

日本語複合名詞の展開・単文化手法

納 富 一 宏¹ ・ 石 井 博 章¹

¹ 情報工学科

A Method of Extracting and Creating a Simple Sentence from Complex Noun in Japanese

Kazuhiro NOTOMI¹⁾, Hiroaki ISHII¹⁾

Abstract

In this article, we propose two methods of analyzing a complex noun in Japanese, which is used to treat as unknown word in proofreading processing and so on. These words decrease performance and precision of proofreading software. Our methods can create a simple sentence from a complex noun with characteristic surface information and morpheme information of that. Therefore, it assumes that a parser in syntax is able to re-analyze in detail these unknown words at the next time. We implemented a CNE system (Complex Noun Extractor and sentence generator) on a Win32 environment with these methods. It only has a simple rule table of sentence generation and morpheme dictionary with 972 entries, but it can analyze correctly about 80% of 297 words. It was confirmed that CNE system is useful to analyzing complex nouns.

Key Words : Complex Noun, Morpheme Analysis, Syntactic Analysis, Case Analysis, Japanese

1. は し が き

日本語の造語作用は、文字種別の多様性、および統語的膠着性から、頻繁に起こり得る現象として観測される。もっとも一般的な形態は「複合語」として表出する。

具体的には、専門的な内容に関わる文書、たとえば科学技術文書などでは、和語、漢語、およびカタカナ表記の外来語の混成形態としての複合語、特に複合名詞が多く見受けられるという点をあげることができる。

これらの複合名詞に注目した場合、自然言語処理における形態素解析、および構文解析の結果は、そこから得られる情報量に依存しており、単純に名詞、または名詞連接として扱うことは、後段の意味解析パーザに大きな負担をかける要因となるため問題視されている。この問題は、通常、解析辞書に当該の複合名詞エントリが存在しないことにより発生する。

したがって、こうした状況に対面する機会が多い場合、専門用語辞書の利用は必須であると考えられる。しかしながら、最初に述べた日本語の特性から言えることは、

専門用語辞書により対応するには困難な場合が多い。

本稿では、この問題に対処するために、複合名詞を構成する語要素を統計的な手法を用いて意味的・統語的に分類し、任意の複合語に対して語要素を抽出する手法、および抽出要素からの単文生成手法について検討する。また、これらの手法を用いた複合名詞の展開および単文生成アルゴリズムを提案することで、日本語文書校正支援システムへの応用を考える。

続く2章、3章では複合名詞解析に関する2つのモデルを提案し、本手法による解析アルゴリズムとシステム実装について述べる。また、4章では評価実験の結果を示し考察を行う。

2. 複合名詞モデルと単文生成モデル

2.1 複合名詞の例

一般に、複合名詞は、形容詞一名詞、副詞一動詞などの係り受け関係（修飾一被修飾）を表現したものが多く、通常、転成名詞が構成要素の一部となっている。表2.1に例を示す。

表 2.1 複合名詞の例

意味構造	例
AをBする	写真撮影, 音楽鑑賞, 現状維持, 特許申請, 海外旅行, 宇宙開発, 問題提起
AをBする人・物	新聞配達, 嘘つき, 各駅停車
AがBする	動脈硬化, 学生運動
A的なB	現代用語, 古代文明, 基本方針, 誇大妄想
AとなるようなB	問題発言
A的にBする	自動生成, 高速変換
A的にXするB	高速道路
Aを目的とするB	捜査令状, 実現方法, 読書週間
Aすることが可能なB	携帯電話, 留守番電話
AしてからBする	整列乗車
Aで(場所)Bする	現地解散, 現地集合, 海外公演, 大学教育
Aに属するB	大学教授

これらの語は一意に解釈可能な語(熟語)であるため, 本来, 1語として辞書に記載されるべきだが, 熟していない語でも複合化される場合があり, すべてを辞書登録で対応することはできない。

2.2 名詞複合化過程—省略と転成—

複合名詞は名詞や形容詞, 動詞などの自立語が複合化され, 全体として1単語となった語系列として定義することができる。この複合化過程は, 統語論的には文表現の簡略化を目的とした省略形式の一つである。一般に省略の対象となるのは付属語要素であり, 助詞や用言の活用語尾などをあげることができる。省略に伴い品詞転成が生じる。

