

# 係り受け経路長を利用した新聞記事の自動簡約

山形 究 福富 諭 高木 一幸 尾関 和彦

電気通信大学

{ q-chan, fukutomi, takagi, ozeki } @ice.uec.ac.jp

## 1 はじめに

現在、情報技術の発展により世の中には膨大な電子化テキストが存在し、WWW等を通じて大量の電子テキストを入手する機会も多くなった。その結果、読み手の負担を軽減し、短時間で的確に内容を把握することを支援する自動要約技術が求められ、様々な手法が提案されている[1, 2]。これまでに主に検討されてきた要約手法の一つに、重要文抽出がある。これはテキストから文単位で抽出を行い、要約を生成する手法である。この手法では抽出した文そのものの自然性や意味には問題がないという利点があるが、さらに短い要約結果が必要な場合には対応出来ない。そこで我々、は抽出した重要文を更に圧縮する文簡約について研究してきた[3, 4]。ここで文簡約に一定の成果を得たが、生成した簡約文の自然性が十分でないなどの問題が残っていた[4]。本研究では、簡約手法に係り受け経路長[5]と呼ぶ考え方を導入し、“毎日新聞全文記事および54文字データベース”[6]から人間の要約傾向を学習することで、自然性の高い簡約文を生成することを目的とする。

## 2 重要文抽出

新聞記事を要約するために、まず原記事中から重要な文を抽出する。重要文抽出に関しては様々な手法があるが、本研究では、記事の第1段落第1文を重要文として抽出した。この方法を採用した理由は、第一に新聞記事の性質上、記事の最初に重要な情報を含む文(統括する文)が現れる傾向にあるため[4]、第二に抽出した文中に指示詞が含まれていて、かつその先行詞が抽出文中に存在しないなどの問題を避けるためである。

以降、重要文として抽出された文を“原文”と呼ぶ。

## 3 文簡約の定式化

文簡約を、文を複数の文節からなる列と捉え、「原文からできるだけ“良い”部分文節列を抽出すること」と考える。この部分文節列の“良さ”としては、

- 原文の持つ情報をできるだけ保持していること、
- 日本語として構文的にできるだけ自然であること、

という2つの要素が考えられる。そこで、この“良さ”をこの2つの要素それぞれに対応した評価関数の値の和として考える。本研究では、a)のための評価関数として各文節の文節重要度を、b)のための評価関数として2文節間の係り受け整合度を利用する。a)、b)に対応した2つの評価関数は以下のように定義する。

- 原文を複数の文節からなる列  $w_0 w_1 \dots w_{M-1}$  とし、その中の長さ  $l$  の部分文節列  $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$  を考える。文節  $w_m$  の重要度を表す関数  $q(m)$  が与えら

れたとき、この部分文節列の重要度をそれらの総和  $\sum_{i=0}^{l-1} q(k_i)$  として定義する。

- 文節  $w_m$  が文節  $w_n$  に係るときの係り受け整合度  $p(m, n)$  が与えられたとき、その総和が大きな値となる係り受け構造を持つ文節列は、日本語として文法的に自然性が高いと考えられる。部分文節列  $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$  上の係り受け構造は、係り文節番号を、受け文節番号に対応させる写像

$$c: \{k_0, k_1, \dots, k_{l-2}\} \longrightarrow \{k_1, k_2, \dots, k_{l-1}\}$$

によって表される。ただし、 $c$  は

- 後方単一性:  $k_m < c(k_m)$  ;
- 非交差性:  $m < n$  ならば  $[c(k_m) \leq k_n]$  , または  $c(k_n) \leq c(k_m)$  ] .

