インタラクティブなシステムの評価をどのように考えていくか

新田 恒雄
（株）東芝 マルチメディア技術研究所
e-mail: nitta@sp.mmlab.toshiba.co.jp

はじめに コンピュータとの対話が文字メディアだけでなく、音声・映像を含むマルチメディアを通して行われる時代を迎えている。これに伴って、複雑化する対話を円滑に進めるための新しいユーザインタフェース（マルチモーダルＵＩ）の枠組みが求められている。マルチモーダルＵＩは広範な技術を含むため、個々の技術の他、統合システムの評価が大きな課題である。ここでは、音声言語を中心とする対話システムの評価方法を例に、議論すべき論点を考察する。

1. 評価の視点

表１は対話システムの評価目的を、結果を利用する人の立場から整理したものである。評価目的はこのように評価主体によって異なるが、性能向上には互いの結果をフィードバックすることが大切である。Ｂ．にはＵＩの評価を含み、またアプリケーションに密着していることから、未だ評価手法が確立されていない。

表１ 対話システム評価の目的

評価主体		評価目的
Ａ．	個別プロセス開発者 ：	個々のプロセスを診断的に評価したい *(Diagnostic Evaluation)*
Ｂ．	システム開発者 ：	システム全体を総合的に評価したい *(Assessment)*

2. 診断的評価からの検討項目

ここでは、個々のプロセスを評価する立場から、評価尺度と留意すべき項目を列記する。表２は対話・システムの各評価プロセスに対する評価尺度と、その背景となるモデルをまとめたものである。表で音韻認識から構文解析までは、音声認識システムの評価の枠組みにあり、実用的な評価手法の提案もなされている。一方この境界を越えると、個々のアプリケーションを視野に入れた評価が要求され、評価手法も模索の段階にある。

表２ 対話システムの評価 − 診断的立場から

	プロセス		評価尺度	背景となるモデルの例
Ａ．	音韻認識	：	音韻認識率	音韻モデル
Ｂ．	単語認識	：	単語認識率	単語モデル
Ｃ．	構文解析	：	解析成功率	構文モデル
Ｄ．	意味解析	：	解析成功率	意味モデル
Ｅ．	対話解析	：	対話文理解率	対話モデル
			対話の（不）自然性	応答（戦略）モデル
				ユーザ（適応）モデル

実際の評価にあたっては、利用者／音響環境／対話環境／タスク／評価指数の諸条件を明確にする必

要がある。ここでは、各条件ごとに留意すべき項目を挙げる。

イ）利用者に関して　：利用者の熟練度／協力度，許容される発話の態様（孤立単語～自由発話）
話者適応の利用（可／不可）

ロ）音響環境に関して：音声伝送特性（周波数特性（マイク，距離，経路），残響特性　…　）
騒音環境（ＳＮ比：～１５ｄＢ～６ｄＢ～０ｄＢ～）

ハ）対話環境に関して：インタラクションの深さ，対話の制御レベル，マルチモーダル環境の整備度
許容される自由発話のレベル
（非言語音／別称・短縮形の使用／倒置／言い直し／言い淀み／…）

ニ）タスクに関して　：語彙数（小／中／大），複雑度（perplexity／構文／意味／タスク）
異なるタスクへの適応性，一度に対応できるタスク数

ホ）評価指数に関して：認識性能　（認識率　リジェクト率）
累積False-alarm rate（単語スポッティング）
応答時間（実時間性），誤り訂正能力

3.　総合的評価からの検討項目

ここでは、システム全体のＵＩを評価する立場から評価項目と留意すべき項目を列記する。表３はＵＩ評価で行われている手法を参考に、対話システムの評価目的と評価項目および背景となるモデルをまとめたものである。

