4R-02

未知語認識機能を有する音声対話システムの構築とデータ収集

大塩 幹 † 武田 龍 † 駒谷 和範 †

† 大阪大学 産業科学研究所

1. はじめに

未知語の認識は，音声対話システムにおいて，未知語を含む発話の返答や語彙の獲得という観点で重要な技術である．ここで未知語とは，システム内部の語彙リストに登録されていない語を指す．例えば，商品名や人名，新語などが未知語の例としてあげられる．従来の単語辞書ベースの音声認識では未知語は認識できない．

本研究の目的は，聞き返し質問により未知語を認識する音声対話システムの実現である．我々は以前ユーザ返答パターンを用いた未知語認識手法を開発した [1]．この手法では，未知語を含むユーザ発話に対してシステムから聞き返し質問を行い，数ターンかけて未知語を認識することを前提としている．聞き返し質問に対するユーザ返答に対して未知語認識を行う．未知語認識の際には，ユーザ返答パターンを用いるジョイントモデルと，音節単位の音声認識と単語分割を直列に組み合わせたモデルの 2 つのモデルを用いる．この手法について，別の音声対話システムで収録したデータでは評価が行われている．しかし，この手法を用いて音声対話システムの開発を行っていなかったため，音声対話システムに用いた際の性能は評価されていない．我々はこの未知語認識手法を対話システムで用いた際の性能について評価する必要がある．

本稿では，未知語認識機能を有する音声対話システムの構築と，それを用いたデータ収集の計画について述べる．前者では，複数のモデルを用いる音声認識部の構造と，対話の中で未知語認識を行うための対話遷移について詳しく述べる．後者では，データ収集の目的を明確にし，実験条件や評価手法について検討する．

2. 音声対話システムの構成要素の前提知識

本章ではシステム構築に必要となる End-to-End ASR，単語分割，ジョイントモデルについて述べる．

End-to-End ASR は，式 (1) のように音声波形 $\mathbf{X}$ から事後確率が最大となるテキスト列 $\mathbf{c}$ を推定するモデルである．出力テキスト列 $\mathbf{c}$ には，音節単位（仮名文字）や文字単位（仮名，漢字，数字やアルファベットを含む文字）など複数の単位が考えられる．認識結果の文字列に加え，音声認識スコアも推定される．

$$\hat{\mathbf{c}} = \mathrm{argmax}_{\mathbf{c}} p(\mathbf{c}|\mathbf{X}) \tag{1}$$

単語分割は，式 (2) のようにテキスト列 $\mathbf{c}$ から単語境界 $\mathbf{z}$ を推定し，単語単位に区切られたテキスト列を出力するモデルである．音声認識同様，単語単位のテキスト列に加え単語分割スコアも推定される．

$$\hat{\mathbf{z}} = \mathrm{argmax}_{\mathbf{z}} p(\mathbf{z}|\mathbf{c}) \tag{2}$$

ジョイントモデルは，ユーザ返答パターンを用いて，式 (3) のように音声波形 $\mathbf{X}$ から音節列 $\mathbf{c}_{syll}$ と単語境界 $\mathbf{z}$ を同時に推定するモデルである．

Construction of Spoken Dialogue System with OOV Word Detection and Data Collection: *Miki Oshio, Ryu Takeda, and Kazunori Komatani (Osaka Univ.)*

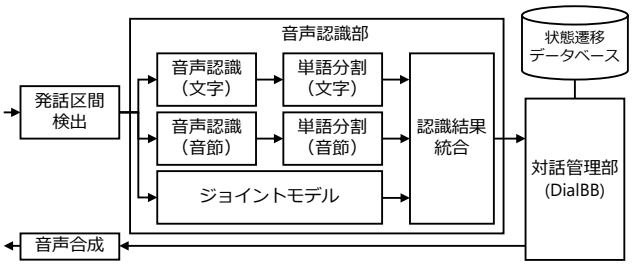

図 1: 対話システムの全体像

$$\hat{\mathbf{c}}_{syll}, \hat{\mathbf{z}} = \mathrm{argmax}_{\mathbf{c}_{syll},\mathbf{z}} p(\mathbf{c}_{syll}, \mathbf{z}|\mathbf{X}) \tag{3}$$

