自然言語処理 126－10
(1998. 7. 24)

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

加藤　直人
ＮＨＫ放送技術研究所
〒157-8510 東京都世田谷区砧 1-10-11
katonao@strl.nhk.or.jp

あらまし 日本語ニュースを自動要約する際の言語知識となる，局所的要約知識を自動獲得する手法について述べる．局所的要約知識は置換知識と置換条件からなり，文字，単語，文節レベルでの言い換えを規定する．

本手法では原文とその要約文とのペアからなるコーパスを使って要約知識を自動獲得する．自動獲得では，はじめに原文中の単語と要約文中の単語のすべての組み合わせに対して単語間の距離を計算し，ＤＰマッチングによって最適な単語対応を求める．その結果より，置換知識は単語対応上で不一致となる単語列として得る．一方，置換条件は置換知識の前後ｎグラムの単語列として得る．

ＮＨＫニュースを使って局所的要約知識の自動獲得を行い，要約知識を評価する実験を行った．

キーワード　自動要約，コーパス，自動獲得，言語知識，日本語ニュース，ｎグラム

A new approach to acquiring linguistic knowledge for summarizing parts of news sentences and its evaluation

Naoto Katoh
NHK Science and Technical Research Laboratories
1-10-11 Kinuta Setagaya-ku Tokyo 157-8510 Japan
katonao@strl.nhk.or.jp

Abstract This paper proposes a new approach to acquiring linguistic knowledge that plays an important role in summarizing parts of news sentences. The linguistic knowledge, which is composed of transformation knowledge and transformation condition, can provide linguistic constraint of transforming characters, words, Bunsetsu-phrases in summarizing Japanese sentences.

The proposed method analyzes original news sentences and the human-summarized ones by Japanese morphological analyzer, and aligns words in the original sentences with words in the summarized ones by DP matching based on distances between the words. Transformation knowledge is acquired as the result of the difference and transformation condition is extracted as n-gram words located near transformation knowledge.

We acquired linguistic knowledge from NHK news corpus and conducted a series of experiments to evaluate the linguistic knowledge.

key words　automatic summarization, corpus, automatic acquisition, linguistic knowledge, Japanese news, n-gram

1　はじめに

インターネットの普及も手伝い，最近は電子化された文字情報を簡単にかつ大量に手にいれることが可能となってきている．このような状況の中で，必要な情報だけを得るための技術として文章要約は重要であり，計算機で自動的に行なえること（自動要約）が望まれる．

我々が最終的な目標と考えている自動要約は，人間が文章要約をする場合と同様，骨格となる文を大域的に探し，それらの文をそれぞれ局所的に縮約するというものである．本稿では前者を大域的要約，後者を局所的要約と呼ぶ．大域的要約は見方を変えれば，文章全体を見渡して不要な文を削除することが主であるが，文よりも小さい単位（日本語の合には節や文節）の削除も含める．一方，局所的要約とは日本語の場合に主に文字，単語，文節レベルの要約であり，その前後の情報のみで言い換えを行うことをいう．本稿では自動要約への第一歩として，局所的要約に焦点をあてる．以下，単に「要約」と書いた場合には局所的要約を指すものとする．

従来，局所的要約を扱った研究では，新聞記事を対象にしたもの [山本 95]，テレビニュース番組の字幕作成を対象にしたもの [若尾 97] がある．これらはいずれも，どの文字／単語／文節をどのように言い換えるか（置換知識），また，どのような場合に言い換えるか（置換条件）という要約知識を人手で作成していた．したがって，要約対象を拡大したり要約精度をあげるためには，要約知識を人手で増やさなければならないという困難がつきまとう．

本稿では局所的要約知識を自動的に獲得する手法について述べる．本手法では原文とその要約文とのペアからなるコーパスを使って要約知識を自動獲得する．自動獲得では，はじめに原文中の単語と要約文中の単語のすべての組み合わせに対して単語間の距離を計算し，ＤＰマッチングによって最適な単語対応を求める．その結果から置換知識は単語対応上で不一致となる単語列として得られる．一方，置換条件は置換知識の前後ｎグラムの単語列として得られる．

