日本語連体修飾節と被修飾名詞間の関係の解析

阿辺川 武[†]　　奥村 学[‡]

概　要

本稿では，日本語の連体修飾節と被修飾名詞の関係を決定付ける要因についていくつかの要素を想定し，それらを用いて連体修飾節内の用言と被修飾名詞に格関係が存在するか，いわゆる内の関係であるか外の関係であるかの判別を行った．従来の手法では，主に動詞と名詞の共起関係を用いて，内の関係の可能性を調査するのみで，外の関係であるかを積極的に求めることはできなかった．本稿では，共起関係に加え，外の関係をとる度合をいくつか定義し，それらの情報を用いて判別する手法を提案する．実際の連体修飾節を用いて判別実験を行った結果，従来の格フレームを用いた手法よりも高い正解率を示すことがわかった．そして格フレームを用いた手法で出力された結果を本提案手法と統合することによって，さらなる精度向上を実現することができた．

Analysis of Japanese relative clauses

Takeshi ABEKAWA[†]　　Manabu OKUMURA[‡]

Abstract

In this paper, we propose a new method of analyzing Japanese relative clauses. Japanese relative clause modification should be classified into at least two major semantic categories: case-slot gapping and head restrictive. In previous methods, only the information for judging a clause to be such as case-frames, and cooccurrence information between nouns and verbs is taken into account. Our proposed method also takes into account the information for head restrictive. In the result of experiments, we could yield higher accuracy than previous methods of using case-frames. Moreover we got higher accuracy by combining our method and case-frame method.

1　はじめに

日本語における名詞の修飾形式は多岐に渡るが，その中の 1 つに用言を含む節が名詞を修飾する連体修飾節がある．連体修飾節は，構文的には用言が連体形で名詞を修飾する形をとるが，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．本稿では寺村 [14] 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節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 [8]，文章要約 [11]，文分割 [10] 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および人手による知識が用いられてきた [1, 15]．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顕在する．また格フレームでは格スロット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であることが多い．網羅性の欠点を解消するために自動的に格フレームを構築する手法 [4] なども存在するが，これらの手法は，最初に内の関係の可能性を考慮し，内の関係でないとき外の関係であるという消去法の上に成り立っている．そのため外の関係を高精度で判別できない．

本研究では，連体修飾節と被修飾名詞の関係の解析にあたり，関係を決定付ける要因としていくつかの要素を想定し，これらの要素を用いて内/外の判別を行う手法を提案する．関係を決定付ける要素は，名詞・動詞の共起関係にとどまらず，外の関係を表す指標も用いる．これにより従来手法の欠点であった外の関係の判別を高精度で出来るようになった．また要素の多くはコーパスから統計的に求められる要素であることから，網羅性の問題にも対処することができる．

[†]東京工業大学大学院 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
abekawa@lr.pi.titech.ac.jp
[‡]東京工業大学 精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology
oku@pi.titech.ac.jp

2 関連研究

2.1 格フレームを用いた手法とその問題点

連体修飾節の解析手法の 1 つに格フレーム辞書を用いた手法がある [1]．格フレームとは，用言と取り得る格要素に対する意味的制約を記述したもので，制約の多くはシソーラスの階層構造に対応した意味属性という形で記述されている．格フレームを用いた解析手法にはいくつかの問題があり，ここでは実際に解析の流れを示しながらその問題点を分析していく．図 1 は内/外の関係の判別を，藤本らの手法 [1] を参考に独自に実装したアルゴリズムである．

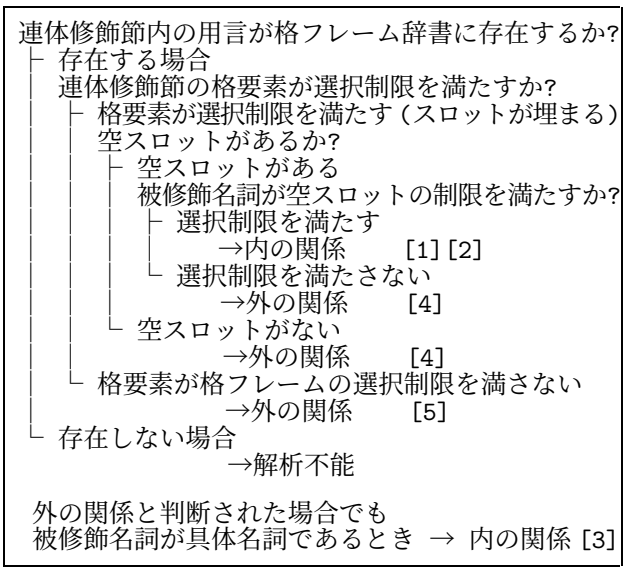

図 1: 格フレーム手法の流れ

最初に連体修飾節中の用言に対する格フレームが辞書に存在しない場合，この手法では解析できない．人手により構築された辞書では収録数に限界があり，この問題を解決するためには新たに対応する格フレームを構築するより他はない．

