3O4-OS-04b-3

事物指向アクティビティに関連する対話における発話文選択

Utterance Selection in Dialogue related to Things-oriented Activities

渥美雅保
Masayasu Atsumi

創価大学理工学部情報システム工学科
Dept. of Information Systems Sci., Faculty of Sci. and Eng., Soka University

This paper describes a method for utterance selection in dialogue related to things-oriented activities. The method learns to generate descriptive texts from things-oriented activity descriptors using a encoder-decoder LSTM neural network and to select utterances about related topics on a contextual topic network. Through experiments using an image dataset with action annotations and captions and web-scraped corpora, it is shown that topics of utterances follow things-oriented activity transitions based on the learned descriptive text generator and utterance selector.

1. はじめに

日常空間において人のアクティビティを理解することは人を自律的に支援するロボット等の機械にとって必要な機能の 1 つである．また，人のアクティビティの認識は会話のトリガーとなり，会話を介しての人への働きかけは人とロボットの共生を促進する．人のアクティビティは物や事，即ち事物に働きかけるアクションの集合もしくは系列として捉えることができる．本研究では，人のアクティビティを事物とアクションの関係，及び人と事物の関係の集合 (系列) で捉えて，これを事物指向アクティビティと呼ぶ．筆者は，[渥美 14, Atsumi 14] において，人の物体に働きかける動作を物体指向アクションと呼んで，人の物体指向アクションの RGB-D 映像から物体とアクションの確率的意味ネットワークを学習することにより，物体指向アクションとそのコンテキストを与えるアクティビティを認識する方法を提案した．また，[渥美 15] において，物体指向アクションの認識を介した非タスク指向対話のトピックを管理する方法を提案した．ところで，事物指向アクティビティの認識からトピック対話を展開する過程では，それら認識を描写する説明文を生成することが認識の確認と共有，及びトピックの選択において有効になってくる．そこで，本論文では，事物指向アクティビティを描写する説明文を生成し，それを用いてそのアクティビティに関連するトピックを選択し発話文を生成する問題を扱う．

画像からその説明文を生成する研究は，画像を物体セグメントの意味グラフに変換してそのグラフから文を生成する方法 [Yao 10] に端を発し，最近では，畳込みニューラルネットワーク (CNN) が出力する画像ベクトルを再帰ニューラルネットワーク (RNN)[Karpathy 15] や Long Short-Term Memory 型 RNN(LSTM)[Vinyals 15] の入力として用いて文を生成する方法などニューラルネットワークを用いる方法が多く研究されている．本研究では，事物指向アクティビティに対して，それに含まれる事物とアクションの間と人と事物の間の関係の意味記述 (事物指向アクティビティ記述) から説明文の生成を行う方法を提案する．ある系列から別の系列を生成するニューラルネットワークとして Seq2Seq モデル [Sutskever 14] や自己エンコーダ・デコーダモデル [Cho 14] がよく知られている．本方法では，事物指向アクティビティ記述からその説明文を LSTM をエンコーダとデコーダに用いた事物指向アクティビティ記述のエンコーダ・デコーダにより生成する．そして，その生成説明文を LDA(Latent Dirichlet Allocation)[Blei 03, Hoffman 10] に基づく文脈トピックネット [渥美 15] に入力することで，トピック管理のもとで発話文系列を生成する．図 1 に，事物指向アクティビティに関連する対話のモデルの概要を示す．このうち，本論文で論ずるのは「事物指向アクティビティ記述からの説明文生成」と「文脈トピックネットによる関連トピック発話」の部分である．

