音声対話システムのための
統計的言語理解モデルの構成とその学習

須藤 克仁　中野 幹生

日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
〒 243-0198 神奈川県厚木市森の里若宮 3-1
{sudoh,nakano}@atom.brl.ntt.co.jp

概要

音声対話システムで用いる言語理解部として，重み付き有限状態トランスデューサ (WFST) による統計的言語理解モデルを利用し，そのモデルをシステムが対話を通じて自律的に学習するための手法を提案する．また，システムがラベルなしデータを使って学習するための手法について検討する．学習サンプルの選択基準としては，対話終了後に対話全体の情報を利用して音声認識・言語理解の正しさを評価するための信頼性評価法 (対話後信頼性評価法) を利用する．

1　はじめに

音声対話システムを構築するためには，音声認識用の音響モデルや言語モデル，言語理解のための規則，さらに対話制御規則や応答生成規則など，さまざまな知識を記述する必要がある．しかし，これらの知識はシステムで扱うタスクのドメインによって大きく異なるため，新しいシステムを構築するたびに記述する必要があり，非常にコストが大きい．一方，近年の統計的モデルを用いた音声認識・自然言語処理研究の進歩により，大量のタグつきコーパスを用意すれば統計的モデルとして知識を記述することができるようになってきている．ところが，大量の音声データを書き起こしたり，その書き起こしにさらにタグを付与するといった作業もコストが大きいため，統計的モデルの作成に必要な書き起こしやタグ付与の労力を極力少なくすることが求められる．

我々の研究グループでは，音声対話システムに対話を通じてさまざまな知識を訓練させるための研究を進めている [1]．こうした研究の成果により，音声対話システムを構築する際の書き起こしやタグ付与などの労力を削減できることが期待される．本稿ではその一環として統計的言語理解モデルを扱い，システムが対話を通じてモデルを自動学習するための枠組みと，そこで利用されるモデルの構成について述べる．

以下，2 節では対話を通じた音声対話システムの知識学習について概略を述べる．続く 3 節では本稿で提案する重み付き有限状態トランスデューサ (WFST) を用いた言語理解モデルの構成法を述べる．4 節では言語理解モデルの自動学習の方法について述べる．5 節では提案手法に関する実験結果を示す．6 節はまとめである．

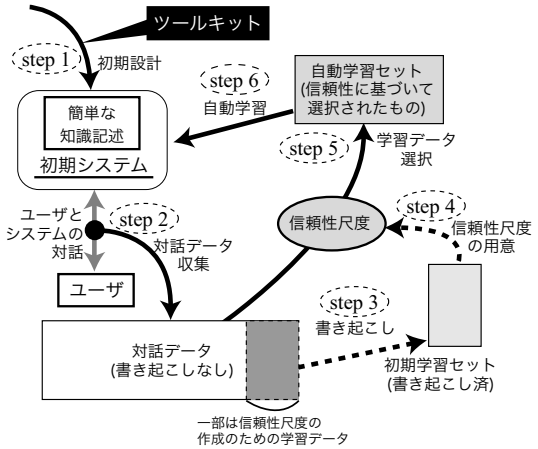

図 1: 本研究における知識学習の枠組み

2　対話を通じた音声対話システムの知識学習

本研究では，書き起こしやラベルが付与されていない未処理の対話データから，言語理解モデルなどの知識の学習を行うことを考える．ここで，音声対話システムが対話を通じて得ることができる未処理の対話データには，音声認識や言語理解の誤りが少なからず存在するという問題がある．本研究では，対話データの中から音声認識結果の信頼性の高いもののみを選別して学習に利用することで，この問題を回避する．

信頼性の高いデータのみを選別するための基準としては，音声認識研究で用いられる信頼性尺度 (confidence measure) を利用する．信頼性尺度とは，1 つまたは複数の音声認識に関わる特徴量をもとに，音声認識結果の信頼性を評価値 (confidence score) として評価するためのものである．本稿では，我々が提案した対話後信頼性尺度 [2, 3] を利用する．

これらのアイデアをもとに，本研究では，音声対話システムが対話を通じて知識を学習していくためのシステム構築の枠組みとして，以下のような手続きに基づいて音声対話システム構築を行う．(図 1 を参照)．

1. ツールキットを用いて，簡素な知識記述のみから初期システムを設計する．
2. システムとユーザとの対話データを収集する．
3. 対話データ中のユーザ発話を書き起こし，シス

テム改良のためのデータセットとする (このデータセットを以下「初期学習セット」と呼ぶ).

