マルチモーダル入力、マルチメディア出力の案内システム：MMGS

森元逞　竹澤寿幸

{morimoto, takezawa}@itl.atr.co.jp
ATR音声翻訳通信研究所

試作したマルチモーダルな情報案内システムについて述べる。　人からの入力は音声と手書きジェスチャによるマルチモーダルとし、機械からの出力は、３次元図形を中心としたマルチメディア機能を積極的に活用している。 また、人との自然な対話インタラクションを行えるよう、照応処理や音声認識誤りなどに起因する曖昧さの解消を行うことにより、協調的な応答を行う。

Multimodal / Multimedia Guidance System : MMGS

Tsuyoshi MORIMOTO, Toshiyuki TAKEZAWA
ATR Interpreting Telecommunications Research Laboratories

A multimodal/multimedia guidance system is described. A user can input with the combination of speech and hand-written gestures, and the system outputs response with the combination of speech, three-dimensional graphics and other information. The system can interact with a user cooperatively by resolving ellipses/anaphora and some ambiguities such as those caused by speech recognition errors.

1. はじめに

近年、音声で対話できるシステムが実現されてきている。さらに親しみ易さや、使い易さの向上を狙いとして、音声だけでなく複数のモダリティを用いたマルチモーダルなシステムもいくつか提案されている[1]~[4]。

本論文では、我々が開発しているマルチモーダルな情報案内システムについて述べる。本システムを開発するにあたり、我々は以下の2点を基本的な方針とすることとした。

(1)ユーザインタフェース

人からの指示は、音声ならびに、表示された画面を指し示しながら行うことができるように

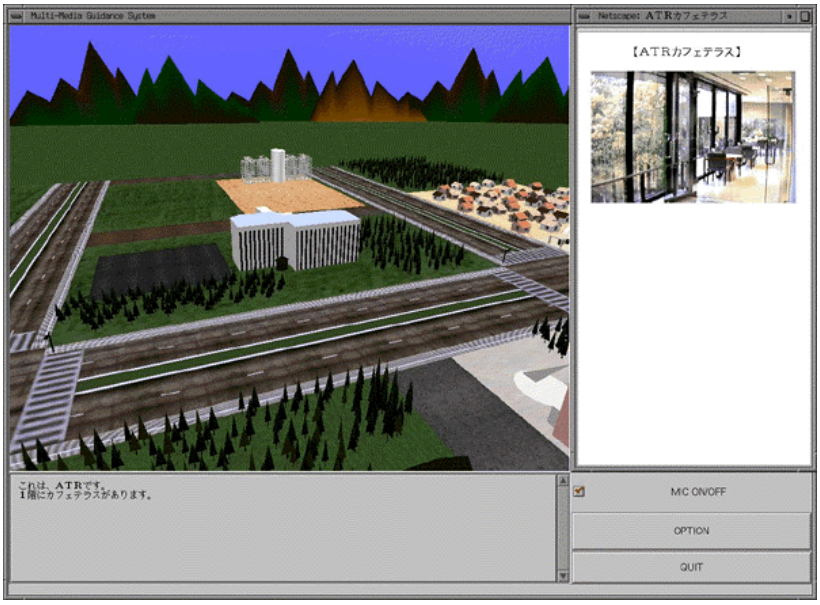

図１　システムの表示画面例

する。機械からの応答は、音声と画像情報を用いる。特に画像としては、人が直感的に理解しやすい3次元(3D)図形を用いる。またその表示にあたっては、視覚的な効果を充分に考慮する。さらに、案内対象物に関し、その他の情報(写真などの静止画情報など)があれば、それも積極的に表示する。以上を要約すると、「人からはマルチモーダル入力を可能とし、機械からの出力は、機械の持つマルチメディア機能を積極的に活用する」と言える。