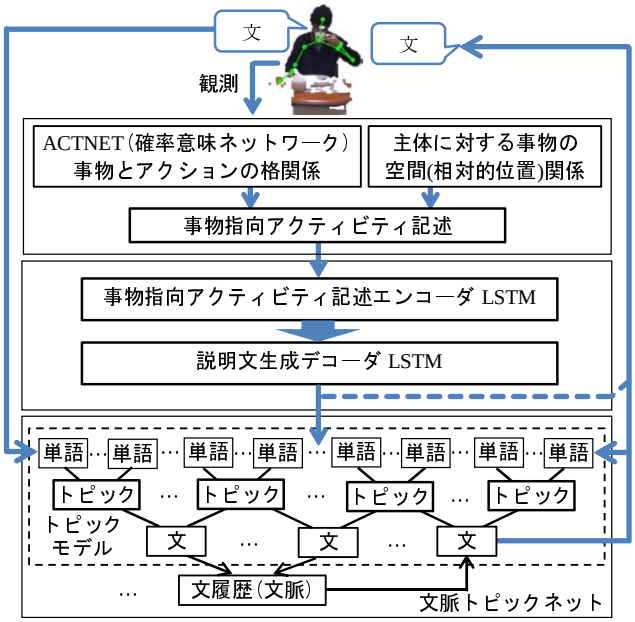

図 1: 事物指向アクティビティに関連する対話のモデル

以下，2. で事物指向アクティビティの意味記述枠組，3. で事物指向アクティビティ記述からの説明文生成，4. で事物指向アクティビティに関連するトピック発話管理，5. で実験による評価について述べる．

2. 事物指向アクティビティの記述

事物指向アクティビティを，事物とアクションとの間の格関係と事物と主体との間の空間関係を用いて記述する．このと

連絡先: 渥美雅保，masayasu.atsumi@gmail.com

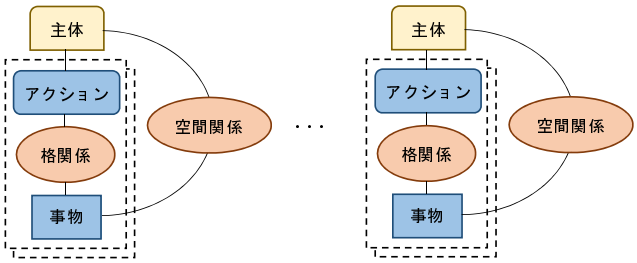

図 2: 事物指向アクティビティ記述枠組

き，ある主体のアクションは，その主体のアクションと事物の格関係と主体と事物との空間関係を用いて表される．この主体のアクションを事物指向アクションと呼ぶ．これより．事物指向アクティビティの記述は事物指向アクション記述の集合で表される．図 2 に事物指向アクティビティ記述の概念図を示す．これは，ある時点，即ち画像または映像フレーム，もしくはある時区間，即ちフレーム区間に含まれる事物指向アクション集合を記述している．形式的には，事物指向アクティビティは次のように記述される．

$$\begin{aligned} &things\_activity := things\_action\{;things\_action\}^*. \\ &things\_action := [things, \{rel\}^*,]\{action\}^+[, subject] \end{aligned} \tag{1}$$

事物指向アクティビティ ($things\_activity$) は 1 つ以上の事物指向アクション ($things\_action$) により記述され，事物指向アクションは事物 ($things$)，関係 (rel)，アクション ($action$)，主体 ($subject$) により記述される．ここで，事物とアクションは WordNet[Miller 95] の同義集合 (synset) により与えられる名詞と動詞の意味素である．関係は格関係か空間関係のいずれかで，格関係は動詞に対する名詞の格，空間関係は主体に対する事物の相対位置を表す主に副詞の同義集合 (synset) により記述される．事物とアクションの意味素とそれらの間の格関係は，⟨ 事物意味素, 格, アクション意味素 ⟩ の格 3 つ組として確率意味ネットワーク ACTNET[渥美 14, Atsumi 14] から求めることができる．主体は単一の場合は省略可能である．5. の実験で用いた事物指向アクティビティ記述の例を示す．ここでは，格関係は省略されている．

$$\begin{aligned} &sandwich.n.01, light\_contact.r.cw\ ahead.r.01, \\ &\quad hold.v.00\ eat.v.00\ touch.v.00\ ;\ sit.v.00. \\ &laptop.n.01, light\_contact.r.cw\ ahead.r.01, \\ &\quad touch.v.00\ look.v.00\ use.v.00\ ; \\ &\quad chair.n.01, full\_contact.r.cw\ below.r.01, \\ &\quad touch.v.00\ use.v.00\ ;\ sit.v.00\ . \end{aligned} \tag{2}$$