また，ＮＨＫニュースを使って局所的要約知識の自動獲得実験を行い，その有効性を検証する実験も行ったのでその結果についても述べる．

以下，２章では我々が使用している原文－要約文コーパスについて簡単に述べる．３章は本稿の中心であり，コーパスから要約知識を自動獲得する手法について説明する．４章では自動獲得された要約知識を使って，いくつかの要約知識に対して評価実験を行った結果について述べる．５章では今後の課題について述べる．

2　原文－要約文コーパス

原文－要約文コーパスについて簡単に紹介する．我々は原文にＮＨＫニュース原稿，要約文にＮＨＫ文字放送を使用している．ＮＨＫニュース原稿とは，主にＮＨＫ総合ＴＶ（ＧＴＶ）のニュース（例えば，「７時のニュース」）でアナウンサーが読む原稿の元になるものであり，電子的にも保存されている．一方ＮＨＫ文字放送ニュースとは，ＧＴＶの電波に多重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．文字放送は市販のデコーダーと受信ソフトにより計算機に取り込むことが可能である．

ＮＨＫニュース原稿は扱う話題によって違いはあるものの，大体５～６文から構成されており１文が長い．文字放送のニュースはテレビ１画面中に収まるように作成されており，ほとんどの場合が２文で構成されている．文字数で比較すると，ＮＨＫ文字放送はＮＨＫニュース原稿の約２０％に要約されている．

NHKニュース原稿，NHK文字放送の例を図１に示す．図１をみると，NHK文字放送の第１文はNHKニュース原稿の第１文から，第２文は第２，３文から要約されていることが分かる．このとき，NHKニュース原稿の第４，５，６文はNHK文字放送中では省略されている．また，対応が付けられる第１文同士を詳細に比較すると，次のような要約が行なわれているのがわかる．

「ごみの焼却場などから出る」（連体節）
→「φ」（φは省略を表す）
「有害物質のダイオキシン」
→「有害物質ダイオキシン」
「摂取基準を引き下げること」
→「摂取基準引き下げ」
「受けて」→「受け」
「国内の基準」→「国内基準」
「なりました」→「なった」

３　コーパスからの要約知識の自動獲得

３．１　要約知識

我々の要約知識は置換知識と置換条件の２種類からなる．置換知識とは原文の単語列を別の単語列に置き変えるという知識であり，例えば，次のようなものである．

【置換知識例】
「の／体助」→「φ」

また置換条件とは，置換知識を適用するか否かの条件である．すなわち，置換知識は必ず適用できるわけではなく，適用してはいけない場合もある．例えば，「日本の銀行」中の「の／体助」は省略できない．

以下では，置換知識と置換条件をコーパスから自動的に獲得する手法について具体的に説明する．

タイトル：朝用・ダイオキシン基準見直し
日付：1998年06月04日01時30分
本文：
文１：*ごみの焼却場などから出る有害物質のダイオキシンについてWHO・世界保健機関が先週、人体への摂取基準を引き下げることを決めたのを受けて厚生省も国内の基準を見直すことになりました。*
文２：*WHOはダイオキシンが人体に入っても健康に影響が出ないとされる量をこれまで１日あたり体重１キログラムにつき、１０ピコグラム、つまり１兆分の１０グラムと定めていて、厚生省もこの基準を取り入れています。*
文３：*しかしWHOのヨーロッパ事務局などの専門家会議は先週ジュネーブで開いた会合で、ダイオキシンが乳幼児に及ぼす影響や「環境ホルモン」として生殖機能に与える影響などを考慮してこれまでの「１０ピコグラム」を引き下げて「１から４ピコグラム」を望ましい基準とすることを決めました。*
文４：このため厚生省は今月中にも専門家による検討会を作り、国内の基準をさらに厳しくすることを検討することにしたものです。
文５：これとあわせてごみ焼却場から排出される基準や大気中の基準などについても見直すとともに食品についても安全性の目安となる基準が必要かどうか検討することにしています。
文６：また厚生省ではダイオキシンとよく似た構造を持ち、毒性のある「コプラナーＰＣＢ」についても今後、ダイオキシン類として規制の対象に加えることを検討することになりました。

