自然言語処理　138－2
（2000. 7. 18）

テキスト構造に基づく要約生成制約条件の検討

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

あらまし　近年、自動要約等のテキスト処理においてテキストの構造情報を利用する技術への期待が高まっている。しかし、テキスト構造が自動要約にどのような有益な情報を提供するか、また、自動要約にはどのようなテキスト構造の表現形式が有益であるかは明らかになっているとはいえない。そこで本研究では、実際に人間に自然な要約を作成してもらい、実験的なテキスト構造解析タグ付け体系によりテキストを構造解析した結果と比較することで、テキストの構造と人間の要約との間にどのような相互関係を見出すことが可能か調査を行った。この結果、テキスト構造と要約の生成との間に興味深い関係があることを確認できた。

キーワード　テキスト構造、自動要約、文生成

Role of Text Structure for Automated Summarization: Clues for Sentence Combination

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

Abstract

To investigate the relationship between text structure and its summary, we examine human-generated summaries. We prepare manually analyzed coherence structure of each source text of the summaries. From manual analysis on alignment between sentences in a source text and those in the summary, some roles of text structure in summary generation are suggested.

key words　text structure, automated summarization, sentence combination

1 はじめに

要約は人間の自然言語を用いた知的な活動の一つと考えられる。自動要約はこの知的な活動を計算機によって実現しようとする試みである。このような自動要約の研究には様々なアプローチが存在するが、より一般的と思われる高次の言語情報を利用し、人間にとってより自然であり、有益な要約の生成を目指す研究がある。例えば、Ono ら [1] や Marcu[2] の研究では、Mann と Thompson が提案した修辞構造理論 (RST: Rhetorical Structure Theroy) [3] に基づいて、要約を行うテキストを木構造に解析し、その解析情報を用いて重要部分選択する試みがある。具体的には、解析されたテキストの構造木の位置に基づいて、節や文に相当するテキストの部分の重要度を点数付けするものである。この手法において、テキストの構造に対してどのような点数の付け方が適当であるかは完全に明らかになっているわけではない。しかし、情報検索分野で古くから研究されてきた語に対する統計的な重要度の重み付けを基本にする手法と、言語学的枠組みで研究されてきた談話構造を有機的に結合させる試みとして、この手法に対する期待は高い。

自動要約における数多くの関連研究では、大まかな処理過程を、テキスト中から重要だと思われる文や節を抽出する過程と、抽出した文や節を再構成し出力する要約生成過程との 2 つに分類することが多い。テキスト構造は、要約における重要文や重要節の抽出だけでなく、生成する要約をより自然にする上でも役に立つことが予想される。重要部分抽出は人間に重要文や節を同定させ、自動要約システムの評価とすることも多いが、例えば、難波ら [4] の研究のように、抽出した重要部分の集合を人間が読みやすいように書き換えてやる必要性が指摘されている。このように自然な要約生成を実現させるためには、純粋にテキストの節や文をとりだすだけではなく、その部分要素に関連した構造上の情報も必要であるように思われる。この生成の側面に対し、Mani ら [5] は、テキストの結束性 (cohesion) にもとづく構造と RST の構造の 2 つの構造が要約の過程にどのように影響を及ぼすかを検討し、この 2 つの構造が重要部分抽出に有益な役割を果たすことを確認し、さらに、RST の構造は重要部分の抽出だけではなく、要約生成の過程でも有益であるという予想をたてている。

以上のようにテキストの構造が表現している文間の修辞的な関係や関連性を要約における重要部分抽出と生成の 2 つの過程に対してどのように応用してゆくかについて、すべてが明らかになっているわけではない。我々は、このような問題を明らかにするため、実際に人間に要約を作成してもらい、その要約と元テキストの構造とを比較する実験を行い、要約作成に対してテキスト構造をどのように有益に利用できるかを検討した。本稿ではその結果を報告する。

