Vol.2010-FI-97 No.10
Vol.2010-NL-195 No.10
2010/1/29

複数の対話システムからの応答候補文を用いた最適応答文選択手法の性能評価

今 井 健 太[†1]　ジェプカ ラファウ[†1]
荒 木 健 治[†1]

本稿では，ユーザの入力に応じ複数の既存システムによる応答候補を評価し適切な応答文を選択するシステムを提案する．応答文は人手で記述されたルールや既存システムがどのドメインを対象としているかといった情報を用いずに，生成された応答候補をそれぞれ複数の方法で評価することによって選択を行う．実験システムを用い，応答の選択精度評価実験，ユーザとの対話による印象評価実験を行った．その結果，選択精度実験では 55.0%の精度を，印象評価実験ではいずれの評価方法においても既存システムを単独で用いた場合や応答候補から無作為に出力した場合に比べ高い評価を得た．

Performance Evaluation of Method for Selecting Best Replies from Candidate Sentences from Different Dialogue Systems

Kenta Imai,[†1] Rafal Rzepka[†1] and Kenji Araki[†1]

In this paper, we propose the system which selects an adequate reply from response candidates generated from different dialogue systems. It is selected by analyzing the response candidates without using information on manually set rules or on domains particular system treats with.In the accuracy evalation experiment weachieved a level of 55.0% and in the impression evaluation experiment our system got the higher score than the other systems.

†1 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

1. はじめに

古くから自然言語処理分野では，多種多様な人対コンピュータの対話システムが研究，開発されてきた[1)]．また，自然言語による対話は人間が自然に扱うことのできるコミュニケーション手段である．そのため，近い将来の高度情報化社会において身の回りに無数に存在する情報機器・システムを制御するためのインタフェースとして，対話システムの活用が期待されている[2)]．

対話システムにより自然な対話を実現するには多種多様なユーザの発話に対して柔軟に自然な応答を生成することが要求されるが，現在存在するシステムでは単独であらゆるユーザの発話に自然な応答を返すことはできない．また，対話の際に想定される状況やタスクは無数にあるので，現在のように対話システムを単独で用いることであらゆるユーザの発話に自然な発話を返すことはできない．

そのため，ある限定されたタスクについては適切に応答できるが，それ以外の話題に関しては対応できない，対話を継続させることはできるが新たな話題を提供することはできない，話題を広げることはできるが意味の通じない応答を生成することがある等といった対象分野が限定されたシステムとして使用されているのが現状である．しかし，単独ではあらゆる入力に対応できなくても，複数のシステムがそれぞれの対象分野を受け持って対話を行えば多くの要求に応えることが可能であると考えられる．

1.1 関連研究

複数の状況に対応することのできる対話システムとして Ueno らのシステム[3)] や神田らのシステム[4)] がある．Ueno らのシステムは京都市の観光案内を行うシステムである．観光案内というタスクの中で，ユーザモデルや状況モデルによってプランニングを行い観光エージェント，バスエージェント，レストランエージェントの 3 つの状況に特化したエージェントを使い分ける．神田らのシステムはドメイン選択問題を，応答すべきドメインが，1 つ前の応答を行ったドメイン，音声認識結果に対する最尤のドメイン，それ以外のドメインのいずれかという判別問題ととらえ対話履歴から得られる特徴量を利用してドメインを選択するシステムである．しかし，これらのシステムはタスク指向型対話内での複数の状況に対応したものであり，自由な対話に対して応答を生成することはできない．

また，Ueno らのシステムは複数のエージェントを人手により記述されたルールを用いて使い分けている．そのため，新しいエージェントを追加するときにはエージェントごとに人手でルールを記述する必要がある．さらに，そのルールがすでに記述されているルールに干

