Vol.2021-NL-251 No.2
Vol.2021-SLP-139 No.2
2021/12/1

論理否定を含む質問音声を理解する音声言語獲得エージェント

豊田 啓介[1]　篠崎 隆宏[1]

概要：人間が音声を習得する能力をコンピュータ上で再現するためには，音声とそれに紐づく概念を理解すること，対話における相手の発話音声を認識すること，状況認識に基づいて発話内容を決定しそれを発声することが必要である．我々はこれまでの研究で，音声画像データから視覚情報に接地された単語概念を自動獲得し，状況（状態）に応じて適切に単語発話を行う音声言語獲得エージェントを実現した．しかしエージェントが発話を行う状態は，入力画像とエージェントの内部状態のみにより規定されることが想定されていた．本研究では，音声質問を状態に加える拡張を行う．シミュレーション実験により，エージェントが対話試行を通して質問音声と画像，および自身の内部状態に基づいた適切な発話を学習することを示す．さらにタスク設定として音声質問に否定表現を含めた実験を行い，学習時には存在しない入力の組み合わせが与えられた場合でも否定概念の理解のもと適切な返答を行えることを示す．

1. はじめに

現在の音声認識手法はラベル付けされたデータを用いた教師あり学習が多くを占めている．そのため，学習データの少ない言語への対応が難しく，また言語の意味内容の理解もできていない．我々はラベル付けされているデータを用いず，人が言語を獲得する様子をコンピュータ上のソフトウェアエージェントで次のような手順で再現することを目指した．はじめにエージェントに画像とそれに対する説明音声を同時に入力することで，人が目の前にある物体に関して説明を聞いている状態を再現する．その後，対話相手に相当する環境とやりとりを行い音声言語を獲得する．エージェントは 2 つの食べ物を見せられたあと何らかの発話をし，発話したものがそのどちらかの食べ物の名前であった場合に対話相手に相当する環境からその食べ物がエージェントに与えられる．もし渡されたものが欲しい食べ物だったらエージェントにとって嬉しいことであり，欲しくない食べ物を渡されたり何も渡されなかった場合は嬉しくないことである．それを繰り返すことでエージェントは欲しい食べ物の名前を覚え正しく発話できるようになる．先行研究ではエージェントは 2 つの食べ物の画像を見せられ，2 つの画像のうちより欲しいものを常に答えるものだったが，本研究では ‘Which do you want?’ と ‘Which do not you want?’ の 2 つの質問音声をエージェントに与え，‘not’ の有無を含めて質問の内容に応じて適切な応答ができるよう拡張を行った．

[1] 東京工業大学
Tokyo Institute of Technology, Tokyo, Japan
http://www.ts.ip.titech.ac.jp

2. 先行研究

人間は音声情報と視覚情報を組み合わせて言語を獲得していると考えられる．それはラベル付のされていない音声と画像のペアデータを用いた教師なし学習と考えることができる．音声と画像を統合利用する研究として，[1] では音声と画像のそれぞれの特徴量からもう一方を予測する手法を提案されている．[2] ではペアとなる音声と画像が同じ特徴量を出力するようにモデルを学習する手法が提案されている．

また，人間の赤ん坊が親とのやりとりを通じて言語を覚えて成長していく過程を強化学習と捉えることができる．Gao らはソフトウェアエージェントが 3 次元空間上を移動し原点を目指しながら音声言語を獲得する手法を提案した [3]．up, $down$, $left$, $right$, $forward$, $backward$ のどれかを正しく発話できたらその方向へ進んで行くというストーリーで，より最適な移動ができるようエージェント自身が強化学習を用いて学習をしていくというものである．

3. 音声言語獲得エージェント

我々はすでに強化学習を用いた音声言語獲得手法を提案してきた [4]．ここではベースとなる音声言語獲得エージェントを説明する．エージェントは内部状態としてランダムに初期化された色についての嗜好を持っている．そし

