

文節間係り受け整合度と文節重要度を用いて自動簡約した 日本語文の主観評価

諸岡 祐平 小黒 玲 高木 一幸 尾関 和彦

電気通信大学

〒 182-8585 東京都調布市調布ヶ丘 1-5-1

Tel. 0424-43-5374

E-mail: { m_yuu, rei, takagi, ozeki } @ice.uec.ac.jp

あらまし 現在,主に検討されている要約手法はテキスト中から重要文を抽出するものである。しかし,目的によっては,一文毎の簡約が必要と考えられる。われわれは,与えられた文から何らかの意味で,最適な部分文節列を抽出することにより,一文毎の簡約を行う方法を研究している。そのため,文簡約を「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た。本稿では,この自動簡約において必要な文節間係り受け整合度と文節重要度を頻度統計から推定する方法について述べ,さらに「提案した手法に基づいて自動簡約した日本語文」「手で簡約した日本語文」「無作為に簡約した日本語文」の3通りの簡約文をそれぞれ複数の被験者に主観評価させた結果について報告する。実験結果から,文節間係り受け整合度と文節重要度の文簡約における有効性が確認された。

キーワード 文簡約,文節間係り受け整合度,文節重要度

Subjective Evaluation of Japanese Sentences Automatically Compressed by Using Inter-phrase Dependency and Phrase Significance

Yuhei MOROOKA Rei OGURO Kazuyuki TAKAGI Kazuhiko OZEKI

The University of Electro-Communications

1-5-1 Chofugaoka, Chofu, Tokyo 182-8585

Tel. 0424-43-5374

E-mail: { m_yuu, rei, takagi, ozeki } @ice.uec.ac.jp

Abstract Most text summarization methods reported so far are based on the idea of selecting significant sentences from a text. However, depending on the purpose, it is sometimes necessary to shorten, or compress, each sentence in a text. We formulated the problem of sentence compression as an optimization problem of extracting a subsequence of phrases from the original sentence that maximizes the sum of inter-phrase dependency score and phrase significance score. In this paper, we report a result of subjective evaluation for the quality of sentences compressed by using the proposed algorithm, comparing it with compression by a human and random compression. The result shows the effectiveness of inter-phrase dependency and phrase significance in automatic sentence compression.

Keywords sentence compression, inter-phrase dependency, phrase significance

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、短時間での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]。市販の翻訳ソフトの中には、翻訳機能の他に要約機能を装備するものが増えてきている。それらは、ウェブ上の文書を見るとき、ブラウザ上の内容の重要部分を要約・翻訳する機能を有する。

現在、要約文の作成・表示は、概ね重要文抽出と呼ばれる手法を使って実現されている。この重要文抽出という手法では、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。作成される要約文は、テキスト中の色々な箇所から抽出された文の集合である。そのため、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、文節重要度と文節間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、文節間係り受け整合度と文節重要度を具体的に設定し、このアルゴリズムに基づいて得られた文簡約結果に対し、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) 作成された簡約文が原文の持つ情報をできるだけ保持していること、
- b) 作成された簡約文が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点が考えられる。そこで、この“良さ”をこの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 $w_0 w_1 \dots w_{M-1}$ 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。もし、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

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

で計ることができる。ここでは、これを部分文節列の重要度を計る評価関数とする。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 $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$ 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

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

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 $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$ の日本語としての構文的な自然性の程度を

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

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 $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$ の“良さ”を計る評価関数 $g(k_0, k_1, \dots, k_{l-1})$ を次のように定義する [5]。

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

ここでは評価関数 $g(k_0, k_1, \dots, k_{l-1})$ 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 $w_0 w_1 \dots w_{M-1}$ の部分文節列 $w_{k_0} w_{k_1} \dots w_{k_{l-1}}$ 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問題として定式化する. そして, この問題を効率良く解くアルゴリズムが提案されている [5]. α は文節間係り受け整合度にかかる重みである.

3 データベースと被験者

3.1 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 [7], 構文解析システム KNP [8] 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.

