自然言語処理　132－7
(1999. 7. 22)

話題の階層構成に基づく文書自動要約: 本一冊を一頁に要約する試み

仲尾 由雄
富士通研究所
〒211-8588 川崎市中原区上小田中 4-1-1
nakao@flab.fujitsu.co.jp

あらまし 本1冊のような長い文書を1頁程度に要約する手法を提案する。1頁という要約の長さは、現在の計算機ディスプレイ上で1画面に提示できる量の上限に相当する。1頁という量は、本稿が対象とする文書の大きさから見ればごく短い量であるので、要約に取り込む話題を厳選し、かつ、それぞれの話題内容をなるべく簡潔に表現する必要がある。そこで、話題の階層構成を参照して適切な粒度の話題を選び、それぞれの話題の導入部から集中的に文を抜粋するという要約手法を考案した。逆に、1頁という量は、段落区切りを設けずにベタ詰めで提示するには長すぎる量でもある。提案手法では、要約の出力量に応じて適度な大きさの話題のまとまりごとに、要約を区切って出力することで、要約の読みやすさを改善している。

キーワード 文書要約, 話題構成, 語彙的結束性, 文章構造解析, 話題抽出

An Algorithm for Text Summarization base on Thematic Hierarchy Detection: How to Generate a One-page Summary of a Long Text

Yoshio Nakao
Fujitsu Laboratories Ltd.
4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, 211-8588 Japan

Abstract This paper presents an algorithm for text summarization using thematic hierarchy of a text. Its main purpose is to generate a one-page summary of a long text of several tens pages, which helps the user to skim an electronic book on a computer display. It is required for the one-page summarization that topics of appropriate grading be extracted, and that the selected topics be expressed as short as possible. The algorithm selects thematic textual units of appropriate size from the thematic hierarchy of a text. Then it extracts sentences from every leading part of these units, and outputs a summary with thematic boundaries that separate groups of sentences by their related topics so that the user can quickly understand outline of the summary.

Keywords Text Summarization, Thematic Hierarchy, Lexical Cohesion, Text Structuring, Text Decomposition

1 はじめに

本稿は、長い文書の要約に適した要約手法を提案する。提案手法の主たる目標は、数十頁超の文書に対して、1頁程度の要約を作成することにある。

白書のような数十頁におよぶ報告書の場合、骨子をひとまず把握しておこうとしている利用者にとっては、1/4程度にまとめた通常の要約ではなく、1頁で主要な話題の骨子のみを取り上げた要約の方が利用価値が高い。あるいは、電子図書館が所蔵する本の中から読みたい本を選ぶことを考えても、1頁程度の要約は、計算機ディスプレイの1画面に表示できるので利用価値が高い。例えば、開架式の図書館の書棚の前で本の頁を繰って得られる程度の情報を、1頁の要約にまとめることができれば、読みたい本の選別や発見の支援に役立つと考えられる。

このような考えに基づき、本稿では、数十頁超の長い文書を1頁に要約するという課題を設定した。以下、このような要約に関し、第2章で問題点を整理し、第3章で話題の階層構成に基づく要約手法を提案し、第4章で81頁の調査報告書を要約した結果に基づき議論を行う。

2 長い文書の要約に関する問題点

数十頁を超える文書を1頁程度に要約する場合の問題点について検討する。

第一の問題は、このように原文に比べて極端に短い要約は、要約に取り込む話題を厳選しないと作成できないことである。例えば、新聞記事からの重要文抜粋実験[1]によれば、それぞれの話題に対して最低3文程度(120～150字程度)抜粋しないと内容の把握が難しい[1]。よって、1,500字程度(A4判1頁程度)の要約を作成するのであれば、要約対象の文書から10個程度以下の主要な話題を厳選して抽出しなければならない。

従来の自動要約研究の多くは、新聞の社説や論文など、全体を貫く論旨の流れのはっきりした文章を対象にしてきた(例えば[2])。あるいは、複数記事をまとめて要約する研究(例えば[3])であっても、何らかの一貫した流れ(事件経過など)に沿う文章を対象にしてきた点に変わりはない。言い換えれば、ひとつの談話の流れに沿った文章を対象に、要約研究が進められてきたといえる。

しかし、白書などの長い文書では、文書全体を貫く論旨の流れが存在するとは限らず、ある論旨に沿って記述された複数の文章が、緩やかな関連性の下に並べ置かれていることが多い。このような集合的文書を1頁程度に要約するためには、大局的な話題構成を認定して、要約に取り入れるべき話題を選択/抽出する必要がある。すなわち、それぞれの談話の単位(修辞的な文章構造)を要約する前に、個々の談話の単位を包含する大きな話題のまとまりを認定し、要約に取り入れるべき適切な話題のまとまりを選択しなければならない。

