

## 既存の和語動詞の格フレームを利用した サ変動詞の格フレーム獲得

橋本順子      峯 恒憲      雨宮真人

九州大学大学院 システム情報科学研究科 知能システム学専攻

〒 816 福岡県春日市春日公園 6-1

Tel:092-583-7615

E-mail:{junko,mine,amamiya}@al.is.kyushu-u.ac.jp

あらまし

従来、格フレーム獲得は大量の用例を用い、統計的手法を用いてデータを処理することによって行なわれてきたが、サ変動詞の場合、数の多さと、新語・造語の多さから大量の用例を集めることは困難である。また、サ変動詞には、漢字や和語動詞の連用形などの要素から作られるものが多く、多くの場合、和語動詞に言い替えが可能である。そこで本稿では、サ変動詞を構成する漢字から言い替え可能な和語動詞を抽出し、その和語動詞の格フレームを利用してサ変動詞の格フレームを獲得する方法を提案する。また獲得した格フレームを、タグなしの用例を用いて改良する方法を示す。

キーワード      格フレーム, サ変動詞, コーパス, 知識獲得

## Case Frame Acquisition of SAHEN Verbs by Using Precreated Case Frames of Japanese Basic Verbs

Junko HASHIMOTO      Tsunenori MINE      Makoto AMAMIYA

Dept. of Intelligent Systems, Graduate School of Information Science  
and Electrical Engineering, Kyushu University

Kasuga Kouen , Kasuga-shi , Fukuoka 618, JAPAN

Tel:092-583-7615

E-mail:{junko,mine,amamiya}@al.is.kyushu-u.ac.jp

Abstract

A large amount of sample sentences are required in order for case frame acquisition. However, collecting sentences that include Japanese SAHEN verb is hard work because the number of SAHEN verbs are huge and lots of coined words are created as such SAHEN verbs.

Most SAHEN verbs consist of several elements such as a noun word, an adjective word, and a Japanese basic verb called a WAGO verb. Taking account of this feature, we assume that a case frame of a SAHEN verb is similar to one of a WAGO verb that is a component of the SAHEN verb.

This paper proposes a method to acquire case frames of SAHEN verbs using case frames of WAGO verbs that have already obtained before. Experimental results show the effectiveness of our method.

key words      Case Frame, Sahen Verb, Corpus, Knowledge Acquisition

# 1 はじめに

自然言語処理において、意味解析は困難な問題であるが、格フレーム情報を用い、動詞と格の関係を限定することで、構文解析結果を絞り、ある程度意味的に整合性のある文章を得ることができる。そのため、より精度の良い、大規模な格フレームの獲得が望まれている。

格フレームを機械的に獲得する場合、通常、コーパスなど形態素情報を付加（タグ付け）されたデータを大量に用意し、統計的手法などを用いてデータを処理する方法が採られる。しかし、大量の、しかもタグ付けされたデータを用意するには非常に手間と時間が必要であり、特にサ変動詞のように、種類が多く、新語・造語が多い場合、そのようなデータをあらかじめ集めておくことは困難である。

このような特徴を持つサ変動詞を分析すると、より細かな要素から構成されていることが多いことが分かる。例えば、サ変動詞「登山する」は、「登」と「山」からなり、それぞれが「登る」「山」を表している。このようにサ変動詞では、構成する漢字が意味に密接に結び付いている [1, 2]。また、多くの場合、和語動詞を用いて言い替えることができる。

これらの特徴から、既に作られている和語動詞の格フレームをサ変動詞の格フレームとして利用することが考えられる。例えば、「登山する」の格フレームとして、「登る」の格フレームを利用する。和語動詞の格フレームは基本的な語彙からなるため、サ変動詞に比べ獲得がそう困難ではない。

そこで、本稿では、既存の和語動詞の格フレームを利用してサ変動詞の格フレームを獲得する方法を提案し、その有効性を検討する。またタグなしの用例を利用して、獲得した格フレームを改良する方法についても述べる。

