LE-009

ロボットとの対話によるマルチメディアブログ創作と評価

奥村 明俊† 池田 崇博† 西沢 俊広† 安藤 真一† 安達 史博†
Akitoshi Okumura Takahiro Ikeda Toshihiro Nishizawa Shin-ichi Ando Fumihiro Adachi

1. はじめに

ブログは，関連市場も含めると約1377億円に拡大し，事業者から様々なサービスが提供されている[1]．最近では，テキストだけでなく，映像や音声，動画などのマルチメディアコンテンツを取り込んだブログが急増している．ブログのマルチメディア化には，1)録音した音声や録画した動画に見出しとなるキーワードをつけるもの，2)テキストのブログに対して関連する映像や音声，イラストなどのコンテンツを盛り込むものがある．マルチメディアブログは，臨場感豊富にメッセージを伝えることができるが，WEB やコンテンツ検索などに慣れていないユーザにとって，魅力的なものを作成することは容易ではない．手馴れたユーザにとっても，1)の場合，携帯電話などを用いてメッセージを入力した後，他のユーザから検索・閲覧されるための見出しとなるキーワードの付与に手間がかかる．2)の場合も，メッセージに関連するコンテンツの検索，編集・コーディネートに手間がかかり，作成コストに課題がある．さらに，せっかく作成したブログであっても他者からのコメントやトラックバックがないと持続しにくい．最近，自動的にブログにコメントやトラックバックするエージェント的なペットが登場しているが，ユーザの気分を汲み取ったコメントにはなっていない．これらの課題を解決するため，ユーザがロボットに話しかけ対話することで，ロボットがユーザの発話内容を解析し，関連するコンテンツを探してロボットからのコメントとともにマルチメディアブログを創作するシステムを提案する[2]．このシステムは，ノート PC と同程度のハードをベースに開発された対話型ロボット PaPeRo(パペロ)[3][4]上にプロトタイプとして構築される．システムは，作成コストの軽減や作成意欲の増進といった作成者の観点と，作成されたブログの魅力といった閲覧者の観点から評価する．作成者に対して使い勝手や出来栄えなどのヒアリング調査を行い，閲覧者に対して作成されたブログの感想などをアンケートで調査する．本稿では，まずマルチメディアブログ創作手法について述べ，次に全体システム構成を説明し，システム動作例と評価結果を示す．

2. マルチメディアブログ創作

2.1.創作処理の概要

本システムは，ロボットとの対話によって入力されたビデオメッセージから，見出し用キーワードやメッセージの内容に関連するコンテンツ(イラスト，音楽など)と，ロボットからのコメントを含むマルチメディアブログを創作する．提案システムの処理の流れを図1に示す．システムは，入力されたビデオメッセージを蓄積し，発話音声を抽出し，映像発話音声認識によって発話テキストに変換して見出しとなるキーワードを抽出する．次に，自然言語文検索により発話テキストと関連するマルチメディアコンテンツを，予め指定された WEB やディレクトリから検索する．これらの機能によってユーザのマルチメディアブログ作成コストを軽減する．さらに，発話テキストの表現やモダリティからユーザの心的状態を推定し，それにあわせたコメントを生成する．そして，入力ビデオメッセージ，見出し用キーワード，検索されたマルチメディアコンテンツおよびコメントをコーディネートしてブログを創出する．

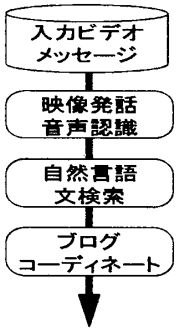


図1:処理の流れ

2.2.映像発話音声認識

ブログ映像の中では，さまざまな単語や表現を含む話し言葉を扱う大語彙連続音声認識が必要である．大語彙連続音声認識は，一般に多くの処理能力やメモリ等のリソースを必要とするが，対話や動作など様々な処理を行う小型ロボットにおいて話し言葉を認識するためには，省メモリでコンパクトに動作する大語彙連続音声認識が必要である．そこで，PDA 程度の端末でも動作する，コンパクトでかつスケーラブルな大語彙連続音声認識フレームワークを用いる[5][7]．本フレームワークは，以下の手法によりコンパクトかつ高速な処理を実現している．

(1) 音響モデル・距離計算

記述長最小基準に基づく効率的な分布数削減[8]，対角共分散行列の共有化による分布の簡易化，分布の木構造化に基づく出力確率の高速計算[9]によって，音響モデルの使用メモリ量と距離計算の計算量を削減する．

