同調的音声対話システムの開発

柏岡秀紀　　翠 輝久　　大竹清敬　　堀 智織　　中村 哲
NICT MASTAR プロジェクト/ATR SLC
{hideki.kashioka, teruhisa.misu, kiyonori.ohtake, chiori.hori, satoshi.nakamura}@{nict.go.jp, atr.jp}

1. はじめに

様々な情報がネットワーク上に溢れ，携帯電話をはじめとするモバイル機器などによるネットワークアクセスが容易になり，様々な環境で利用可能になりつつある．現在，多種多様な情報を提供する端末とのコミュニケーションは，キーボード，タッチパネル等を利用した入力とディスプレイへの情報の表示が，主な入出力を担うインターフェイスである．これらの入出力だけでは，多様な利用者の千差万別の要求を満たすための情報端末とのインターフェイスとしては不十分であり，大きな障害となっている．我々は，いつでもどこでも誰とでもコミュニケーションを行える基盤技術の一つとして，人間の最も自然なインターフェイスの一つである音声を中心とした同調的対話の実現を目指し，京都観光案内を対象に音声対話システムのプロトタイプを構築している．開発中のプロトタイプシステムは，1)音声を中心とした入出力機構，2)対話制御機構，3)多様な情報へのアクセス機構，という主に３つの機構から構成されるシステムと捉えられる．実際に構築しているシステムには，可搬型の機能を重視したシステム（可搬型システム）と音声以外のインターフェイスとの統合を重視したシステム（大画面ディスプレイ対話システム）との 2 つのシステムを構築している．

また，これら対話システムの構築のために，対象となる対話のデータ収録を行っている．収録したデータを利用し，対話の状態遷移を学習することで対話制御を行う機構の構築を試みる．タスクとしては，一日の京都観光の計画立案とし，適切な対話戦略を学習するため，実際に観光業務を行っているプロのガイドと利用者である観光旅行者との対話を複数の環境で収録している．本稿では，我々の目標としている同調的対話について考察を加え，現在構築しているプロトタイプシステムについて，可搬型対話システム，および大画面ディスプレイ対話システムの概要について述べるとともに，収録している京都観光案内対話コーパスについて紹介する．

2. プロトタイプシステム

現在開発しているプロトタイプシステムの概略を図 1 に示す．中の対話制御部と書かれている対話制御機構，その左側の音声を中心とした入出力機構，右側の多様な情報へのアクセス機構から構成されている．プロトタイプシステムは，音声を中心としたインターフェイスによる対話システムとして開発している．基本機能には，画像などの音声以外のインターフェイスによる入出力に対応できるシステム構成となっている．また，京都観光案内をタスクとしていることから，システムの知識として，Wikipedia に含まれる京都に関するページをあらかじめ DB 化し，知識源の一つとして処理できるようにしている．さらに，一般の Web Page の信憑性・信頼性を示す検索機構である WISDOM[1]を用いて Blog や SNS などに書き込まれている評判情報を利用できるようにしている．旅行者の発話から抽出される検索キーワードに関連するキーワードの連想検索も，京都検定に関するテキストを利用し実現している．この検索された連想キーワード列をリストとして表示することにより，普段，意識しないキーワードや，思いもよらぬキーワードを見いだすことができ，対話における話題の広がりを得ることができる．あらかじめ DB 化した情報や評判情報以外の情報に関しては，一般的な WEB 検索を利用し，システムから何らかの情報を利用者に提供できるように設計した．

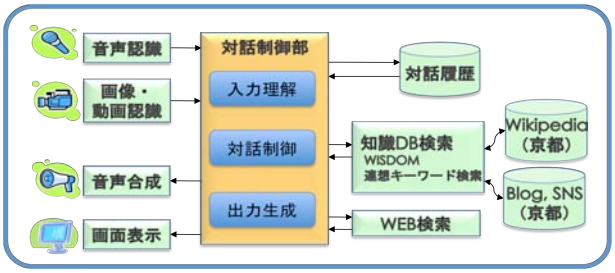

図 1 対話システム概略

システムを開発するにあたり，タスクは，京都観光に関する対話を取り上げた．現在のプロトタイプシステムは，履歴の管理を行いつつ基本的には一問

一答を行うものとなっている．しかしながら，観光計画の立案には，様々な知識処理や多様な対話行為を実現する基本的なシナリオの組み合わせによる状態遷移を制御する必要がある．このような制御機構を実現するため，図 2 に示す対話制御部を現在構築中である[2]．この図に示される“シナリオ”を重み付き有限状態トランスデューサ(Weighted Finite-State Transducer: WFST)により記述し，システムの内部状態と入力信号による状態遷移によって対話を制御する．

