日本語文における名詞並列の構造解析

2C－2

木下知貴　飯島正　関洋平　原田賢一

慶應義塾大学理工学部

e-mail: kazuki@hara.cs.keio.ac.jp

1　はじめに

自然言語処理における困難な問題の一つに，文中に含まれる並列構造の範囲を決定することがある。一般に並列構造を含む文は，その係り先の候補が複数あることが多いため，誤って解析されやすい。並列構造を正しく認識すれば，文の構造をより簡単にでき，構文解析での誤った解析を減らすことができる。

2　日本語文における並列構造

ここでは [5] にならって，一文中に同等の機能をもつ単語列を複数個並べたものを**並列構造**と定義し，並列の単位となっている単語列を**並列要素**と呼ぶ。並列要素には大きく分けると以下のように３種類のものがある。

- 名詞並列 — 並列要素が名詞句であるもの
  (1) 白い壁に赤と青のペンキを塗った。
- 述語並列 — 並列要素に述語が含まれるもの
  (2) 僕は英語も読めるし，フランス語も読める。
- 部分並列 — 並列要素が上記２つ以外のもの
  (3) 私はビールを，彼はジュースを注文した。

大部分の並列構造には，その存在を示す表現 (上の例文中の太字部分) があり，これらを**並列標識**と呼ぶことにする。名詞並列の場合は「と」「や」「, (読点)」等，述語並列の場合は連用中止法や「～し」等がよく用いられる。部分並列には，特徴的な並列標識はほとんどないが，(3) のように付属語の一致などが見られることもある。また，並列要素のうち，並列標識の前にあるものを**前部要素**，後ろにあるものを**後部要素**と呼ぶ。

3　関連研究とその問題点

ここでは，並列構造に関する研究を整理して述べる。

1. 形態的・構文的類似性の利用
   品詞の一致や接尾辞の一致など，意味に立ち入らない簡単な処理を行う [4]。
2. 意味的類似性の利用
   階層的な意味分類より，意味的類似度の高いペアを並列要素とするもの [3] などがある。

日本語の並列構造を解析する手法は，意味的類似性を利用するものが主流となりつつある。しかし，

(4) お見舞に花束やみかんなどの果物を持って行った。

のような例を考えた場合，意味的類似性を利用する方法では，「花束」—「みかん」の類似度と，「花束」—「果物」の類似度とを比較するわけだが，どちらの方が類似度が大きいと定めるのは難しい。また，

(5) 太郎と太郎の弟が遊んでいるのを見かけたよ。

のような例において，類似度を用いると，「太郎」—「太郎」の類似度の方が「太郎」—「弟」の類似度よりも大きいため，誤って認識されてしまう。

このように類似性によるアプローチでは解析できない並列構造があるため，解析精度に上限ができてしまう。

4　本研究でのアプローチ — 接続チェック

前節で述べた並列要素間の類似性よりも本質的な特徴として，**各並列要素は前後の文脈と矛盾なく接続する**ことがある。

そこで，新たなアプローチとして，

> **並列要素になり得る候補を求め，それらに対して前後の文脈と接続可能かどうかをチェックすることにより，並列構造の範囲を定める**

ことが考えられる。

なお，ここでは対象を次のように限定する。

- 名詞並列のみを扱う。
- 一文中に複数の名詞並列標識が現れる文を除く。

また，入力と出力はそれぞれ以下のように仮定する。

入力：正しく解析された形態素列

出力：並列要素をまとめた後の形態素列

An Approach for Coordinate-Structural Analysis of Japanese Noun Phrases.

Kazuki KINOSHITA, Tadashi IIJIMA, Yohei SEKI and Ken'ichi HARADA

Faculty of Science and Technology, Keio University

3-14-1 Hiyoshi, Kouhoku-ku, Yokohama 223, Japan

解析手順

ここでは, 並列要素の範囲を推定するための解析手順について, 例文 (4) を用いて説明する。

Step1. 文の前方から**並列標識**を検出する。

お見舞に花束やみかんなどの果物を持って行った。

Step2. **標識の直前の単語**を前部要素の末尾 (W1) とし, その語彙情報を保持する。

W1
お見舞に**花束**やみかんなどの果物を持って行った。
普通名詞

Step3. 標識の後方から読点や文末に到達するまでの間にある W1 と**同じ品詞の単語** W2(複数) を探し, 各 W2 に対し **Step4.** を繰り返す。この例では、 W2 として「みかん」と「果物」が見つかる。

W1 W2 W2
お見舞に**花束**や**みかん**などの**果物**を持って行った。
普通名詞 普通名詞 普通名詞

Step4. 並列標識から W2 までの部分を W1 で置き換えて**接続チェック**を行い, 接続可能であれば並列標識から W2 までの単語列を後部要素として記録する。

1. 「みかん」を W2 とした場合

接続チェック『A などの B』

花束 などの 果物を持って行った。

A に現れる語は B の例示であり, 概念階層で言えば, B が A の直接の上位概念となる。

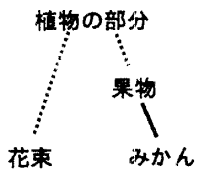

図 1: 概念階層

A(花束) と B(果物) の EDR 概念辞書 [1] による概念階層は図 1 のようになっており, 果物は花束の直接の上位概念とはならず, 接続不可能となる。

2. 「果物」を W2 とした場合

接続チェック『A を B(する)』

花束 を 持って行った。

動詞 B の目的語として A が適切であるかを調べる。このとき, 動詞の格フレーム情報を利用する。

EDR 日本語動詞共起パターン副辞書 [1] によれば, 「持って行く」の格フレームは以下のようになっている。

(agent) が (object) を (goal) へ (に) 持って行く。

ただし, object の部分に入ることができるのは以下のものを表す名詞句のみである。

object : 動物, 植物, **植物の部分**, 金属, 人工物, etc.

「花束」の上位概念である「植物の部分」が object : に含まれているので, 「花束」は「持って行く」の目的語として適切であり, 接続可能であると言える。したがって, **「みかんなどの果物」という部分を後部要素として記録する。**

Step5. 前部要素および後部要素に対して前部要素の直前の句が**接続可能か調べる。** 前部のみに接続可能であれば, その句を前部要素に含めて再び **Step5.** を行う。そうでなければ, 現在の前部要素を記録して **Step6.** へ。

接続チェック

お見舞に 花束 や みかんなどの果物 を持って行った。
前部要素 後部要素

「お見舞に」 は用言に係るので両要素への接続は不可能。
⟹ 「お見舞に」は前部要素には含まれない。

Step6. ここまでの処理で得られた前部要素と後部要素のペアを作る。複数のペアが得られた場合は、ヒューリスティックスにより優先度をつけて結果を出力する。

5 おわりに

本研究の提案である接続チェックを用いた並列構造解析システムを主に Prolog により実装を行った。接続チェックを行う際に必要な情報は EDR の辞書 [1] および IPAL 動詞辞書 [2] から抽出している。

参考文献

[1] 日本電子化辞書研究所 (EDR) : "EDR 日本語単語辞書, EDR 概念辞書, EDR 日本語動詞共起パターン副辞書," 1995.

[2] 情報処理振興事業協会 (IPA) : "計算機用日本語辞書 IPAL." 1994.

[3] 黒橋 禎夫, 長尾 真 : "長い日本語文における並列構造の推定," 情報処理学会研究報告, 91-NL-86-2, 1991.

[4] 長尾 真 : "画像と言語の認識工学," コロナ社, 1989.

[5] 首藤 公昭, 吉村 賢治, 津田 健蔵 : "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.