発話シーンの共有信念に基づく推定とその評価

木村 優志[*1]　作元 佑輔[*1]　田口 亮[*1]　桂田 浩一[*1]　岩橋 直人[*2*3]　新田 恒雄[*1]
[*1] 豊橋技術科学大学　[*2] 情報通信研究機構　[*3] 国際電気通信基礎技術研究所

1. はじめに

人間同士の対話では，相手から伝えられた言語情報だけではなく言外の情報を推定することによって，円滑なコミュニケーションを実現する場合が多い．言外の情報を推定するための条件として，Grice は，協調の原則や発話の背景知識を話者同士で共有する必要があると指摘した[Grice 75][石崎 01]．本稿では，言外の情報として発話場面を対象に，ロボットが人間との共有信念に基づいて発話場面を推定する手法を提案する．

2. ロボットの基本機能

本稿の実験では，ロボットが人との対話を通して予め共有信念を形成した後，人の発話から発話場面を推定する．本節では，共有信念の形成，ロボットの行動と発話理解・生成について述べる．

2.1 共有信念の形成

まず，人がロボットに対して語意を教示する．図1のように，人がロボットにオブジェクト o を見せながら単語 w_o を発話することで，それらの条件付確率 $p(o|w_o)$ を学習させる[Iwahashi 04]．また，動作を表す単語 w_m も同様に，動作 u をロボットに見せ，対応する単語 w_m を発話することで条件付確率 $p(u|o_t,o_l,w_m)$ を学習させる[羽岡 00]．ここで，o_t はトラジェクタ（動かすオブジェクト），o_l はランドマーク（動作の基準点となるオブジェクト）である．

次に，図 2 のように，人がロボットに文を発話し，ロボットがその発話通りにオブジェクトを動かす（またはその逆の）インタラクションを行う．この過程でロボットは，文法だけでなく，動作とオブジェクトの関係（箱は「乗せる」という動作のランドマークになりやすいなど）や，どのくらい曖昧な発話でも正しく相手に伝わるか(成功確率)なども学習する．

2.2 人の発話の理解とロボットの行動

ロボットは場面 v で人の発話 s を聞くと，下の式(1)を用いて発話 s に相応しい動作 a を求める[Iwahashi 06]．$\psi(s,a,v)$ の各項は音声や，動きを表す単語等の各信念を表している．これらの信念は確率モデルとして表現されており，人との対話を通して学習される．ロボットは発話 s が与えられると，その場面で可能な全ての動作に対して $\psi(s,a,v)$ を計算し，その中から $\psi(s,a,v)$ が最大となる動作 a を求める．

$$\begin{aligned}\psi(s,a,v) = \max_{l,z}\{\ & \gamma_1 \log p(s|z) && \text{[音声]}\\ & +\gamma_2 \log p(u|o_t,o_l,W_M) && \text{[動き]}\\ & +\gamma_2(\log p(o_t|W_T)+\log p(o_l|W_L)) && \text{[オブジェクト]}\\ & +\gamma_3 \log p(o_t,o_l|W_M)\ \} && \text{[動き-オブジェクト関係]}\end{aligned} \quad (1)$$

o_t: 動かすオブジェクト，o_l : 動作の基準点，u: 軌道，
a: 動作(u, o_t)，W_T : o_t を表す単語
W_L : o_l を表す単語，W_M : 動作を表す単語
z: 発話の意味構造 $z=\{W_T, W_L, W_M\}$，
$\gamma_{1,2,3}$: 共有の確信度

2.3 ロボットの行動と発話生成

ロボットは，目的となる動作 a を与えられると，式(2)に基づいて発話 s_r を生成する．$f(d(s,a,v))$ は，発話 s が人間に正しく理解される確率（成功確率）を出力する関数であり，

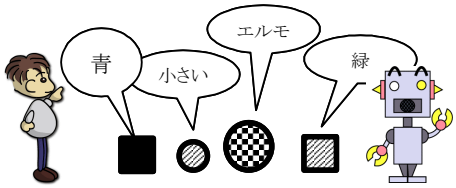

図 1 語意学習

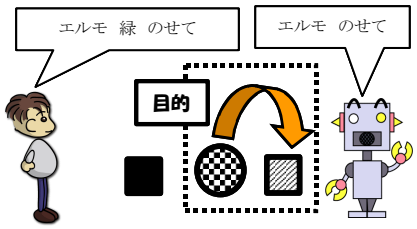

図 2 発話と動作の関係を学習

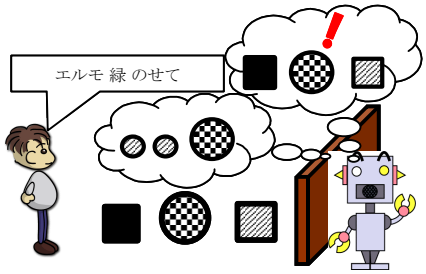

図 3 発話場面の推定

ε は発話を生成するときの目標確率を示す．$d(s,a,v)$ は発話 s の曖昧性を表している．$d(s,a,v)$ が 0 に近い場合，発話 s が曖昧であることを意味し，負値の場合は発話 s が不適切であることを意味する．曖昧さと成功確率の関係は予め定義しておくことができないため， 実際の対話を通して調整される．