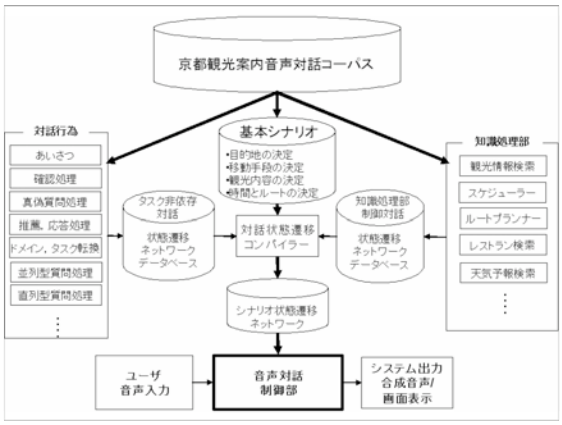

図 2　対話制御機部

実際には，以下の２つのタイプの対話システムを構築している．

1. 可搬型システム
   小型で持ち運びできるシステムであり，主として音声によるインターフェイスを実現している．
2. 大画面ディスプレイ対話システム
   多様な利用者の発信している情報を利用し，音声以外の入出力情報を統合することを目的とした対話システムである．

2.1. 可搬型システム

VAIO-U を利用したシステムであり(図 3)，システムへの入力は音声を主としたもので，ディスプレイへの出力および音声による応答により対話を実現する．音声入力は，Push-To-Talk のスタイルを利用している．システムからの情報提示に利用している画面情報は，画面左側に上から，利用者の発話に含まれていた検索キーワードおよびシステムの発話内容，検索キーワードから連想される連想キーワード列，評判情報がそれぞれ表示される．画面右側では，検索キーワードによって検索された WEB-Page を表示している．対話の履歴を管理することで，一部，検索の絞り込みをおこなっている．発話に含まれる検索キーワードに，事前に準備された DB の主要キーワードが含まれていなければ，それまでの主要キーワードを引き継ぎ，検索を行うようにしている．より適切な検索キーワードの管理を行うには，基本シナリオに従ったキーワード等の対話状況の管理が必要である．現在開発中の対話制御機構においては，WFST の状態に応じたキーワードの管理が実現できると考えている．

図 3　可搬型システム

図 4 大画面ディスプレイ対話システム

2.2. 大画面ディスプレイ対話システム

大画面ディスプレイ対話システムでは，52 インチ大型ディスプレイを使用し，ディスプレイの左右および下部に計 3 台のカメラを設置し，正面に立った利用者を捉え顔や視線の方向を推定し，その情報を対話制御に利用している(図 4)．また，利用者に画面上の注目すべき場所を明示しシステムとの対話をより円滑に進めることを目的としてガイドエージェントを表示し，お辞儀などのいくつかの動作をさせている．顔向き・視線の情報により，利用者が表示

画面のどの部分に興味を持っているかを判断し，一定時間以上注視しているようであればシステムから注視している画面に表示されている情報の詳細な情報を画面に表示し，詳細な説明を必要とするかを音声により問いかけるようにしている．音声入力には，ディスプレイ上部の指向性マイクを利用している．

3. 同調的対話

対話システムには，テキストでチャットするような対話システムや，音声入出力による音声対話システムが考えられる．音声対話システムでは，入力された音声発話を理解し，システムが適切に応答することが期待される[3]．しかし，現実の対話では，音声のみの入力で円滑に自然な対話が行われているのではない．様々な情報に反応することで，各話者が相互に理解しあい対話が成立する．ここでは，各話者が積極的に対話しようとする状況での対話を，特に同調的対話と呼ぶ．このような状況では，より円滑な対話が成り立つと考えられる．ここで考えられる同調性は，多様な側面を持っている．相槌や身振りなどによる対話の自然さもその一面として捉えることができる[4,5]．このような同調性をいくつかの側面で分けると，以下のようなものが考えられる．

