3E3-7

人間－ロボット間の共有信念に基づく発話場面の推定

Inference of situation based on Shared Belief between human and robot

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]
Masashi Kimura Yusuke Sakumoto Ryo Taguchi Koichi Katsurada Naoto Iwahashi Tsuneo Nitta

[*1] 豊橋技術科学大学 Toyohashi University of Technology
[*2] 情報通信研究機構 National Institute of Information and Communications Technology
[*3] 国際電気通信基礎技術研究所 Advanced Telecommunications Research Institute International

We propose a method to infer the surrounding situation of a human through speech interaction between a robot and the human. In this paper, the robot is assumed to be put on a blindfold, and can communicate with the human only through speech. To infer the surrounding situation, at first, the robot tries to share some beliefs (such as lexicon, grammar, and so on) after several turns of interaction with the human. After completing to share the beliefs, the robot listens to the human's utterance and infers his situation based on the beliefs. In the inference process, the robot remembers some situations that it has ever seen. Then it calculates confidence measures of the situations based on easiness to understand the utterance, success probability in understanding the utterance, contribution to appearance of the word, and so on. The experimental results show that contribution appearance plays an important role in inferring detailed situation from human utterance.

1. はじめに

人間同士の対話では，伝えられた言語情報の理解だけではなく，話者の置かれた状況などの言外の情報を推定することによって幅広いコミュニケーションを実現している．

例えば，AがCに手紙を送ろうとしており，それをBが知っているとして，次のような対話がなされたとする．

A:「Cさんはどこに住んでいるのですか？」
B:「フランスの南のどこかです．」

ここで，Bの発話の状況を推定することで，その発話の内容以上の事は知らないということが言外の情報としてAに伝わる．このように，人間同士の対話では，伝えられた言語情報を直接的に理解するだけではなく，発話の裏側を推定することで，幅広いコミュニケーションを実現している．こうしたコミュニケーションを人とロボットの間で実現することが本研究の目的である．

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両者で共有する必要があると指摘した[Grice 75][石崎 01]．そうした背景知識や文脈の共有が無ければ，A，Bの発話が与えられても，そこから言外の情報を推論することはできない．近年，岩橋らによって，人とロボットが共に経験を重ねることで信念を共有し，その共有された信念（共有信念）に基づいて，発話を理解・生成するための手法が提案された[Iwahashi 06]．先行研究では，人とロボットが同じ場面を共有していることを前提としていたが，本稿ではその手法を応用し，人の発話だけからロボットが話者の場面を推定する手法を提案する[作元 2007]．提案手法では，共有信念に基づいて発話に対する場面の適切さを表す尺度を確信度として求め，これが最も高い場面を推定結果とする．以下，2 章では本研究の問題設定を説明し，3 章では先行研究の発話理解・生成のメカニズムについて説明する．4 章では発話場面を推定するために，発話の理解のし易さや，成功確率，単語の発話目的を達成するための寄与度などに基づいた 3 種類の場面推定手法を提案する．続いて 5 章で先述の3種類の手法について場面推定の実験を行い，発話の理解のし易さだけでなく，単語の寄与度を考慮することで，より詳細な場面を推定できることを述べる．6 章で本稿の内容をまとめる．

連絡先：木村優志，豊橋技術科学大学，〒441-8580 豊橋市天伯町雲雀ヶ丘１－１，(0532)44-6890，kimura@vox.tutkie.tut.ac.jp

2. 問題設定

本稿の実験では，ロボットが人との対話を通して共有信念を形成した後，人の発話だけから発話場面を推定する課題を扱う．この節では，場面推定課題の問題設定について述べる．

2.1 共有信念の形成

まず，ロボットは人の教示から語意を学習する．語意学習では，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．たとえば，図 2 の矢印で示したような動作を「乗せる」と解釈する場合，中央の丸がトラジェクタとなり，右の四角がランドマークとなる．

