自 然 言 語 処 理　135－13
（2000．1．28）

重要文と要約の差異に基づく要約手法の調査

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

[*1] （株）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

文書中から重要文を抽出する技術が開発されてきているが、人間の作成した要約とはかなりのギャップがある。我々は、人間の作成した要約の性質を知るために、新聞記事を対象に100人の被験者から自由作成の要約、重要文抽出、重要文をつなぎ合わせた要約を収集し、２種類の要約の差異に基づく調査を行なった。２種類の要約について、ツールと人手によって原文との対応付けを行ない、この対応結果から、重要文と要約での原文情報のカバーの違い、要約の原文との対応率、語順の傾向、指示詞と接続詞の使用について調べた。

A Comparison of Summarization Methods based on the Differences between Important Sentences and Summaries

MOTINUSI Masako[*1], OGINO Shiho[*2], OHTA Kimiko[*3], ISAHARA Hitoshi[*3]

Software Research Center, RICOH Co., Ltd.[*1]
Tokyo Research Laboratory, IBM Japan Ltd.[*2]
Communications Research Laboratory, M.P.T.[*3]

Several techniques have been developed to identify important sentences in documents. However, a large gap still exists between the results from such important sentences by machine and summaries written by human readers. We conducted a set of evaluations in order to better understand these differences.

The target documents were newspaper articles, and 100 people were asked to do each of the following:

1. freely construct a summary
2. identify the important sentences in the document
3. create a summary using the important sentences

We then analyzed the results, both by automated analysis and human inspection, by comparing:

A. the freely constructed summary with the original document
B. the summary based on the important sentences with the important sentences themselves.

The comparisons are reported as: differences in coverage of document information between important sentences and summaries, correspondence ratio between the summary and original document, word order bias, and usage of conjunctions and demonstrative words.

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が、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。

要約という言語活動は、日々人間が何らかの形で行なっているものと考えられるが、人間の言語理解や言語生成とも深く関わるためそのメカニズムをとらえることは難しい。

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。

人間の言語活動としての要約について、佐久間 [1] らは、原文を「残存認定単位」という節や句に近い単位に分割し、要約での残存傾向や要約時にマークさせた手がかり個所との対応傾向で分析した [2]。邑本 [3] は、認知心理学の立場から要約を人間の理解過程に関わる活動として要約データ収集実験を行なっている。単文に近い単位を分析単位とし、原文と要約間の表現の対応と変換を１２種類に分類し、考察している。

重要文から要約へ近づくための研究として、文の部分を出力・生成する試みや、抽出した文のまとまりを読みやすく書き換えるなどの研究も行なわれてきている。特に、重要文やその一部分をもとに言い換える実際的な研究も始まっている [4]。

我々は、人間の行なう要約の性質を知るために、現状の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。

今回の実験設定は以下を目的としたものである。

- 重要文、重要文を経た要約との違いから要約の性質を明らかにする

  ゴール[1] となる要約の性質を、現状技術 (重要文抽出)、次ステップとして考えられる重要文や文の部分をつなぎあわせた要約[2] との違いから明らかにしたいと考えた。

- 現状技術の不足点の段階的把握とデータ収集

  ３種類のデータは現状とゴールとなる要約と、その中間地点をたどることになり、現状技術の不足点を段階的に把握しつつ、データの収集ができると考えた。

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が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ステップずつ独立に行なった。結果を返送した被験者に対してだけ、次の作業指示を与えた。

2.2 対象記事と被験者

100 人の被験者に作業を行なわせた。期間は約１か月。対象は、毎日新聞 96 年度版 [11] の３記事を対象にした。

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、論理的展開や意見の対比はあるが会話等の表現の含まれたやや長いもの (総合)、論理展開をあまり感じさせないもの (芸能) を選択した。

要約率は、強い制限にならないように幅をもたせ、指示時にもその幅におさまらなくともよい旨を伝えた。要約率は事前に試行して決めた。

重要文抽出では、あらかじめ文の終了個所にマークをいれたものを用意し、被験者に、重要と判定した文に下線をひかせた。重要文の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。