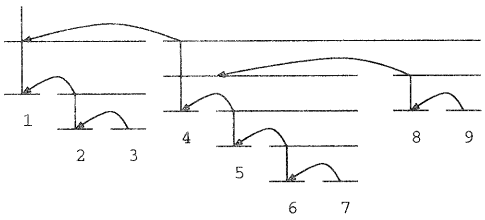

図 1: テキスト構造の木構造モデル

2 構造解析実験

我々は要約とテキスト構造の関係を研究するため、本実験に先行して日本経済新聞の報道記事 32 記事、合計 500 文に対して 3 人の被験者によるテキストの構造解析タグ付け実験を行った [6] 。本節ではその実験結果を簡潔に紹介する。新聞報道記事を実験対象に選んだ理由は、記述の目的がはっきりしており、テキストの長さも短く、人間による重要情報の選択のゆれが少ないと考えたからである。

テキストの構造解析は RST を参考にし、実験的な構造解析スキームを作成する。RST は前節で紹介した研究 [1][2] においてもテキスト構造のモデルとして利用されている。例えば、9 文から構成されるテキストの構造を RST によって解析し、構造を提示した例が図 1 である。図 1 中の数字は文の番号を示しており、文と文は修辞関係と呼ばれる抽象的な関係により結び付けられ、テキスト全体は木構造に解析される。修辞関係については、結び付けられた要素間のうち、どちらをより内容的な中心とみなすかによって、矢印がその関係方向が決まる。(本研究ではより中心的な側を関係先，他方を関係元と呼ぶ)

RST は自動要約に応用することを前提として提案されたわけではないために、人間による判断の個人差を許容するなど、計算機でテキストを扱うモデルとして利用するには課題がある。例えば、Moore と Pollack[7] が指摘するように、部分領域間に選択された修辞関係の種類によっては同時に多重解析が可能で

あり、また、テキスト中の修辞関係により結ばれた2つの部分領域のうち修辞関係の方向が逆になってしまう例があることが指摘されている。実際、我々が予備的な実験としてオリジナルのRSTを用いてテキストの構造分析をしてみたところ、RSTでは15種類以上の関係種類を提案していることもあり、関係種類の選定の差異によって構造の同定がゆれてしまうことが分かった。そこで、本研究ではテキストの構造を決定する上で、修辞関係よりも関係方向の決定が重要と考え、具体的なテキスト構造解析の手順を以下に記す通りに簡略化した。

まず、テキスト中の個々の文に対して、その文にもっとも関連がある文を1つ選ぶ。次に、このようにして選んだ文対のそれぞれの文に対してテキスト全体における重要性の度合いを比べ、その文間に重要性についての順序（主従）関係を付ける。この関係付けをテキストのすべての文に対して、循環構造をもたないようにする制約、及び、最も重要性の高い文がテキスト中で1文のみになるという制約を課してテキストの構造解析を行う実験を行った。このようなテキストの重要性に対しての順序関係は、テキストの構造という抽象的な制約を加えた重要文抽出と考えることもできる。関係種類については、まずはおおまかに主従関係と並列関係を区別しておき、将来的にその関係を詳細化する方向性を考えている。

以上のような構造解析実験において、被験者がある文を他のどの文に関係付けるかの傾向について、関係距離と被験者間の一致の傾向を示したのが表1である。関係距離はある文が文書の先頭に向かって直前に関係付けられたときを1として、いくつ前の文に関係付けられるかを示したものである。この表から、被験者の関係付けの傾向として、ある文に対してほとんど制約なくテキスト中の任意の文に対して関係付けを許したにもかかわらず、ある文がその直前の文に関係付けられる事例が多く、直前の文以外に関係付けられたものに比べて、一致率も高いことが観察できた。

