ウェブを利用した関連用語の自動収集

佐々木 靖弘[†] 佐藤 理史[‡]

[‡]京都大学工学部電気電子工学科
[†]京都大学大学院情報学研究科知能情報学専攻
sasaki@pine.kuee.kyoto-u.ac.jp, sato@i.kyoto-u.ac.jp

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、その用語が他のどのよう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、その分野の中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、始めて意味を持つ。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。

我々はこれまで、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「関連用語収集システム」を実現してきた[1]。今回は、このシステムを用いて、ジャンルの異なる5つの分野の用語に対しての関連用語収集実験を行ったので、その結果について報告する。

2. 関連用語収集システム

我々が作成した関連用語収集システムは、(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。本システムの構成図を図1に示す。

2.1 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 が指し示す分野 D_t を代表するコーパス S_{D_t} を作成する処理である。

本システムでは、ウェブを利用した次のような方法でコーパスを作成する。

(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」という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。

サーチエンジンとしては、Goo と Infoseek を用いる。文抽出では、用語 t を含む文のみを抽出しているが、前後 n 文を一緒に抽出するという方法も考えられる。

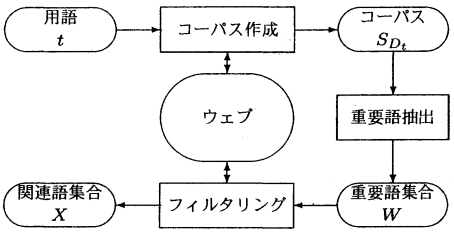

図1 システム構成

2.2 重要語抽出

関連用語収集の第2ステップは、第1ステップで作成したコーパスから重要語を抽出する処理である。この処理は、いわゆる重要語の抽出 (ATR) としてよく研究されている[2),3)]。本システムでは、各種提案されている手法の中で、日本語に対してよい精度が得られている中川の方法[4)] を採用し、これを一部修正したものを実装した。

この方法は、次の2つのステップからなる

(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 に含まれる候補語を、その構成単語の造語能力に基づき得点付けし、得点の高いもの上位 N 個を取りだして、これら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。

表1の左半分に、「再生医療」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。

2.3 語構成による重要語の分類

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、語構成の観点から次の5つのタイプに分類できる。

タイプ0 与えられた用語それ自身
タイプ1 t を含む単語列 (t 以外)
タイプ2 t の部分単語列 (t 以外)
タイプ3 タイプ2のいずれかを含む単語列
タイプ4 t の構成単語を含まない単語列

表 1　関連用語収集の結果

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	$H(x)$	$H(t \wedge x)$	$a(x \to t)(\%)$	$a(t \to x)(\%)$	採用
523.31	531	再生医療	0	-	-	-	-	
76.36	11	再生医療研究	1	190	-	-	-	√
63.20	11	再生医療技術	1	124	-	-	-	√
47.12	19	再生医療分野	1	137	-	-	-	√
44.72	25	幹細胞	4	11940	1710	14.32	30.95	√
40.80	20	臓器移植	4	25536	788	3.09	14.26	√
38.71	4	医療技術	3	35099	677	1.93	12.25	√
37.27	5	移植医療	3	3392	281	8.28	5.09	√
35.85	11	再生医療ビジネス	1	56	-	-	-	
35.53	6	再生医療関連	1	70	-	-	-	
34.88	34	ES 細胞	4	4358	1166	26.76	21.10	√
32.30	6	研究開発	4	138589	-	-	-	
31.61	2	細胞医療	3	122	50	40.98	0.90	√
28.45	14	日本再生医療学会	1	211	-	-	-	√
27.93	0	医療	2	1213134	-	-	-	
24.47	16	人工臓器	4	5432	339	6.24	6.14	√
24.27	4	再生医療事業	1	42	-	-	-	
23.66	0	研究	4	1880768	-	-	-	
23.42	7	再生医学	3	2639	695	26.34	12.58	√
23.11	3	細胞移植	4	5262	456	8.67	8.25	√
23.02	2	臓器再生医療	1	14	-	-	-	
22.95	3	臓器移植医療	3	334	17	5.09	0.31	√
21.28	3	再生医学研究	3	90	-	-	-	
20.26	3	ゲノム医療	3	330	65	19.70	1.18	√
18.47	0	再生	2	570255	-	-	-	
18.11	8	遺伝子治療	4	13684	1130	8.26	20.45	√
17.89	0	細胞	4	300699	-	-	-	
17.61	2	血管再生医療	1	36	-	-	-	
17.58	2	再生臓器	3	69	-	-	-	
17.31	1	再生医療研究開発	1	6	-	-	-	

