3H2-OS3-09

視覚情報を用いて対話を行うシステムの試作

Construction of Prototype Dialogue System Using Visual Information

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] 岩橋 直人[*5] 伊東幸宏[*6]
Makoto Kondo 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

[*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

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

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

[*6] 静岡大学
Shizuoka University

We describe about a prototype dialogue system using visual information taken by CCD camera. This prototype system integrates two types of semantic representations and their frameworks. One is able to generate a variety of language expressions from visual information with user's focus in their target scene. Another is able to interpret a variety of semantic relations leaded by a variety of language expressions. In this paper, we explain about the overview of the prototype dialogue system which supports a dialogue sharing a scene in sight.

1. はじめに

人間・コンピュータ間で共通の視覚情報を対象としたコミュニケーションを行うことを考えた場合，人間が視覚で捉えた事柄を言語で自由に表現できるのと同様の能力をコンピュータが持つことが重要な課題だと考えられる．通常，自然言語処理システムでは，ユーザからの自然言語入力やシステムに予め構造的に記述された知識に関して問題解決を扱う場合が多かったが，近年ではカメラ等から得た視覚情報を取り扱う技術についても関心が高まってきている．我々は，コンピュータ上において視覚情報を多様な言語表現で表現するための意味表現形式と意味表現から言語表現を生成する手続きについて検討してきた[高木 1987][麻生 2010]．そして，その枠組みに則ったシステムを構築し，人間が作る文とシステムの出力文とを比較することで，提案意味表現形式と生成手続きの有効性を評価した[野口 2010]．

また一方で，自然言語の処理に関して，自然言語が同一の意味内容を非常に多様な言語表現で記述することができるのに対して，システム側はその多様性に対応することが困難という問題がある．我々はこの問題に対して，文を構成する各依存関係を「属性名＝属性値（属性は属性値である）」という均一な「断定」表現の形式で記述する意味表現方式を用いて，一定の手続きを通して文の意味を解釈・蓄積する方法を検討し[Takagi 2006]，その枠組みに則った対話システムを評価した[野口 2008]．

本研究では，これらの意味表現形式及び処理手法を用いて，視覚情報を用いて対話を行うシステムの実現を目指す．本稿では，非常に単純なオブジェクトで構成されたシーンを対象として対話を行うシステムを試作した．3 章，4 章でその枠組みについて報告し，最後に 5 章でまとめと今後の課題について述べる．

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 3-5-1，053-478-1457，noguchi@inf.shizuoka.ac.jp

2. 関連研究

自然言語と同時に視覚情報を取り扱う代表的なシステムとしては，Winograd らによる SHRDLU システム[Winograd 1972]が知られている．このシステムでは，予め設定した仮想的な積木の世界に関して，人間とシステムとの間で対話を行い，積木の操作を行うことができる．近年では，新山らが仮想世界上のロボットを音声入力によって操作するシステム傀儡を開発し[新山 2001]，特にユーザの視点の違いが照応解決の問題に与える影響について検討している．

また，幾つかの研究が仮想的な世界の中のオブジェクトの照応表現を生成することを目的として行われている[Dale 2009][Albert 2008][Viethen 2008]．最近では，特定の仮想世界を対象とした照応表現の生成能力を競うチャレンジが実施されており[Albert 2009][Byron 2008]，実際に実装されたシステムを用いた比較・評価が行われている．

久野ら[久野 2006]は，サービスロボットを対象として，カメラで撮影した実画像中のオブジェクトの照応解決の問題を扱っている．そこでは，ユーザが言及した自然言語文からオブジェクトの属性に関する情報を読み取って，それに対応するオブジェクトを実画像中から照応する方法や，実画像から認識したオブジェクトの情報から照応表現を生成して，ユーザに問い合わせて照応解決を行う試みがなされている．また，視覚情報と言語表現との対応を学習するシステムの研究も行われている[Iwahashi 04][Roy 02a][Roy 02b]．

これらの研究で検討された，あるいは新たに提起された課題は，本研究でも共通しているが，いずれの研究においても言語表現の文体の多様性に関してはあまり扱っていない．本研究では，対象とする世界は強く限定しているが，文体の多様性の扱いについては特に意識して取り組む．

3. システムの概要

本稿で試作したシステムの処理の流れを図 1 に示す. 本システムにおいてシーンは, PointGrey 社 Bemblebee2 及び SwissMesa 社 SwissRanger SR4000 の 2 台のカメラで撮影している. また, 本システムでは, ユーザとシステムとは同じシーンを同一方向から参照している状況を想定している.

