3O4-OS-04b-4

多層マルチモーダルLDAと隠れセミマルコフモデルを用いた概念・語彙・文法の相互学習

Mutual Learning of Concept and Language Using Multilayered Multimodal LDA and HSMM

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

In this paper we propose the use of hidden semi Markov model (HSMM) for grammar learning. Since the HSMMs can segment phoneme sequences into words unsupervised and , the system can learn language from scratch. The mMLDA, in conjunction with the HSMM-based grammar learning, makes it possible for the system to verbalize the scene observed in daily life. This paper also examines a bootstrapping method for learning concepts and language iteratively. Some promising results were obtained by the proposed method.

1. はじめに

ロボットが観測したシーンを言語で表現する能力は人とコミュニケーションをするために重要な能力である．また，言語獲得のメカニズムに構成論的に迫ることも重要な課題である．本稿では，スマートハウスをロボットと捉え，センシングした情報から人の日常生活における活動を言語化することを考える．これは，観測した内容を言語で報告する監視システムや，日記のようにユーザの活動を自動的に記録するシステム等に応用することもできる．本稿では，システムが取得可能なマルチモーダル情報から，教師なし学習によってボトムアップに言語を学習することで，シーンを文章で表現する手法を検討する．

筆者らは従来シーンを文章化するためのモデルとして，multilayered Multimodal Latent Dirichlet Allocation (mMLDA) [Attamimi 14, Attamimi 15] を提案した．しかし，教示文を形態素解析器を用いて単語分割し，助詞を除いた単語情報を使用しているため，生成される文章は助詞が含まれない不自然なものになってしまうという問題があった．また，一般に形態素解析器には教師あり学習が用いられ，学習コーパスの多くは書き言葉，新聞データを対象として作られている．そのため，ユーザ独自の表現などに対応することができない．本稿では，これらの問題を解決するために，Hidden Semi Markov Model (HSMM) [Uchiumi 15] を用いる．HSMM を用いることで，与えられた文から教師なしで単語分割を行い，同時に品詞 (概念クラス) の推定を行うことが可能である．品詞の推定を行うことによって，統語情報を考慮した確率的文法を獲得できる．さらに文献 [Attamimi 15] では，各概念に対して同じ単語情報を与えて学習を行う．つまり，物体概念の学習を行う際に，動きや場所を表す単語の情報も与えて学習を行っていることになる．提案手法では，HSMM の品詞の推定結果を用いて各概念と単語の結び付を更新し繰り返し学習することで，概念と言語の学習精度を相互に向上させる．

関連研究として，深層学習を用いた画像に対する説明文生成の研究を挙げることができる [Vinyals14, Fang 14]．これらの手法は，非常に高い精度で説明文を生成できるため近年注目されているが，end-to-end 学習のため，内部でどのような学習がなされているかを把握することが困難であるという問題がある．例えば，どのような語意が獲得され，どのような文法が表現されているのか，なぜうまく文が生成されるのか（されないのか）といったことを把握することが難しい．このことは，システムの性能改善という意味だけでなく，言語獲得のモデルという点でも問題となると考えられる．

連絡先: 長井 隆行，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，tnagai@ee.uec.ac.jp

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像を，図 1 に示す．まず，言語学習について説明する．システムは，人の行動を観測し，人物，視覚，動き，場所情報を得る．マルチモーダル情報は，2.3 節に述べる信号処理を行い，Bag-of-Features (BoF) モデルとして扱う．ユーザが与える教示文は，2.2 節に述べる HSMM を用いて単語分割を行い，Bag-of-Words (BoW) モデルとして扱う．処理したマルチモーダル情報及び単語情報を mMLDA により学習し，多様な概念を形成する．同時に単語と形成された概念との結びつきを概念と単語の相互情報量を基準として学習する．その際，全ての概念との相互情報量が小さい単語を機能語と判定する．相互情報量による結びつきは，HSMM の単語分割と確率的な文法を学習する際の初期値として利用する．本稿ではさらに，HSMM の出力結果に従って単語情報を更新する．更新した単語情報を用いて mMLDA で概念を形成し，再度 HSMM で文法を学習することで，概念・文法を相互に学習することが可能である (2.5 節を参照)．

