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

Meaning Acquisition from Interaction Data under Pressure

左祥 ZUO Xiang　北川 憲 KITAGAWA Ken　林口 円 HAYASIGUCHI Madoka　小野 広司 ONO Kouji　荒木 雅弘 ARAKI Masahiro　岡 夏樹 OKA Natsuki

京都工芸繊維大学
Kyoto Institute of Technology

We discuss an algorithm that learns the meaning of action and evaluation instructions in a navigation task. In the navigation, a human navigator freely gives an agent instructions via a microphone. From these speech data several candidates of meaningful phrases are automatically extracted and each of them is associated with the meaning of the instruction based on their co-occurrence, recurrence, and statistical significance. For verifying our method, we conduct an experiment in a labyrinth task and demonstrate its potential.

1. まえがき

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，近年高い関心が寄せられている．これまで，言葉の意味の獲得の研究には，多くの蓄積がある（例えば [3] と [6]）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，動き等を指し示す働き）を扱ってきた．Roy[3] は，お母さんと赤ちゃんがいくつかのおもちゃで順次遊ぶ場面において，赤ちゃんに向けた発話と，おもちゃを撮影した画像を結びつけることでおもちゃを指す言葉を獲得するモデルを提案した．Yu と Ballard[6] は，ユーザに自分がやっている作業を説明しながら作業をさせるというタスクを用い，腕の位置や目の動きといった情報から，ユーザの行動の特定，注目している物体の抽出を行い，これらを音声と結びつけることで意味学習を行った．

言葉の参照的な意味でなく，機能的な意味（聞き手に影響を与える働き）も持っており，この両方の意味を理解できることが重要であるが，機能的な意味の獲得の研究は限られている．機能的な意味の獲得に関するこれまでの研究として，鈴木ら [5] は，行動の事後指示と行動の評価という異なる社会的働きを持つ言葉が混在する中での意味学習を目指した．鈴木らは，迷路を移動するロボット（ただし，実機ではなくシミュレータを使用）を教示により誘導するタスクを設定し，ある状況でどのような行動が適切であるかの知識をロボットが利用できることを前提として，ロボットによる教示意味の学習モデルを提案した．鈴木らの研究では，学習が難しくなることを避けるために，行動の事後指示と行動の評価は計算機が与えることとした．これに対して，岡ら [2] は，計算機の画面上の迷路を移動するエージェントを実際に人が声で誘導するタスクを用いた．人が誘導する場合は，行動の事後指示は行われず，発話は主に‘次の行動の指示’，または，‘直前の行動の評価’ の 2 種類であることが実験から分かったため，この 2 種類の異なる働きをする言葉の意味を獲得することを目指した．ただし，岡らは，人による発話の音声信号そのものは扱わず，発話を書き起こした文字列を入力として用い，また，音声中の繰り返しの検出も人手で行った．これに対して，本研究では，岡らの研究において人手で行っていた部分を自動化し，意味学習の全体を自動化することを目標とする．また，本研究では，時間的に切迫した状況で，人が迷路中のエージェントを誘導するタスクを用いるが，このタスク設定は本研究の特徴である．時間的に切迫した状況で自然な教示を受け入れ可能とすることは，例えば車の運転中のような，余分な認知負荷を人に与えてはいけない状況や，緊急事態が発生するかもしれない状況における人と機械のインタラクションを実用レベルに高めるためには，避けて通れない技術課題であると考えている．

連絡先: 左祥
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御所海道町
TEL,FAX:075-724-7477
e-mail:d8821502@edu.kit.ac.jp

2. 発話データの収集実験

本研究では，計算機の画面上の迷路を用いて，人が赤ちゃん型のキャラクターを声でミルクの所（ゴール）まで誘導するタスクを設定し，インタラクションデータを収集した．本論文では，このキャラクターをエージェントと呼ぶ．

