人工知能学会研究会資料
SIG-SLUD-B402-06

雑談対話中の発話文に対する多面的評価の分析

Multi-aspect Evaluation for Utterances in Chat Dialogues

杉山 弘晃 [1*]　目黒 豊美 [1]　東中 竜一郎 [1,2]
Hiroaki Sugiyama [1]　Toyomi Meguro [1]　Ryuichiro Higashinaka [1,2]

[1] NTT コミュニケーション科学基礎研究所
[1] NTT Communication Science Laboratories
[2] NTT メディアインテリジェンス研究所
[2] NTT Media Intelligence Laboratories

Abstract:
The evaluation measures for chat-oriented dialogue systems are required in order to effectively improve such systems. Some studies have evaluated systems with several arbitrarily defined measures; however, it is not examined whether their measures are appropriate. We analyze evaluation measures for chat-oriented dialogue systems through the semantic differential. Our analysis shows that evaluation measures are clustered into four factors for each evaluator. The factors consist of two common factors, one resemble factor between evaluators, and one personal factor. We also develop an automatic evaluation system that estimates each evaluation measure defined in the semantic differential. Our experiment shows that the developed system estimates most of the scores with the similar correlation coefficients as between human evaluators.

1　序論

近年，従来のタスク指向の対話システムとは異なる，雑談を行う対話システムに注目が集まっている [大西 14, Ritter 11, Wong 12, 東中 14]．雑談対話は，エンタテインメントやカウンセリング目的のみならず，ユーザの潜在的な要求を引き出したり，ユーザと良好な関係を構築する上で重要である．

雑談対話システム研究を進める上での課題の一つが，評価尺度の設計である．タスクの遂行を目的とする対話システムでは，タスクの達成率や達成にかかる時間など，客観的に測定可能な評価尺度が考えられる [Janarthanam 08, Walker 97]．一方，雑談対話システムは明確な目的を持たないため，こうした客観的な評価尺度の設定は難しい．それゆえ，雑談対話システムの評価には，ユーザ満足度などの主観的な尺度が用いられてきた [Sugiyama 13, Meguro 10]．しかし，従来の研究で用いられてきた尺度は設計者によって恣意的に設定されたものであり，どのような尺度が雑談対話システムの評価に必要かを詳細に分析した例は知られていない．雑談対話システムの評価に必要な尺度を分析し，かつ各尺度の評価値を自動的に推定することで，効率よくシステムを改善することが可能になる．

そこで本研究ではまず，人が雑談対話中の発話を評価する際に注目する尺度を，SD(Semantic Differential) 法によって明らかにする．SD 法とは，評価対象に対して複数の形容詞対を提示し，各形容詞対のどちらに近い印象を受けるかを評価することで，評価対象についての印象を詳細に分析する手法である．なお本研究では，ある単文を入力とした場合の応答文を分析の対象とする．応答文について 2 名の評価者が評価し，評価者間の相違点や，入力文ごとの評価尺度の変動について報告する．また同時に，SD 法で付与した各形容詞に対する評価値の自動推定システムを構築し，推定精度と課題について述べる．

章立ては以下の通りである．2 章で関連研究を示し，3 章で応答文の作成と評価値の付与について説明する．4 章で得られた評価値の分析を，5 章で評価値の自動推定の実験を説明し，6 章で結論を述べる．

2　関連研究

対話中の発話の主観的な印象を分析する試みとして，音声特徴 [Nishimura 08] やターンテイキングの情報 [Nishimura 07] など，非言語情報が及ぼす影響に着目して分析したものが多い．例えば Maat ら [Maat 10] は，システムのターンテイキングの戦略がユーザに与える印象を，SD 法を用いて分析している．彼らは因子分析を用いて，SD 法で評価を付与した 27 種類の形容詞対を 8 種類の因子に分類し，そのうち 5 つ（agreeableness, assertiveness, conversational skill, rapport,

