# 最大エントロピー法による発話理解のための効率的モデル構築

谷垣宏一[†] 渡邉圭輔[†] 石川泰[†]

本論文では，最大エントロピー法による意味理解モデルの構築を効率的に行う学習方式を提案する．最大エントロピー法では，モデルに取り込む学習データの特性を規定するために，複数の素性を用いる．これらの素性群は，通常設計者が与えた候補集合から選択する形式で決定される．従来の素性選択方式は，尤度基準による最適素性の探索と，素性のモデルへの追加とを繰り返すというものであった[6]．しかし，尤度計算に要する計算量が大きいことから，意味的識別問題でしばしば生成される冗長で大規模な候補集合に適用した場合，モデル構築に多大な時間を要するという問題があった．本方式は，従来の素性選択方式に対し，素性の生起確率に関するモデルの推定誤差を検定する処理を導入することにより，尤度改善効果が小さい素性を効率的に検出し，候補から棄却する点を特徴とする．これにより尤度計算処理が削減され，高速にモデルを構築することが可能となった．本方式を，音声対話システムにおけるユーザの発話意図を識別するモデルの学習に適用し，評価実験を行った．実験の結果，従来方式の識別精度を劣化させることなく，学習時間を 18%に削減することができた．また，頻度や相互情報量に基づく予備選択を素性候補に適用する方法と比較して，同一学習時間で安定して良い識別性能が得られ，学習効率化における本方式の有効性を確認することができた．

# Efficient Training Algorithm for Maximum Entropy Semantic Modeling

Koichi Tanigaki,[†] Keisuke Watanabe[†] and Yasushi Ishikawa[†]

This paper proposes an efficient feature selection algorithm for maximum entropy modeling, to cope with the problem of heavy computation required to construct models for semantic processing. The proposed method introduces hypothesis tests into the previous algorithm of successive feature selection [6] that is completely resort to liklihood gain using Newton method. The tests for estimation errors of probability about feature occurrences efficiently detects those features that take less effect at liklihood gain. By rejecting the ineffectural features from candidate set, the computational cost to calculate liklihood gain is highly reduced, so that the algorithm can efficiently generate maximum entropy models. The proposed method is applied to generate the model that discriminates the intention of speaker, given the word sequence of utterance to our spoken dialog system. Experimental results show that our method cuts the training time down to 18% with no decline at discrimination performances, compared to the previous algorithm. It is also showed that our method performs stably better at discrimination error rates in the equal training time, than those naive methods of pre-selection based on frequency/mutual information.

## 1. はじめに

近年の音声認識技術の進展にともない，情報検索をはじめとする各種タスクを音声認識を介して実現する，音声対話システムへの期待が急速に高まっている．音声対話システムにおける音声言語処理は，一般にタスクへの依存性が強いことから，従来人手で記述した文法を利用する方式が広く用いられてきたが，近年は統計モデルを意味理解に適用した音声言語処理方式もさかんに研究されている（たとえば文献 1)~5)）．意味理解に統計モデルを用いることで，次のような効果が期待できる．

- 決定的なルールとしては利用できない表層—意味の選好的関係を，確率的な規則として利用することができる．
- 設計者の内省では網羅しきれない多様な言い回しを被覆する規則を，コーパスから自動抽出することができる．

我々は，音声対話システムのための自由発話理解実現を目指し，統計的な意味理解方式を検討している．また，学習データのスパースネスに対処するため，最大エントロピー法[6] の適用を検討している．スパースネスの問題に関しては，音声認識に用いられる N グ

† 三菱電機株式会社情報技術総合研究所
Information Technology R&D Center, MITSUBISHI Electric Corporation

ラム言語モデルなどにおいても，しばしば話題として取り上げられてきた．これは，対象とするタスクで十分な規模のデータ（言語表現のバリエーション）を収集することが困難なためである．これに加えて，意味理解モデルの学習においては，通常人手を介して意味的なアノテーションを行う必要があるため，学習データの収集はさらに高コストとなる．最大エントロピー法は，1) 未学習の事象に最も一様な確率分布を仮定すること，2) 種々の抽象度の規則を混用できるため先見的な知識を反映させやすいこと，などの特徴を持つため，限られた学習データから頑健なモデルを構築する際に有効な手法といえる．

ところで，最大エントロピー法では，モデルに取りこむ学習データの特性を規定するために，「素性関数」（あるいは単に「素性」とも呼ばれる）を複数用いる．通常，識別に有効な素性の組合せを人手で決定することは困難なため，モデルに用いる素性を自動的に決定する方法が必要となる．Berger らは，あらかじめ生成した素性の候補集合をもとに，尤度の増分を最大化する素性の探索と，モデルへの追加とを繰り返す，逐次的な素性選択アルゴリズムを提案した[6]．しかし，尤度の増分を計算するためには，ニュートン法などを適用して素性に対する重みパラメータを設定する必要がある．このパラメータ計算は繰返しごとにすべての素性候補に対して行われるため，大規模な候補集合に適用する場合，モデル構築に要する計算量は深刻な問題となる．