さらに、構造解析実験の結果を観察したところ、テキスト全体に対する被験者の一致率は高くなかったものの、3人の被験者が全く別の関係先を選ぶことは少なかった（1テキストあたり1,2の文をのぞけば、多数決で関係先が特定できる）。また、関係距離1の関係は、この多数決で決定したテキスト構造上で無秩序に分布するのではなく、連続して平均3文が関係距離1で結びつけられ、まとまりを形成することがわかった。我々は、このまとまりを一種の談話セグメントではないかと考えている。具体的に先ほどの図1を用いて説明すると、文番号からなる関係距離1でまとめられた部分木(1,2,3),(4,5,6,7),(8,9)は談話セグメントと考えられ、ある程度人間の解析と一致しやすい。他方、それぞれの談話セグメントの先頭文が図1では4→1、8→4と長距離の関係付けが存在するが、この部分は人間によって解析にばらつきが出やすい。

表1: 距離による関係付けの傾向

関係距離	1人でも関係付けした数(A)	うち、2人以上で一致した数(B)	(B/A)
1	352	272	0.773
2	113	47	0.416
3	53	20	0.377
4	36	14	0.389
5以上	85	29	0.341

また、任意の隣接文間に対して関係距離1の関係付けがあるか否かを2値分類することは、計算機を用いてもある程度再現することが可能である。我々は機械学習を用いて、多数決で決定した木に現れる隣接関係の有無に対して形式段落の情報を用いなくても69.7%の正解率（交差検定を用いて評価）を得た。

本稿では、以降、このテキスト構造解析実験において多数決で決定したテキスト構造を正解として使用し、要約とテキスト構造の関係を検討する。

3 要約事例について

3.1 要約の作成

本研究では2人の被験者に文や節での重要部分抽出を前提としない要約を作成してもらった。この2人の被験者は2節で説明したテキスト構造解析実験を行った被験者とは別の被験者であり、区別のため本稿では要約作成者と呼ぶ。要約を作成してもらった記事は、テキスト構造解析に用いたものと同じ記事から20記事を選んだ。

要約は1つの記事に対して、元記事の文字数の最大40%程度に要約してもらったものである。要約の指示として、「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。なお、題名と形式段落は要約作成者が参照できないように元記事からこれらを取り除いて実験を行った。基礎的なデータとして、元記事および要約作成者A、Bの要約の長さを表2に示す。

表 2: 要約の長さに関する基礎データ

	元記事	要約	
		作成者 A	作成者 B
平均文字数	882.6	342.0	320.4
平均文数	16.1	7.15	6.85

要約率を 40%と比較的高率にした理由は、極端な言い換えを抑え、文節レベルで元テキストと要約の対応付けをとりたかったからである。実際、両要約作成者がほぼ指示通りに、テキストを要約率 40%で要約したことから、高度な言い換えと思われる例が少ない要約事例を作成することが出来た。この結果、固有名詞を保存する制約から元テキストのどの部分から要約が生成されているかを比較的容易に見極めることができた。

3.2 元テキストと要約の対応付け

Jing ら [8] や加藤ら [9] の研究では要約事例と要約元テキストの対から要約に用いられる知識を機械学習によって獲得する手法を提案している。我々の研究の最終的な目的も、それらの研究と同様にテキストを要約する操作を計算機により実現することである。そのため、要約作成は計算機によっても遂行可能な操作に細分できると仮定する。元テキストと要約との対応付けは、人間がテキストに対して行っている操作をこのような観点から分析する上での基本となる。

元テキストと要約を対応付ける上で、テキストを構成するどの単位を基準として対応付けを行うかを決めることが必要である。我々は、要約生成過程の最小単位を文節とし、要約生成を以下のような操作過程に分解したいと考えている。

- 文節の抽出→並べ替え→書き換え→要約文として再構成

つまり、文節に対して以上のような操作がどのように行われているかを検討することが本稿の実験の目的である。しかし、元テキスト中に現れる要素と要約文中に現れる要素を文節単位で対応付けるには文の意味を無視する点からも、対応付けに対する手間が増大する点からも良い分析手法とは考え難い。そこで本稿では、まず重要部分の抽出の過程として文を基準とした抽出を考え、元テキストと要約事例の対応付けを行う。対応付けの類型は以下の 2 つのタイプを想定し、その後に要約生成の過程として、抽出された文の中の文節がどのように要約に利用されるかを調査する。

