読みやすさの向上と冗長性の排除を考慮した重要個所抽出型要約

望月 源[†], 奥村 学[‡†]
[†]北陸先端科学技術大学院大学 情報科学研究科 motizuki@jaist.ac.jp
[‡]東京工業大学 精密工学研究所 oku@pi.titech.ac.jp

[概要]
本研究では，generic で informative な要約の作成を目指し，「指定された要約率の範囲で，元のテキストの情報をできる限り含めること」と「作成された要約が文章として自然で読みやすいこと」の実現のための重要個所抽出による要約作成を行なう．本稿では，重要個所抽出の際に，構文情報と語彙的結束性の情報を考慮して，同じ内容を表わす語の繰り返しによる冗長性を抑える手法，文としての意味を維持するために必要な他の部分を補完する手法，内容的に一貫性のある読みやすい要約を作成する手法について述べる．また，指定された要約率の範囲で内容がどの程度保持できているかを人間による重要個所抽出型要約との比較で評価する．

キーワード　テキスト自動要約，　語彙的結束性，　重要個所抽出

Summarization by Extracting Important Parts Considering Readability and Redundancy

[†]MOCHIZUKI Hajime, [‡†]OKUMURA Manabu
[†] School of Information Science, Japan Advanced Institute of Science and Technology
[‡]Precision and Intelligence Laboratory, Tokyo Institute of Technology

Abstract
In this research, we aim to develop a summarization system which is capable of making an informative and generic summary. We require the system to have two abilities; to include as much information as possible on the original text in the summary of the specified rate, and to produce natural sentences with higher readability which can be exchanged for the original text. We adopted the method of extracting important parts of sentences for our first summarization system. Since important parts are smaller than important sentences, it can be considered that a fine-grained summary can be produced by using these parts. In this paper, we describe an automatic summarization method of extracting important parts which result in an improvement of readability and an exclusion of redundancy. We also make a content-based summary comparison by measuring the similarities between the summaries extracted by some automatic summarization methods and those extracted by human subjects.

key words　automatic text summarization,　lexical cohesion,　important parts extraction

1　はじめに

本研究では，その要約を読んで元のテキストの内容をすばやく把握できるような要約を作成するために，元テキストの複数の話題で構成される generic で，元テキストの代わりとなりうる informative な要約の作成を目指す．具体的には「指定された要約率の範囲で，元テキストの情報をできる限り含めること」と「作成された要約が文章として自然で読みやすいこと」の実現を目標とする．自然な要約の作成は要約研究全体におけるこれからの大きな課題でもある [4]．

これまでのテキスト自動要約の研究では，主に重要文抽出の手法が用いられてきたが，本研究では，重要個所抽出による要約作成手法を採用する．重要個所抽出 [6, 5, 8, 2] (または不要個所除去 [11, 10]) では，文単位よりも細かく要約のサイズを変化させたり，同じ要約率でも，要約に含める部分を細かく入れかえたりすることが原理的に可能である．そのため，指定した要約率の範囲で，冗長個所を削除し，他の有用な部分を含めることによって，より多くの情報を含めたり，内容の重要度はそれほど高くないが，要約の意味的なつながりを維持するために最小限必要な部分を加えたりといった調整ができ，重要文抽出よりも自然な要約作成の実現が期待できる．

しかし，単に重要個所を抽出するだけでは，文の断片の連続となり，意味が通じなかったり，読みにくい要約が作成される可能性がある．また，重要度を各部分ごとに独立に計算すると，同じ内容が繰り返し出現し，冗長な要約となるおそれがある．つまり，重要個所抽出の利点を生かし，自然な要約を作成するための工夫が必要になる．

本研究では，重要個所抽出の際に，1 文内から一部分を抽出した場合に，文としての意味を維持するために必要な他の部分を構文解析結果を利用して補完する．これにより，意味がわからなかったり，読みにくい文が作成されることを避ける．また，文中の各部分の重要度計算には，語彙的結束性に基づく語彙的連鎖の情報を用いる．語彙的連鎖は，テキスト中に存在する話題の種類や範囲の認定に利用でき，文中の各部分に含まれる話題の重要度を，そこに含まれる連鎖の重要度から計算することができる．ただし，全ての語彙的連鎖が話題を強く表わすわけではなく，ある話題を表わす他の連鎖を補助するような連鎖も実際には存在する．そこで本研究では，各連鎖を構成する語の数と，構文解析の結果からわかる複数連鎖どうしの関係を考慮して各連鎖の重要度を計算する．