Vol.2010-FI-97 No.10
Vol.2010-NL-195 No.10
2010/1/29

渉する場合には，すでに記述されているルールを書き換えることも必要である．

それに対して，神田らのシステムは複数のドメインの選択を抽象化した 3 つのクラスの選択問題と捉えることで複数のドメインを使い分けることができる．ただし，選択すべきドメインが不明な場合には聞き返しを必要とするなどの問題もある．また，1 つのドメインに対応させられるエキスパートシステムが 1 つに限られるため，候補となるシステムが多くなるにつれて，使い分ける複数のシステムの目的が類似している場合や，同じ状況で同じ入力がなされた場合にもそのときに各システムの出力した応答によって選択するべきシステムが変わる場合など，ドメインごとにシステムを対応させることが難しい状況が発生するという問題点がある．

1.2 本研究の目的

自由な対話は話題の限定された対話に比べて，ある話題を受けてそれと関連のある別の話題へと次々に主題が移り変わることが多い．また，雑談の中で一時的に明確な目的が生まれることによりタスク指向型対話へと変化をし，目的が達成されることで再び目的を持たない非タスク指向型対話へ戻るということが起こる場合もある．したがって自然な対話を単一のシステムのみによって行うことは難しい．また，その問題を解決するために複数のシステムを使い分ける方法としてルールやドメインによる分類を用いた場合にはルール記述やドメインに対応させるシステムの選択のコストがかかる．

そこで本稿では，ユーザの入力に対し複数の既存対話システムにより生成された応答候補の中で最も適したものをルールやどのシステムによって生成された候補かという情報を用いることなく選択することで，多様な入力に対して，より柔軟に応答を行うことのできる実験システムの開発を行った．この実験システムによる選択と人手による選択とを比較することによるシステムの選択精度評価実験とユーザと対話を行うことによる印象評価実験を行った結果について述べる．

2. システム概要

本システムの処理の流れを**図 1** に示す．本システムの処理は応答候補生成部と応答候補評価部の 2 つの部分からなる．応答候補生成部ではユーザの入力から既存の各対話システムが応答候補を生成する．応答候補評価部では，各システムにより生成された応答候補の応答文としての妥当性を様々な観点から評価する．

2.1 応答候補生成部

本システムでは応答候補を生成する既存の対話システムとして以下の 3 つのシステムを用いる．

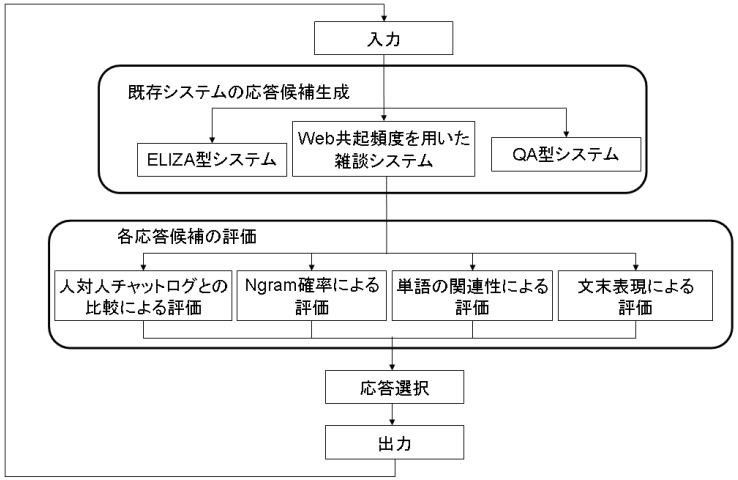

図 1 処理の流れ
Fig. 1 Flow of this system

- 木村らの研究[5] で用いられた ELIZA 型対話システム (システム A)
- 樋口らの Web による単語共起頻度を用いた雑談システム[6](システム B)
- 亀山らの研究[7] に用いられた質問応答システム (システム C)

次に各システムの詳細を述べる．

2.1.1 ELIZA 型対話システム

ELIZA は，ユーザの入力から得たキーワードとテンプレートを用いることで対話を行うシステムである．今回用いた ELIZA 型対話システムも同様にユーザ入力文中に含まれるキーワードを一時的に記憶し，用意されたテンプレートに当てはめることで応答文を生成する．

表 1 の一番目の例ではユーザの「雰囲気的には良い」という入力から「雰囲気」と「良い」というキーワードを抽出,「形容詞＋名詞＋ですか？」というテンプレートに当てはめることで応答を生成している．

2.1.2 Web による単語共起頻度を用いた雑談システム

樋口らのシステムは常に情報が更新され続ける巨大なオープンドメインである Web の利

Vol.2010-FI-97 No.10
Vol.2010-NL-195 No.10
2010/1/29

表 1 既存システムの応答例
Table 1 Example of existing system's responding.

