文章構造と要約事例の対応付けにもと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
E-mail: {kazuh-ta, matsu}@is.aist-nara.ac.jp

1 概要

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が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を用いて、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。

2 はじめに

自動要約における数多くの関連研究では、大きくわけて、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、それら抽出した文や節を再構築し出力する要約生成過程との2つの処理過程を仮定する。

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をする試みを行った。具体的には、節や文に相当するテキストの部分が、解析されたテキストの構造木のどの位置に相当するかに基づいて、重要度の点数付けをを行うものである。その際の点数付けとしてどのようなものが適当であるかが完全に明らかになっているわけではないが、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は違い、部分要素の重要度を文書の構造を考慮にいれた上で決定できる可能性があるため、この手法に対しての期待は高い。

これらの方法を含め重要文や重要節の抽出は、自動要約研究において自動要約を実現する技術としてある程度の成功を収めている。また、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。しかし、重要な文や節のみを抽出する手法には、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再構成してやる必要性などの問題が指摘されている。このような処理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情報を、重要部分抽出とは別系統に表現しておく必要が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な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はないかという予想をたてている。

表 1: 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	860.8	334.6	316.2
平均文数	15.8	7.3	7.0

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動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、テキスト構造が要約にどのような影響を与えるかを調査した。

3 要約事例について

3.1 要約の作成

本研究では2人の被験者に要約を作成してもらった。この2人の被験者は4章で説明する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。実験に用いる記事は、新聞報道記事であり、あらかじめテキスト構造解析がなされた記事の中から10記事を選んだ。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。

要約は1つの記事に対して、長さが異なる、長短2つの要約を作ってもらった。長い要約は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、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。これに対し短い要約は、元記事と長い要約を見ながら、長い要約を文字数で約半分程度の長さに要約してもらったものである。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。

今回は、元テキストから直接生成された長い要約について行った調査を報告する。このため、本稿では全体の最大40%の文字数に要約された長い要約を単に要約と呼び、元記事および要約作成者A、Bの要約の長さを表1に示す。[1]

3.2 元テキストと要約の対応付け

元テキストと要約がどのように対応付けされるかについてはいくつかの類型があると考えられるが、本研究では以下の2つのタイプにまとめた。

1. 元テキストをそのまま、もしくは短くして要約中の1文にする。
2. 元テキストの2つ以上の文を要約中の1文にする。

上の2つのタイプは、要約事例中の文を基準に考えており、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。

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、ベクトル空間モデル上のコサイン類似度を利用した。具体的には、元テキスト中の文と要約文書中の文を、それぞれの中で使われている名詞を内容語とみなして、その出現数を要素とするベクトルで表現しておき、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対応付けるようにした。ベクトルの要素のもととなる名詞は形態素解析を用いて自動的に抽出した。この結果、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、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。現在、この対応付けの自動化を検討中である。

[1]長い要約から短い要約への書き換えの性質については現在調査中であるため、その結果については別の機会に報告する。

表2: 対応付けの結果

要約文生成の類型	要約作成者	
	A	B
1つの元文を1文にした	41	54
複数の元文を1文にした	32	16

表3: 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

要約中の各文が元テキストの文からどのように生成されたかを、上記の方法で対応付けした結果を表2に示す。表2を見ると、両要約作成者ともに元テキスト中の元文を短い文にする形で要約を行っている事例の多いことがわかる。他方、特に要約作成者Aで複数の元文をまとめあげて要約中の1文にしている例も相当数あることが判明した。さらに、表2での表記は要約事例中の元文からまとめあげられた文の数で評価しているが、元テキスト側から見れば、元テキスト中の複数の文から要約中の1文へまとめあげられるため、元テキスト中から抽出される文の数は多い。実際に、要約作成者Aの要約中の32文は元テキスト中の65文がまとめあげられたものであり、要約作成者Bの16文は元テキスト中の34文がまとめあげられたものである。

4 要約生成条件の検討

以降、3章で得られた要約と元テキストの対応付けの結果を元に、テキストの構造と要約との間の関係を考察する。

4.1 構造解析実験

我々は、本研究に先行して、日本語の新聞報道記事32記事、合計500文に対して複数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]。その際、

表 4: まとめあげ元文の位置関係とテキスト構造上の位置タイプとの一致

元文の位置	要約作成者 A	要約作成者 B
隣接文間	24(23)	12(11)
非隣接文間	8(7)	4(2)
計	32	16

テキストの構造モデルとして用いたのは、文と文との関連性と、テキスト全体における重要性の相違という 2 つの観点から RST を簡略化したものである。具体的には、テキスト中の任意の文に対して、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）関係を関係付ける。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こと、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。

