自 然 言 語 処 理 147−4
（２００２．１．２１）

自動文節対応付け手法を用いた要約生成操作の調査

竹内和広、松本裕治
奈良先端科学技術大学院大学 情報科学研究科
〒 630-0101 奈良県 生駒市 高山町 8916-5
Tel. 0743-72-5246
E-mail: {kazuh-ta, matsu}@is.aist-nara.ac.jp

本研究では、要約文とその要約文を作成するために使用された表現を含む元文とを自動的に対応付ける手法を用いて、人間が要約文を作成する上で、要約元となった原文をどのように再構成するかを調査した。対応付けに用いた手法は、 かかり受け構造の解析結果を利用し、要約文とその対応文との間の対応付けを文節単位で行うことができる。また、要約文１文に対して、要約元文章中の複数文を対応付けすることを許して対応付けが可能である。調査した対象は、複数の作業者が新聞の社説を要約したデータである。このデータに対して、対応付け手法を実際に適用した。対応付けの結果、要約文中で文節対応付けができなかった文節が、どのように作成されたかを、計算機でも処理可能な操作を主眼に分類・整理し、考察した。その結果、要約において、要約元文章にまったく現れないような新しい表現が使われることは少なく、複数の元文から１つの要約文を作成する文結合と、単文節の言い換え操作を中心に要約生成が行われていることがわかった。

KEYWORDS: 自動要約，自動対応付け，言い換え

Sentence Reconstruction in Summary Generation: An Investigation using Automated Alignment

Kazuhiro Takeuchi and Yuji Matsumoto
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takamaya, Ikoma, Nara 630-0101, JAPAN
Tel.+81-743-72-5246
e-mail:{kazuh-ta, matsu}@is.aist-nara.ac.jp

In this paper, we investigate operations in summary generation. In order to align summary expressions with their corresponding original expressions in source text, we propose an automated alignment algorithm based on dependency structure of sentences. Our algorithm detects not only one-to-one sentence alignments, but also one-to-n sentence alignments. We apply the algorithm to human made natural summaries, and analyze the results of the alignments. We, then, find most of the summary expressions are kept their dependency structure in the original sentences and confirm one of the opration called "sentence combination," in which more than two source sentences are used to generate a summary sentence, plays an important role in summary generation. Furthermore, we characterize paraphrases in summary generation.

KEYWORDS: automated summarization, automated alginment, paraphrasing

1　はじめに

近年の電子化文章の増大により、計算機を利用して人間の文章処理を支援する研究がなされるようになった。そのような研究のひとつに、計算機に文章を要約させる試みがある。

計算機による要約の試みでは、文章中の重要と思われる部分を抽出することを中心に研究されてきた。しかし、要約は人間の高度に知的な作業であるため、計算機により重要と認定された部分を列挙するだけではなく、要約文章の結束性、構成などの点で課題があることが認識されてきている [1][2]。

このような課題を解決するためには、実際に人間が作成する要約が、要約の元になった文章のどのような表現を用いて要約を作成するかを対応付けし、計算機でも実現が可能な操作として整理・分類することが有益である。

このような対応付けを自動的に行う試みとして、例えば、Marcu [3] は論文とそのアブストラクトのように、要約とその元文章が組になっている文章集合から、要約の各文が要約元文章のどの文から生成されたかを、コサイン類似度を用いて自動的に対応付ける手法を提案している。また、日本語の自動要約の研究では加藤らが DP マッチングの手法を用いて、局所的な要約知識を自動的に抽出する研究を行っている [4]。加藤らの研究では、放送原稿とその要約を使用しているため、要約文書は元文原文の残存率が高く、語や文節レベルの言い換えといった局所的な要約知識の獲得に効果をあげているが、人間が行う、より一般的な要約作成に必要な知識獲得を行うためには、より複雑な文再構成操作を仮定した、対応付け手法の拡張が必要となってくる。

本研究では、このような背景から、人間が作成した要約に対して、要約元文中で統語的な依存関係のある表現が、同様な依存関係をもって要約で使われているか否かを基準に文・文節対応付けを行い、要約生成において、どの程度の文再構成操作を仮定することが適当であるかを調査した。

2　要約における文再構成操作

2.1　調査対象データの収集

人間が作成した要約を調査する対象データとして、毎日新聞社説 90 記事に対して要約したものを用いる。要約を行ったのは 3 人の作業者で、それぞれ 90 記事をすべて要約する。したがって、調査対象の要約は全部で 270 要約となる。要約の長さは文字数で元記事の約 40 %なるよう指定した。要約を行う際に作業者に与えた指示として、「全体のあらすじと著者の主な主張がわかるように要約する」、「固有名詞はできるだけ原文の表現を用いる」という 2 つの制約を課した。

