2F4-OS-01a-6

# 多層マルチモーダルLDAとHMMを用いた文法の学習

Learning of Grammar Using Multilayered Multimodal LDA and HMM

安東 裕司 *1 Yuji Ando
アッタミミ ムハンマド *1 Muhammad Attamimi
中村 友昭 *1 Tomoaki Nakamura
長井 隆行 *1 Takayuki Nagai
持橋 大地 *2 Daichi Mochihahii
小林 一郎 *3 Ichiro Kobayashi
麻生 英樹 *4 Hideki Asoh

*1電気通信大学 The University of Electro-Communication
*2統計数理研究所 The Institute of Statistical Mathematics
*3お茶の水女子大学 Ochanomizu University
*4産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

It is desirable for the robot to express their intentions through language in communication with humans. Regarding this, we have proposed multi-layered multimodal latent Dirichlet allocation (mMLDA) to enable the formation of various concepts and inference using the grounded concepts. The mMLDA, in conjunction with a grammar learning, makes it possible for the robot to verbalize the scene observed in daily life. The problem that is remain unsolved in the previous work is the handling of functional words. In order to deal with this problem, we propose to use Bayesian HMM for grammar learning. The experimental result reveals that the Bayesian HMM improves sentence generation performance by incorporating syntactic information.

## 1. はじめに

近年，高齢化社会に伴い，日常生活を支援するロボットの必要性が高まっている．しかし，ユーザとロボット間で円滑にコミュニーケションを取ることができず，利用の妨げになっている．この問題は，ロボットが観測するシーンを言語化できないことが原因の一つである．そのため，ロボットが観測する情報を自律的に言語化する能力が必要である．この際に言語をボトムアップに学習できれば，よりユーザに適応したコミュニケーションが実現できると考えている．

実環境においてロボットが観測する情報は，マルチモーダルな情報である．語意獲得は，ロボットが経験することによって取得するマルチモーダルな知覚情報のカテゴリ分類を基盤としており，Latent Dirichlet Allocation (LDA) をマルチモーダル情報の分類に適用した Multimodal LDA (MLDA) [Nakamura 09] によって実現できる．実際，人間の言葉もそのようなカテゴリに基づいており，ロボットがカテゴリ分類により概念を獲得することで，言語の理解・産出が可能になると考えられる．さらに著者らは文献 [Attamimi 14] において，動作を含む様々な概念とその階層性を考慮した概念モデルである multilayered MLDA (mMLDA) を提案し，概念間とその関係を教師なしで学習し，語意だけでなく文法を獲得できる可能性を示した．しかし，獲得した文法は助詞を考慮していないため，日本語として不自然な文が生成されるという問題がある．

そこで本稿では，mMLDA と Hidden Markov Model (HMM) を用いることで多様な概念を形成し，同時に語意や助詞を含む文法を獲得することで，観測したシーンに対して自然な文の生成を行う手法を提案する．mMLDA は多層の MLDA で構成されており，下層の MLDA では下位概念である，物体，動き，場所の概念がそれぞれ形成され，上層の MLDA ではこれらの概念を統合する上位概念が形成される [Attamimi 14]．このモデルを用いることで例えば，下位概念としてジュースという物体概念や物を口に運ぶという動き概念，ダイニングという場所概念などが形成される．上位層ではこれらの関係性が学習され，「飲む」という動作概念が形成される．これにより，ジュースを見ることでそれを口に運ぶ「飲む」という動作や，その「飲む」という動作が「ダイニング」という場所で行なわれやすいといった未観測情報の予測を行うことが可能となる．また，mMLDA は単語と概念の結び付きを学習することが可能であり，相互情報量を用いることで各単語に対応する，物体，動き，場所といった概念クラスの推定が可能である．しかし，文献 [Attamimi 14] では助詞など直接的に概念を表現しない語は扱えないという制約があった．つまり，学習の際に形態素解析が利用できる，もしくは教示文には助詞などが含まれていないという不自然な前提が必要であり，生成される文章にも助詞などが含まれないため，日本語としては不自然なものとなってしまう．