3. 事物指向アクティビティの説明文生成

画像，映像フレーム，またはフレーム系列から得られる事物指向アクティビティ記述から説明文を生成するための LSTM を用いたエンコーダ・デコーダを，画像に対する事物指向アクティビティ記述と説明文集合のペアのデータセットから学習により構築する．図 3 に，5. の実験で用いた画像に対する事物指向アクティビティ記述と説明文集合のペアの例を示す．事物指向アクティビティのエンコーダ・デコーダは，図 4 に示すように，事物指向アクティビティ記述のエンコーダ LSTM と説明文生成のデコーダ LSTM から構成される．事物指向ア

事物指向アクティビティ記述:
spoon.n.01 , light_contact.r.cw ahead.r.01 ,
hold.v.00 touch.v.00 use.v.00 ;
bowl.n.01 , near.r.01 ahead.r.01 , use.v.00 ;
sit.v.00 pose.v.00 ;
dining_table.n.01 , light_contact.r.cw below.r.01 ,
touch.v.00 use.v.00 .

説明文:
a man who is eating some food out of a bowl .
a man is sitting in a break room eating a meal .
a man eating a bowl of stew with some flat bread .
a man in a blue shirt eating a meat and bread dish from a bowl .
a man is sitting down at a table , eating his stew and tortillas .

図 3: 画像・事物指向アクティビティ記述・説明文集合の例

クティビティ記述エンコーダでは，事物指向アクティビティ記述 $d_1, \ldots, d_N$ を埋込み層が $embed(d_i)(i = 1, \ldots, N)$ によってベクトル系列化し．エンコーダ LSTM はその系列を再帰的に符号化してその結果をデコーダ LSTM に渡す．説明文生成デコーダでは，符号化された事物指向アクティビティ記述をもとに，デコーダ LSTM の出力を復元層が $disembed(h_{d,j})$ によって単語 w_j に復元し，それを埋込み層で $embed(w_j)$ によりベクトル化してデコーダ LSTM に渡す処理を再帰的に繰返すことにより，説明文の単語列 $w_j(j = 1, \ldots, M)$ を生成する．ここで，エンコーダ LSTM とデコーダ LSTM には一般的な LSTM[Gers 00] を用いている．学習時には，事物指向アクティビティ記述の符号化結果に対して，説明文生成デコーダでは教師説明文の単語列を入力し，予測される説明文の単語分布列との間の誤差を逆伝播する．

4. 関連トピック発話管理

4.1 文脈トピックネット

文脈トピックネットは，事物指向アクティビティに関連する対話の文脈をトピックの確率分布に基づいて管理し，その確率分布のもとで発話文を選択するのに用いられる．文脈トピックネットは，事物指向アクティビティの説明文集合，及び事物に関連するトピック文章の集合から生成され，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遷移グラフから構成される．ここで，トピック文書は段落に分けて収集され，それにより，文履歴とそれに続く文の遷移グラフは段落ごとに分割される．

いま，文集合 (コーパス) を $S = \{s_i | i = 1, \ldots, N_S\}$，文 s_i に含まれる単語集合を W_{s_i}，全文の単語集合 (辞書) を $W = \bigcup W_{s_i} = \{w_j | j = 1, \ldots, N_W\}$ 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

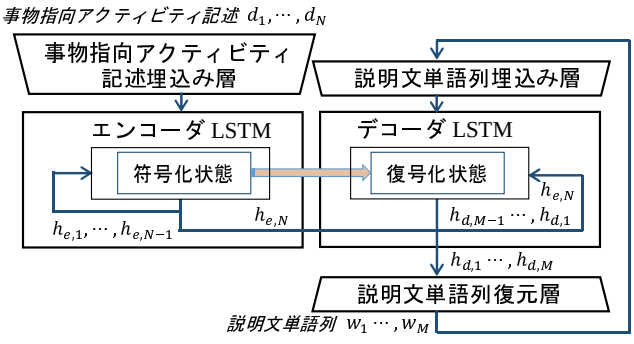

図 4: 事物指向アクティビティのエンコーダ・デコーダ

単語確率分布 $p(w|t)$ と各文のトピック確率分布 $p(t|s)$ を LDA を用いて求める．また，文履歴のトピック確率分布を，文履歴 $\vec{s}$ を長さ $|\vec{s}|$ の文の列とするとき，$p(t|\vec{s}) = \frac{\sum_{s\in\vec{s}} p(t|s)}{|\vec{s}|}$ により求める．文履歴 $\vec{s}$ とそれに続く文 s' のペア $(\vec{s}, s')$ は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，

$$p(\vec{s}|t) = \frac{\frac{\sum_{s\in\vec{s}} p(t|s)}{|\vec{s}|} \times p(\vec{s})}{p(t)} \tag{3}$$