また，語彙的連鎖を利用して重要個所抽出を行なうと，重要な話題がある程度繰り返し抽出されることが予想される．これは，要約内の内容の結束性を保つためには有効であると考えられるが，作成される要約の冗長性を排除し，できるだけ多くの話題を含めるためには，過度の繰り返しを避ける工夫が必要になる．本研究では，要約を作成していく過程で，既に抽出された部分に含まれる話題に関係する語彙的連鎖の重要度を調整し，同じ話題が何度も抽出されることを避ける．

以上のことから，本稿では，我々は語彙的連鎖の情報に構文解析の情報をあわせて，重要個所抽出の際に，できるだけ情報を落とさずに，要約の読みやすさの向上と，冗長性を排除する手法について述べる．

2　要約作成手法

本節では，要約作成の基本方針とアルゴリズムについて述べる．

2.1　基本方針

要約作成の基本方針は次のようにする．

- 細かな要約率への対応

  指定された要約率にできるだけ近い要約の作成と，含める情報の細かな調整を可能にするため，抽出する最小単位を 1 文内で意味のある複数単語のまとまりとして作成する．これを，基本単位と呼び，基本単位ごとに重要度の計算を行なう．

- 話題の結束性を考慮した重要度計算

  各基本単位の重要度は，基本単位内の内容語が属す語彙的連鎖の重要度を元に計算したスコアに，その基本単位自身の文内における構文的な重要度による重みをかけて計算する．これにより話題として重要であり，1 文内でも構文的に重要である個所 (基本単位) が選ばれやすくする．

- 読みやすさを保つための補完処理

  抽出された要約が文の断片の連続となるのを避けるため，ある重要な基本単位を抽出した場合に，文としての意味を保持するために必要な他の基本単位を補う．ただし，補完された他の基本単位の重要度も考慮する必要があるため，まず，重要度の高い，いくつかの基本単位を抽出の候補とし，それぞれ補完を行なう．次に，補完後のそれぞれの文の重要度を計算し，最も重要度の高いものを要約に含める．

- 冗長性を排除するための，基本単位の重要度の再計算

  要約の冗長性を排除し，できるだけ多くの情報を要約に含めるために，1 つの重要個所 (補完後の 1 文) が選択された後に，既に要約として抽出され

た個所に含まれる話題に関する語彙的連鎖の重要度を下げる．次に，再計算された語彙的連鎖の重要度を用いて，まだ選択されていない文に属す，各基本単位の重要度を計算し直す．

2.2 アルゴリズム

2.1節の方針に基づく要約の作成アルゴリズムは次のようになる．なお，個々の実装に関する詳細は3節でまとめて説明する．

1．テキストの形態素解析，構文解析

形態素解析には，juman を用い，その出力を knp によって構文解析した結果を情報として使用する．

2．テキスト中の語彙的連鎖の作成と重要度の計算

我々の語彙的連鎖作成手法 [12] により，テキスト中の語彙的連鎖を作成し，各語彙的連鎖の重要度を計算する．重要度計算の詳細は3.4節で述べる．

3．各文内の基本単位の計算

本研究では，基本的に 1 つの内容語とそれに付属する機能語から基本単位を作成する．詳細については3.1節で述べる．

4．基本単位の重要度計算

基本単位の重要度は，内容語の語彙的連鎖に基づく重要度と，文内における構文的な重要度によって計算する．詳細は3.5節で説明する．

5．抽出候補となる基本単位の選択

重要度の上位 N 位までの基本単位を抽出候補として選択する．

6．他の基本単位の補完

N 個の各抽出候補について，その基本単位を抽出する場合に，文としての意味を成り立たせるために必要な，同一文中の他の基本単位を構文解析結果の係り受けの情報を利用して補完する．詳細は3.3節で述べる．

7．補完後の文の重要度の計算

N 個の各抽出候補について，補完された文全体での重要度を，その文に含まれる各基本単位のスコアを元に計算する．詳細は3.6節で述べる．

8．重要箇所の選択

N 個の中で，全体での重要度が最も高い候補を重要個所として抽出する．

9．要約率の判定

抽出個所が要約率に達しているかどうかを判定する．もし達していれば 11 に．そうでなれば 10 に進む．

10．語彙的連鎖の重要度の変更

抽出した個所に含まれる内容語の属す語彙的連鎖の重要度を減らし，4 に戻る．重要度を減少させる方法についての詳細は3.7節で述べる．

11．要約の出力

抽出した重要個所を元のテキストの出現順に並べて，出力する．

3 要約作成システムの構築

2.2節のアルゴリズムに基づく要約作成システムの実装に必要な詳細を示す．なお，現在の実装では，ここで説明する手法のいくつかは，パラメータ化して，切りかえて使用できるようになっている．

