1I5-OS-09b-3

マルチモーダルLDAとベイズ階層言語モデルを用いた物体概念と言語モデルの相互学習

Learning of Object Concept and Language Model Using MLDA and NPYLM

中村友昭 *1*2 Tomoaki Nakamura　西原成 *2 Joe Nishihara　長井隆行 *2 Takayuki Nagai　船越孝太郎 *1 Kotaro Funakoshi　長坂翔吾 *3 Shogo Nagasaka　谷口忠大 *3 Tadahiro Taniguchi
岩橋直人 *4 Naoto Iwahashi

*1（株）ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co., Ltd.

*2電気通信大学 The University of Electro-Communications　*3立命館大学 Ritsumeikan University　*4京都大学 Kyoto University

Humans develop their concept of an object by classifying it into a category, and acquire language by interacting with others at the same time. Thus, the meaning of a word can be learnt by connecting the recognized word and concept. We consider such an ability to be important in allowing robots to flexibly develop their knowledge of language and concepts. Accordingly, we propose a method that enables robots to acquire such knowledge. The object concept is formed by classifying multimodal information acquired from objects, and the language model is acquired from human speech describing object features. We propose a stochastic model of language and concepts, and knowledge is learnt by estimating the model parameters. The important point is that language and concepts are interdependent. There is a high probability that the same words will be uttered to objects in the same category. Similarly, objects to which the same words are uttered are highly likely to have the same features. Using this relation, the accuracy of both speech recognition and object classification can be improved by the proposed method. However, it is difficult to directly estimate the parameters of the proposed model, because there are many parameters that are required. Therefore, we approximate the proposed model, and estimate its parameters using a nested Pitman–Yor language model and multimodal latent Dirichlet allocation to acquire the language and concept, respectively.

1. はじめに

事物のカテゴリ分類は，人間の認知機能において重要な役割を果たしていることが指摘されており [Rosch 99]，またこのようなカテゴリが概念を形成しており，概念と単語が結びつくことで，我々は単語の意味を理解することができる．すなわちロボットにおいても，このような経験をカテゴリ分類する能力を持つことは非常に重要であると考えられる．

そこで著者らは，これまで LDA (Latent Dirichlet Allocation)[Blei 03] を拡張したマルチモーダルカテゴリゼーションを提案し，複数のモダリティを用いることにより，より人間の感覚に近いカテゴリを形成することが可能となることを示した．さらに，人の発話を音節認識器で認識し，Nested Pitman-Yor Language Model(NPYLM)[Mochihashi 09] を用い教師なしで音節列を単語に分割し，切り出された単語を概念と結びつけることで，その語意の学習を行った [Araki 12]．

しかし，これまでの研究では単語分割において言語モデルである NPYLM は扱っていたものの，音声認識には音節認識のみを用い，言語モデルは扱っていなかった．すなわち，学習することで物体概念は形成できるものの，物体名を正しく認識ができず，またロボットの発話も誤りが多いものであった．言語と概念は密接に関わっており，これらを同時に学習出来ることができれば，音声認識の精度を向上させることができ，さらに誤りの少ない物体概念を形成することが可能となる．そこで，本稿では，これまでのマルチモーダル LDA（MLDA）を拡張することで，言語モデルと物体概念を相互に学習可能なモデルを提案する．図 1 が提案手法の概要である．ロボットは，物体から取得可能なマルチモーダル情報と，その物体の特徴を教示

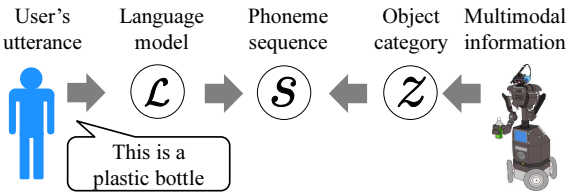


図 1: 提案手法の概要

するユーザー発話から物体概念と言語モデルの学習を行なう．教示音声は言語モデルを用いて，文字列へと変換される．さらに，この文字列は，物体の特徴を表していると考えられるため，物体のマルチモーダル情報から形成される物体カテゴリからも生成される．ここで重要な事は，単語はロボットが形成した概念を意味しており，単語と概念が独立しているのではなく相互に関係している点である．すなわち，同じカテゴリに含まれる物体には，同一の単語が与えられる可能性が高く，また逆に同じ単語が与えられた物体は，共通する特長を有している可能性が高いと言える．ロボットは概念形成と音声認識を行う際に，このような情報を利用することにより音声認識の精度と，分類精度の両方を高める事が可能となる．