本システムは, カメラからシーンの動画像情報を受け取って, そこに含まれる情報の一部を自然言語文に変換して表現するモジュール(右上点線枠内)[野口 10]と, シーンに関する対話を行う対話モジュール(下点線枠内)[野口 08]とで構成している. この 2 つのモジュール間では, シーンに変化が生じた場合に行われる“シーンの変化通知”, “シーンの情報を表す言語表現の生成リクエスト“, 及び”その生成結果“がやり取りされる. ユーザの入力は対話モジュールに対してキーボード入力で行う. システムからの出力はテキストと音声で出力する.

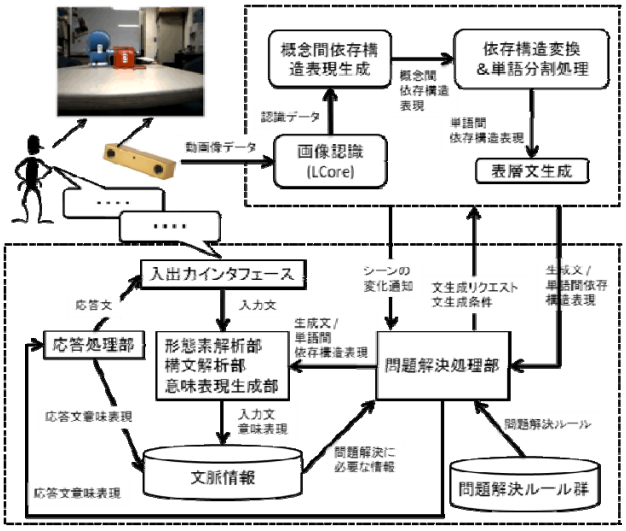

図 1：システムの処理の流れ

システム全体のインタフェースを図 2 示す. 左側のウィンドウに現在認識しているシーン動画と認識中のオブジェクトの情報が表示される. ユーザは入力文を右側のウィンドウ上側(図 2 中の対話データ部分を表示している部分)から入力し, システムは応答文をテキストと音声で出力する. 文脈情報に蓄積された入力文・出力文・シーン情報を表す文の意味表現は右側のウィンドウの下側に表示することができる.

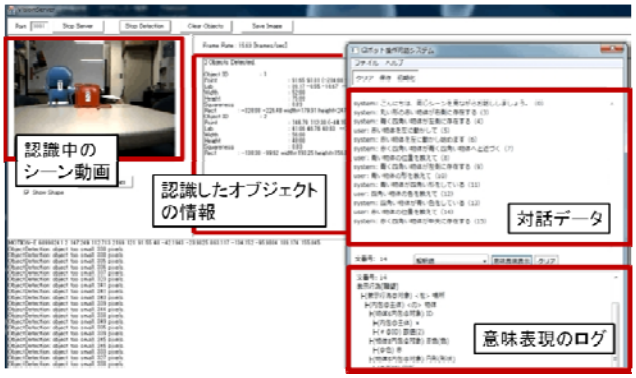

図 2：システムのインタフェース

3.1 視覚情報を言語表現で表現するモジュール

シーンは 2 台のカメラによって常時撮影される. カメラから取得した動画像データは, LCore[岩橋 2009]のシステムによって解析され, 「色」「形」「大きさ」「幅」「高さ」「位置の変化」の情報を含む認識データに変換される.

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. 最後に表層文生成処理において, 単語間依存構造表現から, 割り当てられた単語を 1 次元列に並べ替えることにより, 文を生成することができる.

本モジュールでは, 主現象(主節動詞)の選択, 現象の属性の言及・不言及及び言及の仕方, 現象に参画する対象の言及・不言及及び言及の仕方, 対象の属性の言及・不言及及び言及の仕方, 単語の選択に関して, 文法的にあり得る候補を網羅的に生成することができる.

(1) 画像中に含まれるどのオブジェクトについて自然言語文を生成するか？
(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？
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？
候補：色・形・大きさ・幅・高さ・位置の変化・速度・時間・距離

3.2 対話モジュール

ユーザの入力文は, 入出力インタフェースを通して「形態素・構文解析・意味表現生成部」に送られ, 意味表現へと変換され文脈情報として蓄積される. 文脈情報は「属性名＝属性値(属性は属性値である)」という均一な「断定」表現の形式で記述する. 文脈情報に新たな意味表現を追加するたびに先行文脈意味表現との意味の突き合わせを実施し, 問題解決処理部から文脈を考慮した意味の参照を可能な状態を保つ.

