Web と携帯端末向けの新聞記事の対応コーパスからの文末言い換え抽出

岩越　守孝[†]　増田　英孝[†]　中川　裕志[††]

本研究では，数十文字程度の長さで携帯端末向けに配信されている新聞記事と数百文字程度の長さの Web 新聞記事の両者を約 3 年に渡って収集した．こうして収集したコーパスから文末表現の縮約などの言い換え表現の抽出を機械的に行った．まず，Web から収集した携帯向け新聞記事と Web 新聞記事からなるコーパスに対して記事単位の対応付けを行い，次に文単位の対応付けを行った．次に携帯向け記事文の文末の表現を形態素解析を用いて抽出し，その文に対応する Web 新聞記事の文を集める．そして Web 新聞記事の文の文末から形態素ごとに言い換え先表現を抽出し，それに対して頻度等を用いた得点付け，および必要な名詞を欠落させてしまう不適切な言い換えの除去を行うことにより言い換え表現の抽出精度向上を図った．

キーワード:　言い換え，携帯端末，*Web*，文末表現

Extraction of Paraphrasing Pattern by Aligned Corpora of Web and Mobile Terminal News Articles

Moritaka Iwakoshi[†] , Hidetaka Masuda[†] and Hiroshi Nakagawa[††]

We have collected both Web news-paper articles of several hundreds of characters, for three years and their counter parts distributed for mobile terminals, which consist of fifty to a hundred characters. Then, we extracted a number of candidates of paraphrases of the final part of sentences from them automatically. At first we have aligned these two types of corpus first at article level, then at sentence level. Next, we extract the final part of mobile article sentences using morphological analyzer, and collect their counterpart expressions of Web article sentences. Finally, we extracted the candidates of morpheme sequence from the final part of Web article sentence, then we propose the combination of two methods for them in order to improve the extraction accuracy of the sets: 1) ranking based on frequency, branching factor and length of string, and 2) filtering to remove inappropriate expressions which eliminate semantically indispensable nouns.

KeyWords:　*Paraphrase, Mobile terminal, Web, Sentence final part*

† 東京電機大学工学部, School of Engineering, Tokyo Denki University
†† 東京大学情報基盤センター, Information Technology Center, The University of Tokyo

1　はじめに

最近，種々の応用を睨んで言い換えの研究がさかんになっている (乾 2002; Inui and Hermjakob 2003)．例えば，語彙的言い換えの研究 (Yamamoto 2002) は種々の応用に役立つ．また，機械翻訳の前処理や評価 (Kanayama 2003)，情報検索，質問応答，情報抽出の柔軟性を上げること (Rinaldi, Dowdall, Kaljurand, Hess, and Molla 2003; Shinyama and Sekine 2003)，年少者や初心者向けの教科書やマニュアルを読みやすくする，などは直接的に役立つ応用である．似た研究としては聾唖者に理解し易いテキスト言い換えもある (Inui, Fujita, Takahashi, Iida, and Iwakura 2003)．また，非母国語話者が理解しやすいように簡易な言い方に言い換えることも有意義である．こういった目的のためには，国語辞典を用いた用言の言い換え (鍛冶，川原，黒橋，佐藤 2003) や普通名詞の言い換え (藤田，乾，乾 2000) などが役立つ．

一方，要約も言い換えの応用分野として有力である．従来の文書要約は重要文の抽出が主体であった (Mani 2001)．しかし，抽出した文をさらに短縮することを目指す場合には言い換えが役立つ．例えば，

例文 1:　本法案が衆議院本会議で審議が始まった。

を

例文 2:　本法案、衆議院本会議で審議。

というような言い換えが考えられる．実際にこの例文 2 のような短縮された表現はテレビの字幕あるいは列車の字幕ニュースなどでよく見かける．このような応用は文書表示を行う端末の多様化からみても有用さが増してくる．Web ページは従来からパソコンの大画面への表示を想定して作られていた．しかし，携帯電話や PDA の普及により 100 文字程度の小画面への表示を念頭におくテキストも増加している．このような画面へ表示するコンパクトなテキストは多くの場合短縮された表現である．このような短縮を自動的に行うために言い換え表現を収集することは意義深い．

新聞記事の場合，重要な文は記事の先頭に現れることが多いという性質を利用して抽出できるが，画面が小さく表示文字数に限りがあること，短い時間で読むことができることなどを考慮すると，さらに縮約が要請される．後に詳しく述べるが，よく使われるのは，上記の例文 2 に見られる体言止めのような文末の短縮表現である．また,「国会で審議へ」という文末の助詞止めも多く使われる．このような縮約した文末表現は従来から字幕放送で用いられている．しかし，通常の書き言葉の文末である終止形を体言止めや助詞止めに変換する規則は，これまでほとんど手作りであった (安藤，今井 2001)．

このような文短縮を目的とした言い換え表現を言語の実際の使用例から自動収集するための言語資源として Web に配信されている新聞記事と，これに対応した内容を携帯電話向けに発信している新聞記事に注目する．これらは毎日数十記事発信され，長期間にわたって蓄積すれば大量の言語資源となる．すなわち，同じ内容が数十文字程度で構成された携帯端末向けの新聞

記事と数百文字程度で構成されている Web 新聞記事が対応付けられれば，ある言語表現とその短縮表現の対応データとして使える．この対応付けコーパスを用いれば，多様な文末表現の縮約のための言い換え表現を機械的な手法で抽出することが可能になる．ここで留意しなければならないのは，この研究で目的としている言い換えは「Web 記事の文 → 携帯端末向け記事の文」という方向性を持つ点である．実際には，書き手がこの方向で作業しているかどうかは不明である．しかし，縮約のような言い換えによって短縮された記事を作ることは技術的に可能であっても，その逆方向の言い換えは困難である．よって，この方向性を前提として研究を進める．なお，以下では必要に応じて，言い換え操作の対象になる Web 記事の文からの抽出表現を「言い換え元表現」，対応する携帯端末向け記事の文からの抽出表現を「言い換え先表現」と呼ぶ．

さて (乾 2002) は言い換えの研究にいくつかの問題を提起している．それらに対して，この研究ではいかなる解決策を採っているかをまとめることによって，本論文の構成を述べる．

言い換え事例をどのように集めるか

この問題に対しては，1)Web 上から得られる言い換え表現獲得のための言語資源として Web 新聞記事と携帯端末向けの新聞記事を用いること，2) この両記事コーパスを文単位で対応付ける方法の提案と実験的評価，を行って対処している．具体的には 2 節において，研究で使用した記事データについて，および Web 記事と携帯記事の対応付け，さらにそこから文単位での対応付けを行う方法について述べる．このような対応付けコーパスを用いる言い換え事例収集は多くの研究 (Brazilay and McKeown 2001; 関根 2001) があるが，本研究での新規性のひとつは対象としている言語資源にある．

どの表現を言い換えるか

この問題は，これまでの言い換え研究の中心課題のひとつであった．特に類似した表現の対をコーパスから探し出すことは重要なテーマで，多くの研究 (Murata and Isahara 2001; Torisawa 2001; Terada and Tokunaga 2001) がなされた．我々の場合，3 節において述べるように，対応付けられた文からなるコーパスを利用して Web 記事文の文末を縮約する携帯端末向け文の文末の言い換え表現を獲得することに的を絞っている．よって，言い換えるべき場所は Web 記事文の文末のうち，本論文で述べる方法で抽出した言い換えにおける言い換え元の表現が出現した場合と限定できる．

可能な言い換えの網羅的生成と，生成された候補の評価

(乾 2002) では，この問題は上の問題の一部と位置付けられているが，本研究では網羅性の確保はその困難さから諦めた．代わりに文末表現に限定し，どのような範囲の形態素列を切り出せ

ば正しい言い換え表現を抽出できるかという問題に絞って扱う．3.3 節で言い換え表現の抽出について説明し，その抽出結果に 3.4 節で説明する得点付けを行うことによって正しい言い換え表現を取得する．3.5 節では，その結果の言い換え表現のうち必要な名詞を削りすぎた不適切な言い換えを除去するフィルタリングについて述べる．これらの 3 節に提案する手法の実験評価を 4 節で述べる．

