用言と直前の格要素の組を単位とする格フレームの自動構築

河原 大輔[†]　　黒橋 禎夫[††]

本稿では、生コーパスから格フレームを自動的に構築する手法を提案する。格フレームの自動構築における最大の問題は、用言の用法の多様性をどのように扱うかということである。本研究では、用言と直前の格要素の組を単位としてコーパスから格要素と用言の用例を収集することにより、用言の用法の多様性を扱う。さらに、用法に違いはないが、直前の単語が異なるために別の格フレームになっているもののクラスタリングを行う。得られた格フレームを用いて格解析実験を行い、その結果を考察する。

キーワード： 格フレーム, 生コーパス, クラスタリング, 格解析

Case Frame Construction by Coupling the Predicate and its Closest Case Component

Daisuke Kawahara[†]　　Sadao Kurohashi[††]

This paper describes a method to construct a case frame dictionary automatically from a raw corpus. The main problem is how to handle the diversity of verb usages. We collect predicate-argument examples, which are distinguished by the verb and its closest case component in order to deal with verb usages, from parsed results of a corpus. Furthermore, we cluster and merge predicate-argument examples which do not have different usages but belong to different case frames because of different closest case components. We also report on an experimental result of case structure analysis using the constructed case frame dictionary.

KeyWords: *case frame, raw corpus, clustering, case structure analysis*

1　はじめに

日本語には語順の入れ替わり、格要素の省略、表層格の非表示などの問題があり、単純な係り受け解析を行っただけでは文の解析として十分とはいえない。例えば、「ドイツ語も話す先生」という文の場合、係り受け構造を解析しただけでは、「ドイツ語」と「話す」、「先生」と「話す」の関係はわからない。このような問題を解決するためには、用言と格要素の関係、例えば、「話す」のガ格やヲ格にどのような単語がくるかを記述した格フレームが必要である。このような格フレームは文脈処理 (照応処理、省略処理) においても必須の知識源となる。

† 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University
†† 東京大学大学院情報理工学系研究科, Graduate School of Information Science and Technology, The University of Tokyo

これまで、重要な用言の典型的な格フレームについては、人手で辞書をつくるということも試みられてきた。しかし、格と同じ振る舞いをする「によって」、「として」などの複合辞があること、「〜が〜に人気だ」のように名詞+判定詞にも格フレームが必要なこと、専門分野ごとに用言に特別な用法があることなどから、カバレージの大きな実用的な辞書をつくるということは大変なことであり、人手による方法には限界がある。

そこで、格フレーム辞書をコーパスから自動学習する方法を考える必要がある。しかし、格フレームの学習には膨大なデータが必要となり、現存するタグ付きコーパスはこのような目的からは量的に不十分である。そこで、本論文では、格フレーム辞書をタグ情報が付与されていない大規模コーパス (生コーパス) から自動的に構築する手法を提案する。

格フレーム辞書を生コーパスから学習するためには、まず、生コーパスを構文解析しなければならないが、ここで解析誤りが問題となる。しかし、この問題はある程度確信度が高い係り受けだけを学習に用いることでほぼ対処することができる。むしろ問題となるのは用言の用法の多様性である (これはタグ付きコーパスから学習する場合にも問題となる)。つまり、同じ表記の用言でも複数の意味、格要素のパターン (用法) をとり、とりうる格や体言が違うことがあるので、用言の用法ごとに格フレームを作成することが必要である。本論文では、これに対処するために、用言とその直前の格要素の組を単位として用例を収集し、それらのクラスタリングを行うという方法を考案した。用言とその直前の格要素の組を単位とするというのは、「なる」や「積む」ではなく、「友達になる」「病気になる」、「荷物を積む」「経験を積む」を単位として収集するということである。用言とその直前の格要素の組を単位として考えると、用言の用法はほとんど一意に決定される。この組み合わせは膨大になるので充分な量のコーパスが必要であるが、本研究では生コーパスから収集するので問題にならない。クラスタリングは、用法に違いはないが、用言の直前の単語が異なるために別の格フレームになってしまう用例をマージする処理である。

2　格フレーム構築の種々の方法

我々の提案する格フレーム辞書の自動構築の過程は以下のとおりである (図 1 の点線で囲まれた部分)。

1. コーパスのテキストに対して、KNP(黒橋, 長尾 1994) を用いて構文解析を行い、その結果から、ある程度信頼できる用言・格要素間の関係を取り出す。ここで取り出すデータを**用例**と呼ぶ。
2. 抽出した関係を用言と直前の格要素の組ごとにまとめる。このようにして作成したデータを**用例パターン**と呼ぶ。
3. シソーラスを用いて、用例パターンのクラスタリングを行う。この結果できたものを**用例格フレーム**と呼び、本研究ではこれが最終的に得られるものである。以下では「荷物」、「物資」、「経験」などの格要素になる単語を**格用例**、用例格フレームにおけるある格の格

