OCR文字認識におけるルビ文字の影響

2L－1

岸本 頼紀† 曹 宇†† 佐藤 匡正†

†島根大学大学院理学研究科 ††島根大学総合理工学部

1．序論

印刷物をシステムに取り込む際に、文字認識機能OCRが効果的に利用できる。しかし、この認識は完全とは言えず、誤認の問題がある[1]。誤認は資料の汚れや皺、字体という物理的な要因によるものもあるが、文字寸法の異なる文字が混在すると、文の読取り誤認が顕著に表れるという経験をする。この状況の典型的な例として、ルビ文字がある。この状況を把握すべく、ルビ文字を含む文章の認識における誤認の特性について分析を試みたので報告する。

2．分析方法

(1) 考え方

ルビ文字は、読者の便に配慮して難解な漢字の読みを補助するためのもので、通常漢字の横に置かれる。字体は小型のいわゆる「ルビ(ruby)活字」が用いられる。文字の認識においては漢字とルビを一体化した文字として認識する可能性がある。このような誤認が実際にどのように生ずるかについて把握する。

(2) 読取り条件

①装置

A社製スキャナとB社製のOCRソフトウェアを使用する。ここでは、特定社の製品についての評価が目的ではないので意図して社名は示さない。

②読取り条件

原稿は参考文献2（縦書き）を用いる。ページ数は141である。総文字数は94579、ルビを有する文字数は2822である。

③ルビ無し文字の読取り

紙面上で加工するのではなく、ルビ付き文字の画像データを加工編集してルビ文字を除去することとする。

3．分析結果

ルビの有無の違いによる文字認識の正当さの差異を比較分析する。また誤認のうち、OCRの特徴的誤りである区切り誤りについて分析する。また、1ページにおける文字数に対してルビを有する文字数にばらつきがあるため、ルビを有する文字のみを調査対象とする。

なお、正認率や区切り誤りは参考文献1による。

(1) 正認率

ルビを含む文字の正認率及びルビを除いた文字の正認率をそれぞれ図1，2に示す。これらで、横軸は正認率を、縦軸はその正認率の頻度を示したものである。ここで正認率は文献1の定義による。また、図1においてはルビの誤認については無視している。図1，2を比較して、ルビを含む場合は、平均正認率は37.5%であるのに対して、ルビを含まない（除去した）ものは86.2%である。この正認率は文献1の結果と概略等しいから、ルビがあると正認率が格段に（44%程度）に落ちる。

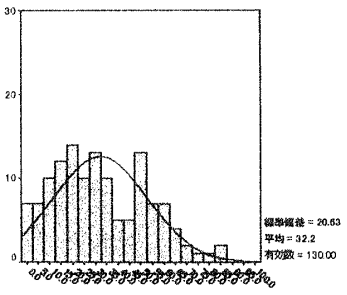

図1 ルビを含む文字の正認率

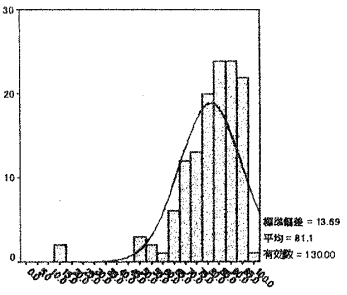

図2 ルビを取り除いた文字の正認率

(2)句切り誤り

漢字や仮名の構成範囲を認識し誤り、ルビを含めて１文字に判断したり、１文字を複数文字に分解誤読は区切り誤りと言う。この誤認について、漢字、平仮名に分けて述べる。

①漢字の構造に見られる句切り誤り

■文字構造による分類

表１に、被認文字の誤認結果の構成を示す。表１では、被認文字の構成を縦に並べ、この横方向に誤認の様子を並べてある。

文字の認識において、構造上の誤認特性がみられる。被認文字が１つの塊、例えば「上」「今」のような場合、認識結果文字が２塊に誤認される割合は83.4%である。また、被認文字構造が２塊である場合は、認識結果文字のうち60.5%が２塊の構造をもつ誤認であり、37.0%が「粥」「蹴」のような、偏と旁の間にもう１つ部をもつ３塊に誤認されている。

表１文字構造の特性

被認文字	認識結果(%)				合計
	平仮名	1塊	2塊	3塊	
平仮名	82.6	3.3	12.4	1.7	100
1塊	6.5	7.2	83.4	2.9	100
2塊	0	2.5	60.5	37	100
3塊	0	7.1	86.9	6	100

■ルビ文字の認識状況

ルビの誤認状況を表２に示す。表２は、地文とルビについての誤認の組み合わせを示したものである。

表２ ルビ付き文字の認識状況

認識		
地文	ルビ	比率(%)
○	○	31.1
×	○	1.2
○	‖	8.3
×	‖	46
○	×	4.9
×	×	0.7
区切り誤り	まとめ	2.9
	分解	3.1
その他		1.7

凡例
○ :正認
×:誤認
‖:不認

全体のうち、46.0%が「**ルビは認識せず・本文は誤認識**」している。これはルビ文字が本文文字の部首を成す偏と旁として認識されるためと推定される。

②平仮名にみられる句切り誤り

平仮名の場合の区切り誤りを表３に示す。

表３ 平仮名の誤認結果

出力結果	割合(%)
い	18.2
か	7.4
その他(平仮名)	5.8
が	47.1
その他(濁音・半濁音)	4.1
その他(漢字)	14

被認文字が平仮名の場合、被認文字に関わらず「が」と認識している割合が高い（47.1%)。これは、ルビを「が」の右にある「ゞ」の部分と誤認していると推定される。平仮名がルビをもつ場合は、その直上の漢字に対するルビが、下にはみ出している場合が大半である。

③その他の誤認

その他の誤認として、複数行を１行にまとめて認識している誤認が、全行数2154行に対して62行ある。これはルビを取り除いた場合には見られない誤認であり、ルビ文字により行間が狭くなったことに起因する誤認であると推定される。

４．結論

OCR 文字認識におけるルビ文字の影響について分析した結果、次のことが分った。

1)ルビを有する文字は、ルビをもたない文字に比べて誤認の可能性が高い。

2)この誤りは、ルビ文字を地文の文字と一体化して認識することによって生ずることが多い。

以上

参考文献

[1] 佐藤 匡正「利用者からみたOCR認識誤りの分析」電気・情報関連学会中国支部第49回連合大会、講演論文集(1998)

[2]荒俣宏「帝都物語８」角川書店