

# 辞書定義文を用いた二字漢語の言い換え表現の生成

萩行 正嗣

黒橋 禎夫

京都大学大学院 情報学研究科

{hangyo,kuro}@nlp.kuee.kyoto-u.ac.jp

## 1 はじめに

自然言語の意味処理を行なうためには、意味の基本単位である単語に対して意味情報を与える必要がある。しかし、コストや一貫性の問題から、意味情報を与える語はできるだけ少数の基本的な語にとどめたい。現在、JUMAN<sup>1</sup> 基本語辞書には約 30,000 語の語彙が登録されているが、この語数は意味情報を記載するには多過ぎる。そこで、語の意味を他の語で表現することで意味情報を記載する語数を減らすことを本研究の目標とする。

日本語においては漢字の造語力により多彩な二字漢語が存在し、JUMAN 基本語辞書においても、半数の約 15,000 語が二字漢語である。二字漢語の中には、個々の語構成漢字の意味を考慮することで、言い換え表現を生成できるものがあり、そうすれば、その語の意味情報を記載する必要がなくなる。例えば、「高山」が「高い山」と表現することができれば、「高山」に意味情報を記載する必要がなくなる。

このように、本手法ではまず二字漢語に着目し、その語構成漢字と辞書定義文から、言い換え表現を自動で生成する。またその過程で逐次的に語構成漢字の言い換え表現を獲得。

## 2 語構成漢字を利用した言い換え

ある語構成漢字  $a$  と  $b$  から成る二字漢語  $ab$  の意味を  $sem(ab)$  とすれば、そこには多くの場合、

$$sem(ab) = sem(a) + sem(b) + \alpha \quad (1)$$

の関係が成り立つと考えられる [2]。これは、語構成漢字が組み合わされて単語化される際に、それぞれの漢字の意味が組合わされ、さらに意味の付加が行なわれることを表わす。例えば、以下のような二字漢語とその辞書定義文を考える。本論文において、辞書定義文中の二重下線は語構成漢字との対応を表わす。

1. 高山: 高い 山。
2. 花見: 花、とくにサクラの 花 を 見て 楽しむ こと。
3. 指貫: 裁縫する時、針の頭を押すために 指 にはめる、皮や金属で作った輪。

「高山」では語構成漢字から推測される意味と大きく変わらず、 $\alpha$  の意味的大きさはほぼ 0 といえる。「花見」では単純に語構成漢字から推測される「花を見る」に「サクラ」や「楽しむ」といった要素が加わっており、これらが  $\alpha$  となる。「指貫」においては、語構成漢字から推測されるような意味とはまったく異なっており、式 (1) のような関係が成り立っているとは言い難い。さらに、「硝子 (がらす)」のように当て字であるもの、「八朔 (はっさく)」のように種別名であるものなど語構成が構成的でないものも存在する。

上記のような例から、式 (1) が成り立つような語においては、辞書定義文中に対応する表現が存在すると仮定することができ、これを利用することで二字漢語の言い換え表現が生成できると考えられる。また、辞書定義文を利用することで、語構成が構成的なものとそうでないものを区別することができ、語構成から意味をある程度推測できる語のみを言い換えるの対象とすることができる。

例えば以下のような語とその辞書定義文を考える。

- 水鳥: 川や湖の 水辺 にすむ 鳥。
- 駄本: 内容に価値がない、くだらない 本。

「水鳥」の場合には、「水辺」が「水」に対応し、「鳥」が「鳥」に対応している。さらに、「水辺」と「鳥」の係り受け関係を利用することで、辞書定義文から「水辺にすむ鳥」という言い換えを生成する。以降の例では、二重下線が、語構成漢字と辞書定義文の対応がとれている部分、二重下線を含む、下線部が生成された言い換え表現を示す。また、語構成漢字は単純に語構成漢字を単語化したものとは限らない。例えば、「駄文」では「駄」は「くだらない」という意味を表す (本

<sup>1</sup><http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>

論文では、このような漢語特有の語構成漢字の意味を語構成漢字言い換えパターンと呼ぶ)。そこで、語構成漢字の言い換えパターンについての学習を平行して行なう手法も提案する。

