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

諸岡 祐平　小黒 玲　高木 一幸　尾関 和彦

電気通信大学

{ m_yuu, rei, takagi, ozeki } @ice.uec.ac.jp

1 はじめに

今日，電子化されたテキストが世の中に溢れ，様々な情報をインターネット等から入手する機会が多くなった．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必要性が高くなり，大量の関連情報から必要な部分だけを抽出する種々の自動要約技術が提案されている [1].

現在，要約文の作成・表示は，概ね重要文抽出法と呼ばれる手法を使って実現されている．ここでは，文書中の文を抽出の単位として考え，それぞれの重要度を計算する．そして，要約は，重要度の高い順に文を取り出すことで作成する．この手法は，抽出した複数の文間のつながり (結束性) の問題を別にすれば処理が簡単であるという利点がある．しかし，要約の目的によっては，更に要約率を上げるため，個々の文を簡約することが必要となることがある．そのため，特にニュース字幕作成を目的として，表層文字列の変換を行い，1文の文字数を減らすなどの研究 [2, 3] が行われている．また，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も提案されており，削除選択に係り受け関係を考慮することで，原文の部分的な係り受け構造の保持を図る方法 [4] も研究されている．我々はこれまで，文簡約を「原文から，文節重要度と係り受け整合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，それを解くための効率の良いアルゴリズムを提案してきた [5]．本稿では，係り受け整合度と文節重要度を具体的に設定して，このアルゴリズムに基づいて得られた自動簡約文に対し，主観評価を行った結果について報告する．

2 文簡約アルゴリズム

われわれの提案する手法において文簡約とは，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．ここで，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，原文を複数の文節からなる列 $w_0 w_1 \ldots w_{M-1}$ 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

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

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

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

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．

(1) 後方単一性: $k_m < c(k_m)$.
(2) 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

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

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，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 [5].

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

式 (4) 中の α は係り受け整合度にかかる重みである. ここでは評価関数 $g(k_0,k_1,\ldots,k_{l-1})$ を用いて, M 文節からなる原文を l 文節に簡約する問題を, 文節列 $w_0w_1\ldots w_{M-1}$ の部分文節列 $w_{k_0}w_{k_1}\ldots w_{k_{l-1}}$ の中で, 関数 $g(k_0,k_1,\ldots,k_{l-1})$ を最大にするものを求める問題として定式化する. また, 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 [5].

3 データベース

本研究では, 京都大学テキストコーパス Version3.0[6] を用いた. このテキストコーパスは, 毎日新聞 1995 年 1 月 1 日から 17 日までの一般記事約 2 万文, 1 月から 12 月までの社説記事約 2 万文, 計約 4 万文に対して, 京都大学の形態素解析システム JUMAN, 構文解析システム KNP で自動解析を行い, その結果を人手で修正したものである. このコーパスには, 各文の係り受け構造と各文節を構成する形態素の品詞情報が付与されている.

4 パラメータの推定

4.1 係り受け整合度

4.1.1 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, 219 種類の係り文節と 118 種類の受け文節に分類された.

4.1.2 係り受け規則の推定

係り受け規則は, 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なる.

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

4.1.3 係り受け整合度の推定

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.

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

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 [7].

4.2 文節重要度

4.2.1 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**で文節の最後に
  - — 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.

4.2.2 文節重要度の推定

文節重要度は，文簡約において重要な要素である．本実験では，まず，人手による簡約実験を行った．被験者には，各文を簡約率80%，65%，50%，35%，20%の5通りにそれぞれ簡約するよう指示した．ここで，簡約率は原文中の文節数に対する簡約文中の文節数の割合である．以下に，$q(i)$ の計算手順を示す．

1. 原文中の文節クラス i の出現頻度 $C(i)$ を求める．
2. 簡約率 k %のとき，簡約文中の文節クラス i の出現頻度 $C(i,k)$ を求める．
3. 簡約率 k %の簡約文における文節クラス i の残存率 $R(i,k)$ の計算:
   $R(i,k) = C(i,k)/C(i)$.
4. 残存率 $R(i,k)$ の正規化:
   $F(i,k) = R(i,k)/\sum_i R(i,k)$.
5. 5通りの簡約率に対する $F(i,k)$ の平均化:
   $F(i) = (\sum_k F(i,k))/5$.
6. 文節クラス i の文節重要度 $q(i)$ の計算:
   $q(i) = \log F(i)$.

