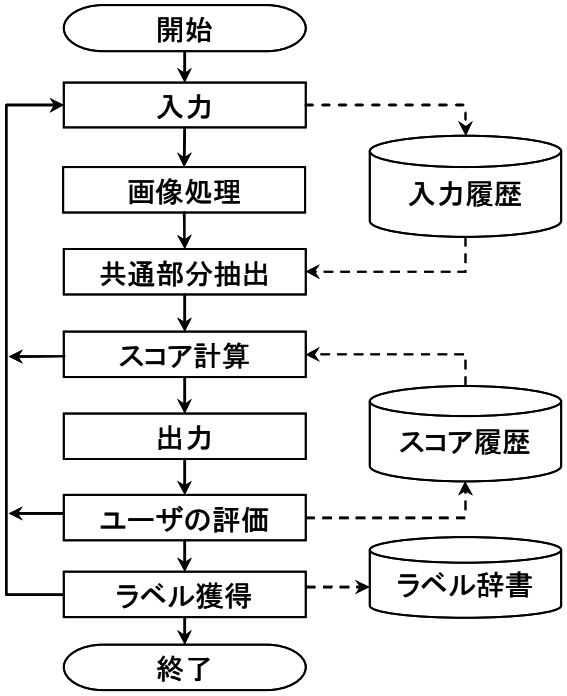

図 1　システムの処理の流れ

2.1 入力

入力は画像と文の対である．入力画像は Web カメラ（USB-CAMCHAT2／アイ・オー・データ機器．有効画素数：30 万画素）からキャプチャされた画像（以降画像 P と呼ぶ），入力文は画像 P を見せながらユーザが幼児に話かける発話 1 文（以降

文 S と呼ぶ）である．入力画像は，ユーザが自由に被写体を選び撮影するものである．また，画像のサイズは 320×240 ピクセルで，なるべく被写体全体が画像内に収まるように撮影する．

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，ユーザによって表記に揺れが生じることと，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．

2.2 画像処理

過去に同じ被写体が写った画像が入力されたかどうかを判断する．ここでは，エボリューション・ロボティクス社の”ERSP3.1（Evolution Robotics Software Platform）”[4]に含まれる”ERSP ビジョン”を用いた．”ERSP 3.1” は，ロボット製品の作成を目的とした総合開発プラットフォームで，”ERSP ビジョン”は照明や物体の位置が管理されていない現実的な環境の中でも，ロボットや装置が２次元と３次元の物体を認識することができる画像認識ツールである．

2.3 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．

2.4 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．基本スコアとは，その共通部分のラベルとしての確からしさを表した値であり，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ほど高いスコアを与えられる．基本スコアの計算式は式(1)のようになる．

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

(1)式で，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はたらく係数，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．

2.5 出力

2.4 で述べた方法で求めた基本スコアが閾値 8.0 を超えた共通部分は，画像 P に含まれる事物のラベルに適している可能性が高いと判断され，テキストで出力される．

2.6 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，入力する．

- 「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

幼児がこれらのキーワードを完全に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情報の代わりにキーワードを用いることとする．

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．具体的には，基本スコアに係数 β を乗ずる．係数 β は，予備実験から，ユーザの評価が「じょうず」の場合は 1.5，「おしい」の場合は 0.8，「ちがうよ」の場合は 0.2 とした．

2.7 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 30.0 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．

3. 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 2 に示す．なお，Microsoft Visual Studio .NET 2005 を用いてこのインタフェースの構築を行った．