関連研究として，視覚情報のみを用いた物体カテゴリの教師なし学習に関する研究 [Sivic 05, Fergus 03, Fei-Fei 05, Wang 09] や，ロボットが物体に触れた際の音を用いた研究 [Sinapov 11] 等，単一のモダリティを用いた研究は数多く行われている．しかし，人間がカテゴリ分類する際には単一のモダリティだけではなく，複数のモダリティを用いていると考えられるため，より人間の感覚に近い分類を実現するためにはマルチモーダルな情報が必要である．また，ロボットによる語意学習の

連絡先: 中村友昭，電気通信大学，東京都調布市調布ケ丘 1-5-1，naka_t@apple.ee.uec.ac.jp

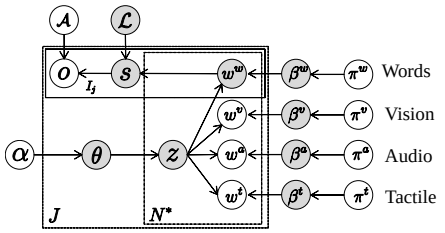


図 2: 言語と物体概念のグラフィカルモデル

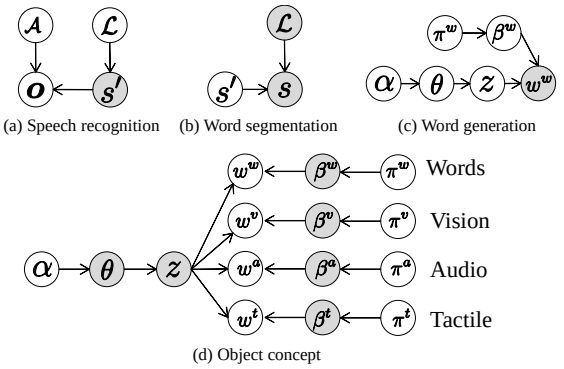


図 3: 近似モデル

研究もいくつか行われている [Roy 02, Iwahashi 06, 田口 10]. 我々はモダリティ間の未観測情報の予測が物体理解の基礎となると考えている [Araki 12]. しかし，これらの研究では，そのような未観測情報の予測は考慮されていない.

2. 言語と物体の統合モデル

図 2 が，言語と物体概念を統合したグラフィカルモデルであり，灰色で示されたノードは未観測ノードを表している．図中の o が人から教示される音声であり，この音声を $\mathcal{A}$ をパラメータとする音響モデル，$\mathcal{L}$ をパラメータとする言語モデルにより，認識した結果が s である．さらに，認識結果 s を，言語モデル $\mathcal{L}$ を用いて単語へ分割し，Bag of words(BoW) 表現へと変換したものが単語情報 w^w であり，さらに，w^v，w^a，w^t はそれぞれ物体から得られる視覚情報，聴覚情報，触覚情報を示している．各情報の詳細については後で述べる．また z は物体のカテゴリを表している．さらに，w^v，w^a，w^t，w^w は，それぞれ β^v，β^a，β^t，β^w をパラメータとする多項分布から発生する．これらの多項分布は，それぞれ π^* をパラメータとするディリクレ事前分布に従う．また，カテゴリ z の出現確率分布を表す多項分布のパラメータを θ とする．このパラメータは，ハイパーパラメータ α により決まるディリクレ事前分布に従う.

