自 然 言 語 処 理　149－18
音声言語情報処理　41－9
（２００２．５．２４）

電話音声による列車時刻問合せシステムの評価

山崎一孝　伊東伸泰　西村雅史

日本 IBM 株式会社　東京基礎研究所
〒242-8502 神奈川県大和市下鶴間 1623-14
Email: yamasakk@jp.ibm.com

あらまし　トランスクリプションと統計パーサーによって，発話を認識し意味解析を行い，スロットフィル対話を行う列車時刻問合せシステムを作成した．対話制御はフォームを用いて行われる．システム構成はハブを介してシステム要素が結合するクライアント・サーバー型になっている．構成の概略を紹介し，次に, システムのタスク依存部分の作成方法について述べる. 収集した問合せ文から言語モデルを作成し，新聞記事などから作成した汎用言語モデルで補間する．同じ問合せ文に構文木を付与したものから，クラッサーとパーサーの統計モデルを作成する．このようにして得たシステムの音声認識部と意味解析部の精度を評価したので，報告する．

Development and Evaluation of Telephony Railway Information System

Kazutaka Yamasaki, Nobuyasu Itoh, Masafumi Nishimura

Tokyo Research Laboratory, IBM Japan
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502
Email: yamasakk@jp.ibm.com

Abstract　We developed an automatic railway information system whose domain is slot filling task. Using query examples, tri-gram language model and the two statistical models for the classer and the parser are obtained. Experiments are done to obtain speech recognition accuracy, classer accuracy, and parser accuracy.

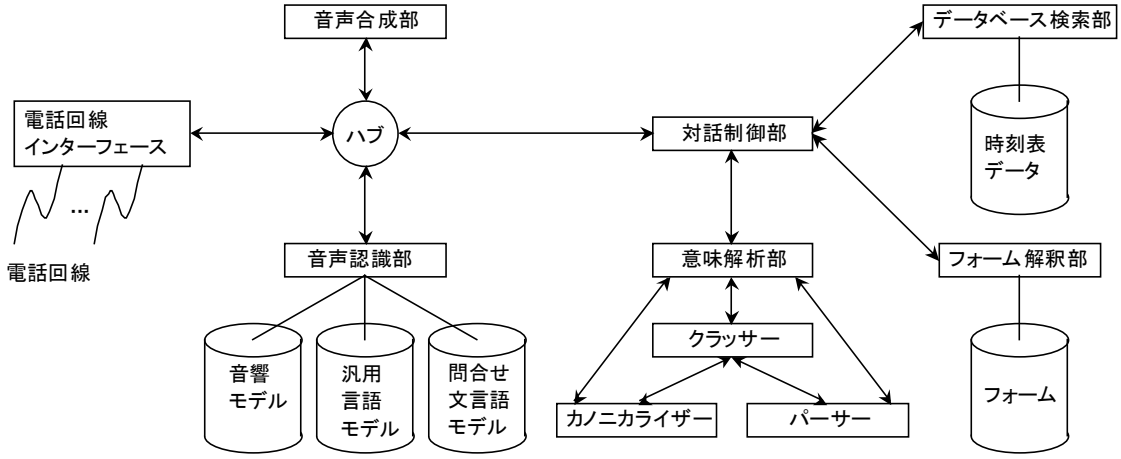

図 1：システム構成

1 はじめに

電話音声による情報提供を行う商用サービスが始まり，天気予報やニュースなどを音声コマンドによって選択し，最新の情報を聞くことができるようになった．このようなメニュー選択方式に基づくものよりも複雑な処理を行うアプリケーションにおいては，ユーザーが伝えるべき情報が多くなり，発話が長くなる．同時に，不要語が混入する可能性が高くなり, 言い回しも多様になるため, 文法制御による認識は困難になる．そこで，*n*-gram モデルを言語制約として用いる方法[Davies99][Lamel99]や, 文法と *n*-gram モデルの両者を言語制約として併用する手法[Souvignier00][鹿島 00]が対話システムにおいて用いられている．

