E-042

符号誤り訂正を用いた携帯端末向け日本語入力手法の有効性について

Effectiveness for Fast Japanese Input Method Using Bit Error Correction on Mobile Terminal

鈴木 悟史† 松原 雅文† Goutam Chakraborty† 馬淵 浩司†
Satoshi Suzuki Masafumi Matsuhara Goutam Chakraborty Hiroshi Mabuchi

1. はじめに

現在，日本において携帯電話は，通話機能だけではなく電子メールや Web ページの閲覧をはじめとする多機能な端末として用いられ，携帯電話上での日本語入力の機会と必要性が増大している．しかし携帯端末は，それ自身が小型であることを求められ，通常のフルキーボードよりはるかに少ないキー数となることは必至である．このキー数の少なさを補うために，現在主流となっている文字入力方式では，1 文字の入力に多くの打鍵数を必要としており，迅速な入力が困難であるという問題を抱えている．

この問題を解決するために,「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」[1] が提案されている．この提案手法では，文字情報縮退方式という入力方式を利用し，変換の際に問題となる曖昧性をニューラルネットワークで解決しようとした．この提案手法の実験結果は，有効性を示唆するものであったが，入力が単語単位という制約や，ニューラルネットワークの規模が定まらないという問題があり，実用化は困難であると考えられる．

そこで，本研究では入力を固定長で分割した数字列とすることで，ニューラルネットワークを利用しながらも，先行研究の問題点を解消し，より実用に適した，携帯端末向け日本語変換手法を提案する．本稿では，ニューラルネットワークによって変換された結果に対し，符号誤り訂正を行うことで，変換精度の向上を目指しており，その有効性について検証を行った．

2. 先行研究

2.1 文字情報縮退方式

文字情報縮退方式のキー配置は，現在主流である文字循環指定方式と同じである．このため，利用者は新たにキー配置を覚え直す必要が無い．

具体的な入力方法を「大会（たいかい）」という例を使い説明する.「たいかい」を文字情報縮退方式により入力する場合，4121 の順にボタンを打鍵する．そして入力された数字列を「大会」に変換する．このように，入力したい仮名が含まれる数字を 1 回，または濁点半濁点の場合は 2 回打鍵するだけで，1 文字の入力が可能である．よって従来の文字入力方式に比べ迅速な入力が可能となる．

しかし，母音情報が縮退されているため入力数字列が非常に多くの曖昧さを含むという問題がある.「たいかい」を表わす 4121 は，他に「とうけい」や「つうこう」など $5^4 = 625$ 種類の仮名文字列に対応している．また，日本語は漢字への変換も必要となるためさらに多くの量の候補が存在する．このため，縮退された数字列を日本語に変換する手法 (以下，数字漢字変換手法)[2] が，文字情報縮退方式を用いる上で重要となる．

2.2 ニューラルネットワークを用いた変換手法

曖昧さを解決するために，先行研究ではニューラルネットワークを利用し，その高度な学習機能を活用することで数字漢字変換を行った．誤差逆伝播法のニューラルネットワークで学習を行い，入力に単語の数字列と，その数字列の前後に入力された数字の出現頻度を与えることで，日本語の文字コードが出力される仕組みである．

実験はクローズドデータで行われ，その結果，ノード単位では 97.3[%]，単語単位では 62.0[%] の正解率となり，有効性が示唆されたが，入力される数字列が単語単位であるという前提のもとに実験が行われていた．しかし，実際の入力においては必ずしも単語単位で区切って入力を行うとは限らない．よって，どのような長さ，区切りの数字列にも対応することが可能な変換手法が必要である．

3. 提案手法

3.1 処理の概要

本手法における数字漢字変換の流れを図 1 に示し，図 2 を例に挙げて説明する．まず，はじめに入力数字列を受け取り，固定長による分割処理を行う．分割処理は，はじめに入力数字列の先頭から固定長のサイズ分数字を取り出し，次に入力数字列の 2 番目の数字を先頭とし，固定長のサイズ分数字を取り出す．以降，3 番目，4 番目…末尾まで各数字を先頭としながら固定長で数字を取り出していくことで入力数字列を分割する．例では入力数字列である 41210213139 を 6 数字の固定長で分割している．

分割が終わったらニューラルネットワークによる変換を行う．ニューラルネットワークの入力に対して，分割された数字列と，その数字列の前に出現した 4 数字を係り受け情報として与える．その入力を使いニューラルネットワークから「日本語の文字コード」(以下，日本語コード) または「文字ではないことを示すコード」(以下，非文字コード）のいずれか一方の出力を受け取る．この受け取る際に，符号誤り訂正を行うことで変換精度を高める．例では日本語コードの「大会」,「を」,「開催」,「する」と，非文字コードの「FFFF」にそれぞれ変換している．

最後にニューラルネットワークの出力結果である日本語文字コードと非文字コードを使い，合成を行うことで，入力された数字列の変換を完了とする．例では変換結果から文章を合成し「大会を開催する」を出力している．