特に意味的な識別問題では，扱う候補集合が，冗長で大きなものになりやすい．これは，学習データのスパースネスに起因して，一元的な統計量に対応する素性だけでは十分な識別性能を得ることが難しく，多様な観点から生成した種々の抽象度を持つ素性を候補に用いるのが有効なためである．したがって，効率的な素性選択方式の開発は，最大エントロピー法による意味理解を実現するうえで重要な課題である．

これまでに，素性選択を効率化する方式として，以下が提案されている．Printz は，ニュートン法の代わりにサンプリングした点の補間曲線を用いることで，パラメータの決定と尤度の増分計算を高速化する手法を提案している[7]．この方法で削減される演算量は，おおむねサンプリングする点の個数とニュートン法の反復計算の回数との差分に比例するが，適切な近似精度を得るためにはサンプリングする点を減らすことはできないため，大幅な高速化を実現することは難しい．Mikheev は，素性ラティスと呼ぶネットワークを用いた学習方式を提案している[8]．ネットワークのノードは網羅的な素性の組合せに対応しており，ノード上で学習データの頻度分配を行うことにより，数値解析的な手法に頼らず，高速にモデルを構築することを可能とした．しかし，この方法は，組合せ爆発の問題から，多数の素性候補を扱う問題に適用することができない．一方，次数 N の異なる単語 N グラムを素性として，N グラム言語モデルを最大エントロピー法で生成する場合においては，素性間の共起関係が明らかなことを利用して，素性間で共通の計算を省く方法や，素性が成立しないことが自明な単語系列に対し計算を省略する方法が提案されているが[9),10]，こうした手法は必ずしも一般の問題に適用できるわけではない．

本論文では，冗長性を有する素性の候補集合から，高速にモデルを構築することを可能とする素性選択方式を提案する．本方式は，従来の尤度基準による逐次的な素性選択方式[6] に対し，素性の生起確率に関するモデルの推定誤差を検定する処理を導入することで，尤度改善効果の小さい素性を効率的に検出し，候補から棄却する点を特徴とする．これにより，計算量の大きい尤度計算処理を削減し，短時間でモデルを構築することが可能となった．本方式の有効性を，発話意図を識別するモデルの学習・理解実験の結果から示す．

## 2. 最大エントロピーモデルによる発話意図の識別

発話に対する意味理解処理は，しばしば，与えられた発話 $u$ を，ある意味的な観点から規定した $M$ 種類のカテゴリ $Y = \{y_1, \ldots, y_M\}$ に分類する問題と見なすことができる．発話 $u$ の最適なカテゴリ $y^*$ は次式で求められる．

$$y^* = \arg\max_{y \in Y} p(y|u) \tag{1}$$

本論文では，発話 $u$ が単語の系列として与えられるものとする．識別カテゴリとしては，音声対話システムに対するユーザ発話の要求や応答タイプの分類を用い，これらを発話の「意図」と呼ぶ．意図としては，システムの対象タスク（横浜市内のホテル検索・予約）に合わせて，「予約要求」「比較要求」「値要求」「値応答」「承認」「否定・拒絶」などの 19 種類を規定した．式 (1) の条件付き確率 $p(y|u)$ は，最大エントロピー法でモデル化する．

### 2.1 最大エントロピー法における素性の役割

最大エントロピー法では，モデルに取り込む学習データの特性を，「素性」と呼ばれる 2 値関数 $f_i$ の集合 $F$ により規定する．それぞれの素性は，発話 $u$ と意図 $y$ の組 $(u, y)$ に対して条件判定を行い，条件に

合致した場合には 1（発火）を，非合致の場合には 0 を返す．

$$F = \{f_i : (u, y) \mapsto \{0, 1\}, i = 1, 2, \ldots\} \quad (2)$$

いま，素性の集合 $F$ が何らかの方法により決定されているとする．このとき，最大エントロピーモデルは，$F$ により規定される制約条件 $C_F$ を充足する，最も一様な（エントロピーが最大の）確率分布 $p^*(y|u)$ として，次式で与えられる．

$$p^* = \arg\max_{p \in C_F} H(p) \quad (3)$$

式 (3) でエントロピー最大化の条件となる $C_F$ は，次式 (4) の制約等式群で表される．式中，$p()$ はモデルにより推定される確率を表し，$\tilde{p}()$ は学習データにおける観測確率を表す．式 (4)〜(6) に示すように，各素性 $f_i$ は，注目する事象（素性 $f_i$ の発火条件に合致する発話 $u$ と意図 $y$ の組）の生起確率に関し，モデルで推定される期待値 $p(f_i)$ が，学習データでの観測値 $\tilde{p}(f_i)$ に等しくなるよう，制約を加えるために用いられる．

$$C_F \equiv \{p | \forall f_i \in F,\ p(f_i) = \tilde{p}(f_i)\} \quad (4)$$

ここで，

$$p(f_i) = \sum_u \tilde{p}(u) \sum_y p(y|u) f_i(u, y) \quad (5)$$

$$\tilde{p}(f_i) = \sum_{u,y} \tilde{p}(u, y) f_i(u, y) \quad (6)$$

なお，式 (3) を解くことで得られる最大エントロピーモデルは，各素性 $f_i(u, y)$ とその重み $\lambda_i$ をパラメータとする次の形式で表される[6]．

$$p(y|u) = \frac{1}{Z(u)} \exp\left(\sum_i \lambda_i f_i(u, y)\right) \quad (7)$$