n-gram といった緩い言語制約を用いる場合，話者の意図を認識結果から取り出すために，構文解析が用いられる．例えば，入力文に最も近い意図を求めるために，あらかじめ用意しておいたテンプレートの構文木と比較し意図を求める方法[Kurohashi00]や，統計パーサー[Magerman93]によって解析木を生成し意味ラベルを得る手法が提案されている．後者では確率モデルを用いているため，音声認識システムと組み合わせることが容易であり，統計モデル作成のためのコーパスを増やすことに注力すれば，さまざまな言い回しに対応することができるという特徴を持つ．

ここではトランスクリプションと統計パーサーによって発話を認識し意味解析を行う列車時刻問合せシステムを作成したので，その構成と性能について報告する．

2 システム概要

図 1 に示す本システムは[Davies99]に従い構成した．電話回線インターフェースに接続されたハブが音声合成部，音声認識部，および対話制御部にネットワークを介して接続されている．これら 3 要素はそれぞれ別個の計算機で動作可能な独立したプロセスとして実装されている[Zue00]. 回線数が多い場合にも，複数の計算機によって負荷を分散させることができる構成になっている．

システムに電話が着信すると，ハブは対話制御部を呼び出して最初のメッセージを得る．次に，ハブが音声合成部を呼び出し，ユーザーが最初のメッセージを聞く．次に，ハブが音声認識部を呼び出し，ユーザーが発話する．その結果，認識単語列 W が得られる．次に，ハブは対話制御部を呼び出し，単語列 W を渡す．対話制御部は，意味解析部，フォーム解釈部，データベース検索部を呼び出し，ユーザーへの次のメッセージを決める．

2.1 音声認識部

システムに対するユーザーの問合せを正しく認識するためには，問合せ文コーパスから言語モデルを作成することが望ましいが，そのために十分な例文を集めることはコストを要するため実施できないことが多い．この場合，さまざまなコーパスから作成された汎用言語モデルと，問合せ文言