3.1 基本単位の計算

基本単位は，原則として 1 つの内容語とそれに付属する機能語として計算する．knp の出力の文節にあたる単位が基本であるが，次の場合は例外として扱う．

- 名詞+連語 (例：出勤の+ため)

形の上では内容語であるが，実際には内容のほとんどない語も存在する．例えば，助詞的表現「～において」の「おいて」や，「～のため」の「ため」などである．このような内容語は機能語相当語として扱い，直前の基本単位に連結し，1 つの基本単位とする．

機能語相当語の判別には，IPA コーパス [7] の分類と，形態素解析システム茶筌用の IPA 品詞体系版辞書 version2.0 の Postp.dic 内の「連語」を参考に決定する．IPA コーパスでは「助詞」の分類のなかで，「動詞」と「格助詞」の組み合わせで格助詞に相当するような働きを持つものを「連語」として分類している．また，茶筌の Postp.dic 内にも同様な意味での「連語」が用意されている[1]．本研究では，これらの連語の中から 75 種類の組み合わせを機能語相当語として選択している．

- 複合名詞

複合名詞はすべて 1 つの基本単位とし，付属語がつく場合には，複合名詞と付属語で 1 つの基本単位とする．

- 副詞的名詞が続く場合 (例：帰る+場合)

副詞的名詞も内容語のような意味を持たず付属語のように扱われるため，直前の基本単位に含める．

- 「」や “” に囲まれた語 (例：「安全な国」)

「」や “” で囲まれた部分は，引用であったり専門用語であったりするため，分割せずに 1 つの独立した部分として扱う．

3.2 基本単位のタイプ分類

1 つの基本単位が，他の基本単位との間で係り受け関係になる場合の性質の違いによって，基本単位を以下の 4 種類に分類する．この分類は，3.3節の補完や3.5節の基本単位の重要度計算の際に処理の切り替え

[1] なお，juman 3.6 の Rengo.dic はここでいう連語とは内容が異なるため，考慮しない．

の基準として使用する．なお，ここで「N 系」は「体言」か「体言に助詞がついたもの」であり活用しないものを意味し，「V 系」は「動詞」や「形容詞」や「助動詞」など活用するものを意味する．

- N 系+V 系，例：「手薄+という」
  N 系で受け，V 系で係るタイプの基本単位
- V 系+N 系，例：「帰る+場合に」
  V 系で受け，N 系で係るタイプの基本単位
- N 系+N 系，例：「日本+が」，「日本社会+は」
  N 系で受け，N 系で係るタイプの基本単位
- V 系+V 系，例：「撃たれ」，「撃った+という」
  V 系で受け，V 系で係るタイプの基本単位

3.3 他の基本単位の補完

要約としてある 1 つの基本単位を選択した場合に，文全体として意味が通じるようにするために他の基本単位を補完する．

補完する基本単位は，基本単位間の係り受け関係と，格関係に基づいて決定する．係り側の補完，受け側の補完それぞれについて，3.2節の 4 つの分類の組み合わせによって以下のように処理を分ける．

3.3.1 係り側の補完

最初の基本単位を出発点とし，係り側になる各基本単位を以下により補完する．1 つの基本単位が補完された場合はその基本単位を起点にし，補完が行なわれなくなるまで再帰的に補完を続ける．

1. 受け側の基本単位が「N 系+V 系」，「N 系+N 系」タイプで，係り側の補完候補が「N 系+N 系」もしくは「V 系+N 系」の時，以下の場合は補完する．
   - 係り側が「の」で終わる場合
   - 基本単位どうしが並列関係にある場合
2. 受け側の基本単位が「N 系+V 系」，「N 系+N 系」タイプで，係り側の補完候補が「N 系+V 系」もしくは「V 系+V 系」の時，以下の場合は補完する．
   - 受け側が形式名詞で始まる場合
   - 受け側が副詞的名詞で始まる場合
3. 受け側の基本単位が「V 系+N 系」，「V 系+V 系」タイプで，係り側の補完候補が「N 系+N 系」もしくは「V 系+N 系」の時，以下の場合は補完する．
   - 係り側が必須格（「が」，「を」，「に」）か，提題助詞（「は」，「も」）または「で」で終わる場合

3.3.2 受け側の補完

抽出候補として選択した最初の基本単位を出発点とし，まず直接受け側になる基本単位の補完を行なう．次に補完された基本単位を起点にし，受け側になるその先の基本単位を以下の条件により補完する．この処理は補完が行なわれなくなるまで再帰的に続ける．

