複数物体が存在する環境下での共同注意を用いた ロボットによる語意学習

長野 匡 隼* 中村 友 昭*

Learning Word Meanings Using Joint Attention and MLDA in Environments with a Plurality of Objects

Masatoshi Nagano* and Tomoaki Nakamura*

Humans can learn word meanings by associating objects with words even in an environment with a plurality of objects by using joint attention, which is an ability to detect a target object that others pay attention to. In this paper, we propose a method for robots to learn word meanings using joint attention and co-occurrence of objects and words, which is modeled by multimodal latent Dirichlet allocation (MLDA). A target object is detected by using MLDA and joint attention, and MLDA is updated by the detected object. This updated MLDA can improve the accuracy of the target object detection.

Key Words: Joint Attention, Word Meaning Acquisition, Unsupervised Learning, Multimodal Latent Dirichlet Allocation

1. は じ め に

人は、複数の物体が存在している環境の中で、養育者などから発せられる単語と物体を結びつけて、物の名前を学習している。その際に、共同注意が語意の獲得を促進していることが指摘されている [1]。共同注意とは養育者の顔や視線の向き等から、見ている対象を特定する能力のことである。共同注意によって養育者が注目している物体を特定することができ、養育者が発している単語と物体を結びつけることが可能となる。それにより、その単語が指示する範疇である概念を学習している。

一方で、ロボットによる語意の学習に関する研究も多く行われているが、学習環境が限定されているという問題があった。従来の語意学習の研究では、ロボットに物体を教示する際に、ロボットに提示される物体は容易に物体領域を抽出できる環境や [2]、また、単語の指示対象の物体（指示対象物体）が一意に定まるような環境を想定していた [3]。すなわち、これらの手法では、複数の物体が乱雑に置かれた環境（複数物体環境）において、指示対象物体を一意に定めることができず、また容易に物体領域を抽出できるとは限らないため、語意の学習が困難である。そこで本稿では、共同注意と特徴の共起性を用いることで、複数物体環境下で単語の指示対象物体を推定し、語意を学習する手法を提案する。提案手法では、共同注意と物体領域検出器により、ロボットが取得した物体領域の候補から Multimodal Latent Dirichlet Allocation (MLDA) [4] を用いて一つを選択する。選択された画像と教示発話の特徴量の共起性に基づいて MLDA によりカテゴリ分類することで、画像と単語を結びつけ、語意

を学習する。この指示対象物体の選択とカテゴリ分類を繰り返し、MLDA のパラメータを更新することで、画像選択とカテゴリ分類の精度を向上させる。このように提案手法では、確率モデルである MLDA を利用することで、学習プロセスを確率的に定式化している。一方で、深層生成モデル [5] を利用することも考えられるが、この定式化をネットワークの構造としてそのまま組み込むことは困難であり、また学習には多くのデータを要するといった問題がある。

2. 提 案 手 法

本稿では、ロボットが学習対象の物体が存在する複数のシーンを観測し、人がそのシーンに含まれる物体名を教示することを想定している。また、ロボットは共同注意能力と、物体の領域を検出する能力はすでに獲得していることを想定している。本稿では、ロボットが教示者の顔の向きを推定し、教示者が注意を向けている物体方向にカメラを向ける機能を共同注意能力とした。また、汎用的な物体領域検出器により、画像から物体と予測される領域を複数検出する機能を物体検出能力とした。そのため、顔向き推定の誤差により必ずしも指示対象物体が正確にシーンの中心には存在するとは限らず、また、指示対象物体の一部の領域や異なる物体の領域が検出されることも起こり得る。ただし、シーン内に指示対象物体が存在しないなどの共同注意が不完全な状況は想定していない。

Fig. 1 が提案手法の概要である。初めに、教示者は指示対象物体の方向に顔を向け、物体名を含む文章（教示文）を発話する。単語情報としては、ロボットが観測しているシーン s に対して、指示対象物体の名称を含む教示文（例：「ぬいぐるみがあるよ」）を用いる。この教示文を Bag of Words (BoW) 表現に変換して得られる単語発生頻度ヒストグラムを、シーン s の単語情報の特徴量 w_s^w とする。

