並列構造を含む日本語文の構文解析と単文分解

北川 創一[1]　池本 浩次[1]　田辺 利文[2]　吉村 賢治[2]
[1]福岡大学大学院 工学研究科　[2]福岡大学 工学部
{td212004, td212001}@cis.fukuoka-u.ac.jp
{tanabe, yoshimura}@fukuoka-u.ac.jp

概要

本研究では，並列構造を含む日本語文の構文解析における並列構造の推定と単文分解の手法を提案する．構文解析は，日本語句構造文法(JPSG)に基づく単一化文法を用いてチャート法で行う．並列構造のうち名詞並列と述語並列については CFG の枠組みで処理し, CFG の枠組みで記述できない部分並列に対して特別な処理を行う.従来の手法[1]では入力文が並列構造を含む場合，単文に分解して構文解析を行っていたが，本稿では単文に分解せずに並列構造の情報を保持したまま解析し，解析結果から単文に分解する手法について提案する．

1 はじめに

並列構造は、二つの文を共通部分でまとめて一つの文にしたときに生じる同等の機能を持つ単語列の対である．並列構造を用いることで，複数の文を一つの文として表現し，文の冗長性を無くした簡潔な記述を可能としている．

日本語の構文解析において，並列構造の処理に関する先行研究の手法[2][3]では，文節間の類似度を求め，最も類似した二つの文節を並列構造として推定する. この手法は,様々な並列構造を含む文に対して高い精度が得られているが，「太郎がタンクに水、フィルタに豆を入れる」のように並列構造が不完全な文節を含む場合に対して誤った推定をしてしまう. 本稿では文節ではなく非終端記号列の比較をすることで並列構造を推定する手法を提案する．

２章で述べるように基本的な並列構造は，名詞並列，述語並列と部分並列に分類される．このうち名詞並列と述語並列については JPSG の枠組みで記述可能であるため，本研究では部分並列に対してのみ特別な並列構造処理を行う． 従来の手法[1]では入力文が述語並列や部分並列を含む場合には構文解析が失敗した後で並列構造の推定を行い，入力文を単文に分解してから構文解析をやり直していたため一度作成した構文情報が無駄になっていたが，本稿では単文に分解せずに並列構造の情報を保持したまま解析を継続し，解析結果を単文に分解する手法について提案する．

2 並列構造

並列構造は，次のような三つの特性を持つ.

- 並列構造は同じ形式の部分木の列である並列要素で構成される.
- 並列要素を接続する働きを持つ文字列である並列キーが存在する.
- 文から並列キーと片方の並列要素を取り除いても意味が通る.

また，並列構造には，次のような種類がある.
例文では{と}で囲まれた文字列が並列要素である.

a) 名詞並列：並列要素が名詞句であるもの
ex.『太郎が紅茶に {砂糖} と {ミルク} を入 re ru』

b) 述語並列：並列要素が動詞句であるもの
ex.『太郎が {お菓子を食 be} 、{コーヒーを飲 m } ru』
ex.『{太郎がお菓子を食 be} 、{次郎がコーヒーを飲 m} ru』

c) 部分並列：並列要素が部分木の列であるもの
ex.『太郎が {タンクに水}、{フィルタに豆} を入 re ru』

さらに，上記の並列を組み合わせた並列構造として，次のような種類がある。

d) 入れ子並列：
並列要素の中に並列構造を含むもの
ex.『太郎が{ (コーヒーに砂糖), (紅茶にミルク)を入 re}, {お菓子を食 be} ru』

e) 多重並列：三つ以上の並列要素を含むもの
ex.『太郎が{コーヒーに砂糖}, {紅茶にミルク}, {水に氷} を入 re ru』

本稿では a)~c)を対象とし，特別な並列構造処理を行うのは c)のみである．

3 単一化文法

単一化文法は，CFG の非終端記号を素性構造で表現し，素性構造間の単一化によって言語の文法的制約を表す文法である．代表的な日本語の単一化文法である日本語句構造文法(Japanese Phrase Structure Grammar, JPSG)[4][6][6]に基づいて，その原理と素性構造を拡張する．JPSG に示す一つの書き換え規則と素性構造の単一化に関する原理で構成する．

$$M \rightarrow C\ H \qquad (1)$$

ここで，M,C,H はそれぞれ素性構造で M を親，C，H を子と呼ぶ．特に，H は日本語における右側主要部の規則から主辞(head)と呼び，左側の子 C と区別する．C と H は，補語構造，付加構造，等位構造[1]の関係にあるときに結合し，原理に従ってそれぞれの素性構造から M の素性構造が計算される．

