格フレーム辞書の漸次的自動構築

河原 大輔[†]　　黒橋 禎夫[†]

本稿では、格フレーム辞書を漸次的に自動構築する手法を提案する。カバレージの高い格フレーム辞書を構築するために、大規模コーパスから徐々に確からしい情報を抽出する。まず、コーパスを構文解析し、構文的曖昧性のない述語項構造のみを抽出・クラスタリングすることによって、1 次格フレーム辞書を得る。次に、1 次格フレーム辞書を用いてコーパスを格解析し、新たに分かる確実な情報を抽出し、2 次格フレーム辞書を構築する。このように徐々に新たな情報を加えていくことによって、高次格フレーム辞書を構築する。結果として得られた格フレーム辞書は、二重主語構文、連体修飾の外の関係、格変化といった複雑な言語現象を解析することを可能にする。新聞記事 26 年分、約 2600 万文のコーパスから格フレーム辞書を構築し 2 種類の評価を行った。1 つは、得られた格フレームを人手で評価するものであり、もう 1 つは得られた格フレーム辞書を用いた構文・格解析実験による評価である。これらの結果、本手法の有効性が確かめられた。

キーワード：　格フレーム, コーパス, 格解析

Gradual Fertilization of Case Frames

Daisuke Kawahara[†]　　Sadao Kurohashi[†]

This paper proposes an automatic method of gradually constructing case frames. First, a large raw corpus is parsed, and initial case frames are constructed from reliable predicate-argument examples in the parsing results. Second, case analysis based on the initial case frames is applied to the large corpus, and the case frames are upgraded by incorporating newly acquired information. Case frames are gradually fertilized in this way. We constructed case frames from news paper articles of 26 years, consisting of 26M sentences. The case frames are evaluated by hand, and furthermore evaluated through syntactic and case analysis. These results presented the effectiveness of the constructed case frames.

KeyWords:　*case frame, corpus, case analysis*

1　はじめに

計算機で文章を理解するためには、少なくとも、文章においてどの単語とどの単語がどのような関係をもっているかを明らかにする必要がある。このような単語間の関連性を解析するためには、人間がもっている常識のような幅広い知識が必要となる。

[†] 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

このような知識のひとつとして格フレームと呼ばれるものがある。格フレームとは、用言とそれがとる格要素の関係を記述したものであり、例えば「積む」という用言の格フレームのひとつとして次のようなものが考えられる。

(1)　{ 従業員, 運転手, ···} が　{ 車, トラック,···} に　{ 荷物, 物資 } を　積む

我々は、大量のコーパスを構文解析し、その解析結果から構文的曖昧性のない用言と格要素の関係を抽出、クラスタリングすることによって、格フレーム辞書を自動的に構築する手法を提案した (河原 黒橋 2002)。格フレーム構築における大きな問題は用言の多義性であり、「積む」の場合では「荷物を 積む」「経験を 積む」のような意味の異なる表現を区別する必要がある。我々の手法は、用言とその直前の格要素を組とすることによってこの問題に対処している。例えば「積む」の場合、「荷物を 積む」「物資を積む」「経験を 積む」「体験を積む」などの組で扱うことによって、上記 (1) の格フレームと下記 (2) の格フレームが別々に構築される。

(2)　{ 選手, 従業員, ···} が　{ 経験, 体験 } を　積む

上記の手法は、構文的曖昧性のない用例のみを用いるために、基本的に格助詞の付属している格要素を収集している。このため、得られる格フレームは、次に挙げる下線部のような複雑な言語現象には対処できないという問題がある。

(3)　象は 鼻が長い

(4)　さんまを焼く けむり

(5)　自民党の 支持を得る

(3) の文は二重主語構文であり、下線部は 2 つ目のガ格 (「外のガ格」) である。(4) の「けむり」は、「焼く」に対してガ格、ヲ格などの直接的な関係をもたず、「外の関係」と呼ばれる関係をもっている。(5) は、人間ならば「自民党から 支持を 得る」と解釈でき、「の」と「から」が交換可能 (ノ格とカラ格の格変化) となっている。

本研究では、これらの現象の解析を実現するために、コーパスから信頼性の高い情報を段階的に抽出し、格フレーム辞書の自動構築を行う。まず、コーパスを構文解析し、確信度の高い述語項構造のみを抽出・クラスタリングすることによって、1 次格フレーム辞書を得る。次に、1 次格フレーム辞書を用いてコーパスを格解析し、新たに分かる確実な情報を抽出し、格フレーム辞書を高度化する。格解析によって新たに抽出される情報は、二重主語構文、連体修飾の外の関係である。また、格変化の問題に対処するために、格フレームの格間の類似性に基づいて、格のマージを行う。

2　対象とする表現

本章では、格フレーム構築や格解析において問題となる表現について説明する。

日本語では、「が」「を」「に」といった助詞が体言-用言間の関係を表している。しかし、次に挙げる表現の関係は表層に現れない。

係助詞句 (「は」「も」といった係助詞が付属した句)

(6) a. 車は速い　(ガ格)
b. 本も読んだ　(ヲ格)

被連体修飾詞 (連体修飾を受ける句)

(7) a. 速い車　(ガ格)
b. 読んだ本　(ヲ格)

上例の下線部が用言に対してもつ関係はそれぞれの右の括弧内に示した格である。

係助詞句、被連体修飾詞が、単純なガ格、ヲ格などではない場合を以下に示す。

二重主語構文
「〜[係助詞] … 〜が [用言]」という形の表現で、係助詞句がガ格となり、用言がガ格を 2 つもつ場合

(8) この車はエンジンがよい

この例において「車がよい」といえるので「車」の格はガ格である。従って、「よい」は「車」と「エンジン」の 2 つのガ格をとり、「車」は外のガ格である。本論文では、このような外のガ格を「ガ 2 格」と表す。

