要約文生成における照応処理

大塚 敬義† 内海 彰‡ 廣田 薫†
† 東京工業大学大学院 総合理工学研究科 知能システム科学専攻
‡ 電気通信大学 電気通信学部システム工学科
otsuka@utm.dis.titech.ac.jp, utsumi@se.uec.ac.jp, hirota@hrt.dis.titech.ac.jp

1 はじめに

単語の出現頻度や，単語間のつながりの情報を用いることによって重要箇所を抽出する従来の自動要約手法の問題点として，抽出した文中に代名詞などの照応表現が含まれている場合，その先行詞が要約文中に存在する保証がないことが指摘されている [1]．つまり，照応表現を含む文や節が要約文として抽出されたとき，それらの照応表現の先行詞が要約文から脱落すると，要約文の結束性が損なわれ，読み手が要約文の内容を誤って理解するなどの問題がある．

しかし従来の自動要約の研究において，このような照応によって生じる問題を積極的に解決しようとする研究は少ない．たとえば代表的な要約文生成システム [2] では，重要文として抽出される文の冒頭に「これ」「その」などの指示語が出現した場合，直前の文と併せて両方の文を採用するという単純な処理しか行っていない．この方法では，要約文が長くなり，なおかつ直前の文内に先行詞があるとは限らないという欠点がある．また先行詞が所在する文を直前の文に限定せず，簡単に推定して文単位で抽出を行う研究 [3] もあるが，その指示語が文頭ではなく文中に出現する場合は先行詞の補完処理が行われないという問題がある．

そこで本研究は，新聞の解説記事や社説を題材として，照応表現を含む文が重要箇所として抽出された場合に原文章から先行詞を語句単位で推定し，その先行詞が重要文に含まれない場合には照応表現を先行詞で補完することによって，結束性を保った要約文の生成を行う手法を提案する．さらに本研究ではアンケート実験を行い，提案手法の有効性を検証する．

2 要約文生成システムの概要

本研究で作成した自動要約システムは，以下に示す5つの処理部から構成され，以下に示す順に処理が行われる．

1. **文章入力部**
   原文章の各文を句点ごとに切り出す．
2. **形態素解析部**
   日本語形態素解析ツール ChaSen を用い，入力文を形態素解析する．このとき普通名詞，固有名詞，サ変名詞，副詞的名詞には分類語彙表[1][4] の索引番号を付与しておく．なお多義語に対しては，その語義の数だけ索引番号を付与する．
3. **重要文抽出部**
   まずは，原文章中のすべての普通名詞，固有名詞，サ変名詞を出現頻度で降順にソートし，以下の式を満たす単語 w を重要語とする．

$$i_w < [n \times 0.3 + 1] \qquad (1)$$

   なお上式において，i_w および n はそれぞれ出現頻度における単語 w の順位，出現頻度が最少回数の単語の順位を表す．
   次に，文章中の各文に含まれる重要語の総数を各文の重要度として与え，重要度の高い順に重要箇所として要約文に採用する．ユーザがあらかじめ指定した要約率を超えた時点で重要文の抽出を終える．
4. **照応表現解析部**
   重要文抽出部で抽出された文に含まれるすべての照応表現について，指示詞の係り先を近似的に同定した後，先行詞の同定を行う．同定された先行詞が重要文に含まれていない場合には，照応表現を先行詞で補完する結束処理を行う（この部分の詳細は次章で述べる）．なお，本研究で対象とする指示語は，連体詞形態の指示語（「この」「それらの」など）のみとする．
5. **要約文出力部**
   得られた要約文を整形して出力する．

