短文分割の自動要約への効果

福島 孝博[†] 江原 暉将^{††} 白井 克彦^{†††}

TV ニュース原稿は、新聞記事に比べて1記事中の文数が少なく、1文当たりの文字数も多い。このため、自動要約としての重要文抽出を行うと、文単位で選択が行われる為、情報の欠落が大きい。本論文では、記事中に現れる長文を分割出来る条件を設定し、条件に合う場合は、短い文に分割するという処理（短文分割処理）を行った結果が自動要約の基本的技術にどれだけ影響・効果があるのかを調べた。短文分割は、基本的に、動詞、形容動詞と述語名詞の連用文節を分割の対象とした。また、分割の自動要約に対する影響については、評価の尺度として、各文の重要度による順位付けと文字数圧縮（不要部分削除）を用いた。文順位付けの評価では、テキスト中の各文を人手及びシステムによって、その重要度に応じて順位を付けたものを対象とした。人手により重要と判断された文が、短文分割により分割された場合に、その分割された文は、どのような順位となると判断されるのかを調べた。その結果、短文分割により分割された重要文は、分割後の順位差において「3」以上離れる場合のほうが、順位差が生じない場合、つまり順位差が「1」の場合より多くあり、短文分割の効果が見られた。次に、記事中の重要文だけではなく全部の文を対象として、人手とシステムによる順位付けについて短文分割前後での変化をスペアマンの順位相関関係係数を用いて比較した。その結果、短文分割をすることにより、スペアマンの係数が0.0318～0.065増加し、文の順位が、人とシステムにおいてより近いものになることが判明した。最後に、文字数圧縮での評価では、不要部分を特定し、文字列を削除または言い換えを行う文字数圧縮処理において、短文分割を行う前後での変化を調査した。短文分割により削除される文字数は増え、文字数圧縮後の文字数を元記事の文字数で割る圧縮率において、2.71%～2.78%減少することが判明し、短文分割が文字数圧縮に良い効果があることが分かった。

キーワード: 短文分割, 自動要約, 文字数削減

Partitioning long sentences for text summarization

TAKAHIRO FUKUSHIMA[†], TERUMASA EHARA^{††} and KATSUHIKO SHIRAI^{†††}

It is known that there are fewer sentences and the sentences are longer in a TV news text compared with those in a newspaper article. If we would like to summarize such TV news texts by selecting important sentences, since each sentence is rather long, we end up with losing a good amount of information by omitting a whole sentence. Therefore, we adopt a method in which we partition a long sentence into shorter sentences before summarization. To evaluate how the partitioning affects text summarization, we select two basic measures for text summarization, and examine how they vary before and after the partitioning of long sentences. The two measures are first ranking of sentences in the text by their importance, and second, the number of characters removed from the text by applying the same set of rules for shortening

and deleting the text. All the sentences in the text are ranked by their importance by hand and by sentence extraction system. First, we examine how the ranks of the sentences judged important by human vary before and after the partitioning. We found that there are more partitioned important sentences whose difference in ranking is greater or equal to three (3) than those whose difference in ranking is one (1). This suggests that the partitioning is good for sentence extraction. Then, we compare the rankings of the human and the system for all the sentences in the text using Spearman's rank correlation coefficient. We found that the coefficient increases between 0.0318 and 0.065, which means the rankings of the human and the system for the partitioned texts are more similar than those for the original texts. Lastly, we investigate how the partitioning affects the shortening the text. Here we found that the number of characters that are deleted increases for the partitioned texts and a compaction ratio (the number of characters of the shortened text divided by the number of characters in the original text) decreases by 2.71 percent to 2.78 percent. It shows that the partitioning long sentences makes shortening method work better.

KeyWords: *Partitioning sentences, text summarization, shortening text*

1 はじめに

近年テキストを自動的に要約する技術に関する研究が国内外で盛んになって来ている (Hovy and Marcu 1998; 奥村, 難波 1998). 自動要約に関する研究の歴史は, 古く 1950 年代後半から研究されているが, 対象のテキストから重要な部分を抜き出して要約とする重要部分特定の手法が中心であった. テキストの内容を理解しての自動要約は, 難しくまだ現実的なシステムを作成するに至っていないのが現状である. また, 最近, テキストの重要部分に注目するのではなく, 不要部分を特定し, 言い換え及び削除により, 要約を行う研究も出てきている.

本研究の目的は, 長い文を短い文に分割する処理 (今後「短文分割」と呼ぶ) を行ない, その短文分割の自動要約手法への影響を調査することである.