談話の単位を包含する大きな話題のまとまりは、文書の論理構造(章や節など)と深く関連するので、その認定を書式解析(例えば[4])により行うことも考えられる。しかしながら、書式解析処理は、処理対象を限定すれば容易に実現できるものの、汎用性に問題がある。つまり、書式はある種類の文書における約束事であるため、文書の種類毎に経験的な規則を用意しなければならないという問題点がある。また、同じ章の下に並んでいる節であっても、節間の関連の程度が大きく異なる場合もあり、文書の論理構造と話題の階層構成とは必ずしも一致しない。

そこで、本稿では、書式解析ではなく、語彙的結束性という一般性の高い言語現象に基づく話題の階層構成の認定手法[5, 6]を利用して、要約作成を試みる。

第2の問題は、話題を表す要約内容(抜粋文)の理解し易さと結束性をいかに確保するかにある。例えば、話題毎に分けた部分でもまだ大きすぎる場合に、重要語が多く出現する文を抜粋すると、たまたま論の半ば付近の文が抜粋されて、要約が理解不能になってしまうことがある。詳細な議論を行っている箇所を前提となる説明なしに抜粋してしまうと、読者には何を議論しているのかが掴めない可能性が高いからである。また、1%程度以下の極端に短い要約を作成する場合、重要な文の中から少数の文を選択しなければならないため、要約が関連性のない文の羅列になってしまう可能性も大きくなる。

例えば、図1は、[5]で提案した方法に従って作成した1500字程度の要約の1部、4.3節に対応する部分[2]を示したものである。この要約は、話題の階層構成に基づき、各話題に特徴的な語を統計的にもとめ、それを多く含む文を抜粋して作成した。要約中の文は、重要なキーワード(概念)を含んでいるようにも見えるが、4.3節でどういう位置付けで取り上げられているのかが分からない。そのため、これだけからでは、この文書がどういった種類の文書であるのかの推測はついても、文書の内容を自信をもって推測することができない。

そこで、本稿では、文書全体の流れが理解できるよう、話題の導入部に絞って要約を試みることとした。

第3の問題として、長い文書を要約する場合、必然的に要約結果の分量も大きくなり、結果として、読みにくい要約になってしまうという問題もある。例えば、百頁の本を要約した場合、1%に縮めても、要約は1頁になってしまう。1頁の文書は、少なくとも数段落にわけ、見出しなどを付与し、内容の区切りの目印をつけない限り、読みづらい。長い文章の提示法に関しては、自動生成し

[1] 見出し1文に本文から抜粋した2～3文を提示すれば、雑談の話題として提供できる程度には理解できた気になれる。

[2] 全体の1/6程度。要約対象は、本稿でも要約実験に用いている81頁の調査報告書。第3章参照。

4.3 ネットワーク上の検索サービス

…また検索精度を高めるために、高頻度語は検索の対象としない、タイトルや見出しに含まれる語に重みをつける、などの工夫がなされている。

…また、検索サービスが収集したページ数が膨大になるにつれて、ヒット数も膨大になってきたため、すばやく必要な情報を探すために、よりわかりやすい自動抄録作成技術が必要となる。…

…tf・idf 方式とは、単語に分割された文章の各単語の重要度を、その単語が文書中に出現する頻度 tf と、その単語を含む文書が文書集合中に出現する頻度の逆数 idf の積によってその単語の重要さを数値化する手法である。…

図 1: 旧手法による要約結果 (一部)

た見出しを付与して話題階層を表示する手法も提案されている [7]。要約は、その本来の目的からいって、通常の文書より短い時間で内容を把握したいという要請が強いと考えられる。1 頁程度以上の長さに要約するのであれば、要約の内容が一目で把握できるような提示形式が強く望まれるといえる。

3 話題の階層構成に基づく要約手法

本稿で提案する要約手法は、要約の単位として適した大きさの話題のまとまりを抽出し、それぞれの話題の導入部から集中的に文を抜粋して要約を作成する手法である。具体的手順は以下の通りである。

