

## 語彙空間の構造に基づく専門用語抽出

辻河 亨† 吉田 稔‡ 中川 裕志‡

† 東京大学大学院 総合文化研究科

‡ 東京大学 情報基盤センター

E-mail: {tjkawa, mino}@r.dl.itc.u-tokyo.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp

文書中から用語抽出を行う手法としては様々なものが提案されているが、多くはコーパスにおける用語候補の使用における統計量を利用するものであった。

本稿では、これまで光が当てられていなかった用語を構成する語基間の関係について、そのグラフ構造に着目する。まずグラフにおける局所的ないし大域的構造を反映する用語抽出の指標を新たに提案し、その上で提案した用語抽出のいくつかの指標、およびそれらの組み合わせを、複数のコーパスを用いて実験的に評価した結果について報告する。

## Term Extraction based on the structure of terminology set

†Tohru Tsujikawa ‡Minoru Yoshida ‡Hiroshi Nakagawa

†University of Tokyo, Graduate School of Arts and Sciences

‡University of Tokyo, Information Technology Center

E-mail: {tjkawa, mino}@r.dl.itc.u-tokyo.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp

Many methods to evaluate termhood of word sequences extracted from documents have been proposed. The majority of them use the statistics of term candidates appearing the corpus. The ignored side, however, is the graph structure whose nodes are basic terms that consist of complex terms. That is exactly what we focus on in this paper. We propose term extraction algorithms which utilize this graph structure. More precisely, the proposed methods assume the whole set of compound terms as a graph and applies web mining method to term extraction. The nodes are element words, the edges coincide with modifying relations between words in the definition of graph, and the importance of terms are ones used to find web communities or in ranking pages by Web search engines. Finally we experimentally evaluate the proposed algorithms with terminology dictionaries on Web.

### 1 はじめに

専門用語辞典は初心者にとって種々の専門分野への取掛かりになるのはもちろん、その分野の専門家にとっても必携の言語資源である。このような専門用語辞典の基礎となる専門用語集合をコーパスから自動的に抽出し重要度の順にソートするアルゴリズムは、過去に多くの研究がなされてきた [1] [2]。

この処理の要は、「重要」と思われる順にソートする

ための指標である。重要度の判断は最終的には「専門家」に委ねられるが、統計処理による指標を導入し、その結果が重要度の順に並んでいれば人手による作業の負担を軽減できる。

重要度の指標は TFIDF [7]・C-Value [5]・FLR [9] など、先行研究によって既に多くの提案がなされているが、指標そのものと、指標の精度を評価する手法には改善の余地がある。

さて、このような改善における着目点のひとつに、従来の用語抽出方法で利用されてこなかった（用語候補集合内の）複合語を構成する語基の間の係り受け関係がある。この点を組織的に利用することを本稿では提案するが、そのためには用語候補集合の数理的モデル化が必要である。我々は、このために語基をノードとし、語基の間の係り受け関係をエッジとするグラフ構造を用いる。なお語基とは、それ以上分割できない基本的な単語を意味する。このグラフを以下では語基グラフと呼ぶ。

グラフ構造によるモデル化は他の分野では活用され成果をあげている。たとえば、Web の検索エンジン Google では Web ページをノードとし、それらの間のリンクをエッジと見たグラフ構造において、そのグラフを表す接続行列の固有値を利用した Web ページの重み付けを導入し成功した [3]。また、同じように接続行列を用いた HITS アルゴリズム [8] も Web コミュニティの発見で成果をあげた。本稿では、このような実績のあるアルゴリズムを語基グラフに適用する。これについては第 3 章で述べる。

さらにこれらに加え、従来から知られている用語性を計る指標を組み合わせていくつかの新規な用語性の尺度を提案する。また、正解集合を持つ唯一のコーパスとして従来用いられてきた TMREC コーパス [6] での精度比較にとらわれず、Web 上の用語辞典もテストコレクションに用いて実験的に評価を行った。これらの結果について第 4 章で述べる。

## 2 用語抽出の既存指標

用語には語基 1 つのみから成るものと、複数の語基から構成されるもの（複合語）が存在する。例えば「専門用語抽出」という用語は、「専門」「用語」「抽出」の 3 つの語基から成る。専門的文章には複合語の専門用語が出現することが多く、用語抽出において複合語の扱いは重要である。

既存指標のうち TF や DF は語基と複合語の関係情報を利用しないのに対し、C-Value [5] や LR [9] [10] はこの情報を利用している。

本章では、まずこれら既存の指標について説明する。なお、複合語  $W$  を、語基を  $w_i$ 、語基数を  $n$  とし、 $W = w_1w_2\dots w_n$  と記述する。以下、語基 1 つの用語も  $n = 1$  の複合語として表記する。

