ＯＣＲの認識誤り訂正における学習の効果

7C−7

浜田和彦*　村木一至**　森義和***

*ＮＥＣ技術情報システム開発　**ＮＥＣ　Ｃ＆Ｃ情報研究所　***(株)日本電子化辞書研究所

１．はじめに

日本語ＯＣＲでは、十分な認識正解率が得られず、ＯＣＲ認識結果を自動訂正する装置が考えられてきた。筆者らも、自動訂正する手法として、認識誤り（コンフュージョン）マトリクスや、形態素解析を組み合わせる方法を提案してきた[1]。筆者らの提案する手法では、人出による訂正履歴や、文字パターンをページ単位で学習することで、ＯＣＲの持つ認識誤りの癖を吸収し、一連のテキストでは繰り返し同じ修正をユーザに行なわせないという効果が期待できる。

本稿では、提案する手法における学習の効果を検証するため、試作したシステムにおいて、学習を行なわない場合の認識正解率と、学習を行なった場合の正解率の差を報告する。

２．ＯＣＲ誤り訂正とユーザインタラクション

ＯＣＲが市場に出始めて、パタン認識アルゴリズムの強化だけでなく、シグナルパターン外の対象の特性を用いて認識率を向上させる方法が取り上げられている。

たとえば、日本語の単語辞書を用いた単語分割尤度や字種などの言語的制約を認識選択に用いて行なう方法が報告されている[2-3]。また、文字連接頻度（マルコフ、ダイグラムetc.）、単語出現頻度などの統計量を用いた認識誤り検出や自動訂正の効果も報告されている[4-5]。

これらの研究報告では、良い効果が得られたと報告されているが、いずれにしても最終的には人手での訂正を必要としている。この人手による訂正は、同じ様な訂正を繰り返し行なうような事態がしばしば発生し、その結果、”体感認識率”も下がることになる。

Effect Brought by Learning Function of OCR Error Correction

*HAMADA Kazuhiko, **MURAKI Kazunori, ***MORI Yosikazu

*NEC Scientific Information System Development, Ltd., **NEC Corp.,

***Japan Electronic Dictionary Research Institute, Ltd.

***本成果は森義和がNEC技術情報システム開発在任中に得られたものである

３．ＯＣＲ誤り訂正における学習

現在のＯＣＲ自動訂正システムでは、１００％の正解率を得ることはできず、どうしてもページ単位での訂正が必要である。そこで、ユーザが訂正を終了した時点でのテキストを正解文字列とし、それを教師データとして学習を行なう。

３．１．文字ごとの誤り確率の学習

同一フォントで印刷された一連のテキストでは、全てのページにおいて、同一文字は同一の間違いを犯しやすいと推測できる。そこで、ユーザが訂正した後の文字列と、ＯＣＲが出力した文字列の差を文字単位でとることで、コンフュージョンマトリクス内の誤り確率の調整を行なうことができる。

３．２．文字パターンの学習

もう一つの学習方法は、正解文字列の中の文字パターンを記憶することで行なう。このシステムでは、英数字、カタカナが連続する部分をテーブルに登録する。登録された文字列が次ページ以降において候補文字の列として得られたなら、他の候補より正解の確率は高くなる。例えば、「ＮＥＣ」と「ＮＥｃ」という文字列のつながりとが候補として得られたとき、テーブル中に「ＮＥｃ」が存在すれば、「ＮＥＣ」よりは「ＮＥｃ」の方が確からしそうである。このテーブルを使用することで、ヒューリスティック規則では解決できない文字の訂正を行なうことができる。

４．システムの構成

試作システムの構成を図１に示す。ＯＣＲの認識結果はパターン認識の結果得られた複数の文字候補が、ＯＣＲの中に持つ文字辞書との距離値とともに与えられる。次に、文字コスト計算ではコンフュージョンマトリクス内のデータと、文字パターンテーブルを参照しながら、文字ごとのコスト値を得る。次にユーザによる文字訂正を行なう。その結果をもとに、コンフュージョンマトリクスと文字パターンテーブルを更新する。

