E-012

# 携帯端末向け日本語入力手法における類似度を考慮した単語変換精度について
# Evaluation of Word Translation Accuracy on Japanese Input Method for Mobile Terminal

鎌田 竜也† 松原 雅文† Goutam Chakrabortyt 馬淵 浩司†
Tatsuya Kamada Masafumi Matsuhara Goutam Chakraborty Hiroshi Mabuchi

## 1. はじめに

近年，携帯電話の普及による利用者の増加に伴い携帯端末上で日本語の文章を入力する機会が増加している．しかしながら，携帯電話等の小型端末においては，その端末自身の大きさの制約上，フルキーボードのような多くのキーを備えることができない．そのため，現在の携帯電話における一般的な日本語入力手法においては，このキーの少なさを補うために多くの打鍵数が必要となり，迅速な入力が困難である．

この問題を解決する日本語入力方式として文字情報縮退方式がある．この方式は表 1 のように 1 つのキーに複数の文字を割り当て 1 回の打鍵で 1 つの文字入力を完了させるものである．例えば，「携帯（ケイタイ）」という文字列を入力する場合，一般的な方式では，4 + 2 + 1 + 2 = 9 回の打鍵数が必要であるのに対し，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，入力された数字列は，入力者が意図した仮名文字列以外にも対応することとなり，結果的に多数の日本語文が存在し，非常に多くの曖昧さを持つ．

その曖昧さを解消し単語を変換する手法として本研究ではニューラルネットワークを用いる．個人使用の多い携帯端末においては，利用者ごとに変換する単語に偏りがあると考え，その変換の偏りを学習に反映させる．また，一般的なかな漢字変換においては，日本語の係り受け関係等は変換精度の向上に有効であると考えられるが，この手法における入力は数字列であり，一般的な規則を適用することは困難である．そこで，ニューラルネットワークによって変換結果の決定に必要な情報を自動的に学習し，これを用いて変換を行うことで，変換精度の向上を目指している．さらに，データ間における類似度を算出し，その値を学習に反映させることで，更なる変換精度の向上を目指している．

表 1: キー割り当て

| | | |
|---|---|---|
| 1<br>あいうえおー | 2<br>かきくけこ | 3<br>さしすせそ |
| 4<br>たちつてと | 5<br>なにぬねの | 6<br>はひふへほ |
| 7<br>まみむめも | 8<br>やゆよゃゅょ | 9<br>らりるれろ |
| *<br>（半）濁点 | 0<br>わをん | #<br>句読点 |

†岩手県立大学ソフトウェア情報学部

## 2. システム概要

本システムは，使用者が文字情報縮退方式によって入力した数字列を，類似度を考慮してニューラルネットワークに学習させ，変換を行い，変換結果を出力するものである．ニューラルネットワークへの入力変数としては，変換対象となる数字列の前後 10 文字中に含まれる各数字列の出現頻度と，変換対象数字列を用いることとした．日本語においては，形容詞や副詞等の，意味に影響を与えない語順の入れ代わりが発生することを考えて，この影響を受けないように各数字列の出現位置は与えずに，出現頻度のみを与えることとする．入力ノード数は 44 で，このうち各数字列の出現頻度が 12 ノード，変換対象文字列が 1 文字 4 ノード，最大文字数を 8 文字として 32 ノードである．

出力としては，文字コードを出力するものとした．本システムにおいては，単語の文字に相当するそれぞれの文字コードを 2 進数で出力することとする．ニューラルネットワークにより出力された値を閾値により，“ 0”，“ 1”に変換し，出力結果とする．出力される単語の最大文字数は 2 文字としており，1 文字を 16 ビットで表現しているので出力層のノード数は 32 となる．本手法のニューラルネットワークを図 1 に示す．なお，中間ノード数は 64 とした．

本システムは，ニューラルネットワークに対する効率的な学習を行うために，データ間の類似度を算出し，類似度の高いデータを優先的に学習する．ここで類似度は，2 データ間の cos の値を用いる．2 データ間の出現頻度と変換対象文字列のデータ，あわせて 44 次元のベクトルの cos の値を算出している．以下に，式を示す．

