自動構築した格フレーム辞書に基づく省略解析の大規模評価

河原 大輔
東京大学 大学院情報理工学系研究科

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

{kawahara, kuro}@kc.t.u-tokyo.ac.jp

1 はじめに

日本語の文章では格要素が指示詞となったり、省略されることが頻繁に起こる。例えば、図1の記事では、「進め」のガ格、「完了した」のガ格とヲ格などが省略されている。この例の省略を解析するためには、「{県, 会, ···} が {作業, 構想, ···} を 進める」、「{議会, 国, ···} が {作業, 処理, ···} を 完了する」のような知識が必要となる。このような知識は格フレームと呼ばれるものであるが、我々は、大規模格フレーム辞書を生コーパスから自動的に構築する方法を提案している [1]。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に基づく省略解析システムを提案する。

省略・照応を高精度に解析する技術は、自動要約、機械翻訳、質問応答などの言語処理アプリケーションを高度化するために必要であり、これまでさまざまな手法が提案されてきた。中岩らや村田らは、人手で作成した規則による省略解析を提案している [6, 5]。これらは高い精度を実現しているが、1文単位で独立した文や物語文を対象としているため、一般の文章で利用するには規則の修正が必要であると思われる。一方、吉野らや関らは、機械学習や確率モデルによる省略解析を提案している [4, 3]。これらは、省略関係の正解を付与した数十記事の新聞記事から学習を行っているが、精度はあまり高くない。この原因のひとつとしては、コーパスの量が、素性の選定、学習を行うには少なかったことが挙げられる。

実際に使える省略解析を実現するには、省略の現象を詳細かつ大規模に調査を行う必要がある。我々は、文章中の単語間の様々な関係をタグづけしたコーパスの作成を行っている [2]。このコーパスで対象とする関係は、用言・サ変名詞に対する格関係、名詞間の関係、および共参照であり、用言に対する格関係には、省略の指示対象が含まれている。本論文では、このコーパスを用いて省略現象の大規模調査を行い、日本語の省略を解析する手法を検討する。その検討に基づき、格フレーム辞書を用いた省略解析システムを試作し、コーパスを解析して評価を行った。

一昨年五月からモザンビークで国連モザンビーク活動の一員として活動してきた自衛隊輸送調整中隊は五日、帰国の途に就く。
モザンビークでの選挙が終了し、任期満了に伴う [φが] 撤収。
首都マプトに駐屯する自衛隊によると、通常の輸送調整業務を昨年十二月末まで [φが] 続けながら三カ月 [φが] [φに] かけて撤収作業を [φが] 進め、予定通り [φが] [φを] 完了した。···

図 1: 記事の例 (ϕ の箇所が省略されている)

2 省略現象の調査

本章では、省略現象の調査を行うコーパスの概要と、調査の結果、考案した位置カテゴリとその順序について述べる。

2.1 コーパスの概要

本論文では、関係コーパス [2] を用いて省略現象の調査を行った。このコーパスは、新聞記事に対して文章中の単語間の様々な関係をタグづけしたものである。対象としている関係は、用言・サ変名詞に対する格関係、名詞間の関係、および共参照である。用言に対する格関係には、省略されている格要素 (ゼロ代名詞と呼ばれる) の指示対象が含まれている。

関係コーパス中の379記事、3,695文を対象としてゼロ代名詞の調査を行った。用言 (動詞、形容詞、名詞+判定詞) は11,173個出現し、そのうちゼロ代名詞を格

表 1: ゼロ代名詞の格分布 (上位 8 個まで)

格	回数 (割合)	格	回数 (割合)
ガ	4,316 (65.1%)	ガ 2	120 (1.8%)
ニ	1,022 (15.4%)	外の関係	80 (1.2%)
ヲ	459 (6.9%)	ト	71 (1.1%)
デ	163 (2.7%)	カラ	49 (0.7%)

表 2: 先行詞の出現位置分布 (5 文前まで)

位置	回数 (割合)	位置	回数 (割合)
同一文	2,512 (50.2%)	3 文前	233 (4.7%)
1 文前	1,208 (24.1%)	4 文前	133 (2.7%)
2 文前	508 (10.2%)	5 文前	86 (1.7%)

