# 句表現要約手法に基づく要約システム

上田 良寛　岡 満美子　小山 剛弘　宮内 忠信
富士ゼロックス株式会社
{ueda, mamiko, koyama, miyauchi}@rsl.crl.fujixerox.co.jp

## 1 はじめに

情報検索の結果には検索意図に適合しない文書も含まれるので、通常は要約などを利用してふるい分けを行う。このように要約を indicative な目的に用いる場合、できるだけ短時間で正確な判断ができることが目標になる。

現状の自動要約は、単語の頻度や文の出現位置などの情報を用いて文をスコア付けし、スコアの高い文をピックアップする「抄録」と呼ぶべき手法を採用しているものが多い。この手法では、本文で用いられた文がそのまま用いられるため、次のような問題が生じる。

- 文が長く複雑になりがちで、文構造を頭の中で再構成しながら読む必要がある。
- 離れた場所からピックアップされた文は文間のつながり(結束性)が悪い。

このような文を読むのにはストレスを伴う。我々は、このようなストレスをできるだけ軽減するような要約、直感的な言い方になるが、「読む」のではなく**「一目で分かる("At-a-glance")」要約**が欲しいと考えた。これが、この研究を開始した動機である。

At-a-glance 要約のひとつの実現手法として、「句表現要約手法」を考案した。本論文では、この概念と、アルゴリズムについて述べ、この方法に基づいて開発を進めているプロトタイプを簡単に紹介する。また、タスクベース評価方法について言及する。

## 2 句表現要約の概念と要約手法

At-a-glance 性を重視したテキストの一例に、電車の中吊り広告で見られる雑誌広告がある。ここで示される記事の見出しは、その記事本体を読むか否かを判断するための情報で、まさに indicative 要約になっている。この特徴をまとめると

- 長さが短い
- 構成が単純

我々は、この単純さ、短さを「句」という言葉に代表させる[1]。句表現要約は、重要概念(単語)を含んだ短い句の並びを列挙することによって、「読む」負担を読者に与えずに、文書に記載される概要が把握できるようにするものである。

これを達成するために、**単語と単語の関係を基本単位として、それらを組み立てる**方法を採る。ここで関係としては、係り受け関係を採用している。係り受け関係は、構文関係を単語(特に自立語)間の関係のみに絞り、複数の単語からなる構成要素(節など)の関係は扱っていない。

### 要約手法

詳細は第 3 章で説明するが、ここでは句表現要約の概略を図 1 に例を示しながら説明する。基本的には次の 4 つのステップからなる。

(1) テキストを解析してその中に現れる単語と単語の関係を抽出する。
(2) 重要な関係をコアとして選択する。
(3) 意味的なまとまりを持たせるために必要な関係を補完する。
(4) 選択されたサブグラフから句を生成する。

まず、入力される文書中の文はそれぞれ関係解析がなされ、(b)のような基本関係を組み合わせたグラフを形成する。一つ一つのアークとその両端のノードが基本関係を構成している。基本関係のそれぞれの構成要素は、名詞連続も含む。

次に、重要な関係をコア関係として抽出する。ここでは、網のかけられた単語とそれを結ぶアークがコア関係として選択されたものとする。

コア関係だけでは、意味の特定度が低く、ふるい分けの情報としては不十分であるので、特定度を上げる為に必要な関係(ここでは二重線で囲んだ要素)を補完する。

このようにしてネットワーク中で選択された要素から、次に示すような短い句(日本語の場合、形式は基本的には文と変わらない)を生成する。

**モエギ社は環境保護技術をアカネ社に供与する**

このようにして組み立てられた短い句を複数並べることにより、文書の概要を把握させることが本方式の基本的考えである。次章では、アルゴリズムの全体像を示し、個々のステップを検討する。

