自 然 言 語 処 理 140－3
（2000．11．21）

携帯電話情報サービスのための新聞記事要約の研究

徳永　秀和　　青江　順一
高松高専制御情報工学科　徳島大学工学部知能情報工学科
E-mail：tokunaga@takamatsu-nct.ac.jp , aoe@is.tokushima-u.ac.jp

最近、携帯電話への情報配信のニーズが高まっており、新聞社では、記事を要約して配信している。この要約は、制限文字数内に抑える必要があり、要約率（要約文字数を元記事の文字数で割った値）が２０％から４０％である。したがって、本論文では、文節に重要度を付け、文字数をコストとし、コストを制限文字数以内とするナップサック問題として、要約文を作成する方法を提案する。そして、簡単な実験結果を示し、妥当な要約文が作成されたことを報告する。

A Study on Summarizing Newspaper article for a Service to The Portable Telephone

Hidekazu TOKUNAGA　and　Jun-ichi AOE
Dept. of Electro-Mechanical Systems Engineering, Takamatsu National College of Technology
Dept. of Information Science & Intelligent Systems, University of Tokushima

At present, a newspaper company provides the article summarized by the human being for the portable telephone. The number of the characters is restricted to the summary sentence which we make. And, a summary rate(The number of the characters of the summary sentence / The number of the characters of the article) is 40% from 20%.We propose how to make a summary sentence as a knapsack problem that a "bunsetsu" was considered goods. Then, an easy experiment result is shown, and it is shown that a suitable summary sentence could get it.

１．はじめに

最近、携帯電話が急速に普及し、Ｉモードなど携帯電話のインターネットサービスへの情報配信の需要が高まっている。こうしたなか、多くの新聞社が、新聞に印刷する前に、記事を携帯電話のインターネットサービスに配信している。また、地方の新聞社は、携帯電話以外にもケーブルテレビの文字放送にも、記事の配信を行っている。このような、携帯電話や文字放送への配信では、１記事あたりの文字数が制限されることになる。た

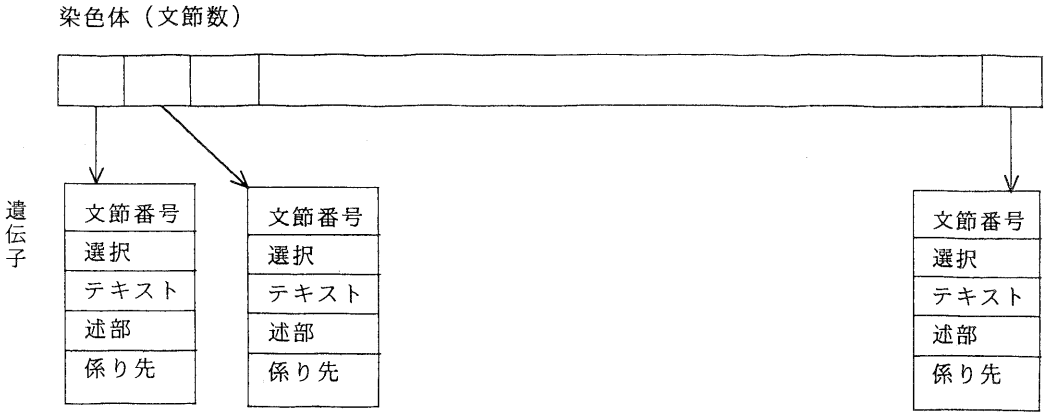

図１　染色体、遺伝子の構造

とえば、四国新聞社が高松ケーブルテレビに提供する文字放送用記事の場合は１３０文字以内と規定されている。したがって、記者が新聞用に書いた記事を制限された文字数に要約する必要がある。現在、この要約作業は、人手で行われ、かなりの労力を要しており、自動化することが求められている。よって、我々は、新聞記事を制限文字数に自動要約する手法を開発することにした。

現在、開発されているテキスト要約手法は、参考文献１[1]に詳しく紹介されているが、これらの要約と本要約の関係を簡単に述べる。開発されている要約は、要約率の大きな応用を目的として、重要文を抽出するものと、ニュースの字幕など小さな要約率を応用として、一文からの削除部分を決定するものの２つに大別される。本研究は、要約率が２０％から４０％であり、この２つの中間的な場所に位置し、これら２つの手法の両方が必要となる。すなわち、重要文を抽出するとともに、その文のあまり重要でない文節を削除し、より多くの文を選択できるようにする必要がある。

