文字数削減によるニュース文の要約システム

5N－1

加藤　直人*　浦谷則好**
*NHK放送技術研究所　**ATR音声翻訳通信研究所
katonao@strl.nhk.or.jp　uratani@itl.atr.co.jp

1　はじめに

大量の文章を手にいれることが容易となってきている現在，自動要約が注目されている．実際，市販のワープロや機械翻訳システムにはその機能の一部として自動要約が組み込まれているものもある．しかし，これらの自動要約の多くは文章中に含まれる単語の出現頻度，文章中での文の位置，手がかり語などに基づいて文の重要度を計算し，重要文を抽出することによる要約である．したがって，要約は文単位で行われるので大胆な（すなわち，要約率の設定幅が大きい）要約に向いている．

これに対して本稿で述べる自動要約は文章中の単語列を削減したり，より短い単語列で置換することによる要約である．このような文字数削減による要約では細かい要約率の設定が必要な要約（例えば，日本語ニュース音声の字幕化）に向いている．しかし，どのように文字数を削減するかという知識（要約知識）が必要となり，従来は人手で作成していた[若尾 97]．

本稿では日本語のテレビニュースを対象にした，文字数削減による要約システムについて述べる．本システムの要約知識は自動的に作成されている．

2　要約システムの概要

要約システムの概要を図1に示す．本システムに文を入力すると，その要約された文が出力される．入力は文のみならず文章でもよい．要約処理はまずはじめに入力の形態素解析を行う．次に（後述する）要約知識の適用を行い，要約候補がコスト付きで出力される．図1の例では要約候補として「総理大臣→首相」，「委員会→委」などが出力されている．最適要約計算では希望の要約率（＝要約された文章の文字数／元の文章の文字数）の中で，要約候補のコストの和が最小となる組み合わせを求める．要約率を変更すると最適要約計算のみを再度実行し，形態素解析や要約候補の処理を行うことはしない．

本システムの特徴は要約知識と最適要約計算にある．これらについて若干詳しく説明する．

2．1　要約知識

要約知識は要約の精度を左右する重要なものである．要約知識は大量にあることが望ましいが，人手で作成するのは困難である．

本システムでは要約知識をコーパスより自動的獲得している[加藤 98a]．我々の要約知識は置換規則と置換条件の2つから構成されているが，現在置換規則は約1,200，置換条件は15,000ある．置換規則と置換条件の例を図2に示す．

置換規則は原文の単語列をどのような単語列に置換するかを規定する知識である．例えば，次の例は連体助詞の「の」を省略するという置換規則である．

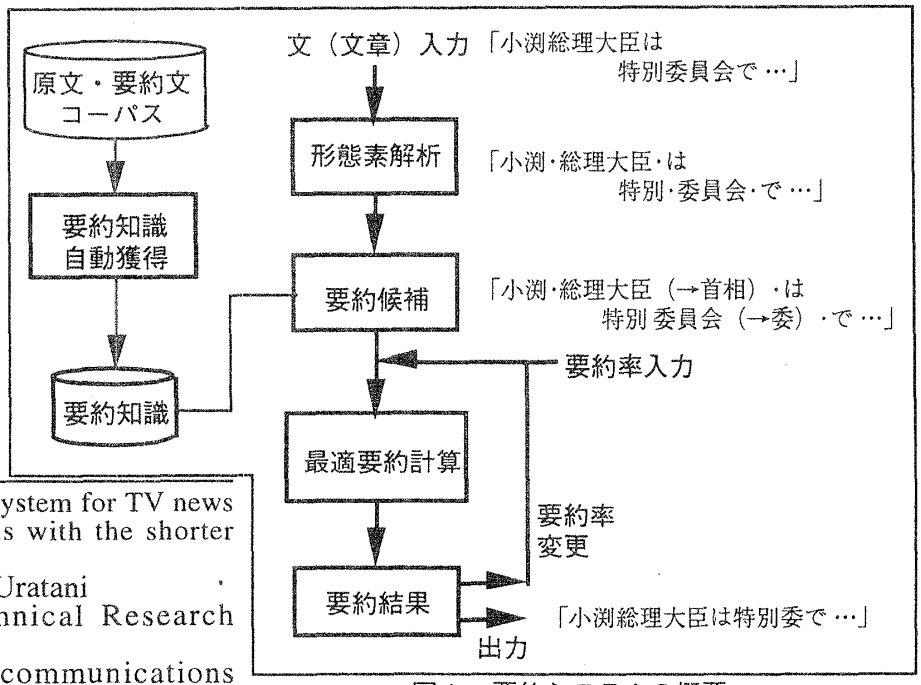

図1　要約システムの概要

An automatic summarization system for TV news sentences by replacing words with the shorter words.
Naoto Katoh and Noriyoshi Uratani
*NHK Science and Technical Research Laboratories
**ATR Interpreting Telecommunications Research Laboratories