（ａ）NHKニュース原稿（原文）

タイトル：厚生省ダイオキシン国内基準見直しへ
日付：1998年06月04日
本文：
文１：有害物質ダイオキシンについてWHO＝世界保健機関が先週人体への摂取基準引き下げを決めたのを受け厚生省も国内基準を見直すことになった。
文２：WHOは１日当たり体重１ｋｇにつき「１０ピコグラム＝１兆分の１０ｇ」を引き下げて「１～４ピコグラム」を望ましい基準とした。

（ｂ）NHK文字放送（要約文）

図１　NHKニュース原稿とNHK文字放送の例

3．2　置換知識

本手法では，はじめに原文－要約文コーパスのそれぞれの文を形態素解析し，単語単位に分割する．原文中の単語と要約文中単語のすべての組み合わせに対して単語間の距離を計算し，その距離に基づいてＤＰマッチングを取る．単語間の距離は式（1）のように３つの場合に分けて定義した．

【単語間の距離】

$$distword(w_i, x_j) = distword(c^o{}_i / p^o{}_i, c^s{}_j / p^s{}_j) \quad (1)$$

$$= \begin{cases} \lambda_1 distchar(c^o{}_i, c^s{}_j) + \lambda_2 distpos(p^o{}_i, p^s{}_j) & (1a) \\ \quad if\ w_i \neq \phi\ かつ\ x_j \neq \phi \\ \quad かつ\ ContWord(p^o{}_i) = ContWord(p^s{}_j) & \\ 2.0 & (1b) \\ \quad if\ w_i \neq \phi\ かつ\ x_j \neq \phi \\ \quad かつ\ ContWord(p^o{}_i) \neq ContWord(p^s{}_j) & \\ 1.5 & (1c) \\ \quad if\ w_i = \phi\ または\ x_j = \phi & \end{cases}$$

ここで，ϕは省略を表す記号であり，$w_i = \phi$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_iが内容語であるかないかをその品詞（p）から判定する関数であり，式（２）で定義する．

【内容語判定関数】

$$ContWord(p) \quad (2)$$

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

式（１）の中で，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を重み付け（$\lambda_1 + \lambda_2 = 1$）して計算する．表層文字列間の距離は内容語同士の意味的な距離であり，シソーラスを使って計算される．実際には角川類語新辞典 [角川] の分類番号の一致する桁に基づき，式（３）で計算している．

【表層文字列間の距離】

$$distchar(c^o{}_i, c^s{}_j) \quad (3)$$

$$= \begin{cases} 0 & if\ c^o{}_i = c^s{}_j \quad (3a) \\ 0.2 & if\ 上位３桁のみが一致 \quad (3b) \\ 0.4 & if\ 上位２桁のみが一致 \quad (3c) \\ 0.6 & if\ 上位１桁のみが一致 \quad (3d) \\ 1.0 & otherwise \quad (3e) \end{cases}$$

また，品詞間の距離は，式（４）のように定義した．

【品詞間の距離】

$$distpos(p^o{}_i, p^s{}_j) \quad (4)$$

$$= \begin{cases} 0 & if\ p^o{}_i = p^s{}_j \quad (4a) \\ 0.5 & if\ p^o{}_i と p^s{}_j は人手で指定したもの \quad (4b) \\ 1.0 & otherwise \quad (4c) \end{cases}$$

