人の動作を対象にした確率的言語生成への取り組み

†小林 瑞季　　†小林 一郎　　‡麻生 英樹
†お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース
‡産業技術総合研究所知能システム研究部門
†{kobayashi.mizuki, koba}@is.ocha.ac.jp, ‡h.asoh@aist.go.jp

1　はじめに

センサなどによって観測される情報の殆どは時系列データであり，ビッグデータを扱う時代においては，観測された時系列データの中から有益な情報を取得し，その内容を理解する手法の開発が重要となる．時系列データの分析方法には，トレンドの予測や複数データ間の相関関係の分析など様々な方法が存在する．一方で，時系列データの内容を理解するには可視化などの手法が用いられている．しかし，ロボットなど複数のセンサによって取得された時系列データの情報に基づき状況を認識する必要がある場合，取得された情報をより抽象度の高いレベルで観測されたデータを表現する必要がある．そのことに着目し，我々は観測された時系列データの振る舞いを言語で説明する手法の開発を目指し，その一つとして Kinect から得られた動画像の情報を入力とした確率的なテキスト生成手法を提案する．

2　視覚情報の言語化の枠組み

本研究の概要を図 1 に示す．まず，Kinect[1] がもつ人の骨格を追跡するライブラリとパーティクルフィルタを用いることで，人と物の動きを時系列データとして取得する．取得された時系列データはいくつかの次元圧縮作業を行い，データと自然言語の仲立ちをする中間表現とともにデータベースに格納される．その後，データベース内に蓄積された時系列データと中間表現の対応関係を機械学習することで，動作判別器を生成する．テキスト生成に用いられる言語資源は，人の動作の表現を被験者実験によって収集し，それぞれの中間表現に対してバイグラムモデルを構築する．これにより中間表現を選択すると，その中間表現に対応したバイグラムモデルが選択され，そのモデルに動的計画法を適用することで，人の動作を表現するもっともらしい語の組み合わせから文を生成することができる．

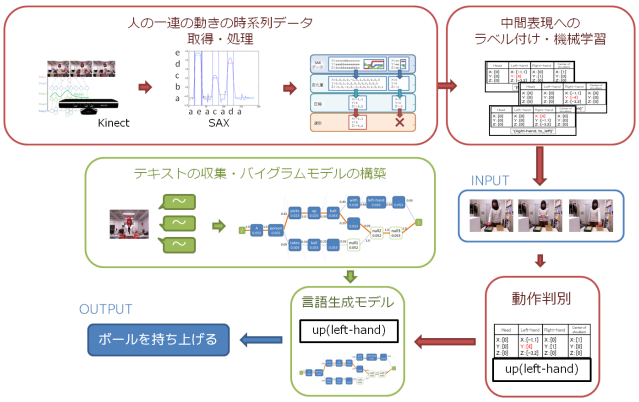

図 1: 動画像を入力とする確率的テキスト生成の枠組み

2.1　時系列データの取得と処理

2.1.1　時系列データの取得

人間の動作の時系列データは，Kinect カメラ [1] を用いて取得する．Kinect の開発元である MicroSoft 社は，人間の骨格を推定できる標準ライブラリも提供しており，そのライブラリを用いると人の関節の 3 次元情報を推定することができる．本研究では，RGB 画像と深度センサー，またそれらを用いた人物の関節位置推定も用い，RGB 動画像と人物の肩の右手・左手・右肘・左肘・肩の中心の 5 箇所の xyz 座標の時系列データを取得する．また，物体の動きの時系列データは，パーティクルフィルタ [2] を用いることで取得する．

2.1.2　時系列データの処理

人の骨格と物体の色を追跡することで得られた時系列データは，Symbolic Aggregation approXimation (SAX) [3] を使い，文字列に変換する．

本研究では動作判別の精度を高めるため，一般にデータを等間隔に分割するところを，各データに動的計画法を用い，尤もらしい区切りを取得することで，よりデータに沿った文字列を取得する．