て 2 枚の食べ物の画像を見せられたとき，より好きな色に近い食べ物を欲しくなる．エージェントは何らかの発話をするが，その発話が 2 枚の食べ物の画像のどちらかだった場合，その食べ物が与えられるという設定である．

学習は，観察学習と対話学習を組み合わせて行う．観察学習では音声を覚える単語学習と，音声と画像の関係の学習を行う．単語学習は食べ物の名前を含む連続音声を ESKMeans [5] を使い分割をし単語辞書を作成する．ESKMeans による単語分割を行うだけでは辞書サイズに比例して探索範囲が拡大され，後の強化学習を用いた対話学習の効率が大幅に落ちてしまう．そこで，Triplet loss 学習をもとにした音声と画像を接地する手法 [6] を利用し，エージェントが発話をする単語を語彙から選ぶ際に視覚情報を活用する．

対話学習では，エージェントは 2 枚の食べ物の画像を見た上で，単語辞書に基づいて発話をする．その発話内容が見せられた食べ物のどちらかを表すものであるなら，環境によってその食べ物が与えられる．エージェントが欲しい食べ物の名前を正しく発話できた場合，欲しい食べ物を与えられたという状態であるため，それは正の報酬であると考える．一方で欲しくない食べ物の名前を発話をした場合はほしくない食べ物が与えられ，関係ない発話をした場合には何も与えられていない状態となり報酬は発生しない．これを繰り返すことで欲しい食べ物の発音を覚え発話できるようになる．

4. 提案手法

先行研究での対話学習は，常に見せられた食べ物のうち欲しい食べ物を答えるというものだった．そこで今回は，図 1 のように質問音声を入力して質問内容に応じて音声で応答するように拡張した．質問内容は ‘Which do you want?’ と ‘Which do not you want?’ の 2 つで not の有無で答える内容が反転する．2 つの質問からランダムに 1 つが選ばれエージェントに与えられる．そしてその質問内容に応じて適切に応答させる．内部状態 s_{in} で好きな食べ物の色をランダムに定義し，その色により近い食べ物を常に欲しいものとした．‘Which do you want?’ と聞かれたときには欲しい物を，‘Which do not you want?’ と聞かれたときには欲しくない食べ物を答えることを期待する．

先行研究では 2 枚の画像を観察学習の際に使用した ResNet-50 [7] で特徴量に変換した x_{I1}，x_{I2} と内部状態 s_{in} で式 1 のように状態を定義した．

$$s = x_{\text{I1}} \oplus x_{\text{I2}} \oplus s_{in}. \tag{1}$$

本研究では質問音声からスペクトログラムを抽出し，観察学習に用いた ResNet-50 で画像特徴量と同じ次元の音声特徴量 x_a を得る．なお，音声特徴量抽出の際の ResNet-50 では入力チャンネル数を 3 から 1 に変更している．観察学習の際, ResNet-50 はランダム初期値から教師なし学習をしており，ラベル付きデータは一切用いていない．x_a を加えて定義し直した状態は

$$s = x_{\text{I1}} \oplus x_{\text{I2}} \oplus x_{\text{a}} \oplus s_{\text{in}}, \tag{2}$$

と表される．そうして求めた状態 s から式 3 を用いて発話を決定する．

$$Q(s, a; \theta) = \sigma(g_c(s))[a]. \tag{3}$$

θ はランダムに初期化された重みであり，σ はソフトマックス関数である．g_c は複数層からなる全結合のニーラルネットワークで構成されている．最も Q の値が大きくなる発話 a が選択される．