次に格フレームの格スロットに対する選択制限の記述法に問題がある．例えば日本語語彙大系 [2] の格フレームは機械翻訳における動詞の訳し分けのために構築されており，訳し分けが十分にできる範囲で最も一般的になるように格スロットの選択制限が記述されている．その多くは意味属性もしくは特定の単語そのもので記述されているが，一方でどんな名詞をも許容できる格スロットが存在する．つまり，その格スロットが空いていれば，被修飾名詞がいかなる種類の名詞であっても内の関係とみなしてしまう．

Baldwin[15] は，格フレームの他に外の関係になりやすい名詞や意味別に分類した動詞など人手で作成された様々な属性を定義し，機械学習手法を用いることで精度の高い結果を残している．

2.2 コーパスを用いた解析

前節で挙げた問題点の 1 つである網羅性の問題を解決するために大規模なコーパスから自動的に格フレームを抽出する研究が行なわれている．河原ら [4] は，大規模コーパスから動詞と直前の格要素の組を単位として格フレームを自動構築し，得られた格フレームを用いて，内/外の関係の判別と内の関係の場合はどの格に挿入できるかの実験を行なっている．

また，村田ら [7] は，コーパスから名詞と動詞の頻度を求め，それぞれが高頻度で出現するにもかかわらず対象となる動詞・名詞対の同時出現頻度が低いとき，外の関係であるという手法を提案している．

これまで述べてきた手法は，格フレームを用いた手法にせよ，コーパスを用いた手法にせよ，基本的には動詞と名詞が共起し得るかという内の関係の観点からの解析手法である．しかしこれでは，偶然共起しなかった動詞・名詞対や，共起関係はあるが外の関係であるという事例などを正しく解析することはできないという問題がある．したがって内の関係の判別率に比べ，外の関係の判別率は大きく低下している．それに対し，本研究では名詞の持つ外の関係をとる度合や，名詞を修飾する複数の要素といった共起関係以外の要素を考慮に入れた解析を行なっており，外の関係の判別率の向上をめざしている．

3 提案手法

3.1 内/外の関係を決定する要素

人が連体修飾節と被修飾名詞における関係を判断するとき，どのような要素を考慮して判断しているのだろうか．被修飾名詞の性質，動詞・名詞間の共起関係，連体修飾節の格スロットの充足度など多岐に渡る要素を考慮し，最終的に与えられた文脈すべてを考慮して判断していると思われる．しかし機械的な処理でそのすべてを考慮することは現在困難であるので，ここでは内/外の関係を判断する 7 つの要素を提案する．

3.1.1 名詞の内/外の関係をとる度合い

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ほぼ等しい．一方，外の関係をとる名詞「意向」「事実」「用

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

意」では，名詞の出現頻度が低いにも関わらず動詞異なり数は連体修飾関係の方が多い．これは最初に述べた，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る．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 $P_k(v|n) = \frac{f_k(n,v)}{f_k(n)}$，連体修飾関係で共起する確率を $P_m(v|n) = \frac{f_m(n,v)}{f_m(n)}$ 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．

2 つの確率分布 $P_k(v|n)$，$P_m(v|n)$ 間の類似度の計算には式 (1) の Jensen-Shannon divergence[3] を用いる．確率分布 p, q があるとき，Jensen-Shannon divergence は次のように定義される:

$$J(p,q) = \frac{1}{2}\left[D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2})\right]. \qquad (1)$$

ここで $\frac{p+q}{2}$ は，2 つの確率分布 p，q の平均である．また，$D(p||q)$ は Kullback-Leibler divergence で次の式 (2) により定義される:

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}. \qquad (2)$$

2 つの確率分布に差異があるほど，式 (1) の値は大きくなり，その名詞の持つ外の関係度が高いといえる．本研究ではこの距離を名詞の外の関係度として利用する．

3.1.2 動詞と名詞の共起関係

連体修飾節中の動詞と被修飾名詞の共起関係が内/外の関係に影響を与えることを示し，共起関係を尺度として算出する方法を説明する．まず次の例を見ていただきたい．ここでは連体修飾節中の格要素は考慮に入れずに，単純な動詞と名詞の関係に着目する．

