

書き換えによる抄録の読みやすさの向上

難波 英嗣, 奥村 学

Email: {nanba,oku}@jaist.ac.jp

北陸先端科学技術大学院大学 情報科学研究科

[概要]

これまで、抄録を自動的に作成する様々な手法が提案されてきた [9]。その多くは、テキスト中の文を1つの単位とし、それらに何らかの情報を付与し、その重要度で順序付け、重要な文を選択し、それらを寄せ集めることで抄録を作成する。しかし、こうして作成された抄録の問題点として、Paice[6]は、1) 抽出した文中に代名詞など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、2) テキスト中の色々な箇所から抽出したものを単に集めているため、抽出した複数の文間のつながり(首尾一貫性)が悪いことなどを指摘している。

このような読みにくさを解消する1つの方法に抄録の書き換えがある。抄録中の欠落情報を補ったり不要な箇所を削除するといった書き換えを行うことで文間のつながりが良くなり、抄録の読みにくさが解消されと考えられる。本研究では、まず心理実験で数人の被験者に抄録を読みやすく書き換えてもらい、抄録の読みにくさの要因を5種類に分析した。また読みにくさを解消する方法を検討した。次に分析結果に基づき抄録を読みやすく書き換えるシステムを計算機上に実装した。

[キーワード] 抄録, 読みやすさ, 書き換え, 結束性

Producing More Readable Extracts by Revising Them

NANBA Hidetsugu, OKUMURA Manabu

School of Information Science, Japan Advanced Institute of Science and Technology

(Tatsunokuchi Ishikawa 923-1292 Japan)

Abstract

Several techniques have been proposed for summarizing a text[9]. Each sentence in a text is considered as a unit for extraction, and summarization system calculates the importance of each sentence using several extract methods. Then several sentences are extracted according to their importance. However, Paice[6] points out the problem of the coherence of extracts as follows: 1)If anaphors are included in the extracts, their antecedents are not always included in the extracts. 2)Extracts are incoherent, because they are constructed from several sentences(phrases) which are partially extracted from texts.

In order to make the extracts more readable, first, we have to clear the reasons of incoherence of the extracts. We let some subjects revise the extracts. Then, we classify the reasons of incoherence into 5 categories. Furthermore, we make rules for revising the extracts, and evaluate them by comparing the extracts revised by rules with those by subjects. Then, we find that our rules are effective to revise readability of the extracts

Key Words extract, readability, revision, cohesion

1 はじめに

これまで、抄録を自動的に作成する様々な手法が提案されてきた [9]。その多くは、テキスト中の文を1つの単位とし、それらに何らかの情報を付与し、その重要度で順序付け、重要な文を選択し、それらを寄せ集めることで抄録を作成する。しかし、こうして作成された抄録の問題点として、Paice[6]は、1) 抽出した文中に代名詞など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、2) テキスト中の色々な箇所から抽出したものを単に集めているため、抽出した複数の文間のつながり(首尾一貫性)が悪いことなどを指摘している。

このような読みにくさを解消する1つの方法に抄録の書き換えがある。抄録中の欠落情報を補ったり不要な箇所を削除するといった書き換えを行うことで文間のつながりが良くなり、抄録の読みにくさが解消されると考えられる。

本研究では、まず心理実験で数人の被験者に抄録を読みやすく書き換えてもらい、抄録の読みにくさの要因を分析する。また要因毎に読みにくさを解消する方法を検討する。次に分析結果に基づき抄録を読みやすく書き換えるシステムを計算機上に実装し、システムの評価を行う。

本稿の構成は以下の通りである。2節では関連研究について述べる。3節では、抄録の読みにくさの要因について述べる。4節では、3節で述べた調査結果を基にした抄録書き換えシステムの実現方法について説明する。5節では、4節で述べたシステムの評価を行い、結果について考察する。

