208 帰納的学習を用いた数字漢字変換手法における繰り返し学習の有効性について

松原雅文 荒木健治 栃内香次
北大工

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，携帯電話の上で日本語を入力する機会と必要性が増大している．しかしながら，携帯電話はそ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，少数のキーを用いて，そのキー数より多くの文字種を入力する必要があるため，入力に要する打鍵数が多くなり迅速な入力は困難である．

これに対して迅速な日本語入力を可能にするために，我々は「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」を提案している[1]．本手法は，文字情報縮退方式[2]により入力された数字列を，漢字かな混じり文に変換するものである．入力に，かなの母音情報が失われた数字列を用いるため，入力文字列は非常に多くの曖昧さを有している．しかし，本手法においては，帰納的学習による高い適応能力により，この曖昧さを排除し，正しい変換が可能となっている[3]．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た．

しかし，実用的には，さらなる変換精度の向上が望まれる．これを実現するために，本稿では，繰り返しによる学習を提案する．同一のデータに対して，本手法の一連の処理を繰り返すことにより，対象に強く適応することができる．そのため，新たな知識をまったく与えることなく，変換精度の向上が可能である．

本稿では，本手法の概要，および評価実験により確認した，繰り返し学習の有効性について述べる．

2 帰納的学習による数字漢字変換手法

本手法の処理過程をFig. 1に示す．変換処理，校正処理，学習処理，フィードバック処理の順である．入力に用いられる数字のかなとの対応関係をTable 1に示す．これにより，少数のキーのみを使うにもかかわらず，迅速な入力が可能である．本手法による変換例をFig. 2に示す．Fig. 2に示されるように，本手法においては，まず，意図する日本語文のかなに対応した数字列を入力する．入力された数字列は，変換処理でセグメント辞書と隣接文字列辞書を用いて漢字かな混じり文に変換される．変換が正しく行われなかった場合，校正処理が行われる．人手により変換結果を訂正する過程である．学習処理では，入力数字列と校正済み変換結果との比較から，語に相当するセグメントを獲得する．同時に数字列，校正済み変換結果の全文字列を隣接文字列辞書に登録する．ここで登録された情報により，隣接する文字列を考慮した変換が可能となっている．フィードバック処理では，正変換，誤変換されたセグメントはその情報をセグメント辞書に持ち，次回からの変換に役立てられる．すなわち，正変換であったセグメントはその尤度を上昇させることにより，次回からの変換時に，より高い優先順位が与えられる．逆に，誤変換であったセグメントはその尤度を下げることにより，次第に使われなくなっていく．

このように，変換処理，学習処理，フィードバック処理を繰り返し，変換精度が向上するのと同時に，対象，または使用者に合わせた辞書が生成されていく．

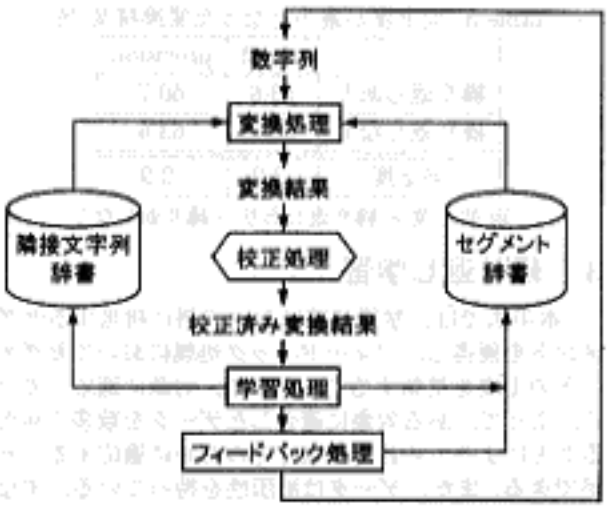

Fig. 1 処理過程

Table 1 数字とかなの対応関係

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	#:句読点

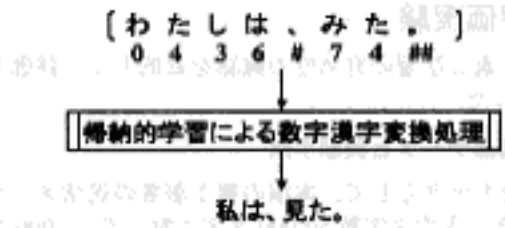

Fig. 2 変換例

Table 2 向上度が最大となった変換精度 [%]

	recall	precision
繰り返しあり	73.3	67.8
繰り返しなし	63.5	54.0
向上度	9.8	13.8

Table 3 向上度が最小となった変換精度 [%]

