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，言語理解に踏み込まずとも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 95] [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，なるべく人手を使わずに自動要約することを目標としている．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，原文をあまり縮約しなくともよいのであれば，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．わずかの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45%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，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．

タイトル：鹿・Ｈ２五号機は来月二十日打ち上げ
日付：1998年01月29日17時46分
本文：
S1：搭載する衛星の性能試験中のミスで部品を取り替える必要が生じたため打ち上げが延期されていた国産の大型ロケットＨ２（エイチに）の５号機は来月（２月）２０日に鹿児島県の種子島宇宙センターから打ち上げられることになりました。
S2：Ｈ２の５号機は当初、通信放送技術衛星「かけはし」を搭載して来月１３日に打ち上げられる予定でしたが今月に種子島宇宙センターで行われた「かけはし」のテスト中に衛星の一部に誤って過大な電圧がかかってしまったため宇宙開発事業団ではＨ２の５号機の打ち上げを延期するとともに、衛星の部品を新しいものに取り替える作業を行っていました。
S3：その結果、修理作業はきょうで全て終わり、宇宙開発事業団ではＨ２ロケット５号機の打ち上げを来月２０日の午後４時５５分に行うことを決めました。
S4：今回打ち上げられる通信放送技術衛星「かけはし」はデジタル方式によるハイビジョン放送や衛星どうしでデータのやりとりをする技術などの開発に向けた試験を行うことになっています。

(a) 原文（＝ＮＨＫニュース原稿）

タイトル：Ｈ２－５号機来月２０日打ち上げ
日付：1998年01月30日
本文：
S1：部品取り替えで打ち上げが延期されていた国産大型ロケットＨ２－５号機は来月２０日午後４時５５分に種子島宇宙センターから打ち上げが決まった。
S2：Ｈ２－５号機は通信放送技術衛星「かけはし」を搭載、各種開発試験を行う。

(b) 要約文（＝ＮＨＫ文字放送ニュース）

図１　原文と要約文の例

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，他の例を見ても同様の場合が多かった．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にその距離を計算し，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（後述するように実際にはしきい値以下の）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．頻度統計をとり，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う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^o_i/p^o_i（wは表層文字列，pは品詞），要約文中の単語をw^s_j/p^s_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．

$$dist(w^o_i/p^o_i,\ w^s_j/p^s_j) \tag{1}$$

$$= \begin{cases} \lambda_1 dist(w^o_i, w^s_j) + \lambda_2 dist(p^o_i,\ p^s_j) & \text{if } w^o_i/p^o_i \neq * \text{ かつ } w^s_j/p^s_j \neq * \text{ かつ } ContWord(p^o_i) = ContWord(p^s_j) \quad (1a) \\ 2.0 & \text{if } w^o_i/p^o_i \neq * \text{ かつ } w^s_j/p^s_j \neq * \text{ かつ } ContWord(p^o_i) \neq ContWord(p^s_j) \quad (1b) \\ 1.5 & \text{if } w^o_i/p^o_i = * \text{ または } w^s_j/p^s_j = * \quad (1c) \end{cases}$$

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定する関数であり，式(2)で定義する．

$$ContWord(p) \tag{2}$$

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

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，共にそうではない場合であり，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．

$$dist(w^o_i, w^s_j) = 1.0 - \frac{|w^o_i \cap w^s_j|}{|w^s_j|} \qquad (3)$$

$|w|$はwの文字数，$|w^o_i \cap w^s_j|$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，

$dist($自由民主，自民$) = 1.0 - 2/2 = 0$

となる．

一方，品詞間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

$$dist(p^o_i, p^s_j) = \begin{cases} 0 & if\ p^o_i = p^s_j \\ 1.0 & otherwise \end{cases} \qquad (4)$$

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，詳細にすればさらによい結果が得られることが期待できる．例えば，式 (3) ではシ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マッチングを取り，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（今回は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