実験参加者は 20 代前半の工学系の大学生／大学院生で男性 2 名であった．発話データの収集は防音室の中で行い，指向性ヘッドセットを用いて行った．スタート地点を出発してからゴールに到達するまでを 1 試行とし，合計 4 試行を行った．エージェントは全ての地点における最適行動（ゴールまでの最短の道筋）を知っているが，各試行での最適行動の確率を順に 0.2，0.4，0.6，0.8 と変化させた．それ以外の確率では，ランダムな行動（最適行動を含む）を取る．行動確率を変化させた理由は，エージェントが学習しているかのような印象を実験参加者に与えるためである．ここでの行動とは，上，下，左，右いずれかの移動が可能な方向への 2 秒ごとの移動である．各実験参加者が各試行に要した時間と発話数を表 1 に示す（発話数の数え方は 3.1 節に示す）．

3. 提案する意味学習アルゴリズム

3.1 提案アルゴリズムの概要

本研究では，実験参加者の発話の中で，下記の種類の発話の意味をエージェントが獲得することを目指す．1 つの意味を持つ発話は複数種類あることを想定している．学習対象とする発話の意味は次の 6 種類である：エージェントが次に進むべき方向を示す行動教示 4 種《上／下／左／右に進め》（以後は，略

表 1: 各試行における所要時間と発話数

実験参加者	各試行に要した時間と発話数								合計発話数
	第 1 試行 (0.2)		第 2 試行 (0.4)		第 3 試行 (0.6)		第 4 試行 (0.8)		
	時間 (秒)	発話数	時間 (秒)	発話数	時間 (秒)	発話数	時間 (秒)	発話数	
A	176	89	98	43	71	28	61	20	180
B	135	54	79	31	51	18	38	12	115

号《↑／↓／←／→》を用いる），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 種《適切だった》《不適切だった》（以後は，略号《○／×》を用いる）である．また，これら 6 種類以外の意味を持つ発話の存在も想定しているが，それらには意味が付与されないことを目指している．

提案するアルゴリズムの流れは以下の (1)〜(5) である．このうち (4) と (5) については，次節以後に詳述する．

1. **発話抽出**
   連続な音声信号から発話を抽出する．抽出の方法は，エージェントの行動タイミングおよび 300ms 以上の無音区間で発話を区切り，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，2 秒間隔である．エージェントは，2 秒毎に次の位置に瞬間移動する．なお，正確には，人の反応時間を考慮して，行動タイミングから 300ms 遅れたタイミングで区切っている．

2. **音節認識**
   抽出された発話は Julius-3.4[1] を単音節認識器として用いて音節列へと変換する．音響モデルは，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使用した．言語モデルは，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. **状況と共起する発話を集める**
   本研究では，参照的な意味でなく機能的な意味を扱うため，音声との共起関係を調べる対象として，具体的な事物でなく，抽象的な「状況」を考える必要がある．本研究で扱う状況には 2 種類あって，1 つは，次に取るべき行動についての状況《↑／↓／←／→》，もう 1 つは直前の行動の評価についての状況《○／×》である．本論文では，意味とはこの状況との対応であるとの立場をとるため，状況の表記も意味と同じ表記《》を用いる．
   本研究では，次に取るべき行動についての状況は，直前の行動から次の行動までの間，成立しているとする．また，直前の行動の評価についての状況は，直前の行動から次の行動までの間，成立しているとする．このように定義すると，各時刻それぞれは，同時に 2 つの状況に属することに注意されたい．したがって，各発話は，2 つの状況 1 つは，次に取るべき行動についての状況，もう 1 つは直前の行動の評価についての状況）と共起することになる．
   このステップでは特定の状況と共起する発話を集めて，次のステップである類似区間抽出処理の対象とする．なお，人の反応の遅れを考慮して，状況の区切りに対して，発話の区切りは 300ms 遅らせた共起の判断をした．

		β					
		ん	あ	い	う	え	⋯
α	ん	14620	105	381	323	98	
	あ	12	2284	46	22	43	
	い	57	30	10245	64	889	
	う	23	7	47	1246	14	⋯
	え	12	17	308	20	2919	
	⋮			⋮			⋱

