2004－NL－160　(7)
2004／3／4

話題の流れを保持する自動要約

市丸 夏樹*,　飛松 宏征†,　日高 達‡

本稿では，論説文を読みやすい informative な要約文へと自動的に要約するための手法を提案する．まず，文章を段階的に段落分けし，階層的な結束構造を構築する．その際，文間，段落間の繋がりを計測するために，従来の語彙結束性に代えて話題間の連想による話題の流れのよさを表す値を用いる．論説文中の段落列に着目すると，導入部を表す段落から展開部を通って結びへ至る大きな話題の流れが見られることが多い．そこで，我々の要約システムはその大きな流れこそが書き手の主張を表す文脈の本流であるとみなす．最後に，流れを構成する文とその流れへの寄与度が高い文を抽出することによって，流れのよい読みやすい要約文が得られる．

Text Summarization by Subject Flow Analysis

Natsuki Ichimaru*,　Hiromasa Tobimatsu†,　Toru Hitaka‡

In this paper, we propose a method of automatic summarization to produce easy-to-read, informative summaries of editorial articles. Firstly, the document is segmented hierarchically to form a multi-layer paragraph structure. Instead of lexical cohesion, associative relations between subjects are used to measure subject-flow consistency between sentences and between paragraphs. In a paragraph, the most significant subject-flow stream runs from an introductory subparagraph to conclusive one through some intermediates. Thus, our system regards this flow as the contextual main stream of the document. Finally, several sentences which have greater contribution ratio in the flow are extracted gradually, to form a multistage summary which has maximum consistency.

1　はじめに

情報技術の普及によって WWW を始めとする大量の情報を誰もが日常的に検索・閲覧することができるようになった．検索エンジンのページでは原文を読むかどうかの判断に用いるための indicative な要約が付いたインデックスが提示される．しかし検索された本文が長い場合，現状ではユーザは文書内で文字列検索等を用いてさらにキーワードの出現個所を一個所一個所チェックせざるを得ない．こういう時に長すぎる本文が自動的に要約されたら便利に違いない．

従来の重要文抽出法 [3] で出力される要約文は，文の間に繋がりがなく文章としては読みにくいものになりがちである．そこで本研究では文と文との間の連想による繋がりに着目することによって，原文の代わりとなる informative な要約として用いることができるような読みやすい自然な要約文を生成することを目指す．

文間の繋がりを捉えようとする従来研究 [3, 4] としては，連続して用いられる同じ単語間の語彙結束を用いる手法，あるいはシソーラスを用いて同義語・類

*九州大学システム情報科学研究院
Graduate School of Information Science and Electrical Engineering, Kyushu University
†九州大学システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
‡九州大学名誉教授
Emeritus Professor, Kyushu University

義語等にまで拡張された単語間の繋がりを用いる手法，テキストコーパスから収集した単語間の共起関係を用いる手法などが提案されている．しかし，字面が共通した単語の語彙結束による場合は，単語表記や表現の多様性によって，文間で繋がるべきところが繋がらないことが多い．一方，大量のテキストコーパスから収集した共起関係や大規模なシソーラスを用いた場合は，本来繋ぐべきでないところまで繋がってしまうため，質の良い結果が得られなかった．

KeyGraph[9] は，高頻度語だけでなく他の単語に支えられて文章全体に渡って述べられているような語もキーワードとして抽出するものであり，大変良いキーワードを抽出できることが知られている．KeyGraph はコーパスやシソーラスではなく，解析対象文中の単語の共起頻度を利用するという点で重要である．新規の文章から新しいキーワードを発見するためには，その文章内の文脈における単語の共起関係が必要なのである．なお，既に KeyGraph を応用した要約手法が提案されているが [5]，それは文間の繋がりを直接的に扱うものではないようである．

本研究では，要約対象文自体から名詞の共起関係を抽出し，文間の話題の流れを用いて導入から展開，結びに至る文脈の流れの本筋を抽出し，できるだけ周りとの繋がりがよい文を取り出すことによって，文間の繋がりのよい要約文を抽出する手法を提案する．

2 文と文の間の繋がりの強さ

2.1 文章中の話題

いま，N 個の文 s_i からなる文章 $D = \{s_1, \ldots, s_N\}$ が与えられたものとする．まず各文を形態素解析し，文中の一般名詞，固有名詞，サ変名詞 (サ変動詞語幹) を抽出しておく．形態素解析器としては茶筅 [8] を使用している．以下，これらの抽出された名詞の組 (w, w') を話題と呼ぶ．

