人間の動作データを利用したヒューマノイドロボットの言語理解と動作生成

加藤 敦樹　　指導教員：小林 邦和

1 はじめに

ヒューマノイドロボットが人間とのコミュニケーションで動作を行うためには, ロボットが自身の動作を言葉の意味と対応付けて理解する必要がある. Yamada ら [1] は, 深層学習モデルを用いて言語とロボットの動作の統合表現を獲得するモデル Paired Recurrent AutoEncoder (PRAE) を提案している.

このような言語と動作の統合学習において, ヒューマノイドロボットに多様な言語理解と動作生成をさせるためには, 利用するロボットごとに大量の学習データを収集することが必要である. しかしながら, ロボットごとに言語と動作を組み合わせた学習データを用意することは時間的・労力的に困難である.

そこで, 本研究では, ヒューマノイドロボットの言語と動作の統合学習における学習データの増加を目的として, 人間の動作データをヒューマノイドロボットの動作データに変換して学習に利用する手法を提案する.

2 提案手法

提案手法の概要を図 1 に示す. 提案手法は, (1) 人間の動作からヒューマノイドロボットの動作への変換, (2) 言語と動作の統合学習, の 2 段階で構成される. ヒューマノイドロボットが人間に近い身体構造を持つ特性を利用して, 人間の動作をヒューマノイドロボットの動作のデータリソースとする手法である.

2.1 動作データの変換

人間の動作データからヒューマノイドロボットの動作データへの変換には, Master Motor Map[2] の MMMMotionConverter を用いた. MMMMotionConverter は, ヒューマノイドロボットの体の各所に仮想マーカーを設置し, モーションキャプチャの対応する各マーカーの位置との差分が最小かつロボットの力学的制約を満たすような最適化によって動作を変換する.

2.2 言語と動作の統合学習

言語とヒューマノイドロボットの動作の統合表現を獲得するために PRAE[1] を用いた. PRAE は, Reccurent AutoEncoder(RAE) を言語用と動作用の 2 つ用意して, 同じ意味を持つデータの中間表現が近くに配置されるように学習することで言語と動作の統合表現の獲得が行える手法である. 一方の Encoder から出力される特徴量をもう一方の Decoder に入力することで, 動作から説明文, 説明文から動作の双方向の変換を行うことができる. 動作 RAE の Encoder には関節角度の時系列データが入力され, 同じ時系列データを Decoder が出力するような恒等写像として学習される. 言語 RAE も同様に, Encoder に動作説明文の時系列データが入力され, 同じ時系列データを Decoder が出力するように学習される.

損失関数は言語 RAE の損失, 動作 RAE の損失, 言語と動作の中間表現の分布を近づけるための共有損失によって構成される. 言語 RAE の損失は次の式で表すクロスエントロピーである.

$$L_{description} = -\sum_{t=1}^{T_d-1}\sum_{w}^{W} x_{t+1}(w) log y_t(w)$$

ここで, T_d は入力文章の系列長, W は語彙数, x は入力単語の教師ラベル (1 or 0), y は単語の出現確率を表す. また, 動作 RAE の損失は次の式で表す二乗誤差である.

$$L_{motion} = \sum_{t=1}^{T_m-1} ||j_{t+1} - \hat{j}_{t+1}||_2^2$$

ここで, T_m は入力動作の系列長, j は予測の動作, $\hat{j}$ は実際の動作を表す. さらに, 言語と動作を対応づけるため, 各中間表現を近づける制約としてバッチサイズを N, 言語と動作の中間表現をそれぞれ $\{z_i^d | 1 \leq i \leq N\}$, $\{z_i^m | 1 \leq i \leq N\}$ とした時, 次の式で示される損失関数を加えられる.

$$L_{share} = \sum_{i}^{N} \psi(z_i^m, z_i^d) + \sum_{j}^{N}\sum_{j\neq i} \max\{0, \Delta + \psi(z_i^m, z_i^d) - \psi(z_i^m, z_j^d)\}$$

ここで, ψ はユークリッド距離を, Δ は閾値を示すパラメータを表し, 初項は対応する内部表現を近づけ, 第二項は対応しないものを遠ざける. 最終的に, PRAE は次の式で示される損失関数を最小化するように学習することによって言語と動作の対応, すなわち, 統合表現を獲得する.

$$L_{total} = \alpha L_{description} + \beta L_{motion} + \gamma L_{share}$$

なお, α, β, γ は各損失を調整するための重みのパラメータである.