表 2：各動作に対する生成文の上位 3 文

	生成文	尤度
1	• 左手, を, 上げる, 。, null_5, null_6, null_7, null_8, EOF	1.76e-12
	• 左手, を, 上げる, 。, null_4, null_5, null_6, null_7, null_8	1.57e-12
	• 左手, を, ひじ, を, 上げる, 。, null_4, null_5, null_6	1.19e-14
2	• 右手, を, 上げる, 。, null_4, null_5, null_6, null_7, null_8	8.63e-13
	• 右手, を, 上げる, 。, null_5, null_6, null_7, null_8, EOF	5.40e-13
	• 右手, を, すこし, あげる, 。, null_4, null_5, null_6, null_7	3.98e-15
3	• 両手, を, 下ろす, 。, null_4, null_5, null_6, null_7, null_8, null_9	2.91e-14
	• 両手, を, 下ろす, 。, null_5, null_6, null_7, null_8, null_9, EOF	2.68 e-14
	• 両手, を, 同時に, 下げる, 。, null_4, null_5, null_6, null_7, null_8	1.49e-16
4	• ボール, を, 持ち上げる, 。, null_5, null_6, null_7, null_8, null_9, EOF	2.64e-14
	• 左手, で, ボール, を, 持ち上げる, 。, null_6, null_7, null_8, null_9	2.03e-14
	• ボール, を, 左手, で, ボール, を, 持ち上げる, 。, null_6, null_7	2.77e-15

SAX によって変換して得られた文字列から動作とみられる個所を取り出す．ここでは，ある動画像データ中の全ての文字列において 3 つ前の文字から変化がなければ「動きがない」，変化があれば「動きがある」とみなす．

その後,「動きがある」とみなされた個所の文字列を変化量に変換し，圧縮する．これは同じ動作でも位置や速さによっては文字列がある一定の間隔でずれたり文字列の長さが変化したりしてしまい，同じ動きとして学習されないためである．また，より特徴的な動作を抽出するために，圧縮された変化量のうち最大の大きさが閾値以下を示す動きは取り除く．

2.2 動作へのラベル付けと機械学習

時系列データと自然言語文は、人の動作の意味内容を示すラベル（ここでは「中間表現」と呼ぶ）によって対応関係が表される．観測された時系列データからその内容を表すのにふさわしい中間表現が選ばれるように，その対応関係を対数線形モデルを用いて学習する．また，それぞれの中間表現に対してその意味内容を自然言語で表現するための言語資源（バイグラムモデル）を予め用意し，それを用いてテキスト生成が行われる．

表 1：動作内容を表す中間表現

action	中間表現	意味
up	“up(joint, null)”	upward movement
down	“down(joint, null)”	downward movement
pick	“up(joint, object)”	pick up movement
put	“down(joint, object)”	put movement

2.3 バイグラムモデルによるテキスト生成

本研究では，バイグラムモデルを用いたテキスト生成を行う．それぞれの中間表現に対しバイグラムモデルを構築するために，被験者実験を行い特定の動作を説明する様々な自然言語表現を集めた．これにより，観測された時系列データに対して特定の中間表現が与えられたとき，言語資源として中間表現に対応するバイグラムモデルが選択されテキスト生成が行われる．個々の単語を出現の組合せに基づく確率を基準に文生成を行う場合，一般に文長が長い方が低い確率となってしまう．そのため，本研究では，文の長さに依存しないテキスト生成が行えるよう，バイグラムモデルに null ラベルを導入する．null ラベルは，文中の単語として扱われ，他の単語と同じようにユニグラムとバイグラムの構成要素とみなすことで長文においても短文と同等に尤もらしい生成文として採用することができる．

最後に，構築されたバイグラムモデルに動的計画法を適用することで尤度が最も高くなる単語の組み合わせを選ぶことでテキストを生成する．

3 実験

3.1 実験仕様

言語化の対象となる動作を「左手をあげる」「左手を下げる」「右手を上げる」「右手を下げる」「両手を下げる」「両手を上げる」「ボールを取る」「ボールを置く」の 8 つの基本動作から成ると定義する．テストデータには,「左手をあげる」「右手を上げる」「両手を下げる」「ボールを取る」の順で動作を行った Kinect 動画を使用し，それぞれの動作に対し自然言語での説明文を生成することとする．2.2 節で作成した動作判別器を交差検定で評価した結果，正答率が 90.9%となった．

