自然言語処理 153－8
（2003．1．20）

ウェブを利用した関連用語の自動収集

佐藤 理史†, 佐々木 靖弘‡

†京都大学大学院情報学研究科知能情報学専攻
‡京都大学工学部電気電子工学科

与えられた専門用語に対して、その用語と関連する用語をウェブテキストを利用して収集する方法を提案する。提案方法は、コーパス作成、重要語抽出、フィルタリングの3ステップから構成される。コーパス作成では、サーチエンジンを利用して、与えられた用語を説明するテキストをウェブから収集し、その用語に対するコーパスを作成する。次の重要語抽出では、このコーパスから、中川の方法を利用して重要語を抽出する。最後のフィルタリングでは、得られた重要語の中から、関連用語としてふさわしいものを、ウェブのヒット数を利用して、選択する。

Automatic Collection of Related Terms from the Web

Satoshi Sato†, Yasuhiro Sasaki‡

†Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
‡School of Electrical and Electronic Engineering, Kyoto University

This paper proposes a method of collecting related terms of a given technical term from the Web. The proposed method consists of three steps. The first step, compiling corpus step, collects texts that contain the given term by using search engines. The second step, automatic term recognition, extracts technical terms from the corpus by using Nakagawa's method. These extracted terms become the candidates for the final step. The final step, filtering step, selects the related terms from the candidates based on the search engine's hits.

1. はじめに

「ある用語を知る」ということは、その用語が何を意味し、どのような概念を表すかを知ることである。それと同時に、その用語が他のどんな用語と関連があるのかを知ることは非常に重要である。特定の専門分野で使われる用語(専門用語)は、その分野のなかで孤立した用語として存在することはない。その分野で使われる他の用語に支えられ、その関連を土台として、はじめて意味を持つ。それらの用語間の関連を把握することは、「その専門分野について知る」ことでもある。

ある分野で使われる用語を集めた専門用語辞典においては、それぞれの用語の項目に、用語の説明と関連用語が記述されるのが普通である。関連用語は、用語の説明の文章内に参照記号を付加することによって示されるとともに、参照語として独立に記述されることも多い。このように、辞典の構成という点からも、用語の説明と関連用語は、用語の理解を助ける2つの基本的な要素となっていることが確認できる。

これまで、我々は、「ウェブを使って用語の意味を知る」ことを支援するシステムとして、用語説明探索システムを実現してきた[1)]。このシステムは、サーチエンジンと情報抽出の技術を用いて、ウェブ上に存在する用語説明を見つけ出し、これを整理した形で提示する。このシステムは、いわば仮想的な用語辞典を目指したものであるが、従来の辞典が提供する2つの要素のうち、提供できているのは用語説明のみであり、関連用語を提供することは実現されていない。

ある用語に対して、関連する用語を見つけ出す機能を実現することは、次の2つの点において有用である。第一に、上記のシステムにおいて欠落していた関連用語に関する情報を提供することを可能とする。第二に、ある

用語から出発して特定の分野で使われる専門用語の集合を収集することによって、特定の分野の用語集・用語辞典を自動的に生成する道を開く。

このような背景より、本研究では、ウェブを利用して、与えられた用語の関連用語を見つけ出す方法について検討する。以下では、まず、「関連用語の収集」という問題を定義し、その解法について検討する。次に、本研究で作成した関連用語収集システムの概要を述べ、システムを構成する3つのモジュールについて説明する。最後に、本システムを用いた実験とその結果について述べる。

2. 「関連用語の収集」問題とその解法

2.1 「関連用語の収集」問題

本研究が対象とする「関連用語の収集」問題を、以下のように定義する。

> **入力** 専門用語 t
> **出力** t に関連する専門用語 (関連用語) の集合 $X = \{x_1, x_2, \ldots, x_n\}$
> ここで、n は、10 を一つの目安にする。

ある用語に関連する用語の集合は、「関連」をどう捉えるかによって、いかようにも変化する。そこで、まず、求めるべき用語集合の大きさを先に決めてしまい、その大きさを見合うように「関連」の捉え方を決めるという方法を採用する。$n = 10$ という目安は、用語の説明と一緒に提示することを想定して決めた数である。

問題をこのように定義するとき、収集すべき用語 x は、

(1) 専門用語であり、かつ
(2) t と関連する

という条件を満たすものとなる。以下では、これらの条件について検討する。

2.2 「専門用語」とは

「専門用語とは何か」ということは、ターミノロジーの中心的問いの一つである。Kageura と Umino[2] は、重要語抽出 (automatic term recognition; ATR) の手法を整理するために、unithood と termhood という2つの概念を提示している。

unithood the degree of strength or stability of syntagmatic combination or collocations

termhood the degree that a linguistic unit is related to domain-specific concept

しかし、最近になって、影浦は、「専門用語とは何か」をより直接的に議論した論文[3]において、次のような定義を与えている。

- 専門用語とは専門用語として使われるものである (p3)
- 専門用語は、もっぱら／特権的に／主に、特定の専門分野で使われる語彙的単位である (p6)