### 2.1 Term Frequency: TF

TF は最も基本的な指標で、文書中における用語の頻度（出現回数）を利用する。

$$tf(W) = (\text{全文書中に } W \text{ が出現した頻度}) \quad (1)$$

$tf(W)$  が大きい用語ほど重要と考えられるが、 $tf(W)$  が大きすぎる語（機能語など）は逆にストップワードとして除く必要がある。

$tf(W)$  は、 $W$  がより長い複合語中に含まれる形で出現した回数を含む。 $W$  そのものが単体で出現した回数は  $f(W)$  と定義する。

$$f(W) = (\text{全文書中に } W \text{ が単体で出現した頻度}) \quad (2)$$

### 2.2 Document Frequency: DF

DF はコーパスが複数の文書から成る場合に使われる指標で、用語が何個の文書中に出現したかを示す。

$$df(W) = (W \text{ が出現した文書数}) \quad (3)$$

DF は一般的には逆数を TF と積を取って指標にする（“ $TF \cdot IDF$ ”）が、この場合は以下の式を用いることが多い。

$$tf \cdot idf(W) = tf(W) \times \left( \log \frac{\text{総文書数}}{df(W)} + 1 \right) \quad (4)$$

### 2.3 C-Value

C-Value は、TF に語基数と部分文字列の性質を取り入れた指標であり、以下のように定義される。

$$t(W) = (W \text{ を部分文字列とする用語の頻度})$$

$$c(W) = (W \text{ を部分文字列とする用語の種類数})$$

とするとき<sup>\*1</sup>、C-Value の値は

$$C\text{-val}(W) = (n - 1) \times \left( tf(W) - \frac{t(W)}{c(W)} \right) \quad (5)$$

と定義される。

C-Value は  $n = 1$  のとき 0 になり、適切な重要度を示さない<sup>\*2</sup>。 $n = 1$  にも対応する手法が MC-Value 法 [11] である。

$$MC\text{-val}(W) = n \times \left( tf(W) - \frac{t(W)}{c(W)} \right) \quad (6)$$

<sup>\*1</sup>  $t(W)$  には、 $W$  としての出現頻度を含まない。

<sup>\*2</sup> もともと C-Value は collocation 抽出のための手法であり、複数の語から成ることを前提にしている。

## 2.4 LR

LR は、語基の左右にどれだけ語基が来やすいかを利用した指標である。詳細は 3.2 節で説明する。

## 3 グラフ構造に基づく指標

複合語では、主要部が後ろに来る (いわゆる head final) 日本語や英語の場合、先行する語基が後続の語基を修飾する関係が成立している (図 1)。

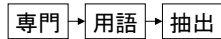


図 1 複合語内の係り受け関係

複合語が 3 つ以上の語基からなる場合、係り受け関係は曖昧になる。したがって複合語を構成する語基列から係り受け関係を決定する手段を講じる必要がある。本稿では、先行する語基が直後の語基を修飾していると近似する。用語候補の複合語集合は、語基間に係り受け関係を与えると、語基をノード、係り受け関係を有向のエッジとする有向グラフという構造を持つ。本文中に出現する全ての複合語を語基に分解して生成した有向グラフのことを、語基グラフと呼ぶことにする (図 2)。

語基グラフは通常の有向グラフの性質を満たしているため、有向グラフに対する数理的な解析手法を適用することができる。

### 3.1 ハイパーリンク構造との対応

周知のようにハイパーリンク構造は、Web ページをノード、リンクを有向のエッジとする有向グラフである。このハイパーリンク構造から情報を取り出す手法として、PageRank や HITS が知られている。これらの手法はサーチエンジンの検索結果として得られたページを「より重要な順に」ソートするためのアルゴリズムである。

語基グラフと Web ハイパーリンク構造はいずれも有向グラフであり、対応付けが可能である。

- Web ページ ⇒ 語基
- ハイパーリンク ⇒ 係り受け関係

と対応づける (図 3)。

語基グラフを用いた用語重要度の指標として、LR・HITS・PageRank について詳細に説明する。先取りし言えば、この 3 つの手法は、語基グラフにおける着目範囲の広さの違いとして捉えることができる。