	recall	precision
繰り返しあり	63.6	60.7
繰り返しなし	66.8	63.6
向上度	-3.2	-2.9

※ 向上度 = 繰り返しあり − 繰り返しなし

3 繰り返し学習

本手法では，学習処理において語に相当するセグメントを獲得し，フィードバック処理においてセグメントの尤度を更新することにより，対象に適応していく．よって，ある対象に適合したデータを数多く与えることにより，それだけ強くその対象に適応することができる．また，データは局所性を持っている，すなわち，最近出現したセグメントは今後も出現する可能性が高いことから，直前のデータに特化した辞書を生成することにより，次回からの変換精度の向上が期待できる．そこで，直前のデータに強く適応することを考える．そのために，直前のデータを繰り返しシステムに与えて学習することにより，現在の対象に特化した辞書を生成する．これにより，新しい知識をまったく与えることなく，変換精度の向上が可能である．

4 評価実験

繰り返し学習の有効性の確認を目的とし，評価実験を行った．

4.1 実験データと実験手順

入力データとして，本稿の第1著者の送信メールを用いた．入力文字数50,000文字に対して，1,000文字単位で学習，変換を行った．すなわち，直前の1,000文字に対して，従来手法どおり繰り返しを行わずに1回のみ学習を行った場合と，本稿で提案する繰り返し学習を行った場合とで生成されたそれぞれの辞書を用いて，次の1,000文字を変換している．変換結果の評価は再現率 (recall) と精度 (precision) により行った．それぞれ式(1)(2)に示す．なお，学習の繰り返し回数は3回とした．

$$recall = \frac{\text{正変換文字数}}{\text{校正済み変換結果文字数}} \quad (1)$$

$$precision = \frac{\text{正変換文字数}}{\text{変換結果文字数}} \quad (2)$$

4.2 実験結果

繰り返し学習の適用により，繰り返して学習を行わない場合に比べて，向上度が最大，最小となった変換精度をTable 2, 3に示す．向上度は，繰り返し学習を行った場合と，行わなかった場合の変換精度の差を表している．よって，繰り返し学習により変換精度が低下した場合，向上度は負となる．向上度が最大となったのは，入力文字数17,000〜18,000のデータを学習データとし，18,000〜19,000文字のデータを変換した場合であり，recall=9.8，precision=13.8ポイントの向上が確認された．逆に，向上度が最小となったのは，入力文字数29,000〜30,000のデータを学習データとし，30,000〜31,000文字のデータを変換した場合であり，recall=3.2，precision=2.9ポイントの低下が確認された．なお，50,000文字の実験全体においては，繰り返し学習を行った場合に平均で，recall=1.0，precision=1.3ポイントの向上が確認された．

4.3 考察

向上度が最大となった場合の学習データ (17,000〜18,000) は，大半が "研究" に関するものであった．変換対象となるデータ中 (18,000〜19,000) にも "研究" に関するデータが数多く含まれていたため，この対象に特化した学習が有効に働いたものと考えられる．向上度が最小となった場合の学習データ (29,000〜30,000) は，大半が "アルバイト" に関するものであった．変換データ (30,000〜31,000) は，大半が "飲み会" に関するものであり，学習データ中にも，同じ "飲み会" に関するデータが含まれていた．しかし，"飲み会" に関するデータは少数であり，繰り返し学習により "アルバイト" に強く適応してしまったために，"飲み会" に対する変換精度が低下したものと考えられる．

5 まとめ

本稿では，帰納的学習を用いた数字漢字変換手法における繰り返し学習の有効性について述べた．同一のデータを繰り返しシステムに与えて学習することにより，現在の対象に強く適応することができる．実験の結果から変換精度の向上が示され，繰り返し学習の有効性が確認された．

今回の実験では，繰り返し学習の有効性の確認を目的とし，繰り返しの対象となる学習データは直前のデータに限定した．しかしながら，学習データは変換データにできる限り適合していることが望ましい．そのためには，変換データに依存して，学習データを切り替える必要があると考えられるが，これは今後の課題とする．

参考文献

[1] 松原，荒木，桃内，栃内：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，東田，林，奥，村上：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，D-6-6，pp.102，March 1997.

[3] 荒木，高橋，桃内，栃内：帰納的学習を用いたべた書き文のかな漢字変換，電子情報通信学会論文誌 (D-II)，Vol.J79-D-II，No.3，pp.391-402，March 1996.