「電総研道案内対話音声コーパス(1998)」

ETL Spoken Dialog Corpus (Town Guidance Task, Japanese)

この言語資源（DVD）は、言語資源協会（GSK）から公開するにあたり、オリジナルのコーパスデータについて、ファイル形式を変換し、ファイル構成を変更するなどしたものである。

本解説書は、コーパス作成者によるオリジナルの Readme を基に、データの現状に合わせ、言語資源協会（GSK）が修正、増補した。オリジナルの Readme は、Readme_original.txt として本 DVD に収録している。

1　概要

このコーパスは、Wizard of Oz (WOZ) 法によって、自動推論エンジンを実装した機械と人間との間の、道案内についての対話を記録したものである。人間と機械の間の自然なやりとりを可能にさせる要素、たとえば、発話の番の交換・うなずき・割り込み・割り込みへの適切な対応などを分析できるように設計されている。33 名の話者による 162 対話のデータを含んでおり、対話データは全部で 1000 分以上になる。本コーパスは、音声データ・書き起こし・発話の始端と終端・発話の意味表現からなる。

なお収録データに関する詳しい記述は、以下の文献を参照されたい。

[1] Katunobu Itou, Tomoyoshi Akiba, Osamu Hasegawa, Satoru Hayamizu, and Kazuyo Tanaka: A Japanese spontaneous speech corpus collected using automatically inferencing Wizard of OZ system, The Journal of the Acoustical Society of Japan (E), vol.20, No.3, pp. 207-214, 1999.

[2] 伊藤克亘, 秋葉友良, 長谷川修, 速水悟, 田中和世, "音声対話システム構築のための実対話データ収録実験", 情報処理学会研究報告, SLP-02-6, pp. 35-42, 1994.

2　コーパスの基本設計

ここでは、人間と機械のやりとりを可能にする要素を明らかにするためのデータを収録することを目的とした。現状程度の実システムでは、余りにも制限が大きく対話の流れや発話のバリエーションが得られないため、ここでは、ユーザの発話を聞き取り意味表現に変換したものをオペレータ（Wizard）が入力し、問題解決と発話生成の部分を対話プログラムが担当する WOZ 方式のシステムを構築した。

このシステムでは道案内を題材に対話をおこなうようになっているが、道案内システム

の開発を目的とはしていないので、道案内の戦略などの最適性などは問題にしない。そのため、システムの応答は合成音声のみとし、地図などは表示しない。また被験者に対してシステムの利用方法などの事前の指導は極力おこなわないものとした。

3　データ収録

3.1 概要

データの収録実験では、渋谷のレストラン、デパートなどの道案内に関する対話を、被験者一人につき 5 対話ずつ実施した。収録は 1994 年 2 月から 3 月にかけて、40 名の被験者を募集しておこなった。本 DVD に収められているのは男性 18 名、女性 15 名の計 33 名分で、大半は 20 代の学生である。一人を除いて対話システムの利用経験はなかった。

収録は、説明の開始からアンケートの終了まで、ほぼ 1 時間 30 分を要した。実際にシステムと対話している時間は 1 名あたり 17 分から 68 分で、平均約 33 分であった。

3.2 収録環境

収録環境を図 1 に示す。オペレータ側と被験者（ユーザ）側はそれぞれ別の部屋であり、オペレータがいる部屋は被験者から全く見えないようになっている。

オペレータ側には、被験者の音声を流すスピーカ(a)と、被験者をモニタするディスプレイ(c)を用意した。対話システムの操作は、ワークステーション(b)のキーボードを利用しておこなった。

被験者側には、被験者をモニタするためのビデオカメラ(d)を 3 台設置した。ワークステーション(f)には、課題表示用のディスプレイとシステムの応答を流すスピーカー(e)を取り付けた。マイクはスタンド付きのハンドマイク(g)とし、このマイクへの入力と、音声合成器の出力を、DAT の別チャネルに同時に収録した。

