頻度情報を利用した相対名詞の判別

阿辺川 武　　奥村 学
東京工業大学大学院 総合理工学研究科　東京工業大学 精密工学研究所
abekawa@lr.pi.titech.ac.jp　　oku@pi.titech.ac.jp

1　はじめに

日本語では名詞を修飾する形式は多岐に渡るが，その 1 つに連体修飾節による修飾形式がある．連体修飾節内の動詞と修飾される名詞の関係を見ると，名詞が動詞に対して格要素となり得る関係と，格要素になり得ない関係に大別される．寺村 [6] は前者を「内の関係」，後者を「外の関係」と呼んでいる．以降では，名詞を修飾する節を連体節，修飾される側の名詞を主名詞と呼ぶことにする．

「外の関係」は連体節と主名詞の意味的関係から，さらに 2 つに分類することができる．

a）北海道を旅行した記憶
b）今朝学校へ行く前

a) は，一般に内容補充と呼ばれる形式で，連体節が主名詞の内容を補充・説明している．一方 b) は，連体節の表す概念に対し，主名詞が相対的な地点・時点を表す形式である．上記の例では「学校へ行く」という事柄に対して，相対的に「前」の時間を表している．本論文では b) のような，連体節と主名詞が相対している関係を相対的関係と呼ぶ．

与えられた連体節と主名詞の関係を求める連体修飾節の解析は，機械翻訳，言い換え，要約，意味解析など多くの処理で必要とされる．従来の相対的関係の解析では，主名詞の辞書項目を引き，その名詞か相対的関係をとれるかどうかの記述によって判断が行われてきた [1]．相対的関係で頻繁に使用される語に対しては，辞書でもその用法が反映されるが，低頻度語や頻出する語でも相対的関係での用法が少ない語に対しては相対的関係についての記述は存在しないことが多い．頑健な解析システムの構築をめざすならば，このような語においても正しい用法を記しておくことは必須である．

本論文では，既存の辞書に存在する相対的関係をもつ名詞に特有な性質を大規模コーパスの頻度情報から抽出し，その性質を素性とした機械学習を用いて従来の辞書に存在しない相対的関係を持つ名詞を見つけ出すことを目的とする．

2　相対名詞

連体修飾節の相対的関係において，時間や空間を軸にとり，連体節と主名詞の関係を簡単に図示すると下図のようになる [4]．この例では，連体節の表す内容が

基準点を示し,「まえ」と「あと」が基準点に対する前後の事象を示している．また基準点が一定の長さを持つ場合は，その基準点の内側を表す「途中」「間」のような名詞も存在する．本論文では，このような基準点に対し相対的な意味を持つ名詞を「相対名詞」と呼ぶことにする．

連体節と相対的な関係を持つ名詞は，名詞節全体として表す意味により以下のような分類ができる．

時間に対する相対名詞	徹夜した**翌日**
	学校を卒業した**以後**
空間に対する相対名詞	彼が座っている**隣**
	火事が起きた**近く**
事象に対する相対名詞	円が下がる**一方**で
	遊びに行く**代わり**に

この他にも時間・空間の両方の用法で使える名詞に「前」「直後」「中」などがある．

本論文では，基準点に対して相対的関係を持つ名詞を相対名詞とするが，この他に相対的関係には連体節と主名詞との間に因果関係が存在する形式がある．

c）彼が行動を起こした結果
d）工場で働いた報酬

因果関係では，連体節が前提条件となり主名詞がその後に生ずる結果を示している．因果関係の主名詞として出現する名詞の多くは，e), f) のように連体節の内容により内容補充の関係や内の関係をとることができる．

e）誰も付いてこなくなる（という）結果
f）会社からもらった報酬

したがって因果関係の場合，主名詞を見ただけでは相対的関係であるかは決まらず，連体節の意味内容と比

較して初めて関係が求まる．一方，先で述べた相対名詞を主名詞とする連体修飾節は，連体節がどのような内容であれ，主名詞を見ただけで相対的関係であるかが決まる．つまり相対名詞は，名詞そのものが持つ相対的意味合いが他の名詞より強い名詞といえる．

以上より，本論文において判別しようとする相対名詞は，連体節の内容に関わらず常に相対的関係をとる名詞とし，連体節の内容により内容補充や内の関係となるような名詞は判別の対象外とする．

3　相対名詞の判別

相対名詞は，それ以外の名詞と比較して特有な性質がある．本手法ではそのような性質を，大規模コーパスの頻度情報から素性として抽出し，機械学習器を用いて相対名詞を判別する．