1. 係り側の基本単位が「V 系+N 系」，「N 系+N 系」タイプで，受け側の補完候補が「N 系+N 系」もしくは「N 系+V 系」の時，以下の場合は補完する．
   - 係り側が「の」で終わる場合
2. 係り側の基本単位が「V 系+N 系」，「N 系+N 系」以外の場合，受け側が存在すればすべて補完する．

3.3.3 受け側で補完された基本単位の係り側の補完

3.3.2節で，新たに受け側として補完された基本単位については，その基本単位を起点として，3.3.1節の規則により，係り側の補完を行なう．

3.4 語彙的連鎖の重要度計算

語彙的連鎖の重要度計算方法について述べる．語彙的連鎖 i の重要度 lc_scr_i は，連鎖 i の基本スコア $lc_basescr_i$ に，連鎖 i が他の連鎖との構文的関係からどのくらい重要であるかを示す重み lc_w_i をかけて計算する．

- 語彙的連鎖の基本スコア
  1 つの語彙的連鎖（連鎖 i）の基本スコア $lc_basescr_i$ は次のように計算する．

  $lc_basescr_i =$ 連鎖 i を構成する語の数

  ただし，同一文に出現する連鎖構成語は 1 回しかカウントしない．

- 各連鎖の重要度に応じた重み付け
  構文解析の結果を利用して，連鎖 i と他の語彙的連鎖との関係を参照し，以下の 4 種類の重みを計算する．語彙的連鎖 i を構成する語が，
  - ヘッドになる割合に応じた重み (lc_w1_i)，
  - トピックになる割合に応じた重み (lc_w2_i)，
  - 名詞である割合に応じた重み (lc_w3_i)，
  - 修飾語あるいは被修飾語となる割合に応じた重み (lc_w4_i)．

  をそれぞれ計算する．それぞれの割合に応じた重みは，割合が 0.1 未満の時，0.4，0.3 未満の時，0.6，0.5 未満の時，0.8，1.0 未満の時，1.0，1.0 の

時，1.1 とする．重み lc_w_i は，lc_w1_i から lc_w4_i のどれかを単独あるいは組み合わせで用いて計算する．組み合わせは 15 通りあるが，今回の実装では以下の 6 種類の組み合わせのどれかを選んで用いる (パラメータ A1)．

1. $lc_w_i = lc_w1_i$
2. $lc_w_i = lc_w2_i$
3. $lc_w_i = lc_w1_i * lc_w2_i$
4. $lc_w_i = lc_w3_i$
5. $lc_w_i = lc_w4_i$
6. $lc_w_i = lc_w1_i * lc_w2_i * lc_w3_i * lc_w4_i$

- 連鎖のスコアと重みの統合
  語彙的連鎖 i の重要度 lc_scr_i は，基本スコア $lc_basescr_i$ に，パラメータ A1 で選択した連鎖 i の重み lc_w_i をかけて計算する．

  $lc_scr_i = lc_basescr_i * lc_w_i$

3.5 基本単位の重要度計算

基本単位の重要度計算方法について説明する．基本単位 i の重要度 bu_scr_i は，基本単位内の内容語が属す語彙的連鎖の重要度を元に基本スコア $bu_basescr_i$ を計算し，基本単位の文内における構文的な重要度に応じた重み bu_w_i を，その基本スコアにかけて計算する．

- 基本単位の基本スコア
  基本スコアの計算手法を 2 種類用意し，どちらかを選んで用いる (パラメータ B1)．以下で，内容語 j のスコアは，語彙的連鎖 j のスコアとして計算する．

  1. 基本単位内の内容語のスコアの平均とする．
     $bu_basescr_i =$
     基本単位内の内容語のスコアの平均
     $bu_basescr_i =$
     $\sum_{j=1}^{n}$ 内容語 j のスコア$/n$
     ここで，n は，内容語の数である．
  2. 基本単位内で最も重要であると考えられる内容語を主の内容語として，他の内容語と分けてそれぞれの平均を足す．

     $bu_basescr_i =$ 主の内容語の平均スコア $+$
     他の内容語の平均スコア
     ここで，
     主の内容語の平均スコア $=$
     $\sum_{j=1}^{m}$ 主の内容語 j のスコア$/m$
     他の内容語の平均スコア $=$
     $\sum_{k=1}^{o}$ 他の内容語 k のスコア$/o$

     主の内容語は，複合名詞のヘッドや，機能語の直前の内容語とする．通常は主の内容語は基本単位内に 1 つであるが，「 」内を 1 つとする基本単位の場合は，複数存在する場合もあるため，平均する．