$$s_r = \arg\min_s \left| f(d(s,a,v)) - \varepsilon \right|^2 \quad (2)$$

$$f(x) = \frac{1}{\pi}\arctan\left(\frac{x-\lambda_1}{\lambda_2}\right) + 0.5 \quad (2-1)$$

$$d(s,a,v) = \psi(s,a,v) - \max_{A\neq a}\psi(s,A,V) \quad (2-2)$$

3. 共有信念に基づく発話場面の推定

本研究では，図3に示すように，人が目隠しをしたロボットに発話をする場面を考える．ロボットは，人間の発話 s_h に対する場面 v の確からしさを確信度 $Conf(s_h, v)$ として計算する．ロボットは，これまでに経験した全ての場面に対して確信度を計算し，それが高い場面を推定結果として想起する．本稿では，以下の 3 種類の確信度を提案し，それぞれの有効性を比較する．

3.1 発話と動作の適切さを用いた確信度 Conf1

$Conf1(s_h, v)$ は，2.2 節で説明した $\psi(s_h, a_h, v)$ を用いて計算する．具体的には，過去に経験した各場面 v に対して $\psi(s_h, a_h, v)$ を最大にする a_h を求め，$\psi(s_h, a_h, v)$ の値を確信度 $Conf1(s_h, v)$ とする（式(3)）．この確信度を用いることは「ロボットが最も適切な動作を取れるよう，人は発話する」という仮定の上で場面を推定することを意味する．

$$Conf1(s_h, v) = \psi(s_h, a_h, v) \quad (3)$$

3.2 発話の曖昧さを用いた確信度 *Conf2*

Conf2(s_h, v)は，発話と動作の適切さに加えて，発話の曖昧さ $f(d(s_h,a,v))$を考慮したモデルである(式(4))．これは「人は曖昧な発話はしない」と仮定し，場面を推定することを意味している．

$$Conf\ 2(s_h,v)=\psi(s_h,a_h,v)\cdot f(d(s_h,a_h,v)) \qquad (4)$$

3.3 単語の寄与度を用いた確信度 *Conf3*

Conf3(s_h, v)では単語の寄与度を利用することを考える．単語の寄与度とは，発話内のある単語が発話の目的を達成するのにどれほど寄与しているかを表した物である．ここでは，単語の発話によるロボットの行動の成功率の増加をその単語の寄与度とする．具体的な手順を以下に示す．

まず，人の発話 s_h から動作以外の単語 w_i を一つずつ抜いた発話 $S_g = \{ s_{g1}, \cdots, s_{gi}, \cdots, s_{gn} \}$ を生成する．そして，人の発話 s_h の成功確率 $f(d(s_h,a_h,v))$ と，生成した発話 s_{gi} の成功確率 $f(d(s_{gi},a_h,v))$ との差を計算し，それを単語 w_i の寄与度 $E(w_i,a_k,v)$ とし，式(5)で計算する．$d(s,a,v)$ が負値（発話が曖昧である）のときは式(5-1)のように $d(s,a,v)$ を 0 にする．

$$E(w_i,a_h,v)=f(d'(s_h,a_h,v))-f(d'(s_{gi},a_h,v)) \qquad (5)$$

$$d'(s,a,v)=\begin{cases} d(s,a,v) & (d(s,a,v)\geq 0) \\ 0 & (d(s,a,v)<0) \end{cases} \qquad (5-1)$$

確信度 *Conf3*(s_h,v)は，寄与度の平均と発話と動作の適切さとの積とする（式(6)）．これは「人は無駄な単語を発話しない」と仮定し，場面を推定することを意味する．

$$Conf\ 3(s,v)=\psi(s,a,v)\frac{1}{n}\sum_{i=1}^{n}E(w_i,a,v) \qquad (6)$$

4. 実験

ロボットには予め以下の単語を学習させた上で，「エルモ　緑色　のせて」と発話して場面を想定させた．

【色：3 種類】　赤色，青色，緑色．
【大きさ：2 種類】　大きい，小さい．
【物の名前：9 種類】箱，エルモ，グローバー，ダンボ，バーバ，カーミット，プーサン，ラッコ，ミカン．
【動作：7 種類】　のせて，飛び越えて，持ち上げて，近づいて，離れて，下げて，回して