1. 話題の階層構成の認定
   同一語彙の繰り返しによる語彙的結束性の分析に基き、話題の境界位置を仮設定する (境界位置は文境界とは無関係に設定される)。
2. 話題境界の確定と話題導入文の認定
   仮設定した話題の境界位置を微調整して文境界にあわせ、境界位置から始まる文を境界文と認定する。そして、境界文の後ろにあり、後続の話題との関連度の高い文を話題導入文と認定する。
3. 適切な粒度の話題の選択と境界文・話題導入文の出力
   要約の出力量と原文書の大きさに応じて、要約に取り入れる話題を選択し、選択した話題に対応する境界文と導入文を抜粋対象文として抽出する。そして、抽出した抜粋対象文を話題毎にまとめ、読みやすく成形して出力する。

以下、要約例を交えながらこれらの処理について説明する。要約対象文書としては、(社) 電子工業振興協会『自然言語処理システムの動向に関する調査報告書』(平成 9 年 3 月) 第 4 章「ネットワークアクセス技術専門委員会活動報告」(pp. 117–197) を用いた。この文書は、4.1 節から 4.4 節の 4 節からなり、1,440 文に延べ 19,311 語の内用語[3] を含んでいる。

3.1 話題の階層構成の認定

提案手法では、まず、[5, 6] で提案した話題構成認定手法に基づき、文書中の話題の階層構成を認定する。

この手法では、まず、[9] にならい、文書中の各位置の前後に、求めたい話題の大きさ程度 (文書全体の 1/4〜段落程度) の窓を設定し、その 2 つの窓に出現する語彙の類似性を測定する。類似性は、次に示す余弦測度 (cosine measure) で測定している。

$$sim(b_l, b_r) = \frac{\Sigma_t w_{t,b_l} w_{t,b_r}}{\sqrt{\Sigma_t w_{t,b_l}^2 \Sigma_t w_{t,b_r}^2}}$$

ここで、b_l, b_r は、それぞれ、左窓 (文書の冒頭方向側の窓)、右窓 (文書の末尾方向側の窓) に含まれる文書の部分であり、w_{t,b_l}, w_{t,b_r} は、それぞれ、単語 t の左窓、右窓中での出現頻度である。本稿では、この値を結束度と呼び、また、結束度に対応する窓の境界位置によって結束度を並べたものを結束度系列と称することにする。

次に、上記の結束度を、ある刻み幅 (窓幅の 1/8) で窓をずらしながら測定して、文書の冒頭から末尾に至る結束度系列を求める。そして、結束度系列の極小点を手がかりに話題境界を認定する。この際、結束度系列の移動平均をとることで、任意の大きさの話題のまとまりを選択的に検出できるようにし、また、極小となる移動平均の値に対して大きく寄与している文書中の範囲 (窓幅の 1/2〜1 程度) を求めて境界位置の候補区間を作成している (詳細は [6])。

以上の操作を、窓幅を変えて行うと、大きな窓幅では大きな話題の切れ目に、小さな窓幅では小さな話題の切れ目に対応する、境界候補区間が認定できる[4] 。

図 2 は、要約対象文書中の話題構成の認定結果である。図中、2 重矩形の外側の矩形は境界候補区間[5] であり、内側の矩形は仮境界位置 (結束力拮抗点:最も境界位置らしい点) である。縦軸は、結束度系列の計算に用いた窓幅である。点線は、要約対象文書中の節の開始位置であり、長い程大きい節と対応する。なお、節の開始位置は、比

[3] 日本語形態素解析ツール jmor[8] を使って切り出した名詞・動詞・形容詞。[5] の時に比べ、数が増えているのは、URL 相当の英字列の結合機能などを省いて要約エンジンに取り込んだ事情などによる。

[4] 調査報告書や新聞の特集記事などをテストデータとして用いた実験 [6] によれば、窓幅程度の大きさの話題の 7 割程度 (再現率) は、境界候補区間内に含まれることが期待される (適合率は 5 割程度)。

[5] 本稿では境界位置の調整処理を設けたため、この段階では、境界候補区間を [5, 6] より広く (約 2 倍) 設定している。

較の目安として示しただけであり、提案手法では節の開始位置に関する情報は全く用いていない。

図2で上方に位置する境界候補区間ほど、大きな話題に関する境界に対応する。以降、それぞれの窓幅による境界候補区間を以下の境界データ $B(i)(i = 1, 2, \ldots, i_{max})$ によって区別して参照する。