3.1 素性

素性には，表 1 に示すものがある。pos や pform，gr は複数個の中の一つを値とする多値素性であり，また，三つ共に主辞素性である為，head 素性の値である素性構造を構成する要素となる．

表 1　素性名と値

素性名	値	備考
pos	v, n, p, …	品詞
pform	ga, wo, ni, …	格助詞
gr	sbj, obj	述語との関係
sem	任意の形式的表現	意味表現
subcat	素性構造の集合	下位範疇化
adjacent	素性構造	下位範疇化 隣接を要求
adjunct	素性構造	修飾先の主辞

JPSG の代表的な原理に主辞素性の原理(head 素性)と下位範疇化素性の原理(subcat 素性，adjacent 素性)があり，基本的な日本語文の記述に必要な原理として，意味素性の原理(sem 素性)，付加素性の原理(adjunct 素性)がある．

4 構文解析

4.1 チャート法

チャート法は，文脈自由文法による構文解析手法の一つである。チャートは，項(term)と呼ばれるデータ構造〈i, j, C → α・β〉の集合である．この手法による構文解析は，二つの項を結合し，その二つの項よりも長い単語列が導出可能であること意味する項を新たにチャートに追加して，最終的に単語列全体が導出可能であることを意味する項がチャート内にできれば，構文解析が成功したと判断する．

本研究では，ボトムアップで幅優先に木を成長させるアルゴリズムを利用し，完全な部分解析木に対応する項を表す弧(不活性弧)のみを用いてチャートを作成し解析を行う．(図 1)

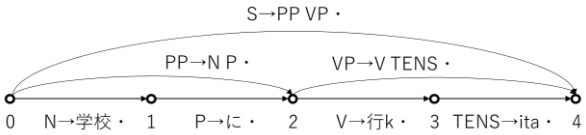

図 1　チャートのグラフ表現（一部）

本研究における先行研究[7]では，構文解析で CYK アルゴリズムを用いて行っていたが，チャート法に比べて，拡張性に乏しい．空所(gap)に対する特別な語彙項目 ε の仮定を行う必要があった場合，CYK アルゴリズムで生成された表を再生成しなければならない．しかし，チャート法は，項を用いることで操作する為，特別な処理に対して容易に拡張できる．

本研究では，この項の基本的なデータ構造を(2)，特別な並列構造処理を行う為の項を(3)とした．

$$\langle i, j, fs, pos, [subcat, adjacent], [adjunct] \rangle \qquad (2)$$

$$\langle i, k, fs, [subcat, adjacent], [adunct], l, j, [rcsi, rcsj], [lcsi, lcsj] \rangle \qquad (3)$$

ここで，i, j, k, l は単語の境界位置の番号であり，i は始点，j は終点，fs は素性構造，pos は品詞である．[subcat, adjacent]はその項が要求する品詞のリスト，[adjunct]はその項が修飾する品詞のリストである．末尾の要素[rcsi, rcsj]，[lcsi, lcsj]は，それぞれ後方と前方の並列要素の始点と終点の対である．

[1] 並列構造の処理により別途行う為，使用しない。

4.2 構文解析の流れ

本稿で述べるシステムでは形態素に分割された入力文に対し構文解析を行う． ただし形態素は学校文法ではなく音韻論的文法に基づいたものである[6]. 4.1 のアルゴリズムは並列構造を含まない単文の構文解析を行うアルゴリズムである．このアルゴリズムで解析に失敗した場合，文中に並列構造があるものとみなして並列構造の解析を行う．構文解析の流れは以下のようになる．

1. 入力文をチャート法で解析する
2. 成功すれば解析終了
3. 解析に失敗した場合，ε が挿入されていなければ ε を挿入して１へ，ε が挿入済みである場合は４へ
4. 既に並列構造の処理を行っていれば解析失敗，そうでなければ並列構造の推定を行い１へ

4.3 並列構造の推定と解析

並列キーは「、」などの単語や記号であることが分かっており個数も限定されるため，並列キーは網羅的に登録できる．並列構造を構成する個々の並列要素は，並列キーの前方と後方にそれぞれ一つずつ存在することが分かっている．