により求められる．

4.2 トピック管理と発話文選択

文脈トピックネットを用いた文脈のトピック確率分布の管理は，事物指向アクティビティの説明文から計算される BoW(Bag of Words)，及び人の発話文と自らの発話文から計算される BoW の系列に対して遂行される．事物指向アクティビティの説明文，及び発話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，それが事物指向アクティビティの説明文から計算されたか発話文から計算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．

$$q = \sum_{l=1}^{N_Q} \big(w(q_l) \times d(q_l) \times q_l \big) \tag{4}$$

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．

$$d(q_1) = 1.0$$
$$d(q_l) = \begin{cases} d_U \times d(q_{l-1}) & if\ l > 1 \wedge type(q_{l-1}) = U_{BoW} \\ d_O \times d(q_{l-1}) & if\ l > 1 \wedge type(q_{l-1}) = O_{BoW} \end{cases} \tag{5}$$

この統合された文脈の BoW に対して文脈トピック確率分布 $p(t|q)$ が求められ，これと式 (3) 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．

$$p(\vec{s}|q) = \sum_{t\in T} \big(p(\vec{s}|t) \times p(t|q) \big) \tag{6}$$

このとき，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，それとペアをなす文 s'_* が発話文として選択される．

5. 実験

5.1 事物指向アクティビティの説明文生成実験

事物指向アクティビティに関して，その記述は映像フレーム系列からのみならず画像からも得ることができる．そこで，画像から得られる事物指向アクティビティ記述から説明文を生成する実験を，Microsoft の COCO[Lin 14] と COCO-a[Ronchi 15] データセットを用いて行った．COCO データセットは，画像に対して物体のカテゴリラベルとセグメント情報，及びキャプション文が付けられた画像データセットである．キャプション文は画像を説明する文で，各画像に 5 つのキャプション文が付けられている．COCO-a データセットは，COCO の画像内の人のビジュアルアクションのアノテーションデータセットである．ビジュアルアクションとは視覚的に識別可能なアクションのことで，COCO データセットの 12 のスーパカテゴリに分けられた 80 の物体カテゴリに対して 145 のビジュアルアクションが VerbNet[*1] を参考に選ばれている．また，人と物体の空間関係と距離，感情もアノテーションとして与えられていて，本論文での提案記述 (図 2) に COCO-a アノテーションは似ている．本研究では，これらデータセットから，屋内に有ることが多い 7 つのスーパカテゴリの 46 個のカテゴリの物体に対する 1 人の主体によるアクションが含まれる画像を選んだ．画像の枚数は 719 枚，それらの説明文の総数は 3596 である．また，事物指向アクティビティ記述は，COCO-a のアノテーションを 2. で述べた記述枠組に変換したものを用いた．図 3 の画像と説明文は COCO データセットの画像とそのキャプション文で，事物指向アクティビティ記述は COCO-a のアノテーションを (1) の形式に変換したものである．

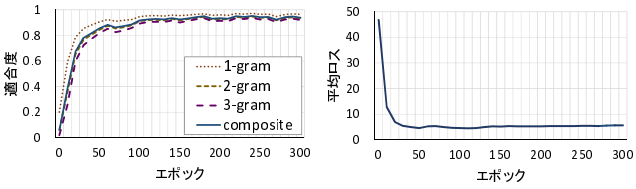

図 5: 説明文生成学習の結果

説明文生成の性能評価指標として，事物指向アクティビティ記述に対して生成される説明文 (予測文)$s_k(\in S)$ と与えられた説明文集合 (教師文集合)$E_k = \{e_k\} \in E^*$ の整合性を n-gram 一致度に基づいて測る式 (7) の指標を用いる．いま，n-gram x について，$m(s_k, x)$, $m(e_k, x)$ をそれぞれ x が s_k，e_k に現れる回数とする．また，$g(n, s_k), (n = 1, 2, 3)$ を s_k に含まれる n-gram の数とする．このとき，

