3B2-4in1

雑談を通した物体と単語の学習

Learning Objects and Words through Natural Conversation

石田 卓也 *1 Takuya Ishida　山本 一馬 *1 Kazuma Yamamoto　岩橋 直人 *1 Naoto Iwahashi　イェ チョウ トゥ *1 Ye Kyaw Thu　中村 友昭 *2 Tomoaki Nakamura
長井 隆行 *2 Takayuki Nagai　国島 丈生 *1 Takeo Kunishima

*1岡山県立大学 Okayama Prefectural University　*2電気通信大学 The University of Eloctro-Communications

This paper proposes the new approach to symbol grounding research. Previous approaches have been focusing on the learning of physically grounded knowledge without any prior linguistic knowledge, which leads to strained interaction between humans and machines. In contrast, the proposed approach aims to realize the learning through natural conversation. Based on the proposed approach, we show concrete algorithms and their implementation, and conduct experimental evaluation.

1. はじめに

近年，雑談システムの研究が盛んである (e.g. [前田 10, 大西 14, 目黒 14, 木村 15])．雑談を行うコミュニケーションロボットが数多く開発されている (e.g. [ペッパー 14, オハナス 15, ロボホン 16, ユニボ 17])．これらのロボットは，雑談の他，人間の要求に応じて，天気予報を調べたり，電話をかけたりと，さまざまなドメインの対話をすることができる．従来研究のほとんどは，眼前の事物とは関係のない，たとえば「どんな映画が好きですか」といった発話（以降，非接地発話と呼ぶ）によるコミュニケーションをいかに実現するかという問題に焦点を当てたものである．

一方で，眼前の事物に関する発話（以降，接地発話と呼ぶ）の言語表現および概念をロボットに学習させる，言語獲得，記号接地，記号創発に関する研究が盛んである [岩橋 03, Iwahashi 07, Taniguchi 16]．従来研究のほとんどは，記号的知識をまったく持たないロボットがいかに記号的知識を学習するかという問題に焦点を当てたものである．

こうした研究の二極化の流れの中で，人間とロボットの間で，非接地発話によるコミュニケーションが既に成り立っている状態から，接地発話により，言語と眼前の事物との対応付けをいかに学習するかという問題は，少数の研究事例 [Holzapfel 08] があるものの，ほとんど取り扱われてこなかった．実際，従来の雑談システムでは，言葉と眼前の事物との結び付きを学習する対話をすることはできない．たとえば，人間がリモコンをロボットに見せながら「これはリモコンだよ」と教えて，「リモコン」という単語と‘リモコン’という物体を結び付けて学習することができない．人間が発した接地発話を，ロボットが適切に理解するためには，言語情報と実環境情報を統合して処理する必要がある．雑談の中に接地発話と非接地発話が混在する可能性を考える必要がある．雑談システム研究において，接地発話をいかに取り扱うかは重要な課題である．接地発話と非接地発話とが混在する雑談を扱った研究として，人間同士の雑談からロボットに向けられた物体操作指示発話を検出する手法を提案した Zuo らの研究 [Zuo 10] があるが，このような領域の研究が希求されている．

連絡先: 岩橋直人，岡山県立大学情報工学部，岡山県総社市窪木 111，0866-94-2001，iwahashi@c.oka-pu.ac.jp

図 1: 人間とロボットが雑談している様子

そこで我々は，ロボットが人間との雑談を通した自然なインタラクションにより，言語と実世界事物の対応付けを学習できるロボットの実現を目指している．本稿では，接地発話として物体教示発話のみを対象にして，ロボットが人間との雑談の中から，人間の物体教示発話を検出し，その発話が対象としている物体とそれを指示する単語を学習する手法について述べる．

2. 提案手法

2.1 概要

本研究において，ロボットは人間とテーブルを挟んで向かい合った位置に配置され，雑談が行われる．ルールベース型 (e.g.[Wallence 09]) の我々が開発した雑談システム [山本 16] を用いる．雑談中に人間はテーブルの上の物体を把持したり指差しながら，非接地発話に加えて物体教示発話も行うものとする．インタラクションの様子を図 1 に示す．雑談の中から，人間の発話の音声情報と，ロボットのカメラから得られる，人間とテーブル上にある一つ以上の物体を含むシーン画像情報の組が，本稿で提案する手法で用いられるデータとして収集される．

提案手法は，音声情報とシーン画像情報の組の集合を入力データとし，物体と単語の組の集合を出力するものである．本手法は，図 2 に示すように連続する二つのプロセスから構成される．一つ目は物体教示発話検出プロセスである．人間の発話情報と身体動作情報から，発話が物体教示発話であるか否かを

