ウェブを利用した専門用語集の自動編集

佐々木　靖弘　　佐藤 理史　　宇津呂　武仁

京都大学大学院 情報学研究科
sasaki@pine.kuee.kyoto-u.ac.jp, {sato, utsuro}@i.kyoto-u.ac.jp

1. はじめに

本研究は、ウェブを利用して特定分野の専門用語集を自動的に編集することを実現することを目的とする。用語集の編集という問題は、ある目的 (編集方針) を持った用語の集合を設定する問題であり、特定の分野に関する文書からその分野の重要語を抽出する重要語抽出問題とは本質的に異なる。重要語抽出問題では、抽出される用語のそれぞれが対象文書において重要であることが求められるが、専門用語集編集問題においては、見出し語として選定された用語集合が、**集合として価値がある**ことに意義がある。

専門用語の自動編集は、以下のような手順で行う。これは、実際の人手による専門用語集の編集方法をモデルとしている [1),2)]。

(1) 専門用語集の編集方針を決定する。ここでは、どのような用語を見出し語とするかや、用語集のサイズなどを定める。

(2) (1) で決定した編集方針により、見出し語のそれぞれが満たすべき要求仕様が決定する。そのような仕様を満たす用語を、見出し語の候補語として収集する。このとき、収集する用語は、最終的に作成される用語集のサイズより十分に多くなるようにする。

(3) (1) で決定した編集方針により、見出し語集合が全体として満たすべき要求仕様が決定する。(2) で収集した候補語の中から、そのような仕様を満たす見出し語集合を作成する。

以下、本論文では、2 節で専門用語集の編集方針と、それによって定まる見出し語の性質について説明する。次に、3 節では見出し語の候補語の収集方法について述べ、4 節では見出し語の選定方法について説明する。

2. 専門用語集の編集方針

専門用語集を編集するには、まず最初に、誰のための、そして何のための用語集を編集するのかという、用語集の編集方針を決めなければ、具体的な編集作業に取り掛かることができない。編集方針を決定することにより、どのような用語を見出し語とするかや、見出し語の数、用語集の全体像などが見えてくる。例えば、ある専門分野についてあまり詳しくない人が、その分野について手っ取り早く理解することができるような用語集を作るのであれば、用語集のサイズはそれほど大きくなく、収集すべき見出し語はその分野を代表するような用語に限定されることになるだろう。一方、専門分野の専門家がリファレンス・ツールとして利用できるような用語集を作るのであれば、用語集のサイズはずっと大きくなり、見出し語として、分野の代表的な用語だけではなく、より専門的な用語を選ぶことが必要となろう。

「専門用語集の自動編集を実現する」という本研究の目的からすれば、あらゆる編集方針に応じて、望むがままの専門用語集を編集できるようになることが理想であるが、いきなり全てに対応するのは困難である。そこで本論文では、用語集の編集方針を、**ある専門分野 $\mathcal{D}$ に関して、その分野の初心者が、分野の全体像が把握できる**ような用語集を編集することに定める。

さて、用語集の見出し語集合を $H = \{h_1, h_2, \cdots, h_N\}$ (N は用語集のサイズ) とすると、H が満たすべき性質は、次の 3 つの性質に分類される。

(i) 個々の見出し語 h_i が満たすべき性質
(ii) 見出し語集合 H の部分集合が満たすべき性質
(iii) 見出し語集合 H 全体が満たすべき性質

以下では、上記の編集方針の場合の、これらの性質について検討しよう。

(i) 個々の見出し語 h_i が満たすべき性質

専門分野 $\mathcal{D}$ の専門用語集を編集するのであるから、個々の見出し語 h_i が満たすべき性質としては、**分野 $\mathcal{D}$ に関連し、専門用語である**ことが求められる。

(ii) 見出し語集合 H の部分集合が満たすべき性質

H の部分集合が満たすべき性質としては、H 中の適当な部分集合の見出し語どうしを見比べたときに、**概念・表現のレベルで粒度が揃っている**ことが求められる。これは、編集方針によって定まる性質というよりは、有用な用語集に一般に求められる性質である。例えば、「自然言語処理」分野の用語集で、「形態素解析」が見出し語となっているときに、「構文解析」が見出し語とならず、「構文解析器」が見出し語となっているような用語集は、用語の選び方に統一感がなく、良い用語集とは言えない。

