情 報 学 基 礎 57−11
自 然 言 語 処 理 136−11
（2000. 3. 22）

# 段落中の接続関係と段落間の重要度配分による文章要約

小堀 誠† 田村 直良††

† 横浜国立大学 工学研究科 電子情報工学専攻
†† 横浜国立大学 教育人間科学部 情報認知システム講座
{kobori,tam}@tamlab.dnj.ynu.ac.jp

重要文を抜き出す「抄録」と呼ぶ手法を採用した自動要約では、原文で用いられた文をそのまま出力するため、文が長くなりがちで、文間の意味的なつながりが悪い。また、短い文章の場合、良い結果は得られないといった問題点もある。そこで本研究では、一般文章に対応できる自動要約を目指し、文章全体の意味（構造）解析を行わず、文章中のパラグラフの重要度、パラグラフ間の類似度に応じて要約率を配分する。そして、パラグラフ内のみの句、節、文の接続関係による構造解析から得られた情報をもとに、一文ごとに必要な文節の抽出を行い、これらを連結して要約文章を作成する手法を提案する。

# Automatic summarization based on the rhetorical structure of each paragraph and the order of importance between paragraphs

Makoto Kobori† Naoyoshi Tamura††

†Department of Electrical and Computer Engineering,
Yokohama National University

††Department of Information and Cognition Systems
Faculty of Education and Human Science
Yokohama National University

{kobori,tam}@tamlab.dnj.ynu.ac.jp

In this paper, we present a summarization method for domain independent texts. Many researches on automatic summarization are presented for rather formal texts such as newspaper editorial. Moreover, the relation between adjacent sentences is not always coherent in the extracted summary. Our system, firstly, allocate the summarization ratio to each paragraph. Then the system constructs the rhetorical structure of the paragraph, in which complex sentences are decomposed into simple sentences, the sentences into clauses, and so on. The summarization is based on the extraction of important sentence clauses, which are decided according to the decision tree with parameters on the structure and superficial characteristics, trained by C4.5.

## 1 はじめに

本研究では、一般文章を対象とした要約システムの実現を目指す。文章を構成するパラグラフごとに、節、句、文間の関係を考慮して構造化を行い、構造上の特徴をパラメタとして用いた判定法によって重要文節を抽出し、要約文を生成する手法を提案する。

近年のインターネットやCD-ROMといった電子媒体の発達により、電子化された文書が大量に存在する。これらの文書から必要な情報を効率よく得るために、情報検索等の研究が行われ、文書の主旨を短時間で把握するために、キーワード抽出、文章抄録、文章要約といった研究が行われている。このとき、大量の文書を高速に処理するためには、深い意味解析を行わずに文章の表層的特徴から解析を行う方法が有効である。

一般に文章要約とは原文の大意を保持したまま、短い文章で表現する事を意味し、抄録とは文章から重要と思われる文を抽出することを意味する。これまで抄録の範疇で、照応処理などによって文整形を行ない、要約文を生成するシステムは扱いやすいという理由から数多く研究されてきた。

山本ら [9] は、照応、省略、語彙による結束性など多くの談話要素から重要文を選択していく論説文要約システム (GREEN) を発表している。このシステムはヒューリスティックスにより要約文を生成している。

比留間ら [3] は新聞社説を対象に、計算機により文章構造を解析し、得られた文章構造上の情報を基にした判定式により重要文を抽出、さらに照応解析処理、一文内の冗長部分削除による整形を行う要約システムの研究を行った。

難波ら [6] は抄録手法によって作られた要約文をより読みやすいものにするため、心理実験によって読みにくさの要因を分析し、その結果に基づき、抄録中の欠落情報を補ったり不要な箇所を削除することによって書き換えを行ない、抄録の読みやすさを向上させる研究を行なっている。

しかし、抄録手法を用いて作られた要約文は原文をそのまま用いるため、一文は長くなりがちで、離れた場所から文を選び隣接させ一つの文章とするため、原文章の結束性が崩れ、文間の隣接関係が不自然になる場合がある。また、比留間らの手法では文章の構成を意識したもので、人間の文章抄録作成の流れに似たものと思われるが、新聞社説のみを対象としているため、汎用性に欠ける。

