日本語文における並列構造の簡便な推定法および推敲支援への適用

2J－1

山村広臣　菅沼明　牛島和夫
九州大学工学部

1　はじめに

近年，機械翻訳や文章校正支援に代表される自然言語処理技術の発展は着実である．しかし，いまだ困難な問題はいくつも残されている．その中の一つに並列構造の解析がある．並列構造を含む文は，並列要素の認識において曖昧性を含みやすいので，並列構造を誤って認識してしまうことが多く，後の構文解析や意味解析にも解析誤りが生じる．また，並列構造を含む文を文章推敲の立場から考えても，文の書き手と読み手の間に食い違いが生じやすい．

本研究は，文章推敲の立場から並列構造を推定し，書き手と読み手の間に食い違いが生じやすい並列構造を指摘することを目的としている．本稿では，名詞句の並列構造（**名詞並列**と呼ぶ）の推定と，節の述語を除いた一部分（非名詞句と呼ぶ）の並列構造（**部分的並列**と呼ぶ）の推定について述べる．また，並列構造の推定を推敲支援に適用するために，ユーザに煩わしさを感じさせない待ち時間で処理をしたいという要求がある．そのため，本手法では，大規模な辞書の使用，単語の意味を反映した解析，形態素解析を行なっていない．

2　並列構造の推定

並列構造の存在を示す語を**並列のキー**と呼び，その前後の並列要素を**前置要素**，**後置要素**と呼ぶことにする．本研究では，並列のキーとして「[読点]，[中点]，と，や，かつ，だけで（は）なく，および，または，ならびに，あるいは，もしくは」をとりあげている．

2.1　処理の概要

読み手は，意味的に類似もしくは対比している単語に注目して並列要素を決定することが多い．しかし，このような単語が複数存在したり，存在しなかったりする場合もある．このような場合でも，構造的に類似している並列構造は読み手に正確に伝わると考えられる．また，実際に，並列構造の前後の文節列は構造的に類似していることが多い．そこで本研究では，構造的類似性に基づいて名詞並列と部分的並列を推定する．構造的類似性に基づく推定とは，例えば，「．．AのBとCのD．．」という名詞並列があれば，（AのB）を前置要素，（CのD）を後置要素とする処理を行なうことを意味する．

名詞並列と部分的並列を推定する手順は，以下のとおりである．

1. 字種情報を利用して，文を仮の文節に分割する．
2. 仮の文節に文節の性質を付与し，文節間の係り受け関係を決定する．ただし，文節の性質や文節間の係り受け関係は，並列要素の構造的類似性を比較するための情報である．
3. 並列構造となり得る最長範囲を求め，その範囲内で並列要素を決定する．

1. 2. については，文献[1]を参照してほしい．3. については，部分的並列，名詞並列の順に推定する．

2.2　部分的並列の推定

部分的並列の例として，以下の文がある．

(例) 人口は10倍、情報利用は40倍になるであろう。

本研究では，部分的並列の特徴に着目して並列要素を推定する．部分的並列の前置要素と後置要素は，ほとんどの場合，構造的に完全に一致する．部分的並列は，このように構造的な一致のみで決定できると考えられ，また，名詞並列と違ってどのような文節列も並列要素の候補となり得るので，名詞並列より先に推定する．推定手順を以下に示す．

1) 部分的並列の最長範囲を求める．
2) 部分的並列の前置要素と後置要素の候補を得る．
3) 前置要素と後置要素の候補の間で，構造的に完全に一致する候補があれば，それを並列要素とする．

1) に関して，部分的並列の最長範囲は確実に言葉の切れ目になる文節までとする．そこで，前置要素の最長範囲は，読点もしくは文頭から並列のキーまでとする．後置要素の最長範囲は，並列のキーから読点もしくは句点までとする（図1-(a)参照）．2) に関して，最長範囲内の格要素を含む文節列を候補とする．ただし，格要素を2文節以上含む文節列を対象とする（図1-(b)参照）．3) に関して，文節の性質が「格要素または主題要素」である文節のみをマッチングさせ，候補間の構造が完全に一致する文節列が存在すれば，部分的並列の並列要素とする．図1の例文では，前置要素の候補（前1）と後置要素の候補（後1）の構造が完全に一致する（図1-(c)参照）．

2.3　名詞並列の推定

名詞並列の推定手順は，部分的並列と同様に，1) まず名詞並列の最長範囲を決定し，2) 前置要素と後置要素の候補の中で，主に構造的類似性に基づいて並列要素を決定する．

1) に関して，前処理で得られた文節間の係り受け関係を用いて，名詞句となり得る最長の文節列までを名詞並列の最長範囲とする（図2-(a)参照）．

2) に関して，構造的類似性だけでは，正確な並列要素を推定できない．名詞並列を含む文には，読み手が並列

A Simple Method to Analyze Coordinate Structures in a Japanese Document, and its Application to a Writing Tool.
Hiroomi YAMAMURA, Akira SUGANUMA, Kazuo USHIJIMA.
Department of Computer Science and Communication Engineering, Kyushu University.