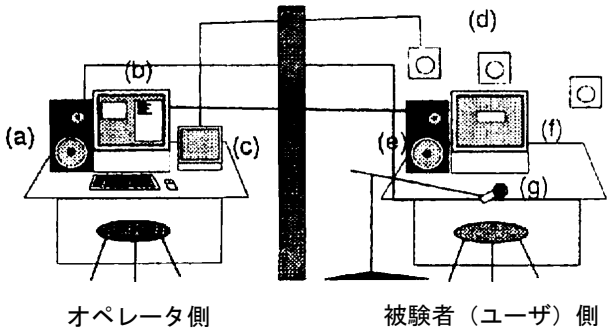

図 1　収録環境

3.3 被験者への教示

収録に先立ち、被験者には別室で図 2 の文章を見せた。システムが知っている内容や扱える言い回しなどは一切教えないものとした。また収録中にメモを取ることも禁止した。

今回、使っていただくシステムは、渋谷の道案内をするシステムです。ディスプレイに案内できる場所（飲食店、デパートなど）の一覧が表示されています。このシステムには、これらの場所について、タウンマップから抽出した情報が入力されています。

このシステムを使って、5 つの課題を順番におこなっていただきます。課題に必要な情報は、全て、このシステムから聞き出せるようになっています。

各々の課題は、10 分程度です。また、課題と課題の間には、数分間程度休憩があります。また、実験の最初に、カメラ・マイクの調整と学術データとしての写真撮影（証明写真程度）をおこないます。終了後には、簡単なアンケートをおこないます。なお、最後の課題では、それまでの課題で説明した場所について質問します。

図 2　被験者指導用紙

対話は、課題ごとに文章をディスプレイに表示し、その指示に従って開始してもらった。課題の例を図 3 に示す。

［課題 1］ あなたからシステムに話しかけて下さい。 そして、東急ハンズの場所をたずねて下さい。
［課題 2］ システムから、はなしかけます。 それから、喫茶店、ファーストフードなどの店の行き方を、いくつかお尋ね下さい。
［課題 3］ あなたからシステムにはなしかけて下さい。 そして、ＱＵＩＮＣＡＭＰＯＩＸ（カンカンポア）、ダル・ボロニェーゼ、ＯＲＲＢ（オーブ）の場所を尋ねて下さい。これらの店の駅からの距離、取り扱っている物などから、自分の好みの店を探してみて下さい。
［課題 4］ 自由に、何か所でも構いませんから、お尋ね下さい。（案内できる場所の一覧を下に表示します。しばらくしても一覧が表示されない場合は、「出ません」とマイクにむかってしゃべって下さい。）
［課題 5］ これまでに、システムがあなたに案内した場所のうち、よく覚えている 2 か所について、システムに行き方を説明して下さい。

図 3　課題の一覧（例）

被験者によってキーワードや固有名詞は変更している。被験者から話しかける課題では、なかなか話しかけない場合はシステムから対話を開始した。［課題 5］はそれまでと違いシステムから被験者に場所を尋ねるものとした。この課題 5 に関しては、対話プログラムを使わず、オペレータが応答文の決定までおこなった。

4　対話システムの概要

4.1 処理の流れ

収録に用いた WOZ システムでは、ユーザの発話を聞き取り意味表現に変換する部分を人間（オペレータ）が担当し、問題解決と発話生成の部分は対話プログラムが担当した。

プログラムへの入力は、例えば「東急ハンズの場所を教えて」という発話の場合、以下のような意味形式に変換される。

```
want(A,inform_ref(B,A,C)),[i(B),you(A),isa(C,location),
location(D,C),name(D,E),value(E,ハンズ)]
```

“i” はシステム、“you” はユーザを表す。“A” や “B” など大文字アルファベット一文字の語は変数である。実際の収録時には、入力にかかる時間を短縮するために、

```
rrp(ハンズ)
```

という短縮型のコマンドを用いた。このコマンドは 83 種類用意したが、利用されたのは付録の表 1 に示す 33 種類だった。各収録対話で入力されたコマンドと、それに対応する意味表現についても、テキストデータとして本コーパスに収められている。

