自動構築された格フレーム辞書に基づく省略解析

河原 大輔　　黒橋 禎夫
京都大学大学院情報学研究科

{kawahara, kuro}@pine.kuee.kyoto-u.ac.jp

1　はじめに

日本語の文章では格要素の省略が頻繁に起こる。例えば、以下の文章では、「加える」のガ格とニ格、「展開」のガ格が省略されている。このような省略を解析することが、文・文章理解を行うためには必要である。

> ロシア南部チェチェン共和国の首都グロズヌイに進攻したロシア軍は三十一日、首都中心部を装甲車などで攻撃、大統領官邸など数カ所が炎上した。
> ロシア側は首都制圧の最終段階に入ったとみられる。
> グロズヌイからの報道では、ロシア軍は激しい空爆と砲撃を [ϕが] [ϕに] 加えた 後、装甲車部隊が大統領官邸付近に進出。
> 同官邸前などでドゥダエフ政権部隊と激しい市街戦を [ϕが] 展開 している。…

英語では、格要素が省略されることはほとんどないが、代名詞化がおこる。代名詞の指示対象の解析では、人称や単数・複数といった代名詞のタイプを手がかりとすることができる。しかし、日本語では代名詞化ではなく省略が一般的であり、省略されている格を見つけるという処理から始める必要がある。その後に、その省略が何を指しているかという解析を行う。例えば、上の「加える」では、まずガ格、ニ格が省略されていることを認識し、次にその省略の指示対象を推定する処理を行う。それらの指示対象の候補として、「ロシア軍」、「首都中心部」、「最終段階」、「報道」などの単語が挙げられるが、その中でもっとも適当なものを選ぶという処理である。

このような処理を行うためには、用言がどのような格をとるか、またその格がどのような単語をとるのかを記述した格フレーム辞書が必要となる。例えば、「加える」について以下のような格フレームが利用できれば、上の例の省略の指示対象の解析が可能となる。

	格	格フレームの用例
加える	ガ	軍, 勢力, 派
	ヲ	砲撃
	ニ	拠点, 基地, 地帯

「加える」のニ格は「拠点」、「基地」といった単語をとりやすいので、候補の中でこれらにもっとも類似しているものを選ぶことにより、省略されているものが「首都中心部」であるとわかる。

これまで大規模で実用的な格フレームは存在しなかったが、我々は、大量の生コーパスから自動的に大規模な格フレーム辞書を構築する手法を提案した [2]。この格フレームは、用言の直前の格要素と用言を組にして作成したもので、用言の用法ごとに詳細に分類されている。

本論文では、この格フレーム辞書を用いて、格解析により格要素の省略を認識し、さらにその指示対象を推定する方法を提案する。まず、省略情報を付与したコーパスを作成し、省略の現象について調査を行った。次に、その調査に基づき、格フレーム辞書を用いた省略解析システムを試作し、作成したコーパスを解析して評価した。

2　格・省略情報を付与したコーパスの作成

省略の現象を調査するために、格・省略情報を付与したコーパスを作成した。このコーパスは、用言に対して直接係り受け関係をもっている格要素と、省略されている格要素を記述したものである。現在までに、京都大学テキストコーパス中の新聞 20 記事 200 文に対して格・省略情報のタグを付与した。タグを付与する作業には、京都大学テキストコーパスプロジェクトの GUI ツールを、格・省略情報のタグが付与できるように変更し、それを利用した (図 1)。

省略のタグとしては、用言 (動詞、形容詞、名詞+判定詞) とサ変名詞に対して、省略されている格の指示対象を与えた。指示対象の単位は基本的には文節内の自立語部分であるが、必要に応じて文節内の一部分

