# 音声対話システムにおける想定外発話の文法検証を用いた対話行為推定に基づくヘルプ生成

池田 智志　駒谷 和範　高橋 徹　尾形 哲也　奥乃 博
京都大学大学院 情報学研究科 知能情報学専攻

## 1. はじめに

電話などのインタフェースを介した音声対話システムでは，事前に使用法を教示されていない一般ユーザがシステムを利用する．システムの想定する前提とユーザが思い込んでいる機能とのずれから，ユーザの発話がシステムの受理できない想定外発話となる場合が少なからずある．想定外発話は音声認識誤りの原因となり，システムの誤動作を引き起こす．想定外発話は，初心者が音声対話システムを使い始める際の障壁のひとつである．

本稿では，ユーザ意図 (対話行為) に応じたヘルプを順番に提示することで想定外発話に対処する．本研究ではまず，言語理解文法をユーザの対話行為に対応させて定義し，Weighted Finite State Transducer (WFST) に基づく文法検証を用いてユーザの対話行為を推定する．ここで課題となるのは，想定外発話は音声認識結果の情報に欠損があることも多く，一発話のみから正しい推定結果を得るのは困難な点である．そこで，対話行為の推定結果と既知度 [1] や対話文脈とを統合してヘルプ候補の提示順序を決定する．既知度や対話履歴などを考慮することで，対話行為推定の高精度化が期待できる．また，ヘルプ提示順序の考慮により，一回では正しいヘルプを提示できなかった場合も，適切なヘルプ提示に少ない対話ターン数でたどりつくことが可能になる．

## 2. ヘルプ提示戦略の要求条件

本研究では，以下を満たすヘルプ提示が目的である．

1. ユーザの対話行為に応じたヘルプの提示
   ユーザのタスク遂行に必要な言語表現をヘルプとして提示することで，想定外発話に対処する．このためには，想定外発話に対しても頑健に対話行為を推定できる枠組みが不可欠である．
2. 多様な情報を提示するためのヘルプ提示順序決定
   同じような内容のヘルプを連続して提示するのは効率的ではない．異なるヘルプを多数提示することで，一回のヘルプ提示精度の低さを補えるからである．そのため，ヘルプ候補を順序付けし，優先度順に異なるヘルプを提示する枠組みが必要である．

本研究が目指す対話例を図 1 に示す．従来手法 [2] では，音声認識結果のみを考慮するため，情報の欠損の多い想定外発話には対処できない．同じ音声認識結果に対しては，毎回同じヘルプが提示されるため，$S2_1$ のように誤ったヘルプ提示をくり返してしまう．本研究では，文法検証結果や既知度等の考慮により提示すべきヘルプの順序付けを行う．これにより，仮に S1 と同じ誤った推定結果が得られても，$S2_2$ のように一度提示したヘルプを避けることで，提示するヘルプが多様になる．また，ヘルプを信頼できる順に提示できるため，少ない対話ターンで正しいヘルプ提示にたどりつくことができる．

Help Generation Strategy Based on Dialogue Act Estimation by Using Grammar Verification in Spoken Dialogue Systems: *Satoshi Ikeda, Kazunori Komatani, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

**U1:** おすすめのスポット を教えて (観光施設の検索)
(下線部が語彙外．対話行為推定結果は「寺社の情報の取得」)
**S1:** 施設情報を調べるには，「清水寺の住所を教えて」などと言って下さい．
**U2:** おすすめのスポット を教えて (観光施設の検索)

いずれか
- $S2_1$(従来手法): 施設情報を調べるには，「清水寺の住所を教えて」などと言って下さい．
  (U1 と同じ推定結果．S1 と同じヘルプ．)
- **$S2_2$(本手法):** 施設タイプで観光施設を検索するには，「神社を検索して」，「博物館を教えて」などと言って下さい．
  (対話行為推定結果:
  × 1-best: 施設情報の取得 (既に提示)
  ○ 2-best: 施設タイプを指定して検索 …)

図 1: 本手法により可能となる対話

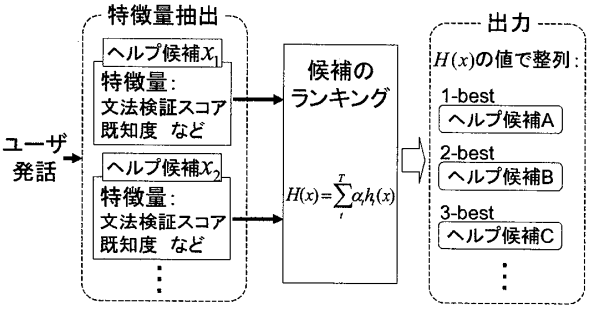


図 2: ヘルプ提示手法の概略

## 3. 文法検証と優先度学習に基づくヘルプ提示

本手法の概略を図 2 に示す．まず，WFST に基づく文法検証を利用して，想定外発話からも頑健に対話行為推定の特徴量を取得する．次に，抽出した特徴量に対して，RankBoost[3] に基づき，ヘルプ提示順序を学習する．