- 動作タイミングによる同調性
  「相槌」や「うなずき」，発話のオーバラップなどに見られる同調性
- 表層表現による同調性
  相手に応じて表現（言葉）をかえることや表現のくだけ方がかわるなどの同調性
- 対話制御(戦略)による同調性
  提示する情報の内容と順序などによる同調性
- 信念共有による同調性
  共有知識，信頼性などによる同調性

これらの同調性は，互いに独立なものではなく，相互に複雑に関連している．また，いずれかの同調性が満たされたからといって，対話全体が円滑に自然になるものでもない．個々の特徴的な同調性を特定の状況下で実現し，統合的な処理機構を対話制御として行うことで，人同士の同調的な対話の一部をシステムとの対話で模倣することが可能と考える．

4. コーパス収集

京都観光案内対話コーパスは，京都観光案内のエキスパートガイド 3 名(男性 1 名，女性 2 名)が模擬旅行者に対して京都市内一日観光の計画立案を行う 2 者による対話である[6]．現在までに，対面での対話，非対面での対話，Wizard of Oz(WOZ)形式での対話を収録している．1 対話は約 30 分である．

対面での対話では，ガイドが，ガイド自身の持つ知識，準備されているガイドブック，地図，WEB 上の情報を利用し，旅行者に対して情報を提供，一日の旅程を作成していったものである．ガイドは，ヘッドセットマイクを使用し音声を収録していた．旅行者は，スタンドマイクあるいはヘッドセットマイクのいずれかを用いて音声を収録している．旅行者は，20 歳代から 50 歳代の 114 名(男性 57 名，女性 57 名)を対象として収録した．

非対面での対話では，ガイドと旅行者の間での情報の授受は，音声およびディスプレイ上の表示に限定されている．音声は，ガイド，旅行者ともに，ヘッドセットマイクを使用して収録した．非対面の対話では，対面対話の収録に参加したガイド 1 名(女性)に限定して収録した．旅行者は，20 名を対象としており，収録時間は約 10 時間となる．

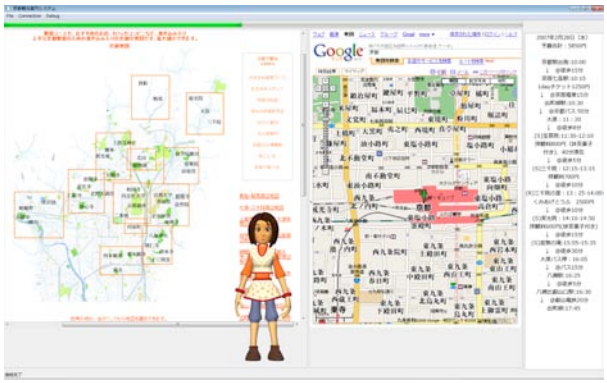

図 5　WOZ 収録システム画面

WOZ 形式での対話では，旅行者に情報を提示するディスプレイ上に，ガイドエージェントを表示した場合と表示しない場合の対話を収録している．ガイドエージェントを表示した画面のサンプルを図 5 に示す．旅行者は，表示／非表示のときともに 20 名を対象としている．非対面同様，収録時間は，表示／非表示，各々約 10 時間となる．WOZ 形式での収録では，システム側に，ガイド，およびタイピストを用いて収録した．システムの発話は，一旦ガイドが発話したものをタイピストがテキストとして入力し，その入力テキストを音声合成したものである．定型的な応答に関しては，ショートカットキーを定義し，比較的早く応答できるように工夫している．しかし，ガイドが直接応答している対話に比べ，システムの応答までに時間がかかっている．