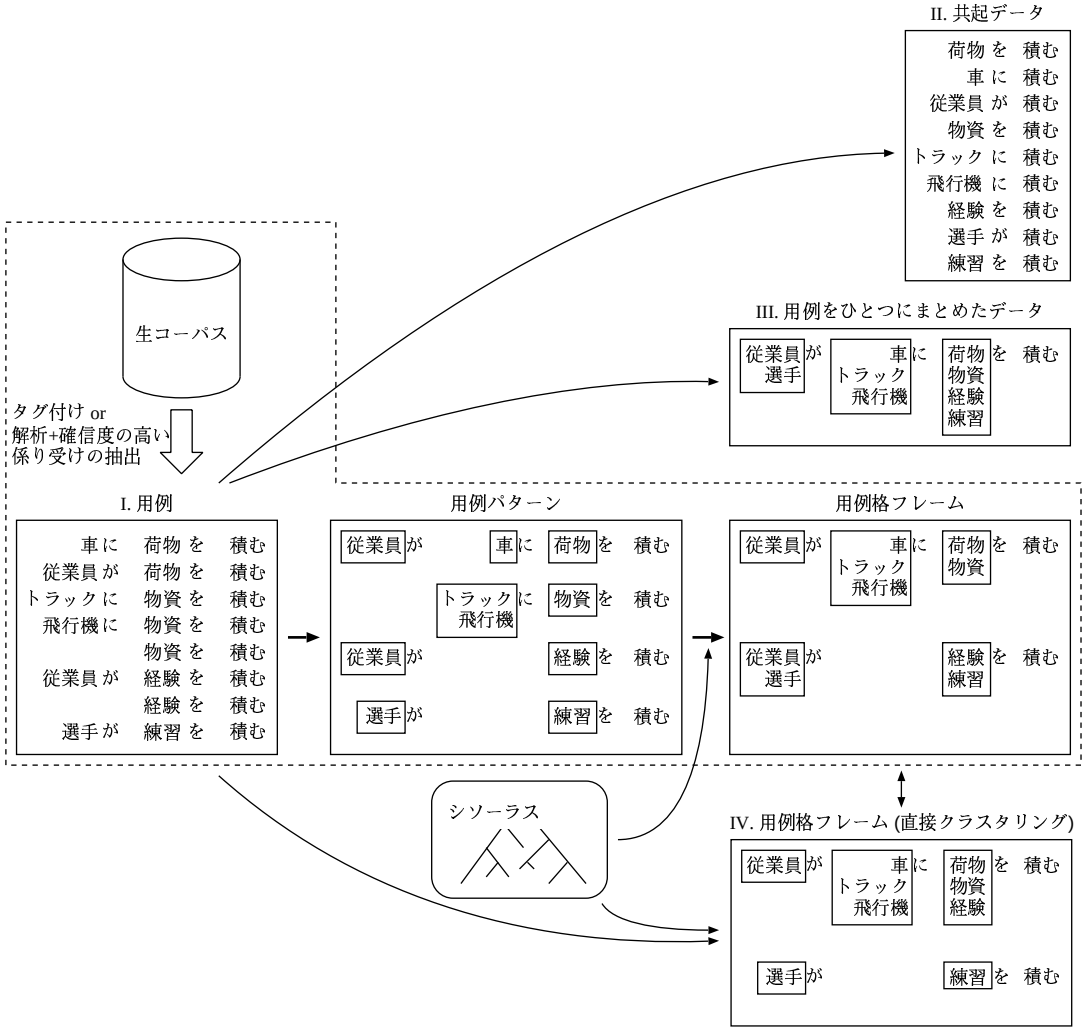

図 1: 格フレームに関連するさまざまなデータ処理

用例の集合、例えば「積む」の 1 つめの用例格フレームのヲ格の格用例集合 {「荷物」, 「物資」} を**格用例群**と呼ぶ。**格要素**は格用例と格の組である。

次に、格フレームに関連するさまざまなデータ処理を図 1 に沿って議論する。

まず、図 1 の I の用例をそのまま個別に使うことが考えられるが、この場合データスパースネスが問題になる。

(1) a. 車に　　荷物を 積む
　　b. トラックに 物資を 積む

例えば、この 2 つの用例がコーパスにあったとしても、「車に物資を積む」という表現が妥当であるかどうかはわからない。

一方、用例を二項関係に分割すると、図 1 の II のような共起データを作ることができる。これは統計パーサによって用いられているデータ形式であり、データスパースネスの問題を回避することができる (Collins 1996)。しかし、その副作用として用言の用法の多様性の問題が生じる。

(2) a. 車に 荷物を 積む
　　b. 　　経験を 積む

例えば、この 2 つの用例から「車に 積む」、「荷物を 積む」、「経験を 積む」という共起データが得られるが、これらのデータだけでは「車に経験を積む」のような間違った表現を許すことになる。

また、図 1 の III のように用例を単純にまとめたものも、もっている情報は共起データと同じであり、やはり用言の用法の多様性が問題となる。

これに対して、本手法で得られる用例格フレームでは、用言とその直前の格要素を組にして扱うという方法で、用法の多様性の問題を解決しつつ、データスパースネスにも対処している。

一方、用例を直接クラスタリングすることによって用例格フレームを作成する方法も考えられる (図 1 の IV)。この方法でも、用言の用法ごとに分かれた用例格フレームが得られるので、我々の作成する用例格フレームに近いといえる。しかし、この方法ですべての格要素を等しく扱うと、用言の用法にあまり関係しない格要素 (用言の直前ではない格要素) が類似していることによって、用言の用法の異なる用例がひとつの用例格フレームにマージされてしまうことがある。

