

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

佐藤 大[†] 岩越 守孝[†] 増田 英孝[†] 中川 裕志[‡]

[†] 東京電機大学工学部 〒 101-8457 東京都千代田区神田錦町 2-2

[‡] 東京大学情報基盤センター 〒 113-0033 東京都文京区本郷 7-3-1

E-mail: {sdai,iwakoshi}@cdl.im.dendai.ac.jp, masuda@im.dendai.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp

概要

本研究は、数十文字程度の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 Web 新聞記事一年分の対応付けされたコーパスを用いて文末表現の縮約などの言い換えパターンの抽出を機械的に行うことを目的としている。まず携帯向け新聞記事と Web 新聞記事で文単位の対応付けを行う。携帯文からは言い換え候補集合をつくり、Web 文で文末からの文字列マッチングを行い言い換えパターンの抽出を行った。

Extraction of Paraphrasing Patterns from Aligned Corpora of Web and Mobile Terminal News Articles

[†]Dai SATO [†]Moritaka IWAKOSHI [†]Hidetaka MASUDA [‡]Hiroshi NAKAGAWA

[†] School of Engineering, Tokyo Denki University

2-2 Kandanishiki-cho, Chiyoda, Tokyo 101-8457, Japan

[‡] Information Technology Center, The University of Tokyo

7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan

E-mail: {sdai,iwakoshi}@cdl.im.dendai.ac.jp, masuda@im.dendai.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp

Abstract

The target of this work is the system which extracts paraphrasing patterns from Web newspaper articles of several hundreds of characters and their counter parts distributed for mobile terminals, which consist of fifty to a hundred characters. The extracted paraphrasing patterns are to be useful for transforming Web articles into shorter mobile articles.

At first we developed an aligned corpus of Web articles and mobile articles of one year period. From them, we extract a number of candidates of paraphrasing patterns. Then, we try to find paraphrasing patterns by character-by-character matching from the end to the beginning of articles. We found that the point where a number of string branch increases gives us the good paraphrasing patterns.

1 はじめに

最近、言い換えの研究がさかんになっている [1][2]。国語辞典を用いた用言の言い換え [3] や普通名詞の言い換え [4] などさまざまなものがある。言い換えの研究には多くの応用が考えられる。た

例えば、機械翻訳の前処理、年少者や初心者向けの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。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。異なる視点として要約という応用分野も有力である。従来の文書要約は重要文の抽出

が主体であった [5]。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、「本法案が衆議院本会議で審議が始まった」を「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。実際にこのような言い換え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の進行からみても有用さが増してくるであろう。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パターンを収集することは意義深い。このような文短縮を目的とした言い換えパターンを自動収集する言語資源として 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が注目される。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。さらに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、多様かつ大量な文末表現の縮約などの言い換えパターンを機械的な手法で抽出することが可能になる。

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	16693	38.7
	上記以外	8795	20.4
助詞		8079	18.7
動詞		6633	15.4
助動詞		2460	5.7
その他		508	1.2
合計		43168	100

本稿では、上記の言い換えパターンを自動的に抽出する方法の提案とその結果について述べる。まず、2 章では研究で使った記事データについて述べる。3 章では Web 記事と携帯記事の対応付け方法について述べる。4 章では、文毎に対応付けられた Web 記事と携帯記事から通常表現と短縮表現の対である言い換えパターンを抽出する方法について述べる。5 章で実験結果を示し、6 章で結論を述べる。

2 対象とする新聞記事データ

本研究では、我々が収集した対応付けコーパスを用いる [6]。このコーパスは、インターネット上に配信されている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。

Web 記事は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。

今回は 2001 年 4 月 26 日から 2002 年 3 月 31 日までに収集した合計 22147 記事を対象に研究を行った。

3 文単位の対応付け

3.1 対応付けの方法

本研究では 2 章で述べた Web 記事と携帯記事のコーパスを用いる。まず記事同士の対応付けは以下のようにして行う。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。この対応付けを行うと、Web 記事の言い換え、要約の言語データとして携帯記事を対応させる対応付けコーパスができる。携帯記事中の名詞と Web 記事中の名詞を調べ、