問題解決処理部は問題解決ルール DB から対応する問題解決ルール群を読み出した後, 文脈情報から問題解決に必要な情報を参照してルールを解釈する. 問題解決ルール群は if-then 形式で記述されたルールの集合である. 問題解決処理部は必要に応じてルールを実行し, ユーザの入力文に対応する応答文意味表現を生成して応答処理部に渡す. 応答処理部は, 応答文意味表現から応答文の文字列を生成し, 入出力インタフェースを介してユーザに提示する. 同時に, 応答文意味表現を文脈情報として蓄積する.

シーンの情報は, 問題解決処理部が「視覚情報を言語表現で表現するモジュール」から生成文／単語間依存構造表現として取得する. 問題解決処理部は取得した生成文／単語間依存構造表現を変換し, 文脈情報として蓄積する. その結果, シーンの情報についても, 入出力文と同等の意味表現の形式で文脈情報として蓄積されることになる. シーンの情報を取得し文脈情報として蓄積するタイミングは, 対話の中でシーンの情報を必要とした時点と, シーンの側の様子が変化した時点としている. シーンに含まれるどのような情報をどのような形式の自然言語文で取得するかについては, 対話モジュール側から「視覚情報を自然言語に変換するモジュール」に文生成リクエストを出す際

に，文生成条件を付けることで対応している．その選択方法について次章で述べる．

4. 生成条件の選択

視覚情報は非常に多くの情報を含んでおり，状況に応じて情報を取捨選択した上で取り扱う必要があると考えられる．本システムの一部を構成する「視覚情報を自然言語文に変換するモジュール」では，主現象，その現象に参画する対象，その対象の属性という観点から，そのシーンの中でフォーカスしている箇所を指定し，その指定に基づいてシーンの情報を表す自然言語表現を生成することができた．しかし，それは人間が明示的に指定しており，状況に適した表現を自動的に選択して生成できる訳ではない．そこで，本試作システムでは，実際に人間が用いる表現を参考にして「シーンに関する対話を行う対話モジュール」が視覚情報から文脈情報へと取り込む情報を選択できるよう拡張することにした．

状況に応じて適切な表現を選択する基準として，本試作システムで扱うシーンを対象に以下の仮説を設定し，被験者を募って文を収集してその結果を分析した．

- そのシーンの中でオブジェクトを一意に特定するために必要不可欠な属性が優先的に言及される
- そのシーンが話者間で共有されているかどうかが文の選択に影響を与える

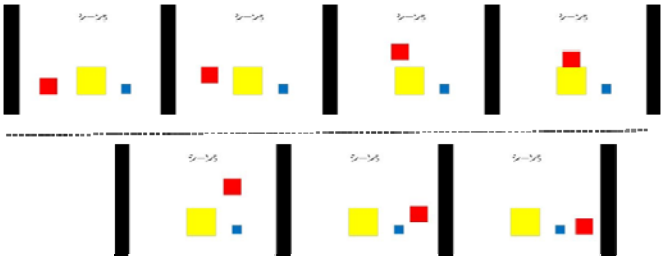

図 3：被験者に示したシーンの例

被験者は情報系の大学生 9 名である．これらの被験者に対してシーンを提示し，そのシーンの内容を表現する文を記述してもらう方法で文例を収集した．被験者に提示したシーンは，パワーポイントのアニメーション機能を使って作成した．各シーンは最初に 3 つのオブジェクトが並んだもので，各オブジェクトは色・形・大きさの属性を持つ．用意したシーンにおけるオブジェクトの動作の例を図 3（実際のシーンはアニメーション表示）に，各シーンの属性の組み合わせを表 1に示す．

表 1：属性の組み合わせ

	1	2	3	4	5	6	7	8
色	○	○	○	○	×	×	×	×
形	○	○	×	×	○	○	×	×
大きさ	○	×	○	×	○	×	○	×

○：シーン中のオブジェクトをこの属性の情報から一意に特定可能

×：シーン中のオブジェクトをこの属性だけでは一意に特定することは不可能

シーン中のオブジェクトを属性の情報によって一意に特定できるか否かを基準に組み合わせを設定し，表 1の通り 8 パターンを設定した．更にこれらの各パターンに対して 2 シーンを用意し，合計 16 シーンを用意した．更に，形属性に関して，言葉で言及しにくい形のシーンを 2 パターン用意した．内 1 シーンは色属性でオブジェクトを一意に識別可能なシーンを設定し，もう一方のシーンは色・大きさが同じに設定した．このようにして，被験者に提示するシーンとして合計 18 パターンを用意した．