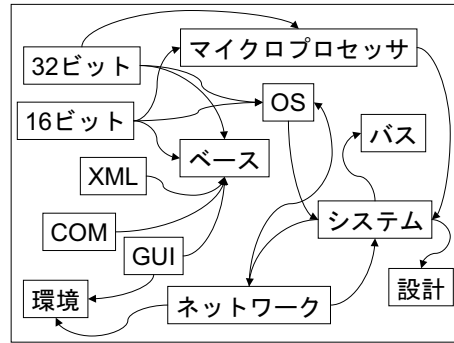


図 2 語基グラフの例

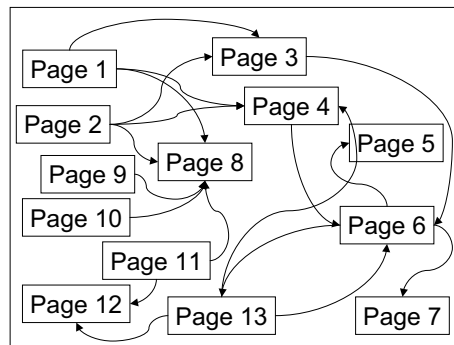


図 3 図 2 を Web ハイパーリンク構造に変換

### 3.2 LR

LR は、個々の語基について、その左右に来た語基の延べ頻度 (接続頻度 LR) または種類数 (接続種類 LR) から語基の重要度を求め、それらを用いて複合語の重要度を算出する手法である。

$$\begin{aligned}
 l(w_i) &= (w_i \text{の左に出現した語基の頻度 or 種類数}) + 1 \\
 r(w_i) &= (w_i \text{の右に出現した語基の頻度 or 種類数}) + 1 \\
 lr(w_i) &= \sqrt{l(w_i)r(w_i)}
 \end{aligned}$$

とすると (図 4)、 $W$  の LR 法による重要度  $LR(W)$  は以下の式で表される。

$$LR(W) = (lr(w_1)lr(w_2) \cdots lr(w_n))^{1/n} \quad (7)$$

語基を Web ページと考えると、 $l(w_i)$  はページに張られているリンク (in-link) の数、 $r(w_i)$  はページから出しているリンク (out-link) の数に相当する。 $lr(W)$  は in-link や out-link が多いほど重要な Web ページであるという仮定 [4] を数値化したものにあたる。

### 3.3 HITS

HITS は、全ての Web ページについて “Hub” と “Authority” の 2 種類の性質があると仮定し、Hub の度合

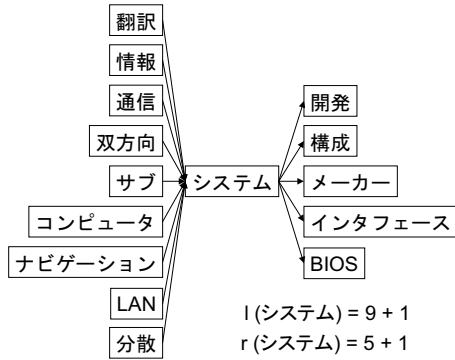


図4 LRによる語基の重要度

語基	Hub 値	Auth 値
32 ビット	0.576	0.000
16 ビット	0.576	0.000
マイクロプロセッサ	0.053	0.407
OS	0.053	0.520
バス	0.000	0.001
システム	0.001	0.151
設計	0.000	0.001
ベース	0.000	0.699
XML	0.247	0.000
COM	0.247	0.000
GUI	0.328	0.000
環境	0.000	0.229
ネットワーク	0.318	0.001

い(以下 Hub 値)値が大きいほど優秀なリンク元, Authority の度合い(以下 Auth 値)が大きいほど優秀なリンク先と考える。Hub 値・Auth 値は, 以下のように計算される。

1. 最初は全てのページ ( $i = 1, 2, \dots, n$ ) について

$$\begin{aligned} hub_i &= 1.0 \\ auth_i &= 1.0 \end{aligned}$$

2. Hub, Auth 値を更新

$$\begin{aligned} auth_i &= \sum_j hub_j \\ &\quad (j = \text{pages links to page}_i) \\ hub_i &= \sum_k auth_k \\ &\quad (k = \text{pages linked from page}_i) \end{aligned}$$

3. Hub, Auth 値を正規化

$$\begin{aligned} auth_i &= auth_i / \sqrt{\sum_{j=1}^n auth_j^2} \\ hub_i &= hub_i / \sqrt{\sum_{j=1}^n hub_j^2} \end{aligned}$$

4. 収束するまで 2~3 を反復

語基グラフに HITS アルゴリズムを適用すると, Hub 値が高くなるのは Auth 値の高い単語を頻繁に修飾する語基(リンクを多く出しているページ)で, Auth 値が高くなるのは多くの Hub 値の高い単語後から修飾される語基(リンクを多く受けているページ)である(表 1)。