得点 Sim

$$= 3 \times (\text{Web 記事見出しと } M \text{ の一致数}) \\ + (\text{Web 記事本文と } M \text{ の一致数}) \quad (1)$$

の値が 35 以上の場合を対応が付いたとすることにより、現在までの実験で 80 % 近い携帯記事を 100 % の精度で対応付けができた。ここで M は携帯記事の名詞を表している。そこで、この方法で 1 年分の対応付けコーパスを作成した。

次にこの対応付け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。これも名詞の出現頻度によって文対応を行った (図 1 を参照)。以下に対応付けの方法を示す。

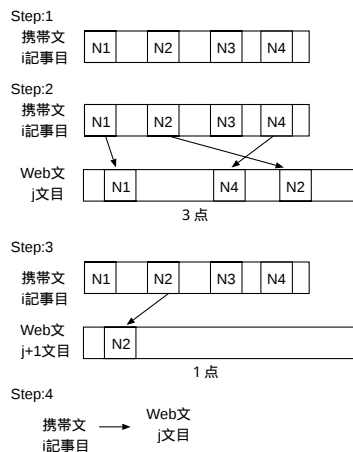


図 1: 文単位の対応付け方法

Step:1 $i=1$

携帯記事の i 文目を形態素解析し名詞を抽出

Step:2 $j=1$

Web 記事の j 文目と Step:1 で抽出した名詞
の一致数をカウント

Step:3 $j=j+1$

Web 記事の最後の文になるまで Step:2 を繰
り返す

Step:4 もっとも多く名詞がマッチした文を対応
文とする

Step:5 $i=i+1$

Step:1 に戻る

まず携帯記事の一文目を形態素解析 [7] して名詞だけを取り出す。次に、Web 記事の一文目と取り出した名詞とでパターンマッチングを行い、何個の名詞がマッチしたかをカウントする。この動作を Web 記事の最後の文まで行い最も多くの名詞が一致した文を対応文とする。さらに携帯記事の二文目以降も同様のプロセスを行うことにより、それぞれに対応する文を抽出することが可能となる。もしマッチした名詞の数が同じ場合には、より前に出てくる文を優先して抽出する。ここで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 文とする。

表 2: 携帯記事の 1 記事あたりの文数

携帯記事の構成	記事数	割合 [%]
一文	1801	8.1
二文	19668	88.8
三文	676	3.1
四文	1	0.0
合計	22147	100

3.2 対応付けの結果

約一年分の対応記事コーパスから、43171 組の対応文を抽出した。携帯記事のほとんどが二文で構成されていること (表 2 を参照) 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。詳細を表 3 に示す。

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。

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

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%である。

表 4: 候補集合

抽出単語	頻度
発表	675
示す	650
安	581
高	501
会談	483
表明	454
逮捕	429
死亡	296
決定	286
方針	264
調査	256

4 言い換えパターン抽出方法

3章により抽出した携帯記事の文(以下、携帯文)とWeb記事の文(以下、Web文)を用いて言い換えパターンの抽出を行う。例えば、

携帯文: コンピュータウイルス感染防止に有力な方法が判明。

Web文: コンピュータウイ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「判明」が「分かった」に言い換えられていることが、人間が見れば判断することが出来る。

このような言い換えパターンを自動的に抽出するために以下の方法を試みた。

Step:1 候補集合の作成

Step:2 候補集合の単語を文末に含む携帯文とそれに対応するWeb文を抽出

Step:3 Web文で文末からの文字列マッチング

4.1章から4.3章でそれぞれのStepを詳しく説明する。

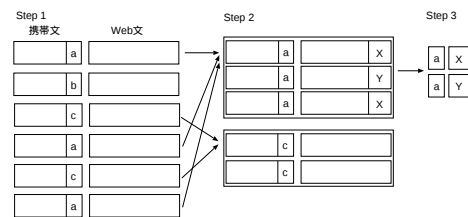


図 2: 言い換えパターン抽出アルゴリズム

4.1 候補集合について

まず、言い換えのもとになる単語の候補を抽出する。本研究では、言い換えのもとになる単語を候補集合として、あらかじめ抽出する。この候補集合は携帯文を形態素解析して文の最後にくる単語を抽出して出現頻度でまとめたものである。ただし「へ」、「か」など助詞で終わる単語については、助詞だけでは意味が通らず、言い換えのもととしてはふさわしくないため、最後の1単語ではなく、助詞の前の単語をあわせて、例えば「派遣へ」、「影響か」といった2単語で抽出をした。また出現頻度が1の単語はStep3において、対象となる文が一文なので、文字列のマッチングを行うことが出来ないため、候補集合から除外してある。抽出した単語の例を表4に示す。合計3289種類の単語を抽出した。

