既存のアニメキャラクタを用いたテキスト対話システム構築手法

Construction Method of Text-based Agent Dialogue System using Existing Animation Characters

石井 亮 [1*] 東中 竜一郎 [1] 水上 雅博 [1] 片山 太一 [1]
光田 航 [1] 川端 秀寿 [2] 山口 絵美 [2] 安達 敬武 [2]
富田 準二 [1]
Ryo Ishii[1] Ryuichiro Higashinaka[1] Masahiro Mizukami [1] Taichi Katayama [1]
Koh Mitsuda [1] Kawabata Hidetoshi [2] Emi Yamaguchi [2] Noritake Adachi [2]
Junji Tomita [1]

[1] 日本電信電話株式会社 [2] 株式会社ドワンゴ
[1] NTT Corporation [2] DWANGO Co., Ltd.

Abstract: 既存アニメキャラクタを用いたアニメーション付きテキスト対話システムを効率的に構築する手法の提案と構築したシステムの評価結果を報告する．具体的に，発話生成のために，まずキャラクタらしい発話ペアデータを効率的に収集する「なりきり質問応答」という方法論を利用し，次に得られたデータから任意のユーザ発話に対して，キャラクタらしい応答文を自動生成する手法を提案する．動作生成では，人間の発話に対応する身体動作のデータを用いて，任意の発話文から身体動作を自動生成する手法を提案する．これに加えて，少量のキャラクタ独自の動作を加えることでそのキャラクタらしいアニメーションを試みた．実システムを構築し，ユーザ評価を行った結果，提案手法を用いた対話システムは，応答の質が高く，キャラクタの個性が反映されていると高く評価され，有用性が示唆された．

1 はじめに

スマートフォンによる音声エージェントやコミュニケーションロボットの実用化に伴い，エージェントやロボットと自由な対話をするための対話技術の必要性はますます高まっている．特に近年では，エンターテイメントやカウンセリングを目的とした雑談対話システムの構築に関する研究 [7] が盛んにおこなわれており，その実サービスの展開にも着目が集まっている．雑談対話システムのサービス展開の一つとして，我々は既存のアニメキャラクタの CG キャラクタエージェントを用いて，自然な動きを伴い，テキストで雑談が可能な対話システムの実現を目指している．

自然な動きを伴うエージェント対話システムを実現するためには，主にユーザの発話に対するシステムの発話生成，発話に沿った動作生成の 2 つの要素技術が必要である．このような技術を構築する上で，既存のキャラクタの持つ個性を違和感なく反映することが重要である．個性に沿った発話や動作が行えないと，ユーザに違和感や不快感を与えるだけでなく，そのキャラクタが認知されている個性そのものを壊す危険性もある．そのため，適切な個性を持つ発話・動作生成を実現することが必要である．しかしながら，これまで実在するキャラクタの個性を反映した対話システムの構築は我々の知るかぎり実現されていない．

そこで本研究では，既存のアニメキャラクタの個性を持つ対話システムを構築するという新しい試みに取り組む．システムを実現するための要素技術として，アニメキャラクタの個性を持つ自然な発話・動作生成手法を提案する．

具体的に，個性を持った発話生成のために，“なりきり質問応答”という個性を持った良質な発話と応答ペア（以後，発話ペアと呼ぶ）データを収集する手法 [3] を利用し，そこから得られたデータを用いて，任意のユーザ発話に対してキャラクタらしい応答文を自動生成する．動作生成では，人間の発話テキストに対応する身体動作データから，任意の発話文から身体動作を自動生成する手法を提案する．さらに，これに加えて少量のキャラクタ独自の動作を加えることでそのキャラクタらしい動作生成を試みる．

これらの提案手法の適用先として，ライトノベル“俺

*連絡先：日本電信電話株式会社 NTT メディアインテリジェンス研究所
〒 239-0847 神奈川県横須賀市光の丘 1-1
E-mail: ishii.ryo@lab.ntt.co.jp

の嫁が可愛いわけがない”のキャラクタである“新垣あやせ”を題材に，提案手法を利用して対話システムを構築する．“俺の嫁が可愛いわけがない”は，日本で500万部以上を販売した人気ライトノベル作品であり，アニメ化もされている．新垣あやせはメインキャラクタではないが，彼女の性格は“ヤンデレ”と呼ばれ，精神的に不安定で，一度精神状態が乱れると感情の捌け口として極端な暴力的な行動を行うことが特徴的な人気キャラクタである．

構築した対話システムを用いてユーザ評価を行い，発話生成，動作生成の有用性とともに，構築した対話システムの有用性を，応対の良さとキャラクタの個人性を反映できているかという観点から評価を行う．

以降，2，3章では発話生成，動作生成手法を述べる．4章では対話システムの構築について，5章では構築したシステムを用いたユーザ評価について示す．最後に，6章で得られた結果を考察し，7章で本論文のまとめを述べる．

2 発話生成

2.1 アプローチ

©Tsukasa Fushimi/ASCII MEDIA WORKS/OIP2 ©BANDAI NAMCO Entertainment Inc. Copyright©2017 Live2D Inc.

図 1: 新垣あやせのなりきり質問応答を用いたデータ収集のためのサイト画面.