我々は、影浦の説得力あるこの論文に同意し、専門用語の定義としてこの定義を採用する。

この定義を採用すると、「ある用語 x が**専門用語である**かどうか」は、「用語 x が**専門用語として使われているか**どうか」を判定することに帰着される。つまり、我々が次に考えるべきことは、「専門用語として使われている」ことの証拠として、どのような証拠を採用すべきかということになる。

「専門用語として使われている」とは、ある集団の人々が、ある分野の特定の意味を表すために、その用語を実際に使っていることと考える。但し、特定の分野によらず一般に使用される用語（いわゆる一般語）は、専門用語とは見なさない。まず、このことに対応する、以下の2つの証拠を採用する。

(1) 特定の分野で広く、または、それなりに使われている。
(2) 一般語ではない。

もし「専門用語として使われている」のであれば、それがどのような意味や概念を表すのかについての説明がどこかに存在してしかるべきである。また、専門用語は他の専門用語との関連に立脚しているので、関連する専門用語があるはずである。すなわち、

(3) 定義や説明が存在する。
(4) 関連する専門用語（関連用語）が存在する。

以上の4つを、我々は「専門用語として使われている」証拠として採用する。なお、これらの証拠を具体的にどう計測するかは、実現上の問題と考え、3節で議論する。

2.3 「関連する」とは

次の問題は、「関連する」をどのように捉えるかという問題である。ここでは、この問題を、関連のタイプと関連の強さの2つに分けて検討する。

「関連のタイプ」の問題とは、2つの用語間の関連に、シンボリックな種類 (= 関係) を導入するかどうかという問題である。用語間あるいは概念間の関係としてどのようなものを設定すべきかということは、古くから多くの議論があったわけであるが、いくつかの基本的な関係を除いては、いまだに決着を見ていない。おおよそ合意できる範囲は、情報検索用のシソーラス[4]の構成で用いられる等価関係、階層関係、連想関係に限られる。このような理由により、我々は、「関連のタイプ」には深入りしない方針を取る。すなわち、(i) 広い意味での階層関係（上位語、下位語）と、(ii) それ以外の関係（関連）、のみを設定する。

「関連のタイプ」を上記のようにほとんど設定しないとするならば、「関連の強さ」という概念を持ち込む必要性が生じる。この「関連の強さ」とは、我々が持っている2つの用語間の親密度/関連度を数値化したものである。このような指標を持ち込まない限り、関連用語の数を十分に ($n = 10$ となるように) 絞り込むことはできない。但し、用語 t に対して、その上位語と下位語は無条件で「t と関連する」と捉え、強さは持ち込まないことにする。

以上をまとめると、「t と関連する」は、以下のいずれかとなる。

(1) tの上位語または下位語である。
(2) tとの関連度の値が十分大きい。
これを実行可能なものとするためには、上位語、下位語の判定条件、および、関連度の計測方法と閾値を与えればよいことになる。

2.4 重要語抽出によるジェネレータの実現

ここまでの議論で、収集すべき用語xの条件と判定方法の基本方針が定まった。しかしながら、ここから導けることは、tとxが与えられたとき、xがtの関連用語となっているかどうかを判定する方法だけである。すなわち、もし、xの候補となる集合が与えられれば、そこから条件を満たす語だけを選びだすフィルターを実現することはできるが、xの候補となる集合をどのように作成する方法については、何も明らかになっていない。つまり、収集すべき用語xの条件を明確化するだけでは、候補のジェネレーターを作ることができないのである。ここで、何らかの発想の転換がもとめられることになる。

関連用語を収集するという問題は、これまであまり検討されてこなかったが、専門用語を収集するという問題は、重要語の抽出として長い歴史がある[2),5)]。重要語抽出の問題設定では、通常、ある分野のドキュメントを集めたコーパスが与えられ、その中から、その分野の重要語を抽出することが求められる。つまり、分野DがコーパスS_Dとして与えられ、そこに現れる重要語が、その分野の専門用語となるということである。これを図式として表すと、次のようになる。