本研究の要約の特徴は、要約率があまり高くなく、文でなく文節の抽出が必要である。したがって、この要約を、文節を１つの品物と考えた、ナップサック問題と考えることができる。よって、各文節の重要度を計算し、文節の文字数をコストとし、コスト制約として要約文字数を導入したナップサック問題を解くことになる。ただし、選択した要約文の評価値は、文節の重要度を加算しただけでは、駄目な場合もある。これは、ある２つの文節が選択された場合、冗長となてしまうことが起こりえるからである。したがって、文節の組み合わせに対して評価点を減点する必要がある。また、制約条件として、文字数以外に文としておかしな文節の組み合わせを禁止する必要がある。このようなナップサック問題の変形の解法として遺伝的アルゴリズム[2]を採用した。

以下、遺伝的アルゴリズムの適用について述べたのち、記事の要約事例を分析し、文節の重要度の決定方法について述べる。そして、最後に実験結果を示す。

２．GA の適用

２．１．染色体（コード化）と初期個体

遺伝子は、文節番号、要約として選択されるか、文節のテキスト、述部であるか、係り先の文節番号を持つものとする。そして、染色体は、文節の数だけの遺伝子を持つものとする(図１)。よって、染色体の要約として選択される遺伝子のテキストを出力すれば、その個体の表現する要約文となる。

初期個体は、各遺伝子の要約として選択されるかを乱数で決定し、染色体のランダムな位置に遺伝子をセットする。そして、後の節で示す制約条件を満たす処理を行う。

２．２．交叉と選択

図２に示すように、交叉させる２点（P1,P2）をランダムに決める。親Ｂの交叉させる遺伝子の文節番号と同じ文節番号を持つ遺伝子を子Ａより

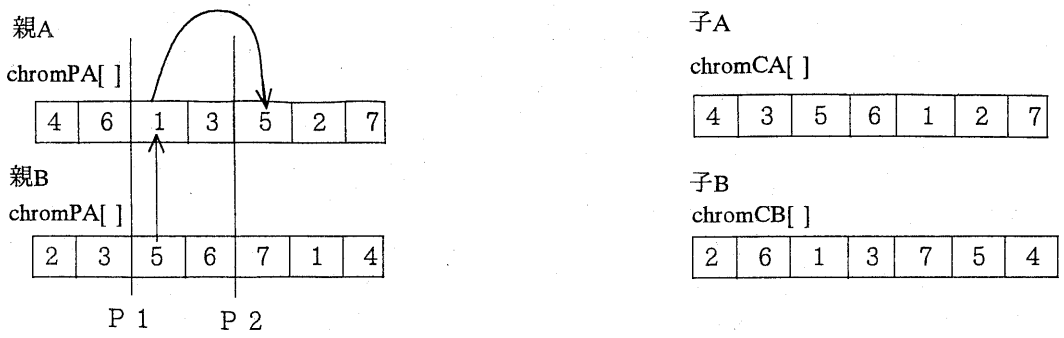

図２　交叉の仕組み

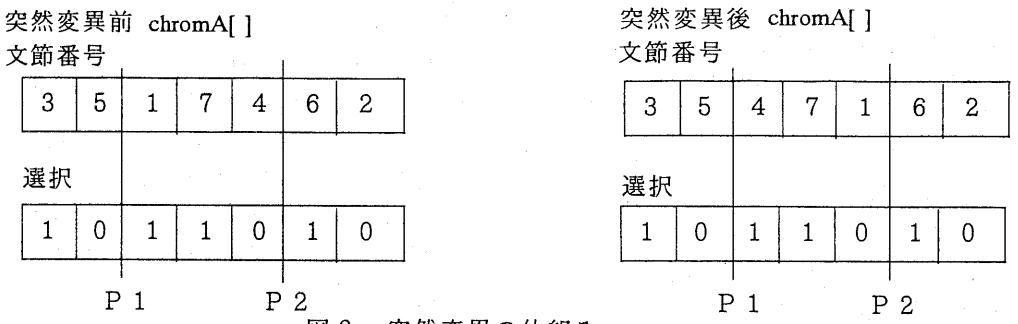

図３　突然変異の仕組み

探し、その位置に、交叉により追い出される遺伝子を移動する。そして、子Aの交差位置に親Bの交叉位置の遺伝子を持ってくる。この操作をすべての交叉位置に行う。最後に後の節で示す制約条件を満たす処理を行う。

