Web 記事と携帯端末向け記事からの文末サ変名詞の言い換えパターンの抽出

佐藤 大 †, 岩越 守孝 †, 増田 英孝 †, 中川裕志 ‡
東京電機大学工学部 †　東京大学情報基盤センター ‡

1　はじめに

最近、多くの応用を見込んで言い換えの研究がさかんになっている [1]。例えば、年少者や初心者向けの教科書やマニュアルを読み易くする、などは直接的に役立つ応用である。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 [2] 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 [3]。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。
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。通常の書き言葉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、これまで人手で行われていた [4]。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。

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 [5]。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。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 組の対応コーパスから抽出した 88333 組の対応文対 (以下、Web 記事から抽出した文を Web 文、携帯記事から抽出した文を携帯文と呼ぶ) を対象に行った [6]。

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。

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	34312	38.8
	上記以外	15875	18.0
助詞		14484	16.4
動詞		16186	18.3
助動詞		6671	7.6
その他		805	0.9
合計		88333	100.0

3　言い換えパターンの抽出方法

本節では対応付けられた携帯文と Web 文を用いて言い換え先表現と言い換え元表現の抽出について述べる。例えば、

携帯文: 自衛隊の派遣を 表明。

Web 文: 自衛隊を今週末から派遣することを 明らかにした。

という対応文がある。このとき文末に注目すると携帯文では 表明 で終わっているのに対して、Web 文では 明らかにした で文が終わっている。文の内容から 明らかにした が 表明 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。本研究で目的としている言い換えは「Web 記事の文 → 携帯向け記事の文」という方向性を持つ。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 言い換え先表現になる携帯文の文末表現抽出

Step:2 言い換え先表現を文末に含む携帯文とそれに対応する Web 文の対を抽出

Step:3 Web 文で文末から文頭方向に文字列走査し、言い換え元表現を切り出す

表 2: 候補集合 (上位 30 種類)

抽出単語	頻度	抽出単語	頻度	抽出単語	頻度
した	2370	**表明**	913	求める	469
高	2002	**死亡**	718	**合意**	429
安	1892	**決定**	648	**検討**	426
」と	1652	いる	625	**批判**	424
発表	1382	いた	600	られる	417
示す	1358	ため	520	**開始**	400
れる	1302	方針	505	**協議**	390
逮捕	1098	見通し	501	語る	375
する	1054	**強調**	485	**要請**	365
会談	992	**判明**	477	**発言**	361

3.1 言い換え先表現の抽出

まず、言い換え先となる文末表現を抽出する。具体的には携帯文を形態素解析して、文末にくる単語を抽出し出現頻度でまとめたものである。このうち出現頻度が上位 30 位までのものを表 2 に示す。本研究では候補集合と呼ぶことにする。ただし「へ」「か」など助詞で終る場合、助詞だけでは意味が通らず、言い換え先としてはふさわしくないので、最後の一形態素ではなく、助詞の前の名詞をあわせて、例えば「派遣へ」、「影響か」といった二形態素で抽出をした。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、文字列の切り出しが行えないので、候補集合から除外してある。この処理で合計 4566 種類の文末表現を抽出した。表 2 を見ると、動詞終止形、助詞、形容動詞語幹 (「高」「安」)、など様々だが、一番多いのはサ変名詞 (表 2 の太文字) であり 15 個で 50%である。文末がサ変名詞は、いわゆる体言止めである。この表でも表されるように頻度が高く、適用される頻度も高いと推測される。そこで本稿では、サ変名詞に対象を絞り、言い換えパターンの抽出および評価については表 2 の太字のうち上位 10 種類のサ変名詞を対象にした。

