論　文

帰納的学習を用いた携帯端末向け機械翻訳手法

松原　雅文[†]　　荒木　健治[†]　　栃内　香次[††]

Machine Translation Method Using Inductive Learning for Mobile Terminal

Masafumi MATSUHARA[†], Kenji ARAKI[†], and Koji TOCHINAI[††]

あらまし　本手法は，携帯電話などの小型の端末を想定した機械翻訳手法である．携帯電話など携帯性を重視した小型の端末では，それ自身の大きさの制約から装備可能なキーの数が必然的に制限される．そこで，本手法においては，この少数のキーのみで入力を可能とするために，一つの入力キーに複数の文字を割り当てることとした．更に，迅速な入力を可能とするために 1 文字の入力を 1 打で行うことにしている．そのため，本手法への入力文字列である数字列は複数の文字列に対応しており，結果として，この数字列に対する翻訳候補も複数存在することになり，あいまいさが増している．しかしながら，本手法では帰納的学習を用いることにより，対象に依存した翻訳ルールを自動的に獲得し，翻訳に利用することができる．このような高い適応能力により，本手法においては翻訳時のあいまいさを極力解消し，正しい翻訳を行うことが可能となる．本手法に基づくシステムを作成し，事前に少量の単語翻訳ルールを与えた日英機械翻訳実験を行った結果，最終的に約 70[%] の翻訳精度及び翻訳効率の高さが示され，本手法の有効性が確認された．

キーワード　帰納的学習，携帯端末，機械翻訳，文字情報縮退方式

1. まえがき

近年，携帯電話を代表とする移動通信が世界的にも急速に普及している．第 1 世代，第 2 世代の移動通信には世界統一標準はなく，世界中どこでも使えるような方式は存在しなかった．しかしながら，IMT-2000 と呼ばれる第 3 世代移動通信方式では世界統一標準を目指しており，これが実現すると，1 台の携帯電話のみで世界中のどこからでも通信を行うことが可能となる [1]．よって，各個人が携帯電話等の端末を，海外においても持ち歩くようになるものと推測される．また，海外旅行者数の増大もあり，他言語を母国語とする世界中の人々と接する機会が増えるものと考えられる．このような背景から，海外においてコミュニケーションを円滑に行うために，携帯電話等の小型の端末で翻訳が可能なシステムの開発が望まれる．一般に，対話によるコミュニケーションにおいては高い即時性が要求される．よって，海外旅行等での会話を円滑に行うためには，高速な機械翻訳システムが必要である．入力方法が煩雑で，入力に要する時間が増大すると，翻訳処理全体としての処理速度も低下することになる．そのため，翻訳処理全体の高速化を実現するためには，迅速な入力方法が必須となる．

小型の端末での機械翻訳システムを想定した場合，その入力方法が問題となる．キーを使わずに入力を行う方法として，音声入力や手書き文字入力が挙げられる．音声入力では，音声の性質上その機密性に問題がある．また，周囲の雑音による影響を受けやすいため，このような環境下で連続音声認識を高精度に行うのは困難であると考えられる [2], [3]．よって，現在の小型端末上で高精度の音声認識を行うのは，その処理性能を考えるとなおさら困難である [4]．

手書き文字入力には，意図する文字をそのまま入力する方式 [5], [6] と専用の記号を用いて入力を行う方式(注1)がある．前者の場合，入力のために特別な訓練がほとんど必要ないという利点はあるが，入力に用いられる文字の種類が後者に比べて多いため，その認識精度や認識速度は不十分である．後者の場合，入力さ

[†] 北海道大学大学院工学研究科，札幌市
Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo-shi, 060–8628 Japan
[††] 北海学園大学大学院経営学研究科，札幌市
Graduate School of Business Administration, Hokkai-Gakuen University, 4–1–40 Asahimachi, Toyohira-ku, Sapporo-shi, 062–8605 Japan

(注1)：パームコンピューティング株式会社の Graffiti など．

れる記号の種類が限定されており，高精度の認識が可能である．しかしながら，その入力を行うためには，専用の記号を覚える必要がある．

一方，キーを用いた入力を考えた場合，携帯電話等ではその大きさの制約から装備可能なキー数が制限される．少数のキーでの入力操作が煩雑になるのを防ぐために，キーへの文字の割当てを独自の仕様にしているものがある [7]．このような入力方式では迅速な入力が可能となるが，その方式独自のキー配置を覚えなければならない．だれもが簡単に使用することができるシステムを構築するためには，特別な訓練を必要とせずに少数のキーのみで迅速に入力が可能な方式を採用する必要がある．

携帯電話は現在広く使用されている携帯端末である．携帯電話はその大きさの制約から装備可能なキー数が限られるが，最低でも 0〜9，#，＊の 12 個のキーは装備しているのが普通である．日本語には約 50 個の仮名が存在するので，これを 12 キーを用いて入力する場合，一つのキーに複数の文字を割り当てる必要がある．一般的には，あ行，か行などの 1 行が一つのキーに割り当てられている．また，一般的な入力方式としては，文字循環指定方式が採用されているが，この入力方式では，1 文字の入力に複数回の打鍵が必要となる．よって本手法では，迅速な入力を可能とするために文字情報縮退方式 [8] を採用し，1 文字の入力を 1 打で行うものとする [9], [10]．これにより，例えば「野球」を入力する場合，文字循環指定方式では，$1+2+5+3=11$ 回の打鍵数が必要なのに対し，文字情報縮退方式では，$1+1+1+1=4$ 回の打鍵数で入力が完了する．しかしながら，入力された数字 1 文字は意図した仮名文字以外にも，それと同一行の他の仮名文字にも対応することになり，結果として入力数字列は多数の日本語文に対応し，あいまい性をもつ．