[1] 電子版の分類語彙表 (増補版) を用いた．

3 照応処理

照応表現解析部では，以下に示すSTEP1,2,3の処理を順に行い，先行詞を要約文に補完する．

STEP 1 指示詞の係り先の同定
STEP 2 指示の分類，先行詞主要部の同定
STEP 3 要約文への先行詞の補完

3.1 指示詞の係り先の同定

STEP 1では，重要箇所として抽出された文の中にあるすべての連体詞形態の指示詞について，指示詞の係り受けの関係を調べる．

ここでは，以下の条件を満たし，かつ指示詞から最も近い名詞句（本研究ではこれを**受け語**と呼ぶ）を指示詞の係り先と同定する．

- 受け語となる部分には，用言の連体修飾句を含めない．
- 受け語となる名詞句は，単一または複数の普通名詞，固有名詞，サ変名詞から構成される．

3.2 指示の分類，先行詞の主要部の同定

図1に示されるように，STEP 2では，すべての指示表現を以下の3つにいずれかに分類し，その分類に応じて先行詞の主要部を同定する．

- **A. 強い限定指示**
  受け語と同一の文字列が探索範囲中に存在する．
- **B. 弱い限定指示**
  受け語と同一の文字列は探索範囲中にはないが，受け語の主要部が探索範囲中にある．
- **C. 代行指示**
  受け語の主要部となる名詞が探索範囲中にない．

図1では，まず照応表現を含む文の直前5文以内の範囲において受け語を探索する．なお，この範囲において同じ受け語をもつ別の照応表現が存在する場合には，探索範囲をその文の直前5文まで広げる．

受け語と同一の文字列が存在する（A. 強い限定指示と分類される）場合には，その中で照応詞から最も近い文字列を先行詞の主要部とする．

受け語の主要部のみが存在する（B. 弱い限定指示と分類される）場合には，それらの名詞に対して，中心化定理 (centering theory) を利用したセンタリングによる重み s_c と照応詞との距離による重み s_d の和（$s_c + s_d$）を得点として与え，得点の最も高い候補を先行詞の主要部とする．重み s_c は18種類の助詞や句読点のみに対し2.0〜3.5までの4段階で与えられ，重み s_d は照応詞が出現する文と，先行詞の候補とが出現する文とが文単位でどれだけ離れているかに応じて-5〜0までの6段階で与える．ただし，同得点の場合は照応詞から最も距離が近い候補を選択する．

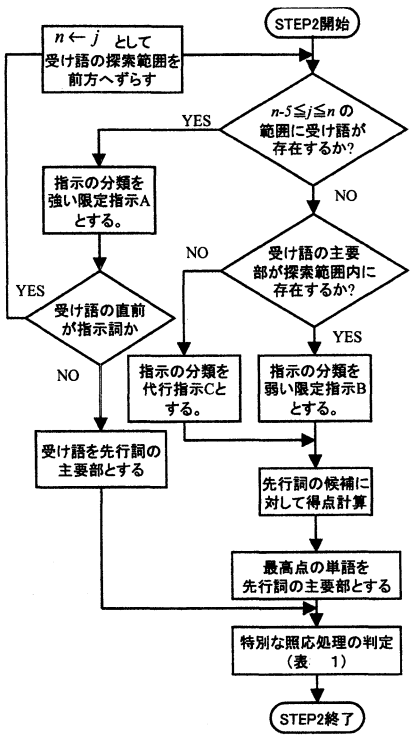

図1 指示の分類および先行詞の主要部を同定するアルゴリズム

受け語の主要部も存在しない（C. 代行指示と分類される）場合には，普通名詞，固有名詞，サ変名詞のいずれかの範疇に属す名詞に対して，上記の2つの重み s_c, s_d および意味情報の重み s_t の和（$s_c + s_d + s_t$）を得点として与え，得点の最も高い候補を先行詞の主要部とする．なお，意味情報の重み s_t は分類語彙表の索引番号を利用し，受け語の索引番号と，先行詞の候補の索引番号とが分類番号や段落番号などの数値がどれだけ一致しているかによって0〜30までの6段階で与える．

ただし図1の処理において指示詞の出現場所や受け語が表1に示す条件と合致する場合は，先

表1 特別な照応処理を要する受け語

(1)	受け語が季節を表す時相名詞 (夏，冬 など) → 3.3 節の処理を行わず照応詞をそのままの形で残す.
(2)	受け語が因果関係や時期を表す副詞的名詞 (ため，結果，場合，点，際など)
	指示詞が文頭に出現する場合 → 前文全体を先行詞と見なして抽出する.
	指示詞を伴い文中に出現する場合 → 図 3.1 の処理で C. 代行指示と判定されたときに限り 3.3 節で先行詞を抽出する際に前文全体を先行詞と見なして抽出する.