## 2 格フレームの類似度計算

### 2.1 格フレーム $H$

本研究で獲得したサ変動詞の格フレームの構造は、元となる和語動詞の格フレームに依存する。

ここでは、元となる和語動詞の格フレームとして、EDR電子化辞書 [3] の日本語コーパス・日本語共起辞書の共起データから東の手法 [4, 5] によって獲得された格フレームのうち、和語動詞に関するものを利用した。以降、東の獲得した格フレームを格フレーム  $H$  と呼ぶ。

### 2.1.1 構造

格フレーム  $H$  は、日本語コーパス・日本語共起辞書の共起データから、動詞+助詞+名詞概念集合の組を抜き出し、それらの各組を動詞語幹・動詞概念をキーとして、助詞ごとに集合的に加算することによって獲得される。その構造は次のようになる。

動詞表記	
動詞概念	
助詞 <sub>1</sub>	名詞概念集合 <sub>1</sub>
助詞 <sub>2</sub>	名詞概念集合 <sub>2</sub>
⋮	

この手法では、共起データから格フレームを自動獲得することができる。また、新たな共起データが得られた場合に、既に作成した格フレームに情報を付加することが容易である。

### 2.1.2 類似度計算の定義

ある文が格フレーム  $H$  のどの格フレームに分類されるかを調べることによって、動詞概念を推定できる。その際、それぞれの格フレームと文との間の類似度の定義を、以下のように格フレームを構成する最小単位である概念間の類似度を基に、次のように定義する。

概念間の類似度

概念  $N_i, N_j$  の類似度

$$\text{sim}(N_i, N_j) = \frac{1}{k} \times Dmax_{ijk} \times \left( \alpha \frac{CS_{ijk}}{S_{ik}} + \beta \frac{CS_{ijk}}{S_{jk}} \right)$$

$$0 \leq \text{sim}(N_i, N_j) \leq 1$$

$$\alpha + \beta = 1$$

$S_{ik}$ :  $N_i$  の上位概念のうち、深さが  $k$  以下のものの数

$CS_{ijk}$ :  $N_i$  と  $N_j$  の共通上位概念のうち、深さが  $k$  以下のものの数

$Dmax_{ijk}$ :  $CS_{ijk}$  のうちの最大の深さ

概念集合間の類似度

概念集合  $Set_i, Set_j$  の類似度

$$\text{sim}(Set_i, Set_j) = \max_{N_i \in Set_i, N_j \in Set_j} \text{sim}(N_i, N_j)$$

スロット (助詞・名詞概念集合の組) の間の類似度  
スロット  $Slot_i, Slot_j$  の類似度

$$\text{sim}(Slot_i, Slot_j) = \begin{cases} \text{sim}(Set_{Slot_i}, Set_{Slot_j}) & \text{if } Case_{Slot_i} = Case_{Slot_j} \\ 0 & \text{otherwise} \end{cases}$$

$Set_{Slot_i}$ :  $Slot_i$  の概念集合

$Case_{Slot_i}$ :  $Slot_i$  の助詞

格フレーム (もしくは文) 間の類似度  
格フレーム  $CF_i, CF_j (i \neq j)$  の類似度

$$\text{sim}(CF_i, CF_j) = \sum_{Slot_i \in CF_i, Slot_j \in CF_j} \text{sim}(Slot_i, Slot_j)$$

## 2.2 EDR格フレーム

### 2.2.1 構造

本研究では、EDR日本語動詞共起パターン副辞書から動詞表記、動詞概念、表層格、名詞概念集合の組を取りだし、EDR格フレームとして比較対象に用いた。その構造は格フレーム  $H$  と同じであるが、名詞概念集合はその格の取りうる名詞概念の全範囲を示したものとなる。

### 2.2.2 類似度計算の定義

概念間の類似度

概念  $EDR.N_i, N_j$  の類似度

$$\text{sim}(EDR.N_i, N_j) = \begin{cases} 1 & \text{if } EDR.N_i \ni N_j \\ 0 & \text{otherwise} \end{cases}$$

