2G1-OS3-8

視覚情報から言語を生成するシステムの試作とその生成文の評価

A Prototype system for Generating a Variety of Language Expressions from Visual Information and Evaluate for the Expressions

野口 靖浩[*1] 麻生 英樹[*2] 高木 朗[*3*2] 小林 一郎[*4]
Yasuhiro Noguchi Hideki Asoh Akira Takagi Ichiro Kobayashi

近藤 真[*1] 三宅 芳雄[*5] 岩橋 直人[*6] 伊東幸宏[*7]
Makoto Kondo Yoshio Miyake Naoto Iwahashi Yukihiro Itoh

[*1] 静岡大学情報学部
Faculty of Informatics, Shizuoka University

[*2] (独)産業技術総合研究所知能システム研究部門
Intelligent Systems Research Institute, AIST

[*3] 言語情報処理研究所
NLP Research Laboratory

[*5] 中京大学情報理学部情報知能学科
School of Information Science and Technology, Chukyo University

[*4] お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

[*6] (独)情報通信研究機構
National Institute of Information and Communications Technology

[*7] 静岡大学
Shizuoka University

We describe about an overview of a prototype system for generating a variety of language expressions from visual information taken by CCD camera, and a rough evaluation for the system. A visual scene can be expressed by using a variety of surface dependency structures and words; however most natural language system can generate only some language expressions which are already prepared. In contrast, by virtue of flexibility of the semantic representation framework by Takagi and Itoh, our system generates a variety of language expressions. Our system makes a concept dependency structure from visual information using the semantic representation framework, transforms the structure and divides the structure into a word sequence. According to a rough evaluation in this paper, our prototype system can generate 335/437 (76.6%) of collected language expressions about 10 visual scenes from 10 subjects.

1. はじめに

人間とコンピュータとが共通の視覚情報を話題にしてコミュニケーションする場面では，コンピュータも視覚情報を言語で自由に表現できることが重要だと考えられる．人間が視覚で捉えた事柄を言語で表現する場合，常に一定の表現で表現する訳ではなく，その時々で異なる語彙，異なる構文構造を用いて表現することができる．例えば，同じシーンであっても，そのシーン中の何を対象とするか，対象としたもののどの属性に視点を置くか，更にはそのものに関係するどの現象に着目するかなどの違いによって，そのシーンを表す言語表現は大幅に異なってくる．

このような問題に対して現状では，表現する対象，属性，現象などをシーンの中から自由に選択した上で多様な表現を生成するための方法が十分に議論されているとは言い難い．例えば，自然言語入力と同時に視覚情報を扱う代表的なシステムとして SHRDLU[Wingrad 1972]が知られており，このシステムは人間との対話の中で言及された物体を，シーン内から特定して操作している．しかしながら，システムが生成する言語表現の構文や用語は限定的で，視点や重きをおきたい情報によって多様な表現を柔軟に生成することはできていない．また，久野の研究[久野 2006]では，実際の物体を画像認識した結果から得られた色，形，大きさ，位置の情報と，システムが持っている「りんご」などに関する知識(色:赤;形:円)を照らし合わせて，物体の特定を試みることができる．更に，一致するものがない場合には，画像から認識した情報を元に「赤色でない丸いものを見つけました」といった文を生成してユーザに問い合わせることもできる．しかしながら，このシステムにおいても，生成される質問文は，予め用意されたパターンに基づくものであり，対象とする視覚情報や利用する語彙，構文構造に対して柔軟性があるとは言い難い．

我々は，実世界を撮影した動画像の内容を多様な語彙，構文構造を用いて表現できる枠組みの構築を目的として研究を進めている．そして，現在までに単純なオブジェクトの限定された属性(色・形・大きさ・幅・高さ，位置の変化)を対象としたシステムを試作した．本システムでは，まず，視覚情報に基づき概念間依存構造表現[高木 1987][高木 1984]を生成する．この概念間依存構造表現は，以下の方針で設計されており，

(1) 概念間依存構造を素直にプリミティブ概念の隣接接続構造として表現する．
(2) 多様な言語表現を生成できるように，概念間依存構造を十分に詳細に記述する．

生成された概念依存構造を単語に分割することで可能な語彙，構文構造の組み合わせで生成する．

本稿では，試作したシステムの概要を説明する．そして，試作システムの現状評価として，あるシーンに関して被験者から収集

連絡先:野口靖浩，静岡大学情報学部，静岡県浜松市中区城北 3-5-1，TEL:053-478-1457，noguchi@inf.shizuoka.ac.jp

した表現と，それと同一のシーンに対してシステムが生成可能な表現との比較結果を報告する．

2. システムの概要

2.1 全体像

本システムの全体像を図 1 に示す．本システムは，撮影した動画像を画像認識し，その動画像に含まれる物体の属性や位置の変化（画像認識により単位時間ごとの座標値として得られる）などに関して言及する自然言語文を生成するシステムである．本システムは大きく「画像認識」モジュール，「概念間依存構造生成」モジュール，「単語分割処理」モジュールから構成されている．