このモデルでは，音声認識結果 s と物体カテゴリ z が単語 w^w によって接続されているため，認識して得られた単語が物体カテゴリに影響し，さらに物体カテゴリから音声認識に影響するモデルとなっており，音声認識・単語の接地・概念獲得などが統合されたモデルとなっている．言語モデル・物体概念獲得は，可観測ノードである音声 o と視覚・聴覚・触覚情報 w^v, w^a，w^t から，パラメータ $\mathcal{L}$, β^*, θ を推定し，隠れ変数である音声認識結果 s，単語情報 w^w，物体カテゴリ z を決定することで可能となる．しかし，このモデルは複雑なため，一度に全てのパラメータを求めることは困難となる.

そこで本稿では，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で学習する手法を提案する．このモデルは，音声認識・単語分割・単語生成・物体概念形成の 4 つに分割することができ，各モデルはぞれぞれ図 3(a)-(b) のようになる．図中の灰色のノードが未観測の推定すべきパラメータを示しており，適当な初期値から初め，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.

1. **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 $\mathcal{L}$ は既知とし，全物体になされた全教示発話 $\boldsymbol{O}$ から，n-best の認識文字列 $\boldsymbol{S}'_{1:N}$ を得ることができる.

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

実験では，音声認識には Julius を用い，Julius 標準の音響モデルを用いた.

2. **単語の分節化**
次に，言語モデルのパラメータの推定を行う．言語モデルのパラメータ $\mathcal{L}$ は，全教示発話 $\boldsymbol{O}$ を生成する確率 $P(\boldsymbol{O}|\mathcal{A}, \mathcal{L})$ を最大化することで得ることができる.

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

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

しかし，$\boldsymbol{S}$ での積分は，あり得る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，ここでは単語列の音響尤度 $P(\boldsymbol{O}|\boldsymbol{S}, \mathcal{A})$ 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，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，言語モデルのパラメータ $\mathcal{L}$ を近似的に計算する.

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

ここでは，NPYLM[Mochihashi 09] により，文字列 $\boldsymbol{S}'_{1:N}$ を単語へと分節化することで，$\mathcal{L}, \boldsymbol{S}_{1:N}$ を計算する.

3. **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.

$$\boldsymbol{W}^w = \underset{\boldsymbol{W}^w}{\operatorname{argmax}} P(\boldsymbol{W}^w|\boldsymbol{O}, \mathcal{A}, \mathcal{L}, \pi^w, \alpha) \tag{5}$$

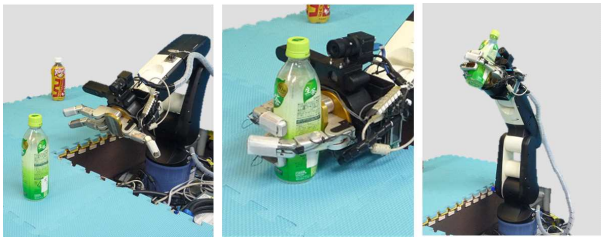
$$= \underset{\boldsymbol{W}^w}{\operatorname{argmax}} \int P(\boldsymbol{O}|\boldsymbol{S}, \mathcal{A},)P(\boldsymbol{S}|\boldsymbol{W}^w, \mathcal{L}) \times P(\boldsymbol{W}^w|\pi^w, \alpha)d\boldsymbol{S} \tag{6}$$

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 $\boldsymbol{O}$ と物体概念から直接確率が最大となる $\boldsymbol{W}^w$ を計算するのではなく，音声認識によってありえる単語列 $\boldsymbol{S}_{1:N}$ を計算し，その中から物体概念によって生成される確率が最大となる単語 $\boldsymbol{W}^w$ を計算する.

$$\boldsymbol{W}^w = \underset{\boldsymbol{W}^w}{\operatorname{argmax}} P(\boldsymbol{W}^w|\boldsymbol{O}, \mathcal{L}, \mathcal{A}, \pi^w, \alpha) \tag{7}$$

$$\approx \underset{\boldsymbol{W}^w \in \boldsymbol{W}_{1:N}}{\operatorname{argmax}} P(\boldsymbol{W}^w|\pi^w, \alpha) \tag{8}$$

ただし，$\boldsymbol{W}_{1:N}$ は $\boldsymbol{S}_{1:N}$ を単語へ分割して各単語列を BoW 表現へと変換したものである．式 (8) の詳細については次章で述べる.



