# 情報エントロピを用いた格フレームの用例の汎化

阿部　裕司†　藤本　浩司‡　小谷　善行‡
東京農工大学工学部情報工学科†　東京農工大学大学院工学府‡

## 1　はじめに

一般に，文の要素（特に名詞）が述語に対して果たす意味的役割を格と呼び，日本語においてはガ格・ヲ格などといったように格助詞を用いて表す．ある述語の格要素としてどのような名詞を使うことができるのかを体系的に表したものが格フレームである．代表的な格フレームは人手により作成されているが，コーパスから用例を収集することにより，格用例データに基づく格フレームを作成することができる[1]．本研究では，自動的に収集された格用例データを抽象化し，格要素を意味カテゴリ（以降単にカテゴリと呼ぶ）で表現した格フレームを生成する手法を提案する．具体的には，河原[1]の手法により自動生成された京都大学格フレーム[3]の格要素を日本語語彙体系シソーラス[2]のカテゴリを用いて抽象化し，カテゴリ格フレームを生成する（図1）．

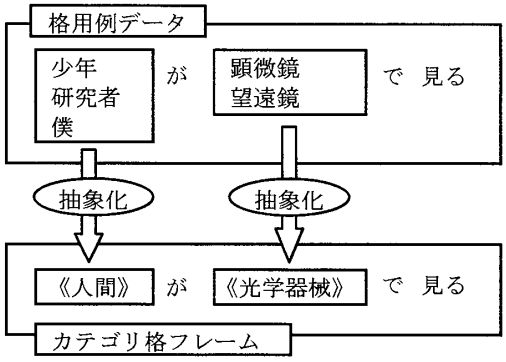


図1　格用例データの抽象化

## 2　格フレーム抽象化システム

本章では，格用例データの一つの格（ガ格・デ格など）に含まれる用例群を抽象化する手順を説明していく．なお，本論文中では日本語語彙体系シソーラス上のカテゴリを《…》でくくって表す．

### 2. 1 個々の格用例のカテゴリ候補の決定

まず，個々の格用例がどのカテゴリに属する可能性があるのかを判別する．個々の格用例に対し，それがシソーラス上のどのカテゴリに登録されているか調べる．なお，一つの格要素は複数のカテゴリに登録されている可能性がある．

### 2. 2 カテゴリへのスコアの割り当て

格用例には，その用例のコーパス内における出現頻度情報が付加されている．格用例の頻度をもとに，カテゴリにスコアを割り当てる．格用例がただ一つのカテゴリに登録されている場合には，その格用例の頻度をラベル付けされたカテゴリのスコアに加算する．格用例が複数のカテゴリに登録されている場合には，格用例の頻度を，それが登録されているカテゴリの数で割った値をそれぞれのカテゴリのスコアに加算する．

### 2. 3　正例および負例の追加

格用例として出現した名詞は適切な用法であると仮定し，それらを正例として扱う．すなわち，2.2 節で計算されたスコアを正例の数として扱う．逆に，格用例として出現しなかった名詞は不適切な用法であると仮定し，それらを負例として扱う．具体的には，日本語語彙体系シソーラスに登録されていて，かつ格用例として出現していないすべての名詞を頻度 1 の負例として扱う．

### 2. 4　抽象化に適切なカテゴリの選択

格用例の抽象化に適切なカテゴリを選択する際，そのカテゴリの包含する範囲内には正例の割合が多く，包含しない範囲には負例の割合が多いことが望ましい．そこで，抽象化に適切なカテゴリを選択する基準として，カテゴリの内側と外側を分離した際の情報エントロピを利用する．情報エントロピの計算式を次ページの式(1)(2)(3)に示す．すべてのカテゴリについて全体の情報エントロピを計算し，全体の情報エントロピの値が最も小さいカテゴリを抽象化に適切なカテゴリとして選択する．

---

Generalization of Case Frame using Information Entropy
†Yuji Abe, Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology
‡Yoshiyuki Kotani, Koji Fujimoto, Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology

$$E=\frac{E_i(P_i+N_i)+E_o(P_o+N_o)}{P_i+N_i+P_o+N_o} \quad (1)$$

$$E_i=-\frac{P_i}{P_i+N_i}\log\frac{P_i}{P_i+N_i}-\frac{N_i}{P_i+N_i}\log\frac{N_i}{P_i+N_i} \quad (2)$$

$$E_o=-\frac{P_o}{P_o+N_o}\log\frac{P_o}{P_o+N_o}-\frac{N_o}{P_o+N_o}\log\frac{N_o}{P_o+N_o} \quad (3)$$