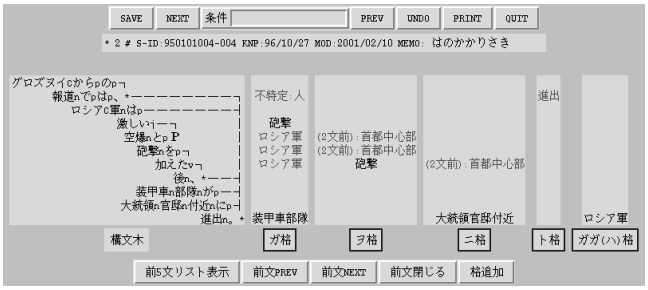


図 1: GUI ツール

の自立語とする。例えば、「首都制圧の」という複合名詞は一文節であるが、「首都」を指示する用言があれば「首都」と記述する。

省略のタグは、用言に対して重要な関係をもっていると判断されるものについて付与した。実際に省略タグを付与した格は、ガ格、ヲ格、ニ格の 3 種類であった。

指示対象の候補の制約

コーパスの作成・調査の過程で、指示対象の存在する範囲はかなり狭いということが明らかになった。指示対象はほとんどの場合、次の範囲内にある。

- 対象用言がある文中
- 前文の重要な要素
- 記事の最初の文の重要な要素

ただし、重要な要素とは以下に挙げるものとする。

- 主節の用言に係る格要素 (すでに決定された省略の指示対象を含む)
- 主節に係る従属節で、かつ「〜ので」のようなスコープの広い従属節に含まれる格要素 (すでに決定された省略の指示対象を含む)
- 「〜は」のように提題助詞が付属する名詞句中に含まれる語

例えば、表 1 の 4 文目の「展開」のガ格は省略されており、この指示対象の候補は表 1 の太字の単語となる。このように、この制約は指示対象の候補をかなり限定することができるので、省略解析において有効である。

タグを付与する際の問題

省略タグをつける際に問題となったことを以下に挙げる。

表 1: 指示対象の候補

1	ロシア南部チェチェン共和国の首都グロズヌイに進攻した**ロシア軍**は三十一日、**首都中心部**を装甲車などで攻撃、大統領官邸など**数力所**が炎上した。
2	ロシア側は首都制圧の最終段階に入ったとみられる。
3	グロズヌイからの報道では、**ロシア軍**は激しい空爆と砲撃を加えた後、**装甲車部隊**が**大統領官邸**付近に進出。
4	同官邸前などでドゥダエフ政権部隊と激しい市街戦を 展開 している。

- 付加的表現の問題
  次のような用言は意味が薄いので、タグを付与する必要はないと判断した。
  - 修飾的な用言
    〜を通じ, 〜における, 決して, 近く
  - モダリティ表現
    思う, 見える, いう, 〜ことになる, 〜する見通しとなる

- 「不特定」問題
  例: 地方公務員法では日本国籍がない人の 任用 を禁じる 規定 はない。

  この例の「任用」、「規定」のガ格はいずれも不特定であると考えられる。これらには「不特定:人」というタグを付与する。
  また、「規定」は動作的な意味ではなく、「規定」という動作の結果を表しているので、ヲ格について考えることには意味がない。そこで、このような単語のヲ格には「不特定:もの」というタグを付与する。

- 「組織・代表者」問題
  例: ポレワノフ副首相は「政府の経済路線を 変更し、企業に対する国家の指導を 強化する ことが必要」と強調するとともに…

  この文章において、下線部の用言の主格は、副首相、国、または政府関連の人々のいずれにもとれるので問題である。本研究では正解を複数与え、そのいずれでもよいとする。

- 同一性の問題
  例: 橋本大二郎知事, 橋本知事

  固有表現は、後の文章になればなるほど短く表現される。このような表現は同一のものとして扱う必要がある。本研究では、それらが省略の指示対象となるときは正解を複数与えている。

3　省略解析のアルゴリズム

2 章で述べた候補の制約を利用し、格フレームに基づく省略解析システムを作成した。省略解析の手順は次のとおりである。

1. 入力文を構文解析、格解析する。
2. 格解析結果から、各用言の省略されている格要素を認識する。
3. 省略の指示対象を推定する。

以下では、図 2 の解析例にそって、2 と 3 の処理を順に述べる。

3.1　省略されている格要素の認識

格解析の基本的な処理としては、用言の用法ごとに用意されている格フレームの中から入力文にもっとも合致するものを選択し、同時に入力文の格要素と格フレームとの対応付けを行う。格解析が終わったときに、格フレームに入力文の格要素と対応付けられていない格があり、それがガ格、ヲ格、ニ格のいずれかであれば、その格が対象用言において省略されていると認識する。