(a) (b) (c)

図 4: マルチモーダル情報取得 (a) 視覚情報，(b) 触覚情報，(c) 聴覚情報

4. **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認識と物体概念を考慮した単語情報 $\boldsymbol{W}^w$ を得ることができた．ここでは，図 2 から，音声認識部，単語分節化部，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，全物体のマルチモーダル情報 $\boldsymbol{W}^v, \boldsymbol{W}^a, \boldsymbol{W}^t, \boldsymbol{W}^w$ を生成する確率を最大とするパラメータ $\theta, \beta^v, \beta^a, \beta^t, \beta^w$ を求めることに相当し，本稿ではギブスサンプリングを用いてパラメータを推定する．処理の詳細は次章で述べる．

$\mathcal{L}$ の初期値として全ての音節が等確率で出現する音節モデルを用い，以上の手順を収束するまで繰り返すことでパラメータを推定する．これにより音声認識と物体概念の双方が影響し合い，音声認識精度と物体概念形成の精度の向上が期待できる．

3. マルチモーダルカテゴリゼーション

ロボットは実際に物体を観察して得られるマルチモーダル情報と，教示発話から生成された単語情報をカテゴリ分類することで，概念の形成を行う．さらに，形成された概念を用いることで，前章の式 (8) の計算をする．図 3(d) がマルチモーダル LDA のグラフィカルモデルである．物体の分類は，図 3(d) のモデルのパラメータを学習データから推定することに相当する．

3.1 物体概念の学習

視覚・聴覚・触覚情報は，図 4 に示したロボットにより取得した．

視覚情報 ロボットはアームの先に CCD カメラと深度センサを搭載しており，観察することで得られる画像を視覚情報として利用する (図 4(a))．各画像から抽出する特徴量として，Dense Scale Invariant Feature Transform (DSIFT) [Vedaldi 10] を用いる．最終的に，これらの特徴ベクトルは，500 の代表ベクトルによりベクトル量子化することで，500 次元のヒストグラムとする．

触覚情報 触覚情報の取得には，アームに取り付けられたバレットハンドと，そのハンドに取り付けられた触覚アレイセンサーを用いる．図 4(b) のように，ロボットが実際に物体を把持することで得られるセンサーの時系列データの近似を行い，その近似パラメータを各センサーの特徴ベクトルとして扱う [中村 10]．さらに，この特徴ベクトルをベクトル量子化することで，15 次元のヒストグラムを触覚情報として用いる．

聴覚情報 図 4(c) のように，ロボットが物体を把持し，振ることで発生する音をロボットのハンドに取り付けられたマイクにより取得し，聴覚情報として利用する．ひとつの物体を観測している間に得られる音声信号をフレームに分割し，フレーム毎に 13 次元の MFCC(Mel-Frequency Cepstrum Coefficient) を計算する．これにより，各フレームは 13 次元の特徴ベクトルとなる．最終的にこの特徴ベクトルも，ベクトル量子化を行い，50 次元のヒストグラムとする．

単語情報 ロボットが物体を観察している間に，ユーザーが各物体の特徴を音声にて教示する．ロボットは認識された音節

図 5: 実験にて使用した物体

列を，教師なしで形態素解析を行い単語へと分割する．最終的に，単語の出現頻度を表すヒストグラムを，単語情報として用いる．

以上のようにして，得られたマルチモーダル情報から，図 3(d) のモデルのパラメータを学習する．学習にはギブスサンプリングを用い，隠れ変数であるカテゴリ z を，事後確率からサンプリングを繰り返すことで，パラメータの推定を行う．

3.2 単語の生成

学習したモデルを利用することで，他の情報から単語が生成される確率である式 (8) を計算することができる．物体の視覚・聴覚・触覚情報 $\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^t$ が与えられた際に，ある単語情報 w^w が発生する確率を次のように書くことができる．

$$p(w^w|\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^t) = \int \sum_z p(w^w|z)p(z|\theta)p(\theta|\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^t)d\theta \quad (9)$$