本稿では，この問題を解決するために，助詞などすべての単語を含んだ形で階層的概念や文法を学習することを考える．これは，文を単語に分解する教師なし形態素解析 [Mochihashi 09] を仮定すれば実現できるため，言語獲得の際にそもそも形態素解析が必要であるという不自然な仮定をする必要がない．ここで問題となるのは，概念に接地しない単語をどのように扱うかということである．本稿では，概念との相互情報量が小さく結びつきの弱いものを機能語として扱い，機能語を含む語順を教示文から学習することでこれを解決する．ただし，相互情報量だけでは単語と概念の結びつきを見出すことは難しい．そこで，この問題を統語情報を利用して解決する．文法の学習には Bayesian HMM を用い，その初期値として相互情報量による単語と概念の結びつきを用いる．最終的には学習した Bayesian HMM を文法として文を生成することで，助詞などを含む自然な文を生成することが可能となる．

関連研究として，文献 [Takano 12] における人の運動のモデル化，運動からの文生成を挙げることができる．しかし文献 [Takano 12] では，物体，動き，場所などの関係性は考慮していない．文献 [Teo 11] において，コーパスを用いた視覚情報

連絡先: 安東 裕司，電気通信大学大学院情報理工学研究科，東京都調布市調布ヶ丘 1-5-1，ando@apple.ee.uec.ac.jp

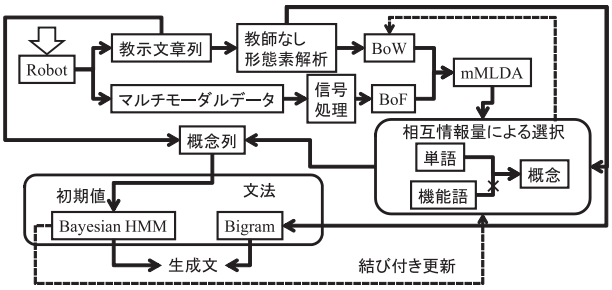


図 1: 言語学習・文生成システムの全体像

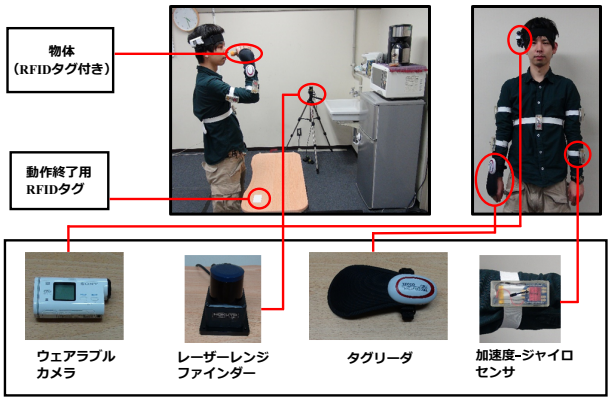


図 2: マルチモーダル情報の取得システム

からの文生成の枠組みが開発されているが，文の構造が単純である．また，動画中の人の動作を説明する文を生成する研究 [Yu 13] や，動画に映る調理の動作を説明する文を生成する研究 [Regneri 13] なども関連研究として挙げることができる．しかしこれらの研究では，視覚情報のみを扱っていること，また，文法が人手によって与えられているなど，本研究におけるボトムアップな言語の獲得とは方向性が異なるものであると言える．文献 [Kawai 14] では，幼児の統語発達を Bayesian HMM を用いてモデル化しており，非常に興味深い結果を得ている．一方本研究では，Bayesian HMM を概念と統合し，文生成に利用することを検討している点が大きな違いである．

## 2. 提案手法

### 2.1 提案手法の概要

提案する言語学習・生成システムの全体を図 1 に示す．ロボットは，人の行動を観測することで次に述べる mMLDA を用いて階層的に概念を形成する．同時にユーザが教示文を与えることで単語と形成された概念との結びつきを相互情報量を基準にして学習する．その際に，機能語が概念とは結びつかないことを閾値によって判定する．しかし，概念との相互情報量だけでは正確な単語と概念の結びつきを見つけることは難しい．そこで，提案システムでは Bayesian HMM を用いて統語情報を用いた文法の学習を実現する．情報量による結びつきは，Bayesian HMM を学習する際の初期値として利用する．一方で単語バイグラムを計算し，マルチモーダル情報から推定される単語を文法を用いて並べなおす際のスコアとして利用する．提案システムにおいても文章を単語に分解するための形態素解析器が必要となるが，文献 [Mochihashi 09] で提案されている教師なし形態素解析手法を用いることができる．