上田ら [7] は句表現要約手法に基づく要約を行なっている。これは重要単語を含んだ短い句の並びを列挙することによって一目でわかる要約を作るものである。しかし、重要単語の列では、文としての機能を持っていないため、要約文としては不十分である。

一般の文章を対象に自動要約を考えた場合、話題や論理の展開の仕方、形式、長さなどの文章の特徴（文章構造）が一様でないため、計算機による文章構造の把握は非常に困難である。そのため、文章構造解析に基づいた抄録手法 ([3]) による要約は有効でない。

そこで本研究では、要約の処理単位をパラグラフとし、パラグラフ間で重用度を配分することにより各パラグラフの要約率を決め、文章全体を要約する手法を提案する。

実際の方法として、文章に出現する名詞の $tf*idf$ 値 [2] からパラグラフの重要度、パラグラフ間の類似度を求め重要度を決定し、要約率を配分する。パラグラフ内の句、節、文の接続関係による構造解析から得られた情報を基に、一文ごとに必要な文節の抽出を行い、要約文章を作成する。重要文節の抽出処理は機械学習システム C4.5[4] を用いて行なう。

機械学習のための訓練、およびシステムの評価のための訓練、評価データは被験者 10 人に対する 1000 編の要約調査による。実験の対象とした文章は、特定の分野によらないシステムの実現の理由から、新聞社説、新聞経済記事、新聞特集記事、本研究室卒論、ウエブ上で公開されている論文、ウエブサイトの情報技術関連記事用いる。

## 2 パラグラフのモデル化

本研究では、パラグラフとは文章中で一つの意味的なまとまりをもった文の集まりであると考える。パラグラフを理解し、要約を行なうには、そのパラグラフ中の文、節、句、文節がどのような構造になっているか把握する必要がある。

ここでは、パラグラフの構造解析に先だって、パラグラフを構成する要素と、要素間の関係を木構造として表現するモデルを示す。

### 2.1 パラグラフの構成

パラグラフは複数の文で構成されており、文は節によって構成され、節は句によって、そして句は意味的な最小単位である文節によって構成されている。

本研究では、パラグラフの要約の立場から、パラグラフ、文、節、文節、句を、以下のように定義する。

**パラグラフ** 一つの意味的なまとまりである。本研究では形式段落で代用する。一つ以上の文によって構成されている。

**文** あるまとまった内容を持ち、形の上で完結した単位である。文には単文と複文がある。

**単文** 単一の述語を中心として構成された文である。

**複文** 中心的な役割を担う文末に述語を含む主節と、それに特定の関係で結びつく接続節によって構成された文である。

**文節** 「用言」または「用言＋助動詞」または「体言＋助詞」。

**句** 述部である文節が最後に位置する文節の列

### 2.2 パラグラフの構造モデル

本研究ではパラグラフの特徴をとらえるために、パラグラフの構造モデルを提案する。

パラグラフを構成する文の構造化として、Mann[5]の修辞構造理論を基に、木構造として表現する手法を用いる。また、文中の節の係り関係について、「原因理由」、「条件」、「順接」、「逆接」、「並列」で結ばれる節を二分木によって構造化し、修辞関係と同等に扱う。

パラグラフの構造は、パラグラフを構成する文が存在する文レベル、文を構成する単文、または複文の節が存在する節レベル、単文、または節を構成する句が存在する句レベル、そして、句を構成する文節が存在する文節レベルの4つの階層からなる。(図1)

**文レベル** パラグラフの最上位レベルである。文レベルの構造は、文どうし、あるいは文と構造化された文のセグメントを修辞関係により関係づけた再帰構造として表す。本研究では、簡単のためにセグメントを衛星、文を核と仮定し、左下がりの構造木とする。修辞関係には以下のものがある。

結論、根拠、一般化、逆接、対比、転換、相反、条件、強調、理由 説明 換言 例示 順接 並列 累加

**節レベル** 文が複文である時、複文中の節間の関係を、節どうし、あるいは構造化された節のセグメントとそれに続く節の節間関係によって結ばれた再帰構造として表す。本研究では、構造化された節が、それに続く節にかかるものと仮定し、左下がりの構造木とする。節間関係は、基礎日本語文法[8]を参考にした。本研究では副詞節のみ扱う。節間の関係として以下のものがある。

