ピンイン情報を併用したオンライン中国語文字認識

2P－6

川又　武典　　丸山 冬樹　　南部 元　　依田 文夫
三菱電機株式会社 情報技術総合研究所

1．はじめに

我々はオンライン手書き中国語文字入力方式の開発を行っている。しかし、文字認識単独では極端な続け字、崩し字等に十分に対応できないため、単語辞書を用いた入力方式を提案した[1]。しかし、この方式においても、1)非単語入力時に改善できない、2)単語入力時でも文字認識の認識候補中に正解文字が入らない場合に改善できない、3)単語の先頭1文字を用いた単語補完では十分に単語候補を絞り切れない、という問題があった。また、誤読時に認識結果の候補中に正解が存在しない場合の修正が困難であった。

オンライン文字認識はオフライン文字認識と異なり、逐次文字を筆記しながら入力するため、筆記情報以外の情報の入力が可能であり、文字の絞り込みに有効な情報を併用することができる。

そこで我々は、容易に入力が可能な、ピンイン（中国大陸における共通語のアルファベット発音表記）における母音情報を併用して認識精度を向上させる入力方式を検討したので報告する。

2．ピンイン情報の分析

今回の文字認識及び単語知識処理実験の対象とした文字は、使用頻度の高い中国語簡体字3,942文字である。この文字セットにおいては、ピンイン（四声は考慮しない）を1種類のみ有する文字は3,584文字(90.9%)である。各文字の有するピンイン種類の平均は1.1となり、各文字の有するピンインの種類は少ない。しかし、ピンインあたり平均10.8文字の候補が存在し、ピンイン単独での文字の確定は困難である。

次に、5種類の単母音情報（「ü」は頻度が低いため「u」に統合した）を用いた場合の絞り込み性能を調査した。結果を表1に示す。なお、複数の母音を有する文字は重複カウントしている。

表1．母音別絞り込み文字数

母音	文字数	母音	文字数	母音	文字数
a	1521	u	1204	o	759
i	1660	e	665		

表1より、「e」を有する文字の絞り込み性能が最も良く、「i」が最も悪い。しかし、最悪でも全文字の4割程度に候補を絞り込むことが可能である。

次に、入力母音の数を考慮した場合（複数の母音を有する文字は、複数の母音をすべて入力する）の絞り込み性能を調査した結果を表2に示す。なお、母音の順序は考慮していない。

表2．母音別絞り込み性能（母音数を考慮）

母音	文字数	母音	文字数	母音	文字数
a	645	a i	415	i o	12
i	779	a u	197	u e	43
u	595	a o	217	u o	239
e	431	i u	179	a i u	22
o	212	i e	213	a i o	124

表2に示すように、母音数を考慮した場合は、一層の絞り込みが可能で、最悪でも全文字の2割以下に絞り込み可能である。

3．文字認識との併用

2章で分析した結果を基に、我々が開発した中国語文字認識にピンインの母音情報を併用した場合の実験を行った。具体的には、正解文字の母音情報による絞り込みを行った結果の候補文字について、従来の文字認識を実行することにより行った。結果を図1に示す。なお、実験には、評価用手書き中国語文字データ200人分を用いた[2]。

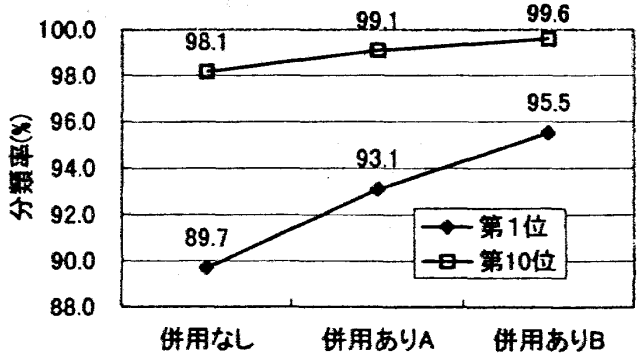

図1．文字分類率

Chinese Character Recognition using Pinyin.
Takenori Kawamata　Fuyuki Maruyama　Hajime Nambu
Fumio Yoda
Information Technology R&D Center, Mitsubishi Electric Corp.
5-1-1 Ofuna, Kamakura, Kanagawa, 247, Japan

