重要文抽出と文簡約を併用した新聞記事の自動要約

諸岡 祐平　江嵜 誠　高木 一幸　尾関 和彦

電気通信大学

{ m_yuu, esaki-m, takagi, ozeki } @ice.uec.ac.jp

1　はじめに

これまで主に検討され，実用されてきた要約手法は重要文抽出法と呼ばれるものである [1]．この手法で作成された要約文は，テキスト中から抽出された重要度の高い文の集合である．そのため，文間のつながり (結束性) の問題を除けば，処理が簡単であるという利点がある．しかし，目的によっては文毎の短縮すなわち，文簡約を行う必要性がある．そこで，重要度の低い文節や単語を削除することによって文を簡約する手法が提案されている．また，そのときに係り受け関係を考慮することによって，原文の部分的な係り受け構造の保持を図る方法も研究されている [2，3]．

本研究では，新聞記事から重要度の高い文を抽出し，更にそれぞれの文を簡約することにより，記事毎の要約を行った．このとき，重要文抽出では，まず見出しと一致度の高い原文を重要視し，次に見出しだけでは重要文の特定が困難な場合，今回使用したデータベースの文章内構造を考慮することで抽出を行った．また，文簡約についてはこれまで,「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」問題として定式化し，効率の良いアルゴリズムを提案してきた [4]．ここでは文簡約アルゴリズムにおける文節間係り受け整合度と文節重要度をデータベースから統計的に推定する方法について述べ，実際に簡約実験を行った結果について報告する．また，文節重要度を $TF \cdot IDF$ で置き換えた場合についても同様に簡約実験を試みた．簡約結果は，実際に携帯端末向けに配信されている要約記事と比較し，要約記事中の文節がどれだけ簡約結果に存在するかを調べることによって評価を行った．

2　データベース

本研究では，新聞記事から重要文を抽出し，それを更に簡約する．簡約結果を評価するためには正解文が必要である．そこで，2002 年 4 月から 2003 年 3 月までの毎日新聞記事と，各記事に対する“ 54 文字要約 ”との対，約 32000 セットを用いた [5]．一つのセットは原記事とそれに対応した“ 54 文字要約 ”のほかにそれぞれの見出しが付けられており，4 つの要素から構成されている．そして，形態素解析システム JUMAN[6] と構文解析システム KNP[7] を用いて解析を行い，形態素情報と係り受け情報を抽出した．表 1 に，実験データの詳細を示す．

表 1: 文セット

セット	文数	用途
A	31201	文節間係り受け整合度と文節重要度の推定.
B	1200	評価実験.

3　重要文抽出

本研究では重要文を抽出する際に，その見出しに着目した．これは，見出しが記事を極端に短縮した要約であり，それに含まれるの個々の単語には重要な情報が含まれていると考えられるからである．しかし，見出しと一致度の高い文を原記事中から抽出する場合，一致度の等しい文が複数存在することがあり，特定が困難である．そのため，本研究では，今回使用したデータベースの文章内構造を調べ，見出し情報だけでは重要文の特定が困難である場合には，その結果を併用した．

3.1　文章内構造

一般に文章の総括形式には，冒頭で総括するもの，結尾で総括するもの，冒頭と結尾で総括するもの，中程で総括するものがある [8]．そのため重要文抽出の方法は，文章の種類によって異なったアプローチを取る必要があると考えられる．

3.2　原記事と“ 54 文字要約 ”のマッチング

“ 54 文字要約 ”と原記事中の全ての文に対しマッチングを行った．ここで，一致度は，それぞれの文中に存在する自立語の一致数で表す．そして，一致数のもっとも大きかった原記事中の文を“ 54 文字要約 ”に対応する文とする．その“ 54 文字要約 ”の多くは 2 文からなっている．以下に示すのは，その第 1 文を対象に調査を行った結果である．横軸は原文の原記事における段落番号を示している．また，縦軸は，全ての原文に対して段落別の相対出現頻度を示したものである．

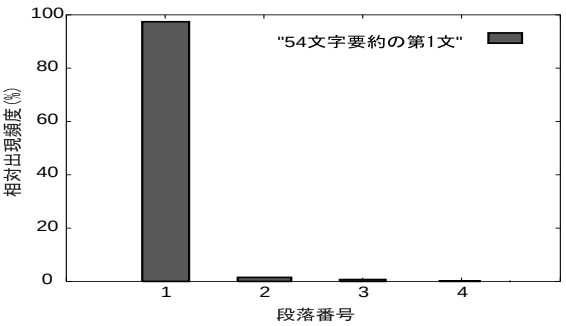

図 1:“ 54 文字要約 ”に対応する原文の段落別相対出現頻度

この結果から，第 1 段落における原文の相対出現頻度が他の段落に比べて明らかに大きいことが分かる．その

ため,“ 54 文字要約 ”に含まれる第 1 文の原文となる文は，原記事の第 1 段落中に高い割合で存在していることが分かる．

図 2 に示すのは，段落を文単位で分割し，段落中のどの位置に“ 54 文字要約 ”の原文が存在するかを調べたものである．この図より，第 1 段落，第 2 段落に共通し