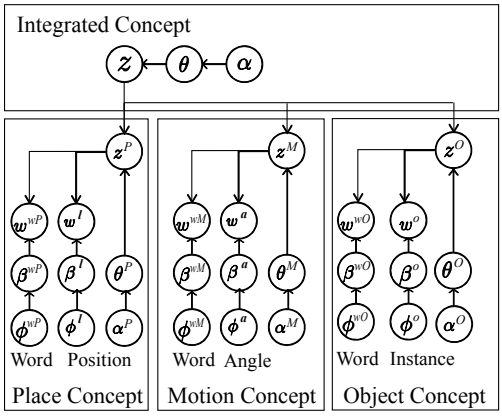


図 3: mMLDA のグラフィカルモデル

本稿では，ロボットがマルチモーダルなデータを取得するために，図 2 のシステムを用いることとする．これは，リアルな行動データから実際に言語を学習し文章を生成できることを示すためのセッティングであり，最終的には画像認識をベースとしたセンシングを目指している．

### 2.2 mMLDA

mMLDA は，下位層に物体，動き及び場所の分類モデルである MLDA を，上位層にそれらを統合する MLDA を配置することによって，物体，動き，場所それぞれの分類を行うと同時に，それらの関連性を教師なしで学習する統計モデルである．図 3 に，mMLDA のグラフィカルモデルを示す．図 3 において，$z$ は統合概念を表すカテゴリであり，$z^O$，$z^M$，$z^P$，はそれぞれ下位概念に相当する，物体，動き，場所カテゴリである．上位カテゴリ $z$ は，下位カテゴリ間の関係性を表現したモデルとなっている．また，$w^o$, $w^a$，$w^l$ は，それぞれ物体情報，物体を扱っている際の人の動き，位置情報である．さらに，$w^{wO}$，$w^{wM}$，$w^{wP}$ は，教示発話から得られる単語情報である．観測情報は図 2 のようなセンサを用いて取得する．以下それぞれの観測情報について簡単に述べる．

### 2.3 下位概念

人が使っている物体は，RFID タグを用いて認識を行う．物体に RFID タグを貼り付け，図 2 のタグリーダで触れることでタグの情報を読み取り，物体を認識することができる．物体情報として，物体番号ヒストグラム $w^o = (o_1, o_2, \cdots, o_{N_o})$ を用いる．ただし，$N_o$ は物体数を表している．$o_*$ は 0 または 1 の値をとり，物体番号 $k$ の物体が観測された場合 $o_k$ が 1 となる．さらに，全ての教示発話を単語分割し，Bag-of-Words（BoW）モデルを用いて表現したものを単語情報として扱う．

人の動きは，図 2 の 3 軸加速度センサ・3 軸ジャイロセンサを用いて，傾きを求めることで計測する．動き情報として，人の動作中の上半身の 5 箇所の 3 軸の傾きを，動作開始から動作終了までセンサの値を用いて取得することを前提とする．また動きの情報は，操作対象となる物体及び場所によって分節することができると仮定している．1 つの動作から複数の 15 次元の特徴ベクトルが得られ，それをあらかじめ計算した 70 の代表ベクトルによりベクトル量子化することで 70 次元のヒストグラムとして用いる．

人の場所は，図 2 の LRF(Laser Range Finder) を複数台用いて推定する．場所情報として，人の動作中の座標を動作開始から動作終了まで取得する．1 つの動作から複数の 2 次元の座標が得られるため，それを予め計算した 10 次元の代表ベクト

ルによりベクトル量子化することで 10 次元のヒストグラムとして用いる．そのため，場所情報として，1 つの人の動作中の場所ヒストグラム $w^l = (l_1, l_2, \cdots, l_{N_l})$ を用いる．

### 2.4 統合概念とパラメータ推定

mMLDA では，各概念を表す隠れ変数 $z$，$z^C \in \{z^O, z^M, z^P\}$ を同時に学習する．学習にはギブスサンプリングを用い，各概念を表すカテゴリ $z$，$z^C$ を，観測データ $w^m \in \{w^o, w^{wO}, w^a, w^{wM}, w^l, w^{wP}\}$ からサンプリングすることで学習する．サンプリングは，$\theta, \theta^C, \beta^m$ を周辺化した事後分布を用いる [Attamimi 14]．

### 2.5 言語学習と生成

#### 2.5.1 相互情報量を用いた単語の予測

本稿では，図 3 に示したように，各概念に教示発話から得られる助詞を含めた全ての単語情報を与えて学習を行う．各概念を表現する適切な単語が存在すると考え，単語とカテゴリの結び付きの強さの尺度として，単語とカテゴリ間の相互情報量を用いる．単語 $w^w$ と概念クラス $c \in$ { 物体概念，動き概念，場所概念 } のカテゴリ $k$ との間の相互情報量は以下の式となる．

$$I(w^w, k|c) = \sum_{K,W} P(W, K|c) \log \frac{P(W, K|c)}{P(W|c)P(K|c)}, \tag{1}$$