(3) a. 従業員が 荷物を 積む
　　b. 従業員が 経験を 積む

例えば、この 2 つの用例は、用法が異なっているが、ガ格の「従業員」が同じであるためにマージされる可能性がある。このような問題があるため、用例を直接クラスタリングする方法では、必ずしもよい精度の格フレームにはならないと思われる。格フレームは辞書として利用されるものであり、精度は非常に高いものが要求されるため、この方法は格フレームの作成には適当ではないと考えられる。

3　関連研究

英語を対象として、生コーパスから格フレームを学習する方法はいくつか研究されてきた (Brent 1991; Manning 1993; Briscoe and Carroll 1997)。英語は格要素が省略されることがなく、問題となるのは格要素が用言にとって必須であるか任意であるかの判定である。この判定は、統計情報を利用して用言と格フレームの関連度を計算することによって行われている。学習する格フレームは用例格フレームのようなものではなく、動詞が名詞句と前置詞句をとるといったパターンである。つまり、用言の用法そのものを収集していると考えられるので、用言の用法の多様性は問題にならない。

日本語では、格フレームを構文情報付きコーパスから学習する方法が提案されている (東, 峯, 雨宮 1996; 宇津呂, 宮田, 松本 1997)。これらの手法は、学習に構文情報付きコーパスを用いているためカバレージの点で問題がある。春野は、意味素を要素とする格フレームをコーパスから学習する方法を提案している (春野 1995)。11 個の動詞を対象とし、新聞 1 年分から人手で抽出した用例を用いているのでカバレージの点では問題ないが、動詞数を増やして実用的な格フレームを作成するのは難しいと思われる。これらの手法で得られる格フレームは、格要素を汎化した意味素を格フレームの個々の要素としたものであり、この点では本研究と異なる。用言の用法の多様性は、それぞれ次のようにして扱っている。東らは EDR コーパスを用いており、動詞についている動詞概念ごとに格フレームを作成している。宇津呂らと春野の手法は、それぞれ機械学習、情報圧縮の手法を用いて意味素の汎化レベルを決定することによって、用例を直接クラスタリングするものである。しかし、前節で述べたように、これらの方法は精度の面で格フレームの作成には適当ではないと考えられる。

4　用例の収集

コーパスを構文解析した結果から、図 1 に示したような用例の収集を行う。質の高い用例を収集するために、コーパスの解析結果から確信度の高い係り受けを抽出する。

4.1　格要素の条件

用例を収集するときに、格、格用例、格要素に以下の条件を設定する。

格の設定

収集する格要素の格として、日本語の基本的な格すべてを対象とする。対象とした格を以下に示す。

ガ格, ヲ格, ニ格, ト格, デ格, カラ格, ヨリ格, ヘ格, マデ格, 無格

これらに加えて、次のものも格として扱う。

時間格　ニ格、無格、カラ格、マデ格で、意味素「時間」(後述) をもっている格要素はまとめてひとつの格にする。これは、格フレームを作成する際には、その用言が時間に強く関係しているかどうかが重要であり、表層格の区別は重要でないからである。

例: 3時に, 来年から

複合辞　格と同じように振る舞う複合辞を、それぞれひとつの格として扱う。

例: ～をめぐって, ～によって,
～について, ～として

格用例の汎化

個別の単語を扱うことにあまり意味がなく、明確な意味を考えることできる格用例はクラスとしてまとめて扱う。この汎化したクラスを以下のように 3 種類設定した。この場合、格用例として単語のかわりにクラスを記述する。

時間

- 品詞細分類が時相名詞の形態素を含む文節
  例: 朝, 春, 来年
- 時間助数辞を含む文節
  例: １９９９年, １２月, ６日, ９時, ３５分, ２３秒
- 「前」、「中」、「後」という接尾辞をもち、自立語がシソーラス上の意味属性「場所」をもたない文節
  例: 会議中, 戦争後, 書く前

数量

- 数詞を含む (助数辞を含まない) 文節
  例: １, ２, 一, 二, 十, 百
- 数詞と、「つ」、「個」、「人」のような助数辞を含む文節については、「<数量>つ」、「<数量>個」、「<数量>人」のように数量クラスと助数辞のペアにして扱う。
  例: １つ → <数量>つ
  ２個 → <数量>個

補文

- 引用節「〜と」、連体修飾+形式名詞 またはそれに準ずる表現 (〜の〜, 〜くらい〜,)
  例: 書くと, 書いたことを, 書くのを,
  書くくらいが

曖昧な格要素の排除

次のような格要素は収集に用いない。

- 提題助詞をもつ格要素と用言の連体修飾先は、表層格が明示されていないので収集に用いない。
  例: その 議員 は〜を提案した。
  〜を提案している 議員 が〜
- ニ格、デ格で副詞的に使われる格要素は、係る用言との関係が任意的であるので収集から除外する。これらの格要素については人手で辞書を作成した。
  例: ために, 無条件に, うえで, せいで

- KNP では、「〜では」、「〜でも」はデ格、「〜には」、「〜にも」はニ格の格要素として扱われるが、副助詞、あるいは従属節の場合もあるので収集の対象から除外する。
  例: 足の１本でも 折ってやろうかと思った。
  育成しないこと には 世界で通用しない。

