音声対話情報案内システムにおける機械学習理論に基づく自律的精度向上の研究

Study on Autonomous Improvement Approaches based on Machine Learning for Speech-oriented Information Guidance Systems

2021008

研究代表者	奈良先端科学技術大学院大学 情報科学研究科	助　教	川　波　弘　道
共同研究者	奈良先端科学技術大学院大学 情報科学研究科	准教授	猿　渡　　洋

［研究の目的］

音声対話に基づく情報案内システムはユーザの負担が少ないことから，万人に優しい情報アクセス方法として期待されている。我々は用例ベースの実環境システムを開発し，大人と子供の発話に対してそれぞれ約 80%，60% の応答正解率を達成しているが，本研究では更なる応答性能の向上のため機械学習によるトピック分類を導入する。また，質問用例選択による応答生成は，一般に質問用例数の増加に伴い性能が向上するため，人手によるユーザ発話の書き起しをデータベースに追加することで精度向上を実現してきた。しかしながらこれは人手による開発コストが甚だしい。そこで性能改善に寄与するデータを自動選択する手法を研究する。

［研究の内容，成果］

我々はこれまで音声情報案内システムの研究開発に従事し，2002 年以来，生駒市のマスコットキャラクターをエージェントとした用例ベース対話システム「たけまるくん」を同市のコミュニティセンターに設置，常時誰でも使える実環境システムとして運用してきた（図 1）。

用例ベースの質問応答システムは，質問例とその応答のペアを用意しておき，ユーザ発話の音声認識結果ともっとも類似した質問用例を選択，その対応する応答を出力する。本研究はシステム入力の音声認識結果を得た後の応答選択部に着目する（図 2）。

これまでは，実環境でのユーザ発話を書き起し，それを質問例に追加していくことで，ユーザが実際に行う質問内容や言い回しに頑健な応

図 1　音声情報案内システム「たけまるくん」（生駒市北コミュニティセンター「はばたき」に設置）

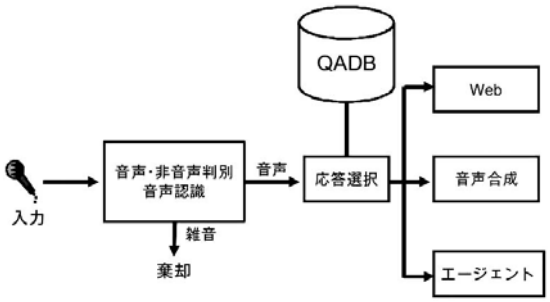

図 2 「たけまるくん」の処理の大まかな流れ（QADB は質問と応答のペアからなるデータベース）

答生成を実現してきた。現在のシステムでは，大人と子供別に用意した音声認識用モデルを用いて認識を行い，認識結果を参照する質問用例データベースも大人用と子供用とを別々に用意することで音響的・言語的に特徴の異なる幅広い年齢層に頑健なシステムを実現した。現行のシステムでは運用開始後の 2 年間で得られたユーザ発話の書き起しを用いた数万の質問用意例を用いて，約 300 種類の応答を行っている。

○ 現在のシステムの課題

本稿冒頭で述べたように収録されたユーザ発話を用いた応答性能を評価実験の結果から，現在のシステムの応答正解率（ユーザの質問総数に対して適切な応答を行った割合）は大人で約 80%，子供では約 60% である。ユーザはシステムが不適切な応答をした場合にはそれを察知して尋ね直すため，運用上の問題とはなっていないが，ユーザの負担軽減のために応答性能の更なる改善が必要である。本研究では，次の 2 つの観点で応答性能の改善に取り組んだ。

○ 研究内容

① 機械学習理論に基づくトピック分類

ユーザ発話ともっとも似た質問用例の選択には類似度という尺度を用いているが，これには認識結果と質問用例とが共有する単語（実際には単語に準ずるより小さな単位の形態素）の数を，文全体の単語数で正規化した値を用いている。しかしながら応答内容に直接影響する単語も，「言い回し」の違いに相当する置き換え可能な単語も対等に評価している。単語の重要度を反映させる方法として人手でキーワードを設定し重要視することも考えられるが構築コストの増大になるばかりでなく，開発者の経験の依存するため客観的な性能評価が困難となる。そこで，単語の頻度ベクトルを特徴量とし，トピック分類規則を機械が自動的に生成する機械学習の手法を導入した。実際の「たけまるくん」の応答は前述のように約 300 種類あるが，それぞれの応答に対応付けられた用例数にはばらつきがあり，学習データサイズも限定されるので，共通する話題（トピック）の応答文をまとめて 27 個のグループに分類した。トピックの例として「あいさつ」「館内施設案内」「生駒市紹介」「交通アクセス」「たけまるプロフィール」等を設定した。実際にシステムを運用する際はこれらのトピックに分類したのち，そのトピック内の質問用例を用いて応答を行うことを想定しているが，ここではトピック分類性能の評価に限定する。主要な特徴量として単語の頻度ベクトル（BOW ; Bag-of-Words）を用い，代表的な識別手法の分類性能を比較調査するとともに，応答性能の自律的改善の可能性を探るため半教師あり学習の有効性を調査した。