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 発話場面の推定

人が目隠しをしたロボットに発話をする場面を考える(図 3)．ロボットはその発話だけから，できるだけ詳しく場面(オブジェクトの配置)を推定する．ここで，ロボットは共有信念の形成過程で

経験した場面を辿りながら，記憶している全ての場面に対して，なされた発話の確信度を計算する．そして，確信度の高い場面を推定結果として想起する．

3. 共有信念に基づく発話理解・生成

提案手法は，岩橋らによる共有信念に基づいた発話の理解・生成のメカニズムを応用したものである[Iwahashi 06]．以下，そのメカニズムについて説明する．

3.1 発話理解

ロボットは場面 v で人の発話 s_h を聞くと，式(1)，(2)を用いて発話に相応しい動作 a_h(動かすオブジェクト o_tとその軌道 u)を出力する．

$$a_h = \arg\max_a \psi(s_h, a, v) \quad (1)$$

$$\begin{aligned}\psi(s,a,v) = \max_{l,z}\{ & \gamma_1 \log p(s|z) && \text{[音声]}\\ & +\gamma_2 \log p(u|o_t,o_l,W_M) && \text{[動き]}\\ & +\gamma_2(\log p(o_t|W_T)+\log p(o_l|W_L)) && \text{[オブジェクト]}\\ & +\gamma_3 \log p(o_t,o_l|W_M)\} && \text{[動き-オブジェクト関係]}\end{aligned} \quad (2)$$

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

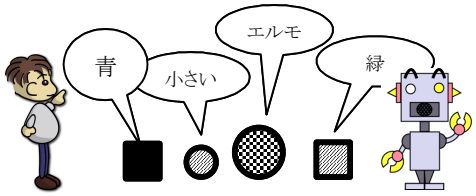


図 1 語意学習

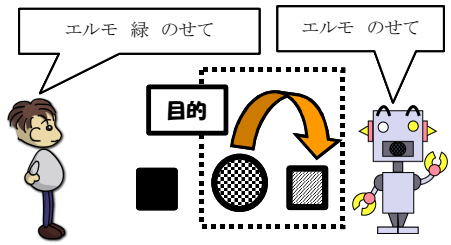


図 2 発話と動作の関係を学習

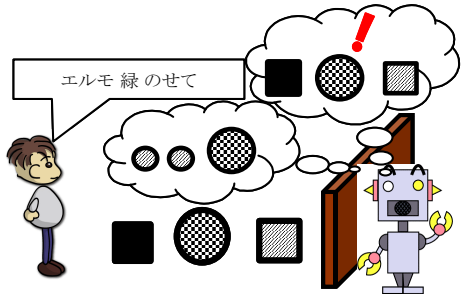


図 3 発話場面の推定

ここで，$\psi(s, a, v)$は発話 s と動作 a の対応の適切さを表している．$\psi(s, a, v)$の各項は音声や，動きを表す単語，オブジェクトを表す単語，動きとオブジェクト特徴の関係といった信念を表している．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_hが与えられると，その場面で可能な全ての動作に対して$\psi(s_h,a,v)$を計算し，その中から$\psi(s_h,a,v)$が最大となる動作を求める．

3.2 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(3)に基づいて発話 s_rを生成する．$f(d(s, a, v))$は，発話 s が正しく理解される確率を出力する関数であり，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．

$$s_r = \arg\min_s \left| f(d(s,a,v)) - \varepsilon \right|^2 \quad (3)$$

$$f(x) = \frac{1}{\pi}\arctan\left(\frac{x-\lambda_1}{\lambda_2}\right) + 0.5 \quad (3-1)$$

$$d(s,a,v) = \psi(s,a,v) - \max_{A\neq a} \psi(s,A,V) \quad (3-2)$$

4. 共有信念に基づく発話場面の推定

本研究では，人の発話 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場面全て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，発話と動作の適切さを用いた確信度 $Conf1(s,v)$，発話の曖昧さを用いた確信度 $Conf2(s,v)$，単語の寄与度を用いた確信度 $Conf3(s,v)$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．