格要素が複合名詞の場合には、もっとも意味的に重要であると考えられる最後の自立語を収集に用いる。

例えば、

(4) ３０日に総理大臣がその２人に賞を贈った。

という文からは、

<時間>:時間格　大臣:が　<数量>人:に　賞:を　　贈る

という用例を得る。

4.2 用言の条件

収集する用言は動詞、形容詞、名詞+判定詞とする。名詞+判定詞として収集する用言には体言止めの名詞も含む。ただし、以下のような用言は収集に用いない。

- 用言が受身、使役、「〜もらう」、「〜たい」、「〜ほしい」、「〜できる」の形であれば、格の交替が起こり、格と格要素の関係が通常の場合と異なるので収集に用いない。
- 「〜で」は、判定詞かデ格かの自動判定が難しいので、KNP が判定詞と認識しても、用言として収集に用いない。
  例: 彼は 京都で、試験を受け… (助詞)
  彼が好きな町は 京都で、… (判定詞)
- 形態素解析において、活用形から原形が一意に決まらない用言は収集に用いない。
  例: あった: ある, あう
  いった: いる, いう
- 用言として用いられているサ変名詞の直後に読点か句点がある場合、そのサ変名詞が受身か能動であるのかを区別することは難しいので、これは収集に用いない。
  例: 世界選手権は約１２００人が出場して福井県鯖江市で 開催。

4.3 確信度の高い係り受けの抽出

コーパスを構文解析した結果から用例を収集するときに問題となるのは、解析結果に誤りが含まれていることである。そこで、誤りの影響を軽減するために、解析の精度が低い係り受けは捨てて、ある程度確信度が高い係り受けを格フレームの収集に用いる。

KNP では、次のような優先規則によって文節の係り先を決定している。

Rule1　文中の強い区切りを見つけることによって、係り先の候補の絞り込みを行う (ここで候補がひとつになるなら、係り先をそれに決定する)。

Rule2　係り先の候補の用言のうち、格要素の係り先にならないことが多い用言を候補から除外する。

Rule3　“読点のない文節はもっとも近い候補に係り、読点のある文節は 2 番目に近い候補に係る” という優先規則に従って、候補の中から係り先を決定する。

用例の収集では、Rule1 は信頼し、Rule2 と Rule3 は信頼しない (多くの場合正しいが、誤っていることもある) こととする。つまり、Rule1 で候補がひとつになり決定される係り受けは用例の収集に用い、Rule2 や Rule3 の処理が適用された係り受けは収集に用いない。

(5) 彼は先生のアドバイスに従って英語を勉強したので、テストのスコアが大きく上がった。

この例では、「〜ので」は KNP によって強い区切りであると認識され、「英語を」の係り先の候補は「勉強した」の 1 つしかないので、この用例が取り出される。「スコアが」の係り先の候補は、「大きく」が Rule2 によって除外されており、「上がった」の 1 つだけであるが、この用例は取り出されない。「アドバイスに」の係り先の候補は「従って」、「勉強した」の 2 つであり、Rule3 の優先規則により係り先は「従って」に決定されるが、この用例は取り出されない。

上の例ではルールがすべて正しく働いていたが、Rule2 によって係り先の候補から除外した用言は、場合によっては係り先になる可能性があるので、このときの用例は収集しないことにしている。例えば、次の例のように、形容詞「早い」の直後に「救う」のような強い用言がある場合、このような形容詞は格要素の係り先になりにくいために、係り先の候補から除外される。

(6) 長女が気づき、家族とともに二人を助けようとしたが火の 回りが 早く救い出せなかった。

この例では、「回りが」は形容詞「早く」に係るのが正解であるが、「早く」は係り先の候補から除外されており解析が誤っている。

また、Rule3 の処理の例を次に示す。

(7) 商工会議所の会頭が、質問に 先頭を切って答えた。

KNP は、「質問に」の係り先の候補として、「切って」、「答えた」の 2 つの可能性を考慮する。この場合、“より近くに係る” という優先規則に従って係り先は「切って」に決定されるが、この解析は誤りである。この例のように、係り先の候補が複数存在すると、係り先に曖昧性があり確信度が低いので、このような用例は収集しない。

京都大学テキストコーパスから確信度の高い係り受けを抽出して、その精度の評価を行った。対象としている格をもつ格要素の係り受けの精度は 90.9%であるのに対し、抽出した確信度の高い係り受けの精度は 97.2%であった。抽出した係り受けは、対象としている格をもつ係り受け全体の 44.0%であった。これより、この処理はかなり効果的であることがわかる。

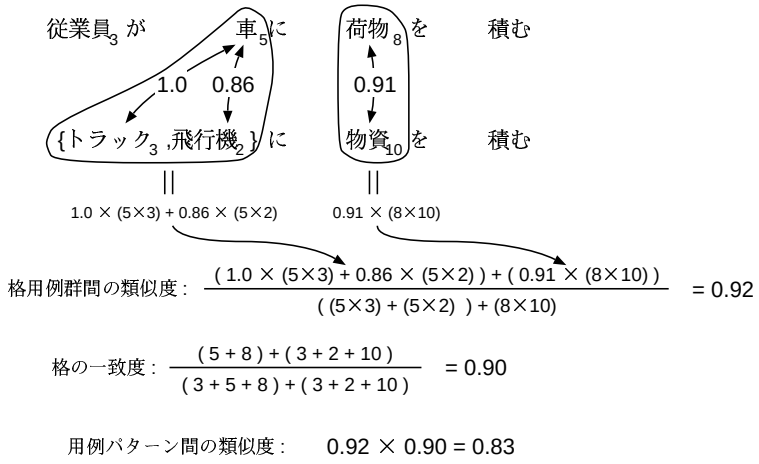

図 2: 用例パターン間の類似度の計算の例 (用例の右下の数字は頻度を示す。)