3 人の要約作成者が作成した要約は要約のべ 2467 文であった。この要約文を調らべてみると、元文書中の文と一字一句違わない形で要約で用いている例は中 692 文しか存在せず、その残りである 1775 文の要約文は人間が何らかの文の再構成・生成操作を行って作成していることが分かった。本論文では、このような要約文がどのように作成されているかを分析する。

2.2　要約文を作成する操作

計算機処理を前提に要約文作成操作を考えた場合、要約元の文章で使われていた表現が、どの程度そのまま要約で用いられるかが、重要な問題となる。

要約研究の多くでは、自動要約を文章中から重要な情報を記述している句や文といった部分を抽出する重要部分抽出の過程と、それを要約として、再構成する過程の 2 つを仮定することが多い。抽出の単位として最も基本的なものは文であり、実際、数多くの要約システムでは文を抽出単位として選んでいる。しかし、文を抽出単位として選んだ場合、文を再構成する過程において文節、句、文といった様々なレベルでの書き換えの多様性を考慮しなくてはならない。

また、さらに再構成の過程の問題を難しくする要素として、要約中の 1 文が要約元文章中の 1 文の表現だけを用いて要約文を作成するわけではないことが挙げられる。例えば、Jing ら [5] の研究では、英語における要約の重要文抽出とその再構成操作を、計算機による実装を見通した諸操作として整理している。彼女らの研究では基本的な操作として、以下の 2 つの操作を挙げている。

- 文短縮 (Sentence Reduction): 元記事の 1 文を短くして要約中の 1 文で表す。
- 文結合 (Sentence Combination): 元記事の複数文をまとめあげて要約中の 1 文で表す

もちろん、この他にも、要約作成者が元文章を完全に理解した上で、元文章中には現れない新しい表現を作成することも当然考えられる。しかし、計算機による要約生成を考える上で、表層情報からわかるレベルの言い換えや文結合を組み合わせて作られた要約事例と、計算機では実現が難しいと思われる操作によって作成された事例とを分け、それぞれの事例がどの程度存在するかを知り、その操作の特質を調査することが必要となる。

3　かかり受け構造を用いた自動的対応付け

本研究では 2 節で述べたような現状を踏まえ、人間が作成した要約を調査する上で、要約元文章中の文の構造と要約の文の構造を踏まえた対応付けを自動化す

ることを考えた。その際の観点は、次のようなものである。

1. 元文章中での文の統語構造はどの程度保存されるのか
2. 要約中に現れる表現はどの程度新しく生成されるのか
3. 文結合はどの単位を基準に、どの程度なされるか

本稿では、このような観点から、文内の統語的構造を表現するかかり受け構造を考慮して文対応付けをとり、複数文との対応付けを考慮して、文対応付けを繰り返す対応付け戦略を考えた。この戦略に従い、我々が行った対応付け処理の概要を図 1 に示し、以降、その具体的な手順を順を追って説明する。

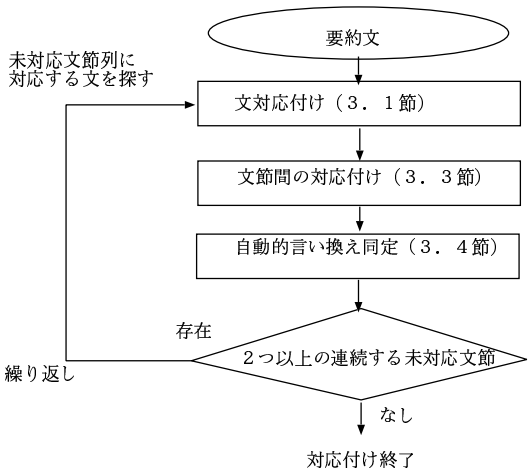


図 1: 対応付け手法の概観図

3.1 文対応付け

本稿が仮定する文対応付けは、文の構造を踏まえた対応付けである。具体的な文の構造の表現形態としては、かかり受け構造を採用した。かかり受け構造は、文中の各文節の修飾関係を木構造で表現したもので、関係付けの交差を許さない。ここで、かかり受け構造を解析した例は図 2 に示し、図中の矢印が文節間の修飾関係となる。かかり構造は、文末の文節以外の各文節が 1 つのかかり先を持つため、木構造となる。ここで、文末の文節をこの木構造の「根」とみたて、どの文節の係り先になっていない文節を「葉」と呼ぶ。また、各文節を木構造中の「節点」と呼び、文節間のかかり受け関係を「枝」と呼ぶ。

また、本稿では、この構造を利用して自動的に対応付けを行うため、かかり受け解析モジュール CaboCha [6] を利用した。

かかり受け構造を利用した対応付けは、文章中で使用された表現の特定を、かかり受け構造上の文脈で制約することにより対応付けを精密化すると同時に以下のような性質の対応付けが期待できる。例えば、「彼女に花をあげる」という文を考えたとき、下の 2 つの