この式を利用することで，物体概念を用い，物体を表現する単語を生成することができる．

4. 実験

提案手法の有効性を検証するための実験を行った．実験では，図 5 に示したペットボトルやぬいぐるみなど 10 カテゴリ 50 個の物体を使用した．また，ユーザーが各物体に関する特徴をロボットへ教示した音声を用い，音声認識には Julius を使用した．実験では，提案手法の性能を評価するため，以下の 3 つの手法で比較した．

A. 教示音声を音節認識した結果 $\boldsymbol{S}_0$ を，NPYLM により単語分割を行った単語列を用い物体概念を学習する手法

B. 言語モデルと物体概念を逐次更新する手法（提案手法）

C. 教示発話を人手で書き起こしして得られた誤りのない文字列 $\boldsymbol{S}_{correct}$ を，NPYLM により単語分割を行い，その単語列を用いて物体概念を学習する手法

すなわち，手法 A がベースラインであり，これまで我々が用いてきた手法である．手法 B が本稿での提案手法であり，手法 C が本手法における理論上の最高性能であると考えることができる．また，提案手法では，$N = 10$ とし，10-best の認識結果を使用した．

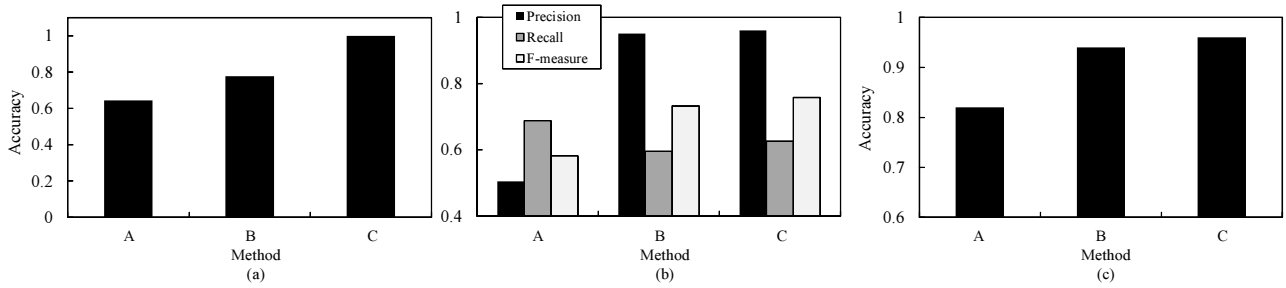


図 6: (a) 音節認識率，(b) 単語分節化の Precision，Recall，F-mesure，(c) 物体分類精度

4.1 教示発話の認識精度

まず，手法 A∼C において，認識された音声の認識誤りについて評価した．手法 C は人手による書き起こしであるため，手法 c で使用した文字列 $\boldsymbol{S}_{correct}$ を正解として，音節の正解精度を計算した．各手法における全教示発話の音節の正解精度の平均が図 6(a) である．この図において，手法 C は正解文字列を用いているため，正解精度が 100%となている．手法 A では，音節認識のみしか用いていないため 64%と最も低い値となった．一方，提案手法である手法 B では，言語的な知識をもたない状態から学習しているにも関わらず，手法 A に比べ 10%以上精度が改善していることが分かる．すなわち，提案手法によってロボットは，物体概念と言語モデルを相互に繰り返し学習することで，より正しい言語モデルを獲得できていることを意味している．

4.2 教示発話の単語分割精度

次に，単語への分節化の性能を評価した．分節化の正解として，形態素解析 (mecab) を用いて誤りのない教示発話を単語へと分割し，誤った分割がなされた箇所を人手により修正したものを用いた．各種法における分節化の Precision, Recall, F 値が図 6(b) である．この結果より，手法 a では，Precision は最も低いが，Recall が最も高くなっている．これは，音節の誤認識により，正しい単語の切れ目が見つけられず，正解よりも短い単語へ分節化する傾向があったためであると考えられる．一方，提案手法である手法 B は，正解音節列を用いている手法 C とほぼ同等の結果となった．Recall が手法 A よりも低くなっているが，これは「おちゃの/ぺっとぼとる」のように，助詞が正しく分節化できなかったためであり，学習データを増やし「おちゃは」や「おちゃが」のような文節を含む教示発話があれば，分節化することが可能である．また，Precision は手法 A に比べて 0.4 以上高く，F 値も約 0.15 高いため，総合的にみても提案手法が有効であるといえる．