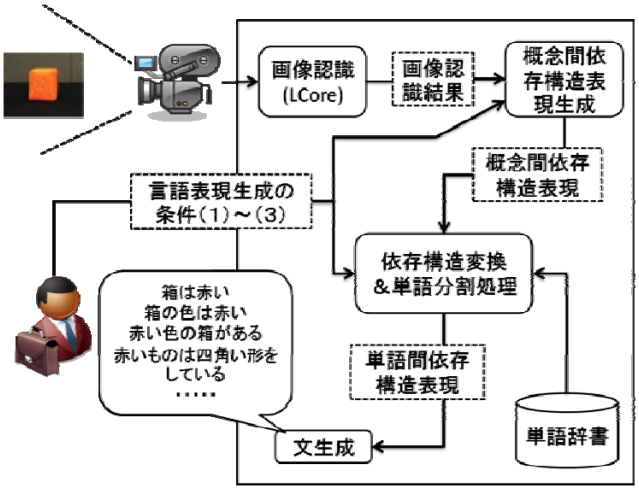

図 1: システムの全体像

撮影された動画像は，画像認識モジュールにて解析され，画像認識結果が出力される．次に「概念間依存構造生成」モジュールでは，画像認識結果を元に，動画像に含まれる概念に関する概念間依存構造表現を生成する．「依存構造変換＆単語分割処理」モジュールでは，概念間依存構造の依存構造を必要に応じて変換し，更に単語辞書中の単語概念に基づいて分割し，単語を割当てて，単語と単語間の依存関係を記述した単語間依存構造表現を生成する．最後に単語間依存構造表現を元に，語尾変化などの処理を行って，文を生成しユーザに提示する．

2.2 インタフェース

あるシーンの内容を言及する場合，そのシーン中の何を対象とするか，対象としたもののどの属性に視点を置くか，更にはそのものに関係するどの現象に着目するかなどの違いによって，様々な表現が存在する．また，用いる語彙，構文構造などの差異によっても，様々な表現が考えられる．そこで，現状のシステムでは，以下の要素の差異によって生じる表現の範囲を対象とすることとした．

(A) 主現象（主節動詞）の選択
(B) 現象の属性の言及・不言及，言及の仕方
(C) 現象に参画する対象の言及・不言及
(D) 対象の属性の言及・不言及，言及の仕方
(E) 単語の選択

本システムのインタフェースを図 2 に示す．そこでは，ユーザが以下の「言語表現生成の条件（1）～（3）」を指定し，その上で，システムが条件に応じた自然言語文を生成するようにしている．

（1） 画像中に含まれる1～3のどのオブジェクトについて自然言語文を生成するか？ [(C)に対応]
（2） どの属性あるいは現象を主に言及する自然言語文を生成するか？ [(A)に対応]
候補：色・形・大きさ・幅・高さ・位置の変化・速度・時間・距離
（3） 各属性に関して，自然言語文中で表現するか，否か？ [(B), (D)に対応]
候補：色・形・大きさ・幅・高さ・位置の変化・速度・時間・距離

ただし，(B)(D)の中の言及の仕方，すなわちどのような依存構造で言及するか，および(E)の単語の選択については，予め生成対象として特定の依存構造，単語を明示的に選択することはせず，全ての可能な候補の組み合わせで文を出力するようにしている．