[1] ここで用いる「句」の意味は、言語学的なものとは異なり、「短さ」を強調するための概念的なものであることに注意されたい。

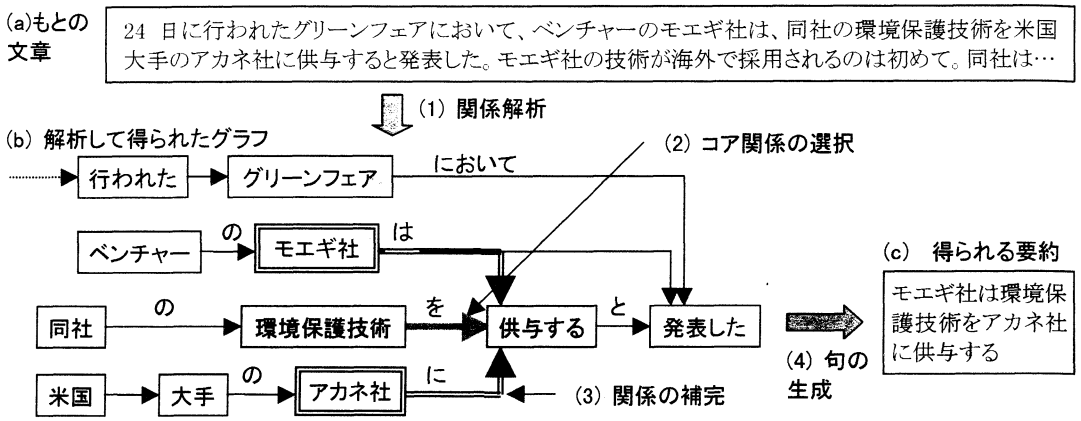


**図 1 句表現要約の概略**

## 3 アルゴリズム

### 基本構造

第 2 章で示したステップは 1 個の句を形成するステップである。これを、最初に設定した条件(要約全体の長さなど)を満たすまで繰り返す。このとき、用いた単語のスコアを一定割合で落とすことにより、同じ単語ばかりが繰り返し出現することを避け、キーワードがその重要度に応じた回数だけ出現することを可能にする。

アルゴリズムの基本構造を図 2 に示す。

以下では、必要に応じて補足説明を行う。

#### 関係解析

原文書を解析して、関係のネットワークとして表現する。ここでは、形態素解析結果の単語列に対して、パターンマッチにより関係を抽出している(Miyauchi, et al. 1995)。

#### 関係スコアリング

すべての基本関係に重要度スコアを付与する。

まず、すべての単語にスコアを付与する。スコア付けの方法としては、一般的な方法である tf*IDF 積(Salton 1989)のみを採用している[2]。

ただし、句の選択において tf の影響が強すぎる傾向が見られる。現在、tf の平方根をとるなどで tf によるスコアの伸びを抑えたり、スコア逓減率の調整などで適切なスコアリング方法を模索している。

関係スコアの計算式は次式で与える。

$$Score = Srel * (W1*S1 + W2*S2)$$

ここで、S1, S2 は関係でつながれる単語(係り側と受け側)のスコアであり、初期値は「単語スコアリング」の項で示したものである。複合語のスコアは、それを構成する単語のスコアの合計、または、複合語自体のスコアのいずれか高いほうを採用する。

W1, W2 はそれぞれに対する重み付けで、係り側、受け側のどちらを重視するかを決められるよう導入したものであるが、現在のところ W1 = W2 = 1 で係り側と受け側を等しく扱っている。

Srel は関係の種類に与える重要度である。動詞の格のように概念の中心的な役割を果たすものは大

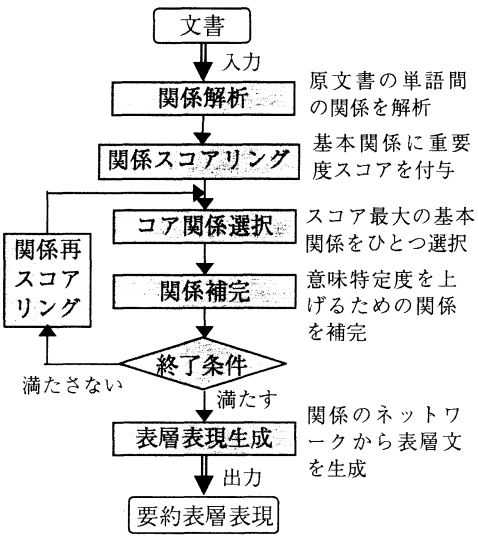


**図 2 アルゴリズム基本構造**

[2] IDF を決めるためには、文書の全体集合を規定する必要がある。WWW 文書の場合、その文書集合は時々刻々と変わっているものではあるが、ある時点で集めた 100 万文書を文書集合として用いている。また、これとは別に新聞記事から DF をカウントしたものも用意している(CD-毎日新聞 95 年版を利用させていただいた)。

きく、名詞の並列のように関係が周辺的と考えられるものは低く設定している(岡他 1999)。また、副詞のように修飾的な意味が強いものは関係そのものを選択しないので、Srel = 0 としている。