図 1: Confusion Matrix の一部．α は入力音声中の音節，β は認識後の音節．

4. **語候補の抽出**
   同じ状況で発話された音節列の中で，繰り返し出現する類似区間を検出し，それを語候補（意味付けを試みる対象）とする．

5. **語候補への意味付け**
   ある語候補に対して，それと類似した語候補を 1 つのグループとする．そのグループに属することと，その音声が発せられた特定の状況が共起する割合の特異性を Fisher の直接法で計算し，特異的に共起しやすい語候補グループと状況のペアが見つかれば，そのグループに属する語候補の意味は，その状況であるとする．

3.2 語候補の抽出

本研究では，発話（この段階までにすでに音節列に変換してある）の中で繰り返し出現する類似区間（語候補）を学習対象として抽出し，意味学習を行う．語候補の抽出処理は，同じ状況で発話された音節列に対してだけ行う．

3.2.1 音節間の類似度スコア

本節では，2 つの音節列間の類似度を計算するために必要となる音節間の類似度を定義する．Julius で認識した各音節の間の類似度を Confusion Matrix を用いて算出する．

本研究で使う Confusion Matrix は，音節 ω_i $(1 \leq i \leq n)$ が入力として与えられたとき，これが音節 ω_j $(1 \leq j \leq n)$ であると識別された回数を要素とする $n \times n$ 行列である．本研究では，日本音響学会の研究用連続音声データベースに収録されている ATR 音素バランス 503 文の音声波形をもとにして Confusion Matrix を作成した．図 1 に作成した Confusion Matrix の一部を示す．

作成した Confusion Matrix より，音節 a と音節 b の類似度を式 1 定義する

$$s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right) \tag{1}$$

ここで，p_{ab} は，音節 a が出現し，かつ，それが音節 b と認識される確率，q_a は，音節 a が出現する確率，q_b は，音節 b という認識結果が得られる確率である．これら p_{ab}，q_a 及び q_b を先に生成した Confusion Matrix から計算し，類似度を求める．類似度 s は，a と b が対になる確率が，偶然に対になるのに比べてどれだけ大きいかを示したものである．

3.2.2 音節列間の類似度を計算するアルゴリズム

前節で求めた音節間の類似度に基づき，Smith-Waterman のアルゴリズム [4] を用いて，音声の類似区間を検出した．

Smith-Waterman のアルゴリズムは，動的計画法の考え方に基づいてローカルアラインメントを行う．具体的には，音

節列 $A = \{a_1, a_2, \cdots, a_m\}$ と $B = \{b_1, b_2, \cdots, b_n\}$ に対して，次のようにして類似する音声区間を求める．

1. $(m+1) \times (n+1)$ 行列 H を作成し，初期値として

$$H_{i0} = H_{0j} = 0 \quad (0 \leq i \leq m, 0 \leq j \leq n) \tag{2}$$

を与える．

2. 続いて，残りのノードのスコアを式 3 に従って埋めていく．

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j) \\ \max\limits_{k \geq 1}\{H_{i-k,j} - W_k\} \\ \max\limits_{l \geq 1}\{H_{i,j-l} - W_l\} \\ 0 \end{cases} \tag{3}$$

3. 全てのノードについてのスコアの計算完了後，閾値を超えるスコアを持つノードからスコアがゼロになるノードまでを逆向きに辿っていくことで，類似音声区間を求めることができる．

ここで，$s(a_i, b_j)$ は前節で求めた類似度である．W_k および W_l は挿入・削除のペナルティであり，式 4 で定義される．

$$W_k = dk \quad (d \text{ は } -0.2 \text{ とした}) \tag{4}$$

そして，二つの音節列の類似度を式 5 で定義する．

$$SS = 2 \times \frac{H}{L_1 + L_2} \tag{5}$$

SS は長さ L_1 と L_2 の 2 つの音節列間の類似度，H は Smith-Waterman のアルゴリズムで計算した同じ音節列間の類似度である．また，今回の実験では，類似音声区間の最低検出区間長は 2 とした．

