3E3-5

# 未登録語の発話を目的としたモデル選択による語彙獲得手法の提案

## Model-Selection Bases Method for Word Acquisition through Natural Unlimited Utterances

田口 亮[*1] Ryo Taguchi
岩橋 直人[*2*3] Naoto Iwahashi
新田 恒雄[*1] Tsuneo Nitta
船越孝太郎[*4] Kotaro Funakoshi
中野幹生[*4] Mikio Nakano

[*1] 豊橋技術科学大学 Toyohashi University of Technology
[*2] (独)情報通信研究機構 National Institute of Information and Communications Technology
[*3](株)国際電気通信基礎技術研究所 Advanced Telecommunications Research Institute International
[*4](株)ホンダ・リサーチ・インスティチュート・ジャパン Honda Research Institute Japan Co., Ltd.

This paper proposes a lexical acquisition method to enable a robot to learn words from spontaneous speech of a human. The robot is given a statistical model of Japanese phonemes, but isn't given knowledge of words. Therefore, at first, the robot recognizes only phoneme sequences. But it can gradually acquire words by statistically analyzing the phoneme sequences and can learn the meanings of the words too.

## 1. はじめに

家庭や街で人の生活を助けるロボットに対する社会的な期待が高まっている. ロボットが実世界で人とコミュニケーションするためには, 多くの言語知識が必要になる. 現在, 多くの対話システムでは, システム開発者が言語知識を用意しておく必要がある. しかし, その全てを網羅することは不可能であり, ロボットが自ら知識を学習していくことが望まれている. ロボットによる言語獲得の先行研究では, 単語単位で発話を区切って教えることで, その意味と音素系列の学習を可能にした[iwahashi07]. しかし, ユーザの使いやすさを考慮すると, 学習や指示に用いられる発話は出来るだけ制約のないものが望ましい. 例えば,「ここは今木さんの研究室だよ.」でも「こっちが今木さんの部屋です.」でも同じように学習できれば, ユーザは特定の形式にとらわれることなく普段の生活のまま自由に発話することができる.

本稿では, ロボットによる案内タスクを対象に, 多様な言い回しでの教示から単語の意味および正しい音素列を学習できる語彙獲得手法を提案する. 本手法は,初期知識として日本語の音素列の統計モデルだけを与え, 単語のモデルは与えない. そのため, 初めは音声を音素列でしか認識できないが, 学習を進めることで, 得られた音素列の統計量を元に, 単語(音素列単位)を切り出すことができるようになる. 自由発話を対象とした単語学習の先行研究として[Gorin99,Roy00]がある. これらの研究では, トピックや画像と音声の対応関係を利用して単語の切り出しと, 意味の学習を行っている. ただし, 両研究共に音声からトピックやおもちゃを認識することが目的であり, 単語の音素列を正しく学習し発話することは目的としていない.

続く2節では具体的な問題設定について述べる. 3節で提案手法について説明し, 4節で評価実験の結果について述べる. 5節は本稿のまとめとなっている.

連絡先:田口亮
〒466-8555 名古屋市昭和区御器所町
名古屋工業大学 19 号館 2 階 226 室
E-Mail: taguchi.ryo@nitech.ac.jp
TEL: 052-735-5552

## 2. 問題設定

具体的な案内タスクは次のようなものである.

まず, 学習フェイズでは, 人がロボットを所望の場所に連れて行き,「ここはスマートルームです.」や「この場所の名前は辻野さんのブース.」などと言って, その場所について教示する. ロボットはそうした発話から, 場所と単語の対応関係を学習する. こうした学習を何度か繰り返した後, 案内フェイズに移る. 案内フェイズでは,「スマートルームへ行ってください.」というような指示に対して, 正しい場所へ人を案内し,「ここがスマートルームです.」と発話することが目的である.

発話はヘッドセットマイクで取得した音声であることを想定する. 場所を表す単語(**キーワード**)や, 場所を伝える際の**言い回し**には制限を与えない. 従ってこのタスクでは, ①異なる言い回しの発話が入力された際に, 正しい場所を出力できること, ②キーワードの正しい音素列が獲得できること, の二つが必要となる. なお, 以下では, 議論に一般性を持たせるために場所を**トピック**と呼ぶ.

## 3. 提案手法

提案手法の概要を図 1,2 に示す.

