手書き用語入力処理における再照合方式

1K-4

大光明直孝　　小黒　雅己　　田中　忠仁

ＮＴＴ情報通信処理研究所

1. まえがき

任意ピッチの手書き漢字で記入された文字列の処理方式として、入力文字列中に用いられる用語が既知であることを前提として用語辞書を用意し、これを利用して文字領域の誤切り出し部分を修正する方式が提案されている[1],[2]。

従来方式における修正処理は、用語照合の結果として得られる全候補用語について、文字切り出しから用語照合までの処理を再実行（再照合）することにより行われるため、高い正読率が得られる反面処理負荷が重くなる傾向であった。

本稿では、再照合の前処理段階で、それまでに得られている候補用語の構成文字と文字照合の結果得られる候補文字との一致状態を評価することにより、再照合の対象とすべき候補用語を絞りこむ方式を提案し、従来方式と比較した評価結果を示す。

2. 任意ピッチ文字列の処理方式

従来の任意ピッチ文字列の処理方式の概念を図1に示す。

この方式では、文字切り出し、文字照合、用語照合、再照合の4処理ステップで手書き漢字文字列を処理する。各処理の概要を以下に示す。

(1)文字切り出し：入力画像情報からＸ，Ｙ方向の黒画素分布、黒画素連結成分等の抽出により、一文字毎の領域を切り出す。

(2)文字照合　：　切り出された文字画像データから文字の特徴を抽出する。抽出した特徴量と文字辞書の特徴量との距離を求め、候補文字を抽出する。

(3)用語照合　：　文字の照合結果をもとに、用語を連想し、上位Ｎ個を候補用語とする（連想統合法[1]）。候補用語の各文字が候補文字列中にすべて存在する場合は、該用語を照合結果として出力する。存在しない文字がある場合には、再照合処理を実行する。

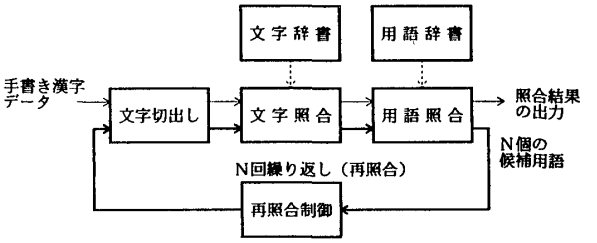

図1．手書き文字列入力の照合方式

(4)再照合　：　完全に一致していないＮ個の候補について、不一致部分の文字切り出し、文字照合等の一連の処理を実行し、再照合で一致した文字の得点を再付与する。この得点の最も高いものを結果として出力する。

3. 候補の絞り込み機能を設けた再照合方式

(1) 基本的な考え方

本提案の再照合方式では、再照合の前段で、それまでに得られる情報を基に、候補用語が事前に推定できる評価尺度を用いて、用語毎に正解となる可能性を評価し、可能性の低い用語を再照合の対象から除外することで、再照合負荷の軽減を図る。

(2) 評価尺度の設定

再照合の前段までに得られる情報には、①連想統合で得られる用語の得点、②用語の文字数、③候補文字と用語構成文字との一致状況（不一致領域の数、一致文字数、一致文字の得点を累積した用語の得点（ＤＰ得点）等）が有る。

これらの情報の中で、”用語の構成文字と候補文字の一致する割合が少ない用語は、正解となる可能性が低い”との観点から、一致文字数に関連深い②と③の情報を用いて、以下のような評価尺度を設定した。

ｉ．文字一致率：Ｒ＝Ｋ／Ｌ

ii．ＤＰ得点　：$P=\sum_{j=1}^{k} S_j$

Ｋ　：　用語の構成文字と一致した候補文字数

Ｌ　：　用語文字数

S_j　：　用語のｊ番目の構成文字と一致した候補文字の得点

Re-collating Method for Handwritten Term Input.
Naotaka OMIYA, Masami OGURO and Tadahito TANAKA
NTT Communications and Information Processing Laboratories

(3) 評価尺度の特性分析

上記評価尺度の妥当性確認のため、手書き漢字で記入された平均10文字の有意文字列、123 例のサンプルデータについての特性分析を行い、以下の結果を得た。

・候補用語を再照合後の最終得点でソートしたとき、各用語の文字一致率とＤＰ得点は、最終得点に対し、単調減少ではないが、漸減傾向を示す。即ち、一致部分の少ない用語が再照合によって高得点になる確率は低い。

