2I3-3

# 階層ディリクレ過程を用いたロボットによる概念と言語の長期学習

西原 成 Joe Nishihara　青木 達哉 Tatsuya Aoki　中村 友昭 Tomoaki Nakamura　長井 隆行 Takayuki Nagai

電気通信大学大学院 情報理工学研究科
Faculty of Infomatics and Engineering, The University of Electro-Communications

In this paper, we propose a method which enables robots to learn object concepts, word meanings and language model in a bottom up manner. That robots simultaneously learn concepts and language, which improves performances of multimodal categorization and speech recognition each other. Moreover, in this paper, we conducted long term learning experiment. The robot learned a lot of objects and words with a human for a month. We examine our method through the long-term experiment, and discuss how the robot was learning concepts and word meanings.

## 1. 緒言

人間の認知において，事物のカテゴリ分類は非常に重要である．著者らはロボットに関してもカテゴリ分類を通じて概念を形成し，これら概念を利用することで，人間と同様に時空間的予測や言語の理解が可能となると考えている．これまで，中村らは自然言語処理で提案された Latent Dirichlet Allocation (LDA)[Blei 03] を，視覚，聴覚，触覚情報や，人から与えられた教示発話 (単語情報) などのマルチモーダル情報を用いた Multimodal LDA (MLDA) へ拡張することで，ロボットが人に近い概念を獲得でき，さらに自身のセンサ情報と単語情報を形成した概念に結びつけることによって，ロボットによる語意の理解が可能となることを示してきた [Nakamura 09]．

しかし，これまでの研究には大きな問題が存在している．一つに音声認識の問題がある．これまでの研究では，ロボットが単語情報を得る際に学習済みの音声認識器や形態素解析器を用いてきた．これは暗にロボットが辞書や言語モデルといった言語的知識を持った状態で概念形成と語意の接地を行ってきたことを意味しており，厳密な意味での語意獲得を行っていないことになる．しかしながら，ロボットがこれら言語的知識を持たずに音声認識を行った場合，その認識結果には多くの音素誤りが含まれてしまう．これら音素誤りはロボットの概念形成に大きな悪影響を及ぼすことが過去の研究によって明らかになっており，音声認識をどのように行うかは重要な問題となる．二つ目の問題として，カテゴリ数決定の問題がある．従来研究で用いられてきた MLDA では，人が事前にロボットが形成する概念 (カテゴリ) の数を与える必要があるが，実環境に存在する物体は多種多様であり，学習前に人が正解のカテゴリ数を与えることは困難である．さらに，今まで行われてきた研究は，ロボットが学習する物体が数十という規模であり，モデルのスケーラビリティに関しては議論されていなかった．ロボットにおいてもより多くの物体を理解し概念を形成する必要があり，スケーラビリティに関する議論が重要である．

本稿では一点目の問題に対しては，学習の過程でロボットが，人から与えられた発話を元に言語モデルを自ら獲得していくことにより，音声認識精度を向上させることを考える．しかしながら，より良い言語モデルを獲得するためには，より良い音声認識結果を得る必要があり，これは鶏と卵の問題である．そこで，ロボットは既に獲得した物体概念を足がかりとして言語モデルを計算することで，徐々に精度の良い言語モデルの獲得を行う．本稿ではこれを “言語と概念の相互学習モデル” と呼ぶ．次に，カテゴリ数の問題を解決し，より柔軟な概念学習を行うために online Multimodal Hierarchical Dirichlet Process (oMHDP) と呼ばれるノンパラメトリックベイズを用いた学習モデルと，上述の言語と概念の相互学習モデルとを組み合わせる．oMHDP は学習データの複雑さに応じてカテゴリ数を逐次的に推定可能なモデルであり，ロボットは事前にカテゴリ数を与えることなく逐次的に概念を形成することが可能となる．

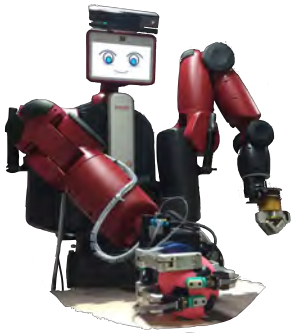

図 1: 本研究で用いるロボットプラットフォーム

本研究では，提案モデルを搭載したロボットと一ヶ月 (百時間以上) の長期間に渡って，実環境において学習を行うことで提案モデルの性能やスケーラビリティを検証する．さらに，長期間の学習におけるロボットの概念形成や語意学習の様子を観察することで，今までの実験規模では見られなかった，ロボットの学習過程において生じた様々な現象を分析する．