機械学習器には決定木を使用し，訓練セットに既存の相対名詞を分類した辞書を用いて学習を行う．そして構築された決定木を使用して判別を行い，既存の辞書に登録されていない相対名詞を見つけ出す．

本節では相対名詞に特有と思われる素性について説明するが，その前に素性を算出する各種頻度情報について定義する．ある名詞 n について，連体節によって修飾される頻度を $f_m(n)$，格関係で出現する頻度を $f_k(n)$ とする．また，名詞 n が動詞 v を含む連体節で修飾される頻度を $f_m(v,n)$，名詞 n が動詞 v に格関係で係る頻度を $f_k(v,n)$ とする．

3.1　判別に使用する素性

「という」の介在割合

g）身長が 2m あるという人
h）デフレを助長するという意見
i）お金を節約すべきだという建前

連体修飾において「という」が動詞と名詞の間に介在する形式はいくつかある．g) のように本来の「〜と言われる」という意味で使用される場合，h) のように引用が少し形式化して発話や思考に関する名詞に対して使用される場合，そしてさらに形式化して i) のように話者の気持ちが含まれるときにモダリティの許容度を上げる場合がある．相対名詞は上記のどの場合にも使うことはできないので連体修飾における「という」の介在割合は低いといえる．名詞 n が連体節により修飾されたとき「という」が介在している回数を $f_{toiu}(n)$ とすると,「という」の介在割合 P_{toiu} は式 (1) で表される:

$$P_{toiu} = f_{toiu}(n) \ / \ f_m(n) \qquad (1)$$

形容詞による修飾の割合

相対名詞は形容詞による修飾の割合が少ない．例えば「広い場所」「近い未来」と表現できるが,「広い横」「近い翌年」のように表現することはできない．一般に相対名詞を主名詞とする連体節では，静的な表現である形容詞よりも動的な表現である動詞が好まれる．奥津 [4] は次のような例文を用いて相対名詞に対して形容詞が使われにくいことを述べている．

j）?物価がこんなに高い前
k）　物価がこんなに高くなる前

名詞 n が格関係で出現したとき，同時に形容詞により修飾された回数を $f_{adj}(n)$ とすると，形容詞による修飾の割合 $P_{adj}(n)$ は，式 (2) で表される:

$$P_{adj} = f_{adj}(n) \ / \ f_k(n) \qquad (2)$$

格助詞が後接する割合

時間的・空間的相対名詞は，時間や空間に対して相対的な意味合いを持つが，名詞節全体としては時間・空間の意味を表すことが多い．一般的に時間・空間の意味合いを持つ名詞に対して格助詞「ニ」「デ」が後接する割合は高いが，格助詞「ガ」「ヲ」が後接する割合は低い．名詞 n に格助詞「ガ」「ヲ」「ニ」「デ」が後接する回数を $f_{\{ga,wo,ni,de\}}(n)$ とすると，格助詞が後接する割合 $P_{\{ga,wo,ni,de\}}(n)$ は式 (3) で表される:

$$P_{\{ga,wo,ni,de\}} = f_{\{ga,wo,ni,de\}}(n) \ / \ f_k(n) \qquad (3)$$

完了時制の割合

l）今朝学校へ行く前にコンビニに寄った．

例文 l) は主節の完了時制に対し，連体節の時制は現在時制である．これは日本語の時制の特色で，英語では完了時制となるところである．日本語では，連体節の時制は主節の表す事象が起こった時点を基準とし，基準より前に連体節の表す事象が完了していれば完了時制になり，完了していなければ現在時制になる．主名詞が「前」の場合には，主節の事象が生ずる「前」なので連体節は現在時制になる．逆に「後」のような名詞の場合は，連体節の表す事象は既に完了していることから完了時制をとることになる．このように時間を表す相対名詞は連体節に完了時制をとるか現在時制をとるかは明確であり，完了時制 (解析ではタ形) の割合を測定することでこの傾向がわかる．名詞 n を修飾する連体節に含まれる動詞の時制が完了時制である回数を $f_{past}(n)$ とすると，完了時制の割合 $P_{past}(n)$ は式 (4) で表される:

$$P_{past} = f_{past}(n) \ / \ f_m(n) \qquad (4)$$

表 1: 素性値の例

素性	間	末	場所	常識	議会
$P_{toiu}(n)$	.0027	.0076	.0044	.5272	.0042
$P_{adj}(n)$	.0039	.0002	.1680	.0304	.0051
$P_{ga}(n)$	.0091	.0029	.0959	.1016	.1791
$P_{wo}(n)$	.0611	.0081	.2846	.3497	.1489
$P_{ni}(n)$	.5217	.6640	.3319	.2064	.3126
$P_{de}(n)$	.3556	.0943	.1720	.1198	.2985
$P_{past}(n)$	.0923	.9788	.3408	.2378	.4158
$E(n)$	5.997	4.661	6.180	4.974	4.812