2.2 連想による話題の流れ

文中で次々に移り変わる話題の繋がりを取り扱うため，話題間の連想関係の利用を試みる．しかし実際の人間の読者が行うような連想を計算機上に実装することは今のところ困難であるから，これを文中で話題となっている名詞同士の共起関係で代用する．

新規の文章中の文脈の流れの中に現れる様々な文と文の繋がりを捉えるためには，やはりその新しい文脈の中で語られる話題間の連携を考慮する必要がある．そこで連想の源となる単語の共起関係は，コーパスなどの他の文章からではなく，要約の対象となる文章自体から獲得したものを用いる．

話題 (w, w') が D 中の同一文中に出現する確率を $rf(w, w')$，D における $rf(w, w') > 0$ なる話題 (w, w') の全体集合を R とおく．話題 (w, w') から D 中のいずれかの文に含まれる共起関係を 1 段階辿って連想される話題 $(v_1, v_2) \in R$ を連想話題と呼び，話題 (w, w') の全ての連想話題の集合を $A(w, w')$ とおく．

$$A(w, w') \stackrel{def}{=} \{(v_1, v_2) | (v_1, v_2) \in R, \forall i.(w, v_i) \in R \vee (w', v_i) \in R\} \quad (1)$$

2.3 文間の話題の流れの良さ

文 s に含まれる話題の集合を $R(s)$，$R(s)$ から連想される話題の集合を $R^a(s)$ とおく．

$$R(s) \stackrel{def}{=} \{(w, w') | w \in s, w' \in s, w \neq w'\} \quad (2)$$

$$R^a(s) \stackrel{def}{=} \bigcup_{\rho \in R(s)} A(\rho) \quad (3)$$

いま前の文 s と後の文 s' の間の話題集合の重なりについて考えると，

1. $R(s) \cap R(s')$ は継続した話題，
2. $\overline{R(s)} \cap R^a(s) \cap R(s')$ は連想によってスムーズに繋がった話題，そして，
3. $\overline{R^a(s)} \cap R(s')$ は s' で新たに加わった話題，

をそれぞれ表すことになる．

後の文に含まれる話題 $R(s')$ の中で 1, 2 の割合が多い場合は文 $s-s'$ 間の話題が連続していると考えられ，逆に 3 が多い場合には話題が急激に転換しているのではないかと考えられる．そこでこのような話題間の連想関係を利用して，2 文間の話題の流れのよさ $F(s \to s')$ を次のように表す．

$$F(s \to s') \stackrel{def}{=} \sum_{\rho \in R^a(s) \cap R(s')} rf(\rho) \,/\, \sum_{\rho' \in R(s')} rf(\rho') \tag{4}$$

$$F(s_1, \ldots, s_n) \stackrel{def}{=} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} F(s_i \to s_j) \tag{5}$$

$F(s \to s')$ は，$R(s')$ に占める $R(s)$ の連想話題の割合であり，文 $s-s'$ 間の繋がりの強さを表す．$F(s_1, \ldots, s_n)$ は，文の並び $s_1, \ldots, s_n$ の中の全ての組み合わせの 2 文間の前の文から後の文への話題の流れのよさの合計であり，段落のまとまりの良さ (結束性) を表す目安となる．

2.4 流れの良さの特性

文間の話題の流れのよさ $F(s \to s')$ は次のような特性を持っている．

- 2 文間で少なくとも 1 つ共通する単語があり，各文中に 2 語以上の話題が存在する時，s' 自体の中の話題による連想によって s' 中の語は全て s から連想され, $F(s \to s') = 1$ となる．
- $R^a(s) \cap R(s') = \phi$ の場合，文 $s-s'$ 間に繋がりがないため $F(s \to s') = 0$ となる．
- それらの中間では，s' 中の話題が s から連想される割合に従う $0 < F(s \to s') < 1$ の値をとる．

つまり $F(s \to s')$ によって，共通の単語を含まない文間の繋がりをも検出できる．文 $s-s'$ が文脈によって繋がるためには，s' に含まれる全ての単語が連想される必要はなく，少なくとも 1 つの話題が連想されればよいものと考えられるから，$F(s \to s') > 0$ のとき 2 文間に何らかの繋がりがあると判断する．