$EDR.N_i$ : 概念の上位・下位関係を利用した集合的な概念記述の最小単位

$N_j$ : 一つ概念 ID

概念集合間・スロット間・格フレーム間の類似度  
格フレーム  $H$  の定義と同様である。

## 2.3 文の分類

ある文の動詞概念を推定する際、まず、文を次のように整形する。

動詞表記	
助詞 <sub>1</sub>	名詞概念集合 <sub>1</sub>
助詞 <sub>2</sub>	名詞概念集合 <sub>2</sub>
⋮	

ここでの名詞概念集合は、EDR電子化辞書の日本語単語辞書を引くことで得られる。

ある文データを  $S$ 、対応する動詞表記の格フレーム集合を  $CFS$  とする。このとき、 $S$  について、全ての  $cf_i \in CFS$  と類似度を取り、最大類似度  $MaxSim(S, CFS)$  とその時の  $cf_i$  のキー集合  $MaxKey(S, CFS)$  をそれぞれ求める。

$$MaxSim(S, CFS) = \max_{cf_i \in CFS} \text{sim}(S, cf_i)$$

$$MaxKey(S, CFS) = \{cf_i \mid \begin{aligned} &MaxSim(S, CFS) = \text{sim}(S, cf_i), \\ &MaxSim(S, CFS) \neq 0, \\ &cf_i \in CFS \end{aligned}$$

このとき、 $S$  の動詞概念は、 $MaxKey(S, CFS)$  の示す格フレームの動詞概念であると推定される。

## 3 サ変動詞の格フレーム獲得方法

### 3.1 サ変動詞の特徴

サ変動詞は、名詞+「する」という構造を持つ動詞である。サ変動詞には和語動詞と同じ意味を持つものや和語動詞に何らかの制限を加えた意味を持つものが多い。そのため、例えば次のような言い換えが可能である。

- 登山する → 山に登る
- 薄笑いする → かすかに笑う
- 問い合わせする → 問い合わせる

ここで、言い替えが可能ならば、取る格も同じであると仮定する。もしこの仮定が正しければ、サ変動詞から言い替え可能な和語動詞を抽出し、その格フレームをそのサ変動詞の格フレームとして用いることが可能である。

### 3.2 和語動詞の抽出

サ変動詞の構造により、以下の三つの規則を用いて和語動詞を抽出する。サ変名詞とは「する」を伴って動詞化する名詞である。

- 規則 a. サ変名詞に含まれる漢字が、ある和語動詞  $W$  に含まれる唯一の漢字である場合、 $W$  をその漢字に対応する和語動詞とする。
- 規則 b. サ変名詞に和語動詞  $W$  の連用形が含まれている場合、 $W$  を連用形部分に対応する和語動詞とする。

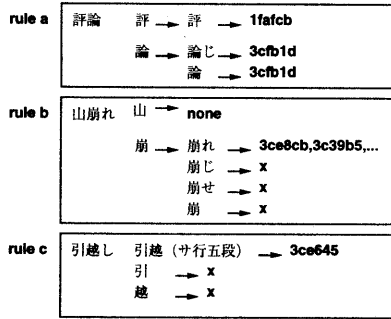


図 1: 和後動詞の抽出

- 規則 c. サ変動詞のサ変名詞部分が、和語動詞 W の語幹に対応している場合、W をサ変名詞に対応する和語動詞とする。

規則はサ変名詞の各部分で排他的に適用する。規則は c、b、a の順に優先される。適用の例を図 1 に示す。

### 3.3 格フレーム S の構造

本研究で獲得したサ変動詞の格フレームを格フレーム S と呼ぶ。格フレーム S は、3.2 節の規則を用いてサ変名詞から和語動詞を抽出し、その和語動詞の格フレームをもとに生成する。このとき抽出した和語動詞は、一般に複数であると考えられる。また、各和語動詞は多くの場合、概念異なりによって複数の格フレームを持つ。従ってそれらの格フレームのキーとして次の二つが考えられる。

- 動詞表記 + 概念 ID をキーとする
- 概念 ID のみをキーとする

それぞれを格フレーム  $S_A$ 、 $S_B$  とする。例を図 2 に示す。

### 3.4 格フレーム S を用いての分類

東は、共起データを格フレーム H に分類することによって評価を行っている [4, 5]。そこでは、日本語共起副辞書の共起データから 2.3 節に示す文データを作成して分類実験を行い、分類された格フレームの概念 ID (推定) と共起データに振られている概念 ID (正解) を照合して正答・誤答の判定を行った。

