音声による対話システムの評価法における一考察

6G－5　山本誠治　山本幹雄　中川聖一
豊橋技術科学大学

1　はじめに

近年様々なタスクを扱った多くの音声対話システムが試作されている。しかしそれらのシステムの評価方法はシステム構成やタスクに依存したものが多く、異なる機関で設計されたシステムに対して公平な評価が行なえるような明確な方法はまだ確立されていない。

本報告では、これらの音声対話システムの一般的な評価方法についての一考察を述べる。更に“Wizard of Oz”法を用いて、音声対話システムの構成法とユーザの評価の関係について考察した結果も述べる。

2　システムの評価法と問題点

ある音声対話システムの性能を考える場合、システムのカバーレイジ、処理時間、正解率（タスク達成率）などは、システムの性能を知る1つの方法である。もし全く同じタスクを扱うシステム同士の比較を行なうのであればこのようなメジャーは有効であるが、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、正解率を単に比較することは出来ない。我々は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ている。

まず最初に考えた方法は、タスクの複雑さを意味表現の数で表すという方法である。これは、対話の仕方や言い回しなどに影響されないという利点がある。しかしどのような意味表現を用いるのか、またどのようにして意味表現の数を求めるのか、というような様々な問題点があげられる。そこで次に別の方法として、タスク内で現れる対話数でその複雑さを表そうと検討した。タスクの複雑さを対話数で考えるのであれば、タスクは対話の集合として簡単に定義できる。しかし表層レベルの対話で考えると、言い回しがタスクの複雑さに考慮され（我々は、言い回しはシステム側の問題と見ている）、深層レベルの対話で考えると先ほどの意味表現の場合と同じような問題点が考えられる。その後種々の検討を重ねた結果、タスクの複雑さを次のように考えることにした。

タスクの複雑さとは、“システムがそのタスクを処理するのがどれほど困難か”という度合を表すメジャーである。曖昧さが大きい対話ほどシステムにとっては処理が困難であると考えれば、タスク内の対話で得られる情報量が大きいほどタスクは複雑であるといえる。そのため我々はタスクを部分対話（1つの情報が得られる最小単位の対話）の集合で定義し、タスクの複雑さを一部分対話当たりの平均情報量で定義した。

＜タスクの定義＞

- タスクとは“ある情報を得るとそれに対する情報を返す”というような部分対話の集合である。
- タスクはシステムと独立である。

＜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 $H(X_i|Y_i)$ とするとタスクの複雑さ TC は、

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

で求められる。ここで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一部分対話当たりの相互情報量で表すことにより、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以下のように定義した。

＜システムの評価法＞

- 目的とするシステムで扱っているタスクの複雑さをTC、テスト文中の部分対話の数をN、達成率（正解率）をCT、全てのテスト文が終了するまでの時間をtとすると、システムの性能SPは、

$$SP = \frac{(TC\times N)\times CT}{t} \qquad (2)$$

で表される。

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、

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のようにして求めるのか。

などである。このような問題点を今後検討していくつもりである。

A Study on an Evaluation Method for Spoken Dialog Systems, Seiji Yamamoto, Mikio Yamamoto and Seiichi Nakagawa, Toyohashi University of Technology

3 Wizard of Oz 法を用いた評価実験

この節では、Wizard of Oz 法を用いた評価実験について述べる。この実験の目的は、音声対話システムにおける対話の傾向を知ることである。対話の傾向を知ることは、音声対話システムの評価法を研究していく上で、重要であると考えられる。

本実験では、システムの対話方法の違いや、タスクの違いによる被験者の振舞い、主観などを検討するため、タスクおよびシステムの構成法を次のように設定し、これらの組合せで実験を行なった。

1. タスクは、“富士五湖周辺の宿泊施設案内” と “航空座席予約” の２つを考える。
2. システムの構成法としては、システム主導型とユーザ主導型を考える。

なお本実験で用いる音声対話システムでは、入力音声の認識から応答文の作成までを人間が行ない、応答の出力だけを機械（音声合成システム）が行なった。

システムの使用方法や実験に対する指示を正確にかつ各被験者に対して同じように説明するため、本実験では、文書により被験者に必要な指示を与え（２度読んでもらう）、ビデオによりシステムの使用方法（実際の実行例の一部）を示した。その後で以下に示すようなシナリオに沿ってシステムを使用してもらい、アンケートに答えてもらった。

- 宿泊施設案内では２泊３日の旅行を行なうと仮定し、その時に利用したい宿泊施設について調べる。
- 航空座席予約では、２ヶ所の都市訪問のために航空チケットを予約をする。

同じタスクを扱うシステムで対話方法を変えて実験を行なう場合、後から使用する対話方法の方が有利になる可能性がある。そこでシステムを使用する順序（ユーザ主導型からするか、システム主導型からするか）を変えて実験を行なった。

4 実験結果

本実験では１つの条件で６人ずつ（システムを使用する順番によりさらに２つに分けられる）、合計１８人で実験を行ない、1329 対話のデータを収集した。実験により得られた結果は以下に示す通りである（user_time とは、システムの応答が終了してから被験者が質問を終了するまでの時間である）。

1. システム主導型システムの方がユーザ主導型システムより対話数が多くなる（表１参照）。
2. ユーザ主導型システムの user_time は、システム主導型システムの user_time よりも長くなる。これは被験者が発話を開始するまでの時間が長くなりがちであったからである。
3. システム主導型の後にユーザ主導型のシステムを使用した場合の user_time の差は、その逆の順序で使用した場合よりも小さくなる。

なお異なるタスクを扱うシステムに対しても同様の実験を行なったが（対話方法は両タスクともシステム主導型）、平均対話数、user_time ともほぼ同じような結果になった。そのため、タスクの違いに関するアンケートからは、特徴的な結果が得られなかった。

表 1: 平均対話数と user_time

タスク	宿泊施設案内		航空座席予約	
sys 主導/usr 主導	sys	usr	sys	usr
平均対話数（文）	33.3	16.3	53.0	31.7
user_time（秒）	8.5	15.0	5.7	11.8

またアンケートにより次のような結果が得られた。

- すぐ使えるという点では、システムに対して何も知らない被験者でも、受身的に対話を進められるのでシステム主導型の方がよい。
- ユーザ主導型の方は被験者の方から自由に質問できるので、ある程度慣れればユーザ主導型の方が使い易い。
- ユーザ主導型の方がシステム主導型よりも知的である（人間に近い）。

5 まとめ

本報告の前半では、異なるタスクを扱うシステムの評価方法について報告した。タスクの複雑さを情報量を用いて定義し、それをもとにシステムの性能を求める方法を示した。先にも述べた通り、この方法にはまだ解決されなければならないいくつかの問題点がある。しかしこのような異なるタスクを扱うシステムの評価方法を検討することは、音声対話システムの研究、作成を行なう上で大変重要である。

本報告の後半では、システムの対話方法やタスクの違いによる被験者の振舞い、主観などを検討した。３つの条件で１８人の被験者により実験を行なったが、使用したシステムに人間が介在していると気づいた被験者は誰もいなかった。最初はこの実験による結果をシステムの評価方法に利用しようと考えていたが、あいにく利用できるような結果を得ることは出来なかった。しかし、1329 対話からなる実験データは、今後の研究に役に立つと考えている。