式中の $Z(u)$ は，$\sum_y p(y|u) = 1$ とするための正規化項である．

$$Z(u) = \sum_y \exp\left(\sum_i \lambda_i f_i(u, y)\right) \quad (8)$$

### 2.2 発話意図の識別に用いる素性

本論文では，単語系列として与えられる発話から意図を識別するために，2 種類の形式の素性を併用する．

**(A)** 意図が $y$ で，かつ，発話中にある部分単語系列 $w_1 \ldots w_n$（$n$ は任意）が生起するときに発火する素性．

**(B)** 意図が $y$ で，かつ，発話中に単語 $w_1$ と $w_2$ がこの順序で，距離 $d$ ($\in \{1, 2, 3\}$) 単語以内に共起するときに発火する素性．ここで，距離 $d$ は単語 $w_1$ と $w_2$ に挟まれる単語の数を表す．

**表 1** 素性の例
Table 1 Example of features.

| | 発火条件 | | | | 観測頻度 / 関連発話数 *1 |
|---|---|---|---|---|---|
| | 意図 | 発話 | | | |
| (A) | | 部分単語系列 | | | |
| A1 | 予約要求 | 予約 | | | 251/270 |
| A2 | 予約要求 | 取り | | | 22/25 |
| A3 | 予約要求 | 予約/を/取り | | | 8/9 |
| A4 | 値要求 | 教え/て | | | 126/133 |
| (B) | | 先行単語 | 後続単語 | 距離 | |
| B1 | 予約要求 | 予約 | し | 3 | 236/246 |
| B2 | 比較要求 | どれ | END | 2 | 23/24 |

*1 観測頻度は，素性の発火条件に合致する学習サンプル（発話と正しい意図の組）の数を表す．関連発話数は，素性の発話に対する条件に合致する発話数を表す．

**表 1** に A，B それぞれの素性の例を示す．たとえば A1 の素性は，意図が「予約要求」で，かつ，発話中に単語「予約」が生起するときに発火する素性を表す．この素性 $f_{A1}$ をモデルに用いると，学習データ $S$（発話数を $n$ とする）中で単語「予約」を含む 270 発話から，モデルが意図「予約要求」を推定する確率的な回数 $n \cdot p(f_{A1})$ は，学習データでの観測頻度 $n \cdot \tilde{p}(f_{A1}) = 251$ 発話に等しくなるよう，式 (4)〜(6) で制約される．

これらの素性は次の方法で生成し，後述の評価実験に用いた．まず学習データから，上記 A および B の形式で観測頻度が 1 以上の素性を生成する．次に，学習データ中でまったく同じ識別情報を与える素性群☆をマージする．マージでは，暫定的に以下 i〜iv の優先順位により代表として残す素性を 1 つ決定し，その他の素性を棄てる．i) 形式 A と B の素性では，A の素性を優先する．ii) 形式 A の素性からは，単語連鎖数 $n$ が小さい素性を優先する．iii) 形式 B の素性からは，距離 $d$ が小さい素性を優先する．iv) その他はリスト中の若い素性を優先する．こうして生成した素性の候補集合に対し，次章で述べる素性選択を適用することで，モデルに用いる素性を決定する．

## 3. 推定誤差の検定に基づく効率的素性選択

以下では，本論文で提案する効率的な素性選択方式について説明する．本方式のベースとなるアルゴリズムは，文献 6) の素性選択方式である．この方式では，素性をまったく持たないモデルを初期モデルとし，あ

☆ ある 2 つの素性において，意図の発火条件に合致する学習サンプル（発話と正しい意図の組）の集合が等しく，かつ，発話の発火条件に合致する学習サンプルの集合も等しいとき，「学習データ中でまったく同じ識別情報を与える」とする．これらの素性は，尤度を基準とする素性選択処理で順位付けができない．

らかじめ規定した素性の候補集合に次の Step 1，2 を繰り返し適用してモデルを構築する．

**Step 1** モデルに対し，候補集合中の各素性を仮に追加して，尤度の増分（近似値）をそれぞれ計算する．追加した素性に対する重みは，ニュートン法で求める☆（近似値）．その結果，尤度の増分を最大化する素性を 1 つ選択する．

**Step 2** 選択した素性をモデルに追加し，モデル中の全素性の重みを改良反復スケーリング法（Improved Iterative Scaling: IIS）により最適値に設定する．（Step1 に戻る）

この方法で学習を行った場合に問題となるのが，Step1 の計算量である．候補集合中のすべての素性に対しニュートン法を適用して尤度の増分を求めるため，繰返しあたり多大な時間を要する．そこで，候補集合の中から，モデルにとって無効な素性を検出し候補から棄却することにより，Step1 の尤度計算対象を削減する方法を考える．