- 短縮型：元テキストの 1 文内の文節だけを用いて要約中の 1 文にする。
- まとめあげ型：元テキストの 2 文以上にわたる文節を用いて要約中の 1 文にする。

先に述べたとおり、要約の処理過程は重要部分抽出と要約生成の過程を仮定することが多い。上の 2 つのタイプは、重要部分抽出の単位を文と仮定し、要約生成にいくつの元文を抽出したかによって分類したものに相当する。例えば、上の短縮型は重要部分抽出である 1 文を抽出しておき、要約生成の過程でその文のあまり重要でない文節を削除して、要約文を再構成する操作と考えることができる。

実際の元テキストと要約の対応付けは、計算機を用いておおまかな対応付けをとっておき、その対応付けを 2 人の対応付け作業者によって修正し、作業者の意見が分かれる部分については、計算機の対応付けを保留するという手法をとった。おおまかな対応付けには、テキスト中の内容語（名詞）の出現回数を要素とするベクトル空間モデル上のコサイン類似度を利用した。これらの名詞は形態素解析を用いて自動的に抽出した。すなわち、このベクトルと類似度を用い、要約文書中の各文に対し最も類似度の高い元テキスト中の文をおおまかな対応付けとした。

上の手順を踏むことで、要約作成者によって元文が完全に書き換えられてしまった要約文の場合には、その要約文で使われた名詞が最もよく出現している元テキスト中の文が対応付けられる。つまり、高度な書き換えなどの理由から対応付け作業者 2 人がその要約文に対してどの元文を要約に用いたかを同定できない場合、この計算機による対応付け結果が保留されることになる。なお、元テキストの 1 文が、上の 2 タイプのいずれか、もしくは両方のタイプで複数の要約事例中の文に対応付けられることは許している。

上記の方法で要約文と元テキストの複数の文との対応をとった後、要約中の各文を生成する際に、上の 2 タイプの操作を元テキストのどれだけの文に適用していたかを表 3 に示す。なお、要約された 20 記事全 322 文に対し、要約作成者 A が要約のために抽出した文は 188 文、要約被験者 B が 166 文である。表 3 を見ると、被験者によってどちらの型の要約操作を好むかについては差異が存在するものの、複数の元文をまとめあげて要約中の 1 文にする操作を相当数行っていることが判明した。つまり、人間の作った要約を観察

表 3: 各タイプに対する抽出文数

要約文生成の類型	要約作成者	
	A	B
短縮型	94	110
まとめあげ型	103	58

表 4: 両被験者が全く要約で使用しない文

テキスト構造	総数 (A)	要約で未使用	
		例数 (B)	(B/A)
関係距離 > 1	75	17	22.7%
関係距離 $= 1$	190	60	31.6%

する限り、テキストの重要部分同定は文を単位に個別に抽出し、それをそれぞれ短縮する操作だけでは要約生成の操作として不十分であることが分かる。

表 3 での表記は要約事例に対応する元文の数を示している。要約事例側から見た場合の文の数は、要約作成者 A の場合、元テキストから抽出された 103 文に存在する文節が再構成され要約中の 49 文にまとめあげられ、要約作成者 B の場合は元テキスト中の 58 文が要約中で 27 文にまとめあげられる。このため、要約事例中の文数で各操作タイプの事例数を比べると、どちらの要約作成者も元テキスト中の元文を短縮型の操作を行って要約を行っている事例が多いことになるが、上述した通り、まとめあげ操作で作成される要約文も決して少なくない。

4 要約生成条件の検討

以上までで説明した通り、テキスト構造と、そのテキストに対する要約事例とを用意し、対応付けを行った。本稿の残りでは、これらの道具立てを用いて、要約作成における、重要部分抽出と、要約生成のそれぞれの過程に対するテキスト構造の役割を検討する。