## 2. 提案手法

### 2.1 マルチモーダル情報

本研究では，図 1 に示した双腕ロボット Baxter と，ロボットに取り付けた各種センサを用いてマルチモーダル情報の取得を行う．本研究ではこの Baxter に Barrett Hand を取り付けることによって，物体を握る・掴むという動作を行う．また，この Barrett Hand の指先に取り付けた触覚センサにより，物体を握った際の触覚情報を，手先に取り付けたマイクロフォンを用いて，物体を振った際に生じる音情報を取得する．また，頭部には RGBD センサが取り付けられており，これによって物体の位置の検出や画像情報の取得を行う．

連絡先: 西原 成, 電気通信大学大学院情報理工学研究科, 東京都調布市調布ケ丘 1-5-1, joe@apple.ee.uec.ac.jp

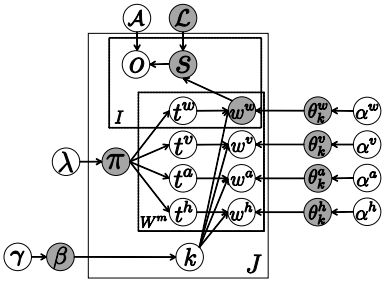

図 2: 言語と概念の相互学習モデル

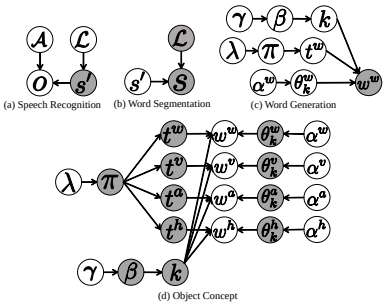


図 3: 言語と概念の近似モデル

本研究で用いるマルチモーダル情報は，視覚情報として頭部に搭載した RGBD センサより取得した物体画像の Convolutional Neural-Network(CNN) の中間層 4096 次元の出力を画像特徴量として扱う．触覚情報は物体を握ったときに生じる触覚センサの値を 15 次元にベクトル量子化したものを，聴覚情報は物体を振った時に生じる音情報の MFCC を 50 次元にベクトル量子化したものを用いる．また，単語情報は節 2.2 で説明する言語と概念との相互学習モデルにより計算される単語情報を用いる．

## 2.2 言語と概念の相互学習モデル

図 2 が，本研究で提案する言語と概念の相互学習モデルのグラフィカルモデルである．灰色で示されているノードが，未観測情報を表している．図中の $o$ が入力される教示発話であり，この音声を $\mathcal{A}$ をパラメータとする音響モデルと $\mathcal{L}$ をパラメータとする言語モデルによって認識した結果が $s$ である．さらに，認識結果 $s$ を BoW 表現へと変換したものが単語情報 $w^w$ となる．つまり，図 2 は，認識結果 $s$ と物体カテゴリ $k$ が単語 $w^w$ によって接続されたモデルになっており，音声認識結果と物体概念形成が相互に影響するモデルとなっている．言語モデル獲得・物体概念形成は可観測ノードである音声 $o$ と視覚・聴覚・触覚情報である $w^v, w^a, w^h$ からパラメータ $\mathcal{L}, \theta^*, \pi, \beta$ を推定し，隠れ変数である認識結果 $s$，単語情報 $w^w$，物体カテゴリ $k$ を決定することで可能となる．

### 2.2.1 近似モデルによる学習

図 2 の未観測ノードのパラメータを推定することにより，ロボットは言語モデルと概念を同時に求めることができる．しかしながら，このモデルは複雑であり，また，推定すべきパラメータ数が非常に多いため，一度に全てのパラメータを推定することは困難である．そこで，このモデルを 4 つのモデルへと分割し，各モデルのパラメータを逐次推定することにより，言語モデルと概念を相互に学習する．分割した各モデルは図 3 のようになる．図 2 と同様に，図中の灰色のノードは未観測情報を示しており，以下の手順を繰り返すことで各パラメータの推定を行う．

