コーパスからの格フレーム半自動獲得のための支援環境の構築

中塚 幸毅
奈良先端科学技術大学院大学 情報科学研究科
mailto:yukita-n@is.aist-nara.ac.jp

大規模なコーパスから格フレームなどの言語知識の自動的な獲得を行うことは、自然言語処理のための辞書を構築する上で非常に重要な課題である。本論文では、構文解析等においてよく使用される動詞の格フレーム獲得の一つの手法として、動詞・格要素の共起用例のクラスタリングを計算機により求め、その結果に基づいて動詞の格フレームを人手により抽出する手法について述べる。また、計算機によるクラスタリングの結果を視覚的に表示し、人手による修正を行うためのＧＵＩを製作し、格フレームを半自動的に獲得するための支援環境の構築を行なった。コーパスより抽出した動詞の共起用例より、実際に格フレームを獲得する実験を通して、本研究において構築した支援環境を用いて格フレームを獲得する手法の有効性について検討する。

1 序論

自然言語処理インターフェイスなどにおいて、計算機による自然言語解析を実現させる場合には、対訳辞書や格フレーム辞書などといった、人間の言語知識を計算機で利用できるように表現した言語辞書の整備や構築が不可欠になるが、従来の言語処理システムにおいては、これらの言語辞書はすべて、人間の手によって構築が行われてきた。しかし、人手による言語辞書の構築には、

1. 膨大な言語知識量に伴う作業コスト
2. 辞書の一貫性の保持の困難性
3. 辞書の拡張の困難性
4. 特定の分野の辞書構築の困難性

などの問題点がある。

しかし、最近になって、大規模な計算機可読なテキスト（コーパス）が大量に出回るようになり、計算機の性能が向上するに伴って、大規模コーパスから計算機を用いて自動的に言語知識を獲得する必要性が主張されてきた。

大規模コーパスより、格フレームを獲得する事を目的とした既存の研究は大きく、コーパスの文章に構文解析を施さずに格パターンを抽出する手法によるもの [Brent93, Manning93] と、構文解析済みのコーパスから格フレームを抽出する手法によるものに分けることが出来る。後者の手法では、統計的手法が多く用いられ、１個の格要素の汎化レベルの学習を行なう研究 [Resnik93, Li95] や複数の格要素に対して汎化レベルの学習を行なう研究 [春野 95] が知られている。

しかしながら、(a) 多義の動詞を複数の用法の分類に関して、どのような観点より分類を行なうかについての判断について、絶対的な基準が存在しない。 (b) 必須格や、任意格について、どのような観点で分類を行なうかについての判断について、絶対的な基準が存在しない。などの理由から、完全な格フレームの自動獲得が実現できていないのが現状である。

以上を踏まえて、本論文では、計算機を用いた格フレームの完全な自動抽出を目指すのではなく、計算機による格フレーム抽出の結果について、人間の手により検証を行い、計算機が判断できなかった部分や計算機の判断が誤ったものであった場合に、人間の手により、その誤りを修正するといった、人手による操作を容易に行えるような支援環境の構築を目指すことを考える。

そして、多義性のある動詞の語義の分類を目的とした格フレーム辞書の半自動獲得について、計算機による用例のクラスタリング及び手動による格フレーム候補選択の手法について提案する。

2 動詞の格フレーム辞書の半自動獲得

2.1 格フレーム・用例のデータ構造

用例・格フレームのデータ構造について述べる前に、格フレーム中の格要素に対応する名詞の意味クラスを記述するのに必要なシソーラスの概略について述べる。

本研究においては、シソーラスのデータ構造を、意味クラスを節点とし、上位・下位概念の関係にある意味クラス間に枝を張った木構造とみなす。

そして、木構造中の意味クラス間に存在する上位概念関係（包含関係）を $\preceq_c$ で定義し、意味クラス c_2 が意味クラス c_1 の上位概念にある（包含関係にある）と言うことを、$c_1 \preceq_c c_2$ と表記し、意味クラス c_1 と