短文分割に関する研究は, 機械翻訳の研究で見られる. 機械翻訳においては, 長文は, 文の係り受け構造の複雑さが増え, 翻訳精度低下の原因の一つとされている. このため, 短文分割を翻訳の前処理として行い, 翻訳の精度を高めることを目的とした研究が行われている. 金ら (金, 江原 1993) は, 長文に現れる連用中止表現, 引用, 連体節, 接続節などを分割点と認定している. また, 木村ら (木村, 野村, 平川 1993) は, 単語数の多い文で, 特定の言語表現を持つ場合に分割を行っている. 特定の言語表現としては, 動詞と助動詞の連用中止表現, 接続助詞の「ので」などがある. 本論文で用いる短文分割手法は, これらの手法と同様のものである.

† 〒151-0064 東京都渋谷区上原 3-2 3-5 オーガストハウス 2 F, 通信・放送機構 渋谷上原リサーチセンター, Telecommunications Advancement Organization (TAO) of Japan., August House 2F, 3-23-5 Uehara, Shibuya-Ku, Tokyo 151-0064, Japan

†† 〒157-8510 東京都世田谷区砧 1-10-11, NHK放送技術研究所, NHK Science and Technical Research Lab., 1-10-11, Kinuta, Setagaya, Tokyo, 157-8510, Japan

††† 〒169-8555 東京都新宿区大久保 3-4-1, 早稲田大学理工学部情報学科, Waseda University, Department of Information and Computer Science., 3-4-1 Ookubo, Shinjuku-ku, Tokyo 169-8555, Japan

一方、関連研究としては Marcu の研究がある (Marcu 1997) . Marcu は、手がかりとなる語句を使って要約の基本となる単位を決定し、談話構造を解析してその結果を自動要約に用いている。手がかり語としては、because などの接続詞などを使っているため、節が基本単位となる場合があり、文より小さな単位を用いての要約を実現している。手がかり語がない場合は、文全体が1つの単位となる。Marcu の手法は、文より小さな単位を扱っているが、短文分割は行っていない。

もう1つの関連研究として、簡易な文構造解析を行い、自動要約に役立てるものがある。構文解析の結果を利用して重要・不要部分を特定し、要約を作成するものである。日本語では三上らが、TV ニュース原稿を題材として、構文解析を行い、文中の要素に重要度を与えて、重要要素や、削除すると文を壊してしまう恐れのある要素を重要として抽出している (三上, 山崎, 増山, 中川 1998) . 英語では、Grefestette の研究があり、自身の開発した parser を使い、構文解析を行い、主節は従属節より重要であり、否定の表現も重要であるなどとして、重要部分を特定して文書の単純化を行っている (Grefenstette 1998) . これらの研究は簡易構文解析処理を行っているのに対し、本研究では、構文解析を行わずに、文字列や品詞の情報のみを利用して短文分割を行っている。

上記の短文分割に関係した研究と比べて、本研究は、短文分割の手法は既存の手法と同様のものを用いており、その短文分割が自動要約の基本的手法にどれだけ効果があるのかに焦点を置いている。

本研究は、聴覚障害者向けにサービスしようとしている字幕付きテレビニュースでの自動要約技術に関する研究の一環であり、自動的にテレビニュース原稿を要約する手法について、重要文抽出、文字数圧縮などをテーマに研究を進めて来ている (Wakao, Ehara, Sawamura, Abe, Shirai 1997 ; Wakao, Ehara, Shirai 1998) .

本稿で題材としているのは、TV ニュース番組の電子化原稿である。ニュース番組原稿は、新聞記事と似ているが、両者を比較した場合、ニュース原稿のほうが1記事中の文数が少なく、且つ一文当たりの文字数が多いことが分かっている (江原, 沢村, 若尾, 阿部, 白井 1997) . ここで重要文を自動的に抽出することにより要約を作成すると、文数が少なく、一文が長いため、どうしても粗い要約となってしまう。この欠点を補正するために、短文分割を行い、その自動要約における基本的技術への効果を評価した。

評価には、文の重要度における順位付けと文字数圧縮を取り上げた。文の順位付けでの評価では、まず、各文を人手及びシステムによりその文の重要度に応じて順位付けを行い、人手により重要と判定された文が、短文分割により分割された場合、分割された文の順位がどうなるかを調査した。次に、記事中の重要な文だけではなく、全部の文を対象として、文の順位付けにおける短文分割の自動要約への影響を調べるために、人手とシステムにより順位付けされた結果の類似度を算出し、短文分割の前後での変化を調べた。この類似度には、スペアマンの順