この結果から、文字一致率またはＤＰ得点が評価尺度として妥当であることがわかる。

但し、分布が単調減少でないという特性から、この評価尺度の絶対値をしきい値として利用することは適切ではない。

上記評価尺度をもとに、｜Xi-Xj｜＜Ｃの条件を満たす用語を一つのグループとしてまとめて、そのグループに相対的な順位(グループ化順位) を付与し、上位のｎ位を再照合対象とする形でしきい値を設定する。

(Xi,Xj:評価尺度の値, C:グループ化の基準値)

(4)評価尺度のしきい値

評価尺度のしきい値を求めるため、しきい値(順位) と正解候補が含まれる数（包含率）の関係を調査した。その結果を図２に示す。

この図から、正解包含率が１００％になる順位、即ち正読率が維持できる順位は、ＤＰ得点の上位 i_1 位または文字一致率の上位 i_2 位である。ここで、i_1, i_2 および候補用語数（Ｎ）の間には、$1 \leqq i_1 \leqq i_2 \leqq N$ の関係が成立することがわかる。

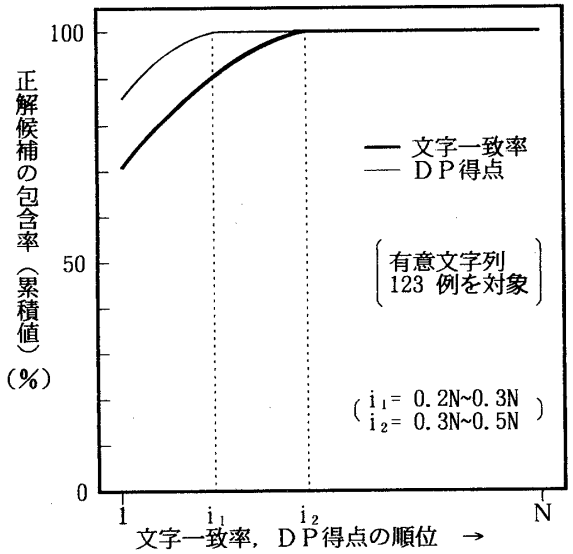

図２．文字一致率，ＤＰ得点の順位に対する正解候補の包含率特性

(5) 負荷削減効果

前述の評価尺度のしきい値に基づき再照合対象として残した用語数と元の用語数を比較する方法で負荷の削減効果を評価した。その結果を図３に示す。

正読率の低下がない評価尺度値 i_1, i_2 における負荷削減率は、i_1 の場合約７０％、i_2 の場合約６５％となり、設定した評価尺度が有効であることを確認した。

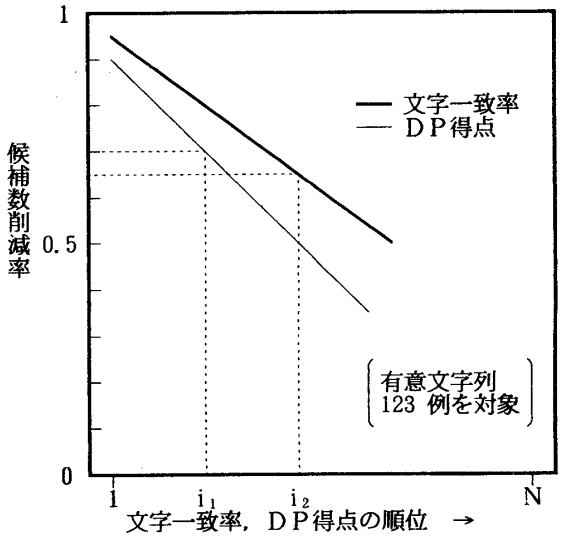

図３．候補数削減率特性

4. むすび

任意ピッチ手書き漢字文字列の処理方式に関して、候補の絞り込み機能を設けた再照合方式を提案した。本稿では、再照合候補削減の評価尺度として、ＤＰ得点または文字一致率が使用できることを示し、正解候補がほぼ１００％含まれるしきい値の設定により、約６５～７０％の候補削減率が得られることを示した。これにより、全体の処理負荷削減効果は、およそ３分の２となる。

(文献)

(1)松尾，佐藤，津田「連想統合型照合による単語あいまい検索法」情処学会，第３４回全国大会，４Ｅ－７（１９８６）

(2)仲林，松尾，津田「用語あいまい検索を用いた手書き文字列入力方式」６２年度人工知能学会全国大会，８－７（１９８７）