意味の差，およびその計算法

この問題はこの論文では人手での評価に頼った．今後の課題である．

言い換え知識の共有

本論文で述べた言い換え知識は文末表現の縮約に役立つが，これを大きくの研究者，技術者に共有する枠組みについても今後の課題である．

2　対象とする新聞記事データとその対応付け

2.1　対応付けの概要

文縮約のための言い換え規則を機械的に取り出すためには同一内容の長短 2 文が大量に必要となる．そこで本研究では，(大森，増田，中川 2003) の手法を利用して Web から長期にわたって収集したコーパスを用いる．このコーパスは，インターネット上に配信されていて，パソコンでの閲覧用に作成されている新聞記事 (以下，Web 記事と呼ぶ) と携帯端末向けに作成されている新聞記事 (以下，携帯記事と呼ぶ) の間で同じ内容のものを自動的に対応付けたものである．さらに言い換え表現抽出のために，その携帯記事中の文 (以下，携帯文と呼ぶ) に対しそれに対応付けられた新聞記事中から同一内容を持った文 (以下，Web 文と呼ぶ) を対応付ける (佐藤，岩越，増田，中川 2004)．

本節以下の実験では 2001 年 4 月 26 日から 2003 年 11 月 30 日までに収集した Web 記事と携帯記事から得た 48075 組の記事から抽出した合計 72203 組の対応文を用いた．Web 記事は通常，数百文字で構成されているのに対して，携帯記事は 50 文字程度で構成されている．また携帯記事は体言止めや文末が助詞で終わる文が多いのが特徴として挙げられる．携帯記事の文末品詞の割合を表 1 に示す．

2.2　記事単位での対応付け

新聞記事の対応付けの方法は以下のようになる．

収集した記事群で 1 日単位に Web 記事と携帯記事の対応付けを行う．まず，両記事群を「茶筅」(松本，北内，平野，松田 2002) で形態素解析する．この結果に対して，携帯記事 i 中の

表 1　携帯記事の文末品詞の割合

品詞		頻度 [個]	頻度 /合計 [%]
名詞	サ変接続	28687	39.7
	その他	11796	16.3
助詞		13397	18.6
動詞		11988	16.6
助動詞		5589	7.7
その他		746	1.1
合計		72203	100.00

名詞と Web 記事 j 中の名詞を調べ，次式のようにこの両者の記事の類似度 $Sim(i,j)$ を計算する．ここで $M(i)$ は携帯記事 i 中の名詞の集合，$Wt(j)$ を Web 記事 j の見出し中の名詞の集合，$Wb(j)$ を Web 記事の本文中の名詞集合とする．なお，携帯記事には見出しは付いていない．

$$Sim(i,j) = 3 \times |Wt(j) \cap M(i)| + |Wb(j) \cap M(i)| \tag{1}$$

2001 年 5 月 10 日から同 8 月 10 日まで毎月 10 日と 20 日に収集した 605 記事について Sim の値と人手でつけた対応付けの正解率の関係を図 1 に示した．

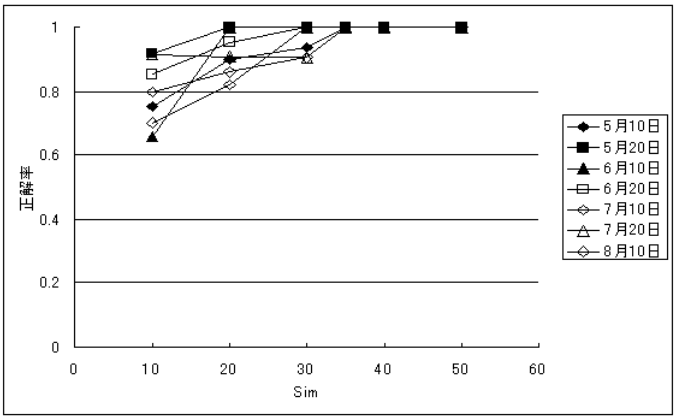

図 1　Sim の値と記事単位の対応付けの精度

この図より，Sim の値が 35 以上の場合を正しく対応が付いたとすることにより，現在までの実験で 481 記事が正しく対応した．すなわち，約 80 %の携帯記事を 100 %の精度で対応付けができた．そこで，この方法，すなわち $Sim \geqq 35$ の条件を満たす記事対を取り出すことにより，約 3 年分の記事対応付けコーパスを作成した．

2.3 文単位での対応付け

記事単位で対応付けられたコーパスにおいて携帯記事を基準として Web 記事から対応文の抽出を行う．これも以下に示すように対応した記事対において共起した名詞の頻度によって行った．具体的アルゴリズムを以下に示す．

文対応付けアルゴリズム

Step:1　$i = 1$

携帯記事の第 i 文を形態素解析し，第 i 文に含まれる全名詞を抽出し，この集合を $Ms(i)$ とする．

Step:2　$j = 1$

Step:3　Web 記事の第 j 文を形態素解析し，第 j 文に含まれる全名詞を抽出し，この集合を $Ws(j)$ とする．

$S(j) = |Ms(i) \cap Ws(j)|$

を求める．

Step:4　$j = j + 1$

Web 記事の最後の文になるまで Step:3，Step:4 を繰り返す．

Step:5　$S(j)$ がもっとも高い Web 記事の文を携帯記事第 i 文の対応文とする．なお，一致した名詞の数が同数の文が複数あった場合は，記事の先頭に近いものを対応文として選ぶ．

Step:6　$i = i + 1$

携帯記事に残った文があれば Step:1 に戻る

Step:5 で名詞一致数が同数の場合に記事先頭に近いものを選ぶのは，新聞記事の場合は先頭に近い文が重要な情報を担うからである．つまり，携帯文に対応する文のうち，より重要な情報を含む文を選択しようという指針を採った．以上の方法で抽出した対応文対のうち，以後，携帯記事から抽出した文を携帯文，Web 記事から抽出した文を Web 文と呼ぶ．

ここまでに述べた方法で抽出したデータのうち，2001 年の約一年分の対応記事コーパスの 43171 組の対応文についての詳細を表 2 に示す．携帯記事のほとんどが二文で構成されていることから，使用した対応付けコーパスの記事数の約二倍の対応文が抽出される．また携帯文は一文が数十文字程度であるのに対して，Web 文はそれよりも長い構成になっているので，二文の携帯文に対して，Web 文が一文で抽出される場合もある．これは携帯記事が二文で構成されているときのみ現れる．

次に今回，抽出した対応文からランダムで 500 組を抽出し対応付けの精度を求めた．

表 2　記事の構成文数と対応文の抽出状況

携帯記事の構成	抽出した時の状態 (携帯文対 Web 文)	抽出した対応文
1 文	1 対 1	1801
2 文	2 対 1	9606
	1 対 1	29732
3 文	1 対 1	2028
4 文	1 対 1	4
合計		43171

$$\text{精度} = \frac{\text{抽出した対応付け正解文数}}{\text{抽出した全文数}} \tag{2}$$

対応付けの正解・不正解は人手で行っており，対応付けられている場合は正解，対応付けられていない場合は不正解として 2 人で行った．2 人の判断が異なった場合には 3 人目が判断して，多数決で正解・不正解を判断する．この方法により評価を行った精度は 92.8 %であった．この精度は対応付けそのものとしては必ずしも十分ではないが，次節で述べる言い換え表現抽出では，さらに頻度などに基づく言い換え表現の重み付けなども行っているため，100 %の精度は必須とは言えない．よって，この方法によって得られた対応文のデータによって，言い換え元表現と言い換え先表現抽出の実験を進めることにした．

3　言い換え表現の抽出

3.1　言い換え抽出の枠組

本節では，2.3 節で述べた方法で抽出した携帯文と Web 文を用いて，言い換え先表現と言い換え元表現の対を抽出する方法について述べる．例えば，

携帯文:　コンピュータウィルス感染防止に有力な方法が 判明。
Web 文:　コンピュータウィルス感染防止に有力な方法があることが、研究所の調査で 分かった。

という対応文があったとする．このとき文末に注目すると携帯文では 判明 で終わっているのに対して，Web 文では 分かった で文が終わっている．文の内容から要約の際は 判明 を言い

換え先表現に，分かった を言い換え元表現に使えることが，人間が見れば容易に判断できる．このような携帯文の文末にある言い換え先表現に対する言い換え元表現を Web 文から自動的に抽出するのは，概略，以下のような方法になる．

