# 最終行動ヒューリスティクスを用いた状況推定による自由発話音声データからの語句意味学習

# The meaning acquisition of phrases from natural speech data by situation estimation based on Final Action Heuristics

小野 広司* 左 祥 伊丹 英樹 尾関 基行 岡 夏樹
Kouji ONO Zuo XIANG Hideki ITAMI Motoyuki OZEKI Natsuki OKA

京都工芸繊維大学 大学院工芸科学研究科
Graduate School of Science and Technology, Kyoto Institute of Technology

**Abstract:** In human-agent interaction, it is very important for agents to understand the functional meaning of utterances from humans, because the agents should act according to the functional meaning of utterances such as instructions to do something and questions about something. We have studied the acquisition of the functional meaning of phrases by an agent which interacts with a human in a maze on a computer screen. In our previous studies, we chose setting in which the agent had knowledge on what to do in a position in the maze. We aim to realize learning without such prior knowledge in this study. The agent estimates appropriate actions by Final Action Heuristics which states that an action finally taken in a state is likely to correct. An experiment demonstrates that the precision and the recall of the learning result by the proposed method compare favorably with those by a method which uses prior knowledge.

## 1 はじめに

言葉の意味の獲得は，心理学，言語学，哲学，認知科学などの分野で古くから関心を持ち続けられてきたテーマであるとともに，工学的にも高い関心が寄せられている．その背景として，高度で複雑な情報機器やソフトウェアの自律的なユーザ適応や環境適応への要求がある．我々は，そのような適応技術の一つとして，ユーザの言葉の意味獲得の研究に取り組んでいる．情報機器やソフトウェアが人の言葉の意味をゼロから獲得することができれば，人間の赤ちゃんに教えるようにして機械に必要な機能を与えることができる．

我々が取り組んできた研究 [左 09] の特徴は, 言葉の“機能的意味” を扱う点にある．言葉の意味獲得の研究には多くの蓄積があるが, そのほとんどは参照的な意味（世界の中の物, 属性, できごと, 関係, 動きなどを指し示す働き）を扱ってきた．しかし, 言葉は参照的な意味だけでなく機能的な意味（聞き手に影響を与える働き）も持っており, この両方を理解できることが重要である [Roy 05]．本研究では “状況” という概念上のオブジェクトを導入することで, 参照的意味と同様の共起に基づいた方法で機能的意味の獲得を実現する手法を提案する．また本研究では, 人間と触れ合う中で自然に得られる情報をもとにして, 独力で言葉の意味を獲得できるシステムを目標としている．そのため我々の研究では自然な連続発話によるインタラクションで得られた音声データを入力として意味学習を行う．この目標を達成するためには, 以下の処理を自動的に行うアルゴリズムが必要となる．

1. 連続音声の中から意味を担う可能性があるフレーズを切り出して語候補とする
2. 切り出した語候補といずれの状況が共起するかを判断して対応付ける
3. 多数ある対応付けられた語候補と状況の組の中で統計的にどの組が信頼できる（信頼できない）か順位付ける

[左 09] の研究では, 以上の処理のうち, 1. と 3. に関しては自動化しているが, 2. に関しては, “状況” は既知であるとして人手で正解を与えていた．

そこで本稿では, 最終行動ヒューリスティクスにより “状況” を自動的に推定する手法を提案する．また人手による正しい “状況” を対応付けた場合と, 本手法によって得た “状況” を対応付けた場合の学習性能を比較し, その影響を調べる．

*連絡先：京都工芸繊維大学大学院工芸科学研究科
〒 606-8585 京都市左京区松ヶ崎橋上町
E-mail:m7622017@edu.kit.ac.jp

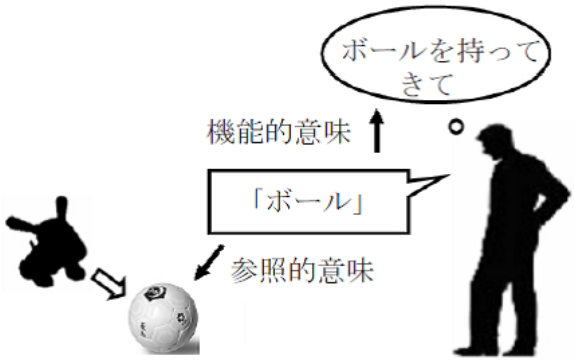


図 1: 言葉の参照的意味と機能的意味

以下, 本稿では, まず 2. で言葉の機能的意味について説明し, これを参照的意味と同様の手法で獲得する方法を説明する. 続いて 3. 問題設定と前述の手順 1 と 3 について概説する. その後 4. では最終行動ヒューリスティクスについて説明し, 5. で実験とその考察を述べ, 6. で結論と今後の展望を述べる.