(iii) 見出し語集合 H 全体が満たすべき性質

分野の初心者が分野の全体像を把握するための用語集を編集するという方針により、H 全体が満たすべき性質としては、H **のサイズ N はそれほど大きくないサイズ (N=20〜50)** となり、H **全体で対象分野 $\mathcal{D}$ をカバーす**

る必要がある。

上の (i) から (iii) の性質の内、(i) に関しては、他の見出し語に関係なく、個々の見出し語 h_i に対して、求められる性質を満たしているかどうかをチェックすることができる。一方、(ii) と (iii) の性質に関しては、いくつかの見出し語候補の中から、(ii) と (iii) の性質を満たす最適な見出し語集合 H を選択する必要がある。

よって、専門用語集の編集法としては、まず十分に多い数の (i) の性質を満たす見出し語候補を収集 (**見出し語候補の自動収集**) し、そこから (ii) と (iii) の性質を満たす見出し語集合を選択 (**見出し語の自動選定**) する。すなわち、

- **見出し語候補の自動収集**では、対象分野 $\mathcal{D}$ に関連する専門用語を十分な数 kN (k=2〜4) 収集する。
- **見出し語の自動選定**では、収集した用語の中から、分野全体をカバーし、概念・表現のレベルで粒度が揃った用語を N 語選定する。

3. 見出し語候補の自動収集

3.1 関連用語収集システム

前節で定めた用語集の編集方針に基づき、まず専門用語集の見出し語の候補となる用語を収集する。これには、我々がこれまでに作成してきた「関連用語の自動収集」システム [3] を利用することができる。関連用語収集システムでは、入力として専門用語を与えると、その専門用語に**関連する専門用語**をウェブから自動的に収集する。これは、本研究で作成する専門用語集の候補語を収集するのに適している。

関連用語収集システムの概要図を図 1 に示す。関連用語の自動収集は、以下の手順で行う。

1. **コーパス作成**　入力用語 s をサーチエンジンのクエリとして入力し、ウェブから s に関するコーパス C_s を作成する。
2. **候補語抽出**　コーパス C_s から名詞および複合名詞を抽出し、関連用語の候補語集合 X を得る。
3. **フィルタリング**　サーチエンジンのヒット数を利用した関連度 $R_{\wedge/\vee}(s,x)$ を用いて、s の関連用語の集合 T を得る。

$$R_{\wedge/\vee}(s,x) = \frac{H(s \wedge x)}{H(s \vee x)}$$

$H(s \wedge x)$：s と x の AND 検索のヒット数
$H(s \wedge x)$：s と x の OR 検索のヒット数

本研究では、サーチエンジンとして goo* を利用している。

関連用語収集システムに「自然言語処理」を入力したときに収集される関連用語を表 1 に示す。入力用語によってバラツキがあるが、関連用語収集システムでは、90%前後の精度で 10 語〜40 語程度の関連用語を収集することができる。

* http://www.goo.ne.jp/

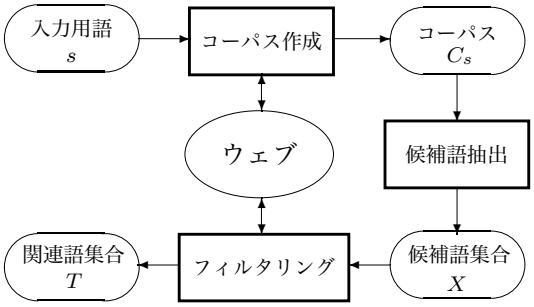

図 1　関連用語収集システムの概要図

表 1　「自然言語処理」の関連用語

$R_{\wedge/\vee}$	関連用語
0.504	言語処理
0.488	自然言語
0.097	形態素解析
0.095	自然言語処理技術
0.093	形態素
0.073	コーパス
0.066	機械翻訳
0.063	言語処理学会
0.059	構文解析
0.049	言語理解
0.043	言語情報
0.042	人工知能
0.038	自然言語処理研究会
0.037	意味解析
0.034	知識ベース
0.032	知識処理
0.030	言語理論

