ニュース速報記事の前文情報との照合に基づく見出し文の言い替え

安達 久博
宇都宮大学 工学部 情報工学科
e-mail: adachi@is.utsunomiya-u.ac.jp

1 はじめに

近年，情報化社会の進展に伴い，以前に比べて様々な文書情報が電子化され蓄積されつつある．同時に，それらの情報，特に文字情報などは，比較的容易にインターネット上から検索できる環境も整備されてきている．このような背景から，最近は，これら膨大な情報資源から個々のユーザが必要とする情報を取捨選択する必要性があり，文書の要約技術に関する研究が各所で盛んに行なわれている (山本, 増山, 内藤 1995, など)．この要約技術の重要な要素技術の一つとして，文の言い換え（パラフレーズ）に焦点を当てた研究についても幾つかの成果が報告されている (佐藤 1999, など)．そこで，高齢者や視聴覚障害者 (乾, 山本, 野上, 藤田, 乾 1999)，日本語を母語としない人達に，日本語文で提供される情報を円滑に獲得できるための支援システムを実現する要素技術となる，単言語内での言い替え（ことばの位相変換）技術 (安達 1999) の一実現方法について検討した．対象として取り上げた題材は，災害時や日常の事件・事故や内外の情勢をいち早く獲得できるニュース速報文の見出し文とした．見出し文は，必要な情報にアクセスするための最初の手がかりと捉えることができる．しかしながら，字数の制約のため，助詞が省略され，難解な複合語による表現や体言止めなどが多用され，日本語を母語とする人も背景知識なしでは理解しづらい文形式と捉えることができる．見出し文は，本文記事への導入の役割と同時に，本来，見出し文だけである程度の情報を把握できる役割を担っているはずである．本稿では，後者の役割に着目し，ニュース速報文の見出し文の省略された部分を復元・補完し，見出し文と前文情報との中間に位置する情報量を確保する文を生成する言い替え手法を提案する．本手法の特徴は，従来の単独の入力文 (章) に体する言い替えや要約方法と異なり，前文記事と見出し文のペアを入力とし，両者間の共通部分を検出し，相違部分を利用し，見出し文を言い替える，あるいは，前文を要約したものを出力とする点にある．そのため，従来手法が，日本語入力文の単語分割，形態素解析処理を前提としていた部分を共通部分列探索で代用し，規則に基づく言い替え処理を相違部分文字列に着目して変換するなどの違いがある．これは，(山本, 増山, 内藤 1996) のように複数の文書を入力とするタイプの言い換え機構といえる．

2 ニュース速報記事の構造と特徴

新聞などのニュース記事の多くは，一般に，見出し文（ヘッドライン）と前文（リード），本文の階層的な論理構造を成している．また，見出し文は前文の要約，前文は本文の要約と捉えることができる．さらに，前文は通常，2～4文で構成されており，特に，最初の文に結論が述べられており，以下に示すように前文の論理の流れが構成されてる．すなわち，紙面の制約により，後半部分の文が削除されても伝達すべき情報の中心部分は，確実に残される構造といえる (小笠原 1991)．

結論 ⇒ 事実の概要 ⇒ 理由、影響、その他

以下，本研究で使用したニュース速報記事の特徴とその構造について概説する．テキスト版として提供されている共同通信社のホームページ[1]の速報文は，速報性のため，頻繁に更新されている．収

[1] http://www.kyodo.co.jp/texthome.html

録記事は20記事であり，図1に示したように，一記事あたり，見出し文の文字数は約20文字，前文の文字数は約250文字で構成されている．また，見出し文は大きく二つに分かれており，実際の新聞紙面上では左側が大見出し，右側が小見出しのように反映されてる．一方，見出し文は，現在形あるいは体言止めが多く，未来を表現する場合は「～へ」の助詞,「いよいよ」や「近く」などの時に関する副詞が用いられる傾向がある．また，文字数の制約により助詞を省略した複合語が多用されているなどの特徴がある．なお，図1に示した例では，前文の第一文を縮訳した形の見出し文であるが，複数の文が要約された形式の見出しも存在する．

