社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－NL－170（17）
2005／11／22

非タスク指向型対話システムの評価法

曽我部 将義*, 鳥海 不二夫, 石井 健一郎
名古屋大学大学院 情報科学研究科

近年，様々な目的から人間とコミュニケーションを行うコンピュータ，すなわち，対話システムへの期待が高まっている．このようなシステムの実現のためには，開発したシステムの性能の評価が不可欠である．しかし，客観的・定量的な評価法がある程度確立されているタスク指向型対話システムに対し，非タスク指向型対話システムにはアンケートなどの主観的な評価法が主に用いられている．本研究では，テキストによって非タスク指向型対話を行う対話システムの客観的・定量的な評価法を提案する．提案手法では，人間同士の対話を理想的な対話と仮定し，人間同士の対話を基準として，人間同士の対話との類似度という観点から，対話がどの程度人間らしいかという評価を行う．提案手法を用いて人間同士の対話と人間と対話システムの対話とをそれぞれ評価したところ，高度な自然言語処理を使っていない対話システムと人間の対話は，人間同士の対話よりも低い評価値となった．

An evaluation method of non-task-oriented dialog system

Masayoshi Sogabe* Fujio Toriumi Kenichiro Ishii
Graduate School of Information Science, Nagoya University

Recently, computerized dialog systems using natural language are highly expected. The system's performance should be evaluated objectively and quantitatively. The objective and quantitative evaluation method for the task-oriented dialog system has already established, however non-task-oriented dialog systems have been evaluated only by subjective method, like a questionnaire.
In this paper, we propose a new criterion which can evaluate non-task-oriented dialog systems objectively and quantitatively. The criterion utilizes similarity between a human-human dialog and a human-machine dialog. In the simulation, we found that a human-machine dialog which doesn't use natural language processing is evaluated as less human than a human-human dialog. And it became clear that a new criterion can evaluate the humanity of a dialog.

1　はじめに

近年，人間と話し言葉でコミュニケーションを行う事ができるコンピュータ，すなわち，対話システムへの期待が高まっている．その期待は，何らかのタスクを達成するために対話を行うタスク指向型対話システムだけではなく，人間の話し相手になり，人間を楽しませるために対話を行う非タスク指向型対話システムにも向けられている．非タスク指向型対話システムには，例えば，老人や一人暮らしの人の話し相手などの需要が見込まれる．より良い対話システムを設計するためには，その性能を評価する尺度が必要になる．タスク指向型対話システムには，タスク達成率やタスク達成コストなどの客観的・定量的な評価法がある程度確立されている [1, 2] が，非タスク指向型対話システムにはアンケートなどの主観的な評価法が主に用いられており，客観的な評価法は確立されていない．そこで，本論文では，テキスト対話を対象とし，非タスク指向型対話システムを客観的・定量的に評価する手法について述べる．

2　提案手法

非タスク指向型対話システムは，人間を楽しませるための対話を目指しているため，対話がうまく行われたか否かを客観的・定量的に評価する事が困難である．

本研究では，人間同士の対話が理想的な対話であると仮定し，人間同士の対話を基準として，対話がどの程度人間らしいか，という人間同士の対話との類似性に着目する．なお，テキスト対話における一回の発言を発話，会話の始まりから終わりまでの発話の集合を対話と定義する．また，基準となる人間同士の対話を基準対話，評価したい対話を被評価対話と呼ぶ．

対話の一部を切り取ったときに，ある発話に対応して次の発話がなされるという部分的な対話の流れを考える．このとき，被評価対話の部分的な対話の流れが基準対話のものに近ければ，被評価対話は人間らしい対話であるとみなすことができる．例えば，基準対話において，「明日の天気を知っていますか？」という発話の次になされる発話としては，「晴れです」「知りません」等が考えられる．このとき，被評価対話にも同様の受け答えがあれば，被評価対話は基準対話との類似部分を持つと考える．このように，基準対話内の部分対話が被評価対話にも頻出すれば，それは基準対話に近い対話であるとする．本研究では，この考えに基づき，被評価対話が基準対話内の部分対話をどの程度内包しているか，という点から，対話の人間らしさの評価を行う手法を提案する．