順接節、条件節、原因理由、逆接節、並列節

**句レベル** 句レベルでは、句どうし、あるいは構造化された句のセグメントとそれに続く句が、句間関係によって結ばれた再帰構造として表す。句間の関係はつぎのものがある。

連体修飾、連用修飾

**文節レベル** パラグラフの最下層レベルで、葉は文節である。文節レベルは文節のリスト構造となっている。

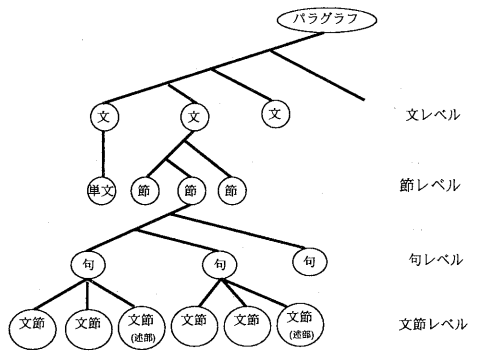


図 1: パラグラフの構造とレベル

## 3 重要文節抽出モジュール

本研究では機械学習システム C4.5[4] を用いて、重要文節の抽出を行なう。重要文節の抽出は、C4.5によって出力された採用不採用の判定、そして要約率の調整は採用不採用判定の確信度に基づく。

機械学習システム C4.5 の訓練には、重要文節抽出の観点から、文レベル、節レベル、句レベル、文節レベルにパラメタを設定する。

**文レベル**

**bunkan** 文間の修辞関係をあたえる
理由、説明、換言なら bunkan_p1
結論、根拠、一般化なら bunkan_p2
例示、順接、並列、累加、選択な

ら bunkan_p3
逆接、対比、転換、相反なら bunkan_p4
条件なら bunkan_p5
強調なら bunkan_p6

**type*** 文末タイプが意見なら iken
叙述なら jojutu
断定なら dantei

**jisei*** 文の時制が現在なら genzai
過去なら kako
未来なら mirai

節レベル

**setukan** 節間の関係が原因理由の関係なら geninriyuu
順接節なら junsetu
逆接節なら gyakusetu
条件節なら jouken
並列節なら heiretu

**inyo** カギ括弧(「」)で囲まれた引用なら、1

句レベル

**kukan** 連体修飾しているなら、rentai
連用修飾しているなら、renyou

文節レベル

**kakujo*** 文節が格助詞を持っているなら、1

**toritate*** 文節がとりたて助詞を持っているなら、1

**ha*** 助詞文節が助詞「は」を持っているなら、1

**bunmatu*** 文節が文末に位置するなら、1

**keisiki*** 文節が形式名詞を持っているなら、1

**meishityp*** 文節が名詞を持っている場合、
組織名なら sosiki
数字なら sosiki
普通名詞なら futuu
固有名詞なら koyuu
サ変名詞なら sahen
副詞的名詞なら fukusi
時相名詞なら jisou
上記に該当しないなら sonota

**nvkyori*** 最短の述部までの距離(正規化により[0..1])

**umekomiv*** 埋め込み文の述部であるなら、1

**impnoun*** 名詞がある場合、その $tf*idf$ 値

**position*** 文中の文節の位置(正規化により[0..1])

## 4 推敲過程

文節選択処理によって、文節が採用され、文が生成される。しかし、文法的に不適切な文が生成されることもあるため、以下のルールに従い、推敲を行なう。推敲は不要文節の削除と必要文節の再採用の処理からなる。

- 原 文において、離れた場所に存在する動詞が、要約文において連続した場合、後半の動詞を削除する。
- 括 弧 () で括られた文節の列は削除する。
- 一 文から一文節の抽出の場合、その一文を削除する。
- 形 式名詞、抽象名詞[1]の直前の文節が抽出されていない場合、その文節を採用する。
- 述 部の直前の名詞句を採用する。
- そ の他
  - 読 点が連続した場合、一つにする。
  - 読 点と句点が連続した場合、句点のみにする。
  - 文 頭に句読点がある場合、削除する。

## 5 文章要約

前章ではパラグラフの要約について述べた。本章ではパラグラフの要約を用いた文章全体の要約について述べる。本研究では、すべてのパラグラフを均一に要約するのではなく、ある観点に基づいて重要度の順位付けを行ない、要約率を配分する。

