6K-06 遺伝アルゴリズムを用いた文字認識後処理*

6K－6

篠沢 佳久† 大駒 誠一‡

慶應義塾大学 理工学部 管理工学科§

1 はじめに

現在さまざまな文字認識の研究が行なわれているが，一文字だけを対象とした認識手法では認識率に限界がある．そこで文字認識システムが出力した文字候補を組合せ，単語もしくは文章として再認識を行なう後処理によってさらに認識率を向上させる研究が行なわれており，郵便葉書の宛先のように実用段階に至っているものも少なくない．

しかし後処理の対象を日本語の一般文章とした場合，文字認識システムが出力する文字候補数が多い，もしくは文章が長いとその組合せにより処理する計算量が大きくなってしまう．そこで本研究においては文字認識の後処理を組合せ最適化問題ととらえ，遺伝アルゴリズムを用いて文字候補の組合せによる計算量の削減と認識率の向上を試みた．

2 後処理の基本的な手法

本研究における後処理は全組み合わせ探索的な方法を基本としている．文字認識システムが出力する一文字に対する文字候補数を M 個，認識した文章の長さを N 個とした場合，N^M 個の文章の組合せが考えられる．その N^M 個の文章一つ一つに対していかに日本語らしいか，自然言語処理の点から適切に評価を行ない，その評価関数を最大にする文章を後処理の結果とする方法である．

3 問題点とその解決案

この手法では N^M 個の文章を評価しなければならないため巨大な計算量がかかる．そこで一つ一つの文字候補をどう組み合わせて文章を作り出すか，組合せ最適化の問題としてとらえ組合せ最適化手法を用いて計算量を削減する．今回は組合せ最適化手法には遺伝アルゴリズムを用いた．

*Post-Processing of Character Recognition with Generic Algorithm .

†Yoshihisa SHINOZAWA

‡Seiichi OKOMA

§Faculty of Science and Technology , Keio University

4 提案する後処理システム

4.1 後処理システムの全体図

提案する後処理システムの全体図は図1に示す通りである．後処理システムはおおまかに分けると文字認識システムからの出力を処理する1. **前処理部**，文章の日本語らしさの評価を行う2. **自然言語処理部**，そしてさらに評価の高い文章を組み合わせていく3. **遺伝操作部**から構成される．

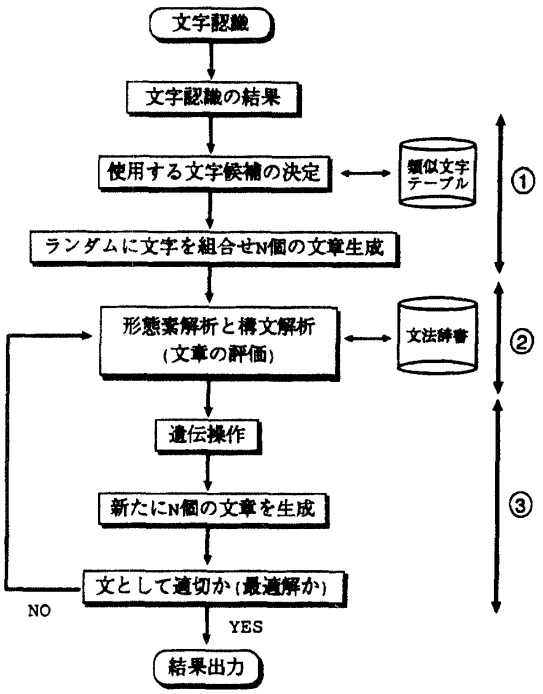

図 1: 後処理システムの全体図

4.2 前処理部

文字認識システムからは通常，候補文字とその文字に対する信頼度が出力される．筆者が作成した文字認識システム [1] は電総研提供の手書き文字データベース ETL8B，956 文字を認識対象とし，一文字につき 10 個の文字候補を出力する．信頼度とは標準文字パターンとのユークリッド距離である．