i: 話題の階層レベル。小さいほど大きい話題に対応。

$B(i)[j]$: 一つの境界データ。以下の値をもつ。

$B(i)[j].bp$: 仮境界位置。

$B(i)[j].range$: 境界候補区間。

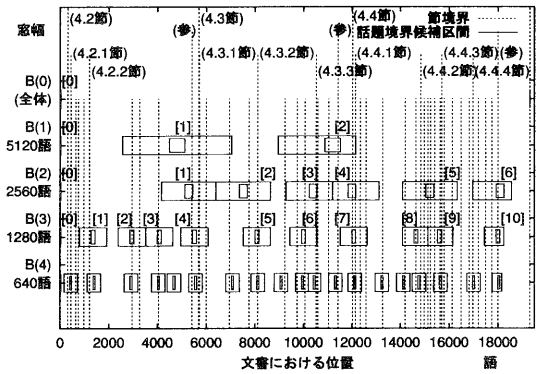

図 2: 話題構成の認定結果

次に、境界データ $b_i \in B(i)$ の仮境界位置と境界候補区間を、その直下の境界データ $b_{i+1} \in B(i+1)$ で $b_{i+1}.bp \in b_i.range$ の条件を満たす境界データを1つ選び、以下のように統合する。

$$b_i.bp = b_{i+1}.bp$$
$$b_i.range = b_{i+1}.range。$$

この操作を $B(i_{max}-1)$ から順に $B(1)$ まで行い、大きな窓幅で認定した大きな話題に関する境界と、小さな窓幅で認定した小さな話題に関する境界が統合することで、話題の階層構成が形づくられる。例えば、図2の $B(1)[1]$ の境界データは、$B(2)[1], B(3)[1], \ldots$ などと統合され、その境界候補区間と仮境界位置が最小窓幅 (40 語) による (候補区間の幅も最も狭い) 境界データ $B[i_{max}]$ の値に置き換えられる。

3.2 境界文・話題導入文の認定

次に、文境界とは無関係に仮設定した話題の境界位置を微調整し、境界位置から始まる文を境界文と認定する。そして、境界文の後ろにあり、後続の話題との関連度の高い文を話題導入文と認定する (図3)。

1. 最下層の境界データを処理対象に位置付ける: $i \leftarrow i_{max}$
2. 処理対象階層の個々の境界データ $B(i)[j]$ のそれぞれに対して、文書の先頭に近い位置の境界データから順に以下の処理を行う。
   (a) 境界文探索範囲の決定
   $i = i_{max}$ の場合には、境界候補区間を境界文探索範囲とし、それ以外は、統合された一回り小さい話題に関する境界データ話題境界候補区間内の $B[i+1]$ の境界文以前の部分を境界文探索範囲とする。
   (b) 境界文の選択
   境界文探索範囲内にある文の中から、順方向相対関連度が正で、かつ、増分が最大となる文を境界文として選択する。
   (c) 話題導入文の選択
   境界文以降の境界候補区間の範囲で順方向関連度が最大となる文を話題導入文として選択する。
3. $i > 1$ であるなら、$i \leftarrow i-1$ とし2から境界文・話題導入文選択処理を繰り返す。

図 3: 境界文・話題導入文選択アルゴリズム

順・逆方向の関連度の計算

境界文・話題文の認定には、順方向関連度、逆方向関連度という2種類の関連度を用いる。順方向関連度とは、境界候補区間内のそれぞれの文と、その直後の話題のまとまりとの関連性の強さを示す値である。同様に、逆方向関連度とは、それぞれの文とその直前の話題のまとまりとの関連性の強さを示す値である。また、順方向関連度と逆方向関連度との差を、順方向相対関連度と呼ぶこととする。

ここで、文 S と話題のまとまり b との関連度 $r_{S,b}$ は、以下の式によって求める。

$$r_{S,b} = \frac{1}{|S|} \sum_{w \in S} \frac{tf_{w,b}}{|b|} \times \log\left(\frac{|D|}{df_w}\right))$$

$\|S\|$	文 S に含まれる延べ単語数
$\|b\|$	話題のまとまり b に含まれる延べ単語数
$tf_{w,b}$	単語 w の話題のまとまり b における出現頻度
$\|D\|$	文書を固定幅 (80 語) 刻みに区切ったブロック数
df_w	単語 w が出現しているブロック数

この式は、[10] で単語の重要度の評価用に取り上げた尺度の一つ (「情報量型複数ブロック $tf \times idf$ 法」) を応用したものである。この尺度には以下のような望ましい

性質がある[6]

(1) 話題のまとまりに特徴的に出現する語が重視される

語の文書全体における出現密度が低いほど、logの部分の値が大きくなるので、文書全体では出現密度が低いが語が文Sと話題のまとまりbの両方に出現した場合、関連度が大きくなる。逆に、文書中でほぼ均一に分布する性質がある機能語などは、ほとんど関連度に寄与しなくなる。