5　用例格フレームの作成

2 章の例文で示したように、用言の用法の異なる用例をひとつの格フレームとしてまとめてしまうと、誤った表現を許す格フレームを作ってしまう。従って、格パターンの異なる格フレームは別々に作成する必要がある。

用言の用法を決定する重要な格要素は用言の直前にくることが多い。また、用言とその直前の格要素をペアにして考えると、用言の用法はほとんど一意に決定される。 そこで、用例を、**用言とその直前の格要素の組を単位としてまとめる**という処理を行い、用例パターン (図 1) を作る。用例パターンの用言の直前の格要素を**直前格要素**、直前格要素の格を**直前格**と呼ぶ。

用例パターンは、ひとつの用言について、直前格要素の数だけ存在している。そのため、次の例のように、用法がほとんど同じパターンまで個別に扱われている。

(8) a.　従業員:が 車:に　　荷物:を 積む
　　b.　　{ トラック, 飛行機 }:に 物資:を 積む

そこで、ほとんど用法が同じ用例パターンをマージするために、用例パターンのクラスタリングを行う。以下では、このクラスタリングの詳細について述べる。

5.1　用例パターン間の類似度

用例パターンのクラスタリングは、用例パターン間の類似度を用いて行う。用例パターン間の類似度は、格の一致度と格用例群間の類似度の積とする (図 2 に類似度の計算の例を示す)。

まず、単語 e_1, e_2 間の類似度 $sim_e(e_1, e_2)$ を、日本語語彙大系のシソーラスを利用して以下

のように定義する。

$$sim_e(e_1, e_2) = max_{x \in s_1, y \in s_2}\, sim(x, y)$$

$$sim(x, y) = \frac{2L}{l_x + l_y}$$

ここで、x, y は意味属性であり、s_1, s_2 はそれぞれ e_1, e_2 の日本語語彙大系における意味属性の集合である (日本語語彙大系では、単語に複数の意味属性が与えられている場合が多い)。$sim(x, y)$ は意味属性 x, y 間の類似度であり、l_x, l_y は x, y のシソーラスの根からの階層の深さ、L は x と y の意味属性で一致している階層の深さを表す。類似度 $sim(x, y)$ は 0 から 1 の値をとる。

用例パターン P_1, P_2 の格の一致度 cs は、P_1, P_2 に含まれるすべての格用例に対する、P_1, P_2 の共通格に含まれている格用例の割合とし、

$$cs = \frac{\sum_{i=1}^{n} |E_{1cc_i}| + \sum_{i=1}^{n} |E_{2cc_i}|}{\sum_{i=1}^{l} |E_{1c1_i}| + \sum_{i=1}^{m} |E_{2c2_i}|}$$

と定義する。ただし、用例パターン P_1 中の格を $c1_1, c1_2, \cdots, c1_l$、用例パターン P_2 中の格を $c2_1, c2_2, \cdots, c2_m$、P_1 と P_2 間の共通格を $cc_1, cc_2, \cdots, cc_n$ とする。また、E_{1cc_i} は P_1 内の格 cc_i に含まれる格用例群であり、E_{2cc_i}, E_{1c1_i}, E_{2c2_i} も同様である。$|E_{1cc_i}|$ などの絶対値記号は頻度を表す。

用例パターン P_1, P_2 の共通格に含まれる格用例群間の類似度 $sim_E(P_1, P_2)$ は、格用例の類似度の和を正規化したもので、

$$sim_E(P_1, P_2) = \frac{\sum_{i=1}^{n} \sum_{e_1 \in E_{1cc_i}} \sum_{e_2 \in E_{2cc_i}} |e_1|\,|e_2|\; sim_e(e_1, e_2)}{\sum_{i=1}^{n} \sum_{e_1 \in E_{1cc_i}} \sum_{e_2 \in E_{2cc_i}} |e_1|\,|e_2|}$$

とする。

用例パターン P_1, P_2 間の類似度は、格の一致度 cs と P_1, P_2 の共通格の格用例群間の類似度の積とし、次のようにして計算する。

$$\text{類似度} = cs \cdot sim_E(P_1, P_2)$$

5.2　クラスタリングの手順

用例パターンのクラスタリングの手順を以下に示す。

1. まず、直前の格要素の出現頻度がある閾値以上あるという条件で足切りを行う。これは、直前の格以外にも格用例がある程度の回数以上出現しているような安定した用例パターンだけを対象にするためである。この閾値は 5 に設定した。
2. **直前格が同じ用例パターンのクラスタリング**
   (a) あらゆる 2 つ組の用例パターンの類似度を計算し、用例パターンの意味属性を固定する。これらの処理は、5.3 節で述べるように繰り返す。