(a) 共鳴する音　(c) 果たす役割
(b) 破壊する音　(d) 警備する役割

(a) と (b) を比較したとき，(a) では「音が共鳴する」と名詞が動詞の格要素になると想起できることに対して，(b) では一般的な解釈において外の関係であると想起しやすい．同様に (c) では「役割を果たす」と想起できるが，(d) は格関係を想起できない．これは，人間の場合，動詞と名詞の意味を考慮しながら，共起できるかできないかを判断しているとともに，実際に「名詞 格助詞 動詞」を想定し，その表現が妥当かどうかにより判断を行なっていると思われる．

人間の場合，総合的な観点から共起関係を捉えることができるが，ここでは従来の手法と同様に動詞・名詞間に格関係が実際に出現しているとき共起関係にあると判断する．つまりコーパスを参照して「名詞 格助詞 動詞」が頻出していれば共起関係があり，出現しなければ共起関係はないと判断する．

共起関係を表わす尺度として，単純に頻度をそのまま使用するのでは全体の頻度が考慮されないため，相互情報量を使用する．相互情報量は，2 つの単語がそれぞれ独立に現れる確率と同時に現れる確率との比を基に共起の強さを測るものである．

ここでは，動詞・名詞間の相互情報量ではなく，その間に介在する格助詞と名詞を 1 単位とした「名詞 格助詞」と「動詞」の間の相互情報量を考える．本研究では格助詞として「が を に で へ と から」の 7 種類を対象にしているが，格要素から被修飾名詞への転出が起こらない格助詞「より」「まで」は除いている [14]．これにより内/外の関係の判別時には，格助詞「まで」「より」を候補から外しておくことができる．

名詞 n がある格助詞 k と同時に出現する確率を $p(n,k)$, 動詞 v の出現確率を $p(v)$，名詞 n，格助詞 k，動詞 v の同時出現確率を $p(n,k,v)$ として相互情報量 $I(n,k;v)$ を次の式により求める:

$$I(n,k;v) = \log \frac{p(n,k,v)}{p(n,k)p(v)}. \qquad (3)$$

この格助詞別の相互情報量を動詞と名詞の共起尺度として利用する．つまりこの値が共起の傾向をあらわし，値が大きい程その格助詞を伴って共起しやすいと考える．共起頻度が 0 のときは，相互情報量の値を求めることができないため，ここでは非常に小さい値 M_{min} を与える．実際に，「噂 聞く」について各格助詞について相互情報量を算出すると表 2 のようになる．

3.1.3 連体修飾節中の格要素

連体修飾節中の格要素が内/外の関係に影響を与える場合がある．

表 2: 「噂 聞く」の相互情報量の値

	が	を	に	で	へ	と	から
頻度	0	52	6	7	0	0	0
相互情報量	M_{min}	-8.13	-10.3	-10.1	M_{min}	M_{min}	M_{min}

(e) 聞いてきた話 → 話を聞いてきた
(f) 落語を聞いてきた話 → (外の関係)

(g) 共鳴する音 → 音が共鳴する
(h) 楽器が共鳴する音 → (外の関係)

(e), (g) では，被修飾名詞が連体修飾節の格要素となる内の関係になるが，(f), (h) では被修飾名詞を挿入しようとする格スロットが既に充足しているので，外の関係となる．内/外の判別では最初に動詞・名詞間共起関係から挿入すべき格スロットが推定されるが，そのスロットが空いていなければ格関係をとることができず，外の関係になってしまう．

ただし，連体修飾節の格スロットがすべて埋まっているとき，すべての名詞が外の関係になるわけではない．それは名詞の性質に依存しており，例えば「娘」では次のように言い換えることができ，内の関係になる．

(i) 東京にいる娘 → 娘が東京にいる
(j) 両親が東京にいる娘 → 娘**の**両親が東京にいる

このような形式を佐藤 [13] は間接限定型と呼んでおり，連体修飾節内の格要素と被修飾名詞がノ格の関係にあるものをいう．これは動詞の格要素となるわけではないが，内の関係の一種として考える．3.1.1 節で説明した外の関係度は，ここで述べた名詞の性質と一致しており，外の関係度の値の低い名詞ほど内の関係をとりやすいといえる．

