

類似用例文の部分的置換による文短縮

牧野 恵 池田 諭史 山本 和英

長岡技術科学大学 電気系

〒940-2188 新潟県長岡市上富岡町 1603-1

E-mail : {makino, ikeda, ykaz}@nlp.nagaokaut.ac.jp

概要

近年、テキストの情報をなるべく減らさずに文を短くし直す文短縮の研究が盛んに行われている。しかし一般のニュース記事から短縮文を得るためには重要箇所の特定や冗長部分の削除等様々な処理が必要となるため容易ではない。そこで本稿では入力文に対して類似した用例文「類似用例文」を選択し、情報の部分的置換を行うことで短縮文を得る用例利用型の文短縮手法を提案する。評価実験では既存研究やリード法との比較を行い、提案手法の優位性を示した。また主観評価では被験者が短縮文に採用した文節と比較を行い、F 値 0.65~0.71 が得られた。

キーワード： 文短縮，用例利用型，「類似用例文」，「新幹線要約」

Example-based Sentence Contraction of Newsflash

Megumi Makino, Satoshi Ikeda, Kazuhide Yamamoto

Department of Electrical Engineering, Nagaoka University of Technology

1603-1, Kamitomioka-cho, Nagaoka-shi, Niigata 940-2188 Japan

E-mail : {makino, ikeda, ykaz}@nlp.nagaokaut.ac.jp

Abstract

There has been an increase of research on sentence contraction, in which textual information are kept as much as possible. Sentence contraction of newsflash requires catching important parts and removing verbose description out of the sentence. In this paper we propose an example-based approach for sentence contraction. Our proposed method prepares collection of contracted examples and use the most similar example to the input, in order to determine deletion parts. The experimental results showed the performance of 0.65-0.71 by F-measure, that outperforms to the conventional method and the baseline.

keywords : sentence contraction, example-based, similar example, newsflash

1 はじめに

現在、携帯端末や電光掲示板など様々な場所でニュース記事を見ることができる。これらのニュース記事は表示する機器の大きさや人間の読むスピード等の制限により、一般のニュース記事に比べ1記事あたりの文字数が少なく、要点のみが述べられていることが多い。これらのような要約を作成するには重要箇所の特定や換言、冗長部分の削除等の処理が必要であり、近年様々な研究が行われている。例えば先行研究 [1] で我々は要約に表れる特徴的な表現のうち文末の体言止めや助詞止めといった表現に着目し、一般的なニュース記事に文末整形を行った。

大森ら [2] は携帯端末向けのニュース作成のために重要度の低い文節を削除することによって短縮文

を得ている。重要度には $tf \cdot idf$ を用いているが、人間が重要語を選択するときは頻度以外の特徴も考慮していると考えられ、重要度の高い文節も削除してしまう場合がある。

三上ら [3] は字幕作成を目的とし、人手により作成した形式的表現を用いて冗長部分の削除を行っている。しかし削除候補が人手で作成した形式的表現に偏ってしまう恐れがある。

また堀ら [4] は単語重要度を最大に、かつ日本語として自然な部分単語列の抽出を動的計画法で解いている。単語単位の抽出では要約率に柔軟だが、要約率が小さいときには日本語として不自然な短縮文を作成することがある。

本稿では入力文に対して類似した用例文「類似用例文」をバタンの比較によって選択し、情報の部分的置換を行うことで短縮文を得る用例利用型の文短

縮手法を提案する。用例文には実際に電光掲示板等で用いられている短縮文を使用する。これらの短縮文は人手により重要箇所を特定し、冗長部分を削除して作成された文である。

従来法では例1のような入力文から短縮文を得る際、どの部分を短縮文に採用したら良いのか判断することは困難であった。しかし提案手法では選択された類似用例文に採用箇所の選択を委ねるため、重要箇所を容易に特定できる。

例1)

入力文：ダイエーの高木邦夫社長(60)は15日午前、決算取締役会の席上で22日付で社長を辞任すると表明

類似用例文：DDIとKDD、日本移動通信(IDO)は16日午後、来年10月1日付で合併すると正式に発表

例えば入力文と選択された類似用例文を比較すると用例文にはワ格が使用されていない。よって短縮文に現れにくい非重要箇所であることが分かり、例2のような短縮文が得られる。