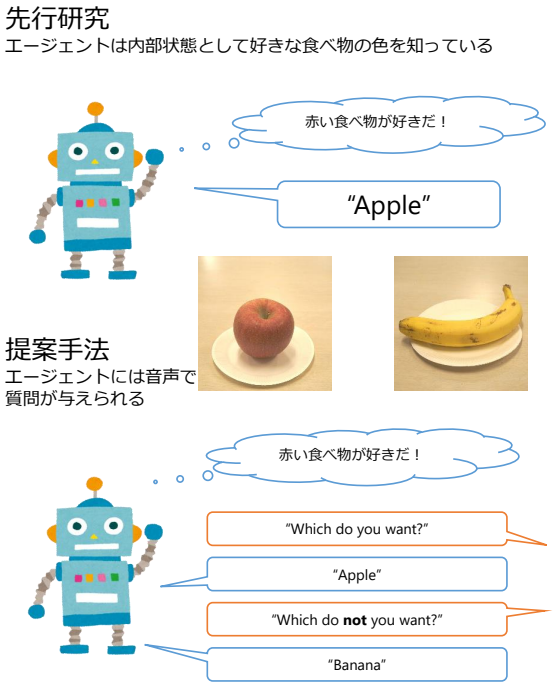

図 1 提案するエージェントのイメージ

5. 実験

5.1 データセット

我々は次の 7 つの食べ物に関するデータを用いて実験を行った．

cherry, green pepper, lemon, orange, potato, strawberry, sweet potato

各食べ物について 120 枚の写真と 4 種類の説明音声を用意した．ランダムに選択した 81 枚の写真を学習に使い，9 枚を検証に用いた．残りの 30 枚を評価に使った．説明音声は Google Text-to-Speech [*1]で生成した合成音声を使用した．説明音声と写真のペアを事前学習に用いたほか，説明音声をランダムにつなげた長時間の音声をもとに ESKMeans

[*1] https://pypi.org/project/gTTS/

表 1 各実験で与えた質問内容

Exp.	質問内容
1	Which do you want? / Which do not you want?
2	Which do you want?
3	Which do not you want?
4	質問音声なし，欲しい食べ物を答えることを期待する
5	質問音声なし，欲しくない食べ物を答えることを期待する

表 2 各実験での正解率

実験の種類	正解率 (%)
UtterQ (want/not want)	90.6
UtterQ (want)	93.8
UtterQ (not want)	93.3
ImplicitQ (want)	93.8
ImplicitQ (not want)	96.1

を用いて単語辞書を作成した．

エージェントに与えられる質問音声も Google Text-to-Speech を用いて生成した．20dB のガウシアンノイズを追加することで，音声発話のデータ拡張を行っている．

5.2 実験内容

先行研究 [4] では質問音声を実際に入力せずに，常に欲しい食べ物を答えることを目的としたエージェントだった．今回はそれと比較するために次の実験を行った．1 つ目から 3 つ目は 1 もしくは 2 種類の質問音声を実際に入力しその質問に適切に応答できるかを確認するものである．4 つ目は先行研究と同様に質問音声を入力することなく，欲しい食べ物を発話することを期待するものである．5 つ目は，同様に質問音声を入力しないものの欲しくない食べ物を発話することを期待したものである．

5.3 実装

実装は PyTorch [8] で行った．発話内容の認識には，スーパーマルチリンガル音声認識モデル [9] を使用した．画像データを K-means クラスタリングをする際のクラスタ数は 24 に設定し，その後その重心を用いて音声語彙のクラスタリングを行う際には，各クラスタに 500 個の音声が含まれるようにした．よって，音声辞書には 12000 個の音声セグメントが含まれることになる．

対話フェーズでは食べ物の画像を 2 枚提示した．その際に，学習時には出現していない食べ物のペアに対しても正しく応答できることを確かめるために ${}_7C_2 = 21$ 個のペアのうち 14 ペアを学習に使い，残りの 7 ペアを検証・評価時に使用した．学習は 7000 エポック行い，各エポックでは 567 個の対話を行った．