一般的に，品質が担保された個性を反映させた発話生成を行うためには，人手で大規模な発話ペアデータを事前に準備して利用する必要あるためコストが高いという課題がある．本研究では，個性を反映した高品質な発話ペアを効率的に構築する手法として，“なりきり質問応答手法”というデータ収集手法を利用する [3]．これは，複数のユーザがオンラインで参加し，特定のキャラクターに質問をしたり，そのキャラクタになりきって質問に回答することで，効率的に質の高いキャラクタらしい発話ペアを収集する手法である．具体的に，ユーザは二つの役割が与えられる．一つは，キャラクタに対する質問（発話）を行う．この質問は，全てのユーザに通知がされる．二つ目の役割は，キャラクタになりきって，その質問に回答を行う．これによって，ユーザの質問を共有利用し，各々の質問に回答を行いあうことで，質問と回答の発話ペアを効率的に収集可能である．また，ロールプレイングするという体験自体が面白いため，ユーザが積極的に参加しやすくなる効果が見込まれる [13]．この手法を用いて，効率的にアニメキャラクタらしい発話ペアデータを収集可能であると考えた．

また，得られたデータから適切な回答を抽出する発話生成モデルを構築するニューラルネットワークを用いた発話生成手法を提案する．

2.2 なりきり質問応答手法を用いたデータ収集

動画ストリーミングサービスであるニコニコ動画[1]では，様々なキャラクタのファンのためのチャンネルサービスを展開している．チャンネルの利用は登録をおこなった購読者に限定されている．本研究では，このチャンネルサービスと連携し，なりきり質問応答を行う掲示板を“新垣あやせ”に関するチャンネルに構築した．図1にサイトのスクリーンショットを示す．ユーザは用意されたテキストフォームから，新垣あやせに自由に質問をすることができる．また，新垣あやせになりきって応答したいユーザはその質問に対して自由に回答を投稿することが可能である．回答時には同時に，その発話に付随する感情をラベリングをおこなった．

ユーザの参加意欲を高めるために，Webサイトでは投稿数でユーザのランキングを示した．さらに，各回答の横には“ポイねボタン”が配置されており，ユーザの回答が新垣あやせらしい（ポイ）と思う場合，この回答に対して，“ポイねボタン”を押下することで，回答の質にユーザ同士の評価を反映させることができるように工夫した．

2017年10月になりきり質問応答を行うWebサイトをオープンし，約90日間にわたりサービスを運用した．合計で333名のユーザが参加をおこなった．収集された発話ペアは，約20日間で10,000を突破し，最終的に15,112という多くの発話ペアが得られた．ユーザはこのなりきり応答サイトに自発的に参加をしており，報酬は支払われなかった．それにも拘らず，このような多くのデータを集めることができたことは，な

[1] http://www.nicovideo.jp/

評価項目	平均値
（Q1）Web サイトの使い勝手の良さ	4.08
（Q2）なりきり応答を楽しめたか	4.53
（Q3）Web サイトを再び利用したいか	4.56

表 1: なりきり質問応答サイトでの体験のユーザ評価結果.

りきり質問応答を用いたデータ収集が有用である可能性が示されたと考えられる.

なりきり質問応答サイトを使用に関する満足度を判断するために，参加ユーザを対象にアンケート評価を実施した．計 36 人のユーザが評価に協力をし，5 段階のリッカート尺度（1〜5 点）に基づいて，表 1 に示した項目について回答をおこなった．表 1 に，ユーザの評価値の平均値を示す．結果を見ると，“なりきり質問応答を楽しめた”，“Web サイトを再び利用したいか”の項目において，4.5 以上の高い評定値が得られた．なりきり応答を利用した Web サービスが，ユーザにとって魅力的なものであったことが示唆された．

次に，得られた新垣あやせの応答発話データの品質を評価するために，参加ユーザによる主観評価を実施した．50 程度の発話ペアに対して，自然さや，キャラクタの個性を反映しているかを評価した．その結果，5 段階のリッカート尺度（1〜5 点）で，約 3.5 以上，4.0 未満の自然性／性格スコアを得た．これは，なりきり質問応答を通して，集められた応答発話データの品質が良好であることを示している．しかしながら，人間のユーザによって応答データが作られたとしても，4.0 以上の評定値を得るのが困難であったことは驚くべき結果であった．すなわち，特定のキャラクタの個性を反映した発話の生成が人間にとってさえも困難なタスクであることが示唆された．

2.3 発話生成器の構築

次に，収集した発話ペアデータを用いて，キャラクタの個性を表す発話生成手法を提案する．収集したデータ数がニューラルネットワークを用いた発話生成モデルを訓練するのに十分な大きさではなかったので [18]，我々は，得られた応答データの中から最適な応答を抽出するアプローチを選択した．すなわち，最も関連性の高い発話ペアの応答を選択するという問題に取り組んだ．本研究では，適切な発話ペアの選択のためにニューラル翻訳モデルを使用する．ユーザ発話に対して，単語のマッチングを取るだけでは不十分な場合があるため，オープンソースの検索エンジンである LUCENE[2]から取得した結果を再ランク付けする新たな手法を開発した．本手法は，多言語質問応答（CLQA）[9] やニューラル対話モデル [18] 等の最近の進歩から着眼を得ている．また，適切な回答候補が上位にランク付けされるように，質問の意味的および意図レベルのマッチングを実施した．こ入力質問 Q が与えられるとき，下記のステップで応答を生成する．

1. Q が与えられると，LUCENE によって，上位 N 件の質問応答ペア $(Q'_1, A'_1) \ldots (Q'_N, A'_N)$ が検索される．なお，Twitter データを用いて検索対象となる質問応答のデータベースを拡張する工夫をしている [20].