$$C(S, E^*) = \prod_{n=1}^{3} \left(\frac{\sum_k \rho(s_k, E_k) c(n, s_k, E_k) + \delta_n}{\sum_k g(n, s_k) + \delta_n} \right)^{1/3}$$
$$c(n, s_k, E_k) = \sum_x \max_{e_k \in E_k} \min(m(s_k, x), m(e_k, x))$$
$$\rho(s_k, E_k) = \begin{cases} 1 & if\ |s_k| \geq \min_{E_k} |e_k| \\ e^{1 - \frac{\min_{E_k} |e_k|}{|s_k|}} & otherwise \end{cases} \tag{7}$$

を適合度と呼ぶ．$C(S, E^*)$ は，1-gram，2-gram，3-gram の適合度の相乗平均である．ここで，$\rho(s_k, E_k)$ は予測文の短さペナルティ，δ_n は修正係数で，$n = 1$ のとき 0，$n > 1$ のとき小さな正数である．また，$|s_k|$，$|e_k|$ は単語数である．

実験では，各画像の事物指向アクティビティ記述と 5 つの説明文それぞれのペアを事物指向アクティビティ記述エンコーダと説明文生成デコーダに繰り返し入力して学習を行った．図 5 に実験結果として，1，2，3-gram とそれらを合成した適合度，及び学習時の平均ロスの学習曲線を示す．事物指向アクティビティ記述埋込み層の記述子ニューロン数は 160，埋込みニューロン数は 100，説明文単語列埋込み・復元層の語彙ニューロン数は 2350，埋込みニューロン数は 600，エンコーダ LSTM とデコーダ LSTM の隠れニューロン数は 400 である．平均ロスは 50 エポックで収束しているが適合度はその後もゆっくり改

*1 http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

善され，50 から 300 エポックの間で 1-gram で 0.06，2-gram で 0.09，3-gram で 0.11 程度良くなっている．これより，この段階でも句を組み合わせて文を生成する学習が進んでいることが示唆される．

5.2 関連トピック発話文選択実験

文脈トピックネットを生成するためにトピック文書をウェブから収集し，説明文集合と合わせてデータセットを作成する．ここで，ウェブの文の場合は，1 つの URL に含まれる文列を文書，その中でブロックタグで区切られた文列を段落とする．また，説明文集合の場合は，各説明文を別の段落の文とする，即ち 1 段落 1 文の文書とみなす．ウェブ文書は，Wikipedia と CNN ニュースサイトを事物カテゴリ名をキーとしてそれぞれ検索，スクレイピングすることで収集し，2 つのデータセットを作成した．Wikipedia データセットの文数は 10072，ニュースデータセットの文数は 5105 である．文脈トピックネットの学習では，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，異なるトピック数のトピックモデルのパープレキシティを比較することにより行い，Wikipedia 文脈トピックネット，ニュース文脈トピックネットともにトピック数を 200 とした．また，文脈トピックネットの文履歴長は 2 とした．

文脈トピックネットを用いたトピック発話文選択実験では，画像の事物指向アクティビティ記述に対してエンコーダ・デコーダにより生成される説明文を文脈トピックネットに入力し，それに対する発話文の列を推論させることを，画像を順に変えて繰返し行って，文脈トピックの変化を計算する．4 枚の画像の事物指向アクティビティ記述の順次入力に対してそれぞれ 3 つの発話文の列を推論・選択させることを，719 枚の画像を用いて 179 回行った．図 6 に，Wikipedia 文脈トピックネットとニュース文脈トピックネットに対して，異なる BoW 系列キュー長 N_Q，及び BoW 系列統合における異なる重み w_O, w_U と割引率 d_O, d_U の設定で行った実験での，文脈トピック確率分布の変動系列を示す．図中，1，5，9，13 時点が説明文が入力された時点で，それ以外が発話時点である．変動値は 2 時点の文脈トピック確率分布間の L1 距離を [0,1] 正規化したものである．$N_Q = 1$ の変動系列は文脈を考慮しないケースで，事物指向アクティビティに関する説明文入力でトピック確率分布は大きく変化するが発話系列中の変動も大きい．$N_Q = 4$ の 4 つの変動系列は文脈を考慮するケースである．このうち，重み $w_O = w_U = 1.0$ かつ割引率 $d_O = d_U = 1.0$ の場合は，前の事物指向アクティビティも発話も同じように次の発話に影響するため，事物指向アクティビティの変化に対してトピックの切り替えが迅速にできていない．これに対して，重みを $w_O = 2.0$，$w_U = 1.0$，割引率を $d_O = 0.1$，$d_U = 0.9$ として事物指向アクティビティの変化に重きを置き，また前のトピックの影響を事物指向アクティビティ説明文入力時点で抑制すると，事物指向アクティビティに関する説明文入力に対して文脈トピック確率分布が大きく変化し，発話系列中の変動は小さくなる．これより，事物指向アクティビティの変化にトピックを追従させた発話文選択が可能となることがわかる．

