事象の認識による発話生成

小南 光
東京工業大学大学院
総合理工学研究科
konan@lr.pi.titech.ac.jp

齋藤 豪
東京工業大学大学院
情報理工学研究科
suguru@img.cs.titech.ac.jp

奥村 学
東京工業大学
精密工学研究所
oku@pi.titech.ac.jp

1 はじめに

世界の状況は、同時かつ刻々と変化しており、人間はこれらの変化を取捨選択や統合をして認識している。人間が認識する状況の変化には以下のような種類がある。

・物体の移動、変形、消失
・観察主体の移動による相対的変化
・観察主体の記憶との相違から生じた変化
・主体物の意思や目的に基づいた行動

これらの状況の変化を端的に表現することは、この情報を伝達する上で重要な役割を担っている。また、日常生活においてもスポーツの実況や音声ガイドなどが活躍しており、状況の変化を伝えることの重要性を確認することができる。この状況の変化を表す端的な表現を自動生成することができれば、それは有益なことである。このような状況変化の認識を扱った研究としては、Ehlert らの研究 [1] や小島らの研究 [2, 3] が挙げられる。Ehlert らは、フライトシミュレータを用いて、飛行機が離陸前や飛行中などの状況の中で、現在どの状況にあるのかを認識するシステムを開発した。小島らは、動画像中における一人の人物の行動の認識を行うシステムを開発した。しかし、これらの研究では、認識する状態数が少数に限られていることや、観察対象が一つという問題点が残っている。

そこで本研究では、状況の変化における多数の事象を認識し、その認識した事象の中からその状況の変化を表現する上で重要な事象を選択、発話するようにすることで、表現する対象を 1 つのオブジェクトに限らずに、状況の変化をよりわかりやすく表現することを目標とする。本研究では、人間の認識過程を模倣し、「認識」「選択」「発話」の 3 段階に分ける。また、本研究では発話形式として実況を用いる。実況形式の表現生成を行う場合、実時間性と発話タイミング、表現内容が問題となるが、人間が何を基準として発話する事象とタイミングを選択しているかを考察し、模倣することでこの問題の解決を図る。

そして、提案手法を実装し仮想空間の状況を変化させた実験を行い、被験者実験との結果の比較を行うことで、提案手法の出力結果を評価する。

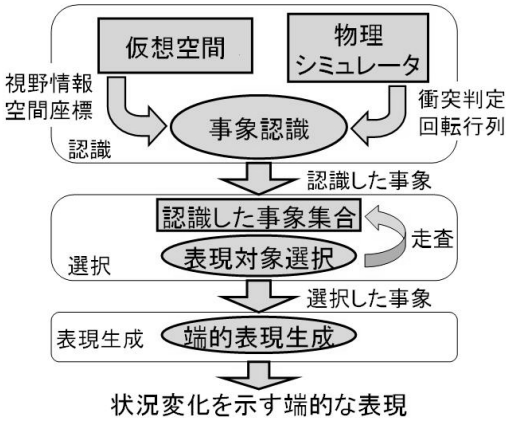

図 1: システムの概略図

2 提案手法

実験環境として [4] で開発され、現在も開発が継続している k3 を用いることとする。この仮想環境には、テーブルやボール、ブロックなどが配置されており、ボールやブロックといったオブジェクトの物理法則に則った様々な運動を観察することができる。この仮想空間を観測し、オブジェクトに関する様々な値を記録することで、その変化から生じた状況の変化 (以下、事象と呼ぶ) を認識することができる。そして、このようにして認識した事象は同時刻に複数存在するので、この中からその状況を表現するのに適した事象を選択し、表現を生成する手法を提案する。

まず、本研究で作成するシステムの概略図を図 1 に示す。システムは、観察対象である k3 に付随する事象の認識モジュールと、認識された事象から表現対象となる事象を選択するモジュール、選択した事象から表現を生成するモジュールの 3 つに分けられる。それぞれについてこれから説明する。

