2004－ＦＩ－74　(4)
2004－ＤＤ－43　(4)
2004／3／26

質問応答技術に基づくマルチモーダルヘルプシステム

浦田 耕二　福井 美佳　藤井 寛子　鈴木 優　酒井 哲也　齋藤 佳美　市村 由美　佐々木 寛

(株) 東芝研究開発センター 知識メディアラボラトリー

ユーザからの質問に対し、映像・音声・取扱説明（テキスト）などで構成される表現力豊かなマルチモーダルコンテンツの検索技術、および、質問内容を理解しユーザが必要としている情報に対して的確に回答する質問応答技術を融合することにより、よりわかりやすい情報提供を実現した質問応答型マルチモーダルヘルプを開発した。このシステムを用いて、オーブンレンジ、デジタルカメラの取扱説明書データ(テキスト 160 ページ、映像 108 分)を登録し、質問データ 123 件による実験を行い、次のような知見が得られた。(1)質問応答技術により取扱説明情報の探索作業が軽減される見込みを得た。(2)映像、音声、取扱説明書の該当ページの表示を併用することにより，取扱説明に関するわかりやすさが向上すること確認した。映像収集、編集について、作業の軽減と質の向上を支援する必要がある。

A Multimodal Help System based on Question Answering Technology

Koji Urata　Mika Fukui　Hiroko Fujii　Masaru Suzuki　Tetsuya Sakai　Yoshimi Saito
Yumi Ichimura　Hiroshi Sasaki

Knowledge Media Laboratory, Corporate R&D Center, TOSHIBA Corp.

We have developed a user-friendly help system by integrating multimodal content retrieval technology and question answering technology. Multimodal content retrieval enables the user to access contents with a rich power of expression such as those comprising video, speech and textual instructions, while question answering enables pinpoint access to the required information. We conducted a preliminary experiment using the manuals of a microwave oven and a digital camera (160 pages of text and 108 minutes of video) as the knowledge source, with 123 questions. Our findings are: (1) Question answering technology enables efficient access to the desired instructions; and (2) Responses in video/audio accompanied by presentation of a relevant manual page helps the user understand the instructions better. However, we need a mechanism for facilitating gathering/editing of video contents and for improving their quality.

1. はじめに

近年、家電や AV 機器の高機能化やネットワーク化が進み、操作が複雑になってきている。また、多機能化が進み、ユーザがすべての機能を使いこなすのが難しくなっている。製品には必要十分な内容の取扱説明書が付属しているが、コールセンターへの問い合わせ事例をみると、取扱説明書に記述されている内容に関する問い合わせも少なくない。今後さらに取扱説明書のデータ量が増えるにつれ、知りたい情報を探せない、操作が複雑で取扱説明書を読んでもわからない、といった問題が増えると考えられる。

筆者らは、映像・音声・取扱説明（テキスト）などで構成されるマルチモーダルナレッジ（MMナレッジ）による表現力豊かなコンテンツの蓄積、検索技術[1]と、質問内容を理解しユーザが必要としている情報に対して的確に回答する質問応答技術[2]を融合し、ユーザに対してよりわかりやすい情報提供を可能とする質問応答型マルチモーダルヘルプを開発した。本システムは、音声により入力された自然言語の意図を理解し、適切なメディアで情報を提供するという特徴をもつ(図１)。