意味クラス c_2 の選言を、 $c_1 \vee c_2$ と表記する。

次に、格フレーム及び動詞の共起用例のデータ構造を次のように定義する。

まず、動詞と動詞に係る格要素の共起データ及び、格フレーム e, f を、動詞 v および、格 p と格要素に対応する名詞の意味クラス c の組の集合として、次のように記述する。

$$e, f \quad = \quad \left[\begin{array}{c} pred : v \\ p_1 : c_1 \\ p_2 : c_2 \\ \cdots \\ p_l : c_l \end{array} \right] \qquad (1)$$

ここに、意味クラス $c_1, c_2, \ldots c_l$ は、用例の場合は、シソーラスにおける葉クラスに相当するものとし、格フレームの場合はシソーラスの任意の意味クラス及びその選言とする。また、格フレーム及び、動詞の共起用例に出現しない格についても、格フレーム及び用例の格要素の値が $\phi(\ =$ 空要素$)$ である格として存在するものと仮定する。省略を表す格要素 ϕ を用いれば、任意格については以下の例の「で」格のように、 ϕ との選言を記述することにより、省略可能であることを示す。

$$\left[\begin{array}{l} pred : \text{飲む} \\ \text{を} : < \text{飲物} > \\ \text{で} : < \text{場所} > \vee \phi \end{array} \right] \qquad (2)$$

2.2 用例・格フレームの包含関係

ある動詞の共起用例及び格フレームが、格フレームに包含されると言う関係 $\preceq_f$ を、次のように定義する。

$e \preceq_f f \quad \Leftrightarrow \quad f$のすべての格 p について、
f の持つ p の意味クラス c_f が、
e の持つ意味クラス c_e を包含する。

ここに、包含関係の例を示す。

$$\left[\begin{array}{l} pred : \text{飲む} \\ \text{が} : < \text{人間} > \\ \text{を} : < \text{飲物} > \end{array} \right] \preceq_f \left[\begin{array}{l} pred : \text{飲む} \\ \text{が} : < \text{動物} > \\ \text{を} : < \text{液体} > \end{array} \right]$$

$$\left[\begin{array}{c} pred : \text{飲む} \\ \text{を} : < \text{ジュース} > \\ \text{で} : \phi \end{array} \right] \preceq_f \left[\begin{array}{c} pred : \text{飲む} \\ \text{を} : < \text{飲物} > \\ \text{で} : < \text{場所} > \vee \phi \end{array} \right]$$

但し、シソーラスによって意味クラスの包含関係 $<$人間$>$ $\preceq_c$ $<$動物$>$, $<$ジュース$>$ $\preceq_c$ $<$飲物$>$ $\preceq_c$ $<$液体$>$ が定義されているものとする。

2.3 基本的な考え

ここで、本研究において提案する手法を述べる前に、この手法についての基本的な考えについて述べる。

まず最初に、具体的に「買う」という動詞について、いくつかの例文を動詞の語義別に分類することを考える。まず、「買う」という動詞は主に、「金銭でものなどをやり取りする」という意味と「ある不快な感情を引き起こす」という意味が考えられる。ここで、「買う」という動詞が前者の意味を取るときの例文としては、

土地を買う, 時計を買う, 食糧を買う, 株を買う, ...

などが考えられ、後者の意味を取るときの例文は、

不評を買う, 怒りを買う, 反感を買う, ...

などが考えられる。以上の例文を見ると、「買う」という動詞は前者の意味では「を」格に物を表す名詞を取り、後者の意味では、「を」格に感情を表す名詞を取ることが分かる。つまり、「買う」という動詞を取る例文を「買う」の語義によって分類する場合には、「買う」に「を」格で係受けする名詞の意味に基づけば良いと考えられる。

そして、他の動詞についても同様に、ある特定の格に注目し、その格に対応する名詞の意味クラスに従って分類すると、適切に用例の類別が行われるのではないかと推測することができる。

2.4 用例のクラスタリング手法

大規模コーパスから抽出された動詞の共起用例のクラスタリングは次の手順によって行う。

1. 初期化 ...

