

オノマトペを言い換える表現の自動収集

古武 泰樹 佐藤 理史 駒谷 和範

名古屋大学大学院 工学研究科 電子情報システム専攻

{y_furuta, ssato, komatani}@nuee.nagoya-u.ac.jp

1 はじめに

オノマトペとは、「ぎーぎー」、「さらり」、「しーん」などの擬音語・擬態語の総称である。日本語は他の言語に比べ、オノマトペが豊富な言語であり、動作のニュアンスや感情を伝える際に頻繁にオノマトペが用いられる。

オノマトペを用いた表現の中には、オノマトペを、それとほぼ等価な意味を持つ語（オノマトペ以外の語）で言い換えることができるものがある。例えば、次のような言い換えが考えられる。

(1) 雨がぎーぎーと（→激しく）降る

このような言い換えを機械的に実現することができれば、日本語の読解支援、および、翻訳支援に役立つと考えられる。

我々は、このような、言い換えの自動収集を実現し、編纂中のオノマトペ用法辞書 [1] に収録した。本稿ではこの自動収集法について述べる。

2 基本方針

オノマトペの言い換えを収集するにあたり、次の3つを考える必要がある。

1. 対象とするオノマトペと用法
2. 言い換えを収集する単位
3. 言い換えを収集する方法

2.1 対象とするオノマトペと用法

先に示した「雨がぎーぎーと降る」では、「ぎーぎーと」を「激しく」に、言い換えることができる。しかし、次の例のように、1語で言い換えることが難しいオノマトペも存在する。

(2) 水をひたひたに（→？）入れる

本研究では、このようなオノマトペは、言い換え収集の対象とはせずに、1語で言い換え可能なもののみを対象とする。

オノマトペは連用修飾用法、連体修飾用法、用言用法などの複数の用法を持つ。それぞれの用法に対して、適切な言い換え語が異なる場合がある。

(3) はっきりと（→明瞭に）見える [連用修飾用法]

(4) はっきりした（→明確な）態度 [連体修飾用法]

(5) 事情がはっきりする（→判明する） [用言用法]

これらの用法の中で、連用修飾用法はオノマトペの典型的な用法である。連用修飾用法の語形は「～と」、「～に」など複数あるが、「はっきりと」のように、「～と」の語形をとるオノマトペは特に多い。そこで、本研究では、この「～と」の語形をとる連用修飾用法を言い換え収集の対象とする。

2.2 言い換えを収集する単位

一般に、オノマトペは多義性を持つ。このため、オノマトペの適切な言い換え語は、オノマトペ単体では一意に定まらず、文脈に応じて、定まるのが普通である。

(6) 雨がぎーっと（→激しく）降る

(7) 資料をぎーっと（→大まかに）読む

そこで、本研究では、オノマトペとオノマトペが修飾する動詞の組に対し、言い換え語を収集する。以降、オノマトペの語形を o 、オノマトペが修飾する動詞を v と表記する。 ov は、「ぎーぎーと降る」のように、 o の直後に v が接続した表現に対応する。このような記法を採用すると、言い換え収集は、与えられた ov に対して、言い換え pv を収集するということになる。以降、 ov を言い換え元表現、 pv を言い換え先表現、 p を言い換え語と呼ぶ。

このような単位で言い換え語を収集する場合、それぞれの o に対してどのような v を選ぶかが問題となる。それぞれの o には、「ぎーぎーと降る」、「ぎーぎーと流れる」のように、それが典型的に修飾する動詞が存在する。そこで、 v として、 o が典型的に修飾する動詞のみを考える。

2.3 言い換えを収集する方法

言い換え元表現 ov において、 o は v を修飾する。同様に、言い換え先表現 pv において、 p は v を修飾する。すなわち、 v を修飾する語の集合を考えたとき、言い換え語 p は、その集合の中に存在するはずである。このような考え方に基づいて、本研究では、 v の修飾語の中から言い換え語 p を選ぶことにより、言い換えを収集するという方針をとる。

3 言い換え収集の概要

前節で定めた方針に基づき、オノマトペの語形 o に対して、次のステップで言い換えを収集する。

Step 1. 動詞リストの作成

オノマトペの語形 o が修飾する典型的な動詞 v を収集する。収集した動詞のリストを $V(o)$ とする。

Step 2. 修飾語リストの作成

各 $v \in V(o)$ に対し、 v を修飾する語のリスト $W(v)$ を作成する。

Step 3. 言い換え先表現の決定

言い換え元表現 ov に対して、 $W(v)$ から言い換えを選択し、言い換え先表現 pv を決定する。

以降の 4~6 節で、それぞれのステップについて説明する。

4 動詞リストの作成

Step 1 では、オノマトペの語形 o が修飾する典型的な動詞 v を収集する。既存のコーパスは、オノマトペの用例の数が少なく、 v の収集には適していない。そこで、動詞の収集には、ウェブを利用する。