4. 初期学習セットを利用して，信頼性尺度を作成する.
5. 4. で作成した信頼性尺度に基づき，対話データ中で書き起こしをしていないものから信頼性の高いものを選択し，システム改良のためのデータセットとする (このデータセットを以下「自動学習セット」と呼ぶ).
6. 自動学習セットを利用して，システムの知識記述を改良し，新しいシステムを作成する.
7. 以下，2,5,6 を繰り返す.

3 統計的言語理解モデルの構成

従来の音声対話システムの大部分では，言語理解のための文法や例文を設計者が作成しなくてはならず，さまざまな言い回しに対応できる頑健な言語理解を実現することは難しかった．そうした問題に対処するためには，統計的な言語理解モデルに基づく手法が有効である．N-gram やベクトル空間モデルなどを用い，発話中の単語の集合から発話意図の推定を行う手法 [4] は自然言語処理の分野でいくつか提案されているが，音声認識誤りへの対処についてはこれまであまり議論されていない．一方，[5] は中間表現の共起確率を利用することで音声認識誤りに対して頑健な音声理解を可能にしている．しかし，音声認識結果の信頼性の高低に関わらず，同じ単語系列に対しては同じ理解結果を返してしまうという問題がある.

本稿では，さまざまな言い回しと音声認識誤りに対して頑健な言語理解のために，言語理解モデルが，初期設計時の規則やドメイン定義に基づいた理解 (規則ベース理解・キーワード抽出) と，対話を通じた学習に基づく統計ベースでの理解とを同時に実現できるようにし，さらに，音声認識の信頼性にも基づいて認識誤りを棄却するための枠組みを導入する.

3.1 実装

本稿では，統計的な言語理解と，規則に基づく言語理解やキーワード抽出などを統合的に扱うためのモデルとして，WFST(重み付き有限状態トランスデューサ)[6] を利用する．WFST を利用することにより，後述する単語棄却パスを導入して信頼性の低い単語仮説を棄却するという枠組みを容易に導入することができるという利点がある．言語理解 WFST は入力シンボル集合をシステムの語彙とし，出力シンボル集合をコンセプトを表現するシステムの内部表現の集合とする,「規則ベースの言語理解部」「統計ベースの言語理解部」「キーワード抽出部」それぞれの WFST を結合したものの閉包として構成される．閉包にすることにより，任意の数のコンセプトを含む任意の長さの単語列を扱うことができる．各部分の WFST の構成例を図 2 に示す.

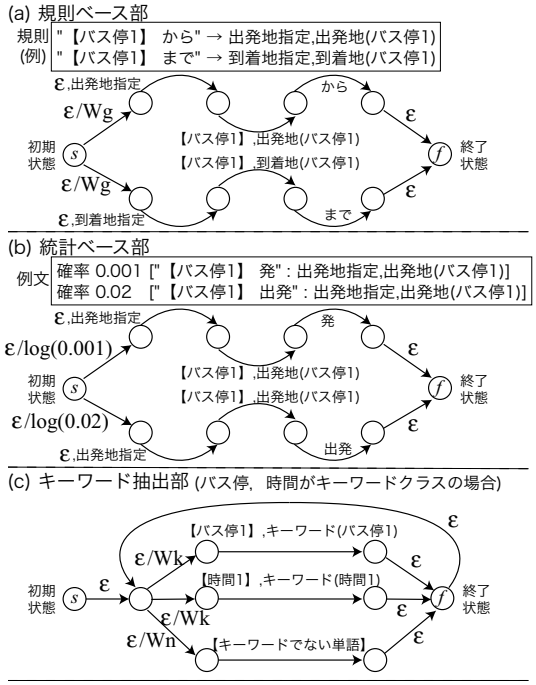

図 2: 言語理解 WFST の作成の例

規則ベース理解部　規則で定義された表現を入力として，対応する対話行為を出力するパスを持つ．このときの各パスに対する重みは個別に定義することが難しいため，定数 w_g をすべてのパスに対する重みとする.