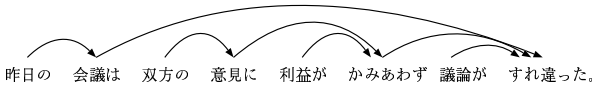


図 2: かかり受け構造の例

例のような「花を」と「彼女に」の出現順が違う文、新たな文節が挿入された文があった時でも、(「花を」→「あげる」,「彼女に」→「あげる」) というかかり関係には変わりがないという性質をもつ。

「花を 彼女に あげる。」
「彼女に、なんとしてでも、花をあげる。」

かかり受け構造を手がかりとした対応付けは、要約文のかかり受け構造の木と、共通のかかり受け構造の部分木を要約元文章中の文から探し出す作業と考えることが出来る。以下、本研究が用いた手法を説明する。

本手法での対応づけは文のかかり受け構造での、葉から根までのすべての経路をもとにする。例えば、図 2 のようなかかり受け構造木から、抽出される経路の集合は

{ [昨日の, 会議は, すれ違った], [双方の, 意見に, かみあわず, すれ違った], [利益が, かみあわず, すれ違った], [議論が, すれ違った] } となる。このような経路の集合を要約文と、対応づけ候補文となる要約元文の各文それぞれに対してすべて抽出する。

要約文 s のかかり受け構造のもつ経路の集合を P_s、要約元文章中の各文のかかり受け構造の経路をすべて集めた集合を P_{cand} として、s に対応付けられる対応文 a を求める。求める手順は、まず、集合 P_s の各経路と集合 P_{cand} の経路のすべての組合せに対して、DP マッチングを用いて、最長共通部分列 (Longest Common Subsequences、以下 LCS) を求め、次に、求めた LCS の集合の中で、最も長い LCS をもつ経路がある文を対応文 a とする。

DP マッチングでは、記号列対を一致させるために各記号をどのように編集するかについて優先順位を設けて LCS を求めるが、今回の実験では、比較する経路の対（それぞれは文節列）におけう各文節の組の編集に対して、「削除しない（文節一致）」、「片方の文節の削除」、「両文節の削除」の順で優先順位を持せて LCS を求めた。文節の一致をどうみるかは 3.2 節に述べる。

3.2 DP マッチング中の文節の比較

文節の言い換えは多様であるため、要約で行われている表現の多様性を見極める上では、例えば、字面でよく似ていておおよその検討がつくものと、シソーラスなどの何らかの辞書を使わないと判断がつかないも

のは区別して扱うことが重要となる。そこで、本稿では、単純な、単文節書き換え規則のみを用意するにとどめ、その結果から、より高度な言い換え事例を収集し、要約で行われている操作の多様性について見通しをたてる方針をとる。

本実験での文節の一致は、文節内の意味的主辞の品詞の下位分類まで含めた一致と原型の一致を文節の一致の基本とする。以下、若干の例外を説明する。

文節の接辞が格助詞の場合は、「が」格が助詞「は」や「も」表現のような取り立て表現になること、もしくはその逆になることを書き換えとしてみとめた。

用言を主辞とする文節については、助詞、助動詞、活用語尾などの文節末要素についてはゆるい制限で言い換えが可能であるとし、主辞の原型が同じであれば、同じ文節とした。

意味的主辞の品詞の一致の例外としては、名詞に格助詞助詞がついていたものが、名詞だけになった場合、また、その逆の場合、下位分類までの一致とはせず、当該語の原型の完全一致をもって文節の一致とした。

また、助詞を接辞とする文節に名詞が複数存在するときは、格助詞が一致しているときに限り、文節内の半数以上の名詞が同じ、もしくは文字ベースで半数以上の文字が一致をもって文節の一致とした。

3.3 文節間の対応づけ

要約文と対応文との間の各文節の対応付けを、3.1節で述べた方法により、要約文の経路と最長の LCS を経路としてもつ文として既に対応づけた情報を用いて行う。具体的には、文対応付けを行った際に利用した、要約文 s と対応文 a の、それぞれのかかり受け構造木の経路すべての組み合わせについて求めた LCS の集合を情報として用いる。ただし、文節対応に用いるのは、元文でなんらかの統語的関係にあったことを仮定しているので、LCS の長さが 1 のものは、この LCS の集合の中から除外した。

具体例を示す。例えば、図 3 の場合は、対応文のかかり受け構造における経路の集合は前節にも述べたように、{[昨日の, 会議は, すれ違った], [双方の, 意見に, かみあわず, すれ違った], [利益が, かみあわず, すれ違った], [議論が, すれ違った]} である。

これに対し、要約文のかかり受け構造における経路の集合は { [会議は, すれ違った], [双方が, かみあわず, すれ違った], [いつまでも, すれ違った], [議論が, すれ違った]} となる。