表1 図2に HITS を適用した例

このことから, 用語  $W$  の HITS を用いた重要度  $HITS(W)$  を以下の式で定義する<sup>\*3</sup>。

$$\begin{aligned} H(w_i) &= (w_i \text{の Hub 値}) + 1.0 \\ A(w_i) &= (w_i \text{の Auth 値}) + 1.0 \\ HA(w_i) &= \sqrt{H(w_i)A(w_i)} \end{aligned}$$

$$HITS(W) = (H(w_1)HA(w_2) \cdots A(w_n))^{1/n} - 1.0 \quad (8)$$

ただし,  $n = 1$  の場合は以下の式を用いる。

$$HITS(W) = HA(w_1) - 1.0 \quad (9)$$

以上から分かるように, Hub 値・Auth 値は, あるノードと直接の係り受け関係を持つノード(親ノード・子ノード)だけでなく, 兄弟ノード, 子ノードの親ノード, の影響を受ける。ノード自身の情報しか用いない LR と比較すると, HITS は語基グラフを兄弟方向に大域的な視点からみる手法である。

### 3.4 PageRank

PageRank は, Web 全体について, 個々の Web ページの重要度がリンク先へ流れ込んでいく計算手法を取る(図 5)。Web コミュニティにおいては, PageRank の値が高いほど重要なページと考えられる。

PageRank を用いた用語  $W$  の重要度  $PAGE(W)$  を以下の式で定義する。

$$P(w_i) = (w_i \text{の PageRank 値}) + 1.0$$

<sup>\*3</sup> Hub 値と Auth 値は値域が 0.0 ~ 1.0 で, そのままでは複合語に 1 つでも Hub 値か Auth 値が 0.0 の語基が含まれると,  $HITS(W)$  は 0.0 になってしまう。この問題を前もって  $H(w)$  と  $A(w)$  を +1.0 することで回避しつつ,  $HITS(W)$  で -1.0 して最小値を 0.0 に戻している。PageRank でも同様の操作を行っている。

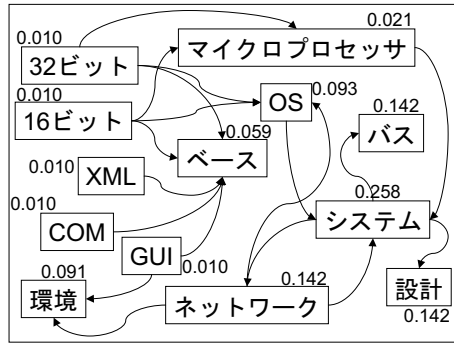


図5 図2にPageRankを適用した例

$$PAGE(W) = (P(w_1)P(w_2) \cdots P(w_n))^{1/n} - 1.0 \quad (10)$$

### 3.5 LR・HITS・PageRankの比較

LR・HITS・PageRankによる用語の重要度計算は、いずれも同じ語基グラフを利用しているが着目する広さが異なっている。

**LR** 語基グラフの個々のノードと、それに繋がるエッジのみの情報を利用しており、最もローカルな着目範囲の手法である。

**HITS** 個々のノードと、それに直接繋がっているエッジとノード、さらにもう1段先に繋がっているエッジとノードの情報までを利用しており、LRよりもグローバルな着目範囲をもつ手法である。

**PageRank** 語基グラフ全体のノードとエッジの情報を利用しており、最もグローバルな着目範囲の手法である。

LR・HITS・PageRankの3種を用いることで、語基グラフのローカルな性質からグローバルな性質までを利用することができる。

## 4 実験

テストコレクションには、Webで公開されているコンピュータ用語辞典を用いた。

- IT用語辞典 e-Words<sup>\*4</sup>
- アスキーデジタル用語辞典<sup>\*5</sup>
- The Business Technology Network (英文)<sup>\*6</sup>
- FOLDOC (英文)<sup>\*7</sup>

<sup>\*4</sup> <http://e-words.jp/> ©Incept

<sup>\*5</sup> <http://yougo.ascii24.com/> ©ASCII Corp.