3.2 収集用語数の拡大

見出し語候補の収集では、見出し語の候補となる用語を最終的な見出し語数の 2〜4 倍程度収集することが求められる。本研究で定めた編集方針から必要となる見出し語数は 20〜50 語であるため、見出し語候補としては 100 語前後収集したい。これには、単純に関連用語収集システムを用いただけでは不十分である。

そこで、収集する関連用語の数を増やす必要がある。これは「関連用語の関連用語は関連用語である」という考えに基づき実現する。具体的には以下のような手順で行う。

ステップ 1　入力用語 s の関連用語を収集する。

ステップ 2　ステップ 1 で収集された s の関連用語の中から、関連度 $R_{\wedge/\vee}$ が高い順に、s と語構成的に包含関係がなく、サーチエンジンでのヒット数が s のヒット数の 3 分の 2 以下の用語を n 語選択する。選択した用語を $s_{c_1}, s_{c_2}, \cdots, s_{c_n}$ とする。

ステップ 3　関連用語収集システムのコーパス作成において、$s_{c_1}, \cdots, s_{c_n}$ のそれぞれの用語と s との AND 検索をクエリとしてコーパスを作成し、作成したコーパスを一つにまとめて、拡大コーパス C_s^+ を作成する。

ステップ 4　C_s^+ から関連用語の候補語集合 X^+ を抽出する。

ステップ 5　s および $s_{c_1}, s_{c_2}, \cdots, s_{c_n}$ のそれぞれの用語と候補語との関連度 $R_{\wedge/\vee}$ を計算し、関連度上位のものを全て関連用語とし、見出し語候補集合 T^+ を得る。

表 2 関連用語収集数の拡大

入力用語	単純収集用語数	拡大収集用語数
自然言語処理	23	87
数理計画法	32	62
情報理論	38	110
論理回路	48	103
認知科学	18	58
言語理論	47	91
パターン認識	28	90

ステップ 2 において、s と語構成的に包含関係がある用語とは、例えば $s =$「自然言語処理」のときの「自然言語」や「自然言語処理技術」などである。また、ヒット数が s のヒット数の 3 分の 2 以下の用語を選ぶ理由は、一般にヒット数が大きいほどその用語が表す概念も大きく、ヒット数が小さければその用語が表す概念はより専門的であると考えられるので、より専門的な用語を選択するためである。n として何語選択するのかには検討の余地があるが、現在は $n = 5$ としている。

表 2 に、「自然言語処理」、「数理計画法」、「情報理論」、「論理回路」、「認知科学」、「言語理論」、「パターン認識」を単純に関連用語収集システムに入力としたときの関連用語数と、本節の関連用語拡大手法を用いて収集された関連用語数を示す。

4. 見出し語の自動選定

4.1 選 定 手 順

次に、収集された見出し語の候補から見出し語集合を選定する処理を行う。2 節で述べたように、ここでは対象とする専門分野全体をカバーし、概念・表現のレベルで粒度が揃った用語を N 語選定する。

いかにして分野全体をカバーするように用語を選ぶかであるが、ここで専門分野の構造というものを考える。専門分野は、おおよそ図 2 のような構造をしていると考えられる。すなわち、分野の中にそれを構成するいくつかのサブ分野が存在する。例えば、「自然言語処理」分野には「構文解析」や「形態素解析」などのサブ分野が存在する。そして、これらのサブ分野を表す用語は、それぞれのサブ分野で用いられる用語との関連が深いはずである。

そこで、より多くの異なる見出し語候補と関連している用語は分野の代表的な用語であると考え、そのような用語を見出し語として選定する。具体的には、以下の手順で行う。

1. 見出し語候補集合 T^+ を収集する際に作成したコーパス C_s^+ のそれぞれの文において、候補語のそれぞれが共起するその他の候補語の異なり数 ($codf$) をカウントする。
2. $codf$ が最も小さい候補語を T^+ から取り除く。
3. 1、2 の手順を、候補語数が $N + \alpha$ になるまで繰り返す。