ここで式 (4b) は，「名詞とサ変名詞」のように，完全一致しないが類似している品詞同士であり，人手で指定した．しかし，現在のところその数は数個と非常に少ない．

単語間の距離に基づいて単語間のＤＰマッチングをとると，図２のように，前の単語列が一致し（単語数$q1$個），一部が不一致となり（p個），その後にまた単語列が一致す

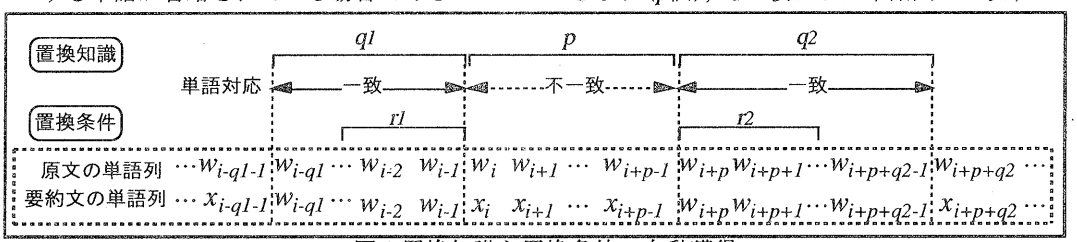

図２置換知識と置換条件の自動獲得

る（$q2$個）という部分が求められる．この不一致となる単語列が置換知識となる．一致する部分が長く，不一致の部分が短いほうが置換知識としての信頼性が高いと考えられるので，式（5）で表される評価関数を定義すると，この値の大きいほうが知識として有効である．

【置換知識の評価関数】

$$f(w_i\, w_{i+1} \cdots w_{i+p-1}, x_i\, x_{i+1} \cdots x_{i+p-1}) = \frac{q1+q2}{p} \quad (5)$$

実際には評価関数式（5）の値が，しきい値（f_0）より大きいものを収集し，さらに収集された置換知識の統計をとり，獲得された個数が多い単語列を有効な置換知識として使用している．

3．3　置換条件

置換条件は，置換知識の前後の単語nグラムで定義されている．置換条件は，原文の単語列が置換される場合 $w_i\, w_{i+1} \cdots w_{i+p-1} \rightarrow x_i\, x_{i+1} \cdots x_{i+p-1}$（正例）とともに，原文の単語列がそのまま保存される場合 $w_i\, w_{i+1} \cdots w_{i+p-1} \rightarrow w_i\, w_{i+1} \cdots w_{i+p-1}$（負例）も収集している．負例を自動獲得する場合にも式（5）による信頼度を使っている．

置換知識の前nグラムを置換前条件，後nグラムを置換後条件と呼び，それぞれのnの値をそれぞれ$r1$，$r2$とおく．すると，要約知識は図3のように表すことができる．

置換知識： $w_i\, w_{i+1} \cdots w_{i+p-1} \rightarrow x_i\, x_{i+1} \cdots x_{i+p-1}$

置換条件：

	置換前条件	置換後条件
正例	$w^1_{i-r1} \cdots w^1_{i-2}\, w^1_{i-1}$	$w^1_{i+p}\, w^1_{i+p+1} \cdots w^1_{i+p+r2-1}$
負例	$w^2_{i-r1} \cdots w^2_{i-2}\, w^2_{i-1}$	$w^2_{i+p}\, w^2_{i+p+1} \cdots w^2_{i+p+r2-1}$
：	：	：
正例	$w^k_{i-r1} \cdots w^k_{i-2}\, w^k_{i-1}$	$w^k_{i+p}\, w^k_{i+p+1} \cdots w^k_{i+p+r2-1}$
：	：	：