<sup>\*6</sup> <http://www.techweb.com/> ©CMP Media LLC

<sup>\*7</sup> <http://foldoc.doc.ic.ac.uk/foldoc/> ©Denis Howe

また、比較のため TMREC タスクコーパスでも実験を行った。

コンピュータ用語辞典では、用語の語義文のみをコーパスとし、用語の見出し語のうち語義文中に出現する用語の集合を正解集合とした。

Web上の用語辞書のような電子辞書をテストコレクションにすることは、

- 擬似的な正解集合を備えた多数のテストコレクションを入手できる。
- 用語辞書であれば何でもテストコレクションにできるため、分野を問わず実験が可能になる。
- 特定の分野について、日本語に限らず様々な言語のテストコレクションを入手できる。

といった長所があるが、

- 重要な用語が必ずしも見出し語（正解集合）になっているわけではない。

といった短所も持つ。

### 4.1 重要度指標の組み合わせ

3章までに示した指標は、いずれも異なった視点から用語の重要度を定義するものであり、組み合わせればより優れた指標となりうる。

本章で示すのは  $tf \cdot idf(W)$ ,  $MC-Val(W)$ ,  $LR(W)$ ,  $HITS(W)$ ,  $PAGE(W)$  (それぞれ式(4)(6)(7)(8,9)(10))と、それらを組み合わせた以下の式の値である。

$$FLR(W) = f(W) \times LR(W)$$

$$FLRH(W) = FLR(W) \times HITS(W)$$

### 4.2 実験結果

それぞれのテストコレクションについて、第4.1節で挙げた値のうち特徴的な傾向を示すものについて実験結果を示す。

日本語コーパスは、形態素解析システム茶筌<sup>\*8</sup>による形態素解析結果から名詞・未知語の連続を抽出し、用語候補とした。ただし、英数字列・接頭語・接尾語については連結する前処理を行った。

英語コーパスについては、ストップワードで区切られた単語列を用語候補として用いた。

用語候補は各手法による重要度の高い順に並べる。ある手法  $M$  の重要度による上位  $x$  個の候補のうち正解

<sup>\*8</sup> 奈良先端技術大学院大学 情報科学研究科 松本研究室 <http://cl.aist-nara.ac.jp/>

集合に含まれる用語が  $y$  個存在することを,  $M$  は上位  $x$  個で正解数が  $y$  個と, と本稿では表記する。本節で示すグラフは横軸が  $x$ , 縦軸が  $y$  または  $y$  の差である。

#### 4.2.1 TMREC における実験結果

TMREC タスクコーパスでの実験結果を示す(図6)。手法間の精度差を見やすくするため, 以後は正解数の HITS に対する差で示す(図7)。

MC-Value の精度が参考文献 [11] と異なるのは, 接頭語・接尾語を隣接する未知語・名詞に連結する前処理を行った結果, 部分文字列のスコアが異なったためである。

HITS と PageRank は語基グラフを大域的に利用するアルゴリズムであることから, LR よりも精度が向上するかの印象を受けるが, TMREC での用語抽出において良いスコアは得られなかった。

例えば, 「システム」のような一般的な単語が, 大域的な情報を考慮したことによって強くなりすぎてしまっている。このことは, 語彙集合におけるグラフ理論的性質の利用における重要な知見である。すなわち, 語彙集合においてはグローバルな情報よりも, 近接するノード間に局在する情報がより有益といえる。

しかしながら, 上位 1000 位までの範囲で比較すると, FLR と HITS の組み合わせである  $FLRH(W)$  は, 既存手法よりも高い精度を示した(図8)。

FLRH の局所的な精度向上は, FLR の上位候補のうち HITS で下位候補となるものが, 積を取ったことで大きく順位を落としたことによる。手法間での順位の違いを利用する研究は既存であるが [12],  $FLR(W)$  が  $LR(W)$  に  $f(W)$  を組み合わせて精度を向上させていることと同様に, 他の指標も有効に組み合わせることで精度向上の余地があると考えられる。

#### 4.2.2 Web 辞書における実験結果

Web 辞書をテストコレクションにした実験結果を示す(図9,10,11)。FOLDDOC は TechWeb と傾向がほぼ同様であったため, グラフを省略した。

TMREC での比較結果と大きく異なり, 各テストコレクションとも TFIDF が最も良い精度を示している。

TMREC コーパスでの正解集合は, 文章中出现する用語のうち一般的な単語を除いた集合に近いものである。本実験で TMREC コーパスからプログラムが抽出した全候補数 17870 のうち 7741 語が正解であった(正解集合は 8835 語)。