ユーザの入力	各システムの出力
雰囲気的には良い	システム A：良い雰囲気ですか？ システム B：隠れ家でも システム C：東京っす
おはようございます。	システム A：おはようございます。 あなたの刺激的なことを話してくれませんか システム B：日本が言ってすか システム C：くらいだったし
広辞苑第五版は いつ発売されましたか。	システム A：続けてください システム B：さっそくですか システム C：今年中だがな

点を生かして，常に現在の社会が反映させることができ，ユーザに新たな情報を提供することのできる対話システムとして開発されたものである．入力文中からキーワードを抽出，得られたキーワードを用いて Web 検索を行い，キーワードとの共起頻度が高い単語を用い応答文を生成，モダリティを付与して出力する．

表 1 の一番目の例ではユーザの入力から「雰囲気」,「良い」というキーワードを抽出，それらを用いて Web 検索を行い，共起頻度の高かった「隠れ家」という単語にモダリティ「でも」を付与して応答文を生成している．

2.1.3 Web を知識源とする質問応答システム

これは入力された質問文に対してシステムがその回答を出力する質問応答システムである．亀山らのシステムは固有名詞的表現，時間表現，数値表現などを含む質問文に対応しており，それぞれ質問文に対応した回答を出力することができる．応答文は入力文中の単語の情報量を基にクエリを生成，Web 検索を行い，検索結果を用いて「名詞＋文末表現」という形で出力される．

表 1 の 3 番目の入力に対する例では「広辞苑」,「発売」などを検索クエリとして Web 検索，時間を表す「いつ」に対応する検索結果「今年中」に文末表現「だがな」を付与して応答文を生成する．しかし，2 番目の発話に対する応答例のように検索結果を得ることができなかった場合は，文末表現のみが出力される．

本来質問応答システムは自由な雑談を行えるものではない．しかし，本研究においてはこのシステムが全ての応答をする必要はなく,「名詞＋文末表現」という形の応答が不自然ではない場合や，雑談中に質問応答システムで回答することのできる質問が現れた場合にはこのシステムによる応答候補を用いることは効果が高いと考えられる．

2.2 応答候補評価部

本システムでは各システムにより生成された応答候補を次の式で表される評価値 $Eval$ によって評価する．

$$Eval = \alpha \times Sim_1 + \beta \times Sim_2 + E + N + R \quad (1)$$

ここで，Sim_1 は人対人のチャットログとの比較による評価，Sim_2 は抽象化を行いチャットログとの類似度を使用した評価である．E は文末表現の対応関係による評価，N は ngram 確率を用いた文法的妥当性による評価，R は入力と各候補と単語の関連性による評価がそれぞれ閾値を越えた場合の加点である (閾値以下の場合は 0)．また，α, β は係数である．

各評価を求める際に形態素解析が必要な場合には形態素解析器 MeCab[?)] を用いる．以下で各評価の説明を行う．

2.2.1 人対人チャットログとの類似度による評価

入力文に対して生成された各応答候補と過去に人対人で行われた対話との類似度を用いて各応答候補を評価する．本システムでは人対人の対話例として IRC (Internet Relay Chat)[?)] のログ約 19 万発話を用いている．