(2)対話のインタラクション

人と機械とが自然な対話のインタラクションを行えるようにするため、「代名詞やゼロ代名詞」による参照が何を示しているのかを理解し(すなわち、照応処理を行い)、適切な応答を行えるようにする。また、音声認識誤りなどのため曖昧性が発生することがあるが、このような曖昧な入力に対しても、不明な部分を照応処理により積極的に補完し、対話を継続させることができるようにする。

以下では、以上のような方針に基づき開発したシステム の構成および具体的な機能を述べる。

2. システム構成と機能

ATRの周辺の道路や建物の案内を行う。画面上に道路、建物、公園などが3D図形で表示される。図1に例をしめす。また本システムの構成を図2にしめす。本システムはSGIワークステーション上に実現されており、音声認識、応答出力、画面更新などをほぼ実時間で行うことができる。

以下では、まず案内情報のオブジェクトによる定義・管理について簡単に述べる。その後、ユーザインタフェース、対話インタインタラクションの2つの観点から、本システムで実現した機能について述べる。

2.1 オブジェクトの定義・管理

質問応答システムでは、質問対象となる事物の意味情報(意味モデル)を用意することが必要である。入力発話の意味と、この意味モデルのマッチングにより、適切な応答を返すことができる。また、3D図形で表示すべきオブジェクトをオブジェクトデータベースで管理する方法が提案されている[5]。本システムでは、基本的にこれらの両者を統合した方式を実現している。

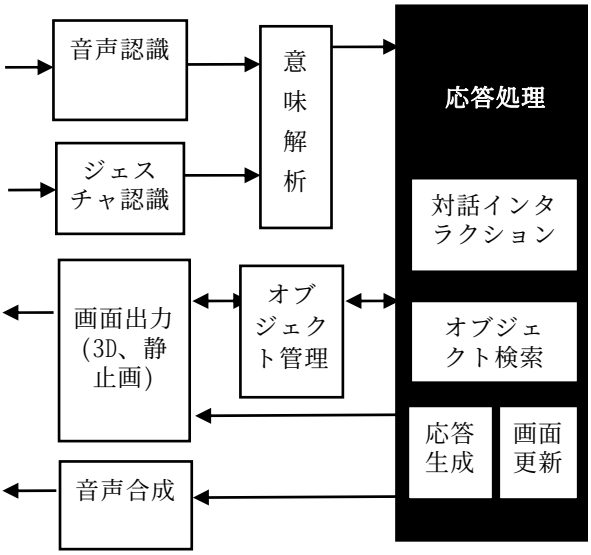

図2 システム構成

すなわち、3D図形で表示されるものについて、その「意味情報」と「空間・形状情報」とをあわせて「オブジェクト」として定義・管理している。オブジェクトはクラスに従って階層的に定義され、名前や各種属性などの意味情報が定義されている。またあるインスタンス・オブジェクトが3D図形として表示すべきものであれば、その空間・形状情報として、「座標」、「形状」、「テクスチャ」などが定義されている。さらに、オブジェクト間にcomposed-of関係があれば、その関係も定義されている。なお以下ではcomposed-of の上位にあるオブジェクトを親オブジェクト、下位にあるオブジェクトを子オブジェクトと便宜的に呼ぶことにする。図3にオブジェクトの定義例をしめす。

2.2 ユーザインタフェース

(1)入力

音声認識用の日本語文法と意味解析規則は一体化して定義されている。すなわち、日本語文法は、文脈自由文法形式で記述されており、各文法規則には素性構造に基づく意味解釈規則が付加されている。

音声認識では、文法規則部のみを用いてHMM-LR手法により認識を行い、認識結果として適用した文法規則番号の系列(すなわち構文木構造)を出力する。意味解析部ではこの規則番号系列から、対応する意味解釈規則を取り出し、対応する素性構造間で単一化処理を行うことに