学習フェイズは図 1 のように二つのステップに分けられる. STEP1 は, 音声を音素列として認識しその統計量から単語辞書を生成する. STEP2 は, STEP1 で生成した単語辞書を用いて音声を認識し直し, 単語と場所の対応関係(意味モデル)や, 単語間の繋がり(文法モデル)の学習を行う. 従って本手法では, 音素認識器が用いる音素列の統計モデルは必要となるが, 対話に使用する単語は予め定義しておく必要がない. 単語辞書の生成は 3.1 項, 意味モデルと文法モデルの学習は 3.2 項で詳細を述べる.

図 2 に示す案内フェイズでは, 音声が入力されると, 学習した知識(単語辞書や意味モデル, 文法モデル)に基づいて, 対応するトピックを認識し出力する. トピックの認識の詳細は, 3.3 項で述べる.

学習フェイズ

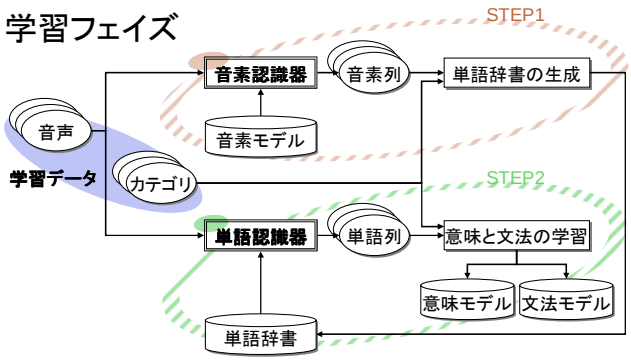


図 1: 学習フェイズの流れ

案内フェイズ

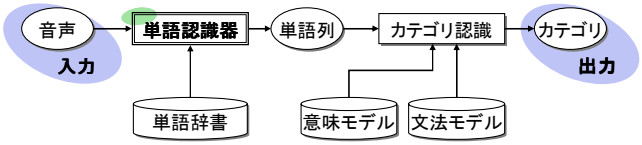


図 2: 案内フェイズの流れ

### 3.1 単語辞書の生成

学習中に得られた全ての音素列を分析し，その統計量に基づいて単語の仮説を生成する．統計量は，音素列 $p_1,\ldots,p_n$ の前に来る音素 $\mathbf{p_0}$ のエントロピー $H(\mathbf{p_0} \mid p_1,\ldots,p_n)$，後ろに来る音素 $\mathbf{p_{n+1}}$ のエントロピー $H(\mathbf{p_{n+1}} \mid p_1,\ldots,p_n)$，トピック z が現れる確率 $p(\mathbf{Z} = z \mid p_1,\ldots,p_n)$，そのエントロピー $H(\mathbf{Z} \mid p_1,\ldots,p_n)$ を使用する．$\mathbf{p_0}$, $\mathbf{p_{n+1}}$, $\mathbf{Z}$ は確率変数であり，$\mathbf{p_0}$, $\mathbf{p_{n+1}}$ は任意の音素，$\mathbf{Z}$ は任意のトピックを値として取る．$H(\mathbf{p_0} \mid p_1,\ldots,p_n)$ と $H(\mathbf{p_{n+1}} \mid p_1,\ldots,p_n)$ を用いる事で，単語の切れ目を統計的に判定できる．また，$p(\mathbf{Z} = z \mid p_1,\ldots,p_n)$ と $H(\mathbf{Z} \mid p_1,\ldots,p_n)$ を用いることで，音素列が意味あるまとまりかどうかを判定することができる．具体的な生成手順を次に示す．

(1) 学習データセットを与え音素列を認識する．得られた音素列の部分音素列の頻度から，音素列の $H(\mathbf{p_0} \mid p_1,\ldots,p_n)$，$H(\mathbf{p_{n+1}} \mid p_1,\ldots,p_n)$，$p(\mathbf{Z} = z \mid p_1,\ldots,p_n)$，$H(\mathbf{Z} \mid p_1,\ldots,p_n)$ を計算する．

(2) 言い回し候補の判定
(2-1) $\{ H(\mathbf{p_0} \mid p_1,\ldots,p_n) + H(\mathbf{p_{n+1}} \mid p_1,\ldots,p_n) \}>0$ かつ $H(\mathbf{Z} \mid p_1,\ldots,p_n)>0$ となる音素列を言い回し集合 E に格納する．
(2-2) 言い回し集合 **E** 内で他の音素列に含まれる音素列を削除する．