原稿受付 2020 年 8 月 18 日

*電気通信大学

*The University of Electro-Communications

■ 本論文は有用性（要素分野）で評価されました。

Fig. 1 Overview of the proposed method

画像情報は、教示者の顔向きを dlib [6] により計算することで指示対象物体の位置を推定し、その方向へカメラを向け、シーン s を撮影する。シーン s に対して、汎用的な物体領域検出器を用いて、物体候補領域画像を M_s 枚抽出する。本稿では、汎用的な物体領域検出器には、学習済みの Single Shot MultiBox Detector (SSD) [7] を用いる。SSD は、その学習データに含まれている物体を検出する手法であるが、ここでは学習物体以外の様々な物体を検出できるようパラメータを調整することで、汎用的な物体領域検出器として利用している。なお、このようにパラメータ調整をすることで、物体領域以外にも多く含まれてしまうが、後述する MLDA を用いる画像選択により、適切な物体領域を選択することが可能である。また、ロボットが指示対象物体方向を向いていることを仮定すると、指示対象物体は画像の中央付近に写っている可能性が高い。そこで、検出された物体の画像上の位置 $\mathbf{x}$ がシーン画像の中心位置 $\boldsymbol{\mu}$ を平均としたガウス分布から生成される確率 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ と、検出器から出力される各領域の確信度の積を計算し、その値が高い上位 N_s 枚の画像を指示対象物体の候補とする。限定された N_s 枚の画像を学習済みの Convolutional Neural Network (CNN) に入力し得られる中間層の 4,096 次元の出力をシーン s の n 番目の候補の特徴量 $\mathbf{w}_{s,n}^v$ として用いる。次に、同じ単語が教示されている指示対象物体の画像は、似た特徴量を含んでいる可能性が高いことを利用して、限定された候補から MLDA により画像 1 枚 $\hat{\mathbf{w}}_s^v$ を選択する。選択された画像と教示発話を MLDA によりカテゴリに分類し、このカテゴリと単語が結びつくことで語意を獲得する。

以上のように、候補からの指示対象物体の選択とカテゴリ分類を繰り返すことで、MLDA のパラメータを更新し、画像選択とカテゴリ分類精度を向上させる。

2.1 MLDA によるカテゴリ分類

画像と単語の結びつけと画像の選択には MLDA を用いる。MLDA のグラフィカルモデルが Fig. 2 である。灰色で示されているノードは可観測ノードを表している。Fig. 2 の w^v , w^w はそれぞれ画像情報、単語情報を示している。 $w^m (m \in \{v, w\})$ は β^m をパラメータとする多項分布から発生する。これらのパラメータは、それぞれ π^m をパラメータとするディリクレ分布

Fig. 2 Graphical model of MLDA

に従う。また z は画像と単語から構成される物体のカテゴリであり、カテゴリ z の出現確率を表す多項分布のパラメータを θ とする。 α は θ の事前分布であるディリクレ分布のパラメータである。MLDA のパラメータ推定にはギブスサンプリング [8] を用いる。ギブスサンプリングでは、 s 番目の物体のモダリティ m の i 番目の情報に割り当てられるカテゴリ z_{mis} を、 θ と β^m を周辺化した式 (1) の条件付確率からサンプリングする。

$$P(z_{mis} = k | z^{-mis}, w^m, \alpha, \pi^m) \propto (N_{ks}^{-mis} + \alpha) \frac{N_{mw^m k}^{-mis} + \pi^m}{N_{mk}^{-mis} + D^m \pi^m} \quad (1)$$

ただし、 D^m はモダリティ m の情報の次元数である。また、 $N_{mw^m k}$ は w^m にカテゴリ k が割り当てられた回数、 N_{ks} は s 番目のすべてのモダリティ情報にカテゴリ k が割り当てられた回数、 N_{mk} はモダリティ m の情報にカテゴリ k が割り当てられた回数である。これらは、 s 番目の物体のモダリティ情報 w^m にカテゴリ k が割り当てられた回数を $N_{mw^m k s}$ として、以下のように計算できる。

$$N_{mw^m k} = \sum_s N_{mw^m k s} \quad (2)$$

$$N_{ks} = \sum_{m, w^m} N_{mw^m k s} \quad (3)$$

$$N_{mk} = \sum_{w^m, s} N_{mw^m k s} \quad (4)$$

また、式 (1) の負の添字は、その情報を除外した残りを表している。つまり、 z^{-mis} は、 s 番目の物体のモダリティ m の i 番目の情報へ割り当てられたカテゴリ z_{mis} を取り除いた残りである。式 (1) のサンプリングを尤度が収束するまで繰り返し、式 (2)~(4) の N_* を算出する。最終的に、パラメータの推定値 $\hat{\beta}_{w^m k}^m$, $\hat{\theta}_{ks}$ は以下のように計算される。

$$\hat{\beta}_{w^m k}^m = \frac{N_{mw^m k} + \pi^m}{N_{mk} + D^m \pi^m} \quad (5)$$

$$\hat{\theta}_{ks} = \frac{N_{ks} + \alpha}{\sum_{k'} N_{k' s} + K \alpha} \quad (6)$$