さらに，トピック分類の概念を拡大し，応答生成の前段で行われるシステム入力の分類，つまり，雑音も含むすべてのシステム入力に対してそれがシステムへの質問（有効発話）かそうでないか（無効入力）を分類する識別実験も行った。

② 質問用例のための発話書き起しコスト削減

これまで質問用例の拡大は，すべてのシステム入力（あるいはその中から音響尤度により雑音と識別されたデータを除いたもの）に対して熟練したラベラ―が聴取による書き起しを行い，システムへの有効な質問か判断し，対応する応答文の ID を付与して行っていた。これは人手がかかる高コストな作業である。そこで「いま書き起そうとしているユーザ発話は質問用例としてデータベースに追加することが性能向上に貢献するかどうか」という観点で選別する方法を研究する。例えば既存の用例と同一のものや著しく類似した質問は追加する必要がなく，また既存の用例と著しく異なる（類似度が低い）新奇な発話も，情報案内タスクから外れた発話や再現性の低い質問であると考えられる。それらをあらかじめ書き起しの対象とするデータから除外しておくことで書き起しコストの削減を目指す。本研究では既存質問用例との類似度，

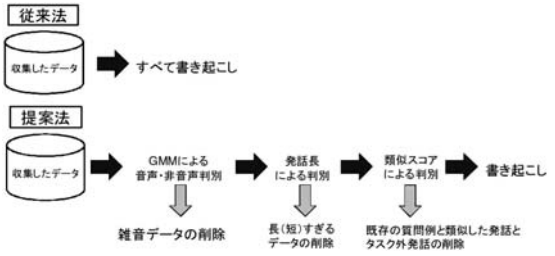

図3　応答性能向上に貢献する（書き起し，追加することが有効な）データの選択のながれ

発話長についての閾値の区間設定，雑音と音声の音響モデルの尤度への重みづけを実験的に設定することで（図3）開発コストを削減する。

○ 本研究の成果

① 音声認識結果の BOW を特徴量とした機械学習によるトピック分類の有効性を確認，子供音声における半教師あり学習での分類性能の改善，および同手法による無効入力棄却性能の改善

一般にユーザ発話は 2〜3 秒程度（1〜10 単語程度）のものが大多数であり書き言葉からなるテキストコーパスを対象とした自然言語処理の手法がそのまま援用できるかどうかは不明であったが，後述の Stacked Generalization（SG）法を適用した結果，27 トピック分類実験において大人発話の F 尺度（大きいほど高精度な分類がされている）は 95.1，子供で 89.2 に達した。（比較のための人手による書き起しデータを学習に用いた識別では結果それぞれ 96.6，96.8。）実験では約 2 年分のデータの音声認識仮説の 1-Best を学習データとし，まず SVM（サポートベクターマシン），pboost 法，最大エントロピー法のテキスト処理で使用される 3 手法の比較調査を行った。次にそれぞれの識別結果を用いて，2 段階目の識別を行う SG 法を導入することで更に高精度な識別ができることを確認した。3 手法を単独で用いた場合と SG 法を導入した場合との性能向上をまとめた論文は 2013 年 4 月の学術誌に掲載された［1］。

次にトピックラベルがまだ付与されていない大量のデータを活用する「半教師あり学習」による識別も検討した。人手でトピックラベルが付与されたデータで学習した識別モデルを用いて未知のデータを識別し，ある条件を満たすとそのラベルを採用して識別モデルを再学習する。識別器として SVM を用いているためトランスダクティブ SVM と呼ばれる。最大 2 年 9ヵ月分のラベルなしデータを用いた実験を行った。約 2 年間のラベル付きデータのみを用いて学習した SVM での分類性能の F 尺度は大人，子供でそれぞれ 93.0，83.5 であった。ラベルなしデータすべてを使用した場合，子供で 84.3 まで改善され教師ありデータのみを用いた学習を超える性能が得られることがわかった。ただし大人の場合は 92.8 程度に達したところで改善が頭打ちとなっており，教師ありデータのみを用いた学習での性能を超えなかった。子供データは全体に大人の倍以上の発話数があるため，発話の多様性とともに教師あり学習のデータ数とも関係していると考えられる。

