Web文書集合からの専門用語獲得

池野篤司[†,†††] 濱口佳孝[†]
山本英子[††] 井佐原均[††,†††]

本研究では，Web 文書集合から専門用語獲得を行った．まず統計的に用語を獲得し，その後，専門用語属性を判別するという手順で処理した．統計的用語獲得に関しては，形態素 n-gram の統計・表層情報を利用し，時間的・物理的コストを考慮した．主に「候補の選定」と「単位性の確認」の工程を経る．候補の選定では，特定の文書に集中して出現する程度（集中度）と，文書集合全体での出現の散らばりの程度（分布度）との双方が極端に高低に寄らないことを条件とした．単位性の確認では，語全体の「単位性」（ひとかたまりの単語と見なせる度合い）が強いことを条件としてさらに候補を絞り込んだ．実験により，専門用語と判断できる複合名詞や名詞句などを獲得できた．専門用語属性判別では，上記手段により獲得された用語を対象に，末尾の構成要素と，属性ラベルが専門用語である構成要素とに着目し，まず単純な属性拡張ルールを適用したところ，適合率は 9 割だったが，再現率が低いという結果を得た．さらに，既判別の専門用語の構成要素を「属性影響語」とする概念を導入し，属性影響語に専門用語属性を仮設定して，再度属性拡張ルールを適用した．その結果，適合率が 6 割に低下したものの，再現率を 9 割近くまで引き上げることに成功した．

Technical Term Acquisition from Web Document Collection

Atsushi Ikeno,[†,†††] Yoshitaka Hamaguchi,[†] Eiko Yamamoto[††] and Hitoshi Isahara[††,†††]

This paper proposes a method to acquire new terms statistically, and also to label them as technical. The process is designed to be simple and fast because the resulting terms are to be added to the dictionary in use for practical systems. Term acquisition process is mainly composed of candidates selection and “unithood” checking. Term candidates are selected on the condition that they appear repeatedly in certain documents and also used widely in the entire corpus. The candidates are further selected by their “unithood,” the appropriateness as a word unit, using the function dependent on the connection strength to the connecting morpheme. The label of a term is estimated by expanding the labels of its constituents. Simple expansion method according to the labels of the last constituent and the constituents labeled as technical, achieved 90% accuracy and 60% recall. Introducing the notion of “label affecter,” named for the constituents of already-discriminated terms, and affecter expansion method raised the recall to 90%, while accuracy became 60%.

1. はじめに

文章を解析する機能を持つ様々なシステムにおいては，日々生み出される新しい用語をいちはやく辞書に取り入れることが精度に影響を及ぼす．従来は，新聞・論文・マニュアルなど，ある程度定められた用語で記述される文章を対象にすることが多かったため，用語が十分に確立されたことを人手で確認して辞書に登録する方法がとられてきており，その方法で大きな問題になることはなかった．

しかし，インターネットの利用が一般化した昨今では，Web ページが重要な解析対象の 1 つである．Web で新しい用語が使われ始めると，ごく短期間で多くのページに広がっていく．つまり，Web においては，用語は急速に確立されるので，自動的な用語獲得の手段を用意することが重要となる．

一方，インターネット上での検索サービスが広く使われるようになり，それにともなって基本的な文書検索だけでなく付加的な機能を提供するものも増えてきた．そのような機能の 1 つに，Webcat Plus[1)]，Cy-

† 沖電気工業株式会社研究開発本部
Corporate Research & Development Center, Oki Electric Industry Co., Ltd.
†† 独立行政法人情報通信研究機構
National Institute of Information and Communications Technology
††† 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University

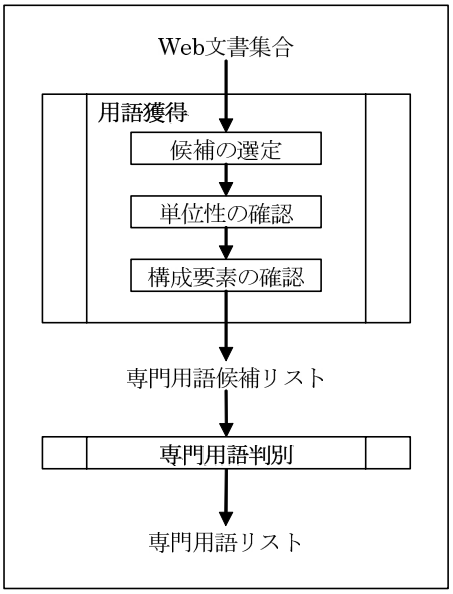


図 1 専門用語獲得の流れ
Fig. 1 Technical term acquisition process.

clone[2]，Bluesilk[☆,3] などで採用されている，関連語を表示する機能がある．特に，Cyclone や Bluesilk では関連語のうち「人名」「地名」などの特定の属性を持つものだけを提示することができるようになっており，多義語の調査や分野別の調査などに有効である．

関連語に関しても，新しい用語を辞書に取り入れて提示できることは重要である．ところが，特定の属性を持つ語だけを提示するためには，属性の分かっている用語が獲得できることが必要となる．

人名や地名などの固有表現に属するような用語の抽出（獲得）は，従来の情報抽出技術でもある程度の精度で対応できるが，特定分野の専門用語などを発見することは困難である．

一方で，従来の統計的用語獲得の手法では獲得される語の属性は意識していないため，様々な属性の語が入り混じって獲得されてしまう．

そこで我々は，Web ページを収集した大規模文書集合からまず用語を統計的に獲得した後に，技術用語などの特定分野の専門用語であるかどうかを判別するという方法により，専門用語を獲得することを試みた．**図 1** に専門用語獲得の流れを示す．まず用語獲得のステップで Web 文書集合から専門用語候補リストを得る．この段階では，リストには様々な属性を持つ用語が含まれている．次に，その専門用語候補リストから，専門用語判別のステップにより専門用語と確定した語のリストを得る，という流れになっている．