表 1 に, 文節間係り受け整合度と文節重要度の推定, および, 文簡約実験のために用意した種々の文セットを示す. これらの文セットの間の包含関係は次の通りである: $B \cap A = \phi$, $C \subset B$, $D \cap A = \phi$, $D \cap B = \phi$.

表 1: 文セット

セット	文数	用途
A	34848	文節間係り受け整合度の推定.
B	200	文節重要度の推定.
C	20	α の推定.
D	200	主観評価のための簡約文生成.

3.2 被験者

本研究では, 文節重要度の推定, α の推定, 最終的な主観評価などに被験者を用いた. そのための被験者グループを表 2 に示す. グループの包含関係は次の通りである: $Y \subset X$, $Z \subset Y$, $\#(W \cap X) = 2$, $Z \cap W = \phi$.

表 2: 被験者グループ

グループ	人数	詳細
X	13	文節重要度の推定.
Y	3	α の推定.
Z	1	人手による簡約.
W	5	最終主観評価.

4 パラメータの推定

4.1 文節間係り受け整合度

4.1.1 文節の分類

文節間係り受け整合度を推定するため, まず, 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, セット A 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, セット A 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.

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

4.1.3 係り受け整合度の推定

本実験では, 文節間係り受け整合度 $p(x, y)$ 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 内の係り受け距離の相対頻度である [9].

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, まず, 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.

4.2.2 文節重要度の推定

文節重要度は文章の要約問題, 文の簡約問題で重要な要素である. 本実験では, 文節重要度 $q(i)$ を以下のように定義した.

$$q(i) = \log F(i) \quad (7)$$

ここで, $F(i)$ は簡約文中に現れる文節クラス i の相対頻度から得られた値である. 本研究では, $F(i)$ を求めるため, 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. この簡約実験に使用した文セットは表 2 のセット B, 参加した被験者は表 2 のグループ X 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 $q(i)$ の定義:

$$q(i) = \log F(i).$$

4.3 重み α

文節間係り受け整合度にかける重み α の最適な値を決めるため, 予備実験を行った. 4.1, 4.2 で得られた文節間係り受け整合度と文節重要度を使って, 文セット C 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 段階で被験者グループ Y に評価させた. 表 3 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, $\alpha = 0.5$ のときに最も良い評価が得られることが確認された.

表 3: 重み α による評価値の変化

α	0.1	0.2	0.3	0.4	0.5	0.6	
平均	3.33	3.34	3.37	3.23	3.40	3.27	

0.7	0.8	0.9
3.24	3.19	3.14

5 簡約文の生成

4.1, 4.2 で得られた文節間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, 文節間係り受け整合度にかける重み α の値は, 予備実験で得られた $\alpha = 0.5$ 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 $\times$ 5 通りの簡約率) の簡約文を生成した.

¹以後, ランダム簡約と呼ぶ.

原文としては表 1 の文セット D を使用した．文セット D に含まれる各文の長さ（文節数）は，7 文節（短文）が 66 文，13～14 文節（中文）が 68 文，18～50 文節（長文）が 66 文であった．

6 主観評価実験

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 1(悪い) から 6(良い) の 6 段階，被験者はグループ W である．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ず，評価には十分な時間を与えた．この作業を文セット D に含まれる 200 文全てに対し行った．評価実験の結果は被験者毎に，簡約率別の平均を求めた．さらに，被験者全体の平均を求めた．

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 つの条件を十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．なお，被験者が評価する際の判断基準を以下のように設定し，被験者に提示した．

- 6: 簡約文として上記の条件を十分に満たしている．
- 5: 条件に不足があるが，簡約文と見なすことが出来る．
- 4: 簡約文としてかろうじて見なせる．
- 3: 簡約文とは言えないが，忠実さ，自然さがかろうじて保持されている．
- 2: 文意が全く通じない．
- 1: 文の意味が変わっている．

表 4: 評価値の被験者に対する標準偏差

	簡約率 (%)				
	80	65	50	35	20
System	±0.36	±0.16	±0.12	±0.12	±0.10
Human	±0.45	±0.29	±0.26	±0.24	±0.17
Random	±0.32	±0.09	±0.14	±0.22	±0.28

