4S-04

対話中のユーザの返答パターンに基づく音声発話中の未知語認識

大塩 幹 † 宗像 北斗 † 武田 龍 † 駒谷 和範 †
† 大阪大学 産業科学研究所

1. はじめに

音声対話システムでは，システム内の単語辞書に登録されていない単語（未知語）は正しく認識されない．システム内の単語辞書に人手で未知語を拡充することもできるが，日々新たな単語が登場している現状では，定期的に人手で単語辞書を更新する必要がある．ユーザの発話からシステムが未知語を自動で認識し単語辞書に登録できれば，人手で単語辞書を更新する手間が省ける．

End-to-End モデルの発展により，音節単位での音声認識（音節認識）と単語分割を組み合わせた未知語認識が可能になりつつある．しかし，音節認識と単語分割を単純に組み合わせるだけでは精度良く未知語認識を行うことはできない．一般的な End-to-End モデルは，特定の状況に特化していない大規模な言語コーパスに基づいて汎用的なモデルを構築するため，コーパスにない音節列に対して正しく単語分割できない．未知語認識を扱う先行研究では，対話の状況や，対象発話以外のシステム発話などの影響が活用されていない [1].

本研究では，ユーザの発話音声からの未知語認識を目指す．具体的には図 1 のように「違うよ，マリトッツォだよ.」というユーザの発話から「マリトッツォ」という未知語を切り出すことを目指す．

提案法は，図 1 のようにシステムが未知語に関する明示的確認を行う状況において，ユーザは決まったパターンで返答することが多い性質を利用する．ユーザ返答パターンモデルを導入し，ユーザの返答に制約を与えることで未知語周辺の単語分割の精度向上を図る．また従来の汎用的なモデルとユーザ返答パターンモデルを併用することで，事前に用意したパターンに適合しない状況にも対応する．

2. ベースとなる手法

2.1 定式化

本研究では，音声信号から対応する音節と単語境界を同時に推定し，内部の単語辞書に含まれない単語を未知語として推定する（図 2 上）．ここでは，音節列は片仮名で表現する．単語辞書はユニークな単語の集合であり，音節列の形で単語は辞書に登録されている．

未知語認識における音節と単語境界の同時推定を次の事後確率を最大化する探索問題として扱う．

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

ここで，$\mathbf{X} = [\mathbf{x}_1,...,\mathbf{x}_T], \mathbf{x}_t \in \mathbb{R}^D$ $(t = 1,...,T)$ は長さ T の入力の音声特徴量系列，$\mathbf{c} = [c_1,...,c_L]$ は対応する長さ L の音節列，$\mathbf{z} = [z_1,...,z_L], z_l \in \{0,1\}$ $(l = 1,...,L)$ は単語境界を表す長さ L の 2 値の指示変数列を表す．音節 c_l の直後に単語境界がくる場合に $z_l = 1$ となり，それ以外では $z_l = 0$ となる．例えば，$\mathbf{c} = [$ア, メ, カ, ゼ$], \mathbf{z} = [0,1,0,1]$ の場合，トークンは「アメ」と「カゼ」の 2 つとなる．

OOV Word Recognition Based on User Response Patterns during Spoken Dialogues: *Miki Oshio, Hokuto Munakata, Ryu Takeda, and Kazunori Komatani (Osaka Univ.)*

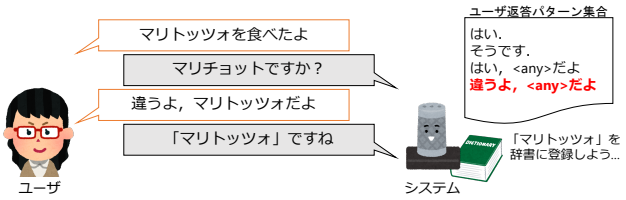

図 1: システムの自動的な未知語認識の理想例

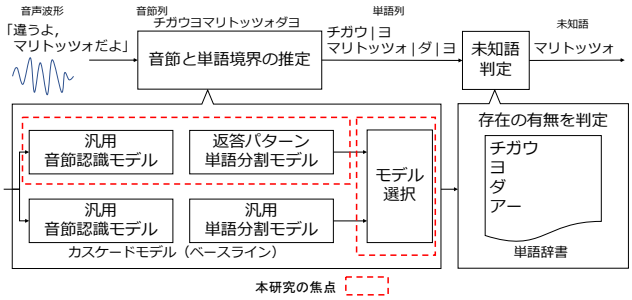

図 2: 未知語認識の枠組みと本研究の焦点

単語辞書 V に「カゼ」のエントリが無い場合,「カゼ」が未知語として認識される．

本研究では簡単化のため，事後確率を次の条件付き確率で近似し，音節認識を介して再定式化する．

$$p(\mathbf{c},\mathbf{z}|\mathbf{X}) \approx p(\mathbf{c}|\mathbf{X})p(\mathbf{z}|\mathbf{c}) \tag{2}$$