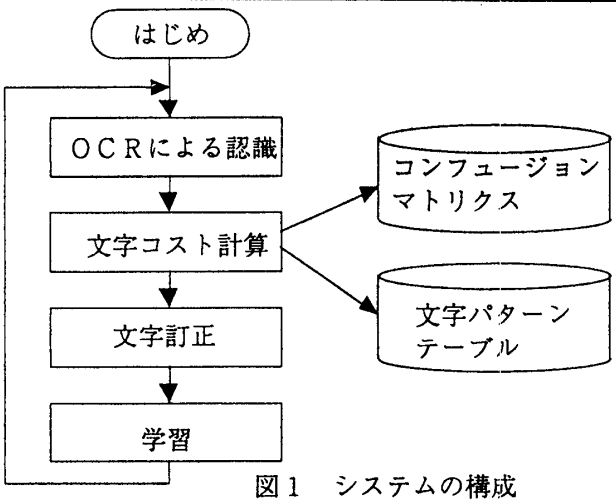

図１　システムの構成

５．実験

学習の効果を明確にするために、以下の３つの方法で実験を行ない、各々に正解率を得て、各々の正解率と学習を行なわなかった場合の正解率との差を求める。

(1) コンフュージョンマトリクスにより与えられた確からしさのみを用いる。

(2) 文字パターンテーブルにより与えられた確からしさのみを用いる。

(3) コンフュージョンマトリクスと、文字パターンテーブルにより与えられた確からを用いる。

実験は１ページ1200から1500文字、約15文程度で、10ページにわたる文書を使用した。正解率の計算は、文字の分離による誤りや、文字の融合による誤りも１つの誤りとして計数して、誤りの合計と正解文字列との百分率をもって正解率とする。ただし、ＯＣＲが原文と無関係なゴミを認識しているものは、正解率計算の対象から除外した。結果を表１に示す。

表１において、1STはＯＣＲの出力結果のそのもので、何の処理も行なわない正解率である。NOはコンフュージョンマトリクスも、文字パターンテーブルの学習も行なわない正解率である。CFMはコフュージョンマトリクスの学習を行なったときの正解率で、CFM-NOは学習を行なわないときとの正解率差である。同様に、文字パターンテーブルの学習した結果(PTN)と学習なしの結果の差(PTN-NO)、コンフュージョンマトリクスと文字パターンテーブル両方学習した結果(ALL)と学習なしの結果の(ALL-NO)を示してある。

表１において、コンフュージョンマトリクスの学習を行なうと、徐々に正解率が向上している。同様に、文字パターンテーブルを使用したときも、ある程度の向上が得られた。さらに、コンフュージョンマトリクス、文字パターンテーブルの両方を用いた場合、最終的には１．３８％の正解率向上が得られた。

６．おわりに

本稿では、ユーザによる同じような訂正の繰り返しを避けるべく、ユーザによる訂正の履歴を学習する場合の効果を計測した。実験の結果により、この方法の有効性を確認することができた。

参考文献

[1] 村木一至、浜田和彦「ＯＣＲの認識誤り訂正に於けるテキスト適合性の評価」電子情報通信学会、言語理解とコミュニケーション研究会予稿集 Oct. 1992

[2] 高尾哲康、西野文人「日本語文書リーダ後処理の実現と評価」情報処理学会論文誌 Vol.30 No.11 Nov. 1989

[3] 伊東伸泰、丸山宏　「ＯＣＲ入力された日本語文の誤り検出と自動訂正」ヒューマンインターフェイス 38-5 Sep. 1991

[4] 相沢輝照、栗田泰市郎「ニュース用語の分類と誤入力訂正の適用」情報処理学会論文誌 Vol.27 No.8 Aug. 1989

[5] 池原悟、白井諭「単語解析プログラムによる日本文誤字の自動検出と二次マルコフモデルによる訂正候補の抽出」情報処理学会論文誌 Vol.25 No.2 Mar. 1984

page	1ST	NO	CFM	CFM-NO	PTN	PTN-NO	ALL	ALL-NO
1	92.21	92.21	92.21	0.00	92.21	0.00	92.21	0.00
2	93.41	95.31	95.11	-0.20	95.31	0.00	95.11	-0.20
3	93.26	94.05	94.12	0.07	94.05	0.00	94.12	0.07
4	88.78	91.40	91.85	0.45	91.43	0.03	91.72	0.07
5	91.87	94.68	94.88	0.20	94.74	0.06	94.88	0.20
6	92.88	95.03	95.06	0.03	95.10	0.07	95.13	0.10
7	89.25	91.98	92.37	0.39	92.05	0.07	92.41	0.43
8	92.42	94.21	94.56	0.35	94.27	0.06	94.63	0.42
9	90.34	93.14	94.12	0.98	93.21	0.07	94.33	1.19
10	89.79	91.03	92.38	1.35	91.07	0.04	92.41	1.38

表１　実験結果