---

[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] 重要文として一度絞らずに原文から直接「できるだけつなぎ合わせて要約」させることもできるが、使用する文数を絞ることで「つなぎ合わせる」作業を確実にさせるという狙いもある

- 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「野茂」と表記する)

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「野球」と表記する)

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「安室」と表記する)

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)

2.3 要約結果

平均文字数、要約率を示す。重要文の括弧は文数である。要約率は、要約文字数/原文 (重要文) 文字数 で算出した。

対象１：社説「野茂よ感動をありがとう」

タイプ	平均文字数	平均要約率	最大文字数	最小文字数
要約１	360.1	27.9	532	287
重要文	456.6	35.4	712(12)	215(5)
要約２	300.6	68.3	426	181

対象２：特集「イチローの球宴登板」

タイプ	平均文字数	平均要約率	最大文字数	最小文字数
要約１	575.8	19.8	825	376
重要文	827.4	28.5	1261(22)	423(7)
要約２	497.6	63.5	832	282

対象３：総合「全国ツアーを始める、安室奈美恵」

タイプ	平均文字数	平均要約率	最大文字数	最小文字数
要約１	369.5	22.3	472	275
重要文	504.2	30.5	766(16)	239(9)
要約２	325.3	67.0	496	172

すべての対象で要約１よりも要約２の方が要約率が高いのは、できるだけ表現を使うという制約と重要文として情報を一度絞っているためである。また、重要文を元にした要約では、もとの文字数より多くなった被験者もいた。

3 要約と原文との対応付け

要約を分析するために、要約を構成する語句が原文のどの個所からとられたものなのかを得る。佐久間 [1]、邑本 [3] は 20〜40 人の要約結果の対応付けを人手で行なっているが、これを効率的に得られるように対応付けツールを作成した。ツールの精度の確認と、実際の分析のために人手による修正も行なった。

3.1 要約と原文の対応付けツール

要約と原文との対応については、加藤 [5] が要約知識の自動獲得を目的に単語の部分一致を考慮したＤＰマッチングを、Jing[6] が、HMMによって対応付けを行なう方法を提案している。

以下のツールでは対応付けの単位や対象品詞を使用者が設定できるようにした。

対応付けの単位、入力　文、文節の単位での対応付けが行なえる。今回は文と文節の両方を行ない、最終的に文節単位での対応付けを出力した。原文、要約ともに形態素解析 [9] と文節生成 [10] を行ない、単語、文節単位に分割したものを入力とした。文節ごとにタグ (文番号-文節番号) を設定する。

入力例

```
1-1
米          米          名詞
大リーグ    大リーグ    名詞
で          で          助詞

1-2
日本        日本        名詞
から        から        助詞

1-3
単身        単身        名詞

1-4
乗り込んだ  乗り込む    動詞
```

対応のスコア　最下位の単位を単語とし、対応元の対応先に対するスコアを、

$$\frac{\text{対応元単語と対応先単語の重複文字数}}{\text{対応元単語の文字列長}}$$

とする。文、文節のスコアは、下位単位のスコアの総計を下位単位数[3] で除したもので計算される。文字列の一致については部分一致、完全一致が設定できる。

[3] 後述の対象品詞の限定がある場合は対象となった単位数

対象品詞　語と語の一致判定に使用されるカテゴリとして品詞、基本形、表記を選択できる。今回は品詞と基本形を使用した。対応付けの対象とする品詞や語、また非対象の品詞、語を指定できる。品詞は細分類のレベルを設定できる。今回は自立語の品詞の大分類を対象品詞とした[4]。

対応付けのステップ　対訳データの対応付けとは異なり、要約と原文との対応が直線関係内におさまるとは限らないため、以下のようなステップで対応付けを行なった。対応付けする範囲(窓)の大きさと対応付けする単位を変え、文と文節の2回の対応付けを行なう。今回の対応付けでは、要約を対応元とし、原文を対応先として行なった。

1. 範囲(窓)を要約、原文の全範囲とし、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、文内を範囲として各文節ごとに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。

