フロー情報を対象にした情報検索システム (3)

4F－8

－ 文章圧縮 －*

柴田昇吾　上田隆也　伊藤史朗　廣田 誠　池田裕治　藤田 稔
キヤノン (株) 情報メディア研究所

1　はじめに

我々は，フロー情報からの情報収集・整理を支援するシステムの研究開発を進めている [1]．情報のライフサイクルを考慮してフロー情報を処理するためには，素早く内容を把握して「いる／いらない」を判断しなければならない．文章は，内容を把握するために読んで理解する必要があり，文章の長さに応じて時間がかかる．従来，紙をメディアとしていた新聞や雑誌などでは，効果的な見出しを用いたり，紙面の構成を工夫したりするなど，さまざまなノウハウを蓄積し，この問題に取り組んできた．

しかし，情報が紙から新しいメディアに変わることで従来のノウハウを変える必要が出てきた．我々は，新しいノウハウの一つとして，情報の受け手である我々が量を圧縮することが有効であると考えた．

本稿では，内容の概要を変えずに量を圧縮する文章圧縮について述べる．

2　文章圧縮の概要

表 1に示すように，文章を圧縮する方法は用途に応じていくつかの種類がある [2]．

表 1: 圧縮方法一覧

種　類	説　明
キーワード	文章から最も重要である単語を取り出したもの． 文章中に現れていない語でも良い．
主題	作品，論文，議論，研究などの中心となる思想． 「何がどうした」のようにまとめる．
大意的要約	文章から話の展開を維持したまま簡略化したもの． 原文の 1/3～1/4 が適量で，原文の引用が多い．
要旨的要約	著者の意図を中立的な観点から抜き出したもの． 原文からのパラフレーズが多く，分量は小さい．
情報抽出	文章から構造化情報を抽出する．

これらの圧縮方法がすべての文章に適しているとは限らない．例えば，新聞の社説では，事実などをもとに論説委員の意志を伝えるのが目的であるので，結論へ至った理由などが必要であり，要旨よりも大意が適している．

また，事件記事は，冒頭部分に結論が置かれ，以下，その説明や補足などが続くような構造になっている．事件記事は，事実そのものを伝えることが目的であるので，圧縮するには，記事の冒頭部分を取り出せば良い．これらは，どちらも大意と言えるが，前者と後者とを区別するため，それぞれを大意 (A)，大意 (B) と呼ぶ．

このように，文章によって適した圧縮方法が異なる．我々は，同じ方法で圧縮できる文章を収集・分類し，文章タイプを定義した．現在，文章タイプは，社説，解説，事件記事，会議案内，新製品記事，その他 がある．そして，図 1に示すように，文章タイプごとに圧縮方法を切り替える文章圧縮を作成した．

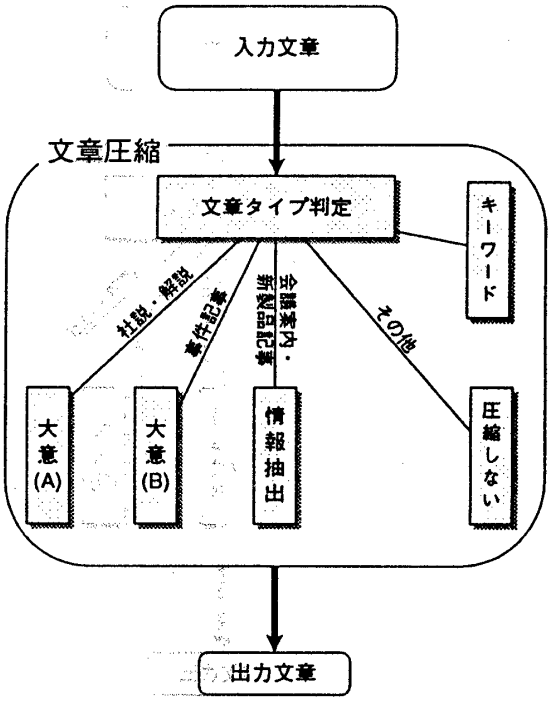

図 1: 文章圧縮の流れ

上述の文章タイプ以外には，会議案内から日時や場所を取り出したり，新製品に関する記事から発売日・企業名・製品名・価格などの情報だけを取り出す情報抽出が

*An Information Retrieval System for Flow-type Information (3) – Compression of sentences –

UEDA Takaya, SHIBATA Shogo, ITOH Fumiaki, HIROTA Makoto, IKEDA Yuji and FUJITA Minoru (Media Technology Laboratory, Canon Inc.)