3.3 語候補への意味付け

ある状況での発話中に類似音声区間（語候補）が見つかったからといって，必ずしもその区間の音声の意味がその状況であるわけではない．ある状況において検出された類似音声区間（語候補）は，その意味がその状況とは対応しないものを 7 割程度含む．このため，何らかの基準により，抽出された状況と対応する意味を持つ可能性が高いもの（語）に絞り込む必要がある．本研究では，候補の選択基準として，Fisher の直接法により算出した生起確率を用いる．

本研究では以下の (1)〜(7) の手順で語候補を抽出し，それに意味を付ける．

1. 意味付け処理の対象である語候補（類似音声区間）の集合 S から，1 つの語候補 x を選択し，プロトタイプとする．

2. プロトタイプ x に対する，集合 S の x 以外の要素（語候補）の類似度を 3.2.2 節で述べたアルゴリズムで計算する．

3. プロトタイプ x との類似度がある基準値より大きい語候補をひとまとめにしてグループを作成する．その際，基準値を変えながら，グループに属する語候補の数が 1 つ、2 つ、3 つ、... 最大値（すべての語候補を含むグループ）となる全てのグループをつくる．

4. 1 つのグループ z_x と 1 つの状況 y を選択し，Fisher の直接法を用いて，ある語候補がグループ z_x に属することとその音声が状況 y で発せられたことの共起頻度パターンの生起確率 P と，観測されたパターン以上に偏ったパターンが観測される確率の総和 P_t を算出する．この総和 P_t は，グループ z_x に属する音声と状況 y の共起度合の特異性を示すものである（詳しくは [2] を参照）．

5. ステップ 4 の生起確率の総和 P_t の計算を，すべてのグループ，すべての状況（今回の評価実験では《↑ ↓←→ ○×》の 6 種類）の組合せに対して行い，P_t が最小の値をもつグループ−状況のペアを暫定的に記録する．

6. 集合 S に属するすべての語候補に対して，上記の（1）から（5）の処理を行い，この過程で暫定的に記録されたグループ−状況のペアの中で，確率の総和 P_t が最小の値をもつグループ-状況のペアをみつける．もし当該グループの P_t が一定の閾値以下であれば，このグループに属する語候補の意味はこの状況であると結論する．そして，このグループに属している語候補は，集合 S から削除する．もし当該グループの P_t が一定の閾値以上であれば，処理が終了する．このようにして，手動で閾値の値を選べるのであれば，これを自動選択も可能である．クロスバリデーションなどの手法を用い，性能が高くなる閾値を調べてそれを選べばよい．閾値の自動選択は今後の課題とする．

7. 集合 S に残っている語候補に対して，語候補がなくなるまで上記の（1）から（6）の処理を繰り返す．

4. 学習性能の評価と考察

提案手法の評価として，2. 節で収集したインタラクションデータを用い，10-fold Cross Validation 法によって実験参加者毎に学習性能を確かめた．

まず，実験参加者 B の意味学習結果の例を表 2 に示す．各行は，あるプロトタイプを中心とした語候補のグループを示している．“音節認識結果” の列はそのグループのプロトタイプとなった音声（音節）の認識結果を示し，“書き起こし” 列はその音声に対応する書き起こしを示す．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の音節が付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（音節）が脱落したことを示す．“類似度の閾値”の列は，そのグループに属するための音節とプロトタイプとの類似度の閾値を示す．新しい音声とこのプロトタイプとの類似度がここに示した閾値を超える場合，この音声はこのプロトタイプを中心としたグループに属すると判断される．“状況”の列は，3.3 節のステップ 6 で得られたグループ-状況のペアの状況を示す．当該グループに属する語候補の意味は，この状況であると結論される．

以下，学習性能を精度と再現率の 2 つの観点から評価する．

実験参加者 B の精度と再現率の計算例を表 3 に示す．“発話” の列は，実験参加者の発話を示す．“正解” の列は，各発話を筆者の一人が解釈して付けた意味である．正解の種類は《↑ ↓←→○×》の 6 つから選んだ．各発話に対し，正解は複数ある．たとえば,「そうつぎもしただよ」という発話には,「直前の行動が適切だった」という評価教示と「下に進め」という行動教示が含まれている．なお,「ひだりうえ」と言う発話は,「左に進め」と「上に進め」という 2 つの行動教示が含まれているとする.「がんばれあかちゃん」と言う発話は，どの教示も