$dist(w^o_i / p^o_i, w^s_j / p^s_j) = 0$ のとき

[単語対応２]　類似している場合

$0 < dist(w^o_i / p^o_i, w^s_j / p^s_j) \leqq 0.5$ のとき

[単語対応３]　その他の場合

$0.5 < dist(w^o_i / p^o_i, w^s_j / p^s_j)$ のとき

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．

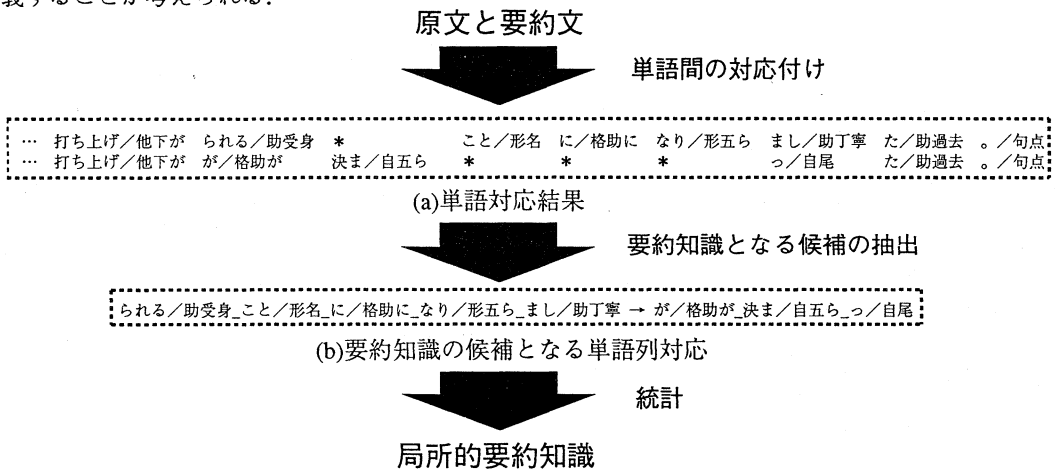

図２　局所的要約知識の自動獲得の流れ

5 実験

９８年１月までに得られた原文と要約文のペア約1,400記事を対象にして，本手法を使って要約知識の自動獲得の実験を行った．

実験結果の中から上位25個を表１に示す．ただし，表層文字列が一致するが品詞が異なる語（例えば，「と／引助と」と「と／格助と」）は除いた．また，単語列間の距離とは各単語間の距離の和である．

表１を見ると妥当な要約知識が得られている．上位には助詞や助動詞の省略が多く現れているのがわかる．例えば，連体助詞「の」は省略され，「なりまし」は「なっ」に言い換えられている．内容語では「総理大臣」が「首相」にと縮約した語に言い換えられている．表３には現れていないが，「日本を訪問している」を「来日中の」という複数文節からの言い換えの例もあった．

表１　本手法で得られた局所的要約知識
（ただし，$\lambda_1 = 0.9$，$\lambda_2 = 0.1$）

頻度	単語列間距離	原文の単語（列）	要約文の単語（列）
3001	1.5	、	[省略]
507	1.5	の	[省略]
469	1.5	まし	[省略]
100	1.5	て	[省略]
94	1.5	を	[省略]
66	2.4	総理大臣	首相
65	1.5	し	[省略]
54	1.5	な	[省略]
51	1.5	が	[省略]
48	1.5	に	[省略]
46	1.5	で	[省略]
45	1.5	は	[省略]
44	1.0	する	の
43	2.4	なりまし	なっ
41	1.0	アメリカ	米
39	1.5	」	[省略]
38	0.1	社民	社
38	0.1	さき	さ
37	0.1	自民	自
36	7.5	ことになりました	[省略]
36	4.5	しました	[省略]．
36	1.0	の	・
35	2.4	りまし	っ
33	2.5	いう	の
26	1.5	「	[省略]

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」は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，大域的要約知識を自動獲得するためには，原文と要約文の２文目以降の文を比較する必要がある．この場合には，節や文がど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ろう．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】

[Kupiec 95] J. Kupiec et al. "A Trainable Document Summarizer", In Proc. of the Fourteenth Annual International ACM SIGIR Conferrence on Research and Development in IR, pp.68-73, 1995

[渡辺 95] 渡辺日出雄「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

[山本 95] 山本ほか「文章内構造を複合的に利用した論説文要約システムＧＲＥＥＮ」自然言語処理，vol.2, No.1, pp.39-56, 1995.

[Zechner 96] K. Zechner "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences", In Proc. of Coling'96, Vol.2, pp.986-989, 1996