2. Q と Q' について，質問タイプ判定と固有表現抽出が実行され，質問タイプと固有表現（体系として，関根の拡張固有表現 [17] を利用）が抽出される．Q によって尋ねられている固有表現が A' の中にどの程度含まれているかを計算し，これを質問タイプマッチスコア（qtype_match_score）とする．

3. 焦点抽出モジュールを用い，Q および Q' から焦点（話題を表す名詞句）を抽出する．もし，Q の焦点が Q' に含まれていたら，焦点スコア（center-word_score）を 1 とし，そうでなければ 0 とする．

4. 翻訳モデルは，Q から A' が生成される確率，すなわち，$p(A'|Q)$ を計算する．また，逆向きの翻訳確率として $p(Q|A')$ も計算する．このような逆向きの翻訳は CLQA においても有効とされている [9]．生成確率は出力側の単語数で正規化する．確率値はレンジの違いなどにより，他のスコアと統合することが難しいため，我々はこの確率をもとに回答候補をランキングし，その順位から翻訳スコア（translation_score，逆向きの場合は rev_translation_score）を得る．具体的には，ある回答候補の順位が r のとき，翻訳スコアを以下で得る．

$$1.0 - (r - 1)/\text{max_rank}, \tag{1}$$

ここで，max_rank は回答候補が取りうるの最大の数である．翻訳モデルは，約 50 万の一般的な質問応答ペアで事前学習したものに，なりきり質問応答で得られた発話ペアで fine-tune することで学習した．逆向きのモデルは質問と応答を入れ替えて同じ処理を行った．学習には opennmt ツールキットをデフォルトの設定で利用している．

5. 意味的類似性モデルにより Q と Q' の意味の類似性を計測する．これには，Word2vec[14] を用いる．まず，Q および Q' のそれぞれについて，単語ベクトルを求めて，平均ベクトルを作成した

[2]https://lucene.apache.org/

後，コサイン類似度を計算し，これを類似度スコア（semantic_similarity_score）とする．

6. これまでのスコアを重みづけで足し合わせ，最終的なスコアを得る．

$$
\begin{aligned}
&\mathrm{score}(Q,(Q',A')) \\
&\quad = w_1 * \mathrm{search_score} \\
&\quad + w_2 * \mathrm{qtypes_match_score} \\
&\quad + w_3 * \mathrm{center\text{-}word_score} \\
&\quad + w_4 * \mathrm{translation_score} \\
&\quad + w_5 * \mathrm{rev_translation_score} \\
&\quad + w_6 * \mathrm{semantic_similarity_score} \qquad (2)
\end{aligned}
$$

ここで，search_score とは，LUCENE による検索結果の順位から得られるスコアであり，式 1 から得られる．$w_1 \ldots w_6$ は重みであり，本研究ではすべて 1.0 である．

7. 前述のスコアによって，回答候補はランキングされ，上位 M 件が出力される．

この手法で，回答候補の抽出を行った際に，場合によっては収集データに含まれる質問データと入力質問が意味的に一致したものが含まれない場合に，上位N件の検索結果に適切な回答候補を含まないことが考えられる．上位 N 個の質問が入力質問と意味的に一致しない場合，答えは不適切である可能性があります．質問の範囲を広くするために，Twitter を使用して質問と回答のペアを拡張しました．私たちの方法は簡単でした．質問回答のペアで 2 回以上出現した回答 A ごとに，Levenshtein の距離（文の長さで正規化）が 0.1 以下の A のようなツイートを検索しました．次に，ツイートが他のツイートと返信関係にある場合は，それらを取得して A と結合し，拡張された質問と回答のペアを形成します． 2 回以上発生した回答に焦点を合わせた理由は，主にクロールの効率性によるものですが，複数回発生した回答は可能性が高いです．問題のキャラクターの特徴であること．村井と綾瀬について，それぞれ 2,607,658 と 1,032,492 の拡張された質問と回答のペアを得ました．

3 動作生成手法

3.1 アプローチ

エージェントやヒューマノイドロボットに適切な身体動作を付与することで，見た目の自然さの向上だけでなく，会話を促進することが示されている．例えば，発話に付随する動作は，発話の説得力を強化し，対話相手が発話の内容を理解しやすくする効果がある [10]．このような背景から，アニメキャラクタの動作をより自然で人間らしいものとするための動作生成手法を提案する．テキスト対話システムでは，システム発話から得られる言語情報を入力として，動作生成を行うことが考えられる．言語情報を用いた動作生成の従来研究では，主に単語情報を用いた，頷きの有無や，限定的なハンドジェスチャ[5, 6, 8] といったごく一部の動作の生成について取り組まれていた．本研究では，多様な言語情報を用いて，より網羅的な全身の動作の生成に取り組む．具体的なアプローチとして，人間の対話時の発話の言語情報と動作情報のコーパスデータを構築し，それらの共起関係を機械学習を利用して学習し，発話の言語情報を入力として動作を生成することを考えた．以降，コーパスデータの構築，動作生成手法とその性能について述べる．

図 2: 2 者対話の様子.

