A-2-1　　意味解析に基づく並列名詞句の構造解析

Coordinate Structure Analysis based on Semantic Analysis

田村直良，田中穂積　（東京工業大学　工学部　情報工学科）

本論文では、自然言語文の意味解析に重要である並列構造解析について述べる。本方式では、並列助詞「と」と連体助詞「の」、名詞のみから成る名詞句を対象とする。評価原理としては、連体助詞「の」の修飾関係が成り立つ導出木について類似度等を考慮した評価関数を用いて、その値の最大なものを意味構造とするものである。名詞の数４以下の５８例について本方式を適用したところ、９５%の正解率で正しい構造決定が行えた。

１．はじめに

自然言語文において、並列構造は、文の曖昧性を増やす主たる要因であり、文の意味解析、意味抽出には、文が持つ並列構造を正確に把握することが非常に重要な問題となる。並列構造決定の手法として、[6]では、構文的な10個の「手がかり」を用いた名詞句の解析手法を述べている。構文的な範囲に納めた理由として、機会翻訳のような応用では、個々の名詞ごとに詳細な意味記述が困難であることをあげている。しかし、おそらく人間は、主として意味解析から並列構造を決定していることであろう。よって、このためにどの程度の解析、意味記述が必要か、その結果どの程度の判断ができるかについて調べることも興味ある題材である。

自然言語文において、並列される文の要素としては、文、用言、体言その他ほとんどの要素が考えられるが、我々の研究では、 簡単のために、並列助詞「と」と、連体助詞「の」、名詞からなる名詞句のみを対象とし、さらに、「と」の出現の個数は、高々一個と制限する。名詞句のみからその並列構造を決定しようとする場合、その名詞句の前後の修飾関係、文や文脈の意味等を考慮する必要があり、また、根本的に曖昧性のある句もあって、一般にはすべての名詞句についての構造決定は不可能である。本研究での目的は、上記の形式の名詞句のみから、連体助詞「の」が結ぶ名詞の意味関係と、並列される名詞の類似度から、人間が判断できる程度に解析できるような解析手法の確立を目指すことにある。

２．並列構造決定の評価原理

この章では、本方式の解析手法の原理について説明する。

２．１　扱う名詞句の形式と意味構造

まず、処理対象とする名詞句の形式と、この形式から得られる意味構造の一般形式を定義する。我々は、名詞が連体助詞「の」のみにより連結されている名詞句の意味構造を、前の名詞が直後の名詞に係るという一次元的な構造に仮定している。こうすると、(2-1)式に対する意味構造は、図１のような構文木で代表される。(2-2)式は、図１に対応している。

［定義］　並列名詞句の形式と意味構造

a)　評価対象とする並列名詞句は、次の形式である。

A_1の…のA_mとB_1の…のB_n　　(2-1)

ここで、各A_i、B_j($1 \leq k \leq m$, $1 \leq l \leq n$)は名詞、「の」は助詞の「の」を表す。

b)　(2-1)式で表される並列名詞句の意味構造を

A_1の…の（A_kの…のA_mと

B_1の…のB_l）の…のB_n　　(2-2)

のように定義する($1 \leq k \leq m$, $1 \leq l \leq n$)。 ここで、「A_kの…のA_m」と「B_1の…のB_l」が並列する名詞句であり(A_mとB_zが並列する名詞である)、「A_1の…のA_{k-1}」は、並列名詞句「A_kの…のA_mとB_1の…のB_l」に係り、さらにこの並列名詞句は、「B_{l+1}の…のB_n」に係る。