ただし，$K \in (k, \bar{k})$，$W \in (w^w, w^{\bar{w}})$ とし，$\bar{k}$ は $k$ 以外のカテゴリを表す．また，$w^{\bar{w}}$ は $w^w$ 以外の単語を表している．相互情報量とは，二つの確率変数の共有する情報量であり，相互依存の尺度である．したがって単語とカテゴリ間の相互情報量が大きい場合，その単語はそのカテゴリを表現しているといえる．

本稿では，単語によって，複数の概念を表す可能性があると考え，式 (1) を用いて求めた相互情報量を単語の各概念クラスに対する重みとして考える．その重みを用いて，次式で単語予測スコアを計算する．

$$\hat{P}(w^w|\boldsymbol{w}^m_{\rm obs}, c) \propto \max_k I(w^w, k|c) P(w^w, k|\boldsymbol{w}^m_{\rm obs}, c), \tag{2}$$

$$P(w^w, k|\boldsymbol{w}^m_{\rm obs}, c) = P(w^{wC}|k) P(k|\boldsymbol{w}^m_{\rm obs}). \tag{3}$$

このように，単語の各概念クラスに対する重みを求め，概念クラス $c$ の単語予測 $P(w^w, k|\boldsymbol{w}^m_{\rm obs}, c)$ の際に重みをつけることで，各概念から生成される単語の予測精度を向上させることができる．

また，単語の各概念クラスに対する重みが全て小さい時，その単語は概念と結びつかないことを意味する．したがって，この単語は助詞を含む機能語であると判断することができる．ただし，この判定は後の統語情報の学習における初期に利用するためのものであるため，それほど厳密に決定する必要はない．

#### 2.5.2 Bayesian HMM を用いた文法の学習

mMLDA を用いることで，観測情報を表現するのに適切な単語を予測することができる．文章を生成するためには，さらに文法を考える必要がある．本稿では，Bayesian HMM を用いて文法の学習を行う．Bayesian HMM は，教示文の単語列を入力として，各単語の助詞を含んだ概念クラスの推定を行う．概念クラスの推定には，ギブスサンプリングを用いる．この時，mMLDA による各単語の概念選択の結果を Bayesian HMM の初期値とする．また，事前にクラスの数を決定する必要がある．ここで文法は，学習後の概念クラスの遷移確率とし，$P(C_t|C_{t-1}) = N_{C_{t-1}C_t}/N_S$ となる．ただし，$C_t$ は文章中の $t$ 番目の単語に該当する概念クラスである．また，$N_{C_{t-1}C_t}$ と $N_S$ はそれぞれ $C_{t-1}$ から $C_t$ に遷移した回数と概念クラス間遷移の総数である．さらに，教示発話から単語のバイグラムを計算し，後で述べる文生成に利用する．

#### 2.5.3 観測情報からの文生成

まず，学習した文法に従い，“BOS” から “EOS” までの概念系列を $N$ 個サンプリングし，$n$ 番目の “BOS” と “EOS” を除いたサンプルを $\boldsymbol{C}^n = \{C^n_1, \cdots, C^n_t, \cdots, C^{T_n-1}_t\}$ とする．次に，概念 $C^n_t$ から，概念と対応した単語を生成する．ここでは，観測情報 $\boldsymbol{w}^m_{obs}$ が与えられたとき，概念 $C^n_t$ と対応した単語の発生確率の高い上位 $K$ 個の単語 $\boldsymbol{w}^n_t = \{w^n_{t1}, w^n_{t2}, \cdots, w^n_{tK}\}$ を用い，全ての単語の集合を $\boldsymbol{W}^n = \{\boldsymbol{w}^n_1, \boldsymbol{w}^n_2, \cdots, \boldsymbol{w}^n_{T_n-1}\}$ とする．すなわち，これらの概念系列・単語から $K^{T_n-2}$ パターンの文を生成することができ，文 $S^n$ が生成される確率を次のように定義する．

$$P(S^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{obs}) \propto \prod_t P(C^n_t|C^n_{t-1}) P(w^n_t|\boldsymbol{w}^m_{obs}, C^n_t) P(w^n_t|w^n_{t-1}) \tag{4}$$