5.4 結果

すべての実験について各エポックでの検証結果を図 2 に示す．Reward はエージェント自身が欲しい食べ物を正しく発話できたときに 1，間違えたときに 0 とし，その平均は各エポックにおける正しく発音できた割合（正解率）と一致する．質問の種類や数、質問音声を入力したかに関わらず最終的には同じくらいの高い精度に収束している．また 7000 エポックの学習を行ったモデルで評価をした結果が表 2 に示すとおりである．

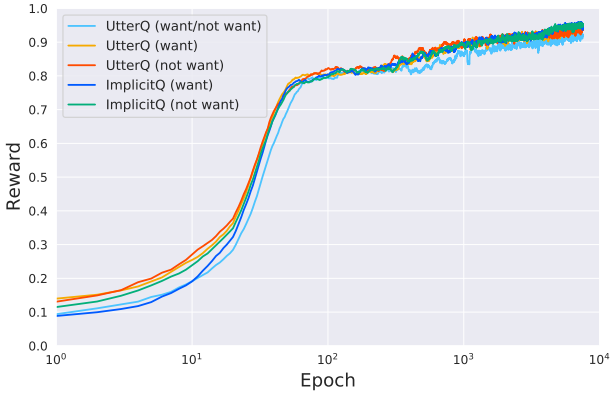

図 2 各実験でエージェントが得た平均報酬の推移

これらの結果より，質問音声を入力した場合にも，‘not’ の有無を聞き分けて適切な応答が可能であることが確認できた．さらに，学習時にはない画像のペアを検証にて用いたものの高い精度を出せていることから，組み合わせに依存せずに応答を決めることができると考えられる．

6. おわりに

我々は Zhang らによる先行研究 [4] を発展させる形で，対話を通じて音声言語を獲得する手法を提案した．事前に言語情報を一切持たないエージェントが，連続音声から語彙を獲得した．対話学習では，2 枚の食べ物の画像とともに質問音声を与えると，質問内容を理解し内部状態を根拠に応答した．エージェントに対し ‘Which do you want?’ と ‘Which do not you want?’ の 2 つの質問音声をランダムに与えたところ，多くの対話で ‘not’ の有無に応じた正しい応答をし，最終的に 90.6%の正解率を得ることができた．これにより論理否定である NOT を理解できたと考えられる．今後の課題として，AND や OR などのこれ以外の論理概念の獲得や，必要に応じて複数の単語を発話をする拡張などが考えられる．

謝辞 本研究は東レ科学振興会の助成を受けたものです．

参考文献

[1] Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B. and Tran, D.: Self-Supervised Learning by Cross-Modal Audio-Video Clustering, *NeurIPS*, Vol. 33, pp. 9758–9770 (2020).
[2] Harwath, D., Hsu, W. and Glass, J. R.: Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech, *CoRR*, Vol. abs/1911.09602 (2019).
[3] Gao, S., Hou, W., Tanaka, T. and Shinozaki, T.: Spoken Language Acquisition Based on Reinforcement Learning

and Word Unit Segmentation, *ICASSP*, pp. 6149–6153 (2020).
[4] Zhang, M., Tanaka, T., Hou, W., Gao, S. and Shinozaki, T.: Sound-Image Grounding Based Focusing Mechanism for Efficient Automatic Spoken Language Acquisition, *Interspeech*, pp. 4183–4187 (2020).
[5] Kamper, H., Livescu, K. and Goldwater, S.: An embedded segmental K-means model for unsupervised segmentation and clustering of speech, *ASRU*, pp. 719–726 (2017).
[6] Harwath, D., Torralba, A. and Glass, J.: Unsupervised Learning of Spoken Language with Visual Context, *NeurIPS*, Vol. 29, Curran Associates, Inc. (2016).
[7] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *CVPR*, pp. 770–778 (2016).
[8] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, *NeurIPS*, Vol. 32, pp. 8026–8037 (2019).
[9] Hou, W., Dong, Y., Zhuang, B., Yang, L., Shi, J. and Shinozaki, T.: Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning, *Interspeech*, pp. 1037–1041 (2020).