以上のような構造解析実験において、被験者がある文について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 3 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、多数決を用いれば構造が決定できること（1 テキストあたり 1,2 の文をのぞけば、関係先が特定できる）がわかっている。また、関係距離 1 の関係はこの多数決のテキスト構造上で単独で定義されるのではなく、平均して 2.0 回連続して定義される傾向が分かっており、文のまとまりを形成するように見える。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。本研究では、この多数決で決定できたテキスト構造を用いて、テキストの構造と要約の関係を調査する。

4.2 テキスト構造と要約生成条件の検討

3 章において、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。このまとめ上げの傾向に、どのような構造的な要因が関係しているかを調べるため、まず、まとめあげられる文が元テキストでどのような位置関係にあるかを調査した。すると、まとめあげられる文は元テキストの中で多くの場合隣接していることが分かった。表 4 に、それぞれの要約作成者がまとめあげに使った文が隣接していたか否かによって整理した事例数を括弧なしの数値で示す。

さらに、このまとめあげられる文が多数決で得られたテキスト構造においてどのような特徴があるかを調査した。表 4 で括弧付きの数値で示したのは、元テキスト中でそれぞれの位置にあった複数の文が、多数決で得られたテキスト構造においても、特徴のある構造位置にあった事例の数である。その際、テキスト構造上において、特徴がある構造位置とみなしたものは次の 3 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。C) まとめあげられる文がテキスト構造上で関係距離 1 の連続するまとまりの中にある。これらのタイプ分けを、まとめあげられる文を斜線の長方形で示し、テキスト構造上の関係付けを矢印で示したものを図 1 に示す。

表 4 において、隣接した文がまとめあげられた場合には、テキスト構造上で意味をもつのは、図 1 の A タイプである直接関係付けられる場合のみである。この表から、元テキストの隣接文の間でまとめあげが起こっている場合、ほとんどの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確認できる。

また、元テキスト中で隣接しない文同士がまとめあげられた場合には、図 1 のすべてタイプの構造位置にあるものが該当する。それぞれの内訳は、要約作成者 A では、A タイプが 5 例、B タイプが 2 例であった。同様に要約作成者 B の 2 事例の内訳は、A タイプが 1 例、C タイプが 1 例であった。

以上の結果から考察すると、要約で元テキスト中の文がまとめられる場合、文が元テキスト上で隣接しない位置関係にある場合でも、テキスト構造上でなんらかの構造位置関係にあることが、まとめあげの条件となっているように思われる。

他方、要約の中で、どのような形の文にまとめあげが行われるかについても、分類を行った。事例数が少ないため、文が隣接文間にあったか否かによる明確な差異は認められなかった。そのため、すべてのまとめあげ事例に対して分類を行った。分類は、まとめあげられる文の中の 1 つで述べられている要素が主題化されて、要約中の一文になるもの。まとめあげられる文の中の 1 つで述べられている要素が連体修飾節にな

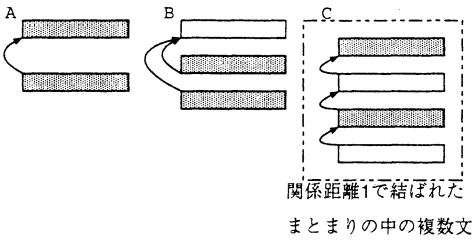

図 1: テキスト構造上の構造位置のタイプ分類

表 5: まとめあげのタイプ分類

	主題化	連体化	複文化	その他
作業者 A	4	2	25	1
作業者 B	7	0	7	2

り、まとめあげられるもの。上の 2 つ以外の形で複文化や重文化されてまとめあげられるもの。そして、対応付け作業者は 2 人一致でまとめあげられた文としたが、言い替えにより元文中の当該部分を明確に同定できないものの 4 つに分類した。その分類結果を表 5 に示す。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、複数の文間の関係を見出して、主に連体修飾以外の複文化をよく行うことが認められる。また、複数の文の 1 つを主題化させ、まとめあげる方法も相当数行うことが分かった。本研究で用いたテキスト構造は RST を自動要約への応用を視座に単純にしたものであるため、2 文間に関係付けがなされていることが、その 2 文間の結束性の高さや詳細な修辞関係を特定することに直接にはつながらないが、このようなまとめあげの分類は、どのようなテキスト構造上の関係が要約への応用上有益かを探る手がかりとなると考えている。

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。現在のところ、テキスト構造と重要情報選択の間の関係については報告するほどのまとまった結果を得ていないが、要約生成の過程において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については、テキスト構造が密接に関係することを確認できた。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。

今後の課題としては、近年盛んに研究されるようになっている文の部分書き換え規則を文のまとめあげとテキストの構造という視点から整理することを考えている。そのためには、文間にテキスト構造の視点から密接な関係が認められる際の関係を自動要約に焦点を合わせて類型分類して詳細化することが必要である。また、統計的手法に基づいて、これらのまとめあげを自動的に行う規則を獲得するためには、要約と元テキストとを対応付ける方法を確立することが必要である。

謝辞

本実験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.