ここでは，$p(\mathbf{c}|\mathbf{X})$ を音節認識モデル，$p(\mathbf{z}|\mathbf{c})$ を単語分割モデルと呼ぶ．これら 2 つのモデルを定義し，事後確率を最大化する $\hat{\mathbf{c}},\hat{\mathbf{z}}$ を見つけることが目的となる．

2.2 単純な手法：カスケードモデル

カスケードモデルでは，まず音節認識を行い，次に認識された音節列に対して単語分割を行って，近似解を得る．

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

音節認識は，深層学習に基づく End-to-End 音声認識の枠組み [2] を利用する．深層モデルにより，確率 $p(\mathbf{c}|\mathbf{X})$ を最大化する解をビームサーチで探索する．

単語分割はベイズモデリングに基づく手法 [3] を適用する．確率分布から $\mathbf{z}$ をサンプリングし，高頻度の分割パターンを解とみなす．

3. ユーザ返答パターンモデルの利用

未知語を汎用的なモデルで正しく未知語認識するのは困難である．そこで，システムからの明示的確認に対する返答という特定の状況を利用する．ユーザ返答パターンを用いて，その状況に特化したモデルを利用することで，汎用的なモデルに比べて精度の高い未知語認識を目指す．

3.1 単語分割モデルのモデル選択

ユーザ返答パターンモデルは，パターンに適合する発話に対しては有効である一方，適合しない発話を扱うことが困難である．パターンに適合しない発話も扱うため，発話に応じて汎用的なモデルと切り替える．

本研究では，音節と単語境界の同時推定を，ユーザ返答パターンモデルと汎用単語分割モデルのモデル選択も同時に行う問題として定式化する．

$$\hat{\mathbf{c}},\hat{\mathbf{z}},\hat{h} = \mathrm{argmax}_{\mathbf{c},\mathbf{z},h}\ p_h(\mathbf{z}|\mathbf{c})p(\mathbf{c}|\mathbf{X}) \tag{4}$$

ここで，$h \in \{g,d\}$ はモデルクラスであり，$\hat{h}$ は最適なモデルクラスである．また g は汎用的な単語分割モデル，d はユーザ返答パターンモデルによる単語分割モデルを表す．モデルに関する最大化は p_g と p_d に基づいた推定を独立に行い，スコアが高い方を選択する（図 2 下）．ここで，p_g を用いた推定は式 (3) で行えばよい．ユーザ返答パターンモデルの定義と推定方法は次節で述べる．

3.2 ユーザ返答パターンモデル

システムによる明示的確認に対するユーザ返答にはパターンが見られ，未知語の出現位置や発話に含まれる語彙には傾向が見られる．ユーザ返答パターンモデルでは，まず事前に与えられたユーザ返答データ（音節列と単語境界）からユーザ返答パターン集合 U を構築する．次に U に含まれるパターンに適合する音節列候補のみ受理するという制約を与える．例えば図 1 では，「違うよ，<any>だよ」というパターンに適合する音節列候補が受理され「違うよ，マリトッツォ<any>だよ」という認識結果を得られている．これは，ユーザ返答パターンモデル $p_d(\mathbf{z}|\mathbf{c})$ を次の関数 $f(\mathbf{z},\mathbf{c})$ でモデル化することと等価であり，音節認識の際の候補集合を制限することに相当する．

$$f(\mathbf{z},\mathbf{c}) = \begin{cases} \text{const.} & \text{if } [\mathbf{z},\mathbf{c}] \in U \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

この時，未知語の部分に関しては，任意の音節列を受け付けるように柔軟性を持たせる必要がある．

任意の音節列で未知語を表現した有限状態トランスデューサ（FST）で，ユーザ返答パターン集合 U を記述した．それを音節認識のデコーディング時の制約として用いることで候補集合を制限した．End-to-End 音声認識にこの制約を組み込むのは，式 (5) を言語スコアとして扱うことにより実現できる．