Step:1　2.3 節で作成された対応文から同じ言い換え先表現を文末に持つ携帯文と Web 文の関連付けする．この詳細は 3.2 節で述べる．

Step:2　Step:1 で関連付けられた Web 文から言い換え元表現の候補を抽出する．この詳細は 3.3 節で述べる．

Step:3　抽出された言い換え先表現の候補それぞれに対し，語彙の分岐数，出現頻度，文字列長から正しい言い換え元表現が高得点になるような得点付けを行い，順位付けする．この詳細は 3.4 節で述べる．

Step:4　Step:3 の結果に対して，精度の向上を図るため，言い換え元表現として不適切な表現を削除する．この詳細は 3.5 節で述べる．

すなわち，この処理では，**言い換え先である携帯文文末の表現をまず決め，それに対応する複数の Web 文の言い換え元表現を推定する** という問題を解くことになる．

3.2　言い換え先表現および対応する Web 文集合の抽出

まず言い換え先表現の抽出方法について説明する．言い換え先表現の抽出のために携帯文を形態素解析し，文末にある 1 形態素を取り出す．これによりサ変接続の名詞であれば「会談」や「表明」といった表現が抽出される．しかし，これだけでは助詞や助動詞，動詞の場合は「も」や「た」,「示す」といった言い換え表現として使用が難しい表現や，意味の範囲が広すぎるために言い換え表現の抽出が困難な表現が抽出されてしまう．その問題はさらに文頭方向にある形態素を続けて抽出することにより解消できる．その結果,「可能性も」や「述べた」,「認識示す」といった言い換え先表現が取り出され，言い換え元表現の抽出も容易になる．

次に図 2 に抽出した言い換え先表現に対する対応文集合を作成する流れを示す．まず図の左側の枠内にあるように，抽出した言い換え先表現に対しその言い換え先表現を文末に持つ携帯文を集める．そして図の右側の枠内にあるように，集めた携帯文に対応する Web 文を集め，それを Web 文集合とする．そのときの Web 文集合の要素数を以下「対応文数」と呼ぶ．3.3 節以降で説明する言い換え元表現の抽出は，ここで作成された Web 文集合の文末から抽出することになる．

この方法により，言い換え先表現として 4617 表現を抽出した．抽出した言い換え先表現のうち頻度が上位 30 位までの表現を表 3 に示す．動詞終止形，助詞，形容詞語幹 (「高」「安」)，など様々だが，一番多いのはサ変接続名詞であり 60 %を占める 18 個である．文末のサ変名詞

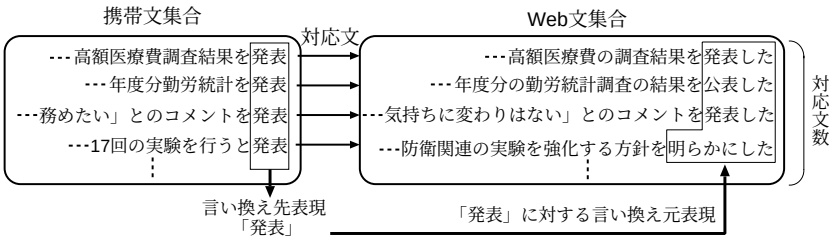

図 2　言い換え先表現の抽出と対応文集合

はいわゆる体言止めであり，この表にも示されるように頻度が高く，結果として適用される頻度も高いと推測される．

表 3　言い換え先表現の例

抽出表現	対応文数	抽出表現	対応文数	抽出表現	対応文数
高	1780	ている	505	協議	315
安	1668	判明	422	可能性も	310
発表	1118	方針	408	要請	292
逮捕	967	見通し	402	れた	286
」と	933	ため	399	発言	282
会談	788	強調	378	指摘	270
表明	735	合意	341	」	264
死亡	629	開始	329	確認	261
決定	538	検討	328	予定	254
いた	513	批判	317	みられる	247

ここで頻度が高かった「高」と「安」であるが，これはほぼ全てが経済の記事からであり，「円高」「円安」などが元となっている．さらに対応する Web 文では「円高・ドル安となった」「円安・ドル高となった」という表現が固定的に用いられている．その特殊性から正解となる言い換え元表現が少なく，例を挙げての説明が困難となるため，以下の説明の際は頻度が次に多い「発表」を用いる．

3.3　言い換え元表現候補の抽出

3.2 節で作成されたデータを用い，言い換え先表現に対応する言い換え元表現を Web 文集合の各文の文末から抽出する．言い換え元表現の抽出には Web 文を形態素解析し，文末から文頭方向に向かって 1 形態素ずつ増やしながら表現を抽出する．言い換え元表現が含まれる長さとして十分な 15 形態素までを使用する．ここでは単純に形態素区切りで表現を取り出すため，言

い換え元表現に適さない表現も数多く抽出されるが，3.4 節で述べる得点付けや 3.5 節で述べるフィルタリングによって排除を試みる．この時点で抽出された言い換え表現の例を表 4 に示す．

表 4　言い換え先表現が「発表」時の言い換え元表現の候補の例

た	れた
した	された
発表した	指定された
を発表した	が指定された
結果を発表した	公表された
調査結果を発表した	から公表された
の調査結果を発表した	発表された
費の調査結果を発表した	日発表された
医療費の調査結果を発表した	が発表された

3.4　分岐数，頻度，文字列長に基づく言い換え元候補の順位付け

3.3 節で抽出された言い換え元表現には言い換えとして適切な表現と言い換えとして不適切な表現が含まれていることになる．そこで，抽出された表現に対して正解が上位に集中することを目的とした順位付けを行う．

順位付けを行うにあたってまず，ある言い換え先表現に対応する Web 文集合において，集合全体として Web 文の文末が持つ特徴を説明する．図 3 に Web 文集合中の Web 文の文末から文頭方向への語の分岐の様子の例を示す．図から前方に向かって形態素が分岐していることが分かる．まず，一番右側に Web 文の一番文末の形態素となる「た」や「する」がくる．さらに 1 つ前方にある形態素を繋げると「した」や「だった」や「発表する」が抽出できる．例えば，「を発表した」に続く形態素は「結果」「表明」「コメント」など 111 種類ある．ここで，ある表現から 1 つ前方の形態素の種類数をその形態素の分岐数と呼ぶ．さらに図 4 に図 3 で示した内容の一部分の分岐数と頻度の関係を示した．このグラフからは「発表した」までは分岐数が小さく,「を発表した」で分岐数が大きく，さらに「結果を発表した」となるとまた小さくなることが分かる．これは，(a) 固定された表現の内部では部分形態素列を長く与えれば与えるほど，直前あるいは直後の形態素が絞り込まれること，(b) ひとたび固定的な表現が終わると，その前後にはいかなる表現も現れることができるようになること，に対応している．よって分岐数が大きい形態素までの形態素列がよい言い換え元の候補であると考えることができる．さらに良く使われる表現ほどその表現は固定的な言い回しで，言い換え先表現と深くかかわっている可能性が高いと考えられる．