表 3: 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
一文	1 対 1	1801
二文	2 対 1	9606
	1 対 1	29732
三文	1 対 1	2028
四文	1 対 1	4
合計		43171

4.2 集合の作成

ここでは候補集合に出現する単語が、文末になる携帯文を抽出する。このとき、3 章の方法で携帯文と Web 文が対応付けられているので、それに対応する Web 文も同時に抽出する。例えば、文末が「判明」の携帯文と、その携帯文に対応する Web 文を抽出すると図 2 の Step3 の状態になる。

これにより、候補集合に出現する単語の言い換えを含んでいる可能性がある Web 文の集合を作ることができる。

4.3 文末からの文字列マッチング

このステップでは Web 文に注目して文末からの文字列マッチングを行う。実際に「表明」が文末になっている携帯文に対応する Web 文における文末からの文字列マッチングの例を図 3 に示す。

まず、一番右側にある「た」が最後の一文字である。次に二文字目までを見ると「した」「めた」「けた」… と続く。ここで人手で見ると「明らかにした」までが抽出できれば、それが「表明」の言い換えになると推測できるが、自動的に抽出するために、マッチングをストップさせる点を求める条件が必要になる。そこで以下のことに着目する。すなわち、一般に単語の内部では、ある文字の後に続く文字の種類数は減少する。例えば、「言い換え」という単語の場合、「言」の後には「い、語、論、説、外、行、質、霊、下、…」などと多数の文字種類が接続しうる。「言い」の後だと、「方、

淀、繕、逃、替」など接続文字種類が減少する。「言い換」まで進むと「え」くらいしか接続する文字がなくなってくる。ところが、「言い換え」という単語が完成すると、その後ろには助詞をはじめとし多くの文字種類は接続することになり、接続文字数は増加に転ずる。なお、接続文字数を以下では分岐数と呼ぶ。文末から文字列を逆方向にたどる場合も同じ傾向が予想される。そこで、以下の方法で言い換えパターンの候補を抽出した。文末からたどった場合に分岐数が減少ないし一定の状態から初めて増加に転じた文字までの文字列を言い換えパターンの候補として抽出する。

この方法で抽出した結果のうち、「判明」、「発表」、「表明」についての言い換えパターン数と人手で付けた正解に対する正解率を表 5 に示す。なお、正解の言い換えの基準は抽出された言い換えパターンをもとの単語に戻しても問題が無いか、である。

この結果では必ずしも自動抽出の正解率としては満足できない。特に「した」などの具体的意味に乏しい短い表現が取り出されている。そこで言い換えパターンの言語的性質について検討してみる。単語あるいは固定的な言い回しには

- 単語内部では分岐数が少なく、単語が終了すると分岐数が増加に転じる。しかし、単語より大きい言語単位を考えると、分岐数そのものが固定した言い回しの程度を表すと考えられる。

し言い回しの発見に寄与すると考えることができる。

5 文末マッチングの分析

文末から文頭方向に向けて文字列マッチングを行い言い換えパターンを抽出するために以下に示す3つについての分析を行う。

a: 分岐数が減少ないし、一定の状態から上昇する点を候補とする。この点での分岐数を以下でのaの値とする。

b: 出現頻度の減少率

$$\text{減少率}^{-1} = \frac{(\text{次に処理する文字の出現頻度})}{(\text{その文字列自体の出現頻度})} \quad (3)$$

c: 文字列長の適切さ

$$f(x) = \frac{1 - e^{-x}}{1 + e^{-x}} \quad (4)$$

xは文字長。この関数により、「た」「した」などの長さ3以下程度の短い表現を削除する。

抽出された単語に対して a, b, c, a×b, a×c, b×c, a×b×c を算出し、それぞれ数値が高い順に並べる。どの組み合わせを用いた場合に、高い精度で単語を抽出することができるかを知るために平均精度を用いた評価を行った。平均精度は以下の式を用いる。

$$\text{平均精度} = \frac{1}{N} \sum_{k=1}^N \frac{k \text{ 位までの正解個数}}{k} \quad (5)$$