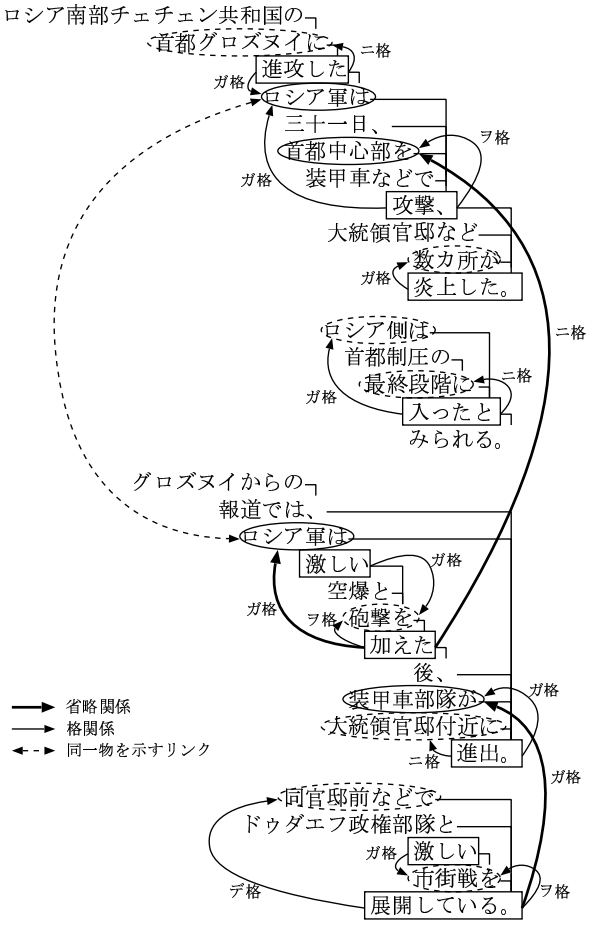
例えば、図 2 の 3 文目の「加える」には以下のような格フレームがある。

	格	用例
加える (1)	ガ	軍, 勢力, 派
	ヲ	砲撃
	ニ	拠点, 基地, 地帯
加える (2)	ガ	警察官, 部員
	ヲ	暴行, 乱暴
	ニ	男性, 相手, 客
加える (3)	ガ	ソ連, 政府
	ヲ	検討, 工夫

「砲撃を加える」という入力側の表現ともっとも合致する格フレームは (1) であり、この対応付けでは、格フレームのガ格、ニ格に対応する入力側の格要素がない。従って、システムはガ格、ニ格が省略されていると認識する。

3.2　省略の指示対象の推定

省略されていると認識された格について、その格フレームの用例群を用いて指示対象の推定を行う。まず、指示対象の候補の範囲について、2 章で述べた制約を設ける。次に、その範囲内にある候補をスコア付けし、



用言	格	格フレームの用例	解析結果†
進攻	ガ	<主体>*	ロシア軍
	ニ	領内	首都グロズヌイ
	デ	地点	-
攻撃	ガ	米国, イラク, 軍	ロシア軍
	ヲ	イラク, 国, 米国	首都中心部
炎上	ガ	<数量>, <主体>*	数カ所
入る	ガ	米軍, 軍隊, 部隊	ロシア軍
	ニ	域, 下, 中	最終段階
激しい	ガ	落ち込み, 突っ込み	砲撃
加える	ガ	軍, 勢力, 派	【ロシア軍】
	ヲ	砲撃	砲撃
	ニ	拠点, 基地, 地帯	【首都中心部】
進出	ガ	企業, 社	装甲車部隊
	ニ	海外, 日本	大統領官邸
激しい	ガ	競争, 争い	市街戦
展開	ガ	軍, 隊	【装甲車部隊】
	ヲ	戦闘, 戦争, 熱戦	市街戦
	デ	全土, 周辺, 市内	官邸

