1R－6

和語動詞の格フレームを利用した サ変名詞の格フレームの獲得

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

1　はじめに

自然言語処理において重要な役割を果たしている格フレームを自動獲得する研究が盛んに行なわれている。獲得方法には大別して人手による方法と機械的に行なう方法がある。後者では、大量の用例を用意し、それを基に何らかの処理を行なって獲得を行なうが、この方法では、出現頻度の少ない語や、新語・造語を扱うことは、困難である。

サ変動詞は、要素の組合せによって構成されていることが多い（ex. 登山する→「登る」＋「山」）ため、種類が多く新語・造語が造られやすい。従って大量の用例をあらかじめ用意しておくことは困難であり、サ変動詞については、用例を必要としない方法を採ることが望ましい。

そこでサ変動詞の特徴に着目する。サ変動詞を構成する要素は、和語動詞、名詞、形容詞などに対応しており、サ変動詞の意味が、それらの構成要素に強く結び付いていることが多い。そのためサ変動詞は多くの場合、和語動詞を用いて言い替えることが可能である。

これを利用して、本稿では、和語動詞の既存の格フレームを用いることによって、サ変動詞の格フレームを作成する方法を提案する。

さらに、そのように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提案する。

2　サ変動詞の格フレーム獲得

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ため、「登山する　→　（山に）登る」といった言い替えが可能である。本稿では、このように言い替えが可能であれば、その和語動詞の格フレームをサ変動詞の格フレームとして用いることができると仮定し、サ変動詞の格フレームを獲得した。以下にその手順を示す。

1. 言い替え可能な和語動詞を抽出するため、サ変名詞に含まれる漢字に着目し、図1のように、同一の漢字を含む和後動詞を抽出する。

2. このように抽出された和語動詞の格フレームを集めたものをサ変動詞の格フレームとする。このとき、「動詞表記＋概念 ID」をキーとした格フレームを S_A、「概念 ID」をキーとした格フレームを S_Bとする。例を図2に示す。

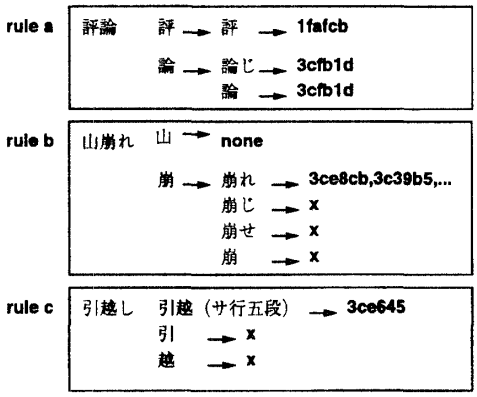

図 1: 和後動詞の抽出

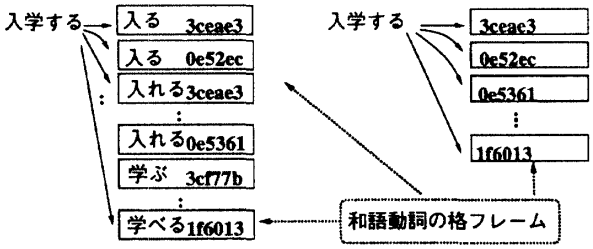

図 2: 格フレーム S_A、S_B

3　実験

3.1　実験方法

本手法で獲得した格フレームの性能を評価するため、EDR日本語共起パターン副辞書 [1] の EDR格フレーム及び EDRコーパスから東の手法 [3, 4] によって獲得された格フレーム H を比較対象の格フレームとして用いた。

サ変動詞の格フレームの基となる格フレームには、格フレーム Hを用いた。

前節の方法により、EDR単語辞書に記されているサ変動詞 18537 語のうち、15863 語について、格フレーム S_A、S_Bを獲得できた。

各格フレームの性能を比較するため、EDR格フレーム、格フレーム H、格フレーム S_A、S_Bに共通して格フ

表 1: 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			
正解率	0.856	0.885	0.698	0.758
適合率	0.850	0.858	0.570	0.515

レームが存在するサ変動詞 1839 個について、それぞれ共起データの分類を行った [2]。その結果を表 1に示す。

獲得した格フレームから不適当な格フレームを削除するため、用例を用いて削減を行なった。毎日新聞 92 年度版～94 年度版を QJP[5, 6] を用いて解析し、サ変名詞を含む文を収集して用例とした。この用例が 100 個以上集まったサ変動詞 100 個を削減の対象とした。

削減は、100 個の用例をサ変動詞の格フレーム Sに分類し、分類された用例の数が閾値 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することによって行なった。閾値を $Th = 5, 10, 20$ と変え、削減前後の格フレームを比較するため、それぞれの格フレームで分類実験を行った結果を表 2に示す。

3.2 考察

格フレーム H、EDR格フレームに比較すると、獲得した格フレームは、正解率・適合率ともに低かった。正解率が高かったサ変動詞では、サ変動詞の意味が、抽出した和語動詞と一致しているものが多かった。また正解率の低かったサ変動詞では、意味の不一致や、目的とした和語動詞の格フレームの欠落、データ不足などが目立った。

ランダムに選んだサ変動詞について、獲得した格フレームを調べたところ、候補として集めた和語動詞の格フレームの中に適当なものがあれば、分類は比較的正しく行なわれていたことが分かった。

また、削減によって、完全にではないが、不適当な格フレームを削除できていた。しかし、閾値を上げ過ぎると適当な格フレームまで削除されてしまうため、閾値の決定には注意が必要である。今回の実験では、閾値 $Th = 20$ で適当な格フレームが削除されてしまう場合があり、$Th = 10$ が適当であると言える。

前述したように、本手法では適切な和後動詞を抽出することができない場合があり、低正解率の原因となっている。この、適切な和語動詞の抽出が今後の課題である。

また S_Aと S_Bでは、S_Bの方が高正解率であった。これは IDのみをキーとしたため、類似した格フレームがまとめられ、整理されたためである。ただし、IDが同一でも格フレームが異なる場合があり、IDのみをキーとすることが一概に良いとは言えない。

表 2: 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)

	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
$\frac{\text{格フレーム数}}{\text{サ変動詞}}$	1.33	13.27	8.00	6.07	4.22	11.94	8.71	7.09	5.55

4 おわりに

本手法で獲得した格フレームには、内容として不正確なものが含まれることがあったが、その原因は主として抽出した和語動詞が不適当なためであった。

より正確な格フレームを獲得するためには、適切な和語動詞の抽出が必要である。今後は漢字の一致だけに頼るのではなく、辞書を用い、意味を考慮した上で和語動詞を抽出することを考える必要がある。

参考文献

[1] 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

[2] 橋本 順子、峯 恒憲、雨宮 真人、「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」信学技報,NLC97-26 pp.55-62,1997

[3] 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

[4] 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」、信学技報,NLC 96-36 pp.39-44,1996

[5] 亀田　雅之、「簡易日本語解析系Ｑ_ＪＰライブラリ使用手引書」、1995

[6] 亀田　雅之、「軽量・高速な簡易日本語解析系Ｑ_ＪＰ」Ricoh Technical Report N0.22,pp.33-40JULY,1996