統計的およびグラフ的素性を用いた専門用語抽出

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

1 はじめに

コンピュータ用語や医学用語といった専門用語を，当該分野の文書群から抽出する専門用語抽出は，用語辞書や索引の作成，検索やクラスタリングのキーワードの生成などに広くその成果を応用できる有用な技術である。

従来，用語抽出は人手によって行われることが多かったが，膨大な人的・時間的コストがかかるため，機械による自動用語抽出への要求は強く存在している。

自動的な専門用語抽出の基礎となるのは，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ための数値的な尺度である。このような尺度としては，TFIDF・C-Value [2]・FLR [5] 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，尺度そのものと，尺度の精度を評価する手法には改善の余地がある。

さて，これらの尺度によって算出された重要度がどの程度正確であるか，言い換えればその分野の専門家が専門用語と判断する用語の集合と，自動抽出によって得られた用語集合とがどの程度一致するかを定量的に評価するためには，用語抽出もとの文書群と，その文書群から抽出されるべき用語集合（正解集合）との対が必要である。正解集合の作成は人手で行わなければならないが，コストが膨大なため，正解集合を持つテストコレクションはほとんど整備されていない。

本稿では，Web 上に公開されている Web 用語辞典を，擬似的な正解集合を持つテストコレクションとして用いる手法を提案し，そのテストコレクションを用いて様々な用語抽出手法の精度比較を行う。また，正解集合を持つ唯一のテストコレクションである TMREC テストコレクション [3] での実験結果と比較し，テストコレクションが持つ性質の差異について報告する。

2 Web 辞書テストコレクション

昨今，インターネットの利用者が増加するに伴い，ネットワークから利用できる用語辞典が多数公開されている。本稿ではこれらの用語辞典を用語抽出のテストコレクションとして利用することを提案する。すなわち，

用語の語義文の本文（説明文）⇒ コーパス
辞書の見出し語の集合 ⇒ 正解集合

とすることにより，辞書の見出し語という擬似的な正解集合を持つテストコレクションを作成する。

本稿ではインターネット上の用語辞典のうち，以下のコンピュータ用語辞典を実験に用いた。

- IT 用語辞典 e-Words [*1]
- アスキーデジタル用語辞典 [*2]
- TechWeb [*3]
- FOLDOC [*4]

比較のため TMREC テストコレクションでも実験を行った。TMREC テストコレクションは人工知能分野の学会発表アブストラクトをコーパスとし，人手により正解集合を作成したテストコレクションである。

それぞれのテストコレクションの規模を示す（表 1）。

	言語	文書サイズ	正解語数	抽出語数
e-Words	日	約 2.0MB	3162	30197
アスキー	日	約 1.5MB	3422	28092
TechWeb	英	約 7.4MB	9799	105626
FOLDOC	英	約 5.5MB	6733	98703
TMREC	日	約 1.0MB	7741	17871

表 1 各テストコレクションの規模

抽出語数とは，該当コーパスの文書群から用語候補として抽出された名詞・未知語列の総数を指す。正解語数とは辞書の見出し語（TMREC の場合は手作業によりあらかじめ抽出された正解集合）のうち，語義文中に出現した（コーパスから抽出可能な）語数を示す。

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用いることは，

- 手作業による正解集合作成が不要
- 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可能

といった長所がある。

特に多言語のコーパスを得られる点は，クロスリンガルな用語抽出実験を行える可能性を有している。

[*1] http://e-words.jp/ ©Incept
[*2] http://yougo.ascii24.com/ ©ASCII Corp.
[*3] http://www.techweb.com/ ©CMP Media LLC
[*4] http://foldoc.doc.ic.ac.uk/foldoc/ ©Denis Howe

3　用語抽出の尺度

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。

複合語では，主要部が後ろに来る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。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。

文書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 1）。

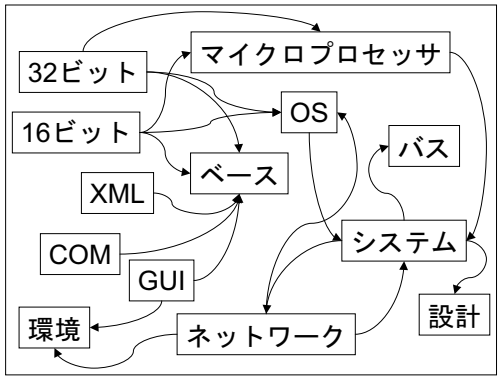

図 1　語基グラフの例

さて，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。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

このとき，Web ページの重要度を算出する手法として，in-link や out-link を用いる手法，HITS [4] や PageRank [1] を用いる手法などがあるが，これらを語基グラフに適用し，用語重要度尺度として LR [5]，HITS，PageRank などを考えることができる。