(b)　用例パターン間の類似度が閾値を越える組について、用例パターンのマージを行う。

3. **直前格を限定しない用例パターンのクラスタリング**
   直前格が同じ用例パターンのクラスタリングでは、次の例のように、格パターンが同じで用言の用法もほとんど同じ用例パターンであっても、直前格が異なっていれば別の用例パターンとなってしまう。

   (9) a. { 物資, 貨物 }:を　　トラック:に　　積む
   　　b. { トラック, 飛行機 }:に { 荷物, 物資 }:を 積む

   このように、直前格が異なっていても格パターンがほとんど同じ格フレームをマージする必要がある。行う処理は、2 の処理で得られた用例パターンのクラスタリングである。類似度、閾値とも 2 と同じものを用いる。2 と異なる点は用例パターンの意味属性の固定を行わないことである。
4. **残りの用例パターンのふりわけ**
   頻度の閾値を越えない用例パターン (残りの用例パターン) をこれまでの処理で作成された用例パターンにふりわける。これまでと同様に用例パターン間の類似度を計算し、類似度が閾値を越え、もっとも類似している用例パターンにマージする。クラスタリング結果に対象としている用言の格フレームがないときは、残りの用例をひとつの格パターンとしてまとめる。

5.3　用例パターンの意味属性の固定

用例パターン間の類似度は、用例パターンの直前格要素の意味属性が大きく影響する。そのため、用例パターンの直前格要素に多義性があるときに問題がある。例えば、「合わせる」の用例パターンのクラスタリングにおいて、用例パターンの組 (手, 顔)[1]と (手, 焦点) がそれぞれマージされる。(手, 顔) は意味属性 <動物（部分）>、(手, 焦点) は意味属性 <論理・意味等> を共通にもつためである。この 2 つの用例パターンの組から結果的に (手, 顔, 焦点) という意味的におかしい組が作られてしまう。この問題は、「手」が複数の意味属性 <動物（部分）>、<論理・意味等> をもち、多義であるにもかかわらず、その多義性をまったく考慮せずに単純にクラスタリングしていることに起因している。

この問題に対処するために、もっとも類似度が高い用例パターンの組から意味属性を固定する処理、すなわち用例パターンの意味の曖昧性解消を行う。この処理は、用例パターンの直前格要素の意味属性を固定することによって、次のような手順で行う。

1. 類似度が高い用例パターンの組 (p, q) から順に、両方の用例パターンの直前格要素 n_p, n_q の意味属性を固定する。固定する意味属性は、n_p, n_q 間の類似度を最大にする意味属性 s_p, s_q とする。ここで扱う用例パターンは、直前格が同じものに限定する。

[1] ここでは、用例パターンを直前格要素で表している。たとえば、「手」は「手:を 合わせる」という用例パターンを意味している。

2. p, q に関係する用例パターンの類似度を再計算する。
3. 閾値を越える用例パターンの組がなくなるまで、この 2 つの処理を繰り返す。

次に、この処理の例を示す。用言「飛ぶ」について、直前格の単語が「声」、「怒声」、「機」、「質問」であり、用例パターン間の類似度がクラスタリングの閾値 (ここでは 0.65 とする) を越える組み合わせが以下の 4 通りであったとする。

(1) 声:<声>　怒声:<声>　0.90
(2) 声:<単位>　機:<単位>　0.78
(3) 声:<声>　質問:<質問>　0.69
(4) 怒声:<声>　質問:<質問>　0.68

この表より、もっとも類似度が高い用例パターンの組は (1) であり、「声」を直前格とする用例パターンと「怒声」を直前格とする用例パターンの類似度が 0.90 となっている。このとき、「声」の意味属性が <声> で、「怒声」の意味属性も <声> のときに、「声」、「怒声」という単語間の類似度、そしてこの用例パターン間の類似度が最大になっている。ここで、「声」の意味素を<声>、「怒声」の意味属性も <声> に固定する。「声」と「怒声」の意味属性が限定されたので、それらの用例パターンに関係する類似度 (2)、(3)、(4) の再計算を行う。再計算の結果、(3)、(4) の類似度は変わらないが、(2) は、

(2) 声:<声>　機:<単位>　0.29

となり、類似度 0.29 は閾値を下回り、結局この用例パターン間のクラスタリングは行われない。

6　必須格の選択

クラスタリングを行った結果得られる用例格フレームについて、格用例の頻度が少ない格は除く。これは、ひとつには構文解析結果の誤りへの対策であり、また頻度の少ない格はその用言と関係が希薄であると考えられるからである。ただし、ガ格についてはすべての用言がとると考え、頻度が少なくても削除せず、逆にガ格の格用例がない場合には、意味属性<主体>を補うことにした。

頻度の閾値は、現在のところ経験的に $2\sqrt{mf}$ と定めている。ただし、mf はその用言においてもっとも多く出現した格の延べ格用例数である。例えば、ある用言について、もっとも多く出現した格がヲ格で、$mf = 100$ であり、ニ格の格用例数が 16 であったすると、このニ格は頻度が 20 未満なので捨てられることになる。

7　作成した格フレーム辞書

毎日新聞約 9 年分の 460 万文から実際に格フレーム辞書を構築した。クラスタリングの閾値は 0.80 に設定した。これは、格パターンが違ったり、意味が違う格フレームが同じ格フレーム