対応付け結果　結果は以下のように、(対応元タグ、対応先タグ、スコア)の形式で出力される。分析にはこれに要約・原文の表記や品詞の情報を加えたものを使用した。

出力例

1-1	2-3	1.00
1-3	2-4	1.00
1-4	2-5	0.8

3.2　人手での対応付け作業

前述のツールによって対応付けしたデータに対して人手の修正を行なった。要約の各語句が原文のどの語句に対応するかについて、すべてが明らかに判断できるわけではなく、判定の難しいものもある。以下の手順で対応付けを行ない、対応の仕方が特殊なものについてコメントを付与した。対応付けの基準と対応の種類(ラベル)は表1である[5]。

- 表記が同じあるいは類似で、文脈(前後の語句)から判断して対応する
- 表記が類似でなくても、意味的に類似かつ前後の文脈から判断して対応する
- 対応付けの精度に関わる現象について以下のラベルを付与する。

一つの対応について複数のラベル付与を許している。

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
野茂	要約1	80.4	73.1	76.6
	要約2	99.5	88.5	91.4
野球	要約1	63.2	58.6	60.7
	要約2	88.0	89.0	88.5
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。ツールでは検出できない対応である。特に要約1の野球の記事は、複数の表現をまとめた対応が多く、ツールでの検出が他対象よりも低くなった。その他には、語句の入れ替え、1文節脱落を考慮しなかったこと、形態素解析の誤り、最初の文対文の対応付けで落ちてしまったものが考えられる。

4　要約と原文(重要文)の対応付けに基づく分析

以降では、対応結果に基づき、

- 重要文と要約の原文情報のカバーの違い

についてまず述べ、

次に2種類の要約を、原文の姿(情報、構成)をどの程度保持しているかという観点から

[4] 動詞と解析された「する」を除外するために、非対象語に動詞「する」等を加えた

[5] 「不明」は以降の対応結果からは除外

- 要約で使用された表現がどの程度原文から得られた(対応付けられた)か。

  原文と対応付けられた要約表現と、原文表現との違いを対応の種類によってみる。

- 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

について述べる。最後に文章の結束性に寄与する指示詞や接続詞の使用傾向について述べる。

4.1 重要文の要約対応個所のカバー率

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、盛り込まれる情報に違いがあると予想される。同一被験者が作成した自由な要約(要約1)と重要文とを比較することで、被験者が選んだ重要文が、同一被験者が作成した要約1の原文の引用個所を、どの程度カバーしているかを調査した。
カバー率は、被験者ごとに以下で算出した。

$$\frac{\text{要約1の原文対応個所を含む文と重要文の一致文数}}{\text{要約1の原文対応個所を含む文の数}}$$

対象	平均カバー率	標準偏差	最大カバー率	最小カバー率
野茂	55.1	16.1	100	20
野球	50.8	15.1	90	5.6
安室	54.5	16.1	100	20

表3：要約1の原文対応個所に対する重要文のカバー率

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、個人ごとのカバー率をみると90〜100%の被験者もおり、被験者によっては重要文で抽出した個所と要約で引用した個所が同じになるケースもある。被験者の多くは、重要文で抽出した以外の個所からも広く情報を拾い集めて要約を作成している。

表4は杜説(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4：文選択しなかった被験者のうち要約1での引用した数

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の中では引用した。

この重要文と要約1の違いから、自由作成の要約と、重要文や重要文を経た要約とでは異なる性質の情報になると想像できる。多くの人が広く情報を集めたものを自由作成の要約として提出していることから、重要文や重要文を経た要約は、文書の代替物としての(imformative な)機能よりも情報の取捨選択(indicative)としての使用や、重要な事柄をごく少ない量で表現する際の使用が望ましいように思える。

4.2 要約と原文との対応

要約で使用された表現がどの程度原文から得られた(対応付けられた)かを対応率として述べる。また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。

