ニュース文を対象にした
局所的要約知識の自動獲得

加藤 直人
NHK放送技術研究所

1 はじめに

文字情報が氾濫する中で、必要な情報だけを得るための技術として自動要約は重要である。自動要約には言語理解が本来必要であろうが、計算機に取り入れることは難しい。しかし、かな漢字変換や機械翻訳などのように、言語理解に踏み込むとある程度実現されている自然言語処理技術もある。

自動要約の技術でも言語理解を導入せずに、表層表現に基づいた手法が従来さまざまな提案されている。例えば手がかり語に基づいて要約する手法 [山本 95]、さらに単語の重要度を統計的に計算し重要文を抽出する手法 [渡辺 95] [Kupiec 96] [Zechner 96] 等である。前者の手法は手がかり語を収集しルール適用条件を人手で作成しなければならなく、後者は要約文章としての自然さに欠けるという難点がある。

我々は手がかり語をコーパスから自動抽出し、後者の方法も取り入れながら、なるべく入手を使わずに自動要約することを目指している。本稿では、その第一歩として原文と要約文から構成される記事コーパスから、局所的要約に必要な知識を自動獲得する手法について述べる。

2 局所的要約

日本語の文章を要約する場合には文字、単語、文節、節、文というレベルで言い換えや削除を行う。大幅に要約するには大域的に文章を見渡してから、節、文レベルの要約が必要となる。本稿では、節、文レベルの要約を大域的要約と呼ぶ。
一方、原文をあまり縮小せずによいのであれば、文字、単語、文節の局所的レベルのみの要約でよい。文字、単語、文節レベルの要約を局所的要約と呼ぶ。

3 原文と要約文のコーパス

要約知識を計算機によって自動獲得するためには、原文と要約文の対応がとれ、電子化されたコーパスが大量に必要となる。我々は原文にNHKニュース原稿、要約文にNHK文字放送ニュースを使った、それぞれの例を図1に示す。

NHKニュース原稿とは、主にNHK総合TV（GTV）で放送されているニュースの原稿である。この原稿は記者がワープロで書いたものであり、電子的に保存されている。

一方NHK文字放送ニュースとは、GTVの電波に重畳され放送されている文字放送（テレビジョン文字多重放送）の番組である。わずかな例外を除いては文字をコードで受信できる。文字放送ニュースは1記事が1画面の中に収まるように作成されているという特徴がある。

ニュース原稿と文字放送ニュースのペア約1,400記事を比較したところ、文字の数では、ニュース原稿が1記事当たり5～6文であるのに対して、文字放送ニュースはほとんどが2文であった。平均的には4.5文程度に縮約されていた。文字数では文字放送ニュースの1文は短く、ニュース原稿が約20%に縮約されていた。

図1の例で、ニュース原稿と文字放送ニュースの本文を具体的に比較してみる。第1文に関して文字放送ニュースからニュース原稿を見るとき、文字放送ニュースの単語列はニュース原稿にほとんど含まれている。逆に、ニュース原稿から見ると、対応がつかない単語列は文字放送中では省略されている。次に第2文を見ると、文字放送ニュースの第2文はニュース原稿の第2文の一部と第4文の一部から構成されており、「各種」以外はニュース原稿中に単語が存在するのがわかる。
4 局所的要約知識の自動獲得

図1の例では文字放送ニュースの第1文はニュース原稿の第1文を要約したものであったが、他の例を見て同様の場合が多いかった。一般にニュース文において第1文はリード文と呼ばれ、ニュースの概要が述べられた重要な文である。局所的要約知識の自動獲得には要約として対応が取れる場合が多い第1文のペアを用いる。自動獲得は原文であるニュース原稿と、要約文である文字放送ニュースの間で単語単位での距離を計算し、DPマッチングにより最適な単語対応付けを求める。その結果一致しなかった（後述のように実際には小さい値以下）の単語列の原文と要約文とのペアを局所的要約知識の候補とする。頻度統計をとることで、頻度が高いものを局所的要約知識とする。以下で本手法について具体的に説明する。

4.1 単語間の距離

はじめに単語間の距離を定義する。単語への分割やその品詞決定は形態素解析プログラムを使用することにより自動的に行っている。形態素解析の結果得られた、原文中の単語を$w_i/p_i$（$w$は表層文字列、$p$は品詞）、要約文中的単語を$w_j/p_j$とする。単語間の距離を式(1)のように3つの場合に分けて計算する。

$$\text{dist}(w_i/p_i, w_j/p_j)$$

(1)

$$\begin{align*}
\lambda_1 \text{dist}(w_i, w_j) + \lambda_2 \text{dist}(p_i, p_j) \\
\text{if } w_i/p_i \neq * \text{ かつ } w_j/p_j \neq * \\
\text{かつ ContWord}(p_i) \neq \text{ContWord}(p_j) \\
= -2.0 \\
\text{if } w_i/p_i \neq * \text{ かつ } w_j/p_j \neq * \\
\text{かつ ContWord}(p_i) \neq \text{ContWord}(p_j) \\
1.5 \\
\text{if } w_i/p_i = * \text{ または } w_j/p_j = *
\end{align*}$$

