重要文と要約の差異に基づく要約データの収集と分析

望主 雅子[*1] 荻野 紫穂[*2] 太田 公子[*3] 井佐原 均[*3]

[*1] リコー ソフトウェア研究所
[*2] 日本ＩＢＭ 東京基礎研究所
[*3] 郵政省 通信総合研究所

masako@src.ricoh.co.jp, shiho@trl.ibm.co.jp, {kimiko, isahara}@crl.go.jp

1 はじめに

近年、生活の場に様々な情報が氾濫し、これらを効率的に取捨選択するための重要文抽出、要約技術が研究されてきている。しかし、人間が行なう要約とはかなりのギャップがあると予想される。

人間の行なう要約を知る一つの手がかりとして、人間の作成した要約がどのような表現で構成されるのかを、元となった原文中の表現との対応で考察する研究がある [1][2]。

我々は、人間の行なう要約の性質を知るために、現状技術との差異に基づき要約データの収集、分析を行なったので報告する。

2 要約収集実験

2.1 実験設定

以下の３段階で重要文、要約収集を行なった。

1. 被験者に新聞記事を読ませ、要約を作成させる
2. 被験者に同じ記事を読ませ、重要文を抽出させる
3. 被験者に、(2) で被験者自身が抽出した重要文をできるだけ使わせ、要約を作らせる

(1) が自由作成の要約 (以下「要約１」と呼ぶ)、(3) が制限を加えた要約 (以下「要約２」と呼ぶ) である。

ゴール[1] となる要約の性質を、現状技術 (重要文抽出) や次ステップとして考えられる重要文をつなぎあわせた要約との違いから明らかにしたいと考えた。

実験ではこれらの３種類のデータを同一被験者が同一対象に対して行なうことで、比較データ間での被験者による重要個所の違いや、言語理解・生成の際の個人差による影響を少なくなるよう試みた。また、被験者には上記の (1)〜(3) の手順を知られないよう、郵送形式で１ステップずつ行なった。

[1] よい要約とは何なのか、また目的・用途によって望ましい要約は異なる。これらは大きな課題だがここでは単純に人が作成した要約とした

2.2 対象記事と実験手順

毎日新聞 96 年度版 [9] の３記事を対象に、100 人の被験者に作業を行なわせた。

要約結果は対象の長さや論理性によって変わると言われている。今回の実験では多くの被験者の要約結果を収集するため、最大でＡ４で１枚半程度の長さにとどめた。被験者の理解のばらつき、誤解を少なくするため一般的な内容の記事にした。

論理的な展開のある社説と、意見の対比はあるが会話等の挿入された構成が複雑で長いもの (総合)、論理展開が希薄で、会話表現を含むもの (芸能) とを選択した。要約率は、事前に試行して決め、強い制限にならないよう幅をもたせた。重要文抽出では、あらかじめ文の終了個所にマークしたものを用意し、ランク付けは行なわなかった。それぞれの対象の文字数と字数制限 (目標要約率) を示す。

- **対象１：社説「野茂よ感動をありがとう」**
  1288 文字,24 文 (以降では「社説」「野茂」と表記)

タイプ	字数制限 (要約率,%)
要約１	300-400 字 (23-31)
重要文	5-10 文 (21-42)
要約２	200-400 文 (15-31)

- **対象２：総合「イチローの球宴登板」**
  2903 文字,59 文 (以降では「総合」「野球」と表記)

要約タイプ	字数制限 (要約率,%)
要約１	400-700 字 (14-24)
重要文	12-21 文 (20-35)
要約２	300-700 字 (10-24)

- **対象３：芸能「全国ツアーを始める、安室奈美恵」**
  1654 文字,40 文 (以降では「芸能」「安室」と表記)

要約タイプ	字数制限 (要約率,%)
要約１	300-400 字 (18-24)
重要文	8-15 文 (20-37.5)
要約２	200-400 字 (12-24)

3 要約と原文との対応付け

分析のために、要約と原文の対応付けデータを作成した [10]。対応付けはまずツールで行ない、さらに人手による修正を行なった。

3.1 要約と原文の対応付けツール

- 対応品詞、スコア、対応の向き
  対応付けを行なう文・文節内の単語の文字列の一致をスコアとし、対応する文節を判定する。一致の判定で対象とする品詞や単語を正規表現風に指定できる (今回は自立語、大分類品詞)。要約の各語句を対応元とし、原文を対応先とした[2]。

