指示文・説明文とロボット動作の対応学習

吉野 幸一郎[1,2,3,a)]　脇本 宏平[1]　中村 哲[1,3]

概要：ロボットが生活の中に入ってくるにつれて、ロボットの動作系列と自然言語による指示文・説明文を結びつける重要性が高まっている。本研究では、ロボットの動作系列と自然言語による指示文・説明文の対応を直接学習することを指向して、ロボットが持つアクチュエータの動作系列やカメラ情報から、行った行動を説明する自然言語文を生成する End-to-End のモデルを構築した。ロボットの動作系列は非常に多くのサンプル系列を持つため、少量の学習データから対応を学習することは難しい。この問題を解決するため、ロボット動作の教師なし分節化、および注意機構を導入して対応学習を行った。実験の結果、提案するモデルは分節化を行わないモデルよりも適切な動作説明文を生成できることが示された。

1. はじめに

生活支援を行うようなロボットが多数開発され、ロボットが人間の命令に従ったり、自身の動作を自然言語で説明できるようになることが期待されている [1]。こうしたロボットの動作と自然言語による指示文・説明文の対応は、ロボットの基本動作を定義し、この基本動作の系列と指示文・説明文の対応を取るように学習を行うことが一般的であった [2], [3], [4], [5]。これに対して、近年進展が著しいニューラルネットワークを用いたモデルに対して系列の対応のみを与え、この対応をニューラルネットワークで直接学習しようとする End-to-End、あるいは Sequence-to-Sequence と呼ばれる考え方が注目を集めている [6], [7], [8]。こうした手法により、ロボットの動作系列そのものと自然言語による指示文・説明文の対応を直接学習することができる可能性がある [9], [10], [11]。

自然言語の説明文・指示文とロボット動作の対応を直接学習しようとする場合、まず自然言語の指示文からロボット動作の系列を学習することが期待される。しかし、ロボット動作の系列は実ロボットが動作可能な範囲という制約が存在し、この制約を満たすような軌道計画を生成する必要がある [2]。これに対し、ロボット動作の系列からその動作に対する説明文を生成するようなタスクは、軌道計画のような厳しい制約は存在しない。そこで本研究では、ロボット動作系列と自然言語の説明文・指示文の対応学習を双方向に行うことを指向して、まずロボット動作の系列を入力とし、自然言語による説明文を出力とするようなシステムの構築に取り組む。

ロボット動作の系列と自然言語の説明文・指示文の系列を対応学習しようとする場合、ロボットが動作を期待される空間、動作は多様であり、限られた学習データから精度良く学習を行う必要がある。また、ロボット動作と自然言語の説明文・指示文それぞれが持つ粒度も問題となる。つまり、ロボットの動作系列は各アクチュエータへの指示のサンプリングレートに従って細かく設定される。しかし、再帰型ニューラルネットワーク（Recurrent Neural Networks; RNN）を用いて系列学習を行おうとした場合、ロボット動作系列のように極端に長いもの入力とすると、少量の学習データからその重み伝搬を正しく学習することが非常に難しくなる。

これに対し、既存研究でのロボットの基本動作のような単位で分節化を行うことで、対応学習をより容易にできることが期待できる。ロボットの基本動作の定義においては、基本動作を人手で定義するような手法 [5] の他に、ロボット動作に頻出するような動作をパターンとして定義し、これを分節化の単位とするような手法が存在する [12], [13], [14]。本研究では、こうしたデータから教師なしで獲得可能な分節化の単位を用い、系列の対応学習の前処理として用いることで、ニューラルネットワークによる対応学習が正しく行われることを期待する。具体的には、ロボットの動作系列上の各点を k 平均法によってクラスタリングし、このクラスタ系列をエントロピー基準でサブワード化することで分節化単位を構成した。

また、ニューラルネットワークを用いた系列の対応学習においても、入力のどの部分が、どの出力に貢献するかの

[1] 奈良先端科学技術大学院大学 先端科学技術研究科
生駒市高山町 8916-5
[2] 科学技術振興機構, さきがけ
[3] 理化学研究所, 革新知能統合研究センター AIP
[a)] koichiro at is.naist.jp