統計ベース理解部　獲得した例文と一致する単語列を入力として，対応する対話行為を出力するパスを持つ．各パスに対する重みは例文-対話行為の組の生起確率の対数である.

キーワード抽出部　システムの語彙すべてを入力とし，特定のキーワードクラスに属する単語に対してはそのキーワードを出力するようなパスと，キーワードでない単語に対しては何も出力しないようなパスを持つ．各単語に対する重みも個別に定義することは難しいので，キーワードのパスの重みは定数 w_k ，キーワードでない単語のパスの重みは定数 w_n とする.

3.2 音声認識誤りへの対処

音声認識誤りに対処するため，音声認識結果のうち信頼性が低いものを除去する．具体的には，音声認識器が出力する信頼性情報つき N-best 音声認識結果を，各単語仮説をある重み (本稿では信頼性スコアの符号を反転させた値) で出力するようなパスと，単語仮説をある重み c_r で棄却するようなパスを持つ WFST(図 3) として表現する．そして，この WFST と言語理解モデルの WFST と合成した WFST 上で最適パスを探索する．このことにより，認識誤りである可能性が高い (信頼性が低い) 単語仮説ほど，より棄却されや

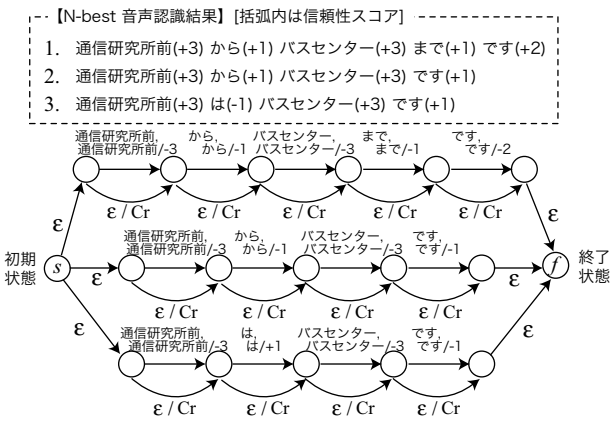

図 3: N-best 音声認識結果に単語仮説棄却パスを加えた WFST の例

すくなり，音声認識誤りに対して頑健な音声理解が可能になる．なお，ここではゼロ中心の対数尤度比 [7] による信頼性スコアを用いる．

4 言語理解モデルの自動学習

3 節で示したモデルの統計ベース理解部を学習させるために，2 節で述べた自動学習の枠組みで学習に必要な例文と対話行為の組を獲得することを考える．本稿では，対話後信頼性評価法に基づく統計的言語モデル学習手法 [2, 3] と同様の手続きにより，例文と対話行為の組を自動的に獲得する手法について検討する．

4.1 対話後信頼性評価法

対話後信頼性評価法 (post-dialogue confidence scoring) とは，従来の信頼性評価法の研究において利用されてきた音響尤度などの特徴量に加え，対話終了後に得られる特徴量を利用することにより，正しい音声認識結果を従来手法より精度よく選択するための手法である．詳しい実装については [2, 3] を参照されたい．対話後信頼性評価法においても，信頼性を計算するための信頼性評価モデルのパラメタは，少量の書き起こし済みデータを利用して最適化される．

4.2 言語理解モデル学習への適用

言語理解モデルの学習サンプル (例文と対話行為の組) は，収集した対話データの各音声認識結果に対して信頼性評価を行い，信頼性の高い (閾値が一定値以上) 音声認識結果を例文として，その音声認識結果に対応する言語理解結果を例文に対応する対話行為として獲得する．そして，WFST による言語理解モデルの統計ベース部に新しいパスを付け加え．重み (例文-対話行為の組の生起確率の対数) を更新すればよい．

