The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

教師なし学習による物体概念および言語モデルと音響モデルの同時学習

Simultaneous Learning of Object Concepts, Language Model, and Acoustic Model using Unsupervised Multimodal Learning

村上 太亮 *1 Hiroaki Murakami　尾崎 僚 *1 Ryo Ozaki　谷口 彰 *1 Akira Taniguchi　谷口 忠大 *1 Tadahiro Taniguchi

*1 立命館大学 Ritsumeikan University

Categorization of objects plays an important role in human cognition. It is important for robots to form the object concepts for communication with humans. The purpose is to enable that robots form such object categories and acquire language. We propose the simultaneous learning of the object concepts, the language model, and the acoustic model using by combining Multimodal Latent Dirichlet Allocation (MLDA) with Nonparametric Bayesian Double Articulation Analyzer (NPB-DAA).

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 [Ashby 05]．このようなカテゴリは事前知識なしで形成され，対話などのコミュニケーションを円滑なものにしている．本稿ではこのように形成される物体のカテゴリを物体概念と捉える．また，概念と単語が結びつくことで，人間は単語の意味を理解することができる [Smith 05]．ロボットもこのように概念の形成と言語の獲得ができることは，人間とロボットのコミュニケーションを円滑にするうえで重要であると考えられる．

教師なし学習を用いた概念獲得に関する研究として，中村らは Multimodal Latent Dirichlet Allocation (MLDA) と Nested Pitman-Yor Language Model (NPYLM) を用いて，物体概念と言語モデルの同時学習を行った [中村 15]．ここで，言語モデルは単語間のつながりを確率的に表現したものである．しかしながら，この研究では事前に学習された音響モデルを用いていた．対して，発話音声のみから音響モデルと言語モデルを同時に学習する手法として，谷口らは Nonparametric Bayesian Double Articulation Analyzer (NPB-DAA) を提案した [Taniguchi 16]．

そこで本研究では MLDA と NPB-DAA を統合することで，言語モデル・音響モデルを事前に与えることなく概念と言語の同時獲得を目指す．本研究の概要を図 1 に示す．ロボットは物体から取得可能な視覚情報と，その物体の特徴を教示する発話から物体概念と言語モデル・音響モデルの学習を行う．これらの物体概念と，言語モデル・音響モデルは単語列を通して互いに影響を与え，より精度の高い概念や言語を獲得できることが考えられる．

本研究で提案するモデルは MLDA と NPB-DAA を統合しており，各モデルで推定されるパラメータの数も多い．ここで，大規模な構造を持つモデル同士を統合するフレームワークとして，中村らが考案した Symbol Emergence in Robotics Tool KIT (SERKET) がある [Nakamura 18]．SERKET は，プログラム的な独立性を保ちながら，その構成単位である基本モデルを階層的に接続することで，大規模な生成モデルとその推論を容易に構築できるようになるフレームワークのことである．SERKET のフレームワークに従って統合モデルを構築することで，それぞれの基本モデルにおいて独立に推定したパラメータを利用しても，モデル全体としてのパラメータを最適化することが可能になる．

