「A の B」型名詞句に対する連体修飾節の係り先の決定

安井 敏　徳久 雅人　村上 仁一　池原 悟
鳥取大学 工学部 知能情報工学科
{yasui,tokuhisa,murakami,ikehara}@ike.tottori-u.ac.jp

1　はじめに

「A の B」型名詞句に先行する修飾表現の係り先の決定には，幾つかの方法が提案されている．[1] は，「N の A の B」型名詞句を対象として，既存の解析手法と，意味属性を素性とした解析手法を，決定木を用いて判定した．[2] は，「AJ ＋ A の B」型名詞句を対象として，形容詞と名詞の意味属性の組み合せ頻度を用いて判定した．しかし，動詞で構成された連体修飾節についての取り組みはない．

本稿では，「連体修飾節 V ＋名詞 A の名詞 B」型名詞句に対する連体修飾節の係り先の判定を行うことを目的とする．

2　連体修飾節の係り先の検討

2.1　格関係の有無による係り先

「内の関係」や「補足語修飾節」として知られるように，連体修飾節の動詞と被修飾名詞の間には格関係がある [3, 4].

例 1:　働く既婚女性の有効サンプル　　（A 係り）

連体修飾節は「働く」，名詞 A は「既婚女性」，名詞 B は「有効サンプル」である．「既婚女性が働く」は自然であり「有効サンプル { が，を，に } 働く」は不自然である．ゆえに，前者に格関係があり，「A 係り」と判断できる．

2.2　格関係の有無によらない係り先

「外の関係」や「内容節」として知られているように，連体修飾節の動詞と被修飾名詞の間の格関係が決定の材料にならない場合がある [3, 4].

本稿では，名詞 A および名詞 B における表現方法，品詞，および，意味情報に着目した決定方法を考察する．

(1)　特殊記号の有無による係り先

括弧等の特殊記号のある名詞が係り先になりやすい．

例 2:　かぎを握る「公明」の意向　　（A 係り）

(2)　名詞の品詞情報による係り先

自立性が低い「形式名詞」は修飾節なしでは，使えないと言われている [4]．このことより，形式名詞は直前の修飾表現が結び付きやすいと考えられる．

例 3:　恵美子さんに計画を打ち明ける時の様子

（A 係り）

形式名詞「時」の直前に連体修飾節があるので両者が結び付き，「A 係り」となる．

例 4:　野球の面白さを愛した人々のこと（A 係り）

形式名詞「こと」の直前に「人々の」という修飾表現があるので両者が結び付くと，連体修飾節は「人々」に結び付くほうが解釈しやすい．

関連して「時詞」や「数詞」を分析する．数詞や時詞は，他を修飾する性質が強いと予想され，連体修飾節はもう一方の名詞に係りうる．例を示す．

例 5:　「カップ獲得」を目指す今回のチーム

（B 係り）

例 6:　前後に取り付けた二台のファン　（B 係り）

ただし，「時詞」にも自立性の低いものがある．

例 7:　北朝鮮政府内でかなり検討を重ねた末の戦術

（A 係り）

また，「数詞」に係りやすい動詞がある．

例 8:　諮問委員を含む十五人の新判事　（A 係り）

(3)　名詞の意味情報による係り先

「発言・思考」に関わる名詞，「感覚」に関わる名詞，「事実，例，状況など」一群の名詞は，「内容節」としての連体修飾節をとることができると言われている [4].