> 分野 D = コーパス S_D(によって表されている)
> 分野 D の専門用語 = コーパス S_D に現れる重要語

さて、分野は、通常、それを指す適切な用語が存在する。これを逆方向に考えて、「用語はある分野を指す」と考えてみよう。「自然言語処理」という用語がある分野を指すということには問題はなかろう。「構文解析」、「文脈自由文法」、「アーリー法」のように、指し示す概念がより小さくなってくるに従って、多少の違和感を感じるが、ある特定の小さな領域(分野)を指すという点では、特に問題はなかろう。このような領域を、マイクロドメインと呼ぶことにしよう。

マイクロドメインも、一応、分野であるから、あるコーパスによって代表されうると考えるのは、妥当である。そのコーパスに現れる重要語は、上記の図式から、そのマイクロドメインの専門用語となる。マイクロドメインは、ある用語tによって表される小さな分野であるから、ここでの専門用語は、tの関連用語とみなすことができよう。以上の議論を図式化すると、次のようになる。

> 用語 t = マイクロドメイン D_t = コーパス S_{D_t}
> 用語 t の関連用語
> = マイクロドメイン D_t の専門用語
> = コーパス S_{D_t} の重要語

この図式に従うと、用語tに対して、そのマイクロドメインD_tを代表するようなコーパスS_{D_t}を求めることができれば、そこから、重要語抽出の手法によって、関連用語の候補集合を得ることができるということになる。

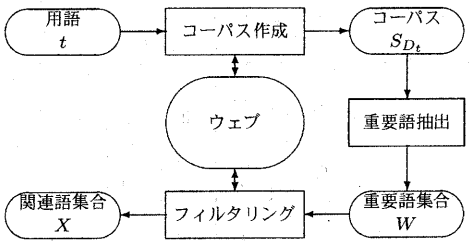

図1 システム構成
Fig. 1 System overview

3. 関連用語収集システム

前節の考え方に沿って、ウェブから関連用語を収集するシステムを作成した。作成したシステムの構成を**図1**に示す。本システムは、(1) コーパス作成、(2) 重要語抽出、(3) フィルタリング、の3つのモジュールから構成される。

3.1 コーパス作成

関連用語収集の第1ステップは、与えられた用語tに対して、それを対応するマイクロドメインD_tを代表するようなコーパスS_{D_t}を作成する処理である。重要語抽出では、コーパスは与えられるものであり、実質的には、人間が注意深く準備するのが普通であった。それに対して、関連用語抽出では、この部分を自動化する必要が生じる。

コーパスをゼロから作成することはできないので、コーパス生成とは、ある与えられたテキスト集合から、その一部を切り出すことになる。ここで、我々は、このテキスト集合としてウェブ文書を利用する。

このコーパスは、関連用語収集の出発点となるので、適切な方法で収集・作成する必要がある。直観的には、コーパスS_{D_t}として、用語tについて書かれたテキストを集めるのがよいことは、ほぼ明らかではあるが、具体的な方法に特に名案はない。そこで、本研究では、用語説明の探索で用いた方法と同一の、次のような方法を用いることとした。

(1) **ウェブページの収集**: 与えられた用語tに対して、「tとは」「tという」「tは」「t」の4種類のクエリを検索エンジンに入力し、得られたURLのそれぞれ上位100ページを入手する。さらに、それらのページに、用語tがアンカーテキストとなっているアンカーが存在する場合は、そのアンカー先ページも入手する。
(2) **文の抽出**:それぞれのページを整形して文に分割し、用語tを(文字列として)含む文のみを抽出する。

ウェブページの収集で「t」以外に3種類のクエリを用

表 1　作成されるコーパス
Table 1　Constructed Corpora

	自然言語処理	形態素解析	構文解析	情報検索	機械学習	同時多発テロ
Goo　t とは	20	21	11	32	4	15
t という	15	14	18	51	4	100
t は	56	41	100	100	12	100
t	100	100	100	100	100	100
Infoseek　t	100	100	100	100	100	100
収集ページ	488	263	300	388	271	461
コーパス (文数)	437	551	805	557	158	531

いるのは、これらの付属語を付加して検索した方が、用語 t を説明する文章が得られやすい、という理由による。検索エンジンとしては、Goo と Infoseek を用いる。但し、Infoseek は、4 種類のクエリに対して同じ答えを返すことが多いので、「t」というクエリのみを用いる。文抽出では、今回は、用語 t を含む文のみを抽出したが、前後 n 文を一緒に抽出するという方法もあり得る。

表 1 に、いくつかの入力用語に対する、収集 URL 数、収集ページ数 (実際に入手できたページ数)、および、コーパスサイズ (文数) を示す。この表からわかるように、コーパスサイズは、おおよそ数百文となる。