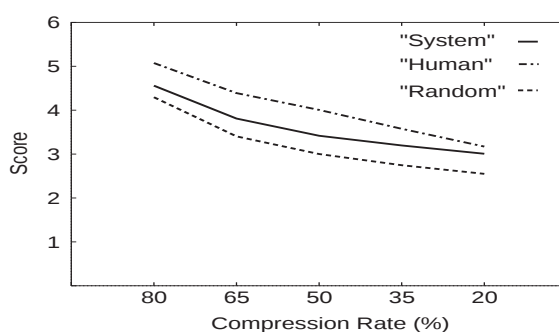


図 1: 総合的評価

図 1 はシステム，人が簡約したとき，そして，ランダム簡約を行ったときの簡約率別の評価の平均値を示したものである．ここで“システム”は提案手法を指している．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，文節間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．

また，簡約率が小さくなるほど，その評価結果は下がっている．表 4 を見ると，被験者間で評価値にかなりのばらつきがあることが分かる．図 2，図 3 に示すのは，評価値の下がり方の傾向にかなりの違いが見られた 2 名の被験者の結果である．被験者 I の場合，簡約率 80% で高い評価を与えているが，簡約率が小さくなるにつれ評価は明らかに低くなっている．一方で，被験者 II の場合では，各簡約率における評価の変動が比較的小さい．簡約率 80% における評価は決して高くはないが，簡約率 20% における評価が非常に低いというわけでもない．このことから，予め評価の判断基準を設定していても，被験者間に明確な個人差が生ずることが分かった．

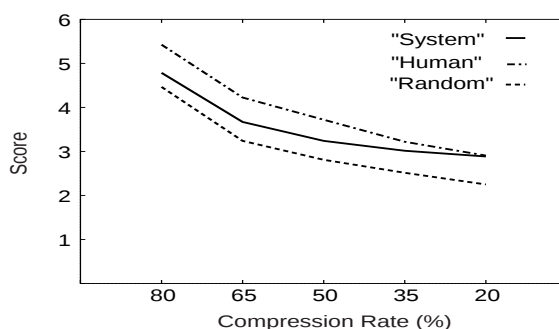


図 2: 被験者 I

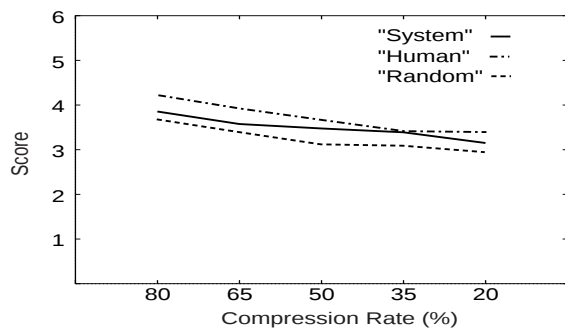


図 3: 被験者 II

6.2 情報の保持に関する評価

簡約文は，原文で重要と考えられる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という点だけについて評価させた．評価基準，及び結果を以下に示す．

- 6: 原文の重要な情報が全て残っている．
- 5: " 十分残っている．
- 4: " かなり残っている．
- 3: " 多少残っている．
- 2: " わずか残っている．
- 1: " 全く残っていない．

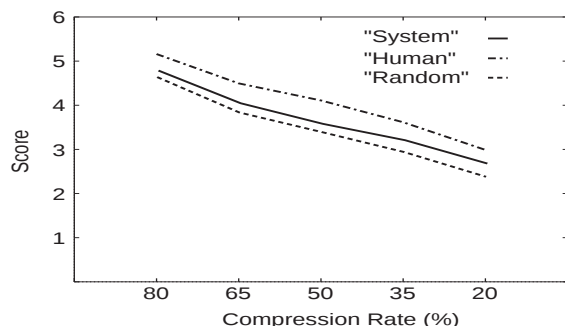


図 4: 情報の保持に関する評価

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，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，文にまとまりがあり，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．判断基準は以下の通りである．

- 6: 全く不自然さが無い．
- 5: わずかに不自然さがある．
- 4: 多少 " ．
- 3: かなり " ．
- 2: 非常に " ．
- 1: 日本語として全く通じない．