   まず、与えられた動詞の共起用例中に出現するすべての格を $p_1, p_2, \ldots p_l$ として、

   $$E = \bigcup_{i=1}^{l} E_{p_i} \ ; \ E_{p_i} = \{e \mid e[p_i] \neq \phi\}$$

   (ここに、$e[p_i]$は、用例eの格p_iにおける
   格要素の値。以下同様。)

   を満たすような用例集合 $E_{p_1}, E_{p_2}, \ldots, E_{p_l}$ を集め、この集合を、軸格を p_i とする、クラスタリングの初期の用例集合とする。

   そして、すべての格 $p = p_1, p_2, \ldots, p_l$ について次の初期化を行う。

   (a) 格 p の存在する用例を集めた集合 E_p について、

   $$\forall e \in E_p \ ; \ e \preceq_f f_p$$

   なる f_p を次の式によって求める。

   $$f_p[p'] \quad \longleftarrow \quad \left\{ \begin{array}{ll} c_g & (\forall e \in E_i(\hat{f}), e[p'] \neq \phi) \\ c_g \vee \phi & (\exists e \in E_i(\hat{f}), e[p'] = \phi) \end{array} \right.$$

   $$c_g \quad = \quad (e \in E_p, e[p'] \neq \phi\text{を満たす全ての}\ e[p']\text{の最小汎化クラス})$$

   $$(\ p' = p_1, p_2, \ldots, p_l\ )$$

(b) 軸格 p について展開を行う格フレーム候補集合 F_p 及び、格フレーム f_p に包含される用例集合 $E(f_p)$ を次のように設定する。

$$F_p \longleftarrow \{f_p\}\ ;\ E(f_p) \longleftarrow E_p$$

2. 反復 ...

集合 $F_p(\ p = p_1, p_2, \ldots p_l\ )$ より、任意に 1 個の格フレーム $\hat{f}$ を選び、以下の操作を繰り返す。

(a) 格フレーム $\hat{f}$ 中の格 p の格要素の意味クラス c が、その子クラス $c_1, \ldots, c_n$ に分割されるとする。

(b) 格フレーム $\hat{f}$ が包含する用例集合 $E(\hat{f})$ を、格 p の意味クラス $c_1, \ldots, c_n$ の制約を満す部分集合 $E_1(\hat{f}), \ldots, E_n(\hat{f})$ に分割。すなわち、$E(\hat{f})$ を

$$E_i(\hat{f}) = \{e_i | e_i[p] \preceq_f c_i\}\ \ (i = 1, 2, \ldots n)$$

なる部分集合に分割する。

(c) 分割された各部分集合 $E_1(\hat{f}), \ldots, E_n(\hat{f})$ について用例を汎化し、格フレーム $\hat{f}$ の下位格フレーム $f_1, \ldots, f_n$ を求める。このとき、格フレーム $f_i\ (i = 1, \ldots, n)$ 中のそれぞれの格 $p'(\neq p)$ の意味クラス c' は以下の基準により求める。用例集合 $E_i(\hat{f})$ 中の用例の格 p' のクラスを汎化した意味クラスを c_g として、

$$c' = \begin{cases} c_g & (\forall e \in E_i(\hat{f}),\ e[p'] \neq \phi) \\ c_g \vee \phi & (\exists e \in E_i(\hat{f}),\ e[p'] = \phi) \end{cases}$$

$$c_g = (e \in E_p, e[p'] \neq \phi \text{を満たす全ての } e[p'] \text{の最小汎化クラス})$$

(d) 以下の式により，格フレーム候補集合 F_p と、$\hat{f}$ の下位の格フレーム集合 $C(\hat{f})$ と、f_i に包含される用例集合 $E(f_i)$ を更新する。

$$F_p \longleftarrow F_p \cup \{f_1, f_2, \ldots, f_n\} - \{\hat{f}\}$$

$$C(\hat{f}) \longleftarrow \{f_1, f_2, \ldots, f_n\}$$

$$E(f_i) \longleftarrow E_i(\hat{f})\ \ (\ i = 1, 2, \ldots, n\ )$$

3. 停止条件 ...

