

音声対話システムにおける誤り原因の階層的分類とその推定に基づく発話誘導

駒谷 和範 福林 雄一郎 池田 智志 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

本研究では、音声によるコミュニケーションエラーの階層性に着目し、適切なヘルプ生成によりユーザ発話を誘導する対話管理手法の開発を目的としている。これにより、事前教示を与えないユーザに対しても新たな音声認識誤りを防止し、タスク達成率の向上を狙う。

この背景として、音声対話システムにおける、インタフェースとしてのアフォーダンス [1] の欠如が挙げられる (図 1)。つまり音声は、メディアの特性として大量の情報を一度に出力するのに適さず、またシステムが受理可能な情報を暗黙に伝えるのも困難であるにもかかわらず、現状では音声認識誤りが生じた際のユーザへのフィードバックが不十分である。実際、著者らが京都市バス運行情報案内システム [2] を一般に公開して実ユーザから得たデータの中には、自分の発話が正しく認識されなかった場合にその原因がわからず、発話を適切に修正できない場合が多く見られた。本研究では、システム側での対処だけでなく、インタラクション相手であるユーザの発話にも影響を与えることで、対話を通じた音声認識誤りへの包括的な対処を目指す。

本研究では、音声対話システムにおけるコミュニケーションエラーを大きく 4 つの階層として定義し、これらを検出してヘルプメッセージを生成することで、ユーザ発話をシステムの受理可能な範囲内へと誘導する。Clark は、言語の使用は参加者間の共同行為であり、かつ話し手及び受け手の行為には 4 つのレベル (Conversation, Intention, Signal, Channel) があることを提唱した [3]。本研究では、これに対応させて音声対話システムのエラーを階層的に分類し、それぞれを検出し対処を図る。この 4 階層を図 2 に示す。従来から扱われる音声認識誤りは、ここでの Signal Level の誤りに相当する。

本稿では、Conversation Level 及び Intention Level のエラーの検出を報告する [4, 5]。Conversation Level 及び Intention Level のエラーは、ユーザの発話がシステムの受理できる範囲外であることに起因する (想定外発話)。まず想定外表現を含む発話を検出できれば、音声認識誤りの誤受理が防げる。さらにこの場合、音声認識結果は信頼できないため、詳しい応答生成に有用な情報を取得するために、トピック推定や発話検証技術を導入した。

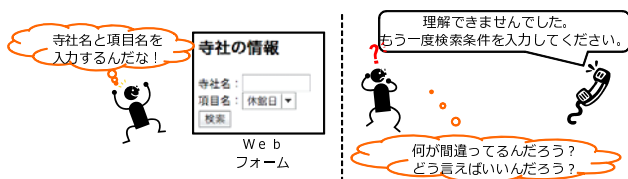


図 1: 音声インタフェースにおけるアフォーダンスの欠如

Hierarchical Classification of Error Sources in Spoken Dialogue Systems and its Applications to Generating Guidance for User Utterances. Kazunori Komatani, Yuichiro Fukubayashi, Satoshi Ikeda, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)

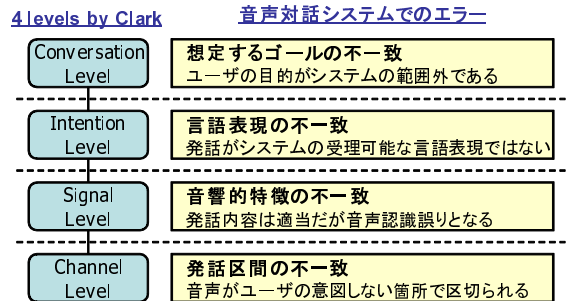


図 2: 音声コミュニケーションにおける誤りの 4 階層

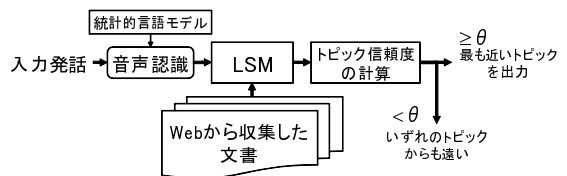


図 3: トピック推定の概略

2. 想定外発話に対するトピック推定 [4]

