慣用的に組み合わされる助詞,用言句のパターン分析による,名詞の意味情報抽出に関する研究*

山澤 美由起　慶應義塾大学政策・メディア研究科†
佐治 伸郎　慶應義塾大学環境情報学部‡　稲葉 陽子　慶應義塾大学 SFC 研究所§
深谷 昌弘　慶應義塾大学総合政策学部¶

1. はじめに

近年構文解析技術の発展などもあり，係り受けを含むチャンクを対象にパターン抽出を試みるテキストマイニングが発展しつつある．しかしその場合扱うデータ構造が形態素解析を利用し単語間の共起関係，相関関係に注目した場合より複雑化するため，現実的に有効な成果を残しているマイニング手法に関して言えばそれほど多くのものが提示されているとは言い難い．

本稿ではこのような流れを受け，日常的な言語使用を反映した大量のテキストデータの集積があることを前提に，テキストマイニングに新しく認知言語論的なアプローチを導入することにより，頻出する係り受け構造から有効な情報・意味を抽出する新たな手法を提案し，その試用結果を示す．本研究では田中茂範，深谷昌弘[1][2]の提案する助詞の操作子機能と用言句の図式構成機能に注目し，テキストデータにおける「名詞＋助詞＋用言句」（以下[句]）という一連の表現を抽出することにより名詞（より正確にはその名詞により指し示される事物）の意味情報抽出，およびそのシステム化を試みた．

2. 目的

慣用的に組み合わされる助詞，用言句のパターン分析を行うことにより，その名詞の意味情報の抽出を目的とする．

3. 設計方針

名詞の意味はその使用される情況によりその都度与えられるものであると考えられる．よって文脈に依存する意味の多義性が生じることは不可避である．しかし，[句]の中に位置付けられたとき，助詞の操作子機能と用言句の図式構成機能によりこの多義性はほぼ消滅する．これにより，その名詞の担う意味情報の解釈のまぎれもほぼ消滅する。したがって，この[句]集合から慣用的に使用されるものを集め，意味解釈作業を施すことにより，広く一般的に共通に理解されている名詞の意味情報の抽出が可能となると考えられる．

本研究ではこのような考えに基づき，意味解釈の段階までも視野に入れ[句]をまとめあげ，解析前（形態素解析前の raw data）とシステムによる解析後のデータを併用して意味解釈を行うようシステムを設計した．

4. システム概要

4-1. 開発環境・動作環境

本システムは Windows 上で Perl 言語を利用して開発された analyze.pl を主軸に構成される．形態素解析には ChaSen[1]を利用した．ハードディスク空き容量 50MB 以上，Windows95 以上を動作環境とする．

4-2. 作業手順

分析用データファイル作成から名詞の意味情報抽出までの作業の流れを図１に示す．
まず、分析用のデータファイル（データファイル(1)）を任意のデータソースより作成する．なお，分析用データファイルの作成については検索エンジンとの連動による，効率的なデータファイル作成用サブシステムの構築が可能である．

---

* Extraction of semantic information based on meaning schema
† Miyuki Yamasawa, Graduate School of Media and Governance, Keio University
‡ Noburo Saji, Faculty of Environmental Information, Keio University
§ Yoko Inaba, Keio Research Institute at SFC
¶ Masahiro Fukaya, Faculty of Policy Management, Keio University

[1] 奈良先端科学技術大学院大学自然言語処理学講座によるフリーの日本語形態素解析器 http://chasen.org/index.html.ja

次に，形態素解析（ChaSen を利用）し、その結果を元に[句]を抽出，その結果は CSV 形式で出力される（データファイル(2)）．また，その結果の各助詞・用言句ごとの集計結果も CSV 形式で出力される（データファイル(3)）．