### 5.1 要約率配分モジュール

本研究では名詞の $tf*idf$ 値[2][2] を用いてパラグラフごとにトピックベクトルを求める。トピックベクトルとは、文章全体に出現する名詞の異なり数を要素数とし、それぞれのパラグラフ内の名詞の $tf*idf$ 値を要素とするベクトルである。トピックベクトルの大きさと内積の値によって、文章に存在するパラグラフの重要順位をもとめ、その順位に基づいて要約率を配分する。

[1] 分類語彙表 [1] による。
[2] 実験では訓練評価データである 1000 のパラグラフを用いた。

#### 5.1.1 トピックベクトル

文章Dを構成する各パラグラフをトピックベクトル $P_1, \cdots P_m$ で表す。mはパラグラフの個数とする。トピックベクトル $P_i$ は

$$P_i = (N_{i1}, N_{i2}, \cdots, N_{in}) \quad (1)$$

で表される。ここでnは全パラグラフに出現する名詞の異なり数とする。また、

$$N_{ij} = \text{パラグラフ}i\text{における名詞}N\text{の}tf*idf\text{値} \quad (2)$$

とする。

#### 5.1.2 パラグラフ間の類似度

トピックベクトル $P_1, \cdots, P_m$ のすべての組合せに対し、式(3)を用いて類似度を計算する。ここで、$V(P_i)$ は正規化された $P_i$ である。この式はベクトル $V(P_i)$ と $V(P_j)$ の内積を表す。$sim(P_i, P_j)$ の値が大きいほど、$P_i$ と $P_j$ は類似している事を示す。

$$sim(P_i, P_j) = \frac{V(P_i) \cdot V(P_j)}{|V(P_i)||V(P_j)|} \quad (3)$$

#### 5.1.3 重要パラグラフの重要度順位付け

パラグラフの重要度の順位付けは、何を重要と考えるかで、方法が異なる。例えば、文章中のすべての話題を網羅したいように順位をつける時と、文章中の主旨に関するものを中心に順位をつける時では異なる。そこで、本研究では以下の3つのモデルを提案する。

**モデル1** パラグラフベクトルのスカラー値の大きい順。
名詞の重要度と出現頻度から求めた順位である。重要なパラグラフを中心とした要約が可能となる。

**モデル2** トピックベクトルのスカラー値が最大のパラグラフを最重要パラグラフとし、上位パラグラフとの距離が近い順。
上位パラグラフすべてから内容が近い順である。これにより、最重要パラグラフの内容を中心とした要約が可能となる。

**モデル3** トピックベクトルのスカラー値が最大のパラグラフを最重要パラグラフとし、上位パラグラフとの距離が遠い順。
上位パラグラフすべてから内容が異なる順である。これにより、文章中に述べられたあらゆる内容について網羅した要約が可能となる。

#### 5.1.4 要約率配分

前述したモデルによって順位づけされたパラグラフに、要約率を配分する。最重要パラグラフの要約率を最大に、最下位のパラグラフの要約率を0とし、線型に要約率を配分した場合、中位のパラグラフは少々削られすぎである。

中位のパラグラフまではより多くの情報を残すのが理想である。そこで本研究では式(4)によって、要約率を求める。この式は重要度の高いパラグラフから情報を落さないように要約率を低く、重要度の低いパラグラフは要約率を高くする。また、3割以下の要約率となったパラグラフに関しては重要な情報がないものと判断し、要約率を0、つまりパラグラフを不採用とする。

$$Rate_{yoyaku} = Y * \frac{log(N+1-i)}{log(N)} \quad (4)$$

$Y$: 最大要約率
$N$: 文章に存在するパラグラフの個数
$i$: パラグラフの順位

## 6 実験と評価

### 6.1 重要文節抽出の訓練

実験にあたって、まず訓練と評価のために、重要文節の選択による要約アンケートを実施した。これは学生10人を被験者とし、新聞の社説(at)、経済記事(ec)、特集記事(sp)、そして、卒業論文(bt)、WWW上にある情報技術関連記事(wi)と論文(wp)からランダムに選んだ1000個のパラグラフを用いて、長さが5割以下で、かつ意味が伝わるようにするために、重要であると思う文節を選ばせた[3]。また、被験者間での重要文節抽出の調査結果を比較するために、被験者全員に、10パラグラフずつ同一