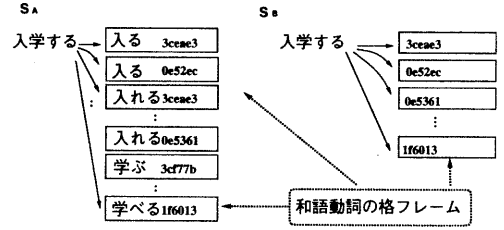


図 2: 格フレーム  $S_A$ 、 $S_B$

本研究でもそれに倣って共起データの分類実験を行った。ただし、獲得されるサ変動詞の格フレームには和語動詞の概念 ID が振られているため、共起データの概念 ID (以下 ID と略す) との一致によって正誤を判定することができない。そこでまず、あるサ変動詞 SN に対する格フレーム H の格フレーム集合  $H_{SN}$  と獲得したサ変動詞の格フレーム集合  $S_{SN}$  の間で類似度を取り、 $h_i \in H_{SN}$  に対して最も類似度が高い  $s_j \in S_{SN}$  を  $h_i$  に対応する格フレームであると見做して  $h_i$  の ID を仮 ID として与えた。正誤の判定は、この仮 ID を用いて行なった。

#### 0. 準備

$kd_k$ 、 $H_{SN}$ 、 $S_{SN}$  をそれぞれ、あるサ変動詞 SN に対する共起データ、格フレーム H の格フレーム集合、格フレーム S の格フレーム集合とする。

#### 1. 仮 ID の割り当て

各  $h_i \in H_{SN}$  について、 $s_j$  のキー集合  $MaxKey(h_i, S_{SN})$  を求める。

ここで、逆関数  $Key(s_j) = MaxKey^{-1}(h_i, S_{SN})$  は、各  $s_j \in MaxKey(h_i, S_{SN})$  に仮 ID として  $h_i$  の ID を与える関数である。

#### 2. 共起データの分類

$kd_k$  について、最大類似度  $MaxSim(kd_k, S_{SN})$  とその時の  $s_j$  のキー集合  $MaxKey(kd_k, S_{SN})$  をそれぞれ求める。

このとき、次のような結果を得る。

類似度ゼロ	$MaxSim(kd_k, S_{SN}) = 0$
類似度非ゼロ	$MaxSim(kd_k, S_{SN}) \neq 0$
正答	$Key(kd_k) = Key(h_i)$ を満たす $h_i \in K_{Key}(s_j)$ が存在 (ただし、 $s_j \in MaxKey(kd_k, S_{SN})$ )
誤答	その他

### 3.5 不要な格フレームの除去

収集した格フレームの中には不要なものが含まれている。このような格フレームを削除するため、タグ付けされていない文データを利用する。

タグ付けされていない文を QJP[6, 7] を用いて解析し、2.3 節で述べた文データの形にする。文データがある程度集まったサ変動詞について、格フレームと文データ間で類似度を取り、分類された文データの数がある閾値  $Th$  以下の格フレームは使用されないものと見做して削除する。以下にアルゴリズムを示す。

アルゴリズム  $B_{SN}$ 、 $S_{SN}$  をそれぞれ、あるサ変名詞  $SN$  に対する文データの集合、サ変名詞の格フレーム集合とする。

1. 各  $b_i \in B_{SN}$  について、 $MaxKey(b_i, S_{SN})$  を求める。
2. 各  $s_j$  が  $MaxKey(b_i, S_{SN})$  に含まれる数を数える。この関数を  $Choice(s_j)$  とする。