図3　置換知識と置換条件

3．4　局所的自動要約

局所的自動要約をする際にはまず，原文の先頭の単語列から順に原文中の単語列（$w_i\, w_{i+1} \cdots w_{i+p-1}$）と，置換知識上の単語列（$x_i\, x_{i+1} \cdots x_{i+p-1}$）との照合を行う．次に照合が一致した場合には置換条件上の距離計算を行うことにより，その置換知識を適用することの良否を計算する．ここで，置換条件上の距離は置換条件上のすべてのデータに対する，原文の単語列の前$r1$グラム（$w_{i-r1} \cdots w_{i-2} w_{i-1}$）と$k$番目の置換前条件（$w^k_{i-r1} \cdots w^k_{i-2}\, w^k_{i-1}$），原文の単語列の後$r2$グラム（$w_{i+1}\, w_{i+2} \cdots w_{i+r2}$）と$k$番目の置換後条件（$w^k_{i+p}\, w^k_{i+p+1} \cdots w^k_{i+p+r2-1}$）との距離の最小値として式（6）で定義する．

【置換条件上の距離】

$$\min_k \left(g(w_i\, w_{i+1} \cdots w_{i+p-1},\, x_i\, x_{i+1} \cdots x_{i+p-1},\, k) \right) \quad (6)$$

$$g(w_i\, w_{i+1} \cdots w_{i+p-1}, x_i\, x_{i+1} \cdots x_{i+p-1}, k) \quad (6a)$$

$$= \mu_1 g^-(w_{i-r1} \cdots w_{i-2}\, w_{i-1},\ w^k_{i-r1} \cdots w^k_{i-2}\, w^k_{i-1}) + \mu_2 g^+(w_{i+p}\, w_{i+p+1} \cdots w_{i+p+r2-1},\ w^k_{i+p}\, w^k_{i+p+1} \cdots w^k_{i+p+r2-1})$$

$$(\mu_1 + \mu_2 = 1)$$

$$g^-(w_{i-r1} \cdots w_{i-2}\, w_{i-1},\ w^k_{i-r1} \cdots w^k_{i-2}\, w^k_{i-1}) = \frac{\sum_{j=1}^{r1} \{weight_1(j) \times distword(w_{i-j},\ w^k_{i-j})\}}{\sum_{j=1}^{r1} weight_1(j)} \quad (6b)$$

$$g^+(w_{i+p}\, w_{i+p+1} \cdots w_{i+p+r2-1},\ w^k_{i+p}\, w^k_{i+p+1} \cdots w^k_{i+p+r2-1}) = \frac{\sum_{j=1}^{r2} \{weight_2(j) \times distword(w_{i+p+j-1},\ w^k_{i+p+j-1})\}}{\sum_{j=1}^{r2} weight_2(j)} \quad (6c)$$

$$weight_1(j) = \alpha_1{}^{j-1} \quad (6d)$$

$$weight_2(j) = \alpha_2{}^{j-1} \quad (6d)$$

（α_1，α_2：定数，$0 \leqq \alpha_1 \leqq 1$，$0 \leqq \alpha_2 \leqq 1$）

ただし，g^-，g^+はそれぞれ，原文と収集された置換前条件，置換後条件間の距離を計算する関数であり，置換知識となる単語列から離れるほど，その影響が少なくなるように*weight(j)*で重み付けされている．さらに，置換前条件と置換後条件の影響の度合いμで重み付けされている．

最終的に式（6）の最小値を与えるものが正例であれば置換知識が適用され局所的に要約される．しかし，置換知識の適用を式（6）で単純に判定してしまうと，負例，すなわち置換知識を適用しない方を解とする場合が多くなってしまう．これは，置換条件の正例が置換しなければならないというものではなく，置換してもよいという程度の意味しか持たないからである．そこで後述する要約知識の評価実験では，あるしきい値（g_0）を決め，式（6）で求められた最小値を与える解が負例であっても正例での最小値がしきい値以下であるならば，正例を解とした．