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．

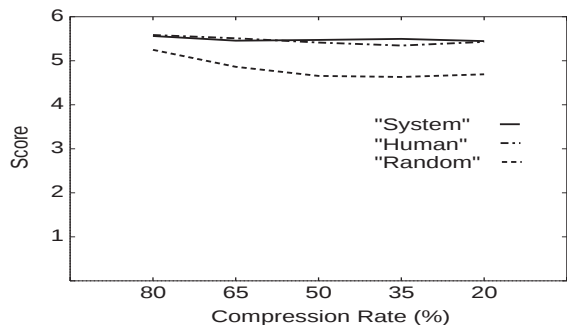


図 5: 構文的自然性評価

図 5 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 簡約例

以下に，本論文の評価実験で使用した簡約文のうち，システムが出力した [自動簡約結果] と [人による簡約結果] [ランダム簡約結果] を示す．最初に原文，それ以降は簡約率 80%，65%，50%，35%，20% の順に簡約例が並んでいる．なお，文中の “ / ” は，文節境界を示している．

“原文 1”は、システムと人との間で簡約文に現われる文節の傾向に違いの見られたものである。ここで原文の主旨となる情報は「北海道内で局地的に雪が強まっている」と考えられる。そして、その情報を補足するように「札幌市南区小金湯」「胆振管内早来町安平」という場所での降雪量が述べられている〔自動簡約結果〕では主旨を補足する情報のみが抽出されていて、原文の主旨がくみ取り難くなっている。また〔自動簡約結果〕と〔人による簡約結果〕を比較したとき、自動簡約結果には“北海道内は/”という文節が一度も現われていない。同じ地名を含む文節としては、“胆振管内早来町安平で/”という文節が4通りの簡約率で現われている。このことから、人は同一文内に似た属性を持つ文節が複数存在するとき、上位の属性を持つ文節を抽出する傾向にあるといえる。

“原文 2”は、システムが人と共通した主旨を抽出できている例である〔人による簡約結果〕と比較したとき、二つの簡約例で抽出されている文節は多少異なる。しかし、その主旨は概ね一致しており、原文の主旨からみても重要な情報が簡約文中にかなり保持されているように考えられる。

ランダム簡約結果を見ると、どちらの例においても明らかに構文的自然性に問題がある。このことは、図 5 にも明確に表れている。

-原文 1-

- 北海道内は/ 六日/ 午後から/ 局地的に/ 雪が/ 強まり、/ 札幌市南区小金湯で/ 同九時までの/ 六時間に/ 一九センチ、/ 胆振管内早来町安平で/ 一六センチなどの/ 降雪を/ 記録した。

〔自動簡約結果〕

- 80%: 六日/ 午後から/ 局地的に/ 雪が/ 強まり、/ 同九時までの/ 六時間に/ 一九センチ、/ 胆振管内早来町安平で/ 一六センチなどの/ 降雪を/ 記録した。
- 65%: 雪が/ 強まり、/ 同九時までの/ 六時間に/ 一九センチ、/ 胆振管内早来町安平で/ 一六センチなどの/ 降雪を/ 記録した。
- 50%: 同九時までの/ 六時間に/ 一九センチ、/ 胆振管内早来町安平で/ 一六センチなどの/ 降雪を/ 記録した。
- 35%: 一九センチ、/ 胆振管内早来町安平で/ 一六センチなどの/ 降雪を/ 記録した。
- 20%: 一六センチなどの/ 降雪を/ 記録した。

〔人による簡約結果〕

- 80%: 北海道内は/ 六日/ 局地的に/ 雪が/ 強まり、/ 札幌市南区小金湯で/ 六時間に/ 一九センチ、/ 胆振管内早来町安平で/ 一六センチなどの/ 降雪を/ 記録した。
- 65%: 北海道内は/ 雪が/ 強まり、/ 札幌市南区小金湯で/ 一九センチ、/ 胆振管内早来町安平で/ 一六センチなどの/ 降雪を/ 記録した。