要素とする用言は 5,550 個あり、ゼロ代名詞は 6,634 個であった。ゼロ代名詞となっている格の分布を表 1 に示す。ゼロ代名詞のうち先行詞が記事中に存在するものは 5,004 個であった。これらの先行詞について、出現する文の位置を同一文、1 文前、2 文前、$\cdots$ に分類した。これを表 2 に示す。一方、先行詞が記事中に存在しないゼロ代名詞は外界照応しており、不特定の人々を指している場合が多かった。

2.2 位置カテゴリの設定とその順序づけ

一般に、ゼロ代名詞の先行詞は、ゼロ代名詞から距離が近いところにある傾向がある。機械学習による省略解析の先行研究 [4] では、先行詞の候補とゼロ代名詞との間の距離 (文数や文節数) を素性に組み込み、この傾向を学習しようと試みている。しかし、省略解析のように、もっともよい候補を選択するタスクに学習器が出力するスコアを用いるのは、直接的ではなく根拠が薄いと思われる。

我々は、ゼロ代名詞と先行詞の位置関係をコーパスを用いて調査し、この調査結果から得た知見を陽に省略解析システムで利用することを考案した。具体的には、ゼロ代名詞と先行詞の位置関係を設定し、ゼロ代名詞に対してどこの位置にある候補が先行詞となりやすいかをコーパスから得る。解析では、この順番に候補を調べ、スコアが閾値を越える最初の候補を先行詞に決めることとする。

ゼロ代名詞と先行詞の位置関係は構造的に捉え、表 3 のように設定した (先行詞は 2 文前までに 84.5% 出現していることから 2 文前までを対象とした)。設定した位置関係を位置カテゴリと呼ぶ。ただし、V_z はゼロ代名詞をもつ用言を示す。「」で囲まれた用言を V_a とすると、先行詞は V_a の格要素である。"並列" と

表 3: 先行詞の位置カテゴリ

対象文		
L_1 : 「V_z の親用言」の格要素		主節
L_2 : 「V_z の親用言」の格要素		
L_3 : 「V_z の親用言」の格要素	並列	主節
L_4 : 「V_z の親用言」の格要素	並列	
L_5 : 「V_z の子用言」の格要素		
L_6 : 「V_z の子用言」の格要素	並列	
L_7 : 「V_z の親体言の親用言」の格要素		主節
L_8 : 「V_z の親体言の親用言」の格要素		
L_9 : 「V_z の親用言の親用言」の格要素		主節
L_{10}: 「V_z の親用言の親用言」の格要素		
L_{11}: 「主節用言」の格要素		主節
L_{12}: 「主節に係る従属節用言」の格要素		
L_{13}: その他要素 (V_z より後)		
L_{14}: その他要素 (V_z より前)		
1 文前		
L_{15}: 「主節用言」の格要素		主節
L_{16}: 「主節に係る従属節用言」の格要素		
L_{17}: その他要素		
2 文前		
L_{18}: 「主節用言」の格要素		主節
L_{19}: 「主節に係る従属節用言」の格要素		
L_{20}: その他要素		

は、V_z と V_a が並列関係にあることを示し、"主節" とは V_a がその文の主節用言であることを示す。例えば、"「V_z の親用言」の格要素" とは、ゼロ代名詞をもつ用言の係り先用言に先行詞が係っている状況を表している。位置カテゴリはそれぞれ独立であり、"主節用言" はほかの主節の位置カテゴリに含まれる用言以外である。

次に、それぞれの位置カテゴリが、どれくらい先行詞となりやすいかをコーパスを用いて調査した。位置カテゴリ L のスコアを次式で計算する。

$$\frac{\text{先行詞が } L \text{ にある回数}}{L \text{ にある先行詞候補の数の和}}$$

これをゼロ代名詞の格ごとに集計し、スコア順に位置カテゴリを並べた。そのうちガ格、ヲ格の順序をそれぞれ、図 2、3 に示す。ガ格のゼロ代名詞は、主節の親用言の格要素を先行詞としやすく、ヲ格のゼロ代名詞は、子用言の格要素を先行詞とすることが多いことがわかる。解析では、この順序で先行詞の候補を調べるが、ヲ格と二格についてはスコアが 0.05 以上の位置カテゴリのみを用いることにした。これは、ヲ格と二格については、先行詞となる可能性が少ない位置カテゴリを考慮しない方が精度がよいことが予備実験によりわかったからである。