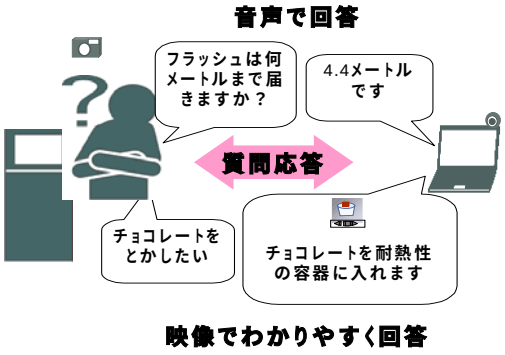

図 1　質問応答型マルチモーダルヘルプ

従来、ユーザが取扱説明書から必要な情報を調べる場合には、まずキーワードを決め、目次、検索ページより該当ページを調べる。次にページの中から該当する記述を探し出すという作業が必要となる。本システムを利用することにより、質問を音声で入力するだけで、該当ページの探索やページ中の該当記述の探索作業を行うことなく、目的の情報を探すことができる。「何メートル」「どこを押すの」といった数量や操作部の名前などに関する質問に対しては、具体的な数値やボタン名などの回答を音声で出力する。また、機器の操作方法に関する質問に対しては、映像や音声を用いて取扱説明書だけでは表現できない細かな操作手順、装置の持ち方、装置とユーザの位置関係など、わかりやすく出力する。そのため、ユーザは取扱説明書を直接開くことなく、基本操作のみならず複雑な操作方法においても、容易に的確な回答を得ることができる。さらにインターネットへ接続することにより、意識することなくメーカー側へ問い合わせることにより、常に最新の情報を取得することが可能になる。

関連研究として、映像の構造化、検索に関しては、ニュース映像を対象に顔認識などによりシーンを検出し、音声認識、字幕認識により映像にメタデータを付加して保存し、自然言語で検索するシステムの研究が行われている[3]。また、マルチモーダルインタフェースに関する研究もさまざま行われており、例えば、自然言語による音声対話技術とペン入力、無線 LAN による位置情報検知を利用したナビゲーションシステムが開発されている[4]。質問応答技術については、新聞記事やオンラインヘルプを対象として、自然言語による質問への回答や、曖昧な質問に対する問い返しなどの研究が行われている[5]。また、NTCIR ワークショップの QAC タスクにおける研究が注目を集めている[6]。

本システムは、映像を含むマルチモーダルコンテンツの構造化技術と質問応答技術の融合により、質問に対して適切な回答を提示しうるメディアを選択することにより、探しやすくわかりやすいヘルプシステムを実現するものである。また、タッチパネル、音声認識、音声合成を併用することにより、一般ユーザが家庭で電化製品を操作しながら取扱説明を調べる利用スタイルに適したマルチモーダルインタフェースを提供する。

本稿では、質問応答技術に基づくマルチモーダルヘルプシステムの開発について報告する。2 章では、質問に対して適切な回答メディアを選択し検索する方法について述べる。 3 章では実験システムの構築について述べ、 4 章では評価実験の結果について述べる。最後に考察とまとめを行う。

2.　適切な回答メディアの選択

本システムは、入力された質問文を解析し、ユーザが必要としている回答にあった出力形態を選択する特徴をもつ。利用した検索技術は、MM ナレッジ検索(2.1 節参照)と質問応答検索(2.2 節参照)である。質問文を解析して適切な回答タイプを判定し、回答タイプにあった検索を行う。機器の取扱に関する質問例を収集し、質問の形態により分類した（表 1)。

表 1　質問形態の種類

1．方法	**一連の操作方法** **～の方法｜やり方｜して欲しい｜するには｜どうやる｜できない**
2．数量	**数字を聞く（時間、量など）** **いつ｜どのくらい｜時間は｜量は**
3．名前	**操作部の名前、場所を聞く** **どこを｜どの｜なにを**
4．機能	**機能の説明、用語の定義など** **って何ですか｜違いはなんですか**
5．状況	**機器の状態を説明し暗に指示を仰ぐ** **～なんですが｜だけど｜してしまう｜なぜ｜しない**
6．YES/NO	**仕様、操作についての YES/NO** **するの｜いいの｜ですね｜ですか｜いいんですか｜使えるの**
7．確認	**操作中に操作が正しいか確認** **こんなかんじでよろしいですか？｜これでＯＫなんですか**