### 3 提案手法

#### 3.1 語構成漢字と辞書定義文の対応付け

以下のような手順で言い換えを生成する。(例として「水鳥」から「水辺にすむ鳥」を生成する。)

1. 辞書定義文を JUMAN・KNP<sup>2</sup>を用いて形態素解析・構文解析する。  
例)  
川 や<P>  
湖 の<P> PARA  
水辺 に  
すむ  
鳥 。
2. 自立語 (以下では定義文構成語とよぶ) のうち各語構成漢字に対応するものを探す。  
例) 川/湖/水辺/すむ/鳥
3. 各語構成漢字に対応する定義文構成語があれば、そこから辞書定義文の主辞に到達する係り受けを言い換え表現とする。  
例) 水辺にすむ鳥

各定義文構成語が語構成漢字と対応するか否かは、以下の三つの基準を利用した。

1. 定義文構成語に語構成漢字が含まれている。JUMAN 辞書に含まれる情報を利用することで、表記のバリエーションも考慮する。  
例) 水 水辺  
視 見る (異表記に「視る」が含まれる。)
2. 柴田らが構築した同義表現辞書 [1] で同義語とされる語に語構成漢字が含まれている。  
例) 陶 焼き物 (陶器が同義表現である)  
鈴 ベル (呼び鈴が同義表現)
3. ブートストラップにより獲得された語構成漢字の言い換えパターンである (詳細は次節で述べる)。  
例) 狡 悪賢い  
卓 テーブル

#### 3.2 ブートストラップによる語構成漢字の言い換えパターンの獲得

単純に辞書の見出しや語の同義語の知識を用いるだけでは、言い換えパターンの獲得は不十分である。そこでブートストラップにより言い換えパターンの獲得を行なう。

例えば、「駄文:くだらない文章」と「駄句:くだらない俳句」という辞書定義文を考える。この場合、「文」と「文章」、「句」と「俳句」が対応する。一方で、二つの辞書定義文を比較した場合には、「駄」が「くだらない」に対応するということが推測される。このようなパターンを辞書全体から学習することで、新たな言い換えパターンの獲得を行なう。これは以下の手順に従う。

1. 二字漢語の言い換え生成を行なう。
2. 未対応の語構成漢字と定義文構成語のすべての組み合わせを言い換えパターン候補として保持する。  
例) 「悪疾:たちの悪い病気」から  
「疾 たち」「疾 病気」  
「疾 風:速く吹く風」から  
「疾 速く」「疾 吹く」  
「痼疾:不治の病気」から  
「疾 不治」「疾 病気」  
「痼 不治」「痼 病気」
3. 個々の語構成漢字  $C_i$  について、 $k_{ij}$  をその  $C_i$  の言い換えパターン候補、 $|k_{ij}|$  を  $k_{ij}$  が出現する見出し語の異なり数とし、 $k_{ij}$  の語構成漢字の言い換えパターンらしさ  $score(k_{ij})$  を以下のように計算する。

$$score(k_{ij}) = \frac{(|k_{ij}| - 1)^2}{\sum_j (|k_{ij}| - 1)^2} \quad (2)$$

- 例) |疾 病気| = 6、|疾 不治| = 2、  
|疾 速く| = 2、|疾 たち| = 1、...  
 $score(疾 病気) = 0.93$ 、  
 $score(疾 不治) = 0.04$ 、...
4. 言い換えパターン候補のうち、 $|k_{ij}|$ 、 $score(k_{ij})$  が閾値より大きいものを語構成漢字の言い換えパターンとする。例) 「疾 病気」を獲得する
  5. 新たに獲得された言い換えパターンを加えて 1. ~ 4. を繰り返す。