2 関連研究

本節では、最初に一般的なテキストの読み易さについて説明し、次に抄録の読み易さとの違いを述べる。また、抄録の書き換えの関連研究についても述べる。一般にテキストの読み易さには次の2種類がある [8]。

- **comprehensibility**(わかりやすさ)
表記の正しさ、語彙の難易、文構造の複雑さなどによる文章の理解
- **legibility**(見やすさ)
活字のフォント・スタイル、行間のバランス、レイアウトなどによる見易さ

legibilityの読みやすさは、主に印刷・出版業界で調査・研究されており、その成果は印刷・製本技術に活用されている。これに対し comprehensibilityの読み易さは、言語学・自然言語処理に関するものである。本研究では comprehensibilityの読み易さを扱う。

comprehensibilityの読み易さに関して、これまで多くの研究がされてきたが [2]、その多くは、人が正しく書いたテキストを入力の対象とし、テキストの読み易さ(難易度)を、語彙レベル(語彙の難易)、構文レベル(文構造の複雑さ)という側面から捉え、数値化するのが中心的であった。一方、抄録は計算機が自動的に作り出したテキストであり、他の観点も考慮して読み易さを考察する必要がある。

そのような観点として、文間のつながりの良さ(文脈レベル)があげられる。Hallidayら [1]、森岡 [15]は、テキスト中の文間のつながりを表す指標をそれぞれ示している。

Hallidayらはテキスト中の文間のつながりを表す指標として、5種類の結束性(cohesion)、すなわち、指示(reference)、代用(substitution¹)、省略(ellipsis)、接続(conjunction)、語彙的結束性(lexical cohesion)をあげている。

森岡は、文間に緊密なつながりのある文章を書くための指標として、つながりを良くする要因を次の5種類に分類している。すなわち、材料の配列²、代名詞の使用、接続語の使用、繰り返し、その他(内容の緊密な関連³)である。

文間のつながりの良さは、抄録の読み易さの重要な要因であり、文間のつながりが悪いと、抄録として読みづらくなると考えられる。Minel[5]は抄録の文間のつながりの良さ(悪さ)を評価をする指標として、1) 先行詞の存在しない照応詞の数を数える、2) 抄録中のある文とつながりの強い原文中の文がどれだけ抄録に含まれているか、という2点をあげている。

抄録の文間のつながりを良くするために、Paice[6]

¹たとえば、'a new one'における'one'、'do so'における'do'などを用いた照応

²材料とは、主題を的確に伝えるために、主題を支え裏打ちするものである。論文における資料(data)、小説における素材(material)に相当する。森岡は材料の選択と、テキスト中でのその配列のしかたが、つながりの良さの要因の1つであると考えている。

³例えば「…それごとコタツに入れて、そして寝た。翌朝は起きると直ぐ彼は…」といったように時間の経過に従って文を書くことと接続の語がなくなるともつながりが保たれる

は、照応詞を含む文の前の数文を要約に追加したり、接続詞は削除したり、動詞の時制や態は調和がとれた流れにしたりすることで、部分的な解決が実現できることを示している。

また、山本ら [16] は、重要文として抽出される文の先頭に指示語や接続詞が出現する場合、あるいは、抽出される文の主語が省略されている場合、単独で要約中に存在すると不自然であると考え、前文も要約に追加することで対処している。

一方、これらの手法では、抄録の読み易さを向上させるために、非重要文を抄録に含めてしまうという点で問題がある。本研究では、抄録に非重要文をそのまま含めるのではなく、書き換えにより抄録を読み易くする。

また、これまでの研究では、抄録の読みにくさに対して体系的な分析がなされてるわけではない。本研究では数名の被験者に抄録を書き換えてもらい、抄録の読みにくさの要因を体系的に分析する。