**音声認識**

図 3(a) が音声認識のモデルである．ここでは音響モデルのパラメータ $\mathcal{A}$ と，言語モデルのパラメータ $\mathcal{L}$ は既知であるとし，各物体に与えられた全教示発話 $o$ から，$n$-best の認識音素列 $\boldsymbol{s}'_{1:N}$ を得る．

$$\boldsymbol{s}'_{1:N} \sim P(\boldsymbol{s}'_{1:N}|O, \mathcal{A}, \mathcal{L}) \tag{1}$$

本研究では音素認識器として Julius を用い，Julius 標準の音響モデルを用いた．

**単語の分節化**

続いて，言語モデルのパラメータ推定を行う．言語モデルのパラメータ $\mathcal{L}$ は，ある物体に対して与えられた教示発話 $\boldsymbol{o}$ を生成する確率 $P(\boldsymbol{o}|\mathcal{A}, \mathcal{L})$ を最大化することで，次式より計算できる．

$$\mathcal{L} = \underset{\mathcal{L}}{\operatorname{argmax}} P(\boldsymbol{o}|\mathcal{A}, \mathcal{L}) \tag{2}$$

$$= \underset{\mathcal{L}}{\operatorname{argmax}} \int P(\boldsymbol{s}|\mathcal{L}) P(\boldsymbol{o}|\boldsymbol{s}, \mathcal{A}) d\boldsymbol{s} \tag{3}$$

しかし，$\boldsymbol{s}$ での積分は，全ての文字の組み合わせの和を取ることを意味しており，直接計算することができない．そこで，ここでは $P(\boldsymbol{o}|\boldsymbol{s}, \mathcal{A})$ は一部の文字列以外の確率は非常に小さく無視できるとして，$\boldsymbol{o}$ の $n$-best の認識文字列 $\boldsymbol{s}'_{1:N}$ のみを用い，図 3(b) のように音声認識モデルと切り離して考える．つまり，教示発話 $\boldsymbol{o}$ を生成する確率を最大とする代わりに，教示発話 $\boldsymbol{o}$ を認識した $n$-best の認識文字列 $\boldsymbol{s}'_{1:N}$ から単語列 $\boldsymbol{s}_{1:N}$ を生成する確率を最大とすることで，言語モデルパラメータ $\mathcal{L}$ を近似的に計算する．

$$\mathcal{L}, \boldsymbol{s}_{1:N} = \underset{\mathcal{L}, \boldsymbol{s}_{1:N}}{\operatorname{argmax}} P(\boldsymbol{s}_{1:N}|\boldsymbol{s}'_{1:N}, \mathcal{L}) \tag{4}$$

本稿では，[Araki 13] によって提案された逐次学習に対応した教師なし形態素解析である Pseudo online NPYLM を用いて，$\mathcal{L}, \boldsymbol{s}_{1:N}$ を計算する．

**単語の生成**

単語の分節化同様，以下の式が計算可能であれば，言語モデルと物体概念の双方を考慮した単語を直接生成することができる．

$$w^w = \underset{w^w}{\operatorname{argmax}} P(w^w|o, \mathcal{A}, \mathcal{L}, w^{v,a,h}, \alpha^w, \theta, \pi, \beta) \tag{5}$$

$$= \underset{w^w}{\operatorname{argmax}} \int P(o|\boldsymbol{s}, \mathcal{A}, \mathcal{L}) \times P(\boldsymbol{s}|w^w, \mathcal{L}) P(w^w|w^{v,a,h}, \alpha^w, \theta, \pi, \beta) dS \tag{6}$$

ただし，$\boldsymbol{w}^{v,a,h}$ は，物体から得られる視覚・聴覚・触覚情報である．しかし，この式においても $\boldsymbol{s}$ で積分することが困難であるため，直接計算することができない．そこで，ここでも $P(\boldsymbol{o}|\boldsymbol{s}, \mathcal{A}, \mathcal{L})$ は，一部の単語列以外の確率は非常に小さく無視できると考え，音声認識の $n$-bset から得た確率の高い単語列 $\boldsymbol{s}_{1:N}$ を利用し，図 3(c) のように音声認識部と単語の分節化部を切り離して考える．すなわち，教示発話 $\boldsymbol{o}$ と物体概念から確率が最大となる $\boldsymbol{w}^w$ を選択するのではなく，音声認識によって得られる $n$-best