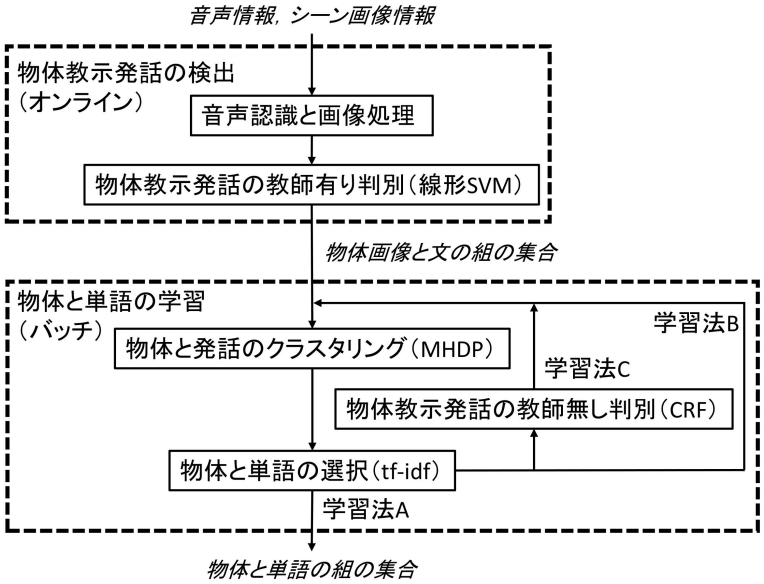

図 2: 提案手法の処理の流れ

判別する．引き続き実行されるもう一つのプロセスは，物体と単語の学習プロセスである．発話がシーン画像情報に含まれる複数の物体の内のどれを指し示すものであるか，同時に，発話中のどの単語がその物体を指し示すものであるかを推測する．以降にそれぞれのプロセスの詳細を記述する．

2.2 物体教示発話の検出

このプロセスは，一発話ごとに物体教示発話か非接地発話かをオンラインで判別するものである．

2.2.1 音声認識と画像処理

発話は大語彙音声認識によりテキストに変換され，さらにテキストが形態素解析される．画像情報は Kinect により取得し，深さセンサー情報と色情報を用いて，画像の中から各物体の画像を切り出す．さらに，人間が物体を把持または指差しているかの情報を抽出する．

2.2.2 物体教示発話の教師有り判別

発話が物体教示発話か非接地発話かの判別は，発話の形態素解析情報，および発話中に人間が物体を把持または指差しているかの情報を入力とした線形 SVM により行う [山本 16]．

2.3 物体と単語の学習

このプロセスでは，物体教示発話検出プロセスにより物体教示発話と判別された発話の集合を用いて，物体とそれを表す単語をバッチ処理で学習するものであり，次の三つの処理から構成される．

物体と発話のクラスタリング: マルチモーダル階層ディリクレ過程（MHDP）[中村 13] を用いることにより，発話情報と物体画像情報の関係性情報を利用し，物体画像と文の組のクラスタリングを行う．

物体と単語の選択: $tf\text{-}idf$ を計算することで，複数の物体と文の中の複数の単語の組み合わせの適切性を評価する．

物体教示発話の教師無し判別: 条件付き確率場（CRF）を用いることにより，発話内のコンテキスト情報を利用し，物体教示発話の教師無し判別を行う．

これらの三つの処理を組み合わせて，図 2 中に示した以下の三通りの学習法を提案する．

学習法 A: 「物体と発話のクラスタリング」と「物体と単語の選択」を行う．

学習法 B: 「物体と発話のクラスタリング」と「物体と単語の選択」を二回繰り返す．

学習法 C: 「物体と発話のクラスタリング」と「物体教示発話の教師無し判別」を行った後，再度「物体と発話のクラスタリング」を行い，最後に「物体と単語の選択」を行う．

2.3.1 物体と発話のクラスタリング

本処理の入力は，切り出し物体画像情報と発話文情報（音声認識単語列情報）の組の集合である．シーン画像中には複数の物体が存在することがあり，各発話が複数の切り出し物体画像情報に対応するので，一つの発話文情報が複数の組に含まれることがあることに注意されたい．この組の集合を MHDP を用いてクラスタリングする．クラスタリングの結果として，切り出し物体画像情報と発話文情報の各組にクラス番号が付与される．