* ガ格の用例がない場合には、<主体>という意味属性を補っている。
† 省略の指示対象として推定された格要素は【】で表している。

図 2: 省略解析例 (サ変名詞は表示していない)

スコア最大のものが閾値*を越えていれば指示対象として決定する。スコアが閾値を越える候補がなければ、ガ格とニ格は「不特定:人」、ヲ格は「不特定: もの」と推定する。

スコアは、基本的には格フレームの用例群と候補間の類似度であり、次の式で計算する。

$$\text{スコア} = \text{重み} \times \text{類似度}$$

それぞれの値の意味は以下のとおりである。

重み: 対象文中の体言で、重要な要素ではないもの、または、処理中の用言より後にあるものは 0.6、それ以外 (前文、記事の最初の文を含む) は 1.0。

類似度: 候補と格フレームの省略格の格用例群との類似度で、NTT の日本語語彙大系 [1] を用いて計算する (最大 1.0)。

以下では、図 2 の 3 文目の「加える」のニ格を例にとって、このスコアを説明する。「加える」のニ格の指示対象の候補と「展開」の格フレームのニ格の用例群 (「拠点」、「基地」、「地帯」) を比較して、スコアを計算すると以下のようになる (類似度が 0 のものは除く)。

候補	重み	類似度	スコア
首都中心部	1.0	1.0	**1.0**
ロシア側	1.0	0.67	0.67
最終段階	1.0	0.75	0.75
報道	1.0	0.43	0.43
ロシア軍	1.0	1.0	1.0
空爆	0.6	0.50	0.30
装甲車部隊	0.6	0.83	0.50
大統領官邸	0.6	0.80	0.48

ただし、「ロシア軍」は「加える」のガ格の指示対象として先に決定されており候補にはならない。この表より、「首都中心部」がもっともスコアが高いので、指示対象として「首都中心部」が推定される。

4 検証

2 章で述べたコーパスを対象として、省略解析の検証を行った。まず、98.6%の省略の指示対象が制限した範囲内にあったので、この制約がかなり有効であることがわかる。次に、省略解析の結果は表 2 に示す。用言はある程度の精度が解析できているが、サ変名詞の解析精度は 50%前後であった。解析において選択された格フレームは、ほとんど用言の用法に一致したものであり、格フレームが有効に用いられていた。

*現在のところ 0.3 に設定している。

表 2: 省略解析の結果

	適合率	再現率
用言	178/259 (68.7%)	178/251 (70.9%)
サ変名詞	64/154 (41.6%)	64/118 (54.2%)
計	242/413 (58.6%)	242/369 (65.6%)

以下では、今後考慮すべき問題点を述べる。

- 格解析と省略解析の相互作用
  格解析を行うときに、直接係り受け関係をもつ格要素を手がかりとして格フレームを選んでいるが、実際には省略の格要素もあるので、それも考慮して格フレームを選ぶ必要がある。そのためには、格解析と省略解析を統合的に行う必要がある。
- サ変名詞の「不特定」の判断
  現在の解析では、スコアが閾値を越える候補がなければ「不特定:人」などと推定されるが、この精度があまり高くない。サ変名詞は、このタグをとることが多いので、サ変名詞の精度がよくないのはこれに起因している。この精度を上げるには、あるサ変名詞は「不特定」をとりやすいといった別の知識が必要であると思われる。

5 おわりに

本論文では、格・省略情報を付与したコーパスを作成し、省略の現象について調査を行った。その結果、省略の指示対象の存在する範囲はかなり制約があることが明らかになった。その制約を利用して、格フレーム辞書に基づく省略解析を行うシステムを作成し評価を行った。この制約と格フレーム辞書はかなり有効であることがわかったが、解析はさらに改良の余地がある。今後は、解析精度を高め、大規模な解析実験を行う予定である。

参考文献

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

[2] 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獲得. 情報処理学会自然言語処理研究会 2000-NL-140, pp. 127–134, 2000.