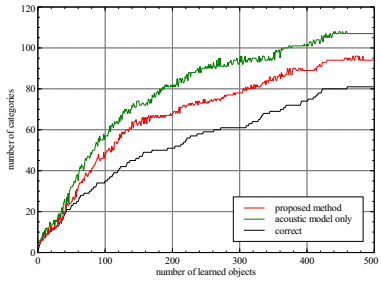


図 4: ロボットの推定したカテゴリ数

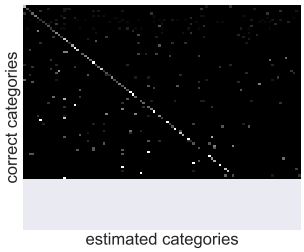


図 5: 比較手法による分類結果 (38.1%)

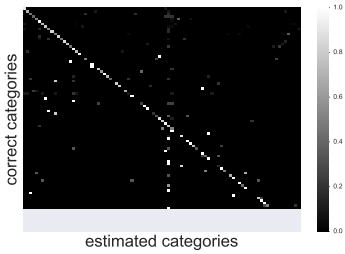


図 6: 提案手法による分類結果 (61.7%)

の単語列 $\boldsymbol{s}_{1:N}$ が物体概念から生成される確率が最大となる単語 $\boldsymbol{w}^w$ を選択する．

$$\boldsymbol{w}^w = \underset{\boldsymbol{w}_n^w}{\operatorname{argmax}} P(\boldsymbol{w}^w|\boldsymbol{o}, \mathcal{L}, \mathcal{A}, \boldsymbol{w}^{v,a,h}, \alpha^w, \theta, \pi, \beta)$$

$$\approx \underset{\boldsymbol{w}_n^w}{\operatorname{argmax}} \frac{P(\boldsymbol{w}_n^w|\boldsymbol{w}^{v,a,h}, \alpha^w, \theta, \pi, \beta)}{\sum_{n=1}^{N} P(\boldsymbol{w}_n^w|\boldsymbol{o}, \mathcal{L}, \mathcal{A}, \boldsymbol{w}^{v,a,h}, \alpha^w, \theta, \pi, \beta)} \tag{7}$$

ただし，$w_n^w$ は単語列 $\boldsymbol{s}_{1:N}$ の $n$ 番目の単語集合を意味する．

**物体概念形成**

以上の手順により，各物体に与えられた音声から物体概念を考慮した単語情報 $\boldsymbol{w}^w$ を得ることができる．ここでは，図 2 から，音声認識部，言語モデル学習部，単語生成部分を切り離し，図 3(d) のモデルとして学習を行なう．学習は，各物体のマルチモーダル情報 $\boldsymbol{w}^*$ を生成する確率を最大とするパラメータ $\theta^*, \pi, \beta$ を求めることに相当する．図 3(d) は oMHDP[青木 15] のグラフィカルモデルである．

逐次学習における言語モデルのパラメータ $\mathcal{L}$ の初期値は，全ての音節が等確率で出現する音節モデルを用いる．新規物体が与えられる度に以上の手順を繰り返すことで，パラメータの推定を行う．これにより，音声認識と物体概念の学習が互いに影響し合い，双方の精度が向上する．

## 3. 実験

実際に長期間，ロボットと人がインタラクションを取りながら学習を行った際の結果について述べる．実験は 1 日 3~5 時間程度の学習を 1ヶ月ほど行い，計 499 個の物体の学習を行った．学習物体には，ペットボトルやお菓子といった日用品の他，ぬいぐるみやガラガラといったおもちゃ，マウスやキーボード，漫画など様々な物体が含まれている．

ここでは，比較として提案手法の他に，音響モデルのみの認識結果から得られた単語情報を利用して学習したもの (acoustic model only) の結果も合わせて述べる．

### 3.1 学習モデルの評価

#### 3.1.1 推定カテゴリ数

図 4 がロボットが推定したカテゴリ数の推移を示している．図中の黒線は人手によって物体を分類したときの物体カテゴリ数の累計を表しており，赤線は言語と概念を相互に学習したもの，つまり提案手法におけるカテゴリ数の推移を表している．緑線は音響モデルのみの結果を単独の oMHDP で学習した比較手法における結果である．