位相関係係数を用いた。

また、文の不要部分を特定して、それを短い表現への言い換えや、削除により、文字数を削減する「文字数圧縮」においても短文分割の前後での圧縮率の違いを算出することにより、短文分割の効果を評価した。

以下に、まず、本研究の対象とした原稿を紹介し、短文分割の条件、短文分割の自動要約の基本的技術への影響について記述する。

2 原稿

題材とした原稿は、NHK 放送データベースの1992年の記事より選ばれた200件のテレビニュース番組の電子化原稿である。これは、NHKにおいて実際に放送されたTVニュースの原稿であり、放送日などの情報がヘッダーとして付けられたデータベースである。記事の大きさは、1記事当たり約500文字であり、1つの記事当たりの文数は、約5文である。200記事平均の1記事の文字数、文数、1文当たりの文字数は表1の通りである。

1記事当たり	
平均 文字数	484.68文字
平均 文数	5.18文
1文当たり	
平均 文字数	93.57文字

表1 対象記事（200記事）の詳細

前節でも述べたが、テレビニュース番組原稿は、新聞記事と比べると、一記事当たりの文数が少なく、一文当たりの文字数が多いという特徴がある。

3 短文分割

短文分割の処理は、条件を満たす長い文を記事中から選び、その文を複数の短い文に分割するものである。分割の条件は、NHK放送データベース（15万件）のテレビニュース原稿のうち、1991年4月1日から1991年6月11日にかけての記事（3,492記事）から、長さが200バイト（100文字）以上の文を無作為に抽出したコーパス（全部で9,205文）を使用し、それらを分析して分割条件を人手で作成したものである。

一方、後出の評価用のテキストは同じ15万件の記事から無作為に選ばれているが、1992年の記事から無作為に選ばれている。分割条件を導き出した記事とは重っていない。

以下に、分割を行う条件を記述し、その後に、分割した結果の詳細を示す。

3.1 分割条件

文の分割を行わない場合をまず記述し、その後に分割を行う条件の例を以下に記述する。分割をする条件は、下記記載のもの以外にも10規則あり、合計で、19規則ある。また、これらの規則は、適用されて生成される文末の時制は、原文の時制を反映した形となる。記載の例は、全て現在形で示してある。

1) 分割をしない場合

- 基本的に、動詞と形容動詞と述語名詞の連用文節と終止文節以外は分割しない。
また、この条件を満たす場合でも、以下の場合には、分割を行わない。
 - 連用文節であっても、連体文節直後の連用文節は分割しない。
 - 直後または2文節以内に連体文節がある連用文節は分割しない。
 - 直後または2文節以内に連用文節がある連用文節は分割しない。
- 用言自立語＋(助動詞)＋接続詞「ば」は、分割しない。例「働きかければ、」
- 「…も(と)あって、」は、分割しない。
- 自立語＋と＋なり＋読点 または、自立語＋に＋なり＋読点は、分割しない。

2) 分割をする場合

- 自立語＋(助動詞)＋「ており」＋読点
自立語＋(助動詞)＋「でおり」＋読点
例「働きかけており、b b¹」→「働きかけております。そして、b b」
- 自立語＋(助動詞)＋「が」＋読点
例「働きかけますが、b b」→「働きかけます。しかし、b b」
- 自立語＋(助動詞)＋「もので」＋読点
例「働きかけるもので、b b」→「働きかけるものです。そして、b b」
- 自立語＋(助動詞)＋「ものの」＋読点
例「働きかけるものの、b b」→「働きかけます。しかし、b b」
- 自立語＋(助動詞)＋「のに対し」＋読点
例「働きかけるのに対し、b b」→「働きかけます。それに対し、b b」
- 自立語＋(助動詞)＋「のに対して」＋読点
例「働きかけるのに対して、b b」→「働きかけます。それに対して、b b」
- 自立語＋(助動詞)＋「にもかかわらず」＋読点
例「働きかけるにもかかわらず、b b」→
「働きかけます。それにもかかわらず、b b」

1 「bb」は、文の一部であり、分割条件を満たす語句の直後に来る部分を示す。

- 自立語 + (助動詞) + 「とともに」 + 読点

自立語 + (助動詞) + 「と共に」 + 読点

例「働きかけるとともに, b b」→「働きかけます. それとともに, b b」

- 「あり」または「強まり」 + 読点

例「強まり, b b」→「強まります. そして, b b」

これらの条件を実際の TV 番組原稿文に適用してみると,

「千葉市に本店がある京葉銀行の成田西支店の女子行員が, 他人名義のカードローンを悪用しておよそ三億円を着服していた疑いが強まり, 京葉銀行ではきょう, この女子行員を懲戒解雇するとともに, 千葉県警察本部に被害を届け出ました。」は, 3 文に分割され,

