2F4-OS-01a-5

物体指向動作認識を伴う対話におけるトピック管理

Topic Management in Dialogue with Object-oriented Action Recognition

渥美雅保
Masayasu Atsumi

創大・工・情報システム工学
Dept. of Information Systems Sci., Faculty of Eng., Soka University

This paper proposes a method of probabilistic topic management for dialogue in conjunction with object-oriented action recognition. In the method, a probabilistic semantic network ACTNET is learned to infer object and action labels from their visual appearance and a probabilistic topic network is learned to manage dialogue context according to the object and action labels. Through experiments using 3D videos and a web-scraped corpus, it is shown that the method works in dialogue to select sentences on topics in accord with object-oriented action recognition.

1. はじめに

日常生活空間において人が何をしているのかをロボットが理解することはロボットが人を自律的に支援するうえで必要不可欠な機能である．また，人が注意している物体やそれに働きかける動作は会話のトリガーとなり，人とロボットの日常会話を介しての共生を促進する．人の動作には物体への働きかけ動作が多くみられる．本研究では，この動作を「物体指向動作」と呼び，物体指向動作の認識を介して非タスク指向の対話を行う問題を扱う．

物体指向動作認識を伴う対話のためには，物体指向動作の認識，対話のためのトピック管理，対話のトピック意味空間への物体指向動作認識のマッピングが必要になる．本研究では，これらに関して次の特徴を持つ物体指向動作認識を伴う対話方法を提案する．まず，物体指向動作認識に関して，物体指向動作を＜対象意味素 (target synset), 格 (case), 動作意味素 (motion synset) ＞の格 3 つ組でラベル付けし，視覚認識カテゴリとこれら意味素をノードとする「アクトネット (ACTNET)」と名付ける確率意味ネットワークを学習することにより，視覚認識カテゴリからその意味素の推論を可能にし，物体指向動作に言語的意味を与える [渥美 14, Atsumi 14]．次に，会話のためのトピック管理に関して，トピックモデルに基づいて文脈のトピック依存構造を学習した「文脈トピック依存ネット」と名付ける確率トピックネットワークを生成し，それを用いて対話のトピック管理のもとで発話文の選択を行う．また，ACTNET から推論された物体指向動作の格 3 つ組を，文脈トピック依存ネットのトピックを構成する単語空間に対話過程を通じてマッピングすることにより，物体指向動作認識を対話のトピック意味空間に埋め込む．そして，Kinect センサーでキャプチャした RGB-D とスケルトンデータからの ACTNET の学習による物体指向動作認識，ウェブからスクレイピングした文章を用いた文脈トピック依存ネットの学習，及びそれらに基づく文脈トピック依存ネットを用いた対話の実験を通じて，本方法に基づく物体指向動作認識を伴う対話を評価する．

関連する研究として，物体指向動作認識に関して，Yao ら [Yao 12, Yao 13] は，静止画像を対象として，物体と人の姿勢を相互にそれぞれの認識のコンテキストとして利用し，さらに，物体に働きかける様々な姿勢を物体が有する様々な機能に発見的に対応づけるモデルを提案している．非タスク指向対話