$$Choice(s_j) = \sum_{b_i \in B_S} In(w_j, MaxKey(b_i, S_{SN}))$$

ただし、

$$In(a, A) = \begin{cases} 1 & a \in A \\ 0 & otherwise \end{cases}$$

3.  $Choice(s_j)$  がある閾値  $Th$  以上である格フレーム  $s_j$  のみを格フレーム集合  $NewS_{SN}$  とする。

$$NewS_{SN} = \{s_j | Choice(s_j) < Th, s_j \in S_{SN}\}$$

4.  $NewW_{SN}$  をサ変名詞  $SN$  の格フレーム集合とする。

## 4 実験

### 4.1 格フレーム $S$ の獲得

EDR 単語辞書に記されているサ変動詞 18537 語について、3.2 節で示した規則 a, b, c を用いてサ変動詞から対

応する和語動詞を抽出し、その和語動詞の格フレームを全て集め、3.3 節で示したキーで統合することによって、格フレーム  $S_A$ 、 $S_B$  を獲得する。結果を表 1 に示す。

### 4.2 格フレーム数の削減

毎日新聞 92 年度版～94 年度版からサ変名詞を含む文を収集して文データとした。文の解析には QJP[6, 7] を用いた。文が 100 個以上集まったサ変動詞 100 個を対象として不要な格フレームを除去し、格フレーム数を削減した。

収集した文データのうち 100 文をサ変動詞の格フレーム  $S$  に分類し、分類された文データの数が閾値  $Th$  以下であるものを不適当な格フレームと見做して削除した。閾値  $Th$  は  $Th = 5, 10, 20$  と変え、実験によって最適値を見つけるものとした。

### 4.3 分類実験

4.1 節、4.2 節で獲得した格フレームを用いて、共起データの分類実験を行った。比較対象として、格フレーム  $H$ 、EDR 格フレームに対しても同様の実験を行った。

EDR 格フレーム、格フレーム  $H$ 、削減前の格フレーム  $S$  の全てに格フレームが存在するサ変動詞 1839 個について 2.5 節に述べた方法で実験を行った。その結果を表 2 に示す。

また格フレーム  $S$  の削減の効果を調べるため、同様の分類実験を削減前後それぞれの格フレームについて行い比較した。これは、削減を行ったサ変動詞の格フレーム 100 個について行った。その結果を表 3 に示す。

評価値には以下で定義する正解率・適合率を用いた。

- 正解率

$$\text{正解率} = \frac{\text{正答数}}{\text{試行数}}$$

- 適合率

$$\text{適合率} = \frac{\sum \text{正答} \cdot \frac{1}{\text{解答数}}}{\text{正答数}}$$

類似度計算には [4] で有効性が示されている  $\alpha = \beta = 0.5$ 、 $k = 3$  を用いた。

### 4.4 評価

#### 4.4.1 獲得数

EDR 単語辞書に登録されていたサ変名詞 18537 語のうち、EDR 格フレーム、 $H$  に登録されていた格フレーム

表 3: 分類実験 (削減後): 格フレーム  $H$ 、 $S_A$ 、 $S_B$ (サ変動詞数: 100, 共起データ数:12409, 正解あり:9605)

	$H$	$S_A$ (削減後)				$S_B$ (削減後)			
閾値 $Th$		削減なし	5	10	20	削減なし	5	10	20
正解率	0.891	0.679	0.689	0.696	0.733	0.756	0.753	0.758	0.766
適合率	0.802	0.580	0.591	0.611	0.651	0.513	0.497	0.507	0.525
格フレーム数 サ変動詞	1.33	13.27	8.00	6.07	4.22	11.94	8.71	7.09	5.55

数は約 3000 語であったが、今回の方法では  $S_A$ 、 $S_B$  ともに 15863 語を獲得できた (表 1)。

表 1: 辞書・格フレームのサ変名詞登録数

	サ変名詞集合	サ変名詞数
EDR 単語辞書	$SN_{All}$	18537
EDR 格フレーム	$SN_{EDR}$	3069
格フレーム $H$	$SN_H$	3550
共起データ	$SN_{KD}$	3550
格フレーム $S$	$SN_S$	15863
$SN_H \cap SN_S \cap SN_{EDR}$	$SN_{SH}$	1839

#### 4.4.2 格フレーム $H$ との比較

実験の結果、EDR格フレームの正解率が85.6% (表 2)、元にした格フレーム  $H$  の正解率が88.5% であるのに対し、格フレーム  $S_A$  の正解率は削減前69.8%であった (表 2)。 $S_B$  は削減前後ともに約76%を示しており、いずれも格フレーム  $H$  を10~15ポイント下回る結果となっている。