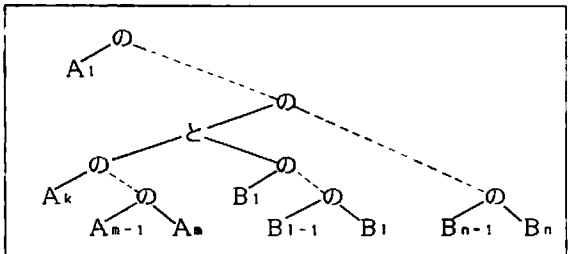

図１　並列名詞句の意味構造

名詞句の並列構造を決定することは、定義１のもとで、k($1 \leq k \leq m$)とl($1 \leq l \leq n$)を求めることに帰着される。

２．２　連体助詞「の」の意味解析

連体助詞「の」による接続を、前の名詞が後の名詞を修飾するもの仮定しているため、名詞句全体では、修飾関係による「意味の流れ」に対応する情報の列ができる。この「流れ」は、導出木の根より出発し、導出木上を左から右へ inorderにtraverseしているが、「と」に対応するノードでは、左右の部分木へ分岐し、このノードで両部分木で得られた情報がマージされて上へ返される。また、導出木の葉の部分、つまり各名詞においては、「の」による修飾関係の解析がなされて、その結果の情報が列に加えられる。この様な処理を形式的に記述するために、属性文法を用いる（図２）。

図２において、ＮＰは名詞句を表す非終端記号、NOUN、「の」、「と」は、名詞、および助詞「の」、「と」に対応する終端記号である。また、非終端記号ＮＰは、相続属性（導出木の根から葉の方向へ流れる属性）ｉと、合成属性（導出木の葉から根の方向へ流れる属性）ｏを

持つ。また、終短記号NOUNは、名詞を構成する文字列に対応する合成属性ｓを持つ。終端記号が表す文字列の領域をＳで、意味の領域をＭで表すと、関数dict、modify、mergeは、それぞれ、以下のような領域で定義される。

dict ： Ｓ → Ｍ
modify： Ｍ × Ｍ → Ｍ
merge ： Ｍ × Ｍ → Ｍ

ここで、dictは、終端記号が表す文字列について辞書を引き、対応する意味を得るための関数である。modifyは、第１引数が表す名詞が、 第２引数が表す名詞句に連体助詞 「の」により修飾されたときの意味を返す関数である。 mergeは、第１引数、第２引数ふたつの意味構造を統合した並列構造を返す関数である。このうち、関数modifyは、両引数が意味がある修飾関係のときのみ定義されているものとする。この制限は、不合理な修飾関係を持つ導出を除外するためにも利用される。

