統計情報を利用した日本語連体修飾節の解析

阿辺川 武　　白井 清昭　　田中 穂積　　徳永 健伸

東京工業大学大学院 情報理工学研究科 計算工学専攻

{abekawa,kshirai,tanaka,take}@cl.cs.titech.ac.jp

1　はじめに

連体修飾節とは，動詞や形容詞などの用言が連体形で名詞を修飾する文節である．連体修飾節には，構文的・意味的関係から2つの異なる関係に分類できる．

(a) さんまを焼く男
(b) さんまを焼く匂い

(a) では被修飾名詞「男」と連体修飾節中の用言「焼く」との間に「男がさんまを焼く」という格関係が成り立つ．一方，(b) では被修飾名詞「匂い」にどのような格助詞を補っても，連体修飾節中に埋めることができない．本論文では寺村 [5] にならい，前者のような関係を「内の関係」，後者を「外の関係」と呼ぶ．これらの関係を求めること，および内の関係において被修飾名詞と連体修飾中の用言の間に介在する格助詞を求めることは，機械翻訳，文章要約，文分割といった様々な処理で必要となる．

従来，連体修飾節の解析には主に格フレームを用いた手法が用いられてきた [1, 6]．しかし格フレームを利用した場合，格フレーム辞書の構築のコスト，網羅性，拡張の非容易性などの問題点が存在する．また格フレームは格要素に対する意味的制約の緩さから，外の関係の解析には不十分なことが多い．

これに対し近年，大規模なコーパスが計算機で利用可能になってきたことから，格フレームの自動構築の研究が盛んに行われている [3]．例えば構文解析されたコーパスから名詞，動詞の共起関係を収集し，シソーラスを用いて同じ格要素として現れる名詞が属する共通の意味クラスを求めることで，格要素に対する選択制約を意味属性で記述した格フレームが構築できる．本研究では，格フレームの自動構築を参考に，名詞動詞対の共起頻度を収集し，内／外の関係の判別，格助詞の付与を行う手法を提案する．

2　従来の連体修飾節解析手法

2.1　格フレームの利用

格フレームは，用言とそれが取り得る格要素に対する選択制約を記述したものである．既存の格フレーム辞書は，格要素の取り得る名詞を列挙したものや，シソーラスを用いて抽象化し，それらを意味属性として記述したものがある．格フレームの格要素を1つの意味属性で記述した場合，比較的抽象度の高い意味属性が用いられる．

連体修飾節の解析では，まず用言に対応する格フレームを見つけ，節内の格要素が格スロットに埋まるかを照合する．その後，被修飾名詞が残りの格スロットに埋まるかを照合する．複数の格スロットへ埋めることができる時，取り得る格要素の名詞との類似度を計算し，一番類似度の高い格スロットへ埋め込む．

2.2　外の関係の解析

外の関係を判別する手法として，「目的」「意見」など外の関係をとる名詞をあらかじめリストアップしておき，被修飾名詞がこれらの名詞であるとき，その連体修飾節の動詞にかかわらず外の関係とする手法がある [1]．しかし，例えば「各人の持つ目的」という連体修飾節があったとき，「各人が目的を持つ」が文として成立するように，外の関係をとる名詞でも内の関係になることもある．このように外の関係をとる名詞でも格助詞の付与を考慮しなければならない．また，格要素に対する選択制約が意味属性で記述された格フレームを利用して解析を行うと，格フレームの格要素は抽象度の高い意味属性で記述されることが多いため，外の関係の名詞であっても格スロットに埋まり，内の関係に判断されるという問題がある．