以上の特徴を踏まえ，言い換えのよさを示す評価関数の構成要素として以下を用いる．

分岐数:　分岐数の大小が言い換え表現句としての切れ目の可能性の大小を表すと考えられる

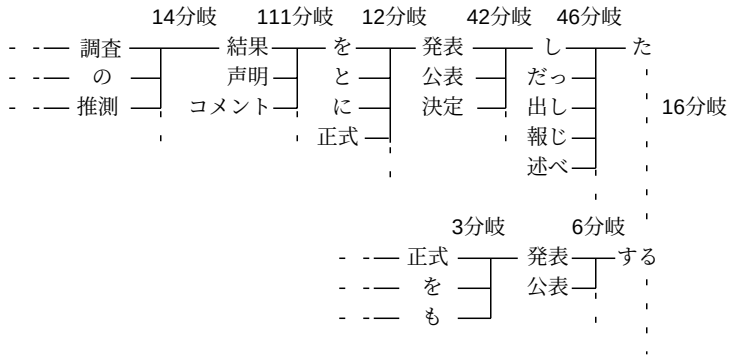

図 3　言い換え先表現が「発表」時の Web 文を文末からみた様子

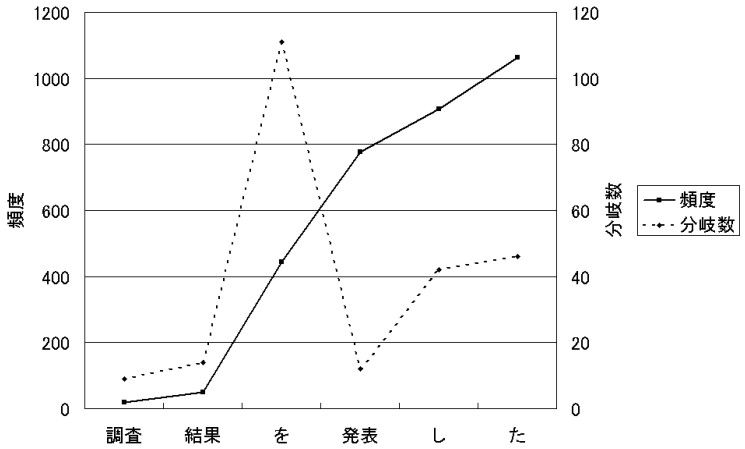

図 4　言い換え先表現が「発表」時の Web 文を文末からみた分岐数と頻度の関係

ので評価関数の構成要素として用いる．

頻度:　良く使われる表現は安定していることを示すので，他の要因と組み合わせて用いることは有益である．

文字列長:　ここでいう文字列長は形態素数ではなく文字数である．言い換え元表現は短過ぎるならば言い回しにならず，長過ぎるならば文脈に依存した表現になってしまう．長過ぎもせず，短すぎもせず，適度な長さの表現を抽出したい．そのため評価関数では log(文字列長 − 1) を用いる．log により長い文字列に対して得点の抑制を，文字列長 − 1 により 1〜2 文字の表現の排除をする効果がある．さて，長さに形態素数を使うという選択肢もある．しかし，もし形態素数を文字列長の代わりに使うと，十分に長くて意味の

ある形態素 (例えば固有表現) が長さ=1 で排除されてしまいかねない．これを避けるために文字数を用いた．

上記の各要素を

$a =$ 分岐数

$b =$ 頻度

$c = \log($ 文字列長 $-1)$

と定義し，評価関数を $a, b, c, a \times b, a \times c, b \times c, a \times b \times c$ の 7 種類を用いて比較実験を行った．対応文数の多かった 100 位までの言い換え元表現の候補のスコアが 1 位の表現を人手で評価し，評価関数の違いによる正しい表現の割合を表 5 に示す．この結果から，評価関数 $a \times b \times c$ を用いた方法が最も精度が高いことがわかる．よって，評価関数 $a \times b \times c$ を用いた得点付けのデータを用いる．

表 5　計算手法の違いによる精度の違い

評価関数	a	b	c	$a \times b$	$a \times c$	$b \times c$	$a \times b \times c$
正解の割合	18 %	5 %	0 %	12 %	46 %	65 %	71 %

さらに言い換え元表現として正しい表現が得点が高くなり，高順位になることを示すために，図 5 に言い換え先が「発表」の場合に $a \times b \times c$ の方法で得点付けをした場合の言い換え元の正しい表現の分布グラフを示す．このグラフから，多くの正しい言い換え元表現が高順位に集中していることがわかる．低い順位にいくつか正しい言い換え元表現がきているが，これは表現の頻度が少なかったために得点が低くなったことが原因である．

3.5　フィルタリングによる不適切な言い換え元表現の削除

言い換え元表現として抽出した表現の中には文の意味として欠落してはならない語を含んでいることがあるため，その語を言い換えによって削除してしまうと意味が通らない文になってしまう可能性がある．このような不適切な言い換え元表現は前節の得点付けによって順位が下位になる場合は採用されないが，収集した記事中でよく使われる表現であれば言い換え元表現として不適切な表現も上位になってしまう．そのような言い換え元表現を削除するためのフィルタリングを行う．そのアルゴリズムは次のようになる．

フィルタリングアルゴリズム

n を言い換え元表現の数とし，言い換え元表現の集合を $\{x_1, x_2, \cdots, x_n\}$，言い換え元表現を

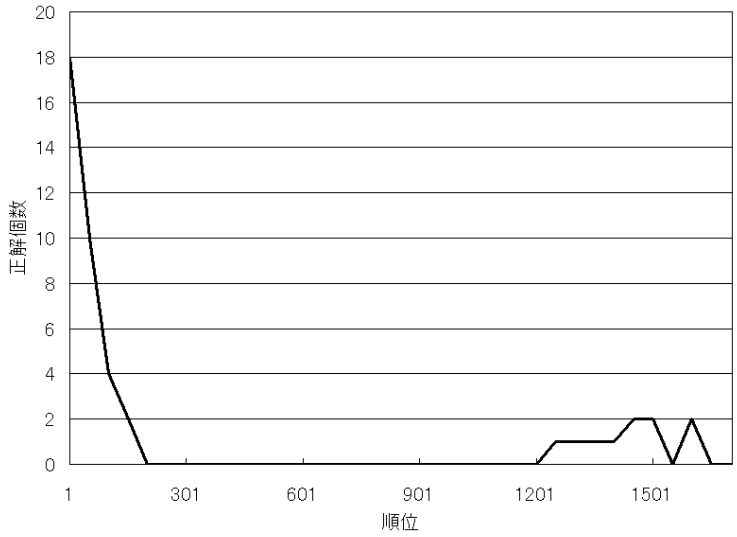

図 5　言い換え先表現が「発表」の場合の正しい表現の分布

$x_i = S_1S_2\cdots S_m$ (S_k は形態素)，言い換え先表現を y，携帯文を $M_1\cdots M_jy$ (M_l は形態素, y は言い換え先表現)，C を名詞とすると，

for($i = 1$, n) {
 if ($C \in \{M_1,\cdots,M_j\} \wedge C \in \{S_1,\cdots,S_m\} \wedge C \notin y$)
 then x_i を言い換え元表現集合から除く
}

なぜなら名詞 C は携帯文に含まれるが，言い換え先表現 y には含まれない．つまり C は携帯文にとって必須の意味を担う．よって C を含む言い換え元表現 x_i を C を含まない言い換え先表現 y に省略することはできない．

具体例を図 6 に示す．ここで C は「声明」となり，削除対象となる言い換え元表現 x_i は「声明を発表した」となる．なお，y は「発表」である．Web 文にも携帯文にも「声明」という語が含まれ，文の内容として必須の語であることがわかる．よって「声明」という意味を削除する「声明を発表した」は「発表」の言い換え元表現として正しくないと考えられるため，言い換え元表現の候補から削除する．

フィルタリングによって 3.4 節で得られたデータがどのように変化するかを表 6 に示す．表中でアンダーラインが引かれているものがフィルタリングによって削除された表現である．こ

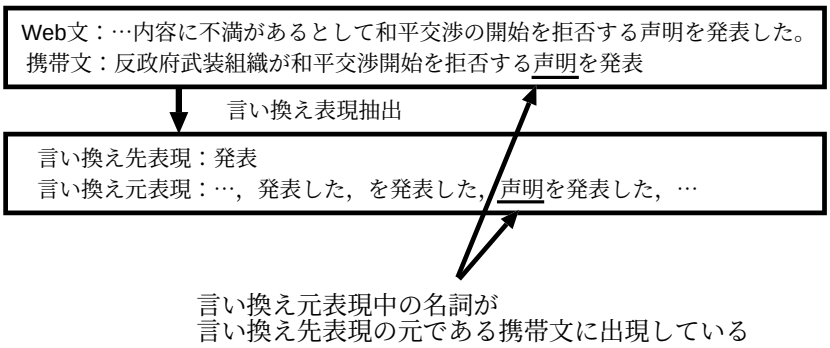

図 6　フィルタリング処理の具体例

の表から，言い換え元表現として用いるには不適切な表現が削除できていることが分かる．

表 6　フィルタリングによる言い換え元表現の削除の例
(言い換え先：「発表」)

を発表した	を明らかにした
と発表した	ことを明らかにした
発表した	たことを明らかにした
すると発表した	調査結果を発表した
たと発表した	明らかにした
したと発表した	策を発表した
結果を発表した	となった
声明を発表した	したことを明らかにした
計画を発表した	する声明を発表した

4　抽出された言い換え元表現の評価

提案した手法で抽出した言い換えの全体に対しての数量的評価を 4.1 節で述べる．4.2 節では抽出した言い換えの典型例についての考察を行う．

4.1　数量的評価

まず言い換え元表現と言い換え先表現の長さについて述べる．3.2 節で述べた方法で抽出した言い換え先表現 4617 個を対象にしたときの言い換え先表現の平均文字数は 2.6 文字，標準偏差は 1.1 で，正しい言い換え元表現の平均文字列長は 5.7 文字，標準偏差は 3.2 であった．さらに言い換え表現先と言い換え元表現の文字列長の差の平均は 3.0 文字で標準偏差は 2.2 であった．

また，表 7 に言い換え先表現の品詞ごとに抽出例をあげる.「ている」や「している」といった言い換え元表現として用いることができない表現が出現しているが，この様な表現を削除することは今後の課題である.