親の選択は、ルーレットホイールアルゴリズム[2]で行う。

２．３．突然変異

突然変異は、図３に示すように、突然変異させる位置（P1,P2）をランダムに決める。そして、遺伝子を逆転させる。このとき、遺伝子の要約として選択されるかの値を、もとの染色体の位置の値となるよう各遺伝子の値を変更する。

３．４．制約条件を満たす染色体の修正

次のような条件を満たすように染色体を修正する。

（１）述部は、名詞句の文節が最低１個ないと選択されない。

（２）係り先の文節が選択されていない文節は選択しない。

（３）要約文が制限文字数を超えない。（染色体の先頭の遺伝子より要約文に追加して行く）

（４）制限文字数以下のなるだけ多い文字数を選択する。（選択されている遺伝子の文字数が制限文字数より少ない場合は、染色体の先頭の遺伝子より、文字数を越えないよう要約文に追加して行く）

（３）、（４）の処理より、染色体の先頭に近い遺伝子ほど、要約文に選択される優先順位が高いことになる。

３．記事要約の特徴

四国新聞社のホームページに記載された２０００年６月の第１週の要約記事７１文を解析し以下のような特徴を見いだした。

３．１．文の重要度についての解析

（１）第一段落は、記事の要約であり、第一段落の文は、最優先に要約文に選択される。

(a)第一段落が制限文字数を超える記事１７個の

主題	選択文の表現する属性
コンテストの募集	問合せ先、応募資格、審査内容
展示会（コンサート）の開催	展示内容、問合せ先、様子
イベントの開催	日程、様子
キャンペーンを行う	趣旨
講習会の開催	様子、日程、参加者
説明会の開催	出席者、
つどいの開催	出席者、趣旨
会議の開催	決定事項、出席者
指導（取り締まり）を行う	場所、指導内容
モデル建築物に指定	基準
花が見頃を向かえた	期間
選挙日が決まった	日程

図４　主題に対する重要な属性

うち、第一段落以外の文が選択された要約文は２個しかない。

(b)第一段落が制限文字数以内の記事５４個のうち、第一段落の文が選択されなかった要約文は１個しかない。

（２）単純に前の方の文から順番に選択されることが多い

(a)７１記事のうち、単純に前の方の文から順番に選択された要約文が４０個ある。

（３）文の順番を無視し、後ろから選択された文は、記事の主題の重要な属性を述べた文である。

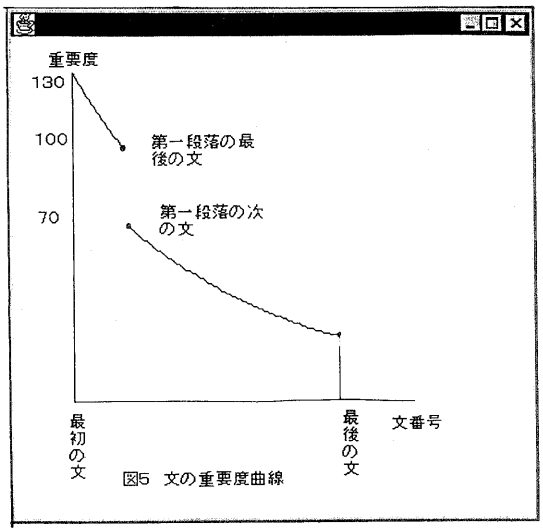

図5　文の重要度曲線

(a)記事主題が、”レンブラント版画展の開催”ならば、選択された文は、”展示内容”を述べた文である。

(b)記事主題が、”写真コンテストの募集”ならば、選択された文は、”応募資格”を述べた文である。

主題、属性を概念化し、主題と選択された文の表現する属性を表にしたものを図４に示す。

３．２．文節の削除についての解析

（１）制限文字数内の第一段落より、削除される文節は、「対話の県政を目指す知事」の「対話の県政を目指す」ような連体修飾が３個、「食事会を通し、孤独にならない生活を呼びかけた。」の「食事会を通し、」ような主文の説明が２個、「研修生を招待して、”交流会を開催”」の「交流会を開催」のような連用の並列の後半部が３個である。