連絡先: 渥美雅保，masayasu.atsumi@gmail.com

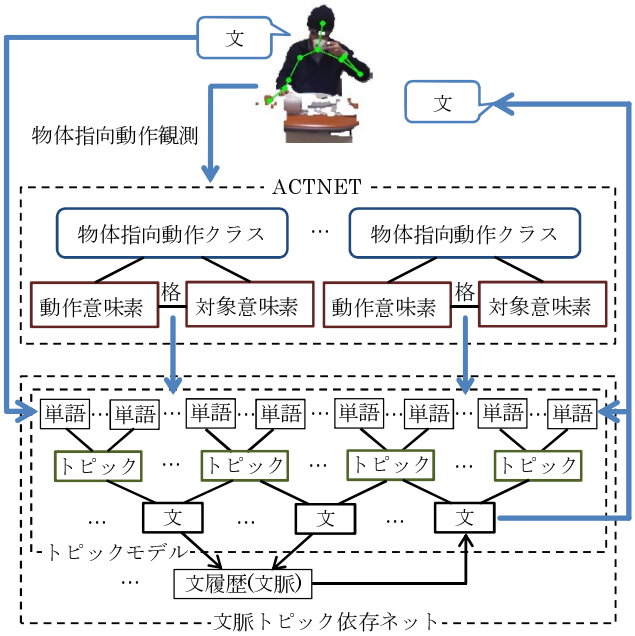

図 1: 物体指向動作認識を伴う対話のモデル

における応答文生成に関して，ウェブ上の文書情報を用いた規則に基づく応答文生成 [柴田 09] やランキング学習を用いた統計的応答文生成 [稲葉 12] 等がある．本研究は，物体指向動作認識を非タスク指向対話の文脈に組み入れている点でこれらの研究とは異なるが，物体指向動作認識に関しても，物体に対する様々な動作の意味を格 3 つ組により確率的に学習して利用している点，非タスク指向対話に関しても，コーパスから学習されたトピック，及びその遷移をもとに発話文が選択される点，が特徴である．

2. 物体指向動作認識を伴う対話

物体指向動作認識を伴う対話のモデルを図 1 に示す．本モデルは，人の物体指向動作の観測，並びに発話する文を入力とし，物体指向動作クラスの認識を介して推論される対象意味素と動作意味素，及び入力文と自らの出力文の BoW(Bag of Words) をもとに文脈のトピックを推定して，その文脈トピッ

クのもとで次に発話する文を選択して出力する．物体指向動作の認識・推論は ACTNET により遂行される．ACTNET の構成と学習，及びそれを用いた認識・推論については 3. 節で述べる．なお，[渥美 14, Atsumi 14] に詳しい記述がある．一般に，1 つ 1 つの動作は一連の動作の中で行われることが多い．本研究では，前者の 1 つ 1 つの動作を「アクション」，後者の一連の動作を「アクティビティ」と呼び，アクティビティがアクションのコンテキストを与えてアクションの認識を促進すると仮定し，ACTNET においてアクションとアクティビティを学習し，また，アクションとアクティビティの関連をそれらの共起関係により表現して認識に利用する．文脈のトピック管理と発話文の選択は文脈トピック依存ネットにより遂行される．文脈トピック依存ネットは，各トピックの単語確率分布と各文のトピック確率分布を与えるトピックモデルをコアに，文履歴とそれに続く文の依存関係を管理する．物体指向動作の格 3 つ組は文脈トピック依存ネットのトピック単語空間へ対話過程を通じてマッピングされる．文脈トピック依存ネットの構成と学習，及びそれを用いた文脈トピック管理と発話文選択については 4. 節で述べる．

3. 物体指向動作の認識

人の動作を身体スケルトンのジョイント点の 3 次元座標の時系列としてキャプチャする．本研究では，両手による物体指向動作を扱うため，肩中心に対する両手の相対 3 次元座標の時系列を利用する．両手の相対 3 次元座標の時系列から，両手の動き特徴量を次の手順により求める．まず，両手の相対 3 次元座標をある間隔で量子化し，量子化された相対位置とその変位の時系列を計算する．次に，それら時系列に対して，アクション，及びその系列であるアクティビティのアノテーションを，それらの開始フレームと終了フレーム，及び格 3 つ組 $\langle \sigma_n[w_n], r, \sigma_v[w_v] \rangle$ を指定することにより付与する．ここで，w_n は動作の対象を表す名詞，σ_n はその意味素，w_v は動作を表す動詞，σ_v はその意味素で，意味素は日本語 WordNet[Isahara 08] の同義集合 (synset) により与えられる．また，r は格表記である．そして，各アクション，及びアクティビティの相対位置と変位の時系列に対して，その時系列が表す動き特徴量を，肩中心を原点として身体周りの 3 次元空間をある大きさで分割したブロックごとの変位ヒストグラムの連結ヒストグラム (格 3 つ組付き動きヒストグラム) として求める．