システムからの聞き返し質問に対する返答という特定の状況に特化した有限状態トランスデューサ（FST）を構築し，それをユーザ返答パターンとして推定の際に用いる．そのため，ジョイントモデルは聞き返し質問に対する返答を前提としている．FST は収集されたユーザ返答データを用いて，それらをすべて受理するように構築される．このモデルの実装は，ESPnet[2] のような言語モデルを用いる End-to-End ASR の言語モデルスコアに，FST の受理を表現することで実現される．

3. 未知語認識可能な音声対話システムの構築

図 1 のように，発話区間検出，音声認識部，状態管理部，音声合成の 4 つのブロックに分けて対話システムを構築した．本システムでは DialBB[3] を用いてルールベースで対話状態を管理する．次項以降では特に重要な部分として，図 1 に示す音声認識部の構成と，対話管理における状態遷移データベースの作成について述べる．

3.1 音声認識部の構成

音声認識部では，図 1 に示すように 3 つのモデルを用いており，それぞれの音声認識と単語分割，その結果の統合を行う．未知語を含む発話に関してはジョイントモデルと音節単位のモデルを用いて未知語認識を行い [1]，未知語を含まない発話に関しては文字単位のモデルを用いる．対話管理部では DialBB を用いるため，各モデルの音声認識結果，単語分割結果，音声認識スコア，単語分割スコアを辞書形式に格納して対話管理部へ入力として渡す．

3.2 対話管理の状態遷移

対話中において，どのタイミングでユーザの発話に未知語が含まれるかを事前に知ることはできない．そのため，ユーザ発話に未知語が含まれていると判断されれば，どの状態からでも未知語認識を実行できるような対話状態遷移を設計する．システム発話によってある程度ユーザの応答は予測できる場合は存在するが，意図しない応答により未知語が含まれる可能性は常に存在する．

状態遷移データベースでは，通常フローと未知語認識フローの 2 つのフローを設計した (図 2)．通常フローの対話でユーザ発話に未知語が含まれると判断した際に，未知語認識フローに遷移し，システムから聞き返し質問をしながら未知語認識を行う．

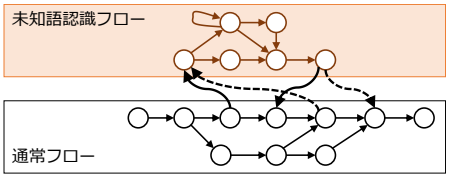

図 2: 通常フローと未知語認識フローの対話状態遷移（図中の○は対話状態，矢印は状態遷移を表す）

ユーザ発話に未知語が含まれるかどうかの判断には，音声認識スコア p_{rec} と単語分割スコア p_{seg} を用いる．本システムでは，文字単位と音節単位のそれぞれのモデルについて，音声認識スコアと単語分割スコアの和を比較する．式 (4) のように両モデルのスコア差 d を計算し，d の正負によって未知語が含まれているかを判断する．

$$d = p_{rec}^{syll} + p_{seg}^{syll} - (p_{rec}^{char} + p_{seg}^{char}) \tag{4}$$

ここで，*char*，*syll* はそれぞれ文字単位，音節単位の音声認識，単語分割スコアであることを示している．未知語が含まれる発話では未知語部分が誤った仮名漢字交じりの文字列で出力されたり，またそれにより不自然な単語分割がなされることがある．そのため未知語が含まれる発話では文字単位でのモデルの音声認識スコアや単語分割スコアが低くなる．

任意の状態から未知語認識を実行できるようにするため，本システムでは DialBB の状態管理ブロックで Subdialogue モジュールを用いた．これを用いることで，ある対話状態において条件を満たしたときに別の対話フローに遷移できる．遷移先の対話フローが終了すると指定した元の対話フローの次の状態に遷移する（図 2）．これにより，ユーザ発話に未知語が含まれると判断した際に，通常フローから未知語認識フローへ遷移する．

4. 未知語を含む対話データの収集

4.1 実験目的

データ収集の目的は 2 つ存在する．1 つ目は未知語認識手法の評価である．我々が開発した未知語認識手法について，発話単位の未知語認識の性能評価は行われている．一方で，未知語が含まれるかの判断などを加味した対話単位での未知誤認識の性能評価は行われていない．構築した音声対話システムを用いてデータを収集することで，対話単位での未知誤認識の性能評価を行う．