本研究の最終成果としては，上述したような情報検索・抽出を行う Web アプリケーションの辞書に対する継続的な用語の供給を行うシステムを想定している．供給すべき用語の決定にあたっての人手の関与は妨げないが，用語の獲得と辞書への供給を短期間のサイクルで回せるようにするために，大量のデータを短時間で処理できることをも目的の 1 つとした．

本稿では，2 章で web ページからの用語獲得の手法と実験，3 章で専門用語判別の手法と実験について述べる．4 章で全体を考察し，5 章で総括する．

2. 用 語 獲 得

2.1 方　　針

2.1.1 獲得する用語

用語獲得に関しては，様々なアプローチから研究が行われている[4)~7)]．中川ら[6)] は，専門用語には複合名詞が多いことを考慮して，複合名詞を獲得対象とする研究を行い，良好な結果を得ている．

我々は，Web アプリケーションの関連語提示機能への適用を考えており，「〜の法則」や「〜を用いた…」のような用語も獲得対象とする．そのため，用語を獲得する段階では複合名詞に限定しない手法を用いる．よって，本研究で獲得する用語は「複合名詞に限らず，名詞句の形を持つ特徴的な文字列」とした．

2.1.2 実行効率の考慮

本研究では，文字列の表層的な頻度分布情報の特徴をとらえることにより，用語らしい文字列を抽出することを試みる．これまでに任意の文字列の頻度情報を得る効率的な手法が提案されている[8),9)]．本研究でもこの手法を利用する．

しかし，1 カ月程度で最新のページ群から用語を獲得しつづけることを想定しているので，さらなる効率化が必要であると考えた．文字単位の n-gram での処理コストは，獲得対象とする用語が長くなるにつれて飛躍的に増大するため，取り扱う列の最小単位を文字ではなく形態素とすることにする．これにより，形態素解析のコストが別にかかることになるが，後の処理対象となる要素数を減らすことができるため，処理全体の作業領域と計算コストを抑えられる効果のほうが大きいと判断した．その結果，処理対象となる文字列は形態素 n-gram となり，獲得される用語は複合語に準じたものとなる．

2.2 手　　法

2.2.1 使用する統計量の定義

本研究で使用した統計量を示す．ここで任意の n-

☆ Bluesilk は（株）三菱総合研究所の登録商標である．

gram を Z と表すとする．このとき，

N：Web 文書集合にある総文書数
$ml(Z)$：Z を構成する形態素数
$cf(Z)$：Web 文書集合中の Z の出現頻度
$df(Z)$：Z が出現する文書の数
$df2(Z)$：Z が 2 回以上出現する文書の数
$idf(Z)$：$logN - log(df(Z))$
$cd(Z)$：Z に連接する形態素の異なり数

を文書集合からの統計量として用いる．

2.2.2 工 程 手 順

本研究では，最初に文書集合から多くの候補を抜き出した後に絞り込んでいくというアプローチをとる．手順としては，まず「候補の選定」を行い，その結果に対して「単位性の確認」を行う．その後,「構成要素の確認」により結果を調整する．

「候補の選定」においては，適当な長さまでの形態素 n-gram をすべて数え上げて，その中から文書集合中の特徴的な用語候補だけを選定する．ところが，ここで得られた結果の用語候補に対しては，語の境界の妥当性に関する判断が行われていないため，ひとかたまりの語と見なすには不適なものも含まれている．そこで,「単位性の確認」において，各用語候補の「単位性」に着目した条件を用いて，さらに候補を絞り込む．

上記手順が主な処理となるが，この段階でまだ除去しきれない不適当な用語候補に備えて,「構成要素の確認」において，特に先頭と末尾の構成要素についての条件を用意して対処することとした．

それぞれの手順について以下に詳述する．

(1) 候補の選定

文書集合中の特徴的な用語の候補としては，ある文書の内容を代表する用語と，分野において特徴的であり，その文書集合を特徴付ける用語との 2 通りの基準で考えることができる．

ある文書の内容を代表する語はいくつかの特定の文書に集中的に出現する用語である．このような用語はある範囲内の $df2/df$ を持つ傾向があると報告されている[10]．この特徴により，$df2/df$ は語分割などに適用されている[11]．そこで，このような集中度を測る指標として $df2/df$ を利用することとした．

その一方で，特定の文書ではなく文書集合全体の特徴をとらえる用語は，ある特定の文書だけに極端に集中するのではなく，適度に散らばって出現する．ただし，あまり多くの文書に出現するものは機能語の類が含まれるので適当ではない．本研究では，このような分布度を測る指標として df/cf を利用することとした．

これにより，文書集合中の特徴的な用語を選び出すためには，集中度 $df2/df$ と分布度 df/cf とが一定の値の範囲に収まっているという条件を課せばよいと考えられる．

本研究での具体的な選定手順を以下に記す．

まずコーパス中の unigram から 10-gram までのすべての形態素列を抜き出す．

次に，その中から，$df2/df$ および df/cf の値が閾値の範囲内に収まっているものを候補として残す．閾値はここでは実験的に $0.2 < df2/df < 1.0$，$0.2 < df/cf < 0.8$ のように設定した．

また，実用上の利点から，この段階で，先頭または最後尾に「助詞」または「副詞」を持つ n-gram を候補から除去することとした．

(2) 単位性の確認

本研究では，形態素列の unithood（単位性）[12]（ひとかたまりの単語として用いられている度合い）を調べるためのスコア関数 $RScore$ と $LScore$ を導入して，さらに候補を絞り込む．単位性を調べる関数の例として C-value[13] があるが，本研究では $RScore$ と $LScore$ に C-value の変形関数をあてはめて実験を行う．

上で選定された候補が下記の 2 つの条件を満たすとき，その n-gram を用語と推定するものとする．

各 n-gram を $S_{1n} = M_1 \ldots M_n$ で表し，M_i は一形態素を表す．S_{ij} を n-gram を構成する形態素 M_i から M_j までの部分形態素列 $M_i \ldots M_j$ とした場合，

