3U-2

複数の言語モデルと言語理解モデルによる音声理解手法のラピッドプロトタイピングへの適用

勝丸 真樹 † 駒谷 和範 † 中野 幹生 ‡ 船越 孝太郎 ‡
辻野 広司 ‡ 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 ‡(株) 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

音声対話システムが一般に広く利用されるには，システム開発初期段階で高精度なシステムを構築する技術，つまり，**ラピッドプロトタイピング (RP)** 技術が必要である．本稿では音声理解部の RP を扱う．この RP により，少ない労力で開発当初からユーザの発話を高精度に理解できる．また，作成したプロトタイプを用いて発話を収集し，より質の高い学習データを得ることで，システムの性能の急速な向上が期待できる．

本稿では，RP 技術の一つとして，システム開発初期段階の学習データが少ない状況でも高精度な音声理解を実現する手法を報告する．音声理解部では，言語モデルと言語理解モデルを用いる．学習データが少ない場合，それらのモデルの学習が十分でないため，単一の音声理解方式による性能は低い傾向にある．そこで，図 1 のように，複数の言語モデルと言語理解モデルを組み合わせることで，その理解結果の中に適切な結果が含まれる可能性を高める．この際に以下の二つの課題がある．

1. **複数の音声理解結果から適切な結果の選択**
   複数の音声理解方式を用いる場合，得られる複数の結果から最終的な結果を求める必要がある．本稿では理解結果の選択を行うモジュールを選択部と呼ぶ．
2. **少量の学習データの適切な分割**
   選択部の学習時には音声理解結果を用いるが，同じデータを用いて音声理解方式と選択部を学習した場合，学習データに対する音声理解結果は多数が正解となるため，適切な選択部が構築されない．

前者の課題に対し，得られた複数の音声理解結果に対して，ロジスティック回帰を用いて信頼度を付与し，その信頼度に基づき適切な理解結果を選択する [1] (2 章)．本稿では主に後者の課題に対処する．具体的には，学習データ増加時のロジスティック回帰式の係数に着目し，係数の変化が小さくなった時点の発話数を，選択部に最低限必要な学習データ量とする．残りの学習データは音声理解方式に割り当てる (3 章)．これにより，少ない学習データでも音声理解方式とロジスティック回帰を適切に学習し，高精度な音声理解を行う．

2. 発話単位信頼度に基づく理解結果の選択

複数出力される音声理解結果から適切な結果を選択する手法について述べる．各発話に対して，N 個の言語モデルと M 個の言語理解モデルの組み合わせによる方式から出力された音声理解結果を i（$i = 1,\ldots, n$）で表す．ただし，$n = \mathrm{N} \times \mathrm{M}$ である．ある発話に対する音声理解結果 i に対し，正解である発話単位信頼度 CM_i を付与する．ここで，音声理解結果が正解とは，当該発話の理解結果が完全に正解，つまり理解結果中に誤ったコンセプトが含まれないことを意味する．次に，最も高い発話単位信頼度が付与された結果を選択し，当該発話に対する最終的な音声理解結果を得る．つまり，選択結果は $\mathrm{argmax}_i CM_i$ となる．発話単位信頼度は，本稿では音声理解時の特徴に基づくロジスティック回帰により算出する．ロジスティック回帰は，音声理解方式 i ごとに以下の式に基づき構築する．

$$CM_i = \frac{1}{1+\exp(-(a_{i1}F_{i1}+\ldots+a_{i7}F_{i7}+b_i))} \quad (1)$$

ここで，学習データを用いて係数 $a_{i1},\ldots,a_{i7}$ と 切片 b_i をフィッティングする．独立変数 $F_{i1}, F_{i2}, ..., F_{i7}$ は表 1 に示した特徴である．特徴量は，平均 0，分散 1 となるように標準化して用いる．

