国語辞典とコーパスを用いた用言の言い換え規則の学習

鍛治伸裕† 河原大輔† 黒橋禎夫‡* 佐藤理史†
†京都大学大学院情報学研究科 ‡東京大学大学院情報理工学系研究科
*科技団さきがけ研究21

{kaji, sato}@pine.kuee.kyoto-u.ac.jp
{kawahara, kuro}@kc.t.u-tokyo.ac.jp

1 はじめに

本稿では，国語辞典を用いて用言を言い換える手法について述べる．言い換えの研究には二つの意義がある．まず一つめは，言い換え処理の研究が，ユーザが望む形態にテキストを自動加工するアプリケーションの開発につながることである．近年の計算機の急速な普及によって，今後はこのようなアプリケーションへのニーズが高まると考えられる．二つめは言い換え処理が，自然言語処理における同義異表記の問題解決につながることである．例えば検索表現を言い換えることによって，テキスト検索の再現率を向上させることが期待できる．

用言の言い換えや国語辞典を用いた言い換えには，様々な先行研究が存在する [1] [2]．しかし，定義文から見出し語の同等句を抽出する処理や，用言が持つ格の表層格を変換する処理は，用言の言い換えに重要な処理であるにも関わらず十分な研究が行われていない．また，語義の曖昧性解消を扱った言い換えの研究も殆んどない．

提案手法では，言い換えられる用言とそれを言い換えた用言の格フレームを対応付けることによって上記の処理を実現する．

2 国語辞典とコーパスを用いた用言の言い換え規則の学習

2.1 国語辞典に基づく用言の言い換え

多くの場合，国語辞典の定義文の文末にある用言(主辞と呼ぶ)と，それに副詞的にかかる文節は見出し語の同等句を形成している．そのため，定義文からそれら二つを取り出せば，見出し語を言い換えることができる．例えば「要求」の定義文は次のようになっている．

要求 こうしてほしいと 強く 求める こと

そこで，主辞「求める」とそれに副詞的に係る「強く」を用いて，(1) のような言い換えができる．

(1) 中止を 要求する ⇒ 中止を 強く求める

2.2 用言の言い換えに必要な処理

しかし実際には，上記のような単純な手法では不十分なことが多い．用言を言い換えるためには，以下のような処理を実現させなくてはならない．

語義の曖昧解消 多義語を言い換える場合，その語義の曖昧性解消が必要となる．例えば「しのぐ」は下のような二つの定義文を持っているので，(2) のように言い換えるには語義の曖昧性を解消しなくてはならない．

しのぐ **1** 耐え忍ぶこと
2 優れていること

(2) 苦境をしのぐ ⇒ 苦境を耐え忍ぶ

同等句の決定 上記のように，「副詞+主辞」が見出し語の同等句を形成している場合が多い．しかし実際には，主辞が持つ格を同等句に含めなくてはならない場合がある．例えば「体得」の同等句は，「身に」を含めた「身につける」である．なお本稿では，用言にかかる「格要素+表層格」を格と呼んでいる．