例2)

短縮文：高木邦夫社長(60)は15日午前、22日付で辞任すると表明

このように用例利用型の文短縮では重要箇所の特定が容易になるだけではなく、利用した類似用例文に近い表現で短縮文が作成することができる。

2 用例文の収集

新幹線車内の電光掲示板で使用されているニュース記事が日経 goo(1) から配信されている。日経 goo は月曜日から金曜日までの週に5日、1日3回のメール配信を行っており、その中の「主なニュース」という部分が実際に新幹線車内で用いられているニュース記事である。本稿ではこのニュース記事のことを「新幹線要約記事」と呼ぶ。例3に配信されている新幹線要約記事の例を示す。

例3)

表題：ハリケーン独立調査委の設置法案、米上院で否決

本文：米上院は14日、ハリケーン被害への米政府対応について調査する独立調査委員会の設置法案を否決。共和党が反対した。

配信されている新幹線要約記事の本文は1~3文で構成される。しかし2文目以降の文は1文目の付加情報であることが多いため、本稿では用例文とし

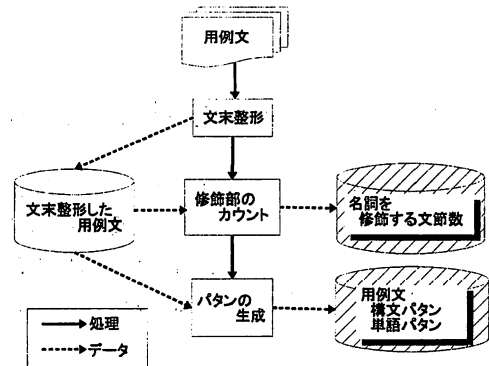


図1 前処理部（用例文に対する処理）

て利用する対象を本文の1文目*1とし、1999年12月~2006年3月の期間で28032文収集した。

3 提案手法

本手法は用例に対する処理を行う前処理部と入力文を与えてから短縮文を出力する短縮文作成部の2つの処理で構成される。各処理の詳細は次節以降で説明する。

3.1 前処理部（用例文に対する処理）

図1に前処理部の流れを示す。本節では短縮文作成部で類似用例文を得る際に必要となる“用例文から生成した2つのパターン”と短縮文を作成する際に必要となる“名詞を修飾する文節数”のデータを作成する。

I. 文末整形

新幹線要約記事に表れる特徴的な文末表現として体言止めや助詞止めが挙げられる。しかし本稿で用いる用例文にはこれらの形式ではない文末も含まれる。例えば、文末が体言止めである「発表」は用例文で1803回用いられているが、文末が終止形である「発表した」も803回用いられている。よって類似用例文の選択が正しく行えるよう、文末の形式を揃える意味で用例文に文末整形を施す。本稿では先行研究[1]の手法で文末整形を行った。

II. パタンの生成

短縮文作成部では入力文と用例文のパタンを比較し、入力文と類似した用例文を選択する。文の類似には構文的な類似と語彙的な類似があるため、構文情報に着目した構文パターンと、単語に着目した単語

*1 例3の下線部。

パタンの2種類を生成する。以下に構文パタンと単語パタンの生成方法について述べる。

【構文パタン】

文末、動詞、格情報を手がかりに文の基本構造を表す部分をパタンに採用する。本稿では構文解析器「南瓜」(2)の解析結果を用いて構文パタンの生成を行った。構文パタンに採用する部分は以下に示す手順で決定する。

Step1 文末の文節を採用する。ただし数字、助詞は「#*2」と表記し、パタンに採用する。

Step2 パタンに採用した文節の係り元文節で動詞、サ変名詞、助詞、助動詞、「こと、もの」はパタンにそのまま採用し、これ以外は「#」と表記して採用する。

Step3 Step2で採用した文節にサ変名詞または「こと、もの」を含む場合、この語を汎化するか判別する。

Step3-1 サ変名詞を含む場合

サ変名詞は名詞的役割と動詞的役割がある[5]。そこでサ変名詞に係る文節が「～の」や「～する」といった修飾語を表している場合、名詞的役割であるとする。名詞的役割のサ変名詞は「#」と汎化する。