人工視覚装置の実用に前進＝米大が動物実験開始へ

9日付の米紙ワシントン・ポストによると、米ジョンズホプキンス大の研究チームは眼鏡に取り付けた小型のビデオカメラと目の中に埋め込む高性能の集積回路によって人工的に視覚を生み出す装置の実用化に向け、近く動物実験を開始する。実用化に成功すれば、網膜の損傷などで目が見えない米国内の約２００万人が恩恵を受けられる。

図 1: 速報文の一例（共同通信社・主要ニュースより抜粋引用）

3 見出し文の言い換え

本稿で提案する手法は，以下に示す二つの処理から成る.（1）見出し文の前文中での近似的照合位置を計算する検索処理（2）照合位置の前後の連接文字情報に基づく文生成処理

3.1 文字列間の近似的照合方法

見出し文の各単語の多くは，前文に含まれているが，必ずしも完全一致検索で照合することはできない．先述したように，見出し文は複合語が多用されるため，前文中では助詞等が挿入されているのに加え，一般に，長い固有名詞などは略称が用いられる[2]．これらを考慮し，本稿では，文字列間の近似的照合方法の一つである最長共通部分列（LCS）を利用する．LCSの計算は，二つの記号列を $A = a_1a_2\cdots a_m$ と $B = b_1b_2\cdots b_n$ とすると，LCSの長さは $p_{i,j}$ を A の先頭の i 個の部分列 $a_1a_2\cdots a_i$，Bの先頭から j個からなる部分列 $b_1b_2\cdots b_j$とすると，以下の漸化式（1）が成り立つ．ここで，$1 \leq i \leq m, 1 \leq j \leq n$ であり $p_{i,0} = p_{0,j} = 0$ とする．すなわち，AとBとのLCSの長さは $p_{m,m}$の値となる．一方，LCSの長さを求める計算式の結果は二次元の行列で表現できるため，図2に示すように，行列の右下隅（p[m][n]）から漸化式を利用して，左上へ逆戻りする形でLCSを求めることができる．

$$p_{i,j} = \begin{cases} p_{i-1,j-1} + 1, & (a_i = b_j) \\ \max\{p_{i-1,j}, p_{i,j-1}\}, & (a_i \neq b_j) \end{cases} \quad (1)$$

```
1 for(i=1;i<=m;i++) lcs[i]=0;
2  while(i>=1 && j>=1) {
3    if (a[i] == a[j]){
4        lcs[i]=1; i--; j--;
5     }
6    else{
7     if(p[i][j] == p[i][j-1])j--;
8     else i--;
9      }
10  }
```

図 2: LCS を求めるアルゴリズム例 (C 言語風)

本研究では，日本語の文字の字種の違いと見出し文の文字の特徴を利用し，見出し文と前文の文字が一致した場合の条件分岐を（1）式に付加して計算する．まず，(1) 双方の字種が漢字の場合は，そのまま計算し長さを 1 加算する．次に，(2) 漢字以外の場合は，直前の文字同士が既に一致している場合もしくは，一文字分だけ先読みし，直後の文字同士が一致できる場合は長さを 1 加算する．もし，前後とも一致しない場合は，その位置での一致は無かったものとみなす．また (3) 見出し文中に出現する句点は照合しないものとする．(4) 見出し文の文末の平仮名文字も照合しない．これらの制約条件は，漢字以外の文字は，見出し文と前文の双方で連続した文字として出現する．助詞と非助詞との平仮名文字のミスマッチを軽減させるなど，不適切なマッチングを減らす方向に働くものと考える．

[2] 例えば,「科学技術庁」は「科技庁」に.