体得 知識やわざを 身につける こと

表層格の変換 用言を言い換える時，大抵はその用言が持つ格の表層格は変化しない．しかし下のように，表層格を変換しなくてはならないケースもある．

(3) 音楽 を 愛する ⇒ 音楽 が 好きだ

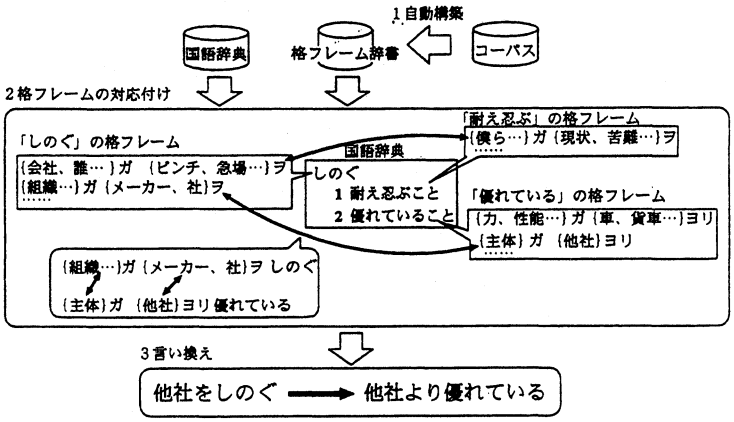

図 1: 格フレームの対応付けに基づく用言の言い換え

2.3 格フレームの対応付けに基づく言い換え

上記のような処理を実現するには，見出し語と主辞がどのような格を持っていて，それらがどう対応しているのかという情報が必要である．しかし下のように，定義文にはそれらの情報が必ずしも明記されていない．

しのぐ　1 耐え忍ぶこと
　　2 優れていること

そのため，定義文だけを利用して (4) のような言い換えを行うことは難しい．

(4) 他社をしのぐ ⇒ 他社より優れている

そこで，見出し語と主辞が持つ格フレームをあらかじめ学習，対応付けすることによって，上記の問題を解決する方法を提案する．提案手法による用言の言い換えの流れは，以下のようになる (図 1)．

1. 河原らの手法 [3] を用いて，コーパスから格フレーム辞書を自動構築する
2. 国語辞典をもとに，学習された格フレーム同士を対応付ける
3. 入力文と類似する見出し語格フレームを選択し，その対応付けに基づいて言い換える

3 格フレームの対応付け

本手法では，まず定義文や例文を利用して，見出し語格フレームの対応先の候補となる主辞格フレームを絞り込んでおく．そして，見出し語格フレームと主辞格フレームの間の類似度計算に基づいて，それぞれの見出し語格フレームの対応先を決定する．

3.1 定義文を用いた絞り込み

主辞格フレームは，その一部だけが定義文と同じ用法を持っている．例えば「落ち延びる」は下のような定義文を持ち，その主辞は「逃げる」である．この時「{ 熱 } ガ { 外部，宇宙 } ニ 逃げる」という格フレームは，定義文での「逃げる」とは用法が異なる．

落ち延びる　戦いに敗れた者が、遠くまで 逃げる

定義文と異なる用法を持つ格フレームは見出し語格フレームの対応先にならないので，定義文を利用してそのような主辞格フレームを取り除き，見出し語格フレームの対応先候補を絞り込む．

主辞の直前格が必須格　定義文中で主辞の直前格 (主辞直前格と呼ぶ) が必須格の時は，主辞直前格が主辞の用法を強く限定しているため，それを用いて絞り込みを行う．ここではガ格，ヲ格，ニ格を必須格とする．

絞り込みの方法は二つのステップからなる．まず主辞格フレームの中から，直前格の表層格が主辞直前格と同じものを取り出す．次にその中から，直前格の格要素が表 1 のような制約を満たす主辞格フレームだけを選ぶ．格要素への制約は，主辞直前格のタイプによって三つに分かれていて，それぞれ，主辞直前格の格要素と全く同じ，類似度が 0．8 以上，意味属性が同じとなっている．ただし，一般概念語とは「人々」「場所」などのように一般概念を表す単語のことである．

表 1: 絞り込みの方法

主辞直前格のタイプ	格要素への制約
格要素が単語一つ	全く同じ
格要素が並列構造	類似度が 0．8 以上
格要素が一般概念語	同じ意味属性

主辞の直前格が任意格　一方，主辞直前格が任意格の場合，それによる強い絞り込はできないので，定義文全体との類似度が 0.8 以上の主辞格フレームだけを選ぶ．類似度は，共通するガ格，ヲ格，ニ格の類似度平均とした．ただし，格の類似度は以下のように定義する．

$$\text{格の類似度} = max\{sim(e_{def}, e)|e \in C\}$$

ここで e_{def} は主辞が持つ格要素であり，e は主辞格フレームの格 C に含まれる用例である．また $sim(e1, e2)$ は日本語語彙大系に基づいて計算した単語 $e1, e2$ の類似度で，1 に正規化されている．

