遺伝的アルゴリズムによる重要文節概念の獲得

大石　亨[a]　西尾　修一郎[a]　藤田　純[a]
遠藤　雅人[a]　奥村　学[c]　難波　英嗣[b]
[a] 明星大学　情報学部　電子情報学科
[c] 東京工業大学　精密工学研究所　知能化工学部門
[b] 広島市立大学　情報科学部　知能情報システム工学科

1　はじめに

電子化されたテキストの増大に伴い、テキストの自動要約の必要性が高まっている[奥村,難波1999,2002]。われわれの研究目標は、人間の文章要約と同様の処理を行うプログラムを進化させることである。人間の行う要約過程は、元文章に対するいくつかの基本的操作の組合せによって再現することができると考えられる。たとえば、

1. E氏はスポーツジムを経営している。
2. ジムには最新のドイツ製マシンが、ずらっと並んでいる。
3. テニスプレーヤーのN氏やバスケット選手のF氏が汗を流している。

という３つの文は、第1文の目的語と、第2文の主語を抽出して（第1文の残りの語句は連体化した上で）付加詞とし、主語を一般化した第3文につけ加えることにより、次の1文に要約することができる。

4. E氏の経営するスポーツジムでは、有名スポーツ選手が、最新のマシンで汗を流している。

このような処理を実現するためには、適用すべき操作を洗い出すとともに、その適用条件をあきらかにしなければならない。しかし、適用すべき操作と条件が決定されたとしても、その操作の適用個所や適用順序の組合せは多数あり、実時間での処理が不可能になってしまう可能性が高い。そこで、前処理としてあらかじめ重要個所を抽出して、問題空間（探索空間）を削減する必要がある。

本稿では、この前処理としての重要個所抽出の問題に、遺伝的アルゴリズム（GA）を適用して行った実験の結果を報告する。

2　GAと概念形成問題

本研究では、重要個所抽出の単位を文節とする。従来の研究では、抽出単位を文とする重要文抽出型のものが多いが、前節の例のように、できるだけ元の文章の情報を保存した要約を実現するためには、文あるいは節では粒度が荒いからである。また、抽出結果をそのまま抄録として用いるのではなく、あくまでも、文として再構成するための材料を提供するための処理として位置付けている。

重要文節の抽出は、要約中に存在する文節の概念形成問題と考えることができ、本質的には文節属性の組合せ最適化の性質を持つ。また、実用的側面においては、問題空間の構造、与えられる事例の数および質、要求される解答の性格など不確定な要素を多く含む。このことから、GAのような適応型探索手法が持つ計算能力および柔軟性が有効に働くと期待できる。

3　アルゴリズム

GAによる選言的概念獲得のアルゴリズムの枠組みは、遠藤らによって提案されている[遠藤,野沢,大内1995]が、実用的な問題には適用されていない。本稿では、第2回NTCIRワークショップの自動要約タスク(TSC1)[奥村,福島2000]で用いられた、30本の新聞記事とその要約(20%と40%の2種類)に対して、若干修正を加えた遠藤らのアルゴリズムを適用した。以下に、概略を示す。

STEP0: コーディング
STEP1: 環境設定（正例集合、負例集合）
STEP2: 正例を用いた初期世代集合の生成
STEP3: 各個体の適応度評価
STEP4: 適応度を基準とした個体の選択

STEP5: 個体対に対する交叉、突然変異
STEP6: 集団の入れ替え
STEP7: 集団の個体数回の入れ替え
➡世代数更新
指定世代数に達していなければSTEP3へ
STEP8: 最大適応度を示す個体を獲得連言集合に追加、この個体が説明する正例を削除、正例集合が空でなければSTEP2へ
STEP9: 獲得概念表示

次に、アルゴリズムの各ステップの詳細について説明する。

STEP0　コーディング

概念形成問題では、各遺伝子は各属性に対応し、染色体は単一連言概念に対応する[遠藤1997]。本研究では、元文章を係り受け解析し、得られた個々の文節を事例として登録する。それぞれの文節には、以下のような属性が与えられる。

- ◆ 位置情報
  - ➢ 文章内での文の相対位置
  - ➢ 文章内での段落の相対位置
  - ➢ 段落内での文の相対位置
  - ➢ 文内での文節の相対位置
- ◆ 文長
  - ➢ 文内文節数
  - ➢ 文内形態素数
- ◆ 主辞の情報
  - ➢ 品詞分類（IPA品詞体系）
  - ➢ 分類語彙表コード[中野1996]
  - ➢ 文章表題での出現の有無
  - ➢ TF（文章内出現頻度）
  - ➢ 活用形
- ◆ 機能語情報
  - ➢ 基本形（辞書見出し）
  - ➢ 品詞分類（IPA品詞体系）
  - ➢ 活用形
- ◆ その他の情報
  - ➢ この文節に係る文節の数
  - ➢ 文節内形態素数