4.2.1 要約の各文節の原文との対応率

要約の各文節の対応率を表5に示す。

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

表5：要約の原文との対応率

自由作成の要約1でも9割近く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、対応の中には作業者ごとのゆれの大きいもの、判断の難しいものがある。前述の人手作業時のラベルの中の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

前述では要約の各文節がどの程度原文の情報と対応付けられるかをみたが、どのような対応をしているかによって、使われた要約手法の難しさが推測できるのではないかと考える。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6は、対応の種類ごとに、対応の種類(ラベル)の出現数で除したものを示す。「複数(別)」は複数対応でかつ別表現になっているもの、「複数(同)」は複数対応でかつ表現は同じもの、「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。

記事	タイプ	ラベル総数	複数 (別)	複数 (同)
野茂	要約１	8385	1.45	1.05
	要約２	6682	0.03	0.52
野球	要約１	14716	1.26	0.32
	要約２	10869	0.44	0.15
安室	要約１	8975	1.46	0.89
	要約２	7034	0.09	0.14

別表現１	別表現２	表記	語単位	ラベルなし
8.57	6.32	1.51	9.10	71.99
2.35	2.45	0.82	3.07	90.75
13.09	4.87	1.25	14.62	64.60
3.13	2.32	0.70	3.39	89.88
13.31	4.60	1.91	11.47	66.36
2.87	1.69	1.22	2.36	91.63

表６：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、対応の種類をみると差がある。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割を占めるが、要約１では65～71%であった。

できるだけ表現を使い、一度情報を絞っている要約２を基準に、自由作成の要約 (要約１) で、各手法 (種類) がどのくらい増えるのかが推測できる。どの種類も要約１の方が多くなっている。自由な要約では、技術的に難易度の高い手法 (複数のまとめ、別表現等) やそれ以外の手法とも、制限を与えた要約に比べ、多く使われる可能性がある。特に要約１は複数表現をまとめる手法が使われているが、制約を与えた要約２では、複数表現のまとめはほとんど使われていない。また、別表現 (文字列非類似、文字列類似) は要約１では２～４倍程度増えていた。これらは重要文をつなぎあわせた要約と、自由作成の要約を分かつ特徴的な現象にあたると考えられる。

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と、対応する原文での文節位置とが、直線関係 ($y = ax$) 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。x 軸に原文の先頭からの文節位置、y 軸に要約の先頭からの文節位置をとり、散布図と回帰式で確認し、調べた。原文の文節の出現順が保持されていれば、この直線関係にのることが予想される。

以下は相関係数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。

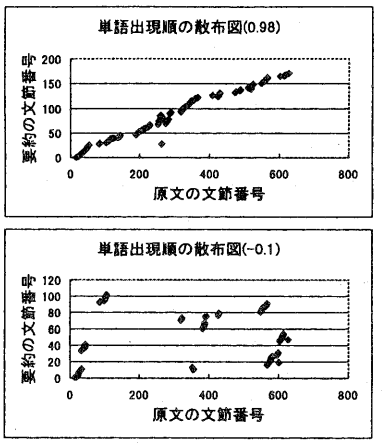

図1: 単語の出現順の散布図例

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７：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。自由作成の要約１でもある程度原文の表現を保持した形で要約を作成していた。しかし、0.5 未満の直線関係にない、語句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。

要約２は「表現をできるだけ使う」という指示での作業だが、表現を利用することから自然と表現の順番も原文に大きく影響を受けたようだ。

単語の順序保持が被験者個人ごとの傾向なのかどうかを、各被験者の対象間での相関係数の差をみると、0.1 未満のものが 57 人であった。半数以上が 0.1 未満におさまっており、語順の傾向は被験者による面が大きい可能性がある。

4.4 原文との対応率と文節の出現順の傾向

前述の文節の出現順傾向は対応付けできる文節に関する位置関係に限られるので、対応率の低いデータでも直線関係にのる場合は高い値になる。

ゆえに、対応率と単語の出現順の２観点を合わせてみることが必要である。図１ (本稿の最後に掲載) に、対応率と単語の出現順の直線関係の値を対象、要約タイプごとの各被験者について散布図にとったものを示す。

