中国語単語知識処理方式の開発

6K－7

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

中国語漢字をコンピュータに入力する手段の1つとして、オンライン手書き文字認識方式の研究が進められている。しかし、1)中国大陸の国家標準（GB）漢字セットで決められている漢字（簡体字）は、6,763文字存在するうえ、旧字の繁体字も姓名・地名等に一部使用されており、認識対象文字数が非常に多い、2)中国ではひらがな、カタカナのような代替表現手段が存在しないため、漢字を非常に速く筆記する必要があり、日本語漢字に比べて続け字等の画数変動が発生する割合が高く、崩れた字形が多い[1]。このため、文字認識方式の高精度化だけでは中国語漢字を効率的に手書き入力するには限界がある。そこで、中国語単語入力効率の向上を目的に、我々は中国語の一般単語辞書を用いた単語知識処理方式、単語連想処理方式を開発した。

本稿では、単語知識処理方式及び単語連想処理方式を用いた中国語入力システムの概要、及び約200人分の手書き中国語文字パターンデータベースを用いた評価結果について報告する。

2．中国語入力システムの概要

図1に今回開発した中国語入力システムの概要を示す。今回開発した方式では、単語の先頭2文字を手書き入力し、文字認識処理により各文字を認識した後、それぞれの文字認識結果の認識候補文字を最大10文字出力する。次に単語知識処理は、それらの認識候補文字の組み合わせにより得られる単語の内、単語辞書中に存在する単語を単語知識処理の結果として出力する。単語知識処理結果の単語の距離値には、文字認識の結果得られた各文字の順位の和を使用した。

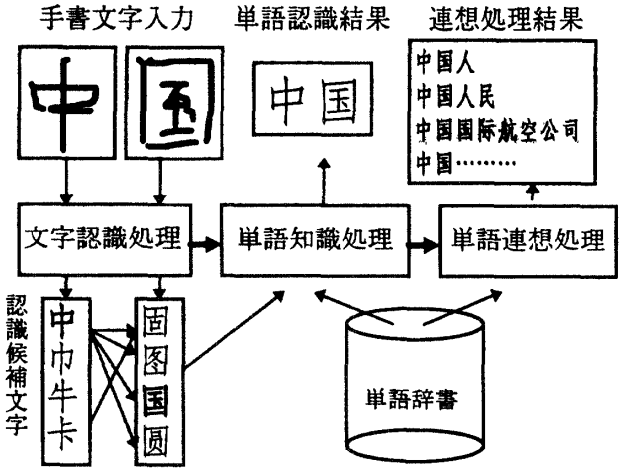

図1．中国語入力システムの概要

最後に単語連想処理では、単語知識処理により得られた単語から始まる3文字以上単語を連想結果として出力する。

ここで、中国語漢字の文字認識処理には、大局的整合法とDPによるストロークの対応付けを併用したオンライン手書き文字認識方式[2]を用いた。

3．単語辞書

単語辞書には、当社が開発したピンイン入力 S/W で用いている単語辞書を使用した。総単語数は、77,258 単語で、平均単語長は 2.2 文字、最長は 12 文字単語である。単語長別の単語数を表1に示す。

表1．単語長別の単語数

単語長	単語数	比率	累積比率
2文字	53630	69.4	69.4
3文字	10271	13.3	82.7
4文字	12467	16.1	98.8
5文字	507	0.7	99.5
6文字以上	383	0.5	100

4．単語知識処理の評価

4．1 文字データベースを用いた評価

音声とは異なり、文字の場合は、文字を1文字、1文字独立して筆記した場合は、各文字の字形は前後の文字の影響を受けにくい。そこで、文字認識方式

Post Processing for Chinese Character Recognition Based on Word Dictionary.
Takenori Kawamata　Fuyuki Maruyama　Hajime Nambu
Fumio Yoda
Information Technology R&D Center, Mitsubishi Electric Corp.
5-1-1 Ofuna, Kamakura, Kanagawa, 247, Japan