Step3-2 「こと、もの」を含む場合

「こと、もの」が含まれる文節の係り元文節は動詞を含むことが多い(例4)。そのため「こと、もの」の係り元文節が動詞の以外の場合、「こと、もの」を「#」と汎化する。

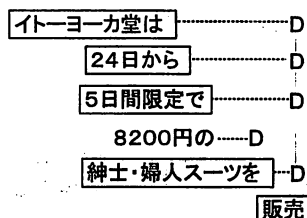
例4)

民主党は/^{*3}27日/午前の/役員会で、/30日からの/
臨時国会の/冒頭に/年金制度改革法の/廃止法案
を/提出する/ことを/決めた/

Step4 Step2, 3で採用した文節に「#」を含まない場合 Step2へ戻る。含む場合は処理を終了する。

原文で構文パタンに採用する文節を図2に示す。図2では、まず文末「販売」をパタンに採用する(Step1)。次に文末に係る文節「紳士・婦人スーツ(名詞句)を(助詞)」を汎化し「#を」として構文パタンに採用する(Step2)。また「#」を含む文節に係る

原文: イトーヨーカ堂は24日から5日間限定で
8200円の紳士・婦人スーツを販売



枠内の文節を構文パタンに採用する

図2 構文パタンに採用する文節の例

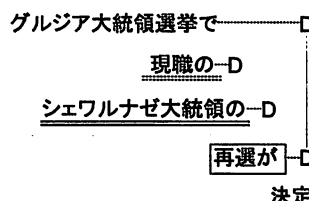


図3 修飾部付加の例

文節はパタンに採用しないため、「8200円の」はパタンに採用しない(Step4)。この手順で生成した構文パタンを以下に示す。

例5)

構文パタン: #は#から#で#を販売

【単語パタン】

単語パタンでは語彙に着目し、固有名詞、数字、記号を「#」と表記し、汎化する。これ以外の語はそのまま採用する。図2の原文から生成した単語パタンの生成例を以下に示す。

例6)

単語パタン: #は#日から#日間限定で#円の紳士・婦人スーツを販売

III. 修飾文節数のデータ作成

短縮文作成部では入力文と類似用例文の情報を置換することによって短縮文を得る。名詞を修飾する部分の長さはその名詞によって異なる。そこで用例文で用いられている「名詞(複合名詞)+格助詞」または「名詞(複合名詞)+係助詞」の文節に対して文節が何重に係っているか平均を取り、四捨五入した数をその名詞に対する修飾文節数として保持する。ただし複合名詞であった場合はその最終形態素に対する修飾文節数として保持する。図3に構文解析の例を示す。この例で「再選が」の係り元文節として「シェワルナゼ大統領の」があるが、この文節にはさ

*2 「#」は語を汎化したものである。ただし「#」の連続は1つの「#」とする。

*3 「/」は文節区切りを表す。

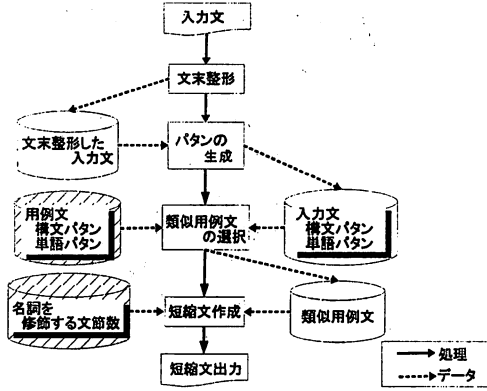


図4 短縮文作成部（入力文に対する処理）

らに「現職の」が係り元文節として存在する。よって「再選」を修飾する文節数は“2”となる。また複合名詞「シェワルナゼ大統領」の最終形態素「大統領」を修飾する文節数は“1”となる。