外の関係
被連体修飾詞と用言との関係を表層格で表すことができない場合

(9) 魚を焼くけむり

「けむり」は、魚を焼くときに発生する「けむり」という意味であり、「けむり」と「焼く」の関係はどのような表層格でも表すことができない。

格変化
同じ意味を表すためにも様々な格が用いられる

(10) a. 社会党が新進党の支持を得る
b. 社会党が新進党から支持を得る

この例では、「新進党」は「得る」とカラ格の関係をもっている。これをノ格とカラ格の格変化と呼ぶ。

(11) a. この車のエンジンがよい
b. この車はエンジンがよい

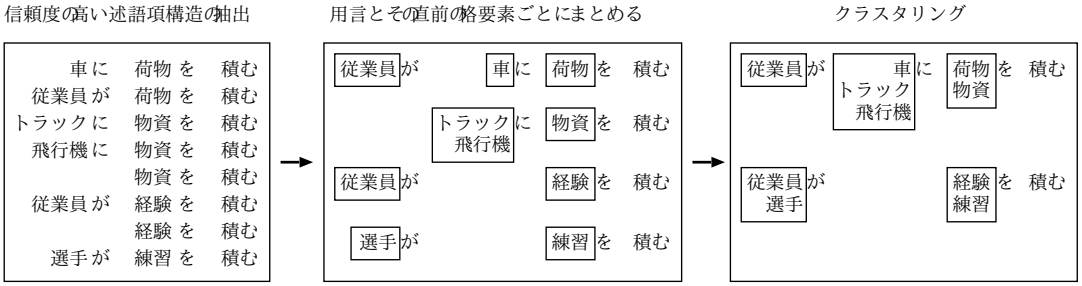

図 1: 1 次格フレーム辞書の構築

例 (8) のようなガ 2 格の係助詞句は、このようにノ格でもいえることが多い。これはノ格とガ 2 格の格変化である。

本研究では、これらの例のように、用言の直前の格要素に係るノ格の要素を用言の格要素と同等に扱う。これらのノ格の要素は、係り先は用言ではないが、上例のような格変化を起こすことがあり、用言との意味的なつながりが強いと考えられるからである。また、用言とその直前格要素の組からの距離を考えると、他の格要素と同じく係り受けの距離が 1 であり、その位置に重要な句が出現することが多い。

3　1 次格フレーム辞書の構築とそれを用いた格解析

我々は、用言の直前の格要素と用言を組にすることによって、大規模コーパスから格フレーム辞書を構築した (河原・黒橋 2002)。この格フレーム辞書を 1 次格フレーム辞書と呼び、本章ではまず、この格フレーム辞書の構築手法の概略を述べる。次に、1 次格フレーム辞書を用いた格解析手法について説明する。

3.1　1 次格フレーム辞書の構築

大規模コーパスから格フレーム辞書を自動構築する際の最大の問題は、用言の用法の曖昧性である。つまり、同じ表記の用言でも複数の意味、用法をもち、とりうる格や用例が異なる。例えば、「トラックに 荷物を 積む」と「経験を 積む」は、用言は「積む」で同じであるが用法が異なっている。用法が異なる格フレームを別々につくるために、我々は、格フレーム収集の単位を用言とその直前の格要素の組とした。「積む」の例では、「荷物を積む」「経験を積む」を単位として格フレームを収集する。さらに、「荷物を積む」「物資を 積む」などかなり類似している格フレームをマージするためにクラスタリングを行う。

この格フレーム辞書構築の手順を以下に示す。

1. KNP (Kurohashi and Nagao 1994b) を用いてコーパスを構文解析し、構文解析結果から構文的曖昧性のない述語項構造を抽出する。
2. 抽出した述語項構造を用言とその直前の格要素ごとにまとめ、その組の頻度が 5 回以上ある述語項構造から (最初の) 格フレームをつくる。ただし、係助詞句と被連体修飾詞は、この段階ではそれらの関係が解析できないため用いない。以後、用言の直前の格要素を「直前格要素」、その格を「直前格」と呼ぶ。
3. 2 でつくった格フレームをクラスタリングし、類似しているものをマージする。クラスタリングは格フレームの類似度 (付録) に基づいて行う。

「積む」という動詞の格フレーム構築例を図 1 に示す。手順 1 における構文的曖昧性のない述語項構造とは、KNP における優先規則によって係り先の候補がただ 1 つしかないということを表す。例えば、次に挙げた例では、下線部が構文的曖昧性のない述語項構造として抽出される。

(12) a. 多くの 練習を積んだが、予戦は…

b. 荷物を積んだトラック が道路を走りながら…

例 (12a) では、「〜が」が強い区切りと認識され、「練習を」の係り先候補は「積んだ」の 1 つのみとなり、この部分が抽出される。例 (12b) では、「荷物を積んだ」という連体修飾節の係り先候補は直後の「道路」しかないので、この部分が抽出される。

京都大学テキストコーパス (Kurohashi and Nagao 1998) を用いて、構文的曖昧性のない述語項構造抽出の精度評価を行った。述語項構造すべての係り受け精度が 90.9% であるのに対し、抽出した述語項構造の精度は 98.3%であった。抽出した述語項構造は、述語項構造全体の 20.7%であった。誤りとしては、並列構造や判定詞を認識できなかったために曖昧性がないと誤って判定され、誤抽出されていることが多かった。

本研究では、新聞記事 26 年分 (毎日新聞 12 年分、読売新聞 14 年分) のテキストから 1 次格フレーム辞書を自動構築した。この辞書には、約 18,000 個の用言が含まれており、1 用言あたりの平均格フレーム数は約 17.9 個である。

3.2 1 次格フレーム辞書に基づく格解析

本章では、1 次格フレーム辞書を用いて格解析を行う手法について説明する。格解析の基本的な処理は、入力文の格要素と格フレームとの対応付けである。格解析は構文解析と統合的に行われる。つまり、入力文のとりうる構文構造のひとつひとつについて格解析を行い、それぞれの構造の妥当性を格解析のスコアで評価し、もっともスコアの高い構文構造、格解析結果を出力する (Kurohashi and Nagao 1994a)。

格解析は文中のそれぞれの用言に対して行われ、格フレームの選択と格の対応付けという 2 つの処理からなる。以下では、それぞれの処理について詳細に述べる。

格フレームの選択

格フレーム辞書には用言ごとに複数の格フレームが存在するので、入力文の用言の用法に合致する格フレームを選択する必要がある。3.1 節で述べたように、用言の用法の決定に対して、用言の直前格要素が重要な役割を果たす。特に、直前格がヲ格、ニ格の場合はその傾向が強い。一方、直前格要素が意味属性<主体>[1]に属する場合、例えば「<主体>が　求める」という表現からは、用言の用法が決まらず、格フレームを選択することができない。これらの点を考慮して、以下の条件を満たす格フレームを選択する。

入力側の条件　入力側の対象用言が直前格要素 C をもち、直前格要素 C と直前格 cm が以下のいずれかの条件を満たす。

- cm がヲ格、ニ格のいずれかである。
- cm がヲ格、ニ格以外で、C が意味属性<主体>をもたない。

格フレームの条件　cm をもち、cm の用例群と C の類似度が閾値以上ある格フレームを選択する。閾値以上ある格フレームが複数ある場合は、類似度がもっとも高い格フレームを選ぶ。

ここで用いる類似度は、格フレームの直前格の各用例と直前格要素の類似度のうちもっとも高いものとする。用例間の類似度は付録の式 (1) を用いる。類似度の閾値は経験的に 0.70 とした。

上記の条件を満たす格フレームが存在しない場合は、入力用言のすべての格フレームを格の対応づけ処理の対象とする。このようにして選択した格フレームのそれぞれについて格の対応づけを行い、最後にもっとも対応づけスコアが高かった格フレームに決定し、そのときの対応付けを出力する。

