4S-8

音声発話の誤分割修復のための連続する発話対の同一発話判定

堀田 尚希 † 駒谷 和範 † 佐藤 理史 † 中野 幹生 ‡
† 名古屋大学大学院 工学研究科 ‡ ホンダ・リサーチ・インスティチュート・ジャパン

1. はじめに

音声対話では，適切な応答内容とともに適切なターンテイキングが重要である．ターンテイキングとは一般的に，二人の話者が交互に発話を行うことである．音声対話システムが満たすべき条件のひとつに「ユーザが話している途中にシステムが話し始めない」ことが挙げられる．

現在の音声対話システムでは，ユーザが発話の途中で言い淀んだ場合に，ユーザの発話の途中でシステムが話し始めることがある．これは言い淀みにより元来一発話であったユーザの発話が，誤って複数の発話区間に分割され，その断片に対してシステムが応答を始めるためである．本研究ではこの現象を発話の誤分割と呼ぶ．

我々はこれまでに，発話の誤分割により生じる2つの問題を，事後的に修復する手法を提案した [4]．

1. ユーザの発話中にシステムが話し始める
   → そのシステム発話を停止する遷移を追加
2. 誤った発話区間に対して音声認識がなされる
   → 発話断片を結合して再度音声認識を行う

本稿では，この修復が必要か否かを，より高精度に判定する．具体的には，2つの発話断片を入力とし，これらが一発話か否かの判定を行う．本手法の概要を図1に示す．一発話である場合は，2つの発話断片を統合し，一つの発話として再度音声認識を行うなどして解釈する．一発話でない場合は，各々の発話断片に対して個別に応答が決定される．

本研究では，発話断片から得られる様々な特徴を用いて，修復が必要か否かの判定を行う．我々は以前，この判定を，2つの特徴のみを使ったルールにより行っていた [4]．我々はこれら以外にも判定に有用な特徴があると考え，それらを用いてより高精度に判定を行うことを目標にする．具体的には発話断片の対から得られる様々な特徴を用いた二値分類問題とし，これを機械学習により解く．

2. 判定すべき正解ラベル

本研究では連続する発話断片 (発話区間検出結果) の対に対して，以下のいずれかのラベルを付与する．

- 元来一発話である発話断片の対
- 元来一発話ではない発話断片の対

我々は元来一発話か否かを判定するために，各発話断片に発話の要点となる情報，つまり検索に必要なキーワードが含まれているかに着目する．元来一発話である発話断片の対は，統合して解釈すべきである．この具体例を

Detecting Originally-Single Utterances from Successive Two Segments for Restoring Incorrectly-Segmented Spoken Utterances: *Naoki Hotta, Kazunori Komatani, Satoshi Sato (Nagoya Univ.), and Mikio Nakano (Honda Research Institute Japan Co., Ltd.)*

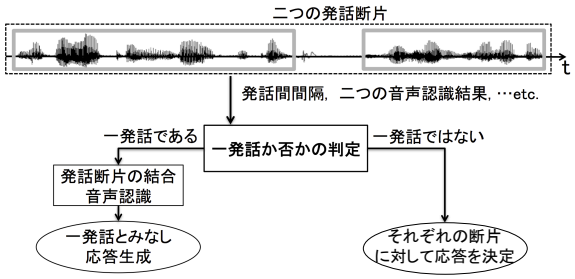

図 1: 手法の位置づけ

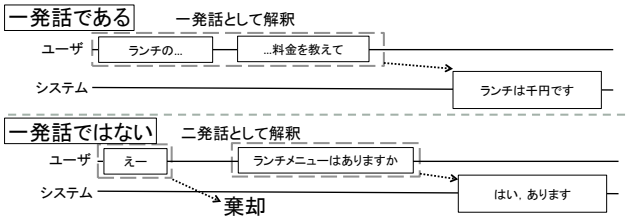

図 2: それぞれの正解ラベルの具体例

図2の上半分に示す．前半断片には “ランチ”，後半断片には “料金” というキーワードが含まれている．この場合，発話断片を結合して再度音声認識を行うのが望ましい．

一方，図2の下半分は，元来一発話ではない発話断片の対である．この発話の前半断片にはキーワードが含まれていない．この場合，発話の統合解釈は必要なく，後半断片のみに対して応答を生成することが望ましい．

3. 決定木による判定

時間的に近接した2つの発話断片に対して，元来一発話であるか否かを判定する．ここでは発話間間隔など計23個の特徴を用いて，元来一発話であるか否かを判定するための決定木を学習する．さらにその決定木が特定のドメインに依存しないことを確認する．

3.1 使用した特徴

決定木学習に用いる特徴として表1に示す18個を用いる．具体的には，Julius[§]の出力から得られる特徴が5つ，言語的特徴が3つ，時間的特徴が5つ，音響的特徴が5つである．この中から，判定に有効であった特徴の一部を，以下で説明する．

(2) 前半断片最終単語の音声認識信頼度

元来一発話である発話断片は，その断片に対する音声認識信頼度が低い．これは，誤分割が発生している場合，前半の発話断片は途中で区切られていることが多く，その部分に対する音声認識結果は信頼できないからである．