この 2 つの経路集合の全て組み合わせの LCS のうち LCS の長さが 2 以上のものの集合は以下のようになる。

{ [会議は, すれ違った], [かみあわず, すれ違った], [議論が, すれ違った]} }

この LCS の集合に基づいて、要約文と対応文のそれぞれについて、この集合の各要素の LCS に含まれる文節、すなわち文節対応がついた文節を 1、未対応の文節を 0 でマークしたものが、図 3 での各文節の下につけられた 0, 1 の数字である。以降本稿では、この対応・未対応を示した 0, 1 の列を対応文および要約文の編集記号列と称す。

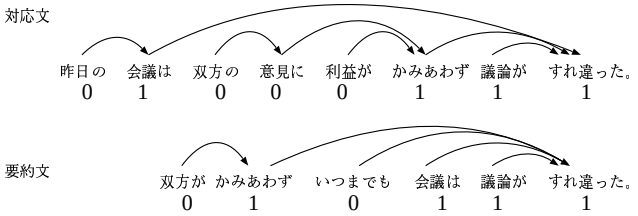
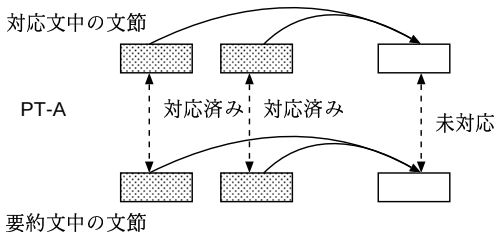


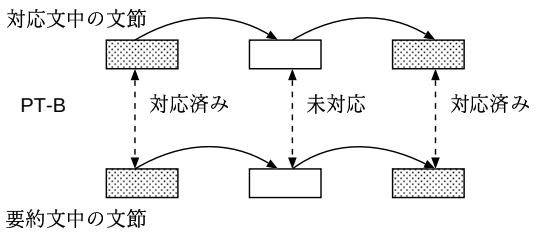
図 3: 文節対応付けの例

3.4 かかり受け構造を手がかりとする文節言い換えの自動同定

文節対応が終わった時点で、より複雑な要約操作を知る足がかりとして、かかり受け情報を利用して確実に単文節の言い換えだと推定できる事例を、文節対応付けがとれているものと仮定し、自動的に収集する。

確実な単文節の言い換えと仮定するのは、3.3 節の文節対応付けが終わった時点で、要約文、対応文中の未対応の文節が、それぞれのかかり受け構造の中で以下のような PT-A、PT-B の位置にある場合である。





図中の四角は単文節を示し、斜線で塗りつぶされた四角は既に文節対応付けがなされている文節である。

図中の PT-A は、未対応の 1 文節が、2 つ以上の対応済み文節のかかり先になっている場合であり、図中の PT-B は、 未対応の 1 文節のかかり先と、その未対応の文節を元文中で直接係り先にしている 1 つ以上の文節すべてが対応済みの場合である。

3.5 文結合操作の同定

2.2 節で述べたように、要約文は常に、要約元文章中の文と 1 対 1 で対応付けられるとは限らない。そこで、本対応付け手法は、要約文に 1 文に対し、複数の要約元文が対応付けられる文結合操作を考慮するため、3.1 節から 3.4 節までの一連の対応付け処理により未対応の文節列が、要約文中に残っている場合、図 1 のように、その未対応の文節列に対応する文を 3.1 節から 3.4 節までの一連の処理を繰り返すことにより対応する。この際、対応付け繰り返し条件は、2 文節以上の未対応文節列とした。

例えば、第一回目の対応付け試行により、以下のような要約文と対応文の組が得られた場合を考える。

要約文「借り上げに 力を 入れるなど 積極的対応を 求めたい。」
元文 1「県には、前例にとらわれず、 積極的対応を 求めたい。」

この例の場合、第 1 回目の対応付け後の要約文の編集記号列は'00011' である。ここで、図 1 の手順にあるように、2 つ以上の連続する未対応文節が文頭にあるため、次の繰り返し処理で、文節列 [借り上げに, 力を, 入れるなど] に対して、3.1 節から 3.4 節までの一連の処理を適用する。元文章中に、例えば、「借り上げに、力を入れ、仮設住宅を確保すべきだ。」といった文ががあれば、この未対応文節に対して、この文が対応付けられ、以下のような対応付け結果となる。なお、要約文に対しての編集記号列は'11011' である。[1]

要約文「借り上げに 力を 入れるなど 積極的対応を 求めたい。」
元文 1「県には、前例にとらわれず、 積極的対応を 求めたい。」
元文 2「借り上げに、力を 入れ、 仮設住宅を 確保すべきだ。」