前述の反復操作は以下のいずれかの条件を満たした場合に停止する。

(a)

$$\forall f' \in F_p \quad ; \quad \{\ f'[p] \text{のシソーラス中での深さが } Th_d \text{ より深くなる}\ \}$$

の条件が成り立つ。(ここに、Th_d は葉クラスに近いシソーラスの階層の深さ。)

(b) 集合 F_p 中のすべての格フレームで、

$$\forall f' \in F_p\ ;\ |E(f')| \leq 1$$

となる。

以上のアルゴリズムによる動詞の用例集合のクラスタリングの手順は、格フレーム f を節点とし、下位の格フレームの集合 $C(f)$ の各要素に向けて、枝を張るような階層的な木構造を用いて表現できる (図 1)。

図 1: 用例のクラスタリングの例

2.5 人手による格フレーム候補の選択操作の手順

そして、計算機によって階層的に行われたクラスタリングの手順を基に、次に人手によって、動詞の格フレームの抽出の手がかりとなる共起用例のクラスタリング結果を求める。

次に述べる作業手順では、2.4 節で述べたクラスタリングの階層的な木構造の枝を、ほぼ深さ優先の手順で辿っていき、それぞれの節点でその節点に割り当てられたクラスタ中の動詞の語義の同一性を判断して、1 つのクラスタ中の用例の動詞の語義が全て等しいと判断できればそのクラスタの格フレームの汎化レベルとする。次に、クラスタ中の用例の動詞の語義が同じもの同士のクラスタを併合する (図 2)。

1. 準備 ...

まず、与えられた動詞について、その動詞が自動詞か他動詞かなどを考慮し、格フレーム獲得の上で、最も適当であると考えられる軸格 p を 1 個選び、初期の格フレーム候補の集合を $F_h = \{f_p\}$ とする。

2. 適切な粒度選択 ...

次の手順に従い、格フレーム中の格要素の適切な汎化レベルを決定する。

(a) まず、F_h より、適切な 1 個の格フレーム $\hat{f}$ を選び、共起用例の集合 $E(\hat{f})$ を見る。

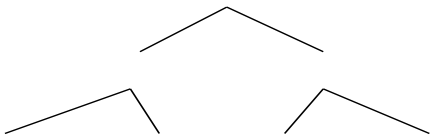

図 2: 格フレーム操作の例

図 3: 支援環境のシステム構成図

(b) 共起用例の集合 $E(\hat{f})$ において、

- すべての $E(\hat{f})$ 中の動詞が、同じ語義によって使用されていると判断できる場合は、格フレーム $\hat{f}$ をそのままにする。
- すべての $E(\hat{f})$ 中の動詞が、同じ語義によって使用されていると判断できない場合は、格フレーム $\hat{f}$ を分割し、F_h を $F_h \longleftarrow F_h \cup C(\hat{f}) - \{\hat{f}\}$ と更新する。

以上の操作を、

$\forall f' \in F_h$; { $E(f')$中の動詞が同じ語義によって使用されていると判断できる }

が成り立つまで反復する。

3. 併合操作 ...

F_h の各要素に包含される用例集合 $E(f_1), E(f_2), \ldots E(f_n)$ $(f_1, f_2, \ldots f_n \in F_h)$ のそれぞれの用例集合中の動詞について、同じ動詞の語義によって使用されていると考えられる用例集合について、次のように併合操作を行う。

$$E'_i \longleftarrow E(f'_1) \cup E(f'_2) \cup \ldots \cup E(f'_l)$$

そして、併合操作の終わった用例のクラスタについて、適当にクラスタ中の用例を包含する格フレーム $f_{E'_i}$ を次のように割り付ける。

$$f_{E'_i} \longleftarrow f'_1 \vee f'_2 \vee \ldots f'_l \quad [1]$$

以上の操作により最終的に割り付けられた格フレームを獲得された格フレームとする。

3 支援環境の構築

この章では本研究において構築した格フレーム辞書獲得のための支援環境の概略について述べる。

3.1 システム構成

支援環境のシステム構成についての全体図を図 3 に示す。システムは、主に、動詞の共起用例の集合のクラスタリングを行なうプログラムと、Tcl/Tk [Ousterhout95] を用いて書かれたＧＵＩ環境インターフェイス clust2.tcl からなる。

まず、大規模なコーパスを構文解析した物及び、解析済のコーパスから、動詞の共起用例の集合を取り出し、その用例集合を用例のクラスタリングプログラムに入力し、階層的なクラスタリング結果を求める。この結果を入力として、ＧＵＩ環境のインターフェイスを起動し、これを通じて人手による作業を行い、格フレーム辞書を獲得する。

3.2 インターフェイスの機能

支援環境におけるＧＵＩ環境インターフェイスについて、概観を図 4に示す。支援環境におけるＧＵＩ環境インターフェイスには、次に挙げる機能がある。

1. 格フレームの表示、閲覧機能 ...