3.2 動作生成のための対話コーパス

2 者対話を対象に，発話言語およびそれに伴う身体の動作データを含む，言語・非言語マルチモーダルコーパスの構築を行った．2 者対話の参加者は，20–50 代の日本人男性・女性であり初対面であった．参加者は計 24 人（12 ペア）であった．参加者は図 2 のように，互いに向き合って着座した．対話内容は，雑談対話に加えて，議論対話，および発話に伴う頷きやハンドジェスチャなどの多様な動作に関するデータを多く収集するために，アニメーションの説明対話を採用した．アニメーション説明対話においては，参加者はそれぞれ，異なる内容のアニメーション（Tom & Jerry）を視聴した後，対話相手にアニメーション内容を説明した．対話相手は，説明者に自由に質問をし，自由に会話を行うことを許可した．発話の収録のために各被験者の胸につけられた指向性ピンマイクを用いた．対話状況の全体的な外観や参加者の様子の収録として，ビデオカメラを用いた．映像は 30Hz で収録された．図 2 はそ

の一例である．

雑談対話，議論対話，アニメーション説明対話の対話時間は，それぞれ 10 分間とし，本研究では前半の 5 分間ずつのデータを使用した．各ペアで，1 つの雑談対話，1 つの議論対話，2 つのアニメーション説明セッションを実施した．よって，各ペア 20 分の対話データを収集し，12 ペアで計 240 分の対話データを収集した．

次に，取得した言語・非言語データを示す．

- 発話： 人手で音声情報から発声言語の書き起こしを行った後，発話内容を確認して文を分割した．さらに，係り受け解析エンジン [4] を利用して，各文を文節に分割した．分割された文節数は 11877 であった．
- 顔向き： 参加者の正面に設置されたビデオカメラから撮影された参加者の正面映像に対して，顔画像処理ツールである OpenFace[16] を利用して，3 次元顔向き情報である yaw，roll，pitch の角度をそれぞれ取得した．それぞれの角度が，10 度以下の時に micro，20 度以下の時に small，30 度以下の時を medium，45 度以上の時を large として分類した．
- 頷き： 人手で映像から頷きが発生した区間をラベリングした．連続的に発生した頷きは 1 つの頷きイベントとして扱った．また，回数を 1 回から 5 回（以上）の 5 段階に分類した．また，頷きの深さについて，OpenFace を利用して，頷きの際の頭部姿勢 pitch の開始と，最も深く回転した箇所の角度の差を算出した．その角度が，10 度以下の時に micro，20 度以下の時に small，30 度以下の時を medium，45 度以上の時を large として分類した．
- ハンドジェスチャ： ハンドジェスチャが行われている区間を人手アノテーションをおこなった．なお，ジェスチャの一連の動作は，下記の 4 つの状態に分類した．
  - Prep：ホームポジションからジェスチャをするために手を上げる
  - Hold：空中に手をあげたままの状態 (ジェスチャ開始までの待機時間)
  - Stroke：ジェスチャを実施
  - Return：手をホームポジションに戻す

  ただし，本研究では，取扱いの簡便さから，Prep から Return までの一連の動作を一つのジェスチャイベントとして扱った．さらに，ハンドジェスチャの種類を，McNeil のハンドジェスチャの分類 [11] を基に，下記の 8 種類に分類した．
  - Iconic： 情景描写，動作を表現するために使われるジェスチャ．
  - Metaphoric： Iconic と同様，絵画的，図形的なジェスチャであるが，指示される内容は抽象的な事柄，概念となる．例えば，時間の流れなど．
  - Beat： 発話の調子を整えたり，発言を強調．手を振動させたり，発話に合わせて手を振る．
  - Deictic： 指差しなど，直接的に方向や，場所，事物を指し示すジェスチャ．
  - Feedback： 他人の発話に同調・同意・呼応して出るジェスチャ．他人の前の発話・ジェスチャに対して，呼応して発言したときに付随するジェスチャ．また，相手のジェスチャを模倣して同じ形のジェスチャを行ったもの．
  - Compellation： 相手に呼びかけを行うジェスチャ．
  - Hesitate： 言いよどみ時に出るジェスチャ．
  - Others： 判断に迷うが，何か意味がありそうなジェスチャ．
- 上半身の姿勢： 本対話シーンでは，参加者は着座しており着座した位置の大きな変化は無かった．そのため，頭部の 3 次元位置を基に，上半身の前後の姿勢を抽出した．具体的に，OpenFace を利用して得られた頭部位置の前後方向の座標位置と中心位置の位置の差を取得した．その位置情報から，上半身の姿勢変化の角度を算出し，それが 10 度以下の時に micro，20 度以下の時に small，30 度以下の時を medium，45 度以上の時を large として分類した．

得られたコーパスデータのパラメータの一覧を表 2 に示す．また，人手アノテーションには ELAN [19] を使用し，上記のすべてのデータを 30Hz の時間分解能で統合した．

3.3 動作生成器の構築

構築したコーパスデータを用いて，単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為を入力として，表に示した 8 項目の動作においてそれぞれ一つの動作クラスを文節ごとに生成するモデルを，決定木アルゴリズム C4.5 を用いて構築した．すなわち，文節毎に，8 つのーションラベルが生成される．具体的に，使用した言語特徴量は下記の通りである．

- 文字数：文節内の文字数

生成対象	クラス数	クラス
頷きの回数	6	0, 1, 2, 3, 4, more than 5
頷きの深さ	4	micro, small, medium, large
頭部姿勢 yaw	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
頭部姿勢 roll	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
頭部姿勢 pitch	7	front, up-micro, up-small, up-medium, up-large, up-micro, up-small, up-medium, up-large
表情	8	happiness, sadness, surprise, fear, anger, disgust, contempt, normal
ハンドジェスチャ	9	none, iconic, metaphoric, beat, deictic, feedback, compellation, hesitate, others
上半身の姿勢	7	center, forward-small, forward-medium, forward-large, backward-small, backward-medium, backward-large

表 2: 動作生成のパラメータ一覧.

