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　　宇津呂 武仁　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
mailto:{yukita-n,utsuro,matsu}@is.aist-nara.ac.jp

1 はじめに

計算機による自然言語解析を実現させるためには、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。一方、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。

言語処理で使用する言語知識の一つである格フレームは、構文解析を行う上において非常に重要である。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、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] 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大きく分けることが出来る。後者の手法では、情報理論的情報圧縮の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、現時点では、(a) 多義性のある動詞の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という判断について、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、基準が存在しない、などの理由から、格フレームの完全な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獲得を目指すのではなく、計算機による格フレーム獲得の結果について、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について述べる。

2 動詞の格フレームの半自動獲得

2.1 基本的なアイデア

格フレームの獲得を行う上で解決すべき問題は、1節で述べたように、多義性のある動詞の語義の類別や必須・任意格の分類などの他、シソーラス上での格の意味制約の汎化レベルの決定の問題が考えられる。本論文においては、最初に動詞の多義性類別を計算機と人手で解決し、その結果から必須・任意格の分類とシソーラス上での格の意味制約の汎化レベルの決定の問題を解決することを考えていく。[1]

そこで、動詞の格フレームの半自動的な獲得において、動詞の多義性の類別を行う手法の基本的なアイデアについて述べる。

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。前者の意味を取るときの例文としては、

「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。

以上より、動詞の用例についてある特定の格（以下、このような特定の格を「軸格」と呼ぶことにする。）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と推測することができる。

[1] シソーラス上での格の意味制約の汎化レベルの決定の問題は、格の意味制約に関して適切なシソーラスが使用できるかどうかに強く依存しており、そのような最適なシソーラスが存在しない現時点においては完全に解決することは難しい。そこで、本論文では現時点において容易に解決できると考えられる動詞の多義性類別の問題を主目的とする。

本論文では動詞の多義性の類別を主目的とした動詞の格フレーム獲得を以下の手順で行う。まず、計算機によって用例集合のクラスタリングを行った後、クラスタリング結果に基づいて人手により格フレーム候補を選択・併合を行う。

2.2 用例のクラスタリング

ここで、計算機による用例のクラスタリングの手法について述べる。具体例として、図 1 の点線で囲まれている動詞「走る」についての用例を「が」格に注目してクラスタリングを行う場合を例を考える。なお、図 1 の左の図のような名詞のシソーラスが既に与えられているとする。

1. 初期化

まず、全ての用例の最小汎化の格フレームを考え、この格フレーム 1 個から成る集合を「現在の格フレーム集合」とする。図 1 の例では、「＜物＞が走る」という格フレームが、点線で囲まれている用例の最小汎化の格フレームであり、「＜物＞が走る」1 個から成る集合を、現在の格フレーム集合とする。

2. 反復

現在の格フレーム集合より 1 個の要素を取り出し、（以下「分割される格フレーム」と呼ぶ）次の操作を施す。

ここで、上位の格フレームの軸格の格要素を Cu とし、 Cu の下位概念の意味制約のクラスを $Cd_1, Cd_1, \ldots Cd_k$ とする。そして、意味クラス $Cd_1, Cd_1, \ldots Cd_k$ を軸格の意味制約とする格フレームを「下位の格フレーム」と呼ぶことにする。

そして、分割される格フレームに包含される用例の集合を分割される格フレームの下位の格フレームに包含される用例の集合毎に分割し、分割される格フレームを現在の格フレーム集合より除き、下位の格フレームを加える。

ここで、「＜物＞が走る」を取り出した場合について考えると、まず、「が」格の格要素の＜物＞の下位クラスは、シソーラスより、＜無生物＞と＜動物＞なので、「＜物＞が走る」の下位の格フレームとして「＜無生物＞が走る」と「＜動物＞が走る」ができ、「＜物＞が走る」に包含される用例の集合をそれぞれ「＜無生物＞が走る」と「＜動物＞が走る」に包含される用例の集合に分割する。そして、現在の格フレーム集合より「＜物＞が走る」を除き、「＜無生物＞が走る」「＜動物＞が走る」を加える。