共起する動詞の偏り

一般に名詞は全ての動詞と一様に格関係で共起しているのではなく，共起しやすい動詞や共起しない動詞など偏りが存在する．内の関係を多くとる名詞では，連体節で共起する動詞に対しても格関係と同様な偏りが見られる．しかし外の関係を多くとる名詞の場合は，連体節の動詞と格関係を伴わずに共起していることから，動詞の種類は格関係よりも多種である．相対名詞も外の関係をとる名詞の一種であり，その傾向は高いといえる．酒井 [5] は，連体節の動詞の偏りを「修飾多様性」とし，動詞ごとに修飾される確率からエントロピーを算出している．エントロピーが高い値であるほど，その名詞は多種多様な動詞の修飾を受けているといえる．名詞 n が動詞 v を含む連体節で修飾される確率を $P_m(v|n) = f_m(v,n)/f_m(n)$ とし，名詞 n を連体修飾で修飾する動詞の集合を $V_m(n)$ としたとき，エントロピー $E(n)$ は式 (5) の式で表される:

$$E(n) = -\sum_{v \in V_m(n)} P_m(v|n) \log P_m(v|n) \qquad (5)$$

最後に「間」「末」「場所」「常識」「議会」の名詞に対するそれぞれの素性値を表 1 に示す.

3.2 低頻度名詞の頻度情報

前節で述べた素性を算出するにあたり，ある一定数以上の頻度を持つ名詞に関しては有効な統計値を得ることができるが，頻度の低い名詞では，正しい統計値が得られない可能性がある．

本手法では，低頻度名詞に対して類似度を計算し，類似度が一番高い名詞とともに統計情報を計算した．類似度を計算する名詞集合には，日本語語彙大系 [2] の意味属性辞書を用い，低頻度名詞と同一意味属性内の名詞を集合とした．意味属性辞書に含まれていない名詞は，コーパス内で頻出する上位 10000 個の名詞を集合とした．また，低頻度名詞であるかを決める基準は，連体節によって修飾された頻度が 100 回未満の名詞とする．

表 2: 類似名詞の例 (頻度は連体節で修飾された頻度)

名詞	頻度	類似名詞	合計頻度
近辺	28	付近	1374
背後	78	裏	714
おかし	9	菓子	548
短縮	35	削減	678

確率分布の類似度の計算には式 (6) の Jensen-Shannon 距離を用いる．2 つの名詞の格関係で共起する確率分布 p，q があるとき，Jensen-Shannon 距離は次のように定義される．

$$J(p,q) = \frac{1}{2}\left[D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2})\right] \qquad (6)$$

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 距離で次の等式 (7) により定義される:

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \qquad (7)$$

低頻度名詞に対し類似名詞を求めた例を表 2 に示す.

4 実験・考察

4.1 訓練データ

本手法は，機械学習の 1 つである決定木により相対名詞の判別を行う．実験の手順について説明する前に，学習に使用した訓練データについて説明する．訓練データには，IPAL 名詞辞書 [3] を用いた．IPAL 名詞辞書には名詞が 1081 語含まれており，意味・用法別に分類されている．その内，相対的関係を持つ名詞の分類には 53 語が含まれている．本実験では，連体節による修飾を受けても相対的関係をとりにくい名詞 (西部，北部) や，因果関係を含む名詞 (形跡，影) を除き，さらに独自に相対名詞であると判断した名詞を加えた計 61 語を利用した．以下に本実験で学習に使用した相対名詞の内訳を示す.

