個性に着目した対話システムの自然性の評価実験

須戸悠太† 高椋琴美†† 谷田泰郎†† 山本和英†
††シナジーマーケティング株式会社 †長岡技術科学大学

1. はじめに

シナジーマーケティング社では，独自に構築している"心のモデル"によって個人の個性を規定し，その個性の違いによる人の行動・興味・感じ方・考え方・人間関係の持ち方などのモデル化を目指している．対話によるコミュニケーションにはこれらに関するヒントが数多くあると考え，個性に着目した対話システムを作成し，その対話システムと人が自由に会話することによって得られる対話データを収集，研究することで，上記への知見を得たいと考えている．

筆者らは，先行研究[1] で作成した対話システムの「個性性」と「自然性」についての評価を行ってきたが，自然性が担保されていないと個性性の評価が難しいと言う課題があった．自然性を上げるためには，①2 文対の用例を超えた文脈，状況，話題の抽出・切り替えなどの仕組みの導入②コモンセンス知識の導入③用例を増やす④辞書やシソーラス，意味素性の整備などの方法などが考えられる．①は課題そのものが難しく，②や④は定量的アプローチによる方法で量を増やすこともできるがその質が問題となり，定性的な方法を選択すると人手によるコストがかかる．そもそも，知識量を増やしてしまうと，個性が埋没する可能性もある．4 つのアプローチの中で，③がコストや課題の難易度の点で一番取り組みやすい．但し，複数の会話データを採用すると用例数は増えるが，複数の個性を混ぜると対話システムの個性が中庸化したり，個性の破綻に違和感を覚えたりすることが過去の実験から分かっており，個性性を確保しながら用例を増やさなければならないという課題があった．用例数と自然性の関係性を確かめ，十分な用例数を確保した上で，対話システムから引き出される情報の量や質の違いと個性の関係性について確認する必要がある．そのための実験も別途計画しているが，本稿では，どれくらいの用例数があれば自然性を担保できるのかの評価実験について報告する．

2. 実験

2.1.対話システムについて

学習データは｛“相手の発話”，“ロボットの発話”｝の 2 文で構成されている．“相手の発話”とは“ロボットの発話”に対する直前の発話である．

ロボットは，非タスク指向型の対話システムである．対話エンジンは 2 つの発話を意味素性と表層表現のベクトルとして学習し，発話の意味素性（発話タイプも含む）と表層表現ベクトルの類似度スコアの高い“相手の発話”を選択し，それに対応する“ロボットの発話”返すというシンプルな仕組みとなっている．

2.2.実験に用いた学習データについて

雑談対話において,用例を増やすには WWW 上から発話文を選択する手法[2] や Twitter データを利用する手法[3] がある．しかし，これらの手法は不特定多数の人が作成した文を使用することになる．前述のように個性を重視するためにロボットの発話に当たる部分は人手で作成した．

実験には 2 種類の個性（type_1, type_2）を使

用した．各個性の発話に用いる学習データはそれぞれ 1 人の作成者が作成した．

① まず，各作成者が内容や発話長などの制約は設けずに一人二役で 1,500 種類の発話対を作成した．

② 次に，対話のパターンを増やすために，クラウド・ソーシングを利用して①の発話に対して想像される直前の発話を収集し，不適切な発話を除いたうえで学習データを作成した．

学習データの用例数は，

(a) ①，②からランダムに選択した 14,000 対

(b) (a) からランダムに選択した 10,000 対

(c) (b) からランダムに選択した 5,000 対

の 3 段階を 2 種類の個性（type_1, type_2）ごとに設定し，表 1 に示すように，合計 6 種類のロボットを作成した．以下，用例数と個性を掛け合わせたそれぞれのロボットを robo-A～robo-F（実験結果の表中は，A～F と簡略化）という名称で表記することにする．

表 1　対話ロボットの分類

学習データ	type_1	type_2
(a) 14,000 対	robo-A	robo-D
(b) 10,000 対	robo-B	robo-E
(c)　5,000 対	robo-C	robo-F

2.3.実験の流れ

被験者は 20 代の男女各 2 名の計 4 名で，6 種類のロボットと 30 分ずつチャット形式で自由に対話してもらった．なお，対話するロボットの順番はランダムに決定した．