（２）５７記事の第一段落以外の文から削除される文節は以下のような数である。

連帯修飾が７、主文の説明が５、連用の並列後ろ部分が８、連用の並列前部分が２、「AやBなど」のような並列句の一部が７、”は”、”も”格の文節が２、”が”格の文節が１、”で”格の文節が５、”に”格の文節が２、”の”格の文節が４

３．３．特別に重要度を上げるもの

電話番号や「　」で囲まれた部分は特に採用されることが多い。

４．文節の重要度の決定

前節に述べた要約文の特徴より、次のように文節の重要度を決定する。

（１）主題文の重要な属性を述べた文を第一段落の次の文として、文の順番を入れ替える。

（２）文の持ち点を次式で求める

(a)第一段落は最後の文が１００となるように以下の式で求める。

100*exp(0.25)*exp(-0.25*(A/B))

Aは、文番号－１

Bは、第一段落の文数－１

だだし、第一段落の文数が１の時は

100*exp(0.25)*

とする。

(b)第一段落以降の文は最初の文が７０となり最後の文が１／ｅとなるように次式で求める。

段落	文	係り	述部	格	並列	テキスト	重要度
1	1	3	-1	1	-1	真鍋知事が	64
1	1	3	-1	3	-1	市町代表らと	44
1	1	4	0	-1	-1	意見を交わす	0
1	1	5	-1	6	-1	本年度の	64
1	1	8	-1	1	-1	「知事と県政を語るつどい」が	128
1	1	8	-1	5	-1	二日、	128
1	1	8	-1	3	-1	丸亀市と山本町を皮切りに	89
1	1	-1	4	-1	-1	スタート。	0
1	2	10	-1	2	-1	対話の県政を	50
1	2	11	0	-1	-1	目指す	0
1	2	15	-1	3	-1	知事に、	70
1	2	15	-1	8	-1	さまざまな	0
1	2	15	-1	1	2	意見や	50
1	2	15	-1	1	1	要望が	100
1	2	-1	4	-1	-1	寄せられた。	0
2	3	23	-1	0	-1	丸亀市では	70
2	3	18	-1	6	-1	大手町の	35
2	3	23	-1	4	-1	ひまわりセンターで	49
2	3	22	-1	1	3	商工、	35
2	3	22	-1	1	2	ボランティア、	52
2	3	22	-1	1	1	まちづくり団体	70
2	3	23	-1	1	-1	などの代表十五人が	70
2	3	-1	4	-1	-1	出席。	0
3	5	32	-1	0	-1	各代表からは	50
3	5	32	-1	1	2	「県立丸亀競技場のマラソンを、ぜひフルマラソンに」	25
3	5	32	-1	1	1	「統合する保健所の設置場所はどうなるのか」	50
3	5	31	-1	1	3	教育、	25
3	5	31	-1	1	2	福祉、	37
3	5	31	-1	1	1	文化問題	50
3	5	32	-1	1	-1	などをめぐり	50
3	5	32	-1	1	-1	二十項目近くが	50
3	5	-1	4	-1	-1	出された。	0
4	6	37	-1	0	-1	真鍋知事は	42
4	6	37	-1	2	-1	フルマラソンについて	42
4	6	37	-1	3	2	「資金が必要で、スポンサーを集めなくてはならない。 公認コースが生かされるよう研究してみたい」	14
4	6	37	-1	3	1	「保健所の統合は決まっているが、場所はは未定。 公平公正に決めていきたい」	29
4	6	-1	4	-1	-1	などと答えた。	0
5	4	44	-1	0	-1	山本町では、	59
5	4	44	-1	3	-1	保健センターに	41
5	4	41	-1	4	-1	公募で	20
5	4	42	0	-1	-1	選ばれた	0
5	4	44	-1	1	2	住民ら十七人と	29
5	4	44	-1	1	1	大橋町長らが	59
5	4	-1	4	-1	-1	出席した。	0
6	7	48	-1	0	-1	会議では、	35
6	7	48	-1	2	-1	ごみ問題について	35
6	7	48	-1	3	-1	「野焼きやごみの不法投棄に、厳しい規制を」との	24
6	7	-1	4	-1	-1	要望があった。	0
6	8	52	-1	0	-1	豊島問題では、	30
6	8	52	-1	1	-1	知事から	30
6	8	52	-1	3	-1	「県民の皆さんに大変迷惑を掛けたが、 解決に向けて進み出したことを理解してほしい」と	21
6	8	-1	4	-1	-1	説明する場面もあった。	0
7	9	55	-1	0	-1	高齢者問題では	25
7	9	55	-1	3	-1	「健康生きがい中核施設などの整備を通じ、 元気で長生きできる町づくりに力を入れたい」と	17
7	9	-1	4	-1	-1	理解を求めた。	0