### 3.1 WFST に基づく文法検証を利用した対話行為推定

ユーザ対話行為を高精度に推定するために，WFST による文法検証を導入する [1]．文法検証に用いる WFST では，音声認識結果と全ての文法に対して重みを計算しマッチングをとり，その累積重み $w$ が得られる．本研究では，システムの受理可能な対話行為に対応して文法を定義し，その文法に対して得られた重み $w$ を得ることで，対話行為を推定する．

WFST の累積重み $w$ は，受理単語に対する正の重み ($w_w$)，文法間違いに対する負の重み ($w_{sub}, w_{del}, w_{ins}$) の和で計算する．$w$ の計算式を以下に示す．

$$w = \sum_{W \in \boldsymbol{W}_{\text{accept}}} w_w + \sum_{W \in \boldsymbol{W}_{\text{wrong}}} (w_{sub} + w_{del} + w_{ins})$$

ここで，$W$ は音声認識結果に含まれる単語，$\boldsymbol{W}_{\text{accept}}$ はマッチングをとった結果 WFST に受理された単語の集合，$\boldsymbol{W}_{\text{wrong}}$ はマッチングをとった結果文法間違いとされた単語の集合である．また，$w_w$，$w_{sub}$，$w_{del}$，$w_{ins}$ は文献 [1] に従い以下とした．

表 1: 使用した特徴量

| | |
|---|---|
| **T1:** | WFST に基づく文法検証スコア |
| **T2:** | 文法検証スコアの信頼度 |
| **T3:** | 受理単語の割合 ($=w_{accept}/W$) |
| **T4:** | 文法検証結果の最大連続受理単語数 |
| **T5:** | 文法検証結果の受理スロット数 |
| **T6:** | 当該対話行為が受理されたのは何発話前か |
| **T7:** | 当該対話行為における現発話までの最大 WFST スコア |
| **T8:** | 当該対話行為がコマンド発話かどうか |
| **T9:** | 当該対話行為が検索条件を指定する発話かどうか |
| **T10:** | 当該対話行為が検索結果の情報を取得する発話かどうか |

$$
\begin{aligned}
w_w &= l(w_{asr})CM(w_{asr}) \\
w_{sub} &= -(CM(w_{asr})l(w_{asr}) + l(w_g))/2 \\
w_{del} &= -(\overline{l(w)} + l(w_g))/2 \\
w_{ins} &= -(CM(w_{asr})l(w_{asr}) + \overline{l(w)})/2
\end{aligned}
$$

ここで，$w_{asr}$ は音声認識結果として得られた単語，$l(w_{asr})$ は $w_{asr}$ の長さ，$CM(w_{asr})$ は $w_{asr}$ に対する信頼度である．$l(w_{asr})$ はモーラ数に比例する値で，語彙中で最も長い単語の長さで正規化している．また，$w_g$ は想定文法の対応する単語，$\overline{l(w)}$ は全語彙の $l(w)$ の平均である．

### 3.2 優先度学習に基づくヘルプ候補順序付け

本研究では，ヘルプの提示順序を決定するために優先度学習を行う．優先度学習とは，複数の候補から正解を選んだり，候補をリランキングするための学習手法である．Web ページの検索結果のリランキングなどに応用され，本研究の目的に適した手法である．

優先度の学習には，RankBoost[3] を用いた．RankBoost は，候補を順序付けするスコア関数を boosting を用いて学習する手法である．スコア関数の値に従って対象を整列することで，候補の順序付けを行う．スコア関数 $H(x)$ は，順序付けの部分的な情報を与える弱順序付け器を線形結合することで学習される．具体的には，$H(x) = \sum_t^T \alpha_t h_t(x)$ と表される．ここで，$T$ はブースティング回数，$\alpha$ は結合重み，$h$ は弱順序付け器，$x$ はランキングの候補である．本研究では，$x$ は提示すべきヘルプの候補となる．弱順序付け器 $h$ は，各特徴量の値を閾値処理することで得た [3]．具体的には，以下の式で表される．

$$
h(x) = \begin{cases} 1 & \text{if } f_i(x) > \theta \\ 0 & \text{if } f_i(x) \leq \theta \\ q_{def} & \text{if } f_i(x) = \perp \end{cases}
$$