手順 3 において、見出し語数 N に α を加えているのは、ここからさらに、見出し語集合の集合としてのバランスを考えて不必要な用語を取り除くためである。例えば、見出し語数が 30 語程度の「自然言語処理」用語集において、「形態素解析」が見出し語となっている場合に、「形態素解析器」を見出し語に含む必要はない。ただし、「形態素」が見出し語となっているからといって、「形態素解析」を見出し語集合から取り除くわけにはいかない。そこで、ある用語 a(例：形態素解析) および、a になにがしかの語が付属した用語 a'(例：形態素解析器) が見出し語集合に存在した場合、a および a' のコーパス C_s^+ における頻度を cf_a および $cf_{a'}$、ウェブにおけるヒット数を $hits_a$ および $hits_{a'}$ とすると、

$$cf_{a'} < 0.1 \times cf_a \quad \text{かつ} \quad hits_{a'} < 0.1 \times hits_a$$

のとき、a' を見出し語集合から取り除くこととする。

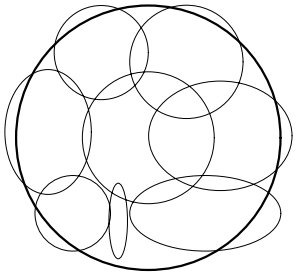

図 2 専門分野の構造

4.2 見出し語集合の構造化

以上の手順により、見出し語の選定を行うことが可能となったが、選ばれた用語を無秩序に提示するだけでは、初心者が専門分野の全体像を把握するのに適した用語集であるとは言えない。そこで、図 2 のような専門分野の構造が見て取れるように、見出し語集合を構造化することを考える。

これを実現するために、**「とは」ヒット数**を利用する。これは、用語に「〜とは」という付属語を付加してサーチエンジンで検索したときの、ウェブにおけるヒット数である。「〜とは」は、用語を説明する際によく用いられる言葉であり、「とは」ヒット数が大きい用語は、ウェブにおいて数多くの説明文が存在する用語であると考えることができる。多くの説明文が存在するということは、逆に言えば、その用語が説明されるべき用語であるということを意味する。

見出し語集合の構造化は、「とは」ヒット数 $toha$、見出し語の選定に用いた共起異なり見出し語数 $codf$、および、用語 x から用語 y への関連度 $R_{x \to y} (= H(x \wedge y)/H(x))$[3] を利用して、以下の手順で行う。

1. 見出し語集合の中で、$toha$ および $codf$ が上位の用語から順に見ていき、両方の値で上位にランクする用語 m 語をサブ分野代表語 $sd_1, sd_2, \cdots, sd_m$ とする。
2. サブ分野代表語に選ばれなかった用語 ot から、$sd_1, \cdots, sd_m$ のそれぞれの用語への関連度 $R_{ot \to sd_i}$ を求め、$R_{ot \to sd_i}$ の値が最も高いサブ分野代表語 sd_i に ot を所属させる。

表 3 「自然言語処理」用語集

大項目	小項目
コーパス	曖昧性 対話システム 自動獲得 自然言語 言語知識 言語情報 言語処理 機械学習 学習アルゴリズム 意味理解 意味構造
構文解析	文脈自由文法 文生成 知識表現 意味表現 意味処理
形態素解析	品詞情報 入力文 日本語文 形態素 データマイニング
機械翻訳	人工知能 自然言語理解 言語理解 機械翻訳システム
意味解析	文脈解析

表 4 見出し語に選ばれなかった用語

ChaSen HPSG JUMAN NTCIR SVM
ニューラルネットワーク パターン認識 意味関係 意味情報
音声言語処理 解析エンジン 概念辞書 各単語
確率モデル 学習手法 機械学習アルゴリズム
機械学習技術 機械学習手法 帰納学習
帰納論理プログラミング 形態素解析システム
形態素解析ツール 形態素解析器 計算言語学 決定木
言語モデル 言語解析 言語処理学会 言語理論
語用論 構文木 自然言語解析
自然言語処理学 自然言語処理学講座
自然言語処理技術 自然言語処理研究
自然言語処理研究会 自動要約 質問応答
質問応答システム 情報抽出 人工知能学会誌
対話処理 単語単位 知識ベース 知識処理
知識発見 茶筌 日本語形態素解析
日本語形態素解析システム 複合名詞 未知語 名詞句
論理プログラミング 曖昧性解消