3.1 物体指向動作の確率意味ネットワーク

ACTNET(図 1) の学習は，アクションとアクティビティの確率意味ネットワークの生成，並びにアクションとアクティビティの共起関係の設定によりなされる．確率意味ネットワークの生成は，格 3 つ組付き動きヒストグラム集合の I-PLCA(Incremental Probabilistic Latent Component Analysis) を用いた確率的クラスタリングによる動きクラスの生成と，動きクラスと格 3 つ組の意味素との結合確率の計算に基づくネットワーク構成により遂行される．I-PLCA により動きのクラス確率分布 $\{p(c)|c \in C\}$，インスタンス確率分布 $\{p(m_a|c)|m_a \in M \times A, c \in C\}$，クラス特徴確率分布 $\{p(f_n|c)|f_n \in F, c \in C\}$，及びクラスの数 $|C|$ が推定される．ここで，C はクラス集合，M は動き系列集合，A は格 3 つ組集合で，m_a は格 3 つ組 a を付与された動き系列 m，即ち動きのインスタンスである．確率意味ネットワークは，動きクラスに関連付けられたクラス確率分布 $\{p(c)|c \in C\}$ とインスタンス確率分布 $\{p(m_a|c)|m_a \in M \times A, c \in C\}$ を用いて，動きクラスと意味素の結合確率の計算に基づいて生成される．図 2

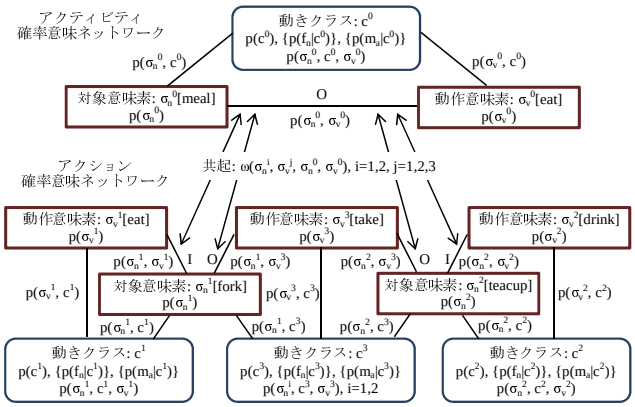

図 2: ACTNET の構成例 (図中の記号は本文を参照のこと)

に，アクティビティとアクションの確率意味ネットワーク，及びそれらの間の共起関係からなる ACTNET の構成の例を示す．ネットワークのノードは，各クラスに対応したクラスノードと格 3 つ組の意味素に対応した意味素ノードからなる．

アクションとアクティビティの間の共起関係は，格 3 つ組 $\langle \sigma_n[w_n], r, \sigma_v[w_n] \rangle$ を持つアクションが格 3 つ組 $\langle \sigma_n^0[w_n^0], r^0, \sigma_v^0[w_n^0] \rangle$ を持つアクティビティに含まれるときに，アクションの対象意味素 σ_n と動作意味素 σ_v のペアとアクティビティの対象意味素 σ_n^0 と動作意味素 σ_v^0 のペアの結合確率を用いて式 (1) により定められる．

$$\omega(\sigma_n, \sigma_v, \sigma_n^0, \sigma_v^0) = \log \frac{p(\sigma_n, \sigma_v, \sigma_n^0, \sigma_v^0)}{p(\sigma_n, \sigma_v) p(\sigma_n^0, \sigma_v^0)} \tag{1}$$

3.2 物体指向動作の認識と推論

アクションとアクティビティの認識と推論では，与えられるアクションの動きヒストグラムの系列に対して，ACTNET を用いて各々のアクションとそれまでの系列が表すアクティビティの対象意味素と動作意味素を求める．アクションまたはアクション系列の動き m に対するアクションまたはアクティビティの動きクラスの認識は，クラスノードの動きクラス c のクラス特徴確率分布 $p(f_n|c)$ とこの動き m のヒストグラム分布 $\hat{h}_m(f_n)$ の類似度を式 (2) により計算し，類似度が大きい動きクラスを求めることによりなされる．この類似度は動きクラスの確信度として用いられる．