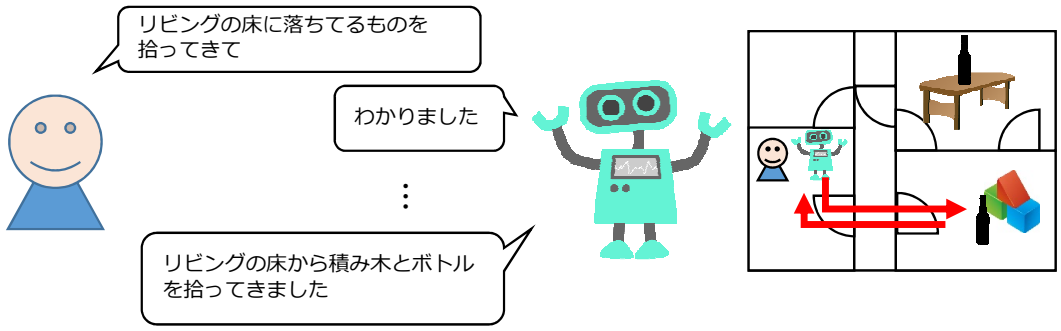

図 1　自然言語指示によるロボット利用の例

重みを学習する、注意機構という枠組みが提案されている [15], [16]。そこで本研究では、この注意機構によって学習された対応が、動作の分節化相当になっていることを期待し、系列の対応学習に注意機構を導入する。さらに、注意機構で学習された対応と、教師なしで行われた分節化がそれぞれ異なる貢献を持つことを期待し、双方を用いるようなモデルも検討する。

実験においては、ロボットシミュレータを用いてロボット動作を生成し、これとクラウドソーシングで付与した説明文との対応学習を行った。評価の結果、提案する分節化と注意機構の利用が、ロボットの動作系列と自然言語による説明文の対応学習に貢献することが示された。

2. ロボットの動作系列と説明文の対応学習

2.1 対応学習の問題設定と関連研究

生活支援ロボットの開発進展にともない、ロボットが家庭環境で人間の補助を行うような様々なタスクが設計されるようになっている。具体的には、屋内で物体を移動させる、特定の場所で動画を撮影する、などのタスクである。本研究では、こうした家庭内環境における補助タスクにおいて、ロボットが人間から自然言語で動作を指示される、あるいは人間に対して自身の行動を説明するといった状況を考える。つまり、これらのタスクではいずれも、ロボットの動作系列とこれに対応する指示文、あるいはこれを説明する説明文が存在し、その対応を取る必要があると考えられる。こうしたタスクの例を図 1 に示す。例では、ユーザは “リビングの床に落ちてるものを拾ってきて” という指示を行い、ロボットはその指示に従い動作を行う。動作終了後に、自身が行った動作の説明として “リビングの床から積み木とボトルを拾ってきました” という説明文の生成を行う。

こうしたシステムを実現するにあたり、これまでの多くの研究では基本動作を人手で抽出し利用していた。これに対し、与えたデータに合わせて動的に基本動作クラスを定義するノンパラメトリックな手法も存在する [12], [13], [14]。しかしこれらの手法はいずれも、ロボット動作の構造化そのものを目的としている。Takano ら [9] は、指示文中

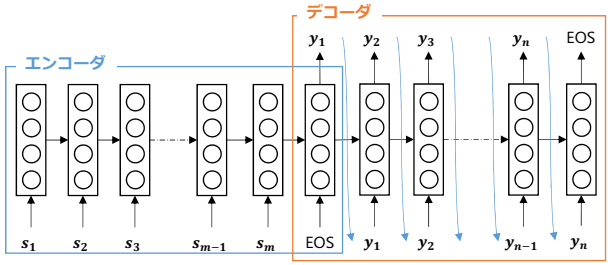

図 2　エンコーダデコーダを用いた対応学習の構成

の意味と言語表現を分割し、言語表現に bi-gram を用いることで対応学習を行った。Yamada ら [11] は Recursive Autoencoder を用い、ロボット動作の埋め込み空間とユーザ発話の埋め込み空間を近づけることによって対応学習を行おうとした。また Plappert ら [10] は、単純なエンコーダデコーダを適用することで対応学習を行っている。しかしこれらの手法はいずれも大量の学習データを要求する。特に分節化を行わずに対応学習を行おうとする場合、新しい環境・新しいロボットにあわせて毎回大量にデータを収集するというのは現実的ではない。そこで本研究では、二種類の分節化手法によって、より少ないデータ量で学習可能な対応学習を目指す。

