社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2004－NL－164 (29)
2004／11／6

検索支援に向けたWeb文書集合からの用語獲得

山本 英子† 池野 篤司‡ 濱口 佳孝‡ 井佐原 均†

† 独立行政法人 情報通信研究機構 〒619-0289 京都府相楽郡精華町光台 3-5
‡ 沖電気工業株式会社 〒541-0053 大阪府大阪市中央区本町 2-5-7 丸紅ビル 4 階
E-mail: † {eiko, isahara}@nict.go.jp, ‡ {ikeno546, hamaguti662}@oki.com

あらまし 本研究では，検索を支援するために，検索キーワードとなりうる用語をユーザに提示することを目的として用語獲得を行う．具体的には，形態素解析の誤りを修正することを含め，単語 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ながら，Web 文書集合からの用語獲得を試みる．実験では，既存の専門用語辞書を追加し解析された形態素を用いた場合と，その辞書を追加せず解析された形態素を用いた場合とを考える．そして，辞書を追加しなかった場合，検討した手法がコーパス中に出現する，専門用語辞書に登録されている専門用語をどの程度再現できるかを測ることによって，本研究で目的としている用語の獲得に有用な手法を推察する．その結果を踏まえ，辞書を追加した場合に獲得された用語を分析し，用語獲得の精度改善について考える．

キーワード 検索支援, Web 文書, 用語獲得, 統計的手法

Term Acquisition from Web Document Collection toward Retrieval Supports

Eiko YAMAMOTO† Atsushi IKENO‡ Yoshitaka HAMAGUCHI‡ Hitoshi ISAHARA†

† National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan
‡ Oki Electronic Industry Co., Ltd.
4F Marubeni Building, 2-5-7 Honmachi, Chuo-ku, Osaka-shi, Osaka, 541-0053 Japan
E-mail: † {eiko, isahara} @nict.go. jp, ‡ {ikeno546, hamaguti662}@oki.com

Abstract In this paper, we try to acquire the terms which support for retrieval from Web document collection, to propose them as helpful words to users. To do this, we use the statistic and the surface information in the corpus. In this experiment, we estimate the terms from n-grams consisting of the morphemes analyzed by the system with or without the technical term dictionaries. In the case without the dictionaries, we guess the method for acquiring terms most successfully by comparing recalls for entries in the dictionaries. Finally, we observe acquired terms in the case with the dictionaries and consider how our term acquisition can be improved.

Keyword Retrieval Supports, Web Document Collection, Term Acquisition, Statistical Method

1. はじめに

インターネットの普及に伴い，Google などを代表とする検索エンジンを使って，ユーザが多くの情報を得ることができるようになった．そのとき，ユーザは検索エンジンにキーワードまたは自然文を入力することで得たい情報を持つ Web ページを検索する．しかし，入力したキーワードが当を得たものでない場合，別のキーワードを考え，再度検索する必要がある．このとき，得たい情報について知識が浅いユーザにとって，別のキーワードを思いつくということは難題である．このような場合，検索対象となる Web 文書集合にある特徴的な用語リストはユーザにとって検索するための有用な情報となりうる．また，その用語リストから用語を選択してもらうことで，ユーザの意図と検索対象コーパスの語彙の違いを緩和することができ，検索精度の向上が期待できる．

本研究では，このような場合に，ユーザの補助となる用語を Web 文書集合から獲得することを目的とする．我々は，この用語を Bluesilk®[7, 8][1]に搭載された検索支援の機能を強化するために利用することを想定

[1] Bluesilk®は(株)三菱総合研究所の登録商標である．

している．Bluesilk®はこれまで，専門用語辞書を追加し，その辞書に含まれる用語を使って検索支援を行ってきた．その用語にはそれぞれ属性ラベルが付けられ，ユーザが指定する属性を持つ用語を提示する形を取っている．しかし，様々なWeb文書集合に，このような用語を人手で追加していくことは困難であり，自動的に用語を獲得することが必要となる．そこで，Web文書集合から特徴的な用語を自動的に獲得し，その用語に属性ラベルを付与し，辞書に追加することを考えた[11]．しかし，解析誤りも通常と比べ多いWeb文書を対象とするには，解析誤りにより分解される用語への対応や，適切な新規語の抽出を行う必要がある．本研究では，このような課題を解決し，検索支援となりうる用語をWeb文書から獲得することを目的とする．