3.2 短縮文作成部（入力文に対する処理）

本節では与えられた入力文から短縮文を作成する処理について説明する。図4に短縮文作成部の流れを示す。

I. 類似用例文の選択

入力文に対して類似する用例文を選択する。本稿ではパターンの比較で類似用例文を選択するため、3.1節で示した用例文に対する処理と同様の方法で入力文から2つのパターンを生成する。比較に用いる類似尺度にはBLEUスコア[6]を用いる。BLEUスコアはシステム出力 c とリファレンス r の類似度を求めることによって評価を行う。以下にBLEUスコアの定義を示す。

$$BLEU(c, r) = BP \cdot \exp\left(\sum_{n=1}^4 \frac{1}{n} \log p_n\right) \quad (1)$$

$$BP = \begin{cases} 1 & \text{if } |c| > |r| \\ e^{(1-|r|/|c|)} & \text{if } |c| \leq |r| \end{cases} \quad (2)$$

$$p_n = \frac{\sum_{gram_n \in c} \min(CT_c(gram_n), CT_r(gram_n))}{\sum_{gram_n \in c} CT_c(gram_n)} \quad (3)$$

BP はシステム出力の長さ $|c|$ がリファレンスの長さ $|r|$ よりも短いときに不当なスコアを出さないために与えるペナルティである。 $CT_c(gram_n)$ はシステム出力に含まれる n -gramの数であり、

$CT_r(gram_n)$ はリファレンスに含まれる n -gramの数である。よって p_n はシステム出力とリファレンスとの間で一致した n -gramがシステム出力に対して占める割合を表す。BLEUスコアは本来、翻訳の自動評価等で用いられるスコアであるが、本稿では幾何平均やペナルティが考慮されている点でパターンの比較にも適していると考えられる。そこで入力文 s の各パターンをシステム出力 c 、用例文 t_i の各パターンをリファレンス r と見なし、次式で入力文 s と用例文 t_i の類似度を算出する。

$$\begin{aligned} \text{sim}(s, t_i) = & \lambda BLEU(d(s), d(t_i)) \\ & + (1-\lambda) BLEU(w(s), w(t_i)) \end{aligned} \quad (4)$$

ただし $d(s)$, $d(t_i)$ は入力文 s 、用例文 t_i の構文パターンを表し、 $w(s)$, $w(t_i)$ は入力文 s 、用例文 t_i の単語パターンを表す。また λ は構文パターン、単語パターンに対するパラメータである。入力文に対する類似用例文 $\hat{t}$ は次式で表される。

$$\hat{t} = \arg \max_{i=1,2,\dots,N} \text{sim}(s, t_i) \quad (5)$$

ここで N は収集した用例文の数を表す。

3.3 短縮文の作成

選択された類似用例文は入力文と構文情報が類似していると考えられるため、入力文の文節を類似用例文の文節と置換し短縮文を作成する。よって置換後は類似用例文が入力文の文節で構成される。以下に作成手順を示す。

Step1 入力文の文末文節と類似用例文の文末文節を置換する。

Step2 入力文と類似用例文の置換した文節の係り元文節を比較し、以下の条件によりいずれかの処理を行う（置換できる文節が無くなるまで繰り返す）。

Step2-1 文節末の形態素が助詞の場合

同一の助詞であればその文節を置換する。ただし入力文の文節に「の」、「など」が含まれる場合は後に修飾部の付加処理を行うため、ここでは置換しない。同一の助詞が見付からなかった場合、人手により作成した換言辞書を用いて助詞の換言を行い、換言後の助詞が同一の場合のみ置換を行う。換言辞書の対象は助詞相当句（21語）と主語の一部となる「は」、「も」、「が」であり、これを助詞1語と換言する。表1に換言辞書の例を示す。

Step2-2 文節末の形態素が助詞以外の場合

文節末の形態素が同じ品詞の場合は置換する。読点の場合、その直前の形態素の品詞を比較し、同じ場合は置換する。

表 1 換言辞書の例

換言対象	換言後
によって	で、から
をもって	で
は	が、も

Step3 構文情報が複雑な場合、主語を表す文節が置換されない場合がある。しかし用例文の多くは主語を表す「は、が、も」を含む文節が存在する。よってこのような文節は重要箇所であると考えることができ、入力文、類似用例文のそれぞれに主語を表す「は、が、も」を含む文節が 1 文節存在する場合は、その文節を置換する。

Step4 類似用例文に入力文以外の文節が含まれている場合、その文節を削除する。

Step5 修飾部の付加を行う。「名詞(複合名詞)+格助詞」、「名詞(複合名詞)+係助詞」の文節を修飾部付加の対象とする。前処理部で保持した名詞に係る文節数を用いる。