名詞は, 意味論における概念表象であることをふまえた上で付属語の働きに注目すると, 付属語は概念間の関係を示していることになる。一般的には以下のように有向グラフで表現することができる。

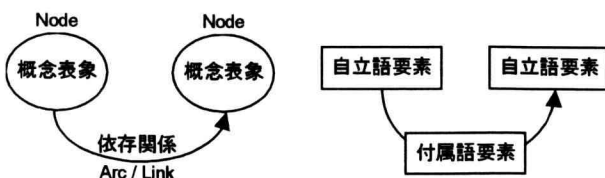


図 2.1 有向グラフによる表現

基本形は図 2.1 のように, 2つのノードと1つのリンクで構成される2項グラフで示される。これら3つの構成要素のうち, ラベル付きリンクで示された付属語要素部分が省略されるためには, これを暗示する図 2.2 のような構造変形を仮定する必要がある。すなわち, 「付属語要素が従属する名詞構造」である。

付属語要素が従属する名詞構造には, 前置名詞側に従属する「左従属」と後置名詞側に従属する「右従属」と

がある。また, 付属語要素が従属可能な名詞は, 統語的に転成名詞となり得る自立語要素でなければならない。

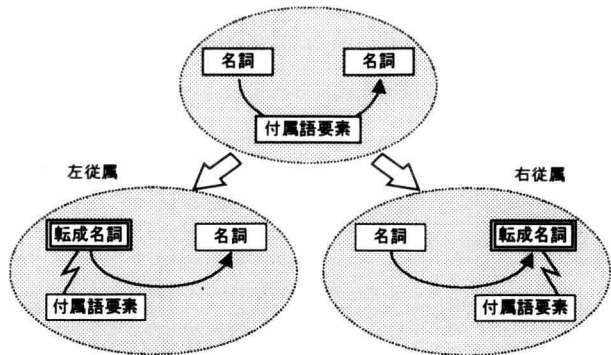


図 2.2 省略に伴う変形（付属語要素の従属）

特定の付属語要素と強い結合をもつ名詞が存在すれば, 付属語要素をキーとしたクラス分類が可能となる。構成要素の特性が大いに関係するものと予想される。

実際の複合名詞における従属付属語要素の展開例を図 2.3 に示す。

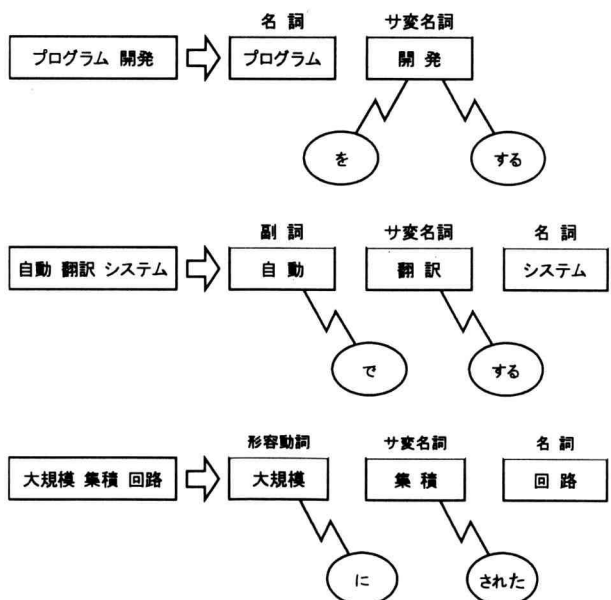


図 2.3 従属付属語要素の展開例

2.3 複合名詞単文化過程—展開と単文生成—

複合名詞モデルは, 名詞の複合化の過程を捉えた表現である。これに対して, 複合化された語から, 複合化過程以前の語要素を抽出し, 意味論的, もしくは統語論的制約から類推できる付属語要素を補うことで, 単文を生成することができる。これを「単文化過程」と呼ぶ。単文化過程の表現を「単文生成モデル」と呼ぶ。

すでに述べた通り, 一般に, 文を構成するそれぞれの文節は, 自立部および付属部に分割することができる。

自立部には、概念表象が含まれ、付属部には、概念間の相互関係（格の依存関係）が含まれる。これら概念表象の共起パターンのうち、依存関係としての付属部要素が従属構造により省略可能であれば、自立部接続が構成できる。

