学術用語シソーラス作成のための語彙抽出にかかわる言語データの検討

荻野 孝野† 影浦 峡‡ 出羽 達也†

†(株) 日本電子化辞書研究所 (EDR) ‡学術情報センター (NACCIS)

1 はじめに

本研究は、学術情報データベースを対象にした[1][2]、知的な検索を可能にする Query Expansion[3] のためのシソーラスデータ構築の検討の一環として行なったものである。ここで想定している検索は、用語の文字列照合による「全文検索」である。

シソーラスデータ構築は、

1) 分野限定の検索対象データベースから索引語を抽出
2) 索引語のシソーラスの構築

という手順で、まず、シソーラスに組み込む索引語の抽出から始まる。ここでは、この索引候補語の抽出部分に焦点をあて、複合的表現と単一語の関連が明確に分析できるように、複合的表現の中で、必要な部分、不要な部分という観点で区分けをした「接辞」の検討を中心にとりあげる。

2 学術用語シソーラス作成手順

対象となる学術情報データベースは、1) キーワードデータ 2) 抄録データの二種類からなる。

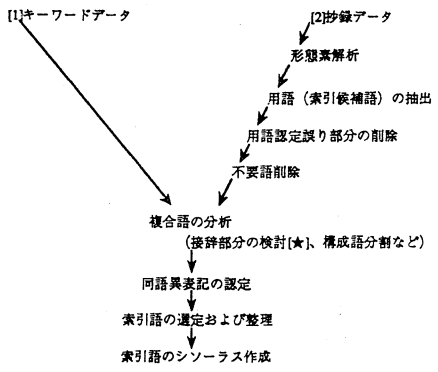


図1：学術用語シソーラス作成の流れ

ここでは、まず、用語が選定されている [1] キーワードデータを対象に中核となるデータを作成し、次に [2] 抄録データ を用いて索引語の補完、拡張を行なう。[1]、[2] は図1に示す通り、抄録データについて形態素切りから用語選択の部分が異なるのみで、用語選択以降は、ほぼ同じ手順で進める。

上記の手順によって、試作した一部を、添付資料1に示す。

3 複合語候補文字列に含まれる接辞相当語の検討

本発表では、図1のうち、★部分「複合語の分析–接辞部分の検討」を中心にとりあげる。なお、通常「語基+語基→複合語」「語基+接辞→派生語」[4] というように、接辞がつくものは、「派生語」とよばれているが、ここでは、便宜的に一つ以上の語構成要素から成り立つ語をまとめて複合語とよぶ。

3.1 接辞相当語の抽出

(1) 学術情報データベースの形態素解析まず、索引候補語を抽出するために形態素解析を行なったが、これは、JUMAN3.1[5] を用い、品詞体系を EDR 日本語単語辞書の品詞体系に変換したものを利用している [6]。データベースは、情報処理学会の学会発表抄録データ 26252 件 [7] を対象にした。

EX.1 形態素解析結果

```
<s id=s1>
  <m id=s1m1 org='データベース' type='普通名詞' orth='データベース'>データベース</m>
  <m id=s1m2 org='管理システム' type='普通名詞' orth='かんりシステム'>管理システム</m>
  <m id=s1m3 org='G-BASE' type='未定義語' orth=''>
  G-BASE</m>
  <m id=s1m4 org='の' type='助詞' orth='は'>は</m>
```

(2) 頻度順形態素リストの作成

(1) で切り出された形態素について、頻度つき異なり語表を作成したものである。

EX.2

17894	システム	普通名詞
11953	処理	サ変名詞
11181	情報	普通名詞
9642	開発	サ変名詞

(3) 複合語の抽出

(1) で切り出された形態素から、名詞または接辞の連続を抽出し、頻度つき異なり語表を作成したものである [6]。ここでは、動詞連用形転成名詞の認定等も行なっている。

EX.3「複合語 & 単語」のリスト

6237	本/稿	連体詞的接頭語/普通名詞
5534	システム	普通名詞
4813	方法	普通名詞
4184	問題	普通名詞

(4) 接辞の抽出

(3) で抽出された複合語レコード (「/」を含む) のうち、構成要素に「接頭語」「接尾語」を含んでいるレコードを抽出する。

EX 4.1 接頭語を含む複合語の頻度順表

6237	本/稿	連体詞的接頭語/普通名詞
2574	本/論文	連体詞的接頭語/普通名詞
2117	本/研究	連体詞的接頭語/サ変名詞

EX 4.2 接尾語含む複合語の頻度順表

1194	有効/性	普通名詞/接尾語
909	利用/者	サ変名詞/接尾語
870	高速/化	普通名詞/接尾語

元データに対して複合相当語がどれぐらい含まれているか、複合相当語に対し、接辞がどれぐらい含まれているかのおおよそを表 1 に示す。

表 1: 元データの形態素数と接辞

	延べ	異なり
形態素数	3112180	44453
複合相当語	317000	133967
接尾語	61192	180
接頭語	28218	38

3.2 接辞相当語の分類

「3.1 接辞相当語の抽出」によって、抽出された接辞を含むレコードについて、接辞の分類を行なう。ここで、接辞の分類は、形態的な接辞の分類 [8] ではなく、索引語として、必要か不要かという観点で区分けを行なう。(1) いわゆる索引語本体から不要な部分を極力排除して、複合語と、その構成部分が単独で使われている場合との明確な関連がつけられるような操作を行なうこと、(2)「~化、~的」のように、接辞がつくことによって意味変換が起こるものを明確にして、体系化の効率をはかること、に着目し、索引語選定に有効な接辞の分類を行なうものである。