$$\beta(c, m) = 1 - \frac{\sum_{f_n} |p(f_n|c) - \hat{h}_m(f_n)|}{2} \tag{2}$$

確信度 β で動きクラス c が求まるとそれから対象意味素 σ_n・動作意味素 σ_v を確信度 $p(\sigma_n|c) \times \beta$，$p(\sigma_v|c) \times \beta$，$p(\sigma_n, \sigma_v|c) \times \beta$ で推論することができる．また，アクションクラス c の確信度を β，アクティビティクラス c^0 の確信度を β^0 とするとき，それらの共起関係を用いて，アクションの対象意味素と動作意味素のペア (σ_n, σ_v) とアクティビティの対象意味素と動作意味素のペア (σ_n^0, σ_v^0) を同時に確信度

$$\begin{aligned}\beta(\sigma_n, \sigma_v, \sigma_n^0, \sigma_v^0|c, c^0) = p(\sigma_n, \sigma_v|c) \times p(\sigma_n^0, \sigma_v^0|c^0) \\ \times (\beta + \beta^0)/2 + \lambda \times \omega(\sigma_n, \sigma_v, \sigma_n^0, \sigma_v^0)\end{aligned} \tag{3}$$

で求めることができる．ここで，λ は共起係数である．また，追加情報としてアクティビティ，もしくはアクションの対象意味素または動作意味素が与えられるとき，アクションの動作意味素または対象意味素を更新された確信度で推論できる．

4. 対話におけるトピック管理

4.1 文脈のトピック依存ネットワーク

文脈トピック依存ネットは，物体指向動作を伴う対話の文脈をトピックの確率分布で管理し，その文脈トピック確率分布のもとで発話する文を選択する．文脈トピック依存ネットは，図 1 に示すように，トピックごとの単語確率分布，各文のトピック確率分布を介した文履歴毎のトピック確率分布，及び文履歴とそれに続く文のトピック遷移グラフから構成される．

いま，文の列を $s_1,\ldots,s_{N_S}$，文集合 (コーパス) を S，文 s_i に含まれる単語集合を $W_{s_i}=\{w_{i,j}|j=1,\ldots,N_{s_i}\}$，全文の単語集合 (辞書) を W とする．また，トピックの集合を $T=\{t_k|k=1,\ldots,N_T\}$ とする．このとき，各トピックの単語確率分布 $p(w|t)$ と各文のトピック確率分布 $p(t|s)$ は，LDA(Latent Dirichlet Allocation)[Blei 03, Hoffman 10] を用いて求められる．また，文履歴のトピック確率分布を，文履歴 $\vec{s}$ を長さ N_H の文の列とするとき，$p(t|\vec{s})=\frac{\sum_{s\in\vec{s}}p(t|s)}{N_H}$ により求める．文履歴 $\vec{s}$ に続く文を s' とするとき，それらのペア $(\vec{s},s')$ は単語空間，及びトピック空間に文脈遷移関係を導入する．与えられたトピックに対する文履歴の確率 $p(\vec{s}|t)$ は，

$$p(\vec{s}|t)=\frac{\frac{\sum_{s\in\vec{s}}p(t|s)}{N_H}\times p(\vec{s})}{p(t)} \tag{4}$$

により求められる．

文集合 (コーパス) は，文章，段落，文に階層化されて収集される．例えば，ウェブ文書の文の場合は，1 つの URL に含まれる文が文章，その中でブロックタグで区切られたブロックの文が段落を構成する．文履歴とそれに続く文のペアは，ある字数以上の長さの文のある個数以上の文の連なりの段落から抽出される．