このように，前処理として固定長で分割することで，どのような長さ，区切りで入力された数字列にも対応することが可能となる．また，固定長のサイズによって入力ノードの数を決めることが出来るため，ニューラルネットワークの規模を自由に定めることができる．

†岩手県立大学, Iwate Prefectural University

図 1: 本手法の変換プロセス

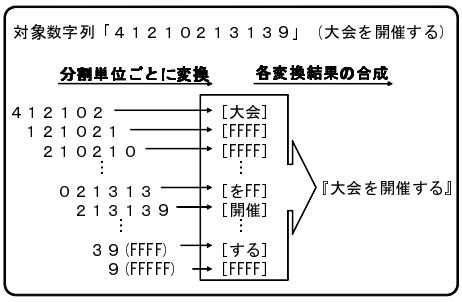

図 2: 本手法における数字漢字変換

3.2 符号誤り訂正処理

ニューラルネットワークによる変換の際に，ノード単位での細かな誤りが発生する．例えば「大会」の「大」のコードを 1 ビット誤ってしまい，コード表の通常利用しない領域を指し示してしまうなど，日本語コードまたは非文字コードのいずれでもないものに変換される場合がある．よって，変換結果が日本語コードまたは非文字コードであるか検査を行い，いずれでもない場合は，コード表と 1 文字ずつの結果を照らし合わせ符号誤り訂正処理を行う．今回は次に示す流れで処理を適用した．なお，これはシフト JIS コードの場合についてである．

1. 変換結果がコード表の正しい範囲に含まれるかを検査
2. コードが範囲外を指している文字を 1 つずつ抽出
3. 5 ビット以下の修正で非文字コードとなる場合は，非文字コードとし，2 へ戻る
4. 全角文字の 1 バイト目を，0x81〜0x83,0x88〜0x9f の範囲へ最小のビット操作で修正
5. 全角文字の 2 バイト目を，0x40〜0xfc の範囲へ最小のビット操作で修正，2 に戻る

4. 評価実験

4.1 実験データおよび実験方法

今回は符号誤り訂正処理の有効性を検証するために，合成処理を除いた，符号誤り訂正処理までの評価実験を，オープンデータを用いて行った．学習データには日本語の論文 [3] である，21,576 字分のテキストデータを用いた．このテキストデータを茶筌で形態素解析し，入力数字列と日本語コードを作成した．そして，作成した入力数字列を固定長サイズ 9 で分割したものに，分割した数字列の前方に出現した 4 数字と，正解データを付加し，1 つの学習データとした．今回は，32,433 個の学習データが生成され，このうち 28,999 個を学習に使い，残りの 3,434 個で変換を行った．実験に使ったニューラルネットワークは，入力ノード数 52 個，中間ノード数 144 個，出力ノード数 144 個で構成しており，学習モデルには誤差逆伝播法を用いた．

表 1: 分割単位の変換正解率

	日本語コード	非文字コード	総計
訂正処理無	14.4[%]	81.4[%]	56.3[%]
訂正処理有	28.5[%]	96.6[%]	69.0[%]

4.2 実験結果および考察

実験結果を表 1 に示す．1 つの分割数字列に対して完全な形で変換に成功したことを表わす，分割単位での正解率は日本語コードが 28.5[%]，非文字コードが 96.6[%]，総計が 69.0[%] であった．以前行った符号誤り訂正処理が無い場合の評価実験と比べると，日本語コードは 14.1 ポイント，非文字コードは 15.2 ポイント，総計では 12.7 ポイント変換精度が向上した．よって，この実験結果から本手法の有効性を示すことができた．

特に，非文字コードへの符号誤り訂正処理の効果が顕著だった．これに対して，日本語コードの変換精度が向上した理由としては，ニューラルネットワークのパラメーター調整の影響が強く，符号誤り訂正処理の効果はそれほど顕著ではなかった．しかし，日本語コードは形態素であるのにも関わらず，単語と記号が混じる，例えば「パターン」を「パタ／ン」としている場合がある．こういった「／」のような，規則性と照らし合わせて不自然な箇所に対し，今回のような符号訂正を適用することで，さらに精度を向上させることが可能であると考えられる．

5. おわりに

本稿では，文字情報縮退方式を実現するために必要な変換手法について，ニューラルネットワークを用いる手法を提案し，符号誤り訂正処理を加えることでより一層の精度向上を目指した．評価実験では，符号誤り訂正処理を加える前よりも変換精度が向上し，本手法の有効性を示すことが出来た．

しかし，日本語コードへの符号誤り訂正処理はまだ不十分であると考えられる．よって今後は，形態素としての不自然さを判定し，これに対しても，符号誤り訂正処理を加えていくことを検討する予定である．

参考文献

[1] 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 (2005)

[2] 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, J83-D-II, No.2, pp.690-702, (2000)

[3] 川嶋宏彰, 松山隆司: 連続状態モデル間の相互作用に基づく多視点動作認識, 情報通信学会論文誌 D-II, J85-D-II, No.12, pp.1801-1812, (2002)