- $R < j \leq n$ の間，$RScore(S_{1j}) \geq RScore(S_{1j-1})$
- $1 \leq i < n - L$ の間，$LScore(S_{in}) \geq LScore(S_{i+1n})$

ただし，R，L は $1 \leq R$，$L < n$ の範囲の数とし，どの部分の連接関係を調べるかを設定するパラメータである．

この式は，ある形態素列が自分より 1 形態素だけ短い部分形態素列よりも単位性が強いことを意味している．そのため，この条件は，用語の候補としてあげられた形態素 n-gram よりも，右（後方）または左（前方）を短くした部分形態素列の方が単位性が強いときに，当該の形態素 n-gram を候補から外す役割を果たす．これにより，単位性の強い形態素列を包含する長い形態素 n-gram は，用語としての境界が妥当でないとして候補から除去される．

R と L に 1 を設定した場合は，形態素列 S_{1n} のすべての部分形態素についてこの関係が成立することを調べることになる．$R = 1$ の場合，S_{1n} を右（後方）から 1 形態素ずつ除去し，つねに長い方の単位性が強いことを確認する．同様に，$L = 1$ の場合，左（前

方）から 1 形態素ずつ除去して確認する．

たとえば，候補 n-gram「まち_づくり_事業」について，それぞれの形態素は M_1=“まち”，M_2=“づくり”，M_3=“事業” であり，$R = 1$，$L = 1$ のとき，$RScore(S_{13})$ と $RScore(S_{12})$ が

- $RScore$（まち_づくり_事業）$\geq RScore$（まち_づくり），

$RScore(S_{12})$ と $RScore(S_{11})$ が

- $RScore$（まち_づくり）$\geq RScore$（まち）

であり，かつ $LScore(S_{13})$ と $LScore(S_{23})$ が

- $LScore$（まち_づくり_事業）$\geq LScore$（づくり_事業），

$LScore(S_{23})$ と $LScore(S_{33})$ が

- $LScore$（づくり_事業）$\geq LScore$（事業）

であるならば,「まちづくり事業」を用語と推定する．

実験では，スコア関数 $RScore$，$LScore$ について，下記の 5 つをあてはめて結果を比較する．

1. $log(ml(Z)+1) * log(cf(Z)) * (1 - 1/cd(Z))$
2. $log(ml(Z)+1) * log(cf(Z)) * cd(Z)/(cf(Z) + cd(Z))$
3. $log(cf(Z)) * cd(Z)/(cf(Z) + cd(Z))$
4. $idf(Z) * (1 - 1/cd(Z))$
5. $idf(Z) * cd(Z)/(cf(Z) + cd(Z))$

これらは，1 の関数を基本とし，その変形により他の 4 種類の関数を作成して得られたものである．

1 の関数は田中らの関数[7] である．この関数は多言語な語彙の用例を調べるツールを開発するために，C-value を基に，文字を単位として定義された関数である．C-value は単語を単位として定義されているが，その単語は名詞が想定されている．しかし，本研究で獲得したい用語は複合名詞ばかりではない．そこで，田中らの関数を文字ではなく，形態素を単位として用いることにした．

1 の関数において，第 1 項 $log(ml(Z)+1)$ が長さの項（Z が長いとき値が高くなる），第 2 項 $log(cf(Z))$ が出現頻度の項（Z が頻出するとき値が高くなる），第 3 項 $(1 - 1/cd(Z))$ が連接する形態素の異なり数の項（異なり数が大きいとき値が高くなる）である．

この基本関数についての懸念は，長さの項と出現頻度の項がともに独立した乗算要素であるため，関数全体における影響が非常に大きいことと，異なり数の項に出現頻度からの影響が補正されていないことである．

そこで，これらの影響を軽減することを目的として変形した 4 つの関数を検討に加えた．

2 番目の関数は，1 の関数における異なり数の項の代わりに，出現頻度の影響を考慮した異なり数の項 $cd(Z)/(cf(Z) + cd(Z))$ を使用した関数である．

また，3 番目の関数は，2 の関数から長さの項を削除した，n-gram の長さをスコアに反映しない関数である．

4 番目の関数は，1 の関数にある長さの項と出現頻度の項の代わりに，長さと出現頻度との影響を緩やかに表す $idf(Z)$ を使用した関数である．

5 番目の関数は，4 番目の関数を変形したもので，2 の関数と同様に，出現頻度の影響を考慮した異なり数の項を使用した関数である．

(3) 構成要素の確認

上で絞り込まれた用語候補について，以下の方法で不適当と思われる用語をさらに候補から外す処理を行う．

- $S_{1n} = M_1 \ldots M_n$ に関して，
  $df(M_1)/N > \alpha$，または $df(M_n)/N > \beta$ であるならば，S_{1n} を除去する．
  ただし，閾値 α および β は 1 形態素の統計量に基づき設定する．

これは，先頭または末尾の形態素が,「お」「ご」などの接頭辞や,「〜について」などの助動詞の活用形で機能語的に用いられる形態素であるような用語候補に対する処理である．このような形態素を含む用語は必ずしも排除しなければならないものばかりではないが，文書集合内での出現の仕方によっては，用語の一部として考えるのがふさわしくない場合もあると考えられる．このような語（形態素）は比較的高い df をとると考えられるので，先頭および末尾の形態素について，閾値より高い df をとる場合に，当該の用語候補を排除するのが上述の式である．閾値はここでは実験的に $\alpha = 0.1$，$\beta = 0.003$ を設定した．

2.3 実　　験

上記 5 種のスコア関数について，単位性を確認する工程で用いるパラメータ R，L を変えて実験を行った．ただし，パラメータ R，L は同じ値 K を設定した．

$K = 1$ の場合，各 n-gram において，すべての部分形態素列についてつねに自分よりも 1 つ長い形態素列の方が「単位性」が強い場合に，はじめて元の形態素列は用語と認められる．部分形態素列の「単位性」の方が強いところが 1 カ所でもあれば，この n-gram は候補から外されるので，最も厳しい条件である．

一方，$K = n-1$ の場合，連接関係が比較的弱いと考えられる先頭（$LScore$ の場合）または末尾（$RScore$ の場合）の 1 形態素を落とした部分形態素列についてのみ,「単位性」の比較をすることになる．1 形態素だけについての確認なので，これが最も緩い条件となる．

たとえば，先の例にあげた候補 n-gram「まち_づくり_事業」については，$K = 1$ の場合，

表 2 各手法によって推定された用語の数
Table 2 Number of estimated terms.