4.1 テキスト構造と重要文抽出過程

今回の実験で用いた要約は、ある程度の制約はあるものの、ほぼ自由要約と考えてよい。そのため、要約作成者間の重要部分選択に対してのゆれは非常に大きかった。しかし、以下の 2 つの点においては両要約作成者に共通して見られる比較的明確な傾向があった。

1. 要約で省かれ難い文の構造的位置
2. 文を連続して抽出すること

以下それぞれの傾向について説明する。

1) 2 人の要約作成者がどちらも要約中に用いなかった文を調べると、構造的に特徴がある位置の文が要約で省かれ難い傾向があった。特徴が見られたのは、多数決で決定した元テキストの構造表示において 1 よりも長い関係距離をもって他の文に関係する文である。具体的な値は表 4 に示す。表中の各テキストの構造位置の総数からは多数決で構造位置が決定できなかった文および構造木の根になる文は省いてある。この表はテキスト構造上で各関係距離を持つ文を 2 人の要約作成者どちらもが要約中で用いない割合を比較しているが、この結果から、関係距離 2 以上で他の文に関係する文は、関係距離 1 で他の文に関係する文に比べて、要約において省かれにくいことが分かる。

2) 抽出される文が単独に抽出されるだけではなく、その文に関連した文も同時に抽出し、要約を生成している傾向があった。我々の研究に関連する先行研究として、望主ら [10] らは、1 記事に対して 100 人分の要約事例を 3 記事分用意し、本研究と同様に文節レベルの対応付けを行う実験を行っている。その実験の結果から、元テキスト中に先に現れる構成要素は、要約中でも先に現れる順序付けの直線的な傾向が報告されている。本研究でも彼女らの研究同様、要約には順序付けの直線的な傾向がみられている。さらに、ある一文の文節が要約中で使用された際に、その文に連続する文が同時に抽出される傾向も見られた。具体数を示すと、要約作成者 A は平均 2.4 文連続して文を抽出し、要約作成者 B の場合も平均 2.1 文連続して文を抽出した。この特徴は、2 章で述べた抽象的な段落に相当する談話セグメントと相関があるように思われ、次節で述べる要約文生成過程における操作の影響が大きいと考えている。

4.2 テキスト構造と文生成過程

4.1 節では、重要部分抽出の過程とテキスト構造との関係を考察した。本節では文生成の過程とテキスト構造との関係を考察する。

要約事例と元テキスト中の文との対応付けをおこなった際に、元テキスト中の複数文がまとめあげられて要約中の 1 文になっている事例が相当数存在した。このまとめ上げ操作で用いられる文の集合は 4.1 節で述べたテキスト上で連続する文を抽出する傾向が特に強い。表 5 は、それぞれの要約中の事例数を、作成者がまとめあげに使った抽出文が隣接していたか否かによって整理したものである。括弧なしの数値で示し

表 5: まとめあげ元文の位置関係とテキスト構造上の位置

元文の位置	要約作成者 A	要約作成者 B
隣接文間	38(34)	19(18)
非隣接文間	11(5)	8(3)
計	49(39)	27(21)

たのは、単に元文が元テキスト上で隣接していたか否かだけの分類事例であるが、括弧内に示した数値はこの分類事例のうちテキスト構造上でも特徴的な位置にあった事例の数である。テキスト構造上で特徴的な位置とは、隣接文間では直接関係付けされていたものを考えた。さらに、非隣接文間において特徴的と考えたのは次の 2 タイプである。