*連絡先：NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
E-mail: sugiyama.hiroaki@lab.ntt.co.jp

rude-respectful）がターンテイキングの戦略と関連することを明らかにしている．また Walker らは，システムと人の間でなされた対話に対し，対話から得られる発話文の長さや発言数などの特徴量について，対話の質を従属変数として重回帰分析することで，タスク対話システムの評価に有用な要素を明らかにしている [Walker 97]．このように，従来の対話システムの主観的な印象を分析する研究では，非言語情報に力点が置かれており，言語的な発話内容と印象の関係性を詳細に分析した例は知られていない．

一方近年，テキストチャット上の雑談対話を扱う研究が増加している [Ritter 11, Higashinaka 14]．テキストチャットでは，韻律やターンテイクの情報が得られないため，発話の言語的な情報に着目して評価を行う必要がある．例えば Sugiyama らは，対話の流れの自然さや対話の有用性のような，主観的な尺度を設定し，システムの評価を多面的に行っている [Sugiyama 13]．しかしながら，それぞれの尺度は設計者の主観で選ばれており，尺度の妥当性については考察されていない．

一方，今井らは，SD 法を用いてシステム単位の印象分析を行っている [今井 10]．彼らは，各形容詞対について付与された評価値の平均値とシステムの定性的な特性との比較を行っている．しかしながら，評価者ごとの共通点や違いについては簡易な分析にとどまっている．また，重視される評価尺度は文単位で変動すると予想されるものの，彼らは対話単位で評価しているため，文単位の評価尺度の変動については考慮されていない．

3　応答文の作成と評価値の付与

本研究では，雑談対話システムが生成する発話について，評価者ごとの共通点や違いと，入力文ごとの評価尺度の変動について分析する．ここでは，少数の入力文について，多数の応答文を比較・分析することで，応答文の印象の要因を分析する．

本研究では，多様な応答文の収集にあたり，杉山らの方法 [杉山 14] を採用する．これは，多数の応答文作成者が，意図的に負例を含む文を作ることで，入力文に対する多様な応答文を収集する方法である．また，各応答文の評価は，総合的な評価と，各 SD 法の形容詞対の評価の 2 種類を行う．本章では，入力文と応答文の作成方法，得られた応答文に対する評価値の付与方法について説明する．

3.1　入力文の収集と応答文の作成

本研究で用いる入力文は，その文単体を見て応答文を作成する必要があるため，文単体で何についての発話であるかが理解できるように書かれていることが求められる．そのため本研究では，Web や対話実験ログから人が記述した発話文を収集し，これらに対して理解しやすさ (了解性) を人手で付与することで，文単体で理解しやすい，了解性が高い文を集める．

次に，各入力文に対して応答文を作成する．このとき，応答文の多様性を確保するため，負例となりうる文を含むように文作成方法に制約をかけた状態で，複数の作成者が文を作成する．文作成の制約として，本研究では，応答文の文字数制約と，入力文のマスクを行う．文字数を制約することで使える表現が制約され，応答文に不自然な表現が含まれる効果が期待できる．また，入力文の一部を隠して作成者へ提示することで，入力文と話題が異なる文が得られると考えられる．

人手で作成した応答文に加えて，検索ベース対話システムやルールベース対話システムなど，既存の対話システムから得られた発話を応答文に加える．現在の対話システムは必ずしも適切な応答を返せていないため，負例と正例が適度に混在した文が得られると予想される．

3.2　評価値の付与

本研究では，各応答文に対し，総合的な評価と，各 SD 法の形容詞対の評価の 2 種類を付与する．作成した応答文の集合に対し，まず総合評価値を人手で付与する．本研究では,「応答文としての自然さ」を評価基準として文のペアごとに優劣を全て人手で評価した場合の，全ての応答文に対する勝率をある応答文の総合評価値とする．すなわち，全ての応答文に対して自然であると判断された応答文は評価値として 1 が付与され，逆に全ての応答文よりも不自然であると判断された文は 0 が付与される．ペアワイズの優劣から得られた勝率は 0 から 1 の間で満遍なく分布しており，最低値の 1 や最高値の 7 と評価されるものの中でも優劣を付けられるという利点がある．一方，ペアごとに優劣を付与すると，評価回数が非常に多くなってしまう問題がある．しかし，本研究では全てに対し評価を付与するが，10%程度のペアのみをサンプリングして評価しても，全体の傾向にはあまり影響しないことがわかっている [杉山 14, Sculley 09]．また，Likert 尺度で付与した自然性との相関が高ければ，Likert 尺度で代用することも考えられる．