また、形態素解析用辞書においても、かなり接辞の含む範囲は広くとられているが、具体的なデータにあたってみると、単に接辞の位置にきているだけの「~自身、~ほど、だけ」などといった、再帰代名詞、副助詞なども、接辞に分類されているものも多いことがわかる。

以上のような観点で接辞相当語として品詞がふられたものを検討し、「3.1 -(4)」で作成したデータを対象に以下の区分を行なう。

1) 接辞以外のもの

1.1) 形態的に接辞の位置にくるが接辞でないもの@

EX.5 1/台/@あたり, コンピュータ/@同士
@今/一歩, @高/・/低/の/アクセント

1.2) 単語認定誤りによる接辞の誤認定%

EX.6 曲線/%あて/はめ, その手/%がかり
もの/%ごと, %高/々

2) 接辞

2.1) 必要な接辞 (意味にかかわる接辞) #

EX.7 定式/#化, 遺伝/#的, 利用/#者, 有効/#性

2.2) 不要な接辞 (意味にかかわらない接辞)*

EX.8 *ご/意見, *各/工程, *他/分野, *同/枠組

3) その他

元データ時の入力ミスで接辞と認定されたと思われるもの $

EX.9 利/$間/者 (←利用者),$項/目 (←項目)
メニュー/階層/$高/造 (←階層構造)
$ま/とめ役, 機/$様/仕様 (←機能仕様)

表 2: 接辞相当語の区分内訳

	接頭語		接尾語	
	異なり	延べ	異なり	延べ
#必要な接辞	16 (42%)	6262 (22%)	24 (30%)	45545 (74%)
*不要な接辞	12	21902	13	10195
@接辞以外の品詞	4	15	46	2734
%単語認定誤り	4	37	31	2156
$その他 (入力ミスなど)	2	2	76	562
	38	28218	180	61192

3.3 意味にかかわる接辞に関するシソーラス上での扱い

次に、3.2で接辞として分類した (#,*) の区分の入ったものについて、意味的な分類を行なったものの一部を下記に示す。

3.3.1 接頭辞の意味区分

#1 程度にかかわる接頭辞
#高/コストルール　形容詞的接頭語/普通名詞
#最/下層　形容詞的接頭語/普通名詞
#小/グループ　形容詞的接頭語/普通名詞
#2 正副にかかわる接頭辞
#準/ミニマックス　連体詞的接頭語/普通名詞
#3 順序、新旧にかかわる接頭辞
#新/仮説　形容詞的接頭語/普通名詞
#前/処理　連体詞的接頭語/サ変名詞
#4 否定にかかわる接頭辞
#非/UNIX/システム　副詞的接頭語/未定義語/普通名詞
#不/稼働/時間　副詞的接頭語/サ変名詞/普通名詞

3.3.2 接尾辞の意味区分

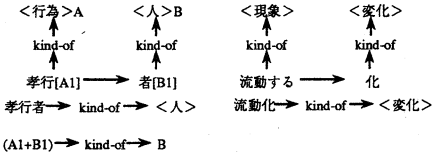
#1 変化にかかわる接尾辞
#化　ベクトル/化　普通名詞/接尾語
#2 様態にかかわる接尾辞
#的　アルゴリズム/的/記述　名詞/接尾語/サ変名詞
#性　膠着/性　サ変名詞/接尾語
#3 用途にかかわる接尾辞
#用　論理/検証/用　普通名詞/サ変名詞/接尾語
#*4 人にかかわる接尾辞
#者　システム/制作/者　名詞/サ変名詞/接尾語
*様　お客/様　普通名詞/接尾語

3.3.3 意味区分とシソーラスとの関係

3.3.1, 3.3.2に示したように意味区分された接辞の中には、接辞が付加されることによって、語基部分の意味に変化を付加するものと、何ら意味変化をもたらさないものがある。前者の例として、例えば「膠着/性」にみられるような「<現象>（サ変名詞）＋”性”（接尾辞）→<状態>」への変化にかかわる接辞、後者の例として、「お客様、お客」（いずれも<人間>）のように意味変化にかかわらない接辞があげられる。

シソーラスでの位置付けでみれば、前者の例は、構成要素である語基部分「膠着」と複合語「膠着性」とは異なる部分に配置される。後者の例「お客様、お客」はともに同じ部分に配置される。このように、意味変換にかかわりのある接辞については、これらの接辞を分類整理することによって、語分割した時の構成語との関連をどうとるか、あるいは、索引語としてシソーラスに配置されていない単語が新たに生じた場合、どの位置に配置するかなどの処理に「NP+VP →VP」といった構文上の変換規則同様に、「構成要素＋接辞意味区分N →複合語意味区分N」のような統一的な意味変換式を持ち込むことを提案するものである。