手法	推定数	手法	推定数	差 (B-A)
A1	1,194 (1,185)	B1	1,778 (1,695)	584 (510)
A2	5,282 (5,226)	B2	10,008 (9,640)	4,726 (4,414)
A3	5,293 (5,233)	B3	6,205 (6,012)	912 (779)
A4	14,173 (13,670)	B4	15,216 (14,347)	1,043 (677)
A5	12,482 (12,064)	B5	13,409 (12,727)	927 (663)

表 1 検討手法：パラメータとスコア関数の組合せ
Table 1 Examined parameters and score functions.

K	関数 1	関数 2	関数 3	関数 4	関数 5
1	A1	A2	A3	A4	A5
n-1	B1	B2	B3	B4	B5

- $RScore$（まち_づくり_事業）$\geq RScore$（まち_づくり），
- $RScore$（まち_づくり）$\geq RScore$（まち），
- $LScore$（まち_づくり_事業）$\geq LScore$（づくり_事業），
- $LScore$（づくり_事業）$\geq LScore$（事業）

のすべてが成立しなければ用語と認められないのに対し，$K = n - 1$ の場合，

- $RScore$（まち_づくり_事業）$\geq RScore$（まち_づくり），
- $LScore$（まち_づくり_事業）$\geq LScore$（づくり_事業）

の 2 つの条件が成立するだけで用語と認められる．

また，上記からも明らかなように，計算コストは，$K = 1$ の場合が最も高く，$K = n - 1$ の場合が最も低い．

以上の点を考慮して，最も条件が厳しく計算コストが高い $K = 1$ の場合と，最も条件が緩やかで計算コストが低い $K = n - 1$ の場合とを選択し，合計 10 種類の手法について検討した．

表 1 に検討手法と以降で用いる識別子を示す．

実験の対象データには，2004 年 1 月時点での t.u-tokyo.ac.jp ドメイン以下に属する，工学部および大学院工学系研究科に関連するページの集合（約 200 MB）を収集したものを用いた．収集の際には，執筆者や内容によるファイルの選別をまったく行っていないため，研究に関係のない個人ページも含まれている．

また，形態素解析は，日本語形態素解析システム『茶筌』[14] に，独自に用意した専門用語辞書を追加したものを利用して解析した結果を用いることとした．品詞情報は，候補の選定時に補助的に用いているだけで，一連の統計処理に関しては品詞情報が付与されていない形態素の列だけを用いる．

2.4 手法についての比較・検討

表 2 に各手法の推定用語数と手法 A と手法 B の推定用語数の差を示す．表中の値は各手法が用語とした n-gram の数である．() 内は，そのうちほかの n-gram の部分形態素列であるものを削除した数である．

2.4.1 パラメータ K に関する比較

$K = 1$ の場合（手法 A）に得られる結果は，$K = n - 1$ の場合（手法 B）においてもすべて得られるが，その結果を検討してみると，A と B がともに推定できる用語は多くの場合複合名詞であった．これに対して，表 2 に示す「差」の部分に相当する B だけが推定できる用語には，「情報_の_可視化」などの間に助詞が含まれるものや，「あいまい_知識_処理_手法」などのひらがなで表記される単語が含まれる場合や「ママ_チャ_リ」などの表層文字種が同じで解析誤りがあるものが観られる．本手法では，候補の選定で，頻度に関して特徴的な n-gram を選ぶので，連接関係が比較的弱いと考えられる先頭または末尾の特徴により用語の範囲を推定する手法 B によって，中間に助詞や頻出する表記を含んでいる n-gram を用語と推定できるためである．

この結果から，パラメータ K を $n - 1$ とした手法 B のほうが，本研究の目的に適していると考える．また，コスト面での有利さも加味される．

2.4.2 スコア関数に関する比較

スコア関数に関して比較するために，専門用語辞書が追加されていない，通常の『茶筌』を用いて形態素解析した場合の形態素 n-gram から用語を推定し，専門用語辞書に含まれ，コーパス中に出現する専門用語について再現率を測定する．

再現する専門用語は，候補の選定条件を満たし，通常のシステムでは 2 形態素以上に分割される専門用語 1,646 個とした．

なお，コーパス中の特定の文書を特徴付ける用語は $df2$ が高い傾向にあることから，用語獲得結果は $df2$ に従ってランク付けしておくものとする．

表 3 に再現率と既存用語含有率を示す．既存用語含有率とは，各手法が推定する用語において，どの程度既存の専門用語を含むかを示す値である．具体的には，獲得された用語数を N，そのうち既存用語であるものの数を M としたときの M/N の値となる．

表 3 より，再現率は推定された用語の数が多いほど，

表 3 専門用語の再現率と既存用語含有率
Table 3 Recall and inclusion rate of existing technical terms.