以下に短縮文作成の例を示す。入力文と類似用例文の同じ下線部は対応した文節であることを示す。

例 7)

入力文：米大統領選の/民主党候補ケリー上院議員は/3 日、/ボストン市内で/演説、/オハイオ州で/すべての/仮投票を/集計しても/「我々が勝つために必要な十分且つ明確な票は得られない」と/言明し、/選挙戦の/敗北を/正式に/認めた/
類似用例文：元建設相の/中尾栄一被告は/11 日、/受諾収賄罪の/初公判で/起訴事実を/全面的に/認めた/

まず「認めた」を置換する(Step1)。次に置換した文節に係る文節、つまり「認めた」に係る文節(下線部分)で助詞が同じ文節「正式に(全面的に)」,「敗北を(起訴事実を)」,「民主党候補ケリー上院議員(中尾栄一被告は)」を置換する(Step2-1)。また「敗北を」,「民主党候補ケリー上院議員は」に係る文節(波線、二重下線)には「の」が含まれているため置換を行わない。さらに残った用例文の文節を削除すると以下の短縮文が作成される(Step4)。

例 8)

短縮文(修飾語の付加前)：民主党候補ケリー上院議員は/敗北を/正式に/認めた/

次に修飾語の付加を行う(Step5)。修飾語の付加の対象は「民主党候補ケリー上院議員は」,「敗北を」の

下線部である。「議員」には平均約 0.7 文節(四捨五入により 1 文節)、「敗北」には平均約 0.3 文節(四捨五入により 0 文節)が係っていたため、「民主党ケリー上院議員は」に係る 1 文節を短縮文に加える。これにより作成される短縮文は以下ようになる。

例 9)

短縮文：[米大統領戦の]*4/民主党候補ケリー上院議員は/敗北を/正式に/認めた/

4 評価実験

4.1 実験データ

本稿では用例文に日経 goo から配信されている新幹線要約記事(1999 年 12 月～2006 年 3 月)の 1 文目 28032 文を用いた。入力するテストデータには NIKKEI NET(3) から配信された Web 記事(2004 年)の 1 文目 100 文を用いた。この Web 記事にはタイトルが付いている。このため新幹線要約記事との間でタイトルが同一の記事を自動的に対応付けして、Web 記事の 1 文目を客観評価で用いる正解データとした。入力文に対応した正解データ(用例文)は 1 件ずつ削除することによりオープンテストを行った。今後、出力上位の複数候補に対して優先度学習等を用いた再ランキング[8]を検討しているため、1 位の類似用例文から作成した短縮文の評価に併せて、上位 5 位の短縮文における評価の最高値も示す。

4.2 要約率

短縮文を作成した 100 文で得られた要約率を表 2 に示す。本稿では入力文中の形態素数に対する出力文中の形態素数の割合を要約率とする。要約率を求めた短縮文は 1 位の類似用例文から作成した短縮文と、上位 5 位の短縮文で ROUGE-1 が最も高かった短縮文である。

表 2 提案手法で得られた要約率

要約率の算出を行った対象	要約率
1 位の短縮文	.416
上位 5 位で ROUGE-1 が最も高かった短縮文	.454

4.3 客観評価

I. 評価方法

客観評価では提案手法と単語抽出により文短縮を行う堀らの手法、文頭から指定した要約率までの単語を抽出するリード法で比較を行った。ただし堀らの手法、リード法は提案手法で得られた要約率 0.416(1 位の類似用例文で作成した短縮文の要約率)

*4 付加された修飾部を表す。

と 0.454(上位 5 位で ROUGE-1 が最も高かった短縮文の要約率)を用いて文短縮を行った。評価尺度には ROUGE スコア [7]を用いる。ROUGE スコアは BLEU スコアを要約の自動評価用に改良したものである。以下に ROUGE スコアの定義を示す。

$$\text{ROUGE-N}(c, r) = \frac{\sum_{gram_n \in r} CT_{match}(c, r)}{\sum_{gram_n \in r} CT_r(gram_n)} \quad (6)$$

$$CT_{match}(c, r) = \min(CT_c(gram_n), CT_r(gram_n)) \quad (7)$$

$N=1$ または $N=2$ の場合に、人間の評価結果に対し高い相関が得られたという報告に従い、本稿ではこの値を用いて以下の 4 手法で評価の比較を行った。

手法 (a) リード法

手法 (b) 既存研究 (単語抽出による文短縮)

