# 言語と動作の統合表現獲得による双方向変換

豊田みのり [1]　林良彦 [1]　鈴木彼方 [1,2]　尾形哲也 [3,4]
[1] 早稲田大学 理工学術院　[2] 富士通株式会社
[3] 早稲田大学 理工学術院/理工総研 [4] 産業技術総合研究所
minori-toyoda@fuji.waseda.jp　yshk.hayashi@aoni.waseda.jp
suzuki.kanata@fujitsu.com　ogata@waseda.jp

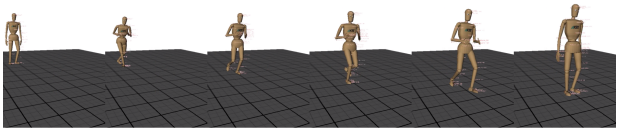
## 概要

ロボットにとって，言語指示から動作を生成する能力，逆に動作を適切に言語化する能力は，日常生活の場で人間と協働する上で不可欠である．本研究では言語と動作の Autoencoder を組み合わせ，さらに単語の事前学習済み分散表現を動作に即して変換することで，言語と動作の双方向変換を実現した．大規模モーションキャプチャデータセットを用いた相互変換の実験結果からは，提案するモデルが，既存手法に比べて一連の動作をより詳細に表す説明文を生成する傾向にあること，言語コーパスのみから学習された分散表現に比べ，より動作に基づいた表現を獲得できることが確認できた．

## 1　はじめに

ロボットにとって，言語指示から動作を生成する能力，自身の行動を伝える能力は人間と日常生活の場で協働する上で不可欠となる．ロボットによる言語理解では，実世界の事物と言語表現を結びつける必要があり，機械学習を用いて実現する場合にはそれらのペアデータセットが求められる．しかし，動作情報等を含めたデータセットは文章のみのデータセットに比べ構築のコストが高い．さらに，実生活において言語も動作も一対一には対応しておらず，使用される可能性のある言語と動作の組み合わせを全て学習させることは現実的でない．したがって，ロボットは未学習の単語を含む指示にも適切に対処しながら，それらを双方向に生成する必要がある．

言語と動作の双方向変換を単一のモデルで行なった研究として，言語と動作のそれぞれの Autoencoder (AE) の中間表現を近づけることで実現した研究 [1] が挙げられるが，データセット内の単語にのみ対応可能であった．未学習の単語に関する研究としては，事前学習済みの分散表現の使用により学習済み単語の類義語の動作を生成した研究 [2] や，言語と動作の相互変換において，動作の内容を反映するように分散表現を変換する学習を行う研究 [3] がある．



someone moves forward in a jog.
a human starts jogging, moving up its hands and stopping the jogging movement
a person jogs for some meters
a person jogs a few steps.

**図 1**　動作と参照キャプション例

後者の研究は，使用文脈が類似している単語には類似した分散表現が与えられてしまうという分布仮説 [4] に起因する課題の解決を試みたものであり，例えば，対義語である"fast"と"slowly"を同一視することによるロボットの想定外の挙動を抑制するなどの実際的な効果が期待できる．しかし，これらの研究で用いられた指示や動作は非常に短く単純であり，実世界に比べ非常に簡易的な設定[1)]であった．

本研究では，より実生活に即した環境での双方向変換を実現することを目的とし，そのためにモーションキャプチャと自然言語の参照キャプションの大規模なペアデータセット [5] を使用する．先述の 1 つのモデルで双方向変換を実現したモデル [1] と，分散表現を変換することで未学習の単語も受容可能としたモデル [3] の 2 つの双方向変換モデルを用い，様々な系列長のより多様な言語や動作へにおける挙動を比較する．後者のモデルでは，動作の情報を含むように変換された単語の分散表現が獲得されるため，その結果についても考察する．

1)　前者の研究 [1] では 1 単語の指示から 3 種類の動作のいずれかを生成した．後者の研究 [3] では BOS と EOS 含めた 5 単語の指示から 12 種類の動作を生成した．後述するモーションキャプチャのデータセットと比べて，言語も動作もバリエーションに乏しく小規模なものであった．