表 3：言語資源を用いた生成文の上位 3 件

動作	生成文	尤度
1	• 右手, を, あげる, 。, null4, null5, null6, null7, null8, null9,	3.43e-30
	• 右手, を, あげる, 。, null5, null6, null7, null8, null9, null10,	1.92e-30
	• 右手, を, 上, に, あげる, 。, null4, null5, null6, null7,	6.54e-31
2	• 右手, を, 下げる, 。, null4, null5, null6, null7, null8, null9,	2.10e-30
	• 右手, を, 下, に, 下ろす, 。, null4, null5, null6, null7,	1.76e-32
	• 右手, を, 上, から, 下, に, 下ろす, 。, null4, null5,	1.85e-33
3	• 左手, を, あげる, 。, null5, null6, null7, null8, null9, null10,	7.72e-31
	• 左手, を, あげる, 。, null4, null5, null6, null7, null8, null9,	4.41e-31
	• 左手, を, 上, に, 挙げる, 。, null5, null6, null7, null8,	7.36e-32
⋮	⋮	⋮
18	• 右足, を, 下ろす, 。, null4, null5, null6, null7, null8, null9,	4.48e-31
	• 右足, を, 横, から, 下げる, 。, null4, null5, null6, null7,	1.48e-32
	• 右足, を, 横, に, ）, 下げる, null7, null8, null9, null10,	1.87e-33
19	• 左足, を, 横, に, あげる, 。, null6, null7, null8, null9,	4.57e-30
	• 左足, を, 横, に, あげる, 。, null7, null8, null9, null10,	1.23e-30
	• 左足, を, 横, に, あげる, 。, null5, null6, null7, null8,	1.17e-31
20	• 左足, を, おろす, 。, null4, null5, null6, null7, null8, null9,	2.18e-31
	• 左足, を, 横, に, ）, おろす, 。, null8, null9, null10,	6.22e-33
	• 左足, を, 横, に, 下ろす, 。, null4, null5, null6, null7,	3.24e-33

3.2 実験結果

構築した動作判別器を用いてテストデータから判別された中間表現は，順に

1. “up((left_hand),null)”
2. “up((right_hand),null)”
3. “down((left_hand,right_hand),null)”
4. “up((right_hand),green)”

となった．次に，選ばれた中間表現に対して予め構築されたバイグラムモデルに動的計画法を適用することで，動作を説明する尤もらしい文を生成する．

結果として，それぞれの動作に対して尤度の高かった上位 3 文を表 2 に示す．実験結果から，人の動作を正確に表現する文が生成出来ていることが確認できた．

4 ドメイン適用による言語資源拡充

前節まで，どの動きに対しても言語資源が十分にあるという前提であったが，実際には言語資源が十分でないことが多い．そこで，動作には組み合わせ構造があると仮定し，ドメイン適用による言語資源の拡充を試みる．ここでは，ドメイン適用に最小二乗推定を用いる．

4.1 最小二乗推定を用いたドメイン適用

要素トピックの確率分布パラメータを $\phi_i (i = 1, \ldots, K)$，観測データの確率分布パラメータを $\psi_i (i = 1, \ldots, M)$ とする．また，どの観測データも，down-front-hand-right といった 4 つの要素トピックの混合であると仮定する．要素トピックの混合行列を，$M \times K$ 行列である A とし，式 (1) の最小化問題を解くことで推定する．ここで，$\Phi = (\phi_1, \ldots, \phi_K)^T$，$\Psi = (\psi_1, \ldots, \psi_K)^T$ である．

