

## 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

鎌田竜也 松原雅文 Goutam Chakraborty 馬淵浩司  
岩手県立大学 ソフトウェア情報学部

### 1 はじめに

近年、携帯電話の高性能化、電子メール利用者数の急増により、携帯電話上で日本語文を入力する機会と必要性が増大している。しかしながら、携帯電話等の小型端末においては、端末自身の大きさの制約から通常のフルキーボードほど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方式においては、このキー数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という問題点がある。

この問題点を解消し、迅速な入力を可能とする手法として「文字情報縮退方式を用いた帰納的学習による数字漢字変換手法」[1]が提案されている。この手法は、文字情報縮退方式[2]により入力された数字列を、「日本語文である」漢字かな混じり文に変換するものである。入力にかなの母音情報が縮退した数字列を用いることにより、迅速な入力が可能となる反面、入力文字列は非常に多くの曖昧さを含んでいることになる。たとえば、「2131」という数字列は「かいさい」や「こうせい」等、約  $5^4 = 625$  種類のかな文字列に対応することになる。

この曖昧さを解消するために、この手法においては、帰納的学習による高い適応能力を利用して、学習、変換を行っている。実験の結果から、個人の送信メールのような、対象が頻繁に変化するデータにおいても、約 80[%] の変換精度が確認されている [1]。

しかし、実用的には、さらなる変換精度の向上が望まれる。日本語における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文字列は数字列であり、一般的な規則を適用することは困難である。そこで、本手法では、ニューラルネットワーク [3] を利用することとした。ニューラルネットワークにより、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により、変換精度の向上を目指している。

本稿では、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について述べる。

### 2 処理概要

文字情報縮退方式により入力された数字列は本手法により使用者が意図した日本語文に変換される。上述したように、例えば数字列「2131」に対応する日本語文字列には多くのものが存在する。しかし、使用者が意図した日本語文字列は一つであり、これをシステムにより自動的に決定する必要がある。

変換候補は、前後の単語との係り受け関係等を利用することにより、ある程度絞り込むことが可能であると考えられる。例えば、後続する文字列が「する」という動詞であることが分かっている場合には、これに係ることが可能な変換候補は「開催」や「構成」等となり、ある程度限定できる。しかし、本手法における入力数字列からこのような形態素、構文情報を獲得することは、その曖昧さのために困難である。そこで、本手法においては、前後の数字列からこの係り受け関係に相当する情報を獲得するためにニューラルネットワークを用いることとした。ニューラルネットワークを利用することにより、前後の数字列から変換に有効な部分を自動的に学習することができるものと考えられる。

### 3 ニューラルネットワーク

本手法の単語変換システムにニューラルネットワークを適用した。これを図1に示す。また、この学習方法としては誤差逆伝搬法を用いた。

ニューラルネットワークへの入力変数には、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。日本語文においては、形容詞や副詞等、意味に影響を与えない語順の入れ替わりが発生することが考えられるので、この影響を受けないよう

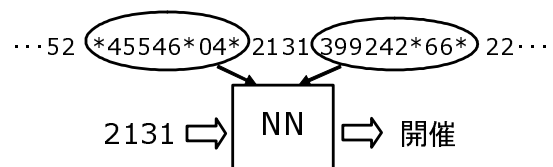


図 1: 単語変換の例

Evaluation of Word Translation Accuracy on Japanese Input Method for Mobile Terminal Using Neural Network  
Tatsuya Kamada, Masafumi Matsuhara, Goutam Chakraborty and Hiroshi Mabuchi  
Faculty of Software and Information Science, Iwate Prefectural University

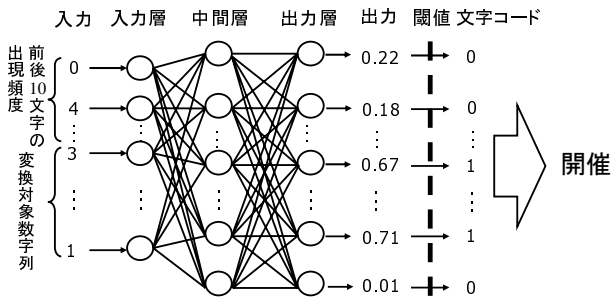


図 2: ニューラルネットワーク

に各数字の出現位置は与えず，出現頻度のみを与えることとする．変換対象数字列の文字数は最大で8文字とし，1文字を4ビットで表すものとする．よって，入力層のノードは，前後10文字中に含まれる各数字列の出現頻度のために12個，変換対象数字列のために32個，合計で44個となる．

本システムにおいては，単語の文字に対応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ため，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2に示す．なお，中間層のノード数は64とした．

また，学習の際に用いる教師データに関しても同様に，正解となる単語に含まれるそれぞれの文字を文字コードとして与えている．

## 4 評価実験

本手法における単語変換精度を確認するために，5種類の数字列を対象とした評価実験を行った．

### 4.1 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1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．すなわち，合計で500文のデータを用意した．

### 4.2 実験手順

学習は500文全てを用いて行う．これらの文を数字列に変換してシステムに入力する．出力された文字コードに対して，教師データを文字コードとして与え，ニューラルネットワークの学習を行う．そして，学習に用いたデータ中から50文を無作為に抽出し，これをテストデータとした．このテストデータに対して対象となる単語の変換を行い，評価を行った．

表 1: 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，碎石，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 2: 実験結果

	正解数	正解率 (%)
ノード単位	1,557 / 1,600	97.3
単語単位	31 / 50	62.0

## 4.3 実験結果

実験の結果を表2に示す．出力されたノード単位での正解率は約97[%]となった．また，出力された文字コードに対応する単語としての正解率は62[%]となった．

## 5 まとめ

本稿では，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について検討した．曖昧な入力数字列の変換候補としては多数の日本語文字列が存在するが，前後の文字列から変換候補の限定が可能であるという考えに基づき，本手法においてはこれをニューラルネットワークにより学習させることとした．クローズドデータを用いて実験を行った結果，ノード単位で約97[%]の正解率を得ることができ，本手法の有効性が示唆された．

なお，実際の使用においては，クローズドデータのようにまったく同一の文を入力することはあまりないものと考えられるので，今後は，オープンデータを用いた実験を行い，本手法の有効性を確認する予定である．

## 参考文献

- [1] 松原 雅文, 荒木 健治, 桃内 佳雄, 柝内 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, “P B 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講演論文集,D-6-6, p.102, March 1997.
- [3] 吉富 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.