頑健な格解析を実現する格フレーム辞書の自動構築

河原 大輔
京都大学 大学院情報学研究科
kawahara@pine.kuee.kyoto-u.ac.jp

黒橋 禎夫
東京大学 大学院情報理工学系研究科
科技団 さきがけ研究 21
kuro@kc.t.u-tokyo.ac.jp

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ことが必要である。このような単語間の関連性を解析するために必要な知識として格フレームがある。我々は、用言とその直前の格要素を組とすることによって、格フレーム辞書を自動的に構築する手法を提案した。しかし、その格フレーム辞書は知識としてまだ不十分なものであり、外の関係、ガガ構文、および、格の多様性を扱うことができない。本研究では、自動構築された格フレーム辞書の知識をさらに増やすことによって、これらの解析を実現する格フレーム辞書を構築する。具体的には、外の関係、ガガ構文の解析を可能とするために、得られた格フレームを用いて再度コーパスを格解析し、外の関係、ガガ構文と認識された用例を格フレームに追加する。また、格の多様性の処理、例えば、「新進党の支持を得る」の「新進党の」を「新進党から」と解釈するような処理を行うために、格フレーム内の格間の類似性を判定し、類似している格の情報を格フレームに与えた。このような再構築を行った格フレーム辞書を用いて、格解析の実験・評価を行い、得られた格フレーム辞書の有効性を確かめることができた。

2 本研究で対象とする表現

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。被連体修飾詞や係助詞句などと用言との関係は表層に現れないが、いままでの格解析でもある程度の精度で解析することができる。しかし、以下で述べる現象は解析が難しく問題となっていた [1, 2, 3, 4]。

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

(1) 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。このような関係は**外の関係**と呼ばれる。

ガガ構文
「〜[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

(2) 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。この場合、「よい」は「車」と「エンジン」の 2 つのガ格をとる。本論文では、この関係を**ガ２格**、このような表現を**ガガ構文** と呼ぶ。

用言の格要素に係るノ格
「N の 〜[助詞] … [用言]」のような形の表現で、名詞 N が用言と関係をもつ場合

(3) 社会党が 新進党 の支持を 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。

(4) 車 のエンジンが よい

例 (2) 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。

本研究では、このような格の多様性を扱う。解析対象としては、用言の直前の格要素に係るノ格のみとする。これは、用言と用言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であり、重要な句はその位置に出現することが多いからである。

上の例のように表層格で解釈できる表現とは異なり、ノ格が用言と直接的な関係をもたない場合がある。

(5) 共和党が 新議会 の主導権を 握る

この例では、「新議会」は「主導権」に対してノ格の関係をもつだけで、「握る」と直接の関係をもたない。例 (5) を連体修飾にすると次のようになる。

(6) 共和党が主導権を 握る 新議会

この例では、「新議会」は「握る」に対して外の関係である。本研究では、この例のように、用言の格要素とノ格の関係をもつ被連体修飾詞についても扱う。

3 提案手法

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 [5]。本研究では、より知識が豊富になるように、この格フレームを再構築する。まず、ガガ構文、外の関係を扱うために、この格フレームを基にしてそれぞれの用例の収集を行い、次に、格の多様性を扱うために格の類似判定という処理を行った。

3.1 ガ２格の用例収集

「〜[係助詞] … 〜が [用言]」のような表現を格解析したときに、格フレームにヲ格などの他の格がなければ、この係助詞句はガ２格であると考えられる。これは、1 回目の格フレームの学習でヲ格などをとるかどうかを学習しているからこそできる処理である。

このような考えに基づき、「〜[係助詞] … 〜が [用言]」のような表現を含む文をコーパスから収集し、そのひとつずつについて以下の処理を行う。コーパスからの収集は、解析誤りの影響を軽減するため、係助詞句が用言に曖昧性なく係る場合のみとした。

1. 係助詞句の係り先の用言を格解析する (4 章参照)。このとき係助詞句については対応付けを行わない。用言の直前格要素が存在せず、格フレームを選択できなければ処理を止める。