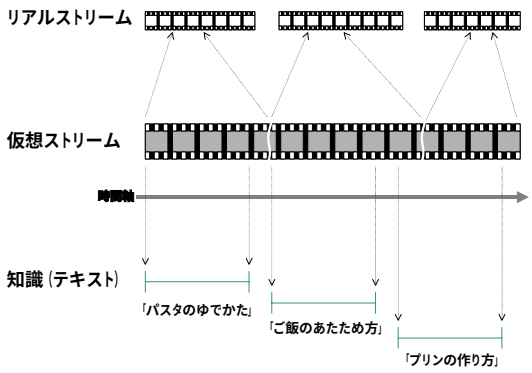

図 2　MMナレッジ概念図

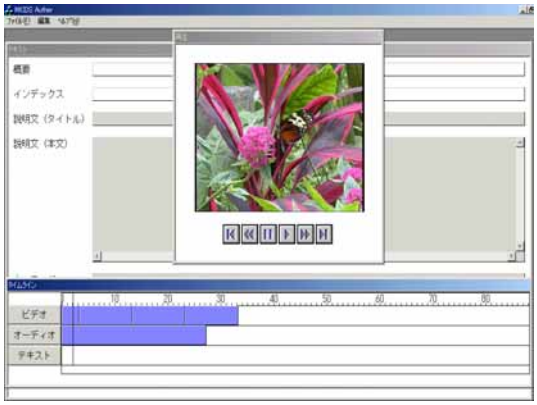

図 3　オーサリングツール

このうち、1.方法に関する質問は、ユーザに対して映像、画像、音声を利用してわかりやすく出力するため、MM ナレッジ検索を行う。2.数量、3.名前、4.機能についての質問は、具体的な回答を出力するため、質問応答検索を行う。5.状況、6.YES/NO、7.確認、および質問形態が不明のものは、現時点で的確な回答を提示するのが困難なため、MM ナレッジ検索により、関連する情報の提示を行う。

例えば、「スパゲッティをゆでる方法」という質問の場合、質問形態「方法」回答タイプ「方法」となり、MM ナレッジ検索に送られ映像データが出力される。

「AC アダプタの重さ」という質問の場合、質問形態「数量」回答タイプ「重さ」となり、質問応答検索で最も確信度が高い回答として「約１５０ｇ」という結果が返される。

「スパゲッティをゆでたい」という質問の場合は、

質問形態「不明」のため、MM ナレッジ検索に送られ映像データが出力される。

2.1.　MMナレッジ検索

ユーザに対して映像、画像、音声を利用して結果を提示するMMナレッジ検索は、知識情報共有システム(KIDS)[7]の知識処理エンジンをマルチモーダルに対応させたものである。複数の映像、音声、画像のファイル（リアルストリーム）から MPEG7[8]により必要部分をつなぎ合わせた仮想ストリームを構成し、その時間軸上に関連づけられたテキスト情報を検索対象とする(図 2)。

取扱説明に関する MM ナレッジデータとしては、

(1)機器の操作方法に関する動画、音声、テキストを組み合わせたデータ

(2)取扱説明書のテキスト及び画面イメージからなるデータ

の 2 種類を想定する。いずれのデータについてもデータ形式として MPEG７を用いて統一的に記述する。

(1)については、取扱説明書に記載されている主な操作項目について映像の撮影を行って作成する。映像の編集及びテキスト情報との関連付けには専用のオーサリングツール(図 3)を利用する。MPEG７への変換はこのオーサリングツールによって自動的に行われる。

(2)については、PDF 形式の取扱説明書から変換した JPEG 画像、及び PDF ファイルから直接抽出したテキストをページ単位で関連付け、スクリプト処理により、ほぼ自動的に MPEG７に変換できる。

2.2.　質問応答検索

質問応答検索で回答として提示するデータは、PDF 形式の取扱説明書などから抽出したテキストより自動抽出する。意味情報を付与した辞書を用いて形態素解析を行い、意味情報と品詞のパタンによって記述した判定ルールにより、数量、操作部の名前、時間表現などの意味クラスを付与した情報を抽出する[9]。回答データにページ情報を付加するためテキストデータをページ単位で区切り、そのページから抽出された回答データを関連づけておく。

入力された質問文に対しても、同様に意味クラス解析を行ったあと、意味情報と品詞のパタンによって記述した質問形態判定ルールと回答タイプ判定ルールを用いて解析を行う。質問文と類似度の高い取扱

説明書のページに付与された回答データのうち、質問文の回答タイプと同じ意味クラス情報を持ち、ヒットワードとの距離が近いものを回答として提示する[2]。回答タイプの体系は意味クラスの体系と同じものである。