表３　対話システムの評価　－　総合的評価の立場から

評価目的		評価項目（手法）
Ａ．　タスク評価	：	タスク達成率，タスク未達の分析（対話破綻，…）
		タスク遂行時間，タスク学習速度
		＜　背景モデル　：ＧＯＭＳモデルなど　＞
Ｂ．　操作性の評価	：	エラー率，反応速度，疲労度
		モニタリング（ＶＴＲ，アイカメラ，呟き…）とプロトコル解析
		＜　背景モデル　：ＭＶＣモデル，マルチエージェントモデル＞
Ｃ．　利用者の主観的評価	：	ヒアリング（インタビュー，アンケート）
		＜　背景モデル　：認知モデルなど　＞

応用システムは、代替技術との比較／ユーザーインタフェース／対話の自然性などの諸観点から比較評価することが大切である。以下に、主な留意点を挙げる。

イ）システムの性能評価に関して　：　代替システムとの比較
既存システムからの移し替え(migration)　－　連続性

ロ）操作性の評価に関して　：　ＵＩの観点からの評価
①　操作の一貫性　　②システムの柔軟性（マルチモーダル，…）
③　対話の協調性（インタラクション）　　④初級者／上級者別評価

ハ）利用者の主観的評価に関して　：　対話の自然性（ユーザの満足度）
自由発話への対応／利用者の意図理解のレベル
応答戦略／インタラクションの深さ／…

<参考文献>：日本電子工業振興協会：ＯＡ機器の標準化に関する調査研究報告書
，４．５　「音声理解システムの評価法」，ｐｐ．１６４－１７０（１９９５．０３）．

音声言語情報処理 7−19−2
（1995. 7. 22）

インタラクティブシステムのユーザビリティ評価

浜田 洋
ＮＴＴヒューマンインタフェース研究所

1．はじめに

ユーザの立場で使いやすいシステムやサービスを実現することが、ユーザインタフェース評価の目的であり、種々の方法が提案されてきている。

従来、コンピュータやシステムとのインタラクションを実現するためのインタフェース“部品”の開発においては、ディスプレイにおける解像度やパタン認識技術における認識率などの指標が用いられてきた。これは、テストデータの質への依存性という問題はあるものの個々の技術性能を的確に表現しており、技術開発の段階では有効な指標であった。しかし、システムの使いやすさという視点で見ると、入出力部品は、サービスやシステムを実現するための一要素に過ぎず、個別“部品”の性能が良くても、サービスとして見たときに人間にとって使いやすいものであるとは限らない。さらに、マルチモーダルシステムになると、“部品”をどのように組み合わせ、統合していくかという新たな問題も加わってくる。

インタラクティブなシステムの開発において、実際に使われる前にユーザの立場から見て最適なシステムを設計することは難しい。現状ではユーザの要求や既存システムの問題点を分析・評価し、早い時期に設計に反映させ、プロトタイプ作成と評価を繰り返すことが最善策と考えられる。

本稿では、インタラクティブソフトウェアのユーザビリティ（使いやすさ）評価に関して、ISO/TC159 人間工学で進められている標準化の動向、および、評価手法の概要について述べる。

2．人間−システムインタラクションの国際標準

ISO/TC159/SC4「人間とシステムのインタラクション」では、ISO9241「VDTを用いたオフィス業務の人間工学的要件」の標準化を進めている[1,2]。この中で、ソフトウェアと人間とのインタラクションに関しては、

□対話の設計原則：人間とコンピュータシステムの対話設計に適用する人間工学的原則
□ユーザビリティ：ソフトウェアの使いやすさの定義と、評価のためのガイダンス
□情報表示法：情報の視覚的な表示法のガイドライン（画面割当方法、レイアウトなど）
□ユーザガイダンス：メッセージ、ヘルプなど、ユーザガイダンス提示法のガイドライン
□対話の設計ガイドライン：メニュー対話、コマンド対話、直接操作対話などの対話技法ごとのインタフェース設計ガイドライン