[§] http://julius.sourceforge.jp/

表 1: 決定木学習に用いた 18 個の特徴

Julius の出力から得られる特徴	**時間的特徴**	**音響的特徴**
(1) 前半断片の平均の音声認識信頼度	**(9) 発話間間隔**	(14) 前半断片の最終部の音量変化
(2) 前半断片の最終単語の音声認識信頼度	(10) 前半断片の silE の長さ	(15) 前半断片第 1 モーラの母音部分 F0 の傾き
(3) 前半断片の言語モデルスコア	(11) 後半断片の silB の長さ	(16) 前半断片全体の F0 レンジ
(4) 前半断片の音響モデルスコア	**(12) 前半断片の発話長**	**(17) 前半断片の最大音量**
(5) GMM を用いた雑音判別結果	**(13) 前半断片の最終モーラの継続時間**	(18) 後半断片の最大音量
言語的特徴		
(6) 音声認識結果の音素 bigram オーバラップ率		
(7) 前半断片のフィラー数		
(8) 後半断片のフィラー数		

表 2: 使用する発話データ

ドメイン	レストラン	世界遺産
総対話数	120 対話	156 対話
総検出区間数	6615	6593
対象とする発話対数	255 対	354 対

表 3: 決定木の判定性能

	レストラン→世界遺産	世界遺産→レストラン
ベースライン	263/354 (74.3%)	210/255 (82.4%)
特徴選択なし	289/354 (81.6%)	214/255 (83.9%)
特徴選択あり	291/354 (82.2%)	217/255 (85.1%)

(5) GMM を用いた雑音識別結果

本研究では発話間間隔が比較的短い発話断片の対を，調査対象としている．このため雑音や咳払いなど，ユーザの意図しない発話や雑音も調査対象に含まれる．これらを除去するために，文献 [1] で提案された，Gaussian Mixture Model (GMM) に基づく雑音識別結果を用いる．

3.2 ドメインに依存しない特徴の選択

決定木学習に使用する特徴は，学習データと異なるドメインでも有効であることが望ましい．本研究では 2 つのドメインにおける発話データを使用し，そのどちらでも有効である特徴を，ドメインに依存しない特徴とみなす．

このために，2 種類のドメインそれぞれのデータに対して特徴選択を行う．具体的には，両ドメインのデータにおいて，10 分割交差検定により決定木を構築する．その結果，ある特徴が 10 個の特徴セットに 1 つでも含まれる場合，その特徴を選択するとする．その後，双方のデータで共通して選択された特徴を，ドメインに依存しない特徴とみなす．

4. 評価実験

4.1 使用データ

評価には，2 つのドメインにおける発話データを用いる (表 2)．これらの発話データは，レストラン検索システム [3] および世界遺産検索システム [2] により収集された発話である．発話の誤分割が発生した可能性がある発話断片の対を調査対象とするため，文献 [4] と同様に，発話間間隔が近接しており，雑音ではない発話断片の対を抽出した．

発話データから繰り返し発話を事前に人手で除外した．これは，繰り返し発話は調査対象としている発話とは現象が異なり，同一の判定基準で分類を行うのは不適切であると考えたためである．なお予備実験により，音声認識結果間の音素 bigram のオーバーラップ率を用いることにより，85%–90%の精度でこれを自動で除外できることを確認している．

以降の実験では，レストランドメインの 255 対，世界遺産ドメインの 354 対の発話断片の対を用いる．

4.2 決定木の判定性能

3.2 節で述べた特徴選択の結果，7 個の特徴が得られた．これらの特徴は表 1 内で太字で示す (2)，(4)，(5)，(9)，(12)，(13)，(17) である．以降ではこれらの特徴を用いて決定木を構築し，その判定性能を調べる．決定木の構築には，Weka[¶] の J48 を使用した．

ベースラインと，特徴選択の有無での決定木の性能を比較する．ベースラインは，以前の研究 [4] で用いた 2 つの特徴，つまり統合後の音声認識信頼度と (9) 発話間間隔を使用した場合とした．“特徴選択なし” は，表 1 に示す 18 個の特徴を全て使用した場合，“特徴選択あり” は，特徴選択により得られた 7 個の特徴を使用した場合である．

決定木が特定のドメインに依存していないことを確認するため，クロスドメインにおける評価を行う．クロスドメインとは片方のデータセットで学習を行い，もう片方のデータで評価をすることである．例えば表 3 において，“レストラン→世界遺産” とは，レストランドメインのデータで学習し，世界遺産ドメインのデータで評価することを表す．

得られた決定木のクロスドメインにおける判定性能を表 3 に示す．まず “ベースライン” から “特徴選択なし” では，両ドメインとも判定性能が向上した．これは，特徴を増やしたことにより判定に有用な特徴が増加したことを示している．また “特徴選択なし” から “特徴選択あり” では，判定性能はほぼ同等，あるいはわずかに向上した．このことから，特徴選択を行わない場合は，データセットに依存する特徴があったと考えられる．

今後は，この判定法を取り入れ，オンラインで発話の誤分割を修復する音声対話システムを実装する．

参考文献

[1] A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari and K. Shikano. Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. Proc.ICSLP, pp.173–176, 2004.

[2] 佐藤隼, 中野幹生, 駒谷和範, 船越孝太郎, 奥乃博. ドメイン外発話が扱え拡張性が高い対話ドメイン選択フレームワーク. 情報処理学会研究報告, Vol. 2013-SLP86, pp.1–8, 2011.

[3] 西村良太, 駒谷和範. データベース検索音声対話システムにおける対話状態の推定. 情報処理学会研究報告, Vol. 2012-SLP-90, pp.1–7, 2012.

[4] 堀田尚希, 駒谷和範, 佐藤理史. ユーザ発話の誤分割に起因する問題を事後的に修復する音声対話システム. 情報処理学会研究報告, Vol. 2013-SLP-96, pp.1–8, 2013.

[¶] http://www.cs.waikato.ac.nz/ml/weka/