要約者ごとに異なる要約戦略の実験による分析

斉藤 喜永子（東京女子大学現代文化研究科, E-Mali: kiekom@twcu.ac.jp)
中川 裕志　（東京大学情報基盤センター, E-Mali:nakagawa@r.dl.itc.u-tokyo.ac.jp)

1. はじめに

大量のテキストを効率よく活用するための技術のひとつである自動要約に焦点をあて，その方法を探る目的で人間による要約の多様性を実験的に分析した結果について報告する．この研究では，要約を「原文の情報を選別し取り込む作業」ととらえ，要約の中に原文が取り込まれる現象を「取り込み」，原文の文節の中で要約へ取り込まれた部分を「取り込み文節」と名づけた．

2.実験

2.1 実験の概要

140名の文系大学生を対象に，1999年12月9日の朝日新聞のコラム「天声人語」(741字)を原文として，150字(20%)程度と250字(34%)程度との2種類の要約率の要約を作成させた．一回目の要約の経験に左右される状態を避けるために，各要約者は用意された2種の要約率のうち，片方の要約率の要約のみ作成した．

2.2 原文と要約文の対応付け

要約への原文の取り込みを評価するために，要約と原文の対応付けにより，取り込み文節を認定した．文を単位とした対応付けをすると，要約文内で文が連結された場合等に対応付けが不正確となる．そのため，文より小さく，かつ，原文内での位置を示す手助けとなる格表示やテンス等の情報を持つ文節を対応付けの単位として使用した．事前に用意した対応付け規則[図1]と教示文「文章中のどの部分を元にして書いたのか，その部分に下線(＿＿＿)を引きなさい.」に従い，要約作成時に要約者がひいた下線を手掛かりとした．更に，未完成な要約，意味が通じない要約，要約で使用されている単語が原文と違いすぎて対応付けできない要約を除外した．

(1)完全一致
　原文の文節の文字列がそのまま要約に取り込まれているもの.
(2)パラフレーズ
　語彙的パラフレーズと文法的パラフレーズに分類
(3)文字列の挿入
　原文の文節のどこに挿入されたかで分類

図1．要約と原文の対応付け規則

この対応付けにより要約は，(1)完全一致(2)パラフレーズ(3)原文にはない文字列の挿入，に分類された．このうち，(1)と(2)を取り込み文節と認定した．(3)を原文と結びつかない要約者の作文と考え，取り込み文節と認定しなかった．この分類の結果を元に，原文の文節が要約ごとに取り込まれていた場合＝1，取り込まれていない場合＝0 とする[表1]のデータが完成した．

表1. 各要約者の取り込み文節のデータ

文節No	原文	要約者No				
		1	2	3	139	140
1	エレベーターへの	1	1	1	0	1
2	不平・不満と	1	1	1	0	1
：						
164	エレベーターに	1	1	0	1	1
165	限らない	1	1	0	1	1

2.3 要約者のクラスタリング

同一の原文から人手で作成される要約に現れる多様性を分類する目的で，要約者のクラスタリングを行った．まず，表1に示す各要約者の取り込み文節のデータから，要約者間の取り込みの類似度(ハミング距離で測る)を求め，クラスタ分析(WARD'S法)を行った．完成したデンドログラムから全体が3グループに分割される階層で被験者を分類した．150字要約の要約者集団と250字要約の要約者集団それぞれに対し，この作業を行い，6つのクラスタを得た．

2.4.クラスタごとの集計

原文を構成する全165文節について，各文節がクラスタ内のどれだけの割合の人の要約で取り込まれたかを示す取り込み率の表を表2に作成した．

表2. 各クラスタの取り込み率

				250字要約		
段落 No	文 No	文節 No	原文	クラスタ1 26人	クラスタ2 9人	クラスタ3 24人
1	1	1	エレベーターの	1.00	0.78	0.88
1	1	2	不平・不満を	1.00	0.89	0.92
			：			
6	28	165	限らない	0.81	0.89	0.54

文節1は，段落1・文1にあり，クラスタ1での取り込み率は1.00
取り込み率=クラスタで当該文節を取り込んだ人数÷クラスタの人数

更に，表3に示す取り込み傾向の塗りわけの基準を用いて，表2を塗り分けた図2を作成した．クラスタごとの取り込まれる傾向の文節と取り込まれない傾向の文節の分布を表した．

表3. 取り込み傾向の塗り分けの基準

取り込み率	文節の分類	塗り分け
0.60以上～1.00	取り込まれる傾向	灰色
0.00～0.20以内	取り込まれない傾向	網掛け

3.結果と考察

3.1.全体的傾向

3.1.1.要約作成にかかわる単位

取り込みが文か分節かを考察するために以下の手順を行った．まず，表1のデータから，文がどれだけの割合が取り込まれたかを示す「文の取り込み率」を求めた．次に，文の取り込み率の高い部分と高くない部分の2種に分類をした．分類の境界となる，文の取り込み率を0.6，0.7，0.8，0.9，1.0の5段階に変化させ，それぞれの段階で，クロス表を作り，互いに従属な二つの比率の検定をした．