3 話題の流れの本筋の推定

まず，文間の話題の流れを用いて導入文や結びの文といった文の役割を推定し，段落内のラフな構造を求めることを考える．

段落 $P = (s_1, \ldots, s_n)$ が与えられたとき，文 s_i への話題の流れの流入量を $CC(s_i|P)$，文 s_i からの流出量を $IC(s_i|P)$ とし，それらの合計によって P 中の流れにおける s_i の寄与度 (重要度)$CR(s_i|P)$ を表す．

$$CC(s_i|P) \stackrel{def}{=} \sum_{j=1}^{i-1} F(s_j \to s_i)/F(P) \tag{6}$$

$$IC(s_i|P) \stackrel{def}{=} \sum_{j=i+1}^{n} F(s_i \to s_j)/F(P) \tag{7}$$

$$CR(s_i|P) \stackrel{def}{=} \frac{CC(s_i|P) + IC(s_i|P)}{2} \tag{8}$$

こうすると $IC(s_i|P)$，$CC(s_i|P)$ はそれぞれ s_i の導入文らしさ，結びらしさを表すことになる．このことから，$IC(s_j|P)$ が最大のもの s_j を導入部 IP，$CC(s_k|P)$ が最大のもの s_k を結び CP と呼び，その間の寄与度が高いものを展開部と呼ぶものとする．IC，CC はそれぞれ段落の先頭，末尾で大きな値をとる傾向があるが，必ずしも第一文と最終文が最大とは限らない．もし IC が最大となる文が複数存在する場合には後順を優先し，CC 最大の段落が複数存在する場合には前順優先とする．

要約対象として主に新聞の社説等の論説文を想定すると，段落 P 中の様々な話題の流れの中で，この導入部から展開部を介して結びへ至る流れこそが作者の意図した文章の主題，すなわちこの段落の文脈の流れの本筋ではないかと考えられる．文章によっては意味的には必ずしも結びが結論とは限らず，導入部が結論の場合も考えられるが，いずれにしろ結論はこの流れの中に含まれるものと思われる．

4 文章の分割と階層化

上記の文脈の流れの検出法は，1 つの段落で構成されたまとまりのよい (結束性が高い) 短い文章に対し

て有効である．しかし文章がある程度以上長くなってくると，段落の区切りなど文脈に隔たりがある部分が現れ，そこで話題の流れが切れてしまうことがある．

読みやすい要約文を得るためには，要約文中に話題の流れの断絶を含まないようにしなければならない．遠く離れた段落内の文が文単位で個別に繋がるとは考えにくいため，複数の段落を持った文章に対しては段落を単位にした繋がりのよさを考える方が自然である．よって本研究では，各段落内での結束性と段落間の結束性を分け，階層的に捉えて取り扱う．

4.1 k-隣接分離法による段落分け

日本語の文章の場合，段落区切りでは字下げすることが通例であるが，新聞記事について調査したところ実際の文章は話題の切れ目よりもかなり細かく段落分けされていることがわかった．そこでここでは，話題の転換点を持った文章を字下げされた段落よりもやや大きめな範囲で自動的に段落分けすることを考える．

まず，文 s_i と文 s_{i+1} の接合点を i で表し，i の k-隣接範囲内の文 $\{s_{i-(k-1)}, \ldots, s_i, s_{i+1}, \ldots, s_{i+k}\}$ 相互の流れについて考える．2 つの文の組み合わせ間の流れには接合点 i を越えないものと，接合点 i を飛び越えるものが存在する．もし i が段落分離点ならば，前者は前後の段落に含まれる局所的な繋がりを表し，後者は段落分離点 i を挟んで前後の段落を跨がった流れを表す．そこでそれらの流れの結束性をそれぞれ $LF(i)$，$OBF(i)$ とおく．

$$LF(i) \stackrel{def}{=} F(s_{i-(k-1)}, \ldots, s_i) + F(s_{i+1}, \ldots, s_{i+k}) \tag{9}$$

$$OBF(i) \stackrel{def}{=} \sum_{m=i-(k-1)}^{i} \sum_{n=i+1}^{i+k} F(s_m \rightarrow s_n) \tag{10}$$