さて，この連体修飾節の格要素の影響をなんらかの尺度として求めたい．ここでは従来の格フレームを用いた手法と同様，すでに充足しているスロットには同じ格を挿入できないという制約を利用する．3.1.2 節で算出した格助詞別の相互情報量は，その格助詞を介在して共起できるかどうかを表しているので，連体修飾節中で格スロットが充足しているとき，この値を非常に小さい値 M_{min} とし，格要素として挿入できないという制約を課すことにする．

ただし，ニ格やデ格のように複数の深層格を持っている格助詞は，1 文中に複数個の表層格として出現する場合がある．また「彼が英語が話せる」のような「がが構文」として複数回出現することもある．本来はこのような場合にも対処できるように深層格を考慮し，深層格のレベルで 1 文 1 格の制約を課すべきである．しかし深層格の解析は困難をともなうので本研究では，これを考慮せず格助詞という表層格で考える．

3.1.4 連体修飾節以外の修飾要素

従来は被修飾名詞が「A の B」の形式や複合名詞のとき，主辞となる名詞だけを見て，それ以外の修飾要素は考慮していなかった．ここでは連体修飾節以外の修飾要素についても考えていく．

(k) 彼に話す目的 → (外の関係 or 内の関係)
(l) 彼に話す旅行の目的 → 彼に旅行の目的を話す

(m) 先生に教えてもらった結果 → (外の関係 or 内の関係)
(n) 先生に教えてもらったテスト結果 → 先生にテスト結果を教えてもらった

動詞を含む節によって修飾される名詞に対して，さらに別の連体修飾要素が付加されているとき，内の関係になりやすい．(k), (l) では内/外の関係の両方の可能性が考えられるが，(m) や (n) のような場合，内の関係に一意に決定される．「目的」「結果」のような外の関係をとれる名詞を考えた時，その名詞単独ではどのような内容を表わしているかわかりにくい．故に連体修飾を用いてその名詞の意味を補充するわけであるが，それは節だけでなく，形容詞や複合名詞のような他の修飾形式によっても可能である．つまり名詞が何らかの修飾形式で修飾されていれば被修飾名詞の意味は限定され，外の関係の連体修飾節で意味を補充する必要はなくなる．そのとき内/ 外が曖昧な場合でも外の関係の可能性はなくなり，内の関係に一意に定まる．

この性質をどのようにして尺度として用いるかであるが，修飾形式や修飾語の内容によりその名詞の意味を限定する度合が異なると考えられる．その度合を正しく見積もることは本研究の範囲を越えるので，ここでは連体修飾節以外の修飾要素が出現しているときは名詞の意味を限定しているとみなす．ただし「現段階」「新方式」のような 1 文字による修飾，つまり接頭辞による修飾は名詞の意味の限定度合は弱いと考え，この形式は対象外とする．したがって 2 文字以上によるなんらかの修飾形式があるとき 1 とし，それ以外は 0 とするような尺度を設定する．

3.1.5 名詞の被修飾の割合

前節の名詞の修飾形式では，実際に連体修飾節に修飾され，さらに他の修飾形式が存在する場合を考慮していたが，ここでは名詞が通常どれだけ修飾を受けているかについて考える．先程も述べたように外の関係をとる名詞の中には，その名詞単独では，どのような

内容を表しているか想定しくい名詞があり，なんらかの修飾形式により修飾を受け意味が限定されるわけである．つまりそのような名詞は常に連体修飾を受け，修飾を受けず単独で出現することは少ないはずである．

ここでは，コーパス中で名詞が出現したとき，なんらかの形式 (連体修飾節，形容詞，ノ格，複合名詞，その他) で修飾されている割合を尺度として求める．例としていくつかの名詞の被修飾の割合を表 3 に載せる．

表 3: 名詞の被修飾の割合

意向	分野	ゴルフ	彼	平均
0.983	0.973	0.341	0.155	0.460

表 3 を見ると「意向」「分野」といった外の関係をとりやすい名詞では被修飾の割合が高く，逆に「ゴルフ」「彼」といった名詞ではその値は低い．本研究ではこの値を「名詞の被修飾の割合」として用いる．

3.1.6 名詞の持つ時制の概念と動詞の時制

被修飾名詞が持つ時制の概念と連体修飾節の時制が内/外の関係に影響を与える場合がある．ここでは簡単に動詞の活用に助動詞の「た」が含まれているとき過去時制とし，それ以外は現在時制としている．