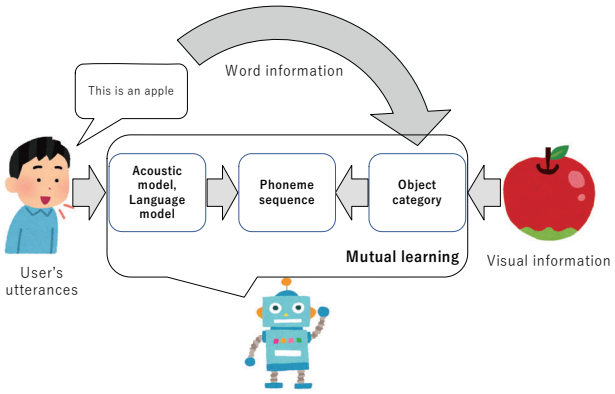

図 1: 物体概念と音響モデル，言語モデルの同時学習イメージ

2. 関連研究

本章では，物体概念獲得と言語モデル・音響モデルの同時学習に関連する各研究の概要について述べる．

2.1 マルチモーダル情報を用いた物体概念の形成

本節では，マルチモーダル情報を用いたカテゴリ分類を行うための手法について述べる．中村らは，Latent Dirichlet Allocation (LDA) を，複数のモダリティから得られるマルチモーダル情報を用いることができるように，Multimodal Latent Dirichlet Allocation (MLDA) として拡張した．ここで，LDA とは文書中の潜在的なトピックを推定するトピックモデルの一種である．中村らは，MLDA を使って複数のモダリティを用いて物体のトピックを推定することで，人間の感覚に近いカテゴリの形成が可能になることを示した [Nakamura 09]．

2.2 物体概念と言語の統合モデル

中村らは，物体概念とその語意を学習する研究において，MLDA と NPYLM を用いて物体概念と言語モデルを同時学習することで，学習する単語の誤りを少なくした [中村 15]．ここで提案されたモデルは，音声認識と物体概念形成が相互に影

連絡先: 村上 太亮，立命館大学 情報理工学研究科，滋賀県草津市野路東 1-1-1，murakami.hiroaki@em.ci.ritsumei.ac.jp

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ている．しかしながら，音響モデルについては事前に学習されたものを用いていた．詳細は [中村 15] を参考にされたい．

また，中村らの研究に関連して，谷口らはロボットの位置情報と音声認識ラティスを利用することで，場所概念と語彙を同時に獲得する手法である SpCoA++を提案した [Taniguchi 18]．場所概念とは，ロボットが持つ位置情報や「キッチン」といった空間の名称に紐づけられる場所に関するカテゴリを指す．

2.3 言語モデルと音響モデルの教師なし学習

言語モデルだけでなく音響モデルも教師なし学習によって獲得可能な手法である NPB-DAA について述べる．NPB-DAA は，二重分節構造を持つ時系列データの生成モデルである Hierarchical Dirichlet Process Hidden Language Model (HDP-HLM) とその潜在変数の推定手法として Blocked Gibbs Sampling を組み合わせたものである．詳細は [Taniguchi 16] を参考にされたい．

3. 提案手法

本章では，教師なし学習によるマルチモーダルカテゴリゼーションと言語モデル及び音響モデルの同時学習手法について述べる．本研究は，マルチモーダル情報からカテゴリ分類と言語モデルの獲得を行う中村らのモデル（2.2 節）を拡張し，NPYLM の代わりに NPB-DAA を用いることで，音響モデルについても教師なし学習によって推定することを可能にする．

MLDA と NPB-DAA を用いたカテゴリ分類と言語モデル・音響モデルの同時学習のグラフィカルモデルを図 2 に示す．ここで，上付き文字 w, v は各モダリティを表し，それぞれが言語情報，視覚情報を指す．$N, S_n, N^w_{ns}, N^v_n, K$ はそれぞれ物体数，n 番目の物体に対する教示発話の回数，s 番目の発話に含まれる単語数，n 番目の物体の画像特徴の出現回数，カテゴリ数を表している．θ^*_k は，カテゴリ k における言語情報の生起確率を表すパラメータであり，β^* をハイパーパラメータとするディリクレ事前分布に従う．$\boldsymbol{o}^v_{ni}$ は，θ^v_k をパラメータとする多項分布から発生する画像特徴量であり，$\boldsymbol{o}^w_{ns}$ は，θ^w_k と L をパラメータとする分布から発生する単語列と仮定する．さらに，z^w_{nsj}, z^v_{ni} は各物体の各特徴が持つトピックを表す．π_n は各 z^*_n の出現確率を表す多項分布のパラメータであり，α をハイパーパラメータとするディリクレ事前分布に従う．$\boldsymbol{y}_{ns}$ は物体に関する教示音声を表し，$\boldsymbol{l}_{ns}$ は音素列を表す．また，A は音響モデルを表し，L は 2-gram 言語モデルを表し，D は言語モデルの単語辞書を表す．