<sup>2</sup><http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html>

表 2: 獲得された語構成漢字言い換えパターンの例

	名詞	$score(k_{ij})$	$ k_{ij} $	正誤	動詞	$score(k_{ij})$	$ k_{ij} $	正誤	その他	$score(k_{ij})$	$ k_{ij} $	正誤
1	頸 首	1.00	8		怨 恨む	1.00	11		冗 無駄だ	1.00	7	
2	患 病気	1.00	6		歿 死ぬ	1.00	6		欧 ヨーロッパ	1.00	5	
3	刹 寺	1.00	6		秀 優れる	1.00	6		永 長い	1.00	5	
4	壘 砦	1.00	6		歆 喜ぶ	1.00	5		翌 次の	1.00	5	
5	艇 ボート	1.00	5		思 考える	1.00	5		爽 さわやかだ	1.00	4	
6	賓 客	1.00	4		飯 召す	1.00	5	×	効 喜々たる	1.00	4	
7	燭 灯し火	1.00	4		情 怠ける	1.00	5		狡 悪賢い	1.00	4	
8	書 手紙	0.99	17		ト 占う	1.00	5		各 それぞれ	1.00	4	
9	他 外	0.98	8		鍊 鍛える	1.00	4		未 まだ	0.99	21	
10	辞 言葉	0.96	21		漏 抜け落ちる	1.00	4		速 素早い	0.96	6	
11	界 社会	0.96	6		贅 贅める	1.00	4		眼 目	0.94	13	
12	壇 社会	0.96	6		氷 凍る	1.00	4		佳 良い	0.94	5	
13	尿 小便	0.96	8		逝 死ぬ	1.00	4		鄙 卑しい	0.94	5	
14	期 季節	0.96	8		否 言う	0.99	12	×	乎 様	0.92	6	
15	誠 真心	0.94	5		報 知らせる	0.97	19		好 良い	0.90	11	
16	撃 敵	0.94	5	×	塊 固まる	0.96	9		寡 少ない	0.90	4	
17	色 様子	0.94	5		名 優れる	0.96	20		陋 狭い	0.90	4	
18	功 手柄	0.94	5		誤 間違う	0.94	7		偵 コッソリ	0.90	4	×
19	疾 病気	0.92	6		英 優れる	0.94	8		劇 激しい	0.89	8	
20	河 川	0.92	8		悦 喜ぶ	0.94	5		巨 大きな	0.88	12	

表 1: 獲得された語構成漢字言い換えパターンの数

反復回数	獲得数
1	267
2	23
3	2
4	1
5	0

## 4 実験と評価

岩波国語辞典に含まれる二字漢語 26,157 語 (表記の揺れは別の語として数えた) を利用して二字漢語の言い換え生成を行った。辞書定義文は各小見出しの一文目のみを使用した。これは二文目以降を利用すると、補足的な説明を言い換えとして生成してしまう場合が見られたためである。また、JUMAN 辞書の基本語に含まれる二字漢語 15,523 語のうち、岩波国語辞典に項目が存在するものは 13,281 語を評価対象とした。 $score(k_{ij})$  の閾値は 0.5、 $|k_{ij}|$  の閾値は 3 とした。

### 4.1 獲得された語構成漢字言い換えパターン

ブートストラップを用いて獲得された語構成漢字の言い換えパターンは反復により表 1 のように変化し、5 回目以降は変化しなくなった。1 回目で獲得されたパターンの例を各品詞ごとに  $score(k_{ij})$  の高い順に表 2 に示す。また、 $|k_{ij}|$  の値、人手による正誤の評価も合わせて示す。表 2 から分かるように、各品詞ともに高い精度で獲得することができた。

また、これとは別に無作為に 50 個のパターンを人手で確認したところ、46 個は正しい言い換えパターンであった。誤っていたものは、「箋 書く」のように本来獲得したい部分 (「紙」) と修飾関係にあるようなものであった。

一方で以下のようなパターンは獲得できなかった。

1. 多義のため個々の言い換えパターンの  $score(k_{ij})$  が低くなったもの  
例) 「屋 家」「屋 店」
2. 語ではなく句の重なりが多く、 $score(k_{ij})$  が低くなったもの  
例) 「漁 魚」「漁 とる」
3.  $|k_{ij}|$  が閾値に満たなかったもの  
例) |房 部屋| = 2 であった。

1、3 は、定義文の主辞との近さや *idf* などを用いて語に重み付けをすることで、確実な語義を選択することができると考えられる。2 は語ではなく句も獲得パターンに含むことで正しいパターンを獲得できると考えられる。

### 4.2 二字漢語の言い換え

岩波国語辞典の二字漢語 26,157 語と JUMAN 基本語彙中の評価対象の二字漢語 13,281 語のうち、生成された二字漢語の言い換えの数は表 3 のようになった。反復回数とはそれぞれ表 1 で示した獲得パターンを利用した場合を示す。この結果からブートストラップを用いて獲得された言い換えパターンを利用すること