総合評価値を付与した後，同一の評価者が，SD 法で用いる各形容詞対について Likert 尺度で評価値を付与する．このとき，総合評価値の評価基準として用いた「応答文としての自然さ」も Likert 尺度で改めて付与し，ペアワイズの勝率との相関について調べる．用いた形容詞対の詳細は，4.1.3 節に後述する．

4　得られた評価値の分析

本章では，収集された文の総合評価値（勝率）と SD 法で付与した各尺度の評価値の関係性について，評価者間での共通点と相違点について分析する．

表 1: 用いた形容詞対，評価者間の相関係数，κ 値，評価者ごとの勝率と各評価尺度の相関

ID	低スコア	高スコア	評価者間相関	κ	勝率との相関 (A)	勝率との相関 (B)
	勝率		0.79	-	-	-
z)	応答文としての自然さ		0.73	0.24	0.86	0.88
a)	こわい	やさしい	0.26	0.12	0.19	0.07
b)	わかりにくい	わかりやすい	0.70	0.19	0.80	0.85
c)	退屈な	興味深い	0.34	0.09	0.30	0.09
d)	感じの悪い	感じの良い	0.49	0.24	0.41	0.21
e)	静的な	動的な	0.39	0.19	0.11	-0.01
f)	近づきにくい	近づきやすい	0.42	0.17	0.51	0.20
g)	古い	新しい	-0.0	-0.0	0.01	-0.01
h)	陽気な	陰気な	0.53	0.31	0.08	0.01
i)	親しみにくい	親しみやすい	0.47	0.19	0.46	0.15
j)	消極的な	積極的な	0.49	0.25	0.31	0.10
k)	つまらない	面白い	0.37	0.16	0.06	-0.00
l)	単純な	複雑な	0.39	0.17	0.01	-0.18
m)	嫌いな	好きな	0.17	0.10	0.14	0.14
n)	わがままな	思いやりのある	0.30	0.20	0.17	0.15
o)	空虚な	充実した	0.17	0.05	0.14	0.05
p)	愚かな	賢い	0.13	0.06	0.13	0.18
q)	にくらしい	かわいらしい	0.24	0.17	0.04	0.00
r)	苦しい	楽しい	0.24	0.11	0.02	-0.02
s)	冷たい	温かい	0.59	0.09	0.18	0.06
t)	機械的な	人間的な	0.73	0.24	0.64	0.68

4.1 実験設定

4.1.1 入力文の収集

入力文を収集するコーパスとして，本研究では，我々が収集した雑談対話コーパス [Higashinaka 14] と Twitter コーパスを用いる．雑談対話コーパスは，のべ 360 名以上の話者から，1 対 1 のテキストチャット形式による雑談を，計 3680 対話，約 13 万文収集したものである．これに，目黒ら [Meguro 10] によって定義された対話行為を付与し，自己開示（自分についての事実や経験などを話した発話），質問，もしくは情報提供に関する対話行為が付与された文を，入力文の候補として抽出した．一方，Twitter コーパスから了解性を高い文を容易に収集する方法として，話題となりうる単語を含む文を Twitter から検索し，そのうち非文でないものをルールで抽出する，稲葉らの方法 [稲葉 14] がある．本研究ではこれを参考に，話題となりうる単語（Google trends 2012 in Japan[1]の各カテゴリで 10 位以上の単語句のうち,「Xperia acro HD」などのように，空白を含まないもの）を含むおよそ 1 億 5 千万ツイートを入力文の候補として抽出した．収集した入力文候補について，筆者ら以外の 1 名のアノテータが，5 段階の Likert 尺度で内容の了解性を付与した．そのうち，最良値の 5（内容は省略なく明確に記述されている）を得た文から，コーパスごとに 5 文をランダムに選び入力文とした．

[1]https://www.google.co.jp/trends/topcharts#date=2012

4.1.2 応答文の作成

各入力文に対し，10 名の応答文作成者が，自由に 3 文，10 文字以上の文を 3 文，10 文字未満の 1 文を 1 つ，計 7 文作成した．このとき，自由入力を含めて文字数は 50 文字以内とした．また，対話中の発話であることを意識し，話を続けたくなるように作成するように指示した．