まず、 o をクエリとしてウェブ検索を行い、得られた URL のソースページから文字列マッチングを用いて o を含む用例（文）を収集する。次に、各用例を形態素解析し、 o の直後が動詞ならば、その語を収集する。

ウェブの用例から動詞を収集しているため、 o が修飾する動詞として、不適切なものが含まれている可能性がある。これに対処するために、収集した動詞の中で、出現頻度が n 未満の動詞は除去する。今回は、用例の形態素解析には MeCab を、形態素解析用辞書には IPAdic を使用し、 $n = 4$ とした。

次に、収集した動詞の中から、典型的な動詞を選ぶ。収集した動詞の中には、「降る」「降り注ぐ」のように、意味が類似した語（類義語）が存在する。これらの語を修飾するオノマトペの意味は同じ（言い換え語が同じ）であると考え、類義動詞をグループ化する。具

体的には、先頭の漢字が一致する語を類義動詞と認定する。このようにして作成した各グループから、最も出現頻度の高い動詞を選び、動詞リスト $V(o)$ を作成する。

5 修飾語リストの作成

Step 2 では、前節で作成した動詞リスト $V(o)$ に含まれる各動詞 v に対し、 v を修飾する語を収集する。収集する語は形容詞および形容動詞に限定する。まず、コーパスから v を含む用例（文）を収集する。次に、各用例を解析し、 v の直前が形容詞、または形容動詞の連用形である場合、その語を収集する。今回、コーパスは青空文庫 [2]、および、現代日本語書き言葉均衡コーパスを使用した。

6 言い換え先表現の決定

Step 3 では、言い換え元表現 ov の言い換え先表現 pv を決定する。 ov は、オノマトペの語形 o と 4 節で作成した動詞リスト $V(o)$ の各動詞 v から作成する。言い換え語 p は、前節で作成した修飾語リスト $W(v)$ から、次の 2 つのステップにより選択する。

6.1 言い換え候補リスト $P(v, o)$ の作成

まず、修飾語リスト $W(v)$ から、 o の言い換え語となる可能性のある語のみを取り出し、言い換え候補リスト $P(v, o)$ を作成する。

本研究で対象としているのは、「〜と」の語形をとり、用言を修飾する用法である。この用法の用例には、次の用例のように、オノマトペとその言い換え語が隣接して共起する現象がみられる。

(8) 雨がざーざーと激しく降る

(9) 山がかすかにうっすらと見える

一般に、一つの動詞が、ほぼ等価な意味を持つ 2 つの連用修飾語を同時にとることはほとんどない。これに対して、オノマトペの場合は、読み手に感覚的な理解を持たせるために、一種の強調を目的として、このような文が許容される。本研究では、この現象を利用して言い換え候補集合 $P(v, o)$ を作成する。具体的には、オノマトペの語形 o と隣接して共起する語 $w \in W(o)$ のみを、 $P(v, o)$ の要素とする。

隣接して共起するか否かの判定には、ウェブ検索ヒット数を用いる。まず、「ざーざーと激しく」、「激しくざーざーと」のように ow , wo を作成し、これらをクエリしてウェブ検索を行い、検索ヒット数 $h(ow)$, $h(wo)$

を取得する。 $h(ov)$, $h(wo)$ のどちらかが 1 以上であれば、 o と w は隣接して共起するとみなす。このようにして作成した言い換え候補リスト $P(v, o)$ は、次式で表される集合となる。

$$P(v, o) = \{w | w \in W(v), h(ov) \geq 1 \vee h(wo) \geq 1\}$$

6.2 言い換え語 p の選択

次に、前節で作成した言い換え候補リスト $P(v, o)$ の中から、言い換え元表現 ov における o の言い換え語 p を選択する。

言い換え語は次の 2 つの条件を満たす必要がある。

1. 同じ文脈で使用できる
2. 同じ意味を持つ

この 2 つの条件を満たす語を選択するために、本研究では、文脈類似度と意味的類似度を考慮した尺度を使用する。

6.2.1 文脈類似度

文脈類似度 $Sim_c(ov, pv)$ は、 ov と pv の文脈がどれだけ類似しているかを表す。 $Sim_c(ov, pv)$ の計算には、単語ベクトルの余弦を用いる。 $Sim_c(ov, pv)$ は 2 つのステップで求める。