2.2 エンコーダデコーダを用いた対応学習

先行研究 [10] で行われているエンコーダデコーダを用いた対応学習のアーキテクチャを図 2 に示す。ここで、s_i は時刻 i におけるロボットの動作の生データで、y_j は時刻 j におけるは説明文中の単語である。それぞれ全体を、以降では S と Y として表す。エンコーダは隠れ層 h_k に時刻 k におけるロボットの動作データ s_k を埋め込む。この埋め込みは、

$$h_i = \sigma(W_{sh}s_i + W_{hh}h_{i-1} + b_h) \qquad (1)$$

によって行う。ここで W_{sh} と W_{hh} は変換対象となるベクトルの次元数に対応する重み行列であり、b_h はバイアス項である。また、σ は活性化関数である。デコーダはエンコーダの埋め込みが終了した時点から説明文の各単語 w_j の生成を開始し、w_{n+1}=EOS が生成されるまで生成を行

う。この入力と生成は、

$$h_i = \sigma(W_{yh} w_{i-1} + W_{hh} h_{i-1} + b_h) \quad (2)$$

$$y_i = softmax(W_{hy} h_i + b_y) \quad (3)$$

によって更新、生成される。W_{yh}、W_{hh} および W_{hy} は重み行列であり、b_h と b_y はバイアス項である。σ、$softmax$ はそれぞれ隠れ層と出力層で用いる活性化関数である。機械翻訳や対話などにおいて用いられるエンコーダデコーダとの違いは、s_i として入力されるロボット動作の生データのサンプル列が、サンプリングレートに応じて非常に長大になり、勾配消失 [17] の問題を生じやすくなることである。こうした問題は学習データのサンプル数を増やせばある程度緩和されるが、ロボットやロボットのタスクが変更されるごとに大量の学習データが要求されることとなる。

2.3 実験環境の設定

本研究では、ロボットが行った動作の説明を行う必要がある状況として、World Robot Summit (WRS)[18] におけるサービスカテゴリの家庭内における片付けの環境を設定した。具体的には、家庭内環境における支援ロボットである Human Support Robot (HSR)[19] が家庭内環境で行った動作をユーザに説明する、あるいはユーザの指示に基づいて動作を行うという状況で実験を行った。実験データの収集には SIGVerse[20] シミュレータを用い、シミュレータ上で生成したロボットの動作に対してユーザの指示文・説明文の付与を行った。

入力となるロボットの動作 S としては 0.3 秒ごとにサンプリングされるロボット上の 9 個の関節角の回転量、およびロボット自身の水平方向の直進・回転を指す移動方向・移動量を表す 12 個の値を用いた。また、ロボット自身が観測可能な情報として、ロボットの手先のカメラで撮影された 160×120 ピクセルの画像特徴量も入力に加えた。なお、画像特徴量は Covolutional Autoencoder[21] によって埋め込み表現に変換した 10 次元のベクトルを用いた。

指示文・説明文 W の付与にはクラウドソーシングを用いた。具体的には、一連のロボットのタスクを表す動画を作成し、その動画でロボットが行っている動作をどう説明するか説明文を付与してもらった。この流れでロボット動作を 50 通り作成し、各動作動画に対して 20 名に説明文を付与してもらった。今回作成したデータセットでは、「取ってくる」「置く」「拾う」「落とす」「見に行く」に対応する動作を、各 10 動画ずつ作成した。[*1] この付与された 1000 文を KyTea[22] によって分かち書きし、W の単語列を作成した。

[*1] ただし、指示文・説明文の付与を行う際にこれらの単語を含むような制約は行っていない。つまり、ワーカーごとに「取ってくる」の動作を「持ってくる」と表現したりするようなバリエーションが存在する。

3. ロボット動作系列の分節化

今回収集したデータは合計で 1000 件と少なく、このままエンコーダデコーダを用いて対応学習を行っても適切な学習を行うことが難しい。しかし、ロボットや環境が変わるごとにこれ以上の学習データを収集するということは現実的ではない。そこで、クラスタリングを用いたロボット動作の各点のクラスタリングと、チャンキングを用いた動作のまとめ上げによって、動作の分節化を行う。また、Sequence-to-Sequence における注意機構の学習結果が分節化相当になることを期待し、注意機構の導入を行う。