並列構造を構成する個々の並列要素は，共に同じ構造を持つ部分木列で構成されると考えられるため，並列キーを中心とした前後の任意の範囲で，部分木列を調べ，同じ構造を持つ部分木列の組を検出し，並列構造の候補とする．**エラー! 参照元が見つかりません。**の場合，前方後方共に後置詞句，名詞が存在し，共に同じ部分木列で構成されているため，「タンクに水」「フィルタに豆」の組み合わせは並列構造であると推定する．

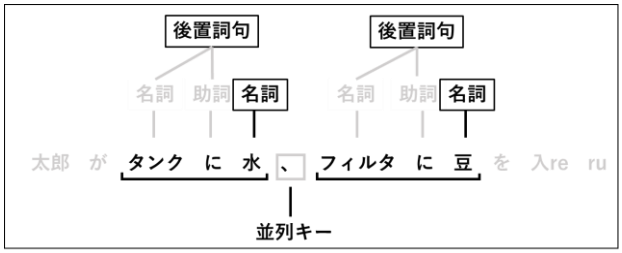

図 2　並列構造の推定の例

先行研究[1]では，並列構造の推定後，単文に分解しチャートを新しく作り直すことで並列構造の構文解析を行っていた．しかし本研究では，推定した並列構造を特別な項(並列項)で表現し，並列構造の解析を行う前までに作成していたチャートを利用する．並列項は並列構造を構成する並列要素中の対応する部分木ごとに作成する．図 2 の場合，「タンクに水」「フィルタに豆」が並列構造であると推定できるため並列項は，先頭の後置詞句である「タンクに」と「フィルタに」を結合したものと末尾の名詞である「水」「豆」を結合した２個を作成し，二つの並列要素の解析を並行して進める(図 3).

構文解析を再開し，並列項と基本の項が単一化される(図 4). 既に作成したチャートと重複した基本の項が存在するため並列項の末尾にある要素([rcsi, rcsj]，[lcsi, lcsj])を用いて推定された並列構造の外側から単一化する条件が必要である．しかし，並列項同士を単一化する場合，条件は必要ない．

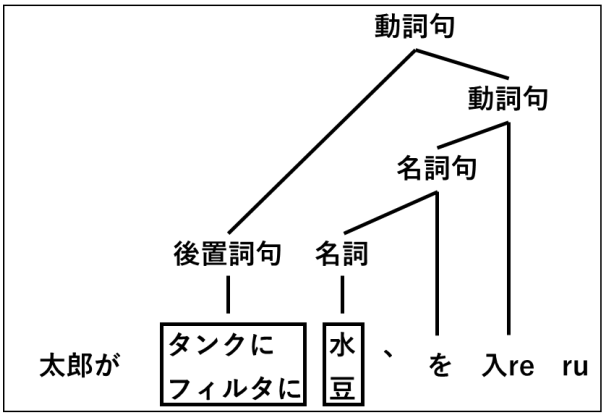

図 3 並列項を作成した後の解析(一部)

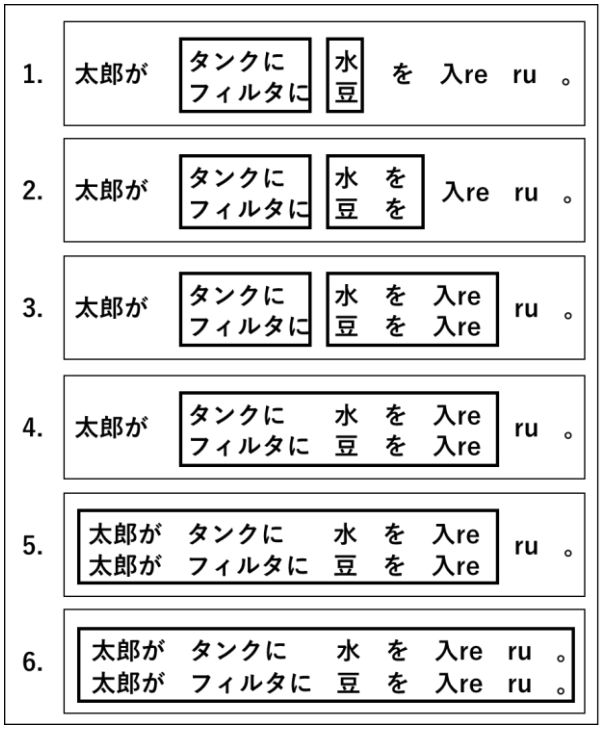

図 4 並列項を作成した後の単一化(一部)

5 述語並列の単文分解

部分並列を含む文に対する単文分解の処理については，図 4 の 6 で処理が完了しているため，ここでは述語並列を含む文の単文分解について述べる．

入力文が述語並列を含む場合，単文分解はチャート法による構文解析が成功した後に，出来上がったチャートにある項の情報を用いて行う．チャートは(2)，(3)式のような項を要素に持つ．述語並列を含む文における並列要素の推定及び単文分解の流れは以下のようになる(図 5).