4.1 発話と動作の適切さを用いた確信度 $Conf1(s_h, v)$

$Conf1(s_h, v)$は，3.1 発話理解で説明した発話と動作の対応の適切さ$\psi(s_h, a_h, v)$を用いて計算される．具体的には，各場面 v において最大となる$\psi(s_h, a_h, v)$を確信度 $Conf1(s_h, v)$とする(式(4))．a_hは式(1)から求められる発話にふさわしい動作のことである．この確信度を用いることは「人は目的の動作を適切に表現する」と仮定し，場面を推定することを意味する．

$$Conf\ 1(s_h, v) = \psi(s_h, a_h, v) \quad (4)$$

4.2 発話の曖昧さを用いた確信度 $Conf2(s_h,v)$

$Conf2(s_h, v)$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(5))．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おり，Grice の協調の原則(話者が目的達成のために当を得た発話を行う)と対応している．

$$Conf\ 2(s_h, v) = \psi(s_h, a_h, v) \cdot f(d(s_h, a_h, v)) \quad (5)$$

4.3 単語の寄与度を用いた確信度 $Conf3(s_h,v)$

*Conf2(s_h,v)*では発話全体としての成功確率 *f(d(s_h, a_h, v))*を利用していたが，*Conf3(s_h, v)*ではさらに詳細な分析を得るために，各場面における単語の寄与度を算出し利用することを考える．ここで単語の寄与度とは，発話内のある単語が発話の目的を達成するためにどれほど寄与しているかを表した物である．例えば，青い箱と緑の箱が置かれている場面で，「青い箱を持ち上げて」と言ったとする．その時，持ち上げる対象はどちらも箱であるため，「箱」という単語の寄与は低い．一方，目的達成に大きく貢献する「青い」の寄与度は高い．ここでは，それぞれの単語を発話に追加することによって変化する成功確率をその単語の寄与度とする．具体的な手順をいかに示す．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 とし，式(6)で計算する．*d(s,a,v)*が負値になる場合，生成した発話が動作に不適切であることを意味する．発話が不適切である場合と曖昧である場合は，ともに発話の目的を達成することができないため，双方を同一視し，式(6-1)のように *d(s,a,v)*を 0 して扱う．

$$E(w_i,a_h,v) = f(d'(s_h,a_h,v)) - f(d'(s_{gi},a_h,v)) \qquad (6)$$

$$d'(s,a,v) = \begin{cases} d(s,a,v) & (d(s,a,v) \geq 0) \\ 0 & (d(s,a,v) < 0) \end{cases} \qquad (6-1)$$

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(7))．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．

$$Conf\,3(s,v) = \psi(s,a,v)\frac{1}{n}\sum_{i=1}^{n} E(w_i,a,v) \qquad (7)$$

5. 実験

本実験では，ロボットに発話だけを与え，その発話がなされたであろう場面を推定させる．与える発話 s_h は，「エルモ　緑色のせて」とした．ただし，音声認識誤りはおこさないと仮定しテキストにより与える．また，ロボットには約 800 場面の中から予め先行研究で使われた，オブジェクトが三つ含まれる場面(161 種類)を候補として与えた．実験では，その全ての場面に対し 3 種類の確信度を求め，その上位 10 種類の場面の傾向から有効性を評価する．実験の際に使用した，式(2)，式(3-1)パラメータの値は表1に示す．また，ロボットには事前に，語意や発話と動作の関係を学習させている．学習した単語は，色 3 種類，大きさ 2 種類，物の名前 9 種類，動作 7 種類である．以下にその単語を示す．

【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

*Conf1(s_h,v)*で推定された場面を図 4，*Conf2(s_h,v)*を図 5，*Conf3(s_h,v)*を図 6 に示す．3 つの図から，推定された場面全てにエルモや緑色に相当する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ψ *(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，各単語が意味する典型的な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」と指示されてもそのどちらのせればよいのか曖昧である．そのた