について記述されている。

（1）対話の設計原則

優れたインタフェースを持つ使いやすい対話ソフトウェアを実現するためには、そのソフトウェアが短期記憶容量やメンタルモデルの形成など人間の認知特性に適合している必要がある。その上でユーザの能力や適性を支援するシステムとすることで、ユーザ中心の対話システムが実現できる。対話システムの設計原則を以下に示す。

□タスクへの適合性：タスクの効率的・効果的達成のためにユーザを支援していること
□自己記述性：対話のステップがシステムからのフィードバックなどにより理解可能であること
□可制御性：目的が達成されるまでの間、ユーザが対話の流れ全体を制御可能であること
□ユーザの期待との一致：タスクに関する知識、教育、経験に対応していること
□エラーに対する許容度：誤った入力が行われても最小限の操作で意図した結果が得られること
□個人化への適合：タスクに対する個別の要求や個人のスキルに対応させるための変更が可能となるようにシステムが構成されていること
□学習容易性：ヘルプやガイダンスを提供することで、学習段階のユーザを支援していること

（2）ユーザビリティ

ユーザビリティとは、ユーザが対話システムを使用する際のインタラクションの質と捉えることができる。ISO9241では、ユーザビリティを、「*ある環境において、特定のユーザが、特定の目*

標を達成するために対話システムを効果的、効率的、満足にしようすることのできる度合い」と定義している。この時、

効果：目標を達成する際の正確さ・完成度

効率：目標を効果的に達成するために用いられたリソース

満足：使用の際の快適さ、受容性

である。ユーザビリティを測定する際には、達成しようとするタスクのゴール、またはサブゴール、およびユーザ、システム、使用環境などの対話システムの使用条件を明らかにした上で、目標を達成する際の効果、効率、満足を表す具体的な尺度で測定する。

使用条件やシステムを利用するタスク、さらに効果、効率、満足で表されるユーザビリティを明らかにすることは、使いやすいシステムを作ることだけではなく、システム開発者とユーザやシステム導入責任者とのコミュニケーションを活性化するという意味でも有効である。

3．対話ソフトウェアのユーザビリティ評価方法

ソフトウェア（インタフェース）の使いやすさの評価法には、いくつかのアプローチがある[3]。ISOで進められている標準化は、使いやすさの考え方と使いやすいソフトウェアを実現するためのインタフェースソフトウェアの作り方のガイドラインである。ここでは、ヒューマンモデルによる評価、チェックリストによる評価、プロトタイピングによる評価、について簡単に述べる。

（1）ヒューマンモデルによる評価

ヒューマンモデルによるインタフェース評価は、人の操作や認知処理過程をモデル化し、人とシステムとのインタラクションにおけるパフォーマンスを予測するものである。ヒューマンモデルとしては、Cardらの提案した知覚システム、認知システム、運動システムの３つで構成される人間情報処理モデル[4]が良く知られている。各プロセッサの間は短期記憶、長期記憶を介して情報のやりとりが行われる。Cardらはさらに、人の認知行動を目標(Goals)、オペレータ(Operators)、方法(Methods)、選択規則(Selection rules)で記述するGOMSモデルを提案している[4]。GOMSモデルを拡張した言語、NGOMSLを用いると、認知行動や運動を記述することで実行時間や作業負荷を予測することが可能となる[5]。

（2）ガイドラインとチェックリストによる評価

数々のインタフェースデザインに関する知見を体系化したものがガイドラインであり[6]、ISO9241でも対話技法ごとにインタフェース設計のガイドラインを作成している。これに対し、チェックリストは、ガイドラインの内容を評価のためのツールとして表現し直したものである。ガイドラインやチェックリストを設計や設計レビューに使用することで、ある程度のレベルの使いやすいソフトウェアが実現可能である。

（3）プロトタイピングによる評価

以上の評価手法は、設計プロセスにおいて行うことを想定した手法であるが、システム全体を通じて具体的な利用状況での使いやすさの評価を行うためには、システムのプロトタイプの作成とそれを用いた評価を行う必要がある。最近では、プロトタイピングや、実使用状況の記録から作業／操作プロトコルを分析するユーザビリティテストを効率的に行うためのツールが開発されている。