2. 選択された格フレームに対応付けられていない格が存在しなければ、係助詞句をこの格フレームのガ２格の用例として、この格フレームに追加する。

(7) 長い 相撲は 足腰に負担が かかる

この例を格解析すると、下表の左側のような「負担がかかる」という格フレームが選択される*。

	格	用例	入力
かかる	ガ*	負担	負担
	ニ	心臓, 体, 足, 足腰, …	足腰

*格フレームの*は、その格が用言の直前の格であることを示す。

例文の格要素「足腰に」と「負担が」の対応がつき、「相撲は」を対応付ける格が存在しない。従って、「相撲」をガ２格の用例として「負担が かかる」の格フレームに追加する。

上記の処理によって、597 用言、15,302 格フレームについてガ２格が作成された。

3.2 外の関係の用例収集

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ので、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収集

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、コーパスから収集した連体修飾を含む文をそれぞれについて以下の処理を行う。コーパスからの収集は、解析誤りの影響を軽減するため、用言が被連体修飾詞に曖昧性なく係る場合のみとした (「〜した A の B」のような表現は収集しない)。

1. 連体修飾節の用言を格解析する。このとき、被連体修飾詞については対応付けを行わない。用言の直前格要素が存在せず、格フレームを選択できなければ処理を止める。

2. 被連体修飾詞が、格フレームの対応付けられていない格のいずれに対しても類似度†が閾値を越えなければ、この格フレームについて外の関係の用例にする。閾値は 0.3 とした。

処理の例を以下に示す。

(8) 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「{ 業務, ビジネス } を 営む」という格フレームが選択される。

	格	用例	入力
営む	ガ	銀行, 会社, 旅行センター	-
	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。

(9) 違法に国際電話業務を 営んでいた 疑い

†本研究の類似度計算には、NTT の日本語語彙大系を用いる。

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収集

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 (8) の「免許」は、5 用言に対して外の関係になっているのに対し、例 (9) の「疑い」は 381 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 128 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,…

これら処理によって、格フレームごとの外の関係の用例が 637 用言、23,094 格フレームに対して得られた。

3.3 格の類似判定

格の多様性を扱うために、ガ 2 格と外の関係の用例を足した格フレームのそれぞれに以下の処理を行う。

1. 格同士の類似度をとる。2 つの格間の類似度は、用例の総当たりの類似度を計算し、それを降順にソートした結果の上から全体の 1/5 個の平均である。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. 類似度が閾値を越える格の組は類似していると判定し、格フレームに類似している格の組を記述する。閾値は 0.8 とした。

この処理を「{ 説明, 釈明 } を 求める」という格フレームに適用した例を以下に示す。

	格	用例
求める	ガ	委員会, 団, 氏, …
	ヲ*	説明, 釈明
	ニ	政府, 学会, 社長, …
	について	経緯, 実態, 状況, …
	ノ	経緯, 理由, 内容, …

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、これらの格が類似してい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、格フレーム辞書全体で 6,461 組得られた。この類似判定の結果は、格要素に係るノ格の解析に用いる[§]。

4 格解析

格解析は、基本的には [6] のアルゴリズムに従って行う。以下では、2 章で述べた表現の格解析手法について説明する。

4.1 被連体修飾詞の解析

被連体修飾詞が「こと」「ため」などの形式・副詞的名詞、および、「3 時」「最近」といった時間表現の場合は外の関係とする。それ以外の被連体修飾詞は、まだ対応のとれていない格フレームのガ格、ヲ格、ニ格、外の関係、ガ 2 格、ノ格と対応付け、その用言のスコアを最大にする対応に決定する。

ただし、外の関係、ノ格に対応付ける条件として、被連体修飾詞と外の関係の用例との類似度が 1.0 である (シソーラスにおいて同じノードに位置する) こととする。これは、格フレームの外の関係の用例と類似度がある程度高い名詞であっても、外の関係であるとは限らないので、このような強い制限を与えている。

被連体修飾詞がガ格、ヲ格、ニ格、ガ 2 格と対応付けられたならば、関係をその格に決定する。外の関係、ノ格と対応付けられたときには、関係を外の関係に決定する。ノ格に対応付けられたときは、被連体修飾詞が用言の直前格要素とノ格の関係をもっていることがわかる。