生成対象	チャンスレベル	提案手法
頷きの回数	0.226	0.428
頷きの深さ	0.304	0.475
顔回転 yaw	0.232	0.329
顔回転 roll	0.297	0.397
顔回転 pitch	0.261	0.378
表情	0.175	0.290
ハンドジェスチャ	0.156	0.303
上半身の姿勢	0.183	0.311

表 3: チャンスレベルと提案手法による推定精度（F 値）

- 位置：文頭，文末からの文節の位置
- 単語：形態素解析ツール Jtag [1] により抽出された文節内の単語情報（Bag-of-words）
- 品詞：Jtag により抽出された文節内の単語の品詞情報
- シソーラス：日本語語彙大系に基づく文節内の単語のシソーラス情報
- 発話行為：単語 n-gram およびシソーラス情報を用いた発話行為推定手法 [2, 12] により文ごとに抽出された発話行為（33 種類）

24 人の参加者のデータの内，23 人のデータを学習に用いて 1 人のデータで評価を行う 24 交差検定法により評価を行った．これにより，他者のデータのみから，どの程度，実際の人間の動作を生成できるかを評価した．性能評価結果である F 値の平均値を表 2 示す．なお，チャンスレベルは最も，正解数の多いクラスを全て出力した際の性能を示している．表 2 を見ると，全ての生成対象で，チャンスレベルよりも大幅に精度が向上した（対応のある t 検定の結果：$p < .05$）．この結果から，発話言語から得られる，単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為を用いた本提案手法は，表 1 に示したような全身の動作生成に有効であることが示唆された．

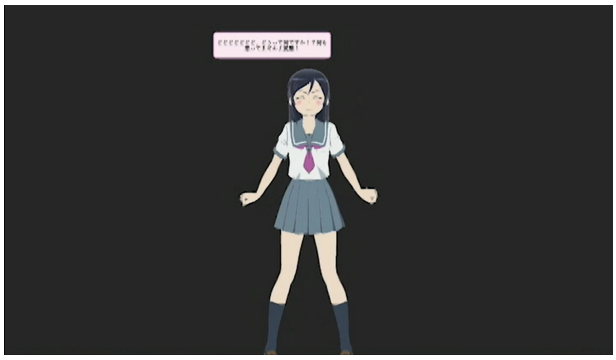

図 3: 対話システムのユーザ提示画面の例：システム発話に含まれる「変態！」のテキスト表示に合わせて，腕を上下して顔を突き出すオリジナル動作を行っているシーン．

オリジナル動作	トリガとなる単語・文の例
回し蹴り	ぶち殺しますよ
腕を組んでドン引きする	気持ち悪い
腕を上下して顔を突き出す	変態！
右手で相手を指差す	通報します

表 4: オリジナル動作と生成トリガとなる単語・文の例

3.4 キャラクタの個性を反映するオリジナル動作の追加

前節までに提案した動作生成に加えて，新垣あやせらしい動作を生成するために，新垣あやせ固有の動作をアニメ中の動作から抽出し，表 4 に示すような 4 つのオリジナル動作を追加した．これらの動作は，新垣あやせのアニメーション作成経験のあるクリエータと共同で選定した．

これらのオリジナル動作に対して，動作を生成するトリガとなる単語および文を設定し，これらの単語，文がシステム発話内に現れた際に，動作生成モデルの出力結果よりもオリジナル動作を優先して生成するようにした．動作数が 4 つと少量であるが，これ以上の特筆するほどの個性的な動作は見当たらなかったため，十

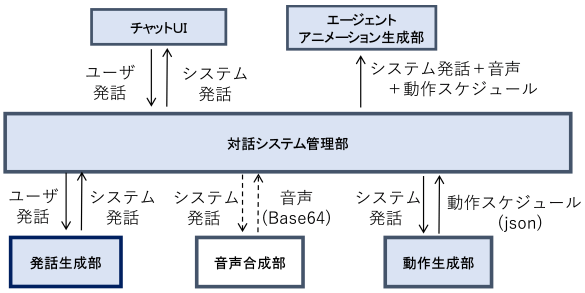

図 4: 対話システムの構成図.

分な動作数であると考えた.

4 対話システムの構築

提案した発話生成と動作生成手法を用いて，ユーザ発話に対して発話と動作で応答可能な対話システムを構築した．システム構成図を，図 3 に示す.

ユーザはチャット UI から，入力テキストを入力する．対話管理部がそれを受け取ると，最初に発話生成部にユーザテキストが送信される．発話生成部からシステム発話を取得した後，システム発話を動作生成部に送信し，システム発話テキストの各文節での動作情報を取得する.

これに加えて，テキスト対話では必須ではないが，音声合成部で発話音声の取得も可能である．本システムでは，リップシンク動作の生成のために音声合成部から得られる音声を利用した.

対話管理部から，システム発話テキストと動作スケジュール，および音声をエージェントアニメーション生成部に送信する．エージェントアニメーション部では，動作スケジュールに従って，エージェントの動作を生成するとともに，発話テキストをキャラクタの上部に吹き出しの中に，最初の文字から等しい時間間隔で順番に表示される．動作アニメーションの生成手段として，UNITY 上で CG キャラクタを作成し，表 2 のリストにある動作リストに対応するアニメーションをリアルタイムに生成した．なおこのとき，表 1 に示した 8 つの生成対象は独立して動作が可能であり，頭部動作についても，全てのパラメータが混合されて生成される．文字の表示のタイミングに合わせて，エージェントの動作アニメーションが生成される．提示画面例を図 3 に示す.