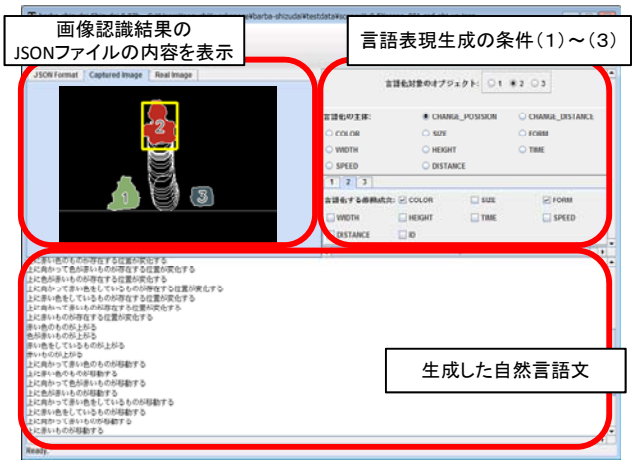

図 2: システム動作図

2.3 画像認識モジュール

本システムでは画像認識モジュールとして LCore[岩橋 2009] を用いている．LCore の画像認識結果を図 3 に示す．この画像認識結果は JSON 形式で出力され，オブジェクトごとに幅（"width"），高さ（"height"），大きさ（"area": 画面上のピクセル数），色（"Lab": L*a*b*表色系での色表現），位置の変化（"xyz": 単位時間ごとのオブジェクトの座標値）を示す．

```
{
  "sceneinfo" : {
    "2" : {
      "width" : 40, "height" : "62", "area" : 1193,
      "pointed" : "no", "prev_moved" : "yes",
      "Lab" : [  66.7652,  -24.2735,  27.4259 ],
      "xyz" : [
        [ 19.815599, 155.977005 ],
        [ 17.22963, 156.578064 ],
        [ 14.811731, 157.140045 ],
              [ 12.811731, 158.50342 ],
        :
        :
```

図 3: 画像認識結果

2.4 概念間依存構造生成・依存構造変換＆単語分割処理・文生成

画像認識結果から言語表現生成の条件（1）「画像中に含まれる1～3のどのオブジェクトについて自然言語文を生成する

か？」で指定されたオブジェクトに関する情報を取り出し，高木らの提案した概念間依存構造表現[高木 1987][高木 1984]に基づく概念間依存構造表現(図 4)を生成する．

更に，言語表現生成の条件(2)で指定された「主に言及する属性あるいは現象」の情報に基づき，指定された現象あるいは指定された属性を主に言及する現象を，主現象とする依存構造へと変換する．例えば，属性「距離」を主に言及する場合には，図 4 左上の連体修飾節「25pixel に等しい(距離)」部分の「等しい」現象が主現象となる以下の概念間依存構造表現へと変換する[野口 2010]．

「・・・・なものが
Y 軸に沿う＋に等しい方向に向かって
5sec に等しい時間をかけて
5pixel/sec に等しい速度で
変化する距離が
25pixel に等しい」．

その後，ヘッドのプリミティブ現象概念から順に単語を割当て，更にその割当てから余った枝に対して再帰的に単語割当てを実施していく．概念間依存構造表現へ割り当てる単語の差異によって，複数の単語間依存構造が生成可能な場合があるため，その全てのパターンを検証している．具体的な手続きに関しては[麻生 2010]を参照．

最後に単語間依存構造表現を再帰的に探索して，単語概念を1次元列に並べた後(例：「上」「に」「向かう」「て」「もの」「が」「移動する」)，語尾変化等を考慮して日本語文を生成する．

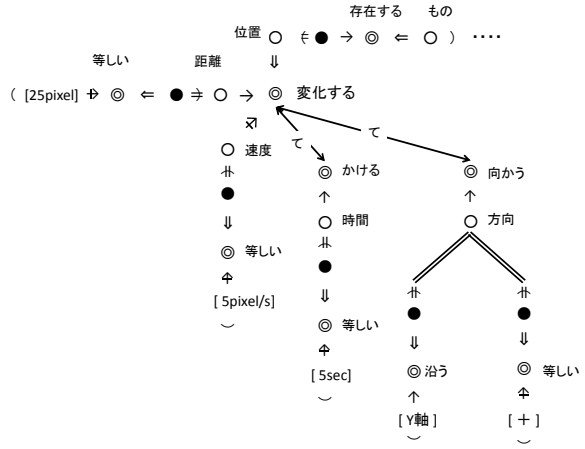

図 4: 概念間依存構造表現例

3. 評価

現状のシステムにおいて，画像認識モジュールの能力が及ぶ範囲で，語彙，構文を違えた表現がどの程度生成可能かを調査した．調査は，被験者から文例を収集し，その文例と現状のシステムが生成可能な文と比較する方法で行った．

被験者は情報系学部の大学生・大学院生を対象とした計 10 名である．被験者には，画像認識結果を示すシーン画像(図 3 で示した JSON 形式の画像認識結果から生成した)を提示し，そのシーン上のオブジェクトに関する日本語表現を用紙に記入してもらった．その際，各被験者はシーンごとに異なる表現を複数記入するものとした．その際，記入できる表現数の上限は設けなかった．

今回，被験者に提示するシーンとして 10 種類のシーンを選択した(図 5)．これらのシーンは，現状のシステムが有する画像認識の能力を考慮して，オブジェクトの色，形，大きさ，幅，高さ及び位置の変化に関する範囲で設定した．