[3] 作成された訓練データの平均要約率は66.7%であった。

の文章を要約対象とした内容のアンケートを実施した。

### 6.2 実験方法

次の実験を行なう。

**実験 1** 構造解析を行なわないパラグラフ要約を要約手法 1、本論文で提案しているパラグラフの構造解析を行ったパラグラフ要約を要約手法 2 とし、両手法の比較検討を行なう。

**実験 2** 要約手法 2 によって生成されたパラグラフの要約文の評価と検討を行なう。

**実験 3** 要約手法 2 を用いて文章要約を行ない、要約文章に関する評価アンケートを実施し、評価検討を行なう。各パラグラフへの要約率の配分は、前章で提案した 3 つのモデルについて行なう。

### 6.3 実験 1 に関する評価

ここではパラグラフの構造化が決定木の精度にどれだけ寄与しているか評価を行う。構造解析を行なわない要約手法 1 では、3章で示したパラメタのうち、* 印のついた構造化と関係のないものを用いて訓練し、実験を行なった。

要約手法 1、要約手法 2 の両手法について、誤り率の比較を行なう。表 1に、手法 1 と手法 2 の枝刈前と、枝刈後の訓練データに対する誤り率 (1)、テストデータに対する誤り率 (2) を示す。また、枝刈後の括弧内の数値は誤り率の推定値である。 訓練データに対する枝刈前の誤り率を比較する。カテゴリ at、 sp、 wi においては、手法 2 の誤り率は手法 1 の約 5 割、カテゴリ bt、 ec、 wp、 all[4]においては、手法 2 の誤り率は手法 1 の約 6 割となっている。誤り率の低下が顕著に表れている。

テストデータに対する枝刈前の誤り率を比較する。手法 2 の誤り率は手法 1 よりも小さい値となっている。

手法 2 のテストデータに対する誤り率を比較する。枝刈後の値は、枝刈前よりも小さな値となった。

[4]all は at,bt,ec,sp,wo,wp を一つにまとめたもの

| | 手法 1 | | 手法 2 | |
|---|---|---|---|---|
| | 枝刈前 | 枝刈後 | 枝刈前 | 枝刈後 |
| at(1) | 25.9% | 28.7%(33.7%) | 13.9% | 19.5%(28.9%) |
| at(2) | 37.4% | 36.5%(33.7%) | 32.0% | 30.9%(28.9%) |
| bt(1) | 36.3% | 37.4%(39.9%) | 23.1% | 27.2%(33.7%) |
| bt(2) | 40.4% | 40.1%(39.9%) | 37.5% | 35.5%(33.7%) |
| ec(1) | 35.9% | 37.3%(40.0%) | 20.9% | 25.5%(33.5%) |
| ec(2) | 41.7% | 40.8%(40.0%) | 37.5% | 36.0%(33.5%) |
| sp(1) | 29.9% | 32.2%(36.8%) | 14.7% | 19.2%(30.1%) |
| sp(2) | 37.8% | 37.4%(36.8%) | 33.1% | 31.7%(30.1%) |
| wi(1) | 30.0% | 32.5%(37.5%) | 16.2% | 20.6%(30.8%) |
| wi(2) | 40.8% | 40.3%(37.5%) | 34.8% | 33.1%(30.8%) |
| wp(1) | 36.7% | 37.9%(39.8%) | 25.2% | 28.8%(35.8%) |
| wp(2) | 40.2% | 40.2%(39.8%) | 38.2% | 37.0%(35.8%) |
| all(1) | 38.0% | 39.3%(40.3%) | 27.5% | 30.9%(36.2%) |
| all(2) | 40.6% | 40.2%(40.3%) | 38.8% | 37.8%(36.2%) |

表 1: 手法 1 と手法 2 の誤り率の比較

以上から、本研究で提案する手法は、文節の選択においてより精度の高い結果をもたらしたと言える。

### 6.4 実験 2 に関する評価

ここでは、要約手法 2 によってパラグラフの要約を行ない、最終的に生成された要約文章に対して検討を行なう。

訓練データを用いた要約文の生成には、実験 1 において最も誤り率の低かった枝刈前の決定木を用いた。

本システムでは、要約率の指定をしない場合、採用と判定された文節をすべて抽出し、推敲処理を行ない、要約文が生成される。要約率を指定した場合、抽出された文節数の割合が指定の要約率を越えた所で、文節の抽出を停止し、推敲処理に渡される。