この開きの一因はサ変動詞当たりの格フレーム数にある。格フレーム  $H$  ではサ変動詞当たりの格フレーム数はわずか1.33であるため、分類をランダムに行ったとしても、 $1/1.33 = 0.75$  の確率で正解可能となる。対して  $S_A$ 、 $S_B$  では削減後でも4~5個の候補の中から分類を行っており、ランダムに行った時でも  $1/4 = 0.25$  の正解率が得られるだけである。

動詞当たりの格フレーム数が多い和語動詞の場合、[4]では、EDR格フレームで59.0%、格フレーム  $H$  で85.2%の正解率を得ており、動詞当たりの格フレーム数が正解率に影響を及ぼすことを示している。またその場合の適合率は50.7%、45.7%であり格フレーム  $W$  と大差ない。

従って、格フレーム  $S$  の正解率・適合率はそう低くはないと考えられる。

#### 4.4.3 キーの違い

$S_B$  は  $S_A$  に比べ、削減前では6%を上回る正解率を出している。これは、 $ID$  をキーとしたことで、次のような利点があったためである。

1. 類似した格フレームの融合による、格フレーム数の減少、違いの明確化
2. 各格フレームのデータ量の増加

表 2: 分類実験 (削減前): 格フレーム  $H$ 、 $S_A$ 、 $S_B$

	EDR	$H$	$S_A$	$S_B$
サ変名詞数	1839			
共起データ数	38316			
正解なし	25827	7977		
正解あり	12539	30339		
類似度ゼロ	8052	4713	4463	3015
類似度非ゼロ	4487	25626	25876	27324
誤答	644	2938	7804	6621
正答	3843	22688	18072	20703
正解率	0.856	0.885	0.698	0.758
適合率	0.850	0.858	0.570	0.515

#### 4.4.4 削減前後の比較

閾値  $T_h$  を 5、10、20 と変えて実験を行った (表 3)。削減前の名詞あたりの格フレーム数 13~12 個を、 $T_h$  を増やすことによって 4~6 個まで削減することができた。このとき、正解率・適合率ともに上昇を示した。従って、この操作は不要な格フレームの除去に効果があることが分かる。

## 5 考察

### 5.1 低正解率の原因

格フレーム  $S_A$  の分類実験で高正解率だったサ変動詞を調べると、比較的是っきりした形で和語動詞の意味が残っている。例えば正解率 100% のサ変動詞には次のようなものがある。

- 出題 → (問題を) 出す
- 許容する → 許す、容れる
- 掘進する → 掘り進む

これらは左辺を右辺に云い換えても違和感を感じず、意味も殆ど変化しない。

一方、正解率の低いサ変動詞は次のように分類できる。

1. 和語動詞の格フレームの欠如  
(ex. 論議、議論、克服)
2. 抽出された和語動詞との意味の不一致  
(ex. 処理、合格、協力)
3. 「××化」  
(ex. 激化、活発化、合理化)

このうち 1 は云い換えが可能であるが、元となる和語動詞の格フレームが不備 (データ不足) であったために正解を出せなかったものである。対して 2 では和語動詞が本来の意味とは異なった意味で使われており、サ変名詞から和語動詞への云い換えができなかった。

3 は「名詞+化」、「形容詞+化」というサ変名詞に極めて多い構造を持つ (18537 語中 199 語) が、

動詞要素 → 化 → 化ける、化す

で格フレームを獲得してしまい、適当なものを得られない。ゆえに「化」については別に取り扱う必要がある。そ

の際、動詞の格フレームよりは、「激化」には「激しい」(形容詞) の、「活発化」には「活発だ」(形容動詞) の格フレームが必要であると思われる。

### 5.2 改良案：動詞成分の抽出方法の改善

1 でデータ数が少ないことがあるのは、サ変名詞が必要とする和語動詞が特殊なものである場合、元となる格フレームのデータが十分である可能性が低いためである。

例えば克服の「克つ」と「克服」では、「克服」の方が一般的に使われ、用例を集めやすい。しかし、この手法では、「克服」の格フレームを得るためにまず「克つ」の格フレームを完備しておく必要がある。