2.1 事象の認識

事象の認識モジュールでは、仮想空間と物理シミュレータから入力情報を受け取り事象を認識する。仮想空間から受け取る情報は、各オブジェクトに関する名前や大きさ、形状の情報、3 次元座標ベクトル、座標ベクトルから算出される移動距離、速度、加速度、カメラと

オブジェクトとの水平面上の距離、カメラ視野におけるオブジェクトの画素数と位置、である。物理シミュレータ [6] から受け取る情報は、オブジェクトの回転行列と他のオブジェクトとの衝突判定結果が格納された配列である。この入力情報をもとに、事象が起きたかどうかを判定する。認識する事象は運動の異なりから 4 つに分けられ、それぞれについて以下で述べる。

また、このモジュールの出力となる認識された事象は以下の要素で構成されるものとする。

- 主語：認識した事象において主動詞の主語となるオブジェクト
- 主動詞：認識した事象を表現する動詞
- 目的語：認識した事象において主動詞の目的語となるオブジェクト
- 事象の開始時間：事象の認識が開始された時間。現在時刻からその事象固有の認識に必要な観察時間だけ遡った時間
- 事象の終了時間：事象の認識が終了した時間。現在時刻が格納される。
- 事象の重要度：認識した事象が状況の変化を伝達する上でどれくらい重要かを表す値。

移動に関する事象

動詞毎に対応する速度と加速度の変化パターンを辞書に登録しておき、これと入力情報の速度と加速度の変化パターンを DP マッチング [5] で照合する。この結果から「移動する」「上がる」「落ちる」「跳ねる」「止まる」を認識する。

衝突に関する事象

入力情報の衝突判定結果をもとに、「ぶつかる」「跳ね返る」「上に乗る」をルールによって認識する。

回転に関する事象

入力情報の各オブジェクトの回転行列とオブジェクトの y 座標をもとに、「転がる」「倒れる」「回る」をルールによって認識する。

状況変化の説明に必要なその他のこと

上記の運動以外に状況変化の説明に必要なこととして、本研究では運動の結果残存、オブジェクトの見え方の変化、オブジェクトの運動方向を扱う。

- 運動の結果残存
  人間は運動の結果が状態として続いている場合に、結果残存として認識している。本研究でもこれを模倣して、「倒れる」「乗る」「止まる」の結果の状態を結果残存として認識する。
- オブジェクトの見え方の変化
  前述の運動には含まれないが、オブジェクトが見えなくなることも説明の上で重要と考え、「見えなくなる」「隠れる」を扱う。
- オブジェクトの運動方向
  「こっちに来る」と「奥へ行く」を扱う。

2.2 表現対象事象の選択

表現対象事象の選択モジュールでは、認識した事象の集合から状況の変化を端的に表現するのに適した事象の選択を行う。表現対象の選択は人間の選択方法を参考にする。人間の場合、何かの基準で重要と考えられる事象が生じたときに逐次発話している。また、一度発話したことをすぐに繰り返して発話しない。さらに、長い間継続していて、それが普通のことと考えられるものについても発話しない。そして、関係がある事象同士は、別々の文で発話するのではなく複文にして 1 度に発話する。この人間の選択基準に基づいて次のルールを作成し、これに則って選択する。

また、選択を行う前に認識した事象集合の整理や、運動が包含関係にある事象の統合を前処理として行う。

(1) 微小時間 (500ms) 毎に認識した事象集合を走査し、事象の重要度が閾値を超える事象がないか確認する。閾値を超える事象が複数存在する場合は、その中で重要度が最大のものを選択する。
(2) 一度選択された事象は、表現生成に使用した事象集合に一時的に保存する。この中に保存されている事象と、主語・主動詞・目的語の全てが一致する事象は、表現対象選択の候補から除外する。
(3) 初期状態や長い間継続している事象といった、認識されてはいても表現はされない事象を、知識として保存しておく。この集合に含まれる事象は、表現対象選択の候補から除外する。
(4) (1) で選択した事象と関係があると考えられる事象が認識した事象集合内にあるか確認する。関係があると考えられる事象の重要度が閾値を超えていれば、これも選択し複文の作成に利用する。