文生成では，まず学習した確率的文法を用いて概念列を生成する．生成された各概念に対して mMLDA により観測情報から単語が生成される確率と HSMM により文法から単語が生成される確率を考慮し，単語の生成を行う．これによって，概念列から単語列を得ることができる．最終的に，単語列の候補を複数生成し，言語モデルを用いて適切な文を選択する．

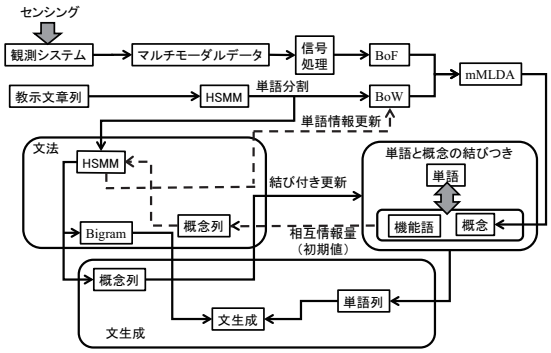

図 1: 言語学習・生成システムの概要

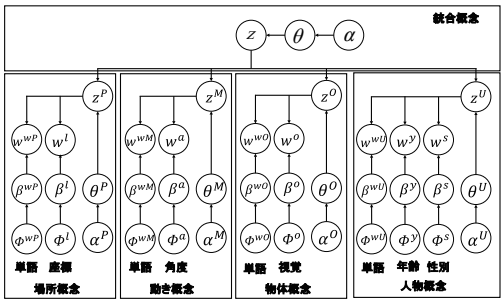

図 2: mMLDA のグラフィカルモデル

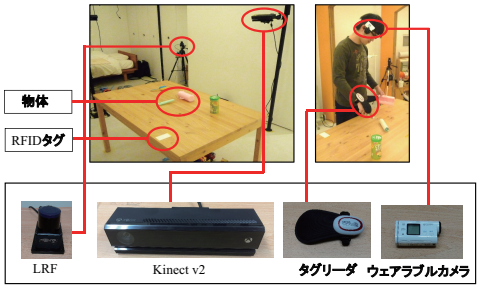

図 3: マルチモーダル情報の取得システム

2.2 HSMM

2.2.1 教師なし形態素解析

形態素解析の問題を，文字列 $s = l_1 l_2 \ldots l_N$ が与えられた際に，s を分割して得られる単語列及び単語列に対応した品詞列の確率 $P(\boldsymbol{w}|s)$ を最大化する問題として考える．ここで，$\boldsymbol{w} = w_1 w_2, \ldots w_M, c_1 c_2 \ldots c_M$ であり，w_n，c_n はそれぞれ単語と品詞を表す．$P(\boldsymbol{w}|s)$ を，次のように置くことで部分問題に分割する．

$$P(\boldsymbol{w}|s) = \prod_{i=1}^{M} P(w_i|c_i)P(c_i|c_{i-1}) \tag{1}$$

ここで，i 番目の単語は i 番目の品詞のみに，i 番目の品詞は $i-1$ 番目の品詞のみ依存すると仮定する．単語の境界は与えれていないため，$\boldsymbol{w}$ についても観測できない．観測できるのは文字列 s のみである．これは，s の部分文字列からなるセグメントを単語候補として品詞が加わった HSMM となっている．

2.2.2 単語分割とサンプリング

学習アルゴリズムとして，動的計画法とマルコフ連鎖モンテカルロ法を組み合わせた手法を用いる．単語分割と品詞列は両方隠れ変数とみなし，同時にサンプリングするため，単語と品詞の同時確率を求める必要がある．ここで，Forward-filtering における前向き確率は (2) 式の再帰式のようになる．ただし，$\alpha[t][k][c]$ は位置 $t-k$ から t までの長さ k の文字列 l_{t-k}^{t} が品詞 c の単語として生成される確率を表す．C は品詞 (概念) クラスの数を表す．

$$\alpha[t][k][c] = \sum_{j=1}^{t-k} \sum_{r=0}^{C} P(l_{t-k}^{t}|c)P(c|r)\alpha[t-k][j][r] \tag{2}$$

$\alpha[t][k][c]$ が求まると，文末から単語分割と品詞を同時にサンプリングすることができる．これを確率に従って，文末から文頭まで繰り返すことで，単語と品詞のサンプリングを行う．