本研究では，負例の応答文を作成するため，上記の文字数制限に加え，入力文の一部を文節単位でマスクして応答文作成者へ提示する．例えば,「何か得意なものはありますか?」という入力文の文節の 60%をマスクする場合,「なにか *** *** ありますか?」のように作成者へ提示される．ここでは，マスクしないものを 6 つ, 全体の 30%をマスクしたものを 2 つ, 60%をマスクしたものを 2 つ用意した．これらをランダムに 10 名の応答文作成者に割り当て，マスク 1 つあたり 1 つの文節が入ることと，そこを想像しながら応答文を作成することを作成者に指示した．以上より，1 つの入力文に対し，マスク無しの $7 \times 6 = 42$ 文，30%マスクの $7 \times 2 = 14$ 文，60%マスクの $7 \times 2 = 14$ 文の計 70 文が得られる．

さらに，人手で作成した応答文に加えて，Ritter らが提案した検索ベースの発話生成手法である，IR-status, IR-response からそれぞれ 10 文，我々が開発したルールベース対話システムから 10 文収集した [Ritter 11, 目黒 14]．IR-status とは，Twitter から入力文に類似したツイート (status) を検索し，in-reply-to 機能で対応付けられた返信ツイート (response) をシステムの発話文として出力する手法である．IR-response は，Twitter 中の返信ツイート (response) から直接入力文に類似する文を検索する手法である．ルールベース対話システムは，入力文との一致を調べるパターンとそれに紐づいた出力文のペアを人手で記述したシステムである．パターンの検索には TF-IDF で重み付けた単語のコサイン類似度を用い，類似度が高い 10 文を応答文へ追加した．最終的に得られた応答文集合は，1 入力文あたり，人手 70 文，検索ベース 20 文，ルールベース 10 文の計 100 文である．

4.1.3 評価値の付与

本研究では，4.1.1 節で得られた 10 個の入力文と応答文集合のペアについて，2 名の評価者が勝率と SD 法の形容詞対の評価値を付与した．ここで 10 入力文のみを対象とした理由は，1 入力文に対する応答文の数が 100 と大きく，ペアワイズの勝ち負けの評価回数が 1 入力文につき 4950 回，加えて SD 法の形容詞対についての付与数が 2000 回と膨大になってしまうためである．

本研究では，SD 法で用いる形容詞対として，非タスク指向型対話システムを SD 法で分析した今井らの研究 [今井 10] によって定義された，20 種類の形容詞対を用いる．この形容詞対を用いた理由は，本研究で対象

表 2: 因子分析によって得られた因子

表 3: a) 評価者 A

Factors (寄与率)	低スコア		高スコア	因子負荷量	Factor4 との相関
Factor 1 (0.164)	親しみにくい	↔	親しみやすい	0.87	0.41
	近づきにくい	↔	近づきやすい	0.79	
	感じの悪い	↔	感じの良い	0.74	
	冷たい	↔	温かい	0.68	
	わがままな	↔	思いやりのある	0.53	
Factor 2 (0.140)	静的な	↔	動的な	0.84	0.28
	消極的な	↔	積極的な	0.71	
	つまらない	↔	面白い	0.71	
	退屈な	↔	興味深い	0.53	
Factor 3 (0.105)	嫌いな	↔	好きな	0.92	0.18
	愚かな	↔	賢い	0.67	
	にくらしい	↔	かわいらしい	0.66	
Factor 4 (0.097)	応答文としての自然さ			0.88	1.00
	わかりにくい	↔	わかりやすい	0.87	
	機械的な	↔	人間的な	0.52	