ここでもし接合点 i が話題の転換点であり，i の k-隣接範囲内に i 以外の転換点が存在しないものとすると，$OBF(i)$ は比較的弱く，$LF(i)$ は相対的に強くなるはずである．よって次の条件を満たす接合点 i を段落分離点として選択する．

【段落分離条件】

1. k-隣接文間の話題の流れ $OBF(i)$ が相対的に小さくなる接合点 i を分離点候補とする．

$$OBF(i) < \theta \cdot LF(i) \tag{11}$$

2. 接続詞や照応による繋がりの分断を防ぐため，直後の文 s_{i+1} が従属文である接合点 i を分離点候補から除外する．

3. 分離点候補のうち，実際に段落を分離する点は k-隣接範囲内に 1ヶ所以下とする．

ただし従属文とは，本稿においては文頭に接続詞相当句または照応語句 (連体詞相当句，第一人称以外の代名詞) が存在する文のことを指す．

語彙結束を用いる手法では本来繋がるべきところで語彙連鎖が切れてしまう場合があるが，本手法による段落分割では連想で文間が繋がっていくため，段落内の繋がりが保存されやすい．ただし判定に用いる前後の文数 k は，$k = 1$ では切れすぎるため，$k = 2$ 程度が適当である．閾値 θ の適値は k によって変動するが，実験によると $k = 2$ の場合 $\theta = 1.5$ 程度が最適なようである [10]．

4.2 階層段落の構築

文章 D 中の各文 s_i を 1 文 1 段落とする段落列を 0 階層とする階層的な結束構造 $(D^0, D^1, D^2, \ldots)$ を考える．$n \geq 1$ の n 階層段落列 D^n の構築が，$n-1$ 階層段落列 D^{n-1} 上の段落分離条件による分割によって行われるものとし，n 階層上の 2 段落間の話題の流れのよさ $F(P_i^n \rightarrow P_j^n)$ を次のように定義する．

$$P_i^0 = s_i \tag{12}$$

$$P_i^n = (P_{s_i^n}^{n-1}, \ldots, P_{e_i^n}^{n-1}) \quad (n \geq 1) \tag{13}$$

$$D^n = (P_1^n, \ldots, P_{N_n}^n) \tag{14}$$

$$R(P_i^n) \stackrel{def}{=} \bigcup_{p \in P_i^n} R(p) \tag{15}$$

$$F(P_i^n \to P_j^n) \stackrel{def}{=} \sum_{\rho \in R^a(P_i^n) \cap R(P_j^n)} rf(\rho) \,/\, \sum_{\rho' \in R(P_j^n)} rf(\rho') \tag{16}$$

n 階層段落 P_i^n は $n-1$ 階層段落列 D^{n-1} の部分列である．ここで挙げた以外の式は前節で述べた $n=1$ の場合に準じて与えられる．階層的な段落分けは D^0 から始めて逐次的に実行し，それ以上分割できなくなった時点で終了する．

5 要約文の抽出

5.1 話題の流れの本筋の抽出

段落列 $P=(P_1,\ldots,P_n)$ においても文の列の場合と同様に，文章の本筋の導入部となる IP と結びとなる CP を準備する．流れの良い要約文を構成するためには，話題の流れの本流の中にある結束した段落列を用いるのが適当であると考えられるため，次の方針で要約文に含める段落を選択する．

1. 段落の前置き (手紙では時候の挨拶など) や追って書き (追伸にあたる部分) を省略するため，IP より前と CP より後の段落を削除する．

2. IP から CP の間にある次の削除条件を満たす段落 P_i を結束性への寄与度 $CR(P_i|P)$ の昇順に削除する．

   【削除条件】

   段落 P_i の削除が許される条件は，段落列 P の初期状態における段落数 n_0，要約率をコントロールする閾値 θ_D を用いて，次の式で表される．

   $$CR(P_i|P) < \frac{1}{\theta_D \cdot n_0} \tag{17}$$

3. 残った導入部 IP〜展開部〜結び部 CP に至る段落列を現在着目している階層段落 P の要約とする．

ただし，接続詞相当句や照応による繋がりを保存するため，1 において IP の先頭が従属文である場合には，IP の前の非従属文で始まる段落までを復活させ P の要約に含める．また，2 においては従属段落を持つ段落を削除候補から除外する．ただしそのような段落も，全ての従属段落の削除後に限って削除を許可する．