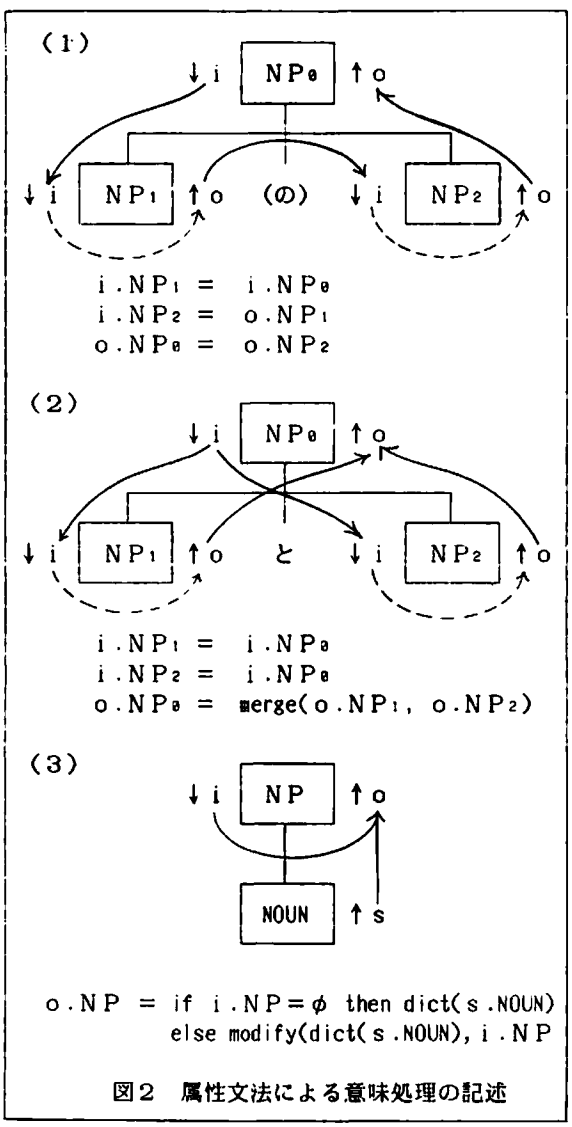

図２ 属性文法による意味処理の記述

２．３ 並列構造評価関数

前節の手法により得られる導出木は、関数modifyにより一応意味のあるもののみ選択される。一般に、一つの名詞句からは複数の導出木が得られるが、これらの中からどれが最も適したものであるかを決定するために、導出木にある値を対応づける並列構造評価関数 evalを導入する。この関数 evalは、図１の形式の導出木に対して、次のような原理に基づいて導出木の意味的な「正しさ」の目安を計算する。

図１において、

(a) 並列する名詞A_nとB_1の類似度が大きいほどのevalの値は大きい。

(b) 並列する名詞A_nとB_1の類似度が大きいときはｋの値が大きいほどevalの値は大きく、類似度が小さいときはｋの値が小さいほどevalの値は大きい。

(a)は、並列する名詞を選択するための手がかりであり(１の決定)、(b)は、並列する名詞の類似度が大きいときは前方にある名詞は両方の並列名詞に係る傾向が強く、類似度が低いときは前方の名詞は「と」の直前の名詞のみに係る傾向が強いとの観察による(ｋの決定)。以上から評価関数evalを次のように定義する。

［定義］ 評価関数 eval

$$eval(t) = k \cdot a_1(A_n, B_1) + a_2(A_n, B_1) \qquad (2\text{-}3)$$

ここで、ｔを図１の形式の導出木とし、A_nとB_1は並列する名詞句である。ｋは並列する名詞句に係る名詞句の長さ、a_1、a_2は類似度を表す関数である。ただし、両関数とも類似度が大きくなるにつれて値が大きくなる単調増加関数であるが、a_2が常に正の値をとるのに対して、a_1では、類似度が小さいときには負の値を取る。

この関数evalでは、「の」による修飾関係については考慮されていない。これについては、第５章で触れることにする。

３．並列構造解析システム

この章では、解析システムの実現について述べる。

３．１ 概要

システムはLangLAB上に構築されている。 評価手順は、１）構文解析、 ２）連体助詞「の」の修飾関係評価、３）類似度および関数evalの評価、と進むが、これらの処理は、入力された単語の列に対して構文解析が非決定的であることから、組合せ的に可能なすべての意味構造を作り出すように行われる。 「の」の修飾関係評価では、２．２で述べたように、 意味情報が導出木をinorderにtraverseする形で流れる。このため、これまでの導出木が下から生成されるボトム・アップ形のパーザでは構文解析と同時に意味情報の評価を行えなかった。我々は、BUP-XGを応用した手法[1]を用いて、構文解析用文法に「の」の修飾関係評価の部分を補強項として記述している。この記述は、同時に不合理な「の」の修飾関係を除外するための制約条件にもなっている。１）～３）の結果得られた意味構造についてevalを評価し、評価値が最大の構造を求める結果として出力する。なお、この方式で必要な名詞の意味記述は、「の」の修飾関係を決定するためと類似度の計算のために必要な、シソーラス上での分

類位置と、「束ね」の名詞であるかどうかの指定だけである。

３．２　「AのB」の解析

「AのB」の解析は、前節で述べたように、解析により得られた意味構造のうち、「AのB」の意味として妥当でないものを除外するために用いられる。ここでの方式は、[3,4]をＤＣＫＲ[5]で実現したものである。概念の検索は、シソーラス[9]をＤＣＫＲで記述したものを用いている。