$$\cos\theta = \frac{X \cdot Y}{|X||Y|}$$

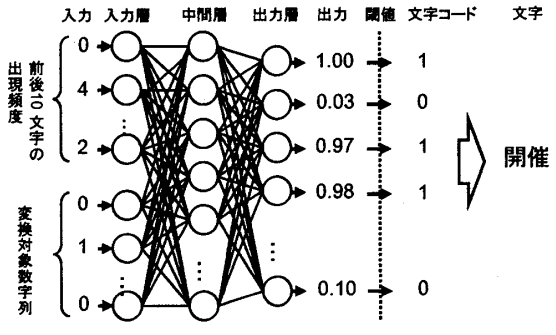


図 1 : ニューラルネットワーク

表 2 : 変換候補

| 数字列 | 単語 |
|---|---|
| 2131 | 開催, 海水, 回数, 改正, 空想 |
| 3031 | 震災, 浸水, 真相, 先生, 戦争 |
| 3132 | 催促, 推測, 成績, 砕石, 正式 |
| 41381 | 大衆, 対象, 対照, 通称, 提唱 |
| 6020 | 反感, 貧困, 変換, 返還, 偏見 |

表 3: テストデータ

| | 数字列 | 単語 | A との類似度 |
|---|---|---|---|
| A | 2131 | 開催 | 1 |
| B | 2131 | 開催 | 0.85 |
| C | 6020 | 返還 | 0.37 |

## 3. 評価実験

データ間の類似度が高いと，ニューラルネットワーク内部で効率的にネットワークが構成されることを確認するために，類似度を考慮した 5 種の数字列を対象とした評価実験を行った．

### 3.1 実験データ

実験データとして，“ 2131”，“ 3031”，“ 3132”，“ 41381”，“ 6020” を含んだ日本語文を用意した．変換候補の一覧を表 2 に示す．それぞれの数字列に対して 5 種類の変換候補が存在し，それぞれの変換候補に対して 20 文のデータを用意した．合計で 500 文のデータで実験を行う．実験に用いた文章は web 上から無作為に選んできたものとする．

### 3.2 実験手順

500 データの中から 1 つ選び，選んだデータと残りのデータの類似度を計算する．このとき，選んだデータをデータ A とする．類似度の高い上位 20 データに対して，学習を強化するために学習時に他のデータより 10 倍多く学習させる．テストデータとして，データ A と同一の変換候補を持つデータ B，データ A と異なる変換候補を持つデータ C を 500 データ中から無作為に選ぶ．テストデータについて，表 3 に示す．残りの 497 データを学習データとしてニューラルネットワークに学習させる．類似度を考慮しない場合の実験を同様の学習データ，テストデータを用いて行い比較し，評価を行う．学習回数を 5000 回，学習率を 0.01 とした．

### 3.3 実験結果

実験の結果を表 4 に示す．データ A，データ B はそれぞれ正解数が最大に達し，単語変換に成功した．反対にデータ C は，正解数が 30 から 16 に低下した．

表 4 : 実験結果

| | 類似度無し | 類似度有り |
|---|---|---|
| A | 28 / 32 | 32 / 32 |
| B | 28 / 32 | 32 / 32 |
| C | 30 / 32 | 16 / 32 |

## 4. 考察

データ A，データ B の変換結果が向上し，データ C の変換結果が低下した．これは，学習を強化した類似度の上位 20 データの中に，データ A，データ B と同じ変換候補のデータが 7 つ含まれていたため，データ A に特化したニューラルネットワークが構成されたのではないかと考える．

## 5. まとめ

本稿では携帯端末向け日本語入力手法における類似度を考慮した単語変換精度について検討した．類似度の高いデータに対して学習を強化することでニューラルネットワークの構築が効率よく行われ，単語の変換精度が向上するという考えに基づき，ニューラルネットワークに学習させ，類似度に基づいて単語の変換精度を比較した．実験の結果から，類似度を考慮しない場合に比べてより良い結果を得ることができ，本手法の有効性が示唆された．

しかしながら，単語変換システムの変換精度としては物足りないと考えるので，更なる変換精度向上を目指すものとする．

## 参考文献

[1] 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”, 電子情報通信学会論文誌 D-II , Vol.J83-D-II , No.2 , pp.690-702, February 2000.

[2] 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”,電子情報通信学会総合大会公演論文集,D-6-6, p.102, March 1997.

[3] 吉冨 康成,ニューラルネットワーク,朝倉書店,東京,2002.

[4] 北 研二,確率的言語モデル,東京大学出版会,東京,1999.

[5] 鎌田 竜也,松原 雅文,Goutam Chakraborty,馬淵 浩司, “ニューラルネットワークを用いた携帯端末向け日本語入力手法における単語変換精度”, 情報処理学会第 67 回全国大会講演論文集, 2J-4, pp.83-84, March 2005.