2.3.2 物体と単語の選択

本処理の入力データは，上述したクラスタリングの結果である．このデータでは，各発話が複数の物体画像情報に対応しており，かつ，各発話が複数の単語を含んでいる．本処理では，各クラスに関するすべての単語の $tf\text{-}idf$ を計算する．$tf\text{-}idf$ の計算は以下のとおりである．

$$tf\text{-}idf(w) = tf(w) \times idf(w) \tag{1}$$

$$tf(w) = \frac{\text{クラスの中の単語 } w \text{ の数}}{\text{クラスの中の総単語数}} \tag{2}$$

$$idf(w) = \frac{\log(\text{総クラス数})}{\text{単語 } w \text{ が入っているクラス数}} \tag{3}$$

本処理は学習法 A,B,C で異なった操作を行う．

- 学習法 A では，各発話文に対して，複数の物体のクラスと複数の単語の組み合わせの中から，$tf\text{-}idf$ が最も高い物体と単語の組を選択する．

図 3: 実験に使用した 10 個の物体

- 学習法 B の物体と発話のクラスタリング処理へのフィードバック過程においては，この tf-idf が最も高い物体と単語の組が属している物体画像と文の組の集合を出力する．
- 学習法 C では，物体教示発話の教師無し判別への入力データとして，*tf-idf* が高い 5 位までの物体と単語の組を選択し，ここで選ばれた単語に物体を指示するという意味の「接地」というラベルを与え，その他の単語には「非接地」というラベルを与える．

2.3.3 物体教示発話の教師無し判別

本処理は，学習法 C の場合のみで実行される．本処理の目的は，前段階のプロセスである線形 SVM による物体教示発話検出によって，誤って検出された発話を排除することである．学習データは，観測特徴量として与えられる発話の形態素情報と，物体と単語の選択の処理で与えられた単語のラベル（「接地」または「非接地」）である．まず，このデータで CRF を学習する．その後，同じ観測特徴量を用いてラベリングが行う．このラベリングにおいて，文内の単語のいづれか一つでも「接地」ラベルが与えれた文は物体教示発話と判別され，繰り返し時の物体と発話のクラスタリングの処理では，これらの発話のみが学習に使用される．

3. 実験

提案手法における二つのプロセス，物体教示発話検出プロセスおよび物体と単語の学習プロセスのそれぞれの有効性を実験により評価した．

3.1 物体教示発話の検出

3.1.1 データ

人間とロボットの雑談を通して，人間による 2000 発話の音声情報とシーン画像情報を収録した．そのうち 200 発話が物体教示発話であった．各シーン中の平均物体数は 2.1 であった．雑談の中で，図 3 に示す 10 個の物体が使用された．正解とした画像クラス数は 7 である．音声認識による単語正解精度は 70.8% であった．

3.1.2 評価結果

提案手法による物体教示発話の検出性能を 10 分割交差検定により評価した結果を表 1 に示す．2000 発話の内，物体教示発話であると分類されたのは 196 発話であった．提案手法により，高い性能で物体教示発話の検出が可能であることが示された．

3.2 物体と単語の学習

3.2.1 データ

上記の実験で，物体教示発話であると分類された 196 発話と，それらに対応した物体画像情報の組（417 組）を学習データとした．各発話の音声認識結果において，物体を指示する単語が正しく認識されていた割合は，35.7% であった．

表 1: 物体教示発話の検出性能 (%)

Precision	Recall	F-measure
87.3	88.2	87.6

表 2: 学習された物体と単語の精度（%）

学習法	P_w	P_c	P_{wc}
A	31 (61/196)	30 (59/196)	10 (19/196)
B	35 (69/196)	57 (112/196)	28 (54/196)
C	46 (66/144)	63 (90/144)	34 (49/144)

3.2.2 評価結果

提案手法の性能を，物体教示発話と分類された文の中で，各発話において正しい単語が選択された確率 P_w，各発話において正しい画像クラスが選択された確率 P_c，各発話において正しい単語と画像クラスが選択された確率 P_{wc} によって評価した．学習法 A，B，C の評価結果を表 2 に示す．性能は，A<B<C の順で高くなっていることが分かる．さらに，CRF による物体教示発話の教師無し判別の評価結果を表 3 に示す．Precision が高いことから，線形 SVM での記号接地発話の誤検出を，ある程度排除できていることが見て取れる．しかし，Recall がそれほど高くないことから，記号接地発話も同時に誤って排除してしまっていることも分かる．したがって，物体教示発話の教師無し判別法の効果が明確に示せたとは言えない．