入力側の格要素と格フレームの格との対応付け

選択された格フレームについて、入力側の格要素と格フレームの格との対応づけを行う。格要素に格助詞が付属している場合は、その格助詞の格に対応する格フレーム側の格に対応づける。被連体修飾詞や係助詞句のように、文中から格がわからない場合は、次表の格それぞれへの対応を試し、対応づけ全体のスコアがもっともよい対応を選択し格を決定する。対応づけ全体のスコアは、各格の対応の類似度を足したものとする。各格の対応の類似度は「格フレームの選択」で用いた類似度と同様に計算する。

係助詞句	：ガ, ヲ
被連体修飾詞	：ガ, ヲ, ニ

ただし、被連体修飾詞が「こと」「ところ」「ため」のような副詞的名詞、形式名詞の場合は、外の関係である場合がほとんどであるので、上記の格には対応させず外の関係と解析する。

以下に格解析の例を挙げる。

(13)　書類は業者に渡した

[1] 本論文では、シソーラスにおける意味属性を<>で表す。

この例に対しては、次の格フレームが合致するため選択される[2]。

	格	用例	入力側格要素
渡す	ガ	長, 妻, 駅員, 事務員	-
	ヲ	<数量>円, コピー, 書, …	書類
	ニ*	業者, 会社, 企業, 銀行	業者

まず、「業者に」は格フレームのニ格に対応する。次に「書類」については、ガ格、ヲ格それぞれの用例群と類似度を計算し、より類似しているヲ格に対応付けられる。

(14)　ドイツ語も話せる先生

この例は格フレームの選択条件を満たさないので、「話せる」のすべての格フレームについて格の対応付けが行われる。その結果、対応づけ全体のスコアがもっともよい次の格フレームとその対応づけに決定される。

	格	用例	入力側格要素
話せる	ガ	親, 教師, 隊員, …	先生
	ヲ*	語, 外国語, 国語	ドイツ語

4　高次格フレーム辞書の構築

1 次格フレーム辞書を用いてコーパスを格解析し、新たに分かる確実な情報を抽出することによって格フレーム辞書を高度化する。高度化した格フレーム辞書を高次格フレーム辞書と呼び、その構築の概略を図 2 に示す。格解析によって新たに抽出される情報は、二重主語構文の外のガ格 (ガ 2 格)、連体修飾の外の関係である。また、格変化の問題に対処するために、格フレームの格要素の類似性に基づいて、格の類似判定という処理を行う。以下では、それぞれの処理について詳説する。

4.1　ガ 2 格の用例獲得

「その問題は彼が図書館で調べている」という例を 1 次格フレームを用いて格解析することを考える。この例に合致する「調べる」の 1 次格フレームは「{ 問題, 課題 } を { 図書館 } で 調べる」であり、「問題は」は、この格フレームのヲ格用例群にマッチするため、ヲ格と解釈することができる。これに対して、例 (8) の「この車はエンジンがよい」という表現について、合致する 1 次格フレームは「{ エンジン } が　よい」である。この格フレームを用いてこの表現を格解析すると、格フレームにガ格以外の格がないことから「車は」はガ 2 格であり、「{ エンジン } が　よい」は二重主語構文をとることがわかる。

[2] 格フレームの格に付属している “*” は、その格が用言の直前格であることを示す。

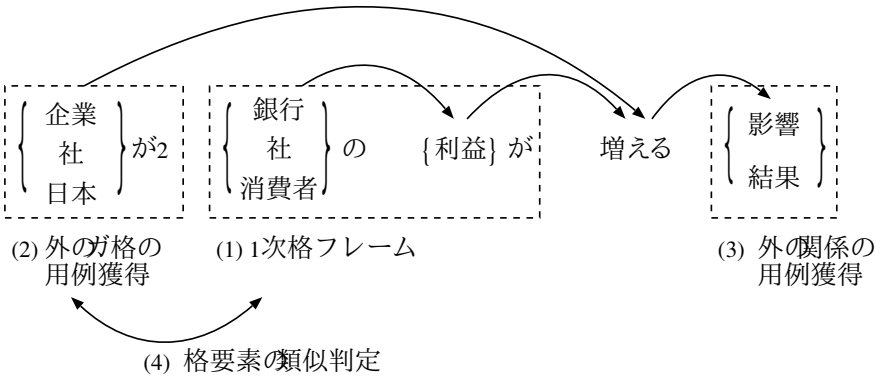

図 2: 高次フレーム辞書構築の概略

このような考え方に基づき、ガ 2 格の用例獲得を行う。まず、「〜[係助詞] …〜が [用言]」のような表現をコーパスから収集する。収集は、1 次格フレーム辞書構築と同様に、係助詞句が用言に曖昧性なく係る場合のみとする。収集したそれぞれの表現について格解析を行い、係助詞句の対応付け可能な格が格フレームになければ、その係助詞句をガ 2 格であると判定し、その格フレームのガ 2 格の用例として追加する。格フレームに対応付け可能な格が存在すれば、その係助詞句はそのいずれかの格である可能性が高く、ガ 2 格の用例としない。

上記の処理によって、次の例文と格フレームからは「相撲」がガ 2 格の用例として収集される。

(15)　長い 相撲は 足腰に負担が かかる。

	格	用例	入力側格要素
かかる	ガ*	負担	負担
	ニ	心臓, 体, 足, 足腰, …	足腰

この処理によって、733 用言、16,431 格フレームについてガ 2 格が作成された。この結果の格フレーム辞書を 2 次格フレーム辞書と呼ぶ。

4.2　外の関係の用例獲得

外の関係の単語は、「焼く」に対する「けむり」のような格フレーム単位のものもあれば、「可能性」「結果」など、どの用言に対しても外の関係になるものもある。それぞれの用例を獲

得するために以下の 2 つの処理を行う。

格フレームごとの外の関係の用例獲得

被連体修飾詞が外の関係のとき、これまでの格フレームには内の関係の用例しかないので、どの格とも類似しないはずである。この考えに基づいて、外の関係の用例を獲得する。まず、連体修飾を含む文をコーパスから収集する。コーパスからの収集は、これまでの格フレーム辞書構築と同様に、用言が被連体修飾詞に曖昧性なく係る場合のみとする (「〜した A の B」のような表現は収集しない)。収集したそれぞれの表現について、2 次格フレーム辞書を用いて格解析を行い、被連体修飾詞が、対応付け可能な格のいずれに対しても類似度が閾値 (0.30 とする) を越えなければ、この格フレームの外の関係の用例とする[3]。類似度は、被連体修飾詞のみからなる格を仮定し、付録の式 (2) を用いて計算する。ただし、次の場合はこの処理の対象外とする。