## 2 言葉の機能的意味

人がロボットと対峙し, ロボットの前にボールが置かれている場面を考えよう (図 1). そこで人がロボットに対して「ボール」と発話したとする. 「ボール」はロボットの前に置かれたボールを指し示している. これが参照的意味である. しかし, この場合, 人がロボットに伝えたいのは, 《ボールを持ってきて欲しい》や《ボールがあるので注意しろ》といったことであろう. このように, 聞き手の行動や心的状態に影響を与える働きとして言葉を解釈したものを機能的意味と呼ぶ.

語意獲得において, 参照的意味に比べて機能的意味を扱うことの難しさは, 言葉と対応づけるべき事象が実空間内に (直接センシングできる形では) 存在しないことである. 参照的意味を扱った Roy らの研究 [Roy 99] では, 言葉と共起する実空間中の事象をその言葉の意味として獲得している. しかし, 機能的意味の獲得では, そのままでは参照的意味と同じ方法は使えない.

そこで本研究では, 以下の考えに基づいて問題を再設計することにより, 参照的意味の獲得と同じ方法論で機能的意味を扱うことを考える.

- 機能的意味を伴った言葉の発話は, その発話の直前のエージェントの行動に対する評価, もしくは, 直後のエージェントの行動に対する指示であると仮定する. これは問題を簡単化するためであり, 本来, 機能的意味の働きかけの対象となる行動は, その発話の直前や直後にあるとは限らないし, 聞き手の行動とも限らない.
- エージェントが行動するタイミングで時間を区切る. 行動 A と行動 B の間の区間は,《行動 A が良かった, または, 悪かったと評価する状況》であり, 且つ,《次に○○の行動をするべき状況》でもあるとする (図 2)

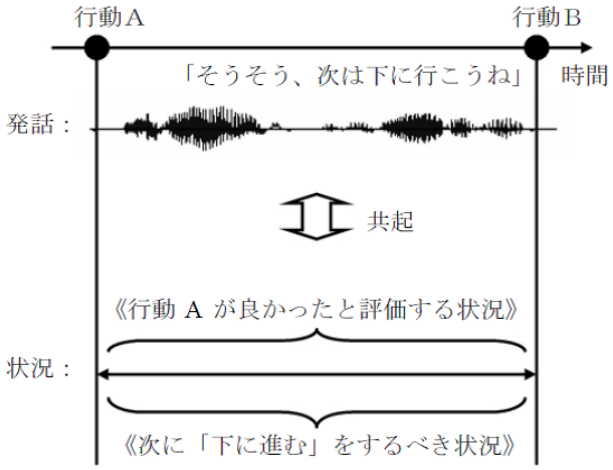


図 2: “状況” の定義

このように考えることで, 参照的意味における実空間中の事象と同じように, “状況” を「発話と共起する概念的な事象」として扱うことができる. 例えば,「みぎ」という言葉が《次に右に動くべき状況》と特異的に共起するのであれば,「みぎ」の機能的意味は「次に右に動いて欲しいという働きかけ」であると推定することができる.

## 3 問題設定と意味獲得アルゴリズム

### 3.1 問題設定

迷路タスクで得られるインタラクションデータを用いて意味学習を行う. 迷路タスクでは, 計算機の画面上の迷路 (図 3) を用いて, 赤ちゃんエージェントをゴール (ミルクのある場所) まで音声で誘導する. エージェントは, 上／下／左／右のいずれかの方向へ 2 秒毎に移動する. 実験参加者は防音室に設置された画面の前でヘッドセットを着用し, 自由に発話してエージェントに教示する.