Web 辞書テストコレクションの正解集合は辞書の見

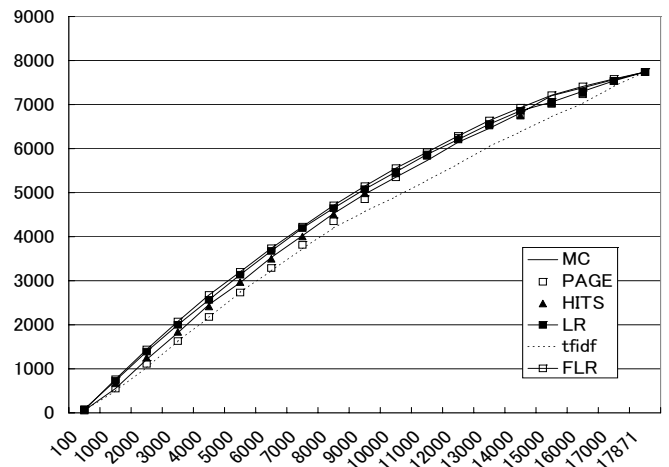


図6 TMREC: 各手法の候補数(X軸)と正解数(Y軸)

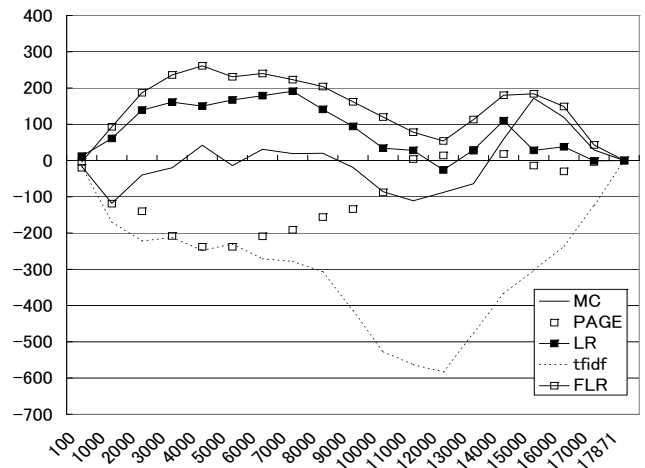


図7 TMREC: 各手法の候補数と HITS との正解数差

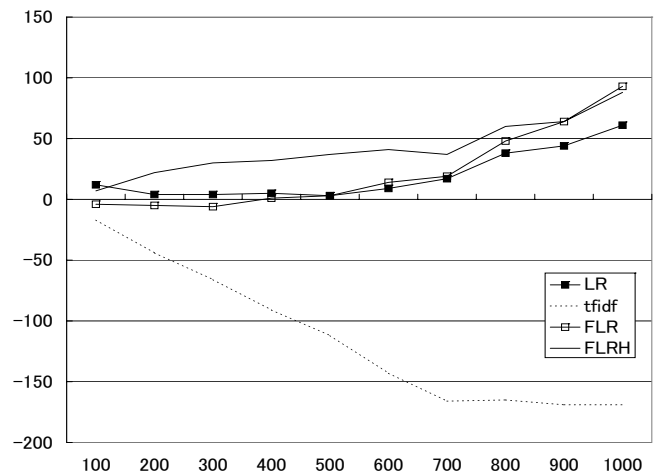


図8 TMREC: 各手法の正解数の差(上位1000位)