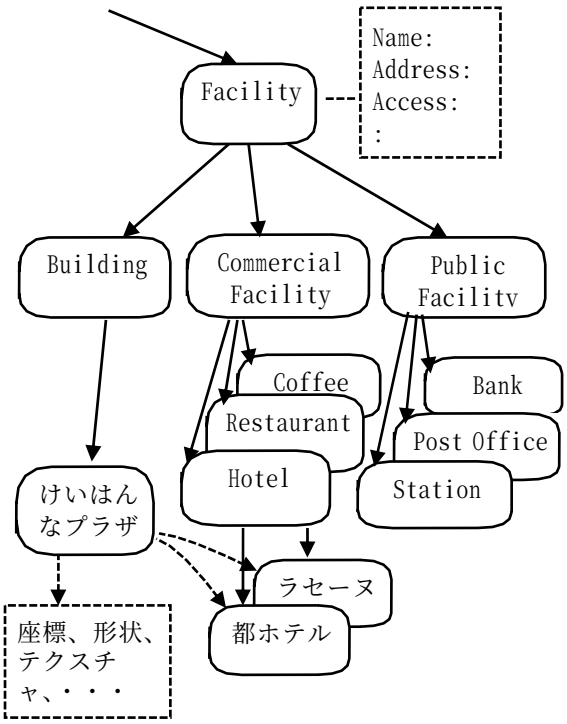

図３ オブジェクトの定義（一部）

より、一発話全体の意味素性構造を求める。

利用者は、各オブジェクトに関し、「これはATRですか」のようなyes/no-質問、「この建物はなんですか?」のようなwh-質問、ならびに「・・・ですね」などのような確認ができる。さらに、「この建物、何ですか?」のような助詞を省略した表現や、「この建物、何?」のような述語まで省略した断片的な表現、「えー」などの間投詞を挿入した表現などもサポートしている。これにより、利用者は自然な発話を行うことができる。さらに、一部に音声認識誤りがあっても、認識できた部分の意味構造を求め、上記の断片的な発話がなされた場合と同様な処理を行うことにより、利用者に協調的に応答することを可能としている。

「この」、「これ」などの指示的表現と同時に、手書きジェスチャで対象を指示することができる。現在サポートしているジェスチャ種別は、「ポインティング」、「サークリング」であり、文献[6]で報告した手法によりジェスチャの形状を認識し、その意味構造を対応する発話の意味素性構造に埋め込む。

(2)出力

SGIのシーン表示機能を用いて3D図形を表示する。また、後述するように、話題となっているオブジェクトが最も良く見え位置まで視点位置を自動的に移動させる。応答時は、応答文を合成音声で出力するとともに、利用者の注意を向けさせるため、対象のオブジェクトをブリンキングさせる。もしあるオブジェクトに関連する他の情報があれば、それも出力する。現在のシステムでは静止画情報のみを対象とし、それを別ウインドウに表示している(図1参照)。遠方にあって画面には表示されていないオブジェクトに関しても質問ができるようにするため、「遠方オブジェクト」を定義している。例えば、「京都はどちらですか? 」のような質問を行うことができ、このような質問に対して、システムは矢印をしめしながら、「こちらの方向です」と応答する。応答文の生成については、現在の回答内容が比較的単純であるため、固定的な文パターンへ変数を埋めこんで応答文を生成する方法を採用している。

2.3 対話のインタラクション

(1) 話題の把握

画面出力、照応処理などを適切に行うため、各時点での話題を把握する。基本的には、「は」でマークされたものに対応するオブジェクトを話題として取り出している。従って、「ATRはどこですか?」では「ATR」が話題であり、「この建物(手書きジェスチャによる指示)はATRですか?」では「指示され建物」が話題となる。

本システムでの問合せは、基本的に「オブジェクト」とその「属性」または「composed-ofによる子オブジェクト」との関係についてである。従って、話題も前者に対する「親話題」と、後者に対する「子話題」という2種類を用意し、それぞれ個別に管理している。