この自立部接続を複合名詞であるとみなせば、順方向の変換過程が複合名詞の生成を意味し、逆方向の変換過程が複合名詞からの単文生成を意味する。単文生成モデルを図 2.4 に示す。

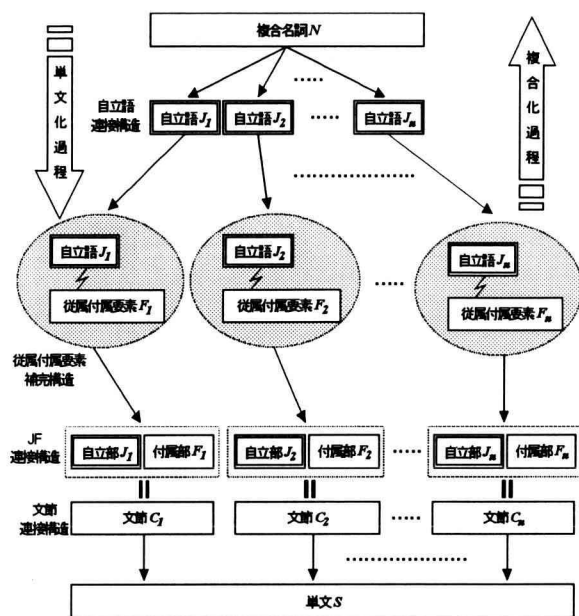


図 2.4 単文生成モデル

以下、図 2.4 について補足する。

任意の複合名詞は、いくつかの構成要素から成る。各構成要素は複合名詞を形成する語基として捉えることができる。この場合の語基とは統語的制約により自立語でなければならない。よって、複合名詞は「自立語接続構造」として認識可能である。

次に、名詞複合化過程より、各自立語は転生名詞であり、付属語要素が従属する名詞構造を保持するものと考えられるため、従属付属語要素を補完した接続構造が得られる。これを「従属付属要素補完構造」と呼ぶ。

さらに、自立語を自立部（J 部）として、また従属付属語要素を付属部（F 部）として読み替えることで「JF 接続構造」となる。JF 接続は仮文節と見なすことができるため、「文節接続構造」が得られる。

よって、以上の変換過程により複合名詞は単文化される。この解釈は、複合名詞生成過程を逆にたどったものである。

2.4 日本語文書校正支援システムへの応用

一般に、文書中に現れた統語的・意味的誤りの検出とその訂正を目的とした校正支援システムでは、形態素解析および構文解析を行うことで表記誤りを検出する。これらのアルゴリズムは解析用辞書、特に自立語辞書に大きく依存したものとなるため、システムの性能は辞書の規模に左右される場合が少なくない（図 2.5 参照）。

ここで、校正支援システムの誤り検出精度を低下させる要因は、「未知語の検出」とであると言える。ここで未知語とは、自立語辞書不記載文字列（単語）を意味する。

自立語辞書に登録されていない単語列は、「正しい表記ではない」とする考え方は、本来、校正の立場では妥当であると言えるが、このアルゴリズムの欠点は、既存の自立語辞書で、あるいは学習機能などでカバーできない専門的な用語や新しい用語には対応できないという点である。

すなわち、正しい用語列もエラーとして扱われてしまう場合があるというのは問題である。校正支援システムの使用する自立語辞書は、10 万語以上のエントリを持つ場合が一般的であるが、ここに含まれないすべての語が誤りであるとするのは乱暴に過ぎる。

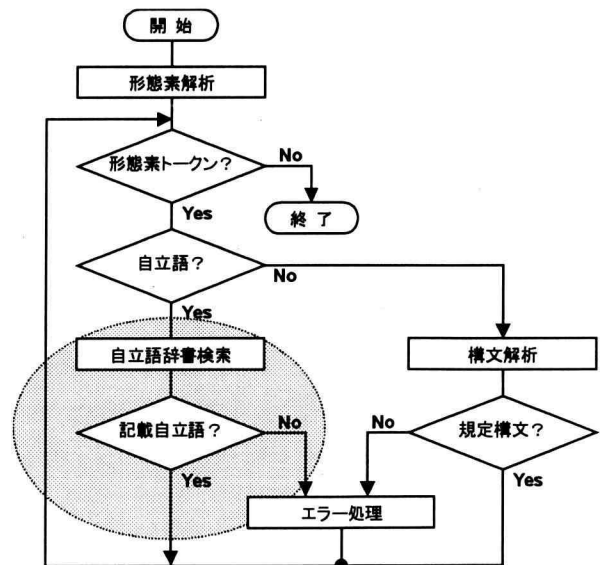


図 2.5 典型的な文書校正アルゴリズム

そこで、本稿で議論している複合名詞の展開・単文化手法を辞書不記載文字列に適用することで、より詳細な解析が実現できる。よって、本来、エラーとして扱われていた不記載文字列のいくつかを救うことにより精度向上が望める。