対話プログラムは、入力からユーザモデルを更新し、プランニングをおこなってシステムの行為を決定する。道案内は、保持している交差点・建物・通り・店舗などの各場所のデータと、それらに隣接する街路などとの関係を用いて、中間地点の数などのスコアに基づき最適な経路を選択しておこなう。また、道案内以外にも、店舗の種類や移動に必要な時間なども答えられるようになっている。

システムの発話は、漢字かな混じり文として生成され、漢字の読みやアクセントをプログラムで付与し、市販の音声合成装置から出力される。

4.2 オペレータの役割

オペレータの主な役割はユーザの発話から対話プログラムの入力を作成するところであるが、ユーザの発話が省略などを含む場合や不完全な場合でも、それらをオペレータが解消することはなく、対話プログラムがそれらの処理をおこなった。したがって、自然言語処理の面から見ると、WOZ システムとしては自動化の度合いが高いと言える。

ただし、発話がプログラムの能力を超える場合には、「理解できません」といった内容の発話を生成するようにした。

また、ユーザの明示的な発話がなくても、例えばモニタを介して肯定するような表情やうなずきが観察できれば、オペレータの判断で「はい」と同等の入力をおこなう、といったことを許した。

課題 5 ではシステムの応答もオペレータが決定しているが、対話がなかなか進まない場合は、「大体でいいですから、説明して下さい」と促したり、「それは何の店ですか」と説明を続けさせたりと、ユーザになんとか発話させるようにした。

4.3 システムの応答設定

対話を円滑に進める要素について調べるために、システムからの間投詞や確認の発話、視覚の存在を示すような発話を、積極的にユーザに投げかけるようにした。

間投詞については、「えーと」と「えー」というふたつの間投詞を説明文の読点の後と発話の最初に任意に挿入できるようにしておき、被験者ごとに、課題によって、生成確率を 0 ないし 0.5 に設定した。

「東急ハンズですね。」といった確認の発話についても、被験者・課題によって、生成確率を 0，0.5，1.0 のいずれかの値に設定して収録をおこなった。

ユーザの発話からシステムの応答までの時間は、最短で数秒、最大で 20～30 秒要した。間投詞と確認の発話を発話の最初におこなう場合は、推論などを介さずになるべく早く応答するようにしたため、通常の応答を生成する場合に比べて、システムの反応時間が短めになっている。

5　本コーパスの構成

本 DVD のファイル構成は以下の通りである。

```
├── Readme.pdf          … このファイル
├── Readme_original.txt … オリジナル（1998 年発行）の Readme
│
├── SPEECH/             … 音声データ
│
├── TRANS/              … 書き起こしテキストデータ
│
└── SEM/                … 意味情報テキストデータ
```

SPEECH，TRANS，SEM の各フォルダに含まれるファイルは、それぞれ以下の規則に従って名前がつけられている。

〈収録日〉-〈被験者番号〉-〈課題番号〉.〈拡張子〉

例えば、ファイル “940216-1-3.wav” には、1994 年 2 月 16 日収録、一番目の被験者、課題 3 の音声データ、が収録されている。

6　ファイルフォーマット

6.1 音声データ

音声データは一つの課題を形成する一つの対話全体を 1 ファイルとした。フォーマットは、サンプリングレート 16kHz、16bit 量子化、ステレオ（2 チャンネル）の wav 形式である。サンプリングレート 48kHz で録音したものをダウンサンプリングしている。左チャンネルにユーザ発話、右チャンネルにシステム発話が収録されている。

6.2 書き起こしテキストデータ

発話の内容を書き起こしたテキストデータを作成した。発話単位は 300msec 以上の無音によって区切られた区間とし、自動的に切り出した。そのため、言語的な単位として適当ではないものも含まれる。

テキストファイルのフォーマットは UTF-8 形式で、各行に一つの発話単位に関する情報を記述してある。各行は、スラッシュ(“/”)で区切られた 7 つのフィールドから構成される。各フィールドの詳細は以下の通り。