3.2　例文を用いた絞り込み

定義文に例文が記載されている場合，例文と似た見出し語格フレームの対応先は，その定義文の主辞格フレームのどれかであると考えられる．そこで，定義文と主辞格フレームの場合と同様の類似度を，例文と見出し語格フレームの間にも定義する．そして，例文との類似度が 0.8 以上である見出し語格フレームの対応先を，その定義文の主辞格フレームに限定する．

3.3　見出し語格フレームと主辞格フレームの対応付け

上記のようにして対応先を絞り込んだ後，見出し語格フレームと主辞格フレームの間に決めた類似度に基づいて，各見出し語格フレームの対応先を決定する．ただし「同等句に含める格」「表層格が変化する格」に関する優先規則をあらかじめ定めておき，これに違反するような対応付けは許さないものとする．

3.3.1　見出し語格フレームと主辞格フレームの類似度

見出し語格フレームと主辞格フレームの類似度は，「格の類似度の重み付け平均」と「格の一致度」の積とする．

以下では，二つの格フレーム F_1, F_2 において，格 $C_{11} \ldots C_{1l}$ が格 $C_{21} \ldots C_{2l}$ に対応していて，F_2 の格 C_{2n+1} は見出し語の同等句に含まれるとする．

$$\begin{array}{l} F_1\text{: } C_{11},\ C_{12},\ \ldots\ C_{1l},\ \ldots\ C_{1m} \\ \qquad\ \updownarrow\ \ \updownarrow\qquad\ \updownarrow \\ F_2\text{: } C_{21},\ C_{22},\ \ldots\ C_{2l},\ \ldots\ C_{2n}\ (C_{2n+1}) \end{array}$$

格の類似度と重み　F_1, F_2 の格 C_{1i}, C_{2i} 間の類似度 $CaseSim(C_{1i}, C_{2i})$ は，以下のように定義した．

$$CaseSim(C_{1i}, C_{2i}) = \frac{\sum_{e_1 \in C_{1i}} |e_1| \cdot max\{sim(e_1, e_2)|e_2 \in C_{2i}\}}{\sum_{e_1 \in C_{1i}} |e_1|}$$

頻出する格は重要度が高いと考えて，格の類似度に加える重みは，その格に出現する用例数の積の平方根とした．格の類似度の重み付け平均 $WeightedCaseSim(F_1, F_2)$ の計算式は，以下のようになる．

$$WeightedCaseSim(F_1, F_2) = \frac{\sum_{i=1}^{l} \sqrt{|C_{1i}||C_{2i}|} \cdot CaseSim(C_{1i}, C_{2i})}{\sum_{i=1}^{l} \sqrt{|C_{1i}||C_{2i}|}}$$

ただし，e_1, e_2 は C_{1i}, C_{2i} が持つ用例で，$|e_1|$ はその頻度である．また $|C_{1i}|, |C_{2i}|$ は，格 C_{1i}, C_{2i} に含まれる用例数である．

格の一致度　F_1, F_2 について「対応付けられた格の用例数/全格用例数」を求め，それらの積の平方根を格の一致度 $Alignment(F_1, F_2)$ とする．計算式は以下のようになる．

$$Alignment(F_1, F_2) = \sqrt{\frac{\sum_{i=1}^{l} |C_{1i}|}{\sum_{i=1}^{m} |C_{1i}|} \times \frac{\sum_{i=1}^{l} |C_{2i}|}{\sum_{i=1}^{n} |C_{2i}|}}$$

格フレームの類似度　以上より，格フレームの類似度は以下のようになる．

$$\text{格フレームの類似度} = WeightedCaseSim(F_1, F_2) \times Alignment(F_1, F_2)$$

3.3.2　同等句に含める格に関する優先規則

定義文の主辞が「する」「ある」などの意味が弱い用言なら，その直前格は必ず同等句に含めるものとする．

逆に，定義文の主辞の直前格以外は同等句に含める場合を考えない．また主辞の直前格であっても，その格要素が一般概念語である場合，直前格が並列構造になっている場合，表層格がガ格，ヲ格，ニ格以外の場合も同等句に含めない．