(2) 主要な話題としてとりあげられている語が重視される

log内の部分は局所的に集中して出現する単語の出現度数を低めに補正した出現密度の逆数の形をとっているため、文S中の語が話題のまとまりbの中で1ヶ所に集中して出現している場合に、関連度が大きくなる。

なお、それぞれの話題のまとまりの境界位置は、境界文の開始位置であるが、境界文が決定していない境界位置に対しては、話題境界の仮位置を用いて上記の関連度を計算する。

境界文・話題導入文の選択例

表1は、要約対象文書の4.4節の開始位置付近(図2の横軸の12,000語付近)の境界文・話題導入文の認定例である。表1で、<外>と印を付けた文の次の文から表の終わりまでが、境界文・話題導入文の候補である。これらの文は、境界候補区間に少なくとも文の一部がかかっている文である。表1の場合、境界候補区間は$[12026, 12060)$の45語幅であり、この区間内に一部でもかかっている文話題文の候補である。

表1で、<境>と印を付けた文は、順方向相対関連度(図の「後-前」の列の値)が正(0.016)であり、かつ、直前(-0.008)からの増分が最大であるので、境界文と認定した文である。<導>と印を付けた文は、境界文以降にある文(この場合は境界文以外で2文)の中で、順方向関連度(「対直後」)が最大(0.023)となっているので、話題導入文と認定した文である。

なお、上位階層の境界データ$\{B(i) : i < i_{max}\}$の境界文の選択に関しては、境界候補区間を絞り込み、それと統合された直下の階層の境界データ$B(i+1)$の境界文までを探索範囲とする。これは、順方向相対関連度がしばしば話題の開始位置より後方で大きくなり、順方向相対関連度の増分を手がかりにしても、話題の開始位置より後方の文が選択される傾向がみられたためである[7]。

3.3 境界文・話題導入文の出力

1. 要約に取り入れる話題の概数Ntを決定する:

$$Nt \leftarrow S_a/S_t$$

ここで、S_aとS_tは、それぞれ、要約の出力量、話題あたりの抜粋量である。
2. Nt個以下の話題のまとまりからなる最下層の話題階層$B(i)$を選択する。
3. $B(i)$の境界データから境界文・話題文を抜粋量の制約(要約の出力量、話題あたりの抜粋量)の範囲で抽出し、話題のまとまりごとに出力する。
4. 抜粋量に余裕があり、$i < i_{max}$(最下層でない)の場合には、$i = i+1$として3の処理を繰り返す。

図4: 境界文・話題導入文の出力アルゴリズム

最後に、適切な粒度の話題を選び、境界文と話題導入文を抜粋して要約を出力する(図4)。ここでのポイントは、要約の出力量と話題あたりの抜粋量とに応じて要約に取り入れる話題の粒度を調整している点と、境界文・話題導入文を話題ごとにまとめて出力している点である。

例えば、1,000字程度の要約を作成する場合、話題当たり150字程度抜粋するのであれば、6〜7個程度の話題を持つレベルを探す。要約対象文書では、表2のような話題構成となっているので、$B(2)$の階層の境界データから境界文と話題導入文を抽出し、境界別に出力する。なお、図5は、$B(2)$の境界だけでなく、その上位階層$B(1)$の境界位置も段付けなどで明示して出力している。

表2: 実験対象文書中の話題の階層構成

話題階層	窓幅	区間数	平均サイズ(語)
B(0)	(文書全体)	1	19,311
B(1)	5,120	3	6,437
B(2)	2,560	7	2,759
B(3)	1,280	11	1,756
B(4)	640	20	966
B(5)	320	42	460
B(6)	160	83	232
B(7)	80	167	115
B(8)	40	330	57

[6] [10]では、上式の$\sum$内の部分の式を用いて、文書内の単語の重要度を評価する実験を行い、評価値が高い順に単語を抽出することで見出しに出現する語(重要語)が効率よく抽出できることを確認した。

[7] 章見出しの直後に節見出しがある場合に、小さい話題に関する境界文として節見出しを、大きな話題に関する境界文として章見出しを認定できる可能性を高めるという意味合いもある。

