2M－2　OCRにおけるルビ文字認字率の改善

曹　宇　　佐藤 匡正
（島根大学大学院総合理工学研究科）

1．序論

書物など印刷物を計算機に取り込む際に、OCR文字認識の機能を利用することによって、便利に実現できる。しかし、誤認識が生じるため手直しは手間がかかり、効率が悪くなる。日本語の文書には異なる寸法をもつ文字が混在しているため、文字認識する際は認字率が低下する[1)]。特に、文にルビ文字を含む文書は顕著に低下する。これは、ルビ文字と地文の字が一括して1つの文字として認識されるからである。例えば、「額(ひた)」は「献」と誤認識されることがある。

そこで、ルビ文字を含む文の認字率の改善を図る方法を考案し、有効性について実験を試みたので、報告する。

2. 改善提案

（1）ルビ文字の分離

ルビ文字付きの文字の認字率が良くないのは、これまでの経験から、字の大きさに極端な差があり、OCR処理ではルビを字と見ないことと推定される。一方、字の大きさが同程度であれば、認字率が低下することはない[2)]。そこで、改善方法として、地文とルビを別箇に認字させることを考える。この場合、ルビをどのようにして地文と区別するかが技術上の問題となるが、文字寸法に着目すれば、識別できる。

ここで文字の寸法とは文字域の有効面積で、文字の外接四角形のことを指す。これまで文字寸法についての調査結果より[3)]、ルビの有効面積／地文の有効面積≦1／4、及びルビの横(縦)有効長／地文文字の横(縦)有効長≦1／2という特性を利用し、ルビ行と地文行の行間域に着目して、ルビ文字と地文が分離できる。

Improvement of Ruby−Character's Recognition Based on OCR Applications
Yu Cao　Tadamasa Satou
University of Shimane, 1060 Nishikawatsu,
Matsue-city, Shimane,690-8504,Japan

（2）システム構成法

新たに、OCR処理の開発は期間及び費用点から難しいので、OCR処理は安価な市販品を流用し、混合した文からルビ文字と地文を分離する。システムを前処理として置き、この処理結果をOCR処理に入力させる方式を実現する。

3．実験

（1）目的

分離したルビの認字率の向上具合、地文とルビを分離する自動化の具合、及び前処理方式の実現性を確かめる。

（2）実験方法

実験に用いた機材及び資料を次に示す。データはルビ文字を含む縦書きの書物を使用する。

- 資料：現代仮名遣い日本語の文庫文(資料1と資料2)*) を対象とし、総ページ数は40で、ルビ文字数は2570である。資料1は10ページ分、資料2は30ページ分とする。
- 解像度：スキャナの解像度は300dpi.
- 認字率：ページごとにOCR処理によって、正当に認識されたルビ数の全ルビ数に対する割合である。

*) 資料1　荒　俣宏：「帝都物語8」角川書店
資料2　壺井　栄：「柿の木のある家」偕成社文庫

（3）実験結果

資料1と資料2におけるルビ文字と地文混在する場合とルビ単独存在する場合にルビ文字のページごとの認字率を図1、図2に示す。両図ともルビ文字単独存在する場合の認字率によって昇順に整列して示す。

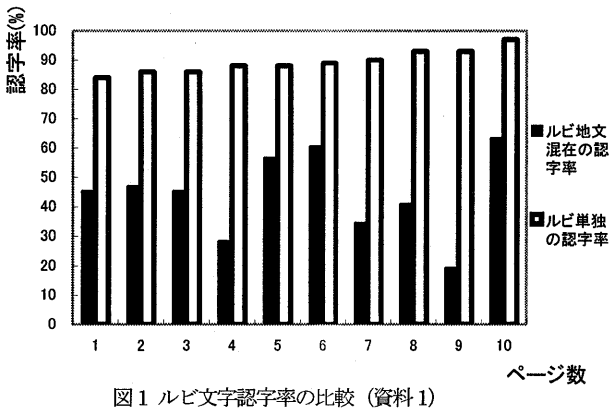

図1 ルビ文字認字率の比較（資料1）

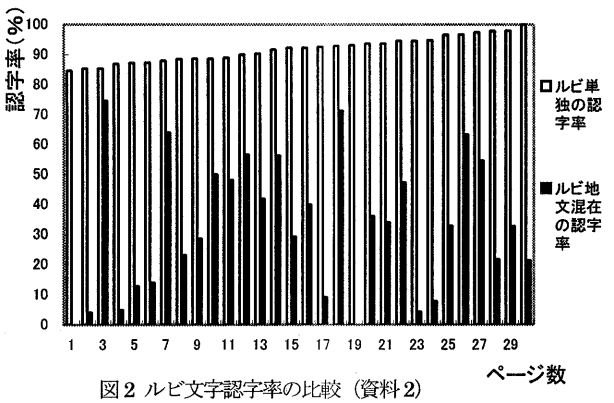

図2 ルビ文字認字率の比較（資料2）

資料1の場合は、ルビ地文混在する場合のルビ文字認字率は19%から63%までの範囲にあり、ルビ文字単独の認字率は83%から97%までの範囲にある。

資料2の場合は、ルビ地文混在する場合のルビ文字認字率は0%から71%までの範囲にあり、ルビ文字単独の認字率は85%から100%までの範囲にある。

4．考察

（1） 認字率の比較

資料1と資料2のルビ文字と地文が混在とルビ文字単独存在の場合の認字率は表2に示す。

認字率＝(認識したルビ／全ルビ数)＊100%

表1 ルビ文字認字率の比較

認字率 / 資料	ルビ地文混在の認字率(%)	ルビ単独の認字率(%)
資料1	45.2	90.1
資料2	29.7	91.2

上の結果から、ルビ文字を地文から分離して、認字する方式は有効であると言える。

（2） 誤り要因

実験結果において、認字率の誤りの要因を次に例挙する。

① 読み取り上の問題

ルビ文字と地文の行間域の空線に着目するため、資料の読み取り上の傾きによって、ルビ文字と地文とうまく分離できないことがある。

② 紙面の物理的な要因

紙面の汚れ、皺また印刷する際にインクによる汚れなどが文字認識に影響を与える。ルビ文字の寸法が小さいため、汚れなどルビ文字の一部と認識される場合がある。例えば、ルビ文字「し」のあたりに黒い点が存在すると、「じ」と誤認識されることがある。

③ 解像度

ルビ文字は寸法が小さいため、地文文字より解像度の影響を受ける。

（3）縦書き文書に対するOCR認識特性

文字寸法の揃った文字においても、縦書き文は誤認識される場合がある。これは、OCR処理の特性と考える。例えば：「う」を「スノ」に、「じ」を「・・し」に、「く」を「ノ＼」に認識される。今回使用したこのOCRの処理特性による誤認識ルビ数が資料の全ルビ数に対する割合は4.3%である。

（4）一般性

ルビ文字認字率の比較によって、資料1と資料2の処理後の認字率はほぼ同じである(90%以上)。よって資料が変わっても高い認字率が得られるという一般性があると確認できた。

5. 結論

本論文ではルビ文字認字率についてルビ文字と地文を分離し別箇に認字する方法を考案し、その効果を実験によって確かめた。効果は顕著で、ルビ文字の認字率は2～3倍改善された。

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) 岸本 頼紀、曹 宇、佐藤 匡正：「OCR文字認識におけるルビ文字の影響」 情報処理学会第59回全国大会講演論文集(1999)

3) 曹 宇、佐藤 匡正：「文字種別による文字寸法の違い」情報処理学会第61回全国大会講演論文集(2000)