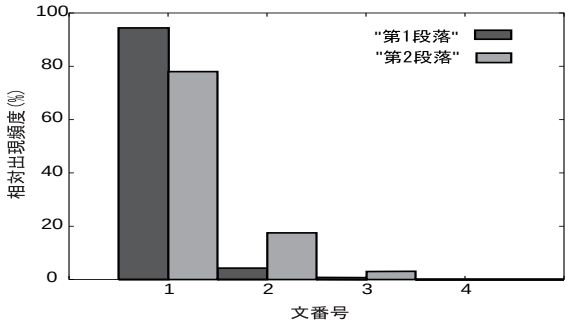

図 2: 段落別に見た原文の文番号別相対出現頻度

て，それぞれ段落の先頭に“ 54 文字要約 ”の原文となる文が現れていることがわかる．図 1 から第 1 段落に原文が集中していることから，第 1 段落に着目してみると，第 1 文に高い頻度で原文が現れていることが分かる．

図 1，2 の結果から“ 54 文字要約 ”に最もよく対応する原文は第 1 段落の先頭部分の文である．また,“ 54 文字要約 ”の多くが 2 文から構成されているため，第 2 文に対しても図 1，2 と同じ調査を行った．その結果，図 1，2 とほぼ同じ調査結果が得られた．

今回の調査より，本研究に用いたデータベースの文章内構造が冒頭統括の形式を取ること，つまり,“ 54 文字要約 ”の原文となる文が記事中の先頭部分に集中していることが確認できた．そのため，重要文の特定に関しては，まず，原記事見出しとの一致度を原記事中全ての文に対し計算し，一致度の高かった上位 2 文を重要文として抽出した．しかし，一致度の等しい文が複数存在する場合は，文章内構造を考慮し，先頭部分に近い文を優先させた．

4 文簡約

われわれの提案する手法において文簡約とは，文を文節の列と捉え，原文からできるだけ“ 良い ”部分文節列を抽出することである．そのためには，この部分文節列の“ 良さ ”を計る評価関数が必要である.，簡約された文の“ 良さ ”を概念的に定義すると，

a) 原文の持つ情報をできるだけ保持している，

b) 日本語として構文的にできるだけ自然である，

という 2 点が考えられる．そこで，文の“ 良さ ”を 2 つの概念それぞれに対応した評価関数の値の和として，以下のように定義する．

まず，原文を文節の列 $w_0 w_1 \ldots w_{M-1}$ と仮定し，その中の長さ l の部分文節列 $w_{k_0} w_{k_1} \ldots w_{k_{l-1}}$ を考える．ここで，各文節 w_m の重要度を表す関数 $q(m)$ が与えられているとすると，この部分文節列の重要度はそれらの総和

$$\sum_{i=0}^{l-1} q(k_i) \tag{1}$$

という評価関数で計ることができよう．

また，文節 w_m が文節 w_n に係るときの係り受け整合度 $p(m,n)$ が与えられているとすると，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 $w_{k_0} w_{k_1} \ldots w_{k_{l-1}}$ 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

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

で表される． このとき， c は次の条件を満たさなければならない．

A) 後方単一性: $k_m < c(k_m)$.

B) 非交差性: $m < n$ ならば $c(k_m) \leq k_n$ ，
または $c(k_n) \leq c(k_m)$.

本研究では，写像 c を用いて，文節列 $w_{k_0} w_{k_1} \ldots w_{k_{l-1}}$ の日本語としての構文的な自然性の程度を

$$\max_c \sum_{i=0}^{l-2} p(k_i, c(k_i)) \tag{3}$$

で計ることとする．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．

以上のような，重要度を表す式 (1) と構文的な自然性の程度を表す式 (3) に基づいて，本論文では文節列 $w_{k_0} w_{k_1} \ldots w_{k_{l-1}}$ の“ 良さ ”を計る評価関数 $g(k_0, k_1, \ldots, k_{l-1})$ を次のように定義する [4].

$$g(k_0, k_1, \ldots, k_{l-1}) \triangleq \begin{cases} q(k_0), \quad l = 1 \text{ のとき;} \\ \alpha\{\max_c \sum_{i=0}^{l-2} p(k_i, c(k_i))\} \\ +(1-\alpha)\{\sum_{i=0}^{l-1} q(k_i)\}, \ 2 \leq l \text{ のとき.} \end{cases} \tag{4}$$

式 (4) 中のα は係り受け整合度にかかる重みである．文簡約問題を評価関数 $g(k_0, k_1, \ldots, k_{l-1})$ を最大にする部分文節列を求める問題として定式化ができる．この問題は，動的計画法の原理に基づき効率良く解くことが出来る [4].