2. Bluesilk®の概要

Bluesilk®は専門性や研究内容に即した効率的な産学連携を支援するシステムとして開発された[7, 8]．Bluesilk®は検索要求のテキストが入力されると，Webページや論文・特許などの文書集合から入力内容に関連する文書を検索する．さらに文書集合から指定された属性を持つ語だけを抽出し，リストアップできる．システムはデータ収集部，情報抽出部，文書検索部，統計処理部からなる．

3. 用語獲得の概要

本研究で獲得する用語はBluesilk®の性能改善のために，情報抽出部と文書検索部で利用することを想定している．利用方法は，情報抽出部では，獲得した用語の中で，固有表現を中心とした属性ラベルを付与できるものはデータベースに登録し，文書検索部では，検索キーワードとしてユーザに提示することである．本研究では，これらの利用方法を考慮し，網羅的に多くの用語を獲得することを目的とする．

3.1. 獲得する用語

ユーザが検索の際に入力するのは多くの場合，得たい情報のキーワードとなる名詞である．複合名詞の専門用語抽出についてはこれまでに高い性能を持つ手法が提案されている[9]．しかし，検索要求の入力としては，複合名詞に限らず，助詞などを含む内容を表す項目やタイトルなどの名詞句も考えられる．たとえば，「ファイルの削除」といった形式のものである．本研究では，このような文字列も検索支援可能な用語として考慮する．したがって，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検索に役立ちうる特徴的な文字列」とした．

3.2. 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捉えることにより，用語らしい文字列を抽出することを試みる．候補となる文字列は形態素を単位とする単語n-gramとした．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[4, 6]．本研究でもこの手法を利用する．しかし，効率的な手法を利用しても，コーパスの規模と計算機のパフォーマンスに実行効率は依存する．そこで，形態素解析システムを利用し，取り扱う列の最小単位を文字ではなく文字列の最小単位を形態素とする．これによって対象となる要素数を減らし，作業領域と計算コストを削減することで，日々増加するWebデータに備えることとした．

4. 獲得手法

4.1. 使用する統計量の定義

ここでは，本研究で使用した統計量を示す．ここで任意のn-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) : log N - log(df(Z))
cd(Z) : Zの左または右に連接する形態素の異なり数

をコーパスからの統計量として用いる．

4.2. 工程手順

用語獲得は1)候補の選定，2)用語の推定，3)用語の絞込みの三つの工程で行う．

4.2.1. 候補の選定

本研究では，候補とすべきn-gramは文書集合からの検索に役立つ，対象文書集合において特徴的なn-gramとする．

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が考えられる．ある文書の内容を代表する内容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[2]．この特徴により，df2/dfは語分割などに適用されている[10]．また，分野において特徴的な用語は文書集合の特徴を捉える用語であるならば，ある特定の文書だけに極端に集中するのではなく，ある程度散らばって出現する．

そこで，集中度を測る指標としてdf2/df，分布度を測る指標としてdf/cfを利用し，候補の選定を行うこととした．また，先頭または最後尾に「助詞」または「副詞」を持つn-gramは候補から除去した．

4.2.2. 用語の推定

本研究では，上で選定された候補のうち，下記の二つの条件を満たすn-gramを用語と推定する．

各n-gramを$S_{0n}=M_0...M_n$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 $M_i...M_j$ とした場合，

- $R<j\leqq n$ の間， $RScore(S_{0j}) \geqq RScore(S_{0j-1})$
- $0\leqq i<n-L$ の間， $LScore(S_{in}) \geqq LScore(S_{i+1n})$

