意味属性を用いた形容詞の係り先解析

車井 登† 池原 悟† 村上 仁一† 大西 真理子‡
†鳥取大学工学専攻
‡神戸日本電気ソフトウェア
{kurumai,ikehara.murakami}@ike.tottori-u.ac.jp

1 はじめに

機械翻訳などでは,言語表現のもつ構造の曖昧性が大きな問題となっている.なかでも日本語名詞句は複雑な構造と多様な意味表現をもっている.

日本語名詞句の解析では,コーパスに基づく方法として,単語間の共起情報を用いて係り先を決定する方法 [1] 等が提案されている.しかし,表層的な情報による解析のため,十分なデータと多くの計算量を必要とする.

最近は名詞の意味属性などによって意味を考慮した解析が行われるようになり,従来の文法的情報の範囲を越えた解析が可能になってきた.日本語名詞句の解析では,「名詞+の+名詞+の+名詞」を対象に一般名詞意味属性体系 [2] を用いた係り受け解析 [3] がある.しかし,形容詞を対象にした研究はあまり行われていない.

本研究では「形容詞+名詞+助詞+名詞」を対象に意味属性を用いた係り受け解析を行い,その有効性を調査する.まず,助詞によって「形容詞+の型名詞句」と「形容詞+並列型名詞句」とに分ける.「形容詞+の型名詞句」の係り受け解析では,「形容詞+名詞」の頻度統計をとり,その値を2単語間の結合力とする.そして,解析対象となる名詞句を「形容詞+名詞A」,「形容詞+名詞B」とし,2つの結合力より強い方を係り先とする [4].このとき,字面での係り受け解析と意味属性による係り受け解析を段階的に行うことで精度の向上を目指す.次に,「形容詞+並列型名詞句」では,形容詞がより抽象的な名詞に係るという仮定のもとに,意味属性の深さから係り先を決定する手法を提案する.このとき,係り受け解析に必要な意味属性だけを選択し,木構造を再構築することで,精度の向上と意味属性の圧縮を行う.

2 形容詞の係り先に関する曖昧性

初めに,「形容詞+名詞句」において最も基本的な形である「形容詞+名詞A+助詞+名詞B」を2つのタイプに分ける.

2.1 形容詞+の型名詞句

「形容詞+の型名詞句」は「形容詞+名詞A+の+名詞B」の形をとり,形容詞と名詞間の係り受け関係と,名詞Aと名詞Bとの間の係り受け関係が存在する.ここで,形容詞は名詞Aに係る場合と名詞Bに係る場合の2つの曖昧性が存在する.以下に例を示す.

- 広い 門 の 下 (広い → 門)
- 長い 北国 の 夜 (長い → 夜)

例では,「広い」は「門」に係り,「長い」は「夜」に係っている.本稿では前者を「A係り」,後者を「B係り」と呼ぶ.

2.2 形容詞+並列型名詞句

一方,「形容詞+並列型名詞句」は「形容詞+名詞A+並列助詞+名詞B」の形をとり,名詞Aと名詞Bが並列に連なっている.以下に例を示す.

- 珍しい 犬 と 猫 (珍しい → 犬, 猫)
- 古い 記憶 と イマジネーション (古い → 記憶)

例では,「犬」と「猫」,「記憶」と「イマジネーション」には係り受け関係はなく,並列に連なっている.並列助詞には,他にも,「と」「や」「か」「とか」「やら」「だの」があるが,「とか」「やら」「だの」については「や」と完全に置き換えが可能である.本稿では比較的多くの抽出が可能であった「と」「や」を対象とする.また,本稿では形容詞が名詞Aに係る場合を「A係り」,名詞Aと名詞Bの両方に係る場合を「A& B係り」と呼ぶ.

3 一般名詞意味属性体系

本研究では頻度統計におけるスパース性解消や,名詞の抽象度の判定に一般名詞意味属性体系 [2] を用いる.一般名詞意味属性体系は一般名詞の意味的用法を約2700の意味属性で表し,最大12段の木構造で構成されている (図1参照).

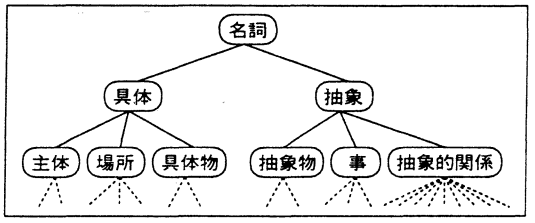

図 1: 一般名詞意味属性体系 (上位 3 段)

4 「形容詞＋の型名詞句」の係り受け解析手法

「形容詞＋の型名詞句」の係り受け解析では, まず字面による解析を行い, 次に意味属性を用いた解析を行う. 以下に具体的な手順を示す.