以下が提案手法の評価の大まかな流れである．まず，前処理として以下の処理を行う．

1. 基準対話を準備する．
2. 基準対話の発話にタグを付与し，対話コーパスを構築する
3. 対話コーパスの各発話ごとに特徴抽出を行う
4. 抽出した特徴ベクトルを用いてクラスタリングを行い，発話クラスを作成する

その後，人間と対話システムによる対話を行い，以下の処理によって評価を行う．

5. 被評価対話の各発話を 4 で定義したクラスに割り当てる
6. 基準対話と被評価対話を比較し，類似度を求める
7. 求めた類似度によって評価を行う

2.1 対話コーパス構築

基準対話の各発話に発話タグと感情タグを付与し，対話コーパスを構築する．付与する発話タグとして SWBD-DAMSL タグ [3] を，感情タグとして Ekman らの基本 6 感情 [4, 5, 6] を用いる．

2.1.1 SWBD-DAMSL タグ

発話の種類を記述するためのタグとして SWBD-DAMSL タグを用い，これを各発話に付与する．例えば，「あなたは夏が好きですか？」という発話に対しては，Yes-No-Question を示すタグと Statement-opinion を示すタグが付与される．ここでは主観的意見を Statement-opinion，客観的叙述を Statement-non-opinion とみなす．主観的意見と客観的叙述の定義については Wiebe らの手法に従い，発話者の主な意図が客観的情報の伝達にあり，かつ主観的な単語が含まれていない場合に客観的叙述とし，それ以外は主観的意見とする [7]．各発話はそれぞれ 1 種類以上の SWBD-DAMSL タグを持つ．

2.1.2 感情タグ

Ekman らの基本 6 感情を感情タグとし，発話から以下のいずれかの感情の生起が感じられる場合，その感情タグを付与する．

- 幸福 (Hapiness)
- 悲しみ (Sadness)
- 驚き (Surprise)
- 恐れ (Fear)
- 怒り (Anger)
- 嫌悪 (Disgust)

各発話はそれぞれ 0 種類以上の感情タグを持つ．

2.2 特徴抽出

対話コーパスの各発話に対し特徴抽出を行い，各発話を d 次元の特徴ベクトル $\boldsymbol{f} = (f_1, f_2, \ldots, f_d)$ で表す．本研究で用いた特徴は Table.1 の通りである．ただし，定型表現には EDR 辞書 [8] を用い，該当する表現が属するカテゴリ内の，その表現と同じ階層にある全ての表現を，定型表現とみなす．例えば，「ありがとう」，「サンキュー」，「サンクス」は同じ概念に属するため，全て定型表現「ありがとう」とみなす．この 82 次元の特徴ベクトル $\boldsymbol{f}$ を KL 展開し，58 次元に次元数を削減した特徴ベクトル $\boldsymbol{f'}$ を，その発話の特徴ベクトルとして用いる．

2.3 クラスタリング

対話コーパス中の各発話をクラスタリングし，発話クラスに分類する．クラスタリング手法には Ward 法 [9] を用い，距離の定義としてユークリッド距離の 2 乗を用いる．クラスタリングの結果として得られるクラスタは，クラス内分散・クラス間分散比