以下、具体的なアルゴリズム、およびシステム構成について述べる。

3. 展開・単文化アルゴリズムの提案

日本語文における自立語部分の解析において、未知語の決定を従来より遅らせることで、より詳細な解析を実行する手法を提案する。最初に 3.1 において、未知語判定を拡張するための「再帰最長一致分割」について述べる。これは複合名詞の展開処理の基本アルゴリズムである。3.2 では、単文化処理を行う上で必要な複合名詞要素の分類を行う。また、3.3 でアルゴリズムの実装について触れ、3.4 でシステム構成を示す。

3.1 未知語判定の拡張 —再帰最長一致分割—

ある非平仮名表記の自立語 w が未知語である条件とは、解析に用いる自立語辞書を D として、

- ① w が D に記載されていない
- ② w が D において「再帰最長一致分割」できない

のいずれかである。

ここで、②の再帰最長一致分割について補足する。

再帰最長一致分割とは、文字列左端から始めて、 D において最長一致検索を行い、得られた部分文字列を w_i としたとき、 w_i 終端の次の文字から同様に w_{i+1} を取得し、 w 終端に達するまで再帰的にこの操作を反復する文字列分割手法である。この手法による可能な分割数は、最小で1となり部分文字列 w_i は w 自身に一致する。また、最大で w の文字数に一致し、部分文字列 w_i は w の左端から i 番目の文字に等しくなる。

②の条件は、理論的には①を含むが、実際のアルゴリズム表現としては、別に扱う方が都合がよい。理由は後述する。

再帰最長一致分割で得られた w の部分文字列 w_i は、複合名詞の「構成要素候補」であると考えることができる。先に述べたように、複合名詞の構成要素を得る処理は、単文化処理の前段に位置する。本稿ではこの処理を「複合名詞の展開」と呼ぶ。

3.2 複合名詞要素の分類

先に述べたように、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捉えることができる。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。

表 3.1 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
		自動詞	「する」を伴ってサ変動詞(自動詞)終止形となる	「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
		形容詞	「的だ」を伴って形容動詞終止形となる	自動, 汎用, 国際
	形容動詞	—	「だ」を伴って形容動詞終止形となる	自然, 完全, 高速

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、形容詞一名詞、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 3.1 のように分類する。

以下、表 3.1 の分類について補足する。

分類は3レベルの統語的属性からなり、それぞれは、①品詞種別、②活用種別、③優先格種別、である。優先格とは、任意他動詞の格フレームのうち、標準的な文において使用頻度の高い格を意味する。

この分類では、意味的属性を扱っていない。あくまでも表層情報のみから成る。複合語の意味論的側面を扱わないため、文脈に依存した複合化過程や自立語要素の欠落・省略を伴う複合化を説明できるものではない。