1. 「京葉銀行の成田西支店の女子行員が, 他人名義のカードローンを悪用しておよそ三億円を着服していた疑いが強まりました。」
2. 「そして, 京葉銀行ではきょう, この女子行員を懲戒解雇しました。」
3. 「それとともに, 千葉県警察本部に被害を届け出ました。」

となる。

上記の条件の下で, 分割は, 対象となる文の長さ (文の長さは文節数により算出) により以下の 2 つの場合に行うこととした。

分割 1	分割前に文が 1 2 文節以上あり, 分割後 5 文節以上の文に分割される場合.
分割 2	分割する文の文節数に制約をかけない場合. つまり, これは, 分割出来る文は, 全て分割しようという場合に相当する.

3.2 分割結果

分割前と分割後の 2 つのケースを 1 記事当たりの文字数, 文数, そして 1 文当たりの文字数で比較してみると表 2 のようになる。

200 記事	文字数/記事	文数/記事	文字数/文
分割前	484.68	5.18	93.57
分割後 (分割 1)	491.87	6.45	76.20
分割後 (分割 2)	493.39	6.75	73.09

表 2 文数, 文字数でみた短文分割の結果

短文分割処理を行った結果, 一記事当たりの文数が約 5 文から約 7 文に増え, 1 文当たりの文字数は, 約 20 文字減ることが確認された。

4 自動要約への影響

短文分割の自動要約への影響を調べるために、まず、重要度に応じて文を並べる、文の順序付けにおける効果を調べた。まず、記事中の重要と判定される文が分割された場合に分割された複数の文の重要度がどう判定されるかを調べた。これは、分割前に重要であると判断された文が、短文分割により分割され、且つ、その分割された文が、重要度の判定において、一方が重要でありもう一方が重要でないといった差が認められる場合、短文分割の要約への効果があると考えられるからである。

また、記事中の重要文だけではなく記事全体として、短文分割が文の順位付けにおいて自動要約の結果にどのような影響があるかを調べた。これには、人手による順位付けの結果と自動要約システムによる順位付けの結果の類似度を、スペアマンの相関関係係数を用いて算出することにより行った。

文順位付けでの評価の対象としたのは、100記事で、短文分割処理で対象とした200記事のTVニュース番組原稿から少なくとも4文以上を含んでいる記事を100選んだ。記事の分野の特定などは一切していない。

次に、文中の不要部分を削除または言い換えて文字数を削減する（文字数圧縮）を取り上げ、短文分割の影響を調べた。この調査の対象とした記事は、短文分割処理で対象とした同じ200記事である。

4.1 分割前後での文順位の変化

対象となる100記事に対して各記事毎に、その記事中の各文の重要度に応じて順位をつけた。順位付けは、まず人手で順位付けを行った。それとは別に、自前の重要文抽出のシステム(若尾, 江原, 白井 1998)を用いて同じ記事の各文に順位を付けた。100記事中、分割1で分割されたものは、75記事あり、分割2で分割が行われたものは28記事である。分割2では、分割された記事の総数が、100記事中85記事で、分割1で分割されたものを再分割したものが18記事、新たに分割されたものが10で、合計85記事が分割されている。人手及びシステムにより文の順位が付与されたのは、分割前の100記事、分割1の結果(75記事)を含む100記事、分割2の結果(85記事)を含む100記事である。

システムで用いられた重要文抽出の手法は、重要語密度法であり、各記事において頻度2以上の自立語を重要語とし、各文の重要語の割合を計算し、それをその文の重要度とした(Luhn 1957; Edmundson 1969)。TVニュース番組原稿は、新聞記事と似ているが、1段落しかなく、見出しも存在しないという前提²であるので、記事中の位置情報(例、第一段落の第一文など)は用いず、単純に頻度の高い自立語を重要語とした重要語密度法を採用した。文の順位付けは、

2 データの出典元であるNHK放送データベースには、見出しが存在するが、ここでは用いなかった。

この重要語密度法で計算された重要度を用いて行われた。

重要とされた文の分割前後での順位の変化の調査には、分割1の結果を使った。分割前の記事中で人手による判定で重要度による順位が1位または2位と判断された文のうち、その文が短文分割を行うことにより分割され、且つ、2文に分割された文、総数30文（分割後60文）について調査をした。