4．むすび

使いやすい対話ソフトウェアを実現するための「使いやすさ」の考え方と、ユーザビリティ評価方法について述べた。インタラクティブなシステムは、人間とシステムとのコミュニケーションの上に成り立っている。従って、機能や性能が優れているだけでなく、そのシステムを使うことが楽しく、繰り返し使いたくなるような、人間にとって魅力的なシステムを実現していることが今後ますます重要になると考えられる。

【参考文献】

[1]浜田　洋, 小川克彦,「人間-システムインタラクションと国際標準 －ユーザビリティと対話の設計原則－」, 人間工学, Vol.30, No.1 (1994)

[2]浜田　洋,「人にやさしいVDT:使いやすいソフトウェア」, 人間工学, Vol.31, 特別号 (1995)

[3]小川克彦,「３つのインタラクションデザイン」, 人間工学、Vol.30, No.1 (1994)

[4]S.K.Card, T.P.Moran and A.Newell, "The Psychology of Human-Computer Interaction", Erlbaum, Hillsdale (1983)

[5]D.E.Kieras, "Towards a Practical GOMS Model Mothodology for User Interface Design", In Handbook of Human-Computer Interaction, Elsevier, North-Holland (1988)

[6]S.L.Smith and J.N.Mosier, "Guidelines for Designing User Interface Software", Technical Report ESD-TR-86-278, Mitre, Bedfored (1986)

音声言語情報処理 7－19－3
（1995．7．22）

音声対話システムの評価法

中川 聖一

（豊橋技術科学大学・情報工学系）

1. はじめに

音声認識システムや対話システムの性能は、最終的には認識率（理解率）や対話達成率で評価されようが、認識アルゴリズムのような中心技術以外に認識率に及ぼす様々な要因があり、これらを含めた評価は難しい。音声認識システムの性能に影響を与える主な要因を挙げれば、[1),2)]

①認識語彙数・パープレキシティ
②特定話者か、多数話者か、不特定話者か
③孤立発声（単語単位）か、文節単位発声か、連続発声か
④朗読音声か、会話音声か
⑤発声環境は？
⑥マイクロフォンの位置はどこか？
⑦音声波の周波数帯域制限は？
⑧その他：（汎用性は？使い勝手は？経済性は？認識速度は？言語モデルのカバー率は？棄却率は？）

一方、音声対話システムは、音声認識システム（ディクテーション止りのシステム）に比べて、はるかにオープンシステムであり、ユーザとシステムとの相互作用のあるシステムであるため、システムの性能に影響を与える様々な要因があり、客観的評価以外にユーザの主観による評価も重要になり、その評価は難しい。

2. 不要語と未知語などの ill-formedness の評価

音声対話システムにおける、入力音声には、間投詞、助詞落ち、倒置、言い淀み、言い直し、体言止め、未知語などの現象を伴い、テキスト入力文と比べて、極めて ill-formed な文となる。

まず、間投詞や言い淀み、言い直し、等の不要語処理および辞書に登録してない未知語処理に対する評価項目を挙げる。[2)]

①不要語・未知語の出現率、
②不要語・未知語の検出率（不要語・未知語の正検出数／不要語・未知語の発声数）、
③適合率（不要語・未知語の正検出数／不要語・未知語として検出された総数）

①が大きいことは、非協力的な発声を多く許しているか、システムのカバー率が小さいかのどちらかであり、前者と後者では評価が逆になる。②と③に関しては、不要語・未知語を任意のモデルの連結で表現することになり、語彙数はみかけ上無限大になるが、話者照合と同じく、ある音声区間が登録単語であるか否かの判定のため、その性能はみかけの語彙数には関係なくなる。このことから、不要語・未知語処理を考慮した認識率は、考慮しなかった場合の認識率に単語照合率を乗じた値になると予想できるが、理論式の導出は難しい。[3)]