$E$ ： 全体の情報エントロピ
$E_i$ ： カテゴリ内部の情報エントロピ
$E_o$ ： カテゴリ外部の情報エントロピ
$P_i$ ： カテゴリ内部の正例数
$P_o$ ： カテゴリ外部の正例数
$N_i$ ： カテゴリ内部の負例数
$N_o$ ： カテゴリ外部の負例数

## 3 カテゴリ格フレーム生成実験

京都大学格フレームに登録されている 34059 個の述語の格要素に対して抽象化システムを適用し，カテゴリ格フレームの生成を行った．実験では，カラ格，ガ格，ガガ格，デ格，ト格，ニ格，ヘ格，マデ格，ヨリ格，ヲ格を抽象化の対象とした．図2，図3に生成されたカテゴリ格フレームの一例を示す．図 2 は妥当な抽象化がなされたと思われる格フレームの例，図3は適切ではない抽象化がなされたと思われる格フレームの例である．

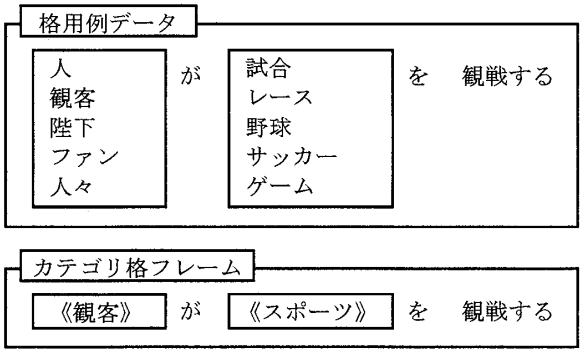


図2 「観戦する」格フレーム

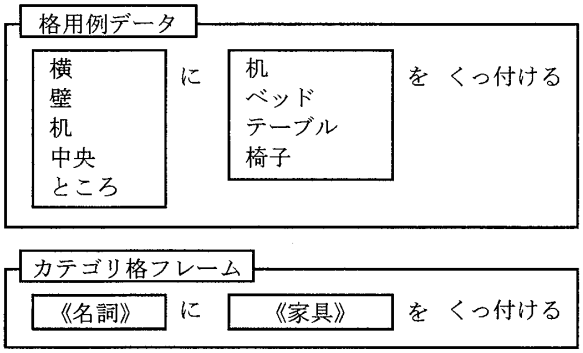


図3 「くっ付ける」格フレーム

## 4 生成された格フレームに対する考察

図 2 の格用例データについて，ガ格の格用例データに含まれている名詞はすべて《人》に登録されている名詞である．しかし，より抽象度の低い《観客》を選んだほうが情報エントロピの値が小さくなるため，《観客》が抽象化のために適切なカテゴリとして選択された．図 2 の格用例データのヲ格に含まれている名詞はすべて，《スポーツ》に登録されているため，抽象化に適切なカテゴリとして《スポーツ》が選択された．

一方で，図 3 においてはニ格を抽象化するためのカテゴリとして《名詞》が選択された．《名詞》は日本語語彙体系シソーラス中のすべての名詞を包含するカテゴリであり，ニ格の格用例データを抽象化するカテゴリとしてふさわしくないと思われる．

ニ格の格用例データに含まれる「壁」という名詞は多義の名詞であり，日本語語彙体系シソーラス上の七つのカテゴリに登録されている．本研究で提案した手法では，格用例の頻度を，その格用例が登録されているカテゴリのスコアに等分割して加算している．そのため，格用例「壁」の頻度が広範囲のカテゴリのスコアへと配分され，汎化に適切なカテゴリとして《名詞》が選択される結果となった．このように，多義の名詞が格用例に含まれている場合，不必要に包含範囲の広いカテゴリが選択される傾向にある．

## 5 おわりに

本研究では，格用例データとして京都大学格フレームを用い，それを日本語語彙体系シソーラスのカテゴリを利用して抽象化する手法を提案した．カテゴリ格フレーム生成実験では，京都大学格フレームに登録されている 34059 個の述語の格要素を抽象化し，それらの述語に対応するカテゴリ格フレームを生成した．

**謝辞** 本研究では，格用例データとして京都大学大学院情報学研究科黒橋研究室が作成した京都大学格フレームを利用した．また，カテゴリ体系として日本語語彙体系シソーラスを利用した．ここに感謝の意を表する．

**参考文献**

[1]河原大輔, 黒橋貞夫:格フレーム辞書の暫時的自動構築, 自然言語処理, Vol.12,No.2,pp.109-131(2005).
[2]ＮＴＴコミュニケーション科学研究所:日本語語彙大系, 岩波書店(1997).
[3]特定非営利活動法人言語資源協会. http://www.gsk.or.jp