７割近くの被験者は右上の相関係数１、対応率100%の近くに位置している。要約１では相関係数１、対応率100%の個所から離れたところに位置する被験者もいるが、要約２では、これらの離れたところに位置する被験者の数が減り、相関係数１, 対応率100の個所に集中している。

但し、ここでは対応の種類を加味していないので、対応の種類 (手法としての難しさ) を尺度に加えることで各被験者のとった手法のスタイルをより明らかにできる。今後、それらを加える必要がある。

4.5 指示詞、接続詞の出現

文間の結束性やつながりを保持、表現するために使用される指示詞と接続詞の出現を調べた。

4.5.1 指示詞の出現

馬場 [8] は要約中に出現する指示詞を原文から保持されたもの (残存,a)、新たに導入されたもの (新規,b)、要約では削除されたもの (c) に分けて分析を行なった。ここでは、対応結果から要約中の指示詞の残存 (a) と新規導入 (b) を調べた。

	要約中総数	残存 (a)	新規 (b)
野茂・要約１	285	170(59.6)	115(40.4)
野茂・要約２	248	211(85.1)	37(14.9)
野球・要約１	278	100(36.0)	178(64.0)
野球・要約２	168	88(52.4)	80(47.6)
安室・要約１	288	144(50.0)	144(50.0)
安室・要約２	246	179(72.8)	67(27.2)

表８： 要約中の指示詞の種類

新規に導入された指示詞は、要約１の方が要約２より多く観測されたが、有意差には至らなかった。

要約１に新出の指示詞が多く見られたが、その出現個所をみると、主張を裏付ける事実の列挙など複数の関連する項目をまとめる個所の直後で「このように」「そうした」「それら」等の語で受ける現象や情報を追加・詳細化する際の使用が目立った。要約１では、原文の多くの材料をまとめ、つなげ、再構成するために使用されている場合が多い。

それに対して、要約２では、多くの材料からまとめ、つなげる作業が少ないため、こうした新しい指示詞があまり使用されなかったと思われる。要約２で残存した原文の指示詞は、その指示詞を含む文自体が重要で、その照応先の情報が重要文として選ばれている場合であった。

要約１、２とも、前述の指示詞以外に、「この野茂投手が」「あの個性的な」といった指示詞自体に明確な照応先はなく (あえて求めるとすれば文脈の外の共有知識)、強調的意味合いもある用法の指示詞がそのまま残る場合があった。原文で表現された主張のうち、筆者の評価や思い入れに関する部分も要約時に保存される場合があるのがわかった。

4.5.2 接続詞の出現

要約中で、原文の接続詞が残存したもの (a) と新たに導入されたもの (b) を調べた。

	要約中総数	残存 (a)	新規 (b)
野茂・要約１	121	45(37.2)	76(62.8)
野茂・要約２	55	23(41.8)	32(58.2)
野球・要約１	189	94(49.7)	95(50.3)
野球・要約２	161	105(65.2)	56(34.8)
安室・要約１	128	46(35.9)	82(64.1)
安室・要約２	118	86(72.9)	32(27.1)

表９：要約中の接続詞の種類

要約で新たに導入された接続詞が要約１の方で多くみられた。情報を拾い集め、再構成していることが予想される。

原文から残存した接続詞をみると、要約１では原文の文章構造の枠組みに寄与する「しかし」「だが」などの接続詞が多く現われたが、要約２ではこれらが少なかった。例えば、以下の段落先頭の「しかし」を要約１では25人の被験者がそのまま使用しているのに対して、要約２では５人であった。

> とりわけ若い世代には「勇気」さえ与えてくれたのではないか。そんな思いを込めて「野茂投手よ、ありがとう」とも、私たちは言いたい。
>
> しかし、同じスポーツでは２カ月前のアトランタ五輪において、日本選手が不振、一部では惨敗と酷評された。「楽しんで競技をした」という若手選手の言葉に対する批判も噴出した。(以下略)

この「しかし」は３段落を逆接的につなぎ、論理展開を明示しているが、文単位でみたときには「しかし」を含む文自体は他の文に比べ重要度は低く判定されていた (100人中７人が重要文として選択)。