Step 1. 用例集合 T_{ov} , T_{pv} の収集

ov をクエリとし、ウェブ検索を行う。次に、検索結果ページのスニペットから、 ov の用例を最大 300 件収集し、用例集合 T_{ov} を作成する。同様の方法で、 pv の用例集合 T_{pv} を作成する。

Step 2. 文脈類似度 $Sim_c(ov, pv)$ の計算

名詞、動詞、形容詞、形容動詞の全単語集合を $X = \{x_1, \dots, x_N\}$ とし、 N 次元単語ベクトル空間を定義する。このベクトル空間において、 T_{ov} と T_{pv} は、それぞれ以下に示すベクトルとして表現できる

$$\mathbf{q}_{ov} = (f_{ov}(x_1), \dots, f_{ov}(x_N))$$

$$\mathbf{q}_{pv} = (f_{pv}(x_1), \dots, f_{pv}(x_N))$$

ここで、 $f_{ov}(x)$ は単語 x の T_{ov} 中の出現回数を表し、 $f_{pv}(x)$ は単語 x の T_{pv} 中の出現回数を表す。この 2 つのベクトルを用いて、文脈類似度 $Sim_c(ov, pv)$ を次式のように定義する。

$$\begin{aligned} Sim_c(ov, pv) &= \cos(\mathbf{q}_{ov}, \mathbf{q}_{pv}) = \frac{\mathbf{q}_{ov} \cdot \mathbf{q}_{pv}}{|\mathbf{q}_{ov}| |\mathbf{q}_{pv}|} \\ &= \frac{\sum_{i=1}^N f_{ov}(x_i) f_{pv}(x_i)}{\sqrt{\sum_{i=1}^N f_{ov}(x_i)^2} \sqrt{\sum_{i=1}^N f_{pv}(x_i)^2}} \end{aligned}$$

6.2.2 意味的類似度

意味的類似度は、 ov と pv の意味がどれだけ類似しているかを表す。意味類似度 $Sim_s(ov, pv)$ を、ウェブ検索ヒット数を用いて次式で定義する。

$$Sim_s(ov, pv) = h(opv)H(o, p, v)$$

$$H(o, p, v) = \log_2 \frac{h(o, p, v)M^2}{h(o)h(p)h(v)}$$

$h(o, p, v)$ は o , p , v の AND 検索ヒット数、 $h(opv)$ は“ opv ”を検索クエリとした検索ヒット数、 M はウェブページの総数である。今回は、 $M = 1.0 \times 10^{12}$ とした。 $h(opv)$ は、 o , p , v が隣接して共起する頻度であり、このような現象が頻繁に現れる p は、 o と等価な意味を持つ可能性が高い (6.1 節を参照)。 $H(o, p, v)$ は、 o , p , v が共起して出現する頻度と、個別に出現する頻度の比であり、3 つの語の共起の強さを表す。

この意味的類似度は、 $h(\cdot) = 0$ となる場合に、計算ができなくなる。これに対処するために、 $h(opv)$, $h(o, p, v)$, $h(o)$, $h(p)$, $h(v)$ には、それぞれ検索ヒット数に 1 を加えたものを使用する。

6.2.3 言い換え語 p の選択

言い換え語 p の選択には、次の尺度を用いる。

$$F(ov, pv) = Sim_c(ov, pv)Sim_s(ov, pv)$$

最終的に、最大の $F(ov, pv)$ をとる語 p を、言い換え元表現 ov における o の言い換え語と決定する。

7 言い換え語の収集結果と評価

本方法を用いて、編纂中のオノマトペ用法辞書のエントリを対象に、言い換えを収集した。辞書が収録するオノマトペの総数は 927 語であり、そのうち、「〜と」の語形をとり、用言を修飾する用法を持つオノマトペは、723 語である。収集した言い換え語は既に辞書に収録されている。辞書のエントリの例を図 1 に示す。

723 個のオノマトペの語形 o のうち、502 個の o に対して動詞 v が収集され、合計 1113 個の言い換え元表現 ov が作成された。このうち、1081 個の ov に対して、言い換え先表現 pv が収集された。得られた 1081 個の pv に対して、それが ov が言い換えとして適切か否かを人手で判定した。判定結果を表 1 に示す。正しい言い換えを収集した割合は、61.2% (662/1081) となる。本方法で収集した適切な言い換え例を、表 2 に示す。