(2) 単語辞書・言語モデル

言語モデルとしては，単語の n 個組の連鎖確率である n-gram モデルを用いている．利用可能なメモリ量や処理能力に応じて単語 2-gram，クラス 2-gram，単語 3-gram を組み合わせている．クラスは品詞をベースに，対象分野に応じて意味的なクラスや自動クラスタリングにより細分化して用いている．

(3) 最適単語列探索

最適単語列探索は距離計算結果と言語モデルを用いて，入力音声に同期して可能性の高い候補のみに絞り込みながら，辞書中の単語との照合を行う．辞書は先頭の共通部分を束ねることで木構造化して圧縮した表現で保持し，動的に展開して用いる．一定間隔ごとのワークメモリのガベージコレクションや，言語モデルの計算結果の再利用などにより，メモリ量，計算量を削減している．

今回，旅行日記作成をタスクとして，旅行会話向け日英音声翻訳システム[5]の日本語認識エンジンをベースに約5万語の大語彙連続音声認識を構築し，発話音声をテキストに変換する．

† NEC 共通基盤ソフトウェア研究所，
Common Platform Software Res. Labs., NEC Corporation

2.3. 自然言語文検索

自然言語文検索は，ユーザによる発話の音声認識結果を検索要求文としてマルチメディアコンテンツの検索を行う．一般に，マルチメディアコンテンツには，十分なインデクスが与えられていることは少ない．そのため，コンテンツが含まれている WEB ページやドキュメントのテキストを手がかりとして検索する必要がある．本システムでは，検索要求文中の自立語をキーワードとしてテキストの検索を行い，その上位結果の近傍に含まれるマルチメディアコンテンツをブログに使用する．

また，検索結果に関するユーザとのインタラクションを無しにブログを創出するためには，自然言語文検索に高い適合率が要求される．本システムでは，Okapi BM25 式[10]による検索モデルをベースとして，適合率を高めるために以下の 3 つの拡張を行い，上位候補からより適切なコンテンツを抽出する[6]．

（1） 係り受け関係にある単語ペアの利用

例えば，同じ「魚」という単語でも，レストランで魚を食べた場合と，水族館で魚を見た場合とでは表示すべきイラストは異なる．そこで，予め単語間の係り受け関係を求めておき，係り受け関係にある 2 つの単語を組にした単語ペアを重みの計算に利用する．具体的には，検索要求文中に含まれる単語ペアがテキスト中に含まれる数に応じてテキストの重みを加点する．

（2） 否定表現と肯定表現の区別

「楽しかった」という表現と「楽しくなかった」という表現のように，否定表現と肯定表現とでは表す意味が正反対になるため，両者を区別して扱わなければならない．そこで，各単語が肯定的に使用されているか否定的に使用されているかを，その単語に付随する付属語によって判別し，検索においては，肯定の単語と否定の単語を異なる単語として扱う．

（3） 同義語の同一視

ユーザの発話内に含まれる単語から確実にマルチメディアコンテンツを見つけるために，同義語辞書を用意し，同義語を事前に特定の単語に統一した上で，テキストの重みを計算する．本システムでは，事前に数百語からなる同義語辞書を用意して，利用している．

2.4. ブログコーディネート

ブログコーディネートは，発話テキストからユーザの心的状態を推定してユーザに共感するコメントを PaPeRo のひととして生成する．そして，そのコメントを，入力ビデオメッセージ，見出し用キーワード，検索されたマルチメディアコンテンツとともに創出する．心的状態は，発話中の表情や態度から推定できる場合もあるが，個人差が大きく推定が困難である．一方，発話中の表現は，人間の感情を推定する重要な手がかりとなる[11]．そこで，感情を直接表現する語彙に注目して心的状態を推定する．ここではモダリティも含めて表層的な発話表現から推定可能な心的状態を定義し，心的状態と発話表現の対応データベースを構築する．このデータベースを 2.3.の自然言語文検索によって，係り受け関係，否定・肯定表現，同義語辞書を用いて検索する．今回，ユーザの心的状態として，喜び，怒り，悲しみ，願望などデフォールト状態も含めて 10 種類を定義し，それぞれに 100 種類ほどの発話表現を対応づけた．そして，心的状態ごとに生成すべき PaPeRo のコメントを用意し，検索結果に合わせてコメントを生成する．コメントは，「やったね」や「残念」といった言語的なものに，PaPeRo の動作を示す GIF アニメをリンクしている．生成されたコメントは，作成されるブログ中に PaPeRo の動作とともに表示される．