そのため、重要文抽出の時点でこの文を選択しなかった被験者はこの段落内の逆接的な内容の文を選択していても、論理展開の明確な逆接の接続詞を用いない、あるいはこの逆接の内容自体をそっくり落とす場合があり、要約２の方が全体として論理展開がシンプルになる傾向があった。これらは、文という単位が論理展

開の単位とは必ずしも合わず、重要文抽出の時点で落ちたことに起因する。しかし、自由作成の要約では接続詞による明示的な論理展開も保存されていることから要約作成という点では内容だけでなく、論理展開を考慮し、それらを補足する手だてが必要ではないかと感じる。

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た。要約と原文の対応付けに基づき以下の分析を行なった。

- 対応率、単語の出現順の傾向から原文の姿をどのくらい保持して要約作業が行なわれたのかを推察した。自由な要約でも、語句の88%程度が原文と何らかの対応を付けることができ68%が原文の表現の順序を保持していた。但し、対応付けには複数表現のまとめなどの対応付け判断が難しく、要約手法としても高度なものがあり (11%程度)、対応率や対応の種類にこれらを考慮することが必要である。

- 自由な要約ではより広い範囲から情報をまとめ、つなぎあわせ、接続詞、指示詞を新たに使い文章を再構成していた。原文の情報や論理構造を保存する傾向にあった。一方、重要文を経た要約は重要文として情報を絞ったこと、論理展開部分が落ちたことで、情報、論理展開ともシンプルになった。

今回は、原文の構成、情報の保持について概観したが、対応の種類ごとの言語現象や対象の論理構造を考慮した分析が必要である。また、新聞記事の異なる紙面での記事の違いまでは分析できなかった。今後は対応結果の見直しなど行ない、これらを進めていく。

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] アンドレイ・ベケシュ1989 「残存認定単位の規定と出現傾向」,「文章構造と要約文の諸相」くろしお出版

[3] 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

[4] 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

[5] 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-」

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

[7] 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

[8] 馬場俊臣 1989 「要約文の指示語使用の特徴」,「文章構造と要約文の諸相」くろしお出版

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

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

[11] 毎日新聞社 1998 「CD-毎日新聞９６年度版」

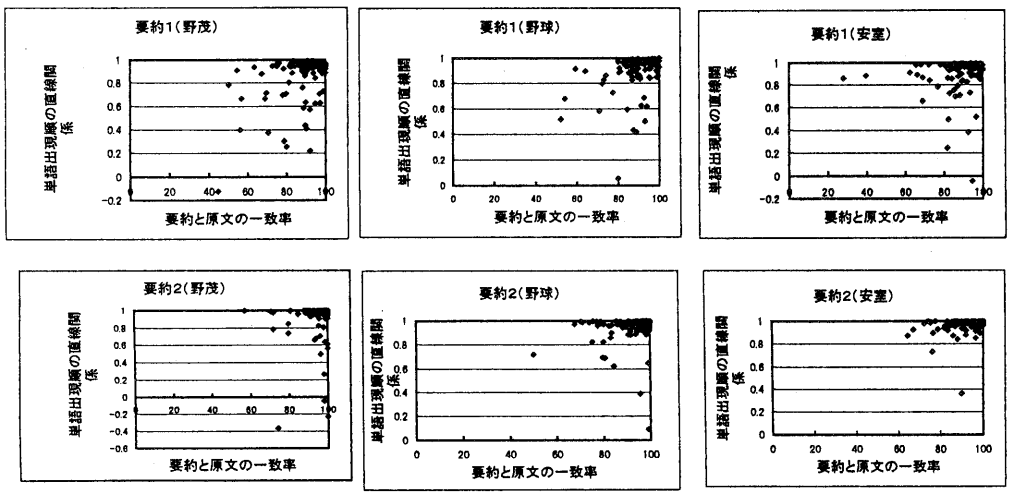

図２：要約タイプ、対象ごとの全被験者の一致率と単語出現順の傾向