2.2 MLDA に基づく画像選択

MLDA では、カテゴリ分類することで、可観測のモダリティ情報からほかのモダリティ情報を予測することができる。つまり、単語情報から、それが表している画像情報を予測することができる。これにより、指示対象物体の N_s 個の候補から、単語が表している物体画像を選択する。シーン s に与えられた単語情報 w^w から、画像情報 w^v が発生する確率は式 (7) となる。

$$P(w^v | w^w) = \int \sum_z p(w^v | z) p(z | \theta) p(\theta | w^w) d\theta \quad (7)$$

ここで、 $p(\theta | w^w)$ は、学習によって得られた $\hat{\beta}_{w^m k}^m$ を固定し、

Algorithm 1 Object detection and categorization by MLDA

```

1: for  $l = 1, \dots, L$  do
2:   for  $s = 1, \dots, S$  do
3:      $\Theta_{-s} = MLDA(W_{-s}^v, W_{-s}^w)$ 
4:      $\hat{w}_s^v \sim P(w_s^v \in \{w_{s,n=1}^v, \dots, w_{s,n=N_s}^v\} | w_s^w, \Theta_{-s})$ 
5:      $W^v[s] = \hat{w}_s^v$ 
6:   end for
7: end for
8:  $\Theta = MLDA(W^v, W^w)$ 

```

w^w に対して式 (1) を用いてサンプリングすることで求めることができる。この確率を利用し、シーン s の教示文 w_s^w が示している指示対象物体の画像 $\hat{w}_s^v$ を N_s 枚の候補からサンプリングすることで決定する。

$$\hat{w}_s^v \sim P(w_s^v \in \{w_{s,n=1}^v, w_{s,n=2}^v, \dots, w_{s,n=N_s}^v\} | w_s^w, \Theta_{-s}) \quad (8)$$

ただし、 Θ_{-s} は、シーン s に関する情報を除いて学習した MLDA のパラメータである。

ギブスサンプリングに基づいた MLDA のパラメータ推定 (2.1 節) と画像選択 (式 (8)) を用いた提案アルゴリズムを Algorithm 1 に示す。ただし、 W^v は候補から選択した画像の特徴量 $W^v = \{\hat{w}_{s=1}^v, \hat{w}_{s=2}^v, \dots, \hat{w}_{s=S}^v\}$ であり、 W^w は各シーンに対して与えられた教示文の単語情報 $W^w = \{w_{s=1}^w, w_{s=2}^w, \dots, w_{s=S}^w\}$ である。 $\Theta = MLDA(W^v, W^w)$ はこれらの情報から MLDA のパラメータ Θ を推定する処理を表している。このように、MLDA のパラメータ推定と、画像の選択を尤度が収束するまで L 回繰り返すことで、画像の選択精度と分類精度を高めることができる。

3. 実験

提案手法の有効性を検証するために、ロボットから指示対象物体までの距離を限定せず、人とロボットが共同注意を行い取得したデータセットを用いて実験を行った。

3.1 データセット

本実験では、共同注意を行うロボットとしてロボットアーム (mover4[†]) のハンド部に RGB-D カメラ^{††} を取り付けてシーンの画像データを取得した。

学習する物体として、Fig. 3 のぬいぐるみとペットボトルをそれぞれ 5 個、お菓子と時計をそれぞれ 6 個、コップ 4 個の合計 26 個の物体のいずれかが含まれている 92 シーンを用いた。実験に用いたシーンの一例を Fig. 4 に示す。シーンの撮影は以下の手順に従い、研究室にて行った。また、すでに置いてある物体は移動させることなく、教示したい物体を含む複数の物体を Fig. 3 からランダムに選択、配置した。

- (1) ロボットが顔向きを推定できる位置に教示者が立つ。
 - (2) 教示者がロボットに搭載されたカメラと向き合い、ロボットが教示者の顔向きを推定。
 - (3) 教示者が指示対象物体方向を向き、その顔向きをロボットが推定。
 - (4) ロボットが教示者の顔向きの変化量を基に、指示対象物体方向にカメラを動かしシーンを撮影。
- 一つのシーンにつき 10 枚 ($N_s = 10$) の指示対象物体の候補を抽出した。また、各シーンに対して、教示者が指示対象物体の名称を含む教示文を Table 1 から一つ与えた。