まず、重要度において1位と判定された文での分割後の2文が、人手によってどのような順位に判定されたかを調べた。結果を図1に示す。

図1 1位文の分割後2文の順位分布

同じく、重要度で2位と判断された文での分割後の文の順位の分布は、図2のようになる。

図2 2位文の分割後2文の順位分布

これらの結果から、分割前の重要度1位文においては、分割後の順位が、1位と2位とが度数が最も多くなるのではなく、1位、3位、2位、4位以降という順になっている。また、2位の文では、度数順に見ると、2位、3及び6位、5及び7位となり、重要度による順位がばらつくことが分かる。

1位と2位の文をまとめて、分割後の文の順位における差の分布を、次に調べた。例えば、分割前に1位の文で、分割後に1位と3位に判定されたのであれば、差は3マイナス1の2となる。この差の分布を30文について調査した。

結果は、図3の通りである。

図3 分割後2文の順位差の分布

分割後の差が「1」の場合は順位が1位と2位などとなっている場合であり、短文分割の効果が見られないと考えられる。他方、順位における差が「2」以上の場合は、分割された文の間に他の文が入って来る場合であり、短文分割の効果があると考えられる。順位の差が「2」の場合は、人手による判定で全体の33% (10/30) を占める。また、順位差が「3」以上の場合は、全体の37% (11/30) を占め、順位差が「1」の場合 (9/30) より多くなっている。このことから、分割を行うことにより、分割された文の順位に差が生じる場合が、生じない場合 (順位差「1」の場合) よりも多くあり、短文分割は、重要度による文順位付けに効果があると言える。

4.2 記事全体での文の順序付けの評価

次に、前節での順位付けの結果を用いて、記事中の重要文だけではなく記事全体として、短文分割が文の順位付けにおいて自動要約の結果にどのような影響があるかを調べた。これは、人手による順位付けと自動要約システムによる順位付けの結果の類似度を、スピアマンの相関係数 (Spearman's rank correlation coefficient) を用いて算出することにより行った。スピアマンの順位相関係数は、2つの順位付けのなされたものの間の相関を計るのに一般的に用いられる統計尺度である (大村 1980)。

スピアマン順位相関係数 (r) の計算は以下のようなものである。 $X = (x_1, x_2, x_3 \dots x_n)$ $Y = (y_1, y_2, y_3, \dots y_n)$, $x_1, \dots x_n, y_1, \dots y_n$ は各文の順位を示し、記事の第1文から第 n 文までの順位を並べたものである X, Y が与えられた時、 r は以下の式で計算される。

$$r = 1 - \frac{6 \times \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

$$d_i : x_i - y_i \text{ for } 1 \leq i \leq n$$

d_i : x_i と y_i での順位における差

n : 記事中の文数

スピアマンの順位相関係数 r は、1 から -1 の間の数値となり、1 であれば2つの順位が完全に一致している場合であり、-1 だと全く逆の順位の並びとなっている場合である。0 の付近だと順位間の相関はなくなることになる。

比較を行ったのは、分割前の記事を用いての人とシステムの文順序付けの結果、分割後の記事を用いた人とシステムの文順序付けの結果である。分割後の記事としては、分割がなされた記事ばかりを集めたもの、つまり、分割1の場合だと75記事、分割2の場合だと28記事を対象としたケースと、それらの分割結果を含む100記事全体を対象とした場合について順位相関係数値を算出した。表3にその結果を示す。

比較対象記事	順位相関係数値
分割前 (100記事)	0.4933
分割1 (75記事)	0.5311
分割1 (100記事)	0.5304
分割2 (28記事)	0.5583
分割2 (100記事)	0.5251

表3 短文分割の文順位付けへの影響 (スピアマン相関係数)

尚, 参考までに, 人間の評価者間での順位相関を調べた。この調査に用いた原稿は, 上記の記事とは違う TV ニュース原稿記事 100 記事を用いた。2 名の人間の評価者間の順位相関係数値は 0.7427 であった。

短文分割後は, 分割前と比べて順位相関係数値が若干ではあるが良くなることが判明した。数値が良くなる原因として以下のことが考えられる。

- 短文分割の結果, 長文の一部であり内容情報を持たない部分が別の文として分割され, 文の順位付けにおいて, 人間にも, システムにも順位が低いものと判断されて, 記事全体の順位付けでより近いものとなった。

例えば, 原文が

「この女子行員は昭和六十年四月に京葉銀行に入社し, 平成元年の四月から成田西支店の貸し付け係, 平成三年十月から成田西支店の出納係を担当しており, ふだんの勤務態度は真面目で上司の信頼もあつかったということです。」