3.1 学習アルゴリズム

本節では，MLDA と NPB-DAA の統合モデルにおける，パラメータの相補的な推定方法について述べる．以下が，提案する学習アルゴリズムである．ここで，NPB-DAA に関する変数の集合として $\mathcal{H} = \{\boldsymbol{y}, A, D, L\}$ と置く．

1. はじめに，NPB-DAA を用いて各教示発話 $\boldsymbol{y}_{ns}$ を単語列 $\boldsymbol{o}^w_{ns}$ に分節化し，言語モデルのパラメータ L，D，音響モデルのパラメータ A をサンプリングする．

2. 以下を一定回数繰り返す．

   A. トピックを考慮した単語列の確率分布 $P(\boldsymbol{o}^w_{ns} \mid \theta^w, z^w_{ns}, \mathcal{H})$ から，単語列をサンプリングする．ここで，$P(\boldsymbol{o}^w_{ns} \mid \theta^w, z^w_{ns}, \mathcal{H})$ を Unigram Rescaling 法 (UR 法) [Gildea 99] を用いて近似的に計算する．UR 法は，トピック依存の N-gram 確率を下式のように近似的に求める手法である．

$$P(\boldsymbol{o}^w_{ns} \mid \theta^w, z^w_{ns}, \mathcal{H}) \approx \frac{P(\boldsymbol{o}^w_{ns} \mid \theta^w, z^w_{ns})}{P(\boldsymbol{o}^w_{ns})} P(\boldsymbol{o}^w_{ns} \mid \mathcal{H}) \quad (1)$$

   ここで，$P(\boldsymbol{o}^w_{ns} \mid \mathcal{H})$ はその組み合わせ種類が非常に大きく，計算困難である．そこで，Sampling Importance Re-sampling (SIR) を導入する．$P(\boldsymbol{o}^w_{ns} \mid \mathcal{H})$ は NPB-DAA から M 個の単語列候補をサンプリングすることで近似される．

$$\boldsymbol{o}^{w[m]} \sim \text{NPB-DAA}(\mathcal{H}) \quad (2)$$

   次に，全単語列候補 $m = 1, \cdots, M$ に対して繰り返す．サンプリングで得られた単語列候補 $\boldsymbol{o}^{w[m]}$ を，Bag-of-Words 表現 $\bar{\boldsymbol{o}}^{w[m]}$ に変換する．$\bar{\boldsymbol{o}}^{w[m]}$ と $\boldsymbol{o}^v$ を用いて MLDA のパラメータ $\boldsymbol{\Theta}^{[m]} = \{\pi^{[m]}, \theta^{*[m]}\}$ および，$z^{*[m]}$ をサンプリングする．

$$z^{*[m]}, \boldsymbol{\Theta}^{[m]} \sim \text{MLDA}(\bar{\boldsymbol{o}}^{w[m]}, \boldsymbol{o}^v) \quad (3)$$

   このとき，各単語列候補の重み $\text{weight}(\boldsymbol{o}^{w[m]}_{ns})$ は

$$\text{weight}(\boldsymbol{o}^{w[m]}_{ns}) = \frac{P(\boldsymbol{o}^{w[m]}_{ns} \mid \theta^{w[m]}, z^{w[m]}_{ns})}{P(\boldsymbol{o}^{w[m]}_{ns})} \quad (4)$$