5 係り受け整合度の推定

5.1 文節の分類

係り受け整合度を推定するため，係り文節と受け文節を次のような文節中の形態素の属性に着目して分類した．

係り文節: 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

受け文節: 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

その結果，係り受け整合度学習文セット中の文節は，198 種類の係り文節と 61 種類の受け文節に分類された．

5.2 係り受け規則の作成

係り受け規則 B は，学習文セット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 に属する文節が受け文節クラス C_u 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定める:

$$B(C_k, C_u) = \begin{cases} \text{真}, & \text{例が存在;} \\ \text{偽}, & \text{例がない.} \end{cases} \tag{5}$$

5.3 係り受け整合度

本実験では，係り受け整合度 $p(x, y)$ を以下のように定義した．

$$p(x, y) = \begin{cases} \log P(x, y), & B(C_k, C_u) \text{ が真;} \\ -\infty, & B(C_k, C_u) \text{ が偽.} \end{cases} \tag{6}$$

ここで，C_k, C_u は，それぞれ文節 x, y が属するクラスである．また，$P(x, y)$ は C_k, C_u，および y が文末文節か否かの別が与えられたときの x，y 間の係り受け距離の相対頻度である [9].

6 文節重要度の推定

6.1 文節の分類

文節重要度を推定するため，文節を次のように文節中の主辞品詞に着目して分類した．

- 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，副詞的名詞，数詞，時相名詞**別に文節の最後に
  — 付属語がつかないもの．
  — 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．

6.2 文節重要度

本研究では，原記事中から“ 54 文字要約 ”に対応すると思われる原文を抽出し，“ 54 文字要約 ”とのマッチングを行った．このとき，文節単位で比較を行い，もし，それぞれの文中に存在する 2 つの文節について主辞となる単語の原型が同じ場合，その文節は同一であるとした．以上の考え方に基づいて文節クラス i の文節重要度 $q(i)$ を次の手順で定めた [10].

1. 原文中の文節クラス i の出現頻度 $C_0(i)$ を求める．
2. “ 54 文字要約 ”中の文節クラス i の出現頻度 $C_1(i)$ 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：$R(i) = C_1(i)/C_0(i)$.
4. 残存率 $R(i)$ の正規化：$F(i) = R(i)/\sum_i R(i)$.
5. 文節クラス i の文節重要度 $q(i)$ の計算：$q(i) = \log F(i)$

7 $TF \cdot IDF$の算出

今回，われわれの提案している文節重要度とは別に，これを $TF \cdot IDF$で置き換えた場合についても同様に簡約実験を試みた．ここでは，文節重要度と同様，文節毎に重みを持たせるために文節中の主辞となる単語の $TF \cdot IDF$ を算出する．$TF \cdot IDF$ 値を求める単語を W_i とする．以下の式により記事 A に出現する単語 W_i の $TF \cdot IDF$ 値: $TF(A,W_i) \cdot IDF(W_i)$ を求める．

$TF(A, W_i)$ は記事 A における単語 W_i の生起頻度である． $IDF(W_i)$ は当日に収集された文書数 N と，その中で W_i が 1 度以上生起する文書数 $DF(W_i)$ に関係し，次のように定義される [11].

$$IDF(W_i) = \log\left(\frac{N}{DF(W_i)} + 1\right) \tag{7}$$

8 評価実験

これまでに得られた文節間係り受け整合度と文節重要度を用いて，評価実験の対象となる簡約実験を行った．簡約方法は，文節重要度のみを用いた場合， $TF \cdot IDF$ のみを用いた場合，文節重要度と $TF \cdot IDF$ を併用した場合，の 3 通りである．ここで，簡約は 5 通りの簡約率: 80%，65%，50%，35%，20%，それぞれにおいて行った．

8.1 評価尺度

それぞれの簡約結果は，原記事に対応する“ 54 文字要約 ”と比較し，その一致度を調べた．ここで“ 54 文字要約 ”との一致度は，それぞれの文中に存在する文節の一致数で表す．