4.1 字面による解析

字面による解析では, まず「形容詞＋の型名詞句」を「形容詞＋名詞 A」,「形容詞＋名詞 B」とに分ける. 次に, その形容詞に対して共起する名詞をコーパス中から抜き出し頻度統計をとる. そして名詞 A と名詞 B の頻度を比べ, より大きい方を係り先とする. ここで, 名詞 A と名詞 B の頻度が 0 の場合, あるいは両者の頻度が同じである場合は次の意味属性による解析で係り先を決定する.

4.2 意味属性による解析

意味属性による解析では以下の手順で係り先を決定する.

1. 名詞句 (形容詞＋名詞) の名詞を意味属性に置き換え, その頻度統計をとる
2. 「形容詞＋の型名詞句」の名詞を意味属性に置き換える
3. 手順 2 で得られた結果を「形容詞＋意味属性 A」,「形容詞＋意味属性 B」に分ける
4. 手順 1 の頻度統計より, 意味属性 A と意味属性 B の頻度を比べ, 頻度の大きい方を係り先とする

ここで, 頻度統計がスパースになることを考慮して, 以下の通りに頻度を計算する. 意味属性体系上の意味属性 A, 意味属性 B の共通の親ノード (深さ d) を基準にして, その子ノード (深さ d＋n) のうち, 意味属性 A と意味属性 B を配下に属する意味属性ノードの共起頻度の和をそれぞれ意味属性 A と意味属性 B の頻度とする (図 2 参照).

さらに, 形容詞は直後の語に係る割合が多いことを考慮して, 形容詞の直近の語に優先度を与える. 具体的には優先度を重み $t(\geq 1)$ で表し, 意味属性 B の頻度が意味属性 A の頻度の t 倍以上ある場合のみ係り先を B とし, 正解率が最大になるように t を定める.

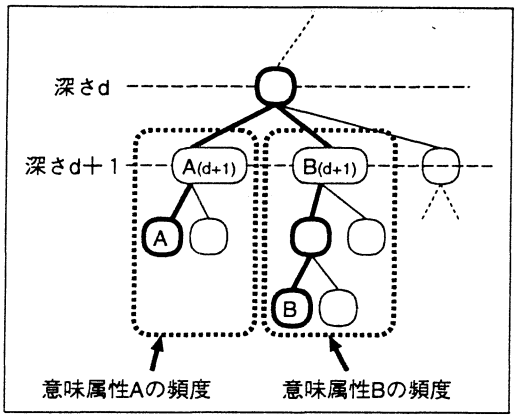

図 2: 頻度の求め方

5 「形容詞＋並列型名詞句」の係り受け解析手法

の型名詞句の場合とは異り, 並列型名詞句の場合, 形容詞は名詞 A に必ず係っている. このことから, 先の頻度統計を用いた解析では名詞 A と必ず共起するので, 正しい係り先が決定できないと思われる. 本研究では「形容詞＋並列型名詞句」の係り先を決定する手法として, 抽象度の概念を導入する.

1. 弱い 者 や ボランティア (弱い → 者)
2. 危ない 家具 や 品物 (危ない → 家具, 品物)

上の例では「弱い」が「者」に,「危ない」が「家具」と「品物」に係っている. 2 つの例文から名詞 A と名詞 B を比べると,「者」は「ボランティア」に比べ, 抽象度が高く,「家具」は「品物」よりも抽象度が低いと考えられる. これらのことから, 形容詞は並列型名詞句に対して名詞 A の方が抽象度が高い場合は A 係り, 抽象度が同じか名詞 B の方が抽象度が高い場合には A&B 係りになると考えられる.

そこで, 意味属性が木構造であることに着目しノードの深さを抽象度とすることで係り先を決定する手法を提案する. まず名詞 A と名詞 B をそれぞれ意味属性 A, 意味属性 B に置き換える. そして意味属性体系よりそれぞれの意味属性の深さを比べその逆数を抽象度とする. 名詞 A の抽象度の方が高い場合は A 係り, 名詞 B の抽象度の方が高いか, 同じ場合は A&B 係りとする.