インタラクションデータとして, 迷路の状態 (各時刻の迷路内におけるエージェントの位置情報) と発話データ (音声) を記録する. 収集した発話には教示 (行動の指示と評価) 以外の内容も含まれており, 教示の際

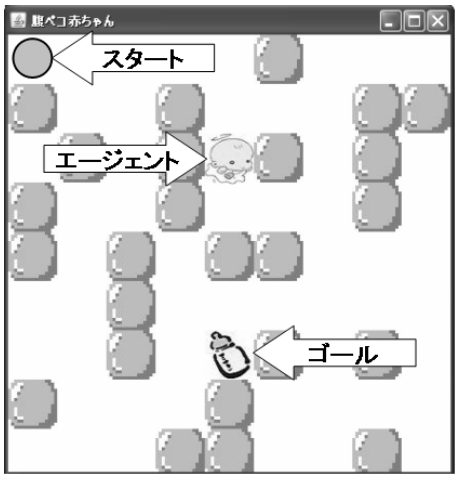


図 3: 迷路タスクで用いた迷路

にも様々な表現が用いられた．例えば,《次に右に動いて欲しい》の意味で「右行け」,「右だよ」,「前へ進め」などの表現が用いられた[1]．また誤った教示が与えられることもあった．

このようなインタラクションデータをもとに，エージェントに以下の 2 種類（6 タイプ）の機能的意味を獲得させる．

**行動教示** 行動教示： エージェントに次に動いて欲しい方向を示す行動教示 4 タイプ :《上に進め／下に進め／左に進め／右に進め》（以下，略号《↑／↓／←／→》を用いる）．

**評価教示** エージェントの直前の行動に対する評価を示す評価教示 2 タイプ :《適切だった／不適切だった》（以下，略号《○／×》を用いる）．

なお，ここに挙げていない意味を持つ発話がデータに含まれることは想定しており，それらの発話には意味が付与されないようにしたい．また，同じ意味を持つ発話は複数あることも想定しており，それらの発話グループには同じ意味が付与されるようにしたい．

ここで発話との共起関係を得るための概念である “状況” を定義する．本研究では，先述の機能的意味と状況は 1 対 1 で対応しているという立場をとるため，状況の種類は獲得させたい機能的意味と同じものになる．同様の理由により，状況の表記も意味と同じ表記《 》及び略号を用いる．

**次の行動に対して働きかける状況** 《次に上／下／左／右へ進むべき状況》（以下，略号《↑／↓／←／→》を用いる）．

[1]本論文では，発話を「 」で囲み，発話の意味を《 》で囲んで示す．

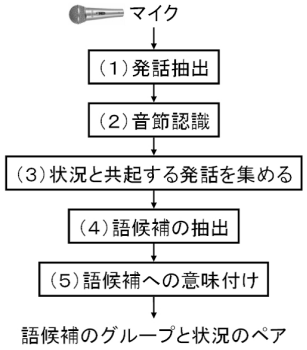


図 4: アルゴリズムの流れ

**直前の行動に対して働きかける状況** 《直前の行動を適切だった／不適切だったと評価する状況》（以下，略号《○／×》を用いる）．

発話データから後述する処理を経て抽出された言葉は，これらの状況との共起に基づいて意味づけされる．つまり，意味づけ対象となる言葉は，上記 6 タイプの状況かそれ以外か（合計 7 タイプ）のいずれかに分類されることになる．

## 3.2 意味獲得アルゴリズムの概要

意味獲得の処理の流れを図 4 に示す．処理は (1)〜(5) の 5 ステップに分けられる．なおアルゴリズムの詳細に関しては [左 09] を参照されたい．

### (1) 発話抽出

まず，連続な音声信号から発話を抽出する．抽出処理は，エージェントの行動タイミングから 300ms 後，もしくは 300ms 以上の無音区間で音声信号を区切り，それを 1 発話とする．行動タイミングとは，エージェントが上／下／左／右のいずれかに動くタイミングであり，本実験では 2 秒間隔に設定している．行動タイミングから更に 300ms 遅らせて区切るのは，エージェントの動きに対して人の反応が遅れることを考慮したものである．

### (2) 音節認識

Julius-3.4 を単音節認識器として用い，(1) の処理によって抽出された発話を音節列へと変換する．本実験では，音響モデルとして Julius ディクテーションキットに付属の不特定話者 PTM トライフォンモデルを，言語モデルとして全音節の出現確率を等確率とした単音節 Uni-gram を使用した．

**(3) 状況と共起する音節列の収集**

次に，3.1 で定義した各状況と共起する音節列を集める．同じ状況に共起する音節列は群にまとめられ，群毎に (4) の処理に渡される．なお, [左 09] の研究では迷路内の各状態における状況は既知であるという前提で行ってきたが, 本稿では後述する最終行動ヒューリスティクスでこれを推定する．

**(4) 語候補の抽出**

同じ状況で発話された音節列群の中で，繰り返し出現する類似区間を検出し，それを意味付けの語候補とする．異なる状況で発話された音節列の間では類似区間の検出は行わない．本研究では，各音節列の間の類似度を行列形式で表現した Confusion Matrix の要素値を元に，二つの音節列間の類似度を Smith-Waterman のアルゴリズム [Smith 81] を用いて計算する．