被修飾側が「A の B」型である場合に同様に成り立つかどうかは不明であるが，少なくとも「意味情報」を頼りに解析を試みる価値はある．

例 9: 「侵略戦争」をめぐる政治家の発言（B 係り）

「「侵略戦争」をめぐる発言」は，「侵略戦争」という話に関連する発言ということが想像ができる．「「侵略戦争」をめぐる政治家」は，想像できなくはないが，相対的に弱い．したがって，「意味情報」を頼りに「B 係り」と判断できる．

3 個別解析方式の提案

3.1 個別解析方式の概要

前章の検討に基づき個別の解析方式を提案する．

- VPM(Valency Pattern Matching) 方式:結合価文法を用いた格関係解析による係り先の決定方式
- VCC(Verb and Case element Co-occurrence) 方式:動詞と格要素の共起頻度に基づく係り先の決定方式
- IPS(Inclusion by Particular Symbol) 方式:特殊記号の存在に基づく係り先の決定方式
- CBS(Co-occurrence Between two part of Speech) 方式:名詞の品詞と品詞の組み合わせに基づく係り先の決定方式
- AAC(Attribute-Attribute Co-occurrence) 方式:名詞の意味情報の組み合わせに基づく係り先の決定方式
- EPA(Exclusion by Particular semantic Attribute) 方式:名詞の意味情報の存在に基づく非係り先の決定方式

VPM 方式で用いる結合価文法は，[5] に収録されている．VCC 方式で用いる格関係の共起頻度は，新聞記事などから収集可能である．IPS 方式の特殊記号は，トレーニングデータ（後述）で出現した（「」，『』，“ ”）とする．

3.2 解析知識の獲得

解析方式に必要な規則を作成する．本稿では，規則をトレーニングデータから作成する．トレーニングデータは，係り先の情報を付与した京大コーパス [6] から「V ＋ A の B」型名詞句のデータを 1,000 件抽出し，作成する．

CBS 規則

トレーニングデータを詳細に分析したところ，次の 3 つの規則が作成できた．

- **規則 1:**もし形式名詞ならば「A 係り」とする．
- **規則 2:**もし次の単語以外の時詞ならば係り先にならない（以来，後，末，瞬間）．
- **規則 3:**もし連体修飾節の動詞が次の単語以外ならば数詞は係り先にならない（含む，除く，次ぐ，当たる）．

3 つの規則から，解析を行う．名詞 A と名詞 B の組み合わせの都合により規則の優先順位が必要となる．トレーニングデータによるテスト結果に基づき，規則 1，規則 3，規則 2 の順位をつけた．

AAC 規則

AAC 規則をトレーニングデータから自動抽出するために，意味属性と係り先の頻度情報を $< S_a : C_a, S_b : C_b >$ という形式で蓄積する．S_x は名詞 x の意味属性コード，C_x は x 係りの頻度である．次の手順で獲得する．

1. トレーニングデータの一つずつについて，名詞 A と B の意味属性コードの組を抽出し，係り先の頻度をインクリメントする．
2. 蓄積したデータから，意味属性の上下関係を確認して，下位属性の組が存在するならばその上位属性の組みを削除する．
3. 係り先が一意，かつ，係り先頻度 C_x が γ 以上のデータを残す．

予備実験において，閾値 γ を決定するために，$1 \leq \gamma \leq 6$ における AAC の正解率を調査した．調査結果により，正解率が高く，かつ，規則数を多いものとし，γ を 1 と設定する．その結果，3,609 個の AAC 規則を作成した．AAC の規則の一例を表 1 に示し，適合例を示す．

例 10: 市場活性化対策を論議する研究会の設置

（A 係り）

表 1　AAC 規則の例

A 係り	（【軍人】，【姿】，6）
	（【会】，【設置】,5）
	（【兵卒】，【姿】,4）
B 係り	（【行政区画】，【機関】，4）
	（【行政機関】，【敬称】，3）
	（【接辞 (人間／単数)】，【成員】，2）

※括弧内は，（名詞 A 意味属性，名詞 B 意味属性，係り先頻度）

EPA 規則

EPA 規則をトレーニングデータから自動獲得するために，意味属性と係り先の頻度情報を $< S_x : C_x >$ という形式で蓄積する．S_x は係り先となる名詞と別の名詞 x