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。

すなわち、タイプ 1 およびタイプ 2 の用語は、「t と関連する」とみなすことができる。

2.4　ウェブのヒット数を用いたフィルタリング

関連用語収集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、関連用語として採用する。

2.4.1　「専門用語」性のチェック

「専門用語として使われている」ことの条件として、本システムでは、次の 2 つの条件の判定方法を実装している。

(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。

これらの条件の判定方法として、我々は、ウェブのサーチエンジンのいわゆるヒット数を利用する。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。ただし、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。

この観察事実を利用して、上記の条件の判定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。

2.5　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。ただし、先に述べたように、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。

本システムでは、サーチエンジンにおけるアンド検索のヒット数を用いて用語間の関連度を計算する。

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。

表 2　システムに与えた入力用語

分野	入力用語
自然言語処理	自然言語処理, 有限オートマトン, 構文解析, 形態素解析, 意味解析, 格文法, 文脈解析, 照応関係, 情報検索, 機械翻訳
日本語	シソーラス, 漢字制限, 形態論, 言語教育, 言語行動, 言語遊戯, 構文論, 混種語, 常用漢字, 類義語
情報科学	工業所有権, 字句解析, 論理演算,QC7 つ道具, 新 QC7 つ道具, バッチファイル, フラクタル, cdmaOne,Ethernet, グローバルアドレス
時事用語	再生医療, グリッド・コンピューティング, 電子ペーパー, 富士山ハザードマップ, コーポレートガバナンス, 京都議定書, ポストゲノム, バイオメトリクス, 道路公団民営化, 偽装牛肉事件
歴史上の人物	卑弥呼, 聖徳太子, 平清盛, 源頼朝, 足利尊氏, 織田信長, 豊臣秀吉, 徳川家康, 坂本竜馬, 伊藤博文

表 3　分野ごとの関連用語収集結果

分野	得られた関連用語数		
	適切	不適切	計
自然言語処理	101 [93%]	8 [7%]	109
日本語	71 [81%]	17 [19%]	88
情報科学	113 [88%]	15 [12%]	128
時事用語	106 [91%]	10 [9%]	116
歴史上の人物	128 [76%]	41 [24%]	169
計	519 [85%]	91 [15%]	610

$H(t)$ = “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 1 の右半分に、「再生医療」の候補語に対するフィルタリングの結果を示す。

3. 実験と検討

3.1 関連用語収集実験

作成したシステムを用いて、関連用語収集実験を行った。入力として、ジャンルの異なる 5 つの分野からそれぞれ 10 語ずつ、総計で 50 語の用語を与え、システムによって得られた用語が関連用語として適切であるかどうかを人手で判断した。システムに与えた入力用語を表 2 に、関連用語収集結果を表 3 に示す。

また、50 語の入力用語それぞれに対して、重要な関連用語を数語設定し、それらの用語が実際に収集できているかどうかを評価した。具体的には、設定した関連用語それぞれに対して、次の 5 種類のいずれに該当するかを調べた。

表 4　分野ごとの重要関連用語収集結果

分野	収集成功	判定失敗	抽出失敗	作成失敗	検索失敗	計
自然言語処理	6	3	14	11	8	43
日本語	7	0	19	5	1	32
情報科学	10	5	27	13	0	55
時事用語	2	0	13	19	5	39
歴史上の人物	18	0	23	1	0	42
計	43	8	96	49	14	210