- 対応付けのステップ：
  1. 要約、原文の全文を範囲とし、文単位で対応付けを行なう。
  2. スコアが既定値以上の文のペアを対象に、文を範囲として文節ごとの対応スコアを計算する
  3. 既定値以上のスコアを持つ文節ペアのうち、複数の対応先候補がある場合は、直前もしくは直後の対応先の文節と隣接している対応先候補を選択する。

3.2 人手での対応付け作業

要約の各語句が原文のどの語句に対応するかについて、人間でも判定の難しいものがある。対応の仕方が特殊なものについてラベルを付与した (表1)[3]。

ラベル	内容
複数	複数の表現をまとめて一表現にしたもの
別表現	いい替えなど別の表現になっている
語単位の違い	文節や辞書登録単位の違い
表記の違い	字種の違いなど表記法の違い
不明	対応がつくかどうか不明なもの

表1：ラベルの種類

3.3 ツール付与の精度

人手で修正した対応結果を正解集合として、ツール付与の結果の精度を調べた。

対象	タイプ	適合率	再現率	F-measure
社説 (野茂)	要約1	80.4	73.1	76.6
	要約2	99.5	88.5	91.4
総合 (野球)	要約1	63.2	58.6	60.7
	要約2	88.0	89.0	88.5
芸能 (安室)	要約1	80.8	64.0	71.4
	要約2	93.8	91.2	92.4

表2 ツール付与の精度

要約2は原文の表現をできるだけ使う実験制約から高くなった。ツールでは文字列の一致をもとに対応を判断するため、別表現や複数の表現をまとめている場合の判断ができない。これらの現象は、全対応中で、要約1で14〜20%、要約2で5%前後であった。

4 要約と原文 (重要文) の対応付けに基づく分析

4.1 重要文と要約の原文対応個所の違い

自由作成の要約 (要約1) と、文という単位で情報を取捨選択した重要文とでは、盛り込まれる情報や作業自体に違いがあると予想される。まず、要約1と重要文の、文単位での原文に対する対応 (選択) 率を以下に示す。

対象	自由要約 (要約1)	重要文
社説 (野茂)	47.1	32.0
総合 (野球)	31.8	27.5
芸能 (安室)	29.9	27.8

表3：原文に対する要約1、重要文の対応率 (文単位,%)

要約の方が対応率が高く、原文の情報がより多く盛り込まれている。しかし、社説では50%近いのに対し、その他の対象では低く、対象によって異なっていた。

また、被験者ごとに、自由作成の要約1を正解とし、要約1が対応した原文の個所を含む文を正解集合とした重要文の再現率、適合率を算出した[4]。

対象	再現率 (%)	適合率 (%)
社説 (野茂)	53.3	78.6
総合 (野球)	53.3	62.2
芸能 (安室)	54.9	59.1

表4 要約1の原文対応個所に対する重要文の再現率/適合率

要約で対応した個所と比べると、重要文はその55%程度をカバーするに過ぎなかった。ただ、再現率が100%の被験者もいた。適合率は再現率よりも高い。多くが選択した文は要約1、重要文である程度一致していた。

表5は社説 (野茂) で、比較的多くの被験者が選択した文を、選択しなかった被験者のうち、要約1では引用[5]していた例である。

文番号	文選択した被験者数	文選択しなかった被験者数	要約1で引用
1	75	25	22
2	39	61	39
16	43	57	28
24	81	19	12

表5：文選択しなかった被験者のうち要約1では引用した数

[2] 原文、要約ともに形態素解析 [7] と文節生成 [8] を行ない、単語、文節単位に分割したものを入力とし、結果は (対応元タグ、対応先タグ、スコア) の形式で出力される

[3] 「不明」は以降の対応結果からは除外。一つの対応について複数のラベル付与を許している

[4] A:要約1の原文対応個所を含む文と重要文の一致文数
B:要約1の原文対応個所を含む文の数
C:被験者が選択した重要文の数
再現率 = A/B, 適合率 = A/C

[5] いわゆる「引用」とは異なるがここでは「用いた」意味で使用

例えば第1文は重要文として100人中75人が選択した重要度の高い文であるが、重要文として選択しなかった25人のうち、22人が要約1の中では引用していた。

以上の現象から、自由作成の要約では広く情報を集めており、重要文は文という単位のため要約に比べ情報が落ちる傾向があり、重要な事柄をごく少ない量で表現する作業になったと思われる。