に対して, 分割後は, 2 文になり,

1. 「この女子行員は昭和六十年四月に京葉銀行に入社し, 平成元年の四月から成田西支店の貸し付け係, 平成三年十月から成田西支店の出納係を担当していました。」
2. 「そして, ふだんの勤務態度は真面目で上司の信頼もあつかったということです。」

となる。分割前のこの文の順位は, 人の判定で, 7 文中 4 位であるが, 分割後の第二文目は, 人の判定及びシステムの判定において順位が 11 文中 10 位と低くなっている。

また, 人手で判断された順位を上から 70% 程度までを見て, それらがシステムの判断した上位 70% と一致しているかも見てみた。分割後の対象としたのは, 分割 1, 分割 2 の双方とも, 100 記事全体である。順位の近さを計るのに今回もスピアマンの順位相関係数を用いた。その結果は以下の通りである。³

比較対象記事 (上位 70%)	順位相関係数値
分割前 (100 記事)	0.5051
分割 1 (100 記事)	0.5056
分割 2 (100 記事)	0.5063

表 4 短文分割の文順序付けへの影響 (上位 7 割での順位相関係数値)

上位 7 割を比較したのは, 現在数少ない字幕付き TV ニュース番組である NHK 教育チャンネルの「手話ニュース 845」において, 原稿に対して付与される字幕は, 原稿の要約となっているが, 文字数において約 7 割程度になっていることを考慮したものである (若尾, 江原, 白井 1997)。

³ 尚, 前述の 2 名の評価者間での上位 7 割における順位相関係数値は, 0.8246 である。

上位7割の文での比較では、分割前、分割後ともほぼ同じ順位相関係数値を残した。これは、上にみたように、分割後の場合、人とシステムで順位の低い文が一致しているケースが多く、その結果、上位7割だけでの比較を行うと、分割前の相関係数値が多少上がり、それに対して分割後の2つの係数値は、多少下がることになり、全体として余り差のない結果となっている。短文分割の順位付けへの影響を総合すると図4のようになる。

図4 文順位付けでの短文分割の影響 (スピアマン相関係数)

4.3 文字数圧縮

記事中の文中の不要な部分の削除や、短い表現へ言い換えることによって文字数を減らす(文字数圧縮)について、短文分割のもたらす影響を調べた。

文字数圧縮規則を、付録に添付した。基本的には、文中、特に、文頭、文末の表現に注目して、文字の削除、または言い換えを行うものである。例えば、「強調しました。」と言う文末であれば、「強調。」とする、「総理大臣」は「首相」とするなどである。

文字数圧縮規則は、聴覚障害者向けであり、字幕付きニュース番組である「手話ニュース845」(NHK教育)の字幕で使われている要約の規則を分析し得られた規則を基礎とし、NHK放送データベース中のニュース番組の原稿を約300記事⁴用いて、更に修正、改良したものである(若尾、江原、白井1998)。

これらの規則を用いて、短文分割を行う前と後での、圧縮率の変化を調べた。この場合、圧縮率は以下の式で算出した。

$$\text{圧縮率} = \frac{\text{圧縮後の記事の文字数}}{\text{元記事の文字数}}$$

⁴ 文字数圧縮規則の改良に用いた記事(約300記事)は、本論文で評価に用いた記事と重複していない。

短文分割を行うと、前述のように記事の文字数は増加するが、それらは中間結果であり、この圧縮率は、あくまで、元の記事の文字数と最終的な記事の文字数から算出されるものである。

まず、元記事 200 に文字数圧縮規則を適用して、圧縮率を計算し、そして、分割 1、分割 2 の処理を行った結果に文字数圧縮規則を適用して、圧縮率を算出した。つまり、3 つの場合での圧縮率を計算している。

1. 短文分割は行わず、文字数圧縮のみを行った場合
 2. 短文分割 1 を行い、その後文字数圧縮を行った行った場合
 3. 短文分割 2 を行い、その後文字数圧縮を行った行った場合
- その結果は、表 5 の通りである。

	200 記事	短文分割による 増加文字数	文字数	圧縮率 (%)
	元記事		96,936	
1	文字数圧縮のみ		91,808	94.71 %
2	分割 1 + 文字数圧縮	1,438	89,182	92.00 %
3	分割 2 + 文字数圧縮	1,742	89,114	91.93 %

表 5 文字数圧縮での短文分割の影響

短文分割を行うと、分割 1、分割 2 を問わず記事がより圧縮されることが判明した。この理由として次の 2 点が考えられる。

- 短文分割により接続詞（「そして」、「しかし」）などが挿入されるが、文字数圧縮の段階で削除される。これにより、短文分割処理による文字数の増加分は、文字数圧縮により相殺される。
- 短文分割の条件に合った長文が、記事中にある場合は、分割が行われ、文数が増える。これにより、文末の数も増加することになり、文字数圧縮規則により文末の文字数を圧縮できる機会が増え、より多くの文字を減らすことが出来ることになる。