3.3 モデル選択基準

音節認識モデル $p(\mathbf{c}|\mathbf{X})$ の尤度を用いてモデル選択を行う．汎用的なモデルとユーザ返答パターンモデルの 2 つの認識音節列に対する音節認識モデルの対数スコアの差を測り，閾値 θ によってモデルを判断する．式 (4) の h の探索を次のように行う．

$$\hat{h} = \begin{cases} d & \text{if } \log(p(\mathbf{c}_g|\mathbf{X})) - \log(p(\mathbf{c}_d|\mathbf{X})) \le \theta \\ g & \text{if } \log(p(\mathbf{c}_g|\mathbf{X})) - \log(p(\mathbf{c}_d|\mathbf{X})) > \theta \end{cases} \tag{6}$$

ここで，$\mathbf{c}_g$ は p_g を用いた認識結果の音節列，$\mathbf{c}_d$ は p_d を用いた認識結果の音節列である．

4. 評価実験

4.1 実験条件

音節認識には ESPnet [4] を使用した．ベースライン手法の汎用的な単語分割モデルには，Nested Pitman-Yor Language Model [3] を用いた．また，モデル選択時の閾値は $\theta = 1.0$ とした．式 (5) における定数は 1 とした．モデルの選択手法の性能を見るために，適切な未知語切り出しを行えるように，手動でオラクル的にモデル選択を行ったケースについても評価した．

表 1: 実験データの規模

データ	発話数	音節数	単語数	未知語数
CSJ	380,872	15,170,650	7,251,909	0
FST 作成	275	3,121	1,421	72
テスト	278	3,217	1,419	89

表 2: 未知語切り出しの評価

モデル	適合率	再現率
ベースライン	0.40 (36/ 91)	0.41 (36/86)
ユーザ返答パターンモデル	0.13 (22/167)	0.25 (22/89)
モデル選択	0.38 (35/ 91)	0.41 (35/86)
モデル選択（オラクル）	0.48 (42/ 87)	0.49 (42/86)

表 3: 未知語認識の評価

モデル	適合率	再現率
ベースライン	0.30 (27/ 91)	0.31 (27/86)
ユーザ返答パターンモデル	0.04 (7/167)	0.08 (7/89)
モデル選択	0.27 (25/ 91)	0.29 (25/86)
モデル選択（オラクル）	0.32 (28/ 87)	0.33 (28/86)

4.2 実験データ

表 1 に実験で用いたデータの規模を示す．表 1 の未知語数は，データに含まれる単語の中で音節認識モデルの学習データに含まれていない単語の数を示す．音節認識モデルと汎用単語分割モデルの学習には日本語話し言葉コーパス（CSJ）を用いた．また，ユーザ返答パターンモデルの作成データとテストデータには，本研究室で収集した海外の食べ物に関する音声対話データを用いた．このうち，システムの明示的確認に対するユーザ返答のみを抽出し，ランダムに 2 分割して，ユーザ返答パターン FST 作成データとテストデータとした．

また，CSJ のユニークな単語の集合をシステムの単語辞書として用いた．単語辞書の規模は 55463 語である．

4.3 実験結果と考察

未知語切り出し，未知語認識という 2 つの観点それぞれについて，適合率と再現率の指標で評価を行った．未知語切り出しは音節列の未知語の開始位置と終了位置の正誤を評価する．未知語認識は得られた未知語が音節認識誤り無く正解データの未知語と一致しているかを評価する．

表 2-3 に結果を示す．提案手法のモデル選択では，ユーザ返答パターンモデルのみに比べて高い精度が得られたものの，ベースラインに比べてどの指標においても 0.1 から 0.3 ポイント性能が下がった．

オラクル的にモデル選択を行ったものは，未知語切り出しの適合率が 0.48，再現率が 0.49 と，ベースラインに比べて 0.08 ポイント程度高い精度を示した．モデル選択の手法は改善が必要である．

参考文献

[1] H. Kamper, et al. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM TASLP*, Vol. 24, No. 4, pp. 669–679, 2016.

[2] S. Watanabe, et al. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE JSTSP*, Vol. 11, No. 8, pp. 1240–1253, 2017.

[3] D. Mochihashi, et al. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proc. ACL-IJCNLP*, pp. 100–108, 2009.

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