表 4: b) 評価者 B

Factors (寄与率)	低スコア		高スコア	因子負荷量	Factor3 との相関
Factor 1 (0.175)	冷たい	↔	温かい	0.78	0.29
	感じの悪い	↔	感じの良い	0.78	
	近づきにくい	↔	近づきやすい	0.77	
	こわい	↔	やさしい	0.76	
	親しみにくい	↔	親しみやすい	0.73	
	わがままな	↔	思いやりのある	0.71	
Factor 2 (0.179)	退屈な	↔	興味深い	0.89	0.21
	単純な	↔	複雑な	0.85	
	つまらない	↔	面白い	0.82	
	空虚な	↔	充実した	0.77	
Factor 3 (0.118)	わかりにくい	↔	わかりやすい	1.00	1.00
	応答文としての自然さ			0.90	
	機械的な	↔	人間的な	0.76	
Factor 4 (0.118)	陰気な	↔	陽気な	0.88	0.06
	苦しい	↔	楽しい	0.77	
	静的な	↔	動的な	0.54	

とする雑談対話システムを包含する非タスク指向型対話システムを対象としており，かつシステムの定性的な傾向を反映するだけの表現力を持っているためである．本研究では，これに「応答文としての自然さ」を加え，7 段階の Likert 尺度で評価値を付与する．

4.2 全入力文を通した分析

用いた形容詞対の一覧と，評価者間の相関係数，κ 値，および各評価者が付与した，各評価尺度と勝率の相関係数を表 1 に示す．勝率や応答文としての自然さは 0.7 以上の高い相関係数を示していたものの，それ以外の尺度は概ね 0.5 以下と，中程度もしくは弱い相関となっていた．また κ 値も，最大でも 0.24 程度と，中程度以下の一致率を示している．総合的な評価は評価者間でおおよそ一致するものの，発話から想起する形容詞は評価者によって異なっていると言える．各評価尺度と勝率の相関係数を調べると，評価者に共通している特徴として，z) 応答文としての自然さ，b) わかりやすい，t) 人間的なの 3 尺度が，勝率と強い相関を得ている．z) 応答文としての自然さは，勝ち負けを付与する際の基準として用いた尺度であるため，評価者の違いに依らず，勝率と強く相関していたと考えられる．また，b) わかりやすさ や t) 人間的な のような，そもそも応答文として成立しうるかを表す尺度が勝率と強い相関を示していた理由として，本研究で用いた応答文が負例を多数含んでいるためと考えられる．

一方，評価者間で異なる特徴として，d) 感じの良さ，f) 近づきやすさ，i) 親しみやすさ が得られた．これらの尺度は，評価者 A のみが高い相関係数を示しており，評価者 B では弱い相関となっていた．評価者 A が，入力文に依らず，親近感を感じる発話を好む一方，評価者 B はこれらの尺度をあまり重要視していない，もしくは入力文によって反応が異なることが示唆されている．

次に，SD 法で得られた評価値を解釈するため，因子分析を行う．因子数は，MAP（最小平均偏相関）に基づき，4 と定める．表 2 に，各因子に関連する形容詞対（因子負荷量が 0.5 以上）と，自然さを含む因子との相関を示す．評価者間で共通している因子として，親近感に関する因子（評価者 A の Factor 1，評価者 B の Factor 1），および自然性に関する因子（評価者 A の Factor 4，評価者 B の Factor 3）が得られている．また，評価者間で類似しているもののやや異なる因子として，評価者 A の Factor 2 ではアクティブさを面白いと評価する因子が，評価者 B の Factor 2 では複雑さを面白いと評価する因子が得られている．一方，評価者間で異なる因子として，評価者 A の Factor 3 では賢さやかわいらしさを評価する因子が，評価者 B の Factor 4 では陽気さを評価する因子が得られているこれらの相違点は，評価者の好みと関連していると考えられる．

4.3 入力文ごとの分析

より詳細に分析するため，入力文ごとに相関分析を行い，勝率との相関係数が全入力文に対する結果と有意に異なる相関係数を示した尺度を調べる．結果を表 5 に示す．A は B に比べて入力文による変動が少なく，概ね入力文に関連する評価尺度の相関が強まっている例が抽出されている．一方 B では，入力文による評価尺度の変動が大きく，全入力文における各評価尺度の相関が小さかった一因と考えられる．特に，s7 や s9 のような，応答に困るような入力文において，複雑さや陽気さに関連する評価尺度と逆相関が見られた．これらについてより詳細に調べると，複雑さが 5 以上の文はほとんど見られず,「おやすみ」のような非常に単純な文に，複雑さとして 1 や 2 が付与されているケースが多かった．すなわち，こうした入力文においては，単純な文がより自然であると判断されている．