A) まとめあげられる文同士がテキスト構造上で直接関係付けられている。

B) まとめあげられる文がテキスト構造上で同一の文に関係付けられている。

隣接文間における関係付けは関係距離が 1 である A のタイプの特別な場合と考えることができる。

表 5 を見ると、特に、隣接文間でまとめあげが起こっている場合、テキスト構造においても当該文間に関係距離 1 の関係付けがなされていることが多いことが確認できる。他方、非隣接文間のまとめあげでは、テキスト構造上の特徴的位置にあったものは全体の半数弱であった。非隣接文間の構造タイプの内訳は、要約作成者 A で A タイプが 5 例、要約作成者 B では、A タイプが 2 例、B タイプが 1 例、計 3 例であった。以上の結果から、元テキスト上の複数文がまとめあげられた場合、特にテキスト構造上で関係距離 1 の関係付けをもっていることが多いことが分かった。このことから、複数の文をまとめあげる上で、隣接文関係における関係付けはまとめあげの重要な手がかりであると考えられる。

4.3 まとめあげ操作とテキスト構造の役割

自動要約を行う上で、テキスト構造が文のまとめあげの手がかりとなると仮定したとしても、まとめあげ操作は単に 1 文を短縮する操作とは異なる文を連結する操作を行わなくてはならない。そこで本節では要約においてまとめあげる操作自体を分類することを試みた。

まず、ここでは便宜上、テキスト構造上で密接な関係をもつ文同士が何らかの形でまとめあげられ、その後に 1 文を短縮する方法と同じ方法で短縮されると仮定する。そのため、まず、1 文から不要な文節を削除して短縮化する操作について、節レベルに対する操作と文節レベルに対する操作に分け、それぞれ以下の操作を仮定した。

節レベル	削除	許可
	順序入替え	許可
	新規挿入	不許可
文節レベル	削除	許可
	順序の入替え	許可
	新規挿入	特定の種類のみ許可
	書換え	主観判断で許可
	それ以外	不許可

ここで、文節の新規挿入は、接続表現、副詞表現、呼応表現のみに限定した。また、文節の書き換えは基本的に主観判断したが、文節内の名詞、動詞といった自立語については一般的に書き換え可能とみなされるものと、同一記事内で明確にわかるもののみとした。例えば、「存在する」→「ある」や、同一記事中で「証券取引委員会（ＳＥＣ）」と説明されている場合の「ＳＥＣ」→「証券取引委員会」、代名詞の復元などを書き換えとして認めた。他方、助詞、助動詞、活用語尾などについてはゆるい制限で書き換えが可能であるとし、削除することも認めた。なお、今回は節レベルの操作の後に文節レベルの操作が行われると仮定した。

以上のような文短縮操作の前提のもとで、複数の元文に存在する節、または文節がどのような形態で 1 文にまとめあげられ、文短縮操作が行われたかを検討した。元文が隣接文間にあったか否かを含め、すべてのまとめあげ事例に対して連体化、主題化、複文化、その他の 4 種類の分類を行った。各分類の定義と事例を表 6 に示し、分類結果は表 7 に示す。表 7 の括弧内に示した数は、当該操作の事例のうち表 5 と同様にテキスト構造上で関係付けられていた事例数である。

この結果から、両要約作成者の指向はあるものの、どちらの要約作成者もまとめあげの際に、主題化、複文化の操作を数多く行うことが認められる。また、主題化、複文化でまとめあげられる文同士はテキスト構造上で関係付けられている場合が多いことがわかる。これに対し、連体化のまとめあげは、元文が隣接文間にある場合と非隣接文間にある場合の割合が要約 A では非隣接文間にあった事例の方が多く、要約 B でも同数である。また、要約事例から考えても、連体化操作はテキストに出現する内容語ごとにその語に対して言及のある文の集合から説明されるべきであり、テキスト構造上の関係付けが主題化、複文化で果たす役割とは別の説明付けが可能であると思われる。

さらに複文化の操作は、まとめあげ元文で提示され

表 7: 各まとあげタイプの分類数

要約	位置	連体化	主題化	複文化	その他
A	隣接	1	9(8)	28(26)	0
	非隣接	3(1)	1(1)	7(3)	0
B	隣接	1	7(7)	9(9)	2(2)
	非隣接	1	1	6(3)	0