3.2　重要語抽出

関連用語収集の第 2 ステップは、第 1 ステップで作成されたコーパスから重要語を抽出する処理である。すでに述べたように、この処理は、いわゆる重要語の抽出 (ATR) として良く研究されている。本システムでは、各種提案されている手法のなかで、日本語に対してよい精度が得られている中川の方法 [6] を採用し、これを一部修正したものを実装した。

中川の方法は、語の造語能力に着目した方法で、「造語能力の高い単語から構成される複合語は、重要語である」という考え方に基づく。この考え方は、前節で検討した我々の関連用語の満たすべき条件とは、直接関係がないため、ジェネレーターの実装法として都合がよい。具体的には、次の 2 つのステップから構成される。

(1)　**候補語リストの作成**：それぞれの文の文節を認識して、名詞を 2 つ以上含む文節を取りだし、その主要部 (付属語等を除いたもの) を集め、候補語のリスト L を作る。

(2)　**候補語の得点付けとそれに基づく選択**：リスト L に含まれる候補語の中から得点の高いもの上位 N 個を取りだし、これを重要語として採用する。なお、現在は $N = 30$ を採用している。

ここで、l 単語からなる候補語 $t = w_1 w_2 .. w_l$ の得点は、次のように計算する。

まず、単語の造語能力を測る指標として、Pre と $Post$ と呼ばれる 2 つの値を導入し、この 2 つの値の相乗平均をとって、単語の造語能力を数値化する。

$$Pre(w, L) = \text{“}L \text{ において、} w \text{ の直前に現れる単語の異なり数”}$$

$$Post(w, L) = \text{“}L \text{ において、} w \text{ の直後に現れる単語の異なり数”}$$

$$ws(w, L) = \sqrt{(Pre(w, L) + 1)(Post(w, L) + 1)}$$

候補語は、一般に複数の単語から構成されるので、ここでも相乗平均をとると、中川の Imp_1 の式が得られる。

$$Imp_1(t, L) = \sqrt[l]{\prod_{i=1}^{l} ws(w_i, L)}$$

我々が用いる得点の定義は、これに出現頻度を加味した、次のような式である。

$$score(t, L) = Imp_1(t, L) \times F(t, L)^{\alpha}$$

$$F(t, L) = \begin{cases} 1 & \textbf{if “}t \text{ が単語 } (l = 1) \text{ の場合”} \\ \text{“}L \text{ における } t \text{ の頻度”} & \\ & \textbf{otherwise} \end{cases}$$

中川のオリジナルの Imp_1 は、用語 t の出現数には全く依存しない☆。しかし、コーパス内の用語の出現数は、その用語がその分野でよく使われるかどうかの一つの指標になっていると考えられる。そこで、$F(t, L)^{\alpha}$ の項を加え、α によって、それをどの程度加味するかコントロールする方法を採用した。予備的な実験に基づき、現在は $\alpha = 0.5$ を採用している。**表 2** の左半分に、「自然言語処理」に対する重要語抽出の結果 (得点、頻度、抽出された用語) を示す。

3.3　語構成による重要語の分類

表 2 を見ると、重要語抽出によって得られた用語、すなわち、用語 t の関連用語の候補は、語構成の観点からいくつかのグループに分類できることに気が付く。これを明確に表現するために、まず、用語 t の前後に総計 i 個の単語を付加することにより構成される単語列の集合を表す記号 $C_i(t)$ を導入する。