更に，被験者から文例を収集する際には，被験者自身と被験者が記述した文の読み手との間の関係を何も指定しない場合と，被験者と被験者が記述した文の読み手とが同じシーンを見ている状況を指定した場合の 2 パターンの状況を設定して，文例収集を行った．被験者から収集した文例は，被験者と被験者が記述した文の読み手とがシーンの情報を共有している場合について 108 文，特に指定しない場合について 211 文である．これらの文例の中でオブジェクトの参照に用いられた参照表現はそれぞれ 415 箇所，674 箇所である．

これらの参照表現の分析結果を表 2，表 3，表 4に示す．

表 2：参照表現中で言及された属性の割合

	指定なし	シーン共有
参照表現数	674	415
形属性	546(81%)	233(56%)
色属性	338(50%)	161(39%)
大きさ属性	68(10%)	33(8%)
位置	133(19%)	139(33%)

表 3：オブジェクトを一意に識別可能であるが言及されなかった属性の割合

	指定なし	シーン共有
形属性	4(1%)	30(15%)
色属性	65(19%)	37(17%)
大きさ属性	123(63%)	82(64%)

表 4：オブジェクトを一意に識別不可能であるが言及された属性の割合

	指定なし	シーン共有
形属性	257(74%)	89(41%)
色属性	82(25%)	3(1%)
大きさ属性	5(1%)	0(0%)

表 2は，参照表現の中で言及されている属性の割合をまとめたものである（参照表現の中で複数の属性に言及する場合があるので割合の和は 100%を超えている）．表 3は，ある属性によってシーン中のオブジェクトを一意に特定することが可能な場合に，参照表現の中でその属性を言及しなかった割合をまとめたものである．表 4は，ある属性によってシーン中のオブジェクトを一意に特定することはできないにも拘らず，参照表現の中でその属性を言及した割合をまとめたものである．

収集した文例から，当初想定した通りあるシーンの中でオブジェクトを一意に特定することができる属性が言及される傾向が確認された．また，表 2，表 3，表 4から，用いられる属性に傾向があり，今回対象としたシーンにおいては，形，色，大きさの順に使われる頻度が高いことが分かる．特に形，色が多く使われる．大きさ属性は相対的な表現になることから考えると，この結果は当初想定したオブジェクトを一意に特定することができる属性が言及される傾向に則ったものと考えられる．また，表 4からは，その属性の情報によってオブジェクトを一意に特定できない場合であっても，オブジェクトを一意に識別しうる可能性が高い属性の情報を言及する傾向が伺える．

シーンの共有・非共有に関しては，シーンが共有されていれば，オブジェクトを一意に識別しうる属性以外を言及する頻度が下がる傾向があった．仮説以外の傾向としては，複数のオブジ

ェクトを言及する際には，それが可能ならば表現を揃える傾向があった．また，動きのないオブジェクトについては「～がある」のようにオブジェクトの存在を言及する文が多かったが，逆に動きのあるオブジェクトについては，「赤いものに向かって動いている箱がある」のように，現象を連体修飾成分で表現して文としては存在を言及する表現はあまり用いられなかった．

本稿で試作したシステムは，ユーザとシステムとが同一の方向からシーンを共有する前提とし，上述の結果を元にルールを設定して，「シーンに関する対話を行う対話モジュール」から「視覚情報を言語表現で表現するモジュール」に対する文生成リクエストの生成条件を制御するようにした．また，話者は相手と単語・構文の似た方法で表現することを好む傾向がある[Zender 2009]などから，対話でユーザの用いた表現を極力利用した文を生成するように条件を制御している．