最近の研究動向としては、認識対象語彙数を数万語に増やし、カバー率を上げる（未知語出現率を下げる）傾向が強い。このためには、確率モデルの作成のための膨大な言語データベース、処理時間が要求される。我々は、語彙数は中規模にして、あとは未知語処理で棄却し、言語処理部で大規模な辞書を用いて復元する手法も有力なアプローチだと考えている。

ill-formedness の発話をどの程度うまく扱えるかどうかを評価する方法としては、機械翻訳システムの評価で提案されている構造処理能力評価法が参考になる。[4)] 例えば、①まず、ある評価対象の音声対話システムのタスクで発声され得る代表的な内容に対して、様々な構文を含めた評価文（勿論、話し言葉）を作成する。次に、②助詞落ち、間投詞、倒置、言い淀み、言い直し、未知語などが単独に生じる文を作成する。③これらが複合して生じる文を作成する。④以上の文を自然に発話したものを対話システムに入力し、システムの応答を評価する。

3. 対話システムの評価項目[1)]

対話システムは、音声認識部、言語理解部、問題解決部（応答文作成部）、対話管理部、音声合成部などから構成される。これらが一体混然となったシステムも多く、各部を独立

に評価するのが難しい場合もある。しかし、少なくとも音声入力から出力までトータルに評価する(Speech Input)以外に、音声入力部の性能が100%と仮定して、テキスト入力から出力までを評価する(Text-Input)ことは可能であろう。Text-Input の場合でも、間投詞、言い直し、言い淀み等をそのまま書き起こして入力する場合と除去して入力する場合が考えられる。音声認識部で間投詞や言い直し部分の検出とその他の認識を同時に行なうシステムが多い。この場合は、テキスト入力として、不要語を除去してもおかしくない。

対話システムの評価は、対話の達成率が基準になるが、それだけではない。要は、自然に少ない回数のやり取りで達成されるかどうかである。[5]情報案内や検索などのタスクにおいては、ユーザの発話回数はせいぜい10回程度であろう。このとき、平均文認識率が80%から85%に向上してもあまり効果はない。それよりも、ユーザに負担をかけない自然対話のやり取りができることの方が重要である。最終的には、ユーザの満足度、即ち、主観評価によらざるを得ない。しかし、満足度と強い相関がある客観的な評価項目が存在すると思われる。[6),7)]

主な評価項目を以下に挙げる[2),8)]

①自然な発話（不要語などの非文、多様な言い回し）が入力できるかどうか
②入力文をどの程度正しく理解できるか？
③システムの入力文のカバー率は？
④システムが理解できなかったとき、どう対処するのか
⑤確認の応答・再入力・言い換えなどの回数はどれぐらい必要か
⑥あいづちや割り込みが可能かどうか
⑦あいまいな入力文に対処できるかどうか
⑧対話の主導権はどちらか
⑨システムの応答文は自然かどうか
⑩システムの音声合成の品質は十分かどうか

この他、対話の評価実験結果に影響を与える項目として、[9)]

⑪ユーザがどの程度、システムに熟知しているか（初心者か専門家か）
⑫入力モダリティとして、音声以外に何が用いられるか（キーボード、マウス、顔画像）
⑬出力モダリティとして、音声以外に何が用いられるか（ディスプレイ、途中結果の表示は？）
⑭システムの応答速度

などが考えられる。

4. タスクの複雑度と情報量

異なるタスクを扱うシステムを比較する場合、システムのカバーレイジ、処理時間、タスク達成率を単に比較することは出来ない。異なるタスク間の比較のために、タスクの複雑さというメジャーを考える必要がある。一例を紹介する。[6)]

