1S－4

OCR誤認識後処理の効率化
-------- 候補単語抽出方法と動詞活用処理を中心に --------

久光徹[+]　丸川勝美[++]　嶋好博[++]　藤澤浩道[++]　新田義彦[+]
[+]日立製作所 基礎研究所　[++]日立製作所 中央研究所

1. はじめに

OCRによる一般文書読み取りにおいては，形態素解析等の自然言語処理を用いて誤認識の修正を行う後処理技術が重要である(サーベイ[1]参照)．文字認識後処理のための形態素解析では，各文字に対して複数出力される候補文字を組み合わせて辞書引きし，抽出された単語候補群から最尤単語列を抽出する．辞書引きと最尤単語列抽出手続きは後処理時間の大半を占めるため[2]，その効率化はOCR後処理における重要な課題である．本報では，単語の先頭2文字をキーとする辞書引き，および，動詞活用処理のための新しい辞書見出しを用いて，後処理精度を改善しつつ，辞書引きと最尤単語列抽出をあわせた効率を大幅に向上できることを示す．

2. 従来方法

OCR後処理は，概略すると次のような手続きから構成される：

A）候補文字調整部：候補文字の削除，追加．

B）候補単語抽出部：候補文字集合と単語辞書を用いて，原文中に出現した可能性のある単語(候補単語)を辞書から抽出する．このとき，候補文字に付与された確信度等から，候補単語の適合度を示す照合コストを計算する．

C）候補単語列抽出部：候補単語集合と文法知識を利用して，最尤単語列を抽出する．接続検定，コスト計算処理，最尤単語列抽出の手続きからなる．

1でも述べたように，本報告では主としてB）について考察する．長さが2以上の候補単語抽出において従来用いられている辞書引き手法は二つに大別できる[1]．すなわち，**候補文字主導型**と，**辞書主導型**の二つである．前者は，候補文字集合内の文字を組み合わせて生成した文字列を用いて辞書を検索する手法であり，後者は，単語の特定位置の文字（一般には先頭文字）を手がかりとして，候補単語をまず辞書から抽出し，これらと候補文字集合との照合により，候補単語を抽出する手法である．

候補文字主導型の辞書引きには，二つの主な欠点がある：

1) 平均の候補文字数が多い場合，及び長い単語が候補文字集合中に潜在する場合，組み合わせによる辞書検索回数が相当増大する可能性がある．

2) 正解文字が候補中に存在しない場合，その補間は困難である．

これに対し，**辞書主導型**の辞書引きは，**候補文字主導型**の欠点を補うために開発された．**辞書主導型**の辞書引きの場合，照合対象単語数が辞書により最初から制限されるため，照合効率の組み合わせ爆発的な悪化は起こらない．また，正解文字が候補中に存在しない場合にも補間が期待できる．

本報告では，比較的認識率が低い場合にも対処するため，補間能力の観点から**辞書主導型**の立場を取りつつ，上記の効率に関する問題を解決する．**辞書主導型**の辞書引き手法を比較する便宜上，第$i_1, i_2..., i_n$番目の文字をそれぞれキーとして各単語にアクセスする手法を $\mathbf{Ext}(c_{i1} \vee \cdots \vee c_{in})$ と書く．$\mathbf{Ext}(c_{i1} \vee \cdots \vee c_{in})$ は，候補単語数の圧縮のため，キーである第$i_1, i_2..., i_n$番目の文字として，候補文字のうち第一候補のみを利用するのが一般的である．これを，記号$\mathbf{Ext}^1(c_{i1} \vee \cdots \vee c_{in})$ で表わす．この記法に従えば，候補文字の第一候補と先頭一文字が一致する単語を辞書から抽出した後に単語照合を行う手法（最も一般的な手法）は，$\mathbf{Ext}^1(c_1)$と書ける．また，候補文字の第一候補の文字を利用し，各単語に，その1文字目と2文字目を個別にキーとしてアクセスできる辞書を用いる候補単語抽出法は，$\mathbf{Ext}^1(c_1 \vee c_2)$ と書ける．4では，この2手法を比較実験に用いる．