**(5) 語候補への意味付け**

(4) の処理で抽出された語候補を元にして, 類似した語候補を一つのグループにまとめ, そのグループに対する意味を付与する．その際 Fisher の生起確率和を基準として [岡 08], 語候補をグループ分けし, ある状況と特異に共起する語候補グループに絞って意味づけを行う．

この処理が必要になるのは, ある状況 A において語候補 B が発話されたからといって，それだけで語候補 B の意味を状況 A であると決定することはできないためである[2]．

# 4 状況の推定方法

本研究では 3.2 の処理において, 発話と対応させるべき “状況” を最終行動ヒューリスティクス（Final Action Heuristics, 以降, FAH と略記する）によって推定する．

## 4.1 最終行動ヒューリスティクス (FAH)

FAH とは「試行錯誤を繰り返して最終的に目標を達成した試行において, ある特定の状態に対する行動は複数回試みられた可能性があるが, その中の各状態で最終的にとった行動は正しかった可能性が高い．また, ある状態で最終的にとった行動とは異なる行動は, 誤っていた可能性が高い」というヒューリスティクスである．

図 5 は, エージェントが試行錯誤を繰り返しながら Start から Goal へ向かっている様子を示している．細く折れ曲がった矢印は通った経路を示している．太い矢印はそのマス目（状態）において最後に取られた行動（最終行動）を示しているが, これは細い矢印が Goal に到達した時点で確定し, それまでは確定しない．すなわち Goal に到達した時点で初めて, それまでに立ち寄ったマス目（状態）における, 取るべきであった行動が推測できる．

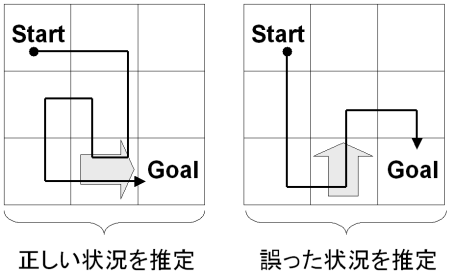


図 5: 最終行動ヒューリスティクスによる状況推定

図 5 の右図を見ればわかるように, FAH では必ずしも正しい推測が行えるわけではない．一般に目標達成に至るまでの道筋が複数あり, その数が多ければ多いほど, FAH による推測が正しい可能性は低くなる．逆に目標達成に至るまでの道筋が一通りならば必ず正しい推測を行うことができる．

## 4.2 FAH を用いた状況推定

本研究において FAH を用いて状況を推定する方法を以下に述べる．

**行動状況の推定方法** スタートからゴールへ向かっている間，エージェントが各時刻に迷路内のどのマス目にいて，どういう行動を取ったのかを記憶しておく．ゴールに到達した時点で，これらの情報からゴールに到着するまでにエージェントが通過した各マス目における最終行動を調べる．各マス目における最終行動を，エージェントがそのマス目にいる時の行動状況とし，エージェントがその場所にいた時の発話と対応付ける．

**評価状況の推定方法** 予め行動状況の推定を行っておく．時刻 $t$ においてエージェントが取った行動を $a(t)$, エージェントがいたマス目を $p(t)$，マス目 p における行動状況を $S_A(p)$ とすると，時刻 $t$ における評価状況 $S_E(t)$ は $a(t-1) = S_A(p(t-1))$ の時に《○》，それ以外の場合に《×》と推定する．

[2][左 09] で収集したデータでは，状況と対応しない語候補が約 7 割含まれていた．

表 1: 各実験協力者の発話数

| | 全発話数 | 正しく状況推定できた発話数 |
|---|---|---|
| 実験協力者 A | 178 | 106 |
| 実験協力者 B | 112 | 94 |

表 2: 精度と再現率の計算

| 発話 | 正解意味 | 検出意味 |
|---|---|---|
| みぎーみぎよし | →, OK | →, →, NG |
| そのままひだりそう | ←, OK | ←, OK |
| だめみぎ | NG, → | ↑, NG |

## 5 評価実験

FAH による状況推定を用いた場合の学習性能と人手で付与した状況を用いた場合の学習性能を比較する．実験参加者として大学院生（男性）2 人に協力してもらい, 学習，および評価は実験参加者毎に行うものとする．なお各実験協力者の発話数と，そのうち FAH によって正しい状況を推定できた発話の数は表 1 の通りである．