5.2 文章全体の要約

文章全体の要約は，階層段落の最上階から 0 階層へと下りながら，段落列に対する要約を各層の段落列に適用することにより作成される．システム全体の動作をまとめると次のようになる．

Step.1 要約対象の文章を段落分離条件により選定される分離点で段落分けし，階層段落を構築する．

Step.2 最上階層から順番に各階層の段落列の要約を求める．このとき上位階層で削除された段落に含まれる下位階層の段落を推移的に削除する．

Step.3 最終的に残った 0 階層の文を順に並べたものを要約文として出力する．

6 考察

本手法を用いて毎日新聞'95[12] の社説 50 文に対する要約文を求め，人手により主観的にチェックしたところ，想定した構造に近い結束構造を持つ文章に対しては比較的読みやすい要約文が得られることがわかった．今後は事件記事等，多様な構造パターンに適応させたいと考えている．

話題間の連想は文間の繋がりを把握するために有効であり，語彙結束の改善案として有望であると思われる．しかし実際の文章中では 50 文章中に 2 文章程度，話題の流れが途切れてしまい，あまりよい要約文が作れないものがあった．その原因は，“イチロー”↔“ICHIRO” のような表記の揺れや，“ポーランド”↔” ポ” のような略記法によるものであった．

これらには別途対処する必要がある．

本手法による要約文は，話題の結束性のみで最適化されているため，必然的に重要文抽出法で言うところのあまり重要でない文を包含することになる．NTCIR2[2] の人手による要約率 10%の正解データに対する文再現率を tf-idf 法を用いた重要文抽出法による要約文と比較してみたところ，本手法を用いて tf-idf 法と同等の文再現率を得るためには，要約率を 10〜20%程度高めにとる必要があることがわかった [11]．読みやすさと重要文の文再現率の間にはトレードオフがあり，今のところ両立は難しいようである．

7　おわりに

文間の繋がりの良い読みやすい要約文を求めるため，文間および段落間の話題の流れの良さを用いて階層段落を構築し，話題の本筋と思われる要約文を抽出する手法を提案した．

今後は既存の重要文抽出法と組み合わせるなどの改良を加えることにより，読みやすさと文再現率とのバランスをとりながら，さらに要約文の質を改善したいと考えている．

謝辞

要約システムを C++で実装して戴きました，田中友也君 (現 (株) 東芝 e-ソリューション社) に感謝致します．

参考文献

[1] Inderjeet Mani 著, 奥村学, 難波英嗣, 植田禎子訳. 自動要約. 共立出版, 2003.

[2] NII. *NTCIR Workshop 3 Meeting Overview*. National Institute of Informatics, 2002.

[3] 奥村学, 難波英嗣. 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, 1999.

[4] 奥村学, 難波英嗣. 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, 7 2002.

[5] 砂山渡, 谷内田正彦. 文章要約のための特徴キーワードの発見による重要文抽出法—展望台システム—. 情報処理学会自然言語処理研究会研究報告 NL135-14, Vol. 2000, No. 11, pp. 103–110, 2000.

[6] 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. 情報処理学会自然言語処理研究会研究報告 NL99-3, Vol. 94, No. 9, pp. 17–24, 1994.

[7] 市丸夏樹, 日高達. 文脈の流れを保持した要約文の自動生成. 平成 14 年度電気関係学会九州支部連合会大会論文集, 第 55 回, p. 628, 2002.

[8] 松本祐治. 形態素解析システム「茶筌」. 情報処理学会誌, Vol. 41, No. 11, pp. 1208–1214, 2000.

[9] 大澤幸生, ネルス E. ベンソン, 谷内田正彦. Keygraph: 語の共起グラフの分割・統合によるキーワード抽出. 電子情報通信学会論文誌, Vol. J82-D-I, No. 2, pp. 391–400, 1999.

[10] 田中友也, 市丸夏樹, 日高達. 文脈の流れを保持した要約文の自動生成-paragraph 分け-. 平成 14 年度電気関係学会九州支部連合会大会論文集, 第 55 回, p. 629, 2002.

[11] 飛松宏征, 市丸夏樹, 日高達. 文脈の流れを保持した要約生成手法の評価. 平成 15 年度電気関係学会九州支部連合会大会論文集 CD-ROM, 第 56 回, 2003.

[12] 毎日新聞社. CD-毎日新聞 '95 年版. 日外アソシエーツ (株), 1995.