手法 (c) 提案手法 (※ 1 位の短縮文)

手法 (d) 提案手法 (※ 上位 5 位で評価が最も高い短縮文)

II. 評価結果

ROUGE スコアによる評価結果を表 3, 4 に示す。表 3 は 1 位の類似用例文から作成した短縮文の要約率を、既存研究、リード法に適用し比較を行った結果である。また表 4 は上位 5 位の短縮文で ROUGE-1 が最も高かった短縮文の要約率を既存研究、リード法に適用し比較を行った結果である。

表 3 リード法 (a)、既存研究 (b) と提案手法 (c) の評価の比較 (要約率 0.416)

	(a)	(b)	(c)
ROUGE-1	.410	.421	.455
ROUGE-2	.233	.186	.279

表 4 リード法 (a)、既存研究 (b) と提案手法 (d) の評価の比較 (要約率 0.454)

	(a)	(b)	(d)
ROUGE-1	.446	.458	.517
ROUGE-2	.254	.208	.332

これらの比較結果より、提案手法である手法 (c)、(d) はリード法 (a)、既存研究 (b) に比べ優位な結果を得た。リード法や既存研究では単語単位の抽出によって文短縮を行うため要約率により柔軟であるが、同じ要約率で文短縮を行った場合、構文や文節の情報を生きた提案手法の方がより自然な短縮文が作成された。よって文短縮ではこのような情報も利用することが重要であると考えられる。

4.4 主観評価

I. 評価方法

被験者 3 人が原文の Web 記事 1 文目を文節区切りにしたものから以下の観点で文節を選択した。

教示 (1) 短縮文を作成する際、必要だと思う文節 (文節数任意)

教示 (2) 指定された文節数の範囲で短縮文を作成する際、必要だと思う文節 (文節数指定)

ただし、教示 (2) における指定した文節数の範囲とは提案手法で出力した短縮文上位 5 位における文節数の最小値と最大値である。主観評価では提案手法で出力した短縮文の文節と教示 (1)、(2) で被験者が選択した文節との比較を行い、再現率、適合率及び F 値で評価した。なお、比較を行う短縮文の対象は、類似用例文を上位 5 位まで出力して、それぞれから短縮文を作成した際に ROUGE-1 の評価が最大になった短縮文である。

II. 評価結果

表 5 に被験者が文節数任意で選択した文節との比較を行った結果を示す。これは人間が考える理想的な短縮文と比較した精度を表す。

表 5 文節数任意の評価で得られた文節との比較

	被験者 A	被験者 B	被験者 C	全体平均
再現率	.564	.600	.599	.588
適合率	.778	.867	.699	.781
F 値	.654	.709	.645	.669

被験者によって多少揺れはあるが、提案手法で選択した文節の約 8 割 (0.781) は人間が考える理想的な短縮文の文節であることが分かる。

次に被験者が文節数指定で選択した文節との比較を行った結果を表 6 に示す。

表 6 文節数指定の評価で得られた文節との比較

	被験者 A	被験者 B	被験者 C	全体平均
再現率	.581	.716	.632	.637
適合率	.648	.786	.637	.734
F 値	.613	.749	.634	.665

これは提案手法で選択した文節の正誤の結果を表す。この結果より提案手法が正解の文節を抽出できたのは約 7 割 (0.734) である。

5 考察

5.1 作成された短縮文の例

作成された短縮文の例を図 5 に示す。入力文の

入力文：

イラク南部サマワで人道復興支援を行ってきた陸上自衛隊第一次派遣部隊の約 110 人が 17 日午前、民間チャーター機でクウェートから北海道東神楽町の旭川空港に到着した