評価は学習した語によって未知の発話の意味をどれだけ正確に推定できるかによって評価を行う．この評価を定量的に行うために再現率と精度を算出する．例えば, 表 2 のようなテストデータがあったとする．「正解意味」とは発話の書き起こし文から人手で判断して付与した意味である．「検出意味」とは学習結果を用いて発話から検出された意味である．

このとき, 再現率と精度を以下の式で求める．

$$\text{再現率} = \frac{RN}{RD} \tag{1}$$

$$\text{精度} = \frac{PN}{PD} \tag{2}$$

ここで $RD$ は全発話中の正解意味数, $RN$ は全発話中の正解意味のうち検出されたものの数, $PD$ は全発話中の検出意味数, $PN$ は全発話中の検出意味のうち正解のものの数を示す．表 2 の例の場合, $RD$ は 6, $RN$ は 4, $PD$ は 7, $PN$ は 5 となる．再現率, 精度という尺度を使用して, 収集したインタラクションデータを用い, 10-fold Cross Validation 法によって評価を行う．

実験参加者 B における意味学習結果の例を表 3 に示す．この学習結果は FAH を用いて推定した “状況” を用いた学習結果であり, Pt が小さい順に 1〜7 番目の語を示している．なお, Pt は Fisher の直接確率計算法によって算出される生起確率和であり, この値が低いほど, 語と状況が特異的に共起している，すなわち，その語の意味がその状況である可能性が高いことを示す．なお，書き起こしにおける括弧の意味は次の通りである．空白の括弧は，認識・分節の際にその箇所に余分な音節が追加されたことを示す．括弧内に文字がある場合は，認識・分節の際に括弧内の文字が脱落したことを示す．

表 3: 実験協力者 B の意味学習結果の例

| 音節認識結果 | 書き起こし | 状況 | Pt |
|---|---|---|---|
| たひ | (し) たし（た） | ↓ | $2.6 \times 10^{-13}$ |
| ほちこ | こっち（） | NG | $6.1 \times 10^{-10}$ |
| にいににいはみ | みぎみぎ | → | $1.8 \times 10^{-9}$ |
| ににりみいいりふ | みぎにみぎみぎ | NG | $5.9 \times 10^{-9}$ |
| ぐえ | うえ | NG | $9.6 \times 10^{-8}$ |
| おみぎー | （）みぎ | → | $1.3 \times 10^{-5}$ |
| ひてそ | きて（） | OK | $3.8 \times 10^{-5}$ |

生起確率和 $P_t$ の閾値を何らかの値に設定した時，閾値よりも低い $P_t$ を持つ語を有効な語とし, テストデータの発話の中からそれと類似する語を検出する．例えば上記学習結果例で閾値を $1.0 \times 10^{-12}$ に決めた場合, 有効な語は「たひ」のみになる．さらにテストデータの発話「したした, したいって」という発話から「たひ」と類似した音節列が検出された場合, その発話には《↓》の意味が付与される．

以上のような考え方を用いて，10-fold Cross Validation 法によって得られる 10 個のデータセットを用いて評価を行う．各データセットではトレーニングデータから語の意味学習を行い，その学習結果を用いてテストデータの発話に意味を付与する．閾値は $1.0 \times 10^{0}$〜$1.0 \times 10^{-49}$ まで 50 通りに設定し, 各閾値毎に（1）（2）式で計算に用いる $RD$, $RN$, $PD$, $PN$ を数える．すなわちある閾値 $t$ に対して, データセット $k$ 番目における上記各数を $RD(k)$, $RN(k)$, $PD(k)$, $PN(k)$ とすると, このデータの再現率 ($Recall(t)$) と精度 ($Precision(t)$) は以下のように計算される．なお 10-fold Cross Validation 法を用いるので k の値は $k = 1, 2, \cdot\cdot, 10$ となる．

$$Recall(t) = \frac{\sum_{k=1}^{10} RN(k)}{\sum_{k=1}^{10} RD(k)} \tag{3}$$

$$Precision(t) = \frac{\sum_{k=1}^{10} PN(k)}{\sum_{k=1}^{10} PD(k)} \tag{4}$$

ただし再現率の分母が 0，又は精度の分母が 0 になる場合はこれらの値を計算せず，無効な評価結果とする．

実験参加者 A と B について, 発話データからこのようにして算出した精度と再現率の関係を図 6 と図 7 に示す．グラフ中の各点は，各閾値における精度と再現率を表しており，閾値が隣り合う二つの点を線でつないでいる．また各系列の上端が最も高い閾値による評価