4.3 予備実験

係り受け整合度にかける重み α の最適な値を推定するため，予備実験を行った．ここでは，4.1，4.2で得られた係り受け整合度と文節重要度を設定し，本手法で日本語文を自動簡約した．文は5通りの簡約率で簡約され，α の値は0.1刻みで変化させた．そして，自動簡約した各文の総合的な印象を1（悪い）から6（良い）の6段階で被験者に評価させた．表1に，α の各値における，簡約文の評価値の平均値を示す．この結果，$\alpha = 0.5$ のときに最も高い評価が得られることが確認された．

表 1: 重み α による評価値の変化

α	0.3	0.4	0.5	0.6	0.7	0.8
平均	3.37	3.23	3.40	3.27	3.24	3.19

5 簡約実験

4.1，4.2で得られた係り受け整合度と文節重要度を用いて，主観評価の対象とする簡約文を生成した．ここで，係り受け整合度にかける重み α の値は，$\alpha = 0.5$ に設定した．自動簡約は5通りの簡約率それぞれにおいて行った．また，比較のため，被験者1名が同じ文セットを簡約し，さらに，無作為に文を簡約する実験[1]も行った．このように，原文から15通り（3通りの簡約方法×5通りの簡約率）の簡約文を生成した．

6 主観評価

被験者に，簡約文の総合的評価，情報の保持に関する評価，日本語としての構文的な自然さに関する評価をさせた．評価は予備実験と同じ6段階とし，被験者には，まず原文を見せ，それから15通りの簡約された文をランダムに提示した．被験者には提示される文がどのように簡約されたものかは教えなかった．

6.1 総合的評価

目標とする簡約文は日本語文として自然であり，かつ，原文の持つ情報を出来るだけ保持していなければならない．最初の評価は，簡約文がこの2つの条件を充分に満たしているかを被験者に総合的に判断させたものである．図1はシステム，人

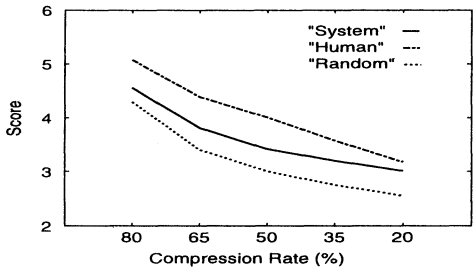

図 1: 総合的評価

が簡約したとき，そして，ランダム簡約を行ったとき，簡約率別に被験者の評価値の平均を求めたものである．ここで“システム”は提案手法を指し，その評価は5通り全ての簡約率においてランダム簡約を上回った．このことから，係り受け整合度と文節重要度が文簡約において有効であると言える．

[1]以後，ランダム簡約と呼ぶ．

6.2 情報の保持に関する評価

簡約文は，原文の持つ重要な情報を出来るだけ保持していなければならない．そこで，被験者に簡約文が原文の持つ情報をどの程度保持しているか，その点に限定して評価させた．

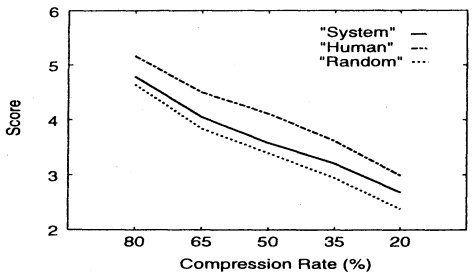

図 2: 情報の保持に関する評価

図 1 の総合評価の結果と同様，システムの評価は 5 通り全ての簡約率においてランダム簡約を上回った．このことからも，文節重要度は文簡約において有効であると言える．複数の被験者に共通して聞かれた感想は，「固有名詞の主語が簡約文で削除されている場合が多少あり，そのときの評価が比較的小さくなる」というものだった．このことから，文節重要度を計算するとき，重要度の差が顕著に表れるような文節クラスの検討が必要であることが分かった．

6.3 日本語としての構文的自然性評価

簡約文にとって，日本語として構文的に自然であることは重要である．そのため，簡約文を独立した文と考えたとき，その自然性を被験者に判断させた．なお，簡約例の中には簡約率の関係で述語が削除されている場合がある．そのため，そうした場合は日本語句としてその自然さを被験者に判断させた．

図 3 の結果を見る限り，システムによる簡約に対して人が簡約した場合と同等と言えるほど良好な評価が得られている．

7 おわりに

係り受け整合度と文節重要度を用いて簡約実験を行い，その主観評価の結果を報告した．被験者

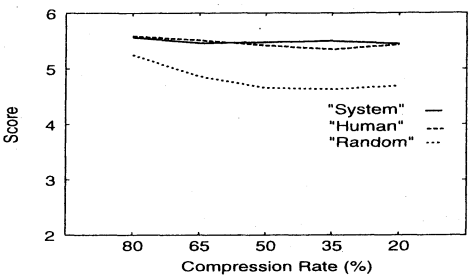

図 3: 構文的自然性評価

に日本語としての構文的自然さに焦点を絞って評価をさせたときには，人が簡約したときと同等の良好な結果が得られた．

また，今後の課題として，重要度の差をより顕著に表すような，文節クラスの分類がある．そのためには，重要文節選択における人のヒューリスティックスを調べ，文節クラスの分類に多く採り入れていくことが有用と思われる．

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] 若尾孝博，江原暉将，白井克彦：“テレビニュース番組の字幕に見られる要約手法，”情報処理学会自然言語処理研究会，97-NL-122-13，pp.83-89(1997).

[3] 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).

[4] 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).

[5] 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).

[6] 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

[7] 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).