4　評価実験

4．1　要約知識獲得実験

ＮＨＫニュース原稿とＮＨＫ文字放送から構成される，原文－要約文コーパスの約３４００記事を使って要約知識を自動獲得する実験を行った．

要約知識のうち，まず置換知識の自動獲得実験を行った．このときの各パラメータの値は次のようにした．

- 表層文字列間距離と品詞間距離の割合（式 (1a) ）

  $\lambda_1 = 0.7$（$\lambda_2 = 0.3$）
- 自動獲得のしきい値（式（5））

  $f_0 = 1.0$

自動獲得された置換知識の一部を表１に示す．表１をみると，妥当な要約知識が得られているのがわかる．上位には「の/体助」や「を/格助を」のように，助詞や助動詞が省略されているのが多い．内容語では「総理/名　大臣/名」が「首相/名」に，「委員/名　会/尾」が「委/尾」に言い換えられている．

表１　自動獲得された置換知識

獲得個数	原文中の単語列	要約文中の単語列
3353	、/読点	φ
2029	の/体助	φ
1334	まし/助丁寧	φ
533	を/格助を	φ
213	に/格助に	φ
207	で/助断定	φ
200	が/格助が	φ
176	」/閉かぎ	φ
168	総理/名 大臣/名	首相/名
163	な/助断定	φ
152	です/助断定	φ
141	や/並助	・/つなぎ
140	し/さ連用	φ
138	する/さ連体	の/体助
132	て/接助	φ
130	など/副助	φ
127	てい/助完了 ます/助丁寧	ている/助完了
123	まし/助丁寧 た/助過去	φ
117	い/形五わ う/自尾	の/体助
111	てい/助完了 ます/助丁寧	φ
106	は/係助は	φ
101	し/さ連用 まし/助丁寧 た/助過去	φ
100	て/接助 、/読点	φ
100	「/開かぎ	φ
89	委員/名 会/尾	委/尾
85	てい/助完了	φ
82	アメリカ/地	米/名
81	なり/形五ら まし/助丁寧	なっ/形五ら
79	を/格助を	の/体助
77	また/接	φ

次に自動獲得された置換知識に対して，置換条件を正例，負例ともに自動獲得した．この際，$r_1 = r_2 = 6$とした．置換知識『「の／体助」→「φ」』に対する置換条件の例の一部を表２に示す．表２をみると，「の／体助」の置換条件はその前後の１単語に大きく依存していることがわかる．