5. 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用いて自然言語で対話を行うことを目的として，非常に単純なオブジェクトで構成されたシーンに関して対話を行うシステムを試作した．ベースとした対話システムは，文を構成する各依存関係を「属性名＝属性値（属性は属性値である）」という均一な「断定」表現の形式で記述する意味表現方式を用いている[野口 08]．本試作システムでは，入出力文に加えてシーンを撮影した得た視覚情報についても，この言語表現を表す意味表現と同等の表現へと変換して文脈に蓄積できるように拡張した．その結果，ユーザの入力文，システムの出力文に加えて，シーンを撮影して得た視覚情報についても，従来システムと同様に「属性＝属性値（属性は属性値である）」という均一な「断定」表現間の意味の比較・統合の枠組みを通して，対話システムを実現することができた．試作システムにおいて，特定のシーンに関してではあるが，実際に対話を行えることを確認できた．

視覚情報に含まれる多種多様な情報から対話の中で話題に上がっている情報を扱う方法として，今回は対象とする問題解決・シーンを限定した上で，ヒューリスティックスを用いた選択を行った．しかしながら，実際にはシーンの状況に加えて，それまでの対話文脈の内容や，扱うタスクの種類などによって，必要な情報・その表現形式は変化し，システム側はそれに柔軟に対応する必要があると思われる．また，話者とシステムとが同一の方向から同じシーンを見ている状況に限定してシステムを試作し，例文の収集・分析においても，シーンを共有している場合は同一方向からシーンを見ている状況を暗黙的に想定していたが，今後は話者とシステムとが異なる角度からシーンを見ている状況など，今後，シーンの捉え方が異なる状況についても分析を進める必要があると考えている．

謝辞

本研究を実施するにあたって，貴重なご意見を頂きました放送大学教養学部文化科学研究科の三宅芳雄先生に感謝いたします．

参考文献

[高木 1987] 高木朗, 伊東幸宏: 自然言語の処理, 丸善, 1987.

[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.

[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：視覚情報から言語を生成するシステムの試作とその生成文の評価，第 24 回人工知能学会全国大会，2G1-OS3-8，2010.

[Takagi 2006] Takagi, A. et al.: Semantic representation for understanding meaning based on correspondence between meanings, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.10, pp.876-912, 2006.

[野口 2008] 野口靖浩, 池ヶ谷有希, 小暮悟, 近藤真, 麻生英樹, 小林一郎, 小西達裕, 高木朗, 伊東幸宏: "文脈への意味の位置付けを用いた対話システムとその評価", 知能と情報(日本知能情報ファジィ学会誌),Vol.20, No.5, pp.732-756, 2008.

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

[新山 2001] 新山祐介, 徳永健伸, 田中穂積: 自然言語を理解するソフトウェアロボット: 傀儡, 情報処理学会論文誌, Vol. 42, No.6, 2001

[Dale 2009] Dale, R., and Viethen, J.: Referring Expression Generation through Attribute-Based Heuristics, Proceedings of the 12th European Workshop on Natural Language Generation, 2009.

[Albert 2008] Albert, G. and Anja B.: Attribute Selection for Referring Expression Generation: New Algorithms and Evaluation Methods, Proceedings of the 5th International Natural Language Generation Conference, 2009.

[Zender 2009] Zender, H., Kruijff, G.M., Kruijff-Korbayova, I.: A Situated Context Model for Resolution and Generation of Referring Expressions, Proceedings of the 12th European Workshop on Natural Language Generation, 2009.

[Viethen 2008] Viethen, J., Dale. D.: The Use of Spatial Relations in Referring Expression Generation, Proceedings of the 5th International Natural Language Generation Conference, 2008.

[Albert 2009] Albert, G., Anja., B and Kow., E: The TUNA-REG Challenge 2009: Overview and Evaluation Results, Proceedings of the 12th European Workshop on Natural Language Generation, 2009.

[Byron 2008] Byron, D, Cassell, J., Koller, A., Dale, R., Striegnitz, K., Moore, J., and Oberlander, J.: Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE), Proceedings of the 5th International Natural Language Generation Conference, 2008.

[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, pp.22-34, 2006.

[Iwahashi 2004] Iwahashi N.: Active and unsupervised learning for spoken word acquisition through multimodal interface, Proceedings of 13th IEEE Workshop Robot and Human Interactive Communication, pp.437-442, 2004.

[Roy 2002a] Roy, D. K., Pentland, A..: Learning words from sights and sounds: A computational model, Cognitive Science, Vol.26, No.2, pp.113-146, 2002.

[Roy 2002b] Roy, D. K.: Learning visually-grounded words and syntax for a scene description task, Computer Speech and Language, Vol.16, No.3, 2002.

[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.