表 7　言い換え元表現の抽出例

会談 (名詞)	可能性も (助詞)	と語る (動詞)	述べた (助動詞)
と会談した	ている	を示した	と述べた
会談した	可能性がある	と語った	述べた
で会談した	可能性もある	と述べた	を示した
について意見交換した	している	語った	を述べた
と相次いで会談した	可能性が出てきた	考えを示した	を明らかにした

次に言い換え元表現の抽出精度について示す．まず精度の評価方法について説明する．精度は人手により評価を行っている．3 人が言い換え元表現について，新聞記事のニュースであることは勘案せずに正否を判定し，2 名以上が正しいと判断した場合を正解，それ以外は不正解としている．なお，グラフはそのままのデータでは見難いため 10〜50 件で平均をとって表示している.

図 7 は全品詞を対応文数の多い順に並べ，それぞれの言い換え元表現の得点付け順位が 1 位になった表現について人手で評価を行ったものである．なお，全体の傾向を把握するために，縦軸の精度を対数とした場合のデータへの当てはめ近似曲線を実線で示した．図 8 には 3.5 節で行ったフィルタリングの有効性についての評価を示す．図 8 は全品詞でフィルタリング前後の精度の対数曲線での近似のみを示したものである．全体では精度が 12 %向上し，特に対応文数が少ない部分ではかなりの精度向上が見られる.

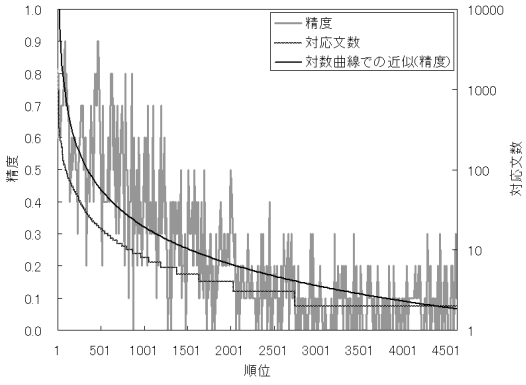

図 7　全品詞のフィルタリング前の精度

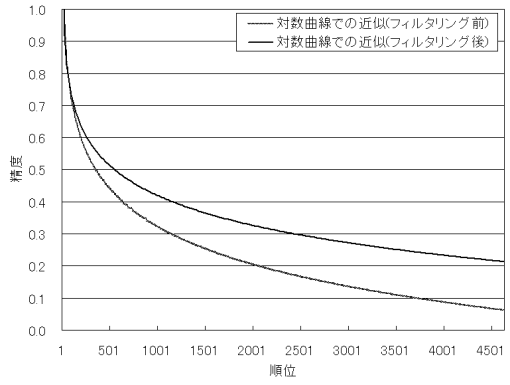

図 8　全品詞のフィルタリング前後の精度の比較

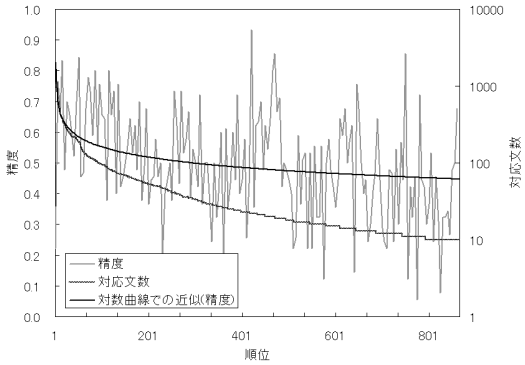

図 9　全品詞のフィルタリング前の精度

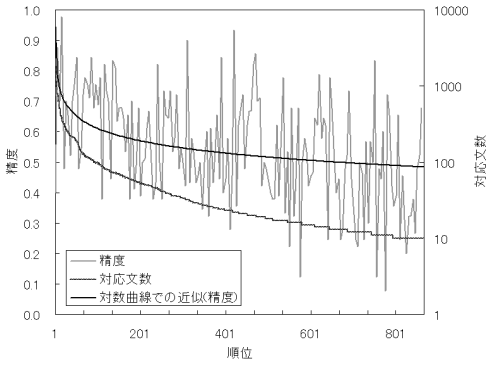

図 10　全品詞のフィルタリング後の精度

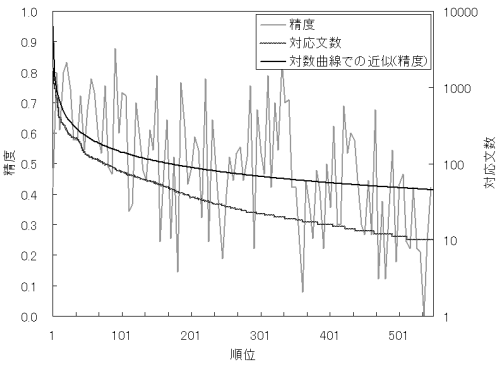

図 11　名詞のフィルタリング前の精度

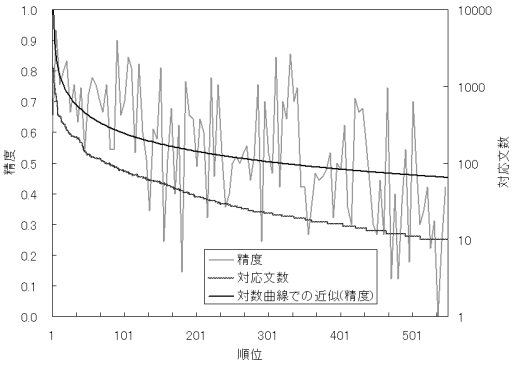

図 12　名詞のフィルタリング後の精度

上記評価方法とは別に以下に述べる評価方法を用いた場合におけるフィルタリング前後の評価を図 9 から図 18 に示す．この評価方法は各言い換え先表現の Web 文集合における対応文数が 10 件になるまでの部分で，言い換え元表現の得点付けによる順位が 3 位までの正否を判定し，全品詞を対象にした場合，名詞，助詞，動詞，助動詞を別々に対象にした場合について，その平均をとったものである．

この図からは，フィルタリングが品詞にかかわらず精度の向上に効果があることがわかる．対応文数が 10 件までということもあり比較的正解となる言い換え元表現が抽出されているため前の全品詞の得点付けが 1 位の場合のデータ程ではないが，フィルタリングにより全品詞において 4 %程度の精度が向上がみられる．

今回用いた言い換え元表現の抽出方法では携帯文の文末にくる表現と Web 文の文末にくる表現が同一内容である必要がある．携帯文の文末が名詞の場合は，抽出した言い換え先表現と