$$J_\sigma = \frac{\sigma_B^2}{\sigma_W^2} \tag{1}$$

によって評価する．σ_W^2 はクラス内分散，σ_B^2 はクラス間分散である．クラスタリングによって得られたクラスタを，発話クラスとする．

Table.1 Features extracted from utterance

特徴	次元数
発話に付与された発話タグ	48
発話に付与された感情タグ	6
文末の？の有無	1
代名詞の有無	1
形容詞の有無	1
数詞の有無	1
接続詞の有無	1
接続助詞の有無	1
感動詞の有無	1
固有名詞の有無	1
フィラーの有無	1
形態素数	1
6W1H の疑問詞の有無	7
定型表現「おはよう」の有無	1
定型表現「こんにちは」の有無	1
定型表現「こんばんは」の有無	1
定型表現「おやすみ」の有無	1
定型表現「ありがとう」の有無	1
定型表現「はじめまして」の有無	1
定型表現「よろしく」の有無	1
定型表現「ごめんなさい」の有無	1
定型表現「どうも」の有無	1
対話全体の中での発話の位置	1
発話者の ID	1
計	82

2.4 発話クラス割り当て

基準対話のクラスタリングによって得られたクラスに被評価対話を当てはめ，発話にクラスを割り当てる．基準対話のクラスタリングによって得られた l 個のクラスの内，i 番目のクラスを $c_i\,(i=1,2,\ldots,l)$，c_i に含まれる k 個の発話の内，j 番目の発話を $u_{ij}\,(j=1,2,\ldots,k)$ とする．このとき，u_{ij} と c_i の重心との距離のうち最大のものを r_i とし，クラスが未知の発話 u の最近傍のクラスが c_i のとき，u と c_i の重心との距離を r_u とする．$r_u \leq r_i$ のとき，u は c_i に含まれ，それ以外の u はリジェクトする．

2.5 評価スコア

提案手法は，対話の流れに着目し，ある対話が人間同士の対話の流れと似た対話の流れを持っていれば，その対話は人間らしいものであるという考え方に基づく．従って，基準対話の部分対話と被評価対話の部分対話がどの程度一致しているかという点から評価を行う．具体的には，様々な発話の繰り返しからなる対話を発話クラスの系列とし，基準対話の系列の部分系列と被評価対話の系列の部分系列との一致率によって，対話に評価スコアを与える．

得られたクラスタリング結果に基づき，基準対話と被評価対話を発話クラスの系列に置き換える．基準対話の系列 D_b と被評価対話の系列 D_e を，それぞれ

$$D_b = (c_{b1}, c_{b2}, \ldots, c_{bm}), \tag{2}$$

$$D_e = (c_{e1}, c_{e2}, \ldots, c_{en}) \tag{3}$$

とする．ただし，c は発話クラスを表す．このとき，D_b，D_e に含まれる長さ t の部分系列 d_b，d_e をそれぞれ

$$d_{b,i,t} = \left(c_{bi}, c_{b(i+1)}, \ldots c_{b(i+t-1)}\right)\,(i=1,2,\ldots,m), \tag{4}$$

$$d_{e,j,t} = \left(c_{ej}, c_{e(j+1)}, \ldots c_{e(j+t-1)}\right)\,(j=1,2,\ldots,n) \tag{5}$$

とする．ただし，$i+t-1>m$，$j+t-1>n$ の場合は，それぞれ $i+t-1=m$，$j+t-1=n$ とする．このようにして得られた系列 D_b と系列 D_e を Fig.1 のように比較し，その類似度を求める．

まず，D_b 中の部分系列 $d_{b,i,t}$ が D_b 中に出現する頻度を $freq_{b,i}$ とし，

$$p_{b,i} = \frac{freq_{b,i}}{m} \tag{6}$$

とする．例えば，$t=3$ で評価を行う場合を考える．ここで，$i=4$ のとき，$d_{b,4,3} = (c_{b4}, c_{b5}, c_{b6}) = (c_\alpha, c_\beta, c_\gamma)$ であるとすると，D_b を $i=1$ から $i=n$ まで順に走査したときに，$c_\alpha, c_\beta, c_\gamma$ が連続して D_b 中に出現する回数を $freq_{b,4}$ とする．これには $d_{b,4,3}$ 自身も含まれる．同様に $d_{b,i,t}$ が D_e 中に出現する頻度を $freq_{e,i}$ とし，