### 3.1 無効な素性

素性の候補を機械的に生成した場合，モデルにとってはほぼ同等の制約として機能する複数の素性が候補中に含まれることがある．たとえば，候補集合として表 1 の素性を用いたとする．A1 の素性は，単語「予約」を含む発話から，モデルが発話意図「予約要求」を推定する確率を制約する．ここで，単語「予約」は，発話中で「… 予約したい …」「… 予約して …」「… 予約お願いし …」などの表現に多く用いられることから，A1 と B1 の素性は，ほぼ共通の発話に対し，同様に意図「予約要求」の推定確率を制約すると考えられる．このように互いに類似した制約を与える素性群では，その中の素性が 1 つ選択されると，その後，他の素性を選択しても推定される確率分布（モデル）はほとんど変化しない．

こうした素性間の類似関係は 1 対 1 に限らず，多対 1 でも存在する．たとえば，表 1 で A1，A2，B1 の素性はいずれも意図「予約要求」が推定される確率を制約する．このうちの A1（単語「予約」を含む発話に対する制約）と A2（単語「取り」を含む発話に対する制約）がすでに選択されているとき，その後，A3（単語列「予約/を/取り」を含む発話に対する制約）を選択しても，モデルとして推定される確率分布はほとんど変化しない．

このように，ある素性によって与えられるべき制約が，モデル中の 1 つないし複数の選択済み素性により，同等な制約として実現されているとき，その素性をモデルにとって無効な素性と見なす．

### 3.2 無効な素性の検出基準

モデルにとって無効な素性を，次の基準で検出する．いま，候補となる素性を $f$ とする．素性 $f$ をモデル $p$ に追加すると，式 (4) においてモデルには新たな制約 $p(f) = \tilde{p}(f)$ が課せられる．ここでもし，素性 $f$ を追加する前に，すでにモデル $p$ がこの制約を十分充足しているならば，素性 $f$ はモデル $p$ にとって無効な素性といえる．

そこで，モデルによる推定値 $p(f)$ と観測値 $\tilde{p}(f)$ との間に有意な誤差がないかを検定する．ただし，学習データ $S$ の中には，素性 $f$ の発話条件に合致しないサンプルが含まれている．これらのサンプル集合では，モデルのいかんにかかわらずつねに $p(f) = \tilde{p}(f) = 0$ が成立し，素性 $f$ はモデルに対する制約として機能しない．したがって誤差の検定は，素性 $f$ の発話条件に合致する学習サンプル $(u, y)$ の集合 $S_f = \{(u, y) | \exists y, f(u, y) = 1\}$ $(\subseteq S)$ において行うこととする．

$S_f$ における素性 $f$ の生起確率に関し，モデルによる推定値 $p_{S_f}(f)$ と観測値 $\tilde{p}_{S_f}(f)$ を，前式 (5)，(6) と同様に次式で定める．式中，$\tilde{p}_{S_f}(u)$ と $\tilde{p}_{S_f}(u, y)$ は，それぞれ $S_f$ における $u$ および $(u, y)$ の観測確率を表す．

$$p_{S_f}(f) = \sum_u \tilde{p}_{S_f}(u) \sum_y p(y|u) f(u, y) \qquad (9)$$

$$\tilde{p}_{S_f}(f) = \sum_{u,y} \tilde{p}_{S_f}(u, y) f(u, y) \qquad (10)$$

また，以下では，$S_f$ のサンプル数を $m$ で表し☆☆，$S_f$ における素性 $f$ の観測頻度を $c$ で表すこととする．$c$ は次式で計算される非負の整数である．

$$c = m \cdot \tilde{p}_{S_f}(f) \qquad (11)$$

素性 $f$ の生起を確率 $p_{S_f}(f)$ で推定するモデルが，$S_f$ と性質の等しい $m$ 個のサンプルにおいて，$f$ の生起を $x$ 回と推定する確率は，2 項分布を用いて次式で求められる．

$$b(x; m; q) = {}_mC_x q^x \ (1 - q)^{m-x} \qquad (12)$$

$$q = p_{S_f}(f) \qquad (13)$$

この 2 項分布を用いて片側検定を行い，観測確率

☆ 厳密に尤度の増分を計算するためには，Step3 同様に，モデル中の全素性の重みを最適値に設定しなおす必要があるが，これにはさらに多大な計算量を要する．そこで，モデルがもともと有していた素性の重みは固定したまま．新たに追加した素性の重みのみを計算することで近似的な尤度の増分を求める．

☆☆ 後述の実験では，学習データ 3,250 サンプル，素性候補 21,096 種類を用い，$m$ の最大値は 1,306，最小値は 1，平均は 39.9 となっている．

$\tilde{p}_{S_f}(f)$ がモデルの推定確率 $p_{S_f}(f)$ よりも低い場合には $x \leq c$ となる確率を，逆に高い場合には $x \geq c$ となる確率を求め，モデルの誤差に対する信頼度 $R$ とする．

$$R = \begin{cases} \sum_{x=0}^{c} b(x;m;q) & if\ \tilde{p}_{S_f}(f) < p_{S_f}(f) \\ \sum_{x=c}^{m} b(x;m;q) & otherwise \end{cases} \tag{14}$$