- 内容語の語彙的連鎖内の出現位置による重み付け
  1 つの連鎖の開始となる内容語に特別な重みをつけるか，つけないかを選択して用いる (パラメータ B2)．

  1. 重みを付けない．
  2. その内容語の出現が，対応する語彙的連鎖の最初の出現である場合に重み付けする．

- 各基本単位の構文的重要度に応じた重み
  基本単位 i の重み bu_w_i は，基本単位のタイプ分類に応じて，基本単位が係り側になる場合の性質の違いによって，N 系で係る場合 (「N 系+N 系」，「V 系+N 系」) と，V 系で係る場合 (「N 系+V 系」，「V 系+V 系」)，それぞれについて以下のように重みを決定する．

  – 「N 系+N 系」または「V 系+N 系」の場合，基本単位の機能語の種類に応じて表 1のように重み付けする．

表 1: 機能語の種類と重み

ハ	1.5	ニ	1.2	ノ	1.0	ノデ	1.0
モ	1.4	デ	1.1	ヤ	1.0	ニテ	1.0
ガ	1.3	ヘ	1.0	マデ	1.0	連語	1.0
ヲ	1.2	ト	1.0	ヨリ	1.0	その他	1.0

  –「N 系+V 系」または「V 系+V 系」の場合，基本単位の係り側の構文的特徴に応じて表 2のように重み付けする．

表 2: 構文的特徴と重み

連用，連格，連体	1.0	文末	1.3	その他	1.0

- 基本単位のスコアと重みの統合
  基本単位のスコアと重みのかけ算とする．
  $bu_scr_i = bu_basescr_i * bu_w_i$

3.6 補完後の文全体の重要度計算

補完後の文全体の重要度 su_scr_i は，その文内で抽出された各基本単位の重要度の平均によって計算する．
$su_scr_i = \sum_{j=1}^{n}$ 基本単位 j のスコア$/n$
ここで，n は基本単位の数を示す．

3.7 語彙的連鎖の重要度の変更

冗長性を排除するため，要約として抽出された基本単位内の内容語が属す語彙的連鎖の基本スコアを減点する．これにより，その語彙的連鎖の重要度が下がり，基本単位の重要度を再計算することで，既に選択された語が繰り返し選ばれ難くなる．

重要度を変更する語彙的連鎖の選択基準と，減点方法を次のようにする．

- 減点対象となる語彙的連鎖の決定
  次の 2 種類を手法として用意し，どちらかを選択して用いる (パラメータ C1)．
  1. 補完されなかった基本単位も含め，重要個所が属す文内の全基本単位を対象とする
  2. 重要個所とその補完個所だけを対象とする
- 減点方法
  連鎖の基本スコアを指定した割合 r だけ減らし[2]，以下のように計算する．
  変更後の $lc_basescr_i = lc_basescr_i * r$

4 予備実験

本研究では，「指定された要約率の範囲で，元テキストの情報をできる限り含めること」と「作成された要約が文章として自然で読みやすいこと」の実現を目標としている．そのため，評価も両方について行なう必要があるが，本稿では，最初の評価として，抽出された要約がどの程度重要な情報を含んでいるかを，人間の被験者が作成した重要個所抽出型要約との比較によって評価し，読みやすさの評価は今後の課題とする．

本研究では，前者を評価する方法として，Donaway らの提案している content based な評価方法 [1] を採用する．この評価方法は，generic で indicative な要約の評価のために提案されたものであり，人間の作成した正解の要約とシステムの要約の直接的な一致度によって評価するのではなく，正解の要約に含まれる情報をより多く含む要約を良い要約と考える．そのため，システムの要約と人間の作成した正解の要約をそれぞれ単語頻度ベクトルとし，ベクトル間のコサイン距離を計り，正解の要約との類似度が高い要約ほど良い要約であるとして評価する．本研究では，各要約を形態素解析し，名詞，動詞，形容詞を対象に，tf*idf の重み付きベクトルで表現する．

Donaway らは，generic で indicative な要約の評価を意図してこの評価方法を用いているが，content based な評価の考え方は，informative な要約における要約の generic 性を評価するために使用できる．そのため，本研究では，我々の目的である重要な情報をどのくらい良く要約に含むことができたかという評価に使用する．

実験用のデータには，筑波大学山本幹雄助教授の作成した人手による重要個所選択による要約データの一部を使用する．このデータは，毎日新聞 CD-ROM'94 年版から抽出した 56 記事を 14 記事ごとの 4 セットに分け，被験者 5 人がそれぞれ 2 セットから 3 セットについて要約を作成したものである．今回の実験では，56 記事の中から死亡記事やイベント開催告知などあまり要約に適さない記事を除いた 48 記事に対する被験者 a,c,d,e の 4 人 (1 人あたり 35 記事から 36 記事)[3]の要約とシステムの要約との比較を行なう．