表 5: 各入力文について，全体の相関係数と有意に異なる相関係数を示した評価尺度 ($p < 0.05$)．**太字**は変動した方向の形容詞を表す

ID	入力文	評価者 A	相関 (全体)	評価者 B	相関 (全体)
s1	そして、ディズニーランドの大混雑も苦手です … 。	**機械的な** ↔ 人間的な	0.38(0.64)	**苦しい** ↔ 楽しい	-0.31(-0.02)
				親しみにくい ↔ 親しみやすい	-0.12(0.15)
				わがままな ↔ **思いやりのある**	0.39(0.15)
				応答文としての自然さ	0.92(0.88)
s2	なにかスポーツをされていますか？	空虚な ↔ **充実した**	0.48(0.14)	機械的な ↔ **人間的な**	0.89(0.68)
		応答文としての自然さ	0.93(0.86)	応答文としての自然さ	0.94(0.87)
				空虚な ↔ **充実した**	0.38(0.05)
				つまらない ↔ **興味深い**	0.40(0.09)
				単純な ↔ **複雑な**	0.14(-0.18)
				わかりにくい ↔ **わかりやすい**	0.91(0.85)
				つまらない ↔ **面白い**	0.25(-0.00)
				愚かな ↔ **賢い**	0.40(0.18)
s3	LAWSON 寄ったらいきなり紅蓮の弓矢流れて噴いた	静的な ↔ **動的な**	0.45(0.11)		
s4	ゴテゴテしいポッキーしか売ってないので別のコンビニ行こう				
s5	iTunes に入ってるの確認したらアニソンとゲーソンとボカロと声優さんとドラマ CD だらけだった	消極的な ↔ **積極的な**	0.51(0.31)	消極的な ↔ **積極的な**	0.35(0.10)
s6	何か得意なものはありますか？	応答文としての自然さ	0.95(0.86)	**機械的な** ↔ 人間的な	0.49(0.68)
		わかりにくい ↔ **わかりやすい**	0.90(0.80)		
s7	文字の攻撃力というものを理解しておかねばならない	わかりにくい ↔ **わかりやすい**	0.89(0.80)	**空虚な** ↔ 充実した	-0.28(0.05)
		機械的な ↔ **人間的な**	0.78(0.64)	**陰気な** ↔ 陽気な	-0.31(0.01)
				単純な ↔ 複雑な	-0.42(-0.18)
				つまらない ↔ 面白い	-0.26(-0.00)
				苦しい ↔ 楽しい	-0.27(-0.02)
				応答文としての自然さ	0.92(0.88)
				嫌いな ↔ 好きな	-0.09(0.14)
s8	和菓子は食べられますか？	**わかりにくい** ↔ わかりやすい	0.69(0.80)	空虚な ↔ **充実した**	0.31(0.05)
s9	日本各地が安定した天気になるよう願いまして、おやすみ	**退屈な** ↔ 興味深い	-0.07(0.30)	**退屈な** ↔ 興味深い	-0.46(0.09)
				つまらない ↔ 面白い	-0.49(-0.00)
				単純な ↔ 複雑な	-0.55(-0.18)
				空虚な ↔ 充実した	-0.35(0.04)
				こわい ↔ **やさしい**	0.35(0.07)
				苦しい ↔ **楽しい**	0.27(-0.02)
				親しみにくい ↔ **親しみやすい**	0.41(0.15)
				冷たい ↔ **温かい**	0.31(0.06)
				愚かな ↔ 賢い	-0.05(0.18)
s10	自分も妹がいますが、気づかずに同じ漫画を買ってきちゃうことはあります。			苦しい ↔ **楽しい**	0.23(-0.02)
				にくらしい ↔ **かわいらしい**	0.23(0.00)

5 多面的評価の自動推定

発話がユーザに与える印象を自動的に推定することができれば，ユーザの感情を考慮した対話システムの改良を効率よく進めることができる．本章では，SD 法の形容詞対に付与した評価値を自動的に推定するシステムを構築し，その推定評価値と人手で付与した評価値の相関を検証する．