類似度による評価には 2 文間の表層的な類似度を評価するため，式 (2) により求められる文類似度 $Sim(A, B)$ を用いた．ここで $D(A, B)$ は文字列 A, B 間のエディット距離，$|A|, |B|$ は文字列 A, B の長さを表す．文字列 A, B 間のエディット距離とは，文字の削除と挿入のみを用いて何度の操作で文字列 A を文字列 B と同一の文字列へ変換できるかを表す値である．例として「東京都中央区」と「京都市北区」という 2 つの文字列間のエディット距離を考える．この場合,「東京都中央区」から「東」,「中」,「央」をそれぞれ削除して，残った「京都区」の「都」と「区」の間に「北」を挿入するという 4 度の操作によって「京都市北区」へと変換できる．そのため「東京都中央区」と「京都市北区」の間のエディット距離は 4 となる．

$$Sim(A, B) = 1 - \frac{D(A, B)}{|A| + |B|} \quad (2)$$

この評価方法では，まずユーザの入力と類似した発話を人対人のチャットログ中から検索し，その入力と類似した発話の次の発話，すなわち入力に類似した発話に対する人間の応答を抽出する．次に 3 つの対話システムにより生成された発話と，先ほど抽出した発話との類

Vol.2010-FI-97 No.10
Vol.2010-NL-195 No.10
2010/1/29

似度を計算する．そうすることで，実際に人間同士で行われた対話に近い候補ほど高い評価値となる．

また，式 (2) の類似度計算方法では共通する文字が含まれているという字面上の類似度を評価することはできるが，文法的・意味的な情報は考慮されない．そこで，評価する際に人対人チャットログ，入力文，応答候補文に対して名詞，動詞，形容詞を変数へ置換，さらに名詞には場所，人名，組織のような 6 つの意味的情報を付与することで抽象化を行った．この抽象化を行ったデータを用いた場合についても同様に評価を行った．

この評価方法において抽象化を行わないものを Sim_1，抽象化を行ったものを Sim_2 と呼ぶ．

2.2.2 文末表現の対応関係による評価

対話において発話への意味の付与や発話者のムードを伝える重要な要素である文末表現を用いた評価を行う．文末に現れる助詞と助動詞の組み合わせを文末表現として，前述した IRC のログ中からある発話に含まれる文末表現と次の発話に含まれる文末表現の組の抽出を行った．その結果，86,977 組の文末表現の組を抽出することができた．これは実際に人間が対話の中で使用した文末表現の組である．

入力文と応答候補との文末表現の組がこうして抽出した文末表現の組の中で出現数が閾値 t_E 以上のものに含まれている場合は，応答候補の評価値 $Eval$ に加点を行う．

2.2.3 ngram 確率を用いた文法的妥当性による評価

単語 ngram 確率を用いて各応答候補が日本語として文法的に正しいかについての評価を行う．各候補文中の全ての単語に対して 3gram 確率を計算し，その中で最も低い値，すなわち最も日本語の文としておかしな部分の 3gram 確率の値が閾値 t_N を上回った場合，応答候補の評価値 $Eval$ に加点を行う．なお，本システムでは ngram 確率の計算に Web 日本語 N グラム第 1 版[10] を使用している．

2.2.4 入力文と各候補中の単語の関連性による評価

システムが入力に対して関連した内容の応答を返すかという観点から評価を行う．評価の対象とする品詞は現在の話題に関する語が現れやすい名詞，動詞，形容詞とする．入力文 A 中に含まれる単語 a と応答候補 B 中に含まれる単語 b の Web ヒット件数を用いた共起頻度のうち最大の値を取るものを $r(A,B)$ とし，式 (3) により計算する．ここで a,b はそれぞれ A,B に含まれる単語，$N(a,b)$ は単語の組 "a b" をクエリとして検索を行った場合のヒット件数，$N(a),N(b)$ は単語 a,b をそれぞれクエリとした場合のヒット件数を表す．本システムでは Web ヒット件数を得るために検索エンジン goo[11] を用いた．

表 2 選択精度
Table 2 Selection accuracy.