を満たす必要がある。本研究では、写像  $c$  を用いて、文節列  $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$  の日本語としての構文的な自然性の程度を  $\max_c \sum_{i=0}^{l-2} p(k_i, c(k_i))$  で測ることとする。ここで、最大化は可能な全ての係り受け構造に対して行う。

以上、a)、b)に基づいて、文節列  $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$  の“良さ”を測る評価関数  $g(k_0, k_1, \dots, k_{l-1})$  を次のように定義した。

$$g(k_0, k_1, \dots, k_{l-1})$$

$$\triangleq \begin{cases} q(k_0), & l=1 \text{ のとき;} \\ \alpha \{ \max_c \sum_{i=0}^{l-2} p(k_i, c(k_i)) \} \\ \quad + (1-\alpha) \{ \sum_{i=0}^{l-1} q(k_i) \}, & 2 \leq l \text{ のとき.} \end{cases} \quad (1)$$

ここで、 $\alpha (0 \leq \alpha \leq 1)$  は文節間係り受け整合度にかかる重みである。評価関数  $g(k_0, k_1, \dots, k_{l-1})$  を用いると、 $M$  文節からなる原文を  $l$  文節からなる文に簡約する問題は次のように定式化することができる。この問題は動的計画法の原理に基づき効率的に解くことができる[3]。この方法を実行するために文節重要度  $q(m)$  と係り受け整合度  $p(m, n)$  を定める必要がある。

文簡約問題

文節列  $w_0 w_1 \dots w_{M-1}$  の部分文節列  $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$  ( $0 \leq k_0 < k_1 < \dots < k_{l-1} \leq M-1$ ) の中で、関数  $g(k_0, k_1, \dots, k_{l-1})$  を最大にするものを求める。

## 4 使用するデータベース

本研究では、新聞記事から重要文を抽出し、更に文毎の簡約を行うことにより要約としている。また簡約の際には、人間の簡約傾向を統計的に学習することにより質の高い簡約を達成することを目指している。これらを実現するために、要約対象文と人間の要約が対になったデータが必要になる。そこで本研究では、“毎日新聞全文記事および

54 文字データベース (2002 年度版) "[6] を使用した。これは 2002 年 4 月から 2003 年 3 月までの毎日新聞の記事と、各記事に対して併せて提供された“ 54 文字要約 ”との対、更にそれぞれの見出しの 4 つ要素から構成されている約 32000 セットを含むデータベースである。このデータベースの 2002 年 5 月から 2003 年 3 月までの 28423 セットを学習用、2002 年 4 月の 2778 セットを評価用として利用した。毎日新聞の各記事は、形態素解析システム JUMAN 3.61[7]、構文解析システム KNP version2.0 b6[8] を用いて自動解析を行い使用した。

## 5 係り受け整合度の学習と推定

### 5.1 係り受け経路長

本論文では、文節同士の距離を表す際に“ 文節間距離 ”と“ 係り受け経路長 ”[5] を利用した。  $M$  文節からなる文節列  $w_0 w_1 \dots w_{M-1}$  の部分文節列  $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$  を考えたとき、“ 文節間距離 ”と“ 係り受け経路長 ”は次のように定義される。

- 文節間距離  
文節間距離  $D(w_s, w_t) = t - s$  ( $0 \leq s < t \leq M - 1$ ) .
- 係り受け経路長  
 $i = 0, 1, \dots, l - 2$  に対して、 $c(w_{k_i}) = w_{k_{i+1}}$  が成り立つとき、部分文節列  $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$  を  $w_{k_0}$  と  $w_{k_{l-1}}$  を結ぶ係り受け経路と呼び、その係り受け経路長を  $DRD(w_{k_0}, w_{k_{l-1}}) = l - 1$  と定義する。また、文節  $w_m$  と  $w_n$  を結ぶ係り受け経路が存在しないとき  $DRD(w_m, w_n) = \infty$  とする。

図 1 の例で考えると、 $D(\text{“この”}, \text{“運休した。”}) = 4$ 、 $D(\text{“台風の”}, \text{“各線が”}) = 2$ 、 $DRD(\text{“この”}, \text{“運休した。”}) = 3$ 、 $DRD(\text{“台風の”}, \text{“各線が”}) = \infty$ 、となる。係り受け経路長には係り受け構造が反映されている。