まず、要約率の指定を行なわない場合について検討する。カテゴリ at、 bt、 ec、 sp、 wi、 wp について要約を行なった時の平均要約率は表 2となった。尚、要約率は原文の文字数と要約文の文字数の割合である。

次に、生成された要約結果について考察する。

| カテゴリ | at | bt | ec | sp | wi | wp |
|---|---|---|---|---|---|---|
| 要約率 | 0.60 | 0.64 | 0.57 | 0.62 | 0.52 | 0.61 |

表 2: 要約システムの平均要約率

結果 1
**(原文)**

釧路では数秒間の細かな揺れの後、激しい横揺れが一分以上続き、がけ崩れで倒壊する住宅もあった。電気、ガス、水道などのライフライン（生命線）にも被害が出た。北海道庁などの集計では、二人が死亡したほか約六百人が負傷した。

**(要約文)**

がけ崩れで倒壊する住宅もあった。二人が死亡したほか約六百人が負傷した。

**要約率**：0.33

(朝日新聞)

結果1の要約文は原文のパラグラフの主旨を端的に表していることが確認できる。要約率も0.33と非常に良い値である。

結果2

**(原文)**

この自然言語を計算機によって処理させようという自然言語処理の最終的目標は、人間と同様に、意味まで正しく理解することにより意志の疎通ができる計算機システムの完成である。しかし、「意味」や「理解」といったことが、人間自身よく分からないこともあって、意味まで正しく理解する計算機システムの実現には、まだかなりの時間を要すると思われる。

**(要約文)**

この自然言語を計算機によって処理させようという自然言語処理の最終的目標は、人間と同様に、正しく理解する計算機システムの完成である。「意味」や「理解」といった、人間自身分からないこともあってかなりの時間を要すると思われる。

**要約率**：0.67

(卒業論文)

要約文では「計算機システム」にかかる修飾語句が変わってしまったが、文の完成度もよく、原文の主旨を明確に表した要約文となっている。

結果3

**(原文)**

ＮＥＣの鈴木祥弘常務は「これからは国際分業の時代」と言った。「作れば売れた」バブル経済が崩壊し、実体に見合った「すみ分け」や再編成が内外で進む中、日本の製造業界は再生を模索している。

**(要約文)**

ＮＥＣの鈴木祥弘常務は「国際分業の時代」と言った。「作れば売れた」バブル経済が、再編成が、日本の製造業界は再生を模索している。

**要約率**：0.69

(朝日新聞)

要約文では、非文が生成された。『「作れば売れた」バブル経済が崩壊し』『実体に見合った「すみ分け」や再編成が内外で進む中』『日本の製造業界は再生を模索している。』と複数の節から構成される文が要約されると、非文が生成される傾向があった。

推敲過程に、文中の後置詞句が採用され、動詞が欠落した場合の処理、そして、節単位の処理がないため、このような非文が生成されたと考えられる。

### 6.5 実験3に関する評価

本研究では、要約文が原文の主旨をどれだけ伝えているか、客観的に評価するために、評価アンケートを作成し実施した。まず、人手により原文の抄録をつくり、その抄録に関する読解問題を作成した。

本研究では前章で提案したパラグラフの重要度順位付けの3つのモデルを用い、一つの文章から3種類のアンケートを作成した。

これらのモデルを用いて作られた要約文章を要約文章A群、要約文章A群で使われたパラグラフを要約率の指定なしで要約したものを要約文章B群とした。被験者には、要約文章A群を読んだ後、要約文章B群を読んだ後、原文を読んだ後に、読解問題を回答させた。そして、要約文に関する印象を問う調査事項にも回答させた。

本研究では3つの文章を用い、評価を行なった。読解問題の正解率を表3に示す。

要約文章A群の文章1の正解率はすべてのモデル

| | 要約文章A群 | | | 要約文章B群 | | |
|---|---|---|---|---|---|---|
| | 文章1 (at) | 文章2 (ec) | 文章3 (sp) | 文章1 (at) | 文章2 (ec) | 文章3 (sp) |
| モデル1 | 26% | 61% | 53% | 62% | 66% | 80% |
| モデル2 | 15% | 59% | 78% | 15% | 76% | 92% |
| モデル3 | 22% | 53% | 61% | 22% | 53% | 94% |