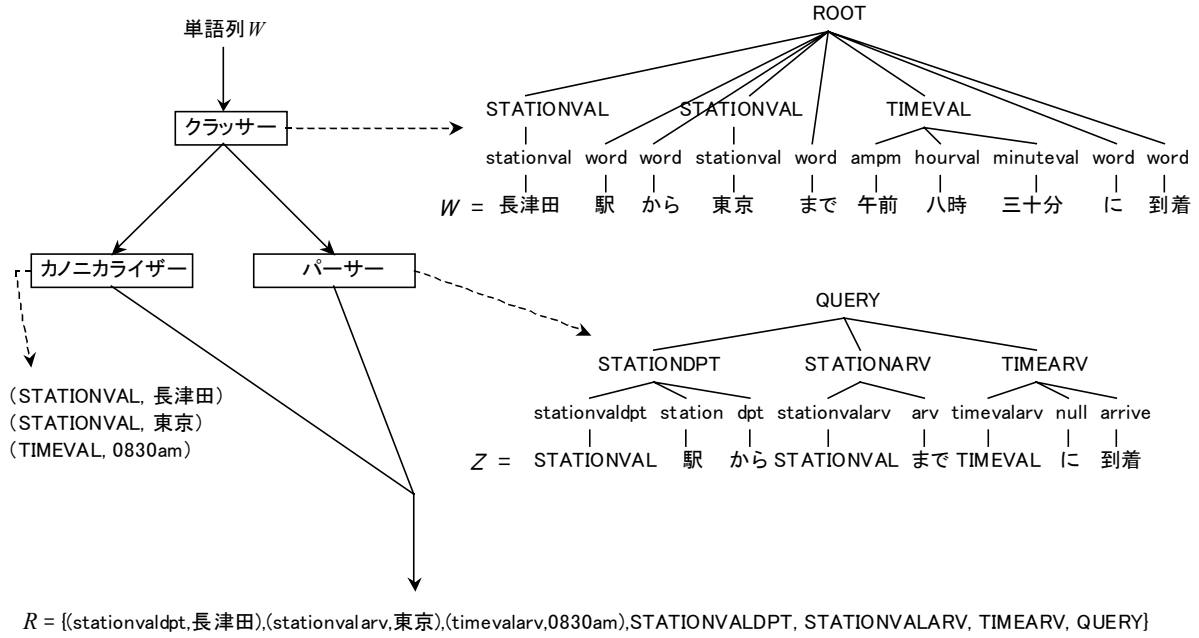

図 2：意味解析部における処理

語モデルを補間して用いることにより，それぞれを単独で用いる場合に比べて精度を改善することができる．本システムでは新聞記事や電子会議室への投稿文など約250M単語から作成した汎用言語モデルを用いている．

2.2 意味解析部

意味解析部では対話制御部が受け取った単語列Wからデータベースの検索条件や肯定，否定など発話の意図を推定する（図2）．

まず，クラッサーが図2の右上に示す解析木を単語列Wから生成し，駅名や時刻などのクラスを抽出する．駅クラスは「長津田」と「東京」であり，時刻クラスは「午前 八時 三十分」となっている．単語の上に付いた小文字は単語の意味を示すタグであり，木のノードに付いた大文字は，その下に含まれる部分木の意味を示すラベルである．Word タグはクラスに属さない単語であることを意味する．

単語列 W の単語と句をクラスラベルで置き換えると，具体的な時刻などによらない表現が得られる．例えば，クラッサーが生成した解析木からは，図2に示す単語列Zが得られる．これを入力として，パーサーは解析木を生成する．得られたタグは単語の意味を表し，出発駅(stationvaldpt)と到着駅(stationvalarv)，および出発時刻(timevaldpt)と到着時刻(timevalarv)が区別されている．意味ラベルはその下に含まれる単語列の意味を表し，出発駅(STATIONDPT)と到着駅(STATIONARV)などが区別されている．

このような意味ラベルの違いは，駅名などの具体的表現によらずに決まる．このクラス表現の揺れをあらかじめ取り除いた状態で解析を行うために，クラッサーとパーサーに分かれて処理を行う構成になっている．なお，クラッサーとパーサーはいずれも同一のアルゴリズムを用いた統計パーサー[Magerman93]であるが，前者はクラスの抽出を行い，後者は意味解析を行っている点が異なる．これら2つはいずれも，解析木Tを生成するために評価関数

$$\overline{T} = \underset{T}{argmax}\, Pr(T|W)$$

に従う．ただし，パーサーの場合はWをZで置き換える．

データベース検索を行うために，カノニカライザーはクラス表現の揺れを標準表現に変換する．例えば，「午前 八時 三十分」と，同一内容を表す

「朝 の 八時半」をいずれも「0830am」に変換する．カノニカライザーへの入力はクラスラベルと対応する単語列の組であり，上の例では（TIMEVAL, 午前 八時 三十分）となる．クラスラベル毎に用意された文法と変換規則に従って標準表現に変換する．

パーサーとカノニカライザーの出力がそれぞれ得られると，両者を合わせてタグと対応する標準表現値の組（attribute value pair, AV 組）を生成する．上の例では（timeval, 0830am）が AV 組である．全 AV 組と全ラベルの集合が意味解析部の出力 R となる．

2.3 対話制御部

ここでは意味解析部へ単語列 W を渡し，ラベルと AV 組の集合 R を受け取る．次に，フォーム解釈部に R を渡す．この値に応じて検索を行うか，あるいは問い返しなどのメッセージを生成するかが決定される．これらの制御規則[Bobrow77]を記述したものが図 3 に示すフォームである．

ここには 2 つのフォーム MAINMENU と QUERY が含まれている例を示した．対話処理中，いずれか一つのフォームに従って制御され，その切り替えはラベルに応じて行われる．