   任意の格フレームの下位の格フレームの展開表示, 格フレームに包含される用例の一覧表示などを行う。

[1] ここで、格フレーム間の選言を次のように定義する。

$$f = f_1 \vee f_2 \Longleftrightarrow f[p'] = f_1[p'] \vee f_2[p']$$

(p' は f_1, f_2 のすべての可能な格を取る。)

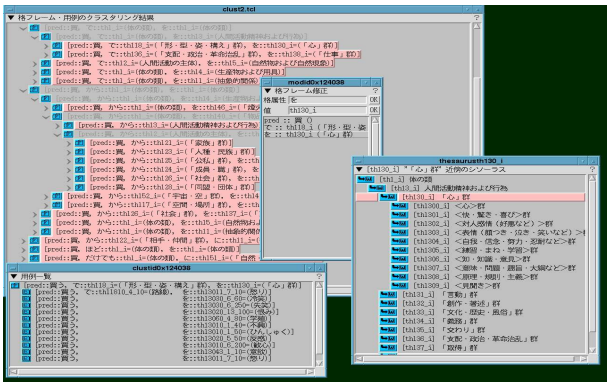

図 4: ＧＵＩ環境インターフェイスの外観

2. 格フレームの修正、操作機能 . . .

   指定した複数の格フレームの併合, 任意の格フレームの格要素の修正などを行う。

3. その他の機能 . . .

   任意のシソーラスの意味クラスを辞書引きした結果の表示、クラスタリング結果の入出力を行うファイルの選択などを行う。

4 格フレームの半自動獲得の実験

本章では、本研究での支援環境を用いて、格フレームの半自動獲得を行う実験について述べる。

なお、本実験においては、シソーラスとして分類語彙表 [国研 93] を使用し、用例を抽出したコーパスについては、構文解析済のコーパスである EDR の日本語共起辞書 [EDR94] から、出現頻度 50 回以上の動詞 835 個について動詞・格要素の共起を合計 153,014 個抽出したものを用いた。この中から、 20 個の動詞を抽出し、以下の実験を行った。

4.1 格フレームの最適な粒度選択に関する実験

20 個の動詞について、本研究で構築した支援環境を用いて、共起用例のクラスタリングを行う。その結果を、ＧＵＩ環境インターフェイスを用いて、 2.5 節の手順の「粒度選択」で述べた操作に従い、格フレームの候補を決定する。

ここで、動詞の用法および、語義について判断を下した基準は、日本語基本動詞用法辞典 [小泉 91] を用いた。この辞典は、日本語の基本動詞 728 個について、語義、用法、文法情報等が記述されている。

そして、今回実験を行った 20 個の動詞について、(a) 粒度選択が完了するまでに必要とする最小の、クラスタの展開回数。 (b) 粒度選択が完了した時の、総クラスタ数の記録を行い、考察を加えた。

表 1: 格フレーム獲得の実験結果

動詞	語義数	軸格	step	粒ク	併ク	棄却
触れる	6	に	16	54	5	4
動く	6	が	21	70	6	20
取る	19	を	30	102	14	32
押す	5	を	12	29	4	5
傾く	4	が	4	12	4	1

4.2 格フレームの併合操作に関する実験

この実験は 4.1 節で述べた実験に続いて行う。

20 個の動詞について、 4.1 節での実験により、粒度選択が完了したクラスタリング結果について、 2.5 節の手順の「併合操作」に従い、格フレームの併合操作を行う。

ここで、用例を見ただけでは、語義の判別が著しく困難であると判断された用例のクラスタは、これを棄却した。

そして、今回実験を行った 20 個の動詞について、(a) 併合操作を行った後のクラスタ数。(本実験において抽出された動詞の語義数) (b) 語義の判別が困難であるとして棄却されたクラスタ数の記録を行い、考察を加えた。