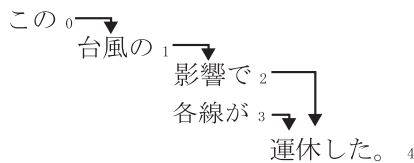


図 1: 例文“この台風の影響で各線が運休した。”の係り受け解析木。矢印が係り受けを表す。

### 5.2 文節対の分類

係り文節  $w_m$  と受け文節  $w_n$  間の係り受け整合度を推定するために、次のような文節中の形態素属性に着目し、2 文節間の係り受け経路長情報を加えて分類した。形態素属性による分類は文献 [4] と同様である。

1. 係り文節  $C_k(w_m)$ : 文節の最後の形態素に着目

- 活用語: 品詞と活用形
- 非活用語:  
助詞: 品詞詳細と表記  
助詞以外: 品詞詳細

2. 受け文節  $C_u(w_n)$ : 文末文節、非文末文節別に、接辞詞を除いて最初の形態素に着目

- 名詞: 名詞連鎖のあと  
判定詞: 品詞詳細  
他の品詞: 品詞詳細
- 形容詞: 品詞詳細と活用形
- 名詞, 形容詞以外: 品詞

3. 係り受け経路長  $DRD(w_m, w_n)$ : 係り受け経路長情報に着目。(  $DRD \neq \infty$  のときのみ)

### 5.3 文節対残存率の学習

本論文では、原文から“ 54 文字要約 ”が作成されていると仮定している。そこで、人間の要約の傾向を学習するために、以下のように文節対残存率の学習を行った。その際は 5.2 の方法で分類した文節対クラス  $(C_k(w_m), C_u(w_n), DRD(w_m, w_n))$  を利用した。

- (1). ある記事セットの原文において係り受け経路を持つ文節対を列挙し、それぞれを文節対クラス  $(C_k(w_m), C_u(w_n), DRD(w_m, w_n))$  で分類する。
- (2). 同じ記事セットの“ 54 文字要約 ”において直接の係り受け関係にある文節対を列挙し、文節対クラス  $(C_k(w_m), C_u(w_n), DRD(w_m, w_n))$  で分類する。直接の係り受け関係にある文節対なので、すべての組で  $DRD(w_m, w_n) = 1$  である。
- (3). (1) の文節対と (2) の文節対で文節対の対応付けを行う。ここでの対応付けは、主辞の原型で比較し、完全に一致すればそれらは同一の文節組とみなす。主辞は“ する ”, “ こと ”, “ なる ”, “ せる ”を除く文節中の最後の自立語とした。(1) で出現した文節対が (2) でも出現した場合、要約を行っても残存した文節対であると考え、その文節対クラスの残存数に 1 を加える。
- (4). (1)-(3) の操作を全ての学習用記事セットに関して行う。
- (5). 文節クラス  $j$  の残存率を以下のように定める。

$$SR(j) \triangleq \frac{\text{文節クラス } j \text{ の残存数}}{\text{文節クラス } j \text{ の出現回数}} \quad (2)$$

### 5.4 係り受け整合度の推定

5.3 で作成した学習結果をもとに、2 文節  $w_m, w_n$  間の係り受け整合度  $p(m, n)$  を次のように定めた。ここで  $j = (C_k(w_m), C_u(w_n), DRD(w_m, w_n))$  とする。

$$p(m, n) = \begin{cases} \log SR(j), & DRD(w_m, w_n) \neq \infty \text{ のとき;} \\ -\infty, & DRD(w_m, w_n) = \infty \text{ のとき.} \end{cases} \quad (3)$$

### 5.5 比較対象とする手法

比較のため、[4] の手法でも簡易実験を行った。この手法は本研究での提案手法と異なり、係り受け整合度の推定において係り受け距離の頻度分布を使用している。ここで係り受け距離とは、直接の係り受け関係にある文節対の文節間距離である。以下に、[4] の手法による係り受け整合度の推定方法を述べる。

### 5.5.1 係り受け距離の頻度分布の学習

文節  $w_m$  と  $w_n$  が直接の係り受け関係にあるとき、それらの文節間距離  $D(w_m, w_n)$  を  $w_m$  と  $w_n$  の“係り受け距離”という。