Fig. 3 Target objects

Fig. 4 Example of scenes

Table 1 Teaching sentences given to target objects (The experiments were conducted in Japanese). ‘/’ represents the boundary between the words

ぬいぐるみ/が/ある/よ	ペットボトル/が/ある/よ
お菓子/が/ある/よ	コップ/が/ある/よ
時計/が/ある/よ	

3.2 評価方法

本実験では、物体領域検出器の確信度の上位 10 個の検出領域を候補とした手法 (W/O joint attention) と、2 章の方法により共同注意を用いて検出した上位 10 個の検出領域を候補とした手法 (W/ joint attention) を比較した。また、それぞれの手法で、候補 N_s から 1 枚の画像をランダムに選択した場合 (Random) と、式 (8) により選択した場合 (Sampling) を比較した。評価指標には、画像の選択精度とカテゴリ分類精度の二つを用いた。

画像選択の成否は、シーン s において各手法で選択された物体領域を O_s 、人手で決めた正解領域を $\hat{O}_s$ としたとき、次式の F 値 F_s が 0.4 以上のとき、選択された領域を正解と定義した。

$$P_s = \frac{|O_s \cap \hat{O}_s|}{|\hat{O}_s|}, R_s = \frac{|O_s \cap \hat{O}_s|}{|O_s|}, F_s = \frac{2P_s R_s}{P_s + R_s} \quad (9)$$

画像選択精度は、各手法において、92 個のシーンのうち正解領域を選択できた割合である。

カテゴリ分類精度は、92 シーンで選択された画像が、正しいカテゴリに分類された割合である。ただしカテゴリ分類精度は、画像選択の成否は考慮せず算出した。

3.3 結果・考察

Table 2 が画像の選択精度とカテゴリ分類精度である。共同注意の有無で比較すると、画像の選択方法にかかわらず、「W/ joint attention」の場合が 2 倍近く「W/O joint attention」より高い画像選択精度となった。これは、 N_s の候補に指示対象物体が含まれている可能性が高くなっていることを意味しており、共同注意が有効に働いていることを示している。また、画像の選択方法を比較すると、「Sampling」による画像選択を行ったときが、「Random」による選択の 2 倍近い精度となった。すなわち、提案手法により、単語情報と画像情報を結びつけることで、指示対象物体を適切に候補から選択できていることが分かる。

カテゴリ分類精度に関しても、ランダムにカテゴリ分類した

[†]<https://cpr-robots.com/education#lightbox/2/>

^{††}<https://www.intelrealsense.com/depth-camera-d415/>

Table 2 Accuracy of the target object detection and categorization

Use of joint attention	Selection methods of image	Accuracy of image selection	Accuracy of categorization
W/ joint attention	Sampling	0.95	0.96
W/ joint attention	Random	0.44	0.39
W/O joint attention	Sampling	0.58	0.49
W/O joint attention	Random	0.23	0.39

Fig. 5 Region of target objects (Green) and estimated objects (Red) **Fig. 6** Transition of categorization accuracy by parameter update

ときに比べ、MLDA を用いて画像選択・分類したときの方が高い精度となった。また、「W/ joint attention」によるカテゴリ分類の精度が、「W/O joint attention」より高いことが確認できる。このことから、共同注意により指示対象物体を含んだ適切な画像が選択されやすくなったことで、カテゴリ分類の精度が向上したと考えられる。

「W/ joint attention」で「Sampling」による画像選択を行った際に、候補から選択された領域の一例を **Fig. 5** に示す。緑枠が指示対象物体の正解の領域であり、赤枠が提案手法により選択された領域である。Fig. 5 (a) は指示対象物体がぬいぐるみであり、カメラから離れた小さい領域であっても、指示対象物体を含む正しい領域が選択できている。また、Fig. 5 (b), (c) のシーンのように物体が多く存在し、既存手法では指示対象物体の特定が難しい環境でも、正しく選択できていることが確認できる。

さらに、**Fig. 6** は、画像選択とカテゴリ分類の繰り返し回数ごとの、カテゴリ分類の精度を示したグラフである。図の青線は「W/ joint attention-Sampling」の結果である。橙色の線は物体領域検出器を使わず正解領域 $\hat{O}_s$ を入手で切り出した画像と単語情報を用いて MLDA によりカテゴリ分類した結果であり、提案アルゴリズムにおける分類精度の上限値と考えることができる。この結果から、画像選択とカテゴリ分類を相互に繰り返すことによって徐々に分類精度が向上し、この上限の精度に近づいており、提案手法が適切に機能していることが分かる。