表 1 実験のパラメータ

パラメータ	γ_1	γ_2	γ_3	λ_1	λ_2
値	0.03333	1	0.5	3.66186	0.588504

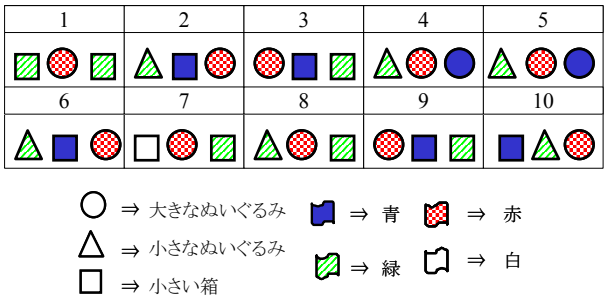


図 4 *Conf1(s_h,v)*に基づいて推定した場面

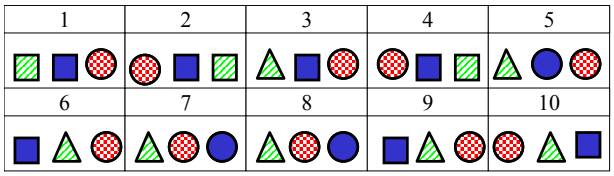


図 5 *Conf2(s_h,v)*に基づいて推定した場面

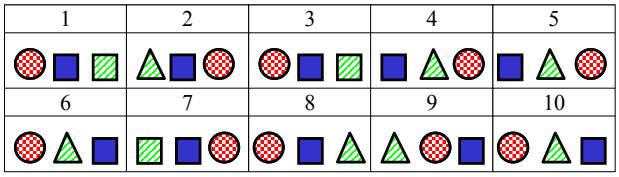
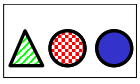


図 6 *Conf3(s_h,v)*に基づいて推定した場面

表 2 図 6 で除外された場面(図 5 では 7 位)

発話		*a*		*Ψ(s,v)*	曖昧さ	成功確率
		t	*l*			
s_h	エルモ　緑色のせて	2	1	625.06	20.27	0.99
s_{g1}	エルモ　のせて	2	1	590.92	2.73	0.18
s_{g2}	緑色　のせて	2	1	603.02	1.97	0.11
				寄与度の平均		0.846

表 2 の場面：



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．$Conf3(s_h,v)$を算出する際に生成する発話 S_g は「エルモ　のせて」と「緑色　のせて」である．表 2 に，図 5 の 7 位の場面の $\psi(s,a,v)$を示す．表 2 において t は動作 a のトラジェクタ，l はランドマークを表し，その数字は場面の中で左から数えた順番に対応している．表 2 から，「エルモ　のせて」や「緑色　のせて」という発話でも，「エルモを左の緑のぬいぐるみにのせる」と解釈できることがわかる(動作 a が t=2，l=1 であるため)．このような「エルモ」や「緑色」を発話しなくても，正しく発話が理解できる場面では，それらの単語の寄与度が低くなり，$Conf3(s_h,v)$では候補から外れる．これは，2.1 で説明した共有信念の形成時の学習でロボットが，「箱や小さなぬいぐるみには物がのせられやすい」という信念を獲得していたためである．そのため，それらの場面では「エルモ　のせて」と言っただけで「エルモ　緑色　のせて」と同じ動作を出力することができる．その結果，図 3 で候補になった場面には，エルモや緑色のオブジェクト以外のオブジェクトとして「青い箱」が含まれる傾向が現れた．他の条件と比べると，推定される場面がより限定的になっていると言える．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．

6. 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と考えている．また，扱うオブジェクトの数や特徴の種類が増えた場合に，如何にして計算量を少なくするかについても課題となる．

今後は，信念の形成から場面の推定までを複数の被験者で行い正しい場面を推定できるかを確認する予定である．

本研究の一部は，文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」の援助を受けた。

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).

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

[作元 07] 作元佑輔，木村優志，田口亮，桂田浩一，岩橋直人，新田恒雄: 共有信念に基づく発話状況の推定，人工知能学会研究会資料 SIG-KBS-A702, pp.81-86, (2007).

[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).