反面、アルゴリズム化には適していると考えられる。

3.3 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 3.1 参照)。

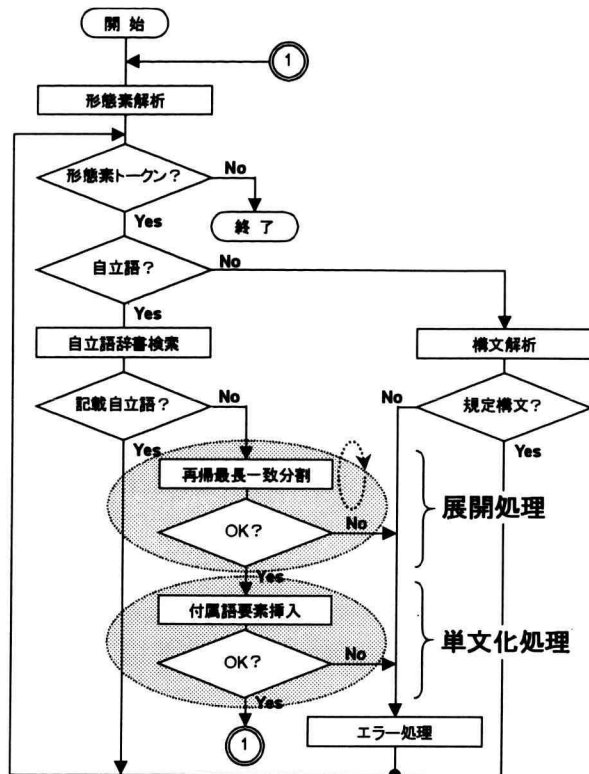


図 3.1 本アルゴリズム（展開処理と単文化処理）

3.3.1 データ表現構造 —JFK 構造—

複合名詞の展開・単文化に用いるデータ表現構造としては、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 構造の定義と概要を図 3.2 に示す。

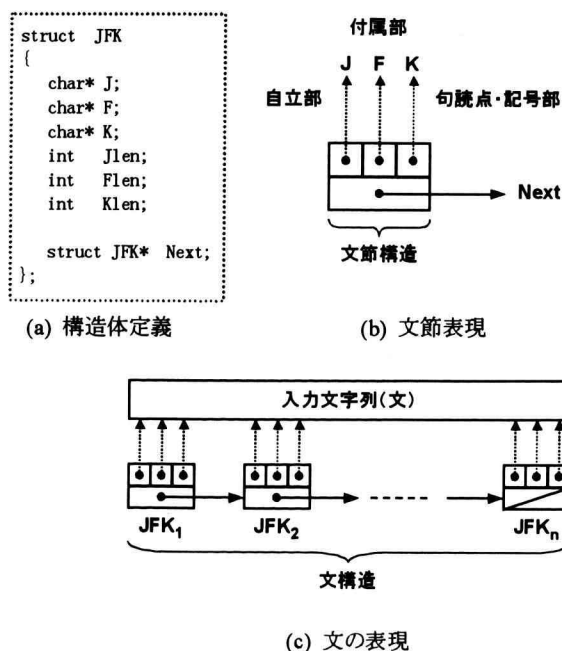


図 3.2 JFK 構造

JFK 構造は、定義にあるように、自立部 (J 部)、付属部 (F 部)、句読点・記号部 (K 部) から成る各パートの開始位置を示す 3 つの文字列ポインタ (J, F, K) とポイントされた部分文字列長 (Jlen, Flen, Klen) を持ち、さらに次の JFK 構造への再帰構造体ポインタ (Next) を持つ。

各 JFK 構造は線形リスト構造をとるので、入力文字列長に合わせた動的メモリ確保が可能となり、メモリ効率が良い。また、入力文字列の文字種別スキャン動作と JFK 分割とが同時に行えるという利点がある。

3.3.2 展開処理（第 1 フェーズ）

展開処理 (図 3.1 参照) では、表 3.1 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 (自立語) 辞書を利用し、さらに 3.1 節で既に述べた再帰最長一致分割を適用して、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。

3.3.3 単文化処理（第 2 フェーズ）

単文化処理 (図 3.1 参照) 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従属付属語要素を挿入する。これらは、展開処理にて決定された属性値をもとに生成規則によって生成される。生成された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 3.2 に示す。

表 3.2 付属語生成規則

		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
P _i	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

以下、表 3.2 について補足する。

展開処理により自立語要素接続が得られるが、これらの要素は表 3.1 に示した分類属性値を持つ。 P_i が前置要素の分類属性値、 P_{i+1} が後置要素の分類属性値である。 P_i と P_{i+1} の交点に示した文字列が付属要素文字列である。

3.4 システム構成

本稿で述べたアルゴリズムを検証するために、Win32 環境上に「複合名詞展開・単文化システム CNE」をインプリメントした。本プログラムは任意の複合名詞の入力により、単文生成を行うことができる。

形態は Windows アプリケーション、開発言語は Borland C++ 4.5J、使用クラスライブラリは Borland OWL2.0、プラットフォームは Microsoft Windows95 または Windows NT4.0 である。