2.3 mMLDA

mMLDA は，下位層に複数の 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，人物，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 2 に mMLDA のグラフィカルモデルを示す．図 2 において，z は統合概念を表すカテゴリであり，z^U，z^O，z^M，z^P，はそれぞれ下位概念に相当する，人物，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，$w^m \in \{w^s, w^y, w^o, w^a, w^l\}$ は，それぞれ人物情報，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，$w^{wC} \in \{w^{wU}, w^{wO}, w^{wM}, w^{wP}\}$ は，教示文から得られる単語情報である．観測情報は図 3 に示すシステムを用いて取得する．パラメータの推定にはギブスサンプリングを用いる．

本稿では，机においてある物体を使用して動作を行うことを仮定する．人物情報は，Kinect v2 で取得した動作直前の人物の画像から，性別・年齢の推定を行う．性別・年齢の推定の結果を基に，データの量子化を行い，2 次元と 7 次元のヒストグラムとする．人が扱っている物体は，Kinect v2 で検出を行う．物体情報として，使用している物体の切り出し画像から畳み込みニューラルネットワーク（CNN）によって，4096 次元の特徴量を抽出し，4096 次元のヒストグラムとして扱う．人の動きは，Kinect v2 を用いて骨格情報を計測する．椅子に座って動作を行うため，上半身部分の骨格情報を用いる．上半身では，17 関節の骨格情報を取得することが可能であり，1 つの関節から 3 軸の傾きを取得する．動き情報として，1 つの動作から複数の 51 次元の特徴ベクトルが得られ，それを予め計算した 70 次元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する．人の位置は，図 3 の LRF 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計測する．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する．

2.4 言語学習

2.4.1 単語予測

提案手法では，図 2 に示したように，各概念に教示文から得られる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，単語と概念の結び付きの強さの尺度として，単語と概念間の相互情報量を用いる．相互情報量とは，二つの確率変数が共有する情報量であり，相互依存の尺度であ

る．従って単語と概念間の相互情報量が大きい場合，その単語はその概念を表現していると言える．一方で，全ての概念に対して相互情報量の値が小さくなる単語は機能語であると考えることができる．提案手法において，相互情報量による判定は文法学習の初期値として利用する．

HSMM によって学習された品詞 (概念クラス) を用いることで，観測情報 $\boldsymbol{w}^m_{\rm obs}$ から単語 w^{wC} の予測を以下のように行うことができる．

$$\hat{P}(w^{wC}|\boldsymbol{w}^m_{\rm obs}, c) \propto \max_k P(w^{wC}|c)P(w^{wC}|k)P(k|\boldsymbol{w}^m_{\rm obs}, c) \quad (3)$$

ただし，$P(w^{wC}|c)$ は HSMM の出力確率を表しており，$P(w^{wC}|k)$ と $P(k|\boldsymbol{w}^m_{\rm obs}, c)$ は mMLDA より計算することができる．また，k は概念クラス $c \in \{$ 人物, 物体, 動き, 場所, 機能語 $\}$ のカテゴリを表している．ここで “機能語’ は，観測情報から mMLDA を用いて予測できないため，$P(w^{wC}|k)$ 及び $P(k|\boldsymbol{w}^m_{\rm obs}, c)$ は一様分布と仮定する．つまり，HSMM によって学習される統語情報のみによって機能語が決定される．

2.4.2 HSMM を用いた文法の学習

先に述べた HSMM は，教示文を単語分割を行い，同時に品詞の推定を行う．この際，mMLDA による各単語の概念選択の結果を HSMM 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する．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 概念・語彙・文法の相互学習

本稿では，mMLDA の単語情報に全て同じ情報を与えて学習を行っている．つまり，物体概念を学習する際に，動きや場所などを表す単語の情報も含まれて学習していることになる．HSMM は，品詞の推定を行っているので，どの単語がどの品詞であるのかを確率的に求めることが可能である．そこで，HSMM の出力確率 $P(w^{wC}|c)$ に従って，単語情報を各概念毎に更新する．それぞれの単語情報を用いて mMLDA で概念を形成し，再度 HSMM で文法を学習することによって概念・文法を相互に学習することが可能となる．