抄録の書き換えに関して、いくつかの研究や調査報告がある [10][3]。抄録の書き換えの目的には、次の3種類があると考えられる。

1. 文の長さを縮める
2. 文のスタイルを変える
3. 抄録を読みやすくする

川原 [10] は、人が要約を作成する時、原文をどのように書き換え (パラフレーズ) しているかについて調査している。その際、言語サイズ⁴という尺度を用い、原文一要約間の書き換えの言語現象を分析している。川原は、1~3すべてを分析対象にしているが、本研究では3に着目し、書き換えパターンの分類を行う。

Maniら [3] は1と3を対象とし、自動的に作成された抄録を書き換えるシステムを構築している。Maniらは書き換えに3種類のオペレータを用意し、それらを抄録に繰り返し適用することで、元の抄録の読み易さを低下させずに要約の informativeness を向上させている。Maniらが informativeness に重点を置いた書き換えを目指しているのに対し、本研究では書き換えによる読み易さの向上を目指す。

本研究では、数名の被験者に抄録を読み易く書き換えてもらい、抄録の読みにくさの要因を分析した。ま

⁴言語サイズとは、語、句、節などの統語的な単位を指す。

た要因毎の書き換え手法について検討した。次節では、それらについて報告する。

3 抄録の読みにくさの要因

3.1 心理実験による読みにくさの要因の調査

抄録の読みにくさの要因を調査するため、心理実験を行った。調査期間は1997年11月~12月、被験者として大学院生12名(1テキストあたり6名の被験者)、対象テキストは日本経済新聞25記事(16文~62文、平均30文/記事)を用いた。実験手順を以下(I,II)に示す。また、実験の際のインストラクションを図1に示す。

- I. 記事中の各文の重要度を3段階(A,B,C)評価。
- II. 評価Aの文(以下、重要文と呼ぶ)のみを用いて要約を作成⁵してもらう。

重要文だけを用いてわかりやすく書き直して要約を作成して下さい。わかりやすく書き直す方法として以下に例示しますが、これ以外の方法を用いても構いません。

- 重要と思われる文中の先行詞が単語である場合は置換する。
- 重要と思われる文中の先行詞がある段落全体にわたる場合、できるだけ短い言葉で置き換える。
- 接続詞を用いてなるべく読みやすく変える。
- 短い文は接続詞等を用いて他の文と結合する。
- 文末は読みやすいように適宜変える。

図1: 心理実験のインストラクション

実験の結果、被験者により全体で343箇所の書き換えが行われた。

3.2 調査結果の分析

343箇所には先に示した抄録修正の目的1~3が含まれていたが、このうち目的3の修正に着目し読みにくさの要因を分類した。その際、川原の研究報告 [10] や Hallidayら [1] の結束性の分類、森岡 [15] の分類を参考にした。また、これらの読みにくさの解消方法と、それに必要な自然言語処理技術についても検討し、表1にまとめた。

⁵修正には以下の3種類のタグを用いる。

- `< add > W1 < /add >`:重要文にW₁を加える。
- `< del > W2 < /del >`:重要文中のW₂を削除する。
- `< para W4 > W3 < /para >`:重要文中のW₃をW₄に置換する。

表 1: 読みにくさの要因と書き換え処理の方法

分類	読みにくさの要因	書き換え処理	必要な技術
ア.	不要な表現/接続詞の不足による文間のつながりの悪さ	接続詞的表現の [削除/付加]	談話構造解析
イ.	構文的な複雑さ	2つの文をつなげる/単文の分割	
ウ.	冗長・不自然な繰り返し	代名詞化/主語の省略/指示詞の付加	
エ.	情報の欠落	主語の補完/照応補完/照応詞の削除 代用の補完/補足情報の付加	照応・省略補完 情報抽出
オ.	不要な副助詞/副助詞の不足による文間のつながりの悪さ	副助詞の [削除/追加]	