なお，対話管理部から，チャット UI にシステム発話を送信し，チャット UI においてもユーザ発話に加えて，システム発話を提示可能である.

5 評価実験

5.1 評価方法

構築した対話システムを用いて，提案手法の有効性を被験者実験によって評価した．評価項目として，提案する発話生成および動作生成手法による対話システムの応答の有用性を評価する．本研究の目的は，既存キャラクタらしい対話システムを構築することにある．評価観点として，応答の良さに加えて，キャラクタ性の再現性について特に着目する.

発話生成の実験条件は下記の 3 条件を設定した.

- **U-AIML 条件**：発話生成をするために利用される一般的な手法である AIML で書かれたルールベースの手法を用いる．具体的に，著者らがこれまで構築した 300K にも及ぶ大規模 AIML データベースを用いる．また日本語では，文末表現はキャラクタ性を示す重要な要素の一つであるため，文末変換手法 [15] を用いて文末表現を“新垣あやせ”らしいものに変換した.
- **U-PROP 条件**：提案手法に発話生成を利用する．なりきり質問応答手法で収集した質問と回答のペアから 20 の発話候補を抽出し，最もランクの高かった発話を生成する．なお，重みパラメータ w_1 ～w_6 はすべて 1.0 に設定した.
- **U-GOLD 条件**：なりきり質問応答手法によって，収集した質問に対するオンラインユーザ（人間）の回答を利用する．質問に複数の回答が与えられている場合は，ランダムに 1 つを選択した.

U-AIML 条件と U-PROP 条件を比較することによって，提案する発話生成の有用性，U-PROP 条件と U-GOLD 条件を比較することで，提案手法と人手による発話生成と比較評価を行う.

動作生成の実験条件は下記の 4 条件を設定した.

- **M-BASE 条件**：リップシンク，表情といったキャラクタの基本動作を生成する.
- **M-RAND 条件**：リップシンク，表情に加えて，全身の動作をランダムに生成する.
- **M-PROP1 条件**：人間のデータを基にした動作生成手法を利用する.
- **M-PROP2 条件**：人間のデータを基にした動作生成手法の利用に加えて，キャラクタ独自の特徴的な動作を少量追加する.

BASE 条件と RAND 条件を比較することによって，全身の動作生成の有用性，RAND 条件と M-PROP1 条

表 5: 主観評価項目.

評価項目	質問文
総合的な応答の良さ	総合的に応答は良いですか?
総合的な応答のキャラクタらしさ	総合的に応答は“新垣あやせ”らしいですか?

件を比較することで，提案する人間の動作データの学習による生成手法の有用性を評価する．また，M-PROP1 条件と M-PROP2 条件を比較することで，提案する人間の動作データの学習による生成に加えて，少量のキャラクタ独自の特徴的な動作を追加することの有用性を評価する．

これらの発話生成の 3 条件，動作生成の 4 条件を組み合わせた 12 条件を実験条件として設定した．

実験方法として，各条件で比較をする際に同一のユーザ発話を設定し，それに対するシステムの応答発話，動作を評価することとした．具体的に，収集したなりきり質問応答データから，10 の質問発話をランダムに抽出した．被験者は，最初に 3 秒間，ユーザ発話テキストを提示し，次にシステムの応答が提示される映像を視聴してもらった．視聴ごとに，被験者は 7 段階のリッカート尺度（1〜7 点）を用いて，対話システムの応答についての印象を主観評価した．具体的な評価項目を表 5 に示す．12 条件ごとに 10 発話の映像が用意されているため，映像視聴と評価は 120 施行となる．順序効果を考慮し，被験者ごとに提示される映像の提示順序はランダムとした．

5.2 評価結果

被験者 7 人による結果実験を実施した．各実験条件における各主観評価項目の被験者ごとの平均値を算出し，さらにその平均値を算出した結果を図 5，6 に示す．

まず，総合的な応答の良さの評定値について，発話条件と動作条件の要因の効果を検証するために二次元配置の分散分析を行った．その結果，発話条件と動作条件の単純主効果が認められ，交互作用は見られなかった（発話条件：$F(2,72) = 3.92, p < .05$，動作条件：$F(3,72) = 3.82, p < .05$）．

次に，発話条件の要因の単純手法が認められため，Bonfferoni 法の多重比較をおこなって，各動作条件下において，どの発話条件間に差がみられるかを検証した．条件の組み合わせが多数存在するため，本稿では U-AIML 条件と U-PROP 条件間，U-PROP 条件と U-GOLD 条件間の 2 つの差についてのみ述べる．

まず，M-PROP1 条件下のみで，U-AIML と U-PROP 条件間の有意傾向が認められた（$p < .10$）．よって，提案する動作生成手法（オリジナル動作無し）を用いたときに，提案する発話生成手法が一般的な手法である AIML による発話生成よりも評価値が高い傾向が示唆された．また，M-RAND と M-PROP2 条件下で，U-PROP 条件と U-GOLD 条件間の有意差が認められた（$p < .05$，$p < .01$）．よって，ランダムな動作生成，提案する動作生成手法（オリジナル動作有り）を用いたときに，提案する発話生成手法よりも人手の応答発話の方が評価値が高いことが示唆された．

次に，動作条件の要因の単純手法が認められため，同様に多重比較により，各発話条件下において，どの動作条件間に差がみられるかを検証した．条件の組み合わせが多数存在するため，本稿では M-BASE 条件と M-RAND 条件間，M-RAND 条件と M-PROP1 条件，M-PROP1 条件と M-PROP2 条件間の 3 つの差についてのみ述べる．