5.1 実験設定

この問題は，入力文と応答文のペアが与えられた場合に，その応答文がどのような印象をユーザに与えるかについて自動的に推定する問題である．本実験では，入力文と応答文のペアは 4.1 節で説明したものを利用し，印象の要素には，SD 法で分析する際に用いた各形容詞対の評価値を用いる．すなわち，推定システムの入力は入力文と応答文のペア，出力は各形容詞対の評価値の推定値となる．

本実験では，評価値を推定する手法として，雑談対話システムの自動評価を目的とした杉山らの研究 [杉山 14] において，最も良い性能を示した手法を用いる．これは，評価対象の応答文とそれ以外の多数の文との類似度を特徴量とし，Support Vector Regression[Smola 04] を用いて応答文の評価値を推定する手法である．本実験では，各応答文 x_i とそれ以外の応答文 $x_j; i \neq j$ との類似度 $s_{i,j}$ を特徴量とし，x_i に付与された評価値 t を正解とする教師データで Support Vector Regression のパラメータを学習する．また，文 x_i と x_j の類似度 $s_{i,j}$ として，Word Error Rate（WER）$e_{i,j}$ に基づく類似度を定義する．WER は Normalized Levenshtein 距離を用いて計算し，$s_{i,j} = 1 - 2e_{i,j}$ に基づいて-1 から 1 の範囲を取る類似度へ変換して用いる．

表 6: 用いた形容詞対，評価者間の相関係数，評価者 A，B との相関（下線は $p < 0.05$ を示す）

ID	低スコア	高スコア	評価者間相関	A との相関	B との相関
	勝率		0.79	0.54	0.49
z)	応答文としての自然さ		0.73	0.54	0.53
a)	こわい	やさしい	0.26	0.19	0.22
b)	わかりにくい	わかりやすい	0.70	0.46	0.50
c)	退屈な	興味深い	0.34	0.37	0.42
d)	感じの悪い	感じの良い	0.49	0.38	0.31
e)	静的な	動的な	0.39	0.41	0.43
f)	近づきにくい	近づきやすい	0.42	0.40	0.30
g)	古い	新しい	-0.0	-0.07	0.14
h)	陰気な	陽気な	0.53	0.39	0.40
i)	親しみにくい	親しみやすい	0.47	0.38	0.28
j)	消極的な	積極的な	0.49	0.48	0.45
k)	つまらない	面白い	0.37	0.26	0.36
l)	単純な	複雑な	0.39	0.24	0.40
m)	嫌いな	好きな	0.17	0.14	0.10
n)	わがままな	思いやりのある	0.30	0.24	0.22
o)	空虚な	充実した	0.17	0.20	0.35
p)	愚かな	賢い	0.13	0.11	0.21
q)	にくらしい	かわいらしい	0.24	0.06	0.07
r)	苦しい	楽しい	0.24	0.17	0.38
s)	冷たい	温かい	0.59	0.29	0.23
t)	機械的な	人間的な	0.73	0.53	0.49

5.2 推定結果と分析

各形容詞対について，推定した値と各評価者が付与した値の相関係数を表 6 に示す．推定値と各評価者の間の相関係数は，評価者間の相関係数が 0.6 以下の場合は，全体的にやや低い傾向を示すものの有意差のない範囲の値となっている．すなわち，人の評価の揺れと大きくは変わらない範囲で，システムも評価値を推定できているといえる．しかし，評価者間の相関係数が高い勝率や z) 応答文としての自然さなどは，有意に評価者間の相関係数のほうが高くなっている．こうした評価者間の揺れが小さい評価尺度については，本実験で用いた応答文との WER ベースの類似度に加え，評価値に影響する要素をより詳細に調べていく必要がある．

6 結論

本研究では，雑談対話システムが出力する発話文に対する評価を，2 名の評価者による SD 法を用いて多面的に分析した．因子分析を通して，両評価者ともに 4 つの因子に分解されること，また評価者間で親近感と自然性の 2 つの因子は共通していることが示された．各評価尺度と総合評価値との相関を調べたところ，自然性と親近感の因子に分類される評価尺度は相関が高かったものの，それ以外の評価尺度については入力文ごとに大きく異なることが示された．また，各評価尺度の自動推定に関する検討を行い，自然性と親近感の因子に関する評価尺度以外は評価者間と同程度の相関を達成した．多様な評価尺度において人と同程度の精度で推定できており，推定手法が頑健であることが示された．展望として，複数ターンから成る対話実験で得られる評価との比較や，入力文数および評価者数の拡充による検証の妥当性向上，言語特徴などの新しい特徴量の導入を進めたい．