これらの結果より、共同注意とモダリティの特徴の共起性を用いることによって、複数物体環境から正しい指示対象物体を選択し、語意の学習が可能であることが示された。

4. ま と め

本稿では、人が言語獲得に利用している共同注意をロボットが使用することで、複数物体環境下で指示対象物体を特定し、語

意の学習が可能であることを示した。従来の研究では、本稿のような複数物体環境は扱わず、各シーンで検出される物体領域は 1 個 ($N_s = 1$) であり、それを Z 個のカテゴリに分類する問題を扱っていた。すなわち、シーン数を S とすると、 Z^S 個の可能性の中から一つを選択する問題であった。一方で本稿では、物体領域候補を複数個 ($N_s = 10$) 検出し、その選択も行っている。すなわち、 $(N_s Z)^S$ 個の可能性の中から一つを選択する問題であり、従来に比べその複雑さが向上している。この問題に対して、従来の特徴の共起性だけを頼りとした方法 (W/O joint attention-Sampling) では、高い精度は達成できなかった。一方で、共同注意も用いた提案手法 (W/ joint attention-Sampling) では精度が大きく向上しており、共同注意の重要性を工学的に示唆する結果となった。

実験では、共同注意として人の顔向きのみを用いて、ロボットが指示対象物体へと注意を向けた。しかし、共同注意には、顔向きだけではなく、視線や指差しなどが存在する。また、本実験とは逆に、ロボットが注意を向けている物体に対して人が注意を向ける共同注意もある。今後、これらを導入することで、より複雑でインタラクティブな学習を実現することを考えている。さらに、実験では、教示文をテキストで与えたが、言語モデルの学習 [9] と統合することで、連続音声と画像から言語と物体概念を学習する手法へと拡張することも考えている。

また本稿では、共同注意の失敗により、シーン内に指示対象物体が存在しない状況は想定していなかった。しかし、実際の語意学習ではそのような状況は起こり得ると考えられる。そこで、このような共同注意が不完全な環境でもロボットが柔軟に学習を可能とする方法 [10] との統合も検討している。

謝 辞 本研究は JSPS 科研費 16H02835 の助成を受け実施したものである。

参 考 文 献

- [1] M. Tomasello and M.J. Farrar: "Joint attention and early language," *Child development*, pp.1454–1463, 1986.
- [2] M. Shridhar and D. Hsu: "Grounding spatio-semantic referring expressions for human-robot interaction," *Robotics: Science and Systems, Workshop on Spatial-Semantic Representations in Robotics*, 2017.
- [3] 石田卓也, 山本一馬, 岩橋直人, Ye Kyaw Thu, 中村友昭, 長井隆行, 國島文生: "雑談を通じた物体と単語の学習", 人工知能学会全国大会, 3B24in1, 2017.
- [4] T. Nakamura, T. Araki, T. Nagai and N. Iwahashi: "Grounding of word meanings in latent dirichlet allocation-based multimodal concepts," *Advanced Robotics*, vol.25, no.17, pp.2189–2206, 2011.
- [5] M. Suzuki, K. Nakayama and Y. Matsuo: "Joint multimodal learning with deep generative models," *International Conference on Learning Representations, Workshop track*, 2017.
- [6] D.E. King: "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol.10, pp.1755–1758, 2009.
- [7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A.C. Berg: "SSD: Single shot multibox detector," *European conference on computer vision*, pp.21–37, 2016.
- [8] T.L. Griffiths and M. Steyvers: "Finding scientific topics," *Proc. National academy of Sciences*, vol.101, no.suppl 1, pp.5228–5235, 2004.
- [9] 中村友昭, 長井隆行, 船越孝太郎, 谷口忠大, 岩橋直人, 金子正秀: "マルチモーダル LDA と NPYLM を用いたロボットによる物体概念と言語モデルの相互学習", 人工知能学会論文誌, vol.30, no.3, pp.498–509, 2015.
- [10] 笹本勇輝, 吉川雄一郎, 浅田稔: "ロボットの語意学習のための主観的整合性に基づくマルチモーダルカテゴリゼーション", 人工知能学会論文誌, vol.29, no.5, pp.436–448, 2014.