表 1: 境界文と話題導入文のの選択例

文の出現位置	関連度 対直前	対直後	後-前	文表記
<外> 12002	0.029	0	-0.029	吉岡誠: "ＳＧＭＬを使いこなす", (株) オーム社, 1996
12008	0.016	0.008	-0.008	吉村賢治 (福岡大学), 日高達, 吉田将 (九州大学): "日本語科学技術文における専門用語の自動抽出システム", 情報処理学会論文誌, Vol.27, No.1, pp.33-40, 1986
<境> 12031	0	0.016	0.016	4.4. 検索エンジン
<導> 12033	0.008	0.023	0.015	ここではネットワークを利用した知的情報アクセスにおける自然言語処理の役割を明らかにするために、情報検索において特に自然言語処理との関連が深い幾つかのテーマについて最新の学術的研究動向を調査した結果について報告する。
12055	0.008	0.015	0.007	以下の各節の報告に共通するテーマは、「ギガバイトあるいはテラバイトに及ぶ膨大なデータから必要な情報を得るにはどうしたらよいか？」という問題である。

4 1頁要約結果の検討

4.1 旧手法の結果との比較

提案手法の要約 (図 5) を旧手法による要約 (図 1) とを比較すると、提案手法の要約の方が抜粋内容に繋がりがあるので文章としては読みやすい印象をうける[8]。どのような場合に読みやすいと感じるのかは今後の検討課題であるが、この印象を支える要因として以下の事柄が考えられる。

第 1 に、旧手法では異なる話題の文がバラバラに抜粋されているのに対し、提案手法の要約には大きな節の冒頭で節の内容を説明している文などが含まれていることである。第 2 に、多くの話題に対して、それぞれ 2 文ずつ以上抜粋があることである。第 2 の点は、見出しの付与とも関連すると考えられる。例えば、提案手法の要約においても、「キーワード抽出」の部分などはやや唐突な印象をうけるが、図 1 に比べると違和感は小さい。この例では、見出しは、実質的に何の情報ももたらしていないが、関連する抜粋内容に添えてあることで、「読みにくい/分からない」という心理的な違和感を緩和していると考えられる。

ただし、要約としては読みやすさだけが重要なわけではないので、提案手法と旧手法の本当の優劣は、利用者がどういう情報を求めているかに依存する。例えば、旧手法の要約に抜粋された文には、「検索精度/自動抄録作成/tf·idf」などキーとなる概念が多く見られる。よって、その分野の専門家が数種類の調査報告書の比較を試みている場合などには、旧手法の要約の方が望ましいことも考えられる。

[8] ごく主観的な評価ではあるが、同僚数人に要約結果の印象を尋ねた範囲では、大体印象は一致していた。

4.2 見出しの抽出状況の評価

提案手法は書式の手がかりは全く使用していないにも関わらず、図 5 は、提案手法の見出しの検出力の高さを示唆している。どの位の検出力があるかについては、稿を改めて報告したいと考えているが、ここでは、要約対象文書中の見出しと境界文との関係に関するデータを紹介する。

表 3 は、1,000 語以上の大きな節の境界がどの話題階層の境界として認定されたかを、節の大きさ順に示している。表中◎がついているのは、節見出しがその階層の境界文として認定されたことを示している。例えば、「4.4 節」の見出しが $B(1)[2]$ の境界文として認定されたことを示している。同様に、○は見出しが話題導入文として認定されたことを、△は見出しが境界文探索範囲に含まれていたが境界文としては選択されなかったことを、×は見出しが上位階層の境界候補区間に含まれていたが統合操作により境界候補区間からはずれたことを示している。無印は、見出しが完全に境界候補区間からはずれていたことを示している。

最大の無印節である 4.2 節が上位階層の境界候補区間からはずれたのは、4.1 節のサイズが 316 語と極端に小さいことに由来する。つまり、話題構成認定手法には、文書の冒頭にごく近い部分から始まる大きな話題の境界位置は検出できないという欠点がある。大きい話題のまとまりの認定では、大きい窓幅を用いているため、文書端で結束度計算用の窓の一方が大きく文書外にはみ出してしまうからである。今後の検討課題の一つである。

話題構成認定手法は、窓幅程度の大きさを認定する手法である。その点では、$B(2)$ は、窓幅 (2,560 語) より若干小さすぎる節境界を境界文に認定している傾向にある。例えば、$B(2)[2]$ では、大きさ 2,411 語の 4.3.2 節の見出し「検索技術の動向」ではなく、その直後の副節 (大きさ 1,120 語) の見出し「(1) キーワード抽出」が境界文とし