3. 提案手法

3.1 辞書引き法

我々は，**辞書主導型**辞書引き手法の一つとして，単語先頭側の連続L文字をキーとして辞書引きを行う手法を考える(但し，キーに用いる文字は，必ずしも第一候補の文字だけではない)．この手法は，次の意味で**候補文字主導型**と**辞書主導型**の融合となっている．すなわち，L＝1のときは$\mathbf{Ext}(c_1)$と一致し，Lが十分大きいときは**候補文字主導型**の辞書引きと一致する．以下では，「先頭L文字を用いる辞書引き法」を $\mathbf{Ext}(c_1 \wedge \cdots \wedge c_L)$ で表わす．

後処理のための辞書引きについて重要なことは，各単語が各辞書引き法に基づく辞書検索により候補単語として補足される確率（以下，**補足率**と呼ぶ）をできるだけ大きくしつつ，効率的に候補単語を抽出できることである．

辞書主導型の場合，対象単語数が制限されるとはいえ，例えば$\mathbf{Ext}^1(c_1)$の場合，一つのキーで参照される単語数は相当な数(実験に使用した辞書では最大

An Efficient Error Correction Method for Japanese Text Recognition
[+]Hisamitsu, T., [++]Marukawa, K., [++]Shima, Y., [++]Fujisawa, H. and [+]Nitta, Y.
[+]Advanced Research Laboratory, Hitachi, Ltd.
Hatoyama, Saitama 350-03, Japan
[++]Central Research Laboratory, Hitachi, Ltd.
Kokubunji, Tokyo185, Japan

750，平均47個)に上り，実際には必ずしも効率的とはいえない．また，正解文字が候補中にない場合に補間が可能であるということは，逆に，候補単語の絞り込みが十分行えない場合，誤置換を引き起こす可能性が増えることを意味する．$\mathbf{Ext}^1(c_1 \vee c_2)$は，高い補足率が期待できる反面，候補単語数が$\mathbf{Ext}^1(c_1)$の倍となり，問題である．

我々は，L＝2，すなわち$\mathbf{Ext}(c_1 \wedge c_2)$が，**辞書主導型**の辞書引きとして最適な方法であることを見い出した．すなわち，$\mathbf{Ext}(c_1 \wedge c_2)$では，各キーに対応する単語数が大幅に減少するため，$\mathbf{Ext}^1(c_1)$と比べて単語照合効率が大幅に改善される．また，簡単な確率モデルに基づく考察から，$\mathbf{Ext}(c_1 \wedge c_2)$は**候補文字主導型**や，$\mathbf{Ext}^1(c_1)$に比べて補足率が高いことが分かる．なお，L≧3とすると，補足率が低減する一方，生成されるキーの個数が$\mathbf{Ext}(c_1 \wedge c_2)$の場合の数倍に増大し，効率上の問題が再び生じる．

3.2 活用形処理

動詞の活用形処理の良否は，日本語形態素解析の効率を大きく左右することが知られている[3]．我々は[3]で提案された，子音動詞語幹末尾子音を屈折接辞先頭側に付加した見出しの辞書を用いることにより(**屈折接辞展開方式**と呼ぶ)，単語抽出，及び最尤単語列抽出の効率化を図った(辞書の一部と，解析例は下の図１を参照)．

entry	comments
消	stem
：	
さな	Negative (s + *ana*)
させ	Causative (s + *ase*)
され	Passive (s + *are*)
した	Past (s + *ita*)
：	

a) 屈折接辞展開方式の辞書見出しの例

入力：消さなかった

標準方式:
消/さ/な/かっ/た

屈折接辞展開方式:
消/さな/かった

b) 解析例

図１

4. 比較結果

日経新聞と日経サイエンスから採った約25000文字からなる記事を用いて比較実験を行った．第一位認識率は91.3%であった．比較には次の4指標を用いた：I) 平均単語照合数(図2)，II) 単語補足率(表1)，III) 最尤単語列抽出効率(図3，指標として，解析表作成時の単語間の接続チェック回数を用いた．詳しくは[3]参照)，IV) 誤認識修正率(図4．但し，誤読修正率＝(C-B)/(C+D+E)，ここに，B：正しい文字を，後処理により誤った文字に置換した数．C：誤った文字を，後処理により正しい文字に置換した数．D：誤った文字を，後処理により別の誤った文字に置換した数．E：誤った文字を，そのまま保存した数)．