事象の重要度

表現の候補を選択する際に最も重要なのが、この事象の重要度である。本研究では、人間が何を基準として事象の重要度を考慮しているかを考察し、それに基づいて事象の重要度として以下の式を提案する。

$$value = \frac{priority * pixel * travel}{distance^2} \tag{1}$$

式 (1) において、$priority$ は、その事象を表現する動詞事態が持つ重要度である。$pixel$ は、仮想空間からの入力情報で、その事象の主語となるオブジェクトが占める視野中の画素数を指す。$travel$ は、その事象において主語となるオブジェクトが移動した距離を指す。$distance$ は、カメラとその事象の主語となるオブジェクトとの xz 平面上の距離を表している。

$pixel$ は、人間が視野において大きく映っているものほど注意を向けやすく、事象を認識し発話することが増えるという経験則に基づいていて、重要度が大きくなるように導入した。$travel$ は、移動量が大きくなるほど事象として認識されやすくなるという経験則に基づき、移動量が大きくなるほど事象の重要度も大きく

なるようにするために導入した。そして、距離が遠ざかるにつれて観察力が低下しあまり事象が認識されなくなることと、ある程度離れてしまうと、それ以上離れても観察力はあまり変わらないことから、$distance$ を導入し、$distance^2$ で割るようにした。さらに、人間は同じ距離にあり、同じ大きさで見えているオブジェクトが「移動する」と「落ちる」の運動として認識されたとき、「落ちる」を先に発話する。このように動詞間にも優先度があると考えられ、これを式に反映させるためのパラメータが $priority$ である。この $priority$ は、ヒューリスティックで値を決定する。

複文対象事象の選択

人間は関係がある事象を別個の文で表現するのではなく、一つの複文で表現している。本研究でも関係がある事象を複文で表現するために、以下のルールを作成した。事象同士の関係として生じた時間が近いことと、片方の目的語がもう片方の事象の主語となる、または主語が同一のときに、本研究では事象間に関係があるとして、複文で表現する。複文表現の対象となる事象は、表現対象事象の選択ルールによって選出された事象をもとに、ルールによって選択する。

2.3 端的表現生成

前節の表現対象事象の選択によって、選ばれた事象を自然言語表現に変換して出力する。その際、時制と文型は以下のように決定する。また、運動方向を表現する場合は事象の主動詞のアスペクトを変更している。

(1) **時制の調整**
単文の場合は選択された事象が、複文の場合は主節となる方の事象が現在も認識されていれば現在形、されていなければ過去形にする。

(2) **文生成**
単文の場合は、「主語が目的語に主動詞」とする。複文の場合は主節となる事象を event、従属節の事象を sub とすると、主語が異なる場合は「sub の主語が sub の目的語に sub の主動詞＋て、event の主語が event の目的語に event の主動詞」、主語が同一の場合は、「主語が sub の目的語に sub の主動詞＋て、event の目的語に event の主動詞」

そして、この手順で生成した表現を出力する。現在のシステムでは、文字をウィンドウに表示している。

3 実験

提案手法の有用性を確かめるために、3 つの評価実験を行った。それぞれについて以下で述べる。

3.1 人の発話とシステムの出力の比較実験

この実験では、本研究で作成したシステムで観察と出力を行ったときの観察対象となるエージェントの視野の画像を撮影しておき、この映像を被験者に見せて被験者の発話した内容との比較を行う。被験者に予め観察対象内にあるオブジェクトに関する知識を説明しておく。その上で、図 2 のような観察対象となる映像を見せ、この映像を見ていない人に映像中で生じたことを理解させるための発話をしてもらった。被験者は 3 人で、映像データはオブジェクトの初期配置や動く順番が異なる 2 種類を用いた。それぞれの被験者の発話を正解として、精度と再現率を求めた結果を表 1 に示す。各データに対する行の上段が精度、下段が再現率を表している。