前処理部では 10 個の文字候補数を削減したり，あるいは正解が文字候補に含まれていない場合，類似文

字テーブル[2]を用いて補うといった処理を行なう．

4.3 文章の評価（自然言語処理部）

前処理部で削除又は補った文字候補から文章をランダムにもしくは次節以降述べる遺伝操作によってN個の文章を作り出し，その文章がいかに日本語らしいかを評価する．その評価方法には，**形態素解析**と**構文解析**という自然言語処理の面から行なった．

1. 形態素解析
   まず形態素解析によって文章をTrie構造型の辞書を利用し形態素単位に区切りその形態素の品詞も合わせて検索し，隣接する形態素どうしが品詞として接続可能かどうかを調べる．

2. 構文解析
   文章の終わりまで形態素解析できた文のみに対して次に構文解析を行なう．構文解析は句構造文法で行ない，各形態素をその品詞情報から上位の句へと還元していく．

3. 評価づけ
   二つの処理の結果から評価づけを行ない数値化する．評価関数には次式を用いた．この評価関数を最小にする文章を後処理結果とする．

$$\text{評価関数} = \alpha \times \text{自立語の個数} + \text{付属語の個数}$$
$$+\beta \times (\text{文の長さ} - \text{形態素解析できた長さ})$$
$$+\gamma \times \text{構文解析時の句の個数} + \eta \sum_{i}^{\text{文の長さ}} D_{ij}$$
$$D_{ij} \cdots i \text{番目の第} j \text{候補文字の信頼度}$$

4.4 文章の組み変え（遺伝操作部）

4.4.1 遺伝操作の流れ

生成されたN個の文章に対し，自然言語処理による評価値をもとに選択淘汰，交叉，突然変異という遺伝操作を繰り返すことによってより日本語らしい（評価関数の値が小さい）文章を作り出していく．

4.4.2 染色体のコーディング

染色体のコーディングは図2に示すように，文の長さ分だけ染色体の配列を用意する．そして文章のi番目の文字候補順位をi番目の染色体の配列に格納していく．交叉マスクは形態素解析もしくは構文解析からの情報を用いて，組み合わされた形態素もしくは句ごとに区切り，１と０の二値によってその境目が分かるようにしておく．形態素解析ができなかった部分に関してはランダムに１と０の値を格納する．

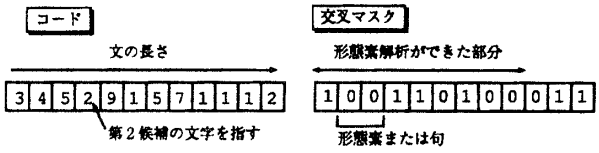

図 2: 染色体のコーディング

4.4.3 選択淘汰

各染色体の適合度はそれに対応する文章の自然言語処理からの評価値とし，適合度の高い染色体（文章）を残すように選択淘汰する．選択淘汰には適合度の上下幅が大きく，スケーリングが困難なためランキング法を用いた．

4.4.4 交叉

二つの染色体をランダムに選び図2に示した交叉マスクを用いて，既に組み合わされた形態素もしくは句をばらさないように交叉させる．現状では形態素の品詞接続関係を無視してしまうような交叉もありうる．

4.4.5 突然変異

染色体上の遺伝子どれか一つを別の文字候補に変えてしまう．交叉マスクから形態素などすでに組み合わされている単語を対象とせず，一文字の単語をその対象とする．以上の遺伝操作を繰り返すことによって評価関数の優れた，すなわち日本語らしい文章のみが次世代に次々と残されていくようになる．

5 まとめ

本研究においては文字認識の後処理を組み合わせ最適化問題としてとらえ，遺伝アルゴリズムを用いて計算量の削減と認識率の向上を試みた．今後は評価関数の見直し，品詞の接続関係を考慮した交叉などによってより少ない組み合わせで質の高い結果が求められるようにする．

参考文献

[1] 篠沢 大駒：手書き文字認識における大分類のための決定木生成法，情報処理学会第54回（平成9年前期）全国大会，Vol. 2-203.

[2] 杉村 利明:候補文字補完と言語処理による漢字認識の誤り訂正処理法，信学論(D-II) Vol.J72 pp.993-1000.