4. 考察

提案手法の第一段階のプロセスにより，高い物体教示発話検出法性能が得られたが，まだ十分とは言えない．性能向上のためには，まず，単語正解精度が 70.8% と低かった音声認識の性能を向上させる必要がある．また，物体と単語の学習実験においても，学習に用いた音声認識結果において物体を指示する単語が正しく認識されていた割合が 35.7% と低かったことがが，最終的に高い性能が得られなかった最大の要因である．本実験において物体を指示する単語の認識性能が，従来の言語獲得や記号創発研究で用いられている音韻認識（音韻認識率約 80%) による単語認識性能に比べて，十分に高いと言えない結果になってしまったことから，物体と単語の学習において，大語彙音声認識と音韻認識のそれぞれの強みと弱みが何であるのかを，より詳細な実験を行うことで整理する必要があると考えられる．

また，物体教示発話の検出が，一段目のプロセスにおける線形 SVM を用いた処理と，二段目のプロセスにおける CRF を用いた処理の二箇所で行われているが，これらを効率良く統合することで，より CRF の性能を発揮させることができるかもしれないと考えられる．

さらに，一段目のプロセスで用いていた人間による物体の把持と指差しの情報を，二段目のプロセスでも利用可能にする

表 3: 物体教示発話の教師無し判別性能 (%)

Precision	Recall	F-measure
95 (137/144)	78 (137/175)	86

ことで，大きな性能向上が期待できる．

最後に，提案手法では，二段目のプロセスがバッチ処理になっているのが，これをオンライン処理にすることが課題である．

5. まとめ

雑談を通した物体と単語の学習という，従来まったくなかった新しい発想の研究アプローチを提案した．本アプローチに基づいた具体的なアルゴリズムを提示，実装し，実験により評価した．今後は，より詳細な評価と改良を行ってゆく予定である．

謝辞

本研究は，JSPS 科研費 15K00244，および，JST CREST「記号創発ロボティクスによる人間機械コラボレーション基盤創成」の助成を受けて実施したものである．

参考文献

[大西 14] 大西可奈子, 吉村健: コンピュータとの自然な会話を実現する雑談対話技術: NTT DOCOMO テクニカル・ジャーナル, Vol. 21, No. 4, pp. 17-21, (2014)

[オハナス 15] www.takaratomy.co.jp/products/omnibot/ohanas/

[木村 15] 木村泰知, ジェプカラファウ, 高丸圭一: Radiobots 型対話システムの提案, 人工知能学会全国大会発表論文集, (2015)

[目黒 14] 目黒豊美, 杉山弘晃, 東中竜一郎, 南泰浩: ルールベース発話生成と統計的発話生成の融合に基づく対話システムの構築, 人工知能学会全国大会発表論文集, (2014)

[Holzapfel 08] Holzapfel, H., Neubig, D. and Waibel, A.: A dialogue approach to learning object descriptions and semantic categories, Robotics and Autonomous Systems, Vol. 56, Issue 11, pp. 1004-1013, (2008)

[岩橋 03] 岩橋直人: ロボットによる言語獲得 －言語処理の新しいパラダイムを目指して－, 人工知能学会誌, Vol. 18, No. 1, pp. 49-58, (2003)

[Iwahashi 07] Iwahashi, N.: Robots That Learn Language: Developmental Approach to Human-Machine Conversations, Human Robot Interaction (N. Sanker, Ed.), pp. 95-118, (2007)

[前田 10] 前田和希, 宋, 國政裕友樹, 豊田博之, 韓東力: 雑談システムにおける話題展開の性能向上, 言語処理学会第 16 回年次大会発表論文集, pp. 250-253, (2010)

[中村 13] 中村友昭，荒木 孝弥，長井隆行，岩橋直人，”階層ディリクレ過程に基づくロボットによる物体のマルチモーダルカテゴリゼーション”，計測自動制御学会論文集, pp.469-478, Vol. 49, No. 4, (2013)

[ペッパー 14] www.softbank.jp/robot/

[Wallence 09] R. S. Wallace: The Anatomy of A.L.I.C.E., Parsing the Turing Test, pp. 181-210, (2009)

[ロボホン 16] robohon.com/

[Taniguchi 16] Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T and Asoh, H.: Symbol Emergence in Robotics: A Survey, Advanced Robotics, Vol. 30, Issue 11-12, pp. 706-728, (2016)

[ユニボ 17] www.unirobot.com/

[山本 16] 山本一馬, 石田卓也, 岩橋直人, Ye Kyaw Thu, 国島丈生: 人とロボットによる雑談から眼前の物体に記号接地している発話の検出法, HAI シンポジウム予稿集, (2016)

[Zuo 10] Zuo, X., Iwahashi, N., Funakoshi, K., Nakano, M., Taguchi, R., Matsuda, S., Sugiura, K. and Oka, N.: Detecting Robot-Directed Speech by Situated Understanding in Physical Interaction, 人工知能学会論文誌, Vol. 25, No. 6, pp. 670-682, (2010)