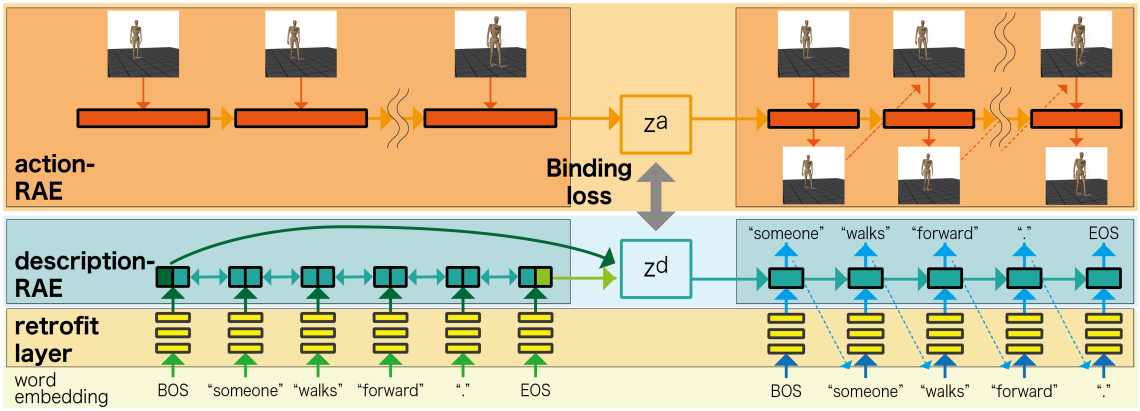



**図 2**　使用モデル：retrofitted Paired Recurrent Autoencoder

## 2　言語と動作の双方向変換モデル

本研究では，言語と動作の双方向変換を実現するモデルとして，単一のモデルで言語と動作の双方向変換を実現したモデル Paired Recurrent Autoencoder (PRAE) [1] と，図 2 に示す未学習の単語も受容可能な統合表現獲得モデル retrofitted Paired Recurrent Autoencoder (rPRAE) [3] を用いる．PRAE は言語用と動作用の 2 つの Recurrent AE (RAE) で構成され，両 RAE は入出力の恒等写像を行うが，言語と動作を結びつけるために両中間表現を近づける制約を与えている．rPRAE は，PRAE に単語の分散表現を動作の情報を含むベクトル表現に変換する非線形層 retrofit layer を加えることで拡張したモデルである．

### 2.1　基本モデル

基本モデルとなる PRAE は動作用 RAE (図 2 上部) と言語用 RAE (図 2 中部) で構成され，共にエンコーダ RNN とデコーダ RNN から成る．本研究において，動作用 RAE ではエンコーダ・デコーダそれぞれにモーションキャプチャによって得られた全身の関節角度と身体の位置，向きを動作情報として入出力する．言語用 RAE でも等しく単語の分散表現を入力し，各単語を出力する．

PRAE は両 RAE の再構成誤差に加え，言語と動作を対応づけるための各中間表現を近づける制約を課している．理想的には 2 つの表現が等しくなり，この表現でそれぞれのデコーダを初期化することで 1 つのモデルで双方向変換が可能となる．バッチサイズを $N$，言語・動作の中間表現をそれぞれ $\{z_i^d | 1 \le i \le N\}, \{z_i^a | 1 \le i \le N\}$ とした時，各内部表現を結びつける損失関数は以下である．

$$L_{bi} = \sum_i^N \psi\left(z_i^a, z_i^d\right) + \sum_i^N \sum_{j \neq i} \max\left\{0, \Delta + \psi\left(z_i^a, z_i^d\right) - \psi\left(z_i^a, z_j^d\right)\right\} \tag{1}$$

ここで $\psi$ はユークリッド距離を表し，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．また $\Delta$ は閾値を示すパラメタである．

### 2.2　分散表現を動作に対応づける変換

事前学習済みの単語の分散表現を，ロボットにとって理解しやすい動作の情報も含む表現に変換するため，rPRAE では言語用 RAE への入力前に非線形層 (retrofit layer, 図 2 下部) を組み込む．この非線形層では分布仮説によって学習された分散表現を動作や状況に合わせたベクトルに変換し，より実世界の事象に即した表現を言語用 RAE に入力する．