4.2 係助詞句の解析

係助詞句が時間表現であるときは時間格とする。それ以外の係助詞句を、まだ対応のとれていない格フレームのガ格、ヲ格、ガ 2 格と対応付け、それが係る用言のスコアを最大にする対応に決定する。ガガ構文の解析は、表層格がガ格の格要素を格フレームのガ格に対応付け、係助詞句を格フレームのガ 2 格またはヲ格と対応付けて、類似度の高い方の格に決定する。

4.3 用言の格要素に係るノ格の解析

格解析で選択された格フレームに、格の類似判定によってノ格と似ていると判断された格があれば、用言

[‡]用言の直前格要素に係るノ格の用例をノ格に集めている。

[§]本研究では、この処理の結果を格要素に係るノ格の解析にしか用いていないが、「〜に対して」がニ格と同じであるというようなことがわかるので、そのような解析にも利用できる。

表 1: 格解析の精度

		連体修飾	係助詞句
S_c	本手法	414/480(86.2%)	349/391(89.2%)
	旧手法	398/480(82.9%)	349/391(89.2%)
S_o	本手法	301/358(84.0%)	307/345(88.9%)
	旧手法	287/358(80.1%)	305/345(88.4%)

表 2: 外の関係の精度

		適合率	再現率	F 値
S_c	本手法	146/197 74.1%	146/160 91.2%	81.8%
	旧手法	151/227 66.5%	151/160 94.4%	78.0%
S_o	本手法	82/116 70.7%	82/92 89.1%	78.8%
	旧手法	88/148 59.5%	88/92 95.7%	73.3%

の格要素に係るノ格をその格へ対応付ける。

5 実験

「関係」タグ付きコーパス [7] を用いて格解析の実験を行った。このコーパスには、格関係の正解が付与されている。コーパス中の 1020 文を格解析し、それに含まれる連体修飾と係助詞句の解析結果を正解と比較して評価を行った。1020 文中の 535 文を参考にしながら、解析システムの調整を行ったので、これをクローズテストセット S_c とし、それ以外の 485 文をオープンテストセット S_o とした。なお、係り受け解析の誤りの影響を除いて格解析の評価を行うために、正解構文構造を解析の入力とした。被連体修飾詞と係助詞句の格解析の精度を表 1 に、被連体修飾詞の中で外の関係の適合率、再現率を表 2 に示す。この表の旧手法とは、外の関係になる単語を記述しておき、被連体修飾詞がその単語であれば、常に外の関係とする解析である。また、旧、本手法ともに、「〜[係助詞] … 〜が [用言]」の表現の係助詞句の対応付け先がなければガ２格と解析している。

結果をみると、旧手法に対して、連体修飾では 3〜4%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、追加したガ 2 格の用例が用いられたときの解析精度がテストセット全体で 4/6 であり、適用例が少なかったためである。また、用言の格要素に係るノ格の解析精度は、セット全体で 2/4、外の関係のノ格の解析精度は、セット全体で 1/2 であった。

6 おわりに

本論文では、外の関係やガガ構文といった難しい表現の解析を実現する格フレーム辞書の構築手法について提案した。構築した格フレームを用いて、外の関係やガガ構文などを含む文を精度よく解析できることがわかった。格解析に後続する処理である照応・省略解析の精度が良くない原因のひとつは、外の関係やガガ構文などの表現が精度よく解析できないことにあるので、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

[1] Baldwin Timothy, 徳永健伸, 田中穂積. 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, 1999.

[2] 阿辺川武, 白井清昭, 田中穂積, 徳永健伸. 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, 2001.

[3] Kentaro Torisawa. An unsupervised method for canonicalization of Japanese postpositions. In *Proceedings of the 6th Natural Language Processing Pacific Rim Simposium*, pp. 211–218, 2001.

[4] 村田真樹. 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, 2001.

[5] 河原大輔, 黒橋禎夫. 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, Vol. 9, No. 1, pp. 3–19, 2002.

[6] S. Kurohashi and M. Nagao. 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, 1994.

[7] 河原大輔, 黒橋禎夫, 橋田浩一. 「関係」タグ付きコーパスの作成. 言語処理学会 第 8 回年次大会発表論文集, 2002.