病気の原因は不明で根本的な治療方法もまだ見つかっていないが，ボツリヌス菌の**毒素製剤**を目の周りに注射してけいれんを抑える**方法**が開発された．この製剤を輸入している「眼科情報センター」でこの治療方法を**紹介**している．
まだそうした案内をする病院は少ない．

図2 原文章における指示分類の例[2]

行詞を同定する際に特別な照応処理を行う．指示分類の例を図2に示す．強い限定指示 (**毒素製剤**)，弱い限定指示 (**方法**)，代行指示 (**紹介**) のようになる．

3.3 要約文への先行詞の補完

STEP 3 では，STEP 2 で推定した先行詞の主要部を含む文が重要文抽出部で抽出されているかを判定する．もしその文が抽出されていなければ，図3に示すように先行詞が用言を含んだ形にするかを判定する．用言を含めないのであれば，図4に示すように先行詞を普通名詞，固有名詞，サ変名詞のみから構成する．

用言を含める場合は図5のアルゴリズムに従い，以下に示す抽出条件 C_v, C_p, C_q を順に満たすように処理を行った後，S への格納処理を終える．

- S に動詞相当語句 (サ変名詞または動詞) を少なくとも1つ格納する (条件 C_v).
- 動詞相当語句の必須格を含む文節を少なくとも1つ格納する (条件 C_p).
- 条件 C_p 成立後さらに前方にある「が」「は」「,」に達する (条件 C_q).

本研究ではヒューリスティックな方法により必須格を「が」「を」「に」としている．また左鉤括弧 (「) や読点 (，) が来た場合は特例として条件

[2]出典 読売新聞 2000 年 9 月 9 日号 (一部改変).

Step 1	S に先行詞の主要部を代入.
Step 2	原文章を形態素解析して逆順に格納したファイル F からデータを読む．第 j 文で先行詞 S を発見すれば Step3 へ.
Step 3	指示の分類が C で，かつ S はサ変名詞句ならば，S の直後のサ変動詞 v_s を S の後方に格納する．$S \leftarrow S + v_s$
Step 4	ファイル F から形態素 m をひとつ読む.
Step 5	m を S の前方へ格納する．$S \leftarrow m + S$
Step 6	m が体言 (普通名詞，固有名詞，サ変名詞) ならば先行詞生成部 **I** へ．そうでなければ先行詞生成部 **II** へ．なお次の処理のためにファイル F から形態素 m を読み出す位置を保持しておく.

図3 先行詞抽出の前処理のアルゴリズム

最初のファイル F の読み出し位置を図 3.3 の処理から受け継ぐ.

Step 1	ファイル F から形態素 m を読む.
Step 2	m が体言 (普通名詞，固有名詞，サ変名詞) ならば m を S の前方へ $S \leftarrow m + S$ のように格納し Step1 へ．m が体言でないか，または m の文中での位置が文頭ならば先行詞 S の抽出を終了する.

図4 先行詞抽出部 **I** のアルゴリズム

最初のファイル F の読み出し位置を図 3.3 の処理から受け継ぐ.

Step 1	形態素 m が動詞相当語句 (サ変名詞，動詞) ならば 条件 C_v を満たす.
Step 2	F から m を読み，m の文中での位置が文頭ならば S の抽出を終了する.
Step 3	m が『が』『は』『,』のいずれかでありかつ条件 C_p を満たしていれば S の抽出を終了する.
Step 4	$S \leftarrow m + S$ とする.
Step 5	m が (『が』『を』『に』『,』『「』) のいずれかであり，かつ C_v を満たしていれば C_p を満たす.
Step 6	Step 1 へ戻る.

図5 先行詞抽出部 **II** のアルゴリズム

C_p を満たすものとする．指示分類が A, B であれば照応詞全体を先行詞で置き換え，指示分類が C であれば指示詞部分のみを先行詞で置き換える．

図2の例文においては，「この治療方法」の先行詞 S は「ボツリヌス菌の毒素製剤を目の周りに注射してけいれんを抑える治療方法」になる (図6).