3. 実験システムの試作

3.1. システム構成

図 4 にシステム全体の構成を示す。

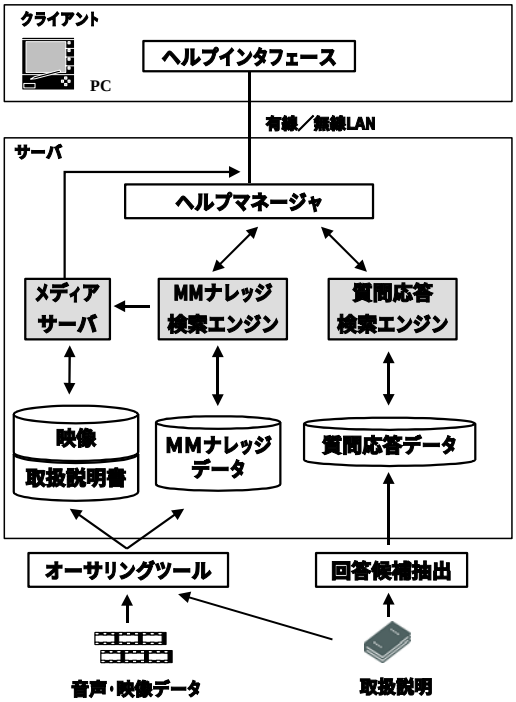

図 4 全体構成

家庭での利用を想定し、家庭内やコールセンターなどに設置したサーバに対して、ユーザの手元のクライアント端末で情報を検索する構成とした。ユーザはキッチンやリビングなど、家電や AV 機器を操作しながら、必要に応じてクライアント端末を利用する。クライアント／サーバ間の通信は無線 LAN を利用するため、クライアントは移動し利用できる。

3.2. クライアント

クライアントにタブレット PC を利用することにより、基本操作を音声入力、ペンで行うことができる。また、環境によりキーボードを利用することもできる(図 5)。

図 5 システム利用イメージ

音声入出力の処理に東芝 LaLaVoice2001[10]の連続音声認識、音声合成機能を使用した。音声認識の精度向上のため、ヘッドセットマイクロフォンを利用し、次の方法により音声認識辞書を強化した。

1. 取扱説明書のテキストから音声認識辞書を自動登録
2. 認識誤りが多い語彙について辞書登録
   また、ユーザの性別により男女の辞書を入れ替えて利用する。

ヘルプインタフェースとして、Microsoft Internet Explorer を利用し、映像と音声は Windows Media Player で再生する。

3.3. サーバ

メディアサーバにMicrosoft Windows Server 2003 Standard Edition の一機能であるWindows Media サービス を利用し、MMナレッジ検索エンジンより指定された映像（Windows Media Video形式）をクライアント上に構成されたヘルプインタフェース上のWindows Media Playerへ配信する。

ヘルプマネージャは質問文の質問形態を解析後、2 章で説明した回答タイプの推定を行い、検索エンジンの振り分けを行う。MMナレッジ検索エンジンもしくは質問応答検索エンジンに対して質問文及び質問形態を送り検索処理命令を出す。

MMナレッジ検索エンジンは、MMナレッジデータに対して検索を行う。結果をヘルプマネージャ及び、メディアサーバへ送る。ヘルプマネージャは画面構成情報をヘルプインタフェースに送る。メディアサーバはMMナレッジ検索エンジンより指定された映像データをヘルプインタフェースへ送る。

質問応答エンジンは、MM ナレッジ検索時と同様にヘルプマネージャより質問文、質問形態を受け取り、回答候補を推定し、質問応答データに対して検索を

行う。検索結果の回答候補と取扱説明書のページ番号はヘルプマネージャに送られる。ヘルプマネージャは画面構成情報をヘルプインタフェースに送る。

3.4. コンテンツ

コンテンツはオーブンレンジ、デジタルカメラの取扱説明書を利用した（表 2）。代表的な操作手順については取扱説明書の中から抜き出し、映像の撮影を行った。映像データは操作方法や機能により分割している。

なお、取扱説明書のテキスト中、用語定義とボタン・絵記号については手動で意味クラス情報を付与し、回答データとして登録した。用語定義などでは表の解析が必要になり、ボタン・絵記号については画像と名称の対応づけが必要になるためである。

映像の作成、オーサリング期間は、準備も含め約３週間必要となった。

表 2　コンテンツの種類

	取扱説明書	映像
オーブンレンジ	48 ページ	31 分(14 データ)
デジタルカメラ	120 ページ	77 分(31 データ)