3. 実験

図 3 に示すシステムを実際にスマートハウスに設置し，人が物体を扱う動作のマルチモーダル情報を取得した．実験では，7 人の女子学生，5 人の男子学生，3 人の大人の男性の計 15 人の被験者が表 1 に示す動作を行った．3 人の女子学生，3 人の男子学生，2 人の大人の男性の計 8 人のデータ (497 シーン) を学習用データとし，残りの 7 人のデータ (442 シーン) をテスト用データとした．また，各動作に対して 10 文ずつ教示文を与えた．実験では，HSMM と同様に品詞の推定が行える Bayesian Hidden Markov Model (BHMM) を比較手法として用いる．BHMM は，予め分割された単語の品詞を推定するため，教示文を形態素解析器で分割した単語を使用する．提案手法である HSMM は，単語分割と同時に品詞を推定するため，BHMM よりも厳しい条件である．BHMM 及び HSMM は供に，品詞数を設定する必要がある．以下では，最も結果の良かった品詞数である 7 (BHMM)，4 (HSMM) を用いた際の結果を示すこととする．

3.1 単語分割

獲得した 1020 文を HSMM を用いて単語分割をした．“じょせいがだいにんぐでおちゃをのむ” という教示文に対して，形態素解析器を用いた場合，” じょせい/が/だいにんぐ/で/おちゃ/を/のむ” という分割になる．これに対して HSMM を用いると，” じょせいが/だいにんぐ/で/おちゃを/のむ” という分割となった．比較すると，提案手法は助詞が分割できていないことが分かる．形態素解析器で分割したものを正解として，HSMM で分割した結果の適合率，再現率，F 値を計算した．適合率，再現率，F 値はそれぞれ，95.73%，44.76%，60.99%であり，適合率が高く，再現率が低いという結果になった．これは，単語の切れ目を入れる数が少ないことを意味する．分割の例からも分かるように，教示文に対して助詞が上手く分割できていない．これは，HSMM で学習した教示文の多くが,「〜が〜で〜を〜する」のような言い回しであったとことが原因だと考えられる．つまり，言い回しが少ないため，助詞がつながってしまった可能性が高い．また，HSMM の品詞数が 4 の場合に全体的な結果が良くなった理由は，助詞が分割されずに 4 つの品詞クラスにマージされてしまったためである．

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

動き	物体	場所	人物	動き	物体	場所	人物
飲む (1)	お茶 (1)	ダイニング (1)	全員 (1,2,3)	投げる (8)	ぬいぐるみ (15)	ソファ(3)	息子 (2)
	ジュース (2)	リビング (4)	娘, 父 (1,3)		ボール (16)	寝室 (5)	息子, 父 (2,3)
	コーヒー (3)	ソファ(3)	父 (3)		クッション (17)		息子 (2)
食べる (2)	クッキー (4)	ダイニング (1)	全員 (1,2,3)	読む (9)	本 (18)	ソファ(3)	娘, 父 (1,3)
	ポテチ (5)	リビング (4)		吹きかける (10)	スプレー (6)	寝室 (5)	息子, 父 (2,3)
振る (3)	コーヒー (3)	ソファ(3)	父 (3)		霧吹き (19)	ソファ(3)	娘, 父 (1,3)
	スプレー (6)	ダイニング (1)	息子, 父 (2,3)	抱く (11)	ぬいぐるみ (15)	ソファ(3)	娘 (1)
	マラカス (7)	リビング (4)	娘 (1)		クッション (17)	寝室 (5)	
	ドレッシング (8)	ダイニング (1)	全員 (1,2,3)	開ける (12)	お菓子の袋 (2 4)	ソファ(3)	全員 (1,2,3)
注ぐ (4)	お茶 (1)	キッチン (2)	全員 (1,2,3)	書く (13)	本 (18)	リビング (4)	息子, 父 (2,3)
	ジュース (2)		娘, 息子 (1,2)	動かす (14)	車 (20)	リビング (4)	息子 (2)
かける (5)	ドレッシング (8)	ダイニング (1)	全員 (1,2,3)		コロコロ (21)		娘, 父 (1,3)
	ソース (9)			拭く (15)	雑巾 (22)	ダイニング (1)	息子, 父 (2,3)
	調味料 (10)	キッチン (2)	娘 (1)		モップ (23)	寝室 (5)	娘, 父 (1,3)
	じょうろ (11)	寝室 (5)		絞る (16)	雑巾 (22)	キッチン (2)	息子, 父 (2,3)
振る (6)	たわし (12)	キッチン (2)	娘, 息子 (1,2)	つける (17)	リモコン (25)	寝室 (5)	息子, 父 (2,3)
	スポンジ (13)						
包む (7)	ラップ (14)	キッチン (2)	娘 (1)				