て認定されていた。これは、「4.3.2 検索技術の動向」の後ろには、節の概要説明などが全くなく、直後から「(1) キーワード抽出」の副節が始まっていることに由来していると見られる。この問題は、話題の導入部から文を抜粋するという手順にも関連する。つまり、4.3.2のように見出しの近傍に概要説明のない見出しを要約に取り入れた場合、現在の手順では適切な話題導入文が抽出できない可能性が高い。これも今後の検討課題である。

表 3: 大きさ別の節境界の認定状況

節の大きさ	節番号	認定状況		
		B(1)	B(2)	B(3)
1,9311	4. 節全体	◎ [0]	([0])	([0])
7,280	4.4 節	◎ [2]	([4])	([7])
6,347	4.3 節	◎ [1]	([1])	([4])
5,368	4.2 節			
4,193	4.2.2 節			(× [1])
2,685	4.4.1 節		× [4]	([7])
2,411	4.3.2 節		△ [2]	([5])
2,398	4.3.1 節		△ [1]	([4])
1,754	4.2.2 節 (1)			× [1]
1,357	4.2.2 節 (4)			◎ [3]
1,312	4.4. 参考文献		△ [6]	([10])
1,272	4.4.3 節		◎ [5]	([9])
1,120	4.3.2 節 (1)		◎ [2]	([5])
1,029	4.4.4 節			
902	4.3.3 節		× [3]	([6])
866	4.4.2 節			○ [8]
464	4.3.2 節 (4)		◎ [3]	([6])
314	4.2.2 節 (2)			△ [2]
82	4.2.2 節 (2)(a)			◎ [2]

4.3 読みやすさ向上の試み

図5には、読みやすさの向上を狙ったいくつかの工夫を施してある。これを題材に、要約の読みやすさについて考察する。

第1の工夫は、原文書中の話題構成従って、要約にも階層構成を導入した点である。これは、$B(2)$ の階層の7つの境界データごとに要約を区切って出力してみたところ、項目の羅列という印象の要約になってしまったことによる。$B(1)$ と比べると、$B(2)$ では、$B(2)$[2](キーワード抽出)、$B(2)$[3](分散検索) を話題として独立し、また、4.4節の参考文献に対応する $B(2)$[6]([青江…]=参考文献) が新たな話題として現れている。これらの話題を、$B(1)$ の3つの大きな話題と同列に扱うと、特に4.3節の中の2つの小話題の粒度が違いすぎるためか、どうしても唐突な印象を与えてしまうようであった。

そこで、$B(1)$ に対応して大きな区切りをつけ、$B(2)$ にしか現れない境界は、図5では、段下げして表示した。また、境界をより目立たせるために、各境界の先頭の文およびその後ろの句点で終わっていない文を見出し扱いで出力した。今回の事例では、$B(1)$ の境界が原文書の大きな話題と対応しているため、この処置は妥当であったと考えている。ただし、上層の話題のまとまりに信頼性がおけない別の事例[9] も見つかっており、上層の話題のまとまりの妥当性の判定が今後の課題となっている。

第2の工夫は、見出しの提示に関し、見出し番号を目立たなく加工している点である。これは、要約結果に4.2節の見出しが欠落していることを目立たなくするために行った。見出し番号をそのまま示すと、要約の読者が4.2節の欠落に気づき、違和感を覚える可能性が高い[10] 。この違和感を緩和し、読みやすさを向上するのが、見出し番号の加工の意図である。ただし、見出し番号の加工には、誤解を生ずる危険性もある。図5で「調査の概要」とまとめてある部分において、3行目末尾の「現在しばしば用…」以降は「4.2 ネットワークアクセスのインタフェース」からの抜粋である。しかし、そのことを知らずに要約を読んだ場合、「調査概要」の一部にしか見えず、アクセスインタフェースに関する詳細な調査が報告されていることを読者が見逃す危険性がある。

5 まとめ

本稿では、数十頁を超えるような長い文書を1頁に要約するという要約処理の課題を提示し、それを実現するための要約手法を提案した。そして、話題の階層構成を参照して適切な粒度の話題を選び、それぞれの話題の導入部から集中的に文を抜粋することで、話題を端的に示す見出し文などを多く含み、かつ、意味的なまとまりの強い要約が作成できることを示した。また、1頁の要約の読みやすさに関しても考察し、要約をいくつかの段落に分けて提示する手法などの検討も行った。

提案手法は、要約の理解しやすさ/読みやすさの向上を狙って発想したものではあるが、第4章で指摘したように、読みやすさに重点を置きすぎると読者に誤解を与える危険性もある。今後は、要約に取り入れる内容の妥当性と要約の読みやすさの両面の向上を目指し、要約手法の改良を行いたいと考えている。