	非 WH 型	WH 型	計
システム A のみ選択	63.1%	14.7%	54.1%
システム B のみ選択	32.2%	35.3%	32.8%
システム C のみ選択	10.7%	58.8%	19.7%
ランダムに選択	35.3%	36.3%	35.5%
本システム	53.4%	61.8%	55.0%
人手による選択平均	73.1%	78.8%	74.2%

$$r(A,B) = \max_{A \ni a, B \ni b} \frac{N(a,b)}{N(a)+N(b)} \tag{3}$$

この共起頻度の最大値が閾値 t_R を越えた候補の評価値 $Eval$ に加点を行う．

3. 選択精度評価実験

3.1 実 験 方 法

本システムの選択精度を評価するために，人手による選択とシステムによる選択の比較実験を行った．本実験で用いた入力は 2.1 で述べたシステム A，B との対話中のユーザの入力から WH 型の質問ではない 149 文 (非 WH 型入力)，WH 型の質問 8 文に NTCIR[12] の QAC タスク[13] の質問文から 26 文を加えた 34 文 (WH 型入力)，計 183 発話である．

これらの入力に対する既存 3 システムの応答計 549 文に対して，人手（理系大学生男子 4 名，女子 1 名）により入力文に対しての応答として適しているかについて 1（適していない）〜5（適している）の 5 段階評価を行った．

この評価では，全ての入力と入力に対する各応答候補の集合は互いに独立であり，前後の発話の文脈は考慮しない 1 ターンのみのものとした．また，評価基準は評価者が対話システムと対話を行っているときに入力と同じ発話をしたと仮定し，それに対してシステムが応答候補のように発話をしたとしたら，そのとき総合的に見てシステムのその発話は評価者にとって望ましいかという点のみで行った．そのため，評価者は応答候補に文法的に間違っている部分や，ユーザの入力とかみ合っていない部分があっても，応答文に含まれる単語が何かを連想させる場合や，かみ合っていないことが面白いと感じさせた場合に高い評価を与える場合がある．

このようにして行った評価の平均値をもとに，1 つの入力に対する 3 つの応答候補の中で

Vol.2010-FI-97 No.10
Vol.2010-NL-195 No.10
2010/1/29

システムによる評価によって最も適した応答として選択された候補と人手による評価の平均によって最も適した応答と評価された候補との一致数の調査を行った．

3.2 実験結果とその考察

本実験では 2.2 で述べた式 (1) 中の各評価基準に対する各係数，閾値，閾値を越えたときの E, N, R の値を変化させて選択精度の評価を行った．その結果，本システムが最も良い選択精度となったのは，Sim_1，Sim_2 の係数が $\alpha = 0.1, \beta = 1$，各評価基準における閾値と閾値を越えた場合の各評価基準の値は 2.2.2 で述べた $t_E = 15$，閾値を越えた場合の $E = 0.2$，2.2.3 で述べた $t_N = 1 \times 10^{-4}$，閾値を越えた場合の $N = 0.1$，2.2.4 で述べた $t_R = 0.05$，閾値を越えた場合の $R = 0.05$ とした場合であった．

そのときの本システムによって選択された応答の精度と，1 つのシステムのみを選択し続けた場合の精度，3 つの応答候補からランダムに選択した場合の精度，人手による選択を他の評価者の平均値と比較した場合の精度をまとめたものが**表 2** である．

表 2 の結果より，また，非 WH 型に対してはシステム A が，WH 型に対してはシステム C が高い精度となった．しかし，システム A は WH 型，システム C は非 WH 型に対しての精度が非常に低く，単独での使用ではそれらに対応できない．それに対し，本システムは WH 型に対しては最も高い精度，非 WH 型に対しては 50%を越える高い精度が得られ，総合的にも 55.0%と全てのシステムの中で最も高い精度を得られた．

このことから，複数の応答候補を評価して適した応答を選択することの有効性が確認された．

4. 印象評価実験

選択精度評価実験に加え，本システムにより応答を選択することによりユーザに与える印象がどのように変化するかを調べるために印象評価実験を行った．

4.1 実 験 概 要

システム A〜C を単独で使用したもの，システム A〜C の応答からランダムに出力するもの，システム A〜C の応答から本システムを用いて選択を行うものの 5 つのシステムとそれぞれ 20 ターン以上の自由な対話を行い，それぞれ評価を行った．