図６　文節の表

70*exp(-(C/D)

Cは、第一段落以降の文番号－１

Dは、第一段落以降の文の数－１

文の持ち点のグラフを図５に示す。

（３）連体修飾、主文の説明部分、連用並列の後半部分の文節は、文の持ち点に０．５を掛ける。

（４)”で”格の文節、は０．７、”の”格の文節に０．５をそれぞれ文の持ち点掛ける。

（５)”ＡやＢなど”の並列句は、前部Ａの部分の重要度を２分の１の値とする。並列句が３個以上ある場合は、最前部を２分の１とし、中間の句は線形補完する。

５．実験

文節の分解、文節の係り先などは、人手で行い、図６に示すような表を作成する。そして、この表より、４章に従い、各文節の重要度を計算する。計算結果は、図６の右端の列に示してある。図６の表を入力とし、３章の遺伝的アルゴリズムによって、最適解を求めた。ここで、個体数は１００、世代数は５０、交差率は０．５、突然変異率は０．１である。図７には、各世代の最大適合率と平均適合率を示す。得られた要約文は「本年度の「知事と県政を語るつどい」が二日、丸亀市と山本町を皮切りにスタート。知事に、意見や要望が寄せられた。丸亀市では商工、ボランティア、まちづくり団体などの代表十五人が出席。各代表からは福祉、文化問題二十項目近くが出された。山本町では、大橋町長らが出席した。」である。また、ホームページに記載されていた要約文は「真鍋知事が市町代表らと意見を交わす平成12年度の「知事と県政を語るつどい」が２日、丸亀市と山本町を皮切りにｽﾀｰﾄ。知事にさまざまな意見や要望が寄せられた。丸亀市では商工などの団体の代表 15 人が出席。山本町では、公募で選ばれた住民ら 17 人と大橋町長が出席した。」である。

６．考察

プログラムが出した要約文と人手で作成した要約文の主な違いは、人手に「公募で選ばれた住民ら 17 人と」があり、プログラムに「各代表からは福祉、文化問題二十項目近くが出された。」がある点である。これは、プログラムでは、連体修飾節が、文字数と重要度の両方の点でかなり不利になるためである。この差異より、人間の要約では、多くの文より細々拾うよりは、絞り込まれた文のなかで文字数の調整を行っていると考えられる。一方、本システムでは、ナップサック問題の最適化を採用したために、より多くの文より文字数の少ない高得点の文節を拾おうとする傾向が強くなっていると考えられる。以上より、より人間に近い要約を得るためには、文字数が多い文節があまりにも不利になりすぎないような工夫が必要となる。

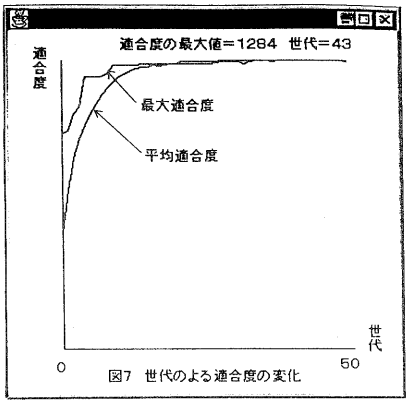

図7 世代のよる適合度の変化

７．おわりに

携帯電話や、文字放送への新聞記事の配信のため、記事を制限文字数内で要約するための手法を提案し、実験より本手法が有効であることを示した。本手法は、文節に重要度を与えたナップサック問題として、要約文作成をおこなった。その結果、人間の要約文より、多くの文より細々と文節をとってしまう、また、文字数の多い文節が不利になりすぎるなどの問題が生じることがわかった。今後は、こられの点を改善するとともに、文節の重要度を機械学習により決定する方法を開発していく予定である。

参考文献

[1] 奥村 学、難波 英嗣“テキスト自動要約擬お術の現状と課題” JAIST Research Report ,IS-RR-98-0010I,北陸先端科学技術大学院大学情報科学研究科（1998）

[2]坂和 正敏、田中 雅博“遺伝的アルゴリズム”朝倉書店（1995）