話題と画面表示の関係について述べる。前述したように、あるオブジェクトが話題(親話題)である場合、そのオブジェクトへ視点を移動させる。また応答を返す場合、応答出力中に対象のオブジェクトをブリンキングさせる。多くの場合、話題となっているオブジェクトと応答対象のオブジェクトは同じものである。しかし、利用者が間違った質問を行った場合は、両者は

必ずしも一致しない。

U: これはATRですか?
S: いいえ、それはけいはんなプラザです。 ATRはこちらです。

この例では、話題は「これ」でマークされた建物である。従ってシステムからの応答の第2発話で言及されている「ATR」への視点の移動は行わず、単にブリンキングを行う(ATRが画面上に表示されていなければ、それが表示される位置まで多少動かした後に)だけである。

(2)照応処理、協調的な応答

具体的な対話の例を用いて説明する。

U1: 喫茶店はありますか?
S1: どの場所でしょうか?
U2: ATRです。
S2: ATRにはカフェテラスがあります。
U3: けいはんなプラザにはないんですか?
S3: けいはんなプラザにはラインという喫茶店があります。
U4: そこにレストランはありますか?
S4: けいはんなプラザの中にですか?
U5: はい。
S5: けいはんなプラザの中には、「えん」、「ラセーヌ」があります。

まず、U1で喫茶店に対する質問が行われた。しかし親であるような喫茶店オブジェクトがなく、すべてが子オブジェクトであったため、S1で親オブジェクトを問合せている。U2での補足により、システムは問合せ内容が確定でき、その結果をS2として応答する。

次に、U3の問合せが行われたが、何があるかが言及されていない。システムはその時点の子話題である「喫茶店」を補完し、応答を行う。この時、補完したものが「喫茶店」であることを確認するという意味も含めて、応答中に「・・・という喫茶店が・・・」を付加している。

U4では「そこ」が参照しているオブジェクトを決定しなければならない。この場合は、親話題から「けいはんなプラザ」であると決定され、またそれが新しい親話題になる。しかし親話題の決定は視点にも影響を与えるため、S4、U5のように利用者に一旦確認を行うこととしている。

(3)断片的な発話、誤りへの対処

利用者の発話自体が断片的であったり、音声認識誤りにより断片しか認識できなかった場合も、上記と同じような方法により、補完や利用者への再確認を行い、対話を継続させる。

U: ここに・・・・(音声認識誤り)
S: それは京阪奈プラザですが、どんなご用件でしょうか?
U: 喫茶店があるかどうかを聞きたいんですが。

ジェスチャの認識誤りについても、同様な処理を行う。

U: これは何ですか? (ジェスチャ認識誤り)
S: どれですか?
U: これです。

3. まとめ

マルチモーダルで入力し、マルチメディアで出力を行う案内システムについて報告した。本システムの基本的なメカニズムは分野非依存である。言語情報、オブジェクト情報、図形情報などを入れ替えることにより容易に他の分野へも移植できる。今後はさらに、モダリティ/メディアの追加、対話インタラクション機能の充実を行うとともに、他の分野への応用についても検討を進める。

参考文献

[1] 竹林: 「音声自由対話システム TOSBURG II」、電子情報通信学会論文誌、VOL.J 77-D II, NO. 8, 1994
[2] 神尾、他: 「マルチモーダル対話システム MultiksDial」、電子情報通信学会論文誌、VOL.J 77-D II, NO. 8, 1994
[3] 長尾: 「マルチモーダルインタフェースとエージェント」、人工知能学会誌、Vol. 11, No.1, 1996
[4] 伊藤、他: 「マルチモーダルインタフェースと協調的応答を備えた観光案内対話システムの評価」、インタラクション '97論文集、1997
[5] 上浦、他: 「3次元空間データベースにおけるデータモデルとアクセス管理機構について」、情報処理学会データベース研究報告、96-DB-109-36、1996
[6] L-Kim、他: 「ATRにおけるマルチモーダル翻訳通信技術」、電子情報通信学会技術報告、SP95-64, 1995