   となり，これは UR 法における右辺第一項とみなせる．計算した重みに比例するように，M 個の単語列から一つ選択する．以上より，NPB-DAA を用いて M 個サンプリングした単語列から，MLDA と NPB-DAA の両方の推定結果を考慮した単語列 $\hat{\boldsymbol{o}}^w_{ns}$ が下式のようにサンプリングされるとみなせる．

$$\hat{\boldsymbol{o}}^w_{ns} \sim P(\boldsymbol{o}^w_{ns} \mid \theta^w, z^w_{ns}, \mathcal{H}) \quad (5)$$

   以上の手順より，n 番目の物体に与えられた s 個目の教示発話 $\boldsymbol{y}_{ns}$ から，単語列 $\hat{\boldsymbol{o}}^w_{ns}$ を得る．

   B. 2-A で得られた各単語列 $\hat{\boldsymbol{o}}^w_{ns}$ を，Bag-of-Words 表現 $\bar{\boldsymbol{o}}^w_{ns}$ に変換する．$\bar{\boldsymbol{o}}^w$ と $\boldsymbol{o}^v$ を用いて MLDA のパラメータ $\boldsymbol{\Theta} = \{\pi, \theta^*\}$ および，z^* を更新する．

   C. 2-A で得られた単語列の集合 $\hat{\boldsymbol{o}}^w$ を用いて，言語モデル・音響モデルの更新を行う．NPB-DAA を用いて言語モデルのパラメータ L，D，音響モデルのパラメータ A を再度サンプリングする．

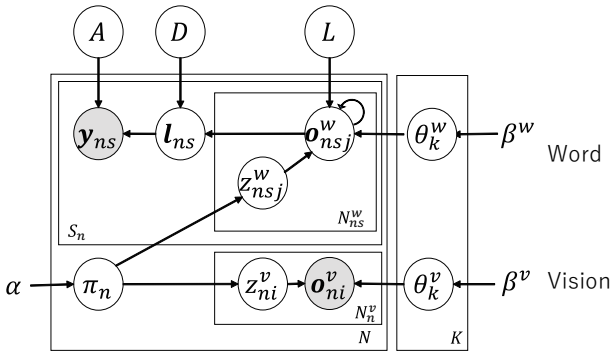

図 2: 物体概念と言語モデル・音響モデルの同時学習のグラフィカルモデル

以上のアルゴリズムで，MLDA のパラメータ Θ と，言語モデルのパラメータ L，D，音響モデルのパラメータ A が得られ，これによって推定された物体カテゴリは，教師なし学習によって推定された単語列を用いて形成される．

4. 実験

4.1 実験 1: カテゴリ分類結果の比較

4.1.1 実験目的

本実験では，正解単語列を言語情報として用いた場合と教師なし学習によって推定された単語列を言語情報として用いた場合のカテゴリ分類結果を比較することで，教師なし学習による単語列推定がカテゴリ分類結果に及ぼす影響を調査する．

4.1.2 実験条件

本実験は中村らのデータセット [Nakamura 17][*1] から一部を抜粋し使用する．この際，物体に対する視覚情報のみを使用し，それ以外のモーダル情報については本稿では用いない．また，各物体に対する教示音声として，新たに子音を含む日本語で発話されたデータセットを作成し，これを用いた．この教示音声は，実験に用いる各物体の特徴に関する全 70 文の発話を無音響空間で収録したものである．また，これらの発話は各音素・単語の生起回数や，単語同士の遷移確率においてバランスがとれるように調整している．本実験で用いる音声特徴量は，フレーム幅を 25 [msec]，フレームシフト長を 10 [msec] として変換された 12 次元の Mel-Frequency Cepstrum Coefficients (MFCC) 特徴量およびその一次微分と二次微分をそれぞれ Deep Sparse Auto Encoder (DSAE) を用いて，8 次元，5 次元，3 次元と段階的に圧縮したのち結合した，合計 9 次元の特徴量を用いる．ここで，MFCC は人間の聴覚特性を考慮した音響特徴量であり，人間の音高に関する知覚的尺度であるメル尺度を使って変換されている．