収集成功　関連用語として収集できた。

判定失敗　重要語抽出において上位 30 位以内には選ばれたが、関連用語としては収集されなかった (フィルタリングにおいて関連用語ではないと判定された)。

抽出失敗　コーパス作成において作成されたコーパスには含まれるが、重要語抽出で上位 30 位に入らなかった。

作成失敗　コーパス作成において収集したウェブページには含まれるが、作成されたコーパスには含まれなかった。

検索失敗　収集したウェブページに含まれなかった。

その結果を表 4 に示す。

3.2 検　　討

表 3 に示したように、入力用語 50 語に対して本システムが関連用語として出力した 610 語中、519 語 (85%) が適切であった。この結果より、本システムが十分に機能することが確かめられた。一方、分野別に見ると、歴史上の人物を入力した場合、得られた用語が関連用語として不適切だと判断されるものが、他と比べてかなり多かった。これらの不適切な用語には、「豊臣秀吉」に対する「豊臣」や「秀吉」など、姓と名が分断された用語が多数含まれる。これらは、用語としては不完全ではあるが、システムにとっての致命的な誤りではないと思われる。

次に、重要な関連用語の収集能力について検討する。表 4 に示したように、入力用語 50 語に対して設定した重要関連用語 210 語中、関連用語として収集できたのは 43 語 (20%) にとどまった。全体の半数近い 96 語は、作成したコーパスには含まれながら、重要語抽出において上位 30 語に含まれなかった。この原因として、次の 2 つの原因が考えられる。

(1) **形態素解析における単語分割の問題**

本システムで用いている重要語抽出方法は、「造語能力の高い単語から構成される複合語ほど重要語である」という考え方に基づく。そのため、単語分割が正しく行なわれていることが、重要語抽出の前提条件となる。しかし、現在使用している形態素解析器 Juman は、特にカタカナ語や頭文字表記の用語などを正しく単語に分割することができない。このため、これらの用語は高得点とならないという問題が発生する。

(2) **作成されるコーパスの問題**

重要語抽出で抽出される用語を観察すると、「造語能力の高い一部の単語から構成される複合語」ばかりが高得点になる傾向があることが判明した。これは、作成されるコーパスが、対象とするマイクロドメインをまんべんなく表しているのではなく、偏りが生じているためであると考えられる。

(i) の問題に対しては、形態素解析器を改良することや、1 単語の用語に対する得点付けの方法を工夫する必要がある。また、(ii) の問題に対しては、2.1 節でも述べたように、コーパス作成時に対象とする文の前後 n 文も一緒に抽出するという方法が考えられる。また、今回実装した方法とは別の方法で関連用語の候補を生成することも考えられる。例えば、本システムでは、ウェブページにおいて用語がどのように記述されているかを全く考慮していない。しかし、ウェブページにおいて、見出し語的に記述されている用語や他のページへのアンカーテキストとなっている用語等は重要語である可能性が高い。そのような用語を関連用語の候補に加えるという方法も有効であると考えられる。

たとえ、重要語として抽出されたとしても、次のフィルタリングによって棄却されるならば、最終的な収集能力は向上しない。そこで、重要語抽出に失敗した 96 語に対して、もし重要語として抽出されたならば、フィルタリングによって関連用語と判定されるのかどうかを調べた。その結果、96 語中 74 語が関連用語と判定された。よって、これらの用語を関連用語の候補として抽出することができれば、重要関連用語収集能力が大幅に向上することが期待できる。

3.3 ま と め

今回の実験により、我々の作成した関連用語収集システムは、かなり高い精度で、正しい関連用語を収集することが確かめられた。しかし、現時点では、重要な関連用語をもれなく収集できているとは言いがたい。これらの用語を収集できるようにするためには、特に、コーパス中に存在する重要な関連用語を、造語能力とはまったく別の観点に基づいて抽出する方法を検討する必要がある。

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) 佐藤理史, 佐々木靖弘: ウェブを利用した関連用語の自動収集, 情報処理学会研究報告 NL-153-8, pp. 57–64 (2003).
2) Kageura, K. and Umino, B.: Methods of automatic term recognition: A review, *Terminology*, Vol. 3, No. 2, pp. 259–289 (1996).
3) Kageura, K. and Koyama, T.: Special issue:Japanese term extraction, *Terminology*, Vol. 6, No. 2 (2000).
4) Nakagawa, H.: Automatic term recognition based on statistics of compound nouns, *Terminology*, Vol. 6, No. 2, pp. 195–210 (2000).