この数字列を英語文に翻訳するためには，数字漢字変換手法 [9], [10] によりいったん，日本語文に変換してから日英機械翻訳を行う方法も考えられる．しかしながら，数字漢字変換手法により入力数字列を日本語文に変換するためには，使用者による校正処理が不可欠である．翻訳処理の途中結果である，この日本語文に対して校正を行うことは，入力に要する労力の増大と，それに伴う処理速度の低下を引き起こすものと考えられる．本手法の目的は携帯電話等の小型の端末において，迅速な機械翻訳システムを実現することである．よって，本手法においては，日本語文に対応した数字列を直接，英語文に翻訳することとした．このように処理を単一化することにより翻訳速度の向上が期待でき，また，処理の多重化による精度の低下を回避可能である．日英機械翻訳には，解析的な翻訳知識をあらかじめ与える手法や統計的な手法 [11]，用例に基づく手法 [12], [13] などが存在するが，いずれも原言語文字列としてあいまいな文字列を対象とはしていない．本手法のように，あいまい性をもつ原言語文字列を直接，目的言語文字列に翻訳する手法は，筆者らの知る限りにおいてほかに存在しない．前述のように，数字列は複数の日本語文に対応しているので，この数字列に対する翻訳候補は更に多数存在することになる．このあいまいさのために，事前に数字列に対する翻訳ルールを与えることは困難であると考えられる．この問題を解消するために，本手法においては帰納的学習のもつ高い適応能力を利用している [14]．本手法で用いる帰納的学習とは，字面情報の共通部分を手掛りに翻訳ルールを獲得し，更に獲得した翻訳ルール同士から再帰的に翻訳ルールを獲得することを指す．この帰納的学習の利用により，対象に依存した翻訳ルールを自動的に獲得可能である．このように対象となる範囲を自動的に限定することができるので，本手法においては翻訳候補の重複を最小限に抑えることができる．対象が変化した場合にも，変化した対象に依存した翻訳ルールを随時学習することができる．更に，翻訳に使用した結果，正しいと判断されたルールはそのゆう度を上げ，誤りと判断されたルールはそのゆう度を下げることにより，対象に適合した正しい翻訳ルールを次回から優先的に選択することが可能となる．このようにして様々な対象に動的に適応可能であるので，本手法は汎用性も保持しており，原則として種々の言語に適応可能である．

本手法は，基本的には翻訳知識を全くもたない状態からでも，使用者が使用する過程で，翻訳，校正，学習を繰り返すことにより翻訳知識を獲得し，次第に正しく翻訳を行うことができるようになる．しかしながら，携帯端末での利用を考えた場合，使用者を限定した利用が予想され，十分な量の学習データが入力される前の初期の段階から，高い精度での翻訳が要求される．そこで，学習効率を考慮し，本手法においては単語単位での翻訳ルールをあらかじめ与えるものとしている．すべての翻訳ルールを与えることは現実的には不可能であり，多大な翻訳ルールを与えることも，翻訳の際のルール重複による翻訳精度の低下，処理量の

表 1　数字と仮名の対応関係
Table 1　Correspondence of number to *Kana*.

1:あいうえおー	2:かきくけこ	3:さしすせそ
4:たちつてとっ	5:なにぬねの	6:はひふへほ
7:まみむめも	8:やゆよゃゅょ	9:らりるれろ
*:（半）濁音	0:わをん	#:区切り記号

※「#」は単語区切り,「##」は読点,
「###」は句点を表す.

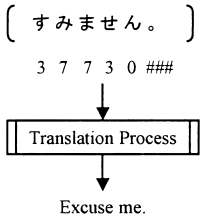

図 1　翻　訳　例
Fig. 1　Translation example.

増加などを引き起こすため好ましくないと考えられる．よって，ここで与える知識は，翻訳における言語間での対応関係を単語単位で決定可能な，必要最小限の翻訳ルールとしている．この少量の知識からでも本手法のもつ学習能力により次第に翻訳精度を向上させることができる．更に，このようにして与えられた正しい知識を利用して帰納的学習を行うことにより，より多くの正しい翻訳ルールを獲得することが可能である．

本論文では，帰納的学習を用いた携帯端末向け機械翻訳手法を提案し，その処理過程を説明する．更に，本手法に基づいて作成したシステムを用いて行った評価実験の結果から，本手法の有効性について述べる．

2.　携帯端末向け機械翻訳手法

本研究の目的は，携帯電話のような小型の端末を想定し，迅速な入力が可能な機械翻訳システムを実現することである．本システム上で，使用者は意図した日本語文の仮名に対応した数字列を 12 キーにより入力する．数字と仮名の対応関係を表 1 に示す．一般的な携帯電話等での文字の割当てと同様の割当てを採用することにより，使用者は特別な訓練なしに入力を行うことができる．更に，1 文字の入力を 1 打で行うことにより，迅速な入力を可能としている．そして，この入力された数字列を直接，英語文に翻訳する．そのため，翻訳候補となるルール数は増大するが，本手法のもつ高い適応能力により，このあいまいさを極力排除する．

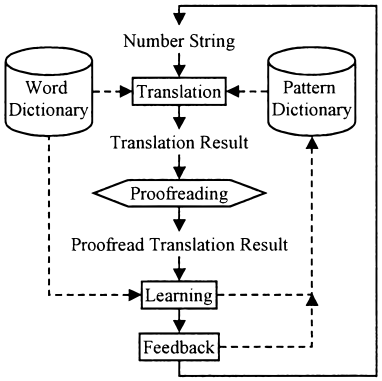

図 2　処 理 過 程
Fig. 2　Procedure.

本手法の翻訳例を図 1 に示す．このように正しく翻訳を行うことが可能なシステムの構築を目指す．

3.　処 理 過 程

本手法の処理過程を図 2 に示す．図 2 に示すとおり，入力された数字列に対して，翻訳，校正，学習，フィードバックの順に処理が行われる．

3.1　入 力 処 理

本手法において，使用者は携帯電話等に装備されている 12 個の数字キーを用いて入力を行う．12 キーでの入力を可能とするために，一つの入力キーに複数の文字を割り当てており，更に，迅速な入力を可能とするために 1 文字の入力を 1 打で行うこととしている．そのため，本手法への入力文字列である数字列は，意図した原言語文字列のほかにも，複数の原言語文字列に対応することになる．

日本語のようにべた書き表記を用いる言語の場合，単語分割の単位は各個人ごとに異なるのが普通である．しかしながら，各個人ごとの単語分割の単位はほぼ統一されているものと考えられる．本手法は，単語分割の単位に基づいて自動的に翻訳ルールを獲得していく．すなわち，各個人ごとの分割単位に合わせて学習を行うことができるので，この分割単位の違いを吸収することができる．また，本手法への入力文字列である数字列はあいまいさをもっているので，システムによりこれを単語分割するのは困難であると考えられる．よって，本手法への入力数字列は使用者により単語分割するものとしている．単語分割の単位は使用者にゆだねられているので，迅速な入力の妨げにはならない．入力の例を表 2 に示す．

表 2　入力処理例
Table 2　Example of the input process.

意図した日本語文	枕を下さい.
使用者による単語分割	枕␣を␣下さい␣.
入力数字列	729#0#24*31###

表 3　パターン翻訳辞書の構造
Table 3　Structure of the pattern dictionary.

数字列	英語	正翻訳度数	誤翻訳度数
α#4#7#11#4*3#2###	May I α ?	5	1
414	get through	3	3
⋮	⋮	⋮	⋮

※ $\alpha, \beta, \ldots$ は変数を表す.

表 4　単語翻訳辞書の構造
Table 4　Structure of the word dictionary.