最適照合位置の獲得

見出し文の文字列 $A = a_1a_2\cdots a_n$ を分離記号（=）の位置 a_i で分割し，部分列 $A_1 = a_1a_2\cdots a_{i-1}$ と $A_2 = a_{i+1}a_{i+2}\cdots a_n$ を得る．また，前文の文字列 $B = b_1b_2\cdots b_m$ とする．なお，B は一般に複数の文で構成され，各文は読点記号（。）で識別可能とする．以下の手順で照合された文字の位置情報を所定の配列に格納する．

(1)$LCS(B, A_1)$ と $LCS(B, A_2)$ を別々に計算し，長さ m の配列 P_1,P_2 に格納する．

(2)P_1 と P_2 をマージしながらソートする．

図 1の例に対する，照合結果を行列で表したものを図3と図4に示す．なお，簡便のため，前文の文字列を行要素とし，見出し文の文字列は列要素として配置してあり，行要素のうち，列要素と一致する部分のみの共通部分列の行列成分を示している．また，図中の行要素には，見出し文字表記，文中での位置情報，前文中での文番号の順番で示した．

lcs = 9		人	工	視	覚	装	置	の	実	用	に	前	進
人	76 1	1											
工	77 1		2										
視	80 1			3									
覚	81 1				4								
装	87 1					5							
置	88 1						6						
の	89 1							7					
実	90 1								8				
用	91 1									9			
実	101 1								8				
実	109 2								1				
用	110 2									2			
人	142 2	1											

図 3: 図 1 に示した例文の LCS 計算過程（1）

3.2 文の生成方法

前節で述べた照合方法により，前文中での見出し文に対応する文字の位置を示す数字列を $T = t_1, t_2, \cdots, t_l$, 前文の文字列を $B = b_1b_2\cdots b_n$ とすると，隣接する t_i, t_{i+1} は $t_i < t_{i+1}$ であり，$t_i + 1 \leq t_{i+1}$ の整列関係にある．文生成処理は，基本的に，隣接する t_i と t_{i+1} の整列関係に基づき，連続している場合は対応する前文の文字を出力し，離れている場合には，以下に示す制約規則により前文の文字を出力し，出力文を生成する．処理過程は，t_i の字種に基づき前方・後方に位置する前文の文字

lcs = 8		米	大	が	動	物	実	験	開	始	へ
米	4 1	1									
米	20 1	1									
大	30 1		2								
実	90 1				3						
動	99 1				3						
物	100 1					4					
実	101 1						5				
験	102 1							6			
開	104 1								7		
始	105 1									8	
実	109 2						1				
米	133 2	1									

図 4: 図 1 に示した例文の LCS 計算過程（2）

を出力するか否かを判断しながら，確定した文字を出力していく機構である．

前方処理 現在注目している t_i の字種を調べ，漢字あるいは英数字なら，異なる字種が現れる位置を前方に探索し，その位置が b_k なら，$b_{k+1}\cdots t_i$ に対応する文字列を出力する．もし t_{i-1} が存在すれば，その直後までとする．以下に処理例を示す

見出し文	前文	出力
期間終了	～の契約期間が	契約期間
９９年物価	１９９９年の消	１９９９年
科技庁	科学技術庁	科学技術庁

後方処理 原則として，t_{i+1} あるいは句読点が現れるまでを出力するが，t_{i+1} が存在する場合は，形式名詞「こと」の直前まで，あるいは，助詞「は」に相当する文字までの出力とする[3]．(以後，本稿では**出力抑制規則**と呼ぶことにする)．最後の要素である t_l の場合は，形式名詞「こと」だけに留意する．

図 5に，上述の前方・後方処理により生成された実験プログラムの生成例を示す．図中で，丸括弧“(”で囲まれた文字は見出し文と前文との一致文字を表し，中括弧“{”で囲まれた文字は出力抑制規則の適用による文字を表し，その他の括弧は後方処理による文字を表す．また，この例では，前方処理は適用されなかったことが分かる．

4 実験と評価

本提案手法の有効性を確認するため，インターネット上で配信されている共同通信社のニュース速報文の見出し文と前文記事を入力対象とする実

[3]文字列「とは」には適用しない．

人工視覚装置の実用に前進＝米大が動物実験開始へ

(米)<ジ><ョ><ン><ズ><ホ><プ><キ><ン><ス>(大)<の><研><究><チ><ー><ム>{は}(人)(工)<的><に>(視)(覚)<を><生><み><出><す>(装)(置)(の)(実)(用)<化><に><向><け>{、}(動)(物)(実)(験)<を>(開)(始)<す><る>。