そして，観測情報に対してサンプリングされた $N$ 個の概念系列と単語から生成した文の中から，生成される確率が高い文を選択する．まず，各概念系列に対して，各概念系列から，式 (4) を最大とする文 $\hat{S}^n$ を決定する．ただし，$S^n$ のパターンは非常に多く単純には決定することができないため，Viterbi アルゴリズムを用いて式 (4) の確率が最大となる文を探索する．ここで，各概念系列のサンプルに対して確率が最大となる文の集合を $\hat{\boldsymbol{S}} = \{\hat{S}^0, \cdots, \hat{S}^n, \cdots, \hat{S}^N\}$ とする．

次に，$\hat{\boldsymbol{S}}$ から生成される確率が最大となる文を選択することで，最終的な生成文とする．しかし，実際には文が長いほど，その確率は小さくなってしまう．そこで以下の様な，調整係数 $\ell(\hat{S}^n)$ を導入する．

$$\ell(\hat{S}^n) = \frac{(L^{\max} - L_{\hat{S}^n})}{\sum_n^N L_{\hat{S}^n}} \sum_n^N P(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{obs}) \tag{5}$$

ただし，$L_{\hat{S}^n}$ は $\hat{S}^n$ の文の長さ，$L^{\max}$ は $\hat{\boldsymbol{S}}$ の中の文長の最大値である．式 (5) を用いて，文の確率を次のように再定義する．

$$\bar{P}(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{obs}) = P(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{obs}) + \omega \ell(\hat{S}^n) \tag{6}$$

ここで，$\omega$ は文の長さを調整する重み付けである．重みを大きくするほど，文が長くなる．従って最終的な文 $S$ は，$S = {\rm argmax}_{\hat{S}^n \in \hat{\boldsymbol{S}}} \bar{P}(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{obs})$ によって求める．

## 3. 実験

図 2 のシステムを用いて人が動作するマルチモーダル時系列データを用いて実験を行った．実験では，表 1 の 16 通りの動作を使用し，文生成を行った．

### 3.1 概念選択

まず，相互情報量を用いて学習データ内の単語の各概念に対する重みを求め，単語の概念を選択した．ただし，各単語についての相互情報量が最も高い概念を，その単語の概念として選択した．また，相互情報量が閾値よりも小さい場合は，その単語を助詞とした．単語選択の結果を図 4 に示す．図 4 において，横軸が各概念を表し，縦軸が単語を表している．図 4(a) は，事前に定義した各単語と概念の結びつきである．図 4(b) は相互情報量のみを用いた時の概念選択の結果であり，正

表 1: 動き，物体，場所データの対応表 (カッコ内の数字はカテゴリ ID).

| 動き | 物体 | 場所 |
|---|---|---|
| 飲む (1) | お茶 (1) | ソファー (3) |
| | ジュース (2) | リビング (1) |
| 食べる (2) | クッキー (3) | ダイニング (4) |
| | ポテチ (4) | キッチン (2) |
| 拭く (3) | 雑巾 (5) | キッチン (2) |
| | ティッシュ(6) | リビング (1) |
| 注ぐ (4) | お茶 (1) | ダイニング (4) |
| | ジュース (2) | キッチン (2) |
| つける (5) | エアコン (7) | 寝室 (5) |
| | テレビ (8) | リビング (1) |
| 振る (6) | お茶 (1) | キッチン (2) |
| | ドレッシング (9) | ダイニング (4) |
| 読む (7) | 参考書 (10) | 寝室 (5) |
| | 雑誌 (11) | ソファー (3) |
| 上に投げる (8) | ボール (12) | ソファー (3) |
| | ぬいぐるみ (13) | 寝室 (5) |

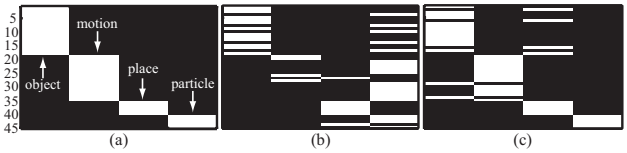


図 4: 概念の選択結果:(a) 正解，(b) 相互情報量，(c)HMM

解率は 57.78%であった．さらに，この時の概念選択の結果を Bayesian HMM の初期値として学習を行い，単語に対する概念選択を行った結果を図 4(c) に示す．Bayesian HMM による概念選択の結果，正解率は 86.67%であった．以上の結果から，相互情報量のみから単語の概念選択を行うよりも，Bayesian HMM で学習することによって，より正しく概念を選択することができることが分かる．