つまり、短文分割を行うことにより、元記事を単に文字数圧縮をした場合よりも文字数をより減らすことが出来る場合が生じることになる。

例えば、3. 1 節の京葉銀行の文だと、原文（1 文のみ）では、文字数圧縮規則により削減されるのは、文末の「届け出ました」が「届け出た」となる 2 文字だけである。ところが、3 文に分割されると、以下のようになる。

1. 「京葉銀行の成田西支店の女子行員が、他人名義のカードローンを悪用しておよそ三億円を着服していた疑いが強まった。」
2. 「京葉銀行ではきょう、この女子行員を懲戒解雇。」
3. 「千葉県警察本部に被害を届け出た。」

第一文で「強まりました」が「強まった」となり、第二文で、「そして、」が削除、「解雇しました」が「解雇」になり、第三文では、「それとともに、」削除、「届け出ました」が「届け出た」となり、3文全体では、原文と比べて、結局、8文字削減されたことになる。つまり、短文分割を行うことにより、より多くの文字が削減されることになった。

5 おわりに

テキスト中の長い文を短い文に分割する短文分割処理の自動要約技術への影響を調べた。自動要約技術としては、重要度による文の順位付け、そして、不要部分を特定し、削除、言い換える文字数圧縮を取り上げた。重要と判定される文で、短文分割により分割される場合の分割文の順位を調べると、順位にばらつきが見られ、短文分割の要約への効果があることが分かった。また、記事中の全文を対象とした順位の評価では、短文分割をすると、人とシステムの順位付けがより近くなることが判明した。次に、文字数圧縮においては、短文分割をすることにより、文字数を削減する機会が増え、より多くの文字が削除出来ることが判明した。短文分割は、文単位の要約を、より小さな単位に分割して要約しようとするものであり、今後は、短文よりさらに小さな単位である文節を基礎とした自動要約技術についての研究を進めて行く予定である。

本論文では、対象をテレビニュース原稿としたが、新聞記事（毎日新聞 1995年版）を用いて実験を行ない、他の分野でのテキストでも、短文分割の効果があるかを調べた。

新聞記事は、記事の文字数が250文字以上ある100記事を無作為に選んだものを用いた。評価には、文字数圧縮において、短文分割の効果があるかを調査した。その結果、元記事を単純に文字数圧縮した場合、圧縮率が97.90%であり、圧縮そのものにあまり効果が見られなかった。これは、新聞記事では、文末が、「です、ます」調ではなく言い切りであること、また、テレビニュース原稿に見られる「…ということです」などの独特の表現が少ないことに原因があると思われる。短文分割後に文字数圧縮をした場合は、圧縮率が97.61%となり、短文分割をしない場合より0.29%下がるに留まり、ニュース原稿の場合ほどには、分割の効果が見られなかった。この実験は、小規模のものであり、TVニュース原稿以外での分野のテキストにおいて短文分割の自動要約への効果を調べるには、より詳細で、大規模な実験が必要であり今後の課題である。

付録

条件	アクション と 例文
文末の動詞がサ変動詞 (但し, 否定の表現は含まない)	そのサ変動詞以降を全て削除する 「強調しました。」 → 「強調。」 「言及しませんでした。」 (適応せず)
文末の動詞がサ変名詞 + 「を」 + 「する」 (但し, 否定の表現は含まない)	そのサ変名詞以降を削除する 「宣誓をしました。」 → 「宣誓」
丁寧助詞の「ます」	「ます」「まし」を削除して適当な文末に 「…なりました。」 → 「…なった。」 「…訪れます。」 → 「…訪れる。」
特定の文末表現	その表現を削除 「ということです」, 「としています」, 「ことにしています」 など
特定の文頭表現	その表現を削除 「一方」「その一方で」「このあと」 など
名詞性語句 + 断定の助動詞「です」で 終る	「です」を削除する 「状況です。」 → 「状況。」 「15 アンダーです。」 → 「15 アンダー。」
特定の表現	意味を変えずに, より短い表現 (語句) に 「総理大臣」 → 「首相」 「最高裁判所」 → 「最高裁」
文中に現れ, 括弧でくくられたカタカナ 文字列	カタカナ文字列を括弧とともに削除 「大洗漁港 (オオアライギョコウ)」 → 「大洗漁港」
省略形がある場合	省略形だけにする 「連合 = 日本労働者組合総連合会」 → 「連合」
括弧でくくられた数字列	括弧とともに削除する 「… 容疑者 (49)」 → 「… 容疑者」
特別な表現 <ul style="list-style-type: none"> 「問い合わせ先」ではじまる文 「電話」ではじまり, その後が数字と括弧だけで構成される文 	削除する