本研究では活性化関数に tanh を用いた 3 層の非線形層を用いている．また，パラメタ数の増加に伴うモデルの表現力を制限し過学習を防ぐため，2 つの RAE と retrofit layer を交互に学習する．

## 3　実験

### 3.1　実験設定

より実世界に近いデータでの双方向変換を実現するため，モーションキャプチャと参照キャプションの大規模データセット KIT Motion-Language Dataset [5] を用いて実験を行なった．

実験は動作情報からの説明文生成，参照キャプションからの全身の動作生成という 2 種類のタスクから成る．文生成の評価は BLEU [6],

**表 1** 言語生成および動作生成結果

| | 言語生成 | | | | | | | 動作生成 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BLEU | SimCSE | $P_{BERT}$ | $R_{BERT}$ | $F_{BERT}$ | 文長 | 種類 | BLEU | SimCSE | $P_{BERT}$ | $R_{BERT}$ | $F_{BERT}$ |
| rPRAE | .2686 | .6671 | .9050 | .9023 | .9035 | 10.21 | 163.0 | .2588 | .6482 | .9051 | .9010 | .9029 |
| PRAE | .2850 | .6886 | .9083 | .9040 | .9060 | 10.04 | 205.7 | .2626 | .6579 | .9062 | .9020 | .9040 |

SimCSE [7] によって得られた文ベクトルの cos 類似度，BERTScore [8] の定量的評価に加え定性的評価を行なった．動作生成の評価は，各動作を評価する一定の尺度がなく動作レベルでの定量的評価が困難なため，生成した動作を共通の学習済みのモデルを用いて逆翻訳した文に同上の定量的評価を行なった．上記の指標を用い，PRAE と rPRAE を比較した．

#### 3.1.1 KIT Motion-Language Dataset

KIT Motion-Language Dataset は 3,911 個 (約 11.23 時間) の動作情報と，それらを説明する 6,278 文の英語の参照キャプションで構成される．

動作情報とは 44DoF (Degrees of Freedom) の全身の関節角度と各 3DoF の身体の位置と向きを合わせた計 50DoF のデータである．本研究では双方向変換の先行研究 [1] に則り 1 秒未満および 30 秒より長いデータを削除し，各動作を 100Hz から 10Hz にダウンサンプリングしている．

参照キャプションの文長は 5〜44 単語である．本研究では各単語の分散表現として 300 次元の事前学習済みの Word2Vec [9] のベクトル[2)] を用いる．この使用により，PRAE でも未学習の単語に対応可能となる．また適切に分散表現が変換されたことを確かめるために，事前学習済みの Word2Vec モデルにない単語を含むキャプションも除いた．

したがって，本研究では少なくとも 1 つキャプションを持つ 100〜3000 ステップの動作と，全ての単語に事前学習された分散表現を持つキャプションのペアデータを用いて学習および評価を行なった．使用した動作は 2721 データ，キャプションは 5〜38 語からなる 5538 文 (1331 語彙) であり，80%を学習に，バリデーションと評価に 10%ずつを用いた．

### 3.2 実験結果

#### 3.2.1 実験１：動作による言語生成

動作用 RAE のエンコーダと言語用 RAE のデコーダを用い，入力動作の説明文を生成した．表 1 に生成文の定量的評価の結果を，AppendixA の図 5 に具体的な生成例を示す．$P_{BERT}$，$R_{BERT}$，$F_{BERT}$ は BERTScore の Precision，Recall，F1 値を表し，文長と種類は生成文に含まれる平均単語数と出力文の異なり文数を表す．ここで，評価用参照キャプションの平均文長は 9.573 単語，異なり文数は 520 であった．