被験者から収集した日本語表現の総数は 481 文である．収集した日本語表現の内，記述が口語体のものに関しては修正を行った．

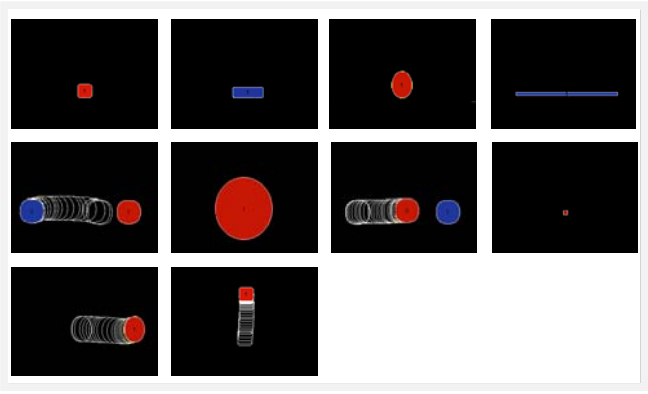

図 5: 被験者に提示したシーン画像

また，次のタイプの表現については，今回の比較対象から除外した．

- 画像認識で扱っていない情報を対象にした表現
  例：「丸いものはふちが白い」
  「角は丸みを帯びている」など
- 否定表現
  例：「赤い正方形のものの大きさはあまり大きくない」
  「青い物体は動かない」 など
- 比喩表現
  例：「赤い円は青い円に引き寄せられている」
  「赤いものが右に流れている」
  「赤色の細長いものが立っている」など
- メタな視点からの表現
  例：「赤い四角形が表示されている」など

今回は画像認識できたものから自然言語表現を生成するプロセスについて評価したいので，画像認識で扱っていない情報を対象にした表現については除外した．否定表現は，ある事象を表す際に，それ以外の事象を表す任意の表現を否定的に用いて表すことができ，そのような表現は無限に生成できるため，現状では扱わないこととした．比喩表現については，これを取り扱うためには視覚情報から得られる情報以外に，それと共通項のある別の物事に関する情報が必要になり，今回の範囲から外れるため除外した．結果，最終的に計 437 文と対象として，現状のシステムで生成可能な文との比較を実施した．

表 1 に被験者から収集した文と現状のシステムで生成可能な文との比較結果を示す．比較する際には，時制，相表現及び語順(複数の連体修飾表現の語順など，順序を変更しても構文的・意味的に問題のないもの)に関しては完全に一致していなくても良いものとした．表 1 の比較結果から，被験者から収集した表現の内 76.6%について，現状のシステムで同等の文を生成可能であることが分かる．現状のシステムで生成できた文例を分析すると，実際に被験者が用いた文例の内，あるオブジェクトの現象あるいは属性に着目した単一節の表現の範囲では，構文構造による表現の差異をほぼカバーすることができた．

表 1: 比較結果

	文数(割合)
現システムで生成可能な表現	335(76.6%)
現システムで生成不可能な表現	102(23.4%)
合計	437

被験者の入力した表現の中で，現状のシステムで生成できなかった表現について分析する．現状のシステムで生成できなかった表現について，その原因は大きく 2 種類に分類することができた．第一に生成した概念間依存構造表現中に当該の情報が含まれているが，生成した概念間依存構造表現が持つ依存構造上，被験者が指定した表現を生成できていないもの，第二に生成した概念間依存構造表現中に当該の情報が含まれていないものである．

(a) 生成した概念間依存構造表現中に当該の情報が含まれている
- 構文構造の変形が必要な場合
  例:「横に移動したのは，赤い楕円である」など
- 複数節構造で表現される場合
  例:「細長い円がゆっくり動いて，止まった」
  「丸いものがあって，赤色をしている」など

(b) 生成した概念間依存構造表現中に当該の情報が含まれていない
- シーン全体を視点に見る場合
  例:「青色の棒がひとつだけある」など
- オブジェクトの属性間の比較を必要とする場合
  例:「高さと幅はほぼ等しい」など
- 複数オブジェクトを主体に取る場合
  例:「2つの丸いものは離れている」
  「2つの丸いものは円に近い」など
- 異なるオブジェクトとの関係を必要とする場合
  例:「赤い物体が青い物体に左から近づく」
  「赤い物体は青い物体の左にある」など