表 1に示す 5 種類の読みにくさの要因のうち、文間のつながりの悪さに関するものはア、ウ、エ、オで、イは構文レベルの読みにくさである。

以下に 5 種類の要因と書き換え手法を例を挙げて説明する。囲み中の例文には各々原文中での位置 (先頭から何文目か) が記してある。また網掛け (■) してある文は評価が B または C の文 (以下、非重要文と呼ぶ) である。

ア. 不要な表現/接続詞の不足による文間のつながりの悪さ

2. 業績が落ち込んでいる中で生産財から消費財へのシフトを進める同社にとって、全体の四割を占める四百人の女子社員の活性化が不可欠。
3. 専任社員を置くことで女性の職域を積極的に拡大していく方針だ。
15. ■ 会社の期待は大きい。このような女性参加のプロジェクトの前例がなく、現在は悪戦苦闘中。
16. しかし目標を二〇〇一年に置いて女性社員にとって満足できる仕事環境、会社にとって有効な人材の活用と両者に利益のある改革を進めている。

文 16 の文頭にある接続詞「しかし」は、原文中の前文 15 にかかる。しかし、文 15 は非重要文であり、文 16 の「しかし」は不要表現であるため、削除した方がよい。この書き換えを計算機上で実現するには、談話構造解析の技術が必要とされる。

これとは逆に抄録中で接続詞的表現が不足しているために、文間のつながりが悪く、読みにくくなる場合も考えられる。その場合、新たに接続詞を加える必要がある。

イ. 構文的な複雑さ

2. (修正前)
榎尾和雄社長自ら「将来の経営の核」に挙げる通信事業の第一弾であり、今後、携帯電話などに手を広げていくための布石でもある。
- ↓
2. (修正後)
榎尾和雄社長自ら「将来の経営の核」に挙げる通信事業の第一弾である。
今後、携帯電話などに手を広げていくための布石でもある。

長い文は複雑な構文になる傾向にあり、それが文を読みにくくする要因のひとつと考えられる [2]。上の例の場合は、長い文を 2 つの単文に分割している。

また、逆に、関連ある短い 2 文はまとめて 1 文にした方がよいという指摘もある [4]。Mathis らは並列関係にある短い 2 文は関係代名詞等でまとめて 1 文にした方が読みやすいと主張している。今回の心理実験結果にも、2 文をまとめて 1 文にする書き換えがあったが、それらは Mathis らの主張するような並列関係の 2 文をまとめる書き換えは含まれていなかった。実験結果では順接や逆接の関係にある 2 文を「ので」や「が」といった接続助詞等で結びつけるものであった。しかし、これらが読み易くするために書き換えているのか、被験者が好む文体に書き換えているかははっきりしていない部分がある。

ウ. 冗長・不自然な繰り返し

同一主語が繰り返し出現すると冗長であり、後者の主語を省略したり [3] 代名詞や代用 (substitution) 表現を用いることで、冗長性を解消できると考えられる。

- 越後製菓が四月に始めた新商品シリーズ「越後ベスト○○」が主婦に人気だ。
- <para 同社>越後製菓</para>は米菓分野で八七年からNTTのキャプテンなどを活用、スーパーなど小売店の売り場から工場に直接注文を受け即日製造し出荷、翌日には店頭と並べるといふ、いわば「ジャストインタイム」方式で売り上げを伸ばしてきた。

上の例では、越後製菓という語が文2と文4に含まれているが、後者が冗長であると考えられ、同社という代用表現に置き換えている。このような処理を計算機上で実現するには、抄録中の固有名詞抽出技術、およびそれらの属性を決定するための知識が必要になる。

また、冗長と考えられる表現を代名詞化すべきか、省略すべきか、代用するのが良いか、状況によって変わってくる。上の例の場合、例に示すような代用表現への置換の他に「越後製菓は」を省略することも可能であるが、「<add>その</add>越後製菓は」と指示詞を付加するのは不自然である。従って、書き換え処理の適用条件も検討する必要がある。

エ. 情報の欠落 (照応・省略・代用等)