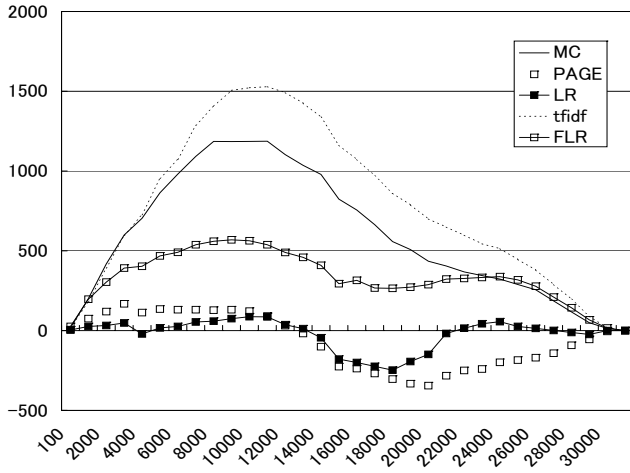


図9 e-Words: 各手法の候補数と正解数の差

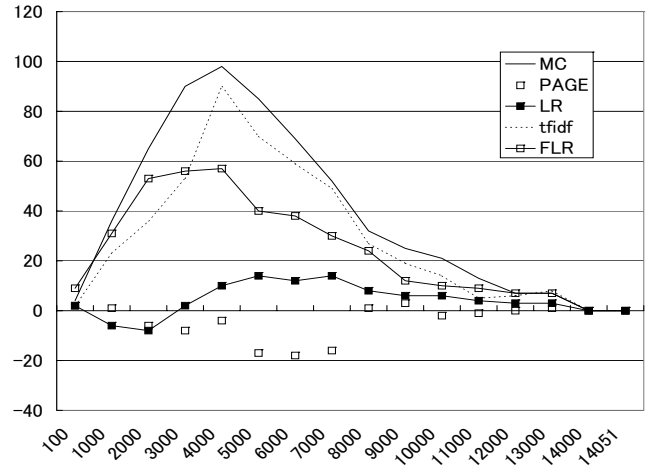


図12 e-Words: 漢字字種用語の正解数差

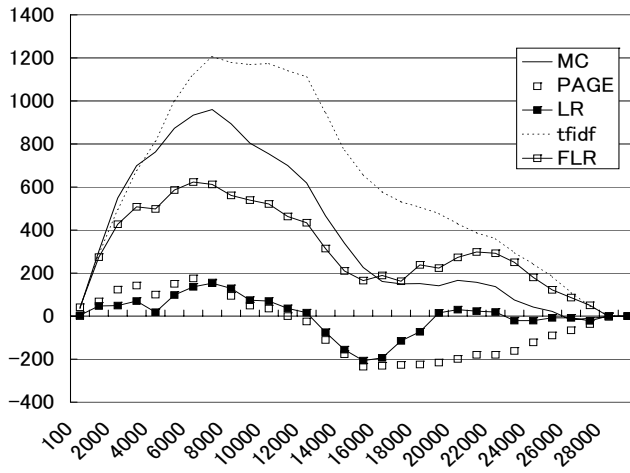


図10 ASCII: 各手法の候補数と正解数の差

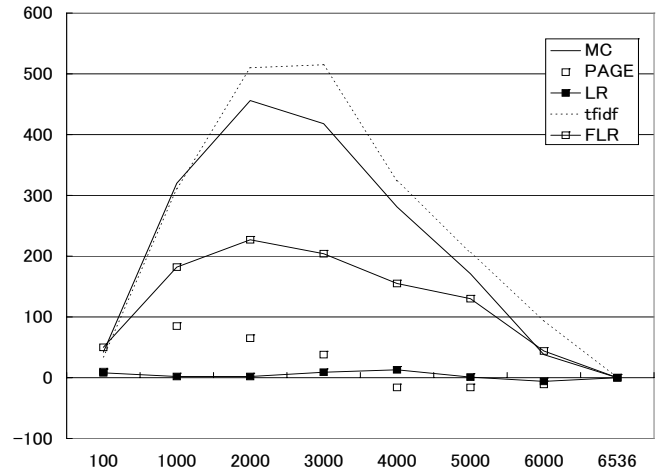


図13 e-Words: カタカナ字種用語の正解数差

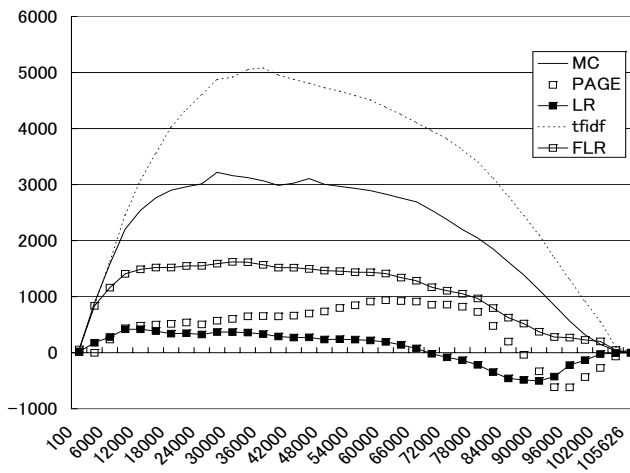


図11 TechWeb: 各手法の候補数と正解数の差

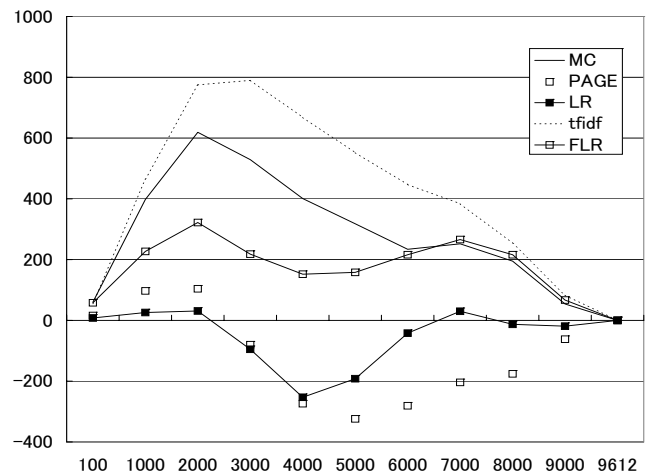


図14 e-Words: アルファベット字種用語の正解数差

出し語であり，用語候補数に対して数が非常に少ない。例えば e-Words では，プログラムが抽出した全候補数 30197 のうち，正解集合に含まれていたのは 3162 語に過ぎなかった<sup>9)</sup>。