現在，これらデータの書き起しデータに対して，形態的なタグを始め，対話行為に対するタグ付けを行っており，対話コーパスからの WFST の学習に利用する予定である．

5. 談話タグ付与

収録したデータに対して，WFST の構築を行えるように発話行為タグ，意味内容タグの府を行っている．タグを付与する単位は，人手により認定された単位ではなく，機械的に認定可能な単位である必要がある．そこで，話し言葉の節単位認定を目的に開発された CBAP[7] を用いて発話の節単位を認定し，タグ付けの単位とする．発話行為タグの設計には，二つの方針がある．一つは，タグ付け単位に対して最も適切なタグを一つだけ付与するアプローチであり，談話タグワーキンググループの発話内行為タグ[8] や AMI コーパスの DA タグなどがこれに該当する．もう一つは，タグ付け単位に対して該当する全てのタグを付与するアプローチであり，DAMSL や，MRDA[9] などがこれに該当する．本研究では，対話の同調性の一つの側面は発話の多機能性にあると考え，発話の情報をできるだけ多く保持しておくために後者のアプローチを採用する．意味内容タグに対しては，意味を記述しようとするために，Lexical Functional Grammar (LFG)[10] や HPSG[11]の枠組みによるアプローチも考えられる．しかしながら，発話が本来断片的であることを鑑みると，より頑健な処理を実現するためには，断片的な表現からその意味クラスを推定できるような枠組みが望ましいと考えた．そのため，特定の意味処理の枠組みを前提にせず，文節単位の依存構造(データ）へ直接意味クラスを付与するアプローチを取る．意味内容タグのための意味クラスには階層構造を持たせた．階層構造の最上位には，約 40 のクラスが存在する．詳細については［12］を参照されたい．

6. まとめ

本稿では，現在我々がプロトタイプとして開発・構築している可搬型システム，およびマルチモーダルシステムについて述べた．これらのプロトタイプシステムは，個別に開発しているのではなく，開発している対話制御機構が，システムの利用できる多様な入力情報を統一的に適切に処理できることを示すとともに，多様なシステムとしての動作を確認し，様々な入出力情報を利用できるような実験環境を整えることを目的として開発している．

また，対話システムを開発するために収録している京都観光案内対話コーパスの概要について述べた．対面対話が約 50 時間，非対面対話が約 10 時間，WOZ 形式の対話が 20 時間(ガイドエージェント有:約 10 時間, 無:約 10 時間)のコーパスとなっている．さらに，コーパスに付与している対話行為タグ，意味内容タグについて説明を加えた．今後は，収録コーパスを利用した対話制御機構の開発に置いて，頑健な対話の実現，および同調的対話を実現するための機構の研究開発を行う予定である．

謝辞

本稿で紹介しているプロトタイプシステムでは京都大学河原研究室の研究成果を，マルチモーダルシステムでは京都大学松山研究室の研究成果を移転・活用しています．また，ガイドエージェントでは，情報通信研究機構 ユニバーサルシティグループの研究成果を移転・活用しています．さらに，評判情報の検索(WISDOM)，連想キーワードの処理は情報通信研究機構 知識処理グループの研究成果を移転・活用しています

参考文献

[1] T. Nakanishi, K. Zettsu, Y. Kidawara, Y. Kiyoki: “Towards Interconnective Knowledge Sharing and Provision for Disaster Information Systems –Approaching to Sidoarjo Mudflow Disaster in Indonesia-,” ICTS2007, pp.332-339, 2007.

[2] 堀，大竹，柏岡,中村. “京都観光案内対話コーパスにおける対話行為に関する研究”，日本音響学会講演論文集, pp.105–106, 2008.

[3] 河原 and 荒木. “音声対話システム”，オーム社, 2006.

[4] Kawashima and Matsuyama. “Interval-based hybrid dynamical system for modeling multimedia timing structures,” In First International Symposium on Universal Communication Proceedings, pp.67–70, 2007.

[5] Kitaoka. “Liveliness of spoken dialog systems – considering response timing and prosodic synchrony”, In First International Symposium on Universal Communication Proceedings, pp.63–66, 2007.

[6] 大竹，堀，柏岡，中村. “京都観光案内対話コーパスにおける対話行為の分析”，言語処理学会第 14 回年次大会発表論文集, pp.159–162, 2008.

[7] 丸山岳彦，柏岡秀紀，熊野正，田中英輝. “日本語節境界検出プログラム CBAP の開発と評価”，自然言語処理, Vol. 11, No. 3, pp. 39–68, 2004.

[8] 荒木雅弘，伊藤敏彦，熊谷智子，石崎雅人. 発話単位タグ標準化案の作成. 人工知能学会誌, Vol. 14, No. 2, pp. 251–260, 1999.

[9] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In Proc. 5th SIGdial Workshop on Discourse and Dialogue, pp.97–100, 2004.

[10] Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, editors. “Formal Issues in Lexical-Functional Grammar”, CSLI Publications, 1994.

[11] Carl Pollard and Ivan A. Sag. “Head-Driven Phrase Structure Grammar”, The University of Chicago Press, 1994.

[12] 大竹清敬　翠輝久　堀智織　柏岡秀紀　中村哲, “統計的手法による対話管理のための発話行為と意味内容タグ”，言語処理学会第１５回年次大会，2009.