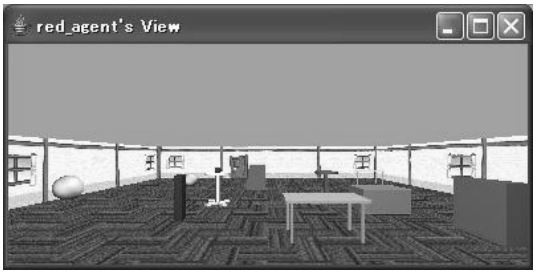

図 2: 被験者に観察してもらう映像の一場面

表 1: システム出力と被験者出力の一致度を調べた結果

	被験者 1	被験者 2	被験者 3
データ 1	$0.333(\frac{14}{42})$	$0.429(\frac{18}{42})$	$0.500(\frac{21}{42})$
	$0.700(\frac{14}{20})$	$0.818(\frac{18}{22})$	$0.913(\frac{21}{23})$
データ 2	$0.313(\frac{10}{32})$	$0.313(\frac{10}{32})$	$0.406(\frac{13}{32})$
	$0.666(\frac{10}{15})$	$0.588(\frac{10}{17})$	$0.650(\frac{13}{20})$
平均	0.323	0.371	0.453
	0.683	0.703	0.782

この表を見ると、再現率に比べ精度が著しく悪いことがわかる。システムが被験者の 2 倍の出力を行っていることがその主な原因と考えられる。システムの方が多く出力する原因としては、システムは人よりも広い部分を同時に観察することができ、認識する事象数が増えたことや、システムが誤った認識をしているために認識した事象数が増えたことが考えられる。また、再現率が最高でも 0.91 で、平均では 0.72 という値しかでなかったことについて考察する。この原因としては、被験者はオブジェクトの動き始めを発話することが多いが、システムでは動き始めの出力をあまりしていなかったことが挙げられる。この結果は、状況変化の理解において動き始めを人間が重要視していることの表れであり、今後これに対応する必要がある。

次に、精度・再現率の他に時系列での評価を行う。これは、精度・再現率を求める際にシステムの出力と人間の発話とで一致すると判定した事象について行う。評価方法は、一致した事象を 2 つ取り出し、その 2 つがシステムの出力と人間の発話とで同じ順序に並んでいるかの確認を総当たりで行うことで評価する。その結果をまとめたものを表 2 に示す。

この表を見ると、時系列順序の一致率の平均は 0.963 と良好で、システムは時系列順序に関して人間と同じように出力することができているといえる。

表 2: システムと被験者の出力における時系列順序の一致度

	被験者 1	被験者 2	被験者 3
データ 1	0.933	0.928	0.948
データ 2	1.000	1.000	0.967

表 3: システムの出力から復元した場合の精度と再現率

	精度	再現率
データ 1	0.826($\frac{38}{46}$)	0.844 ($\frac{38}{45}$)
データ 2	0.806($\frac{25}{31}$)	0.735 ($\frac{25}{34}$)
平均	0.816	0.790

表 4: 人の発話から復元した場合の精度と再現率

	精度	再現率
データ 1	0.917 ($\frac{22}{24}$)	0.489 ($\frac{22}{45}$)
データ 2	1.000 ($\frac{20}{20}$)	0.588 ($\frac{20}{34}$)
平均	0.959	0.539

3.2 システムの認識の正しさを調べる実験

一つ目の実験において、システムが誤った認識をしている可能性があることがわかったので、その確認のために実験を行った。この実験では、仮想空間で 1 つのオブジェクトのみを動かし、それについてシステムが出力した結果が、実際に被験者が認識できるかを評価してもらった。一つ目の実験と同じ 3 人に協力してもらい、システムが出力した 14 事象のうち 13 事象を認識できるという結果を得た。これより、概ね正しく認識できていることがわかったが、「倒れる」と「跳ねる」について少し問題があることもわかった。