文節間係り受け整合度を推定するために、まず、係り文節を 5.2 の 1. と同じ分類で、受け文節を 5.2 の 2. と同じ分類で、文節中の形態素の属性に従って分類している。このとき、係り文節  $w_m$  のクラスを  $s = C_k(w_m)$ 、受け文節  $w_n$  のクラスを  $t = C_u(w_n)$  で表す。そして、 $\Theta(s, t)$  を係り文節がクラス  $s = C_k(w_m)$ 、受け文節がクラス  $t = C_u(w_n)$  であるような係り受け文節対クラスの集合全体を  $\Theta(s, t) = \{(w_m, w_n) | c(w_m) = w_m\}$  とし、 $\Theta(s, t)$  の中で、係り文節と受け文節の距離が  $k$  である文節対クラスの集合を  $\Psi^{(s,t)}(k) = \{(w_m, w_n) | (w_m, w_n) \in \Theta(s, t), d(w_m, w_n) = k\}$  とする。そして、以下の式により文節  $w_m$  が文節  $w_n$  に係る相対頻度  $P(w_m, w_n)$  を計算する：

$$P(w_m, w_n) = \frac{|\Psi^{(s,t)}(d(w_m, w_n))|}{|\Theta(s, t)|}.$$

ここで、 $|\cdot|$  は集合の要素数を表す。

### 5.5.2 係り受け規則と文節間係り受け整合度

係り受け規則は、学習データ中に存在する係り受け関係から作成した論理関数で、係り文節クラス  $C_k(w_m)$  に属する文節が受け文節クラス  $C_u(w_n)$  に属する文節に係る例が 1 つ以上存在した場合に真、存在しなかった場合に偽となる。

$$B(C_k(w_m), C_u(w_n)) = \begin{cases} \text{真, 例が存在;} \\ \text{偽, 例がない.} \end{cases}$$

以上のことを踏まえ、諸岡の手法では 2 文節  $w_m, w_n$  間の係り受け整合度  $p(m, n)$  を次のように定めた。

$$p(m, n) = \begin{cases} \log P(w_m, w_n), & B(C_k(w_m), C_u(w_n)) \text{ が真;} \\ -\infty, & B(C_k(w_m), C_u(w_n)) \text{ が偽.} \end{cases} \quad (4)$$

## 6 文節重要度の学習と推定

文節間係り受け整合度を日本語文としての自然さを測る尺度として位置づけているのに対して、文節重要度を簡約前と後で原文の持つ情報が保持されている程度を測る尺度として位置づけている。本論文では、提案した文節間係り受け整合度の有効性を従来手法 [4] と比較するため、文節重要度に関しては [4] と同じ方法で推定した。

### 6.1 文節クラス残存率の学習

文節クラスの残存率の学習を以下のように行った。

1. 文節の分類  $C_s(w_k)$   
文節を主辞品詞に着目して分類した。以下に、分類の属性を示す。
  - 主辞品詞が名詞の場合は、サ変名詞、普通名詞、固有名詞、形式名詞、副詞の名詞、数詞、時相

名詞 別に文節の最後に  
付属語が見つからないもの。  
助詞以外の付属語がつくもの。  
格助詞がつくもの。  
副助詞がつくもの。  
格助詞、副助詞以外の助詞がつくもの。

- 主辞品詞が動詞、副詞、形容詞、指示詞、連体詞、接続詞、感動詞のいずれかであるもの。
- その他: 形態素解析システム JUMAN の解析結果から、未定義語の存在で自立語を含まないと判断された場合。

### 2. 文節残存率の計算

1 の分類方法で分類した文節クラス  $i = C_s(w_k)$  の文節重要度  $q(i)$  を次の手順で定めた。

1. 原文中の文節クラス  $i$  の出現頻度  $F_0(i)$  を求める。
2. “54 文字要約”中の文節クラス  $i$  の出現頻度  $F_1(i)$  を求める。
3. “54 文字要約”における文節クラス  $i$  の残存率  $r(i)$  の計算:  $r(i) = F_1(i)/F_0(i)$ 。
4. 残存率  $r(i)$  の正規化:  $R(i) = r(i)/\sum_i r(i)$ 。