**コア関係選択**

スコア付けしたすべての関係中から、スコアの最も大きいものを選択する。

**関係補完**

意味的なまとまりを特定する上で不可欠な要素を加える。本稿では、(岡他 1999) に示した補完規則から一部を示す。

- **必須格**

  係り側、受け側のいずれかが用言の場合、必須格に当たる関係を追加する。必須格は動詞ごとに規定されるべきであるが、現在は一律に「が」関係、「を」関係、「に」関係を必須格関係として扱っている。また、係助詞「は」、「も」、格助詞「の」、無形格もこれらを置き換えうるものとして同じ扱いとする。

  例) フーバー社が発売する
  → フーバー社がPDAを発売する
  美しい女性 → 髪の美しい女性

- **用言に修飾される名詞**

  受け側単語が用言で、それによって修飾される名詞がある場合、用言に修飾される名詞は、埋め込み文中でなんらかの格をしめる場合が多い。現在は深い解析を行っておらず、被修飾名詞がどの格を占めていたかまで特定していないが、被修飾語には何らかの重要性があると判断し、用言から名詞への修飾関係を付加している。

  例) PDAを発売する
  → PDAを発売するフーバー社
  PDAを発売する → PDAを発売する計画

- **抽象度の高い名詞への修飾**

  「こと」、「もの」など形式名詞や、「場合」、「時代」など意味の特定度合いの低い名詞の場合、そのままで現れることは少なく、何らかの限定的な修飾句をともなう場合が多い。これらの名詞を受け側とする関係を付加する。

  例) 時代に活躍した → 激動の時代に活躍した

**終了条件の判定**

これまでに作ってきた要約で十分か否かを判定する。現在は、要約で用いた関係の数のみを終了条件にもちいている。今後、要約全体の長さ、単語の数、重要単語のカバー率など、種々の終了条件を選択して指定できるようにする予定である。

**関係再スコアリング**

条件を満たさない場合、もう一度コア関係の選択に戻る。このとき、このループで使われた単語のスコアを落とし、選択されにくくする。現在は、一定の逓減率 R (0 < R < 1)を設定し、使われた単語のスコアに掛けている。新しい単語スコアですべての関係のスコアを計算し直す。

**表層表現の生成**

このようにして、コア関係にいくつかの関係が付加されたネットワーク構造が得られる。このステップにおいては、これらの構成要素を出現順に線形化し、それぞれの表層表現をつなげていけばよい。この際に、係り受け解析では解析用の情報としてしか扱われず、得られた係り受け解析には登場しなかった助動詞、補助動詞、終助詞を付加する。得られた表層表現を出現順に列挙する。

## 4 プロトタイプ

前記アルゴリズムに基づく要約システム X-press (Xerox's Phrase-REpresented Summarization System)のプロトタイプを開発した。要約部分の開発言語は Java、関係解析部分は C++である。関係解析は、キーリレーションに基づく検索システム (Miyauchi, et al. 1995)で開発したものをサーバー化して利用している。開発環境として IBM Visual-Age for Java for Windows[3]を用いている。

要約作成時間は文書の長さに比例し、2000 文字 (A4 文書約 1 ページ)あたり約 900 msec の時間で要約を生成することができる。このうち、関係解析に要する時間は約 95 %を占めており、要約作成では約 40msec である (関係解析サーバ、要約部分ともに CPU が Pentium II 333MHz の場合)。

本論文を要約した結果を付録に付ける。

## 5 評価の方針

句表現要約の目的は、検索結果のふるい分けを速く的確にできるようにすることである。それを評価するために、現在タスクベースの評価を準備している。その方法の概略を示す。

- それぞれ方式の異なる要約を付加した WWW 検索結果を用意し、
- 被験者に、検索要求との適合度を、その要約で判定してもらう。
- 実際の文書でも適合度を判定してもらう。
- それぞれの判定した適合度の一致度を見る。

システム開発前に、アルゴリズムのシミュレーションによって作成した要約を用いて予備実験を行った結果では、重要文ピックアップによる要約よりも、検索結果のふるい分けが的確にできるという結果が得られている。現在この経験をふまえ、システムの

[3] VisualAge は米国 IBM Corp.、Java は米国 Sun Microsystems, Inc、Windows は Microsoft Corp.の商標である。

要約結果を使った評価実験を準備している。

なお、これとは別に、新聞記事をサンプルとして使った特性評価を行った。同じキーワード集合をカバーするのに要する要約全体の長さは、重要文ピックアップの約40%になった。逆に、同じ長さの要約に含まれるキーワード数では、重要文ピックアップは句合成の約61%となった。これらは、重要な概念をコンパクトに表現できることを意味し、短時間での正確なふるい分けに寄与すると考えられる。