- 時間に対する名詞 (5 個) — 以前，最後，最初など
- 空間に対する名詞 (42 個) — 右，奥，横，下など
- 事象に対する名詞 (5 個) — 以下，以上，逆など
- 時間・空間に対する名詞 (9 個) — 前，後，間など

4.2 実験

3.1 節で説明した素性が，実際に相対名詞を判別する素性として有効かどうかを確かめるための実験を行った．実験の手順を図 1 に示す．まず IPAL 名詞辞書に

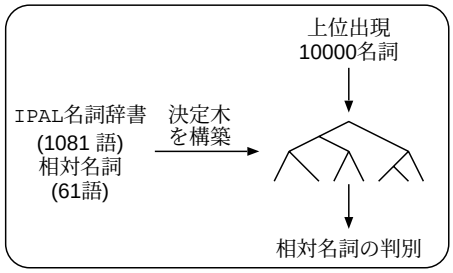

図 1: 実験の手順

表 3: 分類結果

	名詞数	例
相対名詞	131(31.9%)	矢先, 途端, 並び
相対的関係を持つ名詞	107(26.0%)	年末, 社内, 場外
相対名詞以外	173(42.1%)	
– 形式名詞	14	際, 故, ぐらい
– 固有名詞	33	岐阜, パリ, ＮＹ

含まれている名詞を訓練集合として決定木を構築する．そして構築された決定木を用いて，コーパス中で頻出した上位 10000 個の名詞について相対名詞であるかの判別を行う．

実験で用いる頻度情報を獲得するコーパスには読売新聞 11 年分，約 2,100 万文を用いた．各文を KNP を用いて構文解析し，文節の係り受け関係を求める．そして名詞に対する各種頻度を求め，3.1 節で説明した各素性を算出した．

決定木の構築アルゴリズムには C5.0 を使用し，素性には 3.1 節で使用した数値をそのまま利用した．その結果，クローズドテストにおける分類精度は，1.8%のエラー率であった．

構築された決定木をもとに，頻度上位 10000 個の名詞の分類を行なった．分類結果を表 3 に示す．相対名詞と分類された名詞は 411 個で，実際に相対名詞であった名詞は 131 個であった．適合率を計算すると 31.9%となる．

4.3 考察

まず相対名詞を判別するためにどの素性が有効であったかを考える．決定木を構築する素性として 8 つの素性を用いたが，実際に構築された決定木を観察すると完了時制の素性が利用されていなかった．これは訓練集合において，時間に対する相対名詞の数が少なかったことが原因と思われる．

次に，適合率が 31.9%と低い原因について考える．訓練セットにおいて相対名詞の数が少なく，相対名詞以外の名詞をフィルタリングするのに十分な性能が得られなかったと思われる．本論文では，常に相対的関係をとる名詞のみを相対名詞とするという厳しい基準なので適合率が低くなった．相対的関係をとることのできる名詞に基準を緩和すると適合率は 57.9%となる．

また，相対的関係をとらない名詞も数多く誤って分類されたことには，2 つの原因が考えられる．1 つは相対名詞でもない低頻度名詞に対して，類似名詞が相対名詞になってしまった場合である．正しく類似度計算ができていないことが原因である．2 つめは場所を表す名詞，特に都市の固有名詞が多く分類されていたことである．訓練データの偏りと共に今回の実験に使用した素性では，場所を表す名詞と相対名詞を分類するには不十分であった可能性がある．

「年末」や「社内」のように，相対的関係をとることもあるが連体節の内容により他の関係をとる名詞には 1 つの傾向がある．それは「年」「社」のように具体的な意味を持つ漢字に「末」「内」のような相対的意味を持つ漢字が後接して，熟語内で相対的関係が完結しているということである．相対的意味を持つ漢字が後接している場合，1 つの名詞と考えずに個々の形態素として考えた方がよいかもしれない．

5 おわりに

本論文では，動詞を含む連体修飾節によって修飾されたとき常に相対的関係をとる名詞を相対名詞と定義し，頻度情報から算出された素性を元に相対名詞であるかの分類を試みた．低頻度名詞においては，動詞との共起が類似した名詞を用いることで素性を算出した．分類の結果，適合率は 31.9%であった．今後は，類似度名詞の選択手法の改良および，決定木を 2 段階で利用するような決定木の適用手法の改良を行う予定である．

参考文献

[1] 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.

[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] 情報処理振興事業協会. 計算機用日本語基本名詞辞書 IPAL(Basic Nouns)–解説編–. 1996.

[4] 奥津敬一郎. 生成日本文法論. 大修館書店, 1974.

[5] 酒井浩之, 増山繁. 連体修飾表現の省略に関するコーパスからの知識獲得. 言語処理学会 第 8 回年次大会 発表論文集, pp. 627–630, 2002.

[6] 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.