対策としては次のようなものが考えられる。今回は漢字から動詞表記を得るのに、「漢字を一字だけ含み、かつ複合動詞ではない動詞」を *EDR* 単語辞書から自動的に収集して漢字から和語動詞を得る辞書を作成した。これによって、

- 克 → 克つ

などの動詞を得た。この辞書を例えば、適当で一般的な格フレームが得られるよう、次のように改良する。

- 克 → 克つ、勝つ
- 協 → 合う、合わせる

辞書の作成には人手が必要だが、*EDR* 単語辞書に登録された和語動詞の使用している漢字は高々 1800 字であった。しかもその全てが修正を必要とするわけではない。今回の実験で正解率が悪かったものを修正し、再実験を行って悪かったものを再度修正する、と云う手順を数度繰り返せば辞書も十分練れると思われる。

## 6 まとめ

本研究では、サ変動詞を構成する漢字に着目し、その漢字から動詞成分を抽出することによって既存の格フレームからサ変動詞の格フレームを獲得した。評価のため、既存の格フレーム  $H$  と獲得した格フレーム  $S_A$ 、 $S_B$  を用い、共起データの分類実験を行ったところ、 $S_A$ 、 $S_B$  では、*EDR* 格フレーム、格フレーム  $H$  を正解率・適合率ともに下回った。格フレーム  $S$  の低正解率の原因として主に次の二点が考えられる。

- 抽出された和語動詞があまり使われていない (克→克つ)。

- 抽出された和語動詞とサ変動詞間の意味の不一致（協→合わせる）。

これを改善する方法として漢字から動詞集合を得る辞書の改良がある。また複数の動詞要素を含む場合の対処の仕方など、改善の余地は様々にある。

低適合率の原因は動詞あたりの格フレーム数が多かったためである。格フレーム  $H$  でも動詞当りの格フレームが多い場合、同程度の適合率を出している。

格フレーム  $S_A$ 、 $S_B$  には不要な格フレームが含まれるため、そのような格フレームの除去が必要であった。本研究ではこれに新聞から収集した文を用い、 $S_A$ 、 $S_B$  のそれぞれで格フレーム数を削減した。このとき正解率・適合率はともに上昇を示した。

本研究の方法では、EDR単語辞書、既存の格フレーム、サ変名詞の表記、読みを利用して、サ変名詞の格フレームを自動獲得できる。この方法は、格フレームの拡張を容易にし、必要に応じてサ変名詞の格フレームを生成することを可能にする。さらに複合動詞の格フレームへの応用も可能である。

## 7 謝辞

本研究では、毎日新聞 92 年度版～94 年度版並びに EDR電子化辞書第 1.5 版を利用した。研究利用の便宜を計っていただいた関係者の方々並びに開発者の方々に感謝します。また QJP の利用を認めていただいた（株）リコー並びに開発者の亀田氏に感謝します。

## 参考文献

- [1] 仁田義雄、「語構成と文法記述」－漢語動詞をめぐって－、「語彙論的統語論」 明治書院、1980
- [2] 乾裕子 元吉文男 井佐原均、「語構成論に基づいたサ変動詞の分類」 情報処理学会報告 vol.95,no.110,pp.103-110
- [3] EDR 電子化辞書 1.5 版使用説明書（株）日本電子化辞書研究所
- [4] 東 優、「既存の概念辞書を用いた動詞の格フレームの獲得」九州大学大学院システム情報科学研究科修士論文,1997

- [5] 東 優、峯 恒憲、雨宮真人、「既存の概念辞書を用いた動詞語義による文の分類」、信学技法,NLC 96-36 pp.39-44,1996
- [6] 亀田 雅之、「簡易日本語解析系 Q-J P ライブラリ使用手引書」、1995
- [7] 亀田 雅之、「軽量・高速な簡易日本語解析系 Q-J P」 Ricoh Technical Report N0.22,pp.33-40JULY,1996