上の結果が示すように，TMREC コーパスと Web 辞書コーパスではコーパスと正解集合の性質が大きく異なっていると考えられる。TMREC コーパスで高い精度を示す手法が一般的に優れた用語抽出手法であるとは言えず，目的とするコーパス・正解集合に応じて最適な手法を選択すべきである。

#### 4.2.3 文字種別実験結果

Web 辞書テストコレクションにおいては，用語候補が漢字・カタカナ・アルファベットのいずれの字種<sup>10)</sup>であるかによって正解率の傾向が異なるという実験結果が得られた。以下に e-Words での結果を示す(図 12,13,14)。なお，TMREC コーパスでは字種による傾向の違いは見られなかった。

漢文字種の用語では MC-Value が最も精度が良いが，カタカナ字種の用語，アルファベット字種の用語になるにつれ *tf-idf* に逆転されている。コンピュータ用語辞典の見出し語は，用語が漢字かカタカナ・アルファベットかで語としての性質が異なっていると考えられる。

文字種問題は日本語固有であるが，異なる文字種を一括りにして扱うことは，精度を下げる原因となりうる。

## 5 おわりに

本稿では性質の異なるテストコレクションを用いて，異なる側面から用語を捉える様々な用語抽出手法を比較した。

用語抽出の手法を組み合わせることで更なる精度向上の可能性があることが明らかになったが，これらの重要度は性質も値域も異なっており，値の積を取れば全体の精度が向上するといった単純なものではない。

いかなる状況において，どの重要度のどの性質が利用できるか，より詳細に調査し有効な手法を見つけ出したい。

<sup>9)</sup> 後述する漢文字種の用語候補 14051 に対して，漢文字種の正解集合が 281 語しかないことが大きな要因である。漢文字種の正解集合が際だって少ないのは ASCII コーパスでも同様であった。

<sup>10)</sup> 用語の先頭文字が漢字であれば漢文字種，カタカナであればカタカナ字種，アルファベットであればアルファベット字種とした。

## 謝辞

本研究は，科学技術振興財団 (JST) 戦略的基礎研究推進事業 (CREST) の支援のもとに推進されました。

## 参考文献

- [1] *COMPUTERM'98: First Workshop on Computational Terminology*, 1998.
- [2] *COMPUTERM'02: Second Workshop on Computational Terminology*, 2002.
- [3] Sergey Brin and Lawrence Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". *Computer Networks and ISDN Systems*, Vol. 30, pp. 107–117.
- [4] Soumen Chakrabarti. "Mining the Web: Analysis of Hypertext and Semi Structured Data". Morgan Kaufmann, 2002.
- [5] K. T. Franzi and S. Ananiadou. "extracting nested collocations". *COLING*, pp. 41–46, 1996.
- [6] Kyo Kageura. "TMREC Task: Overview and Evaluation". *Proc. of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 411–440, 1999.
- [7] Kyo Kageura and Bin Umino. "methods of automatic term recognition: A review". *Terminology*, Vol. 3, No. 2, pp. 259–289, 1996.
- [8] Jon M. Kleinberg. "Authoritative Sources in a Hyperlinked Environment". *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632, 1999.
- [9] Hiroshi Nakagawa. "Automatic Term Recognition based on Statistics of Compound Nouns". *Terminology*, Vol. 6, No. 2, pp. 195–210, 2000.
- [10] Hiroshi Nakagawa and Tatsunori Mori. "automatic Term Recognition based on Statistics of Compound Nouns and their Components". *Terminology*, Vol. 9, No. 2, pp. 201–219, 2003.
- [11] 中川裕志, 湯本紘彰, 森辰則. "出現頻度と接続頻度に基づく専門用語抽出". *自然言語処理*, 10(1), pp. 27–45, 2003.
- [12] 内山将夫, 井佐原均. "複数尺度の統計的統合とその専門用語抽出への応用". *情報処理学会 研究報告*, pp. 69–76, 2003.