語から見出し語を選定し、サブ分野に分類した結果を表 3 に示す。見出し語数 $N = 30$、$\alpha = 3$、サブ分野代表語数 $m = 5$ とした。表において、**大項目**はサブ分野代表語に選ばれた用語、**小項目**はそれぞれのサブ分野に所属する見出し語である。また、見出し語に選ばれなかった用語を表 4 に示す。

4.3 考　　察

表 3 を見ると、大項目の用語としては「コーパス」、「構文解析」、「形態素解析」などの「自然言語処理」のサブ分野を表す用語が選ばれている。一方、小項目の用語を見ると、「自然言語処理」より大きな概念である「人工知能」が「機械翻訳」のサブ分野に含まれるなど、違和感を感じる部分はあるが、全体としては用語がサブ分野にうまく分類されている。

また、見出し語として選ばれた用語と、選ばれなかった用語を比較してみると、選ばれなかった用語には、「形態素解析器」やその具体例である「ChaSen」、「JUMAN」といった、選ばれた用語と比較して概念の粒が小さな用語が多く見られる。

一方で、選ばれた用語の中に「入力文」や「日本語文」といった、感覚的にはあまり専門用語であると思えない用語が含まれている。これは、例えば「入力文」というのは、一般語である「入力」と「文」が組み合わさった用語であり、「入力文」が表す意味は、「入力」と「文」の意味から合成的に推測できるため、人間の感覚としてはあまり専門用語という気がしない。ところが、実際に**入力が文である**状況というのは、形態素解析や機械翻訳など**言語を処理する**場合に限られるため、「入力文」に関する表層的な情報 (ウェブのヒット数や他の用語との共起情報) から判断すると、「入力文」は間違いなく「自然言語処理」の専門用語である。よって、本論文で見出し語を選定するために用いた指標で、これらの用語を排除するのは困難である。

表 2 の「自然言語処理」以外の入力用語に対しても、見出し語を選定した結果、「自然言語処理」の場合と同様に、対象分野の構造を捉えた用語集を編集することができた。ただ、3.2 節の収集様語数の拡大手順の**ステップ 2** で選択する $s_{c1}, \cdots, s_{cn}$ や、見出し語集合を構造化する際に選択するサブ分野を代表する用語 $sd_1, \cdots, sd_m$ の選び方によって、結果が大きく異なる場合があり、これらの選び方については検討の余地がある。

5. おわりに

本論文では、ウェブを利用して特定の専門分野の用語集を自動的に編集することを実現するための方法について述べた。専門用語集の編集方針を、初心者が対象分野の全体像が把握できるような用語集を編集することに定め、ウェブのサーチエンジンにおけるヒット数や、文中での用語の共起情報を利用することにより、専門分野の構造を考慮に入れた専門用語集を編集することが可能であることを示した。

本研究と関連する研究として、藤井ら [4] の研究がある。これは、ウェブを百科事典のように使うことを目的とした研究であり、百科事典であること、すなわち、網羅的に情報を収集することに主眼が置かれている。一方、我々の研究は、用語集を編集すること、すなわち、有用な用語の集合を選び出すことに重きを置いており、この点が藤井らの研究と大きく異なっている。

本研究の一部は、次の研究費による；特定領域研究「実世界の関連性を投影した語彙空間の構築」(課題番号 16016249)、21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」。

参 考 文 献

1) 長尾真: 辞典形式での専門分野の知識の体系的構成法, 人工知能学会誌, Vol.7, No.2, pp.320–328 (1992).
2) 倉島節尚: 辞書と日本語 国語辞典を解剖する, 光文社 (2002).
3) 佐々木靖弘, 佐藤理史, 宇津呂武仁: 用語間の関連度を測る指標の提案, 言語処理学会第 10 回年次大会発表論文集, pp. 25–28 (2004).
4) 藤井敦, 石川徹也: World Wide Web を用いた辞典知識情報の抽出と組織化, 電子情報通信学会論文誌 D-II, Vol. J85-D-II, No. 2, pp. 300–307 (2002).