### 3.2 文生成

提案手法を用いて獲得した文法を図 5 に示す．図 5(a) は，相互情報量を用いた単語の概念選択結果から得られた文法 (G1) である．図 5(b) は，提案手法である Bayesian HMM で学習して得られた概念クラスの遷移確率を用いた文法 (G2) である．

また，これら獲得した文法を用いて観測情報を基に文生成を行った．図 6 に，生成された文の例を示す．G1 では，概念選択の誤りから誤った文法が学習され，助詞から助詞への遷移や，助詞から EOS に遷移することがある．そのため，図 6 のように誤った文章が多く生成されてしまう．しかし，Bayesian HMM で学習を行った G2 では，必ず概念から助詞，助詞から概念に遷移する．そのため，観測情報から正しい文を生成することができている．しかし，図 6 の右側の例では，誤った文が生成されている．これは，mMLDA によって分類を行った際に，「拭く」という動きが「つける」という動きに誤って分類されたことが原因である．従ってこれは，Bayesian HMM で獲得した文法の問題ではない．

## 4. 結論

本稿では，多様な概念を形成し，自然な文生成を行うために，mMLDA と Bayesian HMM を用いる手法を提案した．mMLDA と相互情報量によって，概念との結びつきの弱いものを助詞として扱い，その後に Bayesian HMM で統語範疇を学習することで，文法を獲得できる可能性があることを示した．また，Bayesian HMM により獲得した文法を用いることで，観測シーンに対して助詞を含んだ自然な文を生成できることを示した．本稿で示した結果は，単語数が少なく，文章も非常に単純なものである．今後，より複雑なデータセットでの評価実験を行い，どの程度の複雑さまで対応可能であるかを検討する予定である．また，ノンパラメトリックベイズの適用により，カテゴリ数や品詞数を自動推定することを検討したい．

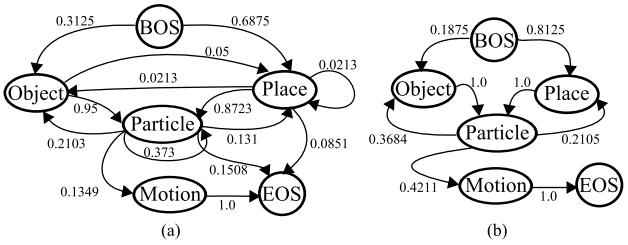


図 5: 獲得した文法:(a) 相互情報量を用いた概念選択，(b) 提案手法

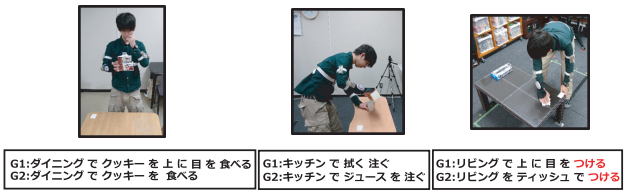


図 6: 観測情報から生成された文の例

**謝辞**

本研究は，JSPS 科研費 26280096 の助成を受けて実施した．

## 参考文献

[Nakamura 09] T. Nakamura *et al.*, “Grounding of Word Meanings in Multimodal Concepts Using LDA, ” in Proc. of IROS 2009, pp.3943–3948, 2009.

[Attamimi 14] M.Attamimi *et al.*, “Integration of Various Concepts and Grounding of Word Meanings Using Multi-layered Multimodal LDA for Sentence Generation, ” in Proc. of IROS 2014, 2014.

[Takano 12] W. Takano *et al.*, “Bigram-Based Natural Language Model and Statistical Motion Symbol Model for Scalable Language of Humanoid Robots, ” in Proc. of ICRA 2012, pp.1232–1237, 2012.

[Teo 11] C. L. Teo *et al.*, “A Corpus-Guided Framework for Robotic Visual Perception, ” in Proc. of AAAI 2011, 2011.

[Yu 13] H.Yu *et al.*, “Grounded Language Learning from Video Described with Sentencest, ” in Proc. of ACL 2013.

[Regneri 13] M. Regneri *et al.*, “Grounding Action Descriptions in Videos, ” Transactions of the Association for Computational Linguistics, vol.1, pp.25–36, 2013

[Mochihashi 09] D. Mochihashi *et al.*, “Bayesian unsupervised word segmentation with nested pitman-yor language modeling,” in Proc. of ACL, pp.100-108, 2009.

[Kawai 14] Y. Kawai *et al.*, “Computational Model for Syntactic Development: Identifying How Children Learn to Generalize Nouns and Verbs for Different Languages,” in Proc. of ICDL-EPIROB, pp.78–84, 2014.