謝辞

実験用文書を提供して下さった電子工業振興協会のネットワークアクセス技術専門委員会の方々に感謝いたします。

[9] 国会の会議録を対象とした場合など。
[10] 同僚に要約結果を見せて指摘された事柄の1つ。

ネットワークアクセス技術委員会 [4. 参照]

調査の概要 [4.1 参照]

…それにともなってインターネットを通じて提供される情報も多種多様化している。…現在しばしば用いられている WWW の情報検索は、ユーザがキーワードを入力してそれを含むページを提示したり、あらかじめ固定された概念階層をユーザがたどって好みの情報にアクセスする方法が多い。…(a)Java Java は、Sun Microsystems 社によって開発された、ネットワークでの利用を主眼においたオブジェクト指向言語である。…(4) 機械翻訳・言語処理技術 (a)WWW における機械翻訳 一昔前なら比較的高性能のワークステーションなどでしか使えなかった翻訳ソフトウェアが、パソコン上で高速に動作するようになった。…

ネットワーク上の検索サービス [4.3 参照]

本節では、WWW 上の検索サービスと電子出版及び電子図書館について、現在行われている各サービスの特徴、技術的なポイント、問題点等を調査すると同時に、関連する研究分野も調査し、将来どのようなサービスが望まれるか、また、そこに必要となる技術は何であるか、についてまとめる。…

キーワード抽出 [(1) 参照]

ネットワーク上の文書をアクセスする方法の 1 つとしてキーワード検索がある。…

分散検索 [(4) 参照]

情報を一ヶ所に集中登録するタイプの検索サービスでは、今後ますます肥大化・多様化していく WWW には対応しきれなくなることが予想される。この問題を解決するためには、各検索サービスが互いに独立して動作するのではなく、相互に連携しあう必要がある。…

検索エンジン [4.4. 参照]

ここではネットワークを利用した知的情報アクセスにおける自然言語処理の役割を明らかにするために、情報検索において特に自然言語処理との関連が深い幾つかのテーマについて最新の学術的研究動向を調査した結果について報告する。…

情報フィルタリング技術の動向 [4.4.3 参照]

情報フィルタリング (information filtering) とは, 動的に変化する情報の集合の中から, ユーザのニーズに合致する情報を取り出す技術である。…

[青江, 92a]

青江順一 (徳島大学): 静的ハッシュ法とその応用, キー検索技法 – I, 情報処理, Vol.33, No.11, pp.1359-1366, 1992.…

(945 字。原文との字数比で 1.0%)

図 5: 要約結果 (1000 字程度)

参考文献

[1] 仲尾由雄：見出しを利用した新聞・レポートからのダイジェスト情報の抽出, 情処研報 NL-117-17, 情報処理学会 (1997).

[2] 住田一男, 知野哲朗, 小野顕司, 三池誠司：文書構造解析に基づく自動抄録生成と検索提示機能としての評価, 電気情報通信学会論文誌, Vol. J78-D-II, No. 3, pp. 511–519 (1995).

[3] 船坂貴浩, 山本和英, 増山繁：冗長度削減による関連記事の要約, 情処研報 NL-114-7, 情報処理学会 (1996).

[4] 土井美和子, 福井美佳, 山口浩司, 竹林洋一, 岩井勇：文書構造抽出技法の開発, 電気情報通信学会論文誌, Vol. J76-D-II, No. 9, pp. 2042–2052 (1993).

[5] 仲尾由雄：文書の意味的階層構造の自動認定に基づく要約作成, 第 4 回年次大会併設ワークショップ「テキスト要約の現状と将来」, pp. 72–79 言語処理学会 (1998).

[6] 仲尾由雄：語彙的結束性に基づく話題の階層構成の認定, 自然言語処理, Vol. 6, No. 4 (1999), 掲載予定.

[7] Yaari, Y.: Texplore – exploring expository texts via hierarchical representation, in *Proc. of CVIF '98*, pp. 25–31Association for Computational Linguistics (1998).

[8] 西野文人：日本語テキスト分類における特徴素抽出, 情処研報 NL-112-14, 情報処理学会 (1996).

[9] Hearst, M. A.: Multi-paragraph segmentation of expository text, in *Proceedings of the 32nd Annual Meeting Annual Meeting of Association for Computational Linguistics*, pp. 9–16 (1994).

[10] 仲尾由雄：文書の話題構成に基づく重要語の抽出, 情処研報 FI-50-1, 情報処理学会 (1998).