今回の実験では，本手法と，重要文抽出の代表的な手法，および全文によってそれぞれ作成した要約と被験者が作成した正解の要約の類似度を計算する．比較を行なった要約作成手法は以下のとおりである．

- 本手法 (共起，角川，同語)
  本手法による要約．ただし，共起，角川類語新辞典 [9]，同語反復の 3 種類の基準が異なる語彙的連鎖を用いる．連鎖は名詞と形容詞によって作成する．また，パラメータを変化させて実験するために，被験者毎のデータを 6 セットに分け，5 セットを訓練，1 セットを評価とする 6 回のクロスバリデーションによって評価する．
- 重要文抽出
  被験者が選択した個所が含まれる文の数 N だけ重要文を抽出する．
  - **BEST**: 被験者が選択した個所が属す文を抽出．
  - **LEAD**: 先頭から N 文を抽出．
  - **TF**，共起, 角川, 同語: 文内の語に重み付けし，点数の高い文を N 文抽出．重み付けに TF を用いる場合と，語彙的連鎖のスコアを用いる場合で 4 種類作成する．
- **FULL**: 全文をそのまま使用．

実験結果を表 3に示す．

4.1 結果

被験者 4 人中，3 人については，類似度で重要文抽出での結果を上回ることはできなかった．しかし，被

[2]現在のところ，割合 $r = 0.2$ に固定している

[3]b は実験テキスト数が他の 4 名より 1 セット分少ないため除外した．

表 3: content based な評価の結果

被験者 a との比較，要求要約率=0.335										
	本手法 (() 内は訓練データでの値)			他の重要文抽出手法						
	共起	角川	同語	BEST	FULL	LEAD	TF	共起	角川	同語
類似度	0.591(0.620)	0.588(0.621)	0.581(0.620)	0.934	0.731	0.656	0.686	0.669	0.706	0.712
要約率	0.371	0.362	0.366	0.405	1.000	0.351	0.462	0.437	0.452	0.450
比率	1.127	1.089	1.114	1.226	3.121	1.070	1.400	1.328	1.373	1.364
被験者 c との比較，要求要約率=0.157										
	共起	角川	同語	BEST	FULL	LEAD	TF	共起	角川	同語
類似度	0.419(0.463)	0.475(0.493)	0.487(0.493)	0.709	0.542	0.484	0.464	0.452	0.448	0.464
要約率	0.192	0.196	0.203	0.390	1.000	0.371	0.492	0.462	0.479	0.478
比率	1.236	1.240	1.274	2.714	7.076	2.589	3.472	3.253	3.386	3.377
被験者 d との比較，要求要約率=0.266										
	共起	角川	同語	BEST	FULL	LEAD	TF	共起	角川	同語
類似度	0.524(0.548)	0.559(0.590)	0.559(0.577)	0.893	0.669	0.669	0.629	0.637	0.638	0.644
要約率	0.290	0.292	0.279	0.369	1.000	0.344	0.458	0.429	0.444	0.441
比率	1.116	1.130	1.091	1.423	4.292	1.367	1.820	1.711	1.768	1.756
被験者 e との比較，要求要約率=0.216										
	共起	角川	同語	BEST	FULL	LEAD	TF	共起	角川	同語
類似度	0.496(0.553)	0.579(0590)	0.586(0.586)	0.814	0.644	0.633	0.603	0.601	0.633	0.636
要約率	0.251	0.246	0.249	0.394	1.000	0.350	0.462	0.437	0.451	0.450
比率	1.191	1.146	1.167	1.950	5.082	1.758	2.325	2.194	2.271	2.261

験者 c の場合は，重要文抽出の場合と，ほぼ同程度の類似度が得られている．本手法の 3 種類の連鎖の中では共起があまり良くなかった．出力される要約の長さを考えると，重要文抽出手法による要約は，本手法よりもかなり長い[4]．今回の評価では長い要約が有利であると考えられるため，要約の長さに大きな差がある場合には直接結果を比較できない．この点を考慮すると，本手法による要約は，重要文抽出に比べて要約の長さが短い割には，それほど情報は減っていないと考えることができ，ある程度の成果が得られたといえる．特に被験者 c のように，要約率が非常に小さい (被験者の平均要約率が 0.157) 場合は，効果が大きかった．