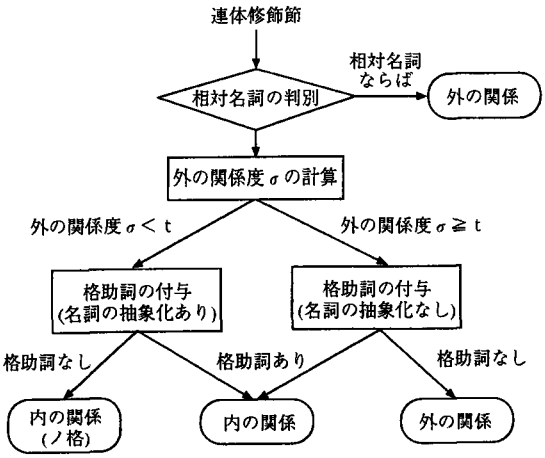

図 1: 解析の流れ

3 提案手法

本節では，提案する連体修飾節の解析手法について述べる．その概要を図 1 に示す．本研究では，連体修飾節の用言が動詞である構文を解析対象とする．

3.1 相対名詞の判別

最初に被修飾名詞が相対名詞かどうかの判別を行う．本研究でいう相対名詞とは「会社を休んだ翌日」「彼が座る隣り」など連体修飾節が表す内容に対して時間的，空間的に相対した概念を表す名詞である．被修飾名詞が相対名詞であるとき，連体修飾節中の用言と格関係を持つことはないため，外の関係であるとみなす．本手法では，相対名詞のリストは人手で作成した．

3.2 外の関係度の計算

外の関係をとる名詞には，連体修飾関係で共起できるが，格関係では共起できない動詞が存在する．例えば名詞「用意」と動詞「走る」では，「走る用意」と連体修飾関係では共起することがあるが，「用意が走る」のように格関係ではどの格助詞を介在させても共起しない．

表 1 は，コーパスから収集した名詞の出現頻度と，連体修飾関係または格関係として共起した動詞の異なり数を共起関係別に集計した結果である．外の関係をとらない名詞「人々」「都市」「ゴルフ」では，出現頻度と動詞異なり数の比が連体修飾関係と格関係とでだいたい等しい．一方，外の関係をとる名詞「意向」「事実」「用意」では，名詞の出現頻度が低いにも関わらず

表 1: 動詞異なり数の比較

	連体修飾関係		格関係	
	出現頻度	動詞異なり数	出現頻度	動詞異なり数
意向	8732	941	14216	677
事実	5454	1448	7301	754
用意	2268	428	2720	74
人々	6681	1367	10026	1998
都市	1172	449	3688	857
ゴルフ	237	116	1692	431

動詞異なり数は連体修飾関係の方が多い．これは先ほど述べた，外の関係をとる名詞は連体修飾関係でのみ共起できる動詞が存在するためであると思われる．したがって，格関係で共起する動詞の頻度分布と連体修飾関係で共起する動詞の頻度分布の差が大きければ大きい程，その名詞は外の関係をとりやすいと考えた．本研究では両者の頻度分布の差を定量的に評価し，「外の関係度」として定義する．

外の関係度を次のように求める．まず格関係で共起する動詞の出現確率を $P_k(v|n) = \frac{f_k(n,v)}{f_k(n)}$，連体修飾で共起する確率を $P_m(v|n) = \frac{f_m(n,v)}{f_m(n)}$ とする．$f_k(n,v)$ は名詞 n と動詞 v が格関係で共起した頻度，$f_k(n)$ は名詞が格関係で出現した頻度である．同様に $f_m(n,v)$，$f_m(n)$ は連体修飾関係の頻度である．そして，式 (1) のように，$P_k(v|n)$ と $P_m(v|n)$ の 2 つの確率分布の差を KL-距離で評価し，これを外の関係度とした．

$$D(P_k(v|n)||P_m(v|n)) = \sum P_k(v|n) \log \frac{P_k(v|n)}{P_m(v|n)} \quad (1)$$

動詞の出現確率の分布に差異があるほど，KL-距離の値は大きくなり，その名詞は外の関係度が高い．

本研究では，被修飾名詞の外の関係度が，ある閾値 t 以上なら外の関係，それ以外は内の関係にあると判定する．ただし 3.4 項で述べるように，ここで外の関係と判断された場合でも，後で内の関係と判断されることがある．閾値 t は，テストセットとは別のデータに対して，t の値を変動させて内の関係か外の関係かの判定を行い，正解率が最も高い時の値を選んだ．