4.2 トピック管理と発話文推論

文脈トピック依存ネットを用いた文脈のトピック確率分布の管理は，物体指向動作認識を介して推論される格 3 つ組から生成される BoW，人の発話による入力文から計算される BoW，及び自らの発話の出力文から計算される BoW からなる系列に対して遂行される．ここで，物体指向動作の格 3 つ組に対する BoW は，アクションとアクティビティそれぞれの対象意味素に対する名詞と動作意味素に対する動詞からなる 4 つの単語に加えて，対話過程を通じて追加された単語から構成される．対話過程を通じた BoW への単語の追加は，物体指向動作認識時に入力された文に対して出力発話文の選択が高い確率値で適切になされたとき，それら入力文に含まれる単語を追加することによりなされる．格 3 つ組の BoW としてこれら単語を対応付けて管理し組み込むことにより，物体指向動作を対話のトピック空間に有効にマッピングできると期待される．

物体指向動作認識，及び入出力文に対して計算される BoW の系列キューを $Q_{BoW}=[q_1,\ldots,q_{N_Q}]$，その長さを N_Q とする．ここで，q_l はキューに追加された BoW で，l の値が小さいものほど最近追加された BoW で，Q_{BoW} はBoW の追加により逐次更新される．また，q_l は，それが物体指向動作認識から生成されたか文から生成されたかのタイプを，それぞれ O_{BoW}，U_{BoW} として持つ．このとき，Q_{BoW} に対して，その要素の BoW をタイプによる重みづけと追加時点による割引率で統合した BoW を次のように求める．

$$q=\sum_{l=1}^{N_Q}\big(w(q_l)\times d(q_l)\times q_l\big) \tag{5}$$

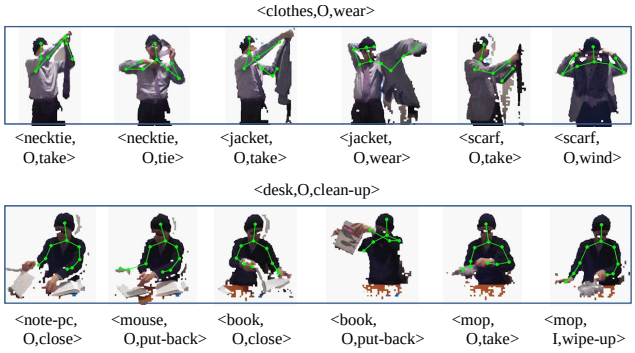

図 3: アクティビティのアクション系列の例

表 1: ACTNET の構成

	クラス数	対象意味素数	動作意味素数	意味素ペア数
アクティビティ	12	4	4	4
アクション	53.5	10	16	27

ここで，$w(q_l)$ は BoW のタイプによる重みで，q_l のタイプ $type(q_l)$ が O_{BoW} のとき w_O，U_{BoW} のとき w_U をとる．また，$d(q_l)$ は BoW の追加時点による割引率で，BoW のタイプが O_{BoW} のときの割引率 d_O，U_{BoW} のときの割引率 d_U を用いて次の規則により求められる．

$$d(q_1)=1.0$$
$$d(q_l)=\begin{cases} d_U\times d(q_{l-1}) & if\ l>1\wedge type(q_{l-1})=U_{BoW} \\ d_O\times d(q_{l-1}) & if\ l>1\wedge type(q_{l-1})=O_{BoW}\end{cases} \tag{6}$$

この統合された文脈の BoW に対してトピックの確率分布 $p(t|q)$ が求められ，これと式 (4) の $p(\vec{s}|t)$ により，文脈の BoW が与えられたときの文履歴 $\vec{s}$ の確率分布は次式により計算される．

$$p(\vec{s}|q)=\sum_{t\in T}\big(p(\vec{s}|t)\times p(t|q)\big) \tag{7}$$

この確率分布 $p(\vec{s}|q)$ の最大値を与える文履歴 $\vec{s}_*$ に対して，それとペア $(\vec{s}_*,s'_*)$ をなす文 s'_* が発話文として選択される．