4.3 物体概念の学習

最終的に各手法によって得られた単語情報を用いて物体の分類を行った．図 5 の分類を正解として，各手法での分類を評価した．図 6(c) が各種法での分類精度である．この図より，手法 A に比べて，手法 B は 10%以上精度が向上しており，正解音節列を用いた手法 C と比較してもほぼ同等の結果となっている．以上のように，提案手法では，物体概念が教示発話の認識や単語分割精度を改善するだけではなく，改善された音声認識や単語分割が物体の分類も改善していることが分かる．

5. まとめ

本稿では，ロボットが取得したマルチモーダル情報と，人からの教示発話を用いてロボットによる概念・語意獲得を相互に学習する手法を提案した．物体概念と言語モデルを相互に学習することで，音声認識精度を改善し，NPYLM による教示発話の分節化の精度，また物体分類精度を向上することができた．今後，我々がこれまで行なってきたオンラインマルチモーダルカテゴリゼーション [Araki 12] へ，この手法を適用することでよりインタラクティブに学習が可能なシステムを構築する予定である．

参考文献

[Araki 12] Araki, T., Nakamura, T., Nagai, T., Nagasaka, S., Taniguchi, T., and Iwahashi, N.: Online Learning of Concepts and Words Using Multimodal LDA and Hierarchical Pitman-Yor Language Model, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1623–1630 (2012)

[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)

[Fei-Fei 05] Fei-Fei, L.: A bayesian hierarchical model for learning natural scene categories, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 524–531 (2005)

[Fergus 03] Fergus, R., Perona, P., and Zisserman, A.: Object Class Recognition by Unsupervised Scale-Invariant Learning, in *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 264–271 (2003)

[Iwahashi 06] Iwahashi, N.: Robots that learn language: Developmental approach to human-machine conversations, *Symbol Grounding and Beyond*, pp. 143–167 (2006)

[Mochihashi 09] Mochihashi, D., Yamada, T., and Ueda, N.: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling, in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Vol. 1, pp. 100–108 (2009)

[Rosch 99] Rosch, E.: Principles of categorization, *Concepts: core readings*, pp. 189–206 (1999)

[Roy 02] Roy, D. and Pentland, A.: Learning words from sights and sounds: a computational model., *Cognitive Science*, Vol. 26, No. 1, pp. 113–146 (2002)

[Sinapov 11] Sinapov, J. and Stoytchev, A.: Object Category Recognition by a Humanoid Robot Using Behavior-Grounded Relational Learning, in *IEEE International Conference on Robotics and Automation*, pp. 184–190 (2011)

[Sivic 05] Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T.: Discovering Object Categories in Image Collections, in *IEEE International Conference on Computer Vision*, pp. 17–20 (2005)

[Vedaldi 10] Vedaldi, A. and Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms, in *ACM International Conference on Multimedia*, pp. 1469–1472 (2010)

[Wang 09] Wang, C., Blei, D., and Fei-Fei, L.: Simultaneous image classification and annotation, in *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 0, pp. 1903–1910 (2009)

[中村 10] 中村 友昭, 西田 匡志, 長井 隆行 : 把持動作による物体カテゴリの形成と認識, 情報処理学会全国大会, 5V-3 (2010)

[田口 10] 田口 亮, 岩橋 直人, 船越 孝太郎, 中野 幹生, 能勢 隆, 新田 恒雄 : 統計的モデル選択に基づいた連続音声からの語彙学習, 人工知能学会論文誌, Vol. 25, No. 4, pp. 549–559 (2010)