表 2: 語義の曖昧性解消の精度

	成功	失敗	精度
ベースライン	60	55	52 %
提案手法	82	33	71 %

表 3: 語義の曖昧性がない用言，又は曖昧性解消に成功した用言の言い換え精度

	成功	失敗	精度
ベースライン	163	24	87 %
提案手法	170	17	90 %

表 4: 実験文全体の言い換え精度

	成功	失敗	精度
ベースライン	147	73	66 %
提案手法	170	50	77 %

3.3.3 表層格が変化する格に関する優先規則

任意的な格は，用言が変わっても表層格が変わらないことが多い．そのため，任意的な格は表層格が異なる格とは対応付けない．ただしここでは，格フレームでの出現頻度が低い格，又は，ガ格，ヲ格，ニ格，ト格，ヨリ格，カラ格，マデ格以外の格を任意的な格とする．

4 実験と考察

例解小学国語辞典と，毎日新聞と日経新聞の計 20 年分から自動構築した格フレーム辞書を用いて格フレームの対応付けを行い，新明解国語辞典に記載されている例文 (220 文) に含まれる用言を言い換える実験を行った．ただし「使う」「作る」などの基本的な用言は，定義文による言い換えは難しく工学的な意味も少ない．このような用言は定義文に頻出すると考え，定義文に頻出する形態素の上位 2000 に含まれる用言は実験対象から外している．

実験の結果を表 2，3，4 に示す．表 2 は，語義に曖昧性がある用言を含む 115 文を対象として，語義の曖昧性解消の精度を求めたものである．ベースラインは，先頭の定義文を選択するという方法を用いた．表 3 は，語義の曖昧性がない用言，又は語義の曖昧性解消に成功した用言の言い換え精度である．ここでのベースラインは，「主辞が「する」「ある」の場合だけ，直前格を同等句に含める」「表層格は変化させない」という方法を用いた．表 4 は実験文全体の言い換え精度である．ここで比較するベースラインは，上記二つの手法を組み合わせたものである．

全ての精度でベースラインを上回っており，提案手法は有効であるといえる．表 5 に，入力文を正しく言い換えることができた例を示す．失敗の最大の原因は，格フレームのデータスパースであった．下のように定義文は，見出し語を無理に定義しようとして，不自然で一般的には用いられない表現になることがある．

ぶら下がる　ぶらりと下がる
例文) 鉄棒にぶら下がる

このような定義文の主辞格フレームをコーパスから自動学習することは難しい．

表 5: 成功例

攻略	**1** 敵の陣地や城をうばうこと
	2 敵を攻めて，負かすこと
(5)	横綱を攻略する ⇒ 横綱を負かす
体得	知識やわざを身につける事
(6)	こつを体得する ⇒ こつを身につける
遠ざける	**1** 遠くへはなれさせる
	2 つきあわなくする
(7)	悪友を遠ざける ⇒ 悪友とつきあわなくする
鳴り響く	**1** 鳴る音が，広く聞こえる
	2 評判が知れ渡る
(8)	ベルが鳴り響く ⇒ ベルの音が広く聞こえる

5 まとめ

本稿では，格フレームの対応付けに基づいて用言を言い換える手法を提案した．これによって，語義の曖昧性解消，定義文からの同等句抽出，表層格の変換など，従来手法では困難であった処理が実現できる．また提案手法は，国語辞典以外のリソースにも適用可能であることから，用言の言い換え一般に利用できる手法であると考えている．

参考文献

[1] 近藤恵子，佐藤理史，奥村学: 「サ変名詞 + する」から動詞相当句への言い換え．情報処理学会論文誌, Vol.40, No.11, 1999.

[2] 近藤恵子，佐藤理史，奥村学: 格変換による単文の言い換え．情報処理学会論文誌, Vol.42, No.03, 2001.

[3] 河原大輔，黒橋禎夫: 用言と直前の格要素の組を単位とする格フレームの自動構築．自然言語処理, Vol.9, No.1, 2002.