図1中、併用ありAは、表1に示す母音数を考慮せずに絞り込みを行った場合、Bは表2に示す母音数を考慮した場合である。母音情報を併用しない場合に比べて、第1位分類率はAで3.4%、Bで5.5%向上している。また、第10位分類率はAで1%、Bで1.5%向上している。以上の実験の結果、対象文字の絞り込みは、文字認識における第1位分類率を大きく向上させると共に、誤読時の候補文字による修正機会を増加させるのに有効であることが判った。

4．単語知識処理との併用

母音情報を併用した文字認識結果に対して、中国語の一般単語辞書を用いた単語知識処理[1]を適用する実験を行った。具体的には、単語辞書中の2文字単語（63,979単語）について、前記200人分の評価用文字データを用いて単語知識処理のシミュレーションを行った。結果を図2に示す。なお、単語分類率は第3位以下同率となる（第4位以下が正解の確率は低い）ため、第3位の結果までをプロットした。

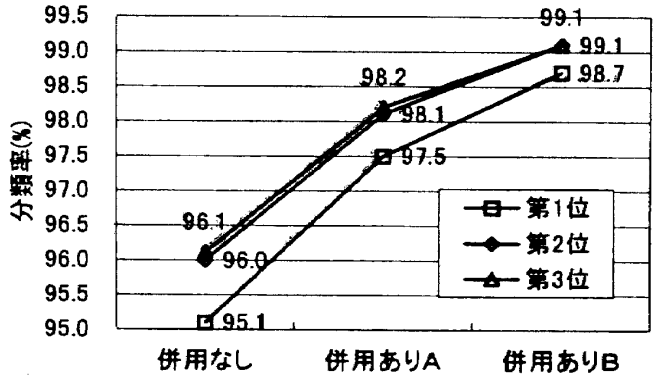

図2．単語分類率

図2に示すように、単語分類率の第1位、第3位共に併用しない場合に比べて向上している。以上の実験の結果、母音情報の併用は単語知識処理における単語の第1位分類率を向上させると共に、単語誤読時の単語候補による修正機会を増加させるのにも有効であることが判った。

次に、非単語（未登録単語を含む）を入力した場合のリジェクト性能を評価するために、中国語の新聞文字データ(1,777 万文字)から 2 文字長の非単語を75,000 個抽出し、同様に単語知識処理のシミュレーションを行った。その結果、リジェクト率は併用しない場合の64.4%に対して、併用ありAで64.8%、Bで66.7%となり、僅かながら向上するが、単純に文字認識における分類率の向上だけでは非単語のリジェクト性能向上には大きな効果がないことが判った。

5．単語連想処理との併用

前記中国語の一般単語辞書中に存在する2文字長以上の単語に対し先頭1文字を筆記して、単語連想処理（2文字以上の単語候補を出力）を行った場合に絞り込まれる候補単語数について調べた。具体的には、2文字目の母音情報を入力しない場合と入力した場合の候補単語数がN個に絞り込まれた単語数の全単語数に対する比率を調査した。結果を表3に示す。なお、括弧内の値は、新聞データによる単語出現頻度で重み付けした値である。

表3．候補単語数別の比率（%）

候補単語数	母音入力なし	母音入力あり	母音入力あり 母音数考慮
1	1.4(0.3)	4.6(2.1)	14.4(8.4)
2以下	3.1(0.9)	9.2(4.1)	26.8(18.3)
3以下	4.8(1.4)	13.7(7.6)	37.1(26.3)
4以下	6.2(7.7)	17.6(10.2)	45.5(33.9)
10以下	15.2(7.8)	38.5(27.8)	72.8(64.4)

表3の結果より、2文字目の母音入力を行うことにより、候補単語数の絞り込みが図れることが判る。また、出現頻度を考慮した場合の比率が考慮しない場合に比べて低くなっており、出現頻度が高い単語は、候補単語数が比較的多いことが判る。

6．まとめ

中国語の母音情報を文字認識処理及び単語知識処理に併用する方式を検討し、評価データにより認識実験を行った。その結果、母音情報を併用することにより、文字分類率、単語分類率が共に大きく向上することが判った。また、2文字以上単語の先頭1文字筆記による単語連想処理において、2文字目の母音情報による単語候補の絞り込みが有効なことが判った。

7．今後の課題

文字認識方式改良による文字認識精度の向上が今後の課題である。

参考文献

[1]川又他："中国語単語知識処理方式の開発"，情報処理学会秋季全国大会(1997)
[2]川又他："中国語オンライン手書き文字データの分析"，信学会春季全国大会(1997)