(o) 春に決定する計画 → (外の関係)
(p) 春に決定した計画 → 計画が決定した

(q) 繰り返された歴史 → (外の関係)
(r) 繰り返される歴史 → 歴史が繰り返される

(o) と (p) および (q) と (r) では連体修飾節の時制のみが異なる．(o)，(q) では外の関係であるという解釈ができるが，(p)，(r) では内の関係でしか解釈ができない．

これは，過去もしくは未来の概念を有する名詞では，一般に外の関係をとるとき，連体修飾節の時制がその名詞の持つ時制と同じでなければならないという制約が働くためである．連体修飾節の時制が名詞の時制の制約と矛盾すると，名詞の内容の説明を行うことができず，格関係を持つ内の関係でしか修飾できない．

この制約の例として時制の概念を持つ名詞「計画」「記憶」について，それぞれを被修飾名詞とする連体修飾節を含む文を 100 文ずつ用意し，時制と内/外の関係を調べた結果を表 4 に載せる．

表 4: 時制と内/外の関係

	計画		記憶	
時制	内の関係	外の関係	内の関係	外の関係
現在	6	89	12	**0**
過去	5	**0**	5	83

時制の観点から見れば,「計画」は現在時制*が多く,「記憶」では過去時制が多く使用されており，名詞の持つ時制の概念と一致している．さらに注目すべき点は，両者ともに外の関係で使用されるとき，常に時制が一方に限定されているということである．例えば「計画」では，外の関係で修飾されるとき，現在時制の事例が 89 例に対して，過去時制の事例は 1 つもない．また，別の言い方をすれば過去時制で修飾されるときは必ず内の関係であるということである.「記憶」に関しても同様なことがいえる．

上記の仮定が事実であるならば，修飾節の時制は内/外の関係を判断するよい指標となり，名詞の持つ時制の概念と逆の時制で修飾されているときは内の関係になるはずである．

次にこの時制による影響を尺度として算出する方法を説明する．先ほど述べたように時制の概念を持つ名詞は，外の関係の連体修飾節で修飾されるとき，両者の時制が一致する場合が多い．つまり過去あるいは現在の時制に偏りが大きいということである．そこでコーパスからこの偏りを求める．名詞が連体修飾されるときの頻度とその時の過去時制の頻度を求め，過去時制のとる割合 R_{past} を求める．ただしこの簡易な方法では，外の関係と内の関係の連体修飾が混在した割合になってしまうので，外の関係で修飾されたときだけの割合が求められない．外の関係で修飾されているかを正確に判断することは難しいが，連体修飾する動詞が格関係でも多く共起する場合は内の関係であるとして，その動詞を使用した連体修飾の頻度を数えないことにした．例としていくつかの名詞について R_{past} を求めると表 5 のようになる．

表 5: 過去時制の割合

計画	記憶	場所	人々	平均
0.032	0.958	0.333	0.422	0.322

時間の概念のない「場所」「人々」では平均値に近いのに対して,「計画」「記憶」では 0 または 1 に近い値を取っており，時制の概念と一致する．この値が平均値から離れる程，その名詞が時制の概念を持つ度合が高いとする．

実際に解析に使用するときは，連体修飾節の時制を考慮する必要がある．ここでは次のようにして過去時制の割合を変換した値 V_{past} を求める．

$$V_{past} \begin{cases} R_{past} - Avg_{past} & \text{現在時制の時} \\ Avg_{past} - R_{past} & \text{過去時制の時} \end{cases} \quad (4)$$

AVG_{past} は全名詞の過去時制をとる割合の平均である.「計画」のような過去時制の割合が小さい名詞では，連体修飾節の時制が過去のとき V_{past} の値が正になり，現在時制のときは V_{past} が負になる．この尺度が正で，値が大きい程，時制の影響が大きいといえる．

*日本語では未来時制がないため現在時制として扱っている．

3.1.7 「という」の前接割合

(r) 給料がアップするという噂
(s) 削減しなければならないという意見

内/外の関係を示す 1 つの指標に連体修飾節と被修飾名詞の間に「という」の介在がある．寺村 [14] は，外の関係をとる名詞を，常に「という」が必要な「命令」「期待」といった名詞,「という」の介在が任意な「事実」「習慣」といた名詞,「という」の介在が不可能な「姿」「写真」という名詞に分類している．また (s) のようにモダリティの許容度を高めるために使用される場合もある．コーパス中である名詞が出現したとき「という」が前接する割合を求めれば，外の関係をとるすべての名詞の識別はできないが，一部の名詞を識別することができる．