ていた主題が、まとめあげられた文でどのように文法的役割が変化するかに基づいて分類することができる。例えば、表6の複文化の例で示した要約文は、まとめあげ元文の各主題がまとめあげられた文でも従属節の主題と主節の主題で表現されている形である。他方、以下のような複文化の例も多数存在する。

要約　百貨店は売上不振が続いているが、初年度4百三十億円の売上を見込み、店舗運営の効率化を徹底する。
元文1　百貨店は売上不振が続いているが、初年度四百三十億円の売上高を見込んでいる。
元文2　物流部門を外部委託する他、パート・アルバイトの比率を高めるなど、店舗運営の効率化を徹底する。

この例は、まとめあげ元文の一方の主題がまとめあげられた要約文でも主題となり、もう一方の元文にはそもそも主題も主格主語も存在しない例である。この特徴は、テキスト構造解析実験で紹介したテキスト構造上の関係付けが計算機で再現する規則の典型的なものである。この点から類推すると、表6の複文化の例でも要約元2文それぞれの主題もしくは主語の関連性に基づいてまとめあげが可能かどうかが決定されると考えられるが、この可能性の有無を構造上の関係付けが表現していると考えられる。また、主題化には表6の主題化の例とは別のタイプとして、例えば、上に示した複文化の事例を

百貨店は、店舗運営の効率化を徹底する。

とするタイプも存在するため、複文化操作の特別な形と考えることもできる。

以上をまとめると、テキスト構造上の関係付けは主に、関係付けられた文対が複文化可能かどうかを決定する手がかりとなっていることが分かる。さらに、今後まとめあげの自動処理を考える上で、関係付けの種類を詳細化することが必要であるが、本節の知見から、少なくとも談話セグメント内の関係付けの詳細化は、複文生成の制約を参考にすることができると考える。

5 まとめ

今回、我々は人間に自然な要約を作成してもらい、それらを用いてテキスト構造と自動要約がどのような相互関係にあるかを調査した。

まず、人手によるテキスト構造の解析では、談話セグメントを構成する連続した部分とその談話セグメント間の関係付けを行う部分で解析の難しさに相違があることが分かり、密接な関連付けが連続して続く談話セグメントについては被験者間で解析が一致することが分かった。

次にこの実験で得られたテキスト構造を用いて要約との関係を分析した。自動要約が2つの過程、すなわち重要部分抽出過程と要約生成過程に分割されると仮定し、それぞれの過程におけるテキスト構造の役割を考察した。

重要部分抽出過程では、談話セグメントの先頭文が要約事例で省き難いことが分かった。また、文抽出は1文ごとに行われるのではなく、隣接する文をまとめて抽出し要約生成することが分かった。この結果からテキスト構造の要約抽出は談話セグメントを単位に行われることが推定できる。

要約生成過程においては、特に複数の文を要約中で1文にまとめあげあげるタイプの要約生成操作ついては、テキスト構造が密接に関係することを確認できた。まとめあげ操作は、人間が談話セグメントにおいて複数の文で表現されている情報を1文で簡潔に表出する上で有益な操作であると考えられる。また、まとめあげ操作が行われた事例を分析したところ、特に主題化、複文化が可能な文対を同定する上でテキスト構造上の関係付けが有益であることがわかった。

今回の実験では、要約作成者は形式段落といった明示的なテキストの構造的な手がかりを知らずに要約を作成したにもかかわらず、このような抽象的なテキストの構造が複数の文のまとめあげと関係していることは興味深い。

要約はそれ自体がテキストであるため、要約においても首尾一貫性 (coherence) が保持されていなくてはならない。本研究で検討したまとめあげ操作は談話セグメント内の首尾一貫性を保持する上では役立つと考えられる。しかし、テキスト全体の首尾一貫性を保つには不十分で、談話セグメント間の関係的意味を考慮する必要がある。今回用いたテキスト構造のモデルでは、この部分に対して人間による解析のゆれが存在するが、重要部分抽出する際に、どの談話セグメントを選択するかを決定する上でも談話セグメント間の関係

