

# 日本語文書校正支援ツールの開発

3Q-1

—共起格情報による助詞要素の訂正—

納富 一宏 石井 博章

神奈川工科大学 情報工学科

## 1. はじめに

あらかじめ定義された構文パターンに合致した文章の誤り検出、および訂正の自動化を目的として研究を進めている[1]-[3]。

日本語文における誤り訂正のうち、いわゆる「てにをは」類としての助詞要素を単文レベルでの共起格情報により解析し、訂正候補を絞り込む手法に付いて述べる。また、本手法を用いた校正支援ツールの実装、およびツールの性能評価に関しても触れる。

## 2. 誤り検出

一般に、日本語文の誤り検出では、形態素レベルの解析による検出を行うが、必然的に自立語表記への対応を目的とする場合がほとんどで、付属語類には対しては、別の統語的手段が必要となる。

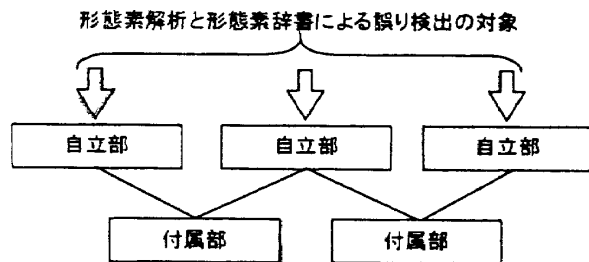


図1. 形態素解析による誤り検出

統語的な順序依存性を重視した検出では、文節単位の構文解析が行われる場合が多く、文単位の誤り検出には問題が残る(例えば、単文の「が」格の重複など)。

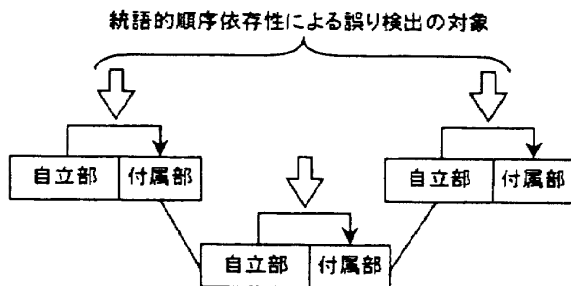


図2. 統語的順序依存情報による誤り検出

したがって、単文単位の格解析による検出を併用しな

ければならない。

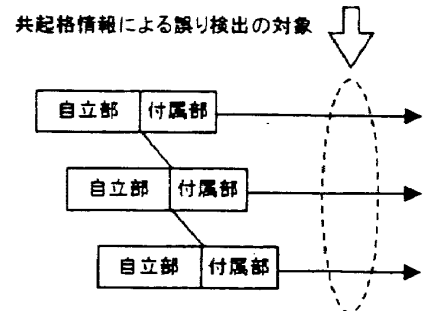


図3. 共起格情報による誤り検出

## 3. 誤り訂正

誤り訂正では、典型的な構文パターンに関してのみ、自動的な訂正が可能である。これは、被解析文において共起した格種別と、あらかじめ構文パターンとして登録されている格フレームとの照合を行うことで、不適当な格を検出し、後述する名詞意味素性などの付帯情報を利用して訂正していくことに相当する。

共起格情報は、特定の述語に対しての格フレームという形で与えられる。したがって、任意の文における本動詞の同定と、その文に出現した格の同定が可能でなければならない。

そこで、述語(本動詞)数が1である文を対象にする。また、被解析文に出現する述語および全ての自立語は同定可能なものとする。

### 3.1 格の同定

被解析文を文節単位の分割し、各文節が以下の JFK 構造[1]、[2]を持つ場合、格の同定は、自立部優先、または付属部優先の2通りの方法が可能である。

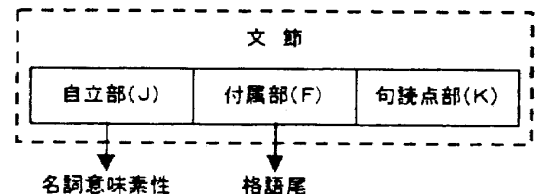


図4. 文節(JFK)構造と格の同定

自立部優先の場合、名詞意味素性および名詞ソーラスが必要であり、全ての名詞要素の同定が可能でなければならない。また、付属部優先の場合は、格語尾から格種別への対応付けが必要である。

後者の方法は、助詞類から格を同定することに相当し、簡便であるため一般によく用いられるが、述語とし

て出現する用言の違いにより、例外を含む(例えば、述語「好きだ」における「が」格は動作主格とならない、など)。

更に、格の同定以外に、助詞要素の自動訂正を目的とした場合、述語一格フレームが被解析文に出現した全ての文節要素に対し、完備でなければならない。

### 3.2 格フレーム辞書

述語一格フレーム辞書の構造を図5に示す。

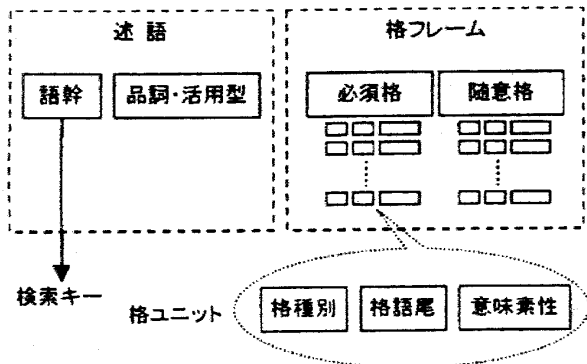


図5. 述語一格フレーム辞書の構造

述語部分は、語幹と品詞・活用の型を持ち、語幹部分を検索キーとする。

格フレーム部分は、必須格フレームと随意格フレームとに分類される。それぞれの部分格フレームには、①格種別、②格語尾、③意味素性、という三つ組みを1セットとする「格ユニット」が複数個保持される。