類似用例文：

中国から出国した朝鮮民主主義人民共和国の一家 5 人が 23 日午前 3 時 45 分、經由地のマニラから韓国の仁川空港に到着

作成された短縮文：

[陸上自衛隊第一次派遣部隊の] 約 110 人が 17 日午前、クウェートから旭川空港に到着

正解：

陸上自衛隊第一次派遣部隊の約 110 人が 17 日、クウェートから旭川空港に到着

図 5 作成された短縮文の例

文末「到着した」は文末整形と施すことにより「到着」と変更された。また入力文の文節「民間チャーター機で」は文末の動詞に係るデ格であるが、選択された用例類似文には表れていない。つまり短縮文には表れにくい表現であることが分かる。これにより正解文とほとんど一致する短縮文が作成できた。

5.2 修飾部付加の効果

修飾部付加の有用性を測るために修飾部を付加する前後で評価の比較を行った。表 7 に修飾部の付加前後での ROUGE スコアの結果を示す。

表 7 修飾部の付加前後における ROUGE スコアの比較

評価対象	評価尺度	付加前	付加後	前後での差
1 位	ROUGE-1	.437	.455	+0.018
(手法 (c))	ROUGE-2	.269	.279	+0.010
上位 5 位	ROUGE-1	.496	.517	+0.021
(手法 (d))	ROUGE-2	.314	.332	+0.018

この結果より、修飾部付加を行う前に比べ付加後では ROUGE-1 で 1.8 ポイント、ROUGE-2 で 1.0 ポイントの改善が見られた。

5.3 主観評価について

主観評価では被験者によって結果に揺れが生じた。その原因として以下の 2 点が挙げられる。

- (1) 修飾部の選択の揺れ
- (2) 文末の選択の揺れ

被験者によってどの部分を修飾部として短縮文に採用するか差異が見られた。具体例を以下に示す。

例 10)

入力文： 輸血用血液への/ウイルス混入問題を/受けて、/厚生労働省は、/何度も/献血する/健康な人

で/つくる/「複数回献血者クラブ」/(仮称)の/創設など/安全な/血液の/確保策を/固めた/

被験者 A： 厚生労働省は、/健康な/人で/つくる/「複数回献血者クラブ」/創設など/安全な/血液の/確保策を/固めた/

被験者 B： 厚生労働省は、/「複数回献血者クラブ」/創設など/安全な/血液の/確保策を/固めた/

被験者 C： 厚生労働省は、/安全な/血液の/確保策を/固めた/

この例では「確保策」という名詞を修飾する部分の長さが被験者によって異なっている。提案手法では用例文の名詞に係る文節数のみを用いて修飾部の付加を行っている。そのため、人間に近い短縮文を作成するには今後さらに検討が必要である。

次に文末選択についての具体例を以下に示す。

例 11)

入力文： 台湾が/1960 年代から/80 年代にかけて、/核兵器の/材料と/なる/プルトニウムを/分離抽出する/実験を/していた/可能性が/ある/ことが/13 日/分かった/

このような入力文では短縮文の文末を「ある」とするか「分かった」を選択するかは被験者によって異なった。提案手法では短縮文の作成で文末 1 文節を必ず短縮文に採用しているが、今後は主動詞判定等を行い、いくつか短縮文候補を出力するなど検討が必要である。

このように被験者が選択した文節には差異が見られたため、入力文に対して採用した文節数も異なった。表 8 に短縮課題 1 題に対して採用した文節数を示す。

表 8 入力文に対して採用した文節数

	被験者 A	被験者 B	被験者 C	提案手法
文節数	7.0	7.8	6.3	5.8

表 8 より被験者によって採用する文節数には差異が見られたが、どの被験者の文節数よりも提案手法が選択する文節数の方は少ないことが分かる。つまり人間が考えた理想的な短縮文にくらべ本手法での短縮文は記述量が少ない。よって今後の課題として要約率可変の文短縮器の作成が挙げられる。

5.4 正しく作成されなかった短縮文について

本手法で正しく作成されなかった短縮文を検証した結果、以下の 2 点が大きな原因として挙げられる。

- (1) 選択した類似用例文
- (2) 修飾部付加の判断

類似用例文の選択では構文的な類似と語彙的な類似を考慮した。しかし以下のような例では選択した類似用例文に問題があり、正しい短縮文が作成されなかった。

例 12)