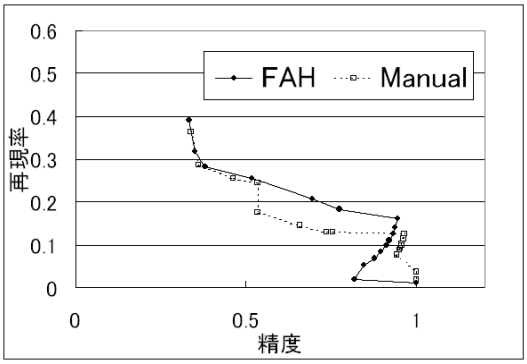


図 6: 実験参加者 A のインタラクションデータからの学習結果

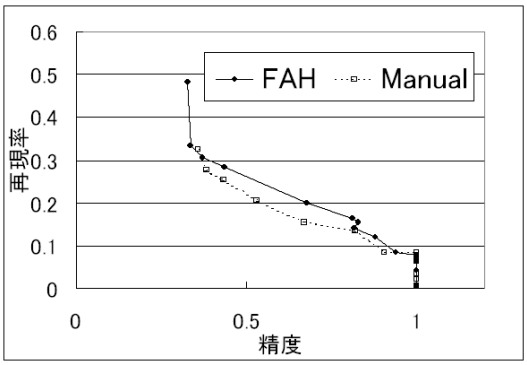


図 7: 実験参加者 B の同上結果

結果，下端が最も低い閾値による評価結果である．（閾値が低いほど，語と状況の共起性が高いものでなければ意味付けされない点に注意されたい．）凡例が FAH となっているものは FAH によって推定した状況を用いて学習を行った結果を示し，凡例が Manual となっているものは人手による正しい状況を用いて学習を行った結果を示す．

図 6，図 7 の結果に対する考察を行う．提案手法では状況を推定するため，人手による正しい状況を用いる場合に比べて状況と発話の対応付けの精度が落ち，最終的な語の意味学習性能も低下することが予想されたが，これらの結果は，今回の実験設定において提案手法による意味学習性能が人手による正しい状況を用いた場合の意味学習性能に劣らないことを示唆している．

しかし実験参加者 A の結果に注目すると，FAH の評価結果グラフが折れ曲がっているのがわかる．これは閾値を低くして厳しい条件で意味付けを行ったにも関わらず精度が低下したことを意味し，3.2 で述べた意味付けがうまく行われなかったことを示す．この原因は，実験参加者 A のインタラクションデータに対しては FAH による状況の推定精度が低かったからだという可能性がある（表 1）．

## 6 結論

本稿では最終行動ヒューリスティクスを用いて推定した状況を用いて語句の意味学習を行う手法を提案した．実験協力者 2 人について提案手法と人手による正しい状況を用いた学習とを比較し，提案手法が意味学習に利用できる可能性のある手法であることを示した．

今後の課題を述べる．現段階では 2 人分のデータによる解析を行ったのみであるので，提案手法がどのような場合にうまく働き，どのような場合にうまく働かないのかなどはまだ確かめられていない．

4.1 で述べたように FAH の性能は，迷路の自由度によっても左右される．

そこで今後はより多くの実験協力者のデータを用いて評価を行い，また，異なった迷路環境における実験も行い，意味学習における FAH の性能を明らかにしていきたい．

## 謝辞

本研究は科研費 (17500093 および 21500137) の助成を受けたものである．

## 参考文献

[Roy 99] Roy, D.: Learning from Sights and Sounds:A Computational Model, *Ph.D.Thesis, MIT Media Laboratory* (1999)

[Roy 05] Roy, D.: Semiotic schemas: a framework for grounding language in the action and perception, *Artificial Intelligence*, Vol. 167, No. 1–2, pp. 170–205 (2005)

[Smith 81] Smith, T. and Waterman, M.: Identification of common molecular subsequences, *J. Mol. Biol* (1981)

[岡 08] 岡 夏樹, 増子 雄哉, 林口 円, 伊丹 英樹, 川上 茂雄：Fisher の直接法を用いたインタラクションデータからの意味学習, 知能と情報 (日本知能情報ファジィ学会誌), Vol. 20, No. 4, pp. 461–472 (2008)

[左 09] 左 祥, 北川 憲, 林口 円, 小野 広司, 荒木 雅弘, 岡 夏樹：時間的に切迫した状況におけるインタラクションデータからの意味学習, 人工知能学会全国大会 (2009)