このように、要約文の編集記号列における長さ 2 以上の未対応文節列'0' の列がなくなるか、そのような未対応文節列に対して、元文章中のかかり関係をもった文節を対応付けられなくなるまで対応付け作業を繰り返す。

4 対応付け手法の適用と要約操作

4.1 対応付け結果

要約の全 2467 文に対して、3 節で述べた自動対応付けを繰り返し適用し、対応付けが出来なくなるなるまで繰り返した結果を表 1 に示す。表中の各回の試行で対応付けられた文数は、当該回の試行以降で対応付けが出来なくなった文数を示す。また、当該の試行で対応付けができなかった文節の数を、未対応文節がない場合、未対応文節が文中の文節の半分より少ない場合、文中の文節の半分以上が未対応の場合の 3 つに場合分けし、それぞれの内訳数を示した。なお、第 1 回

[1] 3.2 節の文節比較規則では「入れるなど」と「入れ、」は同一文節とみなさないため、後ろから 3 番目の文節は未対応となる。本研究では、このような厳しい条件の文節比較に基づいて対応付けをすることにより、要約での文再構成操作の特質を分析する。

表 1: 自動対応付けの適用と文節対応付けの傾向

対応付試行数	対応付文数	要約文中の未対応文節数		
		なし	半数未満	半数以上
1	1598	990	444	164
2	609	141	401	67
3	146	25	118	3
4 以上	19	3	15	1
合計	2372	1159	978	235

ですべての文節が対応付け可能だった 990 文は、元文章中の文をそのまま要約でも用いていた 672 文を含んでいる。

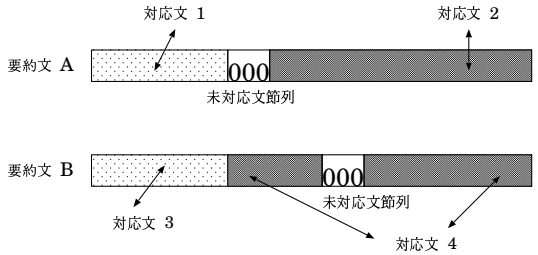
この表 1 の結果から、大部分の要約文は 3 回までで対応付けが終わることが分かる。また、要約文中に文節対応づけがきない文節が残る要約文のうち、文中の半分以上の文節が未対応になる文は 235 文である。表 1 から、文中の半分より少ない文節が未対応になっている文は 978 文であるので、本手法の対応付けで文節対応付けで未対応文節が残る文であってもその 81%は未対応文節は文中の文節の半分以下である。

なお、2467 要約文中、本手法で対応付けできなかった 95 例あった。これらの例は、元文章中の単語レベルでのみ使い、要約者が新しく文を作成したような例であった。

要約文中の未対応文節を含む表現がどのように生成されたかを検討する。大きな分類として、未対応文節が要約文の文頭、文末、文中のどこにあるかで分けることができる

要約文頭、文末の未対応文節は、表 1 に示した要約文中の半数以上の文節未対応で残る場合に多く、その場合、文頭・文末に長さ 3 以上の未対応文節列があることが多い。

文中の未対応文節は、さらに下図の二つに場合分けができる。



二つの場合は、要約に対応付けられた対応文によって場合分けし、2 つの対応文の接合部分に未対応文節がある場合 (図中の要約文 A) と 1 つの対応文のかかり受け構造の中に未対応文節が挿入された形 (図中の要約文 B) がある。図中の要約文 A のような要約文は、この未対応の文節を文結合にともなう言い換えとみなすことができ、要約文 B のような、1 つの対応文のかかり受け構造の中に挿入されている形の未対応文節とみることができる。

表 2: 要約文における文頭の未対応文節

連続する未対応文節の位置	未対応文節列の長さ				
	1	2	3	4	5 以上
文頭	369	157	72	47	50
文末	93	75	31	39	33

以降、要約文の文頭・文末、文中の言い換え、文結合による言い換えのそれぞれの特徴を述べる。

4.2 要約文の文頭・文末の特徴

要約文の文頭と文末で未対応の文節列があるものを、未対応文節の長さによって整理したものを表 2 に示す。表 2 から、文節の未対応には 1 文節のものが多い。単文節の要約文頭の付加としては、「XX は」といった主題表現や、「しかも」や「しかし」といった副詞表現、接続表現が文頭にある場合が多かった。これらの表現は、1 文節に限らず、文頭にあらわれる未対応文節全般にみられ、要約文の生成において、主題、背景情報や付加情報などを付加する副詞表現、接続表現といった文脈上の機能的な要素が重要であることを示している。ただ、対応付けの誤りとして、かかり受け解析で長い距離で依存関係を間違い、対応文に対応する文節があるにも関わらず、未対応になったしまった単文節の例があり、文頭の未対応単文長が 1 である例を必要以上に多くしていた。