$$C_i(t) = \sum_{j=0}^{i} V^j \; t \; V^{i-j}$$

ここで、V は (与えられた) 全単語集合を表す。さらに、以下の記号を定義する。

$$C_*(t) = \sum_{i=0}^{\infty} C_i(t) \;=\; V^* t V^*$$

$$C_+(t) = \sum_{i=1}^{\infty} C_i(t) \;=\; V^* t V^* - t$$

$C_*(t)$ は用語 t を完全に含む単語列の集合を表し、$C_+(t)$ はその集合から t 自身を除いた集合を表す。

☆ 最近の中川の論文 [7] では、我々と同様に、出現数を加味した得点を採用している。

表 2 関連語収集の結果
Table 2 Result of filtering

重要語抽出			フィルタリング					
得点	頻度	抽出された用語 x	タイプ	$H(x)$	$H(t \wedge x)$	$a(x \rightarrow t)$ (%)	$a(t \rightarrow x)$ (%)	採用
138.45	330	自然言語処理	0	3976	-	-	-	
45.30	29	自然言語処理技術	1	592	-	-	-	√
26.14	15	自然言語処理システム	1	107	-	-	-	√
19.75	2	処理技術	3	13882	714	5.14	17.96	√
17.23	-	処理	2	1309586	-	-	-	
17.11	5	自然言語処理研究	1	63	-	-	-	
16.73	3	音声情報処理	3	556	67	12.05	1.69	√
16.26	12	構文解析	4	4323	388	8.98	9.76	√
15.06	3	情報処理学会	3	12636	545	4.31	13.71	√
14.65	8	音声認識	4	14576	499	3.42	12.55	√
13.74	2	音声処理	3	2468	81	3.28	2.04	
13.42	6	研究開発	4	138589	-	-	-	
13.08	12	自然言語処理学講座	1	109	-	-	-	√
13.06	4	情報検索	4	81210	734	0.90	18.46	√
12.14	7	自然言語処理学	1	179	-	-	-	√
12.07	4	自然言語処理分野	1	23	-	-	-	
11.77	5	自然言語処理研究会	1	175	-	-	-	√
11.31	-	技術	4	2092277	-	-	-	
11.30	3	情報処理学会自然言語処理研究会	1	75	-	-	-	
11.04	5	処理過程	3	2150	21	0.98	0.53	
10.78	5	意味解析	4	798	147	18.42	3.70	√
10.78	5	研究分野	4	37776	286	0.76	7.19	√
10.78	2	言語処理系	3	1438	58	4.03	1.46	
9.54	-	情報	4	13476525	-	-	-	
9.49	5	形態素解析	4	1753	394	22.48	9.91	√
9.43	2	音声言語	3	3448	262	7.60	6.59	√
9.25	2	言語解析システム	3	15	-	-	-	
9.19	2	テキスト処理	3	3004	77	2.56	1.94	
8.82	3	自然言語処理研究者	1	8	-	-	-	
8.82	2	計算機言語処理	3	8	-	-	-	

表 3 関連用語のタイプ
Table 3 Classification of related terms

タイプ	定義	説明	例
0	t	与えられた用語それ自身	“自然 言語 処理”
1	$C_+(t)$	t を含む単語列 (t 以外)	“自然 言語 処理 システム”
2	$P_+(t)$	t の部分単語列 (t 以外)	“自然 言語”, “言語 処理”, “自然”, “言語”, “処理”
3	$C_+(P_+(t))$	$P_+(t)$ のいずれかを含む単語列	“言語 処理 系”
4	$V^+ - C_*(P_*(t))$	t の構成単語を含まない単語列	“構文 解析”

次に、用語 t の部分単語列に注目しよう。

$$P_i(t) = \{u | t \in C_i(u)\}$$

$$P_*(t) = \sum_{i=0}^{l-1} P_i(t)$$

$$P_+(t) = \sum_{i=1}^{l-1} P_i(t) = P_*(t) - t$$

$P_i(t)$ は、t から先頭あるいは末尾の単語を総計 i 個取り除くことによってできる単語列である。$P_*(t)$ は t の任意の部分単語列を表す。

ここで、先に導入した関数 C の定義域を用語 (単語列) から用語集合 (単語列の集合) に拡張しよう。すると、ある用語 t に対して、用語 t と共通する部分を持つ単語列の集合は、$C_*(P_*(t))$ と表現できることになる。これをさらに 4 つの部分集合に分割し、表 3 に示すような総計 5 つのグループを考える。これをタイプ 0 からタイプ 4 と名付ける。

このようなタイプを導入する理由は、次のことがほぼ成り立つからである。

(1) タイプ 1 の用語は、元の用語の、広い意味での下位語となる。

(2) タイプ 2 の用語は、元の用語の、広い意味での上位語となる。

これらの事実により、語のタイプ判定に基づく、上位語、下位語の判定が可能となる。

3.4 ウェブのヒット数を用いたフィルタリング

関連用語抽出の最後のステップは、得られた候補の中から条件を満たすものを選ぶフィルタリングの処理である。この処理は、「専門用語」性のチェックと関連性のチェックの 2 つからなり、その両方をパスしたものを、関連用

表 4 Goo のヒット数
Table 4 Estimated pages by Goo

こと	17967090	APEC	15988
人	12558335	万年筆	13901
新聞	1704181	自然言語	10438
JAVA	613378	機械翻訳	5736
概念	311111	構文解析	4570
朝刊	172675	自然言語処理	3798
果物	168137	イージス艦	3292
計算機	88625	テブラ	2438
オブジェクト指向	32734	処理過程	2365
白内障	20288	長期記憶	1306