3. 全体システム構成

PaPeRo は，図2に示すモジュールから構成される．全体制御モジュールが，入出力デバイス制御や認識・検知・合成など各種モジュールを対話動作シナリオに基づいて実行する．入力デバイス制御は，マイク，カメラ，圧力センサーを制御して，音声，映像，タッチを入力情報として全体制御プラットフォームに伝達する．全体制御モジュールは，入力された情報を対話動作シナリオに基づいて，音声認識や顔認識などのモジュールに伝達してモジュールの機能を実行する．モジュールの実行結果は，全体制御モジュールに伝達され，対話動作シナリオに基づいて，さらに各種モジュールや出力デバイス制御部に伝達される．出力デバイス制御部は，各モジュールの出力結果を対象となるデバイスから出力する．

今回，入力デバイスとして，ビデオメッセージ入力用にマイクとカメラ，メッセージの開始と終了のスイッチとしてタッチセンサを利用する．また，出力用デバイスとして，動作のためのモータ，PaPeRo の口や耳の動きを示すライト，声を出すためのスピーカを利用する．これらのデバイスを用いてユーザと対話するシナリオを対話・動作シナリオに追加した．さらにマルチメディアブログを WEB に掲載するための WEB 掲載モジュールを実装した．2 節で述べたマルチメディアブログ創作の機能は，PaPeRo のひとつのモジュールとして実装され，その出力結果が出力デバイス制御を介して WEB 上に掲載される．

4. システムの動作

4.1. 基本動作例

旅行日記作成をタスクとし，イラストや写真 2000 点と音楽 550 点を含む約二千ページを検索対象としてプロトタイプシステムを構築した．その動作例として，ユーザと PaPeRo の対話例を表1に，対話の様子を図3に，作成されたブログ画面を図4に示す．図4A には，PaPeRo の内蔵カメラとマイクから撮影されたビデオメッセージが貼り付けられ，Play ボタンで再生可能である．図4B に，ユーザの発話から見出し用キーワードとして抽出されたヨセミテ，ラスベガス，ルーレットが掲載されている．図4C は，ヨセミテ，ラスベガス，ルーレットのイラストが検索結果として表示されている．図4D には，検索された音楽が貼り付けられ Play ボタンで再生可能である．図4E に，PaPeRo のひととして「そうなるといいね」というコメントが GIF アニメとともに掲載されている．これは，「行きたい」という表現から願望という心的状態が推定された結果である．

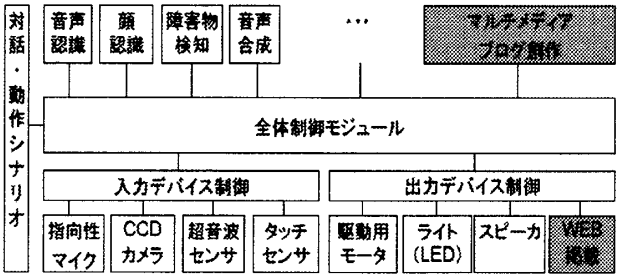


図2: 全体システム構成図

表1：ユーザと PaPeRo の対話例

ユーザ：	ブログ作って
PaPeRo：	はいブログ作っちゃうんですね．頭のボタンを押すとビデオの撮影を始めます．
ユーザ：	（ボタンを押す）
PaPeRo：	終わったらもう一回押してくださいね．では，用意スタート．
ユーザ：	ヨセミテに行きました．そのあとラスベガスでルーレットをしました．また行きたいです．（ボタンを押す）
PaPeRo：	ばっちり撮影できました．これからブログページを作成します．しばらくお待ちください．その間踊らせていただきます．（PaPeRo 踊る）
PaPeRo：	ブログページができました．パソコンや携帯電話でご覧ください．自信作ですよ．

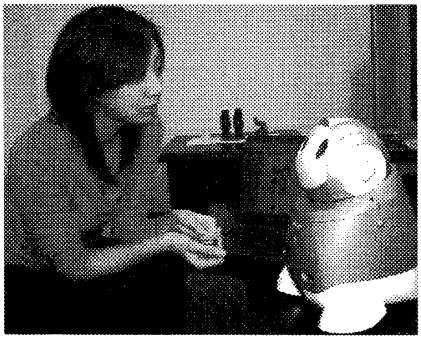


図3：ユーザと PaPeRo の対話の様子

4.2. コーディネートタイプ

本システムは，1）入力ビデオメッセージ，2）映像発話音声認識による発話テキスト，3）発話テキストから抽出した見出し用キーワード，4）検索された関連静止画（写真・イラスト），5）検索された関連音楽・BGM，6）PaPeRo の一言と動作という，最大6種類の素材を組み合わせてブログを創出することができる．提示素材の組み合わせ方（コーディネートタイプ）として，1）関連静止画に発話テキストと見出し用キーワードを加えた静止画ベース，2）入力ビデオメッセージに発話テキストと見出し用キーワードを加えた映像ベース，3）6つの素材すべてを盛り込んだフルタイプを用意した．