5 実験

本稿で提案した統計的言語理解モデルの性能と，対話データを用いた自動学習による性能変化を調べるために，以下の 2 種類のモデルの学習方法に基づいて言語理解モデルを作成し，音声認識結果を入力したときの性能をコンセプト誤り率 (Concept Error Rate:CER) により評価した．

【実験 1】 書き起こし・コンセプト正解ラベル付与済み対話データ (以下，ラベルありデータ) を利用して学習

【実験 2】 書き起こし・コンセプト正解ラベル付与がされていない対話データ (以下，ラベルなしデータ) を利用して自動学習

なお，実験においては，音声認識結果と書き起こしのそれぞれを入力とすることで認識誤りが含まれることによる言語理解結果の影響を調べるとともに，図 3 に示したような単語棄却パスを持つ入力を採用したことの効果を調べるため，

- 書き起こしを入力
- 1-best 音声認識結果 (単語列) を入力
- 1-best 音声認識結果 (単語棄却パスあり) を入力
- n-best 音声認識結果 (単語棄却パスあり: $n \leq 20$) を入力

のそれぞれの場合についての CER を評価している．

実験で使用されたデータは，被験者 150 名と実験用に作成したバス時刻表案内システムとの対話 1,474 対話 (15,464 発話) 分の対話データ (ユーザ音声とその書き起こし，コンセプト正解ラベル) である．システムの語彙数は 307，音声認識エンジンには Julius を用いた．収録した対話における音声認識の平均単語正解精度は 64.97%であった．また，学習する前の初期言語理解モデルは，このシステムで利用される，38 個の簡易な言語理解規則と，バス停名・時刻・曜日などのキーワードを持つモデルである．以下，それぞれの実験の詳細とその結果について示す．

【実験 1】

上記の対話データからランダムに約 2,400 発話分のラベルありデータを選んでテストセットとし，約 700 発話分のラベルありデータを選んで単語棄却パスの生成に用いる信頼性評価モデルの学習セットとした．そして，言語理解モデルの学習セットとするラベルありデータを変化させて言語理解モデルを学習させたときの，テストセットの音声認識結果および書き起こしに対する CER を評価した．その結果得られた，学習セットのデータ量 (発話数) と CER の関係を図 4 に示す．

図 4 のように，入力のタイプを問わず，学習データ数が増えることにより CER が低下していく傾向にあり，提案した言語理解モデルが正解を与えることで学習可能なモデルであることが確認できた．また，1-best(単語列) と 1-best(単語棄却パスあり) との比較により，音声認識の信頼性に基づく単語棄却パスを加

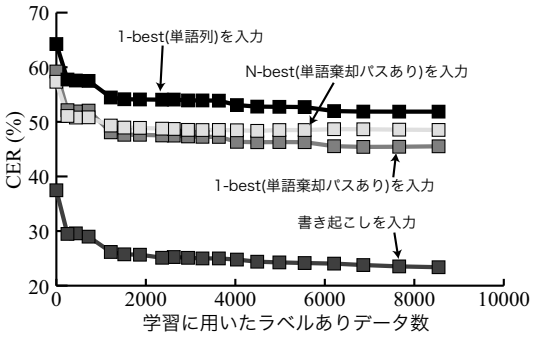

図 4: 言語理解モデルの学習データ量と CER との関係 (ラベルありデータ利用)

えることで，音声認識誤りが含まれるような入力に対する理解の頑健性が向上することも確認できた．一方，単語棄却パスを加えた場合でも，N-best を入力とした場合には 1-best の場合よりも CER が若干高くなっており，複数候補の入力に関しては検討の余地があると言える．

【実験 2】

実験 1 と同様に，全対話データからランダムに約 10% のラベルありデータ (約 2,400 発話分) をテストセットとし，約 5%のラベルありデータ (約 700 発話分) を単語棄却パスの生成に用いる信頼性評価モデルの学習セットとした．その上で，これらのラベルありデータと異なる対話のラベルなしデータ (約 9,000 発話) から言語理解モデルの自動学習セットとするものを信頼性評価法に基づいて選び，ラベルありデータと合わせて言語理解モデルを学習させ，テストセットの音声認識結果 (1-best，単語棄却パスありの場合) に対する CER を評価した．なお，ラベルありデータとラベルなしデータを学習サンプルとして混合する際には重みをつけている (この実験ではラベルあり:ラベルなし=15:1 とした)．また，信頼性評価モデル学習セットとした少量のラベルありデータを用いて言語理解モデルを学習させた場合，ラベルなしデータにもラベルを付与したもの (約 9,700 発話分のラベルありデータ) を用いて学習させた場合との比較も行った．各データセットを入れ替えて 4 セット実験を行った結果得られた CER を，図 5 に示す．