定量的評価においては，PRAE の方が良い結果を示した．用いた評価指標は，いずれもキャプションとの一致度に基づくものであり，PRAE の方がより学習データに忠実な生成を行う傾向にあることが分かる．一方で rPRAE は，文法的にも複雑な表現を用いて詳細な記述を生成する傾向がみられた．これらの多くは意味的には適切であると定性的には評価できるが，現在用いている定量的指標だけでは評価が困難であると考えられる．

rPRAE がより詳細な文章を生成する理由として以下が推察される．rPRAE は分散表現を動作に即した表現に変換するため，近い動作を表す複数の単語が類似した分散表現を持つ．このため，似ている分散表現を持つ多くの単語に対して学習される出力単語の例が増加し，より詳細な説明文が出力されたと推測できる．本研究の文生成においては，greedy に次の単語を選択するため，類似した表現を持つ複数の入力単語に対して，確率の高い単一の単語を出力する．これにより，異なり文種が少なくなったと考えられる．PRAE が rPRAE に比べてやや文長が短くなった理由は，似ている分散表現に対する出力単語が rPRAE に比べて少ないことから，各動作を形容する歩数や方向などの補足情報のうち多く共通している表現を出力したものと考えられる．

#### 3.2.2 実験２：言語による動作生成

言語用 RAE のエンコーダと動作用 RAE のデコーダを用い，キャプションから動作を生成した．表 1 に，rPRAE を用いて生成動作を逆翻訳した文の定量的評価の結果を示す．実験 1 と同じく PRAE の方がやや高い結果となった．これは，今回の評価データには使用文脈の類似により動作生成に悪影響を及ぼす可能性のある単語が含まれておらず，rPRAE がその特性を発揮できなかったことが影響したと考えられる．rPRAE 自体の有効性を示すためにはそのよう

2) https://code.google.com/archive/p/word2vec/

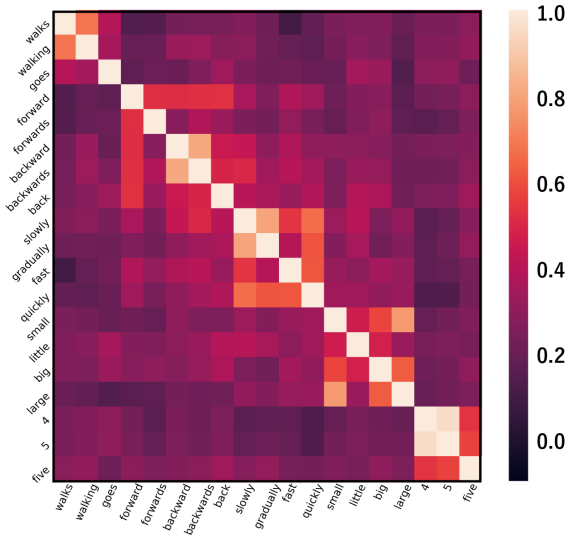



**図 3**　変換前の Word2Vec による各単語の cos 類似度

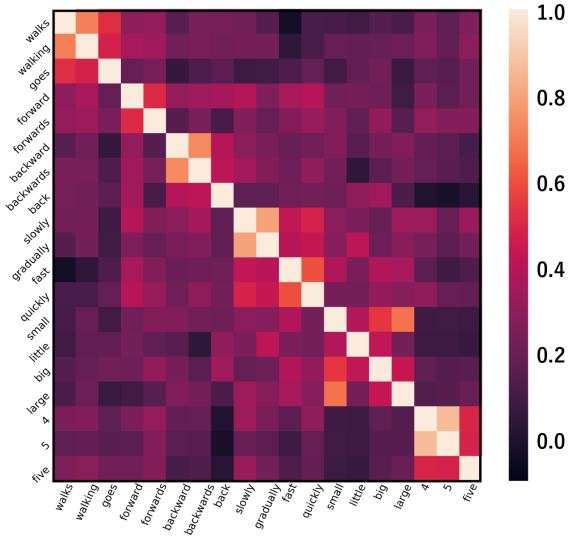



**図 4**　変換された各単語の cos 類似度

なデータを用いて再度検証を行う必要がある．

### 3.3　非線形変換後の分散表現

rPRAE の retrofit layer により動作の情報を含むよう変換した各単語の表現の cos 類似度の例を図 4 に示す．図 3 は変換前の Word2Vec の分散表現である．

これらの図から，特に"forward"と"backward"，"fast"と"slowly"などの文章中での使われ方が似ている対義語が，実際の動作に即し類義語間では高い類似度を，対義語間では低い類似度を示すようになった．