次に、不適切な言い換え先表現を収集した 419 個と、言い換え先表現を収集できなかった 32 個、合計 451

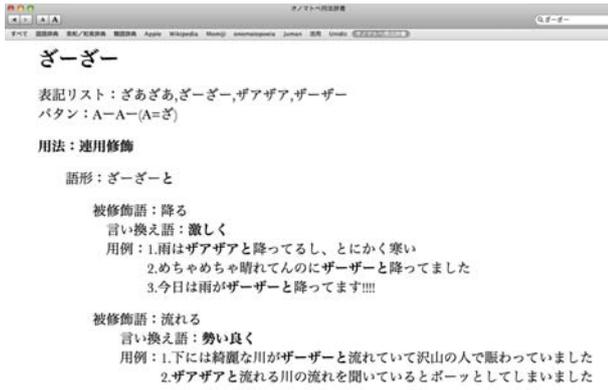


図 1: 辞書のエントリの例

個の言い換え元表現について、適切な言い換え先表現を収集できなかった原因を調査し、4つのカテゴリに分類した。結果を表3示す。

表3において、E1は、言い換え収集のStep 1で、動詞 v として不適切なものを収集したこと起因する誤りである。例えば、オノマトペの語形「ぼいと」に対して、「マフィアぼいと言われる」などの用例から、それが修飾する動詞として「言う」が収集され、言い換え元表現「ぼいと言う」が作成された。

E2は、Step 2で作成した修飾語リスト $W(v)$ の中に、 o の適切な言い換え語 p が存在しなかったことに起因する誤りである。

E3は、修飾語リスト $W(v)$ に、適切な言い換え語 p が含まれていたにも関わらず、Step 3でこの p が選ばれなかったことに起因する誤りである。E3aは、言い換え候補リスト $P(v, o)$ の作成において、 p が言い換え候補から除外されたことに起因する誤りである。一方、E3bは $P(v, o)$ から言い換え語を選択する際に、 p が選ばれなかったことに起因する誤りである。

本方法において、Step 1とStep 2は、言い換え収集の準備に相当し、Step 3で、言い換え元表現 ov の言い換え先表現 pv を決定する。誤り分析の結果から、Step 3の精度 ($W(v)$ が適切な言い換え語 p を含んでいた場合に、適切な言い換え語を選択できた割合) は、80.6%(662/(662+151+8))となり、良好な精度が得られた。

本方法の問題として、適切な言い換え語 p が $W(o)$ に含まれていない ov が多く見られた(表3のE2)。この原因は2つに分けられる。1つは、Step 2で、ある動詞を修飾する語を網羅的に収集できていないことである。この問題は、修飾語の収集元に他の言語資源を追加することで、改善が期待できる。もう1つは、1語で言い換えることが難しいオノマトペが存在するこ

表 1: 言い換えの正否の判定結果

言い換えとして適切	662
言い換えとして不適切	419
合計	1081

表 2: 言い換え例

言い換え元表現	言い換え先表現
あたふたと過ごす	忙しく過ごす
きっぱりと否定する	強く否定する
きっぱりと諦める	潔く諦める
しれつと言う	事も無げに言う
にこにこ笑う	嬉しそうに笑う
にたと笑う	不気味に笑う
にやつと笑う	不敵に笑う
ねちねちと言う	しつこく言う
ぶつと笑う	小さく笑う

表 3: 適切な言い換え先表現を収集できなかった原因

カテゴリ	原因	数
E1	v が不適切	49
E2	$W(v)$ に適切な p が含まれていない	243
E3	a $P(v, o)$ に適切な p が含まれていない	8
	b 適切な p が選択されていない	151
合計		451

とである。本方法は、対象とする o が、 ov に対して、1語で言い換え可能な p を持つか否かの判定を行っていない。この判定は非常に困難であり、今後の研究課題であると言える。

謝辞 本研究では、国立国語研究所が開発した「現代日本語書き言葉均衡コーパス」モニター公開データ(2009年度版)を利用した。ここに感謝の意を表す。

参考文献

- [1] 古武泰樹, 佐藤理史. 2010. 用例に基づくオノマトペ用法辞書の編纂. 言語処理学会第16回年次大会発表論文集, pp.994-997
- [2] 門田裕志. 2007. 青空文庫 全. 青空文庫