ここで，$f_i(x)$ は候補 $x$ の $i$ 番目の特徴量の値，$\perp$ は特徴量の値が得られなかったことを示す．また，$q_{def}$ は $\{0,1\}$ である．弱学習では，$i$, $q_{def}$, $\theta$ の最適値を求める．

本研究で与えた特徴量を表 1 に示す．T1 から T5 を定義することで，文法検証結果がどれだけ信頼できるかを考慮できる．T2 は，全ての候補 $x$ の文法検証スコアの総和で T1 を正規化することで得た．T6 は，対話文脈に相当する．発話が受理されない時，ユーザは同様の発話を繰り返す性質があるため，定義した．T7 は文献 [1] の既知度に相当する．ユーザが知っている文法構造に対するヘルプは提示する必要がないため，定義した．T8 から T10 は当該対話行為の性質を表わした特徴量である．

## 4. 評価実験

### 4.1 評価対象データ

評価には，25 名の話者による想定外発話 418 発話を用いた．これは，マルチドメインシステム [4] を用いて収集された 30 話者 11,773 発話から，文法検証により想定外と判定された発話を取り出したものである．残りの 5 話者分の対話データは，音声認識の言語モデルの改善に使い，評価には用いなかった．また，評価対象ドメインは観光ドメインのみとした．データ収集時には，6 つのシナリオを与え，事前教示は全く与えなかった．

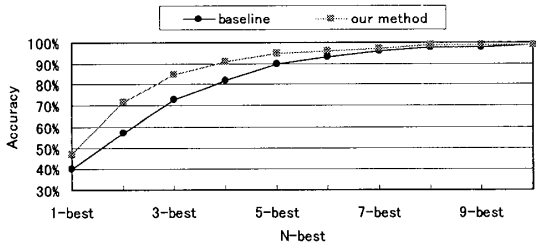

図 3: $N$-best までの対話行為推定精度 ($1 \leq N \leq 10$)

対話行為の正解は，ユーザの発話に関係の強い順に上位 5 つを与えた．判別すべき対話行為の数は 24 である．文法検証で使用する音声認識器には Julius[‡]を用いた．言語モデルは認識文法から生成した例文 10,000 文と，評価に用いない 5 話者 600 発話の書き起こしから学習した 3-gram モデルを用いた．語彙サイズは 3,593 で，単語正解精度は 42.1%であった．単語正解精度が低いのは，対象データが想定外発話だからである．

### 4.2 対話行為推定精度の評価と今後の課題

対話行為の順序付け結果のうち，$N$-best までに正解が含まれる発話の割合を評価した．$N$ の値は，1 から 10 までとした．ベースライン手法は，WFST に基づく文法検証スコアの値のみに基づいてヘルプを提示する手法とした．また本手法において，ブースティング回数は 300 とし，評価は 5-fold cross validation とした．

実験結果を図 3 に示す．全ての $N$ の値において，本手法の方がベースラインより対話行為推定精度が高かった．特に $N=1$ のときは 7 ポイント，$N=2$ のときは 15 ポイント，$N=3$ のときは 12 ポイントと，$N$ の値が小さいときに精度の改善が特に著しかった．これは，本手法が対話行為の推定結果が信頼できる順に対話行為の候補をリランキングしているためである．以上の結果は，本手法が正しいヘルプを提示できるまでに必要な対話ターンを少なくできることを示している．

今後は，ユーザの対話行為の遷移を考慮した対話管理に本手法を組み込む検討を進める．具体的には，POMDP[5] のような確率的枠組みでの統合が考えられる．

**謝辞** 評価データは，HRI-JP の中野幹生氏らとの共同研究において構築したシステムにより収集した．また，科研費，GCOE，SCAT の援助を受けた．

## 参考文献

[1] 福林他. 音声対話システムにおける動的ヘルプ生成を指向した WFST に基づく文法検証によるユーザ知識推定. 人工知能学会研究会資料, SIG-SLUD-A703-09, pp. 45–50, 2008.

[2] G. Gorrell *et al.* Adding Intelligent Help to Mixed-Initiative Spoken Dialogue Systems. In *Proc. ICSLP*, pp. 2065–2068, 2002.

[3] Y. Freund *et al.* An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, Vol. 4, pp. 933–969, 2003.

[4] S. Ikeda et al. Topic estimation with domain extensibility for guiding user's out-of-grammar utterance in multi-domain spoken dialogue systems. In *Proc. Interspeech*, pp. 2561–2564, 2007.

[5] J. Williams and S. Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, Vol. 21, pp. 393–422, 2007.

[‡]http://julius.sourceforge.jp/