8.2 文節重要度 $q(i)$ のみを用いた場合

4.1 節で述べたアルゴリズムに従い文簡約を行った．ここでは，$\alpha = 0.1, 0.9$ それぞれの結果を示す．結果を図 3 に示す．

8.3 $TF \cdot IDF$ のみを用いた場合

文節重要度を$TF \cdot IDF$で置き換えた場合についても同様に簡約実験を試みた．結果を図 4 に示す．

8.4 文節重要度 $q(i)$ と $TF \cdot IDF$ を併用した場合

$q(i)$ と $TF \cdot IDF$ を併用して簡約を行った場合の結果を図5に示す．このとき，文節 w_i の重み $MIX(w_i)$ を以下のように定義した．

$$MIX(w_i) = \log\{F(i) \cdot (TF \cdot IDF(A, w_i))\} \quad (8)$$

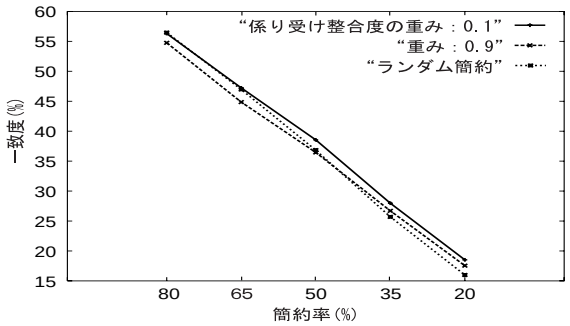

図 3: 文節重要度のみを用いた場合

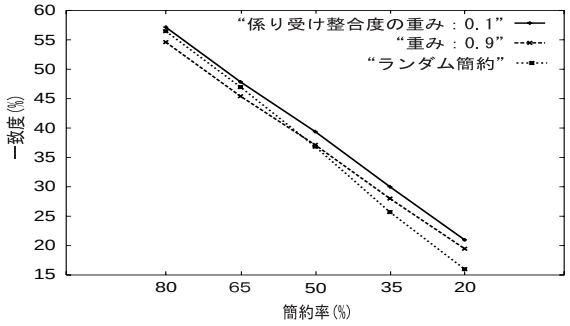

図 4: $TF \cdot IDF$ のみを用いた場合

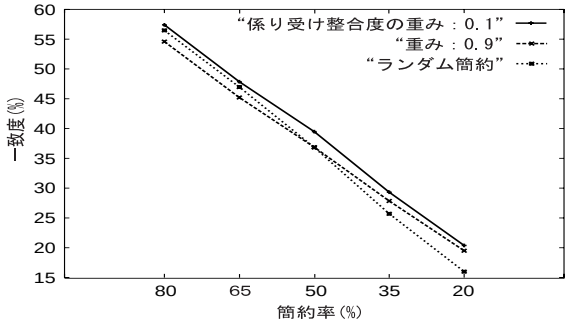

図 5: 文節重要度と $TF \cdot IDF$を併用した場合

今回, (文節重用度と$TF{\cdot}IDF$の併用) > ($TF{\cdot}IDF$のみ) > (文節重用度のみ) > (ランダム簡約) の順に結果が良かった．結論としては，2通りの簡約方法ともランダム簡約の結果よりは良好な値を得られたことから，本手法の有効性を確認できた．

9 おわりに

今回の簡約実験においては，提案手法による簡約文の方がランダム簡約の場合よりも，正解となる54文字要約 ”との一致度が高いということが分かった．このことから，文節重要度と$TF \cdot IDF$が自動簡約において有効であると考えられる．しかし，この評価法では，日本語文としての自然性を評価することが出来なかった．そのため今後の課題として，自然性に対する主観評価を行うことが挙げられる．

10 謝辞

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)」を使用した．また，今回用いた自動簡約プログラムは小黒玲氏によって作成されたものである．

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] 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，**6**(7)，pp.73-92(1999)．

[3] 三上真，増山繁，中山聖一：“ 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，”自然言語処理，**6**(6)，pp.65-81(1999)．

[4] 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(2001)．

[5]「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．

[6] 形態素解析システム JUMAN
http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[7] 構文解析システム KNP
http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[8] 市川孝：“ 国語教育のための文章論概説，”教育出版，1978．

[9] 張玉潔，尾関和彦：“ 文節間係り受け距離の統計的性質を用いた日本語文の係り受け解析，”自然言語処理，**4**(2)，pp.3-19(1997)．

[10] Rei Oguro, Hiromi Sekiya, Yuhei Morooka, Kazuyuki Takagi, and Kazuhiko Ozeki:“ Evaluation of a Japanese Sentence Compression Method Based on Phrase Significance and Inter-Phrase Dependency, ”Proc. TSD2002(LNAI2448),pp.27-32 (2002).

[11] 大森岳史，増田英孝，中川裕志：“ 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，2003-NL-153，pp.1-8(2003)．