ロボットにはオブジェクトが三つ含まれる場面(161 場面)を予め与え，その全ての場面に対し 3 種類の確信度を求め，それぞれ上位 10 種類の場面から有効性を評価した．式(1)，式(2-1)のパラメータの値は，γ_1=0.03333，γ_2=1，γ_3=0.5，λ_1=3.66186，λ_1=0.588504 である．

Conf1(s_h,v)，*Conf2*(s_h,v)，*Conf3*(s_h,v)を用いて推定された場面を図 4，図 5，図 6 に示す．3 つの図から，推定された場面全てにエルモと緑色のオブジェクトが含まれていることがわかる．確信度の計算で利用している $\psi(s_h,a_h,v)$は，発話された単語が，動作に関連するオブジェクトを表す確率が高いほど大きな値になるため，当該オブジェクトを含む場面が推定されやすくなる．

一方，図 4 で 1 位となった場面は，図 5，6 では候補から外れた．この場面では，緑色のオブジェクトが二つあるため，「緑色」と発話されてもどちらであるか曖昧である．そのため，成功確率 $f(d(s_h,a_h,v))$ が小さくなり，*Conf2*(s_h,v)や *Conf3*(s_h, v)では候補から外れた．

また，図 5 と図 6 を比較すると，*Conf2*(s_h,v)では 5，7，8 位となった場面が，*Conf3*(s_h,v)では候補から外れていることがわかる．これは，2.1 節で説明した共有信念の形成時に「箱や小さなぬいぐるみには物がのせられやすい」という信念をロボットが獲得していたためである．この場合，「エルモ　のせて」といった発話でも意図が伝わるため，「緑色」の寄与度が小さくなり，結果として *Conf3*(s_h, v)では候補から外れた．以上から，単語の寄与度を利用することでより詳細な場面を推定できることが確認できた．

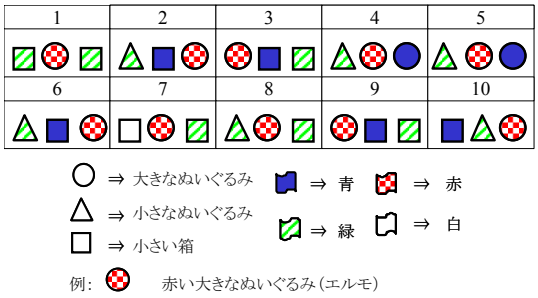

図 4　*Conf1*(s_h,v)に基づいて推定した場面

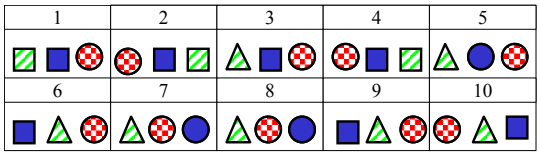

図 5　*Conf2*(s_h,v)に基づいて推定した場面

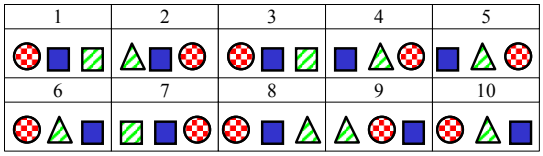

図 6　*Conf3*(s_h,v)に基づいて推定した場面

5. まとめ

本稿では，共有信念を獲得したロボットが，発話だけから，その発話がなされたであろう場面を推定する方法を提案した．実験の結果，発話と動作の適切さだけでなく，発話された単語の寄与度を考慮することで，発話内のすべての単語が推定結果に反映される詳細な場面を推定できることが示された．今後は，より多数のオブジェクトが場面に含まれる場合や，発話に含まれる単語の数を増やした場合などについても実験を行っていきたい．

参考文献

[Grice 75] H. P. Grice: Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and Semantics, Volume 3: Speech Acts, (pp. 41--58). New York: Academic Press, (1975).

[石崎 01]石崎雅人，伝康晴：談話と対話，東京大学出版会，pp.24-28，(2001).

[Iwahasi 04] Naoto Iwahashi. : Active and unsupervised learning of spoken words through a multimodal interface. In: IEEE Workshop on Robot and Human Interactive Communication, pp. 437-442, (2004).

[羽岡 00] 羽岡 哲郎，岩橋 直人: “言語獲得のための参照点に依存した空間的移動の概念の学習”，信学技報，PRMU2000-105, pp39-46, (2000).

[Iwahashi 06] Naoto Iwahashi: Robots That Learn Language: Developmental Approach to Human-Machine Conversations, LNAI4211 Symbol Grounding and Beyond, pp.143-167, (2006).