3.3 格助詞の付与

外の関係度が t 以下の場合，内の関係にあるとして動詞と被修飾名詞の間に埋めるべき格を選択する．具体的には，連体修飾節中の動詞 v と被修飾名詞 n の共

表 2: 格助詞出現確率の例

n	v	$f_k(n,v)$	$P(r\|n,v)$		
施設	オープンする	80	が 0.71	を 0.20	に 0.06

起頻度に着目し，付与すべき格助詞 r を選択する．共起頻度から格助詞出現確率 $P(r|n,v)$ を式 (2) のように定義する．

$$P(r|n,v) = \frac{f_r(n,r,v)}{f_k(n,v)} \quad (2)$$

ここで $f_r(n,r,v)$ は，名詞 n と動詞 v が格助詞 r を伴って共起した頻度である．解析では最初に，連体修飾節中に出現した格助詞は候補から除外する．次に格助詞出現確率順に格助詞を並べ，出現確率の一番高い格助詞を選択する．例えば「海辺にオープンする施設」では，表 2 のような格助詞の候補が出現確率と共に得られる．まず連体修飾節で出現したニ格は候補から削除され，残された格助詞のうち最も高い確率を持つ格助詞であるガ格が選択される．

$f_k(n,v) < 5$ の時は，名詞をシソーラスにより抽象化し，意味属性 c と動詞 v が共起したときの格助詞出現確率 $P(r|c,v)$ を計算する．シソーラスには日本語語彙大系の一般名詞意味属性体系 [2] を利用し，深さ 6 階層までの意味属性を用いた．名詞を抽象化することで，より多くの共起事例を収集することが出来る．例えば「施設」には [建造物] [施設] [労働] と 3 つの意味属性が存在し，それぞれの格助詞出現確率は表 3 のようになる．複数の意味属性で最も出現確率の高い格助詞がすべて同じであれば，その格助詞を選択する．それら異なるときは，多数決で決定する．それでも一意に決まらないときは，それぞれの格助詞の確率を平均し，一番確率の高い格助詞を選択する．

候補となる格助詞がすべて連体修飾節中で出現した場合や，意味属性と動詞の共起が存在せず，格助詞の付与が行えないときは，動詞と被修飾名詞の間に格関係が存在しないかわりに．連体修飾節中の格要素と意味的関係があるとしてノ格を付与する．例えば「底が抜けた鍋」では，文の形にした時「鍋の底が抜けた」となる．

3.4 外の関係をとる名詞の解析

3.2 項で計算した外の関係度が閾値 t を越えるときは外の関係をとる名詞としたが，その場合でも内の関係をとる可能性がある．そのため 3.3 項で述べた手法を用いて，動詞と被修飾名詞の間に埋めるべき格助詞を選択する．ただし $f_k(n,v) < 5$ の場合は，名詞を抽象化した解析は行わず，外の関係とする．なぜなら日本語語彙大系は格フレームの格要素の選択制限のために作成され，内の関係に特化したシソーラスだからである．また，外の関係をとる名詞が格関係で共起する動詞は限られており，それらの共起は，コーパス中で頻出することが多い．したがって意味クラスを用いる前に格助詞を選択できる可能性が高い．

$f_k(n,v) \geq 5$ の場合でも，連体修飾節中で候補となる格助詞がすべて出現し，格助詞を付与できないときは外の関係とする．

表 3: 名詞を抽象化した時の格助詞出現確率の例

c	v	$f_c(n,v)$	$P(r\|c,v)$		
[建造物]	オープンする	156	が 0.56	を 0.26	に 0.15
[施設]	オープンする	64	が 0.70	に 0.14	を 0.14
[労働]	オープンする	44	が 0.67	を 0.13	で 0.12

4 評価実験

本手法で用いる共起情報 $f_k(n,v)$, $f_m(n,v)$, $f_r(n,r,v)$ は，毎日新聞 9 年分の約 900 万文を KNP を用いて構文解析を行い，その解析結果から収集した．格関係での共起頻度に関しては，格が交替する可能性のあるを持つ動詞（使役，受身，可能，難易，～てある）は収集しない．また係助詞を伴って共起する場合も収集しない．連体修飾の共起については，動詞の活用形に関わらず，すべての共起対を収集した．収集した共起情報は，$f_k(n,v)$ が約 560 万組，$f_m(n,v)$ が約 140 万組であった．

テストセットとして，EDR コーパスから連体修飾節を含む名詞句を 1000 個ランダムに選択した [6]．コーパスの構文情報を利用して名詞句を抽出したが，明らかに誤っていると思われるものは人手で修正した．内／外の関係，および内の関係で埋めるべき格助詞は人手で付与した．中には「罪を認める 判決」のように，デ格か外の関係かで迷うものがあった [4]．その場合は外の関係にした．内の関係でどの格助詞を付与するか迷うものは，全ての格助詞を正解とした[1]．受身・使役形の動詞については，格の交替を考慮し，基本形での格助詞を付与した．