システム構成図を図 3.3 に示す。

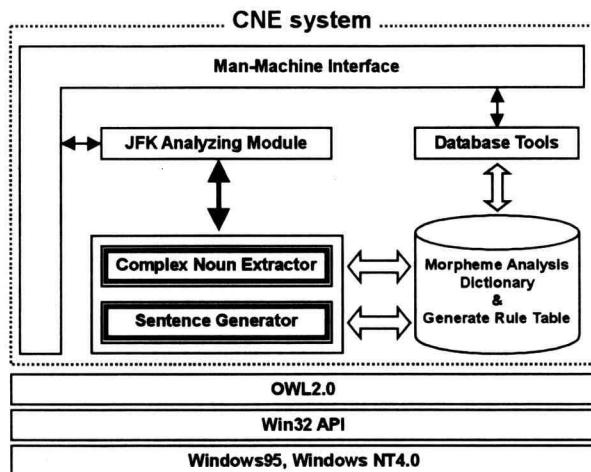


図 3.3 システム構成図 (Complex Noun Extractor System)

以下、図 3.3 について補足する。

CNE は、JFK 解析モジュールと連携して動作し、機能単位では大別して2つの部分、すなわち、複合名詞の展開モジュール、および単文生成モジュールから構成され、各モジュールは共に外部データベースを参照する。外部データベースには形態素（自立語）辞書および単文生成規則が格納されている。また、このデータベースを保守するための専用ツールを備えている。専用ツールは、新規形態素の登録・更新機能、生成規則テーブルの更新機能、外部ファイルからの新規辞書作成機能が含まれる。

4. 評価実験と考察

4.1 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。インターネット上のホームページから入手した

HTML 文書からすべてのタグを排除し、プレーンテキスト (plain text) へ変換後、JFK 分割を用いて自立語候補を収集、その後複合名詞となるサンプルを合計 297 語抽出した。これらの複合名詞に対し、972 語の形態素辞書を用いて複合名詞展開を行い、自立語要素の分割数が 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。単文生成には 6×7 の生成規則テーブルを使用した。

また、生成文が誤りである場合について、エラー分類を行った。エラーは今後のアルゴリズム改良が困難なもののから、①他の自立語要素を必要とする、②意味解析を必要とする、③生成規則の改良を必要とする、という 3 つのレベルに分けた。

本アルゴリズムにより生成された単文の正誤判定結果を表 4.1 に、また、エラー分類別の頻度集計を表 4.2 に示す。さらに、これらに対応するグラフをそれぞれ図 4.1、および図 4.2 に示す。

表 4.1 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%

表 4.2 エラー分類

分割数	他の自立語要素を必要とする	意味解析を必要とする	生成規則の改良を必要とする
2	3	11	7
3	1	9	10
4以上	1	8	9
合計	5	28	26