$$p_{e,i} = \frac{freq_{e,i}}{n} \tag{7}$$

とする．この操作を $i=1$ から $i=m$ まで繰り返し，それぞれの場合の $d_{b,i,t}$ に対する $p_{b,i}$ と $p_{e,i}$ を求める．そして，対話同士の類似度を

$$s = \sum_{i=1}^{m} p_{b,i} \times p_{e,i} \tag{8}$$

と定義する．この類似度 s を被評価対話の評価スコアとする．

3 類似度による対話評価法を用いた対話評価実験

提案手法では対話の人間らしさという観点から評価を行うため，人間同士の対話を被評価対話としたときには評価スコアは高い値を取り，人間と対話システムの対話を被評価対話としたとき，その被評価対話が人間らしくない対話なら，評価スコアは低い値を取るはずである．この考え

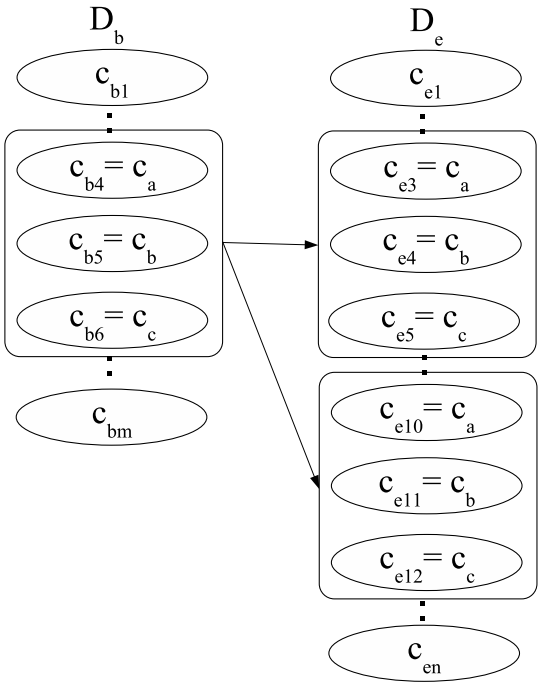

Fig.1　The example of calculating $freq_{b,i}$

に基づいて提案手法の有効性を検証するために，提案手法を用いて人間同士の対話と人間と対話システムの対話の評価を行い，それぞれの評価結果を比較する．

3.1　実験方法

人間と対話システムの対話データとして，本研究室で作成した対話システムである KELDIC [10] と人間との対話データを用いる．対話システム KELDIC は，高度な自然言語処理を用いたものではなく，基本的な仕組みは Eliza [11] にならって設計されている．従って，現状では，人間らしく対話ができるレベルには達していない．人間同士の対話，人間と対話システムの対話ともに，対話の際には Table.2 の制約を課す．

Table.2　Restrictions on a dialog

制約
1 対 1 で交互に発言
話題は制限しない
顔文字や方言は使用しない
文体として書き言葉を使用する

人間同士の対話データ数は 9 であり，人間同士の対話のうち 1 対話を被評価対話とし，残りを基準対話とする．これらの対話データを用い，人間同士の被評価対話と人間と対話システムの被評価対話を評価し，その結果を比較する．クラスタリングによって得られる樹形図を切断するレベルと，対話の部分系列の長さ，人間同士の対話データの中から被評価対話として選ぶ対話データを変えながら，実験を繰り返し行う．

また，提案手法の頑健性を検証するため，KELDIC 以外の別々の対話システム ARISA [12] とししゃも [13] を用いた場合の対話データを用いて同様の実験を行う．