図 2 ユーザインタフェース

ウィンドウ中の２枚の画像のうち，左は Web カメラからのプレビュー，右はキャプチャ済みの画像となっている．この場合，ユーザは赤ちゃんに消し

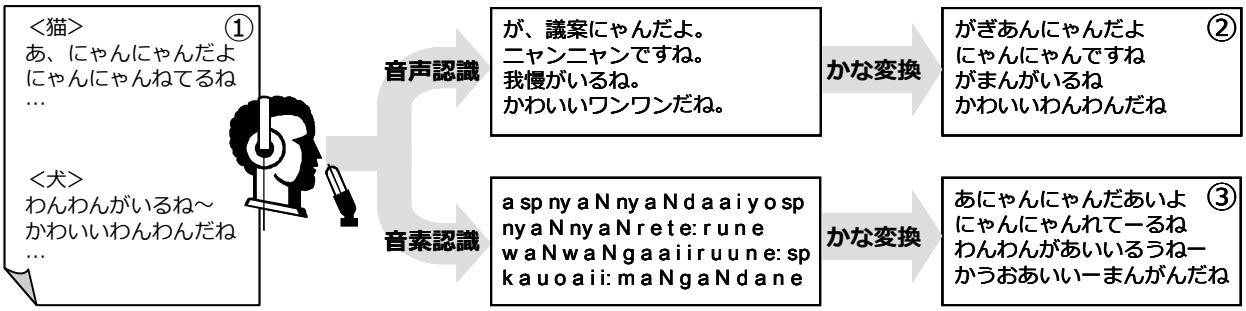

図 3　入力文の収集方法

ゴムを見せながら話しかけているという場面を表している．また，右下の赤ちゃんに付随している吹き出しの内容は **2.5** で述べた出力である．赤ちゃんの画像の下には，「（今キャプチャされている画像は）初めて見る画像である」「前に見たことがある」などの画像処理の結果や，「何か言っています」などのシステムの状態が表示される．

4. 音素認識器

本研究では，音素認識を行うために”Julius 連続単語認識キット”を利用している[5]．これは，言語モデルを使わずに，登録単語の任意の組合わせとして音声認識するための Julius の支援キットである．具体的には，与えられた単語とその読みのリストから，認識辞書や各単語の任意の組合せを許すような擬似 N-gram を生成し，連続単語認識を可能にしている．これを使って単語の代わりに各音素を辞書に登録することで音素認識器を実現している．使用した Julius のバージョンは Julius-3.5.2 である．

5. 実験と考察

5.1 実験方法

今回の実験では，同じ内容の文を，①書き起こしテキスト②音声認識結果(言語モデルあり)③音素認識結果の 3 種類の方法で，作成した名詞概念獲得システムに入力する．そして，音素認識結果を用いた場合，他の 2 つの方法を用いた場合と比較して名詞概念獲得にどのような影響があるかを調査する．

入力文は，3 種類の事物（ボール，りんご，車）に関して「まだ話すことのできない赤ちゃんにその画像に含まれる物を見せながら話しかける」ことを想定した内容である．これらの入力文はアンケートによって 31 名の回答者から事前に収集されたものである．回答者 31 名の内訳は，男性が 7 名（20 代～40 代），女性が 24 名（10 代～50 代）であった．そのうち，子育ての経験がある人は 13 名であった．このようにして収集した文の中から，事物のラベルが含まれる文を 3 類の事物それぞれに対して 20 文ずつ抽出して実験に用いた．

書き起こしテキストは収集した文を全てひらがなに変換したものである．音声認識と音素認識は被験者の発話から同時に行い，それぞれの結果をひらがなに変換したものを用いた．入力文の収集の概念図を図 3 に示す．

5.2 実験結果

ユーザの発話を音声認識・音素認識した結果の例を図 4 に示す．これらの例文は全て実際に入力された文である．なお，ここでは認識結果を出力の通りに記述している．

例文 1：あかいぼーるだ
<音声認識> 赤いボールだ.
<音素認識>あかいぼーろだー

例文 2：ぼーるだよー
<音声認識>ゴール内容.
<音素認識>ぽーおるなえおーおーおうふ

例文 3：まっかなりんごたべたいかな
<音声認識>真っ赤なりんごを食べたいかな？
<音素認識> まかなりんごためたいかなう

例文 4：ちっちゃいぶーぶーだねー
<音声認識>ちっちゃい部分だね.
<音素認識>くひいくはいぶーぶーだんね

例文 5：ぶーぶーまたのろうね
<音声認識>ブームまだなのね.
<音素認識>ぶうぶーまたんあおろおね

図 4　音声認識・音素認識結果の例

次に，3 種類の事物について 3 つの入力方法を用いて名詞概念を獲得させる実験の結果を表 1 にまとめる．表中の”初出力”は，システムが初めて出力をするまでにユーザは何度入力を行ったのか，”初正解出力”は，システムが初めて正しい出力をするまでにユーザは何度入力を行ったのか，”ラベル獲得”は，システムが正しいラベルを獲得するまでにユー