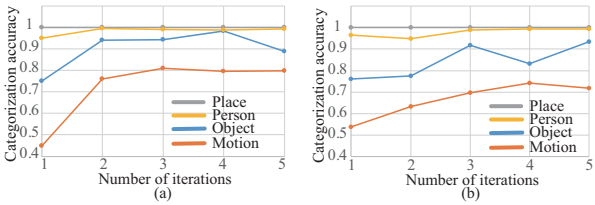

図 4: 相互学習による概念の各概念の分類精度の推移:(a) 比較手法，(b) 提案手法

3.2 概念形成

概念と言語の繰り返し相互学習により，各概念がどれほど正しく形成されたかを検証した．結果を図 4 に示す．図 4 (a) が BHMM を用いた結果，(b) が提案手法である HSMM を用いて相互学習を行った結果である．これらの結果から，BHMM と HSMM には大きな精度の違いがないことが分かる．BHMM は形態素解析器を用いており，単語の分割が正確であるため，繰り返し 2 回でほぼ性能が飽和している．一方 HSMM は，単語への分割が変化するため，分類の性能がすぐに収束しないことが分かる．5 回の繰り返しで大きな差はないが，BHMM がわずかに上回る結果となった．いずれの場合も，動き概念に関する精度が低いのは，人によって動き方が異なることが原因である．例えば，お茶を飲むときに蓋を開けて飲む動作を行う人もいれば，ただ口にペットボトルを運ぶだけの人も存在した．さらには，骨格情報も常に安定して取得できるわけではなく，ノイズが多いデータである．物体概念は，使用した物体を上手く検出することができず，他の物体が検出されていることがあった．また，物体全体を検出することができず，一部しか検

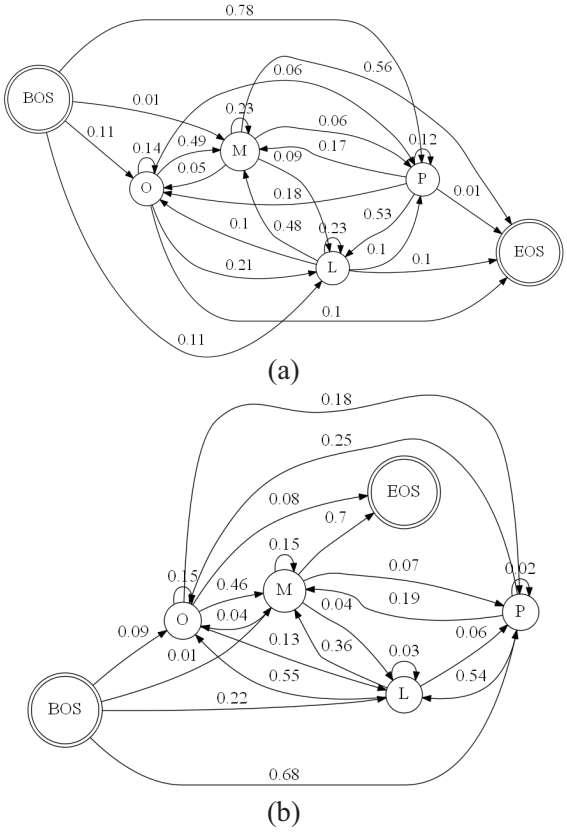

図 5: 学習された文法 (a)1 回の繰り返し (b)5 回の繰り返し

出されていない場合もあり，分類精度に影響が出たと考えられる．こうした様々な影響の中で学習を繰り返すことで，概念同士や言語が相互に影響することで精度を向上させている．