3.2　提案手法における有効なパラメータの検証実験

Fig.2 が人間同士の対話データ 9 対話のうちのある 1 つを被評価対話として選んだときの，残りの基準対話をクラスタリングして得られた樹形図である．

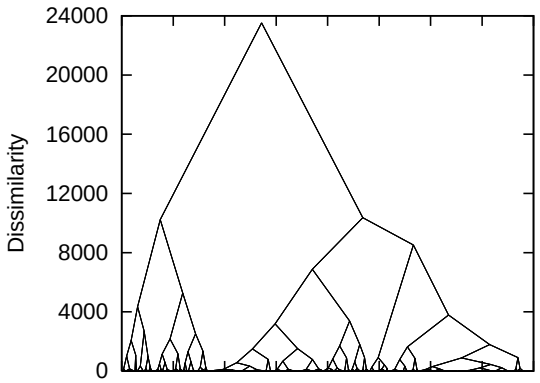

Fig.2　Dendrogram

Fig.3 は，Fig.2 の Dissimilarity が 500 以上のクラスタを別のクラスタとみなした場合の評価スコアのグラフである．すなわち，Fig.3 は，Ward 法の定義により，クラスタ同士の併合後の郡内平方和の増分が 500 以上となるクラスタを別のクラスタとみなした場合である．

グラフの 1 本の線は，被評価対話の評価スコアを平均したものである．本実験では，人間と対話システムの対話データ数は 4 である．従って，人間同士の対話の評価スコアを示す線は，人間同士の対話データのうちの 1 つを他の 8 つの対話の系列を基準としてそれぞれと比較し，その平均をとったものである．人間と対話システムの対話の評価スコアを示す線は，評価用として選んだ人間同士の対話データを除いた他の 8 つの人間同士の対話の系列を基準としてそれぞれと比較し，その平均をとったものである．横軸が部分系列の長さ t，縦軸が評価スコア s の平均値である．このときのクラスタ数は 38 個で，クラス内分散・クラス間分散比は 4.81 である．

Fig.3 より，Dissimilarity が 500 程度でクラスタを分けた場合，部分系列の長さ 5〜10 程度で，人間同士の対話は人間と対話システムの対話よりも高い評価スコアを持つ事がわかる．片側検定を行ったところ，評価スコアの差は

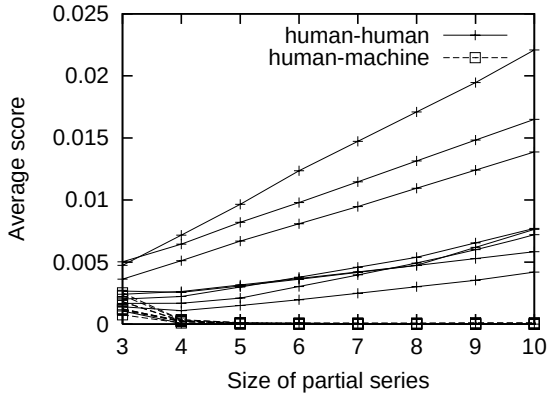

Fig.3 The evaluated score (dissimilarity = 500, dialog system: KELDIC)

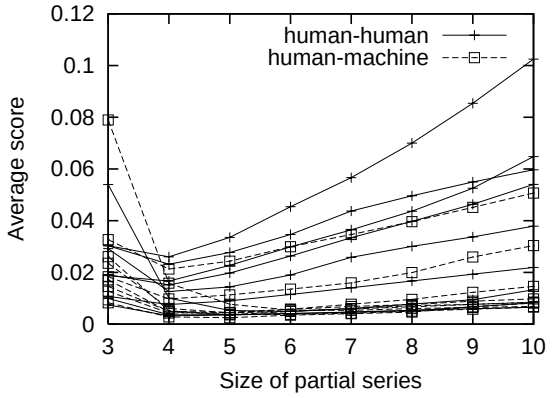

Fig.4 The evaluated score (dissimilarity = 3000, dialog system: KELDIC)

有意水準 0.05% 以下で有意であった．この結果より，人間と対話システムの対話は人間同士の対話よりも評価スコアが低くなるという事ができ，提案手法による人間らしさの評価の妥当性が明らかになった．