文末の未対応文節の事例を調査すると、対応文の文末表現を簡略化した例と、要約者の主観や独自の説明が入れられた例が混在しており、要約文の文末表現は要約者によっての恣意的な度合いが高いと言える。下に文末の変化例をあげておく。

…のケースも六百件以上ある。→…のケースも多い。
…乏しさを弁解する理由にはできない。→…乏しさを露呈した。
…明らかにすべきだ。→…明らかにし、不安を取り除くべきだ。

その他の重要な例として、主題が文末に移動させた構文的変化を原因とする以下のような場合もあった。

大切なことは、金融と財政を 1 つのパッケージにすることだ。
↓
金融と財政をパッケージにすることが大切だ。

4.3 文中の言い換えの特徴

今回用いた対応付けでは、3.4 節で述べたような単文節言い換えの自動同定を行った。この方法で、言い換えを自動的に推定することが出来た例は 121 例であった。この方法で自動的に対応づけた文節言い換えの例をもともとは未対応であった文節を［ ］で囲み、周辺文脈とともに表 3 に示す。表 3 での言い換えは、最も単純なものとして、G1 に分類した例 1 から例 3 のように、漢字表記の違い、誤字などを含め単文節で

表 3: 自動同定した単文節の言い換え例

分類	例番	要約文例
G1	例 1	対応文: 重く、 [つらい] 課題である。 要約文: 重く [辛い] 課題である。
	例 2	対応文: 女性が 生涯に [産む] 要約文: 女性が 生涯に [生む]
	例 3	対応文: 崩れれば、 競争も [強まる] 要約文: 崩れれば、 競争が [活発になる]
G2	例 4	対応文: 電事連が 公に [しないのは] 要約文: 電事連が [公開しないのは]
	例 5	対応文: 反するのではという 問題も [抱える。] 要約文: 反するのではという [問題である。]
	例 6	対応文: 支持率は、 六七%にも [達している。] 要約文: 支持率は [六七%。]

表 4: 文中の言い換えの例

分類	言い換え例
A	対応文: 政治家は、 [韓国の] 発展と 要約文: 韓国政治家は、 [その] 発展と 対応文: 事態の [推移を] 注意して、 要約文: 事態の [推移に] 注意して 対応文: [施設に 一斉に] 家宅捜査を 行った。 要約文: [施設に対して 全国一斉の] 家宅捜索を 行った。
B	対応文: 将来のために [工夫が] 必要だ。 要約文: 将来のために [工夫する ことが] 必要だ。 対応文: 警察組織の [トップが] 銃撃された。 要約文: 警察組織の [トップ、 国松孝痔警察庁長官が] 銃撃された。 対応文: 今週になり、 学校は [ようやく] 隔週 要約文: 今週になり、 [今 ようやく] 学校は 隔週
C	対応文: [支援の ネットワークを] 広げては 要約文: [支援を] 広げては 対応文: 社会党が、[侵略行為・侵略戦争や 植民支配の] 反省の 要約文: 社会党が [侵略行為などの] 反省の 対応文: 実施する [措置予定は 七百件を 超えたが、] 大半は 要約文: 実施する [措置予定の] 大半は 対応文: 青島知事の [こうした 方針や 行動に] 反発した 要約文: 青島知事の [施政方針に] 反発した

帰結する言い換えである。これに対し、表 3 の G2 に分類した例 4 から例 6 は、単純な単文節の言い換えではなく、複数文節の言い換えと考えた方がよいものである。このように、自動同定できた表 3 のような例は、ほぼ間違いなく言い換えとみなせるが、抽出数は 121 例と少ない。これは、3.4 節で用意した要約文と対応文の両方の周辺かかり受け構造の条件が厳しかったことが原因であると考えられる。そこで、以降、この条件でとらえきれなかった要約文中の未対応文節がどのような形で生成されているかを議論する。

要約文中への言い換えのパターンを説明する上で、図 4 に示したような未対応文節列の長さ n, m を導入すると、うまく整理できる。対応文側の未対応文節列の長さが n であり、要約側が m である。

もっとも基本的なものとして、$m = 0$ で $n > 1$ の場合と、その逆の $n = 0$ で $m > 1$ の場合がある。前者は、要約文に対しての文節の追加の追加である。調査をしてみると、このような文節の追加は、単文節で行われることが多く、挿入させる文節の機能的な役