- 2 次格フレームのガ格に用例がないとき
  この場合、被連体修飾詞がガ格に対応付くかどうかが判断できないので対象外とする。
- 被連体修飾詞が「こと」「ところ」「ため」といった副詞的名詞、形式名詞のとき
  この場合の被連体修飾詞はほとんど外の関係であり、KNP では常に外の関係とみなしているため用例収集の対象外とする。
- 被連体修飾詞が意味属性<主体>, <時間>, <数量>のいずれかに属する
  これらの場合は、被連体修飾詞が外の関係をとることが少ないので、あらかじめ処理の対象外とする。次の例では、被連体修飾詞の「人」は<主体>に属し、「営む」に対してガ格である。

  (16)　養畜の業務を 営む 人 が加入資格者とされている。

処理の例を以下に示す。

(17)　売り出し業務を 営む 免許 を取得した。

この例を格解析すると、下表の左側のような格フレームが選択される。

	格	用例	入力側格要素
営む	ガ	銀行, 会社, センター	-
	ヲ*	業務, ビジネス	業務

対応付いていない格はガ格であり、「免許」はガ格のどの用例とも類似度が高くない。従って、「免許」をこの格フレームの外の関係の用例にする。

(18)　違法に国際電話業務を 営んでいた 疑い

この例も、上例と同じ格フレームが選択される。同様に、「疑い」はガ格のどの用例とも類似度が高くないので、「疑い」をこの格フレームの外の関係の用例にする。

用言全体の外の関係の用例獲得

[3] この格解析では、被連体修飾詞の対応づけ先候補としてガ 2 格も考慮するので、ガ 2 格の用例を含む 2 次格フレーム辞書を用いる。

多くの用言に分布する外の関係の単語を、どの格フレームにおいても外の関係になるとみなし、すべての格フレームの外の関係の用例に追加する。例えば、例 (17) の「免許」は、2 用言に対して外の関係になっているのに対し、例 (18) の「疑い」は 303 用言に対して外の関係になっている。そこで、「疑い」は用言全体について外の関係の単語であると考える。

本研究では、100 用言以上に分布している外の関係の単語を対象とした。これは 73 個あり、以下にその例を示す。

可能性, 必要, 結果, 方針, 形, ケース, 考え, 予定, 見通し, 計画, 見込み,…

上記処理によって、格フレームごとの外の関係の用例が 828 用言、34,243 格フレームに対して得られた。この結果の格フレーム辞書を 3 次格フレーム辞書と呼ぶ。

4.3　格の類似判定

3 次格フレームのそれぞれに対して格同士の類似度をとる。類似度が閾値を越える格の組は類似していると判定し、それらの格は交換可能と考える。ただし、(ガ格, ヲ格)、(ガ格, ニ格)、(ヲ格, ニ格) などの基本的な格の組については対象外とする。2 つの格間の類似度は、付録の式 (2) を用いて計算する。類似度の閾値は 0.80 とした。

この処理を以下の格フレームに適用した場合について説明する。

	格	用例
求める	ガ	委員会, 団, 氏, …
	ヲ*	説明, 釈明
	└ノ	経緯, 理由, 内容, …
	ニ	政府, 学会, 社長, …
	ニツイテ	経緯, 実態, 状況, …

「└ノ」で示した格は、用言の直前格要素に係るノ格である。この格フレームではノ格と「について」の用例が似ており、類似度が 0.94 と非常に高い。そこで、これらの格が交換可能であるという情報を格フレームに追加する。

上記の処理により、類似していると判定された格の組が 1,449 個得られた。この結果得られる格フレーム辞書を高次格フレーム辞書と呼ぶ。

5　格フレーム辞書の後処理

本研究では、大規模新聞コーパスから格フレーム辞書を構築している。コーパスは大規模であるが、それでもデータスパースネスの影響を受ける。特に、新聞ドメイン以外の用言 (例えば、「食べる」「座る」のような日常的な用言) においては、とるべき格をもたない不完全な格フレームを含んでいる。そこで、得られた格フレーム辞書に対して、不完全な格フレームの修正・

削除を行う。また、用例頻度の少ない格はその用言との関係が希薄であると考えられるので、用例の頻度がある程度以上ある格のみを選択することによって、格フレームの必須格を選択する。

5.1 不完全な格フレームの修正・削除

これまでの処理で構築された格フレームには、とるべき格をもたない不完全なものがある。例えば、「加熱」という動詞はヲ格をとるはずであるが、一部の格フレームはヲ格をもっていない。この現象は、データスパースネスが原因であり、低頻度の格フレームに多くみられる。この問題に対処するため、用言ごとに格フレーム全体をチェックして、その用言がとることの多い格を調べ、そのような格をもたない格フレームは削除する。ただし、この処理はガ格以外を対象とし、ガ格についてはすべての格フレームがとるとし、ガ格をもたない格フレームには意味属性<主体>を補い格フレームを修正する。

用言が格 cm をとるかどうかの判定は、格 cm の用例を 1 つ以上もつ格フレームが、その用言の格フレーム全体において占める割合 (格割合と呼ぶ) を計算し、この値が閾値を越えるかどうかで行う。格割合は格フレームを構成している述語項構造の頻度をもとに計算する。格割合をガ格以外のすべての格について計算し、閾値を越える格については、その格をもたない格フレームを削除する。閾値は 0.90 とした。

例えば、「引き渡す」という動詞の場合、ヲ格割合が 1314/1342 (0.979) となり、以下に挙げるヲ格をもたない格フレームが削除される。

- { 飼い主 } ガ { 県庁 } デ { 米子 } ノ { 保健所 } ニ* 引き渡す (頻度:7)
- { 台湾, 警察庁,..} ノ { 船 } ニ* 引き渡す (頻度:6)
- { 県 } ノ { 動物園 } ニ* 引き渡す (頻度:6)

このような不完全な格フレームが作られていたのは、格フレームを作るもととなった文の頻度が 6 回から 7 回と低く、省略などによってヲ格の用例がひとつも出現しなかったからである[4]。

高次格フレーム辞書に対してこの処理を適用した結果、1 用言あたり平均 3.2 個の不完全な格フレームが削除された。また、ガ格に意味属性<主体>を補った格フレームは 1 用言あたり平均 7.8 個であった。

5.2 必須格の選択

格フレームごとに必須格を選択する。必須格は、基本的には用例の頻度がある程度以上ある格とし、以下の基準に従って選択する。

- ガ格についてはすべての用言がとると考え、用例が 1 つでもあれば選択する。
- 格割合の高い格は、用例が 1 つでもあれば選択する。

[4] 直前格要素が 5 回未満の述語項構造は、1 次格フレーム辞書構築時に使用していない。

- ガ 2 格は、格フレームに必須と考えられるので、用例が 1 つでもあれば選択する。外の関係は、必須的ではないが格フレームに固有と考えられるので、ガ 2 格と同様に、用例が 1 つでもあれば選択する。
- 直前格の用例の頻度 mf に対して、用例数が $2\sqrt{mf}$ より多い格を選択する。直前格は常に選択する。