また，HDP-HLM の隠れ言語モデルにおけるハイパーパラメータは $\alpha^{LM} = 10.0$ と $\gamma^{LM} = 10.0$ とし，weak-limit 近似の単語上限数を 30 個とした．同様に，隠れ単語モデルにおけるハイパーパラメータは $\alpha^{WM} = 10.0$ と $\gamma^{WM} = 10.0$ とし，weak-limit 近似の音素上限数を 30 個とした．持続時間分布には $\alpha_0 = 200$, $\beta_0 = 10$ のポアソン分布を仮定し，MFCC の出力分布には，事前分布に $\mu_0 = 0$，Σ_0 に単位行列，$\kappa_0 = 0,01$，$\nu_0 = (\text{dimension} + 2)$ に正規逆ウィシャート分布を持つ多変量ガウス分布を仮定した．さらに，DSAE のパラメータは $\alpha = 0.003$，$\beta = 0.7$，$\eta = 0.5$ とした．MLDA のハイパーパラメータは $\alpha = 1.0$，$\beta = 1.0$ とした．上記の条件において，ギブスサンプリング 50 イテレーションを 1 試行とし，MLDA と提案手法をそれぞれ独立に 10 回試行する．

4.1.3 実験結果

検証実験結果に関しては口頭発表にて報告する．

4.2 実験 2: 音素・単語の推定結果比較

4.2.1 実験目的

本実験では音素と単語の推定結果を，NPB-DAA が推定した単語列をもって推定する場合と，UR 法によって選択された単語列をもって推定する場合で比較し，MLDA との統合が音素と単語の推定結果に及ぼす影響を調査する．

4.2.2 実験条件

実験条件は実験 1 と同様である．ギブスサンプリング 50 イテレーションを 1 試行とし，NPB-DAA と提案手法をそれぞれ独立に 10 回試行する．

*1 https://sites.google.com/site/nakatomo1018/sd/mod165

4.2.3 実験結果

検証実験結果に関しては口頭発表にて報告する．

5. まとめ

本稿では，物体の画像とその特徴を述べた教示音声から，その物体が所属するカテゴリ及び教示音声の単語列を，教師なし学習によって推定する手法を提案した．また，提案手法は UR 法によって MLDA と NPB-DAA を統合し，その同時学習アルゴリズムを SERKET アーキテクチャの考えに従い SIR を用いて構築した．

参考文献

[Ashby 05] Ashby, F. G. and Maddox, W. T.: Human category learning, *Annu. Rev. Psychol.*, Vol. 56, pp. 149–178 (2005)

[Gildea 99] Gildea, D. and Hofmann, T.: Topic-based language models using EM, in *Sixth European Conference on Speech Communication and Technology* (1999)

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

[Nakamura 17] Nakamura, T. and Nagai, T.: Ensemble-of-Concept Models for Unsupervised Formation of Multiple Categories, *IEEE Transactions on Cognitive and Developmental Systems* (2017)

[Nakamura 18] Nakamura, T., Nagai, T., and Taniguchi, T.: SERKET: An Architecture for Connecting Stochastic Models to Realize a Large-Scale Cognitive Model, *Frontiers in Neurorobotics*, Vol. 12, No. Jun, pp. 1–16 (2018)

[Smith 05] Smith, L. and Gasser, M.: The development of embodied cognition: Six lessons from babies, *Artificial life*, Vol. 11, No. 1-2, pp. 13–29 (2005)

[Taniguchi 16] Taniguchi, T., Nagasaka, S., and Nakashima, R.: Nonparametric bayesian double articulation analyzer for direct language acquisition from continuous speech signals, *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 8, No. 3, pp. 171–185 (2016)

[Taniguchi 18] Taniguchi, A., Taniguchi, T., and Inamura, T.: Unsupervised spatial lexical acquisition by updating a language model with place clues, *Robotics and Autonomous Systems*, Vol. 99, pp. 166–180 (2018)

[中村 15] 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)