- 50%: 北海道内は/ 雪が/ 強まり、/ 一九センチ、/ 一六センチなどの/ 降雪を/ 記録した。
- 35%: 北海道内は/ 雪が/ 強まり、/ 降雪を/ 記録した。
- 20%: 北海道内は/ 雪が/ 強まり、

〔ランダム簡約結果〕

- 80%: 北海道内は/ 六日/ 午後から/ 局地的に/ 雪が/ 同九時までの/ 六時間に/ 一九センチ、/ 胆振管内早来町安平で/ 一六センチなどの/ 降雪を/ 記録した。
- 65%: 午後から/ 雪が/ 札幌市南区小金湯で/ 同九時までの/ 六時間に/ 一九センチ、/ 胆振管内早来町安平で/ 降雪を/ 記録した。
- 50%: 北海道内は/ 午後から/ 局地的に/ 強まり、/ 札幌市南区小金湯で/ 降雪を/ 記録した。
- 35%: 六日/ 雪が/ 同九時までの/ 一九センチ、/ 一六センチなどの
- 20%: 六日/ 札幌市南区小金湯で/ 胆振管内早来町安平で

-原文 2-

- ロシア政府は/ 一日、/ 首都制圧と/ 発表したが、/ その後も/ 市内では/ 激戦が/ なお/ 続き、/ ロシア側に/ 数百人の/ 死傷者との/ 情報も/ ある。

〔自動簡約結果〕

- 80%: 一日、/ 首都制圧と/ 発表したが、/ その後も/ 市内では/ 激戦が/ なお/ 続き、/ ロシア側に/ 数百人の/ 死傷者との/ 情報も
- 65%: ロシア政府は/ 一日、/ 首都制圧と/ 発表したが、/ ロシア側に/ 数百人の/ 死傷者との/ 情報も/ ある。
- 50%: 一日、/ 首都制圧と/ 発表したが、/ ロシア側に/ 数百人の/ 死傷者との/ 情報も
- 35%: ロシア側に/ 数百人の/ 死傷者との/ 情報も/ ある。
- 20%: 数百人の/ 死傷者との/ 情報も

〔人による簡約結果〕

- 80%: ロシア政府は/ 一日、/ 首都制圧と/ 発表したが、/ 市内では/ 激戦が/ 続き、/ ロシア側に/ 数百人の/ 死傷者との/ 情報も/ ある。
- 65%: ロシア政府は/ 一日、/ 首都制圧と/ 発表したが、/ ロシア側に/ 数百人の/ 死傷者との/ 情報も/ ある。
- 50%: ロシア政府は/ 首都制圧と/ 発表したが、/ ロシア側に/ 数百人の/ 死傷者との/ 情報も
- 35%: 首都制圧と/ 発表したが、/ ロシア側に/ 数百人の/ 死傷者との
- 20%: ロシア側に/ 数百人の/ 死傷者との/

〔ランダム簡約結果〕

- 80%: ロシア政府は/ 一日、/ 首都制圧と/ その後も/ 市内では/ 激戦が/ なお/ 続き、/ 数百人の/ 死傷者との/ 情報も/ ある。
- 65%: 一日、/ 首都制圧と/ 発表したが、/ その後も/ 激戦が/ ロシア側に/ 数百人の/ 情報も/ ある。
- 50%: 一日、/ 首都制圧と/ 発表したが、/ 続き、/ ロシア側に/ 死傷者との/ ある。
- 35%: 発表したが、/ 激戦が/ 数百人の/ 死傷者との/ 情報も
- 20%: 市内では/ ロシア側に/ 情報も

8 おわりに

文節間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．

被験者の感想から，総合的評価では情報の保持と日本語としての構文的自然さを比較したとき，前者を重要視していることが分かった．このことは，総合的評価結果が情報の保持に焦点を絞った評価結果と類似した傾向を示すことにも表れている．したがって，本提案手法の簡約性能をさらに人のそれに近づけるためには，重要度の差をより顕著に表すような，文節の分類について検討する必要がある．そのため，今後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] 形態素解析システム JUMAN
<http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html>
- [8] 構文解析システム KNP
<http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html>
- [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)．