これらの属性は、表1のようにコード化される。たとえば、文章内での当該文節が含まれる文の相対位置（文番号を全文数で割った数値）は、10段階に区分され、1から10までの数値が割り振られる。表中で、NCは、"not care"であり、この属性を評価しないことを表す。NCにはコード0が与えられ、最も一般的な属性を表す。これにより、すべての属性集合は0を根とする属性木として表現されることとなり、後述する一般化交叉、特殊化交叉等の操作が適用可能となる。なお、品詞分類と分類語彙表コードおよび機能語基本形は、それぞれ6桁と11桁の数値および辞書見出しをそのまま遺伝子として登録した。

表 1　属性のコーディング

文章内文位置	1～10, 0=NC
文章内段落位置	1～5, 0=NC
段落内文位置	1～3, 0=NC
文内文節位置	1～10, 0=NC
文内文節数 文内形態素数	1=1-5,2=6-10,3=11-15, ..., 0=NC
品詞分類	6桁(大中小細分類)
分類語彙表	11桁
表題出現の有無	1=有,2=無,0=NC
TF	1=1, 2=2-3, 3=4-5, 4=6-7, 5=8-9, 6=10～, 0=NC
活用形	1=基本形,2=未然形, ... , 0=NC
係る文節の数	1=0, 2=1, 3=2-3, 4=4-5, 5=6~, 0=NC
文節内形態素数	1=1, 2=2-3, 3=4-5, 4=6-7, 5=8~, 0=NC

STEP1: 環境設定

目的概念を獲得するための環境として初期正例集合および初期負例集合を設定する。これは、人間の作成した重要個所抽出要約に現れる文節を正例とし、それ以外の文節を負例として登録したものである。また、獲得概念集合を空とする。

STEP2: 初期世代集合の生成

概念形成問題では、個体の適応度を計算するために用いる正例集合が解の候補でもあることから、正例集合の要素を初期集団の個体とする。

STEP3: 個体の適応度評価

適応度関数は、[遠藤1997]で提案されている以下の式を用いた。

$$Fitness(pos, spc) = \max\left(0, \frac{(1+\sqrt{2})pos - spc}{\sqrt{2}N}\right)$$

ただし、pos, spcは、それぞれ、個体が説明する正例数、個体が説明する事例数を表す変数である。この適応度関数は、与えられたすべての正例を説明し($pos = N$)、かつ説明する事例がすべて正例である($spc = pos$)ような個体に最も高い評価($Fitness = 1$)を与える。また、正例密度(pos / spc)一定の方向と、正例の個数一定の方向で同じ傾きを取るような平面の式となっている。

STEP4: 適応度を基準とした個体の選択

各個体の適応度による存在確率に基づいて、ルーレット選択によって交叉対象となる2個体を選択する。

STEP5: 個体対に対する交叉、突然変異

STEP4で選択された2個体から新しい個体を生成する交叉において、より多くの正例を説明する個体を生成する一般化交叉と、正例密度を上げるためにより特殊な個体を生成する特殊化交叉を行う。前者は、両親の遺伝子よりも一般性の高い属性を持つ個体のうち最も特殊性が高いものを生成し、後者は両親の属性間に一般性に関する包含関係が成立していれば、その特殊な側の値を採用し、成立していなければどちらか一方の属性を採用することによって、両親より特殊な個体を生成する。

突然変異は、ランダムに選んだ属性の値を任意に選択する。

STEP6: 集団の入れ替え

STEP5で得られた2個体を、集団中で最も適応度の低い個体と置き換える。

STEP7: 世代数更新

集団の個体がすべて入れ替われば、すなわち、個体数回の入れ替えが行われたとき、世代が更新されたとする。指定世代数に達するまで、STEP3からSTEP6までをくりかえす。

STEP8: 獲得連言集合に追加

指定世代数に達したとき、集団の中で最も適応度が高い仮説を獲得連言集合に追加する。同時に、追加した概念が説明する正例を正例集合から取り除く。正例集合が空でなければSTEP2へ戻る。

STEP9: 獲得概念表示

獲得連言集合を選言で結んだものが獲得された選言概念となる。

4 実験結果と考察

実験には、第2回NTCIRワークショップの自動要約タスク(TSC1)で用いられた、30本の新聞記事（報道記事15、社説15）を用いた。それぞれの記事を係り受け解析し、各文節を染色体に変換した。正例は、課題A-2（人間の自由作成要約と比較可能な要約）の評価用に人間が作成した重要個所抽出要約に出現する文節である。要約率は20%と40%の2種類がある。