さらに、先行詞が出現場所においてどの格で用言 (V_a) に係っているかを調査した。ガ格のゼロ代名詞に

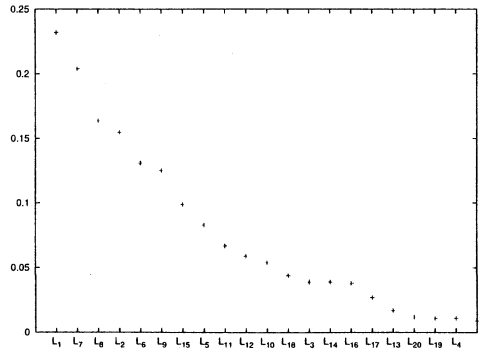

図 2: 位置カテゴリの順序 (ガ格)

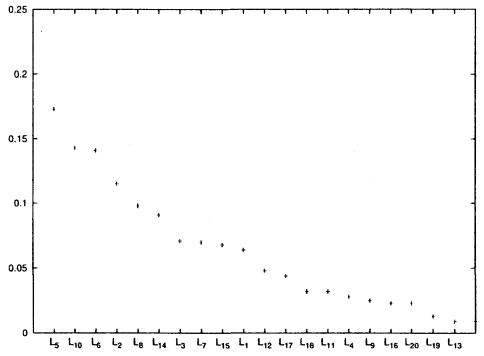

図 3: 位置カテゴリの順序 (ヲ格)

対しては、出現した場所でガ格となっている先行詞を指すことが多く、97.2%がそれを満たしていた。解析ではこの制約を利用することにした。

3 省略解析アルゴリズム

前章で述べた位置カテゴリの順序を利用する、格フレーム辞書ベースの省略・照応解析システムを作成した。解析の手順を以下に示す。

1. 入力文を構文解析する。入力文中の各用言について、文末から文頭に向かって以下の処理を行う。
2. 入力側の表現に合致する格フレームを選択する。その条件として、対象用言の直前に格要素があり、それと類似している格フレームが存在することとする。類似度がもっとも高い格フレームが複数ある場合や、格フレームが選択できない場合は、それぞれの格フレームについて以下の処理を行い、最後にもっとも類似度が高かった格フレームに決定する。類似度は NTT の日本語語彙大系を用いて計算する (最大 1.0)。
3. 格フレームと入力側の格要素との対応をとる。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、次表の格それぞれに対応させ、対応づけ全体の類似度がもっともよい対応を選択し、格を決定する。

係助詞句　　：ガ, ヲ, ガ 2
被連体修飾詞：ガ, ヲ, 外の関係

4. 格フレーム中で対応づけらていない格をゼロ代名詞と認識する。コーパス調査の結果、ゼロ代名詞はガ格、ヲ格、ニ格で 87.4%を占めているので、この 3 つの格を対象にした。また、格要素が指示詞となっていれば照応解析の対象とする。
5. 省略されていると認識された格および、指示詞の格について先行詞の推定を行う。先行詞の候補を、格ごとに設定された位置カテゴリの順序に従って調べ、格フレームの用例との類似度をとる。ガ格の推定時には、候補の出現場所での格がガ格であることを条件とする。類似度が閾値 (現在のところ 0.60) を越える候補があれば、先行詞をそれに決定し処理を終了する。

例として、図 1 の 3 文目の「進め」と「完了した」を考える。それぞれに対して次のような格フレームがある。

	格	用例
進める (1)	ガ ヲ	<主体>, 県, 会, ··· 作業, 構想, ···
進める (2)	ガ ヲ ニ	<主体>, 部 駒 戦, 路線, ···
⋮	⋮	⋮
完了する (1)	ガ ヲ	<主体>, 議会, 国, ··· 作業, 処理
完了する (2)	ガ ヲ	<主体> 償却, 返済
⋮	⋮	⋮