参考文献

- Edmundson H.P. (1969). "New Methods in Automatic Extracting." *Journal of the ACM*, **16** (2), pp. 264-285.
- 江原 暉将, 沢村 英治, 若尾 孝博, 阿部 芳春, 白井 克彦 (1997). "聴覚障害者のための字幕つきテレビ放送制作への自然言語処理の応用." 言語処理学会 第3回年次大会予稿集.
- Grefenstette G. (1998). "Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind." In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, pp. 111-117.
- Hovy E., Marcu D. (1998). "Automated Text Summarization" Tutorial Notes for Text summarization." *the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.
- 金 淵培, 江原 暉将 (1993). "日英機械翻訳のための日本語ニュース文自動短文分割と主語補完." 情報処理学会 自然言語処理研究会報告書 NL-93-3.
- 木村 真理子, 野村 浩一, 平川 秀樹 (1993). "日英機械翻訳前編集における日本語文分割処理について." 情報処理学会 自然言語処理研究会報告書 NL-96-8.
- Luhn, H.P. (1957). "A statistical approach to the mechanized encoding and searching of literary information." *IBM Journal of Research and Development*, **1** (4), pp. 309-317.
- 毎日新聞 (1995). CD-毎日新聞 9 5 版, (株) 毎日新聞社.
- Marcu D. (1997). "From discourse structures to text summarization." In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 82-88.
- 三上 真, 山崎 邦子, 増山 繁, 中川 聖一 (1998). "文中の重要部抽出と言い替えを併用した聴覚障害者用字幕生成のためのニュース文要約." 言語処理学会第四回年次大会併設ワークショップ「テキスト要約の現状と将来」論文集, pp. 14-21.
- 奥村 学, 難波 英嗣 (1998). "テキスト自動要約技術の現状と課題." 北陸先端科学技術大学院大学 情報科学研究科リサーチレポート IS-RR-98-0010I.
- 大村 平 (1980). "統計解析のはなし." 日科技連出版社.
- Wakao, T., Ehara, E., Sawamura, E., Abe, Y., Shirai, K. (1997). "Application of NLP technology to production of closed-caption TV programs in Japanese for the hearing impaired." In *Proceedings of ACL97 workshop, Natural Language Processing for Communication Aids*, pp. 55-58.
- 若尾 孝博, 江原 暉将, 白井 克彦 (1997). "テレビニュース番組の字幕に見られる要約の手法." 情報処理学会自然言語処理研究会, NL-122-13.
- 若尾 孝博, 江原 暉将, 白井 克彦 (1998). "テレビニュース字幕のための自動要約." 言語処理学会併設ワークショップ「テキスト要約の現状と将来」論文集, pp. 7-13.

Wakao, T., Ehara, E., Shirai, K. (1998). "Project for production of closed-caption TV programs for the hearing impaired." In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Coling-ACL98)*.

略歴

福島 孝博: 1990 年米国 State University of New York at Buffalo, 大学院 Computer Science 研究科修士終了. 1990 年から 1993 年米国 New Mexico State University, CRL にて研究員. 1994, 95 年, 英国シェフィールド大学大学院 Computer Science 研究科 Research Associate (研究員). 96 年日本電気 (株) 入社. 同年より通信・放送機構渋谷上原リサーチセンターに出向, 研究員, 現在に至る. 自然言語処理, 情報抽出, 自動要約の研究に従事.

江原 暉将: 1967 年早稲田大学第一理工学部電気通信学科卒業. 同年, NHK 入局. 1970 年より放送技術研究所勤務. かな漢字変換, 放送衛星の管制制御, 機械翻訳, 音声認識などの研究に従事. 1996 年より現職. 本会評議委員. 工学博士.

白井 克彦: 1963 年早稲田大学理工学部電気工学科卒業. 1968 年大学院理工学研究科博士課程修了. 同年同大学理工学部電気工学科専任講師, 1975 年同教授, 1991 年理工学部情報学科教授. 1998 年, 常任理事.

(1998 年 9 月 30 日 受付)

(1998 年 12 月 17 日 再受付)

(1999 年 2 月 19 日 再々受付)

(1999 年 4 月 19 日 採録)