3.3 文法

ここでは，言語と概念を繰り返し相互に学習することで，システムがどのような文法を学習したかを検証する．図 5 に，提案手法の繰り返し 1 回目と 5 回目にシステムが獲得した文法を示す．繰り返し 1 回目では，概念の形成や語彙との結びつきが十分に学習されていないため，文法が複雑になっていることが分かる．一方，5 回の繰り返しで全体の精度が向上し，獲得された文法もある特定の概念間での遷移が起こるように学習されていることが分かる．但し，今回の実験で与えている文章は非常に簡単であるため，今後より複雑で多様な文を与えた時にどのような学習がされるかを検証する必要がある．

3.4 文生成

テスト用データを用いて文生成を行った．生成された文に対して人手による評価を行った．各文に対して，次の 4 つのカテゴリのどれに当てはまるか選んだ．E1: 文法が正しいかつ意味も正しい，E2: 文法は正しいが意味が正しくない，E3: 文法は正しくないが意味が正しい，E4: 文法が正しくないかつ意味も正しくない．結果を表 2 に示す．比較手法である BHMM は，相互学習によって，品詞の推定精度が向上し，E1 が高く，E4 が減少した．提案手法の相互学習前は，E2 が高く，観測シーンに対して誤った意味の文を多く生成している．これは，品詞の推定が大きく誤っていたことが原因である．しかし，相互学習後は，品詞の推定が修正され文生成の精度が向上した．

これまでの実験結果が示していることは，HSMM では単語

表 2: 生成文の評価結果.

	比較手法 (相互学習前)	比較手法 (相互学習後)	提案手法 (相互学習前)	提案手法 (相互学習後)
E1	9.28%	37.33%	37.10%	**40.95%**
E2	12.44%	22.28%	**58.60%**	51.13%
E3	16.06%	**18.55%**	0.45%	4.75%
E4	**62.22%**	21.27%	3.85%	3.17%

分割において助詞がうまく分割できていないこと，そして概念形成の精度は HSMM と BHMM で大きな差がないことであった．これは HSMM が単語分割と同時に品詞の推定を行っており，BHMM が単語分割を与えていることが要因である．しかし，単語分割の問題にもかかわらず，文生成においては HSMM の方が良い結果となった．特に E3 や E4 の文法的な誤りがほとんど見られない．これは，HSMM によって助詞がうまく区切れていないことが逆に文の構造が簡単化し，文法の学習を容易にしていることが要因であると考えられる．つまりこのことは，最終的な文を生成する際には，単語や文法といった複数の要素が絡み合っており，それらがお互いに補い合っていることを意味している．一方で HSMM によって学習された文法では，多様な言い回しを生成することはできない．そのためには助詞を正しく分割し，より複雑な文法を学習する必要がある．より多くのバリエーションを含む教示文を与えた場合に，どのように語彙や文法が変化していくかを今後検証する必要がある．

4. 結論

本稿では，日常生活における人の活動をセンシングし，知能システムのための観測情報から自然な文を生成するボトムアップな枠組みを提案した．提案手法は，多様な概念の形成や予測が可能な mMLDA と，単語分割と同時に統語情報を学習することができる HSMM とを統合したものである．BHMM に比べて HSMM は単語分割と同時に概念クラスの推定を行うため，単語分割の推定精度が劣るが，観測シーンに対して自然な文を生成することが可能であった．今後の課題として，バリエーションを多くした教示文での学習や言語生成を行うことや言語の理解を行うことが挙げられる．また，ノンパラメトリックベイズの適用により，概念数を自動推定することも今後の課題である．

謝辞

本研究は，JSPS 科研費 26280096 及び JST CREST の助成を受けて実施した．

参考文献

[Attamimi 14] Attamimi M. et al., “Integration of Various Concepts and Grounding of Word Meanings Using Multi-layered Multimodal LDA for Sentence Generation,” IROS2014, pp.2194–2201, 2014

[Attamimi 15] Attamimi M. et al., “Learning Word Meanings and Grammar for Describing Everyday Activities in Smart Environments,” EMNLP 2015, pp.2249–2254, 2015

[Uchiumi 15] Uchiumi K. et al., “Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models,” ACL-IJCNLP 2015

[Vinyals14] Vinyals O. et al., “Show and Tell: A Neural Image Caption Generator,” In arXiv:1411.4555 [cs.CV]. 2014

[Fang 14] Fang H. et al., “From Captions to Visual Concepts and Back,” CVPR 2015.