実験は前半・後半に分けて行い，各被験者に 3 体ずつのロボットと対話してもらった．3 体のロボットとの対話が終了した後，それぞれのロボットについての印象をアンケートに記入してもらった．表 2 の表側に示すように，アンケートは質問 10 項目で構成されており，それぞれ 10 点満点の評価になっている．

選択した 10 項目は，設問 1, 2, 3, 7, 8 が会話の破綻や発話の意味の理解を，設問 4, 5, 6 は学習データについて，設問 9, 10 は会話の楽しさや前向きさを聴取するための内容になっている．会話の楽しさなどについては，直接自然性を問うものではないが，自然性に影響を与える可能性があるという理由で項目に含めた．また，アンケートには，自由記述部分を設け，実験終了後にも実験の感想などについてのインタビューを行った．

3. 結果

アンケートの結果を表 2 に示す．No.1 と No.4 は女性，No.2 と No.3 は男性である．No.1 と No.4 は評価点が 6 点前後であったのに対し，No.3 は 3~4 点を中心に点数がつけられていた．No.2 は前

表2 実験結果

	No.1(F)						No.2(M)						No.3(M)						No.4(F)					
	前半			後半			前半			後半			前半			後半			前半			後半		
会話したロボット	F	A	B	E	D	C	D	C	F	B	E	A	B	D	F	A	C	E	A	C	E	F	B	D
1. 会話がつながる	6	7	6	8	7	6	1	1	1	7	8	7	3	3	3	2	2	1	6	6	7	5	5	6
2. あなたの発言の意味を理解している	5	7	7	7	6	6	1	2	1	5	5	5	4	4	5	3	4	2	7	6	8	6	6	6
3. 受け答えが自然である	6	6	6	7	7	6	1	1	1	5	5	5	6	6	6	2	2	2	6	6	7	5	6	6
4. 表現が多様である	8	8	8	8	7	6	1	2	1	3	3	3	4	4	4	3	3	3	6	6	8	6	6	6
5. 言い回しに違和感がない	7	8	8	8	7	7	3	3	2	7	7	7	4	4	4	6	6	6	5	5	6	6	6	6
6. 話題が豊富である	6	7	6	7	6	5	3	2	1	7	7	7	6	6	6	5	5	5	5	5	6	6	7	7
7. 突拍子もない発言が少ない	5	6	5	7	6	5	2	2	2	7	7	5	1	1	1	2	2	1	6	6	6	7	7	7
8. 会話が一方的にならない	4	5	5	6	6	5	3	3	3	6	6	6	5	5	5	2	2	1	5	5	7	7	7	7
9. 会話が楽しかった	4	6	5	5	5	5	0	1	0	9	9	9	4	4	4	2	2	2	5	5	7	6	6	7
10. また話したいと思う	6	7	6	5	5	5	0	0	0	9	9	9	0	0	0	1	1	1	5	5	6	6	6	7

半と後半で評価点が大きく異なっているが，実験後のインタビューによると，前半と後半で会話の形式を変えたとのことだった．

このように，前半と後半で実験への慣れや会話の形式を変えた人もいたという理由から各被験者，前半と後半で分けて標準化を行った．評価点 x を基に平均値 μ から (1) 式を用いて標準偏差 s を求める．なお，n はデータ数であり，今回は $n = 30$ となる．次に（2）式を用いて標準化を行い，そのスコアを求めた．スコアをロボット毎に整理した結果を表 3 に示す．データ数とスコアの合計点をプロットしたグラフを図 1 に示す．

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \qquad (1)$$

$$z = \frac{x - \mu}{s} \qquad (2)$$

4. 考察

図 1 に示した用例数と評価の関係では，全体的に明瞭な相関は見られなかったが，マイナス評価だけに着目すると，学習データの用例数 5,000 より 10,000 の方がマイナス評価は減っている．一方，学習データの用例数 10,000 と 14,000 をではその差はほとんどない．次の個性評価の実験で利用する学習データの用例数としては，5,000 でも十分かもしれないが，マイナス評価が少なくなっていることから 5,000 よりは 10,000 のほうが良