ただし，R,L は $0\leqq R,L<n$ 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．用語の候補である n-gram について，R=0 の場合，先頭 M_0 から後方に一形態素ずつ伸ばした部分形態素列がすべて Rscore に関する条件を満たすかどうかを調べる．そして，L=0 の場合，末尾 M_n の形態素から前方に一形態素ずつ伸ばした部分形態素列がすべて Lscore に関する条件を満たすかどうかを調べる．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_0 = "まち"，M_1= "づくり"，M_2= "事業" であり，R=0，L=0 のとき，

$RScore(S_{01})$と $RScore(S_{00})$が

RScore(まち_づくり)≧RScore(まち)，

$RScore(S_{02})$と $RScore(S_{01})$が

RScore(まち_づくり_事業)≧RScore(まち_づくり)

であり，かつ

$LScore(S_{12})$と $LScore(S_{22})$が

LScore(づくり_事業)≧LScore(事業)

$LScore(S_{02})$と $LScore(S_{12})$が

LScore(まち_づくり_事業)≧LScore(づくり_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 Rscore，Lscore について，下記の 5 つの関数を検討した．

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

3 の関数は田中らの関数である[5]．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[1]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．それぞれの項は長さ，出現頻度，連接する形態素の異なり数に関する項である．本研究では，長さと出現頻度に関する特徴を両方考慮される idf(Z)や，連接する形態素の異なり数は出現頻度に影響されるため，その影響を軽減する cd(Z)/(cf(Z)+cd(Z)) を用いて，4 つの変形関数を作成し，得られた結果を比較する．

4.2.3. 用語の絞込

本研究では，推定された用語について，以下の方法で推定結果を絞り込む．

- $S_{0n}=M_0...M_n$ に関して，$df(M_0)/N>\alpha$ であり，かつ $df(M_n)/N>\beta$ であるならば，S_{0n} を削除する．閾値は一形態素の統計量に基づき，設定する．

5. 実験

実験では，用語を推定する工程で用いるパラメータ R,L とスコア関数の組合せによる 10 種類の手法について，検討した．ただし，パラメータ R,L は同じ値 K を設定し，0 と n-1 の場合を検討した．K=0 の場合，各 n-gram において，すべての部分形態素列について連接関係を調べる．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または末尾の形態素を落とした部分形態素列が候補 n-gram より特徴的な形態素列であるかを調べる．表 1 に検討手法と以降で用いる識別子を示す．

表 1 検討手法：パラメータとスコア関数の組合せ

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

5.1. 実験コーパス

Bluesilk®が対象としているコーパスは，全国の大学が持つ Web サイトから html ファイルを収集したものである．収集の際には，執筆者や内容によるファイルの選別を全く行わないため，研究に関係のない個人ページも含まれている．実験では，このコーパスの一部である 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 200MB)を対象とした．

5.2. 形態素解析

本研究の基となる Bluesilk®は全文検索を行う際に，日本語形態素解析システム『茶筌』[3]に既存の専門用語辞書を追加したものを利用している．実験では，そのシステムで形態素解析された結果をそのまま用いることとした．品詞情報が付与されていない形態素の列だけを用いる．

5.3. 実験結果

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．各手法が用語とした n-gram の数である．()内の数は他の n-gram の部分形態素列であるものを削除した数である．

表 2 各手法によって推定された用語の数

手法	推定数	手法	推定数	差 (B-A)
A1	5293(5233)	B1	6205(6012)	912(779)
A2	5282(5226)	B2	10008(9640)	4726(4414)
A3	1194(1185)	B3	1778(1695)	584(510)
A4	12482(12064)	B4	13409(12727)	927(663)
A5	14173(13670)	B5	15216(14347)	1043(677)

6. 比較・検討

6.1. パラメータ K に関する比較

K=0 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 が共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，多くの用語が獲得できるという点も大きな要因である．

6.2. 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二形態素以上に分割される専門用語 1646 個とした．表 3 に再現率と適合率を示す．

表 3 既存の専門用語の再現率と適合率

手法	推定数	専門用語数	再現率	適合率
B1	7823	1052	0.6391	0.1345
B2	11580	1351	0.8208	0.1167
B3	1921	223	0.1355	0.1161
B4	15068	1405	0.8536	0.0932
B5	16960	1436	0.8724	0.0847

表 3 より，再現率は推定された用語の数が多いほど，高い傾向にあることがわかる．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．