被験者は 9 名（10 代女性 1 名，20 代男性 6 名，20 代女性 2 名，うち 10 代女性は文系大学生，20 代男性 2 名は理系大学生，20 代男性 4 名と 20 代女性 1 名は理系大学院生），評価は**表 3** で示す 6 つの評価基準についての 1〜5 の 5 段階評価，**表 5** で示す 20 項目の形容詞対についての 7 段階尺度の評定尺度法による評価の 2 種類を行った．評定尺度法による評価に用いた 20 の形容詞対は，SD 法[14] においてよく用いられる形容詞対から選出した．以下では各形容詞対についての 7 段階尺度（非常に・かなり・やや・どちらでもない・やや・かなり・非常に）の評定をポジティブな形容詞（表 5 にある形容詞対の右の語）が高くなるように 1 から 7 まで数値化して分析する．

また，式 (1) 中の各評価基準に対する係数，閾値，閾値を越えたときの加点は選択精度評価実験において最も高い精度となった組み合わせとした．

表 3 評価項目
Table 3 Evaluation items.

(A) 対話を続けたいかどうか.
(B) 対話が文法的に自然であるか.
(C) 対話が意味的に自然であるか.
(D) システムの語彙が豊富かどうか.
(E) システムが知識を持っているように感じるかどうか.
(F) システムが人間らしいか.

表 4 各システムの評価値
Table 4 Evaluation value of each system.

	5 段階評価	評定尺度法
システム A	2.54	3.46
システム B	2.60	3.79
システム C	2.90	3.37
ランダムに選択	3.06	4.10
本システムにより選択	3.32	4.42

表 5 評定尺度法に用いる形容詞対
Table 5 Adjective pair used for rating scale method.

こわい	やさしい
わかりにくい	わかりやすい
退屈な	興味深い
感じの悪い	感じの良い
静的な	動的な
近づきにくい	近づきやすい
古い	新しい
陰気な	陽気な
親しみにくい	親しみやすい
消極的な	積極的な
つまらない	面白い
単純な	複雑な
嫌いな	好きな
わがままな	思いやりのある
空虚な	充実した
愚かな	賢い
にくらしい	かわいらしい
苦しい	楽しい
冷たい	暖かい
機械的な	人間的な

4.2 考　　察

各項目の平均点をまとめたものが**表 4**，項目ごとにグラフに表したものが**図 4.2**，**図 4.2** である．

各項目の平均点においては 5 段階評価，評定尺度法いずれも本システムが最も高い評価を得た．評価項目の中でも対話システムの評価において重要だと考えられる「対話を続けたいかどうか」,「苦しい ⇔ 楽しい」,「嫌いな ⇔ 好きな」ではそれぞれ 2 番目に評価の高かったシステムを 0.37 ポイント，0.67 ポイント，0.67 ポイント上回った．また，提案手法

Vol.2010-FI-97 No.10
Vol.2010-NL-195 No.10
2010/1/29

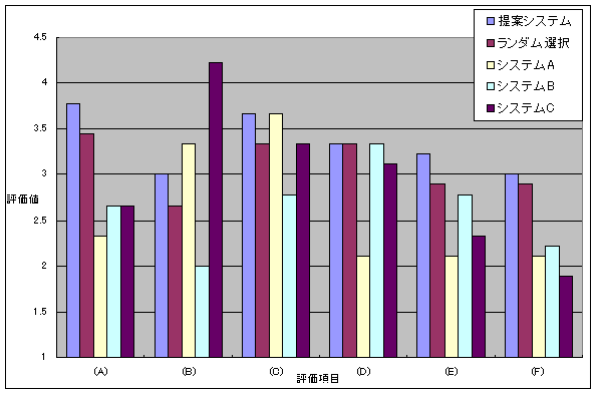

図 2 5 段階評価による評価値平均
Fig. 2 Evaluation value average by five stage evaluation.

