4L−6 日本語ＯＣＲのための誤読修正方式の一検討

紺野　章子　　本郷　保夫

富士ファコム制御　㈱

1．はじめに

近年、日本語文書を対象としたＯＣＲシステムの実用化が進んでいる。日本語文書はその中に含まれる文字種が多く、また、類似字形文字、分離文字等が原因で英文ＯＣＲの認識率程にはまだ到っていない。

そこで、認識精度を向上させるために、日本語文章としての言語的知識を用いて誤読修正を行う方式がいくつか提案されている。[1)2)]

本論文では、我々が先に提案した日本語ＯＣＲ誤読修正方式[3)]についての実験結果の分析と改良のための検討を行う。

2．概要

日本語の文の構造を図1のようにモデル化して捕らえる。ここで、文とは読点によって区切られる範囲、句とは句点によって区切られる範囲とする。また、文節とは文法的には自立語、または自立語＋付属語から成るものである。

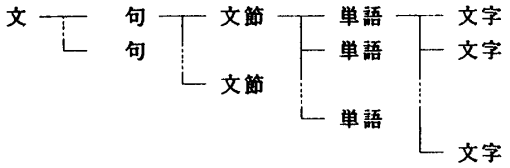

図1．日本語文の構造

文・句といった上位構造に対しては、構文・意味などの高度な知識が必要であるが、現状のＯＣＲは対象文書の種類も多様であり、意味のレベルを扱うのが困難であるため、現段階では、文節レベル以下の構造の最適化により、誤読修正を行う。

処理の流れを図2に示す。

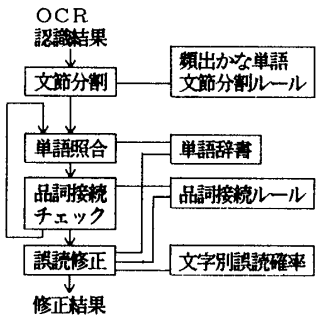

図2．処理の流れ

3．処理とその分析

新聞記事25編(21430文字、認識率平均95.51%)に対して、処理を行った結果と各処理における誤読修正数を表1に示す。ただし、単語照合は、品詞接続チェックを含んでいる。

表1．誤読数と各処理における修正数

候補文字内に正解含む誤読文字数	修正数			
	文節分割	単語照合	誤読修正	合　計
535	23 (4.30%)	206 (38.50%)	83 (15.51%)	312 (58.32%)

以下に各処理についての分析を述べる。

(1)　文節分割

ＯＣＲ認識結果の第1位の文字列に対して、その文字種、頻出かな単語（文章中に頻繁に現れる単語で、文節の切れ目になる可能性の高いもの）を検出し、その並びによって、文節単位に区切る。頻出かな単語の一例を図3に、文節分割ルールの一例を図4に示す。

①　1文字助詞：か・が・て・で・と・に・へ・も・を
②　2文字助詞：から・まで・より・ので・のに
③　2文字名詞：もの・こと・とき
④　連体形語尾：る・た

図3．頻出かな単語の一例

①　助詞「を」の後　　：　これを／下さい。
②　1文字助詞と漢字の間：　将来の／展望
③　連体形語尾と漢字の間：　確認する／手段

図4．文節分割ルールの一例

Analysis and study on Japanese OCR error-correcting method
Akiko KONNO, Yasuo HONGO
FUJIFACOM Corp.

ここで、助詞「を」は単語内の文字としては使われず、しかも我々の文字認識手法では、非常に認識率の高い文字なので、文節の確定度は非常に高いと言える。他の1文字助詞もほぼ認識率が99.9%以上である。また、異文字種間の類似字形文字を持つ文字（力－カ、ロ－ロ等）については、両方の文字種を取れるようにし、前後の文字列により、そのうち一方を選択可能とした。この文節分割によって、75%の文節を正しく分割することができた。しかし、「から」→「か／5」「20%削減」→「20／勿削減」といった誤読に対し誤分割があった。また、ひらがなの長く続く部分は、最長4文節（11単語）が1文節となったところがあった。

（2）単語照合

文節の先頭から単語辞書との照合を行う。単語辞書は約9万語の国語辞書に国内の地名・人名約1万語を加えたものである。

単語照合によって正解単語が照合できた割合は93.82%で、残りは未登録語1.70%、誤読による照合失敗が4.48%であった。

（3）品詞接続チェック

（2）で照合された単語を前後の単語の品詞との接続関係の強度（強，弱，否の3段階に分類）を用いて文法的に正しい単語列を選択する。

（4）誤読修正

（2）での単語照合失敗、（3）での品詞接続チェックによる品詞接続で生じた矛盾、その他誤読を含む可能性の多い単語パターン（1文字単語連続部etc.）の部分について誤読文字を修正する。これは文節内の全候補文字列中の単語から最適な文字列を選択する処理で、単語長、候補順位、連続品詞接続強度、文字別誤読確率（評価用データベースから算出した値）によって評価値を決定する。これにより、修正できた誤読数は表2に示すとおりである。ただし、単語長が単語を決定する大きな要因であることが逆に作用し、「問題でも／ある」→「問題で／もめる」と誤修正した部分もあった。

表2．　文字種別誤読修正率

	誤読数	修正数	修正率
漢　字	325	216	66．46%
ひらがな	123	63	55．26%
カタカナ	38	21	51．22%
その他	49	12	24．49%
計	535	312	58．32%

4．検討

（1）文節型の分類による単語照合の効率化

文節はその末尾に来る単語によって分類可能であり、個々のグループの中で許容される語構成も制限がある。図5に助詞「を」を末尾に持つ文節の解析ルールの一例を示す。文字種等の情報を用いて品詞を制限することにより、単語照合回数を減少させることや未登録語の範囲確定が可能になる。

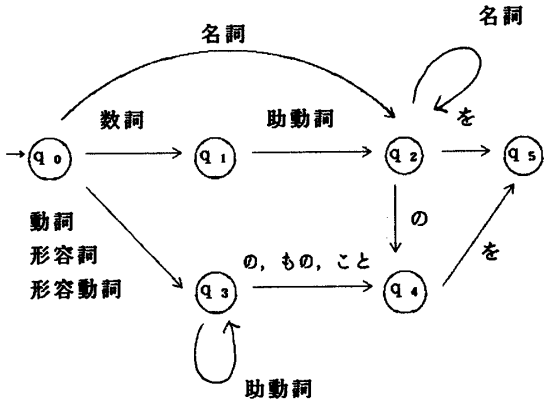

図5．「を」を末尾に持つ文節の解析ルールの一例

（2）頻度情報の利用

現状では、文節内の単語長、品詞接続関係、各文字の候補順位、文字別誤読確率により評価を行っている。しかし、文字毎の確からしさKcは、文字認識時の評価値Cr、類似文字の有無等から求められる文字の安定性Cs、文字の出現頻度Chの関数として、次式のように表される。

$$Kc = Fc\ (Cr,\ Cs,\ Ch)$$

現在の方式ではChを無視しており、Cr，Csについてもその性質を充分把握、利用しているとはいえない。

5．おわりに

本論文では、日本語ＯＣＲの誤読修正方式について、その処理結果の分析と検討を行った。今後は、検討事項に対するレベルアップと、単語辞書の拡充を平行して進める予定である。

1）杉村：候補文字補完と言語処理による漢字認識の誤り訂正手法，信学論Vol.J72-D-Ⅱ(1989)
2）高尾・西尾：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(1989)
3）紺野・本郷・松井：日本語ＯＣＲにおける誤読修正の一手法，電気学会全国大会論文集(1990)