の意味属性コード，C_x は x 係りとならない頻度である．次の手順で獲得する．

1. トレーニングデータの一つずつについて，係り先となる名詞と別の名詞の意味属性コードを抽出し，係り先とならない頻度をインクリメントする．
2. 蓄積データから，意味属性の上下関係を確認して，下位属性の組が存在するならばその上位属性の組みを削除する．
3. 非係り先が一意かつ非係り先頻度が α 以上のデータを残す．
4. 残ったデータに対して，トレーニングデータ 1,000 件を使用して，一データずつ EPA 実験を行い，正解率 β % 以上を最終的なデータとする

閾値 α と β を決定するために，$1 \leq \alpha \leq 6$，$60\% \leq \beta \leq 100\%$ における EPA を用いた予備的な解析実験を行い調査した．調査結果により，正解率が高く，かつ，規則数を多いものとし，α，β を 1，100% と設定する．その結果，252 個の EPA 規則を作成した．規則の一例を表 2 に示し，適合例を示す．

例 11: 関根容疑者が実質的に経営するペット会社の役員

（A 係り）

表 2　EPA 規則の例

意味属性	頻度
【発明】	6
【移動】	5
【役員】	4
【腕】	3

4　決定リストを利用した個別解析方式の統合

6 つの解析方式を決定リスト [7] で統合する．まず，京大コーパスから作成したチューニングデータ 500 件から信頼度を求める．信頼度は，VPM 方式が 62.6%，VCC 方式が 56.8%，IPS 方式が 69.0%，CBS 方式が 78.7%，EPA が 65.5%，AAC 方式が 76.3% であった．ゆえに，決定リストは，CBS，AAC，IPS，EPA，VPM，VCC の順とする．最終的に，どの規則にも当てはまらない場合は，デフォルトの判定規則として「A 係り」と出力する．

5　実験

統合した解析手法の精度を評価する．比較のために，デフォルト規則のみによる判定結果を示す．評価は，クローズドテストおよびオープンテストである．オープンテストは，京大コーパスから作成したテストデータ 500 件を使用する．実験結果の正解率を表 3 にまとめる．

表 3　実験結果

判定手法	クローズドテスト	オープンテスト
提案手法	95.9% (959/1000)	70.2% (351/500)
デフォルト規則	75.4% (754/1000)	77.6% (388/500)

6　考察

6.1　規則削減による統合方式

前述の定義では，デフォルト規則より信頼度が劣る個別解析方式も統合したが，そのような方式は，採用しない方法も考えられる．

そこで，チューニングテストにおいて信頼度がデフォルトを上回る方式のみで決定リストを作成する．CBS,AAC 方式の順で解析を行う．実験を行った結果，正解率は，76.6% 向上した（表 4）．

表 4　CBS，AAC 統合方式の解析実験結果

判定手法	クローズドテスト	オープンテスト
提案手法	95.4% (954/1000)	76.6% (383/500)
デフォルト規則	75.4% (754/1000)	77.6% (388/500)

6.2　個別解析方式の誤り分析

6.2.1　CBS 方式誤り例

CBS 方式の解析誤り例を示す．

例 13: 日本航空８７８便ボーイング７４７ジャンボ機が，札幌を飛び立つ前の点検

（A 係り）

時詞「前」が名詞 A にあるため，CBS 規則 3 から B 係りと出力される．しかし，CBS 規則 3 には含まれていないが，「前」は自立性の低い時詞である．今後，時詞で自立性の低い名詞を，列挙する必要性がある．

6.2.2　VPM 方式解析の誤り例

VPM 方式において，名詞 A および名詞 B どちらにも係り先があり，判定不能となる例があった．

例 12: 国連防護軍に所属するチェコ憲兵部隊の
ツェフ大佐

（B 係り *）

結合価文法から「チェコ憲兵部隊が国連防護軍に所属する」は格関係があり，「チェフ大佐が国連防護軍に所属する」も格関係があると判定することができた．つまり，「AB 係り」と出力される．