格種別は、その述語が有する格の種類(分類名)を保持する。格語尾は、付属語情報としての助詞類の表記が保持される。意味素性は、格語尾に対し、前置される自立語の意味分類を保持する。

### 3.3 意味素性とシソーラス

意味素性とシソーラスの例を図6に示す。

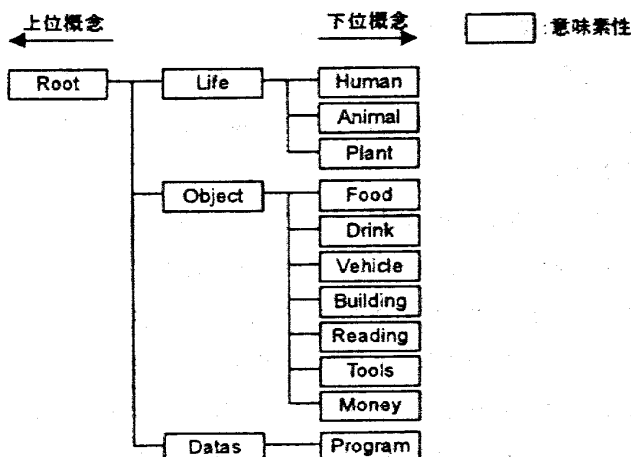


図6. 意味素性とシソーラスの例

意味素性は、シソーラスの構成要素であり、基本的にツリー構造で表現される。被解析文の自立部に出現した全ての名詞類は、この意味素性に対応付けられなければならない。また、格フレーム要素である格ユニット

に含まれる意味素性も、同じシソーラスから選択されなければならない。

図6の例では、トップ概念を Root とし、これを第0階層として、第2階層までを示した。

### 3.4 自動訂正

必須格フレームに含まれる格は、省略される可能性が低く、また、格種別が重複する可能性は更に低い。このため、必須格に対する自動訂正の方法を決めることができる。以下にフローチャートを示す。

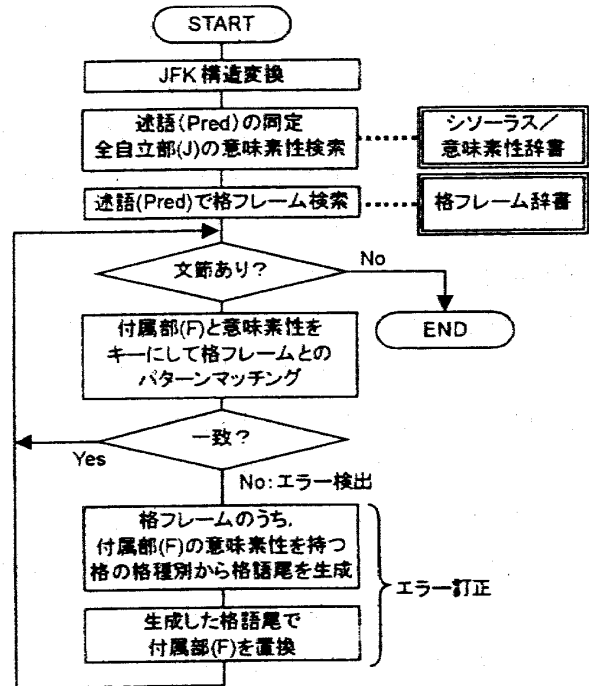


図7. 誤り訂正アルゴリズム

### 4. 評価

正しい日本語文50文に対して、助詞部分を元と異なる別の助詞に変更し、エラーを埋め込んだ。これらの文章を本手法により解析し、自動訂正可能か否か判定した。この結果、ほぼ良好な誤り訂正が可能であった。

ツールはC++を用いてWindows NT上に実装した。

### 5. まとめ

共起格情報による助詞要素の訂正について、述語一格フレーム辞書の構造と名詞意味素性およびシソーラスの例を示し、アルゴリズムを述べた。

本手法は、簡便である反面、付帯情報に大いに依存する。現在、シソーラスの拡張と各種辞書データの拡充を行い、評価を進めている。

### 参考文献

- [1] 納富, 他:「日本語文書校正支援ツールの開発—複合名詞の展開と単文化処理—」, 情処第55回全大, (1997).
- [2] 納富:日本語文書校正支援ツール HSP の開発, 情報処理学会デジタルドキュメント研究会報告, (1997).
- [3] 納富, 他:日本語文書校正支援ツールの開発—動詞格フレームと名詞シソーラスの利用—, 情処第47回全大, (1993).