5. 実験

5.1 物体指向動作認識実験

物体指向動作の ACTNET への学習，及び ACTNET を用いた認識と推論の実験を，Kinect センサーを用いてキャプチャしたビデオデータセットを作成して行った．データセットは，＜洋服, を, 着る＞，＜食事, を, 食べる＞，＜机, を, 掃除する＞，＜報告書, を, 書く＞の 4 つの格 3 つ組でラベル付けされたアクティビティのビデオ映像を含み，それらには合わせて 10 個の物体と 30 の物体指向アクションが含まれる．図 3 にアクティビティの映像の例をアクションのスナップの列で示す．動きのヒストグラム化における身体周りのブロック分けは，身体の近傍の前方と側方をそれぞれ 1 辺 30cm の 9 ブロック，その外側の前方と側方をそれぞれ大きく 9 ブロックと 8 ブロック，後方を 1 つのブロックとする．これよりブロック数は 36 となり，動きヒストグラムの次元は 972 次元である．

4 つのアクティビティの各々に対して 4 つのビデオ映像を用意して 4 分割交差検定により性能評価を行った．表 1 に，学習された ACTNET の構成を示す．ここで，クラス数は I-PLCA により自動的に決められている．表 2 に，ACTNET による認識・推論の評価結果を示す．ここで，括弧内の数値は次善解

表 2: アクティビティとアクションの認識・推論結果

アクティビティ正解率	93.8%(100%)
アクション正解率 (共起なし)	53.3%(59.2%)
アクション正解率 (共起あり)	62.5%(76.7%)
物体名判明時のアクション正解率 (共起なし)	75.8%(85.8%)
物体名判明時のアクション正解率 (共起あり)	83.4%(96.7%)

までの正解率である．また，アクションとアクティビティの共起係数は 0.2 とした．実験結果より，学習された ACTNET によりアクションとアクティビティの認識が可能で，特に，コンテキストを与えるアクティビティとの共起によりアクションの認識性能をあげられること，及びアクション認識のあいまいさが物体が何かの追加情報を用いた推論により解消されることが示された．正解率を下げているのは主に机上で短時間に実行される 5 つのアクションで，これらアクションの正確なスケルトン追跡が Kinect で難しいことが一因である．

5.2 対話におけるトピック管理実験

ウェブからスクレイピングした文集合を用いた文脈トピック依存ネットの学習，及び文脈トピック依存ネットを用いた物体指向動作認識を伴う対話の実験を行った．ウェブからの文の収集は，5.1 の 4 つのアクティビティと 30 個のアクションの格 3 つ組に現れる名詞と動詞をキーとしたウェブ検索によりウィキペディアの 15 個の URL を選び，それらを開始 URL として深さ 2 でクロールして辿ったページをスクレイピングすることで行った．そして，それらから文長 15 文字以上，段落長 6 以上の文集合を抽出することで，URL 数 963，段落数 3880，文数 28051 の DB を作成し，それら文から名詞，動詞，形容詞を抽出することで，単語総数 34687 のコーパスを構築した．これらの文には，4 つのアクティビティとは関係のない文も多く含まれる．

このコーパスを用いた文脈トピック依存ネットの学習に関して，まず，文がトピックによりどの程度特徴づけられているかの評価を，トピック数を変えることによる LDA のパープレキシティを計算することで行った．図 4(a) に異なるトピック数に対する文パープレキシティを相対指標で示す．これより，文脈トピック依存ネットのトピック数を 200 とした．次に，文履歴がトピックによりどの程度特徴づけられているかの評価を，異なる文履歴長に対する文脈トピック依存ネットの文履歴パープレキシティを計算することで行った．図 4(b) に文履歴パープレキシティを同じく相対指標で示す．この結果を参考に，コーパスの平均段落長と対話におけるトピックの継続性も考慮して，文脈トピック依存ネットの文履歴長を 5 とした．