- 消費者のブランド離れ、価格への敏感さで日本に先行している米国で、「デフレ型経営」によって成功している企業がある。
- 自動車のクライスラーとパソコンのコンパクト・コンピューターだ。
- 企業の人員削減が消費意欲を沈滞させ、それがまた企業業績の足を引っ張るといふデフレ型の悪循環に陥りつつある。
- その中で、クライスラーが大健闘しているのは、新製品投入の時期と価格・商品戦略がまさしく低成長時代に適応しているからだ。

文9の「その」の先行詞は文8になるが、文8は非重要文である。このような場合、読みやすくする方法に先行詞を補うことが考えられるが、上の例のように照応詞そのものを削除する書き換えも可能である。このために、照応解析の技術が不可欠であると考えられる。また、この例の場合、照応詞「その」だけを削ると文として不自然なため、「その中で、」という照応詞を含む文節を削除する必要がある。

オ. 不要な副助詞/副助詞の不足による文間のつながりの悪さ

- 今回のド・ムオイ書記長の訪日は両国が相互理解を深めるうえで、よい機会である。
- ベトナムはアジアの安全保障の観点からも今後、重要性を増していく。
- 経済協力の面でも長い目で発展を支援していく姿勢が必要だろう。

この例のように、原文中では並列関係にあるような2文が、抄録中ではその一方しか含まれない場合、副助詞「も」は不要な表現であると考えられる。

逆に、抄録中で並列関係にある2文は、副助詞「も」等を補った方が読みやすくなる場合がある。

4 抄録書き換えシステムの構築

4.1 書き換えシステム

これまで、書き換えに基づく要約生成システムや[7]、抄録修正システム[3]が開発されているが、本研究でも、書き換えにより抄録を読み易く改良する。システムの構造を図2に示す。システムの入力は、テキスト原文と抄録をそれぞれJUMAN[11]により形態素解析し、さらにKNP[12]で文節区切りを入れたものである。テキスト原文を参照しながら、繰り返しルールを適用し抄録を書き換える。どのルールを用いても抄録が書き換えられなくなれば、それを最終的に出力する。

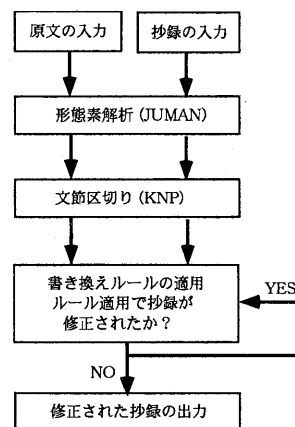


図 2: 抄録書き換えシステム

4.2 抄録書き換えルール

表1のア, ウ, エについて部分的に jperl で実装した。イについては, 読み易くするための書き換えと文体を変えるための書き換えの境界が曖昧なため, 今回は実装していない。オについては, 抄録中で2文が並列関係であるかを判断するには格構造解析を行い, 2文の格要素を比較するといった処理が必要であると考えられるが, 実装する上で考慮すべき点が多く, 実装には至っていない。

また, 今回, 実装する際に UNIX 上で動く談話構造解析ツールが入手出来なかったため, 福本ら [14] の開発したツールのソースを解読し, 部分的に利用した⁶。

ア. 不要な接続詞の削除

アルゴリズムを図3に示す。ルールは, 抄録中の文(以下, 文A)に接続詞が含まれていて, 原文中の前1文(以下, 文B)が抄録に含まれない時, 接続詞を削除する(図3,12)。ここで, 文Bの前文に「例えば」や「ただし」といった接続詞的表現が含まれる場合, 主文はさらに1文前(以下, 文C)となるため(14), 文Cが抄録に含まれているか調べる必要がある(15)。もし, 文Cが抄録に含まれていれば接続詞を削除する必要はないが(18), そうでなければ接続詞を削除する必要がある(16)。また, 文Bが1文前の従文になっていなければ, 文A中の接続詞は不要であり, 削除する必要がある(19)。