最後に，BOW を用いた SVM による無効入力棄却実験を行った結果，特徴量として BOW のみを用いた場合で F 尺度が 89.1，さらに音声・雑音の音響尤度，S/N 比，発話長を加えた場合には 86.6 まで改善された。従来の音響尤度のみを用いた識別手法では 82.2 であったことからは大幅な改善が見られ，その有効性を示すことができた［2］。

② 適切な閾値設定がなされれば，開発データの過半数を除去して書き起しコストを削減しても応答性能の有意な低下は起こらない。（学習データ・開発データにそれぞれ 1 年分（1〜2 万発話）程度のデータが利用できる場合）

既存質問用例との類似度，発話長，音声と雑音の音響モデルの尤度を用いて，新規のデータを書き起すかどうか判定する。すべての開発データを書き起したときの応答正解率を維持したまま書き起しデータ量をどの程度削減することができるか調査した。雑音入力も含め「たけまるくん」で収録した 22ヶ月間のデータのうち，11ヶ月分を初期質問用例および雑音・音声モデル作成の学習データとした。1ヶ月分を評

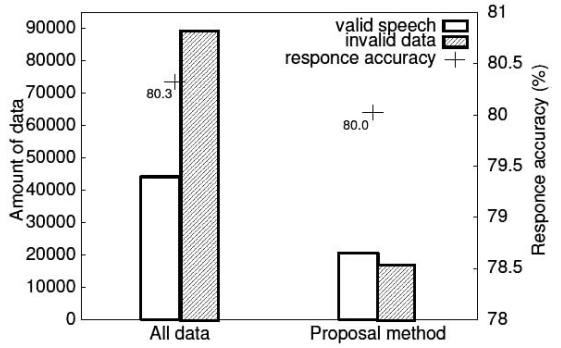

図 4　開発データをすべて書き起した場合（左）と提　案手法でデータを選択した場合（右）のデータ（棒の長さ）と応答正解率（＋記号）の比較。白い棒は有効発話，灰色の棒は無効入力の数を示す。

価データ，10ヶ月分を開発データとした。大人発話と雑音のみを用いて実験を行った。開発データは 16.316 個である。この条件では 0.3% 以内の応答正解率の差は統計的に有意と言えないことから，その範囲内での応答正解率低下は許容するという条件で，いかに書き起しデータ量を削減できるかを調査した。発話時間と類似度スコアの上限と下限，雑音と音声モデルの尤度の重みづけを様々に変えて調査した結果，開発データの 70% を除去しても性能が低下しないことが分かった（図 4）。

［今後の研究の方向，課題］

音声情報案内システムの応答性能改善のために，単語ベクトルを特徴量とした機械学習によるトピック分類と既存用例との類似度に基づく不要データ検出の有効性を示した。これらはいずれも未整理データを活用したものであり，トピック分類のみでなく質問用例拡張も半教師あり学習による自律的改良が期待できる。データ爆発時代～ビッグデータへの関心が高まっている現在，大規模な未ラベルコーパスをいかに活用するかが問われる。今後は自律的改良機能を実環境システムに実装することを目指したい。

［成果の発表，論文等］

［1］ R. Torres, et al. : Comparison of Method for Topic Classification of Spoken Inquiries, IPSJ Journal, vol. 54, no. 2, pp. 157–167, 2013.

［2］ 真嶋温佳他：音声情報案内システムにおける Bag-of-Words を用いた無効入力の棄却，情報処理学会論文誌，vol. 54, pp. 443–451, no. 2, 2013.

［3］ H. Majima, et al. : Spoken Inquiry Discrimination using Bag-of-Words for Speech-Oriented Guidance System, INTERSPEECH, 2012.

［4］ R. Torres, et al. : Topic Classification of Spoken Inquiries using Transductive Support Vector Machine, IWSDS, pp. 231–236, 2012.

［5］ H. Majima, et al. : Evaluation of Invalid Input Discrimination using BOW for Speech-oriented Guidance System, IWSDS, pp. 339–347, 2012.

［6］ 真嶋温佳他：音声情報案内システムにおける Bag-of-Words を特徴量とした無効入力の棄却，情報処理学会研究報告，Vol. 2012-SLP-92, No. 7, pp. 1–6, 2012.

［7］ 真嶋温佳他：音声情報システムにおける最大エントロピー法を用いた無効入力棄却の評価，音響学会講演論文集，3-1-8, pp. 113–116, 2012.

［8］ トーレス ラファエル：Semi-Supervised Learning Algorithms for Topic Classification using Maximum Entropy and Transductive SVM, 音響学会講演論文集，3-P-34, pp. 209–302, 2012.

［9］ 真嶋温佳：音声情報案内システムにおける Bag-of-Words を用いた無効入力棄却モデルの可搬性の評価，音響学会講演論文集，3-9-5, pp. 99–102, 2013.