数字列	英語	正翻訳度数	誤翻訳度数
2161	coffee	5	0
729	pillow	4	2
⋮	⋮	⋮	⋮

表 5　翻訳処理例
Table 5　Example of the translation process.

入力数字列（日本語）	
729#0#24*31###（枕␣を␣下さい␣.）	
パターン翻訳ルール	
数字列	英語
α#0#24*31###	Could I have a α ?
単語翻訳ルール	
729	pillow
翻訳結果	
Could I have a pillow ?	

※ $\alpha, \beta, \ldots$ は変数を表す.

表 2 において，使用者が意図する日本語文は「枕を下さい.」である．使用者は，この意図した日本語文を単語分割（␣）し，その仮名に対応した数字列を表 1 に従い入力する．よって，ここで使用者が入力する数字列は「729#0#24*31###」となる．このように，仮名 1 文字の入力を 1 打で行うことにより，迅速な入力を可能としている．

なお，入力した数字列をそのまま表示すると，使用者による入力の確認が困難になる可能性がある．これを回避するためには，それぞれの数字に対応した子音情報を表示する方法が考えられる．本手法の入力で用いている数字 1 文字は日本語の仮名 1 行に対応している．仮名 1 行は多くの場合，一つの子音に対応しているので，子音情報の表示は容易に実現可能である．例えば，「すみません.」の入力に要する数字列は「37730###」であり，この場合の表示は「SMMSW.」となる．しかしながら，数字列をそのまま表示する場合と比べて，どちらの使い勝手が良いかは使用者に依存するものと考えられるので，この表示方法については使用者により選択可能とする．なお，どちらの表示方法を選択した場合であっても，後述の処理へは影響しないので，本論文中では，入力数字列をそのまま表示するものとしている．

3.2 翻訳辞書

本手法では翻訳のためにパターン翻訳辞書と単語翻訳辞書を用いる．それぞれの辞書の構造を表 3，表 4 に示す．どちらも，日本語に対応した数字列，翻訳結果となる英語文字列，このルールを翻訳に使用した際に正しいと判断された回数，及び，誤りと判断された回数から構成される．パターン翻訳辞書には，**3.5** の学習処理により獲得された，語順を保持した単語列レベルの翻訳ルールが登録されている．このパターン翻訳ルールには変数部分をもつものが存在し，翻訳の際に他の翻訳ルールを代入することができる．単語翻訳辞書には，単語単位での翻訳ルールが登録されている．この単語翻訳ルールは，あらかじめ与えられた，言語間での対応関係が一意に決定可能な翻訳ルールである．

3.3 翻訳処理

入力された数字列は，パターン翻訳辞書と単語翻訳辞書を用いて目的言語文に翻訳される．翻訳は，適用可能な翻訳ルールを検索し，組み合わせることにより行われる．翻訳ルールの変数部分以外のすべての単語が，原言語文中の単語の並びと同一の順番で含まれている場合，その翻訳ルールは原言語文に対して適用可能と判断される．パターン翻訳ルールは語順を保持した単語列レベルの翻訳ルールなので，単語翻訳ルールに比べて，翻訳に使用した際に正翻訳となる可能性が高いと考えられる．よって，まず最初にパターン翻訳辞書中の検索を行い，適用可能なルールが存在しなくなったら，次に単語翻訳辞書を検索する．翻訳の処理手順を以下に示す.

① 適用可能な翻訳ルールを検索
② 検索されたルールを適用し数字列を削除
適用可能な翻訳ルールが存在する間，繰返し
③ 検索ルールの英語を組み合わせて出力

翻訳処理の例を表 5 に示す．表 5 において，入力数字列「729#0#24*31###」に適用可能な翻訳ルールを検索すると，パターン翻訳辞書中

から（α#0#24*31###：Could I have a α ?）が見つかるので，これを適用し入力数字列中の「#0#24*31###」を削除する．次に，残った入力数字列「729」に対して適用可能なルールを検索する．単語翻訳時書中より（729：pillow）が見つかるので，これを適用し，入力数字列から「729」が削除され，翻訳ルールの検索が終了する．最後に，翻訳ルールを適用した順番に組み合わせる．すなわち，パターン翻訳ルールの英語側の変数部分 α に単語翻訳ルールの英語「pillow」を代入することにより，翻訳結果「Could I have a pillow ?」を出力する．

なお，翻訳ルール検索の際に，候補となる翻訳ルールが競合する場合には，以下の順にゆう度評価を行い，適用する翻訳ルールを決定する．

① 抽象度が最小
② 適合度が最大
③ 登録順位が最新

抽象度は，翻訳ルールにおける変数の割合を示している．抽象度が 0 である翻訳ルールとは変数を含まない原文一致となるルールであり，このようなルールは，翻訳に使用した際に正翻訳となる可能性が高いと考えられる．よって，抽象度が低い翻訳ルールの優先順位を高く設定している．翻訳ルールの抽象度を式 (1) に示す．

$$\text{抽象度} = \frac{\text{変数の個数}}{\text{単語数}} \tag{1}$$

例えば，「α#0#24*31###」の抽象度は 1/4=0.25 となる．翻訳候補となるルールの抽象度が同一の場合には，適合度によりゆう度評価を行う．適合度を式 (2) に示す．適合度は，そのルールのもつ正翻訳度数と誤翻訳度数により決定される．正翻訳，誤翻訳度数は **3.6** のフィードバック処理により更新される．

$$\text{適合度} = \frac{\text{正翻訳度数}}{\text{正翻訳度数} + \text{誤翻訳度数}} \tag{2}$$

適合度も同一となる場合には，登録順位が最新のものが現在の対象に最も適合していると考えられるので，これに決定する．このようにして，現在の対象に適合した翻訳ルールを決定する．

3.4 校正処理

翻訳結果に誤りが含まれる場合，人手により校正が行われる．ここで与える情報は，校正済み翻訳結果の字面情報のみであり，文法的な知識を与える必要はない．種々の言語間での翻訳を考えた場合，字面上での正解を与えることができても，文法的な知識を整合性を保って正確に与えることは困難であると考えられる．よって，字面上の正解からシステムが自動的に文法的な知識を翻訳ルールとして学習することが望ましい．そこで，本手法の校正処理では，正解となる字面情報のみを与えるものとしている．

3.5 学習処理

入力数字列と校正済み翻訳結果を用いて学習処理が行われる．ここでは，単語翻訳ルールと字面上の共通部分を利用して翻訳ルールを獲得し，パターン翻訳辞書に登録する．

3.5.1 単語翻訳ルールの利用によるルール抽出

単語翻訳ルールは単語単位で言語間の対応関係が決定している翻訳ルールなので，文中に含まれる単語翻訳ルールは独立性が高い，すなわち，まわりの単語に対して依存性が低いと考えられる．よって，この部分を変数化するものとしている．翻訳時には，この変数部分に他の翻訳ルールを代入することにより，多様な翻訳を行うことが可能となる．また，単語翻訳ルールは人手により与えられた正しい翻訳ルールなので，これを考慮して以下のヒューリスティクスを与える．