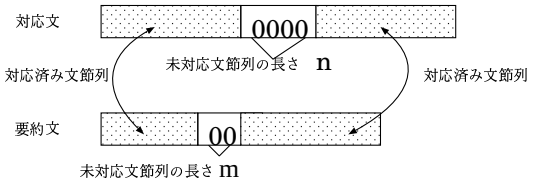


図 4: 要約文中の未対応文節列の長さ

割も、主題、副詞表現、接続表現といったものが多いことがわかった。このような文節追加の例は 113 例あった。

文節の追加以外の要約文と対応文との相違については、0 より大きい n,m に対して、その大小関係に基づいて以下の 3 通りに未対応文節を分類した。

同長表現での入替え: $n = m$ の場合の例を、表 4 の分類 A に示した。例数は 210 例であった。これらの例は、自動同定した言い換えである表 3 の G1 の分類例と同様な、単文節の言い換え、すなわち、$n = m = 1$ が最も多い。$n = m > 1$ の場合であっても、今回収集した要約の中では文節数が 3 を越えるものはなく、単文節の言い換えの連鎖とみなせることが多い。

表現の拡充: $n < m$ で対応文に何らかの文節を追加している例を、表 4 の分類 B に示した。例数は 42 例あった。追加のされ方は、例のように、単なる表現の変化、具体化などがあった。特徴としては、挿入される新しい文節の数が 1 つだけというものが多く、その挿入に伴って、要約文のかかり受け構造が対応文のそれと異なってしまい、双方に未対応の文節ができてしまっていた例が多い。表 3 の G2 の分類例の、ちょうど逆の操作となる。なお、今回調査した要約では、元文章中にない複数の新しい文節を要約文中に挿入する例は少なく、n と m の差が 4 以上のものはなかった。

表現の圧縮: $n > m$ で、要約側で未対応文節が減る例で、表 4 の分類 C に示した。言い換え箇所は 241 例あったが、うち 187 例が、要約文側の未対応文節長 m が 1 のものであった。これらの言い換え特徴としては、表 3 の G2 の分類例に準じる言い換えが多い。また、241 例中、n と m の差が 1 のものが一番多く 101 例、2 番目に多いものが、差が 2 のもので 49 例である。

表現の圧縮方法として特徴的であったものは、

「A が XX する B」 ↔ 「A の B」

といった「の」表現に関わる言い換え事例であった。

他方、要約文側の未対応文節が 3 以上の長い未対応文節である場合、このような言い換えを、複数文節にわたる言い換えが複合的に行われているとみるか、未対応文節列対を一言で言い換えたとみるかには課題が残る。

表 5: 各文結合箇所のタイプ分類

未対応文節	連用接続	主題・主格	連体接続	格要素	その他	合計
なし	176	25	19	10	2	232
あり	319	46	56	11	41	473

4.4 文結合にともなう言い換え

未対応文節をともなう文結合箇所は 473 例あった。文結合が行われる箇所の基準として、まず、表 1 未対応文節がなく 2 文以上が対応付けられた対応付け試行回 2 回以上 (対応付けられた要約文の数は試行回 2 回、3 回、4 回以上でそれぞれ表 1 から 141, 25, 3 文）の 169 文の要約文における文結合の形を分類した。この 169 文の要約文中の文結合部分は 232 箇所であった。ここでは、その結合の諸相を分析するため、文結合のタイプを表 6 のような形で分類してみた。

表 6 の分類を用いて、未対応文節を伴わない文結合箇所を分類した結果を表 5 の上段に示す。未対応文節を伴って文結合がなされる例の分類結果は下段に示す。表 5 をみると、言い換えがなされるかどうかにかかわらず、文結合として最も多いのは、連用接続の形である。

連用接続の例を調べてみると、それらの意味的な接続関係は、背景、理由、連鎖対、結果といった情報を付加するものが多くみられた。すなわち、要約では、中心的な命題に関しての情報だけではなく、その理解を助けるような情報を付加していることが分かる。

連体接続で付加される情報としては、接続対象の事象や主体の背景や、「のような」といった例示が多かった。

文結合の同定には、あいまいさが残る。特に結合される文節数が少なくなったときに、要約元文章中に何度も出現するような慣用表現を文結合の対象とするかどうかには課題が残る。表 5 中でその他に分類した例は、そのような例を含めて、表 6 の例に示した分結合の類型としては適当でないと思われた例数である。

この中には、原文中の「XXX では YYY である」といった構文が、「YYY なのは、XXX ということ。」となるような、構文全体の構造が変化した例も含まれる。このような構文変化は、異なる繰り返し試行において 1 つの要約文に同じ対応文が対応付けられたことを手がかりに、ある程度自動的に判別することが可能である。具体的には表 5 の下段のその他に分類した 41 例中 8 例は、この方法で判断が可能であった。しかし、文結合がおこる文のどちらもにこのような構文変化が起こり、文節の言い換えも組み合わさった文結合もあり、そのような文結合を適当と見るかは、文の構文的なの書き換えをどこまで許すかに依存する。