本研究では、想定外発話への対処として“トピック”を導入する。トピックは『ユーザが本来意図していたドメイン』に対応する言語表現の集合として定義する。ここでドメインとは、あるタスクを行うサブシステムにより受理解釈可能な範囲とする。つまり、あるサブシステムの言語理解部が文法ルールで記述されている場合、その文法で受理可能な発話の集合がドメイン内発話となる。このときトピックを定義し推定することで、システムがその文法で正確に受理解釈できない場合でも「何について話しているか」を推定できる。これにより、想定外発話をその内容に応じて分類し、より詳細なヘルプ生成など適切な応答が可能となる。

我々は、マルチドメイン音声対話システム [6] の各サブシステムに対応させてトピックを定義し、トピック推定を行った。以降、レストラン、観光案内、バス、ホテル、天気 の 5 ドメイン音声対話システムを例として説明を進める。これらに加えて「はい」などどのドメインにも共通する発話をコマンドトピックとする。トピック推定の概略を図 3 に示し、以下で順に説明する。

Webからの学習データの収集: 5 つのトピックに関して、統計的言語モデル作成ツール [7] を用いて、Web から学習データを収集する。まず各トピックごとに人手で 10 個前後のキーワードを検索エンジンに入力する。例えばホテルに関するキーワードは「泊まる」「旅館」などである。また、Wikipedia*から各トピックに関連するテキストを数百文集めて言語モデルを作成する。この言語モデルによるパープレキシティを、検索エンジンで収集し

*<http://ja.wikipedia.org/>

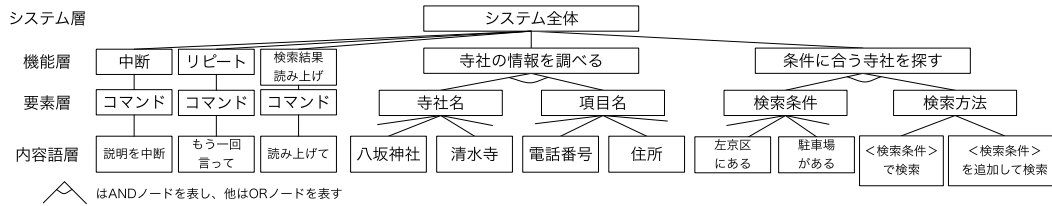


図 5: 京都寺社案内システムのドメイン概念木

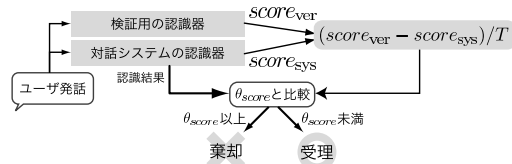


図 4: 発話検証の概略

た各文に対して計算し、値の小さい順に 10 万文を学習データとして取得する。さらにシステムの言語理解用文法から各トピックごとに 1 万文ずつ生成し、学習データに加えた。以上の作業で収集した文書をトピックごとに d 個 ($d = 20$) に分割し、学習文書とした。コマンドトピックの学習データは 175 文を人手で準備した。

Latent Semantic Mapping を用いたトピック推定: 各トピックに対する学習文書集合と入力発話との近さを Latent Semantic Mapping (LSM)[8] により計算し、トピック推定結果を得る。各学習文書での単語の頻度を表す $M \times N$ 共起行列を作成し (M は学習文書集合に現れる異なり単語数, N は学習文書数), その共起行列に対して特異値分解と次元縮約を行って共起行列の階数を k に減じる。本研究で作成した共起行列は, $M = 67533, N = 120, k = 50$ とした。入力発話の音声認識結果と、各文書との距離を k 次元空間上で計算した。ここでの音声認識には LSM の学習データから学習した言語モデルを用いた。

3. 発話検証技術に基づく文法外発話の検出[5]

2 つの音声認識器を利用して、音響尤度差による発話検証を行う。認識器の一方は、音声対話システムが利用するもので、システムが認識・解釈に用いる文法に沿っている。他方は検証用のモデルを持つ認識器で、検証用モデルにはより語彙サイズの大きいモデルを使用する。音響モデルは共通とする。