1. 獲得された用語の部分文字列である用語
2. 部分文字列が獲得された用語

1 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，2 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 の適合率は，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．この値から，どの尺度においても，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わかる．また，B1 が他の手法に比べ，既存の専門用語が占める割合が高い．一般的に，専門用語はコーパス中の特徴を表す用語である．このことから，手法 B1 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文書頻度 df2 の高い上位M件における専門用語の再現率と適合率を表 4 に示す．これは，コーパス中の特定の文書を特徴付ける用語ならば，文書頻度 df2 が高い傾向にあることから，推定結果を df2 の降順にソートしたときの上位 M 件を対象に，同じ数の推定結果について比較を行う．M は B3 で得られた用語数 1921 とした．

表 4 上位 1921 件に関する再現率と適合率

手法	専門用語数	再現率	適合率
B1	415	0.2521	0.2160
B2	397	0.2406	0.2067
B3	223	0.1355	0.1161
B4	378	0.2296	0.1968
B5	337	0.2047	0.1754

これらの値も実験結果すべてを対象としたときの適合率と同様に，既存の専門用語をどの程度再現できるのかを示す値であり，特徴的な用語を獲得する手法の性能を示す値ではない．そのため，既存の専門用語を再現するという問題において手法を比較するための値でしかない．

表 4 より，再現率，適合率ともに B1 が他の手法と比べ，高い値を示す．専門用語は他の用語に比べ，文書を特徴付けるため，文書頻度 df2 も他の用語と比べ，高い場合が多い．このことから，B1 がコーパス中に出現する高い文書頻度 df2 を持つ専門用語を特徴的な用語として他の手法よりも高く評価することを表している．再現できた用語は以下ようなものである．

1. 複合名詞
2. 解析誤りによって分解された用語

1 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，2 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．

7. 考察

5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観られた．

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．

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，学会名や学会誌に関連する用語などが多く観られた．これらは，専門用語辞書や機関名の辞書の強化に利用できる．

Virtual_モード
Voronoi_図
WWW_情報_空間
Web_検索
Windows_版
かな漢字_変換
ペット_ロボット
ロボット_アーム
卓上_全身_型_ロボット
中心_市街地_活性化_法
夏休み_航空宇宙工学_教室
不特定_話_者_音声認識
副_格子_交換_エピタキシー
保証_付き_数値_計算
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_冷却_高速炉
超_臨界圧_軽水_冷却_炉
超_臨界圧_軽水炉
日本_バーチャル_リアリティ_学会
日本_ロボット_学会_誌
日本_音響_学会_誌
日本_気象_学会
日本_建築家_協会
日本建築学会_環境工学_委員_会
日本_原子力_学会_誌

特徴的な文字列から始まる用語

以下に獲得された「まち_づくり」から始まる用語の一部を示す．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，検索支援に役立つだろう．上記に示した「超_臨界圧」も同じように利用できる．

まち_づくり_NPO
まち_づくり_ファンド
まち_づくり_事業
まち_づくり_手法
まち_づくり_条例
まち_づくり_組織

解析誤りによって分解された用語

以下に獲得された解析誤りを含む用語の一部を示す．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を示す．

F_TP_サイト
H_am_ilton_の_定理
P_DF_ファイル
モー_ダル
マイク_ロメ_カニ_クス
お_台_場
き_裂_伝播
二次元_アイ_ソ_パ_ラメ_トリック_要素

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．「Hamilton の定理」もその一つである．実験では，以下のような表現を得た．これらは一部であるが，タイトルもあれば，専門用語と判断できる用語も含まれている．このことから，複合名詞の専門用語だけでなく，句のような専門用語も獲得できることが示唆される．

テンソル_と_ベクトル_の_ドット積
テンソル_の_微分_と_積分
Virtual_Job_Shop_Floor_の_構築
Web_で_の_文字_コード_の_設定
積分_点_の_重み_係数
図書_室_における_情報_サービス_と_業務_電算_化
触覚_フィードバック_を_用い_た_最適_把握_行動
二つ_の_領域_分割_図_の_適合度_評価
レンズ_付き_フィルム_の_分解
揺らぎ_の_非_平衡_緩和