3.3 人の発話やシステムの出力からの復元性を調べる実験

被験者による発話やシステムが出力した結果から、その状況の変化をどの程度理解し復元することができるのかを調べるために、次の実験を行った。仮想空間を模したミニチュアを用意し、一つ目の実験でシステムが出力した表現または、被験者が発話した表現を一つ目の実験とは異なる別の被験者に渡し、そこから想像できる状況の変化をミニチュア上で動かして再現してもらった。出力結果にないことは整合性がとれるように被験者自身に補完して動かしてもらった。もとの映像で認識できた事象を正解として、再現してもらった映像中での事象との一致度を精度・再現率として求めた。この結果を表 3、4 に示す。

表 3、4 を見ると、システムの出力からの復元は、人の発話からの復元と比べると精度は低いが、再現率が高いことがわかる。それも 25%も向上していることから、実験 1 で人とシステムの出力を比較したときの精度が悪かったことは、やむを得ないものであり、システムの方が状況変化を復元するのに有用な多くのものを認識できていると言うことができる。その一方で、システムの出力からの復元の精度は人の発話からの復元の精度よりも 15%ほど低い。一部は、人の発話からの復元の精度が 100%にならなかったのと同じで、出力されていない部分を被験者が補完した事象が間違っていたからだと考えられる。しかし、先述のようにシステムの方が詳しく述べているのならば、この割合は人の発話からの復元よりも小さいと考えられる。そこで精度の低下の主な原因として考えられるのが、余分な出力や誤った出力からの復元による精度の低下である。実際に一つずつ観察されたので、今後これらを修正する必要がある。

4 結論と今後の課題

本稿では、仮想空間を対象として、状況の変化を伝達する端的な表現の自動生成手法を提案した。また、表現対象の選択では、人間が事象の選択において重要と考えていることをモデルとした式を用いた事象の重要度を導入することで、人間と同じような事象の選択を可能にすることを試みた。

提案手法と人間の実況との比較実験を行い、提案手法の結果の人間の発話との一致度は低かったが、状況の復元性の面では人間の発話よりも良好な結果が得られた。これにより、提案手法では状況の復元に有効な情報をより多く表現できているということができる。そしてさらなる有効性の向上に向けて、実験の考察で述べたような事象の認識のルールと認識した事象からの選択ルールの修正や追加を行う必要がある。

また、本システムを発展させて実用化するためには、方向や位置関係に関する表現の追加と、人間の行動の認識と表現が必要であり、今後の課題である。

参考文献

[1] Patrick A.M.Ehlert, Quint M.Mouthaan and Leon J.M.Rothkrantz: A Rule-based and a Probabilistic System for Situation Recognition in a Flight Simulator, in Proceedings of the 4th Int. Conf. on Intelligent Games and Simulation (GAME-ON 2003), pp.201-207, 2003

[2] 小島篤博, 田原典枝, 田村武志, 福永邦雄：動画像における人物行動の自然言語による説明の生成, 電子情報通信学会論文誌, D-II Vol.J81-D-II No.8, pp.1867-1875, 1998

[3] 小島篤博：映像中の人物行動の認識とその自然言語記述に関する研究, 大阪府立大学博士論文, Jul., 2003

[4] 新山祐介, 秋山英久, 鈴木泰山, 徳永健伸, 田中穂積：自然言語を理解するアニメーテッドエージェントのための３次元仮想空間における位置の表現と処理, 第 13 回人工知能学会全国大会論文集, pp.217-220,1999

[5] DP mattching
http://sail.ishikawa-nct.ac.jp/pattern/dp/dp.html

[6] 物理シミュレータ Open Dynamics Engine
http://ode.org/