4.2 パラメータについて

今回の実験では 4 種類のパラメータを変化させて訓練を行なった．訓練データ上で，良い結果を示したパラメータに次の傾向が見られた．

語彙的連鎖の重み (パラメータ A1) は，ヘッドになる割合と，トピックになる割合のどちらか一方，もしくは両者の組み合わせが選ばれる割合が 7 割以上だった．

基本スコアの計算方法 (パラメータ B1) については，主の内容語とそうでないものの平均を別々に計算する方が 7 割以上選ばれた．

1 つの語彙的連鎖の最初の出現位置にある内容語への重み (パラメータ B2) は，ほぼ全ての場合，重み付けする方が選ばれた．

減点対象となる語彙的連鎖の選択方法 (パラメータ C2) は，非抽出個所も含めた文全体とするか，抽出した個所だけにするかについて，はっきりとした差が出なかったが，後者が選ばれる方が多かった．

5 おわりに

本稿では，我々は語彙的連鎖の情報に構文解析の情報をあわせて，できるだけ情報を落とさずに，要約の読みやすさの向上と，冗長性を排除する手法について述べた．今後の課題として次のことがあげられる．

- 他の基本単位を補完する規則の充実
  3.3節で述べた基本単位の補完手法は，基本的な規則であり，より一層の充実を計る必要がある．

- 読みやすさに関する評価の実施
  今回は，読みやすさの面について評価を行なっていない．今後は人間による読みやすさの主観的評価などの手法 [3, 2] により，本手法で作成された要約の読みやすさを評価する必要がある．

- 書き換えによる読みやすさの向上
  現在のところ抽出した要約をそのまま出力しているが，より読みやすい要約を作成するためには，部分的な書きかえも考慮する必要がある [3]．機械翻訳などの文生成技術を考慮し，よりなめらかな要約作成のための書き換えを行なう．

[4] **LEAD** の被験者 a に対する結果だけは異なる．

なお，本手法による要約の出力例を最後に付録として付す[5]．

謝辞

実験データを提供していただいた筑波大学の山本幹雄助教授に感謝致します．北陸先端科学技術大学院大学自然言語処理学講座博士後期課程の難波英嗣君には本研究に対し貴重な助言を頂きました，感謝致します．また，角川類語新辞典の使用を許可下さいました株式会社 角川書店に感謝致します．なお，本研究では，コーパス上の語の共起計算に，日立製作所中央研究所において，IPA 独創的情報技術育成事業に係る開発の成果として開発された連想計算エンジン (GETA) を使用しています．

参考文献

[1] R. Donaway, K. Drummey, and L. Mather. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proc. of NAACL-ANLP 2000 Workshop Automatic Summarization*, pp. 69–78, 2000.

[2] H. Jing and K. McKeown. Cut and paste based text summarization. In *Proc. of the 1st Meeting of the North American Chapter of the ACL*, pp. 178–185, 2000.

[3] H. Nanba and M. Okumura. Producing More Readable Extracts by Revising Them. In *Proc. of the 18th International Conference on Computational Linguistics*, pp. 1071–1075, 2000.

[4] 奥村 学, 難波 英嗣. テキスト自動要約に関する最近の話題. Technical Report IS-TM-2000-001, 北陸先端科学技術大学院大学, 2000.

[5] 岡 満美子, 小山 剛弘, 上田 良寛. 句表現要約の句合成手法. 情報処理学会研究会資料 NL129-15, pp. 101–108, 1999.

[6] 亀田 雅之. 日本語文書読解支援系 QJR の検討. 情報処理学会研究会資料 NL110-9, pp. 57–64, 1995.

[7] 橋本 三奈子，荻野 紫穂，徳永 健伸，元吉 文男，井佐原 均. IPA コーパスの概要. IPAL シンポジウム '95 論文集, pp. 31–44, 1995.

[8] 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文要約アルゴリズム. 言語処理学会第 6 回年次大会発表論文集, pp. 133–136, 2000.

[9] 大野晋, 浜西正人. 角川類語新辞典. 角川書店, 1981.

[10] 石ざこ友子, 片岡 明, 増山 繁, 中川 聖一. テレビニュース番組の字幕作成のための重複部削除による要約. 情報処理学会研究会資料 NL133-7, pp. 45–52, 1999.

[11] 福島 孝博, 江原 暉将, 白井 克彦. 文単純化のための文字数圧縮規則. 言語処理学会第 5 回年次大会発表論文集, pp. 221–224, 1999.

[12] 望月 源, 奥村 学, 岩山 真. 語彙的結束性に基づく語彙的連鎖の計算. Technical Report IS-TM-2000-002, 北陸先端科学技術大学院大学, 2000.