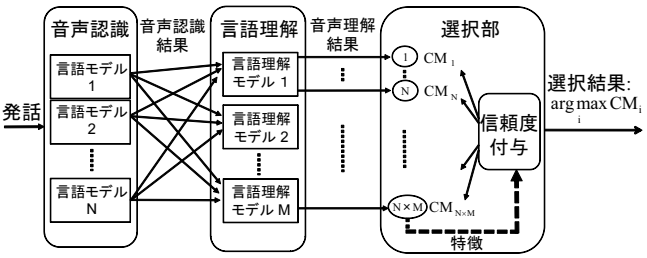

図 1: 本手法における音声理解の流れ

表 1: 音声理解結果 i に関する特徴

F_{i1}	音声理解結果 i に関する音声認識時の音響スコア
F_{i2}	発話検証用言語モデル使用時の音響スコアと F_{i1} の差
F_{i3}	事後確率に基づくコンセプト信頼度の相加平均
F_{i4}	事後確率に基づくコンセプト信頼度の音声理解結果 i 内での最小値
F_{i5}	音声理解結果 i に含まれるコンセプト数
F_{i6}	音声理解結果が得られなかったか
F_{i7}	音声理解結果が肯定・否定発話を表すものか

3. 係数変化量に基づく学習データの分割

学習データを分割し，音声理解方式と選択部とに配分する．方針としては，選択部に優先的に必要最低限の学習データを配分する．これは，学習データが非常に少なく，統計的な音声理解方式の精度が著しく低い場合でも，文法ベースの方式による結果を適切に得る選択部を構築することで，単一手法と比べ同等以上の性能を得るためである．また，残りのできるだけ多くの学習データを音声理解方式に配分し，性能を向上させることで，音声理解部全体の性能向上を図る．

我々は，選択部のロジスティック回帰の係数の変化量に着目することで，選択部に最低限必要な学習データ量を定める．まず，学習データとして使用可能な k_{max} 発話のうち，k 発話，$(k+\delta k)$ 発話を用いて音声理解方式 i に対するロジスティック回帰をそれぞれ構築する．得られた二つのロジスティック回帰から，係数の変化量 $\Delta_i(k)$

Applying Speech Understanding Method Using Multiple Language Models and Language Understanding Models to Rapid Prototyping: *Masaki Katsumaru, Kazunori Komatani (Kyoto Univ.), Mikio Nakano, Kotaro Funakoshi, Hiroshi Tsujino (HRI Japan), Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

を下記の式に基づき算出する．

$$\Delta_i(k) = \sum_j \left| a_{ij}(k+\delta k) - a_{ij}(k) \right| + \left| b_i(k+\delta k) - b_i(k) \right| \quad (2)$$

ここで $a_{ij}(k)$ と $b_i(k)$ はそれぞれ，音声理解方式 i に対するロジスティック回帰を k 発話を用いて構築したときの，特徴 F_{ij} の係数と切片を示している．k を徐々に増加させ，$\Delta_i(k)$ が閾値 θ を下回ったとき，ロジスティック回帰の学習はほぼ収束したと見なす．そのときの k 発話を選択部に，$(k_{max}-k)$ 発話を音声理解方式，つまり言語モデルと言語理解モデルの学習に割り当てる．閾値 θ は，Δ_i が大きく減少した後の値として 10 とした．

本稿では，学習データが不要な文法ベースの音声理解方式に対するロジスティック回帰を用いて係数の変化量を求める．これは，係数の変化量を算出する段階では，音声理解方式に割り当てる学習データ量は未確定だからである．音声理解方式 i ごとに構築するロジスティック回帰の特徴の数は同じである．よって，それぞれのロジスティック回帰で必要となる学習データ量は同等であると仮定し，すべての方式のロジスティック回帰を同じ学習データ量で構築する．

4. 音声理解方式の実装と評価実験

4.1 用いた言語モデルと言語理解モデル

我々はレンタカー予約システム [2] において，2 種類の言語モデルと 4 種類の言語理解モデルを使用できるようにした．本実装時に用いる学習データや，実験時に用いる評価データは，被験者 39 名によるシステムとの対話により収集した 5,240 発話を用いる．5,240 発話のうち 16 名分 2,121 発話を学習データとし，23 名分 3,119 発話を評価データとした．

言語モデルは，文法モデルと N-gram モデルの 2 種類を用いた．文法モデルは，言語理解時に用いる Finite-State Transducer (FST) に対応させて人手で記述した．また，N-gram モデルは，学習データの書き起こしを用いてクラス 3-gram を学習し，作成した．語彙サイズは，文法モデルが 281，N-gram モデルが学習データをすべて用いた場合 420 であり，評価データに対する音声認識精度はそれぞれ 66.3%と 85.0%であった．音声認識器は Julius (ver. 4.1.2) [‡]を用いた．