## 6.2 文節重要度の推定

各文節の文節重要度の推定は、6.1 で学習した結果と文節  $w_k$  の TF-IDF 値  $tf-idf(w_k)$  を利用して行う。 $tf-idf(w_k)$  は次のように定義される。

$$tf-idf(w_k) \triangleq TF(A, W(w_k)) \cdot IDF(W(w_k))$$

ここで、文節  $w_k$  の主辞となる単語を  $W(w_k)$ 、 $TF(A, W(w_k))$  は記事  $A$  における単語  $W(w_k)$  の生起頻度である。 $IDF(W(w_k))$  は当日に収集された記事数  $N$  と、 $N$  の中で  $W(w_k)$  が 1 度以上生起する記事数  $DF(W(w_k))$  に関係し、次のように定義される。

$$IDF(W(w_k)) \triangleq \log \left( \frac{N}{DF(W(w_k))} + 1 \right)$$

これらを利用し、文節  $w_k$  の文節重要度  $q(k)$  を次式のように定める。

$$q(k) \triangleq \log (tf-idf(w_k) \cdot R(C_s(w_k))) \quad (5)$$

## 7 簡約文の主観評価実験

### 7.1 簡約文の生成

これまでに示した方法で文節重要度と文節間係り受け整合度を推定し、評価実験の対象となる簡約文を生成した。簡約文の生成には以下の 3 手法を用いた。これらは文節重要度の推定はすべて式 (5) に基づいて行い、文節間係り受け整合度の推定をそれぞれ異なる方法を用いて行った。

従来手法 評価実験のベースラインとする。文節間係り受け整合度の推定に係り受け距離の頻度分布を使用している。(式 (4))

**提案手法** 本研究での提案手法で、文節間係り受け整合度の推定に係り受け経路長を利用した文節対クラスの残存率を利用している。(式(3))

**結合手法** 提案手法と従来手法を利用した手法である。文節間係り受け整合度の推定を以下のように定める。ここで  $prop = "B(C_k(w_m), C_u(w_n))$  が真、かつ  $DRD(w_m, w_n) \neq \infty$  " とする。

$$p(m, n) = \begin{cases} \log(SR(j) \cdot P(w_m, w_n)), & prop \text{ が真} \\ -\infty, & \text{それ以外.} \end{cases} \quad (6)$$

簡約文の作成は、それぞれの手法に対して 70%簡約, 50%簡約, 30%簡約の 3通りの簡約率について行った。ここで  $x\%$ 簡約とは、原文の文節数を  $x\%$ に削減することである。例えば、10 文節の原文を 70%簡約, 50%簡約, 30%簡約した場合、簡約結果はそれぞれ、7 文節, 5 文節, 3 文節になる。また、文節間係り受け整合度に係る重み  $\alpha$  の値は、文献 [10] を参考にして  $\alpha = 0.5$  と設定した。

## 7.2 主観評価実験

被験者に、日本語としての自然性に関する評価、簡約文の情報の保持性に関する評価、総合的な評価、以上 3つの観点から評価をさせた。評価は 0(悪い) から 5(良い) までの 6 段階で、被験者は学生 10 名である。被験者にはまず原文を提示し、それから各簡約率ごとに 3 手法それぞれで簡約された文を提示し、評価をさせた。この評価を 50 記事に関して行った。その評価値の平均を図 2 ~ 図 4 に示す。