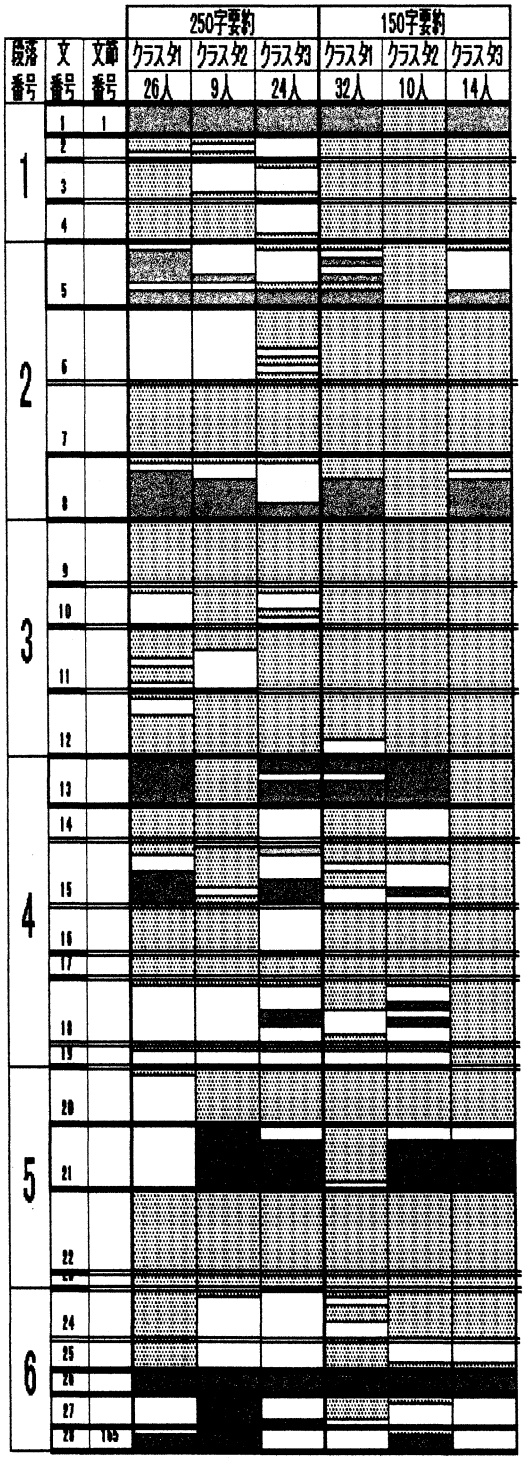

図2. 原文の文節の取り込み傾向の分布
文の境界　文1.5.8.13.21.26.28は3.1.2の核

表4. 取り込み単位の境界を変えたクロス表の作成

表1のデータ → 文の取り込み率へ

段落No	文No	文節No	原文	要約者No 1	2	3
1	1	1	エレベーターへの	1	1	1
1	1	2	不平・不満を	1	1	1
1	1	3	よく	1	0	1
1	1	4	耳にする	1	1	1
1	2	5	いくら	0	0	0

段落No	文No	要約者No 1	2
1	1	1.0	0.8
1	2	0.0	0.0

1.0以上で文単位と認定

段落No	文No	要約者No 1	2
1	1	文	文節
1	2		

0.6以上で文単位と認定

段落NO	文NO	要約者NO 1	2
1	1	文	文
1	2		

クロス表へ

1.0で分類した集計

文	文節	合計
410	722	1132

クロス表へ

0.6で分類した集計

文	文節	合計
840	292	1132

その結果，境界を 0.6，0.7 とすると文単位の取り込みが有意に多くなった($P<0.001$)．0.8 では，有意差は現れず，0.9，1.0 では結果が反転し，文節単位の取り込みが有意に多くなった($P<0.001$)．これは，要約でおきる原文の取り込みは，文節より文に近い単位であることを示唆すると共に，文が完全に取り込まれることは少ないことも示している．よって，今回の実験では，ほぼ文を単位とした原文の取り込みに文節単位の抜け落ちが多少生じたと解釈される．

3.1.2.要約に取り込まれる原文の傾向

要約被験者全体での原文の各文節の取り込み率をもとめ，各段落ごとに図3に示す文節の取り込み率の推移のグラフを作成した．その結果，原文の一部が多くの要約者に取り込まれていた．また，図2でも，複数のクラスタで，灰色の部分と核の部分とが一致している傾向が見える．これらの傾向は，各クラスタの要約文で，取り込まれる原文はある一部に限られていることを示している．これらの文節の取り込み率の推移のグラフより，6割以上に取り込まれている文節を含む文を文単位で抽出するとほぼ意味が通じる要約文が完成する．このため，原文のこれらの文を「核」と名づけた．

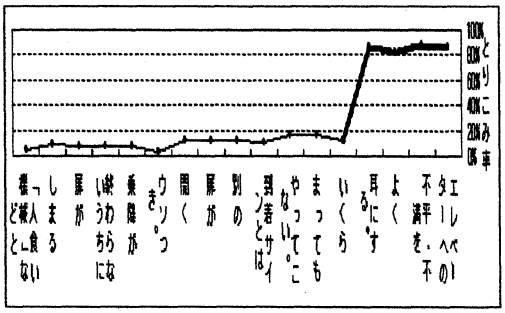

図3. 文節の取り込み率の推移(第1段落)