信頼度 $R$ が低い素性は，モデルの推定値 $p_{S_f}(f)$ に有意な誤差が見られる素性であるから，候補として残すべき素性である．これらの素性は，制約としてモデルに追加することで，尤度の改善が期待できる．逆に，信頼度 $R$ が高い素性は，モデルの推定値 $p_{S_f}(f)$ に有意な誤差が見つからない素性であるから，候補から棄却してよい素性と見なすことができる．

### 3.3 素性の棄却方法

3.2 節で求めた信頼度 $R$ に基づき，候補集合中の素性を棄却する．本論文では，棄却閾値を候補集合中での信頼度順位に設定し，信頼度の低い $N$ 個の素性を候補として残す．棄却方法はこれ以外にも，信頼度に直接閾値を設定する方法や，1 位との信頼度比を用いる方法などが考えられるが，こうした代替手法の棄却効率については 4.3 節で後述する．

**図 1** に，本方式による素性選択処理の流れを示す．本方式では，素性の 2 段階探索と IIS（Inproved Iterative Scaling [6]）によるモデルパラメータの更新を繰り返すことでモデルを構築する．1st サーチを行わない場合は文献 6) の素性選択処理に等しく，図 1 の 2nd サーチが前記 Step 1 に，モデルの更新（IIS）が前記 Step 2 に対応する．1st サーチでは，前述の誤差検定に基づいて，現在のモデルに追加しても尤度改善効果が小さい素性をあらかじめ検出し，一時的に 2nd サーチの対象から除外する．これにより，計算量の大きい 2nd サーチの尤度計算処理を削減し，効率的に学習を行うことが可能となっている．

## 4. 評価実験

本素性選択方式を用いて，2 章で述べた発話意図識別モデルの学習を行い，学習効率を文献 6) の素性選択方式（以下，ベースライン方式と呼ぶ）と比較した．

### 4.1 実験条件

実験は，**表 2** に示すデータを用いて行った．タスクは横浜市内のホテルの検索と予約であり，各発話には，正解とする意図を人手で付与している．学習データとしては，多様な言い回しを抽出する観点から，次の 2 つの方法で収集した計 3,250 発話を用いた．1) ホテル紹介者役の話者との模擬対話で収集した予約者役の話者による 1,461 発話，2) 音声対話システムとの対話で収集した 1,789 発話．評価データ（オープン）としては，実際的な表現に対する識別性能を評価する観点から，WOZ 法で収集した 423 発話を用いた．実験に用いる発話の単語系列からは，前処理としてホテル名や駅名，日時，金額などの固有表現を擬似単語で置換し，間投詞は除去した．素性の候補には，2.2 節の方法で学習データから生成した 21,096 個の素性を用いた☆．モデルの学習は AlphaServer DS20（500 MHz）上で行った．

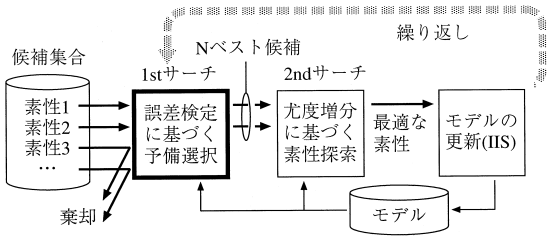



**図 1** 本方式による素性選択処理の流れ
Fig. 1 Proposed feature selection algorithm.

**表 2** 実験データ
Table 2 Experimental data statistics.

| | 学習データ | 評価データ | 合計 |
|---|---|---|---|
| 発話 | 3,250 | 423 | 3,673 |
| 語彙 | 857 | 436 | 957 |
| 語彙（前処理後） | 475 | 261 | 520 |

### 4.2 実験結果

まず，ベースライン方式による学習実験を行った．発話意図の誤り率の変化を**図 2** に示す．評価データの誤り率は，187 素性を獲得したときに最も良い値 5.4%が得られた．以下では，ある実験条件（素性選択方式および $N$ ベストの設定）で素性選択を繰り返した際に得られた最も良い評価データの誤り率を「最小誤り率」と呼ぶ．ベースライン方式の最小誤り率は 5.4%であり，最小誤り率が得られるまでに要した学習時間は，66.1 時間（CPU 時間）であった．

このときの素性の信頼度分布を**図 3** に示す．実線は，素性の全候補を信頼度が低い順に並べたときに，

☆ 候補となる 21,096 個の素性のうち，2.2 節，形式 A の素性は 10,987 個，B の素性は 10,109 個である．また，4.2 節の実験結果では候補から 187 個の素性を選択した際に最も良い誤り率が得られるが，これに含まれる形式 A の素性は 99 個，B の素性は 88 個となっている．

第 $k$ 位となる素性を示している．細い実線の間隔は 1,000 位である．図中の点は，実際に選択された素性を示す．横軸は，上述の最小誤り率が得られた 187 回までをとった．