3.1 クラスタリング・チャンキングによる分節化

ロボット動作系列における各点は、2.3 節に述べた関節角、移動、画像特徴の数値ベクトルを 0.3 秒ごとにサンプリングしたものである。この各点に対して、クラスタリング・チャンキングを行う例を**図 3** に示す。この例では、ロボットの動作系列における各点が二次元ベクトルである場合を示す。まず、このベクトルを k 平均法 [23] によって量子化する。k 平均法は空間上の点を近傍の重心に基づいて任意のクラス数に分類し、分類後各クラスの重心を再度求めることを繰り返して属するクラスを決定するクラスタリング手法である。今回、エルボー法 [24] によってクラス数を 150 と決定した上でクラスタリングを行った。

さらに、バイト対符号化によって量子化された符号列のうち、頻出の部分符号列のサブワード化（分節化）を行った。バイト対符号化は、圧縮率を目的関数として、貪欲に部分符号列の語彙登録を行う手法で、これにより頻出する符号列パターンを 1 語彙として結合することができる。言い換えれば、頻出する動作のパターンを分節化して、基本単位として定義することができる。図 3 の例では、AA、DE というパターンが頻出する符号列として 1 語彙として登録している。今回はバイト対符号化における語彙数を 200 とし、学習データから毎回語彙の学習を行った。この手法を以降では「明示的分節化」と呼ぶ。

3.2 注意機構を用いた分節化

クラスタリング・チャンキングによる分節化では、ニューラルネットワークによる対応学習を行う際の前処理として量子化・分節化を行った。これに対し、ニューラルネットワークによる対応学習の中で、注意機構によって暗黙的に分節化を考慮することができる。注意機構は、入力のどの部分が出力のどの部分に対応するかをゲート機構によって対応学習するもので、限られた個数の出力（単語列）に対して各動作点から注意状態を学習することで、単語に対応する行動のまとまりが学習されることが期待できる。注意機構の例を**図 4** に示す。