語として認定する。

3.4.1 「専門用語」性のチェック

すでに 2 節で述べたように、「専門用語として使われている」ことの証拠として、我々は 4 つの証拠を採用することにした。このうち、3 番目の条件 (定義や説明が存在する) は、用語説明探索システムとの統合によって最終的に実現できるため、今回作成したシステムに組み込まなかった。また、4 番目の条件は、本システムを再帰的に動かせばチェックできるようになるため、システム自身には組み込まなかった。すなわち、実際に実装したのは、次の 2 つの証拠の判定である。

(1) 特定の分野で広く、または、それなりに使われている。

(2) 一般語ではない。

これらの証拠の判定法の実装には、いくつかの方法が考えられる。たとえば、(1) は、コーパス S_{D_t} における用語の頻度を数える、というのが常套手段であり、また、(2) は、一般の国語辞典で調べる、というのが常套手段であろう。しかしながら、

- コーパス S_{D_t} における頻度は、すでに候補生成の際に利用しているので、フィルターの実装に再び同じ値を使うのは望ましくない。
- 国語辞典の見出し語との照合は、表記のゆれという些細ではあるがやっかいな問題に直面する

等の理由により、これらの方法を採用しないこととした。

その代りとして、我々が着目したのは、サーチエンジンのいわゆるヒット数 (その用語が出現するウェブページの概算数) である。Goo のヒット数を調べると、おおよそ以下のような傾向が観察される。

一般語 1 万以上のヒットがある。そのほとんどは、10 万以上である。

専門用語 100 から 10 万あたりに分布する。その中心は、1000 から 3 万あたりである。但し、インターネットやウェブ関連の用語は、例外的に多い場合がある。

用語ではない表現 1 万以下に分布する。そのほとんどは、3 千以下。100 以下は、ほとんど確実である。

表 4 に、いくつかの用語に対する Goo のヒット数を示す☆。

☆ ヒット数は、その時々で若干異なる値が表示される。

この観察事実を利用して、上記の証拠の確認を以下のような方法で実装する。

(1) 特定の分野で広く、または、それなりに使われている。⇒ Goo のヒット数が 100 以上。

(2) 一般語ではない。⇒ Goo のヒット数が 10 万以下。

すなわち、ヒット数が 100 未満か 10 万より多い場合は、専門用語ではないと判断して除外する。

3.4.2 関連性のチェック

関連性のチェックは、関連度とその閾値を定義することによって実現する。但し、先にのべたように、タイプ 1 とタイプ 2 の用語は下位語または上位語とみなし、無条件で「関連する」と判定する。

関連度を計算するための基礎となるデータは、事実上、テキストにおける共起以外に選択肢はない。すなわち、我々に残された選択は、共起を計測するテキストとしてどのようなテキストを採用するか、という点に集約される。候補生成で利用したコーパス S_{D_t} をそのまま利用するのは適切ではないので、全ウェブページをこのためのテキストとして利用する。すなわち、用語 t と候補語 x のいわゆるアンド検索のヒット数を基礎データとして採用する。

まず、以下のような記法を定義する。このうち、後者がいわゆるアンド検索のヒット数である。

$H(t)$ = “用語 t が現れるページ数”

$H(t \wedge x)$ = “用語 t と用語 x が共に現れるページ数”

この 2 つの値を用いて、方向性を持った次の 2 つの関連度を定義する。

$$a(x \to t) = \frac{H(t \wedge x)}{H(x)}$$

$$a(t \to x) = \frac{H(t \wedge x)}{H(t)}$$

$a(x \to t)$ は、「x が現れるページにどれくらい t も現れるか」を表したものであり、逆に、$a(t \to x)$ は、「t が現れるページにどれくらい x も現れるか」を表したものである。これらの値のいずれかがある閾値 Z より大きい場合、2 つの用語は関連していると判断する。本システムでは、$Z = 5\%$ を採用する。表 2 の右半分に、「自然言語処理」の候補語に対するフィルタリングの結果を示す。