ウ. 冗長な主題表現の省略

抄録中の連続する2文の主題表現が同じ場合, 後の文の主題表現を省略する。主題表現とは「名詞+格助詞が」, 「名詞+係助詞は」の2つを考えている。ここで, いくつかの事例を見た結果, 抄録中の隣接する2文の主題表現の名詞が同じ場合, どちらも「名詞+係助詞は」であれば, 後の主題表現を省略しても不自然ではないが, どちらか片方が「名詞+格助詞が」の場合は却って読みづらくなる場合があった。従って, 今回は抄録中の連続する2文の主題表現が「名詞+係助詞は」の時のみ, 後の主題表現を削除した。

エ. 照応詞の削除

この処理は, 照応解析技術が必要とされる。しかし, 実用的な照応解析ツールがないので, 今回は文頭が照

```
# 初期設定
1: open(@ORIGINAL); # 原文の入力 (1行1文)
2: open(@SUMMARY); # 要約の入力 (非重要文は空行)
3: @pre=('例えば','具体的に'...); # 前文が主文の接続詞
4: @flat=('しかし','だが'...); # 主従関係のない接続詞
5: @ref=('こうした','この'...); # 指示詞
6: if("接続詞の削除"){@flat_ref=@flat;}
7: if("照応詞の削除"){@flat_ref=@ref;}
# 書き換えルール $iは原文の文数
8: for($j=1; $j<=$i; $j++){
9:   foreach $conj (@flat_ref){
# @flatの接続詞が含まれる文Aの前の文Bが抄録に
# 含まれていない時, 書き換えルールを適用
12:    if($SUMMARY[$j]~/ $conj/
& $SUMMARY[$j-1]==NULL){
# 原文中の文Aの一文前(文C)が文Bの主文か?
13:      foreach $conj2 (@pre){
14:        if($ORIGINAL[$j-1]~/ $conj2/){
# 文Cが非重要文なら文Aの接続詞は不要
15:          if($SUMMARY[$j-2]==NULL){
16:            "文Aの接続詞の削除";goto AAA;
17:          }
# 文Cが重要文なら文Aの接続は必要
18:          else{ goto AAA; }
}
# 文Bに主文がない場合, 文Aの接続詞は不要
19:      "文Aの接続詞の削除";goto AAA;
}
}
20: AAA:
}
```

図3: [接続詞削除/照応詞削除]のアルゴリズム

応詞である場合について, ヒューリスティックスにより処理した。それは, 文頭が照応詞の場合, 前文に先行詞があるというものである。これにより, 「不要な照応詞の削除」と同様のアルゴリズム(図3)で照応詞の削除が可能になる。照応詞削除の際, 照応詞が含まれる文節を削除した。

5 評価

5.1 評価用データの作成

抄録を読み易く書き換えることは非常に主観的な作業である。本来は書き換えることで, 本当に読み易さが向上しているかを被験者に主観的に判断してもらう等の評価が必要があるが, 今回は, 被験者の書き換えとシステムの書き換えの比較により評価を行った。

前節で述べた書き換え手法を評価するのに, 唯一の正解セットを作成し精度を求めめるのは現実的な評価方法であるとは言えない。しかし, 1つの同じ抄録を複数の被験者に書き換えてもらい, それらが被験者間で一致すれば, その箇所はより必要性の高い書き換えであると考えられる。被験者の書き換えとシステムの書

⁶ICOT フリーソフトウェア

<http://www.icot.or.jp/AITEC/IFS/IFS-abst/049-J.html>

書き換えの比較をする際、被験者間の一致度を考慮することにより、書き換えルールがどの程度必要不可欠な書き換えを実現できているかの評価が部分的に可能になると考えられる。