注意機構は、エンコーダの隠れ層 $h_{e,i}$ とデコーダの隠れ

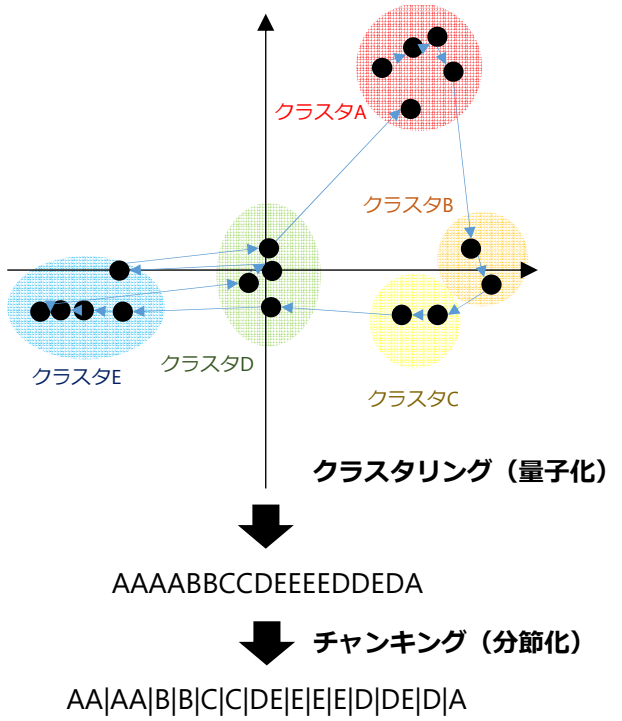

図 3　クラスタリングによる量子化、チャンキングによる分節化

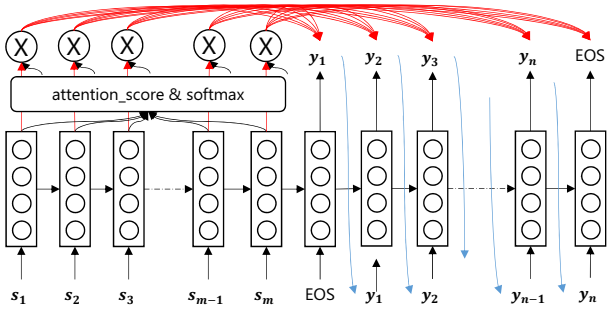

図 4　Attention 機構を持つエンコーダデコーダ

層 $h_{d,j}$ の間で、

$$a_{i,j} = h_{e,i}^T W_a h_{d,j} \tag{4}$$

として計算される。この値を各次元に持つ注意ベクトル a_j をデコーダの各時点に対して注意重みとして用いる。この注意重みは、デコーダのある点に対して、エンコーダの各点から得られた情報をどの程度の割合で利用するかとして解釈できる。今回は、この注意重みが分節化相当になっていることを期待する。すなわち、出力系列に含まれる各単語に対して、類似する動作系列上の点は類似する注意重みを持ち、結果として分節化相当として解釈できることを期待する。以降では、これを「暗黙的分節化」と呼ぶ。

3.3　ハイブリッド分節化モデル

上記で述べた分節化手法は、明示的分節化が入力となる動作系列の情報量のみに着目しているのに対し、暗黙的分節化は対応する言語系列の生成に寄与するという観点から分節化を行っており、それぞれ異なる情報を持つクラスが生成されていることが期待される。そこで、本研究ではこれらの両方を用いるモデルを「ハイブリッド分節化」と呼ぶ。ハイブリッド分節化においては、まず明示的分節化を用いたクラスタリング・チャンキングを行い、このクラスに対して注意機構を持つエンコーダ・デコーダを適用する。

表 1　BLEU スコアによる評価

モデル	BLEU-2	BLEU-3	BLEU-4
分節化なし	0.0649	0.107	0.128
明示的分節化	0.331	0.295	0.264
暗黙的分節化	0.324	0.294	0.266
ハイブリッド分節化	0.339	0.301	0.269

4.　実験

実験では、2.3 節で説明したデータを用いて、ロボットの動作系列 S から説明文の単語系列 W を出力するエンコーダ・デコーダを学習した。条件としては分節化なし、明示的分節化、暗黙的分節化、ハイブリッド分節化の 4 種類のエンコーダ・デコーダを学習し、出力文の比較評価を行う。以下に実験条件の詳細を示す。

4.1　実験条件

データは 50 種類の動作からなるため、このうち 40 種類を学習データ、5 種類を検証データ、5 種類を評価データとして分割する 10 分割交差検証を行った。各エンコーダ・デコーダモデルにおいては隠れ層 160 ノード、1 層の LSTM を用いた。いずれもバッチサイズは 64、ドロップアウト率 0.5、学習率 0.001、weight decay 1e-0.6 を用い、学習の終了は検証データにおける誤差を見て決定した。

4.2　BLEU による自動評価

まず、生成された文の良さを自動評価するため、評価データの動画に付与された参照文との比較を BLEU[25] で行った。今回のデータは 1 つの動画に複数の参照文が付与されているため、出力文を各参照文と比較し、最も高い BLEU スコアを持つ参照文を評価対象とした。BLEU については BLEU-2, 3, 4 をそれぞれ算出し、自動評価とした。

各モデルから生成された出力文に対する BLEU-2, 3, 4 の評価を表 1 に示す。評価の結果、分節化を行わない場合（分節化なし）では BLEU スコアはいずれも低く、クラスタリング・チャンキングを用いた明示的分節化によってスコアが改善することがわかる。また、注意機構を用いた暗黙的分節化によっても同様の効果が見られ、明示的分節化と同程度に各 BLEU スコアが改善していることがわかる。最後に、双方の手法を用いるハイブリッド分節化においては、いずれの手法単体よりもスコアがやや向上していることが見てとれ、それぞれ異なる分節化単位を持つことの有効性が示唆された。

表 2　主観評価によるスコア

Model	a	b	c	d	e	a–c
分節化なし	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
明示的分節化	11.7%	15.0%	33.3%	33.3%	6.7%	60.0%
暗黙的分節化	3.3%	5.0%	30.0%	61.7%	0.0%	38.3%
ハイブリッド分節化	3.3%	5.0%	38.3%	46.7%	6.7%	46.6%