$$\hat{\Phi} = \min_{\phi} ||\Psi - A\Phi||^2 = A^{+}\Psi \qquad (1)$$

4.2 実験

いま，対象とする動作は 20 あり，それらの動作には図 2 のような組み合わせ構造があるとする．

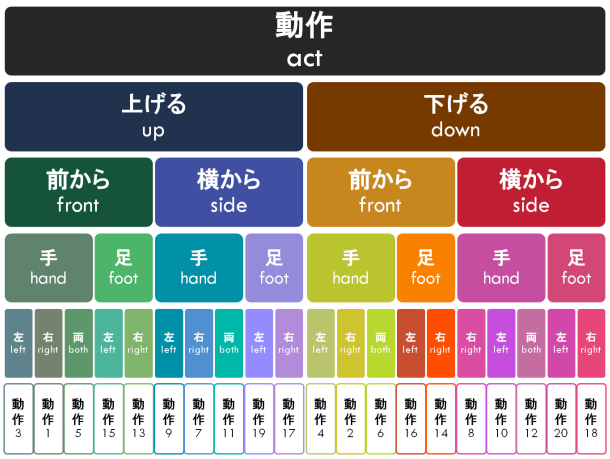

図 2: 実験データにおける組み合わせ構造のイメージ図

表 4：他の動作の言語資源のみを用いた生成文の上位 3 件

動作	生成文	尤度
1	• 右手, を, あげる, 。, null5, null6, null7, null8, null9, null10,	2.38e-31
	• 右手, を, あげる, 。, null4, null5, null6, null7, null8, null9,	1.66e-31
	• 右手, を, 上, に, あげる, 。, null5, null6, null7, null8,	1.69e-32
2	• 右手, を, 下げる, 。, null4, null5, null6, null7, null8, null9,	2.33e-31
	• 右手, を, 下げる, 。, null5, null6, null7, null8, null9, null10,	1.00e-31
	• 右手, を, 上, から, 下げる, 。, null4, null5, null6, null7,	6.85e-33
3	• 左手, を, 上, に, あげる, 。, null6, null7, null8, null9,	1.47e-31
	• 左手, を, あげる, 。, null5, null6, null7, null8, null9, null10,	9.02e-32
	• 左手, を, 上, に, あげる, 。, null5, null6, null7, null8,	1.48e-32
⋮	⋮	⋮
18	• 右足, を, 下げる, 。, null4, null5, null6, null7, null8, null9,	1.86e-32
	• 右足, を, 下げる, 。, null5, null6, null7, null8, null9, null10,	3.79e-33
	• 右足, を, 横, に, 下ろす, 。, null4, null5, null6, null7,	1.63e-33
19	• 左足, を, 横, に, あげる, 。, null6, null7, null8, null9,	5.92e-32
	• 左足, を, あげる, 。, null5, null6, null7, null8, null9, null10,	1.93e-32
	• 左足, を, 横, に, あげる, 。, null4, null5, null6, null7,	2.55e-33
20	• 左足, を, 下ろす, 。, null4, null5, null6, null7, null8, null9,	2.89e-32
	• 左足, を, 下ろす, 。, null5, null6, null7, null8, null9, null10,	6.53e-33
	• 左足, を, 横, に, 下ろす, 。, null4, null5, null6, null7,	6.43e-34

4.2.1 実験仕様

要素トピック数 $K = 9$，観測カテゴリ数 $M = 20$ とする．また，動的計画法を用いて尤もらしい文を生成するには，バイグラムの遷移確率と各単語の出現確率を求める必要があるため，要素トピックの確率分布パラメータと観測データの確率分布パラメータを遷移確率と出現確率の各々用意し，推定を行う．20 ある動作のうち一つの動作の持つ言語資源を削除し，4.1 節のドメイン適用を用いてその言語資源を推定するという実験を全ての動作を対象に行った．

4.2.2 実験結果

20 ある動作うち 6 動作に対して，言語資源を用いて生成された文の上位 3 件を表 3 に，他の動作の言語資源のみを用いて生成された文上位 3 件を表 4 に示す．

4.2.3 考察

表 3 と表 4 を比較してみると，若干表現は違うものの，他の動作の言語資源のみで人の動作を表現する文が生成出来ていることが確認できる．

5 おわりに

本研究では，動画像中の人の動作を表現する確率的言語生成の枠組みを提案した．Kinect ビデオで抽出された人の動作およびパーティクルフィルタで取得された物体の軌跡は，時系列データとして取得され，いくつかの次元圧縮手法を適用することで機械学習に適した形に変換された後，対数線形モデルで人の動作を表す中間表現の対応関係が学習される．また観測された人の動きを表現するために，被験者実験によって集められた自然言語文に基づきバイグラムモデルを構築し，動的計画法を適用することで，文生成に単語数の制限をつけずに自然言語文生成を行うことができた．さらに，言語資源が十分でないときを想定し，最小二乗推定による言語資源の拡充も試みた．

今後の課題として，構文制約などの知識を導入するとともに，より正確にイベントを説明するようなテキスト生成が行えるよう発展させていきたいと考える．

参考文献

[1] http://www.microsoft.com/en-us/kinectforwindows/

[2] 樋口知之: 粒子フィルタ，電子情報通信学会誌，Vol.88, No.12，pp.989-994，2005.

[3] Lin, J. et al. Lin, J., Keogh, E., Lonardi, S. and Chiu, B. : A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, DMKD' 03, 2003.