「完了した」には直前の格要素がないために格フレームは選択されず、後の処理をすべての格フレームについて行うことになる。一方、「作業を 進め (る)」には

表 4: 省略解析結果

適合率	再現率	F
253/714 (35.4%)	253/603 (42.0%)	38.4%

「進める (1)」が合致するのでこれが選択される。次にゼロ代名詞の認識であるが、「完了した」を格フレーム「完了する (1)」を用いて解析している場合には、格フレームのガ格とヲ格に対応する入力側の格要素がないので、ガ格とヲ格が省略されていると認識される。「作業を 進める)」では、ガ格が省略されていると認識される。

先行詞の推定処理では、「完了した」のガ格の候補は L_{15}:中隊, L_{18}:中隊, L_{14}:自衛隊, $\cdots$ であり、ヲ格の候補は L_6:作業, L_{14}:業務, L_{14}:自衛隊, $\cdots$ の順となっている。それぞれ先頭の「中隊 (類似度:0.73)」、「作業 (類似度:1.0)」は格フレームの用例との類似度が閾値を越えているので先行詞に決定される。最終的に、この格フレームの類似度がもっともよいので、この格フレームに決定される。また、「作業を 進める)」のガ格の先行詞も「中隊」に決定される。

4 実験

関係コーパスのうち 60 記事を用いて、省略解析の実験を行った。構文構造は正解のものを与え、それぞれの記事の先頭 10 文までを省略解析システムに入力した。その結果を表 4 に示す。表の適合率、再現率はゼロ代名詞認識、先行詞の推定処理の双方を併せて評価したものである。主な誤り原因を以下に示す。

ゼロ代名詞の認識誤り

適合率が再現率よりも悪い原因は、ゼロ代名詞を余分に認識していることに起因している。

(1) 一方、ドゥダエフ政権側の首都防衛司令官は同日夕、テレビを通じ、首都防衛はうまくいっており、ロシア軍の戦車五十両を破壊したと 発表。

この例では、「発表」の格フレームに対応づけられていないニ格があり、ニ格をゼロ代名詞と認識してしまう。これは格フレームが悪いのではなく、文脈によってその格をとる場合ととらない場合があるためである。これに対処するためには、格のとりやすさを格フレームまたは用言ごとに導入する必要がある。

位置カテゴリが用言のみを対象としていることによる誤り

現在の位置カテゴリは用言に関係する格要素しか扱っていない。

(2) 村山富市首相は年頭にあたり首相官邸で内閣記者会と二十八日会見し、社会党の新民主連合所属議員の離党問題について「政権に影響を及ぼすことにはならない。離党者がいても、その範囲にとどまると思う」と述べ、大量離党には 至らない との見通しを示した。

この例の「至らない」のガ格は「社会党」であるが「首相」と誤って解析される。「社会党」は位置カテゴリ順では上位に来ないためである。位置カテゴリをサ変名詞など用言以外の関係を含んだものにすれば、「離党」に「社会党」が補われ、「社会党」が上位に来るようになる。このように、用言だけでなく文章中のさまざまな関係を用いるようにすれば精度向上につながると思われる。

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用いて省略現象の調査を行った。その調査結果に基づいて、先行詞の存在する位置に順序をつけた。その順序を利用して、格フレーム辞書に基づく省略解析を行うシステムを試作し評価を行った。今後は、このシステムの枠組みに機械学習を統合し、精度向上を目指す予定である。

参考文献

[1] 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.

[2] 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, pp. 495–498, 2002.

[3] 関和広, 藤井敦, 石川徹也. 確率モデルを用いた日本語ゼロ代名詞の照応解析. 自然言語処理, Vol. 9, No. 3, pp. 63–85, 2002.

[4] 吉野圭一, 竹内和広, 松本裕治. 機械学習を用いた日本語ゼロ代名詞照応関係の同定. 言語処理学会第 7 回年次大会発表論文集, pp. 506–509, 2001.

[5] 村田真樹, 長尾眞. 用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定. 自然言語処理, Vol. 4, No. 1, pp. 87–109, 1997.

[6] 中岩浩巳, 池原悟. 語用論的・意味論的制約を用いた日本語ゼロ代名詞の文内照応解析. 自然言語処理, Vol. 3, No. 4, pp. 49–65, 1996.