言語理解モデルは，[1] で用いた FST，Weighted FST (WFST)，Keyphrase-Extractor の 3 種類に，Conditional Random Fields (CRF) [3] を加えた計 4 種類を用いた．CRF による言語理解では，まず，音声認識結果に対して，CRF に基づき意味スロットのみ付与し，次に，その意味スロットに対応する値を音声認識結果を用いて求める．

4.2 音声理解精度の評価

4.2.1 学習データの分割の評価

本手法による学習データの分割を評価するため，本手法と，学習データの分割を行わない場合，音声理解部と選択部に単純に等分に配分する場合とを比較する．学習データとして被験者一名分の 141 発話を用いる．本手法で学習データを分割したとき，音声理解部とロジスティック回帰に割り当てる学習データはそれぞれ，30 発話，111 発話となった．各分割手法ごとのコンセプト理解精度を表 2 に示す．表 2 より，本手法による精度は，分割なしと比較して 3.8 ポイント，単純な分割と比較して 4.4 ポイント高い．これは，本手法に基づく学習データの分割により，適切な選択部が構築できたことを示している．

表 2: 分割方法ごとのコンセプト理解精度の比較 [%]

分割の方法	コンセプト理解精度	Sub	Del	Ins
本手法に基づく分割	77.9	11.9	6.5	3.7
分割なし	74.1	17.4	5.2	3.3
単純な分割	73.5	13.6	9.9	3.0

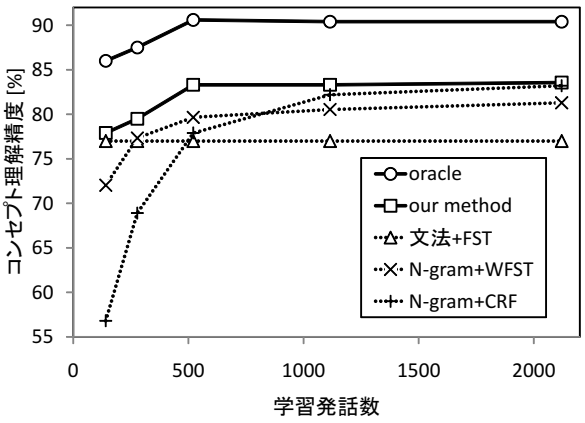

図 2: 学習データ量とコンセプト理解精度の関係

4.2.2 単一の音声理解方式との比較

学習データ量を変化させた時の，本手法と単一の音声理解方式によるコンセプト理解精度を図 2 に示す．ここで，oracle は，人手での最適な理解結果の選択を表す．本手法では，1,000 発話を越える学習データのとき，同じデータを用いて音声理解方式と選択部を学習しても，学習データに対する音声理解結果は正解ばかりに偏らず，選択部も適切に構築できると考え，学習データの分割は行わなかった．比較対象となる単一方式は使用可能な学習データをすべて用いて構築している．単一方式は，2 種類の言語モデルと 4 種類の言語理解モデルの組み合わせから 8 種類あるが，ここでは，データ量を変化させる過程で最も高精度となることがあった 3 つの理解方式の結果のみ示す．表 2 において，oracle による精度はすべての学習データ量において単一の音声理解方式を大きく上回る．これは，複数の音声理解方式を用いることの有効性を示している．また，本手法は，150 から 500 発話程度という比較的少ない学習データでも，単一の音声理解方式より高い精度である．これは，本手法が RP に効果的であることを示している．

5. おわりに

本研究では，複数の言語モデルと言語理解モデルを用いて，少量の学習データでも高精度な音声理解を実現するための手法を述べた．今回，ロジスティック回帰の収束判定に用いた閾値 θ は，$\Delta_i(k)$ の遷移から人手で設定したが，係数の数から自動的に設定できると考えられる．今後，閾値の設定を含めた学習データの自動分割を検討する．

謝辞 本研究の一部は，科研費，GCOE の支援を受けた．

参考文献

[1] M. Katsumaru *et al.* Improving Speech Understanding Accuracy with Limited Training Data Using Multiple Language Models and Multiple Understanding Models. In *Proc. Interspeech*, pp. 2735–2738, 2009.

[2] M. Nakano *et al.* Analysis of User Reactions to Turn-Taking Failures in Spoken Dialogue Systems. In *Proc. SIGdial*, pp. 120–123, 2007.

[3] J. Lafferty *et al.* Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. ICML*, pp. 282–289, 2001.

[‡]http://julius.sourceforge.jp/