図 3 から，繰返し回数の増加にともない，信頼度の分布は全体に高域側（モデルの推定誤差が小さい）へ移行していくが，選択される素性の順位はつねに上位に位置づけられることが分かる．図中，最も順位が悪いのは 67 回目の 2,805 位であり，それ以外はすべて 1,000 位以内である．このことから，信頼度順位は，尤度基準で素性を選択するときの良い見積りを与えるといえる．また，本実験では信頼度の上位 $N \geq 2,805$ 個の素性を 2nd サーチの対象とすれば，ベースライン方式で最小誤り率が得られたモデルと同じモデルが本方式でも得られることになる．

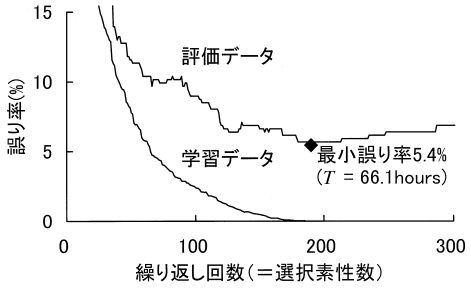



**図 2** ベースライン方式の誤り率
Fig. 2 Training curve of baseline method.

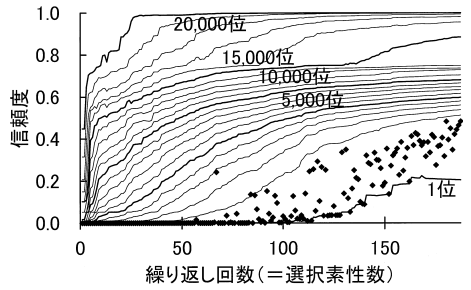



**図 3** ベースライン方式で選択された素性の信頼度
Fig. 3 Confidence scores of the features selected by the baseline method.

本方式による実験結果を**表 3** に示す．表中の学習時間とは，各実験条件でその最小誤り率が得られるまでに要した CPU 時間を，ベースライン（66.1 時間）との相対値で示したものである．実験の結果，3,000 個の候補を 2nd サーチに残した場合，ベースライン方式で得られる誤り率を劣化させることなく，学習時間を 18%に削減できることが確認された．

ところで，本方式により 2nd サーチの対象を絞り込んでも，候補数の割合ほどには学習時間が削減されない．たとえば，獲得素性数がベースラインと等しい N ベスト数 3,000 では，候補数が全体の 14%まで絞り込まれるが，トータルの学習時間は上述のとおり 18%までしか削減されない．本方式のオーバヘッドは，表 3 に示す 1st サーチの計算時間であり，学習時間全体から見て無視できるほど小さいにもかかわらず，オーバヘッドが生じて見える理由は以下である．2nd サーチにおける尤度増分計算はニュートン法を用いることから，素性ごとに計算量が異なる．特に，棄却される素性はもともとモデルの推定誤差が小さいため，重みをニュートン法で近似的に求めると初期値 0 の近くで解が得られることが多く，収束が早かったと考えられる．このように比較的計算量の小さい素性を数多く棄却したため，棄却数の割に学習時間が削減されなかったものと思われる．

### 4.3 棄却閾値の設定対象に関する考察

本論文では，1 段目の探索で棄却に用いる閾値を信頼度順位に設定しているが，これ以外の方法として次の (1)〜(3) が考えられる．本節では，これら代替手法の学習効率について考察する．

(1) 信頼度に閾値を設定．
(2) 1 位候補との信頼度差に閾値を設定．
(3) 1 位候補との信頼度比に閾値を設定：実際には 1 位の信頼度が 0 となることがあるため，以下では信頼度+$\delta$ ($\delta = 1 \times 10^{-12}$) の比とした．

前節で述べたように，本実験でベースライン方式と同じ 187 素性のモデルを得るためには，ベースライン方式で選択された 187 素性の中で順位が最も悪い 2,805 位に閾値を設定すればよい．このとき 2nd サー

**表 3** 信頼度 N ベスト探索による学習時間の削減効果
Table 3 Training time reduction by the proposed method.

| N ベスト数 | 最小誤り率 (%)* | *の獲得素性数 | 学習時間（相対） | 学習時間の内訳 | | |
|---|---|---|---|---|---|---|
| | | | | 1st サーチ | 2nd サーチ | モデル更新 |
| 500 | 6.2 | 199 | 0.04 | 0.00046 | 0.032 | 0.0059 |
| 1,000 | 5.7 | 154 | 0.06 | 0.00031 | 0.053 | 0.0038 |
| 2,000 | 5.7 | 167 | 0.12 | 0.00034 | 0.11 | 0.0069 |
| 3,000 | **5.4** | 187 | **0.18** | 0.00041 | 0.17 | 0.0065 |
| ベースライン | 5.4 | 187 | 1.00 | – | 0.99 | 0.0065 |