表 2: 実験参加者 B の意味学習結果の例（語候補のグループごと）

音節認識結果	書き起こし	類似度の閾値	状況
みいーみ	みぎみ (ぎ)	−0.040254	→
ぎもすどだ	(つ) ぎもしただ	0.006854	↓
うふぎもそ	(そ) うつぎも ()	0.049976	○
ちじゃな	(そっ) ちじゃない	0.018412	×
えよ	(な) いよ	0.131737	×

含まれておらず，正解なしとする．“システムにより付けられた意味” の列では，各発話に対して，システムによって付けられた意味を示す．システムにより付けられる意味の種類も《↑↓←→○×》の 6 つから選ばれる．“★” は，付けられた意味が正しいと表している．“△” は，付けられた意味は正しくないと表している．“□” は，当該意味が検出できなかったと表している．そして，精度は式 6 で定義した．

$$\text{精度} = \frac{\text{★の数}}{\text{★の数} + \text{△の数}} \tag{6}$$

再現率は式 7 で定義した．

$$\text{再現率} = \frac{\text{★の数}}{\text{★の数} + \text{□の数}} \tag{7}$$

表 3 の例では，精度は 63%，再現率は 75%となった．

最後に，実験参加者 A と実験参加者 B に対して，10-fold Cross Validation 法による評価結果を表 4 に示す．二人分のデータでは，平均で 63%の精度と 72%の再現率が得られた．

表 3: 実験参加者 B の精度と再現率の計算例

発話	正解			システムにより検出された意味					
				↑	↓	←	→	○	×
そうつぎもしただよ	↓	○			★			★	
そっちじゃないひだりうえだよ	←	↑	×				△		★
				□		□			
そうつぎもみぎだよ	→	○			△		★		△
								□	
そうそうつぎも	○				△			★	
がんばれあかちゃん							△		
そうつぎみぎにみるくがあるよ	→	○					★		
								□	
つぎはみぎだよ	→						★	△	
そうそうつぎはしたいこうね	↓	○			★		△	★	
そうそうそうつぎもしただよ	↓	○			★			★	
そうつぎもひだりだよ	←	○			△		△	★	
						□			
つぎはしただよ	↓				★				
そうつぎもしただよ	↓	○			★			★	
精度	63%								
再現率	75%								

表 4: 評価実験の結果

実験参加者	精度	再現率
A	60%	63%
B	65%	80%
平均	63%	72%

5. 結論

本論文では，切迫した状況で取得した自然な発話を含むインタラクションデータからの，言葉の機能的意味の獲得アルゴリズムを提案した．また評価実験により，複数の働きを持つ言葉が混在した中で，ある程度の精度と再現率とで，機能的意味を獲得できることを示した．

今後は，より多くの実験参加者で評価実験を行う予定である．また，取得データのオフライン解析でなく，オンライン環境における，提案した学習アルゴリズムの性能を検証することを計画している．

参考文献

[1] 河原達也，李晃伸，“連続音声認識ソフトウエア Julius，” 人工知能学会誌，vol.20，no.1，pp.41–49(2005).

[2] 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).

[3] Roy,D.: “Learning from sights and sounds: a computational model,” Ph.D. Thesis, MIT Media Laboratory, Cambridge(1999).

[4] Smith,T.F. and Waterman,M.S.: “Identification of common molecular subsequences, ”J.Mol.Biol, vol.147, pp.195–197(1981).

[5] 鈴木健太郎，植田一博，開一夫，“自律的な行動学習を利用した評価教示の計算論的意味学習モデル，” 認知科学，vol.9，no.2，pp.200–212(2002).

[6] Yu,C.: “A multimodal learning interface for grounding spoken language in sensory perceptions,” ACM Trans. Applied Perceptions, vol.1, no.1, pp.57–80(2004).