「二つの単語翻訳ルールに挟まれた部分は異なる言語間においてその対応関係を決定できる．」

このヒューリスティクスに基づき，翻訳例中において単語翻訳ルールに挟まれた部分を翻訳ルールとして抽出し，更にその部分を変数化したルールを抽出する．単語翻訳ルールを利用した翻訳ルールの抽出手順を以下に示す．

① 翻訳例に含まれる単語翻訳ルールを検索
② 検索された単語の部分を変数化
含まれる翻訳ルールが存在する間，繰返し
③ 変数に挟まれた部分を抽出
④ 抽出した部分を変数化

この例を表 6 に示す．表 6 の翻訳例に含まれる単語翻訳ルールを検索すると（2161：coffee）が見つかるので，この部分を変数化する．同様に検索を行うと単語翻訳ルール（148：tea）が見つかるので，この部分を変数化する．次に，二つの変数に挟まれた部分（α#2#β：α or β）を翻訳ルールとして抽出する．更にその部分を変数化した翻訳ルール（α#6#122*#4*3#2###：Would you like α ?）を抽出する．このように，単語翻訳ルールを利用してパターン翻訳ルールを獲得する．

3.5.2 共通部分の利用によるルール抽出

2 組の翻訳例あるいはパターン翻訳ルールより翻訳

表 6　単語翻訳ルールを用いた学習例
Table 6　Example of learning by the word translation rules.

単語翻訳ルール	
数字列（日本語）	英語
2161（コーヒー）	coffee
148（お茶）	tea
翻訳例	
2161#2#148#6#122*#4*3#2### （コーヒー␣か␣お茶␣は␣いかが␣です␣か␣？）	Would you like coffee or tea ?
獲得ルール	
α#2#β#6#122*#4*3#2###	Would you like α or β ?
α#2#β	α or β
α#6#122*#4*3#2###	Would you like α ?

※ $\alpha, \beta, \ldots$ は変数を表す.

表 7　共通部分を用いた学習例
Table 7　Example of learning by the common segments.

パターン翻訳ルール	
数字列（日本語）	英語
α#4*#46*2#0#34#4#7#11#4*3#2### (α␣で␣タバコ␣を␣すっ␣て␣も␣いい␣です␣か␣?)	May I smoke α ?
翻訳例	
414#4#7#11#4*3#2### (通っ␣て␣も␣いい␣です␣か␣?)	May I get through ?
獲得ルール	
α#4*#46*2#0#34	smoke α
414	get through
α#4#7#11#4*3#2###	May I α ?

※ $\alpha, \beta, \ldots$ は変数を表す.

ルールを抽出する過程である．2 組の翻訳例あるいはパターン翻訳ルールにおいて，共通となる単語列を決定し，差異部分を翻訳ルールとして抽出し，更にその部分を変数化したルールを抽出する．共通部分を利用した翻訳ルールの抽出手順を以下に示す.

① 2 組の翻訳例あるいは翻訳ルールを選択
② 共通部分を決定
③ 差異部分を抽出
④ 抽出した部分を変数化

この例を表 7 に示す．表 7 において，パターン翻訳ルールと翻訳例が任意に選択されている．これらの字面上の比較から共通部分を決定する．字面が同一である下線部分が共通部分となる．よって，差異部分である（α#4*#46*2#0#34：smoke α）と（414：get through）をルールとして抽出する．更にその部分を変数化した（α#4#7#11#4*3#2###：May I α ?）を抽出する．なお，差異部分が 2 箇所以上ある場合に抽出を行うと，2 箇所以上の変数部分を含む翻訳ルールが抽出されることになる．このような翻訳ルールは，原言語側と目的言語側の変数が 1 対 1 に対応しなくなるので，翻訳に使用する際に正しい箇所へ他の翻訳ルールを代入できない可能性がある．よって，翻訳ルールの抽出は差異部分が 1 箇所のときにのみ行う.

また，ここで抽出された翻訳ルールを用いて，同様の処理により，他の翻訳例あるいはパターン翻訳ルールに対してルールの抽出を試みる．このようにして再帰的な抽出を行っていくと，限りなく抽象化が行われる可能性があるので，再帰的な抽出は，抽出される翻訳ルールの抽象度が 0.5 未満の場合に制限している.

このように，単語翻訳ルールと字面上の共通部分を利用して，抽象度の異なる様々な翻訳ルールを効率良く獲得し，パターン翻訳辞書に登録していく.

3.6　フィードバック処理

翻訳結果と校正済み翻訳結果の字面上の比較から使用した翻訳ルールの正誤を決定し，そのゆう度を更新する．翻訳に使用されたルールのうち，構成するすべての単語がその順番に校正済み翻訳結果に含まれている場合，その翻訳ルールは正しいと判断し，その正翻訳度数を 1 増加させる．それ以外の翻訳ルールは誤りと判断し，その誤翻訳度数を 1 増加させる．この正翻訳度数，誤翻訳度数は **3.3** のゆう度評価に利用される.

このようにして更新された翻訳辞書を用いて次回からの翻訳を行うので，翻訳，校正，学習を繰り返すことにより，次第に精度を向上させることができる．

4. 評価実験

処理過程に基づき実験システムを作成し，その評価のために実験を行った．

4.1 実験データ及び実験手順

実験データを表 8 に示す．本システムを海外旅行中に利用する場合を想定し，旅行者用英会話文 [15]〜 [24] から各場面ごとの翻訳例を抽出し，実験データとして使用した．単語翻訳ルールとして，言語間での対応関係を決定できる名詞 120 単語をあらかじめ与えた．単語翻訳ルールとして汎用的な辞書を与える方法も考えられるが，この場合，本手法の学習能力による翻訳の前に，与えた辞書を利用した翻訳が数多く行われることが予想される．よって，今回の実験においては，本手法の学習能力の確認を目的とし，言語間での対応関係を決定でき，かつ最低限必要と考えられる単語のみを与えることとした．パターン翻訳ルールは文脈により正解が変化するため，単語翻訳ルールのように一意に対応関係を決定するのは困難であると考えられる．よって，パターン翻訳ルールは本手法の学習能力により自動的に獲得していくものとし，初期状態は空とした．この状態から表 8 に示す順に，1 文ずつ翻訳を行っていく．