一方で，大小に関する形容詞では期待した結果は得られなかった．まず，"small"と"little"が中程度の類似度を示したのは，キャプション内での使われ方に起因すると考えられる．"small"は circle などの物理的な形を形容する際に多く使われたが，"little"は程度を表す場合が多かった．これらの指す対象の差から類似度がやや低くなったと推察できる．また"small"と"big"，"large"が高い類似度を示した理由としては，元来の類似した表現に加え，アノテータによって尺度が異なるためと推測する．大人と子供が何かを見たときに感じる大きさが異なるように，自身の身体性により大小の感覚は様々である．本研究では文脈非依存の単語の分散表現を用いているが，文脈としてロボットの身体情報やアノテータの尺度の傾向を与えた言語と動作の統合表現を獲得することで，よりロボットの言語理解に適した表現となると考えられる．

## 4　おわりに

本研究ではより現実に即した設定で言語と動作の双方向変換を実現し，2 つのモデルの比較を行った．1 つは基本となる双方向変換モデル (PRAE) で，もう 1 つは単語の分散表現を動作の情報と対応づける変換を行う統合表現獲得モデル (rPRAE) である．モーションキャプチャと参照キャプションで構成される大規模データセットを用いて様々な系列長の多様な言語や動作への挙動を実験的に比較したところ，今回の実験設定では言語と動作の両方の生成において，前者の双方向変換モデルの方が定量的評価が良いことが分かった．一方で統合表現獲得モデルは，より詳細な文章を生成する傾向があることも確かめられた．提案の rPRAE モデルは，分布仮説に基づく分散表現の課題を解決し，動作との対応付けにより対義語の表現を分離することを狙ったモデルであるため，この目的に適合したデータセットを作成する必要がある．

今回利用したデータセットには視覚情報は含まれなかった．より現実世界に近い状況での双方向変換実現においては，視覚などの他モダリティの統合も研究課題である．また言語生成の評価に関しては，動作情報を含むベクトル表現を活用した評価を行う．

## 謝辞

本研究は，JST ムーンショット型研究開発事業 グラント番号 JPMJMS2031，JSPS 科研費 17H01831 の支援を受けた．

# 参考文献


[1] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. In **IEEE Robotics and Automation Letters**, pp. 3441–3448, 2018.

[2] David Matthews, Sam Kriegman, Collin Cappelle, and Josh Bongard. Word2vec to behavior: morphology facilitates the grounding of language in machines. **2019 IEEE/RSJ International Conference on Intelligent Robots and Systems**, 2019.

[3] Minori Toyoda, Kanata Suzuki, Hiroki Mori, Yoshihiko Hayashi, and Tetsuya Ogata. Embodying pre-trained word embeddings through robot actions. **IEEE Robotics and Automation Letters**, Vol. 6, No. 2, pp. 4225–4232, 2021.

[4] Zellig S Harris. Distributional structure. **Word**, Vol. 10, No. 2-3, pp. 146–162, 1954.

[5] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. **Big Data**, Vol. 4, No. 4, pp. 236–252, 2016.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **Empirical Methods in Natural Language Processing**, 2021.

[8] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In **Advances in neural information processing systems**, pp. 3111–3119, 2013.

## A　動作による言語生成例

3.2.1 節で議論した言語生成結果を図 5 に示す．特に注目すべき表現を赤字で，誤った表現を青字で表記している．表 1 から分かるように，言語生成において rPRAE は定量的評価では PRAE に比べ悪い結果を示したが，図 5 のように詳細な表現を記述を生成する傾向が確かめられた．

(a) は左手で身体の前に円を描いている動作であるが，左右の手を正しく記述しているだけでなく，possibly 以下に補足情報をつけている．使用データセットでは手や指の開閉に関する情報はないため，実際の参照キャプションよりも詳細な例として提示する．(b)，(c) も同じく rPRAE がキャプションに比べ，より細かく時系列に沿った文を生成した例である．それぞれの例の上部に提示した一連の動作から分かるように，接続詞や前置詞を正しく用いた適切な順序の記述が確かめられた．(d)，(e) はキャプションよりも適切と思われる生成が確認できた．(d) は実際の動作からキャプションが誤りであることが確かめられており，PRAE と rPRAE の方が正しい生成ができた．(e) は一見した限りでは不明な動作であるが，rPRAE は抽象的なレベルの表現での生成を的確に行っている．(f) は補足情報の付加による失敗例である．具体的には，動作の向きの情報が誤っている．