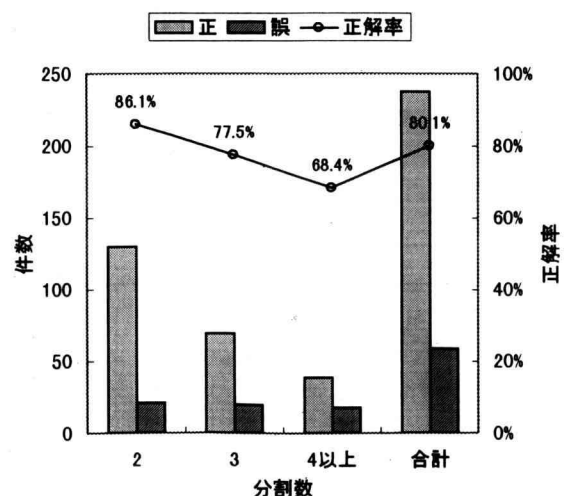


図 4.1 分割数と正解率

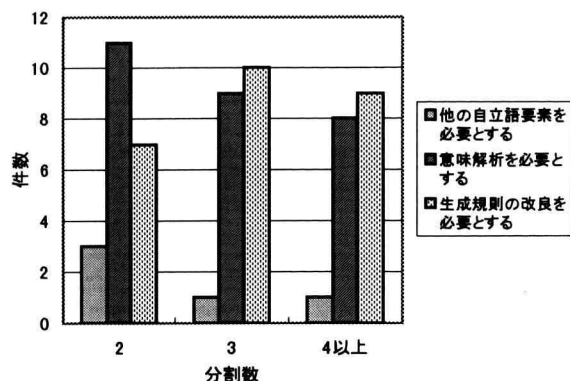


図 4.2 分類エラーの割合

4.2 出力例

本アルゴリズムにより複合名詞から生成された単文の例を以下に示す。なお、矩形で示した部分は自動挿入された付属語要素である。

(正) 生成文の例

入力：産業廃棄物処分場建設 ==>

出力：産業の廃棄物の処分場を建設する

入力：別候補表示機能 ==>

出力：別の候補を表示する機能

入力：暗号化ソフト輸出規制緩和計画 ==>

出力：暗号化するソフトの輸出の規制の緩和を計画する

(誤) 生成文の例

入力：国外流出防止 ==>

出力：国外が流出を防止する

入力：開発途上国 ==>

出力：開発する途上の国

入力：保健医療施策 ==>

出力：保健の医療的に施策する

4.3 考察

図 4.1 から、分割数が大きくなるほど、生成文の正解率（精度）は、ほぼ線形下降することが判る。値としては全体で 80%程度であるため、複合名詞からの単文生成は可能であることが示されたものと考えられる。

図 4.2 からは、分割数が大きくなるほど、他の自立語要素を必要とする複合名詞は減少し、同様に意味解析を必要とするものが減少するという点に注目できる。逆に生成規則を改良すべきエラーへの対処が増加傾向にあることが判る。

分割数が増えるほど、その複合名詞は複雑な統語構造を有するため、今回用いた 6×7 の単純な生成規則（表 3.2 参照）では 70%を下回る（分割数 4 以上の場合）精度しか得られていない。本アルゴリズムは、生成規則が外部データベース中に存在する（図 3.3 参照）ため、今後の改良では生成規則テーブルの拡張と、それに付随する自立語の統語属性および意味属性の細分化により比較的容易に対応ができるものと思われる。

問題は、本来、最も単純であるはずの分割数 2 の複合名詞にある。これらは、自立語の省略を伴う場合が多い。本アルゴリズムでは、入力された複合名詞を分割し、付属語を付加して再結合するという処理を行うが、もともと表層に現れない自立語を類推することはできない。具体例は「高速道路」である。この場合、「高速な道路」という生成文が得られるが、実際には「高速に（車両が走行可能な）道路」という意味である。従って、丸カッコで示した部分を補えないならば、単文化は不可能であると考えざるを得ない。

さらに、意味論的な処理を必要とする例もある。「スピード退院」の場合、「スピードが退院する」という生成文が今回の評価実験で得られたが、意味は「早期に退院する」ということであり、「スピード」という自立語は最終的な単文に現れない。よって、このような場合も対応が困難であると言える。

展開処理、特に再帰最長一致分割に関しては、問題がある場合はほとんどなく、自立語要素抽出の失敗は、わずかに 1 例のみであった。よって、今後の自立語辞書の整備によりほぼ完全な展開が可能であると言える。

また、分割数が最も多い複合名詞の例は、今回では 6 分割のものが 1 例あり、これは以下のように単文化された。

入力：暗号化ソフト輸出規制緩和計画 ==>

出力：暗号化するソフトの輸出の規制の緩和を計画する。

この単文化は判定では正しいものと解釈したが、格助詞の「の」を多用する生成は、構文解析や意味解析では嫌われる場合が多い（格助詞「の」には、所属、存在の時・場所、性質の状態、材料、目的、原因、関係、同格、体言化したことばの主語、という計 9 つの意味があり、同定が困難）ことから、あまり望ましくはないと考えられる。また、上記のように、すべての自立語要素を展開してしまうので、意味的には、「規制緩和」を 1 語で捉えた方がよい場合があるので、制約による展開の制御を考慮しなければならない。

5. むすび

本稿では、複合名詞の展開と単文化処理について、基本アルゴリズムの提案を行った。また、評価実験により提案アルゴリズムの有効性を検証した。

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) 納富一宏：「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) 納富一宏，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) 納富一宏，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)。
- 4) 岡田直之：『語の概念の表現と蓄積』，電子情報通信学会，(1991)。
- 5) 長尾 真 監修：『日本語情報処理』，電子情報通信学会，(1986)。
- 6) 草薙 裕，他：『文法と意味Ⅱ』，朝倉書店，(1985)。
- 7) 市川伸一，他訳：『インダクション』，新曜社，(1991)。