3 実験

3.1 問題設定

問題設定は，人間の動作からヒューマノイドロボットの動作への変換が統合表現の学習に及ぼす影響の評価, 及び, 動作説明文の入力に対する動作生成能力の評価である. そのために, 人間

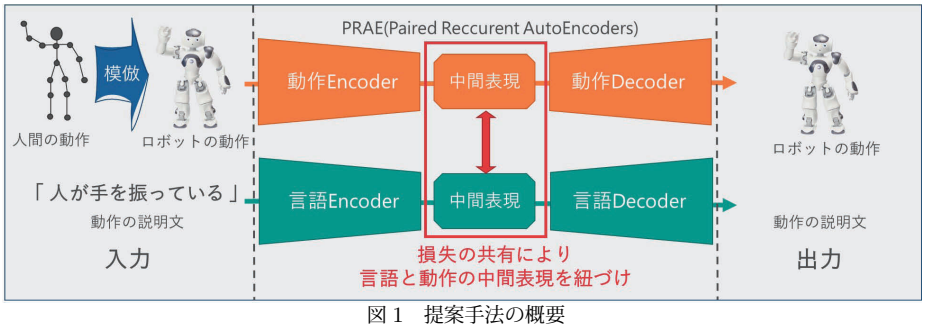


図 1　提案手法の概要

の動作を表す MMM 参照モデル [2] の動作データによる学習と，ヒューマノイドロボット Nao の動作に変換した動作データによる学習結果の比較を行う．

3.2 データセット

学習するデータには，モーションキャプチャから得られた人間の動作 3991 個とその動作の説明文 6353 個で構成される KIT Motion-Language Dataset[3] を用いる．その中から，モーションと説明文の対応があるデータのみを学習に用いた．動作説明文に対しては，事前学習済みの Word2vec[4] を用いて各単語を 300 次元の分散表現に変換を行った．動作に対しては，MMMMotionConverter を用いて人間の動作データからヒューマノイドロボット Nao の動作に変換を行った．学習データは訓練データ 4867 個，検証データ 608 個，テストデータ 609 個の言語と動作のペアの組み合わせで構成される．

3.3 評価指標

文章生成の評価には，BLEU[5]，BERTScore[6] を用いる．BLEU は文脈情報を考慮せず，参照文に含まれる単語との一致度を計算する指標である．BERTScore は，言語モデルである BERT から得られる単語ごとのベクトル表現を用いた指標で，参照文と生成文の Precision，Recall，F1 のスコアを計算する．BERTScore による評価は人間の判断と高い相関があることが示されている [6]．

動作生成の評価については，評価する尺度が決められていないため，定性的に評価する．

3.4 実験結果

3.4.1 動作入力に対する動作説明文生成

動作入力に対する説明文の文章生成の結果を表 1 に示す．表中の $BLEU$ は BLEU のスコアを，P_{BERT}，R_{BERT}，$F1_{BERT}$ はそれぞれ BERTScore の Precision，Recall，F1 のスコアを表している．

表 1　動作入力に対する文章生成結果の比較

動作の種類	$BLEU$	P_{BERT}	R_{BERT}	$F1_{BERT}$
MMM モデル	0.202	0.896	0.899	0.899
Nao	0.156	0.890	0.888	0.889

$BLEU$ スコアは Nao の方が低いが，$F1_{BERT}$ スコアはほぼ同じ値となった．ことのことから，生成される文章に含まれる単語が元となる参照文に含まれる単語と異なってはいるが，意味的に近い文章を生成できていることがわかる．