にならないという基準で設定したものである。従って、格フレームは基本的にはばらばらで、意味がほとんど同じ格フレームを最小限まとめたものになっている。格フレームの例を表 1 に示す。この表では、<主体>、<場所> の意味属性をもつ格用例を【主体】、【場所】という意味属性でまとめて表示している。

71,000 個の用言について格フレームが構築され、用言あたりの平均格フレーム数は 1.9 個、格フレームあたりの格の平均数は 1.7 個、格あたりの平均異なり格用例数は 4.3 個であった。また、クラスタリングによって用例格フレーム数は用例パターン数の 53% になった。

構築した格フレーム辞書をみると、「賛成」のような名詞+判定詞の格フレームや、「ただす」の「について」のような複合辞の格についても得られている。また、「告知する」は、語順の問題への対処が有効に働いて、次の 2 つの分割する必要のない用例格フレームが 1 つにマージされている。

(10) a. <主体>:が 患者:に 告知する
　　b. 同僚:が { 患者, 本人, 家族 }:に 感染:を 告知する

8 解析実験

得られた格フレーム辞書の静的な評価は難しいので、それを用いた格解析を通して評価する。毎日新聞の記事 200 文をテストセットとし[2]、これに対して格解析を行った。格解析は (Kurohashi and Nagao 1994) の方法を用いた。格解析結果の評価は、提題と被連体修飾詞の格を正しく認識できるかどうかで行う。格解析の評価を表 2 に示す。ベースラインは、格フレーム辞書を用いずに、対象の用言がもっていない格をガ格、ヲ格、ニ格の順番に探して最初にみつかった格に決定するという処理を行ったものである。表 2 において、格解析の精度をみるために係り受けの誤りを除いて考えると、本手法では提題が 94%、被連体修飾詞が 78%、ベースラインでは提題が 90%、被連体修飾詞が 67% という精度であり、本手法はベースラインの精度を大きく上回っている。

解析結果の例を表 3 に示す。誤りの大きな原因は、「〜を与える役割」のような外の関係、「業界は〜という特徴がある」といったガガ構文である。この問題の対処は今後の課題である。

9 おわりに

本論文では、用言とその直前の格要素の組を単位として、生コーパスから用例を収集し、それらのクラスタリングを行うことによって、格フレーム辞書を自動的に構築する手法を提案した。得られた辞書を用いて実際に格解析を行った結果、提題、連体修飾の格の解釈をかなり高い精度で行うことができた。従って、実用レベルの格フレーム辞書を構築できたと考えられる。今後、この格フレーム辞書を用いて文脈解析を行う予定である。

[2] このテストセットは、格フレーム辞書の構築には用いていない。

表 1: 構築した格フレームの例 (* はその格が用言の直前の格であることを示す。)

用言	格	用例
買う 1	ガ格	【主体: <数量>人, 乗客, 幹部, 筋, 男性, 資産家, 政府, 銀行, … 】
	ヲ格*	株, 円, 土地, もの, ドル, 切符, 車, 物, 家, 株式, 国債, …
	デ格	【場所: 店, 駅】, <数量>円, 金, 価格, 会社, 仲介, 額, インターネット, …
買う 2	ガ格	対応, 厚生, 絵はがき, 蓄財, シーン, 工作, 禁止, 風刺画, …
	ヲ格*	怒り, ひんしゅく, 失笑, 反感, 恨み, 不興, 憤激, 嘲笑, …
:	:	:
読む 1	ガ格	【主体: 大学生, 首相, 先生, 若者, 女性 サラリーマン】, <数量>割, …
	ヲ格*	本, 記事, 新聞, 小説, 投書, 作品, 書, 文, 文章, 手紙, …
読む 2	ガ格	【主体: <主体>】
	ヲ格	話, <補文>, 意見, 惨状, ニュース, 事件, 記, 経緯, 記事, …
	デ格*	新聞, 本, 本紙, 教科書
読む 3	ガ格	【主体: <主体>】
	ヲ格*	先
:	:	:
ただす 1	ガ格	【主体: 氏, 委員, 議員, 委員長, 党首, 会長, 主席】, 両氏, 副総裁, 喚問, …
	ヲ格*	見解, 真意, 考え, 方針, 問題, 真偽, 意図, 策, 行方, 意向, …
	について	問題, <補文>, 展開, 責任, 影響, 停止, 法案, 見通し, 事例, …
ただす 2	ガ格	【主体: 委員長, 自ら, 業界】
	ヲ格*	【主体: 身】, 姿勢, 姿, 威儀
:	:	:
告知する 1	ガ格	【主体: 医師】
	ニ格*	本人
告知する 2	ガ格	【主体: 同僚】
	ヲ格*	感染, がん
	ニ格*	患者, 本人, 家族
:	:	:
賛成 1	ガ格	【主体: <主体>】
	ニ格*	意見, 考え, 主張, 認識, 論, 立場
賛成 2	ガ格	【主体: <主体>】
	ニ格*	<補文>

表 2: 提題、被連体修飾詞の格解析の評価

		正解	誤り			
			対応付けの誤り	外の関係による誤り	ガガ構文による誤り	係り受けの誤り
本手法	提題	85	3	–	2	13
	連体修飾	50	5	9	–	2
ベースライン	提題	81	7	–	2	13
	連体修飾	43	6	15	–	2

