6J－6

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—

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

専門用語などを多用した文書に見られる、一般的でない複合名詞は、文書校正支援システムにおいて未知語として扱われる。未知語が多く含まれる文書のチェックでは、校正支援のパフォーマンスは低下する傾向がある。これに対応するため、複合名詞が未知語であると判断された場合について、複数の部分文字列への分割（展開）を行い、さらに、適当な付属語要素を補うことで、単文を生成する。生成された単文は、統語構造を持つことから、通常の文書チェック処理を施すことが可能である。

本稿では、複合名詞として表現された未知語の解析を実現する「展開・単文化」手法について述べる。また、本手法を用いた評価実験について触れる。

2．複合名詞要素の分類

複合名詞は、単文における付属語要素の省略と、省略に伴う品詞転成により統語構造を保持したまま一語として表現された文字列として捕らえることができる。

統語構造を有するならば、語の複合化過程を逆にたどることで、最初の単文を推定することが可能である。

例1) 大規模 に 集積 された 回路
例2) 高速 な ネットワーク を 利用 するための 申請書
例3) ファイル へ 出力 する 関数 の 定義
※ [　] :複合名詞要素

これらの例からも分かるように、複合名詞は、形容詞－名詞、副詞－動詞などの係り受け関係（修飾－被修飾）を表現したものが多い。

ここで、複合名詞要素を表１のように分類する。

表1. 複合名詞要素の分類

分類				説明		例
複合名詞要素	名詞	－	－	一般的な名詞		論文，計算機
	サ変動詞	他動詞	対象格優先	「する」を伴ってサ変動詞（他動詞）終止形となる	「を」を伴って対象格をとり、他の格に優先する	開発，利用，設計
			非対象格優先		「から」，「へ」，「で」を伴って源泉格、帰着格、道具格等をとり、対象格に優先する	入力，出力，通信
		自動詞	－	「する」を伴ってサ変動詞（自動詞）終止形となる		行動，変化，発展
	形容動詞	的	－	「的だ」を伴って形容動詞終止形となる		自動，汎用，国際
		－	－	「だ」を伴って形容動詞終止形となる		自然，完全，高速

3．アルゴリズム

複合名詞として表現された未知語の解析のために、次の2つの解析フェーズを設ける。

第１フェーズは、複合名詞要素の抽出を行う。第２フェーズは、抽出された複合名詞要素の出現順序を保持したまま、単文の生成を行う。これら2つの処理をそれぞれ、「展開処理」、「単文化処理」と呼ぶ（図１参照）。

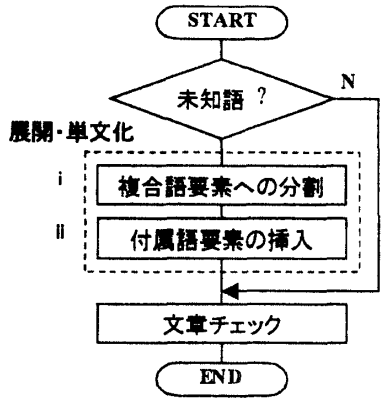


図１.展開(i)と単文化(ii)の文書チェックへの適用

Development of Proofreading Tools for Japanese Document—A Method of Extracting Elements from a Complex Noun and Translating to a Simple Sentence—
Kazuhiro NOTOMI and Hiroaki ISHII
Department of Information and Computer Sciences, Kanagawa Institute of Technology

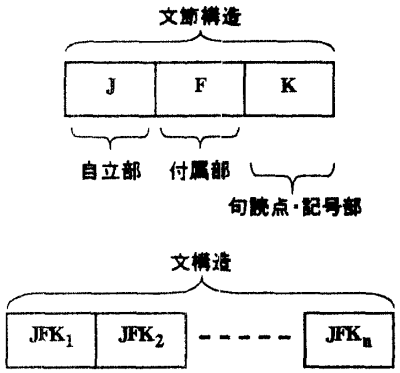
3.1 JFK 構造

複合名詞の展開・単文化に用いるデータ構造としては、我々が以前から提案しているJFK構造[1-3]を用いる。これは、文字種別情報から得られる日本語の文節表現構造であり、文節抽出の際に形態素辞書を必要としないことや、最小限の文字列パターンのみで表現できることから、解析初段のデータ表現に適しているという特徴を持つ。JFK構造を図2に示す。

```
class JFK
{
   char* J;
   char* F;
   char* K;
   int   Jlen;
   int   Flen;
   int   Klen;

   JFK*  Next;
};
```

(a) 定義



(b) 文節と文の表現

図 2. JFK構造

3.2 展開処理（第1フェーズ）

展開処理では、表1による統語情報および格情報を属性値として持たせた形態素辞書を利用し、さらに再帰的に最長一致法を適用して、部分文字列へと分割する。分割された要素をJFK線形リストの各J部へと格納する。この段階では、F, K部に対応する文字列は空である。

3.3 単文化処理（第2フェーズ）

単文化処理では、第1フェーズで抽出された複合名詞要素間に必要な付属語要素を挿入する。複合語要素は、展開処理にて決定された属性値をもとに生成される。これらの付属語文字列はJFK線形リストのF部へと格納する。

次に、各リスト要素を先頭からたどり、それぞれのJFK構造から、J. F部を連結して出力することで、単文が生成される。今回、単文生成に用いた付属語生成規則を表2に示す。

3.4 評価実験

実際に、本手法を用いて複合名詞からの単文生成を行った。Web上のホームページから入手したテキストから複合名詞サンプルを297個収集、972語の形態素辞書を用い、これらを複合名詞展開して、分割数2、3、および4以上のものについての生成文の正誤を求めた。結果を表3に示す。

表2. 付属語生成規則

P_i \ P_{i+1}	−	名詞	サ変他動対象格	サ変他動非対象格	サ変自動	形動−	形動的
名詞	−	の	を	から, へ, で, に, 他	が	の	の
サ変他動対象格	する	する	を	を	が	が	が
サ変他動非対象格	する	する	を	を	が	が	が
サ変自動	する	する	を	を	が	が	が
形動−	だ	な	に	に	に	に	に
形動的	的だ	的な	的に	的に	的に	的に	的に

表3. 評価結果

分割数	正	誤	サンプル数	正解率
2	130	21	151	86.1%
3	69	20	89	77.5%
4以上	39	18	57	68.4%
合計	238	59	297	80.1%

4. まとめ

複合名詞の展開と単文化処理について述べた。本手法では、単純なルールの適用により約300サンプルの複合名詞のうち、80%程度を正しく処理することができた。処理の誤りは、①係り受けの誤り、②助詞選択の誤り、③主語－目的語の混同、が目立った。

統語情報のみでは、意味的な解釈にそぐわない単文化がなされる場合がある。これらを改善するためには、品詞分類の詳細化と名詞シソーラスの積極的な利用を考慮する必要がある。

参考文献

[1] 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).

[2] 納富，他：日本語文書校正支援ツールの開発—複合名詞の統語的検定について—，情処第49回全大，3S-7，(1994).

[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).