なお、この 2 つの指標の大小関係は、関連のタイプとの間に、次のような傾向が見られる。

(1) $a(x \to t) \ll a(t \to x)$: x は t の上位語である場合が多い。

(2) $a(x \to t) \gg a(t \to x)$: x は t の下位語である場合が多い。

4. 実験と検討

4.1 実 験 1

作成したシステムを用いて関連用語を収集する実験を行った。入力用語として 10 個の用語を用い、それぞれの用語から得られた関連用語が妥当であるかを評価した。

表 5 実験 1 の結果
Table 5 Result of Experiment 1

用語 (t)	$H(t)$	コーパス		得られた関連用語数 (タイプ別)				計	不適切な用語
		ページ	文	1	2	3	4		
自然言語処理	3976	488	437	4/1	0/0	3/1	5/1	12/3	処理技術, 自然言語処理講座, 研究分野
形態素解析	1753	263	551	3/0	0/0	2/0	3/0	8/0	
構文解析	4323	300	805	3/0	0/0	5/0	2/0	10/0	
機械翻訳	5443	296	891	4/0	0/0	9/0	0/0	13/0	
情報検索	81210	388	557	10/1	0/0	2/0	0/0	12/1	情報検索演習
クイックソート	1035	274	471	0/0	1/0	0/0	5/1	6/1	ソート済み
機械学習	1090	271	158	0/0	0/0	4/0	4/0	8/0	
白内障	16302	400	1797	6/0	0/0	0/0	5/1	11/1	手術結果
ハロープロジェクト	4540	311	174	0/0	0/0	0/0	7/1	7/1	所属事務所
同時多発テロ	61076	461	531	11/2	0/0	6/0	5/0	22/2	同時多発テロ事件以来, 同時多発テロ発生
計				41/4	1/0	31/1	36/4	109/9	

妥当性の判定では、主に既存の専門用語辞典やウェブテキスト等を利用し、最終的に 2 名の著者が相談して☆、適切か不適切かの 2 段階の評価を下した。

実験結果を**表 5** に示す。それぞれの欄は、"適切な用語の数/不適切な用語の数" を表す。この実験において、関連用語として収集された 118 個のうち、109 個 (92%) が適切と判定された。このことにより、当初想定した典型的な専門用語に対して、提案手法が機能することが確認できた。また、「ハロープロジェクト」や「同時多発テロ」等の当初想定していなかった入力に対しても、ほぼ妥当な関連用語を見つけることができた。

一方、不適切な出力には、(i)「講座」や「演習」といった名詞が接続して作られる複合語、(ii)「手術の結果 → 手術結果」のように「の」などの機能語が省略されて作られる見かけ上の複合名詞、の 2 つの場合が観察される。これらを排除する方法として、末尾の名詞に対するストップワードリストを作成して使用する方法が考えられる。

4.2 実 験 2

次に、ある用語 t から出発して関連用語 (第 1 世代) を求め、さらに、見つかった関連用語に対してさらに関連用語 (第 2 世代) を見つける実験を行なった。但し、関連用語の収集を再帰的に繰り返すと、用語 t との関連が薄れていく方向に進む場合があるので、第 2 世代を見つける際のコーパス収集時には、元の用語 t が存在するという条件を付加して、ウェブページを収集する。たとえば、「自然言語処理」からスタートして、関連用語として見つかった「構文解析」の関連用語を見つけようとする場合は、「自然言語処理 ∧ 構文解析とは」などでウェブページを検索する。

実験では、出発用語として「自然言語処理」を用いた。既に表 2 に示したように、「自然言語処理」に対して関連用語を収集すると、15 個 (内 3 個は不適) の関連用語が得られた (第 1 世代)。この 15 個のそれぞれに対して関連用語の収集をもう一度適用したところ、総計で 171 個、異なりで 86 個の関連用語が得られた (第 2 世代)。

☆ 評価が割れたのは、118 個中 5 個であった。

第 1 世代に不適切な用語が含まれることは避けられないため、残念ながら、第 2 世代をそのまま拡大化された関連用語集合とすることはできない。しかしながら、第 2 世代を求めることにより、第 1 世代の再評価が可能となる。いま、第 1 世代に属する 2 つの用語 x_i と x_j に対して、「x_i の関連用語に x_j が含まれ、かつ、x_j の関連用語に x_i が含まれる場合、x_i と x_j の間に枝を書く」こととすると、第 1 世代の用語集合のグラフ表現が得られる。このグラフにおいて、枝は関連を表すので、枝が密となっている部分 (完全グラフ) は、その中核的部分と考えることができる。

第 2 世代を求めることによって得られる、「自然言語処理」の第 1 世代に対するグラフ表現を、**図 2** に示す。このグラフの最も大きな完全グラフは、{ 形態素解析, 構文解析, 意味解析, 意味処理 } である。これらの 4 語が「自然言語処理」の最も中心的な用語であると推定される。次に大きな完全グラフは、{ 構文解析, 意味解析, 自然言語処理システム }、{ 構文解析, 自然言語処理システム, 自然言語処理技術 }、{ 自然言語処理システム, 自然言語処理技術, 情報検索 }、{ 自然言語処理技術, 自然言語処理研究会, 情報処理学会 } の 4 つである。ここで出てきた 5 語が次に重要な用語となる。残りの 6 語のうち、他の語と連結しない「研究分野」は、1 回目の関連用語収集では関連用語と判定されたが、このグラフからは、関連性は低いと判定される。