(3) キーワード候補の判定
(3-1) $\{ H(\mathbf{p_0} \mid p_1,\ldots,p_n) + H(\mathbf{p_{n+1}} \mid p_1,\ldots,p_n) \}>0$ かつ $\max p(\mathbf{Z} = z \mid p_1,\ldots,p_n)$ となる音素列をクラス毎のキーワード集合 $K_z$ に格納する．
(3-2) クラス毎のキーワード集合 $\mathbf{K}_z$ 内で他の音素列に含まれる音素列を削除する．

(4) 言い回し集合 **E**，およびクラス毎のキーワード集合 $\mathbf{K}_z$ に含まれる音素列を単語辞書に登録する．

### 3.2 意味と文法の学習

3.1 項で生成した単語辞書を用いて学習データの音声を認識する．単語認識器は結果として N 個の単語列候補（N-Best）を出力する．文法の学習では，N-Best の中で最も尤度の高い単語列(1Best)から単語 bigram を計算する．意味の学習では，N-Best に含まれる単語 w とトピック z の共起関係から条件付確率 p(z|w)を計算する．1Best ではなく N-Best を利用することで，認識の揺れを吸収し，頑健なトピック認識を実現する．なお 4 節の実験では N=10 とした．

### 3.3 トピック認識

入力された音声 a とトピック z の共起確率 p(a, z)を次のように定式化する．

$$\begin{aligned} p(a,z) &= \sum_s p(a \mid s)\,p(s)\,p(z \mid s) \\ &= \sum_s p(a \mid s)\; p(s) \sum_{w \in s} p(z \mid w)\,p(w \mid s) \\ &\approx \max_s p(a \mid s)\; p(s) \sum_{w \in s} p(z \mid w)\,p(w \mid s) \end{aligned} \quad \cdots(1)$$

ここで s は単語列を表す．また，条件付確率 p(a|s)は単語認識の尤度，p(s)は文法（単語 bigram の積），p(z|s)は発話の意味を表している．p(z|s)は文書トピック認識の手法の一つである Single random Variable with Multiple Value 法[Iwayama94]と同様のものであり，発話を文書とみなし，発話の意味を発話に含まれる単語が持つ意味の重み和として表現している．

音声 a に対して出力するトピック $\tilde{z}$ は次式で求める．

$$\tilde{z} = \arg\max_z p(a,z) \quad \cdots(2)$$

### 3.4 モデル選択による単語のマージ

3.1 項で生成された単語を MDL（Minimum Description Length）基準によりマージする．これにより，出現頻度の少ない単語がもたらす悪影響の軽減や，分割されすぎた単語の再結合ができる．

## 4. 実験

提案手法の有効性を検証するため実験を行う．本稿では，3.4 項のモデル選択を用いない場合の手法（3.1～3.3 項まで）の評価を行う．モデル選択を用いた実験は今後の課題とする．

男性話者 3 名の音声を用いた．トピック数は 10，言い回しのパターン数は 6 とし，話者一名につき 60 種類の音声を収集した．トピック番号と対応するキーワードを表 1 に，言い回しのパターンを表 2 に示す．話者毎にクロスバリデーション法（59 個のデータで学習を行い，残り1個のデータでトピック認識を行うことを 60 通り行う）でトピック正解率を求めた．その結果，話者 1 が **98%**，話者 2 が **87%**，話者 3 が **93%**の正解率が得られた．トピック認識に使用された単語の数，すなわち 3.1 項の手法で生成された辞書の単語数は，平均 119 語であった．その中には，「学生部屋マ前（gakuseebeyamamae）」や「ア学生部屋の前（agakuseebeyanomae）」，「この場所は (konobashowa)」，「このバシェオア(konobasheoa)」など類似した単語も多く登録されて

いる. 3.2 項の手法ではその中でも, 学習データで認識された(N-Best に含まれる)単語だけ意味を学習する. そのため, 各単語の学習回数にはばらつきがあり, 特に一度しか学習されなかった単語によって, トピック認識を誤ってしまうという例も見られた. この問題については, 3.4 項の手法を用いることで解決できると考えている. なお, 発話全体の音素正解精度は話者 1 が 85%, 話者 2 が 78%, 話者 3 が 80% であった. 本手法は, 音素認識がベースとなっているため, その正解精度がボトルネックとなるが, 80%程度の音素正解精度でも, 高いトピック正解率が得られることがわかった.