ここでは,「という」の他に似た機能を持つ「との」「とする」も含める．名詞が節によって連体修飾されるとき，これら機能語が介在する割合を尺度とする．例としていくつかの名詞の前接割合を表 6 に載せる．

表 6: 「という」の前接割合

意見	噂	場所	人々	平均
0.335	0.246	0.007	0.008	0.007

3.1.8 除外型をとる動詞

(t) 阪神を除いたプロ野球チーム
(u) 怪我人を除く乗客

被修飾名詞が集合的要素で，かつ連体修飾節に集合に関する動詞があるとき外の関係になる場合がある．この型は特別で，今まで説明した外の関係をとる名詞と異なり，具体物を表す名詞，つまり外の関係度の低い名詞が被修飾名詞になるということである．したがって，この型の内/外の関係は名詞よりむしろ動詞に依存する．このような型をとる動詞すべてを列挙することは難しいので，本研究では (t)，(u) のような「除く」の意味を持つ動詞を対象とする．連体修飾節に「除」を含む動詞があるとき 1，それ以外の動詞のとき 0 とする尺度を設定する．

3.2 内/外の関係の判別法

上記で説明した 7 つの要素それぞれが連体修飾節の内/外の関係を決定づけていると考えられるが，この尺度を利用して判別する場合はどのように組み合わせればよいだろうか．各尺度の最適な組み合わせや，重み，閾値を設定してモデルを構築していくことは難しい．そこで，本研究では訓練事例から判別に有効な尺度の組み合わせを自動的に学習する機械学習を用いた手法の 1 つである決定木 [12] を用いる．

決定木は，素性がどの位置で判別に利用され，その時の正解率がどの程度であるかなどを木構造の形で示し，学習された結果が人間に理解しやすい．一方で決定木には素性の種類が増加すると過学習をしてしまうという欠点があるが，今回使用する素性数は比較的少なく，また少ない事例数の枝を枝刈りすれば，過学習の影響は小さいと思われる．

3.3 評価実験

評価実験により，これまでに述べた連体修飾節の内/外の関係を決定する要素の有効性を実証する．

3.3.1 実験方法

本手法で用いる共起情報は，毎日新聞 11 年分，および読売新聞 11 年分の記事に対し，KNP[5] 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，格関係が約 6080 万組，連体修飾関係が約 1240 万組であった．

テストセットとして，EDR 電子化辞書 [9] の日本語コーパスから連体修飾節を含む名詞句をランダムに 1000 名詞句選択した．コーパスに付与されている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．その後，内/外の関係を人手で付与した．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 [6]．その場合は外の関係にしている．日本語語彙大系 [2] に収録されている格フレームを用いた手法と比較するため，テストセット中で，被修飾名詞が名詞意味属性体系に存在しない事例，連体修飾節内の動詞が構文体系に存在しない事例を実験の対象外とした．また提案手法で使用する尺度は統計的に算出するので，被修飾名詞の頻度が 1000 回以下の事例も対象外とした．その結果，内の関係，外の関係の内訳は表 7 のとおりである．

表 7: テストセットの内訳

内の関係	外の関係	計
580 句	169 句	749 句

実験で使用した決定木の導出ツールには C5.0[12] を利用した．学習器に渡す素性として 3 節で説明した 13 種類，外の関係度，相互情報量 (ガ，ヲ，ニ，デ，ヘ，ト，カラ)，被修飾名詞の修飾形式，名詞の被修飾の割合，時制の影響,「という」の前接割合，除外動詞の各素性を用いる．過学習を防ぐため，適用事例の少ないノード (本実験では 10 事例未満) を枝刈りする設定にしている．評価は訓練事例を 5 分割し，交差検定を用いて評価を行なう．評価尺度は，内/ 外の関係を正しく判別した事例数を全体数で割った正解率と，内/外の関係それぞれについての精度と再現率を用いる．

```
┬除外動詞 = 1: 外の関係 (除外型) (22/2)
└除外動詞 = 0:
  ├──外の関係度 <= 0.212: 内の関係 (444/6)
  └──外の関係度 >  0.212:
     ├──被修飾名詞の修飾形式 = 1: 内の関係 (84/17)
     └──被修飾名詞の修飾形式 = 0:
        ├──ヲ格 >  -9.10: 内の関係 (28/4)
        └──ヲ格 <= -9.10:
           ├──名詞の「という」の前節割合 >  0.027: 外の関係 (105/14)
           └──名詞の「という」の前節割合 <= 0.027:
              ├──被修飾の割合 <= 0.735: 内の関係 (25/2)
              └──被修飾の割合 >  0.735:
                 ├──ガ格 <= -13.1: 外の関係 (31/5)
                 └──ガ格 >  -13.1: 内の関係 (10/2)
```