ここで、*は省略を表す記号であり、$w/p = *$は対応する単語が省略されたことを表す。また、ContWordは単語$w/p$が内容語であるかないかを、その品詞から判断する関数であり、式(2)で定義する。

$$\text{ContWord}(p) = \begin{cases} 
1 & \text{if } p \text{は内容語である品詞} \\
0 & \text{otherwise}
\end{cases}$$

式(1b)は、内容語と内容語でない単語が対応する場合であり、他の場合よりも大きい値にした。式(1c)は対応する単語が省略されている場合である。
式 (1a) は 2 つの単語が共に内容語であるか、共にそうでない場合であり、0 から 1 の値を取る。単語間の距離は、表層文字列間の距離と品詞列間の距離（ただし、$\lambda_1 + \lambda_2 = 1$）から計算する。ここで、表層文字列間の距離は単語内の文字列同士で DP マッチングを取った後、一致する文字数から式 (3) で計算する。

$$\text{dist}(w_{i,j}, w_{j,j}) = 1.0 - \frac{\text{hw}_{i} \cap \text{w}_{i,j}}{\text{hw}_{j}}$$  (3)

$h_w$ は単語の文字数、$\text{hw}_{i} \cap \text{w}_{i,j}$ は共通する文字数を表す。

例えば、表層文字列「自由民主」と「自民」は文字「自」と「民」が一致するので、

$$\text{dist}(自由民主, 自民) = 1.0 - 2/2 = 0$$

一方、品詞列の距離は品詞が一致するかしないかの 2 値で表した。すなわち

$$\text{dist}(p_{i}, p_{j}) = \begin{cases} 0 & \text{if } p_{i} = p_{j} \\ 1.0 & \text{otherwise} \end{cases}$$  (4)

今回は、表層文字列間距離も品詞間距離も単純に定義したが、詳細にはばらばらの結果が得られることが期待できる。例えば、式 (3) ではなくシーソーラスを使って単語間類似度を定義したり、式 (4) では品詞分類ごとに細かく距離を定義することが考えられる。

### 4.2 要約知識の自動獲得

要約知識の自動獲得では、単語間距離に基づいて原文と要約文との間で DP マッチングを取る。最小となる単語対応を求める（図 2 (a)）。

次に単語間距離にしきい値（今回 0.5）を設定し、単語対応を次の 3 つに分類する。

- **単語対応 1** 完全一致の場合
  
  $$\text{dist}(w_{i}^{p}, p_{i}^{p}, w_{j}^{p}, p_{j}^{p}) = 0$$

- **単語対応 2** 類似している場合
  
  $$0 < \text{dist}(w_{i}^{p}, p_{i}^{p}, w_{j}^{p}, p_{j}^{p}) \leq 0.5$$

- **単語対応 3** その他の場合
  
  $$0.5 < \text{dist}(w_{i}^{p}, p_{i}^{p}, w_{j}^{p}, p_{j}^{p})$$

単語対応 1 は原文の単語がそのまま保存されている場合である。単語対応 2 は単語レベルでの言い換えになり、局所的要約知識の候補となる。単語対応 3 は単語レベル、文節レベルでの言い換えであり、これらもまた局所的要約知識の候補となる。ここで、文節レベルでの言い換えは、単語対応 3 の連続する単語を集めて単語列として生成する（図 2 (b)）。この際、「(コンマ)」と「* (省略)」の対応は単語対応 3 ではあるが、ストップワードとした。

最終的な要約知識は、これら候補の頻度統計をとり、今回は単純に頻度の大きいものから選ぶことにした。

---

**図 2** 局所的要約知識の自動獲得の流れ

![図 2 局所的要約知識の自動獲得の流れ](image-url)
6 おわりに

NHKニュース原稿と文字放送ニュースを比較することにより，コーパスから局所的要約知識を自動的に獲得する手法について述べた。また，実験により有効性を確認した。

今後は得られた局所的要約知識を使って要約文を作成する実験を行う。しかし，単純に要約知識を適用すればよいというわけにはいかない。例えば，連体助詞「の」を省略される場合もあるし，要約文にそのまま残される場合もある。すなわち，どのような場合に局所的要約知識を適用すればよいかという，知識適用条件が必要となる。我々は要約知識が適用される場合の前後nグラムの文字，単語，品詞等の情報を自動収集することにより，人手を使わずに適用条件を抽出することを考えている。

また，地域的要約知識を自動獲得するために，原稿と要約文の2文字以降の文を比較する必要がある。この場合には，節や句文のような名詞を含むか，どのような接続詞で始まるのか，どのような助動詞，助動詞で終わるのかという知識を得ることが重要となる。これらは，単語の重み付け評価による重要文の抽出や修辞構造による文章構造の解析等の従来技術を参考にして研究を進めていく。

【参考文献】


[渡辺 95] 渡辺日雄「新聞記事の要約のための一手法」言語処理学会第1回年次大会, pp.293-296, 1995