表２　置換知識『「の/体助→「φ」』に対する置換条件の例の一部

	置換前条件（6-gram）	置換後条件（6-gram）
負例	と/格助と なる/形五ら 初代/名 の/体助 行政/名 長官/名	人選/サ名 や/並助 、/読点 ポルトガル/地 の/体助 統治/サ名
正例	マカオ/地 の/体助 トップ/組 と/格助と なる/形五ら 初代/名	行政/名 長官/名 の/体助 人選/サ名 や/並助 、/読点
負例	背景/名 に/格助に は/係助は 、/読点 産業/名 構造/名	転換/サ名 で/助断定 規模/名 が/格助が 縮小/両名 する/さ連体
正例	に/格助に 中国/地 の/体助 国連/組 代表/サ名 部/尾	副/頭 代表/サ名 や/並助 、/読点 外務省/名 の/体助
正例	の/体助 副/頭 代表/サ名 や/並助 、/読点 外務省/名	スポークスマン/片未 など/副助 を/格助を 歴任/両名 し/さ連用 た/助過去
負例	を/格助を めざ/他五さ す/自尾 四/名 者/尾 協議/サ名	中国/地 政府/名 の/体助 代表/サ名 も/係助も 務め/他下ま
負例	ハノイ/地 で/助断定 、/読点 就任/サ名 後/尾 初/名	記者/名 会見/サ名 を/格助を 行/他五わ い/自尾 、/読点
負例	行政/名 改革/サ名 は/係助は 、/読点 橋本/人姓 内閣/名	最/頭 重要/容名 課題/名 であり/助指定 、/読点 早期/名
負例	てい/助完了 ます/助丁寧 が/接助 、/読点 各/頭 党/名	主張/サ名 に/格助に は/係助は 隔た/自五ら り/自尾 が/格助が
負例	べきで/助断定 、/読点 政府/名 案/名 と/格助と 同様/容名	弾力/名 条項/名 を/格助を 盛り込/自五ま む/自尾 改正/サ名
負例	など/形名 、/読点 各/頭 党/名 の/体助 思惑/名	違/自五わ い/自尾 も/係助も 表面/名 化/化尾 し/さ連用
負例	限/他五ら る/自尾 など/形名 一定/サ名 の/体助 制限/サ名	中/副形名 で/助断定 、/読点 代表/サ名 質問/サ名 の/体助
負例	から/格から 1/数 か月/後助 間/別尾 、/読点 拳銃/名	集中/サ名 取り締ま/他五ら り/自尾 を/格助を 行/他五わ う/自尾
負例	社/造 民党/未２ は/係助は 今/副名 の/体助 国会/名	会期/名 中/尾 に/格助に 橋本/人姓 内閣/名 へ/格助へ
正例	し/さ連用 、/読点 社/造 民党/未２ は/係助は 今/副名	国会/名 の/体助 会期/名 中/尾 に/格助に 橋本/人姓
負例	会期/名 中/尾 に/格助に 橋本/人姓 内閣/名 へ/格助へ	閣外/未２ 協力/サ名 を/格助を 解消/両名 せ/さ未１ ざる/助打消
負例	部/尾 として/接助 は/係助は 、/読点 社/造 民党/未２	意見/名 を/格助を 大事/名 に/格助に し/さ連用 て/接助
負例	長/尾 は/係助は 、/読点 「/開かぎ ３/数 党/後助	枠組み/名 に/格助に 留ま/自五ら る/自尾 べきだ/助断定 と/引助と
正例	、/読点 焦点/名 の/体助 ヨルダン川/地 西岸/未２ から/格から	イスラエル/地 軍/名 の/体助 撤退/サ名 問題/名 を中心に/格助他
正例	合意/両名 し/さ連用 た/助過去 日/名 ロ/地 共同/サ名	投資/サ名 会社/名 について/格助他 、/読点 初めて/副 の/体助

４．２　要約知識評価実験

次に４．１で自動獲得された要約知識の評価実験を行った．評価には今回，表１の上位のうち助詞，助動詞の要約に関するものを選んだ．実験は，正例と負例をあわせた全データの90％をランダムに抽出して学習データとし，残り10％をopenテスト用データとした．closedテスト用データはopenテスト用データと同じ量を学習データからランダムに抽出した．

評価実験では，はじめにいくつかのパラメータ（λ_1, α_1, α_2, μ_1）を決定しなければならない．今回，λ_1は要約知識獲得実験で使った値を使い，α_1, α_2, μ_1は置換知識『「な／助断定」→「φ」』を用いていくつかの値で実験を行った．結果を表３に示す．表３のopenテストの結果をみると，表３ (a) は$\alpha_1 = 0.5$, $\alpha_2 = 0.5$にし，μ_1をいくつか値を変えてみた結果であるが，$\mu_1 = 0.8$（$\mu_2 = 0.2$）の場合が一番良い．これは置換前条件の方が置換後条件より重要であることを示している．表３ (b) は$\alpha_2 = 0.5$, $\mu_1 = 0.8$としてα_1の値をいくつか変えた結果であり，$\alpha_1 = 0.4$で飽和しているのがわかる．表３ (c) は$\alpha_1 = 0.4$, $\mu_1 = 0.8$としてα_2の値をいくつか変えた結果であり，$\alpha_1 = 0.4$で飽和している．そこで，今回は以下の実験のパラメータとして$\alpha_1 = 0.4$, $\alpha_2 = 0.4$, $\mu_1 = 0.8$を使った．もちろんもっと様々な実験をすれば，さらに適切なパラメータが求められると考えられる．