では，例文を見ると，「国連防護軍に所属するチェコ憲兵部隊」は，「あるチェコ憲兵部隊」が「国連防護軍に所属している」ということが想像できる．「国連防護軍に所属するツェフ大佐」も，「ツェフ大佐」が「国連防護軍に所属している」ということが想像できる．このことから，これは「AB 係り」と見ることができる．

6.2.3 AAC 方式誤り例

AAC 方式の解析誤り例を示す．

例 14: 侵略と犯罪を美化する日本の行為

（A 係り *）

適応規則から，「B 係り」と判定され，解析誤りとなる．しかし，6.2.2 節と同様に，「侵略と犯罪を美化する日本」と「侵略と犯罪を美化する行為」は，「AB 係り」と見ることができる．

6.3 係り先の曖昧性について

本稿では，京大コーパスの係り先を正解として，解析実験を行ってきた．しかし，誤り分析に見られたように「AB 係り」が存在するため，係り先を一方へ決定した京大コーパスだけでの評価実験では不十分であると考えた．本提案手法は，「A 係り」，「B 係り」が明確な名詞句に絞り，解析実験を行う．

まず，京大コーパスからランダムに抽出した 100 件の名詞句に対して，3 人の分析者が，「A 係り」，「B 係り」，「AB 係り」の判定を行う．次に，係り先が明確な名詞句を選出するために，3 者が同一の係り先とした名詞句を選出する．最後に本提案手法による解析の正解率を求める．

100 件の名詞句のうち，係り先の名詞句なものは，55 件であった．「A 係り」が 31 件，「B 係り」が 24 件であった．この 55 件について，分析者が定めた係り先を正解とすることを「正解タイプ 1」とする．同じくこの 55 件について，京大コーパスの係り先を正解とすることを「正解タイプ 2」とする．

解析実験の結果，「正解タイプ 1」をみると，デフォルト規則では 56% となったことより，係り先の判定の難しさが高いことが分かる．しかし，提案手法では，解析精度は「正解タイプ 2」においても同様である．したがって，「A 係り」，「B 係り」の一意に決定することに関しては本提案手法の性能が確認できたと言える．

今後，「AB 係り」について，調査する必要がある．

表 5　明確な係り先事例に対する正解率

判定手法	正解タイプ 1	正解タイプ 2
提案手法	83% (46/55)	90% (50/55)
デフォルト規則	56% (31/55)	67% (37/55)

7 おわりに

本稿では，「V+A の B」型名詞句において，連体節の係り先が名詞 A と B のどちらになるのか，自動的に判定するため，格関係の有無，名詞の表現方法，品詞，および，意味情報を手がかりとする 6 つの解析方式を作成し，それらを決定リストを用いて，係り受け解析方式を構築した．統合方式の精度は，70.2% であった．個別の方式では，CBS 方式が精度が高かった．単語の自立性あるいは修飾のされやすさの観点が効果を高めたと考察する．

参考文献

[1] 美野秀弥，橋本泰一，徳永健伸，田中穂積．日本語の連体修飾関係に関する研究．言語処理学会第 10 回年次大会発表論文集，pp.600-603，2004.

[2] 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，Vol.2，pp.237-238，1998.

[3] 寺村秀夫．日本語シンタクスと意味 I ～III．くろしお出版，1982～1991.

[4] 益岡隆志，田窪行則．基礎日本語文法．くろしお出版，1989.

[5] 池原悟，宮崎正弘，白井諭，横尾明男，中岩浩巳，小倉健太郎，大山芳史，林良彦．日本語語彙大系．岩波書店，1997.

[6] 黒橋禎夫，長尾眞．京都大学テキストコーパス・プロジェクト．言語処理学会第 3 回年次大会発表論文集，pp.115-118，1997.

[7] Ronald L. Rivest. Learning decision lists. Machine Learning，Vol.2，pp.229-246，1987.