各フォームは slots 部と messages 部に分かれており，前者にはデータベース検索を行うために必要なスロット，および対話制御を行うためのスロットが並ぶ．これらの値が意味解析部から得た R によって決まる．スロット毎に対応すべきタグとラベルが MatchedBy キーワードの後に記述されている．Slots 部の中にも messages 部が存在し，スロット値が未定の場合に生成するメッセージを記述する．

Slots 部の後に置かれた messages 部には，データベース検索などの処理を行う関数名を記述する．図 3 の BERequest は関数名であり，その引数としてスロットが並んでいる．全引数の値が決まれば関数呼び出しを行うので，BERequest は YES スロット値が決まるまでは実行されない．一般には messages 部に複数の関数を記述しておき，得られたラベルの種類に応じて呼び出す関数を変えることもできる．

```
¥begin{form} MAINMENU
. . .
¥begin{messages}
  ¥msg InitPrompt
    Say "こちらは列車時刻案内システムです. . . "
. . .
¥end{form}

¥begin{form} QUERY
¥begin{slots}
  ¥slot STATIONVALDEPART ^MatchedBy: stationvaldepart
     ¥begin{messages}
        ¥msg Prompt:
           Say "出発駅を指定してください";
. . .
¥end{slots}
¥begin{messages}
¥msg BEMsg:
  BERequest {QueryMsg  $YES  STATIONVALDEPART . .
     ¥begin{rclist}
      ¥rc OK ¥msg Prompt: Say "$BE_resp";
. . .
¥end{form}
```

図 3：フォーム

3 システム構築と評価

前節で説明したシステムの各要素はツールキットとして提供され，さまざまなタスクを構築できるようになっている[IBM02]．本節では，音声認識部と意味解析部の中でタスクに依存する部分の構築について述べる．

まず，言語モデルとクラッサーおよびパーサーの統計モデル作成のために問合せ文の収集を行い，さまざまな言い回しを得る．このためには，電話機を用いてなりすまし方式などにより録音したものを書き起こすことが望ましいが，本稿では，被験者によるタイプ入力によって文収集を行った．それを用いて作成したシステムを，実際の対話データによって評価する．

3.1 問合せ文収集

被験者に問合せ文を自由にタイプ入力させると，システムの想定外発話[安達 00]も含まれてしまう可能性がある．そこで，図 4 に示すようなシナリオを約 10 通り用意し，そのシナリオに沿った文を集めることにした．

先頭が S で始まる行はメッセージを表し，U で始まる行は被験者が入力すべきことがらを表す．先頭 A の後の空欄に，被験者が文を記入した．このとき直前の U 行に従った内容を入力するよう

S:こちらは音声による列車時刻案内システムです.

S:出発駅,到着駅,出発時刻または到着時刻を指定して下さい.

U:(指定: 出発駅A, 到着駅B, 出発時刻C)

A:

S:出発時刻C, A駅発, B駅着でよろしいですか.

U:(肯定)

A:

図 4：問合せ文収集用シナリオ

指示したが，言い回しや具体的な駅名と時刻については実際の使用場面を想定しながら自由に入力するよう注意を与え，多様な表現を集める．

被験者 11 人から 939 文を得た. 単語に分割し，のべ 2988 単語, 358 異なり単語を得た. その内，駅名などの固有名詞は 169 個であり，不要語は「あー」の 1 語が 1 回現れただけであった.

これら 939 文を学習用 815 文と評価用 124 文に分け，それぞれ Set B，Set C と呼び，全体を Set A と呼ぶことにする．なお，後で述べる実際の対話データも評価に用いるが，これを Set D と呼び，Set C と区別する．

3.2 音声認識部

前節の Set B から問合せ文言語モデルを作成する．駅名や時刻などのクラスを定義して，クラス言語モデルを作成した．なお，これらのクラスはクラッサーにおいても用いられる．クラスのメンバーに，収集によって現れなかった駅名などを追加し，1997 異なり単語を認識対象語彙とした.