ユーザの発話がシステムの文法に近いかどうかは、この 2 つの認識器の音響尤度差を利用して推定する。ここで、検証用の認識器の音響尤度を $score_{ver}$ 、システムの認識器の音響尤度を $score_{sys}$ 、発話長を T 、閾値を θ_{score} として、以下の式で判別する。

$$\begin{cases} S = (score_{ver} - score_{sys})/T < \theta_{score} & (Accept) \\ & \geq \theta_{score} & (Reject) \end{cases} \quad (1)$$

つまり、尤度差 S が閾値 θ_{score} 未満ならシステムの文法に近い発話として受理し、 θ_{score} 以上ならば遠い発話なので棄却する。図 4 に処理の概略を示す。この手法は、システムの文法に近い発話に対しては、両認識器が音韻的に近い認識結果を出力し、遠い発話に対しては、システム側の認識器が無理に単語列を当てはめることで、検証用の認識器と音韻的に離れた認識結果を出力するという事実を利用している。

4. 推定結果の統合

想定外発話に対して適切な応答を生成するには、想定外発話に対する音声認識誤りを誤って受理しないだけでなく、当該発話からユーザを誘導するのに必要な情報を得る必要がある。2 章で述べたトピック推定は、システム想定外発話の音声認識結果に対して、その内容がシステム内のどの話題に近いかを推定する。つまり、Intention Level のエラーをさらに分類することで、より詳細なヘルプメッセージの生成に役立つ。さらにユーザ発話のトピック推定結果が、システムのどの話題からも遠い場合は、Conversation Level のエラーであることも示唆され、ヘルプ生成に役立てることができる。3 章で述べた発話検証による文法外発話の検出は、Intention Level のエラーの中で、その原因が文法レベルの不一致にあるのか、内容語レベルの不一致にあるのかを推定できる。

これらの情報を、我々が以前に開発したユーザのメンタルモデル推定 [9] と統合する。具体的には、図 5 に示されるようなドメイン概念木上でユーザの既知度を管理し、各推定結果を反映させて更新する。これにより、各ユーザの状態に応じたヘルプ生成が可能となる。各推定結果の精度は 100% ではないため、その信頼度に応じた応答選択も必要である。このような最適な対話戦略の検討や、具体的なヘルプ表現の生成を今後進めていく。

参考文献

- [1] Norman, D. A., 野島 久雄訳: 誰のためのデザイン? - 認知科学者のデザイン原論, 新曜社 (1990).
- [2] 駒谷和範, 上野晋一, 河原達也, 奥乃博: 音声対話システムにおける適応的な応答生成を行うためのユーザモデル, 電子情報通信学会論文誌, Vol. J87-D-II, No. 10, pp. 1921-1928 (2004).
- [3] Clark, H. H.: *Using Language*, Cambridge University Press (1994).
- [4] 池田智志, 駒谷和範, 尾形哲也, 奥乃博: ドメイン拡張性を備えたトピック推定に基づく発話誘導を行うマルチドメイン音声対話システム, 人工知能学会研究会資料, SIG-SLUD-A701-14 (2007).
- [5] 福林雄一朗, 駒谷和範, 尾形哲也, 奥乃博: 音声対話システムにおけるヘルプ生成のためのシステム想定外発話の誤受理抑制, 情報処理学会研究報告, 2007-SLP-65-12, 2007-HI-122-12, pp. 61-66 (2007).
- [6] 神田直之, 駒谷和範, 中野幹生, 中臺一博, 辻野広司, 尾形哲也, 奥乃博: マルチドメイン音声対話システムにおける対話履歴を利用したドメイン選択, 情報処理学会論文誌, Vol. 48, No. 5, pp. 1980-1989 (2007).
- [7] Misu, T. and Kawahara, T.: A bootstrapping approach for developing language model of new spoken dialogue systems by selecting Web texts, *Proc. INTERSPEECH*, pp. 9-12 (2006).
- [8] Bellegarda, J. R.: Latent Semantic Mapping., *IEEE Signal Processing Mag.*, Vol. 22, No. 5, pp. 70-80 (2005).
- [9] 福林雄一朗, 駒谷和範, 尾形哲也, 奥乃博: 音声対話システムにおける発話パターンを教示するヘルプの動的生成, 人工知能学会研究会資料, SIG-SLUD-A601-03 (2006).