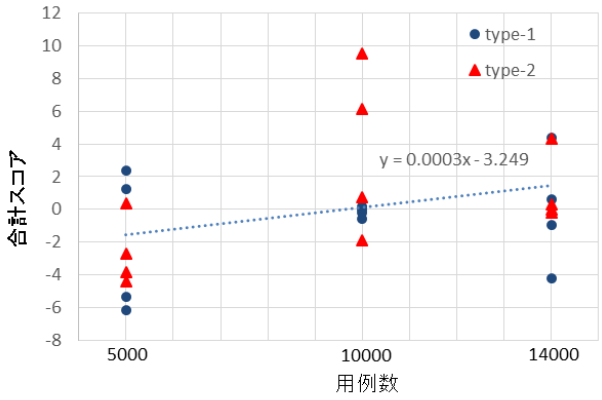

図1 用例数と合計スコア

いのではないかと思われる．

用例数 10,000 の所に極端に評価の高い個所が 2 箇所ある（どちらも type-2 のロボットと喋った女性，No.1 と No.4）．アンケートや対話ログを見ると，No.1 は robo-D ではテレビの話題が，robo-E では手塚治虫の話題が多いとあり，元が同じロボットでも話題が異なっており，違う話題情報を引き出したことが評価に影響を与えていると思われる．対話例を表 4，表 5 に示す．また，アンケートの自由記述の回答部分によると，No.4 は前半の実験（robo-A，C ,E と対話している）の中で robo-E との会話が最も成り立っていたとのことで，定量評価でも高得点となっており，type-1 のロボットに比べて type-2 のロボットを高評価している（robo-E は type-2）．これらのことから，評価が上下する理由として，話題やロボットとの相性の影響もあると考えられる．

表3 標準化

	A				B				C				D				E				F			
	No.1	No.2	No.3	No.4	No.1	No.2	No.3	No.4	No.1	No.2	No.3	No.4	No.1	No.2	No.3	No.4	No.1	No.2	No.3	No.4	No.1	No.2	No.3	No.4
1	0.70	0.30	-0.44	0.04	-0.18	0.30	-0.39	-2.00	-0.20	-0.47	-0.44	0.04	0.82	-0.47	-0.39	-0.38	1.84	0.86	-1.08	1.18	-0.18	-0.47	-0.39	-2.00
2	0.70	-0.82	0.19	1.18	0.70	-0.82	0.14	-0.38	-0.20	0.54	0.83	0.04	-0.20	-0.47	0.14	-0.38	0.82	-0.82	-0.44	2.32	-1.06	-0.47	0.68	-0.38
3	-0.18	-0.82	-0.44	0.04	-0.18	-0.82	1.22	-0.38	-0.20	-0.47	-0.44	0.04	0.82	-0.47	1.22	-0.38	0.82	-0.82	-0.44	1.18	-0.18	-0.47	1.22	-2.00
4	1.58	-1.94	0.19	0.04	1.58	-1.94	0.14	-0.38	-0.20	0.54	0.19	0.04	0.82	-0.47	0.14	-0.38	1.84	-1.94	0.19	2.32	1.58	-0.47	0.14	-0.38
5	1.58	0.30	2.10	-1.10	1.58	0.30	0.14	-0.38	0.82	1.55	2.10	-1.10	0.82	1.55	0.14	-0.38	1.84	0.30	2.10	0.04	0.70	0.54	0.14	-0.38
6	0.70	0.30	1.46	-1.10	-0.18	0.30	1.22	1.25	-1.22	0.54	1.46	-1.10	-0.20	1.55	1.22	1.25	0.82	0.30	1.46	0.04	-0.18	-0.47	1.22	-0.38
7	-0.18	-0.82	-0.44	0.04	-1.06	0.30	-1.47	1.25	-1.22	0.54	-0.44	0.04	-0.20	0.54	-1.47	1.25	0.82	0.30	-1.08	0.04	-1.06	0.54	-1.47	1.25
8	-1.06	-0.26	-0.44	-1.10	-1.06	-0.26	0.68	1.25	-1.22	1.55	-0.44	-1.10	-0.20	1.55	0.68	1.25	-0.20	-0.26	-1.08	1.18	-1.93	1.55	0.68	1.25
9	-0.18	1.42	-0.44	-1.10	-1.06	1.42	0.14	-0.38	-1.22	-0.47	-0.44	-1.10	-1.22	-1.48	0.14	1.25	-1.22	1.42	-0.44	1.18	-1.93	-1.48	0.14	-0.38
10	0.70	1.42	-1.08	-1.10	-0.18	1.42	-2.01	-0.38	-1.22	-1.48	-1.08	-1.10	-1.22	-1.48	-2.01	1.25	-1.22	1.42	-1.08	0.04	-0.18	-1.48	-2.01	-0.38
Total	4.40	-0.93	0.64	-4.19	0.00	0.19	-0.18	-0.54	-6.12	2.35	1.27	-5.33	0.00	0.34	-0.18	4.33	6.12	0.75	-1.91	9.52	-4.40	-2.69	0.36	-3.79