ボツリヌス菌の毒素製剤を目の周りに注射して
(C_q / C_p / C_p / C_v)

けいれんを抑える方法
(C_p / C_v)

図6 先行詞抽出部 **II** による条件 C_v, C_p, C_q の適合例

4 評価実験

4.1 実験手法と結果

新聞の解説記事や社説5編を用いて，システムにより生成された要約文に関するアンケートを実施した．これらの要約文の要約率はおよそ 30 ～ 40 %である．

アンケートに用いる要約文は，以下の手順で作成した．まず単語の出現頻度だけを基準として（つまり，2章の重要文抽出部までの処理で）要約文Tを作成する．この要約文Tに対し照応表現が文頭にある場合のみ，前文をそのまま抽出するという文献[2]の手法で先行詞を補った要約文を，要約文Yとする．要約文Tに対し，本研究の提案手法で先行詞を補った要約文を要約文Oとする．アンケートでは，要約文Yと要約文Oを題材として用いた．

アンケートでは，以下の項目について0～4点までの5段階で評定させた．

実験1：要約文Yまたは要約文Oのみを提示して，

1. 要約文全体のまとまりの良さ
2. 要約文全体の理解のしやすさを評定させる．

実験2：要約文と原文章の両方を提示し，

3. 要約文の内容の正確さを評定させる．

アンケートには大学院生14名が参加し，要約文Yと要約文Oについてそれぞれ7名を割り当てた．なお，実験順序は実験1,2の順に行い，実験2の開始後は実験1のアンケート結果を修正させないようにした．さらに，実験1が終了するまでは，被験者に対して不要な先入観を与えないよう，本アンケートが要約文に関するものであることは伝えなかった．

それぞれの評定項目について，アンケートで得られたデータに対して平均値の差の検定を行った．評価項目1および2については，要約文Oのほうが要約文Yよりも評価が高いという有意な傾向にあった（$t_4 = 0.941, p < 0.2$）．

評価項目3ついては，要約文Oの評価のほうが要約文Yの評価よりも有意に高かった（$t_4 = 1.533, p < 0.1$）．これらの結果は，本研究の提案手法のほうが優れていることを示すものである．

4.2 考察

提案手法による要約文を読んでおらず，従来手法による要約文を読んだある被験者からは，照応詞に対応する先行詞が脱落しているときは，個々の文を理解できたとしても文章全体を理解できない，という意見が挙がった．また提案手法は要約文C,D,Eにおいて，代行指示の先行詞が厳密には正解となってはいないが[3]，しかし評定項目3で高評価になっており，近似解となる先行詞を要約文に補完することで要約文の内容の正確さを向上させていると考えられる．

よって要約文生成という分野では単に重要な文を採用するだけでは不十分であり，文の重要度以外の要素を考えることによって要約文の結束性が向上する可能性があることを本研究で示すことができた．

またある被験者からは，従来手法による要約文は一見して問題が無いように見えるが，実際に原文章と比較すると正確さを欠いている，という意見が挙がった．提案手法ではこうした点を可能な限り改善したことで，提案手法の有効性を示すことにつながったと思われる．

5 おわりに

本研究は要約文生成の過程で生じる照応処理の問題に対して，要約率を抑えつつ要約文全体の結束性を向上させる手法を示した．

またアンケート結果から，要約文において要約率を抑えつつ結束性を維持するためには，接続詞を挿入する処理や，複数の文をひとつにまとめるなどの処理が必要であることがわかった．今後は前者を前者を実現するために自然言語理解の知見を応用し，かつ後者を実現するために自然言語生成や修辞構造理論の知見を採り入れていきたいと考えている．

参考文献

[1] 奥村学，難波英嗣，テキスト自動要約に関する研究動向(巻頭言に代えて)，自然言語処理，Vol.6, No.6, pp.1–26 (1999).

[2] 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*，自然言語処理，Vol.2, No.1, pp.39–55 (1994).

[3] 間瀬久雄，大西昇，杉江昇，説明文の抄録作成について，言語理解とコミュニケーション研究会 NLC89-40, pp.5–12 (1989).

[4] 国立国語研究所，分類語彙表 増補版，(1996).

[3] 先行詞が正解かは，中学校国語教員の免許を持つ第三者が原文章を参照して判定した．