5.1 最適木の構築

抽象度による解析では, 意味属性の木構造が重要な要因となっている. 本研究では形容詞の係り先を決定するための最適な木構造を汎化を用いて構築する手法を提案する. ここで, 本稿での意味属性の汎化とは,

あるノードを削除しその親ノードと子ノードを繋ぐことを示す (図 3 参照).

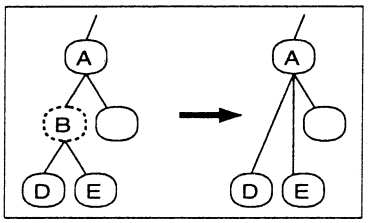

図 3: 汎化の例

最適木の構築では, まずツリー上の全てのノードの貢献度を求める. 貢献度はあるノードが汎化された場合の正解率と現在の正解率との差で表される (式 1 参照). なお, 式中の X は意味属性である.

$$\text{貢献度}(X) = \text{現在の正解率} - X\text{を汎化した場合の正解率} \quad (1)$$

次に, 貢献度が最小のノードを汎化する. そして一つ汎化するごとにツリー上の全てのノードの貢献度を新たに求め, さらに汎化する. このことを繰り返し, 貢献度がすべて 0 より大きい値になるところで汎化を終了する (図 4 参照). ここで, 図中の数字は貢献度を表す.

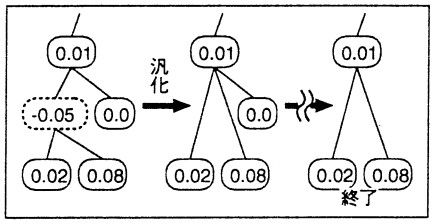

図 4: 最適木の構築

6 実験

4 章と 5 章で述べた手法の精度を評価するために実験を行った.

6.1 「形容詞+の型名詞句」

「形容詞+の型名詞句」ではテストデータとして新潮文庫の小説 100 冊 (約 60 万文) から抜き出した 3541 件のデータに対して人手で正解を付与したものを用いた. しかし, 今回の手法では形容詞毎に頻度統計をとる必要があり, 全ての形容詞を対象とすることはテストデータの量や計算量の面から困難である. そこで比較的多くのデータが抽出できた形容詞「長い」(101 件) のみを使用する. また, 頻度統計用のデータとして, 同じく新潮文庫の小説 100 冊からテストデータを除いた「長い+名詞」(1619 件) を用いた. 結果を表 1 に示す. 表中の ALL A 方式は係り先を全て A 係りとした場合である. 比較のため, 字面のみの解析と意味属性のみによる解析を行った. ここで, 字面のみによる解析では, 頻度による係り先が判定できない場合を全て A 係りとした.

表 1: 「形容詞+の型名詞句」の係り先判定精度

	正解率 (正解数/被適用数)
字面による解析	90% (40/44)
意味属性による解析	70% (40/57)
合計	79% (80/101)
字面のみによる解析	77% (78/101)
意味属性のみによる解析	75% (76/101)
ALL A 方式	73% (74/101)

また, このときの優先度は 7.3 であった. 優先度 t に対する正解率の変化を図 5 に示す. 図より優先度が 7.3～18.0 のとき最も正解率が高かった.

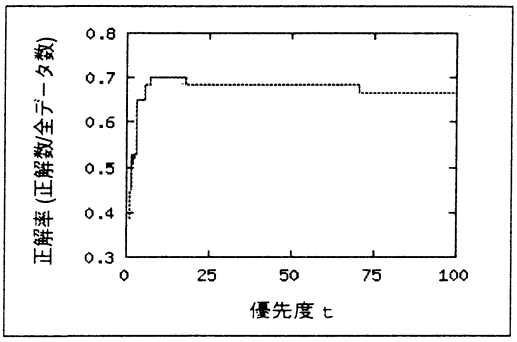

図 5: 優先度と正解率の関係

6.2 「形容詞+並列型名詞句」

「形容詞+並列型名詞句」ではテストデータとして新聞 1 年分 (約 160 万文) から抜き出した「形容詞+名詞+と/や+名詞」(288 件) に対して, 人手で正解を付与したものを用いた. 結果を表 2 に示す. 表中の ALL A&B 方式は係り先をすべて A&B 係りとした場合である.