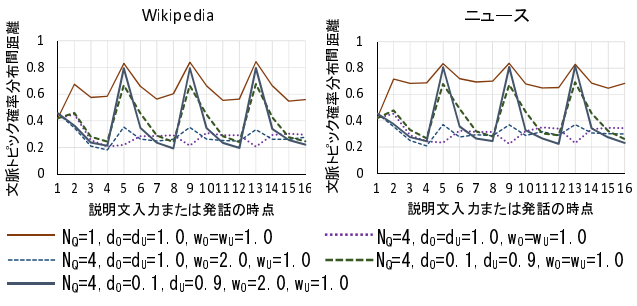

図 6: 発話文選択におけるトピック変動

6. おわりに

本論では，事物指向アクティビティに関連する対話における発話文選択に関して，事物指向アクティビティ記述から説明文を LSTM を用いたエンコーダ・デコーダにより生成し，その説明文からそれに関連するトピックの発話文を文脈トピックネットを用いた推論により選択する手法について述べた．そして，まず，事物指向アクティビティ記述と説明文を伴う画像データセットを用いた実験により，事物指向アクティビティ記述からの説明文生成学習の解析を行い，句を組み合わせた文生成が学習されることを示した．次に，ウェブから収集したトピック文書を用いて文脈トピックネットを学習し，それを用いて事物指向アクティビティに関連するトピック発話文選択の実験を行って，事物指向アクティビティにトピックを追従させた発話文選択が可能であることを示して，提案手法の有効性を確かめた．

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)

[渥美 15] 渥美雅保: 物体指向動作認識を伴う対話におけるトピック管理, 2015 年度人工知能学会全国大会 (第 29 回) 論文集, 2F4-OS-01a-5, 4p. (2015)

[Yao 10] Yao, B.Z., Yang, X., Lin, L., Lee, M.W. and Zhu, S.C.: I2T: Image Parsing to Text Description. Proc. of the IEEE, Vol.98, pp.1485-1508 (2010)

[Karpathy 15] Karpathy, A. and Fei-Fei, Li.: Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR 2015 (2015)

[Vinyals 15] Vinyals, O., Toshev, A., Beigio, S. and Erhan, D.: Show and Tell: A Neural Image Caption Generator, CVPR2015 (2015)

[Sutskever 14] Sutskever, I., Vinyals, O. and Le, Q.V.: Sequence to Sequence Learning with Neural Networks, NIPS2014 (2014)

[Cho 14] Cho, K., Van Merriënboer, B., Gülçehre, Ç. Bahdanau, D., Bougares, F., Schwenk H. and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, EMNLP2014, pp.1724-1734 (2014)

[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 NIPS 23, pp.856-864 (2010)

[Miller 95] Miller, G.A.: WordNet: A Lexical Database for English, Commumications of the ACM Vol.38, No.11, pp.39-41 (1995)

[Gers 00] Gers, F.A., Schmidhuber, J. and Cummins, F.: Learning to Forget: Continual Prediction with LSTM, Neural Computation, Vol.12, No.10, pp.2451-2471 (2000)

[Lin 14] Lin, T-Y., *et. al.*: Microsoft COCO: Common Objects in Context, arXiv: 1405.0312 [cs.CV] (2014)

[Ronchi 15] Ronchi, M.R. and Perona, P.: Describing Common Human Visual Actions in Images, BMVC2015 (2015)