場所を表すデ格など任意的な格でも、その格フレームに閾値以上の頻度の用例が含まれるならば、その格をとる傾向が強いと考えられるため、必須格として選択している。

以下に格の選択の例を挙げる。これは「引き渡す」の格フレームの 1 つで、ガ格は無条件に、ニ格は直前格のため必須格として選択される。ヲ格については、「引き渡す」のヲ格割合が高いために選択される。

	格	頻度	用例
引き渡す	○ ガ	2	生徒, 者
	○ ヲ	6	者, 男, 会社員, 乗客
	○ ニ*	302	署員, 駅員, 隊員
	└ノ	2	消防, 駅
	デ	25	現行犯, 駅, ホーム
	ヲツウジテ	1	駅員
	トシテ	3	現行犯

1 格フレームあたり平均 5.4 個の格が存在していたが、必須格の選択処理によって平均 3.2 個の必須格が選択された。

6　格フレーム辞書の評価

自動構築した格フレームに対して 3 つの評価を行った。1 つは、本研究で新たに獲得できた外のガ格、外の関係、格の類似判定についての人手評価である。さらに、自動構築した格フレーム自体を人手で評価した。また、自動構築した格フレームの有効性を確認するために、構文・格解析実験を行った。

6.1　高次格フレーム辞書構築で得られた用例の評価

ガ 2 格を含む格フレーム、外の関係を含む格フレームそれぞれ 20 個ずつを選択し、それぞれの用例について人手で評価を行った。また、類似している格のペアがある格フレームを 20 個選択し、そのペアが正しいかどうか人手で評価を行った。それぞれについて以下で述べる。

ガ２格の用例評価

ガ２格として収集された用例が、2 章のガ２格の記述に合致するかどうかを人手で判定した。評価対象とした 20 個の格フレームのガ２格中に、367 個、259 異なりの用例が含まれており、そのすべての用例が正しいと判定された。評価を行った格フレームのうちの 2 つを以下に挙げる。

	格	用例
低い	ガ*	背, 身長, 高度, 丈, 高さ
	ガ２	彼, 人, 若者, 私, 男, 族, 僕, ベッド, 馬, おれ, 我が家
続く	ガ*	猛暑, 暑さ, 残暑, 酷暑, 暖かさ
	ガ２	沖縄, 州, 県内, 西日本, 北海道, 界, 船内, 信州, フィリピン

外の関係の用例評価

外の関係として収集された用例が、2 章の外の関係の記述に合致するかどうかを人手で判定した。評価対象とした 20 個の格フレームの「外の関係」の中に、330 個、73 異なりの用例が含まれており、そのうち 298 個 (90.3%)、55 異なり (75.3%) の用例が正しいと判定された。評価を行った格フレームのうちの 2 つを以下に挙げる。

	格	用例
盗む	ガ	子供, 者
	ヲ*	カード, 通帳, 券, 商品券, 免許証, 手帳, 証書, 切手, 小切手, 印紙
	└ノ	名義, <数量>円, 同僚, 銀行, 客, 男性, 女性, 他人, <数量>枚, …
	カラ	車, 宅, 乗用車, 事務所, 財布, 部屋, バッグ, 人, 客, 背広, 金庫, …
	外の関係	疑い, 手口
反対	ガ	遺族, 団体, 知事, 省庁, 大蔵省, 大使
	ニ*	移転, 動き, 移設, 移行, 移管
	└ノ	軍港, 基地, 演習, 病院, 機能, 施設
	外の関係	集会 ×, 会議 ×, 問題, 決議, 宣言, キャンペーン, 請願, …

上側の「盗む」の格フレームについては、「疑い」「手口」が外の関係の用例として獲得されている。「疑い」は一般的外の関係名詞にもなっており、「手口」はこの格フレームに対して獲得された用例である。下側の「反対」の格フレームの外の関係では、「集会」「会議」という 2 つの用例が誤りと判定されている。これらの表現は次の文から収集されている。

(19) a. 都機能移転に反対する都は、きょう十七日午後二時半から、東京体育館で「首都移転に断固 反対する 国民大 集会」を開催する。

b. 同飛行場と那覇軍港の県内移設に 反対する 県民 会議 が主催。

「集会」「会議」は両方とも「反対」に対する正しい関係はガ格である。しかし、両方ともももっとも似ているガ格の用例は「団体」であり、日本語語彙大系では類似度がそれぞれ 0.17、0.14 と高くないため外の関係に判定されている。これを解決するには、よりよいシソーラスを用いるか、ガ格の用例として「集会」「会議」のようなものが収集されるほどコーパスを大きくすることが必要である。

格の類似判定の評価

評価対象とした 20 個の格フレームにおける類似している格のペアが正しいかどうかを人手で評価した。正誤の判断は、格のペアが交換可能かどうかという観点から行った。20 個の格ペアのうち 16 ペアが正しいと判断された。評価を行った格フレームのうちの 2 つを以下に示す。

	格	用例
協議	ガ	長官, 長, 大統領, 幹事, 代表, 首脳, 党首
	ヲ	対応, 問題, 手続き, 情勢, 策
	ト*	大統領, 首相, 外相, 党首, 議員, 司令官, 長官, …
	デ	電話, 官邸, 問題, 本部, キエフ, ディリ
	ニツイテ	問題, 情勢, 対応, あり方, 内容, 撤去, 閉鎖, 改造, …
決める	ガ	民主党
	ヲ*	就任, 再任, 留任
	└ノ	会長, 社長, 長, 学長, <数量>人, 取締役, 部長, …
	ニ	会長, 後任, 長, 監督, 社長, 部長, 学長, コーチ, …

上側の「協議」の格フレームでは、ヲ格と「について」の類似が正しく判定されている。下側の「決める」の格フレームでは、ニ格とノ格が類似していると判定されているが、これは誤りである。ニ格とノ格の用例は確かに類似しているが、別々の役割を担っており、次の文のように両方の格が共起することもある。

(20)　労働組合プロ野球選手会は二十日、都内のホテルで第九回定期大会を開いて役員改選を行い、新 会長 に岡田彰布副 会長 の就任を 決めた。

このような格の共起情報を用いれば、格の類似判定の精度を高めることができると思われる。

6.2　格フレーム辞書の評価

自動構築した格フレーム辞書の評価を行った。20 個の用言について構築されている格フレームを人手で評価を行った。格フレームの評価は、次のような基準をすべて満たす場合に正しいと判定した。

表 1: 格フレームの評価結果

飾る	157/170 (92.4%)	心掛ける	37/41 (90.2%)
固まる	95/98 (96.9%)	折る	31/32 (96.9%)
積む	90/90 (100.0%)	著しい	31/32 (96.9%)
味わう	82/84 (97.6%)	復元	32/32 (100.0%)
恐れる	63/72 (87.5%)	大半+だ	28/31 (90.3%)
問題+だ	57/62 (91.9%)	裏切る	30/30 (100.0%)
寄与	46/50 (92.0%)	冷やす	27/28 (96.4%)
積極的だ	52/53 (98.1%)	散る	27/27 (100.0%)
解く	46/46 (100.0%)	賛成+だ	26/27 (96.3%)
好評+だ	41/42 (97.6%)	躍進	23/23 (100.0%)
計			1021/1070 (95.4%)