手法	推定数	専門用語数	再現率	既存用語含有率
B1	1,921	223	0.1355	0.1161
B2	11,580	1,351	0.8208	0.1167
B3	7,823	1,052	0.6391	0.1345
B4	16,960	1,436	0.8724	0.0847
B5	15,068	1,405	0.8536	0.0932

高い傾向にあることが分かる．これは推定された用語が多いほど，特徴的な用語の範囲が広くなり，再現される既存の専門用語は増加するのは明らかである．再現されなかった用語には以下のようなものが多かった．

(i) 獲得された用語の部分文字列である用語
(ii) 部分文字列が獲得された用語

(i) の場合，たとえば，既存の専門用語「航空_宇宙_工学」より長い「夏休み_航空_宇宙_工学」が獲得される．これはより長い文字列の方が情報量を持つ特徴的な用語として判定されたためである．

また，(ii) の場合，たとえば,「反_磁性_体」「強_磁性_体」「反_強_磁性_体」の部分文字列「磁性_体」,「二_足_歩行_ロボット」の部分文字列「歩行_ロボット」が獲得される.「磁性_体」の場合は，前に多種類の形態素が連接するため，専門用語を作成する部分文字列「磁性_体」が獲得され,「歩行ロボット」の場合は,「二足歩行ロボット」が省略され,「歩行ロボット」と用いられることが多いため,「歩行ロボット」を特徴的な用語として判定された．

表 3 の既存用語含有率からは，どの手法においても，推定した用語のうち 1 割前後が既存の専門用語であることが分かる．また，B3 が他の手法に比べ，既存の専門用語が占める割合が高い．専門用語はコーパス中の特徴を表す用語である可能性が高いので，手法 B3 は本研究での用語獲得に適しているのではないかと推察される．

次に，各手法で推定された用語のうち上位から同数（獲得用語数が最少であった B1 の 1,921 件）取り出して専門用語の再現率と既存用語含有率を比較したものを**表 4** に示す．

表 4 において，再現率，既存用語含有率ともに，B3 が他の手法と比べて高い値を示す．専門用語は文書を特徴付ける語であるため，文書頻度 $df2$ も他の用語と比べて高い場合が多い．したがって，$df2$ の値でランク付けされた上位の用語を多くカバーできている B3 の手法は，専門用語を獲得するという点においては他の手法よりも優位であると期待できる．

再現できた用語は以下のようなものである．

表 4 df2 の上位 1,921 件に関する再現率と既存用語含有率
Table 4 Recall and inclusion rate for terms of high df2.

手法	専門用語数	再現率	既存用語含有率
B1	223	0.1355	0.1161
B2	397	0.2406	0.2067
B3	415	0.2521	0.2160
B4	337	0.2047	0.1754
B5	378	0.2296	0.1968

(i) 複合名詞
(ii) 解析誤りによって分解されてしまった用語

(i) の場合，再現されなかった専門用語と比べ，特徴的な複合名詞と判定された用語である．たとえば,「量子_エレクトロニクス」「量子_井戸」「量子_化」「量子_化学」など,「量子」を単体で用いることが稀であるため，それと組み合わされた専門用語が再現された.「目_詰まり」「角_運動_量」も同様である．

また，(ii) の場合,「アイコ_ン」「ゲイ_ン」「アナ_ラ_イザ」「ジョセフ_ソ_ン_効果」などの解析誤りが多いカタカナ表記を含むものや,「界_磁_巻_線」「透_磁_率」「揚_抗_比」などの人名や未知語と解析される漢字で構成されるものである．これは検討手法により，形態素の並びによって意味を持つ用語であるということを判定できることを示している．

2.5 獲得用語についての考察

2.3 節の実験において用語として獲得された n-gram について考察する．獲得された用語には以下のものがみられた．

(1) 複合名詞
(2) 特徴的な文字列から始まる用語
(3) 解析誤りによって分解された用語
(4) 名詞句
(5) 接頭語が付く用語
(6) 人名

以降，各種類の用語について分析する．

(1) 複合名詞

以下に獲得された複合名詞の一部を示す．英語表記と日本語表記の名詞の組合せも含む複合名詞，同じ先頭文字列から始まる専門用語とみられる用語，などが得られている．これらは専門用語辞書の強化に直接利用できる．

Voronoi_図
超_臨界圧_軽_水冷却原子炉
超_臨界圧_軽水_冷却_減速_炉
超_臨界圧_軽水炉

(2) 特徴的な文字列から始まる用語

「まち_づくり_NPO」「まち_づくり_ファンド」「ま

ち_づくり_条例」など多数獲得された．これらは「まち_づくり」という文字列が特徴付ける用語である．このような用語を特徴付ける文字列によって分類しておけば，アプリケーションのユーザに提示する際に有用であると考えられる．

(3) **解析誤りによって分解された用語**

アルファベットに分解された英単語や，分解されてしまったカタカナ用語，カタカナやひらがな表記と漢字との組合せによる用語が獲得されている．

H_am_ilton_の_定理
マイク_ロメ_カニ_クス
き_裂_伝播

(4) **名詞句**

専門用語は基本的には複合名詞であるが，句の形の用語も存在する．実験では，以下のような，タイトルと思われるもの，専門用語と判断できる用語が獲得されている．

テンソル_と_ベクトル_の_ドット積
シビア_アクシデント_の_伝熱_流動_現象
_における_素_過程

(5) **接頭語が付く用語**

実験では，接頭語「お」や「ご」が連接した用語が獲得された．これらは丁寧語であり，専門性や研究内容に即した文書に現れる用語ではない．したがって，今回の目的に直接は利用できないが，たとえば，このような用語が現れる文書を検索対象から除去するなど，他の種類の用語とは異なる支援に利用できる可能性がある．

(6) **人名**

人名は情報抽出によってもある程度は認識（獲得）できるが，情報抽出の手法上，肩書きなどの人名を認識するためのキーとなる文字列が存在しないものについても本研究の方式で獲得できる．専門用語ではないが，関連語を提示するアプリケーションには有用な情報である．