汎用言語モデルとの補間による効果を調べるために，Set C から 55 文を選び，それを被験者 8 人が読み上げた. 55 文中に未知語は含まれていない．電話回線を通じて録音した合計 440 文の文字誤り率を図 5 に示す.

横軸は問合せ文言語モデルの重みを表し，左端の点が汎用言語モデルだけを用いた場合である．この点と最良の点を比べると，誤り率が半分以下になった．右端の問合せ文言語モデルだけを用いる場合と最良の点を比べると，改善されたのは 0.3%（相対値で 5%）未満である.

3.3 クラッサーとパーサー

解析木導出に用いられる統計モデルを作成するため，3.1 節で得た Set B にクラッサー用の木構造を付与した．3.2 節で定義したクラスがタグに

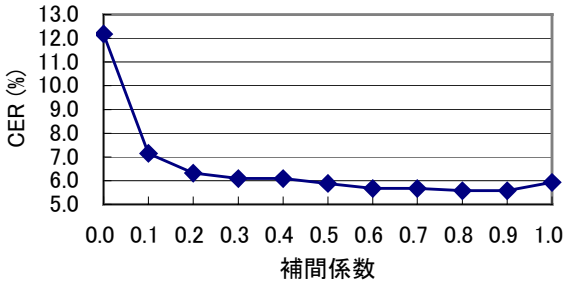

図 5：言語モデル補間係数と文字誤り率の関係

対応するようにした.「新宿 駅」のように「駅」が付いた部分には,「新宿」にだけ駅タグを付与する．時刻表現の時単位と分単位にはそれぞれ時タグと分タグを付与する.「三時半」のように三十分を表す「半」が付いた単語は分割せずに，一単語として半時間タグを付与する．この理由は,「三時 半」と分割するよりも長単位のまま扱う方が,音声認識の精度が高いためである.「午前」「午後」「朝」「夕方」などには午前午後タグを付与する.

なお，クラッサーが学習データに含まれない駅名などに対しても正しい解析木を与えるようにするため，クラッサーにおけるクラスを定義することができる．本稿では，言語モデルと同じクラスを定義した.

クラッサー用に付与した木構造にもとづき, 2.2 節で述べたクラスラベルによる単語の置き換えを行い，パーサーへの入力文を生成する．この文に対して木構造を付与する．そのタグの中でフォームのスロット値を決めるものは, 出発駅, 到着駅, 出発時刻，到着時刻，肯定，否定を表すタグである．これらの情報が含まれている単語列を葉とする部分木を作成し，その根にラベルを付与する.

Set C に対しても同様にして木構造を付与し，Set B から作成したクラッサーとパーサーの評価を行ったところ, 文単位の誤り率はそれぞれ 8.8% と 17.7%であった．クラッサーの誤りには時間タグ，分タグ，半時間タグの取り違えが多いが，その場合にもクラスラベルは正しい．このような例において，パーサーとカノニカライザーの処理に影響を与える誤りは存在しなかった.

パーサーの誤りにおいて頻度が高いものは，出発駅と到着駅が区別できない文（例えば「東京 です」）に対して，出発駅ラベル，または到着駅ラベ

ルを生成してしまう例である．本来はこれらを区

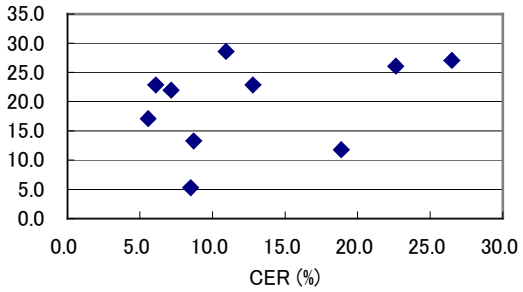

図 6：文字誤り率と意味解釈部の誤り率

別しない駅ラベルをパーサーが生成する．

3.4 対話音声による評価

前節までで得たシステムを用いて対話音声を収録した．被験者は 9 名であり，そのうち 5 名には 3.1 節と同様にシナリオに従うよう指示を与え，残り 4 名には特別な指示を与えなかった．言い淀みや，発話途中の中断なども含む録音データを書き起こし，370 文，のべ 962 単語を得た．未知語率は 4.2%であり，この中には「エー」などの不要語も含まれる．