4.2 評価方法

翻訳結果を 100 文ごとに再現率，翻訳精度により評価した．再現率を式 (3) に示す．

$$\text{再現率} = \frac{\text{正翻訳文数}}{\text{全文数}} \tag{3}$$

再現率は，実験に使用した全文数に対して正しく翻訳できた文数の割合を表している．実験データとして用いた翻訳例を正解の基準とし，これと意味が同じと判断される翻訳結果を正翻訳とした．正解となる翻訳結果には複数のものが考えられるが，ここでの正解は文献の著者により与えられた翻訳例である．このように，文献の著者により与えられた翻訳例を正解の基準とすることにより，この著者を本システムの使用者とみなした場合に行われる正誤の判断を擬似的に再現している．

本手法においては，与えられた単語翻訳ルールと翻訳例中の字面上の共通部分を手掛りに学習を行い，獲得された翻訳ルールを用いて翻訳を行っている．よって，これまでに出現していない単語を含む文を翻訳することは基本的に不可能である．そこで，本手法において翻訳可能な文を網羅文と定義した．網羅文とは，既に出現している単語により網羅されている文，すなわち，その文を構成するすべての単語が既に出現済みである文を指す．よって，理想的には，これまでに出現した単語を組み合わせることにより正しく翻訳できる文である．翻訳精度を式 (4) に示す．

$$\text{翻訳精度} = \frac{\text{網羅正翻訳文数}}{\text{網羅文数}} \tag{4}$$

網羅文のうち，正しく翻訳できた文の占める割合が翻訳精度である．

4.3 実験結果

再現率，翻訳精度の推移を図 3 に，また，それぞれの翻訳率を表 9 に示す．各場面において，再現率，翻訳精度はともに入力文数の増加に伴い上昇し，最終的に再現率で 45[%]，翻訳精度で 70[%] 程度までの上昇が確認された．なお，実験終了時にパターン翻訳辞書に登録されていた翻訳ルール数は 7,556 であった．

4.4 考察

4.4.1 適応能力について

実験の初期の段階である 0〜100 文から翻訳精度が高い値となっている．これは，本手法の学習能力，特

表 8 実験データ
Table 8 Data of the experiment.

場面	文数	入力文字数	入力単語数
機内	300	7,729	2,399
空港	600	15,628	4,889
チェックイン	400	10,587	3,273
全体	1,300	33,944	10,561
1 文平均		26.1	8.1

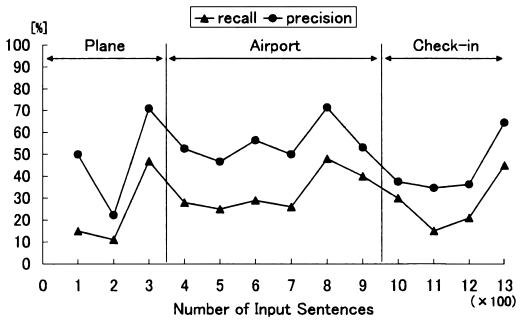

図 3 翻訳率の推移
Fig. 3 Changes in the translation rates.

に単語翻訳ルールを利用した学習が有効に作用した結果であると考えられる．あらかじめ単語翻訳ルールを与え，これに基づき学習を行うことにより，少量の学習データからでも対象に依存した翻訳ルールを獲得し，迅速に現在の対象に適応している．この例を表 10 に示す．表 10 に示す例は 96 文目の入力文である．

表 10 中のパターン翻訳ルール（α#0#24*31###：I'd like a α , please .）は 94 文目の翻訳例（6*19#0#24*31###：I'd like a beer , please .）と単語翻訳ルール（6*19：beer）から獲得された翻訳ルールである．これと単語翻訳ルール（325：fish）を組み合わせた結果，正しく翻訳が行われた．このように，単語翻訳ルールを利用することにより，1 組の翻訳例のみからでも翻訳ルールを抽出可能であり，次回からの翻訳に使用することができる．

その後の対象の変化のために翻訳精度は下降するが，その対象の変化に追従し，場面「機内」において最終的に 70[%] 程度までの翻訳精度の上昇が確認された．その後，入力データが場面「空港」に変化したことにより，これまでに出現していない単語が数多く出現したため，翻訳精度はいったん下降する．しかしながら，以前の対象において獲得された翻訳ルール中にも，現在の対象において正しいと判断される翻訳ルールが含まれているので，翻訳精度の大幅な低下は見られない．その後，また，現在の対象に適応していくことにより，場面「機内」と同様に翻訳精度が上昇し，70[%] 程度までの上昇が確認された．「チェックイン」についても同様である．このように，本手法の適応能力の高さが確認された．

表 9 翻 訳 率
Table 9 Translation rates.

場面	再現率 [%]		翻訳精度 [%]	
機内	24.3	73/300	55.1	59/107
空港	32.7	196/600	56.8	147/259
チェックイン	27.8	111/400	48.0	71/148
全体	29.2	380/1,300	53.9	277/514

表 10 正 翻 訳 例
Table 10 Example of the correct translation.

入力数字列（日本語）	
325#0#24*31###（魚␣を␣下さい␣.）	
パターン翻訳ルール	
数字列	英語
α#0#24*31###	I'd like a α , please .
単語翻訳ルール	
325	fish
翻訳結果	
I'd like a fish , please .	

※ $\alpha, \beta, \ldots$ は変数を表す．

4.4.2 学習の誤りについて

表 11 に学習の誤りの例を示す．表 11 に示す例は 30 文目の翻訳例である．この例では，単語翻訳ルール（1903*：オレンジ），（146*9：アップル）を用いて学習処理が行われる．単語翻訳ルールに挟まれた部分（α#4#β：α juice and β）が抽出されたが，これは誤った翻訳ルールである．このような誤った翻訳ルールは，翻訳に使用されると誤翻訳となるため，本手法のフィードバック処理によりそのゆう度は次第に減少していく．実際，実験終了時にこの翻訳ルールの正翻訳度数は 0，誤翻訳度数は 27 となっていた．このように，誤って獲得された翻訳ルールのゆう度低下が確認された．

4.4.3 翻訳の誤りについて

表 12 に翻訳の誤りの例を示す．表 12 に示す例は 388 文目の入力文である．表 12 中の（α#6#12#β#0#4*3#2*###：α red β .）は翻訳例（043#6#12#010#2*#631#0#4*3#2*###: I'd like red wine.）から獲得されたパターン翻訳ルー

表 11 学習の誤りの例
Table 11 Example of the erroneous learning.

翻訳例（日本語）	
数字列（日本語）	英語
1903*#4#146*9#6#19#73#2*### （オレンジ␣と␣アップル␣は␣あり␣ます␣が␣.）	We have orange juice and apple juice .
単語翻訳ルール	
1903*（オレンジ） 146*9（アップル）	orange apple
獲得ルール	
α#4#β#6#19#73#2*### α#4#β α#6#19#73#2*###	We have α juice and β juice . α juice and β We have α juice .