この文脈トピック依存ネットを用いた物体指向動作認識を伴う対話に関して，物体指向動作認識に基づく発話におけるトピック選択，及びトピック切り替えの評価を行った．実験では，ACTNET が推論するアクティビティとアクションの格 3 つ組の意味素の名詞と動詞からなる BoW に対して，文脈トピック依存ネットを用いた発話する文の推論を，4 つのアクティビティを切り替えながら繰り返した．ここで，BoW 系列キュー長を 5，BoW 系列の統合における重みを $w_O = 2.0$，$w_U = 1.0$，割引率を $d_O = 0.1$，$d_U = 0.9$ とした．そして，これら発話される文のトピックをその内容から判断した結果，58.3% でアクティビティにあった文が選択された．さらに，物体指向動作認識と同時に入力された文に対して適切な応答ができた場合に，入力文の単語を BoW に組み込む実験の結果，より適した発話の選択ができるようになるケースが確認された．また，アクティビティの遷移に伴う発話文のトピック切り替えを，BoW 系列の統合における重みと割引率の設定を変えることで評価した．その結果，BoW 系列の統合における重みを $w_O = w_U = 1.0$，割引率を $d_O = d_U = 0.9$ とタイプ O_{BoW} と U_{BoW} で同じにした場合は 75% の発話で 1 つ前のアクティビティの影響を受けたのに対して，上記の設定では，発話に 1 つ前のアクティビティの影響を受けることは 100% なかった．

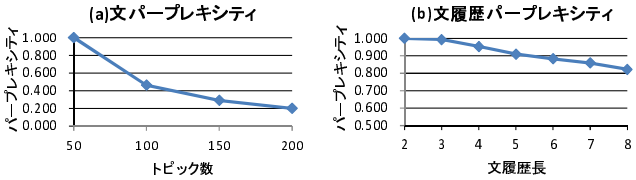

図 4: 文脈トピック依存ネットのパープレキシティ

6. おわりに

本論では，物体指向動作認識を伴う対話に関して，物体指向動作を認識してその格 3 つ組ラベルを確率意味ネットワーク ACTNET に基づき推論する手法，及び物体指向動作認識を伴う対話のトピック遷移を管理して発話文を選択する文脈トピック依存ネットに基づく手法について述べた．そして，Kinect センサーでキャプチャした両手の物体指向動作のビデオ映像，及びウェブ文書のスクレイピングにより構築したコーパスを用いて，物体指向アクションとアクティビティの ACTNET の学習に基づく推論の評価，及び文脈トピック依存ネットの学習に基づく物体指向動作認識を介した対話における発話文のトピック選択とトピック切り替えの評価を行って，これら手法の有用性を確かめた．

参考文献

[渥美 14] 渥美雅保: 物体指向動作の心象と表象の確率的カテゴリゼーション, 2014 年度人工知能学会全国大会 (第 28 回) 論文集, 2I5-OS-08b-5, 4p. (2014)

[Atsumi 14] Atsumi, M.: Learning Probabilistic Semantic Network of Object-Oriented Action and Activity, In: Artificial Intelligence: Methodology, Systems, and Applications, Proc. of 16th. Int. Conf. AIMSA 2014, Lecture Note in Computer Science, Vol. 8722, pp.1-12, Springer (2014)

[Yao 12] Yao, B. and Fei-Fei, L.: Recognizing Human-object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses, IEEE Trans. on Pattern Analysis and Machine Intelligence 34 (9) pp.1691-1703 (2012)

[Yao 13] Yao, B., Ma, J. and Fei-Fei, L.: Discovering Object Functionality, Int. Conf. on Computer Vision 2013 (2013)

[柴田 09] 柴田雅博, 他: 雑談自由対話を実現するための WWW 上の文書からの妥当な候補文選択手法, 人工知能学会論文誌, Vol.24, No.6, pp.507-519 (2009)

[稲葉 12] 稲葉通将, 他: 非タスク指向型対話エージェントのための統計的応答手法, 電子情報通信学会論文誌, Vol.J95-D, No.6, pp.1390-1400 (2012)

[Isahara 08] Isahara, H. *et. al.*: Development of Japanese WordNet, 6th Int. Conf. an Language Resources and Evaluation, pp.2420-2423 (2008)

[Blei 03] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, J. of Machine learning Research, Vol.3, pp.993-1022 (2003)

[Hoffman 10] Hoffman, M. D., Blei, D. M. and Bach, F.: Online Learning for Latent Dirichlet Allocation, Advances in Neural Information Processing Systems 23, pp.856-864 (2010)