「AのB」の解析の方法の概略を以下に示す。

case１.
　Aが、「人ー有意志体」、「社会的集団」、「物」、「行為活動」、「抽象概念」、「時」、「場所」で、
　Bが、「行為活動」のとき、
　Aが格要素、Bが述語の名詞形なる構造とする。

case２.
　Bが「時」、「場所」、「目的」などのとき、
　Aに対しBが相対的な場所、時等を示す構造とする。

case３.
　Aが「物」、「人ー有意志体」、「社会的集団」、「抽象概念」で、
　Bが「属性」、「精神的活動源や心情」のとき、
　BがAの数量、性質等を表す構造とする。

case４.
　Aが、「行為活動」、
　Bが、「人ー有意志体」、「社会的集団」、「物」、「行為活動」、「抽象概念」、「時」、「場所」のとき、
　Aが格要素、Bが述語の名詞形なる構造とする。

case５.
　その他の場合（所有関係、連辞関係、方法関係などの関係をＤＣＫＲにより単純化したもの。詳細については省略する）、
　AとBとがある述語の格要素を表す構造とする。

３．３　評価関数 evalの実現

まず、二つの名詞AとBのシソーラス上の距離を定義する。

［定義］　距離
　名詞AとBのシソーラス上の距離を次のように定義する。
　１．同一の名前のとき、０。
　２．シソーラス中の高さｉの部分木にAとBが含まれるとき、ｉ。

我々は、この距離を用いて関数eval中のa_1、a_2を表１のように決定した。

数値の決定は数回の実験から適当なものを選んだものであるが、a_2の値を大きくとることにより並列される名詞の選択を優先させ、次に「の」の修飾構造を決めようとの意図による。

評価関数 evalは、

$$eval(t) = k \cdot a_1(A_m, B_1) + a_2(A_m, B_1) \qquad (2\text{-}3)$$

のように定義されているが、並列構造の直後の名詞がいわゆる「束ね」の名詞[7]であるときは、評価値が高くなるようにして（具体的には、500を加えている）、このような構造がより尤らしいと判断できるようにしている。

距離	a_1	a_2
0	200	3000
1	110	2000
2	40	1500
3	10	500
4	-150	100
5	-200	0

表１　関数a_1とa_2

４．実験

[8]中の６５例の名詞句を用いて実験を行なったところ、evalの評価値について次の５とおりの結果が得られた。

１．正しい意味構造に対して一番大きい値が与えられ、かつ誤った意味構造に対してはそれより小さい値が与えられているもの　……４０個。
　例）「育てるためのくふうと努力」

２．正しい意味構造が２個あり（人間にとっても曖昧な句）両者に同じ評価値が与えられているもの……４個。
　例）「小学生と高校生の三人の子供たち」

３．正しい意味構造に対して一番大きい値が与えられてはいるが、同時に誤った意味構造にも同じ値が与えられているもの　……４個。
　例）「経済無視と国民生活の悪化」

４．「束ね」の構造として正しい意味構造にのみ一番大きな価が与えられたもの　……９個。
　例）「目と目の間」

５．正しい意味構造に一番大きな値が割り当てられていないもの　……８個。
　例）「世界の世論と米国民」

正しい意味構造に一番大きな値を与えているものを正しい構造決定とするなら、上記実験における正解率は、８８％（１～４）である。また、誤った意味構造に最大の値を割り当てたものを不正解とするならば、正解率は、８２％となる（１，２，４）。さらに、名詞の数が４個以下の名詞句５８例のみに制限するならば、正解率は、それぞれ９５％、８７％であった。上記２については、人間にも曖昧性があると感じられる名詞句の意味構造に同じ評価値が与えられており、興味深いところである。