3.2 言い換え表現の抽出

次に候補集合の単語を文末に含む携帯文に対応する Web 文を抽出することにより、候補集合の単語に対する Web 文の集合を抽出する。

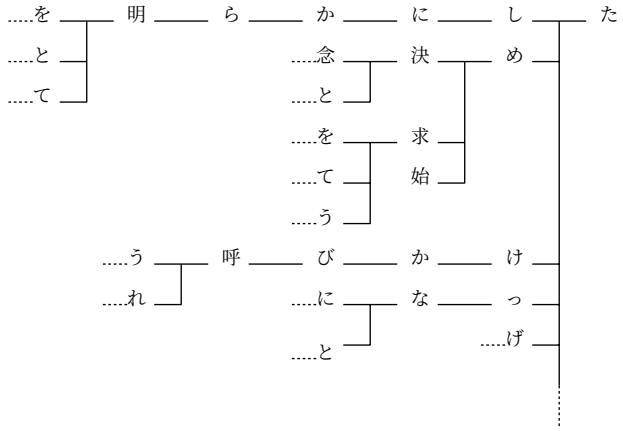

図 1: 文末から文頭への文字列走査の例

3.3 文末からの文字列走査

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」が文末の携帯文に対応する Web 文の集合において文末からの文字列走査を行った例を図 1 に示す。処理の方向は右から左に行っている。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、文字列走査をストップさせる点を求める条件が必要になる。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

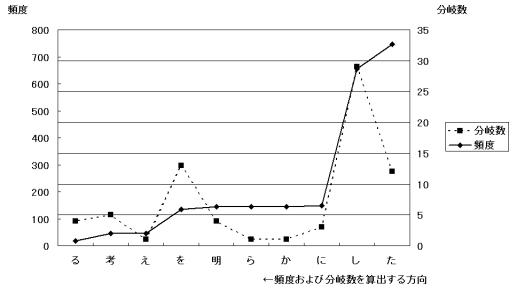

図 2: 出現頻度と分岐数 (表明)

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。

- 出現頻度が多い。
- 文字列の長さが適度である。なぜなら、短すぎると単語や言い回しにならず、長すぎると文脈に依存した表現になってしまうからである。

という性質がある [7] ので、これを利用する。そこで一例として、文末から文頭方向に向かって文字列走査を行う時に、分岐数とそれぞれの文字列における出現頻度をカウントした。候補集合「表明」に対する Web 文の集合において文末から文字列走査を行い「明らかにした」に至るまでの分岐数および出現頻度を図 2 に示す。ただし、文字列走査の処理は右から左に行っている。まず分岐数に注目する。「を」と「明らかにした」の間で分岐数が増加している。次に出現頻度については「を」と「明らかにした」の間を見ると、その前の部分では一定 (頻度の増減が無い) であるが「を」のところで出現頻度が減少していることがわかる。

4 分岐数、頻度、長さに基づく言い換え抽出

文末から文頭方向に向けて文字列走査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: 分岐数が上昇する点を抽出し (ただし 30 文字以内)、そのときの分岐数を a とする

b: 出現頻度を b とする

c: ほどよい長さを抽出するために c を定義する

$$c = \log_e(len - 1)$$

len は文字列長を表す。

上記の方法により抽出した単語に対して $a{\times}b{\times}c$ を算出し、数値が高い順に並べ、次式の平均精度により評価を行う。平均精度は上位から N 位までの候補を考慮する精度の平均値である。

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

図 3 に候補集合 (表 2 を参照) の上位 10 種のサ変名詞に対して、抽出された言い換えパターンの上位 1 位から 20 位までで $a{\times}b{\times}c$ の大きい順に並べた各順位までの精度を示す。この結果を見ると「逮捕」以外では第 1 位の抽出結果は正しい言い換えパターンである。また、「決定」では 7 位までが全て正しい言い換えパターンを抽出している。候補集合の上位 10 種類のサ変名詞の精度を平均してみると 1 位では 90%、3 位までで 87%、5 位までても 76%の精度で抽出でている。実際に抽出された言い換えパターンの例を表 3 に示す。1 位で正しい言い換えパターンが抽出されなかった「逮捕」に注目すると「容疑で逮捕した」が抽出されている。容疑の前には放火、殺人、窃盗など多くの犯罪の種類が続き、分岐数が多くなるので言い換え元の候補として抽出された。しかし「放火容疑で逮捕した → 放火逮捕」となってしまうので言い換えパターンとしてはふさわしくない。