表 3: 格解析の結果の例

(1) 大蔵省は$_{1:ガ格}$ 九日、信託銀行の不良債権の処理を促進するため、一九九五年三月期決算で信託銀行各行が [積み立てている]$_2$ 特別留保金の$_{2:ヲ格}$ 取り崩しを [認める]$_3$ 方針を$_{3:ニ格⇒\ 外の関係}$ [決めた。]$_1$

(2) 特に 日本信託銀行は$_{1:ガガ}$ 不動産融資の焦げ付きで信託勘定の不良債権が [膨らんでおり、]$_1$ 大蔵省は$_{2:ガ格}$「特別留保金を取り崩して不良債権処理を促進することは$_{3:ガ格}$ 顧客保護にも [通じる]$_3$」と [判断した。]$_2$

(3) 新民連は$_{1:ガ格}$ これに [先立ち]$_1$、衆参院議員二十九人が参加して総会を開き、山花氏が、今後の新党問題の 協議は$_{2:ヲ格}$ 準備会で [進める]$_2$ 考えを$_{2:外の関係}$ 表明し、新民連は$_{3:ガ格}$ 事実上、活動停止状態に [なった。]$_3$

(4) 金権選挙追放策の一つとして、戦後 [廃止されてしまった]$_1$ 民衆訴訟による当選無効制度の$_{1:ガ格}$ 復活も$_{2:ヲ格}$ [試みる……。]$_2$

(5) これらの 業界は$_{1:ヲ格⇒\ ガガ}$、比較的外圧を受けにくく、また政治的発言力が強い、という特徴が [ある。]$_1$

(6) 宇宙誕生のなぞや物質の重さの起源に迫ろうという世界最大の素粒子加速器建設計画が、十九カ国が [加盟する]$_1$ 欧州合同原子核研究所の$_{1:ニ格}$ 理事会で本決まりとなった。

(7) そして物質に重さを [与える]$_1$ 役割を$_{1:外の関係}$ [担う]$_2$ ヒッグス粒子の$_{2:ガ格}$ 発見などを目指している。

(8) しかし、日韓正常化の韓国での 歴史評価は$_{1:ヲ格}$、韓国の人々に [まかせるべきであろう。]$_1$

(9) 代表質問を“ 影の内閣 ”として [設置した]$_1$ 政権準備委員会の$_{1:ニ格⇒\ ガ格}$「施政方針演説」と位置付け、政権担当能力をアピールするのが狙い。

下線部は提題または被連体修飾詞を表し、四角形で囲まれた部分は用言を表している。四角形には用言の番号を付与してある。下線部の後に、係り先の用言を示す番号と、格解析によって認識された格を記述し、格解析が誤っているときは ⇒ の後に正解の格を記述した。

参考文献

Brent, M. R. (1991). “Automatic Acquisition of Subcategorization Frames from Untagged Text.” In *Proceedings of the 29th Annual Meeting of ACL*, pp. 209–214.

Briscoe, T. and Carroll, J. (1997). “Automatic Extraction of Subcategorization from Corpora.” In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 356–363.

Collins, M. J. (1996). “A New Statistical Parser Based on Bigram Lexical Dependencies.” In *Proceedings of the 34th Annual Meeting of ACL*, pp. 184–191.

Kurohashi, S. and Nagao, M. (1994). “A Method of Case Structure Analysis for Japanese Sentences based on Examples in Case Frame Dictionary.” In *IEICE Transactions on Information and Systems*, Vol. E77-D No.2.

Manning, C. D. (1993). “Automatic Acquisition of a Large Subcategorization Dictionary from Corpora.” In *Proceedings of the 31th Annual Meeting of ACL*, pp. 235–242.

春野雅彦 (1995). “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.” 電子情報通信学会 言語理解とコミュニケーション研究会 NLC95-11, pp. 29–36.

黒橋禎夫, 長尾眞 (1994). “並列構造の検出に基づく長い日本語文の構文解析.” 自然言語処理, **1** (1).

宇津呂武仁, 宮田高志, 松本裕治 (1997). “最大エントロピー法による下位範疇化の確率モデル学習および統語的曖昧性解消による評価.” 情報処理学会 自然言語処理研究会 97-NL-119, pp. 69–76.

東優, 峯恒憲, 雨宮真人 (1996). “既存の概念辞書を用いた動詞語義による文の分類.” 電子情報通信学会 言語理解とコミュニケーション研究会 NLC96-36, pp. 39–44.

略歴

河原 大輔: 1997 年京都大学工学部電気工学第二学科卒業。1999 年同大学院修士課程修了。現在、同博士課程在学中。構文解析、文脈解析の研究に従事。

黒橋 禎夫: 1989 年京都大学工学部電気工学第二学科卒業。1994 年同大学院博士課程修了。京都大学工学部助手、京都大学情報学研究科講師を経て、2001 年東京大学大学院情報理工学系研究科助教授、現在に至る。自然言語処理、知識情報処理の研究に従事。

(2002 年 1 月 1 日 受付)
(2002 年 1 月 1 日 再受付)
(2002 年 1 月 1 日 採録)