表 1 並列構造 (名詞並列および部分的並列) の推定結果

	並列のキーを 1 つもつ並列構造					並列のキーを複数もつ並列構造					計
実際の並列要素の文節数	1	2	3	4	5 以上	1	2	3	4	5 以上	
正しい並列要素を推定できたもの	254	39	15	7	16	283	68	20	10	13	725
誤った並列要素を推定したもの	35	28	13	13	13	53	46	22	11	20	254
並列のキーとして適当でないもの											42
正解率 (%)	87.9	58.2	53.6	35.0	55.2	85.0	59.6	47.6	47.6	39.4	71.0
並列のキーとして抽出できなかったもの											27

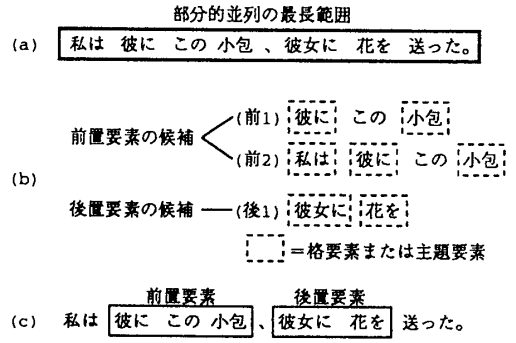

図 1 部分的並列の解析例

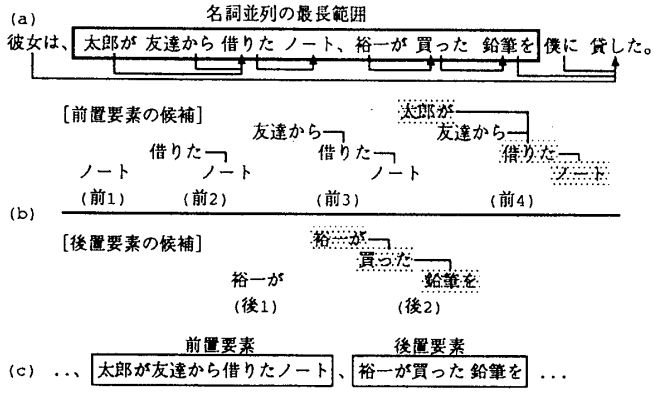

図 2 名詞並列の解析例

要素を容易に決定できる表現がよくある．このような表現を見つけ出すことは，並列要素を決定する表層的な手がかりとなる．そこで，助詞の共起，後置表現，束ねの用法，形態的類似性 [1] という 4 つの手がかりを登録したテーブルを利用する．名詞並列の最長範囲内で，これらの手がかりに字面でマッチする文節が存在する場合は一意に並列要素を決定する．残りの並列要素は，構造的類似性に基づいて並列要素を決定する．この処理は，並列要素の候補間の構造的な類似性を比較することにより，尤もらしい並列要素を決定する．並列要素の候補は，最長範囲内で名詞句となり得るすべての文節列とする (図 2-(b) 参照)．尤もらしい並列要素を決定する優先順位は，構造が完全に一致する候補を並列させる，構造が部分的に最も一致する候補を並列させる，並列のキーの前後の名詞を並列させる，の順である．図 2の例文では，前置要素の候補 (前 4) と後置要素の候補 (後 2) が部分的に 3 文節一致するので，図 2-(c) のようになる．

3 文章への適用

本手法を計算機上に実装し，JICST の抄録文 (299 件，54,858 文字) に対して名詞並列と部分的並列の自動推定を行なった．推定結果の判別については人手で行なった．表 1は，各並列のキーに対する並列要素の推定結果を，並列のキーを 1 つもつ並列構造と複数もつ並列構造に分けて，示した表である．また，実際の並列要素の文節数に対する推定精度も示している．文章を入力し並列構造が得られるまでの解析時間は，SPARC station ELC (CPU：SPARC/33MHz) 上で，1 万字の文章に対して約 1.3 秒程度であった．

並列のキーを 1 つもつ並列構造の推定精度は 76.4%(正解：331 個，不正解：102 個) であった．並列のキーを複数もつ並列構造の推定精度は 60.5%(正解：101 個，不正解：66 個) であった．また，文章中に存在する全並列のキー 1006 個のうち 979 個 (再現率：97.3%) を抽出できた．

4 推敲支援への適用

並列構造を含む文を文章推敲の立場から考えると，並列構造の範囲やその前後の文節間の係り受け関係に曖昧性を含むことが多いので，読み手と書き手の間に食い違いが生じやすい．本節では，並列構造の推定を推敲支援に適用する方法について述べる．本研究では，i) 問題となりそうな並列構造があればそれを提示する，ii) 本手法を用いて得られる情報を提示する，ことを推敲支援への適用方針とした．

i) に関して，書き手と読み手の解釈に食い違いが生じやすい並列構造を一般的に規定するのは難しい．しかし，並列要素を決定する表層的な手がかりが存在する並列構造を分かりやすい並列構造とし，それ以外の並列構造を提示することにした．ii) に関して，次の 2 つの推敲支援を行なっている．一つは，本手法を用いて並列要素の最尤候補，または部分一致候補，または全候補のいづれかを書き手に提示することである．もう一つは，「並列要素の最長範囲が長い」といった並列構造の問題点を提示することである．

5 おわりに

本研究では，大規模な辞書の使用，単語の意味を反映した解析，形態素解析を行なわずに，名詞並列と部分的並列の推定を行なった．今後の課題として，並列要素を推定する規則の充実，述語並列への拡張を考えている．

参考文献

[1] 山村，菅沼，牛島：日本語文における名詞句の並列構造の推定および推敲支援への適用，情報処理学会自然言語処理研究会，111-2，(1996)