※ $\alpha, \beta, \ldots$ は変数を表す．

表 12　翻訳の誤りの例
Table 12　Example of the erroneous translation.

入力数字列（日本語）	
22#6#12#41#0#4*3#2*### （ここ␣へ␣行き␣たい␣ん␣です␣が␣．）	
パターン翻訳ルール	
数字列	英語
α#6#12#β#0#4*3#2*### α#12#β#4*3#2*###	α red β . I'd like to β α .
翻訳結果	
here red .	

※ $\alpha, \beta, \ldots$ は変数を表す．

ルである．この翻訳例において意図した日本語文は「私は赤ワインがほしいんですが．」である．ここで，重複する翻訳候補として，（α#12#β#4*3#2*###：I'd like to β α .）が存在していた．このルールは，翻訳例（36*1#649#74*#12#41#5#4*32*###：I'd like to go to the Savoy Hotel .）から獲得された．この翻訳例において意図した日本語文は「サボイホテルまで行きたいのですが．」である．よって，正しく翻訳するためには（α#12#β#4*3#2*###：I'd like to β α .）が選択されるべきであったが，ゆう度評価のために抽象度を比較した結果，抽象度が低い（α#6#12#β#0#4*3#2*###：α red β .）が選択された．そして，単語翻訳ルール（22:here）が α に代入され，β は未翻訳となり「here red .」が出力されて誤翻訳となった．このように，数字列のもつあいまいさによる誤翻訳が存在する．

翻訳ルール（α#6#12#β#0#4*3#2*###：α red β .）は場面「機内」から獲得されたルールであり，現在の対象である「空港」に適合していないと考えられる．しかしながら，本手法のゆう度評価では，抽象度の低い翻訳ルールを優先的に選択するようにしているため，このような誤りが発生した．この問題を解消するためには，適合度を優先的に用いてゆう度評価を行うことが考えられるが，この場合，抽象度が高いルールを選択してしまう可能性があり，翻訳精度の低下が懸念される．よって，抽象度と適合度からなるゆう度評価関数を導入することが考えられる．しかしながら，このようなゆう度評価関数を導入するにあたっては，最適なパラメータを設定する必要があり，これは今後の課題とする．

4.4.4　再現率について

前述のように，網羅文とは，その文を構成するすべての単語が既に出現済みである文を指す．この網羅文数の全体に占める割合は，39.5[%]（514/1,300）であった．そのため，本手法の再現率がこの値を超えることはない．入力データ数の増加に伴い網羅文の割合も増加するので，本手法によって翻訳可能となる文数も増加するものと考えられる．しかしながら，本手法は辞書容量が小さい携帯端末などでの個人使用を前提としているため，データ量の増加による解決は望ましくない．また，理想的には翻訳可能な網羅文であっても，正しい翻訳ルールが獲得されていなかったために翻訳できなかった文が存在していた．

このような問題は，我々が以前に行った研究である，数字漢字変換手法 [9], [10] においては発生頻度が低い問題である．数字漢字変換手法においては，入力された数字列を日本語文に変換している．この入力数字列と日本語文との間には，単語単位で 1 対 1 の対応関係が必ず存在するのに対して，翻訳における原言語文と目的言語文との間には，同様の対応関係は必ずしも存在しない．よって，数字漢字変換手法では，1 セグメント(注2)単位でルールを獲得し変換に利用できるのに対して，本手法では複数単語から構成されるルールを獲得する必要がある．このように，語順を保持する必要がある本手法の翻訳ルールは，同様の単語から構成される文であっても，語順が異なる文には適用不可能であり，また，このような翻訳例からは翻訳ルールの獲得も不可能である．そのため，本手法の翻訳精度は，数字漢字変換手法の変換精度に比べて低い値となっている．

そこで，翻訳可能となる文を増加させるために，抽象度の高い翻訳ルールをより多く抽出することが考えられる．本手法では，**3.5.2** の共通部分を利用した学習において抽出するパターン翻訳ルールを差異部分が 1 箇所のものに制限している．この制限を緩和し，差異部分が 2 箇所以上となる翻訳ルールも抽出することにより，抽象度の高い翻訳ルールをより多く獲得することができる．しかしながら，差異部分が 2 箇所以上の場合に抽出を行うと，2 箇所以上の変数部分をもつ翻訳ルールが獲得されることになる．前述のように，このような翻訳ルールにおいては原言語と目的言語の間で変数部分の対応関係を一意に決定できなくなるので，翻訳の際に他の翻訳ルールを誤った箇所に代入してしまう可能性がある．正しい箇所に代入を行うために，単語 n-gram により隣接する単語列を利用するこ

（注2）：数字漢字変換手法において，システムが自動的に獲得する文字列の単位であり，理想的には単語に相当する．

とが考えられる．単語のつながりを考慮することにより，複数個ある変数部分の中から正しい代入箇所を決定することができる．このようにして単語 n-gram を導入するためには，ゆう度評価との融合を検討する必要がある．また，本手法は小型端末での使用を想定しているので，その処理量の増加についても考慮する必要がある．

4.4.5 原文一致について

原文一致文とは，翻訳対象となる文が既に翻訳例中に存在している文を指す．よって，原文一致文の翻訳は，過去の翻訳例をそのまま適用することにより正しく行うことができる．最も単純に機械翻訳を行うには，このような翻訳例を記憶しておき，それをそのまま利用すればよい．このように翻訳を行った場合の再現率，翻訳精度を表 8 と同様のデータで評価した．その結果，全体の再現率，翻訳精度はそれぞれ，13.1[%], 33.1[%] であった．原文一致文は 1 単語でも異なる文には適用できないので，その適用範囲が非常に限定される．これに対して，本手法では抽象度の異なる様々な翻訳ルールを獲得し適用可能なので，再現率で 16.1 ポイント，翻訳精度で 20.8 ポイントの向上が確認された．

4.4.6 数字のあいまい性について

本手法における入力数字列は，意図した日本語文のほかにも複数の日本語文に対応しており，あいまい性が増大している．しかしながら，実験で用いた旅行者用英会話文のような対象が限定したデータにおいては，分野を限定しない一般的なデータに比べて，数字列のあいまいさが抑制されているものと考えられる．これを確認するために，本手法における入力数字列をモデル化(注3) し，そのあいまいさの評価を行った．評価結果を表 13 に示す．表 13 中の「一般的なデータ」は，EDR 日本語コーパス [25] より 10,561 単語を含む 457 文を無作為に抽出したデータである．本実験データの 1 文当りの平均単語数は 8.12 であったので，これを用いて，本実験データ 1 文当りの多重度 MU_T，及び一般的なデータの多重度 MU_E は以下のようになる．

表 13　単語に対する数字列のあいまいさ
Table 13　Ambiguity of number for words.