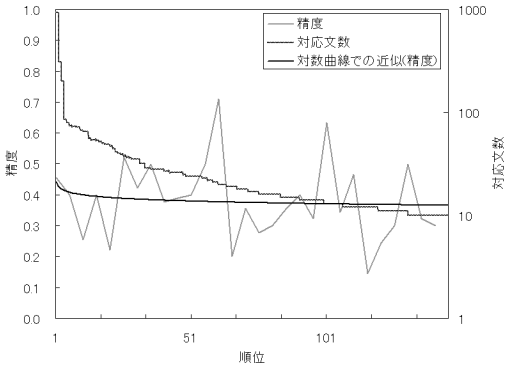

図 13　助詞のフィルタリング前の精度

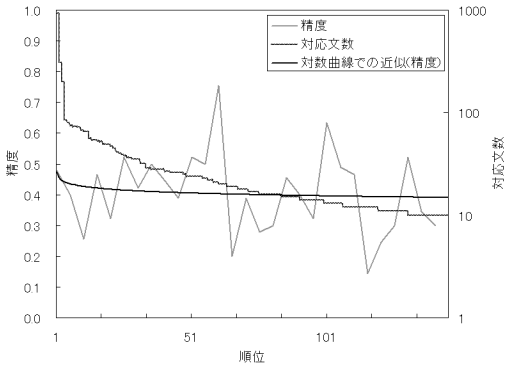

図 14　助詞のフィルタリング後の精度

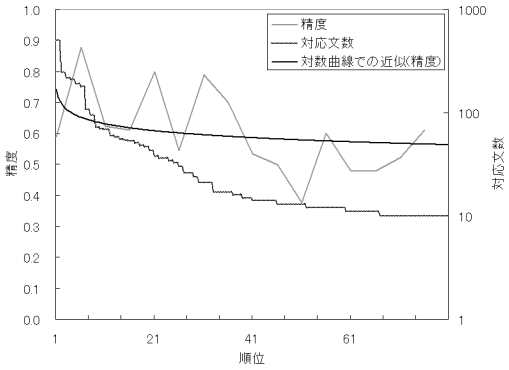

図 15　動詞のフィルタリング前の精度

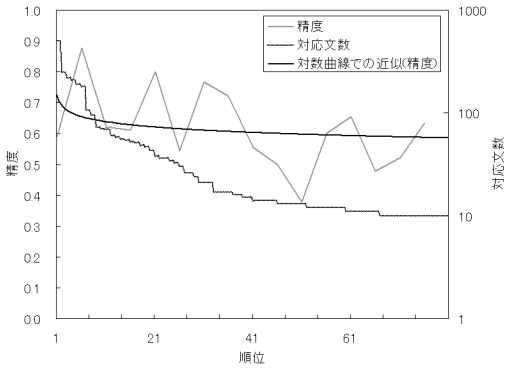

図 16　動詞のフィルタリング後の精度

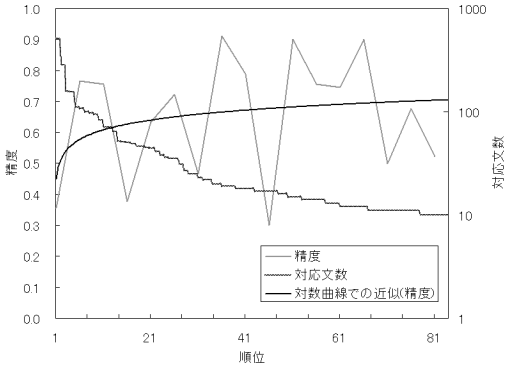

図 17　助動詞のフィルタリング前の精度

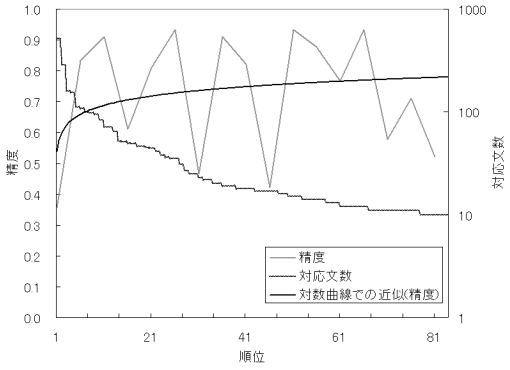

図 18　助動詞のフィルタリング後の精度

言い換え元表現が同一内容の表現がくることが多いため図 12 のように高い精度を得られたが，携帯文の文末が助詞や助動詞の場合では言い換え先表現と同一内容の抽出すべき言い換え元表現は文末よりかなり前にあることが多いという特徴がある．そのため，図 14 や図 18 のようにどのような順位でも精度が低いという結果が得られた．この問題の解決は今後の課題となる．

最後に「精度」と「対応文数」の関係について図 19 に示す．用いるデータは図 7 と同じ全品詞を対応文数の多い順に並べ，それぞれの言い換え元表現の得点付け順位が 1 位になった表現について人手で評価したものである．この図からは，言い換え先表現に対する対応文数が少ないと精度が低く，対応文数が増えるにつれ対数関数的に精度が向上していくことが分かる．つまり，この手法で正解となる表現のより高い抽出精度を求めるならば，精度は対応文数に対し指数関数的な数のコーパスを集めなければならないということであり，これには相当な困難が伴う．よって，今後は構文構造や意味内容を利用する精密な手法を用いることによる言い換え表現抽出を行う必要がある．

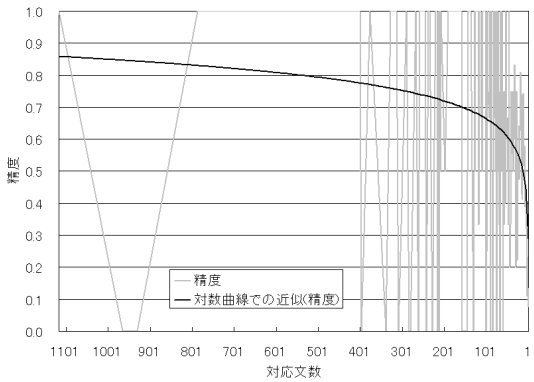

図 19　対応文数と精度の関連

4.2　言い換え例についての考察

言い換え先表現の各々に対応する言い換え先表現についての言語学的考察は興味深いものである．しかし，3.2 節で述べた方法で抽出した言い換え先表現 4617 個全体を対象にすると，各言い換え先表現に対して多い場合は 100 種以上，少ない場合でも 10 種近い言い換え元表現が抽出されているため膨大な労力が必要である．よって相当に長期にわたる研究が必要であるので，別の機会に譲りたい．しかし，典型的な例について言語学的な考察をしておくことは，抽出された言い換え先，言い換え元の性質を窺う上で意味がある．よって，この節では，抽出した全言い換えのうちの相当数を観察した結果，筆者が得た典型的な言い換えパターンについての例

示と考察を行うことにする．なお，以下の例では「言い換え先表現 ← 言い換え元表現 (言い換え先表現に対する表 5 の $a \times b \times c$ による順位)」という形式で言い換えを記述する．

(1) 文末用言の省略による体言止め，など

以下に例を示す．

(1-a)　発表　←　発表した (3 位)
(1-b)　見通し　←　見通しだ（1 位）
(1-c)　見通し　←　見通しを明らかにした（6 位）
(1-d)　見通し　←　見通しを示した (17 位)
(1-e)　高　←　高で取引を終了した (18 位)
(1-f)　事故　←　事故となった (3 位)

文末用言の省略の結果，体言止めになる場合では，(1-a) に見られる「した」（＝「する」の過去形）を省略してサ変名詞のみを残して体言止めにする場合が多い．(1-b) の場合は「である」の言い切りの形である「だ」の省略だが，これも同じようなタイプである．(1-c)(1-d) は形式的用言である「する」や「だ」のような機能語的な用言の省略ではなく，内容語を伴う用言句「明らかにした」「示した」の省略である．これは形式的には導けない言い換えであり，今回のようなコーパスからの抽出データを用いて明らかになった言い換えである．ニュースの文の言い換えとしては普遍性を持つと予想できるが，その適用範囲についてはより深い考察を必要とする．(1-e) はニュース特有の言い換えで，株価や為替レートについての報告となる．これは，株価，為替などの分野でしか成立しない言い換えである．(1-f) はサ変ではない名詞「事故」の場合である．この場合はこの例に見られる「となる」のほかに「になる」などいくつかの典型的用言が省略候補になると予想されるが，それを網羅的に調べることは大規模データを用いての実験となるため今後の課題である．

文末の用言句が省略されても体言止めになるとは限らない．例えば次のような例がある．

(1-g)　盗まれる　←　が盗まれていた (1 位)
(1-i)　盗まれる　←　が盗まれていたことが分かった (2 位)

(1-g) は「いた」という完了を表す動詞接尾辞の省略であり，これは文法的には大きな変化だが，大方の意味は保存されている．(1-i) は「いたことが分かった」という部分の省略である．この部分は「こと」で体言化し，それによって客観化（言語学的には命題化）を行った後，記者ないし記者が直接取材した人の判断である「分かった」が連接している．文を日本語学で使われる 命題 + モダリティ という構造で捉えると，命題の部分だけを単独で取り出すという言い換えである．ニュース記事が少ない文字数で事実，すなわち言語学的には命題，を伝えるも

のと考えれば，その言い換えの構造は理解しやすいものであろうし，ニュース記事の言い換えとしては普遍性を持つ．この問題については，後に助詞止めのところでもう一度議論する．