獲得連言を登録する指定世代を50、登録連言数の上限を50個として実験を行った。世代が進むにつれて獲得される概念は特殊化し、説明事例数が減少するからである。したがって、実行世代数の上限は2500世代となる。交叉確率を0.7、突然変異確率を0.9としたときの結果を表2に示す。評価尺度は、以下に示す再現率(R)、精度(P)、F値である。

$$\text{再現率 (R)} = \frac{\text{獲得概念が説明する正解文節数}}{\text{正解文節数}}$$

$$\text{精度 (P)} = \frac{\text{獲得概念が説明する正解文節数}}{\text{獲得概念が説明する文節数}}$$

$$\text{F値} = \frac{2 \times P \times R}{P + R}$$

表 2 30 cross-validationの結果

要約率(%)	再現率(%)	精度(%)	F値
20	18.7	36.7	0.226
40	62.3	49.1	0.541

表2は、30記事のうち29記事を訓練データ、残りの1記事を評価データとし、評価データを入れ換えて行った30回の試行の平均値である(30 cross-validation)。なお、評価データ1つごとに、乱数の種を変更して10回の試行を行ったうちの最良値を平均している。

実験に用いた30記事には、15本の報道記事と15本の社説が含まれている。そこで、このジャンルごとに15 cross-validationを行った結果を表3と表4に示す。

表 3　報道記事

要約率(%)	再現率(%)	精度(%)	F値
20	37.7	42.7	0.388
40	69.7	50.2	0.565

表 4　社説

要約率(%)	再現率(%)	精度(%)	F値
20	20.9	30.2	0.238
40	61.3	45.7	0.520

表1より、要約率が20%のとき、特に再現率が低いことがわかる。これは、正例数に対して負例数が増えるにしたがって、獲得される概念が負例を説明しないように特殊なものとなり、評価データを説明する確率が低くなるためである。一方、ジャンル別では、報道記事(表3)において、再現率、精度ともに向上しており、この差は統計的に有意である(t検定,p=0.01)。獲得された概念を見ると、文または段落位置の属性が1または2であり、その他の属性がすべて0というものが多い。このことは、文章の先頭付近の文を抽出していることにほかならず、報道記事におけるリード文の概念を獲得したといえる。これに対し、社説の場合は、位置情報にばらつきがみられ、どちらかといえば、文章の末尾の部分(文位置が10、段落位置が5)でかつ長い文に含まれる文節を表す概念が獲得されている。ただし、文長の属性が制約となり、評価データとマッチしなかったために、再現率の向上は報道記事と比較してわずかなものとなった。

一般的な傾向として、早い段階で獲得される概念は、説明正例数が多い、一般的な概念であり、位置情報・文長の情報のみ属性値が保存されている場合が多い。これらは位置情報による文抽出と変わりがないが、興味深い例として段落内文位置が1のものがある。これは、文章論でいうところのトピックセンテンスの概念である。世代数が進むにつれて、獲得概念は特殊なものとなるが、係り文節数の多い、主辞の品詞が動詞で機能語が助動詞であるような文節と、主辞の品詞が名詞または名詞概念であり、機能語品詞が助詞であるような概念が目立つ。さらに、機能語基本形で獲得概念中に保存されているのは、「は」「が」「を」「の」等の助詞と助動詞「た」のみであり、主題および格成分と、文末の動詞概念が獲得されている。

5　おわりに

本稿では、遺伝的アルゴリズムを用いて要約中に現れる重要文節の概念を獲得する実験を行い、TSC1で用いられた人手抽出データによる評価を行った。要約率が小さいときに再現率が低いという問題があるが、この問題には、[奥村,原口,望月1999]で提案されているような、負例をサンプリングした訓練データを用いる等の改善手法が有効であると考えられる。今後、属性の選択方法の見直しや抽出単位の拡張をしていく必要がある。

謝辞

係り受け解析に奈良先端大松本研究室で開発されたフリーソフト「cabocha」を使わせていただきました。記して感謝いたします。

参考文献

奥村学,難波英嗣.(1999) テキスト自動要約に関する研究動向.自然言語処理,Vol.6,No.6,pp.1-26.

奥村学,難波英嗣.(2002) テキスト自動要約に関する最近の話題. 自然言語処理, Vol.9, No.4, pp.97-116.

奥村学,福島孝博.(2000) NTCIR Workshop2の新しいタスクの紹介. 情報処理学会誌, Vol.41, No.8, pp917-920.

奥村学,原口良胤,望月源.(1999) 決定木学習を用いたテキスト自動要約手法に関するいくつかの考察, 情報処理学会第59回全国大会講演論文集, 第2分冊, 5N-2, pp.393-394.

遠藤聡志,野沢慎吾,大内東.(1995) 遺伝的アルゴリズムによる選言概念獲得アルゴリズム.人工知能学会誌,Vol.10,No.1,pp105-113

遠藤聡志.(1997) 遺伝的アルゴリズムの概念獲得問題への応用.北野宏明編『遺伝的アルゴリズム3』,産業図書,第10章,pp245-268

中野洋.(1996) 「分類語彙表」形式による語彙分類表（増補版）.