図 5 のように，ラベルなしデータであっても選択的に学習に用いることで，若干 CER が低下している．セット 2, セット 3 ではわずかな効果にとどまっており，安定した学習効果を得ることがさらなる検討課題と言える．

6 おわりに

本稿では，WFST を利用した音声対話システムの言語理解モデルの構成方法と，言語理解の際に音声認識誤りを棄却するための枠組みを提案し，さらに，ラベルなしデータを用いたモデルの自動学習手法について検討した．そして，単語棄却パスを利用することで音声認識誤りが言語理解に与える悪影響を低減でき，ラベルありデータ，信頼性に基づいて選択したラベルなしデータを利用して，提案した言語理解モデルの性能を若干向上させることができることを確認した．

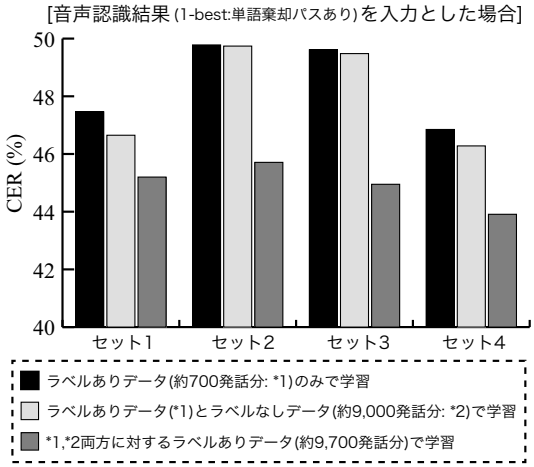

図 5: ラベルなしデータを学習に利用した言語理解モデルの CER

今後は，言語理解モデルの自動学習のための信頼性評価法の改良に加え，WFST の重み決定方法についてのさらなる検討や，言語理解モデル学習への能動学習の適用などを予定している．

謝辞

この研究を進めるにあたり日頃よりご指導いただいている牧野昭二メディア情報研究部長，並びに研究に関して議論させていただいているマルチモーダル対話研究グループの諸氏に感謝する．

参考文献

[1] 中野幹生, 東中竜一郎, Matthias Denecke, 須藤克仁, 宮崎昇, 堂坂浩二. 対話ログによる訓練が可能なインタラクティブ音声理解システムの枠組. 言語処理学会第 10 回年次大会, 2004.

[2] 須藤克仁, 東中竜一郎, 中野幹生, 相川清明. “反省型”信頼性尺度に基づく書き起こしなしデータを用いた言語モデル学習. 情処研報 SLP-45, pp. 83–88, 2003.

[3] Katsuhito Sudoh and Mikio Nakano. Post-dialogue recognition confidence scoring for improving statistical language models using untranscribed dialogue data. In *ASRU 2003*, pp. 447–452, 2003.

[4] 白木将幸, 伊藤敏彦, 甲斐充彦, 中谷広正. 自然発話分における統計的な意図理解方法の検討. 情処研報 SLP-50, pp. 69–75, 2004.

[5] 政瀧浩和, 谷垣宏一, 匂坂芳典. 統計処理による入力文からの中間表現への変換を用いた音声言語理解. 信学論, Vol. J82-D-II, No. 2, pp. 169–177, 1999.

[6] Mehryar Mohri. Finite-state transducers in language and speech processing. *Computational Linguistics*, Vol. 23, No. 2, pp. 269–311, 1997.

[7] Timothy J. Hazen, Stephanie Seneff, and Joseph Polifroni. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech and Language*, Vol. 16, No. 1, pp. 49–67, 2002.