5. システム評価

システムを構成する個別機能の性能評価は，旅行会話向け日英音声翻訳システム[5]や音声入力によるテキスト検索システム[6]などで行ってきた．今回，作成者と閲覧者に対してシステム全体に関する評価を行う．作成者に出来上がったブログに関するヒアリング調査を行い，閲覧者にアンケート調査を行う．

5.1.評価方法

（1）作成者評価

10 名のユーザが，旅行や旅行以外の日記文を一人平均 20 文程度入力してブログを作成し，その結果に対してヒアリング調査する．高い音声認識精度（単語正解率 85%以上 95%以下）の結果作成されたブログに対する意見と，低い認識精度（単語

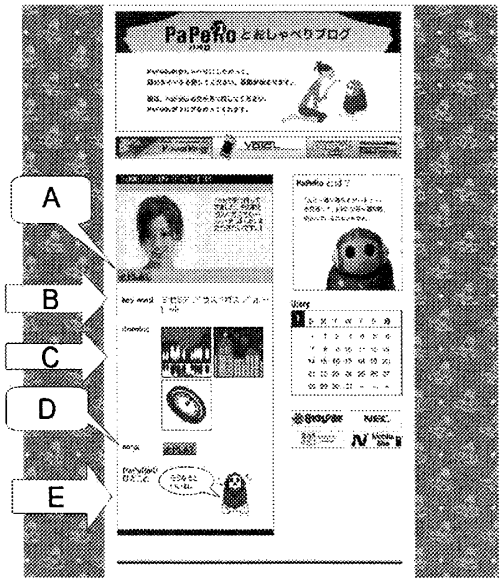


図4：作成されたブログ画面

正解率 50%以上 85%未満）で作成されたブログに対する意見に分け，さらにユーザのブログ作成経験の有無により分析する．

（2）閲覧者評価

176 名（男性 100 名，女性 76 名）の閲覧者に，旅行とレストランに関して作成した2種類のブログを，3つのコーディネートタイプ（静止画ベース，映像ベース，フルタイプ）で提示して，タイプ毎の好感度と6つの素材の効果，改善要望を調査する．

5.2.結果と考察

（1）作成者評価

表2のヒアリング結果から，以下の 3 点が考察される．

-ロボットのブログ作成とコンテンツ拡充への期待

認識精度が高い場合，手間をかけずに簡単にブログが作成でき，ブログ作成経験の有無に関わらずユーザの満足度は高い．現在，十分な検索インデクスが付与されたマルチメディアコンテンツは少ないが，提案システムによりインデクスが付与されたマルチメディアブログが普及することになる．その結果，マルチメディアブログと検索可能となるマルチメディアコンテンツが相乗的に普及・拡充していくことが期待される．

- 音声認識誤りに対する頑健性と寛容性

認識精度が低い場合でも，自立語が正しければ，まずまずの検索結果を得る．また，PaPeRo がブログを作るという本システムは，誤りも面白さとなりユーザに許容されることもある．ただし，必ずしも面白い誤りとなるわけではないので，見出しキーワードの修正や全体レイアウトの編集インタフェースは必要である．

- PaPeRo との対話への期待

認識精度に関係なく，システムとしての面白さや完成度を高めるために，PaPeRo とのやりとり，対話バリエーションや機能の充実が求められている．PaPeRo は，ユーザの発話を解析して内部的に保持しており，以前のブログ内容を反映したコメントやトラックバックが可能である．また，対話シナリオを追加することで，ビデオ撮影中にユーザの発話を解析して，PaPeRo からユーザに質問してブログとすることも可能である．

表 2:ヒアリング調査結果

		ブログ作成経験		
		有り(5 名)	無し(5 名)	共通
音声認識精度	高	・検索できるマルチメディアコンテンツの拡充と普及に期待	・PaPeRo とならブログを作ってみたい	・話すだけでブログが作れるのは楽 ・PaPeRo のひとことが面白い
	低	・見出しキーワードや検索結果，全体レイアウトを編集したい	・検索結果はまずまず． ・誤認識の予想外のイラストも面白い	・誤りも PaPeRo なら許せる ・誤認識テキストが表示されているとビデオ再生したくなる
	共通	・PaPeRo のトラックバックが欲しい ・対話バリエーションの充実	・PaPeRo に曲の選択理由を聞きたい	・PaPeRo のコメントに以前のブログの解析結果を反映