これらの例の多くは意味的には適切であると定性的に評価できるものの，現在用いている定量的指標だけでは評価が困難である．また，補足情報を必要以上に加えることで (f) の例のように不適切な文となってしまう場合がある．ここから，歩くや走るなどの動作の種類のみを出力し，確証度の低い補足情報は出力しない，といった出力される情報の粒度などを制御できることで，より実用的な場面での使用が期待できる．

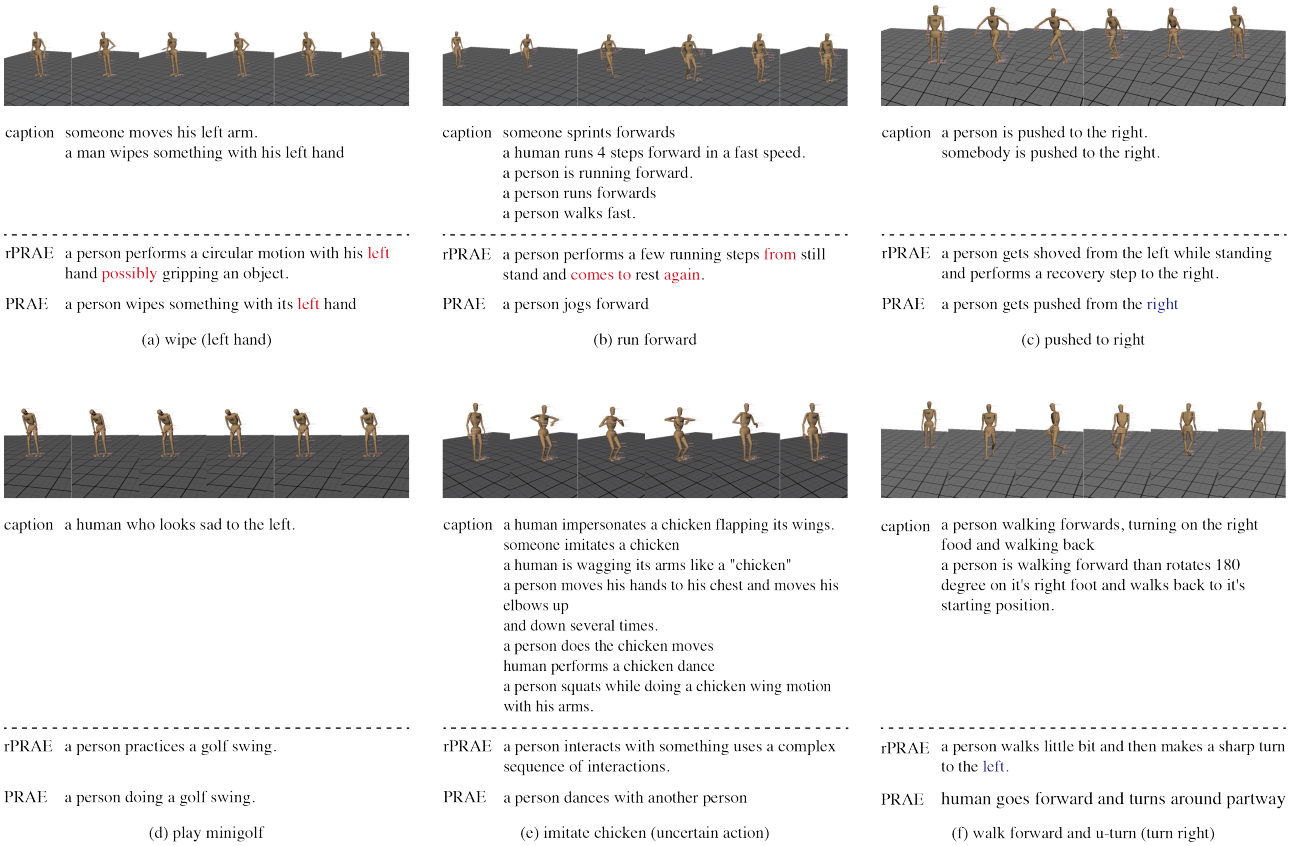



**図 5**　説明文の生成例