3. 停止条件

前述の反復操作は、用例のクラスタリングの階層数が葉クラスに近いシソーラスの階層の深さ程度に深くなるか、現在の格フレーム集合中の全ての格フレームで、それが包含する用例数が 1 以下となった場合に停止する。

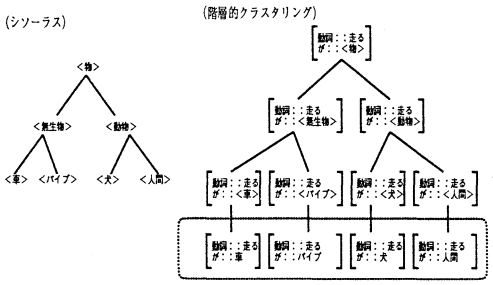

図 1: 用例のクラスタリングの例

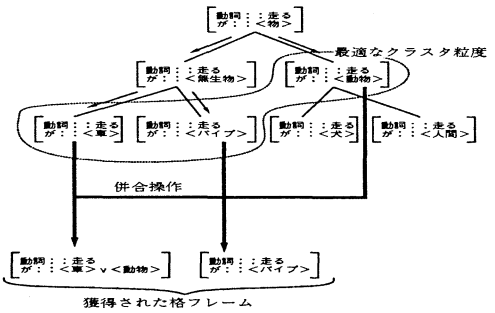

図 2: 格フレーム操作の例

2.3 人手による格フレーム候補の選択・操作

計算機によって階層的に行われたクラスタリング結果を基に、人手により格フレームの最適な汎化レベルを選択し、同じ語義の動詞を持つ格フレーム同士を併合する。ここでは、例として 2.2 節の階層的クラスタリングをもとに動詞「走る」を「物が高速に動く」という語義と「何かが線状に延びる」という語義に類別することを考える。

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を 1 個選び、その軸格によるクラスタリング結果の最上位の格フレーム 1 個から成る集合を格フレーム候補の集合とする。動詞「走る」の例を考えると、動詞が自動詞であることなどを考慮して「が」格を軸格とするのが適当であると考えられる。そして、「＜物＞が走る」からなる集合を格フレーム候補の集合とする。

2. 格フレームの最適な汎化レベルの選択

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。

まず、格フレーム候補の集合より、適切な 1 個の格フレームを選び、その格フレームに包含される用例の集合を見る。用例集合中の全用例の動詞が、同じ語義によって使用されていると判断できる場合は、格フレームをそのままにし、そうでない場合

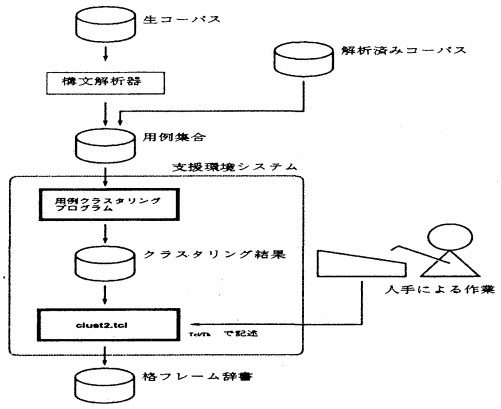

図 3: 支援環境のシステム構成図

は、その格フレームを下位の格フレームに分割し、格フレーム候補の集合からその格フレームを除き、下位の格フレームを追加する。

ここで、動詞「走る」の場合の作業の過程において、「＜無生物＞が走る」という格フレームを考えると、これに包含される用例は「車が走る」「パイプが走る」で、これは前述の2個の「走る」の語義が混在していると判断されるので、この格フレームは下位の格フレームに分割する。一方、「＜動物＞が走る」という格フレームは「犬が走る」「人間が走る」という用例を包含し、これらはいずれの「走る」も「物が高速に動く」という語義であるので、この格フレームはこれ以上の分割をやめる。

以上の操作を、用例集合中の全用例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まで反復する。

3. 同じ語義の格フレームの併合

2. の操作で選択された最適な汎化レベルの格フレームについて、動詞が同じ語義を持つ格フレーム同士を選言演算により併合する。動詞「走る」の例で、最適な汎化レベルとして図 2 の点線で囲まれたものが求められた場合では、「物が高速に動く」と言う語義で用いられている「＜動物＞が走る」「＜車＞が走る」を併合して「＜動物＞∨＜車＞が走る」とする。