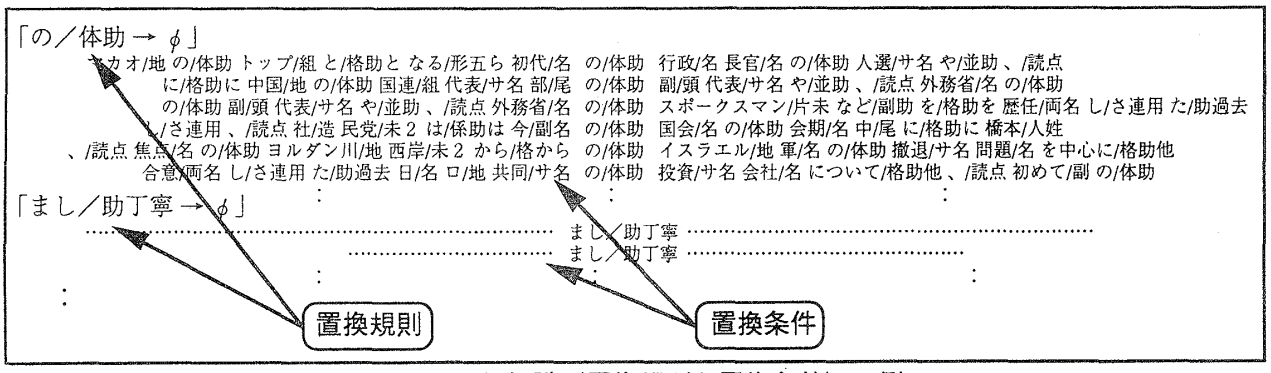

図２　要約知識（置換規則と置換条件）の例

【置換規則の例】

「の／体助 → φ」（φは空を表す記号）

一方，置換条件とは置換規則の適用の良否を数値化したもの（置換規則のコストと呼ぶ）である．置換規則はその前後の単語列によっては適用の良否が決まる．例えば「日本の銀行」に先の置換規則を適用して，「の／体助」を省略することはできない．そこで置換規則のコストは，置換規則の前後の単語列と，あらかじめ獲得しておいた置換条件との距離から計算している．この際にシソーラスを使うことにより，獲得しておいた置換条件とは完全に一致しない場合でも意味的に近い置換条件を適用することを可能としている．

２．２　最適要約計算

１文中にはいくつもの要約候補が出現するが，希望の要約率の中で置換規則のコストが最も小さい組み合わせを使った方がよい要約文であると期待できる．

我々は，希望の要約率の中で，置換規則のコストを最小化する要約候補の組み合わせを求めるアルゴリズムを開発した[加藤 98b].

我々のアルゴリズムは２パスから構成されている．第２パスがメインであり，第１パスは第２パスでの計算量を削減するための準備をする．第１パスでは文末から文頭に向けて計算し，文末からその単語までに最大削減可能な文字数を計算する．第２パスでは文頭から文末に向けて計算し，文頭からその単語までの最適な要約候補の組み合わせを求める．

２．３　要約システム

要約システムは現在 Machintosh 上でインプリメントされている．要約j実行例を図３に示す．図中上段が元の文章であり，下段が要約された文章である．要約個所を明示するために，元の文章も要約文章も当該個所は色を変えて表示するようになっている．処理速度は文章の長さにもよるが，要約率を変えるのに従いほぼリアルタイムでで要約できる．

３　おわりに

日本語のニュース文を対象にした，文字数削減による要約システムについて述べた．現在，さまざまな文章を使って要約実験を行っている．経験的には要約率８５％ぐらいが違和感のない要約文章を作ることができる要約率の限界である．今後は実際にどのくらいまでの要約率が可能かどうかの定量的な実験が必要となろう．また，従来の自動要約（重要文抽出による要約）とも組み合わせてより高度な自動要約システムを研究していく予定である．

参考文献

[加藤 98a]加藤直人：ニュース文要約のための局所的要約知識獲得とその評価，電子情報通信学会，言語理解とコミュニケーション研究会，NLC98-16 (1998).

[加藤 98b]加藤直人：ニュース文を対象にした局所的自動要約手法，情報処理学会第57回全国大会，6R-9 (1998).

[若尾 97]若尾孝博他：テレビニュース番組に見られる要約の手法，情報処理学会，自然言語処理研究会, NL122-13 (1997).

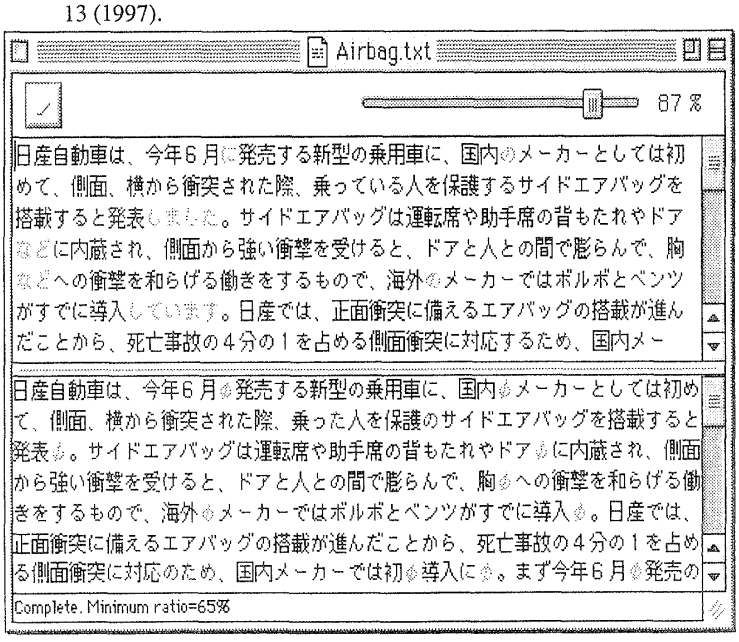

図３　局所的自動要約の例