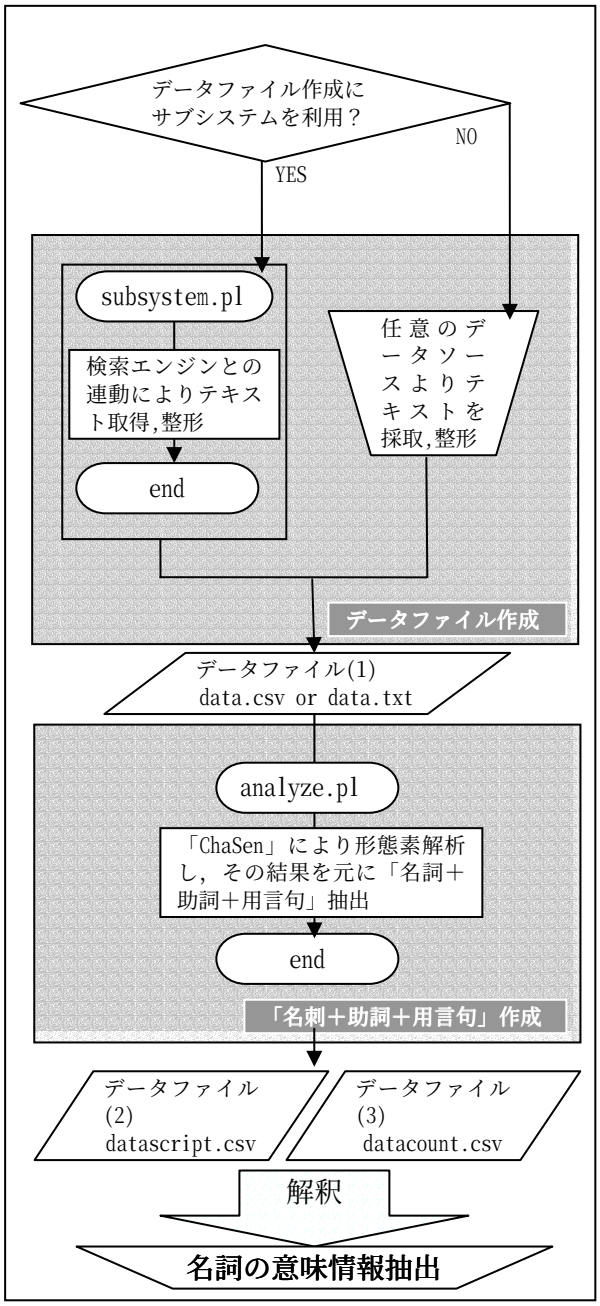

図１　作業手順

5. 試用結果

ある新聞社の記事のアーカイブを利用して「情報」についての意味情報抽出を実験的に行った．1993 年 4 月 1 日から 2001 年 11 月 29 日までの投書記事のうち「情報」という名詞を含む 1667 件について analyze.pl による[句]の抽出を行った．その結果 947 種類の[句]が抽出され，それらのうち 102 種類が反復して抽出されていた．

他の例としては「知識」や「心」についても同様の操作を行ってみた．その結果，反復して抽出された[句]について，出力されたデータファイルは傾向別に分類できるものとできないものとが存在していた．

分類できないものについては出力結果の形態等に工夫が求められるものと考えられる．

6. まとめ・展望

本研究では，助詞の操作子機能と用言句の図式構成機能に注目し，テキストデータにおける[句]を抽出し，そのパターンを分析することにより，広く一般的に共通に理解されている名詞の意味情報の抽出，およびそのシステム化を試みた．

今後の課題としては，システムの出力結果の直感的理解を支援するような，結果の可視化が挙げられる[3]．

また，このシステムの応用としては主として，人々が様々な事物をどのようにとらえているかを探る WEB 社会調査が考えられる．その一例として，形容詞・形容動詞を含む[句]に着目したイメージ調査などを挙げることができる．

7. 謝辞

本研究を進めるにあたり貴重なご助言をいただいた，慶應義塾大学石崎俊教授，同研究室所属岡本潤氏，同大学深谷研究室所属岡田智靖氏，二宮朋子氏，同大学理工学研究科所属江木啓訓氏に謝意を評す．

8. 参考文献

[1]田中茂範，深谷昌弘：＜意味づけ論＞の展開情況編成・コトバ・会話，紀伊国屋書店(1998)．
[2]深谷昌弘，田中茂範：「コトバの＜意味づけ論＞　日常言語の生の営み，紀伊国屋書店(1996)．
[3]馬青，神崎享子，村田真樹，内元清貴，伊佐原均：日本語名詞の意味マップの自己組織化，情報処理学会論文誌，Vol.42，No.10，pp2379-2390(2001)．