3.5. 動作例

図 6 に MM ナレッジ検索の画面を示す。

ユーザは音声により質問フィールドにテキストを入力する。質問文の入力終了後に検索ボタンを押すことにより、ヘルプインタフェースはヘルプマネージャに対して質問文を送る。ヘルプインタフェースでは音声により入力された「フラッシュを使いたい」という質問文、「フラッシュの設定方法」が録画された映像データ、MPEG7 で記述された映像に対してのナレーションを出力する。

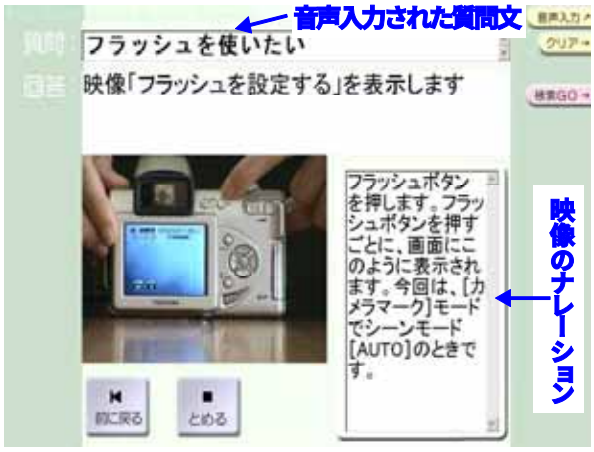

図 6　クライアント画面（映像）

図 7 に質問応答検索を行った画面を示す。

音声により入力された質問文「一回でゆでられるパスタは何グラム」に対して質問応答検索処理が行われ、ヘルプマネージャより画面構成がヘルプインタフェースに送られる。検索結果として、回答の「１００ｇ」の表示、音声合成による読み上げ、さらに取扱説明書のページを画像として出力する。

なお、図６と図７では割愛しているが、画面の下方には、第２位と第３位の検索結果を表示する。

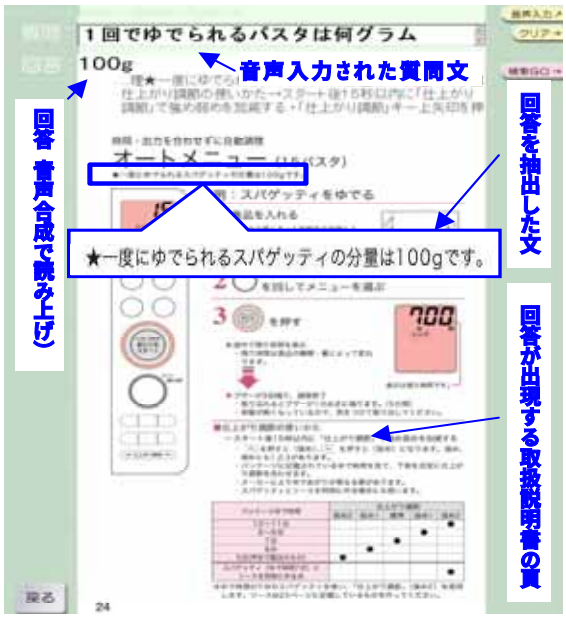

図 7　クライアント画面（質問応答）

4. 評価実験

取扱説明の探索作業が軽減されるか検証するために、本システムの検索精度の評価を行った。
質問データは 123 文準備した。内訳としてMMナレッジ検索を利用して映像、画像、音声での出力を必要とする質問 50 文、質問応答検索を利用して数量、名前、機能についての具体的な回答を必要とする質問 73 文である。この質問文を学習データとして、意味クラス判定ルール、質問形態判定ルール、回答タイプ判定ルール、音声認識辞書の拡充を行った。質問応答に関する判定ルールは、もともと新聞記事用に作られたものであったため、操作部の名前、付属品など取扱説明書に特有の表現について、意味クラスと判定ルールを新たに追加した。

実験システムで前述の既知の質問データに対して評価を行ったところ、3 位以上の結果に対しMRR[1]が 0.65 となった。

5. 考察

既知の質問 123 文に関しては一定の精度が得られ、取扱説明情報の探索作業が軽減される見込みを得た。未知の質問に対して検索精度を上げるために、さらに質問文を収集しルールを拡充していくことも必要である。また、回答メディアの選択に関して質問形態を７つに分類し、そのうち４つに関して回答タイプを切り替えたが、適切な分類であったかの調査と、残り３分類について適切な回答タイプとその実現方法の検討が必要である。