適している．

なお，キーワードは，文章タイプにかかわらず全ての文章について抽出している．

文章圧縮は，文章のレトリックを考慮して行なう処理である．本来，レトリックは，文章を解析し文章構造から導き出すものであるが，大量の文章を短時間で処理するのは難しい．そこで，我々は，文章タイプを決めることで，文章のレトリックを近似的に決めることができると考えた．対象とするフロー情報を新聞の記事に限れば，文章の質が高く文章タイプに適したレトリックを用いているので，この手法が可能となる．今後は，フロー情報の種類が増えることを想定し，文章構造などを利用する方法へ改良して行く予定である．

3 各機能の特徴

以下に文章圧縮の各機能の特徴を述べる．

3.1 文章タイプ判定

文章タイプは，新聞であれば，タイトルや紙面などから判定できるが，今後，多種のフロー情報にも適応できるように，文章の内容のみで判定できることが望ましい．そこで，以下のような判定基準を用いて文章タイプを判定した．

付属語 事件や新製品記事などは事実を述べているので，ムードやモダリティを必要最小限に抑えている．一方，社説などはムードやモダリティを盛り込むので，終助詞や助動詞が多くなる．

また，格については，例えば，主語の省略が解説記事では少なく事件では多いなど，「が，は，を」の比率が異なる．

副詞 客観的に述べている新製品記事などには少ないが，主観の入る社説などには多い．

接続詞 時間順序で述べる事件記事などには接続詞が少ない．

代名詞 同様に新製品や事件記事には少ない．

これらの構成比率を調べ，各文章タイプの構成比率と比較し，最も類似したものを文章タイプとした．

3.2 キーワード抽出

文章の内容が把握できるようなキーワードを抽出することを目指した．我々が抽出したキーワードの特徴は以下の通りである．

長めの表現 複合名詞や“N の N”のパタンでは，長めのパタンをキーワードとした．例えば，「日本，アメリカ，赤字，黒字」より「日本の赤字」，「アメリカの黒字」とした方が内容を推測しやすいためである．

格役割の考慮 動詞の格のうち，格助詞「が，を」や副助詞「は」で接続している語句を優先させた．

文章中の位置の考慮 文章の最初の文や段落の最初の文は重要であることが多いので，その文中の語句を優先させた．

3.3 大意抽出

ここでは，図 1の大意 (A) について説明する．大意は，文章中の全文を評価し，重要である文を残し，それ以外を捨てることで作成した．各文の評価は，例えば文章タイプを社説とした GREEN [3] では，

- 文章の最初の文や段落の最初の文は重要であるといった**文章中の位置**
- 逆接でつながる文は重要であるといった**文の連接関係**
- 特定のパタンを含む文は重要であるといった**特定の表現**

によって行なう．我々は，さらに，

キーワード 前述の方法で作成したキーワードを含む文は重要であり，逆に，含まない場合は不要であるとみなして評価する

を評価項目に加えた．

また，大意として採用した文は，「A と B と C」を「A など」というように，並列表現を圧縮するなど，文の単純化を行なった．

4 おわりに

本稿では，フロー情報を対象にした情報検索システムの文章圧縮について述べた．この方法によって適切な圧縮方法が選択される精度は 78 %であった．そして，大意 (A)，(B) については量を 30〜40 %に圧縮することができた．今後は，

- 圧縮方法を充実させ圧縮できる文章タイプを増やす
- 文章タイプ判定，大意・キーワード抽出を同一処理内ですませることで速度を向上させる

ことにより，より大量の情報を圧縮し，情報洪水を緩和させる．

参考文献

[1] 上田他：フロー情報を対象にした情報検索システム (1) −概要−，情報処理学会 第 50 回全国大会 4F-6, 1995.

[2] 佐久間編：文章構造と要約文の諸相，日本語研究叢書，No.4, 1989.

[3] 山本他：文章内構造を複合的に利用した論説文要約システム GREEN，情報処理学会，自然言語処理研究会 99-3, 1994.