1. チャートから，述語並列における並列キーに該当する項を取り出す.
2. 1 の項の前後から動詞句の項を取り出す.
3. 取り出した項から，並列要素に該当する動詞句を取り出す(図 6).
4. 入力文から取り出された並列要素と述語並列の並列キーの項を除いた共通要素の項を取り出す.
5. 共通要素及び並列要素の項から，単文を生成する(図 5 下).

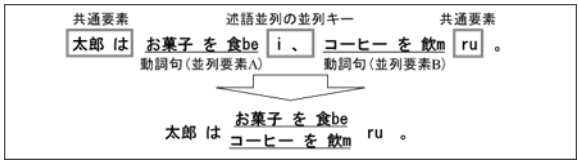

図 5 述語並列の推定及び単文分解

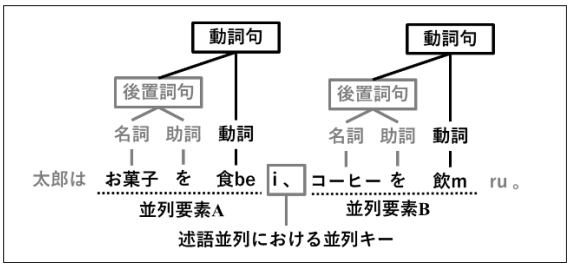

図 6 並列要素となりうる動詞句の例

図 7 上のような例文の場合，並列要素 A が「太郎がお菓子を食 be」，並列要素 B が「コーヒーを飲 m」となるような二つの並列要素が推定される，並列要素 A に主語がついている理由は，「太郎が」の部分木が並列要素 A の部分木に含まれており，共通要素として識別されないからである。その為，並列要素 B を構成する単文は，「誰かが」は「太郎が」を包含すると推定され，「(誰かが) コーヒーを飲 m ru 。」となる。

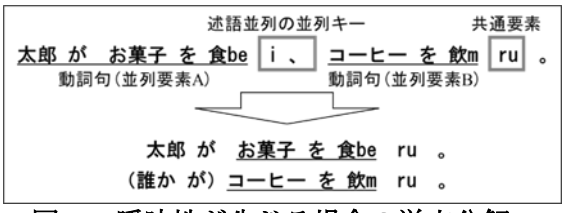

図 7 曖昧性が生じる場合の単文分解

6 おわりに

本稿では，並列構造を含む日本語文の構文解析における並列構造の解析と単文分解の手法を提案した．今後の課題として，現時点ではシステムの検証に必要な語彙項目しか単語辞書に登録していないため，網羅的な解析実験を行うために単語辞書の拡充が必要である．

謝辞

本研究は JSPS 科研費 JP17K00116 の助成を受けたものである．

参考文献

1. “日本語単一化文法における並列構造の解析”. 内野 皓介. 田辺 利文. 乙武 北斗. 吉村 賢治. 火の国情報シンポジウム 2022. 2022.
2. “長い日本語文における並列構造の推定”. 黒橋 禎夫. 長尾　真. 情報処理学会論文誌. 第 33 巻. 第 8 号. 1022-1031. 1994.
3. “並列構造の検出に基づく長い日本語文の解析”. 黒橋　禎夫. 長尾　真. 自然言語処理. 第 1 巻. 第 1 号. 35-37, 1994.
4. 自然言語, 郡司　隆男, 日本評論社, 1994.
5. “HPSG のもとづく日本語文法について―実装に向けての精緻化・拡張”. 大谷 朗. 宮田 高志, 松本 裕治. 自然言語処理. 第 7 巻. 第 5 号. 19-49. 2000
6. “日本語単一化文法による形態素解析と構文解析の融合”. 吉村 賢治. 福岡大学工学集報. 15-21. 2017.
7. “単一化文法を用いた日本語文の構文解析における並列構造の処理”. 中村 健. 乙武 北斗. 田辺 利文. 吉村 賢治. 情報処理学会第 81 回全国大会. 2019.