表3: 原文章の特徴と平均要約率徴

において低い正解率であった。しかし、要約文章B群の文章1では、モデル1の正解率のみ高い。よって、要約文章A群の文章1は適切な文章が生成されなかったが、モデル1の重要度配分は効果があったと考えられる。文章2ではモデル1と2が、文章3はモデル2が効果的であった。

つぎに、調査事項による評価を5段階評価した。要約文章A群に関する調査事項による評価を表4に、

要約文章B群に関する調査事項による評価を表5に示す。

要約文章A群は、文としての適切さの評価が低かった。要約率を設定した時に、不適切な文が生成されているためと考えられる。

要約文章B群は、要約文章A群と比べると、要約文章としてより良いものとなっていることがわかる。

| | |
|---|---|
| 大意の把握 | 2.8 |
| 読みやすさ | 2.3 |
| 文としての適切さ | 2 |
| 原文との一致度 | 2.7 |

表 4: 要約文章A群の評価

| | |
|---|---|
| 大意の把握 | 4.3 |
| 読みやすさ | 3.4 |
| 文としての適切さ | 3.4 |
| 要約文章A群との比較 | 4.4 |
| 原文との一致度 | 3.4 |

表 5: 要約文章B群の評価

## 7 おわりに

一般の文章に対応できる自動要約のため、意味的なまとまりであるパラグラフに注目し、パラグラフごとに要約率を決め、パラグラフ単位で要約を行なう手法を提案、実装し、検討した。要約処理は機械学習システムC4.5[4]を用い、重要文節を抽出することで実現した。

決定木に用いるパラメタは、必須格の有無、名詞のタイプといった文節の特徴に関するものを選んだ場合と、さらに文間、節間の修辞関係、句間の修飾関係など、パラグラフの統語構造上の特徴に関するものも選んだ場合の2種類の訓練を行ない、誤り率によって、決定木の評価を行なった。この結果から、構造上の特徴もパラメタとして使用する方が、使用しない場合よりも良い結果となることがわかった。

そして、パラグラフの要約結果を原文と比較し、検討した。要約率を指定せずに要約を行なった場合、要約文は原文の6割程度になり、また、要約文章として最も出来の良いものとなった。要約率を5割以下に設定し要約すると、非文ができ、理解不可能な要約文章が生成されることがあった。

さらに、多パラグラフから成る文章の要約のために、パラグラフの重要度付けのモデルを3つ提案し、これらのモデルを用い要約率を配分、そして個々のパラグラフ要約を行ない、要約文を生成した。多くの研究は再現率、適合率によって、評価を行なっていたが、要約文章にただ一つの正解があると仮定したものである。本手法のように要約文章の場合、再現率、適合率によって精度を測ることは難しいため、読解問題と被験者の感想を問う評価アンケートを実施し、評価結果を用いて検討した。その結果、要約される文章の種類によって、有効なモデルが異なることがわかった。また、この方法によって作られた要約文が原文の大意を表すことが可能であることが、確認できた。

今後の課題として、指示詞の照応、要約率を設定した場合の非文生成の回避方法があげられる。

## 参考文献


[1] 林大. 分類語彙表 国立国語研究所. 秀英出版, 1966.

[2] Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. *In SIGIR '93*, pp. 59–68, 1993.

[3] 比留間正樹. 重回帰分析による新聞社説の段落分割に関する研究. 横浜国立大学, 卒業論文, 1997.

[4] J.R. キンラン. AIによるデータ解析. 株式会社トッパン, 1995.

[5] W. C. Mann and S. A. Tompson. Rhetorical structure theroty: A theory of text organization. *ISI-Report*, 1987.

[6] 難波英嗣, 奥村学. 書き換えによる抄録の読みやすさの向上. 情報処理学会研究報告 99-NL-133, pp. 53–60, 1999.

[7] 上田良寛, 岡満美子, 小山剛弘, 宮内忠信. 句表現要約手法に基づく要約システム. 言語処理学会 第5回 年次大会 発表論文集, pp. 361–364, 1999.

[8] 益岡隆志, 田窪行則. 基礎日本語文法. くろしお出版, 1989.

[9] 山本和英, 増山繁, 内藤昭三. 文章内構造を複合的に利用した論説文要約システム green. 情報処理学会研究報告, Vol. 99, No. 3, 1994.