[5]角川類語新辞典による語彙的連鎖を使用し，パラメータは，A1=3,B1=1,B2=2,C1=1，要約率は 23.3%である．なお，誌面の都合上改行はしていない．

付録

[原文](毎日新聞'94 年版 CD-ROM, 記事 ID0010800)
南アフリカで制憲議会選挙に向けての選挙戦が本格化してきた。十二日深夜には選挙に参加する政党登録が締め切られた。白人政権与党「国民党」、最大の黒人解放組織「アフリカ民族会議」(ＡＮＣ)など十九の政党が登録を済ませた。四月二十六日から三日間行われるこの選挙は、人種隔離(アパルトヘイト)政策全廃の締めくくりとなるもので、黒人有権者が南ア史上初めて一票を投ずる。それは同時に新生南アのスタートでもある。このように歴史的な意味も大きい選挙だけに、全人種、全政党の参加が望ましいのだが、ズールー族主体の黒人右派組織「インカタ自由党」(ＩＦＰ)、アパルトヘイト政策廃止を不満とする白人右派組織「アフリカーナー人民戦線」(ＡＶＦ)など四政党が登録を見送り、選挙ボイコットの構えを崩していない。大変残念なことだ。国民党及びＡＮＣはこれら四党で作る連合組織「自由同盟」と、政党登録締め切り期限ぎりぎりまで選挙参加を求めて交渉を続けたが、合意に達しなかった。最大の対立点は、自由同盟が民族自決と強固な地方分権を要求、他方国民党、ＡＮＣなどは、中央政府が強い権限を持つ国家を目指しているところにある。私たちは自由同盟の要求の方に無理がある、と言わざるを得ない。なぜなら、新生南アが目指すのは人種の融合・和解であるはずだからだ。自由同盟の要求はこれに逆行するものであり、アフリカーナー(大陸欧州系白人)の準独立国家をというＡＶＦの要求などは、アパルトヘイトの発想そのものである。このような要求は、国際社会からもその正当性を認められないだろう。国民党、ＡＮＣ側は今後も交渉を続け、なお選挙参加の道を開いておくという。自由同盟側にも今少しの柔軟性を求めたい。デクラーク大統領、マンデラＡＮＣ議長はすでに選挙遊説を始めているが、両者ともに頭の痛い問題を抱えている。その第一は、選挙ボイコット勢力の出方である。ＩＦＰのブテレジ議長は十三日の党集会で「流血と死」を予告している。ＩＦＰとＡＮＣは過去四年間、死者一万四千人を出す政治抗争を続けており、投票日が近づくにつれて緊張がますます高まることが憂慮される。また今年に入ってすでに四十件以上のテロ事件が発生している。そのほとんどがＡＶＦ傘下の白人による犯行とみられている。歴史的な選挙から暴力を排除するために、デクラーク大統領、マンデラ議長には粘り強い交渉の継続と効果的な取り締まりという、難しいかじ取りが求められる。マンデラ議長の場合は黒人有権者を投票所に行かせること、そして間違いのない投票をさせることが、なかなかの問題だと現地からの報道は伝えている。なにしろ投票など初めての人たちだし、加えて黒人有権者の識字率は低いからだという。さらに選挙で生まれるはずの黒人主導政権への過剰期待を、あらかじめ戒めておくのもマンデラ議長の仕事だ。一票を投ずることでパラダイスが出現するわけではないからだ。いずれも容易ではない。しかし私たちはこれまで何度か、南アの人たちの自助能力を高く評価してきた。遠かった「夜明け」を目前にして、デクラーク大統領、マンデラ議長を指導者とする国民党、ＡＮＣが、再びその高い自助能力を発揮することを切に期待する。

[要約]
南アフリカで制憲議会選挙に向けての選挙戦が本格化してきた。十二日深夜には政党登録が締め切られた。黒人有権者が一票を投ずる。国民党及びＡＮＣは連合組織「自由同盟」と、政党登録締め切り期限ぎりぎりまで選挙参加を求めて交渉を続けたが、合意に達しなかった。最大の対立点は、自由同盟が民族自決と地方分権を要求、他方国民党、ＡＮＣなどは、権限を持つ国家を目指しているところにある。国民党、ＡＮＣ側は選挙参加の道を開いておくという。デクラーク大統領、マンデラＡＮＣ議長は選挙遊説を始めているが、問題を抱えている。ＩＦＰのブテレジ議長は十三日の党集会で「流血と死」を予告している。ＩＦＰとＡＮＣは投票日が近づくにつれて緊張が高まることが憂慮される。