入力文： 北朝鮮による拉致被害者、曾我ひとみさんの夫で元米兵のジェンキンスさんは 6 日昼、入院先の都内の病院で、前日に引き続き在韓米軍に所属する独立法務官のカルブ陸軍大尉と面会した

類似用例文： 曾我ひとみさんの夫で元米兵のジェンキンスさんが自らの訴追問題で、11 日にも在日米軍のキャンプ座間」に出頭へ

作成された短縮文： ジェンキンスさんは [都内の] 病院で、面会

この例では語彙的な類似に偏りが生じたため、文節の置換を正しく行うことが出来ず、不十分な短縮文が出力された。よって語に重みをおいて類似用例文を選択することや類似度を測る尺度の再検討が必要である。

次に修飾部の付加が不十分であった短縮文の例を示す。

例 13)

入力文： 松下電器産業は世界で最も消費電力が少ない飲料用の自動販売機を開発した

類似用例文： 三菱電機は日本電池と人工衛星用のリチウムイオン電池を開発した

作成された短縮文： 松下電器産業は [飲料用の] 自動販売機を開発

この例では下線部に修飾部付加を行うため、用例から「最終形態素が「機」である複合名詞＋格助詞または係助詞」に係る文節数の平均を求めた。その結果、修飾文節数は「1」であったため、入力文で「自動販売機を」に係っている文節「飲料用の」を付加した。複合名詞によっては用例中に出現しづらいため最終形態素のみを用いて修飾文節数を求めたが、作成された短縮文の情報は不十分であった。そのため複合名詞を扱う場合、修飾部の付加についてはさらに検討が必要である。

6 おわりに

新幹線要約記事を用例文として用い、入力文に対して選択された「類似用例文」を利用することで短縮文を作成する手法を提案した。「類似用例文」は入力文、用例文の双方からパタンを生成し、BLEU スコアの比較によって得た。短縮文は入力文と「類似用例文」の文節を置換することによって作成した。評価実験では既存研究やリード法との比較を行い、

提案手法の優位性を示した。今後の課題として、修飾文節の厳密な判断や要約率可変の文短縮などが挙げられる。

謝辞

本研究の一部は、科学研究費補助金費補助金基盤 (a) 「円滑な情報伝達を支援する言語規格と言語変換技術」課題番号 16200009 によって実施した。

使用した言語資源及びツール

- (1) 日経ニューズメール, NIKKEI-goo,
<http://nikkei.goo.ne.jp/>
- (2) 構文解析器「南瓜」, Ver0.52,
<http://chasen.org/~taku/software/cabocha/>
- (3) 日経 Web ニュース, NIKKEI NET,
<http://www.nikkei.co.jp/>

参考文献

- [1] 山本和英, 池田諭史, 大橋一輝. 新幹線要約のための文末整形. 言語処理学会論文誌, Vol.12, No.6, pp.85-112(2005)
- [2] 大森岳史, 増田英孝, 中川浩志. Web 新聞記事の要約とその携帯端末向け記事による評価. 情報処理学会研究報告, NL-153-1, 情報処理学会 (2003)
- [3] 三上真, 増山繁, 中川聖一. ニュース番組における字幕生成のための文内短縮による要約. 言語処理学会論文誌, Vol.6, No.6, pp.193-200(2004)
- [4] 堀智織, 古井貞照. 単語抽出による音声要約生成法とその評価. 電子情報通信学会論文誌, Vol.J85-D-II, No.2, pp.200-209(2002)
- [5] 山本和英, 大橋一輝. 「サ変名詞 + 名詞」の複合名詞への換言. 言語処理学会論文誌, Vol.12, No.3, pp.19-42(2005)
- [6] K.Papineni, S.Roukos, T.Ward and W.J.Zhu : BLEU : a Method for Automatic Evaluation of Machine Translation. *proc. of the 40th ACL*, pp.311-318(2002)
- [7] Chin-Yew Lin : Looking for a Few Good Metrics : ROUGE and its Evaluation, *proc. of the 4th NTCIR Workshops*, pp.1-8(2004)
- [8] 牧野恵, 平尾努, 山本和英, 磯崎秀樹. 優先度学習を用いた文短縮手法. 言語処理学会第 12 回年次大会, pp.1095-1098(2006)