パイプライン_における_亀裂_発生_時_の
_ガス_減圧_特性
ショ_糖_溶液_を_用い て_比重_分_画し_た
_実_下水処理
噴流_による_矩形_容器_内_自励_スロッシング
_の 発生_機構
噴流_による_自励_スロッシング
シビア_アクシデント_の_伝熱_流動_現象
_における_素_過程

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，ユーザへの提示や，辞書の強化にも利用できないが，このような用語が現れる文書を検索対象から除き，検索対象となる文書を絞ることに役立ち，他の種類の用語とは異なる支援に利用できるだろう．

人名

Bluesilk®は専門性や研究内容に則した研究者を探索する必要がある．実験では，専門用語や研究項目などのほかに，直接検索性能に係る多くの人名が獲得された．獲得された人名に属性ラベルを付与することで，効率的かつ正確な検索を実現でき，Bluesilk®の検索性能向上が見込まれる．

その他

上記の 6 種類の用語の他に，的，階，回，次，間，系などの接辞から始まる用語が結果として得られた．これは，それらの接辞の前に連接する形態素が数やサ変名詞などで，種類が多いため，接辞以降の n-gram が定型文字列となり，つながりが強い文字列と観られる．また，このような定型文字列がコーパス中で統計的に特徴を持つものであったため，検索に役立つ用語ではないかと判定された．これは，接辞であるという情報を使うことで除くことができる．

また，本手法では，同じ文字種の形態素が連接するならば連結するという規則を使用していないが，分解された同じ文字種の形態素を含む n-gram を獲得している．これは，特徴的な n-gram であれば，連結規則を使用せずに獲得できることを示している．連結規則を使用することによって，用語推定の工程にかかる時間的コストを削減できる．ただ，この規則は言語に依存する．

8. まとめ

本研究では，検索を支援するために，検索キーワードとなりうる用語をユーザに提示することを目的として Web 文書集合から用語獲得を行った．本手法は，形態素解析の誤りを修正することを含め，単語 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．そして，既存の専門用語辞書を利用しない形態素解析結果から用語を獲得した場合の既存の専門用語の再現に関する比較評価から，検討した手法のなかで手法 B1 が本研究で目的とする用語の獲得に適しているのではないかと推察した．実際は，獲得した用語への属性ラベル付与についても評価する必要がある．この評価が目下の課題である．そして，その評価を踏まえ，実験結果を詳細に分析し，正確な用語を獲得するための知識を得ることで，用語獲得性能を改善していくことが今後の課題である．

謝辞

本研究では，情報通信研究機構の内山将夫氏によって作成されたシステムを利用させていただきました．深く感謝致します．

文　献

[1] K. Frantzi and S. Ananiadou, "Extracting Nested Collocations.", COLING96, pp41--46, 1996.

[2] K. W. Church, "Empirical Estimates of Adaptation: The chance of Two Noriega's in close to p/2 than p^2," Coling2000, pp.180--186, 2000.

[3] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原 正幸，"日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書"，Dec. 2000.

[4] M. Yamamoto and K. W. Church, "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.

[5] 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，自然言語処理研究会，Vol.152, No.17, pp.115--121, Nov. 2002.

[6] 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.

[7] 中村達生，産学連携支援ツール(Bluesilk)の仕組み，情報管理，Vol.46, No.7, pp.455--462, Oct. 2003.

[8] 産学連携支援ツール Bluesilk, http://www.bluesilk.biz/

[9] 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.

[10] Y. Takeda, E. Yamamoto, K. Umemura, "Determining Indexing Strings with Statistical Analysis," IEICE Transactions on Information and Systems, Vol. E86-D, No.9, pp.1781--1787, Sep. 2003.

[11] 池野篤司，濱口佳孝，山本英子，井佐原均，"統計的に獲得された用語への属性ラベル付与"，情報処理学会 研究報告，自然言語処理研究会，Vol.164, No.31, (電子情報通信学会言語理解とコミュニケーション研究会), Nov. 2004.