図３に解析結果の一部を示す。

５．問題点および限界

本方式では、かなりの単純化を行っているために、問題点もある。この章では、前章で述べた３、５について考察する。

１．関数a_1とa_2の選び方の違いにより生じる不正解
　例：「若さと政務次官の肩書」
　解：「（若さと政務次官）の肩書」

これは、関数a_1とa_2の選び方により生じる不正解である。a_1とa_2を帰れば一つの不正解は訂正できるが、その他の不正解が新たに生ずる恐れがある。統計的手法

などで最適のa_1とa_2を決定できるだろうが、それだけではこの種の不正解を完全に取り去ることはできない。

２．矛盾する修飾関係を受け入れたことにより生じる不正解

例：「反政府の仏教系と政府支持のカトリック系労組の分裂メーデー」

解：「反政府の（仏教系と政府支持のカトリック系労組）の分裂メーデー」

本方式では、修飾関係の分類までは行っているが、属性スロットのチェックなどは行っていない。これに対処するとすると、名詞の意味記述がシソーラス上の分類関係だけでは収まらなくなってしまう。

３．長い名詞句

例：「企業の中の技術革新の進展と職種の変化」

解：「（企業の中の技術革新の進展と職種の変化）」

本方式では、名詞の数に応じて意味構造の組合せ数は、$o(n^2)$で増える（式2-2においてkとlの選び方）。これに対して、評価関数は単純過ぎる。

４．evalでは、「（AとB）のC」と「（AとBのC）」の区別がつかない

例：「横滑りと視界の良くないこと」

解：「（横滑りと視界）の良くないこと」と
「（横滑りと視界の良くないこと）」

本方式では、「（AとB）のC」と「（AとBのC）」については、「AのC」、「BのC」が同一の修飾関係であるかどうかでしか両名詞句の区別がつかない。「の」の修飾関係をevalに導入することが必要となろう。

６．まとめ

本報告では、意味解析を用いた並列名詞句の構造決定の手法について述べた。本方式は、「の」の修飾関係が成り立っている意味構造に対して、並列する名詞の類似度等を基にした評価関数evalを用いて意味構造に正しさの順位をつけるものである。

本方式により６５例の名詞句について構造決定を行ったところ、正解に対して一番大きな評価値が与えられなかったものが８例（１２％）であった。また、名詞句中の名詞の数を４以下の５８例に限れば、３例（５％）であった。

［参考文献］

１．奥村，田村，徳永，田中：ＢＵＰ系自然言語解析システム上でのトップ・ダウンな情報の制御について，本大会予稿集，1986.

２．首藤，吉村，津田：日本語技術文における並列構造，情報処理学会論文誌，Vol.27，No.2，1986.

３．島津，内藤，野村：日本語文意味構造の分類－名詞句構造を中心に，自然言語処理研究会，47-4，1985.

４．島津，内藤，野村：助詞「の」が結ぶ名詞の意味関係のsubcategorization，自然言語処理，53-1，1986.

５．田中，小山，奥村：知識表現言語ＤＣＫＲとその応用，コンピュータ・ソフトウェア，Vol.3，４号，1986.

６．長尾，辻井，田中，石川：科学技術論文における並列句とその解析，自然言語処理研究会，36-4，1983.

７．水谷，田中：語の並列結合子，計量国語学，No.63，1972.

８．電総研：新編日本語品詞列集成，1979.

９．内田裕士，荻野孝野：ICOT Technical Memo，　to appear.


```
<<意味構造>>

No. 1
並列構造
1540, 0+の, [くふう#3, individual_of(くふう)], [努力#4, individual_of(努力)]

意味
[くふう#3, (case5@d:[育てるため#2, individual_of(育てるため)], individual_of(くふう))],
[努力#4, (case5@d:[育てるため#2, individual_of(育てるため)], individual_of(努力))]

解析木
 |-名詞句
    |-名詞句
    |   |-名詞
    |       |-名詞 -- 育てるため
    |-助詞 -- の
    |-名詞句
        |-名詞句
        |   |-名詞
        |       |-名詞 -- くふう
        |-助詞 -- と
        |-名詞句
            |-名詞
                |-名詞 -- 努力
```

図３　出力例