4.3 検 討

今回行なったのは、作成したシステムが設計どおり動作するかを調べる予備的な実験であるが、この範囲においては、システムはうまく動作することが確認できた。次のステップでは、より多様な用語に対して、関連用語をうまく見つけることができるかどうかを調査する必要がある。また、今回は、収集された関連用語の精度のみを調査したが、重要な関連用語がもれなく収集されているか (カバレジ) についても、調査が必要である。

提案手法において、現在までに判明している問題点には、以下のものがある。

- カタカナ表記の専門用語をうまく見つけることがで

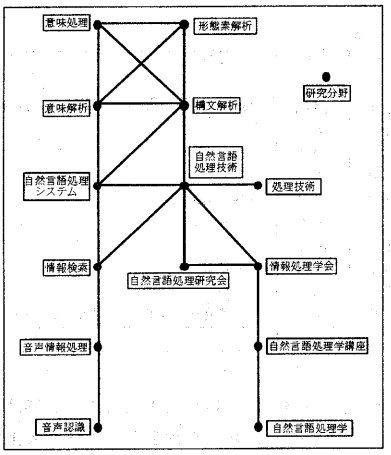

図 2 「自然言語処理」の第 1 世代のグラフ
Fig. 2 Graph of first generation of "natural language processing"

きない。

- 頭文字表記の専門用語をうまく見つけることができない。

この問題は、重要語抽出における問題である。現在の重要語抽出の方法は、複合名詞を対象とした方法であるが、カタカナ表記の複合名詞は、形態素解析プログラムによってひとつの名詞と判定されるため、その得点があまり高くならない。この根本原因は、使用している形態素解析プログラム Juman がカタカナ列を単語に分割しないことにあるが、現実的にはどのような方法を用いてもカタカナ語の単語分割の精度はあまり期待できないため、重要語抽出アルゴリズムにおいて何らかの対処が必要と考えている。また、頭文字表記（いわゆるアクロニム）の専門用語も、形態素解析プログラムで 1 語となるため、上記と同じ問題を抱える。これに対しても対処が必要である。

4.4 関連研究

本研究と最も関連が深いのは、重要語抽出の研究 [2),5)] である。重要語抽出では、出発点となるコーパスが与えられ、そこに含まれる専門用語（重要語）を網羅的に抽出することが求められる。しかしながら、本研究では、このコーパスが与えられない（動的に生成しなければならない）点と、網羅的ではなく、非常に関連が深い代表的な用語のみを収集する点が大きく異なる。

コーパスから関連する語のペアを抽出することは、これまでコーパスからの知識獲得の一貫として、多くの研究がある [8)]。しかし、これらの研究においても、「まずコーパスありき」であり、どのようなコーパスを使用すべきかということについては、全く検討されていない。関連する語のペアを見つけるための指標（関連度）に関しては、補完類似度 [9)] など、各種の指標が提案されている [8)]。これらの指標を用いることも可能であるが、関連性の強さを確認するという用途には、我々の用いた非常に簡便な指標で十分であると考える。

謝辞 本研究の一部は、科学研究費補助金特定領域研究 (2)「ウェブを情報源とした用語辞典の自動編集」(課題番号 14019050) によって実施した。

参考文献

1) 桜井裕, 佐藤理史: ワールドワイドウェブを利用した用語説明の自動生成, 情報処理学会論文誌, Vol. 43, No. 5, pp. 1470–1480 (2002).
2) Kageura, K. and Umino, B.: Methods of automatic term recognition: A review, *Terminology*, Vol. 3, No. 2, pp. 259–289 (1996).
3) 影浦峡: 「専門用語の理論」に関する一考察, 情報知識学会誌誌, Vol. 12, No. 1, pp. 3–12 (2002).
4) Aitchison, J. and Gilchrist, A.: *Thesaurus Construction: A Practical Manual, 2nd Edition*, Aslib, The Association for Information Management (1987).
5) Kageura, K. and Koyama, T.: Special issue: Japanese term extraction, *Terminolgy*, Vol.6, No.2 (2000).
6) Nakagawa, H.: Automatic term recognition based on statistics of compound nouns, *Terminology*, Vol. 6, No. 2, pp. 195–210 (2000).
7) Nakagawa, H. and Mori, T.: A simple but powerful automatic term extraction method, *Computerm2: second workshop on computational terminology*, pp. 29–35 (2002).
8) Manning, C. D. and Schütze, H.: *Fundations of Statistical Natural Language Processing*, The MIT Press (2002).
9) 山本英子, 梅村恭司: コーパス中の一対多関係を推定する問題における類似尺度, 自然言語処理, Vol. 9, No. 2, pp. 45–75 (2002).