[1] 2 つの格助詞を選んだとき，それぞれの格助詞を正解とする文の数を 0.5 とした．

表 4: 内／外の関係の判別

	正解率
ベースライン (すべて内の関係)	79.9%
従来手法 (外の関係をとる名詞→外の関係)	81.3%
提案手法	88.8%

4.1 評価結果

テストセット 1000 個を内／外の関係に分類した結果を表 4 に示す．このとき，内の関係は，正しい格助詞が付与されているかどうかは考慮しない．

内の関係は 1000 個中 799 個存在することから，常に内の関係と判断したときの正解率は 79.9%となり，これがベースラインとなる．従来の手法とは，被修飾名詞が外の関係をとる名詞ならば，常に外の関係であるとする手法である．テストセット中の名詞が外の関係をとる名詞か否かは，人手で判断した．

格助詞の付与まで行った評価結果を表 5 に示す．評価尺度は次の通りである．

$$\text{精度} = \frac{\text{正しく解析された連体修飾節数}}{\text{本手法で解析した連体修飾節数}}$$

$$\text{再現率} = \frac{\text{正しく解析された連体修飾節数}}{\text{正解の連体修飾節数}}$$

$$F\text{-値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$

4.2 考察

本手法では，外の関係をとる名詞を外の関係度という尺度から判別したが，この尺度では判別できない名詞が存在した．1 つは出現頻度が低い名詞で，代表的なものは形容詞派生名詞「勤勉さ」「弱さ」などである．もう 1 つは，外の関係をとる名詞でも，連体修飾された時に内の関係として用いられることの方が多い名詞である．例えば「色」「音」など認識に関係ある名詞，「報酬」「返事」など因果関係を伴う名詞などがあった．

格付与の実験では，ガ格やヲ格に比べ，その他の格の精度が低い．これはその他の格がガ格やヲ格よりも多く選択されていることを表す．ガ格，ヲ格は文脈中で省略されることや，提題として係助詞とともに出現することが多く，共起頻度には反映されないことが原因と考えられる．そこでガ格，ヲ格を他の格助詞より優先して選択する手法を試したところ，全体の精度は 65.3%から 72.3%に向上した．

表 5: 全体の評価結果

	節数	精度 (%)	再現率 (%)	F-値
全体	1000	65.3	—	—
内の関係	799	63.6	64.0	63.8
ガ格	425	83.5	64.5	72.8
ヲ格	245	81.3	69.4	74.9
ニ格	68.5	44.5	71.5	54.9
デ格	29.5	8.53	23.7	12.6
ト, カラ, ヘ格	16	20.0	68.8	31.0
ノ格	15	11.1	6.67	8.33
外の関係	201	72.8	70.6	71.7

5 おわりに

本論文では，コーパスから収集した共起情報を利用し，連体修飾節を解析する手法を提案した．まず内／外の関係のどちらをとるかを「外の関係度」により仮に決めた．さらに用言と被修飾名詞の間に埋めるべき格を選択した．その結果，連体修飾節の内／外の関係の判別，格の解釈を高い精度で行うことができた．しかし，従来の格フレームを利用した解析手法も有効であると思われるので，今後両手法を統合した手法を考える予定である．

参考文献

[1] S. Ikehara, S. Shirai, A. Yokoo, and H. Nakaiwa. Toward an MT system without pre-editing – effects of new methods in ALT-J/E –. In *Proc. of the Third Machine Translation Summit*, 1991.

[2] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林. 日本語語彙大系. 岩波書店, 1997.

[3] 河原, 黒橋. 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. 自然言語処理, Vol. 8, No. 1, pp. 127–134, 2000.

[4] 丸元, 乾. 連体修飾を受ける体言の格構造の復元・コーパスに基づく「内の関係」の分析. 言語処理学会 第 6 回年次大会 発表論文集, pp. 16–19, 2000.

[5] 寺村. 連体修飾のシンタクスと意味—その 1～その 4—. 「日本語・日本文化」4 号～7 号, 1975–1978.

[6] Timothy Baldwin. The analysis of Japanese relative clauses. 修士論文, 東京工業大学 大学院情報理工学研究科, 1998.