(1) **通し番号**

(2) **境界情報**

切り出された区間のユーザ発話の音響的な状態を示す以下のフラグ。

“00”　音響的に問題ない

“01”　先頭に誤りを含む（システム発話の重畳など）

“02”　末尾に誤りを含む

“03”　両方に誤りを含む

“99”　対象外の発話（システム発話）

(3) **話者**

“U” または “S” で、それぞれ “ユーザ”、“システム” を表す。

(4) **開始時刻**

(5) **終了時刻**

収録開始からの msec。ユーザ発話に対してのみ記述されている。

(6) **書き起こし（かな漢字）**

(7) **書き起こし（ローマ字）**

発話単位中にオーバラップした相手側の発話は、“{ }” で囲んで記述されている。

6.3 意味情報テキストデータ

オペレータが選択した、ユーザの発話に対応するコマンドと、その意味表現が記述されている。ファイルのフォーマットは、各行に一つのコマンドに関する情報を記述した UTF-8 形式のテキストファイルで、拡張子は “sem” である。

各行は、スラッシュ(“/”)で区切られた 3 つのフィールドから構成される。各フィールドの詳細は以下の通り。

(1) **通し番号のリスト**

対応する発話単位のリスト。書き起こしテキストデータ中での通し番号が、コンマ(“,”)で区切って記述される。発話単位を無音区切りとしたため、一つの意味表現が複数の発話単位にまたがって構成されている場合がある。

(2) **コマンド**

オペレータが入力したコマンド。コマンドの意味は付録の表 1 を参照のこと。

(3) **意味表現**

コマンドに対応するユーザ発話の意味表現。形式は以下の通り。

〈主要意味表現〉, [〈付加情報〉, . . .]

意味表現中に現れる “A” や “B” など大文字アルファベット一文字の語は変数を表す。同一意味表現中に現れる同じ名前の変数は同じ値を取る。意味表現に使用した構成要素の意味は付録の表 2 を参照のこと。

付録

表 1 コマンド表

コマンド名	対応する発話
Yes	はい
No	いいえ
rrp(P)	P の場所を教えて
rrp(P1,P2)	P1 から P2 への行き方を教えて
Rrp	その場所を教えて
rrpc(C)	そのクラス C の場所を教えて
rrs(P)	P は何の店か
Rrs	それは何の店
rrS(S)	種類 S の店には何があるか
rrt(P1,P2)	P1 から P2 までの時間を教えて
rrt(P)	P からの時間を教えて
Rrt	時間を教えて
rrd(P1,P2)	P1 から P2 までの距離を教えて
rrd(P)	P からの距離を教えて
Rrd	距離を教えて

コマンド名	対応する発話
rrnc(C)	そのクラス C の名前を教えて
rrf(P)	P は何階にあるか
rri(P)	P はどの建物にあるか
Rri	それはどの建物にあるか
rrI(P)	P にはどんな店があるか
rrI	そこにはどんな店があるか
rrSclose(S,P)	P の近くにある種類 S の店には何があるか
rrSclose(S)	その近くにある種類 S の店には何があるか
rii(P1,P2)	P1 は P2 の中にあるか
riclose(P1,P2)	P1 と P2 は近いか
rieq(P)	それは P か
rieqc(C,P)	そのクラス C は P か
krp(P)	P の場所は知っています
It	ありがとう
R	（直前のコマンドを繰り返す）
Goodbye	（対話を強制終了）
Unknown	（処理対象外）
Error	（入力誤り）

表 2　意味表現の構成要素

構成要素	内容
want(A,S)	A が文 S を望む
inform_ref(A,B,O)	A が B に O を知らせる
inform_if(A,B,S)	A が B に文 S の真偽を知らせる
inform(A,B,YN)	A が B に真偽値を伝える
isa(O,C)	O はクラス C のインスタンスである
<属性>(O, P)	O の<属性>は P である
	<属性>は “name”（名前）, “location”（場所）, “shop_type”（店の種類）, “floor”（階数）, “source”（出発点）, “destination”（到着点）など。
value(O,<文字列>)	O の値は<文字列>である
setof(E, S, Set)	文 S を満たす E の集合は Set である