また、一文あたりの長さは、重要文ピックアップが 50.05 文字であるのに対して、句表現では 18.42 文字となり、重要文ピックアップの 37%となった。ある週の週刊誌の広告における表題の長さを調べてみたところ、A 誌 18.4 文字、B 誌 15.4 文字、C 誌 20.6 文字であった。この点では想定する要約に近いものができたといえる。

## 6 関連研究

(Zechner 1996)をはじめとして、多くの研究が重要文ピックアップという方法を採用しており、その中で、どのように重要文を選択するかを問題にしている（奥村・難波 1998）。我々は、重要文ピックアップでは、長い文が選択されがちで、読む負担が大きいことを問題にし、関係を組み合わせて句を合成することによる要約の作成手法を提案した。ここでは、我々の視点で類似研究をまとめる。

まず、短い文にするという点では、文の言い替え、不要な修飾語の削除で文を短縮するという方向がある。(若尾他 1998）、(三上他 1998）は、ＴＶニュースにおいて、聞かせるための原稿から、読ませるための字幕を作成することを目的としている。この informative という性質上、情報をなるべく落とさないようにすることが必要で、あまり短くはできない。Indicative を目的とする場合はもっと短くできるはずであるが、文の中心構造を残し、修飾部分を減らす方向では、句表現要約ほど短い要約を作ることはできない。

文を合成するという立場が似ている研究としては、(Hovy and Lin 1997)、(近藤・奥村 1996)などがある。これらは、シソーラスなどを用いて複数の単語を上位概念で置き換えた文を構成することをねらっている。文章全体の意味を短い表現で置き換えることは要約の目指すところではあるが、単語レベルでの置換では適用範囲が限られるし、より大きな単位の置換が可能になると、知りたかったことが抽象化されすぎて見えなくなる問題も生じるだろう。

我々の句表現要約と同じく、語と語の関係をベースに要約を作るものには、(長尾他 1997)がある。これは、GDA(Global Document Annotation)という、著者が言語情報をあらかじめ文書中にタグとして付与しておくことにより、要約など文書の種々の機械的処理を可能にしようという試みである。必須格などの重要な関係を追加していく点など、手法に類似性があるが、目指す要約の形は At-a-glance を目指したものではない。

## 7 おわりに

本論文では、At-a-glance 要約の概念を提示し、この概念の具体化である句表現要約手法のアルゴリズムを示した。開発したプロトタイプを紹介した。また、タスクベース評価の方針を示した。

現在プロトタイプのチューンナップを続けている。また、本文中でも述べたように、実際のシステムの要約結果を用いた評価実験を行う予定である。

## 参考文献


Hovy, E. and Lin, C. Y. (1997). "Automated Text Summarization in SUMMARIST." *Proc. the Intelligent Scalable Text Summarization*, 18-24.

Miyauchi, T., Oka, M. and Ueda, Y. (1995). "Key-relation technology for text retrieval." *Proc. the SDAIR'95*, 469-483.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley.

Zechner, K. (1996). "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences." *Proc. COLING-96*, 986-989.

岡, 小山, 上田 (1999). "句表現要約の句合成手法" 自然言語処理研究会 129-15, 101-108.

奥村・難波 (1998). "テキスト自動要約技術の現状と課題." 言語処理学会第 4 回年次大会ワークショップ「テキスト要約の現状と将来」論文集, 80-87.

近藤・奥村 (1996). "言い替えを使用した要約の手法." 自然言語処理研究会 116-20, 137-142.

長尾, 橋田, 宮田 (1997). "GDA (Global Document Annotation)タグを用いた文書の要約に関する一考察." シンポジウム「実用的な自然言語処理に向けて」.

三上, 山崎, 増山, 中川(1998). "文中の重要部抽出と言い替えを併用した聴覚障害者用字幕生成のためのニュース文要約." 言語処理学会第 4 回年次大会ワークショップ「テキスト要約の現状と将来」論文集,14-21.

若尾, 江原, 白井 (1998). "テレビニュース字幕のための自動要約." 言語処理学会第 4 回年次大会ワークショップ「テキスト要約の現状と将来」論文集, 7-13.


## 付録：要約結果

本論文をプレーンテキストに変換し入力とした。

...**文をスコア付けし、スコアの高い文**...
...**句表現要約手法」を考案した**
**コア関係選択**
**受け側単語が**...**修飾される名詞**...
...**係り受け解析では解析用の**...**終助詞を付加する**
...**重要文ピックアップによる要約**...