＜タスクの複雑さの定義＞

・ある部分対話の出力（応答）Y_i が得られる直前の入力 X_i の不確かさの程度を $H(X_i)$、出力 Y_i が得られた直後の入力 X_i の不確かさの程度を $H(X_i \mid Y_i)$ とするとタスクの複雑さTC は、

$$TC = \frac{1}{I}\sum_{i=1}^{I}\{H(X_i) - H(X_i \mid Y_i)\} \quad (1)$$

で求められる。ここでは、I はタスク内の部分対話の数である（I で正規化しない定義もありうる）。

・タスクの複雑さはシステムと独立である。

このようにタスクの複雑さを部分対話当たりの相互情報量で表わすことにより、システムの性能を単位時間当たりに得られる情報量として定義できる。

しかしこの定義の方法もいくつかの問題点を抱えている。それは、主に次の２点である。

1. 部分対話をどのようにして定義するのか。
2. 部分対話の情報量をどう求めるのか。

参考文献

1)中川 聖一:音声認識・理解システムの評価とデータベース、電子情報通信学会誌、Vol. 73, No. 12, pp. 1304-1310(1991)

2)中川 聖一:音声認識システム・音声対話システムの評価法、学振第152委員会「文字言語・音声言語の知能的処理」第33回資料(1993. 7)

3)A. Kai and S. Nakagawa:Relationship among recognition rate, rejection rate and false alarm rate in a spoken word recognition, IEICE Trans. Inf. & Syst. Vol. E78-7, No. 6 (1995)

4)成田 一編、天野、村木:こうすれば使える機械翻訳、バベル・プレス(1994)

5)新美、小林:音声認識の誤りを考慮した対話制御方式のモデル化、情報処理学会、音声言語情報処理、研究報告、SLP5-7(1995. 2)

6)山本（誠）、中川:音声対話システムの構成法とユーザ発話の関係の模擬実験による評価、情報処理学会、音声言語情報処理学会、研究報告、SLP5-9(1995)

7)小林、中島、新美:音声模擬対話における対話制御と快適性について、情報処理学会、音声言語情報処理、研究報告、SLP6-2(1995. 5)

8)E. Grbino, P. Baggia, A. Ciaramella and C. Rullent :Test and Evaluation of a Spoken Dialogue System, Proc. ICASSP, Vol. II, pp. 135-138(1993)

9)N. M. Fraser:A standard reporting framework for interactive dialogue systems, ESCA Workshop on Spoken Dialogue Systems, 当日配付資料(1995)

音声対話システムの評価法について

古井 貞熙
NTTヒューマンインタフェース研究所
東京工業大学大学院情報理工学研究科
e-mail: furui@splab.hil.ntt.jp

はじめに

音声対話システムの評価法の確立の重要性を説くのは容易であるが、それを具体的に提案するのは極めて難しい。一方、その難しさを論ずるのは極めて容易である。先月6月にデンマークで開かれたESCA Workshop on Spoken Dialogue Systems -- Theory and Applications -- でも、評価法のセッションが持たれ、種々の議論が行なわれたが、結論は得られなかった。その中で、タスクを決めなければ評価などできるものではない、などという不毛な議論もあったが、筆者の興味を引いた提案として、英国のDRAのRoger Mooreなどを中心として活動しているEAGLES Spoken Language Working Groupの「研究報告の標準的枠組み」[1]の提案があった。ここではその要点を紹介して、議論の種にしたい。なお、この提案の詳細は、9月のEurospeechに合わせて、"EAGLES Spoken Language Handbook"として出版される予定である。

「研究報告の標準的枠組み（A Standard Reporting Framework）」の提案

この基本的考え方は、音声対話システムを厳密に比較評価する方法を確立するのは当面困難であるので、その基本的な第一歩として、システムの開発者がそのシステムに関する実験結果を、標準の枠組みで発表するようにしたらどうか、というものである。すなわち、何を、どのように報告すべきかのガイドラインを決めよう、というものである。完璧を期すのではなく、音声対話システムの性能と評価条件を報告するための最低限の共通の標準を作ろうという姿勢で作られている。