4.3 実験結果

20 個の動詞についてのクラスタの展開回数と、総クラスタ数を表 1 に示す。表 1 では、「動詞」の欄に、今回、実験を行なった 20 の動詞を、「語義数」欄は、日本語基本動詞用法辞典に記述されていた語義の個数を示し、「軸格」の欄に、その動詞において格フレームの獲得に最も適切であると判断された軸格を示した。

また、「step」の欄には最適なクラスタの粒度を選択するのに最低必要なクラスタの展開数を、「粒ク」欄に最適なクラスタの粒度に分割された時のクラスタ数を示す。そして、「併ク」欄に併合操作後のクラスタ数を示し、「棄却」欄に併合操作の時に棄却したクラスタ数を示す。

4.4 考察

まず、クラスタの最適な粒度選択の実験においては、今回はクラスタの分割で注目する格 (軸格) を 1 個に固定する手法を用いたが、この手法において、一部分の動詞を除いて、ほぼ有効に動詞の語義を抽出出来ていることが分かった。

しかし、動詞「張る」については軸格を「を」格に設定しただけでは「氷が張る」という例で使用される語義が適切に抽出できず、もう 1 個の「が」格を軸格

に設定して操作を行う必要があった。このような動詞は他に「かかる」「引く」があった。

そして、本研究で提案した手法により、動詞に係る格要素に対応する名詞の最適な汎化レベルの選択の作業を、「用例中の動詞はすべて同じ語義かどうか」という二値的な判断作業によって行うことが可能となり、従来の格フレーム辞書の構築作業と比較して、労力が大幅に軽減されることが示された。

5 結論

本研究では、解析済コーパスより取り出した用例を１個の軸格に注目しながら、階層的にクラスタリングを行うアルゴリズムを提案した。また、そのクラスタリングの結果を基に、人手により動詞の多義性を類別した格フレームの獲得手法を提案した。

そして、格フレームを半自動的に獲得するための支援環境の構築を行ない、実験の結果、提案した手法による動詞の多義性類別の有効性が示された。

最後に本研究の今後の課題について述べる。

1. 格フレーム辞書の妥当性の評価 ... 実験によって構築した格フレーム辞書を、現実の構文解析に使用した場合の解析性能について評価を加える。

2. 統計的手法の導入についての検討 ... 既存の格フレーム自動獲得に関する研究では、統計的手法の有効性が示されている。これらの統計的な手法を、本研究のアルゴリズムに導入し、人手による判断の手数の軽減について検討を加える。

3. 人手の判断結果の再利用についての検討 ... 前に行ったクラスタの分割における判断結果を、後の方に行うクラスタの分割の判断において、何らかの形で参照し利用する機構の実装などについて検討を加える。

4. ＧＵＩ環境インターフェイスの機能強化 ... 支援環境システムのパッケージ化及び、配布に向けて、共起用例を抽出した原文表示や、軸格表示などのＧＵＩ環境インターフェイスの機能強化について検討を加える。

参考文献

[Brent93] Brent, M. R.: From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax, *Computational Linguistics*, pp. 243–262 (1993).

[EDR94] EDR 日本電子化辞書研究所: EDR 電子化辞書評価版 (第 2.1 版), (1994).

[Li95] Li, H., Abe, N.: Generalizing Case Frames Using a Thesaurus and the MDL Principle, *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pp. 239–248 (1995).

[Manning93] Manning, C. D.: Automatic Acquisition of a Large Subcategorization Dictionary from Corpora, *Proceedings of the 31st Annual Meeting of ACL*, pp. 235–242 (1993).

[Ousterhout95] Ousterhout, J. K. (西中芳幸, 石曽根信訳): Tcl & Tk ツールキット, ソフトバンク社, (1995).

[Resnik93] Resnik, P.: Semantic Classes and Syntactic Ambiguity, *Proceedings of the Human Language Technology Workshop*, pp. 278-283 (1993).

[小泉 91] 小泉 保, 船城 通雄, 本田晶治, 仁田 義雄, 塚本秀樹: 日本語基本動詞用法辞典, 大修館書店, (1991).

[国研 93] 国立国語研究所:‘分類語彙表, 秀英出版, (1964).

[春野 95] 春野 雅彦: 最小汎化とオッカムの原理を用いた動詞格フレーム学習, 情報処理学会研究報告, *Vol. 95, No. 69 (95–NL–108)*, pp. 29–36 (1995).