書き換えルールを評価するためのテキストとして、日本経済新聞 1993 年から 15 記事、1995 年から 16 記事、計 31 記事を用いた。1 記事につき 6~7 名の被験者が記事中の各文の重要度を 3 段階評価 (A,B,C) し、過半数 (4 名以上) の被験者が A と評価した文を抄録として用いた。1995 年の 16 記事については、野本ら [13] の要約心理実験の結果を利用して頂いた。記事の原文は平均 25.2 文、抄録は平均 5.1 文 (平均要約率 20%) である。

書き換えルールを評価するためのデータとして、1 つの抄録に対し 5 人の被験者に読みやすく書き換えてもらったものを用いる。書き換えの際のインストラクションとして、3.2 節の 5 種類の読みにくさの要因を例示し、説明した。その結果、全部で 188 回の書き換えが行われた。このうち読みやすさに関するものが 167 回あった。表 2 に書き換え処理と、処理毎の書き換えの回数を示す。書き換えとして最も多かったのは接続詞的表現の付加 (61) で、次いで補足情報の付加 (26)、照応詞削除 (15)、接続詞的表現の削除 (11)、主語の補完 (11) であった。

5.2 書き換えルールの評価

被験者間の書き換えの一致度には、次の 3 段階がある。

- L1: 書き換え位置も書き換え方法も全く同じ
- L2: 書き換え位置が同じで、書き換え手法が分類レベルで同じ
- L3: 書き換え位置が同じで、書き換え手法が分類レベルで異なる

表 2: 処理毎の被験者による書き換えの回数

	書き換え処理	回数
ア	接続詞的表現の削除 (11)/付加 (61)	72
イ	2 つの文をつなげる (2)/単文の分割 (6)	8
ウ	代名詞化 (5)/主語の省略 (3)/指示詞の付加 (8)	16
エ	主語の補完 (11)/照応補完 (8)/照応詞削除 (15)	34
オ	代用の補完 (2)/補足情報の付加 (26)	28
カ	副助詞の削除 (4)/追加 (5)	9
計		167

このうち、L1 と L2 のレベルを被験者間で書き換えが一致していると考え、ルールにより書き換えられた抄録の結果と評価用データを比較した。

ルールの評価は以下に示す 2 通りで行った。

- 評価 1: 2 人以上の被験者の書き換えが一致したものと、ルールによる書き換えの一致度による評価
- 評価 2: 1 人以上の被験者の書き換えと、ルールによる書き換えによる一致度の比較による評価

また、評価に以下の式を用いた。

$$Recall = \frac{\left(\begin{array}{l} \text{ルールで書き換えられた箇所のうち} \\ \text{被験者と一致した箇所数} \end{array} \right)}{\text{被験者が書き換えた箇所数}}$$

$$Precision = \frac{\left(\begin{array}{l} \text{ルールで書き換えられた箇所のうち} \\ \text{被験者と一致した箇所数} \end{array} \right)}{\text{ルールで書き換えられた箇所数}}$$

評価 1, 2 の結果を表 3 に示す。

表 3: ルールと被験者の書き換え箇所の一致度

書き換え	評価 1		評価 2	
	Recall	Precision	Recall	Precision
ア (のべ 11)	2/2	2/5	4/4	4/5
ウ (のべ 3)	0/0	0/0	0/1	0/1
エ (のべ 15)	4/5	4/7	4/5	4/7

表 3(ア) について、評価 1, 2 共に被験者が削除した箇所はルールでもすべて削除された (Recall=1)。評価 2 で書き換えが一致しなかった箇所は、談話構造解析の代わりに用いたヒューリスティクスに当てはまらないケースであった。

(ウ) について、評価 2 で書き換えられなかった箇所は、「固有名詞+が」の次に「固有名詞+は」が現れ、被験者は後者を削除したが、ルールでは削除されなかった。

(エ) について、ルールにより削除された箇所は 7 箇所あり、そのうち 4 箇所は評価 1, 2 共に一致していた。一致しなかったもののうち 1 箇所は、箇所自体は被験者の削除箇所と重なっていたがルールによる削除箇所が短かった⁷。残りの 2 箇所は先行詞が原文の 2 文前