3. 専門用語判別

3.1 方　　針

継続運用する応用アプリケーションに対して最新の辞書データを追加していくことを想定しているので，精度を保ちながらも処理全体をできる限り単純・高速な（コストの低い）ものにする必要があるため，本研究で用いる手法では，用語リストだけを受け取り，原文に関する情報は利用せずに処理することを試みる．

3.2 専門用語の定義

専門用語の定義については，獲得された用語をどのように利用するかによって定義が変わる可能性があることが，いくつかの議論[5),15)]から読み取ることができる．

本研究においては，関連語のうち専門用語であるものとしてユーザに提示して発想を支援する目的に利用することを想定している．そのため，ユーザに提示して有用な語という観点で，分野全般にある程度の知識がある人間の評価者が認定できたものを専門用語であるとする．なお今回の対象データは工学系ホームページであるので，専門用語は技術用語であるとする．

以降の実験において正解集合として利用する専門用語（技術用語）の判定には以下の基準を設けた．

- 工学系用語にある程度なじみがある人間の判定者1名による判断で作成する．
- コンピュータ関連用語は研究内容と直接の関連がなくても専門用語とする．
- 組織名・書籍名など他の属性を付与するほうが妥当な場合は専門用語としない．
- 迷った場合は根拠となった文書，またはその他の文書において，用語の利用例を確認して判断する．

3.3 判別の必要性（予備実験）

以降の実験ではB3の手法によって獲得された用語候補を対象とする．

ドメインが限定されたテキストから獲得された用語候補は，そのほとんどが専門用語であることも考えられる．そのため，最初に，一般的な固有表現抽出手法によって容易に固有表現であると判断できる用語がどれくらい含まれるかを確認する予備実験を行った．

(1) **用語候補に対する固有表現判定**

まず，開発済みの固有表現抽出器[16)]を適用したところ，6,205語中1,190語は固有表現であると判定された．語境界の間違いや一般用語を固有表現と判定した間違いなどにより精度は約9割であったが，これらの中に専門用語が含まれてしまう間違いはなかった．この結果により，以降の実験では，処理対象を残る5,105語に絞ることにした．

(2) **専門用語の比率確認**

さらに，残った語のうち専門用語がどれくらいを占めるかを正解集合と比較して確認したところ，5,015語中2,805語のみが専門用語である，という結果が得られた．このことは，対象用語をそのまま辞書化するのは現実的ではなく，専門用語か否かの識別を行う必要があることを示している．

3.4 手　　法

獲得された用語は複数の要素語が結合された形をしている．用語全体が専門用語であるかどうかを判定す

るにあたって，用語を構成する各要素が専門用語であるかどうかの情報を利用することにした．

各要素についても，予備実験のときと同様に，既存の固有表現抽出器や専門用語辞書とのマッチングにより，固有表現か専門用語かそれ以外の一般的な語であるかを判別した情報を付与しておくものとする．ここで付与される，人名・地名などの固有表現の分類に専門用語という分類を加えたものを，以降の説明では一括して属性ラベルと呼ぶことにする．

提案手法は，用語のいずれかの構成要素の属性ラベルを全体に反映させる単要素属性拡張をまず行う．次に，それでも属性ラベルが付与できなかった用語を対象として，再度拡張を試みる．このとき，単要素で拡張できた用語の構成要素から選択した属性影響語を用いる．以下に詳細を記す．

3.4.1 単要素属性拡張

ここでは，構成要素の属性ラベルを用語全体に波及させる単要素属性拡張により，専門用語がどの程度判別できるかを確認する．

処理対象の用語リストに対し，以下の属性拡張ルールを順に適用することによって，専門用語と判定されたものの再現率と適合率を調べる．

(1) **末尾属性拡張ルール**

用語を構成する末尾要素の属性ラベルを，用語全体に割り当てる．一般に日本語の複合語においては，末尾要素が複合語全体の品詞・意味を規定するという考えに基づく．

以下の場合はこのルールを無条件に適用する．

(a) 末尾以外の要素には特定の属性ラベルが付与されていない場合

(b) 末尾要素属性と同じか同類の属性ラベルを持つ要素のみが存在する場合

末尾要素の属性ラベルと，他の要素の属性ラベルが異なる場合には，相互の相性を記したルールにより割り当てるかどうか判断する．

(2) **専門用語属性拡張ルール**

専門用語は固有表現よりも用語全体に与える影響が大きいであろうという仮定に基づいて，(末尾以外に) 専門用語属性を持つ構成要素が存在する場合に，上記のルール (1) と同様の条件で用語全体の属性ラベルを割り当てる．

3.4.2 属性影響語による拡張

先のルール未適用となった用語は属性に関する決定的な情報を持たないため，既知の情報を用いて，未適用用語に関連する情報を補完する必要がある．そのため，先の実験で判別できた用語の構成要素から「属性影響語」を選択する方法を提案する．

(1) **「属性影響語」の抽出**

特定の属性を選択し，その属性ラベルを付与された用語の集合から頻出する構成要素をリストアップして，それらを「属性に影響を与える語（属性影響語）」であるとする．

(2) **属性の仮設定**

属性影響語のうち，現在は属性ラベルを持っていない語（固有表現でもなく専門用語でもない語）に対して，(1) で選択した特定の属性を一時的に設定する．

(3) **拡張ルールの適用**

属性影響語の属性を設定した状態で，3.4.1 項の各ルールを再度適用する．

3.5 実　　験

3.5.1 単要素属性拡張の結果と検討

表 5 に専門用語の判別に関する結果を示す．

ここでの再現率は，「専門用語と判別されてかつ正解であった語数（C）」を，「母集合中における人手で判定した専門用語数（専門用語正解の総数）（T）」で割ったものである．以降の実験結果においても同様である．

ここで属性拡張ルール (1) または (2) が適用できた場合，専門用語判定の適合率は 9 割程度であった．辞書登録前の人手でのチェックは必要であるとしても，これらのルールは実用に耐えうる有効なルールであったといえる．

一方で，どちらのルールも適用できなかったものが半数以上残り，その中に専門用語と判別されるべき語が多く含まれていること（再現率約 44%）から，この 2 つのルールだけでは十分ではないことが分かる．

末尾属性を採用して正しくラベルが付与できた例を以下に示す．

例 1).