図 5: 図 1 の例文を入力とした実験プログラムの出力結果

験を行なった．実験対象データは 2000 年 1 月 11 日から 18 日までの 50 記事とした．言い換えの評価は，(佐藤 1999) が論じているように，客観的な評価が難しい．本稿で提案した手法は，見出し文と前文との文字列照合が正しく行なわれれば，生成処理を全く適用しない場合でも，照合文字列を含む文を直接出力することで，見出し文は言い換えられたこととみなすこともできる．そのため，評価基準としては，一記事あたり，最適な照合が行なわれたか否かの評価と生成処理は出力抑制規則により削除された結果，非文とならず，かつ平易な表現に変換されたか否かで評価を行なった．この評価は主観で行なった．表 1に実験結果を示す．

表 1: 実験結果

最適な照合	誤った照合	照合精度
44	6	88%
最適文	非文	生成精度
40	10	80%

照合の誤りの最も多い原因は，見出し文と前文間で文節の交換がある場合である．例えば，図 6 見出し文が「欧州でも米国流が過熱」に対して，前文中では「米国流の敵対的買収が欧州で過熱」の場合，図 2のアルゴリズムは LCS の長さが 5 であるため,「米国流」が抽出されない，一つの解決策としては，見出し文中の抽出できなかった部分列のみを入力として，再計算することで解決できる．なお，照合に洩れがあった場合でも，一部は生成処理で補完される．生成精度の評価が低い原因は，見出しが「雌雄の産み分けが容易」に対して,「雌雄の判別が容易」のように，逆に難解な表現に置換した 2 件と複合語あるいは体言止め表現のままの例が 5 件，非文と判定されたものが 3 件であった．このことから，見出し文中の言い換えるべき表現と前文中の表現が完全一致の場合は，本手法では，言い換えができない．すなわち，置換した対応関係を知識として蓄積，利用できれば，当該の前文記事中には存在しない場合もこの置換表を探索することで，例えば,「事件が続発」が「事件が相次いでいる」に変換可能となる．これらは今後の課題とする．

lcs = 5		欧	州	で	も	米	国	流	が	過	熱
米	9 1					1					
国	10 1						2				
流	11 1							3			
欧	19 1	1									
州	20 1		2								
で	21 1			3							
過	22 1									4	
熱	23 1										5
米	79 1					4					

図 6: 複数の LCS が存在する場合

5 おわりに

本稿では，附属語の省略による難解な複合語や体言止めなどの表現が多用される傾向がみられる，ニュース速報記事の見出し文を前文記事との文字列照合により，省略された語句を復元することで言い換える方法を提案した．実験の結果，有効性を示す妥当な結果が得られた．一方，実用レベルにするためには，照合法の改良と言い換え表の自動獲得など残された課題は少なくない．しかし，本格的な言語処理機構を用いない，簡便かつ有用な実用システムの実現性を例示できたと考える．

参考文献

安達久博 (1999). “手話通訳のためのニュース文の話しコトバへの変換処理.” 技報, 電子情報通信学会.

乾健太郎, 山本聡美, 野上優, 藤田篤, 乾裕子 (1999). “聾者向け文章読解支援における構文的言い換えの効果について.” 技報, 電子情報通信学会.

小笠原信之 (1991). 実例で学ぶ日本語新聞の読み方. 専門教育出版.

佐藤理史 (1999). “論文表題を言い換える.” 情報処理学会論文誌, **40** (7), 2937–2945.

山本和英, 増山繁, 内藤昭三 (1995). “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, **2** (1), 39–55.

山本和英, 増山繁, 内藤昭三 (1996). “関連テキストを利用した重複表現削除による要約.” 電子情報通信学会論文誌, **J79-D-II** (11), 1968–1972.