取り込み率＝当該文節を取り込んだ人数÷要約者の人数
太線部分が『核』

3.1.3.要約戦略の類型化

図2では，250字要約では150字要約の方が，白く残る部分が少ない．この部分は，被験者間のばらつきを反映した部分と考えられる．全要約者の取り込み文節を，核に分類される文節と核以外に分類される文節とに分類し，その構成比を図4のグラフとした．

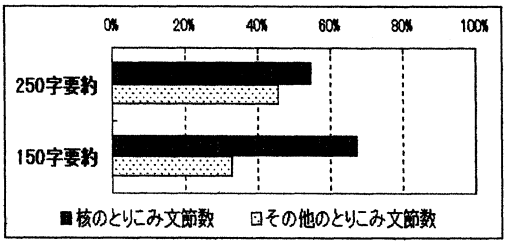

図4. 要約率の違いと核の取り込みの割合の関係

150字の要約では，「核の取り込み」の割合が増え，ばらつきにあたる「その他の取り込み」が減少している．この傾向より，要約率を下げることは，被験者の間にみられるばらつきを減少させる効果があると解釈できる．

3.2.クラスタごとの位置情報による分析

まず，クラスタ分析により作成された分類が，文節の取り込み方で有意な分類であるかを検定した．それぞれのクラスタでの各段落からの取り込み文節を数え上げ，クロス表を作成し，χ二乗検定を行った．その結果，有意差が認められた($P<0.001$)．この結果は，各要約率の3つのクラスタにおいて原

文での取り込みの分布が等しくないことを示している．これは，クラスタによる分類が取り込まれる原文の分布の偏りを反映していると解釈される．

次に，クラスタごとの取り込まれている原文の偏りを位置的に把握する目的で，各クラスタの要約の原文からの各段落の取り込みの偏りをもとめた．その結果，取り込み文節の位置的な偏りに，①散在型②中抜き型③後半型の３種が観察された．

取り込みの偏り＝
段落の取り込みの取り込み全体での割合÷原文での段落の割合

①散在型

このタイプは，段落からの取り込みを原文での位置による特徴づけができない．図２の250字要約１と150字要約１から核の取り込みの分布を比較すると，共通して文の前の方に位置する核が取り込まれ，後ろのほうは取り込まれず，結論の核が取り込まれている．更に，150字要約では，第４段落と第６段落の核からの取り込みが減っている．最初のほうの核から順に取り込む方法で要約を行い，字数が埋まりそうになると，結論まで話題をとばす方法で要約率を低めると解釈される．

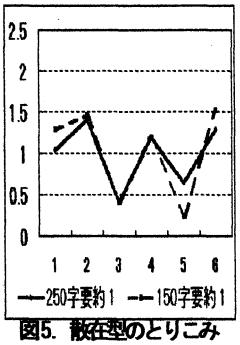

図5. 散在型のとりこみ
図中の番号は図２でのクラスタの番号．図6-図７でも同様の記述方法．

②中抜き型

このタイプの段落からの取り込みは中央の２つの段落で落ち込む．導入部分と結論部分のみで構成される要約文である．図２の250字要約２と150字要約３を比較すると，第６段落の核の周辺から取り込みが減少している．このタイプの要約文は，文のはじめと終わりの核を取り込む方略を使用し，要約率を低めるために，核の周辺に余分に取り込まれていた部分を削る方法をとっていると解釈される．

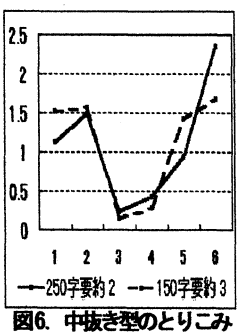

図6. 中抜き型のとりこみ

③後半型

このタイプの段落への依存度が第４段落以降の後半部分で高まる．一般論から結論を中心とした要約である．図２の250字要約３と150字要約２を比較すると，150 字要約でのリード文の欠落が見られる．よって，このタイプの要約文は，文の後半の核を取り込む方略を使用し，要約率を低める方法として余分に取り込まれていた前半部分を減らしていると解釈される．

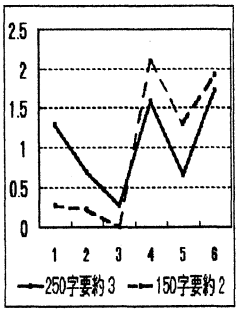

図7. 後半型のとりこみ

謝辞

本研究に有益なアドバイスをしてくださいました東京女子大学の指導教員西原鈴子教授，データの収集に協力してくださった学習院大学の高瀬誠先生に感謝致します．

参考文献

佐久間まゆみ編(1989)『文章構造と要約の諸相』くろしお出版

永野賢(1986)『文章論総説』朝倉出版

邑本俊亮(1998)『文章理解についての認知心理学的研究』風間書房

奥村学・難波英嗣(1999)「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』言語処理学会p1-p26