表３『「な／助断定」→「φ」』の実験結果

(a) パラメータの決定（$\alpha_1 = 0.5$, $\alpha_2 = 0.5$）

α_1	α_2	μ_1	closed(%)	open(%)
0.5	0.5	0.0	98.9	73.9
0.5	0.5	0.2	100	74.6
0.5	0.5	0.4	100	74.4
0.5	0.5	0.6	100	75.1
0.5	0.5	0.8	100	75.6
0.5	0.5	1.0	99.8	74.2

(b) パラメータの決定（$\alpha_2 = 0.5$, $\mu_1 = 0.8$）

α_1	α_2	μ_1	closed(%)	open(%)
0.1	0.5	0.8	100	75.0
0.2	0.5	0.8	100	76.8
0.3	0.5	0.8	100	76.8
0.4	0.5	0.8	100	78.6
0.5	0.5	0.8	100	78.6
0.6	0.5	0.8	100	78.6

(c) パラメータの決定（$\alpha_1 = 0.4$, $\mu_1 = 0.8$）

α_1	α_2	μ_1	closed(%)	open(%)
0.4	0.1	0.8	100	76.8
0.4	0.2	0.8	100	76.8
0.4	0.3	0.8	100	76.8
0.4	0.4	0.8	100	78.6
0.4	0.5	0.8	100	78.6
0.4	0.6	0.8	100	78.6

表４　各局所的要約知識に対する実験結果

局所的要約知識		正例の獲得個数	負例の獲得個数	closed (%)	open (%)
原文中の単語列	要約文中の単語列				
の／体助	φ	2029	5087	99.7	74.8
を／格助を	φ	533	3968	99.9	81.3
に／格助に	φ	213	2470	99.9	90.5
で／助断定	φ	207	967	100	85.2
が／格助が	φ	200	1732	99.7	86.6
な／助断定	φ	163	403	100	78.6
し／さ連用	φ	140	869	100	95.9
する／さ連体	の／体助	138	655	99.9	79.2
て／接助	φ	132	321	100	73.3

次にこのパラメータを使って他のいくつかの置換知識に対しても評価実験を行った．ただし，$g_0=0.1$とした．結果を表４に示す．

表４のopenテストの結果をみると，置換知識によって精度のばらつきがあるはあるものの，概ね良好な結果であると思われる．今後は誤りに対する分析を行い，精度の向上をはかる予定である．

５　おわりに

原文－要約文コーパスより局所的要約知識を自動獲得する手法について述べた．また，ＮＨＫニュース原稿とＮＨＫ文字放送から構成されるコーパスを使って，局所的要約知識を自動獲得する実験を行った．さらに要約知識の評価実験を行い，良好な結果を得た．

今後の研究の方向は２つある．１つは局所的要約に関するものである．今回は評価実験の第一歩として特定の局所的要約知識にのみ着目したが，今後は文全体の局所的要約を試みたい．その際には，適用する要約知識間で競合が起こることが予想される．そこで，ある要約率の中で最も信頼性が高い局所要約を求めるアルゴリズムが必要なる．我々は信頼度の評価関数を最適化することにより，最適な局所的要約を求めるアルゴリズムを研究している．

もう１つの方向は，大域的要約に関するものである．これには要約には現れなかった元のニュースの文（例えば，図１の文４，５，６）や節（図１の文２，３中の節）を，文や節の削除手法の研究の評価用データとして使っている．現在，従来の評価関数（例えば，tf法やtf*idf法）を使ってどのくらいの精度で削除できるかを実験中である．

これらの詳細については稿を改めて報告したい．

【参考文献】

[角川] 大野ほか「角川類語新辞典」，角川書店，1997.

[若尾 97] 若尾ほか「テレビニュース番組に見られる要約の手法」，情報処理学会研究報告，NL-116-23, pp.137-142, 1997.

[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.