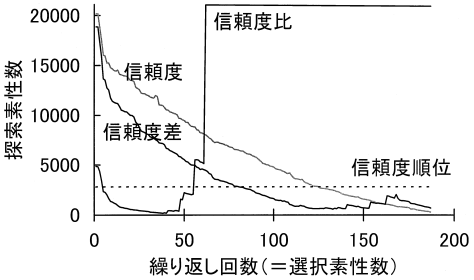



**図 4** 信頼度閾値の設定方法による探索効率の比較
Fig. 4 Comparison of training efficiency by confidence measures.

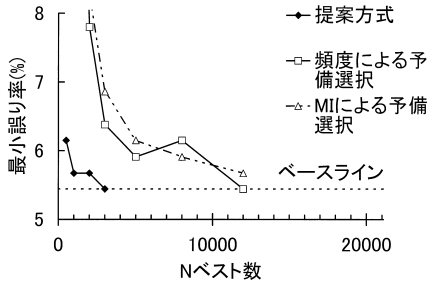



**図 5** N ベスト数による誤り率の比較
Fig. 5 Error rates compared by N-best size.

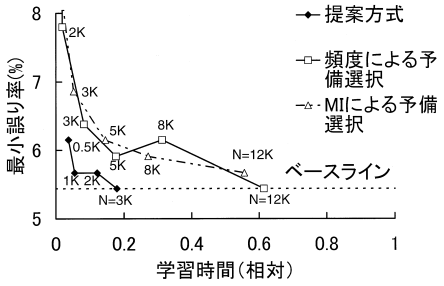



**図 6** 学習効率の比較
Fig. 6 Comparison of training efficiency.

チでは，合計 2,805 個 × 187 回 = 524,535 個の候補に対し，尤度の増分を計算することになる．

同様にして，手法 (1)〜(3) により，ベースライン方式と同じ 187 素性のモデルを得ようとした場合，2nd サーチの対象となる素性数は**図 4** のよう変化する．187 回の合計では，信頼度順位に閾値を設定した場合に比べ，(1) 信頼度に閾値を設定した場合に 2.2 倍，(2) 信頼度差を用いた場合には 1.4 倍，(3) 信頼度比を用いた場合には 5.2 倍の候補が 2nd サーチの対象となり，いずれの代替手法も学習効率が劣化する．

手法 (1)，(2) における効率の劣化は，図 3 に示す信頼度分布が学習の初期には低域に集中し，素性選択を繰り返すことで高域に移行していくことに起因する．棄却閾値は図 3 右側領域に合わせて設定されるため，左側領域で効率が悪くなる．また，手法 (3) に関しては，素性選択を繰り返したときに，図 3 中央付近まで 1 位の信頼度が 0 またはそれに近い小さな値を維持するため，信頼度比にきわめて大きな閾値を設定する必要がある．図 4 では，信頼度比に $10^7$ のオーダで閾値を設定しており，繰返し回数が 61 を超えた時点から棄却能力が完全になくなった．適切な正規化を行わない限り，信頼度比を棄却閾値として用いることは難しい．このように代替手法 (1)〜(3) の棄却性能が劣化する原因は，本信頼度の特性によるものである．実験条件によっては，信頼度順位と信頼度差との間で効率が逆転する可能性もあるが，おおむね同様な傾向が一般に観測されるものと考えられる．

## 5. 静的な予備選択方式との比較

学習に要する計算量を削減する単純な方法としては，候補となる素性に対し，あらかじめ，頻度や相互情報量に基づく予備選択を適用する方法が考えられる．本章では，これらの予備選択方式の学習効率を本方式と比較する．

頻度/相互情報量による予備選択方式の評価は，次のようにして行う．4 章の実験で候補として用いた 21,096 個の素性から，あらかじめ，それぞれの基準により上位 $N$ 個の素性を選択しておく．これら $N$ 個の素性を候補として，ベースライン方式により学習を行う．$N$ の値としては，4.2 節で本方式の評価に用いた 500，1,000，2,000，3,000 に加え，5,000，8,000，12,000 の場合を評価した．頻度の 5,000 ベストとは頻度 5 以上の素性に相当し，頻度の 8,000 ベストは頻度 3 以上の素性に，頻度の 12,000 ベストは頻度 2 以上の素性に相当する．

実験結果を**図 5** に示す．ベースラインと比べて，相互情報量による予備選択を行った場合には，$N \leq 12,000$ で最小誤り率が劣化する結果となった．同様に，頻度による予備選択を行った場合には $N \leq 8,000$ で劣化する結果となった．また，両方式とも，$N \leq 3,000$ では最小誤り率が急激に劣化した．これらの予備選択方式と比較して，本方式では，$N$ を小さくしていったときの最小誤り率の劣化が緩やかであることが分かる．また図 5 には示していないが 4.2 節で述べたように，本方式では $N \geq 3,000$ ($\geq 2,805$) とした場合に，最小誤り率がベースラインより劣化しない．以上から，本方式は，$N$ ベスト数の設定に関して比較的安定した識別性能が得られる効率化方式であるといえる．