提案手法，比較手法ともに正解より多くのカテゴリの推定がされた．これは，観測情報にはノイズが含まれているためであり，特に本研究のように音声認識に誤りや分割誤りが含まれる場合は学習データが複雑になり，推定カテゴリ数が多くなった．しかし，提案手法は比較手法よりも推定カテゴリ数が少なく，またその増加傾向も約 150 物体目付近から正解である黒線と一致している．提案手法は言語と概念を相互に学習することで，後述するように音声の認識精度が高くなり，また概念による適切な単語の生成ができているため単語情報のノイズを軽減し，より人の感覚に近い概念の学習ができたことがわかる．また，カテゴリ数の増加が収束した後も，新たなカテゴリに分類すべき物体が教示されると適切にカテゴリ数を増加させながら学習ができており，そのような学習の様子が実験を通じて数回生じていることから，今後も学習を続けることで，適切にカテゴリ数を推定しながら学習できることが予想される．つまり，提案したモデルはノンパラメトリックベイズの枠組みによって高いスケーラビリティを持つことができており，また言語と概念を相互に学習することによって，より人に近い感覚でカテゴリ数を推定できたと言える．

#### 3.1.2 ロボットの形成した概念

図 5 および図 6 が全物体を学習後に学習物体である 499 個全てを分類した際の結果を示している．縦軸は正解のカテゴリを，横軸は実際の分類が行われたカテゴリをそれぞれ表しており，各要素が対角線上に高い確率 (白色) で割り当てられるほど人の分類と一致していることを意味している．提案手法は全物体に対して 61.7%の分類ができており，比較手法よりも良い分類ができていることがわかる．これは，比較手法の言語情報に音素誤りや分割誤りが非常に多く，概念が正しく形成できなかったのに対して，提案手法は言語と概念を相互に学習することによって，より良い認識結果を得ることができたため，概念形成もより正しく行えたことを意味している．

### 3.2 未知物体の認識

449 番目に学習したモデルを用いて，最後に教示した 50 個の物体を未知物体として扱うことで，提案手法における未知物体に対する性能の検証を行った．

#### 3.2.1 未知物体の分類精度

図 7，図 8 が比較手法と提案手法による未知物体の分類精度を表した混同行列である．未知物体においても，提案手法の方がより人間に近い分類ができており，より適切な概念が形成されていることがわかる．

#### 3.2.2 未知物体に対する音節認識精度

図 9 は，未知物体に対して与えられた教示発話を各手法で音声認識した際の精度を示している．音声の認識精度は人が書き下した文章を正解として，その文章との編集距離を求めることで算出した．灰色の部分は認識対象の 50 物体の学習が含まれる領域である．緑線は比較手法における音声認識精度を示しているが，これは音響モデルのみの認識を行っているため，

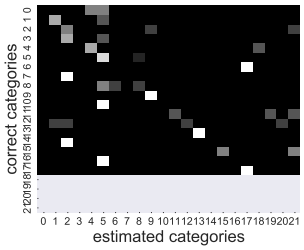


図 7: 比較手法による未知物体の分類結果 (46.0%)

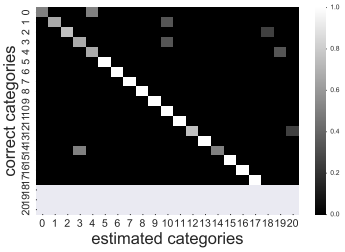


図 8: 提案手法による未知物体の分類結果 (86.0%)

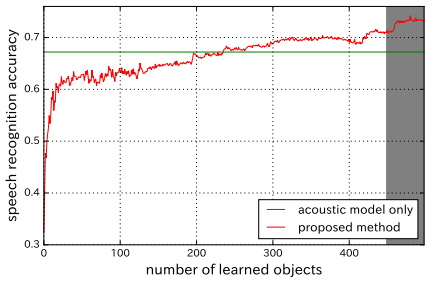


図 9: 学習に応じた音節認識精度の変化

その精度は学習段階によらず一定になる．一方，言語モデルを学習の度に更新する提案手法は，学習初期段階では学習データが少ないため比較手法よりも精度が低くなっているもの，学習が進むにつれて徐々に精度が上昇し，5%ほど精度が高くなった．今回用意した未知物体は学習後半である 400 物体以降に教示されたものが多いため，400 物体を過ぎたあたりで大きく精度が向上し，灰色の領域まで含めると最終的には 8%ほど精度が向上する結果が得られた．つまり，言語と概念を相互に学習する提案モデルにより，ロボットはより精度よく人の教示発話を認識することが可能になった．