3.2 節において文字誤り率の最小値を与えた補間係数を用いて，文字誤り率を測定したところ，指示グループ 5 名の平均は 7.2%，無指示グループ 4 名の平均は 19.4%となり，2 グループの間で 2 倍以上の差が生じた．

次に，意味解釈部の精度を評価するために．AV 組の誤り個数を数えた．書き起こし文に対して AV 組の正解を作成し，実験結果と比較する．正解，または実験結果のいずれか一方にだけ含まれる AV 組の数を誤り個数とし，それが正解個数に占める割合を求めた．結果を図 6 に示す．

文字誤り率との相関は明らかではなかった．この原因は，正しい認識結果に対して誤った解析木をパーサーが生成していることだと考えられる．

4 おわりに

トランスクリプションと統計パーサーによって発話を認識しその意味解析を行う問合せシステムを作成した．次に，そのシステムを用いて対話音声を収集した．被験者の中で，シナリオ指示グループと無指示グループの間に 2 倍以上の文字誤り率があった．この結果は，指示を与えない被験者からのデータを学習コーパスに加え，言語モデルを改善する必要があることを示唆する．今後はパーサーの統計モデルも改善し，実用に供することができるシステムを作成する予定である．

参考文献

[Bobrow77] D. Bobrow, R. Kaplan, M. Kay, D. Norman, H. Thompson, and T. Winograd. GUS, A frame-driven dialog system, Artficial Intelligence, Vol. 8, pp. 155-173, 1977.

[Magerman93] D. Magerman. Parsing as Statistical Pattern Recognition. IBM Technical Report No. 19443. 1993.

[Davies99] K. Davies, R. Donovan, M. Epstein, M. Franz, A. Ittycheriah, E. Jan, J. LeRoux, D. Lubensky, C. Neti, M. Padmanabhan, K. Papineni, S. Roukos, A. Sakrajda, J. Sorensen, B. Tydlitat, and T. Ward. The IBM conversational telephony system for financial applications. Proc. the 6th Eurospeech, Vol. 1, pp. 275-278, 1999.

[Lamel99] L. Lamel, S. Rosset, J. L. Gauvain, S. Bennacef. The LIMSI ARISE system for train information. Proc. ICASSP, pp. 501-504, 1999.

[安達 00] 安達，駒谷，河原．音声対話情報検索システムにおける想定外の発話の分析とその対処．人工知能学会研究会資料，SIG-SLUD-A001-2, 2000.

[鹿島 00] 鹿島，河原．複合的言語制約に基づくキーフレーズスポッティングによる対話音声理解．電子情報通信学会技術報告，SP2000-114, 2000.

[Kurohashi00] S. Kurohashi and W. Higasa. Dialogue helpsystem based on flexible matching of user query with nagural language knowledge base. Proc. the 1st ACL SIGdial Workshop on Discourse and Dialogue, pp. 141-149, 2000.

[Souvignier00] B. Souvignier, A. Kellner, B. Rueber, H. Schramm, and F. Seide. The thoughtful elephant: strategies for spoken dialog systems, IEEE Trans. SAP, Vol. 8, No. 1, 2000.

[Zue00] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hanzen, and L. Hetherington. JUPITER: A telephone-based conversational interface for weather information. IEEE Trans. SAP, Vol. 8, No. 1, 2000.

[IBM02] http://publib.boulder.ibm.com/voice/pdfs/white_papers/WSVS20.pdf