2 つ目はユーザ返答データの収集である．ユーザ返答パターン FST の作成に用いるデータを収集し，より多様な返答パターンを受理する FST の構築を目指す．未知語のタイプやドメインによる返答パターンの差を分析するために，複数タイプの未知語や異なるドメインを設定し，対話データを収集する．

上記の目的のための条件や評価尺度などについて次節以降で検討する．

4.2 実験条件

システムからの未知語確認質問に対するユーザの返答を収集する．効率的に未知語に関する質問応答の対話を収集するために，ユーザが用いる未知語はあらかじめユーザに指定する．何も教示を与えない状態では未知語が発話される可能性が低く，目的に即したデータ収集が効率的に行えない．

未知語認識の挙動やユーザ返答の傾向を分析するため，ここでは複数のタイプの未知語を設定する．まず，そもそも世の中に存在しない語を設定する．アストルテ，ヌフロなどが例としてあげられる．これは，今後大規模言語モデルを用いる際に，確実に語彙に含まれない単語として用いるために収集する．次に，世の中に存在する語として以下の 2 タイプを設定する．1 つ目は海外のスイーツなど，現状のシステム語彙には存在しないが世の中には存在する語である．ザッハトルテやヨウルトルットゥなどが例としてあげられる．2 つ目は日本のお菓子の商品名など，一部に既知語を含む語である．シゲキックスやピュレグミなどが例としてあげられる．

表 1: スイーツに関する対話シナリオの例（S はシナリオ発話，U はユーザ発話，**太文字部分**が未知語）

S1	最近何か変わったものを食べましたか？
U1	最近だと**マリトッツォ**を食べました．
S2	マリトですか？
U2	いえ，**マリトッツォ**ですね．
S3	マリトッツォですか？
U3	そうです．
S4	マリトッツォですね．分かりました．

表 2: 人名に関する対話シナリオの例（S はシナリオ発話，U はユーザ発話，**太文字部分**が未知語）

S1	あなたの周りで下のお名前の読み方が難しい方はいますか？
U1	**シアン**ちゃんっていう名前の子がいます．
S2	サンさんですか？
U2	シアンちゃんですね．
S3	シアンさんですか？
U3	そうですシアンちゃん．
S4	シアンさんですか． ちなみに漢字ではどのように書くんですか？
U4	星空の星に心って書いてシアンって読みます．
S5	なるほど，分かりました．

対話シナリオは，スイーツに関する対話（表 1）と読みが難しい人名に関する対話（表 2）の 2 つのケースを設定する．主としてデータを収集するのは前者のスイーツに関する対話データである．これは未知語のスイーツ名を含む発話に対して聞き返し質問を行う数ターンの対話である．異なるドメインでの未知語認識の性能を評価するために，後者のシナリオでは人名に関する話題を設定する．今回の未知語認識手法では音節単位で認識しており漢字表記などについては考慮していない．未知語の表記について今後検討を行うために，シナリオの後半では漢字表記を聞き取る対話も収集する．

4.3 評価尺度

発話単位と対話単位の 2 種類の観点で評価を行う．発話単位の未知語認識では，ユーザ返答ごとに正しく未知語認識できているか判定する．表 1 の例では，U1 に対して誤って認識しており，U2 に対しては正しく認識しているため，正解率は 50% (1/2) となる．対話単位での未知語認識では，対話の最終的な結果が正しく未知語認識できているか判定する．同じく表 1 の例では，最終的には正しく「マリトッツォ」を認識できているため，正解率は 100% (1/1) となる．今後実際に収集したデータにおいて，これらの尺度に基づき評価を行う．

参考文献

[1] M. Oshio, et al. Out-of-vocabulary word detection in spoken dialogues based on joint decoding with user response patterns. In *Proc. APSIPA ASC*, pp. 1753–1759, 2023.

[2] S. Watanabe, et al. ESPnet: End-to-end speech processing toolkit. In *Proc. Interspeech*, pp. 2207–2211, 2018.

[3] 中野幹生, 駒谷和範. DialBB: 情報技術の教材を指向した対話システム構築フレームワーク. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 96, p. 39, 2022.