**図 6** は，図 5 と同じ実験結果を示すもの（$N$ ベス

ト数によるプロット）であるが，横軸には学習時間をとってある．学習時間とは，4.2 節と同様，最小誤り率が得られるまでに要した CPU 時間のことであり，ベースラインとの相対値で示している．図中，グラフが左下に位置するほど，効率の良い学習方式といえる．本方式は，4.2 節で述べたように，もともと計算量の小さい素性を棄却する傾向があるため，$N$ ベスト数あたりの計算量が頻度や相互情報量による予備選択方式より大きくなるが，こうした計算量の増分を考慮しても，図 6 から，本方式の学習効率が優れていることが分かる．

## 6. おわりに

意味理解処理に最大エントロピー法を適用するうえで障害となる既存学習アルゴリズムの計算量の問題を解決するため，効率的な素性選択方式を提案した．本方式は，従来の尤度を基準とする逐次的な素性選択方式[6] に対し，モデルの推定誤差を検定する処理を導入することにより，尤度改善効果が小さい素性候補を効率的に検出し，候補から棄却する点を特徴とする．これにより，尤度増分計算に基づく素性の探索処理を削減し，冗長性を有する素性の候補集合から高速にモデルを構築することが可能となっている．

本方式を，音声対話システムにおけるユーザの発話意図を識別するモデルの学習に適用し，評価実験を行った．実験の結果，従来方式の識別精度を劣化させることなく，学習時間を 18%に削減することができた．また，頻度や相互情報量に基づく予備選択を素性候補に適用する方法と比較して，同一学習時間で安定して良い識別性能が得られ，学習効率化における本方式の有効性を確認することができた．

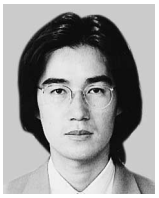
今回の実験では，性能劣化を来たすことなく，素性の候補を 3,000 個程度まで絞り込むことが可能であった．これは本実験で候補として用いた素性全体のおおむね 1/7 にあたるが，互いに強い相関を持つ素性がより多く候補中に含まれる場合，棄却可能な割合はさらに増えるものと期待される．

一方，3,000 という数値自体は，本実験条件において最も順位の悪い素性に合わせ設定したものであり，一般性は低い．順位のばらつきは，誤差検定により尤度の増分を見積もる際の精度限界によるものだが，さらに，実際の尤度増分計算に用いられる近似の特性が影響した可能性もある．今後，これらの関係を明らかにし，探索候補数を適切に設定/制御する方法を検討する必要がある．

## 参 考 文 献


1) Ramaswamy, G.N. and Kleindienst, J.: Hierarchical feature-based translation for scalable natural language understanding, *Proc. ICSLP2000*, pp.506–509 (2000).
2) Potamianos, A., Riccardi, G. and Narayanan, S.: Categorical understanding using statistical ngram models, *Proc. Eurospeech99*, pp.2027–2030 (1999).
3) Papineni, K.A., Roukos, S. and Ward, R.T.: Feature-based language understanding, *Proc. Eurospeech97*, pp.1018–4074 (1997).
4) Bellegarda, J.R. and Silverman, K.E.A.: Towards unconstrained command and control: Data-driven semantic interface, *Proc. ICSLP2000*, pp.258–261 (2000).
5) Chu-Carroll, J. and Carpenter, B.: Vector-based natural language call routing, *Computational Linguistics*, Vol.25, No.3, pp.361–388 (1999).
6) Berger, A.L., Della Pietra, S.A. and Della Pietra, V.J.: A maximum entropy approach to natural language processing, *Computational Linguistics*, Vol.22, No.1, pp.39–71 (1996).
7) Printz, H.: Fast computation of maximum entropy/minimum divergence feature gain, *Proc. ICSLP98*, pp.2083–2086 (1998).
8) Mikheev, A.: Feature lattice for maximum entropy modeling, *Proc, ACL/COLING98*, pp.848–854 (1998).
9) Wu, J., et al.: Efficient training methods for maximum entropy language modeling, *Proc. ICSLP2000*, pp.114–118 (2000).
10) Goodman, J.: Classes for maximum entropy training, *Proc. ICASSP2001* (2001).


(平成 13 年 11 月 15 日受付)
(平成 14 年 4 月 16 日採録)



**谷垣 宏一**（正会員）

平成 4 年東北大学工学部情報工学科卒業．平成 7 年同大学大学院情報科学研究科修士課程修了．同年三菱電機（株）情報技術総合研究所入社．平成 9 年より平成 11 年まで国際電気通信基礎技術研究所（ATR）に出向．音声認識，言語理解の研究に従事．日本音響学会会員．

**渡邉　圭輔**

平成 3 年東京工業大学工学部情報工学科卒業．平成 5 年同大学大学院総合理工学研究科修士課程修了．同年三菱電機（株）情報電子研究所（現，情報技術総合研究所）入社．音声認識，音声対話の研究に従事．日本音響学会，日本認知科学会各会員．

**石川　　泰**（正会員）

昭和 55 年東京工業大学工学部電気電子工学科卒業．昭和 57 年同大学大学院修士課程修了．同年三菱電機（株）入社．音声合成，音声対話の研究に従事．日本音響学会，電子情報通信学会，人工知能学会，米国音響学会各会員．