図 2: 構築された決定木

3.4 実験結果

以上の実験条件で評価実験を行った結果を表 8 に載せる．また，実験の比較対象として，すべて内の関係とした場合 (Baseline) と 2.1 節で述べた格フレームを用いた手法も同様に表 8 に掲載する．

表 8: 実験結果

		内の関係		外の関係	
	正解率	精度	再現率	精度	再現率
Baseline	0.774	0.774	1.000	-	-
格フレーム	0.830	0.868	0.921	0.657	0.521
提案手法	0.902	0.931	0.942	0.794	0.762

参考としてテストセットのすべての事例を用いて構築された決定木を図 2 に載せる．各ノードに分類規則が記述され，枝においては分類結果およびその規則により分類された事例数 (適用事例数/誤り数) を表している．

3.5 考察

構築された図 2 の決定木を見ると，外の関係の特別な形である除外型を除き，内/外の判別に一番寄与した素性は外の関係度であった．外の関係度がある閾値以下であるときには，98.6%(=438/444) と非常に高い精度で内の関係であった．これにより連体修飾する動詞と格関係で共起する動詞の分布が類似しているほど内の関係をとりやすい名詞であるという仮定が正しいことがわかる．この尺度は名詞だけを見ており，連体修飾節内の動詞を考慮していない．つまり連体修飾節の内容にかかわらず，一意に内の関係であると決定できるということである．

図 2 の決定木によると，被修飾名詞の修飾形式による判別が続く．仮定の通り被修飾名詞以外の修飾形式がある場合は内の関係になると判断されているが，この規則による正解率はあまりよくない．やはり修飾する要素により名詞を限定する度合が異なるため，修飾要素の有無のみを判断の材料とするべきではないと思われる．

その他，適用事例数でみると「という」の前節割合に基づく規則による判別が多い．この規則により分類された事例の被修飾名詞を見ると，すべて外の関係をとり得る名詞であった．誤りの多くは共起関係の欠如のためであり，動詞には「超える」「異なる」「優れる」のような比較の意味がある動詞が多いように見うけられた．

構築された決定木には，時制の影響による規則が出現しなかった．これはこの素性が判別に有効でないことを示すが，原因として適用事例が少ないこともあげられる．テストセット中に時制の概念を持つ名詞が少なく，その上，時制の概念と逆の時制が連体修飾節に出現する事例がテスト事例に存在しなかったためである．

最後に Baldwin[15] による手法との比較を行う．Baldwin の手法では内/外の関係の判別の他に，内の関係では連体修飾節のどの格になるかまでを一度に解析している．テストセットも異なるため一概に比較はできないが，格解析を含め全体で 89.3%という非常に高い精度を実現している．これは決定木の素性として，人手で構築された多くの知識を用いているが，冒頭で述べたように網羅性の欠如という点で問題がある．また外の関係の判別については人手による素性を用いても有効な規則が導けなかったと述べており，本手法で提案した素性を組み合わせることができれば，さらなる精度向上が見込まれると思われる．

4 格フレームを用いた手法との統合

4.1 確度の高い出力結果の利用

3.4 節の実験結果より，格フレームを用いた手法は提案手法よりも正解率が悪いという結果になった．そこで格フレームを用いた手法について，適用された規則ごとの内/外の判別の正解率を表 9 に載せる．それぞれの規則の番号は 2.1 節の図 1 における規則の番号と対応している．正解率を見ると適用された規則ごとに正解率が異なることがわかる．したがって格フレームを用いた手法の中で確度の高い事例のみを効果的に用いれば，提案手法の精度をさらに向上できると考えられる．

表 9: 適用規則ごとの正解率

		規則単独の正解率		追加後の全正解率	外の関係度を除く	
結果が	1. 制約がすべての名詞を許容	0.849	(319/377)	0.902	0.600	(81/135)
内の関係	2. 制約が単語 or 意味属性	0.904	(178/197)	0.902	0.691	(38/55)
	3. 被修飾名詞が具体物	0.902	(37/41)	0.902	0.500	(4/8)
結果が	4. 被修飾名詞が制限に適合しない	0.795	(70/88)	**0.905**	**0.860**	(49/58)
外の関係	5. 連体修飾節が制約に適合しない	0.391	(18/46)	0.901	0.643	(18/28)
	規則の番号を追加	0.830	(622/749)	0.902	0.671	(190/283)