表 1　名詞概念獲得実験結果

		テキスト	音声認識	音素認識
ぼーる	初出力	4	4	10
	初正解出力	4	8	13
	ラベル獲得	6	10	15
	入力文のラベル含有率	20/20(100%)	15/20(75%)	9/20(45%)
りんご	初出力	4	4	4
	初正解出力	4	6	5
	ラベル獲得	6	7	8
	入力文のラベル含有率	20/20(100%)	17/20(85%)	13/20(65%)
くるま	初出力	4	7	4
	初正解出力	7	-	8
	ラベル獲得	8	-	11
	入力文のラベル含有率	20/20(100%)	0/20(0%)	7/20(35%)

ザは何度入力をしたのか，”入力文のラベル含有率”は，用意した 20 文の入力文のうち，正しいラベルが含まれている文が何文あるのかを表している．また，「くるま」の「音声認識」の項目において，“-（ハイフン）”が記入されている部分は，正しい出力が一度も行われなかったため，データが得られなかったことを意味する．

5.3 考察

まず，音声認識と音素認識の結果について考察する．両者の最大の違いは言語モデルを用いているか否かであるが，音素認識には言語モデルを用いていないため，認識結果に文法的な誤りが多く見られる．

一方，音声認識では多少認識誤りがあるものの，文法的には概ね正しく，日本語として成り立った文が得られた．しかし，「くるま」のラベルとして用いられる「ぶーぶー」という単語は 20 文全てにおいて正しく認識されず，「部分」や「ブーム」と誤認識された．これは，辞書に幼児言葉である「ぶーぶ」が登録されていなかったことが原因であると考えられる．

次に名詞概念獲得実験の結果について考察する．入力に音声認識・音素認識結果を用いた場合には，テキストを用いた場合よりもラベルを獲得するまでに多くの入力が必要である．これは，音声認識誤りが含まれるため，学習が効果的に進まなかったことを表している．

ここで最も注目すべき結果は「くるま」の項目である．音素認識を用いた場合は，他の 2 つの事物よりも認識結果は低いものの，ラベル獲得には成功している．しかし，音声認識では正しいラベルを獲得することができなかった．これは，音声認識結果に「くるま」を表す「ぶーぶ」という文字列が一度も含まれなかったためである．上述したように，今回用いた音声認識器に含まれる単語辞書には「ぶーぶ」などの幼児語が登録されておらず，本来は認識結果を改善するためにはたらく言語モデルが弊害をもたらしたことが確認された．また，音素認識は単語を予め与えていないため，実際に幼児が言語獲得を行う状況により近い．したがって，音素認識結果からの名詞概念獲得は効果的である．

6. まとめ

名詞概念獲得システムへの入力を，テキスト，音声認識結果，音素認識結果とした場合の比較実験を行った．実験の結果，本システムは音素認識結果による入力が与えられた場合にも有効であることが示された．

一般的に，音声認識よりも単語単位の言語モデルを使用しない音素認識の方が認識精度は低くなるが，その言語モデルが認識の柔軟性を奪う危険性もあることがわかった．したがって，家庭内で利用される対話システムには固有名詞や新語，造語に対応する能力が必要であるため，単語単位の言語モデルを使わない音声認識を取り入れる必要があると考えられる．

今後は，実際に大人が幼児に話しかけている発話文を収集し，それらを入力した場合にも本システムでの名詞概念獲得が可能であるかを評価する予定である．

参考文献

[1] 松井 俊浩 , 麻生 英樹 , John Fry , 浅野 太 , 本村 陽一 , 原 功 , 栗多 喜夫 , 速水 悟 , 山崎 信行 : オフィスロボット Jijo-2 の音声対話システム , 日本ロボット学会誌 , Vol.18 , No.2, pp. 42-149 , 2000.

[2] 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.

[3] 内田 ゆず , 荒木 健治 : 画像に対する発話からの名詞概念の獲得 , 情報処理学会研究報告 , 2006-NL-176 , pp.81-86 , 2006.

[4] エボリューション・ロボティクス社 : http://www.evolution.com/products/ersp/

[5] 大語彙連続音声認識エンジン : http://julius.sourceforge.jp/