1. 「直前格要素群+用言」によって用法が固定されている。つまり、異なる格パターン・意味をとる表現が混ざっていない。
2. その格フレームの用法に対して必須的な格が抜けていない。例えば、他動詞であるのにヲ格が抜けていれば誤りとする。
3. 直前格以外の格の用例として、不適格と思われるものが混ざっていても、その格フレームを誤りとはしない。これが直前格の場合は、条件 1 に違反するので誤りとなる。

表 1 に、格フレームの評価を行った 20 用言とその精度を示す。例えば、「折る」の場合、格フレームは 32 個あり、そのうち 31 個が正しいと判断された。全体では、1070 個中 1021 個、つまり 95.4%の格フレームが正しかった。「折る」の各格フレームの評価を表 2、「冷やす」の評価を表 3 に示す。表の左端に、格フレームが正しいければ “○”、正しくなければ “×” を記した。

誤りと判断された 49 個の格フレームは、すべて条件 1 に違反しており、意味の異なる表現が同じ格フレーム内に混在していた。例えば、「冷やす」には「{ 感, 下落,..} ガ { 心理, 景気, マインド } ヲ* 冷やす」のような格フレームがあり、シソーラスにおいて「景気」と「心理」「マインド」が類似しているために、本来意味の異なるこれらの表現が混ざった格フレームとなっている。この問題に対処するには、よりよいシソーラスを用いるか、コーパスを解析するときに語義曖昧性解消を行っておく必要がある。また、「折る」の格フレームの 1 つに、「{<数量>人, 女性,..} ガ {<補文>, 調整,.. } ニ { 頭, 首,.. } ノ { 骨, ろっ骨,..} ヲ* 折る」があるが、これには複数の意味が含まれる。つまり、物理的に体のどこかの骨を折る意味と、慣用句的な「苦労する」という意味の「骨を折る」が混ざっている。これは現在の手法では区別するのは難しく、用言の直前だけでなくより広い範囲を単位にするような処理をする必要がある。

表 2:「折る」の格フレームの評価

×	{ <数量>人, 貴彦, 川瀬, 宏人, 少年, シェフチェンコ, 栄一, 和久, めぐみ, 紀,.. } ガ　{ <補文>, 車, 探し, 調整, 助手席, オートバイ, 実現, 首, 裏付け, 作り,.. } ニ　{ 頭, 首, 足, 腰, 胸, 右足, 肩, 左足, 鼻, 腕,.. } ノ　{ 骨, ろっ骨, 鎖骨, 背骨 } ヲ*　折る
○	{ 円高, 上昇, 急騰, 打ち切り } ガ　{ 可能, 兆し, 構造 } 外の関係　{ 景気, 回復, 話, 経済, 消費, 高, 朗読, 番組, 基調, 姿,.. } ノ　{ 腰 } ヲ*　折る
○	{ 家, 英太郎 } ガ　{ 筆 } ヲ*　折る
○	{ ディレクター, 課長 } ガ　{ 指 } ヲ*　折る
○	{ 速球 } ガ　{ シュート, 速球,<数量>回, スライダー, 球 } ニ　{ バット } ヲ*　折る
○	{ 児童 } ガ　{ 鶴 } ヲ*　折る
○	{ 紙, 鶏肉, 縦, のり巻き, カバー, バナナ, ボールペン, 布, 錦, 口,.. } ヲ　{ <数量>つ } ニ*　折る
○	{ <数量>本, 紙 } ヲ*　折る
○	{ 座, 上座, 前 } ニ　{ ひざ } ヲ*　折る
○	{ 母, 卒業生,<数量>人 } ガ　{ 千羽鶴 } ヲ*　折る
○	{ 自分, 人, 義母, 男性, 娘,<数量>人 } ガ　{ 足 } ヲ*　折る
○	{ 線香, 枠, シート, タオル, 割りばし, 全幅, 紙,<数量>,<数量>判 } ヲ　{ 半分 } ニ*　折る
○	{ 被告,<数量>人 } ガ　{ 重傷 } 外の関係　{ 左腕, 腕 } ヲ*　折る
○	{ ポン, 男性, 祐 } ガ　{ 首 } ヲ*　折る
○	{ 生徒, 男子 } ガ　{ ツル, つる } ヲ*　折る
○	{ 生徒 } ガ　{ 前歯 } ヲ*　折る
○	{ 右足 } ヲ*　折る
○	{ 木, 花, 草木, 庭木, 黄金, 神木, 周囲,<数量>本 } ノ　{ 枝 } ヲ*　折る
○	{ <数量>人 } ガ　{ <数量> } ヲ*　折る
○	{ 歯 } ヲ*　折る
○	{ 折り紙 } ヲ*　折る
○	{ 勉,<数量> } ガ　{ 重傷, 大けが } 外の関係　{ 左足 } ヲ*　折る
○	{ 様子 } 外の関係　{ 千羽づる } ヲ*　折る
○	{ 大たい骨 } ヲ*　折る
○	{ 棋聖 } ガ　{ 棒 } ヲ*　折る
○	{ 年寄り } ガ　{ <数量>羽 } ヲ*　折る
○	{ マスト } ヲ*　折る
○	{ 体 } ヲ*　折る
○	{ 遮断機 } ヲ*　折る
○	{ ペン } ヲ*　折る
○	{ 紙, 封筒, ナプキン } ヲ　{ 状 } ニ*　折る
○	{ 牛 } ガ　{ 事故 } 外の関係　{ 脚 } ヲ*　折る

6.3　解析実験

自動構築した格フレームの有効性を確認するために、構文・格解析実験を行った。この実験には、構文構造、格・省略関係、共参照関係、名詞間の関係の正解が人手付与されている「関係コーパス」(Kawahara, Kurohashi, and Hasida 2002) を用いた。コーパス中の 94 記事 (1100 文) に対して、自動構築した格フレームに基づく構文・格解析を行い、その解析結果をコーパスに含まれる構文構造、格関係と比較することによって評価を行った。