これらから，3で提案した$\mathbf{Ext}(c_1 \wedge c_2)$と**屈折接辞展開方式**の併用は，1) 単語照合手続きを$\mathbf{Ext}^1(c_1)$と比べて84%削減し，2) 単語の補足率を約3%改善し，3) 最尤単語列抽出効率を20%近く改善し，4) 通常の活用処理を用いる$\mathbf{Ext}^1(c_1)$と比較して，誤読修正率を14.3%改善した．以上から，提案方式は，OCR後処理の効率/精度改善に極めて有効であると考える．

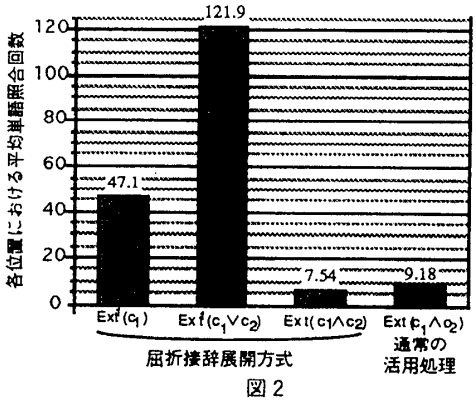

図２

単語抽出方式	$\mathbf{Ext}^1(c1 \vee c2)$		$\mathbf{Ext}^1(c1)$		$\mathbf{Ext}(c1 \wedge c2)$	
	通常の活用処理	屈折接辞展開方式	通常の活用処理	屈折接辞展開方式	通常の活用処理	屈折接辞展開方式
単語補足率	98.33%	98.89%	96.11%	95.45%	98.68%	99.07%

表１

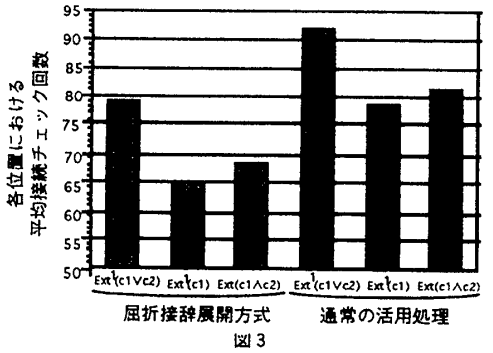

図３

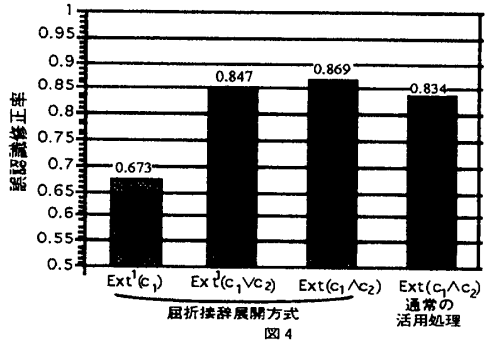

図４

5. おわりに

本報では，1) 候補単語抽出方法，2) 動詞活用処理の工夫，により，形態素解析技術を利用したOCR後処理を，単語補足率，誤認識修正精度を向上させつつ，大幅に効率化できることを示した．

参考文献

[1] 西野 文人: 文字認識における自然言語処理，情報処理, Vol. 34, No. 10, pp1274-1280 (1993)

[2] 高尾 哲康 他:日本語文書リーダ後処理の実現と評価，情処論, Vol. 30, No. 11, pp1394-1401 (1989).

[3] Hisamitsu, T. et al.: "An Efficient Treatment of Japanese Verb Inflection for Morphological Analysis", in Proc. of COLING '94, pp194-200 (1994)