4.1.1 重要文/要約だけで選択された文(個所)

要約1にだけ選択(引用)された文(個所)に関して

事実の列挙、詳細化した情報が多く該当し、要約1には広く様々な情報が盛り込まれていたことがわかる。

また、社説で話題が大きく展開する接続詞を含む文が該当していた。これは、話題展開する後半部分の先頭に位置する接続詞だが、その接続詞の属する文自体の重要度が低いため、選択されなかった。論理展開の単位が文単位でないことに起因する。

このように重要文抽出では論理展開を考慮できないので、重要文をもとにした要約2では、情報として重要(要約1、重要文とも80%以上が選択)な文が論理展開を明示する文や直前の根拠となる文がないために落ちるという現象が見られた。

芸能、総合では主題を表す表現が分散し、短く区切られた文が該当した。特に後続文との結び付きが強い、体言で終了する文に多くみられた。

重要文としてだけ選択された文(個所)に関して

記事の主題に関わる会話内容、口語的表現、段落先頭の小見出しに近い表現は重要文としてだけ抽出されていた。要約1ではこれらは選択されなかった。

4.2 要約と原文との対応

4.2.1 要約の各文節の原文との対応率

要約で使用された表現がどの程度原文と対応付けられたかを調べた。要約の各文節の対応率を表6に示す。

記事	タイプ	対応率(数/文節数)
社説(野茂)	要約1	88.2(7178/8129)
	要約2	95.1(6347/6677)
総合(野球)	要約1	90.0(11628/12917)
	要約2	91.4(10266/11228)
芸能(安室)	要約1	88.3(7387/8365)
	要約2	91.9(6734/7326)

表6：要約の原文との対応率

自由作成の要約1でも9割近くが原文と何らかの対応をつけることができることがわかった。

但し、対応の中には対応付けの判断の難しいものがある。複数の表現に対応するもの、別表現のうち文字列が類似しないものの割合が、全文節中で要約1では7〜11%、要約2では2.4%程度あり、この分、対応率が下がる可能性がある。

4.2.2 要約と原文との対応の種類

対応付けられた表現について対応の種類によってどのような要約手法がとられたかをみる。単純に文字列を引用したものか、複数をまとめたものかでとられた要約手法の難しさは違ってくる。表7に、対応の種類(ラベル)ごとの割合を示す[6]。

記事	タイプ	ラベル総数	複数	別表現1	別表現2	表記	語単位	ラベルなし
社説(野茂)	要約1	8385	2.50	8.57	6.32	1.51	9.10	71.99
	要約2	6682	0.55	2.35	2.45	0.82	3.07	90.75
総合(野球)	要約1	14716	1.58	13.09	4.87	1.25	14.62	64.60
	要約2	10869	0.59	3.13	2.32	0.70	3.39	89.88
芸能(安室)	要約1	8975	2.35	13.31	4.60	1.91	11.47	66.36
	要約2	7034	0.23	2.87	1.69	1.22	2.36	91.63

表7：対応の種類の割合(全対応のラベル数で除算)

前述の対応率では要約1、要約2とも比較的高かったが、ラベルの付与されない通常の文字列一致が要約2では9割を占め、要約1では65〜71%であった。できるだけ表現を使うという制約の要約2を基準に、自由作成の要約(要約1)では、各手法(種類)がどのくらい増えるのかが推測できる。

4.3 要約と原文での文節の出現順の傾向

要約の表現の順序が原文の表現の順序をどの程度保持しているかを対応付けられた原文と要約の文節の出現順序に着目して調べた。被験者ごとに、要約中での文節位置(y軸)と対応する原文での文節位置(x軸)とが、直線関係にどのくらいのるのかを、対応付けられた文節間の位置の相関係数によって調べた。