3 支援環境の構築

本節では、格フレーム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。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（Tcl/Tk で作成）を起動し、これを通じて人手による作業を行い、格フレームを獲得する。

表 1: 格フレーム獲得の実験結果（抜粋）

動詞	語義数	軸格	step	汎ク	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
張る	12	を	15	47	9	4
.	–	が	3	7	1	5

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** (図 4 の左上図), **2. 格フレームに包含される用例の表示、閲覧機能** (図 4 の左下図), **3. 格フレームの修正、操作機能** (図 4 の中央図), **4. シソーラスの意味クラスの辞書引き結果の表示** (図 4 の右下図)

4 格フレームの半自動獲得の実験

本節では、3 節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。

4.1 実験

本実験においては、シソーラスとして分類語彙表 [国研 64] 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 [EDR95] 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、共起用例のクラスタリングを行った。その結果を、ＧＵＩ環境インターフェイスを用いて、 2.3 節の手順に従い、格フレームの候補を決定し、併合操作を施した。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。

4.2 実験結果

20 個の動詞のうちの 5 個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」欄に実験を行なった動詞を、「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」欄に、その動詞において格フレーム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」欄には最適な格フレームの汎化レベルを選択するのに必要なクラス

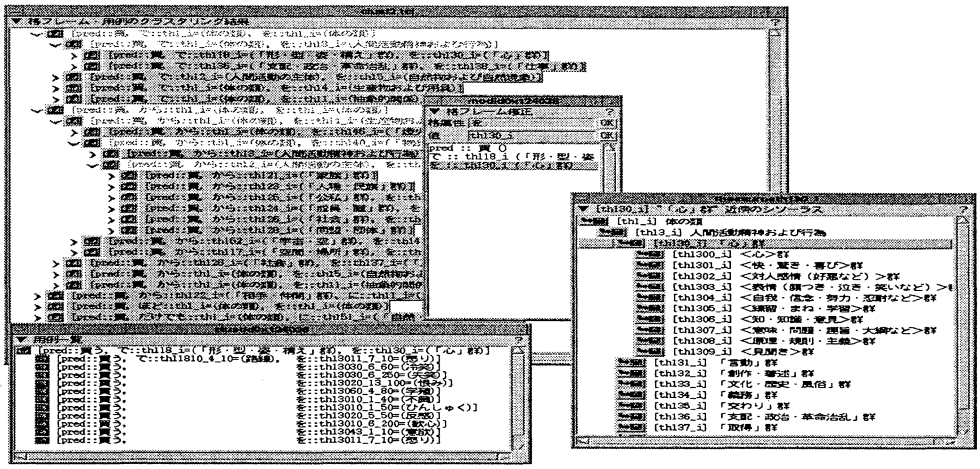

図 4: ＧＵＩ環境インターフェイスの外観

タの展開数を、「汎ク」欄に最適な格フレームの汎化レベル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」欄に併合操作時に棄却したクラスタ数を示す。なお、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、もう一つ「が」格を軸格に設定して操作を行う必要があったため、「を」格と「が」格による実験結果を併記した。

4.3 考察

まず、格フレームの最適な汎化レベル選択の実験においては、格フレーム候補の選択で注目する格を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」「かかる」「上げる」「引く」については１個の軸格を設定しただけでは語義が適切に抽出できず、複数の軸格を設定する必要があった。

そして、本論文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、「用例集合中の全用例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、労力が軽減されることが示された。

5 おわりに

本論文では、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について述べた。そこ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い、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を主眼として格フレームの獲得を行った。

そして、格フレームの半自動獲得の実験を行った結果、提案した手法による動詞の多義性類別の有効性を確認した。

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. 人手の判断結果の再利用についての検討 ... 以前に行った最適な汎化レベル選択における人手の判断結果を、後に行う汎化レベル選択の判断において、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。

謝辞

なお、本研究は「IPA 創造的ソフトウェア育成技術」の研究支援の一環として行ったものである。

参考文献

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

[EDR95] EDR 日本電子化辞書研究所: EDR 電子化辞書, (1995).

[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).

[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 64] 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).