倒立振子　　＜ラベル：専門用語＞

倒立（一般名詞）+振子（専門用語）

また，専門用語属性を採用して正しくラベル付けできた例を以下に示す．

例 2).

弾塑性解析　　＜ラベル：専門用語＞

弾塑性（専門用語）＋解析（一般名詞）

一方，誤ったラベルを付与してしまった例には以下のようなものがあった．

例 3).

原子炉実習　　＜ラベル：専門用語＞

原子炉（専門用語）＋実習（一般名詞）

表 5 専門用語判別の結果
Table 5 Technical term discrimination results.

	ルール適用語数 (W)	うち専門用語と判別された数 (E)	うち正解であった数 (C)	専門用語正解の総数 (T)	専門用語の再現率 (C/T)	専門用語の適合率 (C/E)
(1) 末尾属性	864	745	669	672	—	—
(2) 専門用語属性	612	612	553	553	—	—
未適用	3,539	—	—	1,562	—	—
合計	5,015	1,357	1,222	2,787	0.4385	0.9005

表 6 属性影響語を利用した専門用語判別の結果
Table 6 Results after using "label affecters".

	ルール適用語数 (W)	うち専門用語と判別された数 (E)	うち正解であった数 (C)	専門用語正解の総数 (T)	専門用語の再現率 (C/T)	専門用語の適合率 (C/E)
(1) 末尾属性	1,476	1,476	975	975	—	—
(2) 専門用語属性	641	641	378	378	—	—
未適用	1,422	—	—	209	—	—
合計	3,539	2,117	1,353	1,562	0.8662	0.6391

3.5.2 属性影響語による拡張の結果と検討

ここでの属性影響語の抽出は，属性ラベルが専門用語であるものを対象に行った．一時的に属性ラベルを付与する属性影響語として 556 語が得られた．ルール適用後の結果を**表 6** に示す．

この手法により，3 章の単要素属性による拡張実験の際に取りこぼした専門用語のうち約 87%を救済できたことは評価できる．

しかし，その一方で適合率は大きく低下した．属性影響語には，専門用語として妥当と考えられるものも多数あるが，一般名詞としてよく使われるものも多く含まれていることが原因であると思われる

属性影響語の例を以下に示す．

- （高頻度語）
  解析，構造，システム，制御，工学，計算，情報，光，特性，ファイル，冷却，熱，実験，など
- （低頻度語）
  連立，列，力学，稜線，面積，未知数，法線，溶液，補強，濾過，攪乱，炉心，励起，など

属性影響語による拡張により正しくラベルが付与できた例を以下に示す．

例 4).
ウェーブレット変換　＜ラベル：専門用語＞
ウェーブレット（未知語）
+変換（一般名詞 → 専門用語）

例 5).
マイクロマシン　＜ラベル：専門用語＞
マイクロ（一般名詞 → 専門用語）
+マシン（一般名詞）

一方，誤ったラベルを付与してしまった例には以下のようなものがあった．

例 6).
システム研究　＜ラベル：専門用語＞
システム（一般名詞 → 専門用語）
+研究（一般名詞 → 専門用語）

3.6 考察と課題

3.6.1 属性影響語による拡張

属性影響語の概念は，今まで気づかなかった重要な（「用語の属性」に影響を与える）部分文字列の発見を意味している．

本研究での影響語の選定にあたっては，用語リスト中での出現頻度のみを基準として一定の効果を得た．用語候補集合中での idf を用いて選別した実験も実施しているが，効果があまり得られないことが分かっている[17]．さらに選定方法について今後実験を重ねる予定である．

また，本研究では単純に，属性を持っていない構成要素（形態素）に属性ラベルを一時的に付与するという手法をとったが，ここにもまだ様々な選択方法を試すことが可能である．たとえば，対象となっている形態素が，属性を持っている長い形態素の一部分になっている場合は，さらに属性影響語としての重要度を増すように調整することも考えられる．

3.6.2 属性判別困難語

属性影響語の概念を導入し，それがある程度うまく働いたとしても，誤ったラベルを付与してしまったりする用語や，判定の手がかりを持たない用語がまだ存在することが分かってきた．全体の判別性能を向上させるため，まず今回は現状まだ手がつけられていない問題点について検討している内容についてふれる．

(1) 専門用語と一般名詞からなる用語

たとえば「都市工学/実習」「直角/方向」などは専門用語と判定したくない．一方で「空隙/構造」は専門用語であると判定したい．ところが，現在与えられている情報は，いずれも「専門用語＋一般名詞」であるということだけである．

末尾要素を重視するという観点からは,「実習」と「構造」との間に何らかの違いを見出せなければ判別は困難である．あるいは，同じ専門用語属性であっても,「都市工学」と「空隙」との間に専門用語としての強さの違いが存在すると考えるべきだろうか．

属性影響語の導入にともなう議論はここでもあてはまるかもしれないが，ここまでの結果からは，統計情報は決め手とするには弱いように思われる．この問題に関しては，今後，用語の語構成の果たす役割について詳細に分析する予定である．

(2) 再現率の向上に関して

「公開/鍵」（および「秘密/鍵」）という暗号技術に関する専門用語は，どちらの構成要素も一般名詞であるので，それだけの情報から，これを技術用語と判定するのは不可能である．この語を知らない人が，これを技術用語と分かるのは周囲の語（共起語）との関係を見て初めて判断できるのではないだろうか．

現時点では，計算コストが高いため，共起語を判別に用いることは考えていない．しかし，どれくらいの割合でこのような問題を生じる用語が存在するかを把握した結果，対応が必要になる可能性があるので，今後も検討を継続する．