現在のシステムでは，主に言及する属性あるいは現象を明示的に指定している．そして，指定された現象あるいは，指定された属性を言及する現象が主現象となるような依存構造で概念間依存構造を生成するようにしている．そのため，「横に移動したのは，赤い楕円である」のように，同じものを二つの異なる表現（「横に移動したの」「赤い楕円」）に分けて表現することや，複数節あるいは複数文に分けて表現することができていない．このような表現を生成するためには，概念間依存構造表現の変形が必要となるが，ある1文で表現される内容を複数節あるいは複数文に分けて表現するバリエーションは無数に存在するため，予めどのような表現を生成するかを決定する必要がある．

また，現在のシステムでは，ひとつのオブジェクトに関する色，形，大きさ，幅，高さといった属性や位置の変化といった現象に着目して概念間依存構造表現を生成している．しかしながら，「青色の棒がひとつだけある」のような表現は，シーン全体の中で同様のオブジェクトがいくつあるかを認識していないと生成できない表現である．同様に，「2つの丸いもの」といった表現もシーン中に存在する同様のオブジェクトを認識している必要がある．また，「高さと幅はほぼ等しい」においては，あるオブジェクトの「高さ」と「幅」という二つの属性に着目する見方が必要である．このように，シーンの見方を単純にシーン中の1オブジェクト，更にはそのオブジェクトの 1 属性に注目する見方から拡張して，画像認識結果から様々な見方の概念間依存構造表現を生成できるようにする必要がある．更に，「赤い物体が青い物体に左から近づく」のように，他のオブジェクトとの相対的な関係（この場合位置関係）が必要な場合もあり，これらの問題を解消するために，概念間依存構造を生成する段階で画像認識結果を捉える視点を増やす必要がある．

4. まとめ

本稿では，コンピュータが実世界を撮影した動画像を自由に自然言語で表現できるようにことを目的として作成したシステムについて述べた．提案したシステムでは，画像認識によって得られた色，形，大きさ，位置の変化などの情報から概念間依存構造表現を生成し，主節の主現象の選択に応じた依存構造の変形を行った後，単語辞書中の単語を割り当てて分割するという汎用的なアルゴリズムによって，様々な語彙，構文を用いた言語表現の生成を行う．また，第 3 章では，いくつかのシーンに対して，現状のシステムが生成可能な表現と人間が生成可能な表現とを比較し，現状のシステムが対応可能な表現の範囲と対応できない表現の範囲について分析を行った．分析の結果，現状のシステムでは，あるひとつのオブジェクトに着目して単一の節で表現する範囲では，被験者が言及したほとんどの表現を生成することができたが，ひとつ以上のオブジェクト，あるいは同一オブジェクトのひとつ以上の属性に着目した場合や，複数節あるいは複数文で表現する場合には対応できていないことが明らかになった．

今後の課題としては，この分析で明らかになった問題点を解消することが挙げられる．特に複数節あるいは複数文での表現は，今回対象としたシーンに限らず，より複雑な事柄を表現する上で必要不可欠な要素なので詳しく検討したい．その上で，扱えるオブジェクトの種類や現象の拡張を検討する予定である．例えば，現状では単一のオブジェクトが関与する現象のみを対象としているが，複数のオブジェクトが関与する現象を扱うようにしたい．また，現状のシステムでは全ての可能な文を出力しているが，シーンの状況や別途対話などから得られた情報などを利用して，冗長な表現を除き，より適切な表現だけを選択する枠組みを検討して行きたい．

参考文献

[Winograd 1972] Winograd, T.: Understanding Natural Language, Academic Press, New York, 1972

[久野 2006] 久野義徳: サービスロボットのための視覚と対話の相互利用，情報処理学会論文誌，Vol.47, No. SIG15 (CVIM 16), pp.22-34, 2006

[高木 1987] 高木朗，伊東幸宏：自然言語の処理，丸善，1987

[高木 1984] 高木朗，伊東幸宏，六沢一昭，北岡和憲，清水正朗，小原啓義：二次元図形世界における視覚情報からの日本語文の生成，電子情報通信学会論文誌 D, vol.J67-D, no.2, pp.216-223, 1984

[岩橋 2009] 岩橋直人: LCore: 言葉と動作によるコミュニケーションを学習するロボットの知能化技術，第 12 回情報論的学習理論ワークショップ，2009

[野口 2010] 野口靖浩，麻生英樹，高木朗，小林一郎，近藤真，三宅芳雄，岩橋直人，伊東幸宏：視覚情報から言語表現を生成するシステムの試作，第 58 回言語・音声理解と対話処理研究会(SIG-SLUD)技術報告，pp.43-48, 2009

[麻生 2010] 麻生英樹，野口靖浩，高木朗，小林一郎，三宅芳雄，岩橋直人，伊東 幸宏: 視覚情報から多様な言語表現を生成するための意味表現形式，第 24 回人工知能学会全国大会 2G1-OS3-7，2010