各話者のデータから獲得したキーワードの例として, 生成した辞書に含まれる単語のうち最もキーワードに近い単語を表 3〜5 に示す. キーワードの音素正解精度を求めた結果, 話者 1 が **78%**, 話者 2 が **71%**, 話者 3 が **81%**であった. その誤りのうち約 7 割が脱落誤りであった. これは, キーワードが途中で切れていることを意味する. 例えば「スマートルームの入り口」を「スマトルムの(sumatorumuno)」で分割してしまう例があった. ただし, この単語が実際の音声認識に使われた結果を見ると, 「この場所は, スマトルムの, い, り, うち(konobashowa, sumatorumuno, i, ri, uchi)」というように, 他の単語を用いて分割された残りの部分を補っていた. この傾向を利用すればキーワードの再結合が可能であると考えている.

## 5. まとめ

ロボットによる案内タスクを対象に, 多様な言い回しでの教示から単語と場所トピックの関係や正しい音素列を学習できる語彙獲得手法を提案した. 実験の結果, トピック正解率が平均 93%, キーワードの正解精度が平均 77%であった. トピック正解率に関しては, 十分に学習されない単語が悪影響を与えており, キーワードの正解精度に関しては 7 割が脱落誤りとなっていることがわかった. この問題を解決するため, 現在, 統計的モデル選択に基づき, 辞書に登録されている単語をマージする手法の検討を行っている.

## 謝辞

発話データの収集にご協力いただいた皆様に感謝します.

## 参考文献

[Iwahashi07] Iwahashi, N : “Robots That Learn Language: A Developmental Approach to Situated Human-Robot Conversations,” In Sankar, N. ed. Human-Robot Interaction, pp.95-118, I-Tech Education and Publishing (2007).

[Gorin99] Gorin, A. L., Petrovska-Delacretaz D., Wright, J. H. and Riccardi, G. : “Learning spoken language without transcription,” Proc. ASRU Workshop, Colorado, 1999.

[Roy00] Roy, D.: “Integration of speech and vision using mutual information.” In Proc. of ICASSP, Istanbul, Turkey (2000).

[Iwayama94] Iwayama, M. and Tokunaga, T.: “A probabilistic model for text categorization: Based on a single random variable with multiple values,” In Proc. of the 4rth Applied Natural Language Processing Conference (ANLP), pp. 162-167 (1994).

表 1:トピック番号と対応するキーワード

| トピック z | キーワード X | トピック z | キーワード X |
|---|---|---|---|
| 1 | 会議室の前 | 6 | 竹内さんのブースの南 |
| 2 | 辻野さんのブース | 7 | 工作室 |
| 3 | フロアの真ん中 | 8 | アシモの部屋 |
| 4 | 学生部屋の前 | 9 | スマートルーム |
| 5 | お茶飲み場 | 10 | スマートルームの入り口 |

表 2:言い回しのパターン

| | |
|---|---|
| X の所に行って | 今から X へ行って |
| X へお願い | この場所は X |
| ここは X です | ここの名前は X |

表 3:話者 1 のデータで獲得したキーワードの例
(辞書中の単語のうち最もキーワードに近い単語を表示)

| | |
|---|---|
| 会議室の前 | kaigihitsunomae |
| 辻野さんのブース | tsuzinosanobu |
| フロアの真ん中 | hurowanomang |
| 学生部屋の前 | agakuseebeyanomae |
| お茶飲み場 | anobiba |
| 竹内さんのブースの南 | isanobusunominami |
| 工作室 | wakoosakushitsu |
| アシモの部屋 | ashimomoheya |
| スマートルーム | sumaatorumu |
| スマートルームの入り口 | atorumunoiribichi |

※「ん」は「ng」として表記する.

表 4: 話者 2 のデータで獲得したキーワードの例

| | |
|---|---|
| 会議室の前 | kaigishitsu |
| 辻野さんのブース | atsuzinosanggabuusu |
| フロアの真ん中 | wadomangnaka |
| 学生部屋の前 | gakiseebeea |
| お茶飲み場 | ochangnobiba |
| 竹内さんのブースの南 | uchisangnobuusunomigaami |
| 工作室 | koosateshi |
| アシモの部屋 | ashimonoheeya |
| スマートルーム | imaatoruu |
| スマートルームの入り口 | owasugaatoruumenoirigu |

表 5: 話者 3 のデータで獲得したキーワードの例

| | |
|---|---|
| 会議室の前 | aigishitsunomae |
| 辻野さんのブース | tsuzinosangnobu |
| フロアの真ん中 | hurowanamangnaka |
| 学生部屋の前 | gakuseebeyanoma |
| お茶飲み場 | ochanomibae |
| 竹内さんのブースの南 | nobuusunaminami |
| 工作室 | koosakushi |
| アシモの部屋 | ashimonoheyae |
| スマートルーム | sumatorumu |
| スマートルームの入り口 | sumatorumunri |