図1は相関係数0.98(直線関係にある)と-0.1(直線関係にない)の散布図の例である。

[6] 「複数」は複数表現をまとめたもの、「別表現1」は文字列の類似が半数未満のもの、「別表現2」は文字列の類似が半数以上のものである。

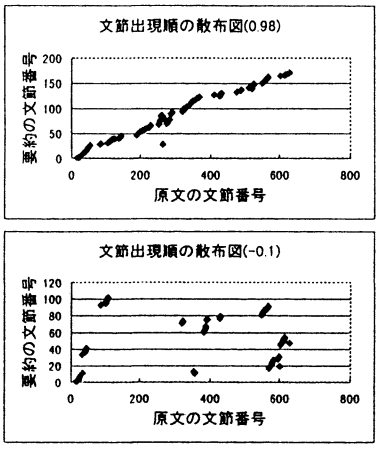

図1: 単語の出現順の散布図例

以下に相関係数ごとの該当被験者数を示す。

設定	0.9 以上	0.9-0.7	0.7-0.5	0.5 未満
野茂・要約 1	68	16	8	8
野茂・要約 2	84	6	5	5
野球・要約 1	67	22	7	4
野球・要約 2	84	10	4	2
安室・要約 1	69	24	3	4
安室・要約 2	91	8	0	1

表 8 : 相関係数 (直線関係) と該当被験者数 (100 人中)

要約 1 の設定では被験者の 67〜69%が、要約 2 では 84〜91%が原文の表現の順番を保持した形で要約を作成していたのがわかる。しかし、0.5 未満の文節の順序を大幅に変更した要約も 20〜33%あることがわかった。「表現をできるだけ使う」要約 2 では表現を利用することから自然と表現の順番も原文に近くなったようだ。また、被験者ごとに相関係数の差は少なく、語順の傾向は被験者による面が大きい可能性がある。

5 まとめ

新聞記事 (社説、総合、芸能) を対象に 100 人の被験者に対して自由作成の要約と、重要文、重要文をできるだけ使った要約を収集し、要約と原文の対応付けに基づき、分析を行なった。

- 重要文と自由要約の違い
  自由な要約ではより広い範囲から情報をとり、原文の情報や論理構造を保存する傾向にある。重要文では、要約に比べ詳細化した情報と論理展開部分が落ちた。そのため重要文を経た要約では情報、論理展開ともシンプルになった。

- 要約のスタイルの概観
  要約を、原文との対応率と文節順序の保持率をあわせてみることで被験者ごとの要約のスタイルが推測できる (図 2)[10]。

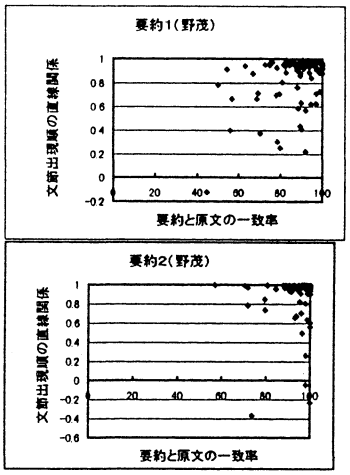

図 2 : 記事(社説・野茂)の全被験者の一致率と文節出現順の散布図

要約 1 、要約 2 とも、7 割近くの被験者は右上の相関係数 0.8 以上、対応率 80%以上に位置していた。要約 1 では離れたところに位置する被験者もいるが、要約 2 では、これらの被験者の数が減る。これに対応の種類 (手法の難しさ) を加えることで各被験者のスタイルをより明らかにできる。

参考文献

[1] 佐久間まゆみ編 1989 「文章構造と要約文の諸相」くろしお出版

[2] 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

[3] 難波英嗣、奥村学 1999 「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-133-8

[4] 加藤直人 1998 「ニュース文を対象にした自動要約-局所的要約知識の自動獲得-」言語処理学会第 4 回年次大会ワークショップ

[5] H.Jing , K.R.McKeown. 1999 「The Decompositon of Human-Written Summary Sentences」 In Proceedings of SIGIR'99

[6] 春野雅彦 1999 「辞書と統計を用いた対訳アライメント」情報処理学会研究報告,96-NL-112-4

[7] 黒橋禎夫, 長尾 真 1999 「日本語形態素解析システム JUMAN version 3.61」

[8] 黒橋禎夫 1998 「日本語構文解析システム KNP version 2.0b6」

[9] 毎日新聞社 1998 「CD-毎日新聞 9 6 年度版」

[10] 望主雅子、荻野紫穂、太田公子、井佐原均 2000 「重要文と要約の差異に基づく要約手法の調査」情報処理学会研究報告,2000-NL-135-13