まず，U-AIML 条件下のみで，M-BASE と M-RAND 条件間の有意傾向が認められた（$p < .10$）．よって，AIML を用いた発話生成を用いたときに，ランダムに動作生成することは，動作をさせないときに比べて評価値が低い傾向が示唆された．また，U-AIML，U-PROP，U-GOLD のすべての条件下で，M-RAND 条件と M-PROP1 条件間の有意差または有意傾向が認められた（$p < .01$，$p < .01$，$p < .10$）．よって，発話生成の条件によらずに，ランダムに動作生成するよりも提案する動作生成手法（オリジナル動作無し）した方が評価値が高いこと（または高い傾向）が示唆された．加えて，U-PROP 条件下で，M-PROP1 条件と M-PROP2 条件間の有意傾向が認められた（$p < .10$）．よって，提案する発話生成を行った際に，提案する動作生成手法にオリジナル動作を追加するよりも，追加しない方が評価値が高い傾向であることが示唆された．

次に，総合的な応答のキャラクタらしさの評定値についても同様の分析を行った．その結果，発話条件と動作条件の単純主効果が認められ，交互作用は見られなかった（発話条件：$F(2,72) = 2.61, p < .10$，動作条件：$F(3,72) = 5.63, p < .01$）．

次に，Bonfferoni 法の多重比較を用いて，各動作条件下において，どの発話条件間に差がみられるかを検証した．まず，M-RAND，M-PROP1，M-PROP2 条件下で，U-AIML と U-PROP 条件間の有意差または有意傾向が認められた（$p < .10$，$p < .10$，$p < .01$）．よって，ランダム，提案する動作生成手法（オリジナル動作有無によらず）を用いたときに，提案する発話生成手法が AIML による発話生成よりも評価値が高いこ

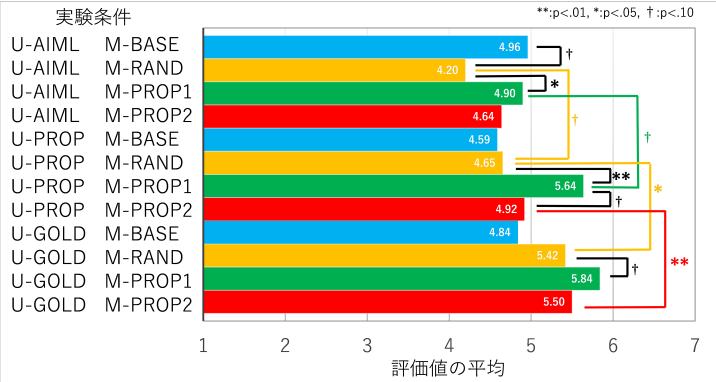

図 5: 各実験条件における，“ 総合的な応答の良さ ” の印象の平均主観評価値．

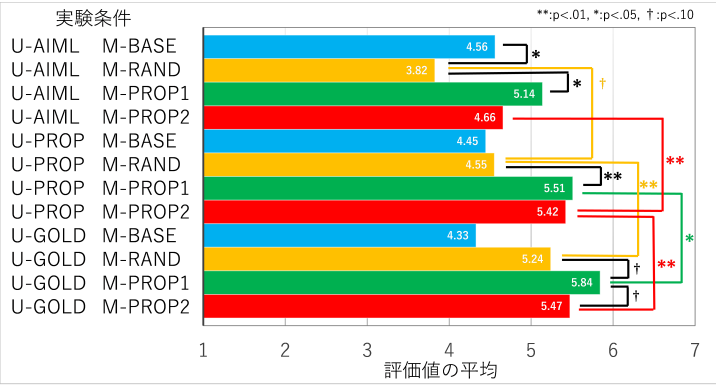

図 6: 各実験条件における，“ 総合的な応答のキャラクタらしさ ” の印象の平均主観評価値．

と（または高い傾向）が示唆された．また，M-RAND と M-PROP1 条件下で，U-PROP 条件と U-GOLD 条件間の有意差が認められた（$p < .05$，$p < .05$）．よって，ランダムな動作生成，提案する動作生成手法（オリジナル動作無し）を用いたときに，提案する発話生成手法よりも人手の応答発話の方が評価値が高いことが示唆された．

次に，各発話条件下において，どの動作条件間に差がみられるかを検証した．まず，U-AIML 条件下のみで，M-BASE と M-RAND 条件間の有意差が認められた（$p < .05$）．よって，AIML を用いた発話生成を用いたときに，ランダムに動作生成することは，動作をさせないときに比べて評価値が低いことが示唆された．また，U-AIML，U-PROP，U-GOLD のすべての条件下で，M-RAND 条件と M-PROP1 条件間の有意差または有意傾向が認められた（$p < .05$，$p < .01$，$p < .10$）．よって，発話生成の条件によらずに，ランダムに動作生成するよりも提案する動作生成手法（オリジナル動作無し）した方が評価値が高いこと（または傾向）が示唆された．加えて，U-GOLD 条件下で，M-PROP1 条件と M-PROP2 条件間の有意傾向が認められた（$p < .10$）．よって，人手で作成した発話を行った際に，提案する動作生成手法にオリジナル動作を追加するよりも，追加しない方が評価値が高い傾向であることが示唆された．

6 考察

評価結果から，応答の良さ，キャラクタらしさのどちらの項目においても，発話生成と動作生成が印象に影響を与えることが認められた．また，提案手法を用いて構築した対話システム（U-PROP × M-PROP1）の評定値は，1-7 の評定値において，応答の良さで 5.83，キャラクタらしさで 5.98 と高い値であった．