### 3.3 長期学習に見る記号創発

一ヶ月という長期に渡る大規模実験を通じて，非常に興味深いロボットによる創発現象が見られた．それは学習の過程で殆どの物体に共通する特徴を表現するカテゴリが生成された点である．図 10 は最終的にロボットが学習したモデルの学習物体全体に対する各カテゴリの推定確率を示した行列であり，縦軸が各物体に，横軸が各カテゴリに対応している．カテゴリ番号はロボットが学習の過程で推定したカテゴリの順番に対応しており，右に向かうほど学習後期にロボットが生成したカテゴリである．61 番のカテゴリに注目すると，このカテゴリはほとんどの物体に対して高い確率を示しており，物体全体に共通する特徴を表現したカテゴリになっている．実際，このカテゴリは『これは』『です』『だよ』といったありとあらゆる物体に対して与えられた機能語と非常に強い結びつきが学習され，それら単語が占める割合は約 75%と非常に高いものだった．さらに興味深いことに，このカテゴリはロボットのカテゴリ形成に大きな影響を与えており，図 4 を再び見てみると，このカテゴリが生成された時刻である約 140 物体目付近からロボットの推定カテゴリ数の増加の様子が人の分類の様子と一致していることが分かる．すなわち，全ての物体を一般化するようなカテゴリがロボット自身によって創り出されたことによって，ロボットはより物体の特徴的な部分に注目して新たなカテゴリを生成することができたことが分かる．また，確信度の低い物体はこのカテゴリに割り当てられる確率が高くなり，『これ』

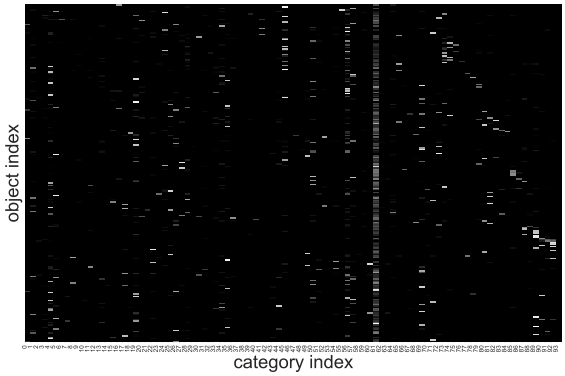


図 10: 全物体に対する各カテゴリの予測確率

『だよ』といった機能語がより想起されるようになる．物体に対して『これ』といった指示語を多用するといった傾向は，人の幼児期においてもよく見られる現象であり，長期に渡る学習を行うことによって，ロボットにおいてもより人間らしい学習プロセスが再現されている可能性を示唆している．

重要なことは，実環境における長期的で大規模な実験を行うことによって初めて見えてきたロボットによる記号創発現象がある ということである．

## 4. 結論

本稿では，ノンパラメトリックベイズの枠組みによる概念学習と言語と概念の相互学習モデルとを組み合わせることで，言語的知識並びに物体に関する知識を持たない状態から概念・語意・言語モデルを同時に学習する手法を提案した．実環境において一ヶ月に渡る大規模な実験を行うことで，提案手法がスケーラブルに概念を形成でき，また言語モデルも適切に更新されることを示した．さらに，長期間の学習を行って初めて明らかになったロボットの興味深い記号創発現象を確認することで，ロボットにおける概念学習の新たな可能性について言及した．

今後の課題として，階層構造などを持ったより複雑な概念構造をオンラインで獲得するモデルへの拡張や，人とのインタラクションフレームワークの拡充，より複雑なタスク環境内での概念構築などが挙げられる．

## 参考文献

[Araki 13] Araki, T., Nakamura, T., and Nagai, T.: Long-term learning of concept and word by robots: Interactive learning framework and preliminary results, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. pp.2280–2287 (2013)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Nakamura 09] Nakamura, T., Nagai, T., and Iwahashi, N.: Grounding of word meanings in multimodal concepts using LDA, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3943–3948 (2009)

[青木 15] 青木 達哉, 中村 友昭, 長井 隆行 : オンラインマルチモーダル HDP による物体概念の獲得, 人工知能学会全国大会，2D5-1 (2015)