(2) 引用の「」と」の言い換え

文末用言の省略の結果，体言止めになる場合は，引用を表す「」と」で終わる例が多数観察された．以下に例を示す．

(2-a)　」と　←　と述べた (1 位)
(2-b)　」と　←　との認識を示した (3 位)
(2-c)　」と　←　と語った（4 位）
(2-d)　」と　←　と報じた (34 位)
(2-e)　」と　←　という (8 位)

引用や報告を表すこれらの言い換えが組織的に抽出できたことは，本論文で説明したコーパスの効果である．ただし，(2-e) の「という」は今日では実際上「と言う」ではなく固定した表現のように扱われることが多く，必ずしも「いう」の省略でなく，語彙的な言い換えとみなすべきかもしれない．これらはニュース記事であれば正しい縮約と考えられるが，(2-a) から (2-d) については，もう少し深い言語学的考察を (5) で述べる．

(3) 助詞の省略など文法構造上の言い換え

文末用言に加え，用言の左方の助詞を省略する場合もある．以下に例を示す．

(3-a)　発表　←　を発表した (1 位)
(3-b)　発表　←　に発表した (7 位)

用言の直前の助詞を省略した場合もある．具体的には (3-a) だと「移転計画を発表した」を「移転計画発表」，(3-b) だと「午後に発表した」を「午後発表」という言い換えになる．しかし，この言い換えは不自然な言い換えになることがある．例えば.「XX 誌に発表した」を「XX 誌発表」というのは不自然である．このような例は正解としなかった．さらに

(3-c)　発表　←　と発表した (2 位)

という「と」を省略する場合は特に不自然さが大きい．例えば,「移転すると発表した」を言い換えた「移転する発表」は非文である．ただし,「移転計画と発表した」を「移転計画発表」とする言い換えは若干不自然な程度である．現象的にはかなり複雑だけに，厳密な言語学的分析は今後の研究を待たなければならない．

(4) 意味の類似した語彙ないし言い回しでの言い換え

言い換えの研究でしばしば対象になるものに語彙的な言い換えがある．この範疇に入る言い換えとしては，まず同じ意味を持つ単語への言い換えがある．次に用言の言い換えの例を示す．ただしこの例では，一つの言い換え先「発表」に対する複数の言い換え元をまとめて ← の右側に「、」で区切って示す．

(4-a)　発表　←　公表した (10 位)，分かった (16 位)，まとめた (24 位)，
示した (43 位)，述べた (58 位)，表明した (60 位)

このような同じ意味を持つ表現がコーパスから機械的に得られ，本提案の言い換え抽出の有効性を示している．しかし，このような言い換えは多数得られているわけではない．というのは，文末に使われる用言の種類は相当に限定されているからである．上記のような同義あるいは類義の用言を網羅的に求めるためには，文末以外の部分からの言い換え抽出が必要であるが，これは本論文の範囲を超える研究テーマである．また，サ変名詞以外の体言は，そもそも文末に出現することが少なく，ほとんど言い換えは得られていない．本研究の方法論的限界といえる．

一方，ニュース記事の言い換えという点で特徴的な例を示そう．(1-e) の例で示した株価，為替のニュースに現れる「高」あるいは「安」であるが，以下のような言い換え元表現が求まっている．

(4-b)　高　←　高で取引を終了した (18 位)
(4-c)　高　←　で取引を終えた (2 位)
(4-d)　高　←　高・ドル安となった (5 位)
(4-e)　高　←　高の８４２４円５１銭で取引を終えた (43 位)
(4-f)　高　←　反発して取引を終えた (24 位)
(4-g)　高　←　続伸して取引を終えた (25 位)
(4-h)　安　←　安・ドル高となった (7 位)
(4-i)　安　←　割り込んで取引を終えた (34 位)
(4-j)　安　←　反落して取引を終えた (43 位)

(4-b) は言い換え元の「で取引を終了した」が省略されているが，これを省略してもよいのは，為替，株価のニュースという背景が読み手にも分かっているからである．(4-c) は，言い換え元が「高で取引を終えた」か「安で取引を終えた」であるのかという情報を無視して「高」とするため誤った言い換えである．このような言い換えも候補に出てきてしまうのが，提案手法の限界である．

(4-d) は興味深い．原文では「円高・ドル安となった」なのだが「円高・ドル安」が為替としては同じ情報の繰り返しであるが慣用化している．ところが，短縮すると「円高」にだけ焦

点を当て,「高」と言い換えられる．これは,「円」を通常使用する日本人を対象にした文章だからであろう．(4-h) の「円安・ドル高」についての言い換えも同様である．このように言い換えは，前後の文章という狭義の文脈だけではなく，文化，国家などを含んで考えなければならないため，扱いが難しいことがわかる．(4-e) は，言い換え元では「XX 円 YY 銭高の 8424 円 51 銭で取引を終えた」という構造なので (4-e) で言い換えれば「XX 円 YY 銭高」となり新聞記事としては許容できるが，明らかに情報は欠落している．よって，一般的な言い換えとしては不適切である．今後，言い換え先，元ともに数値まで含めた言い換え抽出の方法を検討する必要があることが分かった．(4-f)，(4-g)，(4-i)，(4-j) は，深い意味解釈をした上での言い換えである．このような言い換えが抽出できたのは，ここで使っている携帯記事と Web 記事のコーパスが言い換え対象以外の部分を用いて対応付けされていること，対象の言い換えを文末に限定したことの 2 点によって，抽出が可能になったと考えられる．しかし，これらもまた新聞記事ニュースでだけ成立する言い換えである．

上記の「高」「安」は典型的な意味的言い換えが行われていたが，それ以外でも，内容を解釈した上で言い換える例がある．例えば次のようなものである．

(4-k)　事故　←　行方不明になっている (17 位)
(4-l)　事故　←　で止まっていた大型トレーラーに追突 (28 位)
(4-m)　盗まれる　←　窃盗事件として捜査を始めた (92 位)

このような言い換えも普遍的に正しいものではないが，高い圧縮率の要約とみなすことはできる．実際，抽出された言い換え元表現のうち，この例のような長めの表現のうち正しいと判断できるものは，相当な情報の損失を伴う高い圧縮率の要約という性格を持つ．

(5) 助詞止め

言い換え先が「焦点に」のような助詞止めの場合は，言語的には複雑である．まず典型的な例を示そう．

「」と」についての要約は既に一部述べたが，言い換え元表現に存在した認識や引用という記者あるいは記者が直接取材した人の持った事実認識を表すモダリティの表現が省略されたものが多い．この言い換えは新聞ニュース記事であれば成立する言い換えである．以下はそのような例である.「認識を示した」「報じた」は少々長い表現であるが，判断や伝聞のモダリティを表すと考えられる．

(5-a)　」と　←　との認識を示した (3 位)
(5-b)　」と　←　と報じた (34 位)

一方，記者らの事実認識を示すモダリティではなく，記者の取材した事実関係そのものの中

で，登場人物が行った言語行動を省略した以下のような例もある．

(5-c)　」と　←　を言い渡した (10 位)
(5-d)　」と　←　求めた (21 位)
(5-e)　」と　←　をけん制した (39 位)

これらは引用符の前に書かれた内容から,「言い渡す」「求める」「けん制する」という言語行動が十分に予想できるからこその省略である．その意味では普遍性のある言い換えではない．

(5-f)　会談へ　←　と会談することが決まった (4 位)
(5-g)　焦点に　←　焦点となる (1 位)

これらの例は,「へ」や「に」のような方向性を表す助詞は，確定的になった将来の事柄を表す例である．より詳しく調査分析すれば，言語学的には興味深い観察が得られるが，この論文の範囲を越える研究テーマと考える．

さて，終助詞「か」は元来が疑問などのモダリティを意味するだけに言い換え元に興味深いものが多い．

(5-h)　狙いか　←　狙いがあるとみられる (2 位)
(5-i)　原因か　←　が原因とみられる (2 位)
(5-j)　犯行か　←　の犯行とみている (3 位)

(5-h) と (5-i) は言い換え元表現に記者自身の判断が記載されているが，それが「か」という終助詞に凝縮していると考えられる．また，(5-j) は，言い換え元表現において記者ではなく警察などの判断を表している．しかし，警察の判断まで含めたモダリティも「犯行」という事実があれば,「か」という終助詞に凝縮することができることを示している．よって，これらの言い換えもまたニュースであることに依存して成立するタイプであるといえよう．このような，命題にモダリティが後接する日本語の基本構造に基づく助詞への言い換えは，言語学の課題としては興味深いし，大きなテーマであるが，詳細に踏み込むことは，この論文の範囲を越えると考える．