EX.10



また、これらは、派生的な利用であるが、増え続ける造語にどう対応するかにも深くかかわる部分である。ＥＤＲのように概念辞書にリンクした単語辞書という位置付けの場合、「名詞＋接辞→名詞」ということが形態的に可能であっても、合成によって意味の変化が生じ、体系上の位置が変わるのであれば、新語として登録するか、意味的変換式を作成して対処することが必要である。造語力の高い漢語などについて、辞書登録あるいはシソーラス上への位置付けという両方の観点からも、接辞の意味的な分類は必要である。

4 まとめ

本研究は、情報検索用シソーラス作成のために、複合語とそれらの構成要素が単独で出現するものとの関連づけを効率的に行なうための一つのステップとして、接辞の分析を試みたものである。

接辞そのものについては、まとめると次のような傾向が見られる。

1. 分野限定のデータベースを対象とした場合、語の異なりがかなり限定されている。

   接頭辞 (異なり 28)、接尾辞 (異なり 37)

   (参考:EDR 辞書 (文献 [9] では、接頭辞 296。接尾辞 609)

   →分野ごとの接辞の出現傾向の調査が必要である。

2. よく出現する接辞に偏りがある。

   接頭辞：最出語 「本～」延べ 17144 語 (60%)

   接尾辞：最出語 「～化」延べ 11755 語 (21%)

   → 不要な接辞の整理、意味の変換にかかわる接辞の整理を行なえば構成要素の抽出およびシソーラス作成時の効率化が期待できる。

3. 接辞抽出時に認定エラーが起こっている事例数が、接頭辞と接尾辞で異なる。

   接頭辞：接頭辞として抽出された文字列のうちの 0.1%が切り出しエラー

接尾辞：接尾辞として抽出された文字列のうちの3.5%が切り出しエラー

→ 形態素解析が左側を始点とした最長マッチングによって行なわれることに関係すると思われる。接頭辞の場合は、通常最小文字列で辞書照合。最後の候補語となるケースが予想される。接尾辞の認定エラーは、ひらがな混ぜ書きの自立語部分の誤認定が多い。より短い単語を登録するにあたっては、それらを一部に含む単語の辞書登録の充実が必要である。

以上のような検討結果を参考に、今後の実用化にあっては、さらに他分野での調査比較、接辞の意味区分の詳細化を行なう。

○ 添付資料1 キーワードデータからのシソーラス作成

索引語「検索」に関する体系表の部分

[システム]: 検索システム 14
情報検索システム 11

[手法]: 全文検索 68
類似検索 42
文字列検索 21
キーワード検索 20
知的検索 13
:
:

○添付資料2

2.1 接尾辞区分表

2.11 必要な接辞

接尾辞	出現頻度
#化	11755
#的	7678
#性	7415
#間	4642
:	

2.12 不要な接辞 *

接尾辞	出現頻度
*上	5917
*中	2073
*後	312
*前	136
:	

2.2 接頭辞区分表

2.21 必要な接辞#

接頭辞	出現頻度
#高	1465
#再	1149
#非	1128
#超	551
:	

2.22 不要な接辞

接頭辞	出現頻度
*本	17144
*各	3514
*同	493
*全	390
:	

2.13 接辞以外 @

出現頻度	
@ごと	498
@以上	313
@以下	268
@なし	256
:	

参考文献

[1] 情報検索サービス NACSIS-IR : http://www.nacsis.ac.jp/ir/ir-j.html

[2] 大山敬三;「インターネットに適応した全文データベース検索システムの構成」学術情報センター紀要第7号、p. 13-27, (1995)

[3] Lancaster, F. W. Vocabulary Control for Information Retrieval. 2nd ed. Arlington, Information Resources Press. (1986)

[4] 斎藤倫明、石井正彦;「語構成」日本語研究資料集 1-13; ひつじ書房 (1997)

[5] 松本裕治、黒橋禎夫、山地 治、妙木 裕、長尾 真;「日本語形態素解析システム JUMAN version 3.1 使用説明書」; 京都大学工学部長尾研究室、奈良先端科学技術大学院大学松本研究室 (1996) http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/

[6] 出羽達也、清野正樹、安川秀樹、西野文人、辻井潤一「コーパスからの言語知識抽出 (語彙知識・共起知識・概念関係知識)」; EDR電子化辞書利用シンポジウム 論文集 (1997)

[7] Kageura, K., Koyama, T., Yoshioka, M., Takasu, A., Nozue, T. and Tsuji, K. (1997) "NACSIS Corpus Project for IR and Terminological Research", Natural Language Processing Pacific Rim Symposium 1997. 1–4 December, Phuket, Thailand. p. 493–496.

[8] 丸元聡子, 乾 裕子, 荻野孝野; コーパスを用いた名詞と接辞の形態的分類; 言語処理学会第4回年次大会予稿集 (1998)

[9] 日本電子化辞書研究所;「EDR電子化辞書仕様説明書 (第二版)」(1995)