⁷ 「こうした現状をふまえ、…」で始まる文で、書き換えルールは「こうした」を削除したが、被験者は「こうした現状をふまえ、」を削除した。

以内に存在するというヒューリスティクスに当てはまらないケースであった。

5.3 考察

今回作成したルールで、冗長な主語の削除がルールで1箇所削除されなかった場合を除くと、削除の箇所によらずれがあった場合も含めると、被験者が削除した箇所は、すべてルールでも削除されていた。被験者が削除せず、ルールが削除した箇所は、談話構造解析と照応解析のヒューリスティクスに当てはまらなかった場合である。従って、今後これらの解析技術が向上すれば、より被験者の書き換えと一致するルールが実現可能になると考えられる。

6 おわりに

本研究では、抄録の読みにくさについて、心理実験により調査し、その要因を5種類に分類した。またこれらのいくつかを解消するルールを作成し、被験者の書き換えとの一致度を調べることで、ルールの有効性を確認した。

今回は被験者が書き換えた箇所と書き換えルールの一貫度で読みやすさの評価を行ったが、最終的には、書き換えられた要約を人が読んで主観的に判断する必要がある。今後の課題として、(1)元の抄録、(2)ルールにより書き換えられた抄録、(3)2人以上の被験者の書き換えが一致した箇所のみ書き換えた抄録(4)被験者が書き換えた箇所をすべて書き換えた抄録の4種類について、読みやすさを比較する予定である。

参考文献

- [1] Halliday, M.A.K., Hasan, R. "Cohesion in English". Longman, 1976.
- [2] Klare, G.R. "The Measurement of Readability". Iowa State University Press, 1963.
- [3] Mani, I., Gates, B., Bloedorn, E. "Improving Summaries by Revising Them". the 37th Annual Meeting of the Association for Computational Linguistics(ACL-99), 1999.
- [4] Mathis, B., Rush, J., Young, C. "Improvement of Automatic Abstracts by the Use of Structural Analysis". 24 (2), pp.101-109, 1973.
- [5] Minel, J., Nugier, S., Gerald, P. "How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN", Intelligent Scalable Text Summarization, Proceeding of a Workshop, ACL, pp.25-30, 1997.
- [6] Paice, C.D. "Constructing Literature Abstracts by Computer: Techniques And Prospects". Information Processing & Management. 26 (1), pp. 171-186, 1990.
- [7] Robin, J., McKeown, K. "Empirically Designing and Evaluating a New Revision-based Model for Summary Generation". Artificial Intelligence 85, pp.135-179, 1996.
- [8] 乾裕子. 日本語文書可読性評価に関する研究. 平成7年度富士通株式会社委託研究報告, 1996.
- [9] 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, 6, (6), pp.1-26, 1999.
- [10] 川原祐美. "9. 要約文のパラフレーズの様相「文章構造と要約文の諸相」". くろしお出版, pp.141-167, 1989.
- [11] 黒橋禎夫, 長尾真. "日本語形態素解析システム JUMAN version 3.5 使用説明書", 1998.
- [12] 黒橋禎夫. "日本語構文解析システム KNP version 2.0 b6 使用説明書", 1998.
- [13] 野本忠司, 松本裕治. "人間の重要度判定に基づいた自動要約の試み" 情報処理学会自然言語処理研究会報告, 120-11, pp.71-76. 1997.
- [14] 福本淳一. "著者の主張に基づく日本語文章の構造化" 情報処理学会自然言語処理研究会報告, 78-15, pp.113-120. 1990.
- [15] 森岡健二. "文章構成法 -文章の診断と治療-" pp.124-132, 至文堂, 1963.
- [16] 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, 2 (1), pp.39-55, 1995.