具体的な報告枠組みは、Figure 1に示す通りである。報告者はこの表の各欄に数値を記入するようになっており、できるだけ多くの欄に記入することが求められている。各欄の目的と、記入すべき値の範囲については、研究会の当日に説明したい。

[1] Norman M. Fraser: "A Standard Reporting Framework for Interactive Dialogue Systems", Draft (May 1995)

PARAMETER	VALUE
Type of users	
Number of users	
Input modality	
Output modality	
Number of dialogues	
Number of tasks	
Dialogue type	
Average turns per dialogue	
Average dialogue duration	
Average turn delay	
Dialogue success rate	
Task success rate	
Crash rate	

Figure 1: a reporting framework for interactive dialogue systems

インタラクティブなシステムの評価をどう考えていくか

平井 誠
松下電器中央研究所
hirai@crl.mei.co.jp

1．はじめに

広義の対話型システムの性能は、一般ユーザから見たシステムの操作性、透明性、頑健性等の因子に依存するため、その客観的な評価は難しい。特に、音声言語に代表される自然言語対話機能を有するシステムの場合は、自然言語対話自体の構造的な複雑さのためにシステム全体の性能評価は一層困難になる。また、一般に対話型システムは何らかの特定のタスクを遂行するように設計されているため、タスクの異なるシステム間では能力比較が単純にはできない。ここでは自然言語処理における対話モデルの考え方を多少一般化して対話型システムに適用する事により、評価に関するこれらの課題を考察する。

2．能力評価と運用評価

チョムスキーは人間の言語機能を（先天的）能力(Competence)と運用（Performance)に分割し、生成文法では記述対象を能力のみに限定した。その結果、問題が簡略化されたわけだが、対話型システムの評価にもこれと同様の考え方を導入して、評価観点を整理する事ができる。

音声対話システムの（上記意味での）能力はシステムの語彙、文法、認識率、応答発話パターン、タスク・ドメイン知識等に相当し、マルチモーダル対話システムでは、音声以外のチャネル（マウス、ＧＵＩ、キーボード等）と各チャネルで伝達可能な情報形態がこれに加わる。他方、対話型システムの運用とは、対話の各時点でのシステムの反応様式、具体的には発話内容や使用チャネルの選択であり、換言すれば広義の対話管理機能である。視点を変えれば、能力はシステムの静的な側面であり、運用は動的な側面である。

システム機能をこのように２つの側面に分割する事により、特性の異なる機能を独立に評価することができる。能力評価は、システムを構成するサブシステムや種々のデータの個別的な評価であり、数値的あるいは統計的な評価が可能な項目も多いと思われる。この評価はシステム設計者、開発者にとって意味のあるものであるが、システムのユーザには不透明な部分である。他方、運用評価はシステムのユーザに直結したものであり、実際の使用状況においては能力評価より重要であるが、評価法や評価項目に関しては能力評価とは異なった語用論的な観点からのアプローチが必要となる。これについては次節以降で考察する。

１例としてこの２つの観点から認識率を見てみると、いわゆる音声認識率は能力認識率である。システムのユーザにとっては、この能力認識率は意味がなくシステムの挙動からみた全体的な発話の認識率、いわば運用認識率が重要である。

3．一貫性と結束性

運用評価は対話管理機能の評価であり、対話の一般的な制約に基く評価が必要となる。目的指向の協調的な（自然言語）対話においては、常に一貫性(Coherence)と結束性（Cohesion)が保持される必要がある。一貫性とは対話全体が目的達成に向けて合理的に推移する事であり、結束性とは対話内の隣接発話が構文・意味的に関連している事である。これらの制約は自然言語以外のチャネルを持つ対話型システムにおいても保持されるべきであり、機能評価における重要な観点である。

特定のタスクを遂行する協調的対話型システムは概念的にタスク遂行機能と対話機能の２つに分割できる。抽象的には、目的指向型タスクはプランオペレータの集合として定義でき、タスクの遂行は特定の条件下でこれらを例化しながら目標－副目標木を生成する過程として捉え