表 6 の分類例に準じる文結合の、未対応文節を伴う例を表 7 に示す。文結合にともなう未対応文節は、元文の文節を別の語に言い換えや、品詞の変化によ

表 6: 文結合のタイプ分類

連用接続	2 つの文を連用接続にて結ぶ。接続形態は論理接続、付帯状況の説明などがあった。
	対応文 1　また、大学側も 柔軟な 思考で 共同研究に 応じるべきであろう。 対応文 2　特に、 各地の 大学は 地方ニーズに 応える よう 配慮してほしい。 要約文　大学側も 共同研究に 応じるべきで、 特に、 各地の 大学は 地方ニーズに 応える よう 配慮してほしい。
主題表現の追加	結合する文の一方に現れた表現が、他方の文の主題や主格主語になっている。
	対応文 1　北京での 南北の 次官級交渉は、関連問題で 最後の 詰めを 残している。 対応文 2　北朝鮮側は、 コメの 支援問題以外の 政治問題は 今回の 会談では 話し合いたくない立場を 譲っていないようだ。 要約文　北京での 南北の 次官級交渉は、北朝鮮側も、 コメの 支援問題以外の 政治問題は 今回の 会談では 話し合いたくない立場を 譲っていないようだ。
連体接続	結合する文の一方の表現を用いて、他方の文の名詞を連体修飾する。
	対応文 1　新党結成を めぐり 社会党は 大混乱で、 政権基盤は 大きく 揺らいでいる。 対応文 2　首相にとっては 演説どころではなかったのかもしれない。 要約文　新党結成を めぐり 社会党は 大混乱で、 政権基盤が 揺らいでいる 首相にとっては 演説どころではなかったのかもしれない。
格要素の追加	結合する文の一方の表現が他方の文の句の格要素となっている結合。
	対応文 1　観光目的などで 来日して 不法残留する 外国人労働者の 医療問題は、「人道」と「不正」が からみつく。 対応文 2　行政は 「不正」を 重く 見て、 良心的な医療機関ほど 「人道」的に 対応してきた。 要約文　行政は 「不正」を 重く 見て、 不法残留する 外国人労働者の 医療問題に、良心的な 医療機関ほど 「人道」的に 対応してきた。

表 7: 文結合に伴う言い換え

対応文 1	…負担増を求める以上、政府自らも…
対応文 2	阪神大震災での損失を…
要約文 a	…負担増を求め、そのうえ阪神大震災での損失を…
対応文 1	…事情は理解すべきだと思う。
対応文 2	輸入を増やし、不均衡を …
要約文 b	…事情を理解し、内需拡大で 輸入を増やし、不均衡を…

り起こるが、例に挙げた要約文 a と b は、そのような変化だけではなく、結合後の文をよりよくするために、要約文 a では「そのうえ」、要約文 b では「内需拡大で」といった対応文に存在しない文節を追加している。このような例のほかにも、文結合にともなう文節の追加には、主題表現や接続表現の追加などが多くみられた。

5　まとめと今後の課題

本稿では、人間が作成した要約が、要約元文章の表現をどの程度用いて作成されているかを調査した。調査には、要約文および要約元文のかかり受け構造を基準に、文結合を考慮して自動的に文節を対応付けする手法を用いた。その結果、要約生成において、文結合操作が大きな役割を担っていることを確認した。また、要約文中のすべての文節が対応付けできた要約文を除いても、残りの要約文の 81%は、その要約文の文節の半分以上が、元文章中でなんらかのかかり受け関係を持つ表現を用いて生成されていることが分かった。

さらに、対応付けで、未対応のまま残っている文節についても、その多くは、文結合にともなう言い換え、単文節を中心とする言い換え、文脈上の機能的表現の追加という形で分類できることが分かった。

今後の課題としては、今回の実験で収集できた言い換え事例を考慮しつつ、さらに広い範囲の要約に対して適用できる対応付け手法の拡張を考えている。また、その対応付け結果を用いて、それぞれの要約操作に関する知識を自動獲得することを検討したい。

参考文献

[1] H. Nanba and M. Okumura. Producing more readable extracts by revising them. *Proc. of COLING-2000*, pp. 1071–1075,2000.

[2] E.Bloedorn, I.Mani, B.Gates. Improving summaries by revising them. *Proc. of ACL'99*, pp. 558–568, 1999.

[3] D. Marcu. The automatic construction of large-scale corpora for summarization research. *Proc. of SIGIR '99*, pp. 137–144, 1999.

[4] 加藤直人, 浦谷則好. 局所的要約知識の自動獲得手法. 自然言語処理, Vol.6 No.7, pp. 73–92, 1999.

[5] H.Jing and K.R.McKeown. The decomposition of human-written summary sentences. *Proc. of SIGIR'99*, pp. 1–8, 1999.

[6] 工藤拓, 松本裕治. チャンキングの段階適用による係り受け解析. 情報処理学会研究報告 01-NL-142,pp. 97–104, 2001. (http://cl.aist-nara.ac.jp/ taku-ku/software/cabocha/)