で、生成可能な言い換えを 10% 程度増加させることができた。

評価対象語から無作為に抜き出した二字漢語 100 語は、語の構成と提案手法による生成の可否から表 4 のように分類された。これから、JUMAN 基本語彙中において約 1/3 は語構成からは言い換への生成が困難であり、残りの約 2/3 は本手法の対象となる語構成から言い換えが生成可能なものであることが分かる。また、2/3 のうちの半分、全体の約 1/3 が実際に生成できたものであり、これは表 3 の結果と大体一致している。

正しく生成されたものは以下のようなものがあつた。以下で二重下線部の添字は 3.1 節のどの手法で語構成漢字との対応がとられたかを示す。

例) 骨折:体の 骨<sub>1</sub> を 折る<sub>1</sub> こと。  
 凝血:体外に流れ出て 固まった<sub>2</sub> 血<sub>1</sub>。  
 寵臣:寵愛<sub>1</sub> を受けている 家来<sub>2</sub>。  
 英知:深遠な道理を知りうる すぐれた<sub>3</sub> 知恵<sub>1</sub>。

正しく生成されなかったものは以下のようなもので、これらは必要な副詞的、形容詞的、格要素が欠落していた。以下の例で波線部は欠落した要素を示す。

例) 船室:船<sub>1</sub> で乗客の使用に あてる 部屋<sub>2</sub>。  
 金言:金<sub>1</sub> のように、価値の 高い 言葉<sub>1</sub>。  
 取材:作品や報道の 材料<sub>1</sub> をある物事・人から取る<sub>1</sub> こと。

語構成が構成的なもので生成されなかったものは以下のようなものであつた。

例) 小陰:ちょっとした 物陰<sub>1</sub>。  
 勅許:天子の ゆるし<sub>1</sub>。  
 姫松:小さな 松<sub>1</sub>。  
 短所:劣っている ところ<sub>1</sub>。  
 右折:進んでいる車・人などが 右<sub>1</sub> に曲がること。  
 裏話:一般には知られていない、うちわばなし<sub>1</sub>。

一方、語構成からは生成できないものには以下のようなものがあつた。

例) 経験:実際に見たり聞いたりおこなったりして、まだしたことがない状態から、したことがあるという状態に移ること。  
 可能:(まだ実現していないが) 実現の余地があること。  
 実際:単に頭の中で考えるのでなく、われわれの生活する場に臨んでのものであること。  
 矛盾:前に言ったこととあとに言ったことが一致しないこと。

表 3: 生成された言い換え

反復回数	岩波	JUMAN
0	8177	4817
1	9336	5362
2	9418	5394
3	9426	5396
4	9430	5399
辞書全体	26157	13281

表 4: 二字漢語の展開可能性による分類

		自動処理	
		生成される	生成されない
語構成	構成的	正解 28:不適 4	39
	非構成的	0	29

次元:変化するものの状態が個の独立の変数でとらえられる時、そのをこの場合の次元と言う。  
 百計:あらゆる、はかりごと。  
 南無:絶対的な信仰を表すために唱える語。  
 加里:「カリウム」の略。

これらのなかには、「経験」などのように頻度が高い語が存在する。このような語が今後意味情報の付与を行なっていく対象となる。

## 5 おわりに

本論文では、二字漢語の語構成漢字を利用した辞書定義文からの言い換え表現生成を提案した。その結果、JUMAN 基本語彙の二字漢語のうち約 1/3 について言い換を生成することができた。さらにその過程で語構成漢字の言い換えパターンを獲得できた。

今後は、本手法では生成できなかった二字漢語の言い換の生成や語構成が構成的ではない語の意味表現について検討したい。

## 参考文献

- [1] Tomohide Shibata, Michitaka Odani, Jun Harashima, and Takashi Onishi. Syngraph : A flexible matching method based on synonymous expression extraction from an ordinary dictionary and a web corpus. In *Proceedings of IJC-NLP2008*, pp. 787–792, 2008.
- [2] 斎藤倫明. 語彙論的語構成論. ひつじ書房, 2004.