次に、出力メディアとして映像を用いることで、取扱説明書に記載された模式図のかわりに、実機を実際の人間が操作している動きを確認できるようになり、特にデジタルカメラのような細かい操作部を持つ機器の操作やメンテナンス作業の理解を助けることがわかった。また、質問応答検索の結果に関しても、音声での具体的な返答と同時に、取扱説明書の該当ページを表示することでユーザは聴覚・視覚の両方で確認することが可能となり理解度が向上した。今後の問題点として、MMナレッジコンテンツを広く活用するには、映像収集、編集についてのコストを軽減する必要がある。取扱説明書の主要操作についてすべて映像コンテンツを用意するのではなく、評価実験により映像ヘルプが有効な操作を明らかにし、映像化する事項を厳選することを考えている。また、映像の撮影・編集支援ツールにより映像の質の向上とコストダウンを図る。

本システムでは、タブレット PC による音声での質問入力、ペンを利用した入力によってユーザの操作の負担を軽減した。さらに画面を縦型として利用することで、取扱説明書と同様の表示サイズとなり可読性が向上した。問題点として、持ち歩きにくいことが挙げられる。一台の端末を家庭内で持ち運んで使う場合、質問を行う端末の重さ、大きさが重要になる。一方、オーブンレンジの横などに位置を固定して利用する場合、クライアントの重さは問題にならないが、設置スペースが確保できるかどうかが問題になる。また、デジタルカメラや AV 機器のような複雑な表示部を持つ機器の説明は、大きく表示するべきだと考えられる。このように家電、AV 機器にあった端末の大きさや重さ、出力方法について調査する必要がある。

6. まとめ

質問応答技術に基づくマルチモーダルヘルプシステムを開発し、オーブンレンジ、デジタルカメラの取扱説明データに対して質問データ 123 件による実験を行った。既知の質問に関しては一定の精度が得られ、取扱説明情報の探索作業が軽減される見込みを得た。また、映像、音声、取扱説明書の該当ページの表示を併用することにより、取扱説明に関するわかりやすさが向上することを確認した。

今後、実際のユーザによる(1)操作性、(2)出力メディアの適性、有効性、(3)精度、表示速度についての評価を行い、誰もが IT 家電や AV 機器を使いこなせるマルチモーダルヘルプシステムの実用化をめざす。また、家電以外のヘルプやサポートセンター、教育分野への適用も検討していく。

7. 参考文献

[1] 鈴木他: マルチモーダルナレッジ技術の展示案内システムへの適用,人工知能学会誌, Vol.18,No.2 (2003)

[2] Sakai, T. et al.: ASKMi: A Japanese Question Answering System based on Semantic Role Analysis , RIAO 2004

[3] Christel, M. et al.: Collages as Dynamic Summaries for News Video (CMU) Multimedia2002

[1] MRR(Mean Reciprocal Rank)正解が最初に出現した順位の逆数を得点としたもので、全質問にわたって平均したもの[6]

[4] Johnston, M. et al.: Matchless Multimodal Info Access
http://www.research.att.com/news/2001/October/MultimodalAccessToCityHelp.html
[5] 西田・黒橋研究室. ダイアログナビ (Dialog Navigator). (2002)
http://www.kc.t.u-tokyo.ac.jp/msnavi/
[6] Fukumoto, J. et al.: Question Answering Challenge (QAC-1): An Evalution of Question Answering Tasks at the NTCIR Workshop 3, AAAI Spring Symposium: New Directions in Question Answering pp.122-33 (2003)
[7] 中山他: 知識情報共有システム(KIDS)の開発と実践－組織におけるノウハウ共有の促進－, 人工知能学会誌, Vol.16, No.1, pp.64-68 (2001)
[8] MPEG7 Japan 情報規格調査会
SC29/WG11/MPEG-7 小委員会編
http://www.itscj.ipsj.or.jp/mpeg7/
[9] 市村他: 質問応答と, 日本語固有表現抽出および固有表現体系の関係についての考察,情報処理学会自然言語処理研究会研究報告 NL-161, [報告予定] (2004)
[10] LaLaVoice2001,(株)東芝
http://www3.toshiba.co.jp/pc/lalavoice/index_j.htm