まず格フレームの手法で確度の高い規則で出力された結果を，提案手法で用いた決定木に組み入れる方法を考える．格フレームの手法により内/外の判別を行った事例に対して 2 値の素性を付与する．ある規則により出力された事例のみに 1 を付与し，それ以外の事例に対しては 0 を付与する．使用された規則により内/ 外の関係が決定されているので，内/外の関係のどちらであるかは考慮しない．規則別に決定木に組み入れた結果が表 9 の追加後の正解率である．

次に，確度の高い規則のみを用いるのではなく，どの規則が使用されたかを素性として組み入れる実験を行なった．格フレームの手法で適用された規則の番号をそのまま決定木の素性として用いた結果が表 9 の「規則の番号を追加」の行である．

4.2 考察

決定木に格フレームの確度の高い出力結果を素性として追加する手法では，結果的には，4. の規則で外の関係として解析された事例を素性として追加した場合のみが全体の正解率向上に寄与した．規則単独の正解率はそれ程高くないにもかかわらずである．これは提案手法で用いられている素性との事例の重複の差であると思われる．特に提案手法において適用事例が多く正解率が非常に高い外の関係度の素性と関係している．格フレームを用いた手法におけるどの規則でも確度は外の関係度よりも低いため決定木構築では外の関係度よりも下位のノードになってしまい，適用事例は外の関係度により適用された事例以外のものとなる．その中での確度が全体の正解率に寄与すると思われる．表 9 の「外の関係度を除く」にその時の正解率を載せているが，これを見ると 4. の規則が一番正解率が高い．そのため全体の正解率も高くなったものと思われる．

5 まとめ

本研究では，連体修飾節と被修飾名詞の関係の解析にあたり，関係を決定付ける要因としていくつかの要素を想定し，これらの要素を用いて内/外の判別を行う手法を提案した．提案手法は，従来手法の欠点であった外の関係の判別を精度よく行えるようになり．また要素の多くはコーパスから統計的に求められる要素であることから，網羅性の欠如にも対処することができる．

実際の言語処理への応用を考えたとき，内/外の関係の判別の次には，内の関係において被修飾名詞が動詞のどの格要素になるかを決定する必要がある．これは，内/外の判別と同時に求めていく手法と，内/外の関係を判別後，内の関係に対してだけ格解析を行う手法とがあるが，提案手法の性質上，後者の手法を考えている．この手法の実験は，今後の課題としたい．

参考文献

[1] 藤本敬史, 池原悟, 村上仁一, 表克次. 複文における底の名詞と修飾部の内と外の関係の判断規則. 言語処理学会第 8 会年次大会, pp. 679–682, 2002.

[2] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩己, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系 – 全 5 巻 –. 岩波書店, 1997.

[3] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE TRANSCATIONS ON INFORMATION THEORY*, Vol. 37, No. 1, pp. 145–151, 1991.

[4] 河原大輔, 黒橋禎夫. 用言の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. 9, No. 1, pp. 3–19, 2002.

[5] 黒橋禎夫, 長尾眞. 日本語構文解析システム KNP 使用説明書 version 2.0b6. Technical report, 京都大学大学院情報学研究科, 1998.

[6] 丸元聡子, 乾裕子. 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.

[7] 村田真樹, 井佐原均. 頻度に基づく正の例からの負の予測. 情報処理学会研究報告 144-NL-15, pp. 105–112, 2001.

[8] 成田一. 連体修飾節の構造特性と言語処理—日本語らしい表現の機械翻訳と応用技術—. 日本語の名詞修飾表現, pp. 67–126. くろしお出版, 1994.

[9] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第 2 版. Technical Report TR–045, 1995.

[10] 野上優, 藤田篤, 乾健太郎. 文分割による連体修飾節の言い換え. 言語処理学会 第 6 回年次大会 発表論文集, pp. 215–218, 2000.

[11] 大竹清敬, 増山繁. 多重修飾に着目した文内要約: 削除型換言. 言語処理学会 第 7 回年次大会併設ワークショップ, pp. 59–64, 2001.

[12] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[13] 佐藤龍一. 日本語の連体修飾節の意味解析. 修士論文, 東京工業大学 大学院情報理工学研究科, 1989.

[14] 寺村秀夫. 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.

[15] Timothy Baldwin. The parameter-based analysis of japanese relative clause constructions. 情報処理学会研究報告 134-NL-8, pp. 55–62, 1997.