4. 全体の考察と今後

本研究では，用語獲得と専門用語判別について，それぞれ実験を行って評価したが，本来の目的は応用アプリケーションへの辞書データ供給である．そのため，今後は，本研究で得られた専門用語辞書データをアプリケーションに反映することによって，アプリケーションのユーザが良くなったと感じたかどうかに関して，実装を行って実験・評価する予定である．これによって，用語獲得でいくつか試みた関数やパラメータのうちいずれを用いるのが最適かを決定することが可能となる．

また，本研究では，大学の工学系ホームページを対象とした実験のみであったが，今後様々なデータに対して同じ実験を行い，本研究の手法が有効であるかどうか確認していく予定である．

5. ま と め

本研究では，Web 文書集合から専門用語獲得を行った．まず統計的に用語を獲得し，その後，専門用語属性を判別するというステップで処理した．

統計的用語獲得に関しては，形態素 n-gram の統計・表層情報を利用し，時間的・物理的コストを考慮した．実験では，専門用語と判断できる複合名詞や名詞句などを獲得できた．

専門用語属性判別では，与えられた用語の構成要素のうち，末尾要素と，属性ラベルが専門用語である要素とに着目し，まず単純な属性拡張ルールを適用した．さらに，専門用語と判定されたものの構成要素を「属性影響語」とする概念を導入し，属性影響語に専門用語属性を仮設定して，再度属性拡張ルールを適用した．その結果，適合率が 6 割に低下したものの，再現率を 9 割近くまで引き上げることに成功した．

実験の結果，本手法により，応用アプリケーションに供給するための専門用語辞書データを Web 文書集合から獲得できることが分かった．

参 考 文 献

1) Webcat Plus. http://webcatplus.nii.ac.jp/
2) 事典検索システム Cyclone. http://cyclone.slis.tsukuba.ac.jp/
3) 産学連携支援ツール Bluesilk. http://www.bluesilk.biz/
4) 下畑さより，杉尾俊之：隣接文字情報を用いた n-gram 抽出文字列からの名詞句の自動抽出，情報処理学会研究報告，NL-114, pp.13–18 (1996).
5) 久光 徹，丹羽芳樹，辻井潤一：タームの representativeness を測る，情報処理学会研究報告，NL-133, pp.115–122 (1999).
6) 中川裕志，湯本紘彰，森 辰則：出現頻度と連接頻度に基づく専門用語抽出，自然言語処理，Vol.10, No.1, pp.27–45 (2003).
7) 田中久美子，山本真人，中川裕志：web 検索に基づく多言語動的 KWIC，情報処理学会研究報告，NL-152, pp.115–121 (2002).
8) Yamamoto, M. and Church, K.W.: Using Suffix Arrayes to Compute Term Frequency and Document Frequency for All Substrings in a Corpus, *Computational Linguistics*, Vol.27, No.1, pp.1–30 (2001).
9) 梅村恭司，真田亜希子：文字列を k 回以上含む文書数の計数アルゴリズム，自然言語処理，Vol.9, No.5, pp.43–70 (2002).
10) Church, K.W.: Empirical Estimates of Adaptation: The chance of Two Noriega's in close to p/2 than p2, *18th International Conference on*

Computational Linguistics, pp.180–186 (2000).
11) Takeda, Y., Yamamoto, E. and Umemura, K.: Determining Indexing Strings with Statistical Analysis, *IEICE Trans. Information and Systems*, Vol.E86-D, No.9, pp.1781–1787 (2003).
12) Kageura, K. and Umino, B.: Methods of Automatic Term Recognition: A Review, *Terminology*, Vol.3, No.2, pp.259–289 (1996).
13) Frantzi, K. and Ananiadou, S.: Extracting Nested Collocations, *16th International Conference on Computational Linguistics*, pp.41–46 (1996).
14) 形態素解析システム 茶筌. http://chasen.aist-nara.ac.jp/index.html.ja
15) 九津見毅，吉見毅彦，小谷克則，佐田いち子，井佐原均：サポートベクターマシンを用いた対訳表現の機械翻訳辞書登録適切性の自動判定，言語処理学会第 11 回年次大会 (2005).
16) 大沼宏行，池野篤司：ホームページやメールを対象とした質問応答システム，情報アクセスのためのテキスト処理シンポジウム発表論文集，pp.89–95 (2003).
17) 池野篤司，濱口佳孝，山本英子，井佐原均：属性影響語を用いた専門用語判別，情報処理学会研究報告，NL-168, pp.87–92 (2005).

(平成 17 年 10 月 17 日受付)
(平成 18 年 4 月 4 日採録)

池野 篤司（正会員）

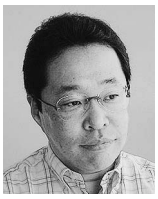
1991 年神戸大学大学院工学研究科システム工学専攻修士課程修了．同年沖電気工業株式会社入社．自然言語処理の研究開発に従事．言語処理学会会員．

濱口 佳孝（正会員）

1991 年名古屋大学大学院理学研究科物理専攻修士課程修了．同年沖電気工業株式会社入社．画像認識，情報検索の研究開発に従事．日本物理学会会員．

山本 英子（正会員）

1998 年豊橋技術科学大学大学院工学研究科情報工学専攻修士課程修了．2002 年同大学院工学研究科電子・情報工学専攻博士後期課程修了．博士（工学）．現在，独立行政法人情報通信研究機構自然言語グループ有期研究員．自然言語処理，情報抽出の研究に従事．言語処理学会，人工知能学会各会員．



井佐原 均（正会員）

1980 年京都大学大学院工学研究科修士課程修了．博士（工学）．同年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所．現在，独立行政法人情報通信研究機構知識創成コミュニケーション研究センター自然言語グループリーダー，同アジア研究連携センター自然言語ラボラトリー長．自然言語処理，語彙意味論の研究に従事．言語処理学会，人工知能学会，日本認知科学会各会員．