表 3:「冷やす」の格フレームの評価

○	{ 業者 } ガ　{ 地震, 粘り, シーン, 渇水, 路肩, 報, 運転, もろさ, 事故, 犯罪,.. } ニ　{ 肝 } ヲ* 冷やす
○	{ 母親 } ガ　{ 方法 } 外の関係　{ 体, 身体, アキレスけん, 患部, 足首, 底, 左腕, アヌビスヒヒ } ヲ* { 氷, シャワー, 水たまり, 氷のう } デ* 冷やす
×	{ 感, 下落,<補文>, 問題, 不安, 懸念, テロ, 報道, 低下, 株安,.. } ガ　{ 心理, 景気, マインド } ヲ* 冷やす
○	{ お互い, 者 } ガ　{ 頭 } ヲ* 冷やす
○	{ 木, 紅茶, 水, 苗 } ヲ　{ 冷蔵庫 } デ* 冷やす
○	{ 悪化, 目減り, 不安 } ガ　{ 消費 } ヲ* 冷やす
○	{ 電機 } ガ　{ 必要, 恐れ } 外の関係　{ 景気, 市場, 春闘, 部屋, 者, 顕著に,<時間> } ノ　{ 感, 空気, ムード } ヲ* 冷やす
○	{ 患部, 硝酸, 水溶液, 体温, 日本酒, 内部 } ヲ　{ <数量>度 } ニ* 冷やす
○	{ 上昇, 混乱 } ガ　{ 家, 企業, 個人, 者, 投資, ビジネス, 消費, ユーザー } ノ　{ 意欲 } ヲ* 冷やす
○	{ レベル } 外の関係　{ 相場 } ヲ* 冷やす
○	{ 温度, 飲料水, 水, 体温, ジュース } ヲ　{ <数量>度 } マデ* 冷やす
○	{ そうめん, 肉 } ヲ　{ 氷水 } デ* 冷やす
○	{ タンク, 足 } ヲ　{ 水 } デ* 冷やす
○	{ 療法, 状況 } 外の関係　{ 熱 } ヲ* 冷やす
○	{ 患部 } ヲ* 冷やす
○	{ 化 } ガ　{ 市場 } ヲ* 冷やす
○	{ ビール } ヲ* 冷やす
○	{ 水 } ヲ* 冷やす
○	{ ＣＤ } ヲ　{ 冷凍庫 } デ* 冷やす
○	{ 観 } ガ　{ 需要 } ヲ* 冷やす
○	{ エンジン } ヲ* 冷やす
○	{ 肩 } ヲ* 冷やす
○	{ 炉心 } ヲ* 冷やす
○	{ 地球 } ヲ* 冷やす
○	{ おなか } ヲ* 冷やす
○	{ 腰 } ヲ* 冷やす
○	{ <補文> } ガ　{ 相場, 全般 } ノ　{ 合い } ヲ* 冷やす
○	{ 蒸気 } ヲ* 冷やす

構文解析実験

自動構築した格フレーム辞書を用いた構文解析の精度を表 4 に示す。ベースラインとは、従来の KNP による構文解析である。精度は、文節の係り先が正しいかどうかを評価したものであり、文末の文節とその直前の文節は評価から除いている。表によると、全体では 0.2%の精度向上がみられ、特に連体修飾節の係り先が 0.8%向上している。これは、格フレームの選択選好によって連体修飾節の係り先が正しく推定されたためである。ベースラインの係り先は正解であったが、本手法によって誤りになった例も少数あり、そのほとんどがシソーラスの不整合によるものであった。

表 4: 構文解析の精度

	全体	連体修飾節	係助詞句
ベースライン	6582/7458 (88.3%)	685/780 (87.8%)	573/689 (83.2%)
本手法	6602/7458 (88.5%)	691/780 (88.6%)	576/689 (83.6%)

表 5: 格解析の精度

	被連体修飾詞	係助詞句
1 次格フレーム	652/784 (83.2%)	561/625 (89.8%)
高次格フレーム	676/784 (86.2%)	560/625 (89.6%)

表 6: 外の関係の精度

	適合率	再現率	F 値
1 次格フレーム	155/176 (88.1%)	155/218 (71.1%)	78.7%
高次格フレーム	182/212 (85.8%)	182/218 (83.5%)	84.7%

格解析実験

本実験では、高次格フレーム辞書を用いて格解析を行い、被連体修飾詞と係助詞句の格解析結果を評価した。高次格フレーム辞書では、ガ 2 格、外の関係といった格が利用できるので、係助詞句、被連体修飾詞の対応づけ先を次表のとおりとする。

係助詞句	: ガ, ヲ, ガ 2
被連体修飾詞	: ガ, ヲ, ニ, ガ 2, 外の関係

ただし、係助詞句の対応付け先が格フレームにない場合はガ 2 格と解析する。

被連体修飾詞と係助詞句の格解析結果を正解と比較して評価を行った。構文解析誤りの影響を除いて格解析の評価を行うために、正解の構文構造を入力した。被連体修飾詞と係助詞句の格解析の精度を表 5 に、さらに被連体修飾詞の中で外の関係の適合率、再現率を表 6 に示す。これらの表の上側には、1 次格フレーム辞書に後処理を行って得られた格フレーム辞書を用いた格解析の精度を比較のために挙げている。1 次格フレーム辞書による外の関係の解析は、従来の KNP によるもので、外の関係になる単語を人手で記述した辞書をあらかじめ用意しておき、被連体修飾詞がその単語であれば常に外の関係としている。

結果をみると、1 次格フレーム辞書による解析に対して、被連体修飾詞では 3.0% 格解析の精度が向上しているが、係助詞句の精度はほとんど変化しなかった。これは、ガ 2 格の用例の適用例が少なかったためである。

7 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が近年活発に研究されてきた (Brent 1993; Ushioda, Evans, Gibson, and Waibel 1993; Manning 1993; Briscoe and Carroll 1997; Korhonen and Preiss 2003) 。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を収集したものではなく、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、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。

連体修飾や係助詞句の解析については、いくつか先行研究があり、そのほとんどは統計情報や機械学習手法を用いたものである (Baldwin, T., 徳永, 田中 1999; 阿辺川, 白井, 田中, 徳永 2001; Torisawa 2001; 村田 2001; 阿辺川 奥村 2004)。阿辺川らは、格フレームを用いる手法は、内の関係でないときに外の関係であるという消去法の上に成り立っており、外の関係を高精度に判別できないと主張している (阿辺川・奥村 2004)。我々の提案手法では、高次格フレームを構築する段階ではこのような消去法を行っているが、高次格フレームを用いた解析時には、内/外の関係両方の用例を考慮している。実際に、1 次格フレームよりも高次格フレームを用いたときの方が精度が向上しており、阿辺川らの指摘している欠点を補うことができたと考えられる。

8 おわりに

本論文では、コーパスから信頼性の高い情報を漸次的に抽出し、格フレーム辞書の自動構築を行った。この格フレーム辞書には、二重主語構文の外のガ格、連体修飾の外の関係、および格の類似性の情報が含まれる。新聞記事 26 年分、約 2600 万文のコーパスから格フレーム辞書を構築し 2 種類の評価を行った。1 つは、いくつかの用言の格フレームを人手で評価するもの

であり、もう 1 つは得られた格フレーム辞書を用いた構文・格解析実験による評価である。これらの結果、格フレーム辞書が精度がよく構築されており、また、それによる頑健な格解析が可能であることがわかった。格解析に後続する処理である省略解析の精度が良くない原因のひとつは、外の関係や二重主語構文などの表現が精度よく解析できないことにある。つまり、外の関係の被連体修飾詞がガ格などとして解析されると、ガ格が省略されていても省略の検出に失敗することになる。今後この格フレームを用いることによって照応・省略解析の精度向上が期待できる。

参考文献

Baldwin, T., 徳永健伸, 田中穂積 (1999). “パラメータによる日本語連体修飾構造の解析.” 情報処理学会 自然言語処理研究会 1999-NL-134, pp. 55–62.