また，Table.3 は，Dissimilarity を変化させた場合の，人間同士の対話の評価スコアの分散，人間と対話システムの対話の評価スコアの分散，両者の評価スコアの t 検定の結果をまとめたものである．t 値の*は有意水準 0.05% 以下である事を表す．Table.3 より，Dissimilarity が大きくなり，クラスタ数とクラス内分散・クラス間分散比が小さくなるにつれ，評価スコアの分散が大きくなり，評価スコアの差が無くなっていく傾向があるといえる．しかし，Dissimilarity が 3000 までなら，提案手法の有効性は保たれているといえる．

Table.3 The change of score with dissimilarity

Dis-s	h-h v ($\times 10^{-5}$)	h-m v ($\times 10^{-5}$)	t-value
500	1.06	0.0001	13.60*
1000	1.08	0.07	9.00*
1500	2.49	0.50	6.65*
2000	7.32	1.18	6.10*
2500	10.80	4.64	5.67*
3000	12.20	4.77	4.94*

この結果より，Dissimilarity 500〜3000 程度でクラスタを分けた時，部分系列の長さ 5〜10 程度で，提案手法は非タスク指向型対話システムの客観的・定量的な評価法として有効な手法であるといえる．

Fig.4 が，Dissimilarity が 3000 以上のクラスタを別々のクラスタとみなした場合の結果であり，このときのクラスタ数は 8 個で，クラス内分散・クラス間分散比は 0.20 である．Fig.4 をみると，Fig.3 よりも評価スコアの差が小さくなっていることがわかる．

3.3 様々な対話システムに対する提案手法の有効性の検証実験

Fig.5，Fig.6 は，それぞれ対話システム ARISA としゃもの対話データを用いて同様の実験を行った結果である．ARISA は KELDIC や Eliza と同じく、あらかじめ返答を辞書に定義しておき、相手の発話に対してもっともらしい返答を選び出して出力することで対話を行う対話システムである。また、ししゃもはあらかじめ学習しておいた人間同士の対話データの中から、相手の発話に対する返答候補をいくつか選び出し、それらをもとにマルコフ連鎖を用いて発話文を生成する対話システムである。

人間同士の対話データ数はそれぞれ 9 であり，これは 3.2 の実験で用いたものと同じものである．人間と対話システムの対話データ数はそれぞれ 2 である．また，Dissimilarity 500 以上のクラスタを別々のクラスタとみなす．基準となる人間同士の対話データが共通するため，Fig.3 と同様に，クラスタ数 38 個，クラス内分散・クラス間分散比は 4.81 である．

Fig.5，Fig.6 の結果より，他のアルゴリズムを用いた対話システムへ適用した場合でも，人間同士の対話は人間と対話システムの対話よりも高い評価スコアを持つ事がわかる．片側検定を行ったところ，この場合も評価スコアの差は有意水準 0.05% 以下で有意であった．

このことから，提案手法が特定の対話システムだけでなく，様々な対話システムに対して有効であり，頑健性がある事が確認できた．

4 おわりに

4.1 まとめ

本論文では，人間と非タスク指向型対話を行う対話システムの性能を客観的・定量的に評価する手法を提案した．

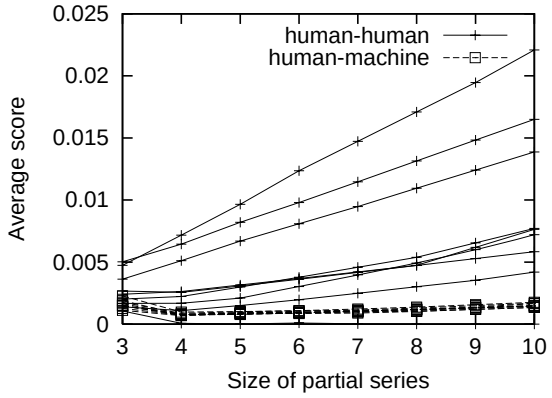

Fig.5 The evaluated score (dissimilarity = 500, dialog system: ARISA)

