2J－9

文書認識における言語情報の活用 (2)
－認識誤りの自動訂正と指摘について－

斉藤 孝広　小川 知也　松井 くにお
富士通研究所

1　はじめに

文字認識は一文字のイメージデータに対して，パターン辞書中の各文字のデータとの類似度を距離値を用いて評価し，その値の小さいものから順に候補文字として出力する．従って，ある文字列イメージの正しい認識結果は，文字が正しく切り出されたとすると，認識時に作成される文字のラティス構造の中の一つのパス (正解パス) となる．本稿ではこの正解パスを言語情報を用いて求めることで文字認識誤りを訂正する方式について報告する．また，訂正が行なえなかった認識誤りに対しては，それを指摘するという誤り指摘機能についても触れる．

2　言語情報による認識誤り自動訂正 / 指摘

認識文字ラティスの中の正解パスを推定する方法として，ラティス中の全てのパスについての形態素解析を行ない，その解析結果中で最も自然なものを正解パスとみなす方法を採用する．ここで形態素解析結果の自然さを判定する評価値として，解析コストという値を導入する．解析コストは接続コストと認識コストの和で表されるとし、その各コストは以下のようなものである．

- 接続コスト (C_m)

  文字列を形態素に分割した場合の，その形態素の並び方の自然さを評価するコストである．例えば，解析結果中の辞書未登録語などは，普通は使われない単語であり，そのコストは高くなる．また，辞書登録単語であっても，「＜動詞＞＋＜動詞語尾＞」という接続は自然であり，文中に高頻度で現れると考えられるのでそのコストは低くなるが，「＜形容詞＞＋＜動詞＞」の接続の場合には，その出現頻度は前の例に比べると少ないと考えられるのでそのコストは高くなる．

- 認識コスト (C_r)

  認識文字の信頼度を評価するコストであって，認識時に最も類似度が高いと判定された文字 (第１位候補) のみによって作成されるパスの認識コストを０とし，それ２位候補以下の文字を選択することで，その選び方に応じてコストが加算される．

この方式においては，第１位候補文字列パス (P_0) 以外のパス (P) が正解パスとして判定される時には以下の式が成立する．

$$C_m(P_0) > C_m(P) + C_r(P)$$

また，接続コストはその文字列の文としての自然さを評価するものであり，文字列中のコストが高い部分 (日本語として不自然な部分) に関しては，その部分に認識誤りが含まれていると考えることができる．そこで，コスト最小パスの中でもコストの高い部分を指摘を行なうことにより，なんらかの原因によって，訂正が行なわれなった認識誤りをユーザに指摘することが可能となる．

3　誤り自動訂正／指摘システム

前節で論じた誤り訂正 / 指摘方式によるシステムを作成した．本システムで行なっている処理を以下に述べる．

1. 前処理

   システムの性能向上のために，入力される文字ラティスに対して以下の前処理を行なう．

   a: 候補文字の絞り込み
   実際には殆んど正解となり得ないような下位候補文字を含んだパスまで形態素解析するという無駄な処理をしないために，各文字の認識結果について距離値による候補文字の絞り込みを行なう．

   b: 類似文字の追加
   候補文字の絞り込みによって，正解文字が認識候補に含まれない場合がある．また，正解の順位が低過ぎて訂正が成功しない場合もある．これらをなるべく回避するために，絞り込みによって正解がカットされる認識結果を多数収集しておき，頻出するものを類似文字デー

Application of linguistic information in document recognition(2)
Takahiro SAITO, Tomoya OGAWA, Kunio MATSUI
Fujitsu Laboratories Ltd.

タベースに登録し，ラティスに適当な順位点と共に追加するようにした．

c: 認識コストの設定

候補順位に応じた順位点を認識コストとし，その後に字種に関するルールなどによって認識コストを修正する．

2. 解析コスト最小パスの出力 (誤り訂正)

文字ラティス中の全てのパスについて解析コストを計算し，そのコスト最小パス文字列を形態素解析結果として出力する．この処理は全パスを効率良く形態素解析する文字ラティス対応型形態素解析エンジンで行なう．

3. 誤り指摘

まず，最小パスの形態素解析結果に対して，コストが高い部分を指摘する．また，第1位候補文字列の形態素解析結果ではコストが高いと判定されなかった部分であるにもかかわらず，最小パスにおいて置換が行なわれた文字については，この自然さが同等であるとみなし，これも指摘することにした．

4 誤り訂正評価実験

本システムの誤り訂正の性能評価実験について報告する．実験はチューニングに用いた認識結果とは異なる文書サンプルを用意し，認識結果の中で文字数に変化が起こっていない (文字切り出し誤りが起こっていないと思われる) 文をシステムに入力し，その訂正 / 指摘性能を集計するといったものである．なお，実験に用いた文書データは新聞の社説である．

表1に実験結果をまとめた．この表における累積認識率とは，前処理で作成された入力文字ラティスに関して，各文字の認識候補文字に正解文字が含まれている率を表している．つまりこの値は，システムが全ての訂正可能な誤りを正しく訂正し，かつ改悪を全く行なわなかった時に得られる認識率となる．

結果を分析してみると，候補に正解が含まれている場合には，ほぼ (90% 前後) 正しく訂正されていた．また，処理後に残った誤りについて分析を行ない，正解が候補に含まれているにも関わらず訂正されなかった誤りは，文の自然さに影響を与えない括弧などの記号の誤りや，「十八」を「十人」と誤るなどの，誤りがたまたま自然な文を作成した場合であった事が分かった．また，改悪文字 (全部で 34 文字) は，正解を含まない認識誤りの影響により起こったものが多かった．

5 今後の課題

本システムの性能をより向上させるための方針をいくつか記す．

1. 類似文字データベースの充実

より多数の認識結果より作成した類似文字データベースにより，システムに入力する文字ラティスの累積認識率を上げる．

2. 距離値の利用

距離値によって認識コストを設定することで，より精度良く修正が行なえると考えられる．また，距離値による指摘ルールを用意することで，誤り指摘性能も上げることができる．

3. 複数切り出しへの対応

切り出し誤りが起きている箇所をシステムが推定し，その部分の別の切り出しパターンによる認識結果を参照することができれば，切り出し誤りをも訂正することができる．

6 まとめ

以上，言語情報を利用して，文字認識誤りを訂正／指摘するシステムについて述べた．本システムは文字ラティス対応形態素解析エンジンを用いて，文字ラティスの中で，日本語としての自然さと認識時に与えられた候補順位を考慮して正解と思われる文字列を選択するものである．また，このシステムの評価実験を行なった．その結果，正解を含む誤りに対しては，大体においてうまく訂正できることが分かった．また，本システムについて，より性能を向上させるための方法についても触れた．

表 1: 誤り訂正結果

記事番号	1	2	3	4	5	6
総文字数	1117	1035	2150	2178	1179	978
処理前認識率	91.3%	90.0%	91.5%	91.7%	90.6%	88.7%
累積認識率	96.3%	95.9%	98.0%	97.1%	95.9%	96.3%
処理後認識率	95.5%	95.2%	97.3%	95.7%	94.2%	94.9%