	単語数（文数）	1 単語当りの平均エントロピー
本実験データ	10,561　(1,300)	0.24
一般的なデータ	10,561　(457)	0.28

$$MU_T = 2^{0.24\times 8.12} = 2^{1.97} = 3.9$$
$$MU_E = 2^{0.28\times 8.12} = 2^{2.27} = 4.8$$

すなわち，8 単語程度で構成される数字列に対して，旅行者用英会話文においては 4 通り程度の解釈が存在し，EDR コーパスにおいては 5 通り程度の解釈が存在することを表している．このように，一般的なデータに比べて，本実験データのように対象分野を限定したデータにおいては，数字列のあいまいさが抑制されているのが確認できる．本手法においては，このように限定した対象に対して動的に適応するというアプローチをとっている．なお，本手法は学習能力を有しているので，対象が変化しても，その対象の変化に追従することができる．以上のようにして本手法は数字列のもつあいまいさを解消し，高い翻訳精度を実現している．

5. 他手法との比較

本手法は携帯電話等の小型端末を想定し，迅速な入力が可能な機械翻訳手法を目指している．小型端末での迅速な入力にこだわらず翻訳精度の向上のみを目指すのであれば，打鍵数，あるいは端末の携帯性を犠牲にして，入力時により多くの情報を与えればよい．すなわち，打鍵数を犠牲にすることにより，現在の携帯電話等で採用されている文字循環指定方式を用いて仮名を明示的に指定することができ，また，端末の携帯性を犠牲にすることにより，パソコン等と同様の大きなキーボードを用いてローマ字入力を行うことができる．一般的な日英機械翻訳手法では，翻訳対象が漢字仮名混じり文なので，入力された仮名文字列を漢字に変換する必要がある．仮名漢字変換を行うためには，変換候補を表示し，意図した変換結果を選択する必要がある．この変換，確定の作業にはそれぞれにキーが必要であり，また，それぞれに最低でも 1 打の打鍵数が必要となる．現在の携帯電話では，多くの場合，1 文節単位で仮名漢字変換が行われている．よって，仮名漢字変換に要する最小の打鍵数は「文節数× 2」で表すことができる．実際の変換単位は使用者に依存するものと考えられるが，ここでは文節数を自立語の数により近似し，仮名漢字変換に要する打鍵数は「自立語数× 2」で表すこととする．これに対して，本手法では入力数字列を直接翻訳するので，仮名漢字変換の

(注3)：数字列のモデル化の詳細については，付録に示す．

表 14 翻訳効率
Table 14 Translation efficiency.

	本手法	文字循環指定	ローマ字入力
入力文字数	33,944	24,683	24,683
仮名入力打鍵数	-	65,499	36,144
仮名漢字変換打鍵数	-	10,334	10,334
入力打鍵数	33,944	75,833	46,478
キー数	12	14	30
入力コスト ($\times 10^3$)	407	1,062	1,394
翻訳正解率 [%]	29.2	70.5	70.5
翻訳効率 ($\times 10^{-6}$)	71.7	66.4	50.6

ための打鍵数は必要ない．しかしながら，本手法においては単語分割記号を入力する必要があるので，その分だけ入力文字数は多くなる．また，ローマ字入力のように入力に必要なキー数が増えると，その分，入力操作が煩雑になり，入力に要するコストが増大するものと考えられる．これを考慮し，入力コストを式 (5) で定義する．

$$\text{入力コスト} = \text{入力打鍵数} \times \text{キー数} \tag{5}$$

少数のキーを用いて少ない打鍵数で入力が完了する場合に，この入力コストは低い値になる．また，入力コスト当りの翻訳正解率を翻訳効率と定義し，これを式 (6) に示す．

$$\text{翻訳効率} = \frac{\text{翻訳正解率}}{\text{入力コスト}} \tag{6}$$

各手法の比較のために，翻訳効率により評価を行った．比較対象とする手法は，文字循環指定方式，及びローマ字入力による一般的な機械翻訳手法である．本手法において，入力のために必要なキーは 12 個のみである．文字循環指定方式においては，仮名を入力するための 12 個のキーに加えて，仮名漢字変換を行うために 2 個のキーが必要となるので，合計で 14 個のキーが必要となる．ローマ字入力においては，ローマ字のための 26 個のキーと句読点のための 2 個のキー，更に仮名漢字変換を行うための 2 個のキーも必要となるので，合計で 30 個のキーが必要となる．評価データとして，**4.** の実験と同様のデータを用いた．なお，一般的な機械翻訳手法の翻訳正解率は，文献 [12] より 70.5[%] とした．これに合わせて本手法の翻訳正解率として再現率 29.2[%] を用いた．評価結果を表 14 に示す．表 14 に示されるように，本手法の翻訳効率が最も高い値となっている．このように，本手法が携帯電話のような小型端末上で，迅速な入力と比較的高い翻訳精度を両立していることが確認された．

6. むすび

本論文では，帰納的学習を用いた携帯端末向け機械翻訳手法を提案した．海外旅行等での会話文の翻訳においては，高い即時性が要求される．機械翻訳システムの入力方法が煩雑で，入力に要する時間が増大すると，翻訳処理全体としての処理速度が低下することになる．よって，翻訳処理全体の高速化を実現するためには，迅速な入力方法が必須となる．本手法においては，携帯電話のような小型の端末において迅速な入力を可能とするため，入力には文字情報縮退方式を採用した．そのため，本手法での入力文字列である数字列は，使用者が意図した日本語文以外にも複数の日本語文に対応し，結果としてこれに対応する翻訳候補も複数存在することになり，あいまいさが増大している．しかしながら，このあいまいさは帰納的学習のもつ高い適応能力により解消することが可能である．すなわち，人手により事前に与えられた単語翻訳ルールと字面上の共通部分を利用して，抽象度の異なる様々な翻訳ルールを動的に獲得し，更に，翻訳結果に基づき翻訳ルールのゆう度を更新することにより，現在の対象に次第に適応する．このような適応能力により，評価実験の結果，約 70[%] の翻訳精度と高い翻訳効率が得られ，本手法の有効性が確認された．