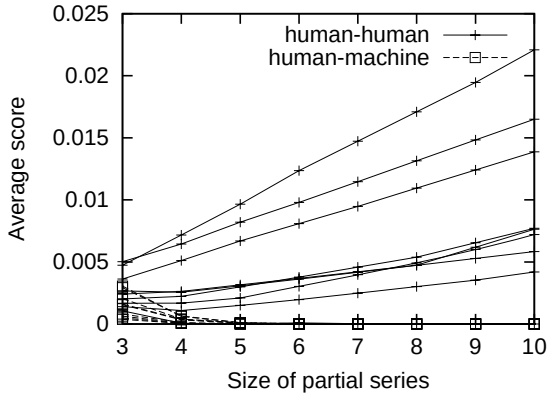

Fig.6 The evaluated score (dissimilarity = 500, dialog system: sixamo)

提案手法では，人間同士の対話を理想的な対話と仮定し，対話を発話クラスの系列と考え，評価したい対話システムと人間の対話が，基準となる人間同士の対話系列に出現する部分系列のパターンをどの程度再現しているかという類似性の面から評価を行った．提案手法の有効性の検証実験を行ったところ，高度な自然言語処理を用いない対話システムと人間の対話は，人間同士の対話よりも低い評価スコアを記録した．また，様々な対話システムを用いた場合でも同様の結果であった．このことから，提案手法が対話の人間らしさの評価法として有効である事がいえた．

4.2 今後の課題

本提案手法を用いて対話システムを評価することで，より良い対話システムの開発に繋げる事ができる．また，改良点として以下のような事柄が考えられる．

人間同士の対話データをさらに増やすことで提案手法の頑健性の向上を図るとともに，評価実験を基に最も有効な部分系列の長さと樹形図の切断レベルを決定する．また，提案手法の評価スコアと人間の主観的な評価結果を照らし合わせる事で，人間が感じる対話の人間らしさと提案手法の評価結果とを比較し，対話の人間らしさの度合いに対応する評価スコアを求められるようにする．また，最終的には，対話に関する制限を無くし，多対多の自由な対話にも対応させる．さらに，各個人それぞれが理想的な対話と考える対話データをコーパスとして用いて評価基準とする事で，各個人に特化した評価を行えるようにする事も可能である．

参考文献

[1] Y. Wilks. (1999). Machine conversations. Kluwer Academic Publishers.

[2] D. Jurafsky. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall.

[3] Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. University of Colorado, Boulder. Institute of Cognitive Science Technical Report 97-02

[4] P. Ekman. (1984). Expression and the nature of emotion. In K. Scherer & P. Ekman (eds). Approaches to Emotion. Hillsdale, NJ: Erlbaum, pp. 319-344.

[5] P. Ekman. (1992). An argument for basic emotions. Cognition and Emotion, 6, 169-200.

[6] P. Ekman. (1992). Are there basic emotions? Psychological Review. 99, 550-553.

[7] J. Wiebe, M. Bell, J. Maples. (1999). Coding Manual for Distinguishing Subjective and Objective Sentences in Text: Draft 2. http://www.cs.pitt.edu/ wiebe/pubs/acl99/codmanv2.ps

[8] http://www2.nict.go.jp/kk/e416/EDR/J_index.html

[9] 宮本定明. (1999). クラスター分析入門 ファジイクラスタリングの理論と応用. 森北出版.

[10] 曽我部将義, 小澤猛志, 石井健一郎. (2004). テキスト対話における対話戦略とその評価法, 電気関係学会東海支部連合大会予稿集, O-414

[11] J.Weizenbaum. (1966). ELIZA: A Computer Program for the Study of Natural LanguageCommunication Between Man and Machine. Communications of the ACM 9(1), pp.36-45.

[12] http://www.nagisanet.com/cgi/index.htm

[13] http://yowaken.dip.jp/sixamo/sixamo.rb.html

[14] 石井健一郎, 上田修功, 前田英作, 村瀬洋. (1998). わかりやすいパターン認識. オーム社.