を開発するために収集した中国語の文字パターンデータベース（6,763文字、200人分）を用いて、同一人物の筆記した任意の単語パターンを作成し、その認識結果に対して単語処理のシミュレーションを行った。

4．2 評価結果

単語辞書中の先頭2文字が異なる単語（63,979単語）について、単語知識処理の評価を行った結果を表2に示す。

表2．評価結果（登録単語）

	1位	2位	3位	10位	エラー	リジェクト
1文字目認識率	90.0	94.1	95.5	98.0	10.0	0.0
2文字目認識率	89.9	93.9	95.5	98.0	10.1	0.0
単語認識率（単語処理前）	81.4	88.7	91.4	96.1	18.6	0.0
単語認識率（単語処理後）	95.1	96.0	96.1	96.1	2.7	2.2

表2より、単語知識処理後の単語候補の第3位までで、第10位までの認識候補文字の組み合わせにより単語知識処理を行った場合の理想単語認識率96.1%に到達することが判った。

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単語抽出し、同様に評価を行った（表3）。

表3．評価結果（非単語）

	1位	2位	3位	10位	エラー	リジェクト
1文字目認識率	90.2	94.4	95.8	98.1	9.8	0.0
2文字目認識率	90.0	94.2	95.6	98.0	10.0	0.0
単語認識率（単語処理前）	81.6	89.2	91.8	96.2	18.4	0.0
単語認識率（単語処理後）	0.0	0.0	0.0	0.0	47.7	52.3

表3より、単語辞書中に存在しない文字列を筆記した場合は、約半分をリジェクトできることが判る。また、単語の距離値がある一定閾値以上のものをリジェクトすることにより、登録単語の単語認識率を変えずに、表3における非単語のリジェクト率を64.4%まで向上させることができる。

5．単語連想処理の評価

単語辞書中に存在する3文字長以上の単語に対し先頭2文字を筆記して単語連想処理を行った場合の候補単語数を調べた結果を表4に示す。

表4より、3文字以上単語の先頭2文字を筆記した場合は、その6割が確定すること（単語の残りの文字の筆記が必要ないこと）、候補単語を一度に6個表示できるようにした場合は、95%の単語が候補単語中に含まれることが判った。

表4．候補単語数の分布

候補単語数	単語数	累積比率
1	14347	60.7
2	4408	79.4
3	1857	87.2
4	1000	91.5
5	475	93.5
6	348	95.0
7以上	1193	100

次に前記中国語の新聞文字データの一部（100万文字）を用いて、単語連想処理の効果を評価した。その結果を表5に示す。

表5．単語連想処理の効果

単語長	度数	文字数	比率	補完文字数	比率
2	334390	668780	66.9	0	0.0
3	16429	49287	4.9	16429	1.6
4	8593	34372	3.4	17186	1.7
5	441	2205	0.2	1323	0.1
6	167	1002	0.1	668	0.1
7	80	560	0.1	400	0.0
8	15	120	0.0	90	0.0
9	4	36	0.0	28	0.0
		756362	75.6	36124	3.6

表5より単語辞書中の2文字以上単語の占める割合は75.6%、3文字以上単語の占める割合は8.7%であり、中国語の文章においては、単語の占める割合が比較的高いことが判った。また、3文字以上単語について、先頭の2文字筆記による単語連想処理結果の候補単語選択により単語を入力した場合は、3.6%の文字の筆記が省略可能（補完文字）で、入力効率が向上する。

6．まとめ

2文字の手書き入力に対し単語知識処理、単語連想処理を行う方式を検討し、大量データで、単語認識率、非単語入力時のリジェクト率、単語入力効率を評価した。その結果、2文字入力による入力方式が効果的に中国語単語を入力できることが判った。

7．今後の課題

文字認識精度の向上及びより高次な知識を用いた知識処理方式の開発が今後の課題である。

参考文献

[1]川又他：“中国語オンライン手書き文字データの評価”，信学会春季全国大会(1997)

[2]川又他：“大局的整合法とDPによるストロークの対応付けを併用したオンライン手書き文字認識”，信学会春季全国大会(1996)