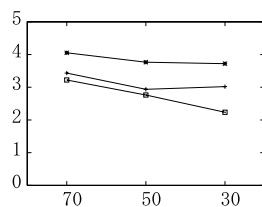


図 2: 自然性の評価

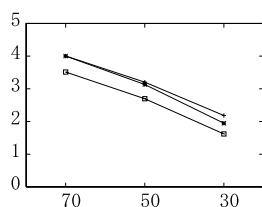


図 3: 重要情報の保持性の評価

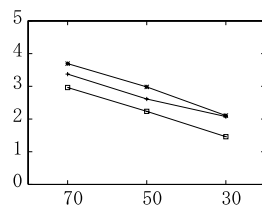


図 4: 総合評価

## 7.3 考察

自然性に関する評価 (図 2) では、3 通り全ての簡約率において、提案手法が従来手法を上回った。このことから、提案手法の係り受け整合度の推定方法が有効であると言える。特に圧縮率の高い 30%簡約の場合でも、提案手法は 50%簡約時と同程度の自然性を確保している。一般的に、高い圧縮率の簡約は文節数が少なくなるために、自然性や重要文節の保持性が下がる傾向があるが、提案手法では下らない。これは、原文の係り受け構造を学習したことで、高い圧縮率の場合でも自然性が確保されていることを示している。また、結合手法は提案手法に比べても高い評価値を得ている。結合手法の場合、係り受け経路長が

同じでも係り受け距離が短いほうに大きな係り受け整合度が与えられる。この違いが自然性の評価の差に結び付いたと考えられる。

重要情報の保持性の評価 (図 3) は各手法で似た変化の傾向を示している。各手法の文節重要度の推定方法が同じであることから、これは妥当である。評価値の差は、係り受け経路長を利用することにより述語にあたる文節が残る場合が多くなったことに因ると考えられる。

総合評価 (図 4) でも、結合手法が最も高く、次いで提案手法、従来手法の順に高い評価値を得た。総合評価の傾向が、情報の保持性の評価を表した図 2 の傾向に近いことから、被験者の感じる簡約の“良さ”は情報の保持性により強く影響されたと考えられる。原文から重要な単語のみを抽出して表示した場合でも文の大意が理解できることを考えると、妥当な結果である。また、図 4 の提案手法による 30%簡約の場合をみたときに、評価値の下降が緩くなっていることがわかる。これは、自然性の評価を表した図 2 の提案手法でみられた傾向であり、総合的な評価においても係り受け経路長が有効に働いていることが読み取れる。

## 8 おわりに

今回、「原文から文節間係り受け整合度と文節重要度の総和が最大になる部分文節列を選択する」方法の枠組の中で、原文の係り受け構造の残存傾向を人間の要約結果から統計的に学習することにより簡約文の自然性を改善することを試みた。そして、得られた簡約文に対して主観評価実験を実施し、提案手法の評価を行った。評価は、従来手法、提案手法、従来手法と提案手法の結合手法に関して行った。その結果、結合手法が最も良い評価値を得、ついで提案手法、従来手法の順に良い評価値を得た。このことから、文簡約問題において係り受け経路長が有効であることが確認された。

今後の課題として、文節重要度の推定法の改善、重み  $\alpha$  の最適値の推定が挙げられる。

## 参考文献

- [1] 奥村学, 難波英嗣, “テキスト自動要約に関する研究動向,” 自然言語処理, 6(6), pp.1-26 (July.1999).
- [2] 奥村学, 難波英嗣, “テキスト自動要約に関する最近の話題,” 自然言語処理, 9(4), pp.1-26 (July.2002).
- [3] 小黒玲, 尾関和彦, 張玉潔, 高木一幸, “文節重要度と係り受け整合度に基づく日本語文簡約アルゴリズム,” 自然言語処理, 8(3), pp.3-18 (July.2001).
- [4] 諸岡祐平, 江崎誠, 高木一幸, 尾関和彦, “重要文抽出と文簡約を併用した新聞記事の自動要約,” 言語処理学会第 10 次年次大会発表論文集, pp.436-439 (May.2004).
- [5] 福富論, 高木一幸, 尾関和彦, “概念距離と係り受けを利用した要約文の文節対応付け,” 言語処理学会第 11 回年次大会, 2005 年 3 月.
- [6] 「毎日新聞全文記事および 54 文字データベース (2002 年度版)」, 毎日新聞.
- [7] 形態素解析システム JUMAN version3.61: <http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html>
- [8] 構文解析システム KNP version2.0 b6: <http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html>
- [9] 京都大学テキストコーパス Version3.0(2000): <http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/corpus.html>
- [10] 諸岡祐平, 小黒玲, 高木一幸, 尾関和彦, “係り受け整合度と文節重要度を用いた自動簡約文の主観評価,” 言語処理学会第 9 回年次大会発表論文集, pp.667-670 (May.2003).