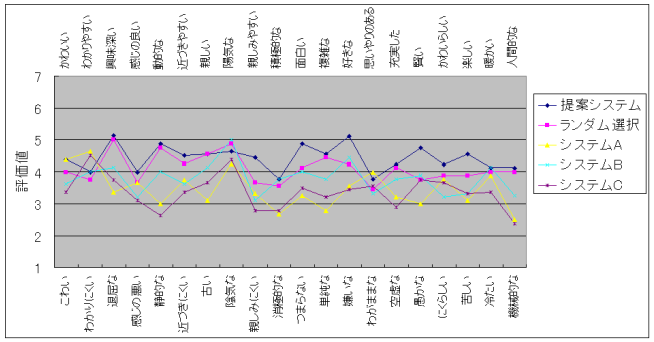

図 3 評定尺度法による評価値平均
Fig. 3 Evaluation value average by rating scale method.

によって適切な応答を選択することで向上すると考えられる「システムが知識を持っているように感じるかどうか」,「愚かな ⇔ 賢い」ではそれぞれ 2 番目に評価の高いシステムを 0.37 ポイント，0.87 ポイント上回った．特に,「親しみにくい ⇔ 親しみやすい」,「嫌いな ⇔ 好きな」,「苦しい ⇔ 楽しい」の 4 つの項目について，5%の有意水準において t 検定を行ったところ，有意差が認められた．これにより，提案手法を用いて複数の応答候補から出力を選択することで被験者に与える印象を向上させられることが確認された．

しかし，5 段階評価においては「対話が文法的に自然であるか」が 1.22 ポイント，評定尺度法による評価においては「わかりにくい ⇔ わかりやすい」が 0.63,「陰気な ⇔ 陽気な」が 0.37 ポイント,「わがままな ⇔ 思いやりのある」が 0.22 ポイント，それぞれ最も評価の良かったシステムを下回った．以下にそれらの本システムの評価が他のシステムを下回った項目についての考察を行う．

- (B) 対話が文法的に自然であるか．

システム A は人手によって作られたヒューリスティックなルールを用いて応答を生成している．システム C は「名詞 + 文末表現」という，極めて単純な形の応答を生成する．そのため，これら 2 つのシステムの応答は文法的に不自然なものが生成されることは少なく，この項目の評価が高くなった．

一方，本システムでは文法的妥当性を評価項目の一つしている．しかし，この基準は最優先されるものではないため，文法が自然なものが必ずしも選択されるわけではない．また，人手による評価も文法的に自然なものが高評価だとは限らない．選択精度評価実験に用いた 549 の応答文中 138 文は文法的に不自然な表現であるにもかかわらず，人手による評価平均が中間地である 3 を越えている．

- わかりにくい ⇔ わかりやすい

この項目については，システムのどの部分についてのわかりやすさについての評価かを明記していない．そのため，被験者の中にシステムの発話内容がわかりやすいかという評価を行ったものとシステムの応答文の生成方法が推測しやすいかという評価を行ったものがいた．後者の基準で評価を行ったものは，この項目の評価が「単純な ⇔ 複雑な」の項目と近い結果となった．その結果，入力文から単語を引用しすることでおうむ返しを生成するシステム A や常に「名詞+文末表現」という単調な形式の応答を生成するシステム C がわかりやすいという評価を受け，応答に多様性のある本システムやランダムに応答を選択するシステムがわかりにくいという評価を得た．

- 陰気な ⇔ 陽気な

提案手法では文末表現による評価を行っている．この評価により丁寧な口調の入力に対しては丁寧な口調の応答が選択されやすくなる．また，被験者は全体的に丁寧な口調で対話を行う傾向にある．そのため，その点を考慮していないランダム選択やシステム B に比較して陰気だという評価となった．本システムよりも他のシステムに高い評価をつけた被験者の

Vol.2010-FI-97 No.10
Vol.2010-NL-195 No.10
2010/1/29

対話ログを見ると，本システムが丁寧語を用いて応答を行った割合が全体の 37.5%だったのに対し，ランダムに選択した場合は 10%，システム B は 5%のみだった．