以上，本論文で述べた携帯記事と Web 記事の対応付けコーパスを用いて抽出された言い換え先と言い換え元表現のうち筆者が典型例と考えるものについて若干の分析を試みてきた．しかし，この分析自体は，大きなテーマであり，本格的な分析は，本論文で述べたような言い換え抽出結果を用いて言語学的に精密に行うことが望まれる．さらに，ここまで述べてきた分析において問題になったのは，言い換えが新聞記事ニュースとしてなら許容されるが，一般的ではないという場合が多数抽出されたことである．このような場合は，既に述べたように，言い換えよりは，要約あるいは縮約という性質を持つ．要約や縮約の正しさは，informative，indicative

の区別に見られるように，目的依存性があるため，正解の決め方が難しい．この論文では，普遍性のある言い換えを正解と考えているが，要約あるいは縮約としての評価も必要であることが明らかになってきている．しかし，そのことは大きな研究テーマであるため，今後の課題としたい．

5　おわりに

本論文では携帯端末向け新聞記事と Web 新聞記事の対応付けコーパスから文末表現に関する言い換え表現の抽出方法を示した．まず，記事対応になっているデータから文単位での対応付けを行った．そしてそこから言い換え元表現の抽出を形態素単位で行い，それに対して分岐数，頻度，文字列長による得点付けし，さらに言い換え表現を要約に適用した時に必要な意味が削除されることを防ぐためのフィルタリングを行うことにより言い換え表現抽出の精度向上を行った．

今回作成した携帯端末向け新聞記事と Web 新聞記事の対応付けコーパスを用いることを想定すると，以下のような課題が残っている．

(1) 抽出された言い換え表現を用いた文縮約を試みることおよびその評価

言い換え元表現と言い換え先表現の組から機械的に言い換えを生成することができる．文字列の単純マッチングを用いて,「〜を明らかにした。」を「〜表明。」へ,「〜を決めた。」を「〜決定。」へと言い換えることができる．実際，我々はこのようなシステムを試作したが，この方法から分かるように予測された結果以上のものは得られない．したがって，抽出した言い換えを言語的に分析して一般化したルールを作成することができれば,「を M(名詞サ変接続) する方針。=> M へ。」というルールで,「〜審議経過を開示する方針。」を「〜審議経過開示へ。」という適用範囲の広い言い換えが可能になると予想される．しかし，言語的分析は，人間が行うにしても，機械学習を利用するにしても，それ自体が大きなテーマであるため，今後の課題としたい．

(2) 名詞以外での言い換え表現の精度の向上

(3) 精度向上を目的としたフィルタリング規則の追加

(4) 文末以外に現れる表現の言い換え抽出の検討

通常，言い換え抽出においては，(乾 2002) で述べられ，また本論文第 1 節でも述べたように，言い換え候補をコーパスから網羅性良く抽出することが大きな課題である．提案手法の場合は，携帯端末向け新聞記事と Web 新聞記事の対応付けコーパス双方の文末表現に限定したことによって，この問題を回避した．しかし，文末以外に現れる表現の言い換えを抽出しようと

した場合は，たちどころにこの網羅性の良い言い換え候補抽出を解決しなければならない．これに関しては多くの研究成果があるが，文末言い換えに限定して機能する本論文での提案とは根本的に異なる方法論が必要となる．したがって，本論文で紹介したコーパスを用いるにしても，新たな研究テーマとして検討する必要があるため，将来的な研究課題となる．

謝辞

本研究の初期の段階で尽力いただいた佐藤大君 (東京電機大学大学院，現在，富士電機情報サービス株式会社勤務) に深く感謝いたします．なお，本研究の一部は，科学研究費補助金　特定領域研究「情報学」，課題番号 16016215 の補助を受けて行われました．

参考文献

安藤彰男，今井亨ほか (2001)．“音声認識を利用した放送用ニュース字幕制作システム．”電子情報通信学会論文誌, **84-D-II**, pp. 877–887.

Brazilay, R. and McKeown, K. (2001). “Extracting paraphrases from a parallel corpus.” *Proceedings of ACL-EACL 2001*, pp. 50–57.

藤田篤，乾健太郎，乾裕子 (2000)．“名詞言い換えコーパスの作成環境．”電子情報通信学会思考と言語研究会予稿集, pp. 53–60.

乾健太郎 (2002)．“言語表現を言い換える技術．”言語処理学会第 8 回年次大会チュートリアル, pp. 1–22.

Inui, K. and Hermjakob, U. (Eds.) (2003). *Proceedings of the Second International Workhop on Paraphrasing*. ACL2003.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). “Text Simplification for Reading Assistance: A Project Note.” *Proceedings of The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications, Workshop of ACL03*, pp. 9–16.

鍛冶伸裕，川原大輔，黒橋禎夫，佐藤理史 (2003)．“格フレームに基づく用言の言い換え．”自然言語処理, **10** (4), pp. 64–81.

Kanayama, H. (2003). “Paraphrasing Rules for Automatic Evaluation of Translation into Japanese.” *Proceedings of The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications, Workshop of ACL03*, pp. 88–93.

Mani, I. (2001). “Automatic Summarization.” *John Benjamins*.

松本裕治，北内啓，平野善隆，松田寛 (2002)．“形態素解析システム「茶筌」version 2.2.9 使用説明書．”奈良先端科学技術大学院大学松本研究室.

Murata, M. and Isahara, H. (2001). “Universal model for paraphrasing - using transforma-

tion based on a defined criteria." *Proceedings of Workshop on Automatic Paraphrasing: Theories and Applications, NLPRS2001*, pp. 47–54.

大森岳史，増田英孝，中川裕志 (2003). "Web 新聞記事の要約とその携帯端末向け記事による評価." 情報処理学会自然言語処理研究会, **153**, pp. 1–8.

Rinaldi, F., Dowdall, J., Kaljurand, K., Hess, M., and Molla, D. (2003). "Exploiting Paraphrases in a Question Answering System." *Proceedings of The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications, Workshop of ACL03*, pp. 25–32.

佐藤大，岩越守孝，増田英孝，中川裕志 (2004). "Web と携帯端末向けの新聞記事の対応コーパスからの言い換え抽出." 情報処理学会自然言語処理研究会, **159**, pp. 193–200.

関根聡 (2001). "複数の新聞を使用した言い換え表現の自動抽出." 言語処理学会第 7 回大会ワークショップ「言い換え/パラフレーズの自動化」.

Shinyama, Y. and Sekine, S. (2003). "Paraphrase Acquistion for Information Extraction." *Proceedings of The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications, Workshop of ACL03*, pp. 65–71.

Terada, T. and Tokunaga, T. (2001). "Automatic disabbreviation by using context information." *Proceedings of Workshop on Automatic Paraphrasing: Theories and Applications, NLPRS2001*, pp. 21–28.

Torisawa, K. (2001). "A nearly unsupervised learning method for automatic paraphrasing of Japanese noun phrases." *Proceedings of Workshop on Automatic Paraphrasing: Theories and Applications, NLPRS2001*, pp. 63–72.

Yamamoto, K. (2002). "Acquisition of Lexical Paraphrases from Texts." *Proceedings of Computerm2 Workshop of COLING2002*, pp. 22–28.

略歴

岩越守孝: 2003 年東京電機大学工学部電気工学科卒業．2005 年同大学院工学研究科情報通信工学専攻修士課程修了．現在，キヤノン株式会社に勤務．本論文は在学中の成果をまとめたものである．

増田 英孝: 1995 年東京電機大学大学院博士後期課程修了．博士 (工学)．東京電機大学工学部助手，講師を経て，同助教授．Web 情報検索，Web マイニングなどの研究に従事．

中川裕志: 1975 年東京大学工学部卒業．1980 年同大学院博士課程修了．工学博士．横浜国立大学工学部講師，助教授，教授を経て，1999 年より東京大学情報基盤センター教授．言語処理学会長 (2004.6 - 現在)，ACL Executive Committee (2002 - 2004) ．計算言語学，Web テキストマイニング，情報検

索，情報抽出などの研究に従事.

(2005 年 2 月 17 日 受付)
(2005 年 5 月 27 日 再受付)
(2005 年 6 月 30 日 採録)