参考文献

[Higashinaka 14] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing, in *Proc. COLING*, pp. 928–939 (2014)

[Janarthanam 08] Janarthanam, S. and Lemon, O.: User simulations for online adaptation and knowledge-alignment in Troubleshooting dialogue systems, in *Proc. LONDIAL*, Vol. 45 (2008)

[Maat 10] Maat, M., Truong, K. P., and Heylen, D.: How turn-taking strategies influence users ' impressions of an agent, in *Proc. IVA*, pp. 441–453 (2010)

[Meguro 10] Meguro, T., Higashinaka, R., Minami, Y., and Dohsaka, K.: Controlling Listening-oriented Dialogue using Partially Observable Markov Decision Processes, in *Proc. COLING*, pp. 761–769 (2010)

[Nishimura 07] Nishimura, R., Kitaoka, N., and Nakagawa, S.: A Spoken Dialog System for Chat-like Conversations Considering Response Timing, *Text, Speech and Dialogue*, pp. 599–606 (2007)

[Nishimura 08] Nishimura, R., Kitaoka, N., and Nakagawa, S.: Analysis of Relationship Between Impression of Human-to-human Conversations and Prosodic Change and Its Modeling, in *Proc. Interspeech*, pp. 534–537 (2008)

[Ritter 11] Ritter, A., Cherry, C., and Dolan, W.: Data-Driven Response Generation in Social Media, in *Proc. EMNLP*, pp. 583–593 (2011)

[Sculley 09] Sculley, D.: Large Scale Learning to Rank, in *NIPS 2009 Workshop on Advances in Ranking*, pp. 1–6 (2009)

[Smola 04] Smola, A. J. and Schölkopf, B.: A Tutorial on Support Vector Regression , *Statistics and computing*, Vol. 14, No. 3, pp. 199–222 (2004)

[Sugiyama 13] Sugiyama, H., Meguro, T., Higashinaka, R., and Minami, Y.: Open-domain Utterance Generation for Conversational Dialogue Systems using Web-scale Dependency Structures, in *Proc. SIGDIAL*, pp. 334–338 (2013)

[Walker 97] Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents, in *Proc. EACL*, pp. 271–280 (1997)

[Wong 12] Wong, W., Cavedon, L., Thangarajah, J., and Padgham, L.: Strategies for Mixed-Initiative Conversation Management using Question-Answer Pairs, in *Proc. COLING*, pp. 2821–2834 (2012)

[稲葉 14] 稲葉通将, 神園彩香, 高橋健一：Twitter を用いた非タスク指向型対話システムのための発話候補文獲得, 人工知能学会論文誌, Vol. 29, No. 1, pp. 21–31 (2014)

[今井 10] 今井健太, ジェプカ ラファウ, 荒木健治：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, 情報処理学会研究報告, Vol. 2010-NL-19, No. 10, pp. 1–7 (2010)

[杉山 14] 杉山弘晃, 目黒豊美, 東中竜一郎：大規模マルチリファレンスに基づく雑談対話システムの自動評価に向けた実験的検討, 人工知能学会 言語・音声理解と対話処理研究会, pp. 1–6 (2014)

[大西 14] 大西可奈子, 吉村健：コンピュータとの自然な会話を実現する雑談対話技術, NTT DoCoMo テクニカル・ジャーナル, Vol. 21, No. 4, pp. 17–21 (2014)

[東中 14] 東中竜一郎：雑談対話システムに向けた取り組み, 第 70 回言語・音声理解と対話処理研究会 (2014)

[目黒 14] 目黒豊美, 杉山弘晃, 東中竜一郎, 南泰浩：ルールベース発話生成と統計的発話生成の融合に基づく対話システムの構築, 人工知能学会全国大会 (2014)