今後は，翻訳精度を更に向上させるために，単語 n-gram やゆう度評価関数の導入について，検討を進める予定である．

文　献

[1] 阪田史郎，“モバイルインターネットの展望，” 情報処理，vol.42, no.12, pp.1193–1197, Dec. 2001.
[2] 西村雅史，伊東伸泰，山崎一孝，“単語を認識単位とした日本語の大語彙連続音声認識，” 情処学論，vol.40, no.4, pp.1395–1403, April 1999.
[3] 伊田政樹，森　弘之，中村　哲，鹿野清宏，“据置き型情報提供端末向き雑音処理を用いた音声入力インタフェース，” 信学論（D-II），vol.J84-D-II, no.6, pp.868–876, June 2001.
[4] 大渕康成，北原義典，小泉敦子，松田純一，畑岡信夫，“マイコン向け音声認識技術を用いた携帯型音声通訳機，” 信学論（D-II），vol.J83-D-II, no.11, pp.2309–2317, Nov. 2001.
[5] 木村義政，小高和己，鈴木　章，佐野睦夫，“携帯型ペン入力インタフェース用個人辞書の学習，” 信学論（D-II），vol.J84-D-II, no.3, pp.509–518, March 2001.
[6] 和泉勇治，加藤　寧，根元義章，“2 乗結合を持つ多層パーセプトロンによる手書き文字の高精度認識，” 信学論（D-II），vol.J83-D-II, no.10, pp.1969–1976, Oct. 2000.
[7] 入鹿山剛堂，“ケータイ文字入力の現状と将来，” 信学誌，

vol.84, no.11, pp.819–827, Nov. 2001.
[8] 佐藤　亨，東田正信，林　智定，奥　雅博，村上仁一，“PB 電話機を利用した日本語入力方式，” 1997 信学総大，D-6-6, March 1997.
[9] M. Matsuhara, K. Araki, Y. Momouchi, and K. Tochinai, “Evaluation of number-Kanji translation method of non-segmented Japanese sentences using inductive learning with degenerated input,” Lecture Note in Artificial Intelligence 1747 (AI’99), pp.474–475, Sydney, Australia, Dec. 1999.
[10] 松原雅文，荒木健治，桃内佳雄，栃内香次，“文字情報縮退方式を用いた帰納的学習によるべた書き文の数字漢字変換手法の有効性について，” 信学論（D-II），vol.J83-D-II, no.2, pp.690–702, Feb. 2000.
[11] P.F. Brown, J. Cocke, S.A.D. Pietra, V.J.D. Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin, “A statistical approach to machine translation,” Computational Linguistics, vol.16, no.2, pp.79–85, June 1990.
[12] 古瀬　蔵，隅田英一郎，飯田　仁，“経験的知識を活用する変換主導型機械翻訳，” 情処学論，vol.35, no.3, pp.414–425, March 1994.
[13] 佐藤理史，“実例に基づく翻訳，” 情報処理，vol.33, no.6, pp.673–681, June 1992.
[14] 松原雅文，荒木健治，栃内香次，“帰納的学習を用いた機械翻訳手法における数字表現の利用方法について，” 情処学 NL 研報，2001-NL-146, pp.47–52, Nov. 2001.
[15] 荒木庸子，Joanna C. Lee, 旅行英会話ポケットブック，日本文芸社，東京，1995.
[16] 旅行会話研究会，海外旅行英会話，実業之日本社，東京，1980.
[17] K.S. Gilbert, ケントのトラベル英会話，実業之日本社，東京，1995.
[18] 石川洋一，トラベル・コミュニケーション研究会，ひとり旅これで十分英会話，実業之日本社，東京，1995.
[19] 前川　裕，アメリカを自由に歩く旅の米会話，池田書店，東京，1994.
[20] William Reed, 困った時のトラベル英会話入門，日本文芸社，東京，1995.
[21] ブックメーカー，海外旅行かんたん英会話ハンドブック，池田書店，東京，1996.
[22] 甲斐順子，ひとり歩きの英語自遊自在，日本交通公社出版事業局，東京，1991.
[23] 地球の歩き方編集室，旅の会話集 2 米語/英語，ダイヤモンド社，東京，1993.
[24] 斉藤晃雄，六ヶ国語会話 1 ヨーロッパ・アメリカ編，日本交通公社出版事業局，東京，1960.
[25] 日本電子化辞書研究所，“EDR 電子化辞書使用説明書，” 1995.

付　　録

数字列のモデル化

本手法では，日本語文を数字列により表している．以下のように，M 単語から構成される入力数字列 N を考える．

$$N = N_1 \sqcup N_2 \sqcup \cdots \sqcup N_m \sqcup \cdots \sqcup N_M$$

N_m は日本語単語に対応する数字列を表しており，スペース（$\sqcup$）により単語分割されている．数字列は日本語仮名文字列の母音情報が縮退したものであり，一つの数字列に対応する単語は一般に複数存在する．この数字列 N_m に対応する単語候補を W_m で表し，この単語集合を以下のように表す．

$$N_m = \{W_m^1, W_m^2, \ldots, W_m^L\}$$

数字列 N_m に対して，L 個の単語候補が存在している．単語候補の発生確率に従うエントロピー H を用いて，単語多重度（PerPlexity）PP を表すと，式 (A·1) のようになる．

$$PP(N_m) = 2^{H(N_m)} \tag{A·1}$$

$$H(N_m) = -\sum_{l=1}^{L} P(W_m^l|N_m) \log_2 P(W_m^l|N_m) \tag{A·2}$$

よって，入力数字列 N の多重度（MUltiplicity）MU を PP を用いて式 (A·3) で表すことができる．

$$\begin{aligned} MU_N &= PP(N_1)PP(N_2)\cdots PP(N_M) \\ &= 2^{H(N_1)+H(N_2)+\cdots+H(N_M)} \\ &= 2^{\overline{H}(N)\times M} \end{aligned} \tag{A·3}$$

以上のように，M 単語から構成される数字列の多重度は，1 単語当りの平均エントロピー $\overline{H}$ より求めることができる．

なお，この 1 単語当りの平均エントロピー $\overline{H}$ は，評価データ K 単語に対して，式 (A·4) で評価している．

$$\overline{H} = \frac{1}{K}\sum_{k=1}^{K} H(N_k) \tag{A·4}$$

（平成 14 年 3 月 20 日受付，7 月 26 日再受付）

松原　雅文　（学生員）

平 12 北海学園大大学院工学研究科電子情報工学専攻修士課程了．現在，北大大学院工学研究科電子情報工学専攻博士後期課程在学中．自然言語処理の研究に興味をもつ．情報処理学会会員．

荒木　健治　（正員）

昭 57 北大・工・電子卒．昭 63 同大大学院博士課程了．工博．同年，北海学園大学工学部電子情報工学科助手．平元同講師．平 3 同助教授．平 10 同教授．平 10 北大・工・電子情報工学専攻助教授．平 14 同教授，現在に至る．自然言語の機械学習と機械翻訳に関する研究に従事．情報処理学会，言語処理学会，日本認知科学会，人工知能学会，ACL, IEEE, AAAI 各会員．

栃内　香次　（正員）

昭 37 北大・工・電気卒．昭 39 同大大学院工学研究科電気工学専攻修士課程了．現在，北海学園大学大学院経営学研究科教授．主として音声情報処理，自然言語処理の研究に従事．工博．情報処理学会，日本音響学会各会員．