3.4.2 動作説明文入力に対する動作生成

動作説明文の入力に対する動作生成の結果を図 2 に示す．学習したデータには動作説明文に'walk' が含まれたデータが多いこともあり，両足を交互に前後する特徴を捉えた動作を生成することができている．'walk' 以外の動作においても，それぞれの特徴を捉えた動きを生成していることも確認できた．しかし，動作の特徴を捉えられていない場合や動作の変換に伴う不自然な動きを含んでしまうこともあるため，動作の変換手法の改善などが課題である．

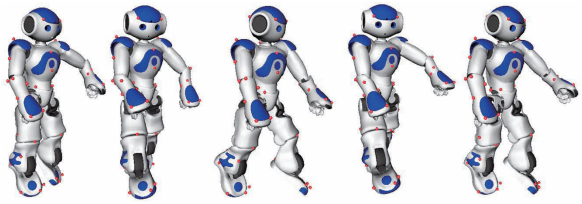
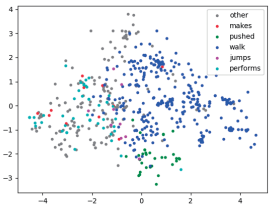


図 2　"A person walks forward."の入力に対する動作生成結果

3.4.3 中間表現の分布

言語と動作の統合表現を獲得できているかを可視化するために，各 RAE の中間表現を PCA 分析で次元削減して，図 3，4 に示すような散布図を作成した．データセット内の頻出動詞 5 個とそれ以外の 6 色で色分けしている．



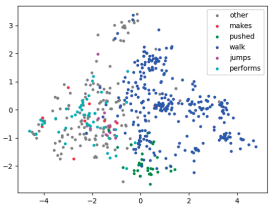


図 3　言語 RAE の中間表現の分布　図 4　動作 RAE の中間表現の分布

色ごとまとまりを見てみると，言語と動作の中間表現が類似した位置に分布されているのが見て取れる．このことから，同じ意味の言語と動作が統合して学習できていることがわかる．

4 おわりに

本研究では，ヒューマノイドロボットの言語理解と動作生成における学習データの増加を目的として，人間の動作データを利用する手法を提案した．実験では，提案手法によってヒューマノイドロボット Nao の動作入力に対する動作説明文生成，動作説明文入力に対する動作生成を行った．その結果，生成された動作説明文は元の参照文と意味的に近い文章を生成しており，生成される Nao の動作は入力された動作説明文の意味を捉えた動作が生成できることが確認できた．このことから，ヒューマノイドロボットの言語理解と動作生成における学習リソースとして，人間の動作データが利用可能なことを示すことができた．

今後の課題は，生成されるヒューマノイドロボットの動作を 2 足歩行でのバランスや実機の耐久性などを加味して実世界で実現することである．

参考文献

[1] T. Yamada, H. Matsunaga, and T. Ogata: "Paired Recurrent Autoencoders for Bidirectional Translation Between Robot Actions and Linguistic Descriptions", in IEEE Robotics and Automation Letters, Vol.3, no.4, pp. 3441–3448 (2018)

[2] O. Terlemez, S. Ulbrich, C. Mandery, M. Do, N. Vahrenkamp, and T. Asfour: "Master Motor Map (MMM) - Framework and Toolkit for Capturing, Representing, and Reproducing Human Motion on Humanoid Robots", IEEE/RAS International Conference on Humanoid Robots (2014),

[3] M. Plappert, C. Mandery, and T. Asfour: "The KIT motion-language dataset", Big Data, Vol.4, No.4, pp.236-252 (2016)

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean: "Distributed representations of words and phrases and their compositionality", Advances in neural information processing systems, Vol. 26 (2013)

[5] K. Papineni, S. Roukos, T. Ward, and W. Zhu: "BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association fo Computational Linguistics, pp.311-318(2002)

[6] T. Zhang, V. Kishore, F. Wu, K. Weinberger, and Y. Artzi: "BERTScore: Evaluating Text Generation with BERT", International Conference on Learning Representations (2020)