ることができる。つまり、タスク記述が対話の一貫性を定義し、タスク遂行機能がこれに従って一貫性を管理していると言える。従って、システム評価の観点から重要な点は、ユーザ発話が一貫性を崩す場合のシステムの軌道修正能力である。

他方、対話機能はユーザとの直接のやりとりであり、対話の結束性管理と言える。システム評価の点で重要なのは次の２点である。第１はシステムの質問に対するユーザの応答の認識である。人間同志の対話における応答発話は直接的なものから間接的なものまで極めて多様であり、これらをどの程度システムが認識可能かが頑健性の１つの因子となる。第２は、この反対の場合で、ユーザの情報要求に対するシステムの応答である。第１点と同様に多様な応答発話が可能であり、システムがどの程度対話文脈やユーザの知識状態に応じた発話を生成可能かが性能を大きく左右する。

結束性管理には、いづれの場合にも隣接発話の結束性に関する構文・意味・語用論的モデルが必要であるが、本格的な研究の歴史は浅く、全体的に満足できるモデルは提案されていないが、現時点で提案されているモデルによるシステム評価は可能であろう。

４．会話公準のマルチモーダル化

対話の一貫性は主としてタスク構造に依存するので、タスク構造が明確であればそれだけ一貫性管理も容易になる。一方、結束性は語彙面ではタスクよりドメインに強く依存するが、応答発話のパターンの多様性はこれらとは独立して対話一般に観察される。従って、その適切なモデル化を行なえば、タスク・ドメインから独立したシステム評価法が導入できる。また、結束性モデルはユーザに許容される応答パターンの範囲を規定する制約でもあり、ユーザから見たシステムの柔軟性や頑健性の判断基準ともなる。従って、システム評価およびシステムの高性能化の両面から結束性モデルの研究が重要となる。

結束性に関する今ひとつの課題は、結束性の概念のマルチモーダル化である。当然ながら、これまでの結束性の研究は自然言語対話の言語表現を対象にしている。しかしながら、言語以外のチャネルを許すマルチモーダル対話システムとの対話を考えた場合は、全てのモダリティを含めた結束性モデルが必要である。

結束性の基本は、Ｇｒｉｃｅの協調原理も基づく４つの会話公準：

１．充分な情報を伝達せよ（量の公準）
２．誤りなく伝達せよ（質の公準）
３．礼儀正しく伝達せよ（関係の公準）
４．簡潔に伝達せよ（様式の公準）

であり、これらは言語以外の手段にも適用可能である。画像、音声、キーボード等を統合した結束性モデルがシステムの対話機能の総合的評価に不可欠である。

５．結論にかえて

機能評価と運用評価という観点から対話型システムの評価法について述べたが、ここで２つの例にこの考え方を適用してみよう。第１はチューリングテストを意識した ＥＬＩＺＡ[1] システムである。このシステムは精神科医のカウンセリングを模擬するものであり、対話自体には明確な目標がなく、ただ自由な対話を自然に続ける。つまり、結束性管理のみを持ったシステムと言える。また、この意味でチューリングテストは結束性管理のテストとも言える。第２は、ソフトウェア科学会の研究会で企画されている「DiaLeague」である。これは、機能評価と運用評価を同時に行っているため、対話機能の評価という点では不明瞭であり、また協調的対話システム同志のリーグ戦という考え方はある種矛盾であり、個々の対話システム単独の機能評価ができない事が難点である。

対話によって、計算機システムではなく人間を評価しようというのが面接や面談であるが、対話システムの運用評価も一貫性と結束性モデルに関するチェックリストを横目にシステムを面接するというのも、最終的には、１つの有効な手法になり得ると思われる。

文献

1) Weizenbaum J.: ELIZA, CACM, Vol. 9, No. 1, '66