表 2: 「形容詞+並列型名詞句」の係り先判定精度

	意味属性数	正解率 (正解数/被適用数)
汎化しない場合	2716	67% (192/288)
汎化した場合	94	92% (265/288)
All A&B 方式	-	86% (248/288)

7 考察

7.1 解析精度について

の型名詞句に対する形容詞の係り先解析では字面のみでは77%, 意味属性のみでは75%であるが, これらを段階的に適用することで79%の精度が得られた. 字面のみの解析で精度が比較的良いのは, 同じ小説内のデータで頻度統計を行ったためだと思われる. また, 段階的に適用することで精度が向上した原因としては, 意味属性による解析では失敗する対象を字面解析の段階で取り除いたためと考えている.

並列型名詞句では, 汎化しない場合と汎化した場合の精度の差が顕著に現れた. 本稿での汎化の手法はある解析に対して大規模なシソーラスから最適なシソーラスを構築する際に有効ではないかと考えている. ただ, 一つのノードを汎化する度に全てのノードの貢献度を求める必要があるため計算量が多くなっている. 今後は効率の良い汎化アルゴリズムを検討していきたい.

本稿の実験によって意味属性が係り受け解析にある程度有効であることが分かった. 今後は例外処理を加えるなどして精度の向上を目指したい.

7.2 汎化の効果について

並列型名詞句での汎化後の木構造を観察すると, ルートから「具体」側のノードが32個,「抽象」側のノードが61個であった[1]. これは, 形容詞と「抽象」の意味をもつ名詞との関係が深いことを表していると考えられる.

7.3 形容詞の観点について

本稿での「の型名詞句」の係り先決定法では形容詞を字面のまま扱うため, 計算量が膨大になるという欠点がある. この点については形容詞を観点によってグループ化することで解決できると思われる.

形容詞は量的観を持つものと質的観点を持つものとに大別できると考えられる. 量的観点を持つ形容詞とは, 主観を程度によって表現できる形容詞である. 量的観点を持つ形容詞はさらに以下のように分類できる.

量的観点を持つ形容詞

- 空間的量の大小 ● 美醜 ● におい
- 速度 ● 新旧 ● 強弱 ● 温度
- 難易 ● 味 ● 音 ...など

また, 質的観点を持つ形容詞とは, 主観を程度による相対的な表現ではなく, 絶対的な主観によって表現する観点である. 質的観点はさらに, 話し手の主観を反映する場合と, 名詞の概念に付随する場合とに分類できる.

質的観点を持つ形容詞

- 話し手の主観を反映する質的観点
- 色に関する質的観点 ...など

並列型名詞句での解析手法では抽象度の概念を導入し, 形容詞が "2つの名詞の抽象度が同程度になるよう" に働いていると仮定した. このことは, 形容詞の観点を考えるうえでは,"ある観点における名詞の意味距離が同程度になる" と考えることができる.

[1] 意味属性体系ではルートから「具体」と「抽象」の2つの意味に分かれる

8 あとがき

本稿では「形容詞＋名詞＋助詞＋名詞」を対象に意味属性を用いた係り受け解析を行い, その有効性を調査した. まず, 助詞によって「形容詞＋の型名詞句」と「形容詞＋並列型名詞句」とに分けた.「形容詞＋の型名詞句」の係り受け解析では,「形容詞＋名詞」の結合力から係り先を決定した. このとき, 字面での解析と意味属性による解析を段階的に行うことで精度が向上することが分かった. 実験の結果, 正解率は79%であった. 次に,「形容詞＋並列型名詞句」では, 抽象度から係り先を決定した. このとき, 係り受け解析に必要な意味属性だけを選択し, シソーラスを再構築することで, 精度の向上と意味属性の圧縮を行った. 実験の結果, 意味属性数が2716から94に減少し, 正解率は92%であった. 以上より意味属性が形容詞を含む名詞句の係り受け解析にある程度有効であることが分かった. 今後は文法的情報を用いたり, 例外処理を加えるなどしていきたい.

参考文献

[1] 佐々木 美樹, 坂本 仁: "文書一括処理による係り受け関係の解析", 言語処理学会第1回年次大会,pp.101-104(1995)

[2] 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦: "日本語語彙体系", 岩波書店 (1997)

[3] Satoru Ikehara, Shinnji Nakai, Jin'ichi Murakami: "Automatic Generation of Semantic Dependency Rules for Japanese Noun Phrase with Particles "no" ",TMI 99,pp.55-65(1999)

[4] 森内 昭雄, 中井 慎司, 池原 悟, 大西 真理子: "「の」型名詞句に対する形容詞の係り先解析", 情報処理学会第57回全国大会,4R-3(1998)

[5] 国立国語研究所: "分類語彙表", 大日本図書 (1964)