表 3　生成例 1

参照文	床のティーポットを取ってきて
分節化なし	ののののののののののののの
明示的分節化	床の上のソースを取って
暗黙的分節化	テーブルの上のソースを取って
ハイブリッド分節化	床にあるソースを拾って

表 4　生成例 2

参照文	寝室の様子を見てきて
分節化なし	ののののののてててててて
明示的分節化	部屋の様子を見てきて
暗黙的分節化	テーブルの上のぬいぐるみを持ってきて
ハイブリッド分節化	キッチンの上の様子を見てきて

4.3　被験者による主観評価

BLEU を用いた自動評価は、参照文との比較であるため生成文の良さとある程度の相関はあるものの、意味的に類似するが語彙が異なる場合（「取ってくる」と「持ってくる」など）の評価を正しく行うことが難しい。そこで今回は、ロボットの動作を表示する動画とその動作に対応して生成された文をユーザに提示し、評価してもらう主観評価実験を行った。実験では、3 名の各被験者に対してそれぞれの手法で生成された 20 文ずつ、合計 80 文を、以下の観点で評価してもらった。

a). 文は動画中のロボットへの指示内容を適切に説明している
b). 一部に誤りを含むものの、文は動画中のロボットへの指示内容を概ね適切に説明している
c). 対象とする物体名などにいくつか誤りが存在するものの、文は動画中のロボットの動作指示を表している
d). 文は文法的に正しいものの、対象とする物体、動作双方に誤りを含む
e). 文法的に正しい文になっていない

被験者は各動画と生成文のペアに対し、上記の選択肢のいずれかを選択した。被験者は 3 名、評価されたサンプル数は各手法ごとに 60 文となる。この評価の結果を 2 に示す。分節化を行わない場合、いずれの生成結果も文法的に意味をなさない文になっていることが確認された。これに対して提案する分節化を適用した場合、特に明示的分節化はその効果が確認された。明示的分節化は、生成結果の 60% において正しい動作説明を生成できており（a–c）、その有効性が確認された。また暗黙的分節化も、分節化を行わない場合と比較して正しい動作説明を生成しているが、その割合は 38.3%であり、明示的分節化と比較して評価が低い結果となった。最後に、ハイブリッド分節化は BLEU スコアでの評価では一番スコアが高かったが、ユーザによる評価は明示的分節化に劣るという結果となった。これらの結果から、提案したクラスタリング・チャンキングを用いる明示的分節化を用いる手法が最も有効で、特に動作動詞の生成には寄与しているものの、動作中の物体など細部においては改善が必要であることが明らかになった。

4.4　生成された説明文の比較

また、表 3、4 に各手法からの生成例とその参照文を示す。まず、分節化を用いない場合多くの例で非文が生成されていた。これは、今回の学習データが非常に少量で、対応学習をそのまま行うことが難しいという今回の仮説に合致する。これに対して分節化を用いた手法はいずれも、意味のある文を生成している。特に動作動詞については正しく生成ができている場合が多いことが確認された。これは、ロボットの動作系列と動作動詞の対応が対応学習により取れているということが考えられる。一方で、動作動詞の対象である物体の名称については誤っているものが多くみられた。これは画像情報が正しく反映されていない問題があると考えられる。この問題については、物体のラベル情報などを用いて事前学習などを行うことで、ある程度改善が可能であると考えられる。

5.　まとめ

本研究では、ロボットの動作系列から、動作を説明するシステムの構築を行った。エンコーダ・デコーダを用いた生成を少量の学習データから行うため、クラスタリング・チャンキングを用いた分節化と、注意機構を用いた分節化、これら双方を用いるハイブリッド分節化を提案・利用した。実験の結果、分節化を用いる手法では、より適切な説明文が生成され、特に動作動詞の生成において有効性が示された。一方で、カメラで捉える物体の名称については課題が見られた。これを物体のラベル情報を用いた事前学習によって改善することは、今後の課題である。また、自然言語による指示文を与えた場合のロボットの動作系列生成についても、今後取り組む必要がある。

謝辞

本研究は JST さきがけ JPMJPR165B、および JSPS 科

研費 JP17H06101 の支援を受けた。

参考文献

[1] Hatori, J., Kikuchi, Y., Kobayashi, S., Takahashi, K., Tsuboi, Y., Unno, Y., Ko, W. and Tan, J.: Interactively picking real-world objects with unconstrained spoken language instructions, *2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 3774–3781 (2018).