表 6: まとめあげのタイプ分類

連体化：	同一対象への言及がある文がまとめられ、まとめあげられる文の１つがまとめあげられた文の中で対象の修飾を行う。
要約例	生産ラインを導入する野洲工場の半導体製造に使っている建屋を活用し、カラーフィルタ用のクリーンルームなどを設ける。
元文１	野洲工場に生産ラインを導入、４月から量産する。
元文２	野洲工場の半導体製造に使っている建屋を活用、カラーフィルター用のクリーンルームなどを設けた。
主題化：	まとめあげられる文中の１文の要素がまとめあげられた文の主題となる。
要約例	紛争処理案は紛争処理機構を設立し、国家間紛争の処理を調停にとどめる案と、拘束力を待たせる案を検討する。
元文１	協定の効果を確実にするのが紛争処理案だ。
元文２	紛争処理機構を設立したうえで、国家間紛争の処理を OECD での調停を含めた協議の場にとどめる案と、最終的な拘束力を持つ仲裁権限をもつ組織にする案を検討する。
複文化：	連体修飾節以外の形をとる複文化。便宜上、重文も含める。
要約例	冷戦時代インドは旧ソ連と親密な関係だったが、米国はパキスタンに軍事援助を進めていた。
元文１	冷戦時代を通じてインドは旧ソ連から武器の大半を調達するなど、親密な関係だった。
元文２	これに対して、米国はパキスタンに軍事援助を進めていた。
その他：	人間が高度な書き換えを行っていると考えられ、上の分類から除外したもの。
要約例	また、内国民待遇加盟国は、他の加盟国のサービス提供者に対して、自国産業と同等の待遇を与えるサービス分野を約束表に記載する。
元文１	内国民待遇加盟国は、約束表に記載したサービス分野について他の加盟国のサービス提供者に対し、内国民待遇を与える義務がある。
元文２	すなわち、その分野では他の加盟国のサービス提供者に対し、自国の同種のサービス提供者に与える待遇より不利でない待遇を与えることを盛り込んでいる。

付けを表現する枠組みを検討することが必要である。

謝辞

本研究の実験データとして新聞報道記事を使用させていただいた日本経済新聞社に心から感謝します。また、要約作成やテキスト構造解析実験に協力してくださった方々にこの場をお借りし、お礼申しあげます。

参考文献

[1] Kenji Ono, Kauzuo Sumita, and Seiji Miike. Abstract generation based on rhetorical structure extraction. In *COLING-94, Vol.1*, pp. 344–348, 1994.

[2] Daniel Marcu. To build text summaries of high quality, nuclearity is not sufficient. In *Intelligent Text Summarization *Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06*, pp. 1–8, 1998.

[3] W. C. Mann and S. A. Thompson: *Rhetorical Structure Theory: A Theory of Text Organization*, Technical Report ISI/RS-87-190, ISI Reprint Series, 1987.

[4] 難波英嗣、奥村学. 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 (自然言語処理研究会）, 99-NL-133,pp. 53-60, 1999.

[5] Inderjeet Mani, Eric Bloedorn, and Barbara Gates. Using cohesion and coherence models for text summarization. In *Intelligent Text Summarization *Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06*, pp. 60–67, 1998.

[6] 竹内和広、松本裕治. 自動要約を視野にいれたテキスト構造解析実験. 情報処理学会研究報告 (自然言語処理研究会）, 99-NL-133,pp. 61-68, 1999.

[7] J. D. Moore and M. E. Pollack: A Problem for RST: The Need for Multi-Level Discourse Analysis, *Computational Linguistics, Vol.18, No.4*, pp.537-544,1992.

[8] H.Jing and K.R.McKeown. The Decomposition of Human-Written Summary Sentences. *In Proceedings of SIGIR '99*, 1999.

[9] 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, 1999.

[10] 望主雅子他. 重要文と要約の際に基づく要約手法の調査 情報処理学会研究報告 (自然言語処理研究会）, 00-NL-135,pp. 95-102, 2000.