「表明」、「判明」、「発表」の3つの単語について抽出された言い換えパターンの平均精度を求めた結果を図6に示す。「表明」については、どの点においても9割近い精度で言い換えパターンが抽出されている。「判明」はc:文字数で他の点と比べて精度が落ちてしまっている。これは「27日、明らかにになった」といった一般的に言い換え

パターンとして考え難いものが上位に多く存在していたからである。「発表」に関しては文末マッチングをする時に候補となる文に、実際には言い換えパターンが含まれていない文が多くあり、そこから単語が抽出されてしまうので全体的な精度が下がったと考えられる。

表 5: 分岐数が初めて増加に転じる点で切り出した文字列の精度(例)

抽出単語	頻度	正解率[%]
判明	13	52.4
発表	24	81.8
表明	23	72.7

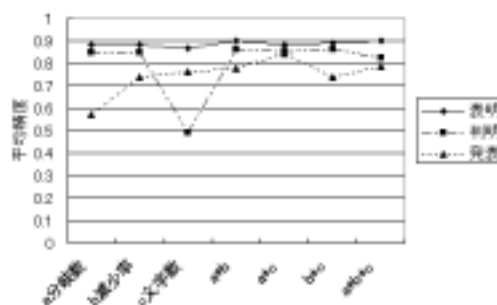


図 6: 平均精度

全体的に見れば高い精度で言い換えパターンが抽出されていると言える。今回の例では a×c の方法、分岐点増加点によって切り出した候補に、文字列の長さを考慮した場合が最も精度が良かった。実際に抽出された言い換えパターンの例を示す(表6を参照)。

6 おわりに

本稿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パターン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の対応付けを行っ

た。そこから言い換えパターンの抽出を文末から文字列マッチングし、分岐数が増加に転ずる点を抽出する方法で言い換えパターンの候補を抽出した。次に分岐数、出現頻度の減少率、単語の文字数を考慮することにより、高い精度での言い換えパターンを抽出することができた。今後は抽出された言い換えパターンを用いた文章縮約を試み検討評価を行う。また、より大規模なコーパスからの言い換えパターンの抽出を行う予定である。

参考文献

- [1] 乾健太郎：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 (2002).
- [2] ACL2003: *Proceedings of the Second International Workshop on Paraphrasing* (2003).
- [3] 鍛冶伸裕, 川原大輔, 黒橋禎夫, 佐藤理史：格フレームに基づく用言の言い換え, 自然言語処理, Vol. 10, No. 4, pp. 64-81 (2003).
- [4] 藤田篤, 乾健太郎, 乾裕子：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 (2000).
- [5] Mani, I.: Automatic Summarization, *John Benjamins* (2001).
- [6] 大森岳史, 増田英孝, 中川裕志：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, No. 153, pp. 1-8 (2003).
- [7] 松本裕治, 北内啓, 平野善隆, 松田寛：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 (2002).
- [8] Tanaka-Ishii, K., Yamamoto, M. and Nakagawa, H.: Kiwi: A Multilingual Usage Consultation Tool based on Internet Searching, *Proceedings of the Interactive Posers/Demonstrations(ACL2003)*, pp. 105-108 (2003).

表 6: 抽出された言い換えパターンの例

候補単語	抽出された言い換えパターン	
	成功例	失敗例
発表	明らかにした 公表した 分かった を示した 確認した と語った	ている 選ばれた 減少した
示す	明らかにした 発表した 述べた 強調した 指摘した	したものだ
会談	協議した 意見交換した を確認した	行った 署名した
表明	示した と語った を決めた	したのは初めて
決定	発表した を固めた に提出した	した
判明	分かった 明らかになった 明らかにした 発表した 断定した	たという された