表4 対話ログ1

No.1	色々見るよ
D	バラエティーだと何見てる？
No.1	月曜から夜ふかしとか
D	アメトークだとどのシリーズがいい？
No.1	アメトーク面白いよね
D	アメトーク好きなの？
No.1	好きだよ
D	眠い時どうしてる？

表5 対話ログ2

No.1	興味ある？
E	あるけどあんまり好きじゃない
No.1	なんだそれ
E	そうそう、だから「リボンの騎士」は宝塚風
No.1	リボンの騎士も手塚治虫だね
E	うん、手塚治虫は偉いな
No.1	手塚治虫リスペクトしまくりだね
E	福山ファンに怒られるぞ

男性，女性で評価を比較すると，女性の被験者の評価には幅がある．前述のように，No.1，No.2 における type-2 との対話の評価が 10,000 と 14,000 で大きく離れている．また，対話数とロボットの発話文の種類数をまとめた表 6 を見ると，男女で対話数，種類数は大きく異なり，いずれも女性のほうが多かった．今回の実験では，被験者数が少ないので何とも言えないが，シナジーマーケティング社の別の調査（対話ロボットに求めるものを自由記述してもらった 1,000 件の調査）で女性の方が男性より対話ロボットの受容性が高いということも分かっている．当然性差による会話の違いもあるのでその影響もあると思われる．

また，表 6 を人ごとに確認すると，30 分間の対話数には幅があるが，ロボットの発話文の種類数は大きく変わらない．これは人によって 30 分間で引き出せる情報量がある程度決まっていて，その違いが表れているのではないかと考えられる．

表6 対話数と発話文の種類数

	No.1		No.2		No.3		No.4	
robo	対話	種類	対話	種類	対話	種類	対話	種類
A	282	213	103	90	142	107	226	163
B	294	216	176	119	109	91	218	153
C	395	220	259	137	120	79	221	152
D	307	199	215	133	112	92	170	139
E	336	219	134	115	131	95	198	155
F	232	172	272	127	129	93	218	158
平均	307.7	206.5	193.2	120.2	123.8	92.8	208.5	153.3

5. おわりに

本稿の実験を通して，個人の評価に対する考え方や会話で引き出された情報などが評価に影響を与えてしまうため，自然性についての定量的な評価は難しいという課題も見つかったが，現状の対話システムで次に計画している個性評価の実験をする場合の用例数は 10,000 程度あれば十分ではないかという結論も得られた．また，対話システムを考える場合，やはり，個性差，男女差を考慮する必要があるという感触も得られた．

アンケートの自由記述に「会話がつながらないのが辛い」という意見もあった．用例数を増やすだけでは自然性を上げるのには限界があり，個性を埋没させないような文脈の理解やコモンセンス知識の導入の検討も必要である．

今後，用例数を 10,000 に設定して，対話システムから引き出される情報の違いと個性の関係性についての実験を行う予定である．実験では，ロボットを 3 体用意して 10 人の被験者に評価してもらう計画になっている．これまでに行ってきた対話実験の内容も含め，個性と対話ロボットの関係をまとめていきたい．

参考

[1] 谷田泰郎, 高椋琴美. 対話ロボットの個性 - 個性に着目した対話ロボットの評価 - . 電子情報通信学会 クラウドネットワークロボット研究会（CNR）, 信学技報 Vol. 115 No. 283 pp. 5-10

[2] 柴田 雅博, 冨浦 洋一, 西口 友美. 雑談自由対話を実現するための WWW 上の文書からの妥当な候補文選択手法. 人工知能学会論文誌 Vol. 24 (2009) No. 6 pp. 507-519

[3] 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. Vol. 29 (2014) No. 1 論文特集「知的対話システム」,「近未来チャレンジ 2012」, 一般論文, 2013 年度大会速報論文特集 pp. 21-3