- わがままな ⇔ 思いやりのある

システム A は単体ではユーザの入力を引用した発話を行い，システムの側から別の話題に転換することもない．そのため，被験者は話を聞いてくれているように感じ，思いやりがあると評価された．しかし，システム A のこの項目に高い評価値を付けた被験者は，総合的には本システムが最も良いと評価している．そのため，この項目の評価が低いことがそのままシステムの評価が低いということにはつながっていないと考えられる．

5. おわりに

本稿では，3 つの既存対話システムにより生成された応答候補を評価することで，ルールを用いずに最適な応答文を選択する手法の提案を行った．評価値を用いて選択を行うことで，各々の既存システムの対象としていない分野の応答に対して大きく精度が上回り，入力全体に対しては全ての既存システムを上回る精度が得られた．

また，ユーザとの対話による印象評価実験では既存の各システムを単独で用いたものや応答候補からランダムに選択を行ったものに比べ，対話の印象に大きくかかわる項目において高い評価を受けた．これらの結果より，ユーザの入力によって複数のシステムを使い分けることが非タスク指向型の対話システムにおいて有効であると考えられる．

本稿では 3 種類のシステムによる応答候補から適した応答の選択を行った．しかし，日常的な会話の全てをカバーするためには 3 種類のシステムだけではなく，さらに多くのシステムの応答候補文から最適な応答文を選択する必要がある．

また，人間が対話の中でどのような応答をするかを決定する際や相手の応答を評価する際には今回用いた評価方法だけでは十分ではなく，周囲の状況や文脈情報，対話者についての情報，感情やユーモアなど多くの情報が複雑に作用していると考えられる．

この問題を解決するためには，日常的な会話をすべてカバーするほど多くの応答候補文から選択を行う際には，今回用いた 4 種類の評価方法だけでは正確な選択を行うには十分ではないと考えられる．そのため，現在より多くの情報を用いることでより的確な選択を行うことが必要となる．今後はその点を考慮し，より多くの対話システムを用い，新しい評価尺度を使用した選択を行う予定である．

参 考 文 献

1) J.Weizenbaum, ”ELIZA - A Computer Program for the Study of Natural Language Communication Between Man and Machine”, Communications of the Association for Computing Machinery, Vol.9, pp.36-45, 1966.
2) Wolfgang Minker, Ramon Lopez-Cozar and Michael McTear, The role of spoken language dialogue interaction in intelligent environments, Journal of Ambient Intelligence and Smart Environments 1 (2009) 31.36.
3) S.Ueno, I.R.Lane, and T.Kawahara.: ”Example-based training of dialogue planning incorporation user and situation models”, Proc. ICSLP, pp.2837-2840, 2004.
4) 神田 直之，駒谷 和範，中野 幹生，中臺 一博，辻野 広司，尾形 哲也，尾形 哲也，奥乃 博，マルチドメイン音声対話システムにおける対話履歴を利用したドメイン選択，情報処理学会論文誌，pp.1980-1989 20070515
5) 木村泰知，荒木健治，桃内佳雄，栃内香次，” 遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法，”電子情報通信学会論文誌，Vol.J84-D-2 No.9 pp.2079-2091, 2001.
6) Shinsuke Higuchi, Rafal Rzepka and Kenji Araki: ”A Casual Conversation System Using Modality and Word Associations Retrieved from the Web”, Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 382-390, Honolulu, USA, October 2008.
7) 亀山恵祐、荒木健治: ”質問応答システムにおける知識源選択規則の自動獲得の有効性について ”, 情報処理学会研究報告 2006-NL-178, pp.85-90, 2007.
8) 工藤拓，MeCab: Yet Another Part-of-Speech and Morphological Analyzer, 2006.
9) IRC. http://www.irchelp.org/irchelp/rfc/rfc.htm.
10) 工藤拓、 賀沢秀人著、「Web 日本語 N グラム第 1 版」、言語資源協会発行
11) goo, http://www.goo.ne.jp/.
12) NTCIR. http://research.nii.ac.jp/ntcir/.
13) QAC. http://www.nlp.is.ritsumei.ac.jp/qac/.
14) 末永俊郎，社会心理学研究入門，東京大学出版会，1987.