また候補集合の上位 100 種類のサ変名詞について $a{\times}b{\times}c$ の一位で抽出された言い換え元候補を評価した結果 84%の精度であった。

5 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。携帯文から、言い換え先の表現となる候補集合を作成し、Web 文から言い換え元表現の抽出を行った。言い換え元表現の抽出には文末からの文字列走査を行い、分岐点が増加する点を抽出しその時の分岐数、出現頻度、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、(1) 抽出された言い換えパターンを用いた文縮約を試みる (2) 文末におけるサ変名詞以外の言い換え抽出評価、(3) 文末以外に現れる言い換えパターンの抽出実験などが課題となる。

表 3: 抽出された言い換えパターンの例

候補集合	抽出された言い換えパターン		
	1 位	2 位	3 位
発表	を発表した	発表した	と発表した
逮捕	容疑で逮捕した	の疑いで逮捕した	逮捕した
会談	と会談した	会談した	で会談した
表明	を表明した	表明した	を明らかにした
死亡	死亡した	人が死亡した	人が負傷した
決定	を決めた	することを決めた	を決定した
強調	を強調した	強調した	と強調した
判明	分かった	で分かった	明らかになった
合意	することで合意した	で合意した	合意した
検討	を検討していることを明らかにした	を検討している	検討していることを明らかにした

参考文献

[1] Inui, K. and Hermjakob, U.: Proceedings of the Second International Workshop on Paraphrasing (IWP2003) (2003).

[2] 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).

[3] Mani, I.: Automatic Summarization, *John Benjamins* (2001).

[4] 安藤彰男, 今井亨ほか：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, Vol. 84-D-II, pp. 877–887 (2001).

[5] 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, Vol. 153, pp. 1–8 (2003).

[6] 佐藤大, 岩越守孝, 増田英孝, 中川裕志：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, Vol. 159, pp. 193–200 (2004).

[7] 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).

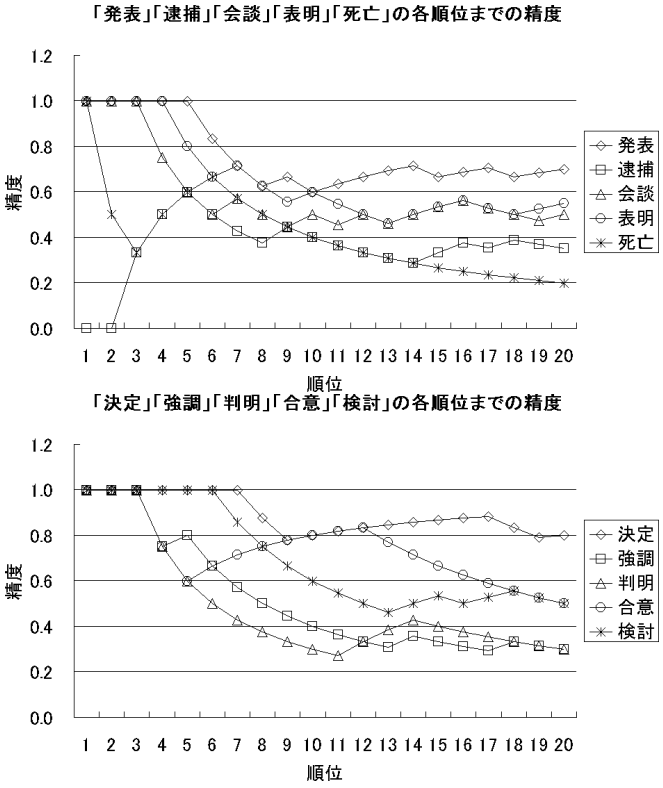

図 3: 平均精度