[2] Sugiura, K., Iwahashi, N., Kashioka, H. and Nakamura, S.: Learning, generation and recognition of motions by reference-point-dependent probabilistic models, *Advanced Robotics*, Vol. 25, No. 6-7, pp. 825–848 (2011).

[3] Kollar, T., Tellex, S., Roy, D. and Roy, N.: Toward understanding natural language directions, *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, IEEE Press, pp. 259–266 (2010).

[4] Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S. and Roy, N.: Understanding natural language commands for robotic navigation and mobile manipulation, *Twenty-Fifth AAAI Conference on Artificial Intelligence* (2011).

[5] Fasola, J. and Mataric, M. J.: Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots, *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, pp. 143–150 (2013).

[6] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734 (2014).

[7] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to sequence learning with neural networks, *Advances in neural information processing systems*, pp. 3104–3112 (2014).

[8] Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E. et al.: State-of-the-art speech recognition with sequence-to-sequence models, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4774–4778 (2018).

[9] Takano, W. and Nakamura, Y.: Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions, *The International Journal of Robotics Research*, Vol. 34, No. 10, pp. 1314–1328 (2015).

[10] Plappert, M., Mandery, C. and Asfour, T.: Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks, *Robotics and Autonomous Systems*, Vol. 109, pp. 13–26 (2018).

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

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

[13] Nakamura, T., Iwata, K., Nagai, T., Mochihashi, D., Kobayashi, I., Asoh, H. and Kaneko, M.: Continuous motion segmentation based on reference point dependent GP-HSMM, *Proceedings of the IROS Workshop on Machine Learning Methods for High-Level Cognitive Capabilities in Robotics* (2016).

[14] Nakamura, T., Nagai, T., Mochihashi, D., Kobayashi, I., Asoh, H. and Kaneko, M.: Segmenting continuous motions with hidden semi-markov models and gaussian processes, *Frontiers in neurorobotics*, Vol. 11, p. 67 (2017).

[15] Luong, T., Pham, H. and Manning, C. D.: Effective Approaches to Attention-based Neural Machine Translation, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421 (2015).

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Advances in neural information processing systems*, pp. 5998–6008 (2017).

[17] Bengio, Y., Simard, P., Frasconi, P. et al.: Learning long-term dependencies with gradient descent is difficult, *IEEE transactions on neural networks*, Vol. 5, No. 2, pp. 157–166 (1994).

[18] Kimura, T., Okugawa, M., Oogane, K., Ohtsubo, Y., Shimizu, M., Takahashi, T. and Tadokoro, S.: Competition task development for response robot innovation in World Robot Summit, *2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*, IEEE, pp. 129–130 (2017).

[19] Yamaguchi, U., Saito, F., Ikeda, K. and Yamamoto, T.: HSR, human support robot as research and development platform, *The Abstracts of the international conference on advanced mechatronics: toward evolutionary fusion of IT and mechatronics: ICAM 2015.6*, The Japan Society of Mechanical Engineers, pp. 39–40 (2015).

[20] Inamura, T., Shibata, T., Sena, H., Hashimoto, T., Kawai, N., Miyashita, T., Sakurai, Y., Shimizu, M., Otake, M., Hosoda, K. et al.: Simulator platform that enables social interaction simulation—SIGVerse: SocioIntelliGenesis simulator, *2010 IEEE/SICE International Symposium on System Integration*, IEEE, pp. 212–217 (2010).

[21] Masci, J., Meier, U., Cireşan, D. and Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction, *International Conference on Artificial Neural Networks*, Springer, pp. 52–59 (2011).

[22] Neubig, G., Nakata, Y. and Mori, S.: Pointwise prediction for robust, adaptable Japanese morphological analysis, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, pp. 529–533 (2011).

[23] Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 100–108 (1979).

[24] Bholowalia, P. and Kumar, A.: EBK-means: A clustering technique based on elbow method and k-means in WSN, *International Journal of Computer Applications*, Vol. 105, No. 9 (2014).

[25] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, pp. 311–318 (2002).