(2)閲覧者評価

- コーディネートタイプ別評価 (グラフ1参照)

フルタイプは，最も好評であり，静止画ベースのものが映像ベースよりも評価が高い．これは，瞬間・受動的に閲覧可能な静止画に対して，映像は積極的に時間をかけて閲覧する必要があること，PaPeRo と向き合って撮影される映像が単調になりがちのためと思われる．PaPeRo がブログとして面白い内容となるようにガイドする機能を検討する．

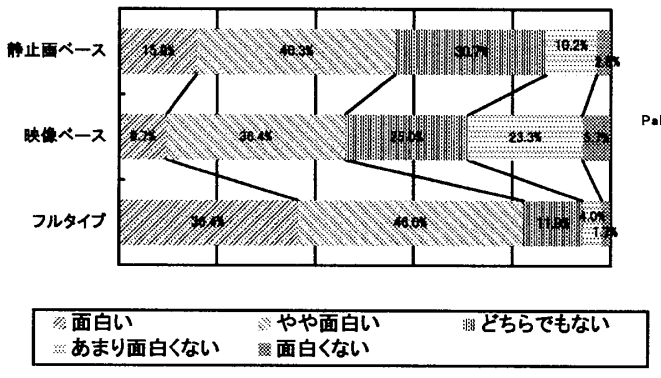
- 素材別評価 (グラフ2参照)

発話テキストと静止画が素材として重要であり，また，映像や PaPeRo の一言に対する必要度も高い．また，映像がキーワードより必要性が高いのは，瞬間・受動的閲覧可能性が優れているからであり，音楽・BGM の必要性が高くないのは，映像と同様に積極的に時間をかけて試聴するインタフェースとしたことが原因と思われる．素材の特性を活かした提示インタフェースが必要である．

- 改善要望

約 4 割の閲覧者から，映像中にテキストや BGM を組み入れる要望があり，約半数の閲覧者から，顔文字や絵文字の利用や，色や書体のバリエーションなどの要望があった．今後，より表現力豊かな提示手法を検討する．

グラフ1:コーディネートタイプ別評価



6. おわりに

ロボットとの対話によるマルチメディアブログ創作システムを，PaPeRo 上に構築して評価・分析を行った．本システムは，作成者にとって，ブログ経験の有無に関わらず簡単にブログを作成する機能を提供し，音声認識誤りも含めて好印象を与えるものであることが判った．また，閲覧者に対してコーディネートタイプと盛り込むべき素材に関する調査を行い，本システムが提供する素材をすべて盛り込んだタイプのマルチメディアブログがもっとも好意的に受け入れられることを確認した．今後，対話機能の充実やインタフェースの改良，マルチメディアオーサリング手法の検討を進め，ブログを創る楽しみと見る喜びを，より多くの人に与える基盤システムとして改良を進めていく．

参考文献

[1] 総務省:ブログ・SNS の現状分析および将来予測,2005/5

[2] 奥村明俊他:“ロボットとの対話によるマルチメディアブログ創作システム”,言語処理学会 13 回年次大会,E1-4,2007

[3] 藤田善弘, “パーソナルロボット PaPeRo の開発,” 計測と制御, Vol.42, No.6 (2003.6)

[4] http://www.incx.nec.co.jp/robot/

[5] 山端潔他:“PDA で動作する旅行会話向け日英双方向音声翻訳システム”, 情処研報, 2002-NL-150-9, 2002.

[6] 池田崇博他:“自由発話音声入力による携帯電話向けテキスト検索システム”, 言語処理学会第 10 回年次大会, pp.109-112, 2004.

[7] 磯谷亮輔 他: “話し言葉認識に向けた基本技術と応用”, 情処研報, 2005-NL-169, pp.109-116, 2005.

[8] 篠田他;「音声認識のための MDL 基準を用いた効果的なガウス数削減」, 信学技報, SP2001-83, 2001-10.

[9] Watanabe et al.; “High Speed Speech Recognition Using Tree-Structured Probability Density Function”, ICASSP-95, pp.556-559, 1995.

[10] S. E. Robertson et al.: Okapi at TREC-3, TREC-3, pp.109-126, 1995.

[11] 中村明 編者:“感情表現辞典”,東京堂出版

グラフ2 : 素材別の評価