被験者が 7 名であり，サンプル数が少なすぎることもあり，全ての動作条件下で差がみられたわけではないが，提案する発話生成が AIML を用いた発話生成よりも，応答の良さやキャラクタらしさの印象を向上させることが示唆された．また，提案手法は，人手による発話生成には及ばないことも示唆された．また，同様に提案する動作生成（オリジナル動作無し）もランダムに生成するものよりも印象を向上させることが示唆された．

狙い通りでなかった結果として，提案する動作生成

手法において，オリジナル動作を追加したものは追加しないものよりも評価値が低い傾向もみられた．実験後に被験者に，聞き取り調査をしたところ，オリジナル動作と他の通常動作との移行時がスムーズでないことや，タイミングが発話文の表示と完全に一致していないという意見があった．今回，作成したオリジナル動作は通常動作に比べて，全体的な動作が大きかったため，他の動作との移行がスムーズにいかなかったことが考えられる．よって，提案する動作生成手法にオリジナル動作を追加する際には，オリジナル動作生成時に前後の動作とスムーズな移行がなされるようなデザイン的な配慮が重要であることが分かった．

また，提案する動作生成にオリジナル動作を入れない動作条件（M-PROP1）において，U-PROP × M-PROP1 条件で，評定値が 6 点弱と非常に高かったことから（応答の良さで 5.83，キャラクタらしさで 5.98），オリジナル動作を入れなくとも良い印象を与えられる応答が生成できる可能性が示唆された．

今後は，追加実験を行ってより多くのサンプル数を扱い，発話条件や動作条件の相互効果の有無についても詳しく検証を行う予定である．

7 結論

既存アニメキャラクタのアニメーション付きテキストチャットシステムを効率的に構築する手法の提案した．発話生成のために，キャラクタらしい発話ペアデータを効率的に収集する「なりきり質問応答」という方法論を利用し，次に得られたデータから任意のユーザ発話に対して，キャラクタらしい応答文を自動生成する手法を提案した．動作生成では，人間の発話に対応する身体動作のデータを用いて，任意の発話文から身体動作を自動生成する手法を提案した．実システムを構築し，ユーザ評価を行った結果，提案手法は，応答の質，キャラクタらしさの印象を向上することが示唆された．

今後は，システム評価の追加実験を実施するとともに，手法の改良を進める予定である．

参考文献

[1] Takeshi Fuchi and Shinichiro Takagi. Japanese morphological analyzer using word cooccurrence -jtag. In *International conference on Computational linguistics*, pp. 409–413, 1998.

[2] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *International conference on Computational linguistics*, pp. 928–939, 2014.

[3] Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, and Nozomi Kobayashi. Question answering technology for pinpointing answers to a wide range of questions. *NTT Technical Review*, Vol. 11, No. 7, 2013.

[4] Kenji Imamura. Analysis of japanese dependency analysis of semi-spoken words by series labeling. In *Proceedings of the Annual Meeting of the Association for Natural Language Processing*, pp. 518–521, 2007.

[5] Carlos T. Ishi, Judith Haas, Freerk P. Wilbers, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of head motions and speech, and head motion control in an android. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 548–553, 2007.

[6] Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Head motion during dialogue speech and nod timing control in humanoid robots. In *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 293–300, 2010.

[7] Ryo Ishii, Ryuichiro Higashinaka, Kyosuke Nishida, Taichi Katayama, Nozomi Kobayashi, and Junji Tomita. Automatically generating head nods with linguistic information. In Gabriele Meiselwitz, editor, *Social Computing and Social Media. Technologies and Analytics*, pp. 383–391, Cham, 2018. Springer International Publishing.

[8] Yuki Kadono, Yutaka Takase, and Yukiko I. Nakano. Generating iconic gestures based on graphic data analysis and clustering. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, pp. 447–448, Piscataway, NJ, USA, 2016. IEEE Press.

[9] Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. Building effective question answering characters. In *Proc. SIGDIAL*, pp. 18–27, 2009.

[10] Manja Lohse, Reinier Rothuis, Jorge Gallego-Pérez, Daphne E. Karreman, and Vanessa Ev-

ers. Robot gestures make difficult tasks easier: The impact of gestures on perceived workload and task performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pp. 1459–1466, New York, NY, USA, 2014. ACM.

[11] David McNeill. *Hand and Mind: What Gestures Reveal About Thought*. University Of Chicago Press, 1996.

[12] Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. Controlling listening-oriented dialogue using partially observable markov decision processes. In *International conference on computational linguistics*, pp. 761–769, 2010.

[13] Morry Van Ments. *The Effective Use of Role Play: Practical Techniques for Improving Learning*. Kogan Page Publishers, 1999.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pp. 3111–3119, 2013.

[15] Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, and Yoshihiro Matsuo. Towards an entertaining natural language generation system: Linguistic peculiarities of Japanese fictional characters. In *Proc. SIGDIAL*, pp. 319–328, 2016.

[16] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, Vol. abs/1503.03832, , 2015.

[17] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *Proc. LREC*, 2002.

[18] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[19] P. Wittenburg, H. Brugman, A. Russel, A Klassmann, and H Sloetjes. Elan a professional framework for multimodality research. In *International Conference on Language Resources and Evaluation*, 2006.

[20] 水上雅博, 東中竜一郎, 川端秀寿, 山口絵美, 安達敬武, 杉山弘晃. なりきり QA データを用いた用例の拡張. 人工知能学会全国大会予稿集, No. 4J2-04, 2018.