Brent, M. (1993). “From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax.” *Computational Linguistics*, **19** (2), 243–262.

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.

Kawahara, D., Kurohashi, S., and Hasida, K. (2002). “Construction of a Japanese Relevance-tagged Corpus.” In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 2008–2013.

Korhonen, A. and Preiss, J. (2003). “Improving Subcategorization Acquisition using Word Sense Disambiguation.” In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 48–55.

Kurohashi, S. and Nagao, M. (1994a). “A Method of Case Structure Analysis for Japanese Sentences based on Examples in Case Frame Dictionary.” *IEICE Transactions on Information and Systems*, **E77-D** (2), 227–239.

Kurohashi, S. and Nagao, M. (1994b). “A Syntactic Analysis Method of Long Japanese Sentences based on the Detection of Conjunctive Structures.” *Computational Linguistics*, **20** (4), 507–534.

Kurohashi, S. and Nagao, M. (1998). “Building a Japanese Parsed Corpus while Improving the Parsing System.” In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pp. 719–724.

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

Torisawa, K. (2001). “An Unsupervised Method for Canonicalization of Japanese Postposi-

tions." In *Proceedings of the 6th Natural Language Processing Pacific Rim Simposium*, pp. 211–218.

Ushioda, A., Evans, D., Gibson, T., and Waibel, A. (1993). "The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora." In *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*, pp. 95–106.

NTT コミュニケーション科学研究所 (1997). 日本語語彙大系. 岩波書店.

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.

村田真樹 (2001). "機械学習手法を用いた日本語格解析 – 教師信号借用型と非借用型、さらには併用型 –." 情報処理学会 自然言語処理研究会 2001-NL-144, pp. 113–120.

河原大輔 黒橋禎夫 (2002). "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." 自然言語処理, **9** (1), 3–19.

阿辺川武 奥村学 (2004). "日本語連体修飾節と被修飾名詞間の関係の解析." 情報処理学会 自然言語処理研究会 2004-NL-160, pp. 9–16.

阿辺川武, 白井清昭, 田中穂積, 徳永健伸 (2001). "統計情報を利用した日本語連体修飾節の解析." 言語処理学会 第 7 回年次大会発表論文集, pp. 269–272.

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.

付録 格フレームの類似度の計算方法

2 つの格フレーム F_1, F_2 の類似度を、格の一致度と用例群間の類似度の積と定義する。その計算方法を説明するために、格フレーム F_1 が格 $C_{11}, C_{12}, \cdots, C_{1m}$ を持ち、格フレーム F_2 が格 $C_{21}, C_{22}, \cdots, C_{2n}$ を持つとし、格 $C_{11}, \cdots, C_{1l}$ が格 $C_{21}, \cdots, C_{2l}$ とそれぞれ一致するとする (下図)。

$$\begin{array}{lccccccc} F_1: & C_{11}, & C_{12}, & \cdots, & C_{1l}, & \cdots, & C_{1m} \\ & \updownarrow & \updownarrow & & \updownarrow & & \\ F_2: & C_{21}, & C_{22}, & \cdots, & C_{2l}, & \cdots, & C_{2n} \end{array}$$

まず、2 つの用例 e_1, e_2 間の類似度 $sim_e(e_1, e_2)$ を、日本語語彙大系 (NTT 1997) を利用して以下のように定義する。

$$sim_e(e_1, e_2) = max_{x \in s_1, y \in s_2}\, sim(x, y) \tag{1}$$

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

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。この類似度 $sim(x, y)$ は 0 から 1 の値をとる。

F_1, F_2 の格 C_{1i}, C_{2i} 間の類似度 $CaseSim(C_{1i}, C_{2i})$ は、それぞれの用例ごとにもっとも似ている相手格中の用例をみつけ、その類似度の平均とし、次式のように定義する。

$$CaseSim(C_{1i}, C_{2i}) = \frac{\sum_{e_1 \in C_{1i}} |e_1| \cdot \max\{sim(e_1, e_2) | e_2 \in C_{2i}\} + \sum_{e_2 \in C_{2i}} |e_2| \cdot \max\{sim(e_1, e_2) | e_1 \in C_{1i}\}}{\sum_{e_1 \in C_{1i}} |e_1| + \sum_{e_2 \in C_{2i}} |e_2|} \quad (2)$$

頻出する格は重要度が高いと考えて、格の類似度に加える重みは、その格に出現する用例数の積の平方根とした。格の類似度の重み付け平均 $WeightedCaseSim(F_1, F_2)$ の計算式は，以下のようになる。

$$WeightedCaseSim(F_1, F_2) = \frac{\sum_{i=1}^{l} \sqrt{|C_{1i}||C_{2i}|} \cdot CaseSim(C_{1i}, C_{2i})}{\sum_{i=1}^{l} \sqrt{|C_{1i}||C_{2i}|}} \quad (3)$$

ただし、e_1, e_2 は C_{1i}, C_{2i} が持つ用例で、$|e_1|$ はその頻度である。また $|C_{1i}|, |C_{2i}|$ は、格 C_{1i}, C_{2i} に含まれる用例数である。

一方、格の一致度 $Alignment(F_1, F_2)$ は、F_1, F_2 について「対応付けられた格の用例数/全格用例数」を求め、それらの積の平方根とする。計算式は以下のようになる。

$$Alignment(F_1, F_2) = \sqrt{\frac{\sum_{i=1}^{l} |C_{1i}|}{\sum_{i=1}^{m} |C_{1i}|} \times \frac{\sum_{i=1}^{l} |C_{2i}|}{\sum_{i=1}^{n} |C_{2i}|}} \quad (4)$$

以上より、格フレームの類似度は以下のようになる。

$$\text{格フレーム } F_1, F_2 \text{の類似度} = WeightedCaseSim(F_1, F_2) \times Alignment(F_1, F_2) \quad (5)$$

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。2002 年同大学院博士課程単位取得認定退学。現在、東京大学大学院情報理工学系研究科 学術研究支援員。構文解析、省略解析の研究に従事。

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。

(2004 年 8 月 30 日 受付)
(2004 年 12 月 15 日 再受付)
(2005 年 1 月 15 日 採録)