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。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捉えられる。

LR　語基グラフの個々のノードと，それに繋がるエッジの本数の情報を利用する，最もローカルな着目範囲の手法。

HITS　個々のノードと，2 つ先までのノードとエッジの情報を利用しており，LR よりもグローバルな着目範囲をもつ手法。

PageRank　語基グラフ全体のノードとエッジの情報を利用する，最もグローバルな着目範囲の手法。

4　実験

2 章であげたテストコレクションについて，名詞・未知語の連続文字列を抽出しストップワード除去等の前処理を施す。こうして抽出された用語候補群を，TF・IDF, MC-Value [5], グラフ構造を用いた尺度である LR, HITS，PageRank，LR と頻度情報を融合した FLR [5] を用いてランク付けした。

それぞれの尺度によるランク上位から一定数を取ったときに正解集合に含まれる語数を「精度」として比較したグラフを示す（図 2，3，4，5，6，横軸が Recall，縦軸が Precision）。

辞書を用いたテストコレクションである e-Words，アスキーデジタル用語辞典，TechWeb，FOLDOC の 4 つについては，言語やコーパス規模によらず同様の結果が得られている。Recall が小さい領域（上位の候補）について精度が高いのは FLR，MC-Value，TF・IDF の順であり，中程度の Recall においては TF・IDF，MC-Value，FLR の順に精度が高くなる。また，グラフ構造のみを用いた用語抽出尺度（LR，HITS，PageRank）は低い精度しか得られていない。

これに対し，TMREC においては全く異なった傾向が現れている。FLR，MC-Value 等が高い精度を示し，LR などのグラフ構造のみを用いた用語抽出尺度が続き，TF・IDF は低い精度しか得られていない。

4.1　正解集合の選択基準の差異

Web 辞書と TMREC で尺度ごとの精度が著しく異なる傾向を示したことには，それぞれの正解集合の選び方の違いが大きく影響していると考えられる。

TMREC の正解集合は非常に大きく，本実験で TMREC コーパスから抽出された候補数 17870 語のうち，正解集合に含まれていたのは 7741 語（43%）に及ぶ。

これに対して，Web 辞書の正解集合はあくまで辞書の見出し語であり，数が非常に少ない。例えば e-Words では，文書から抽出された全候補数 30197 語のうち，見出し語は 3162 語（10%）にすぎない。

TMREC テストコレクションの例を表 2 に示す。太字部分が TMREC で正解集合とされている用語である。これらの正解語は専門用語か否かというならば，TMREC コーパスのもととなっている人工知能分野の文書で固有に使われる，または人工知能分野で固有の意味を持つ用語であり，間違いなく専門用語である。

しかし，TMREC の正解集合は「専門用語を選んだ」というよりは「全ての候補語から明らかに一般的な語を除いた」用語集合に近いと考えられる。TMREC テストコレクションの正解集合は辞書の見出し語からは遠いも

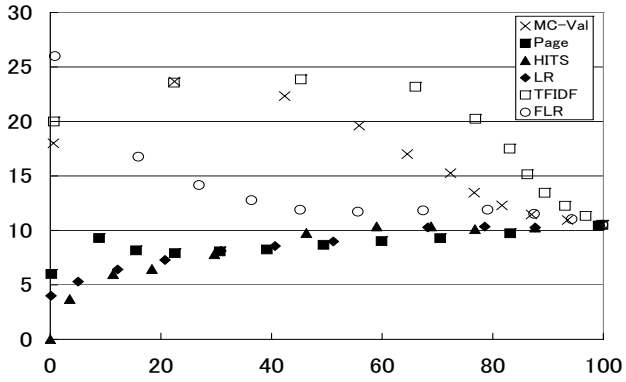

図 2　e-Words での各尺度の精度比較

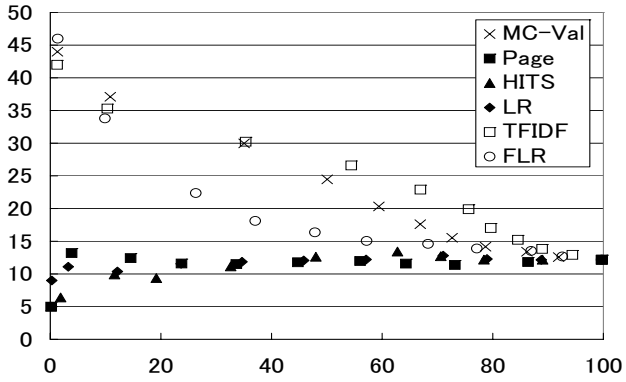

図 3　アスキー用語辞典での各尺度の精度比較

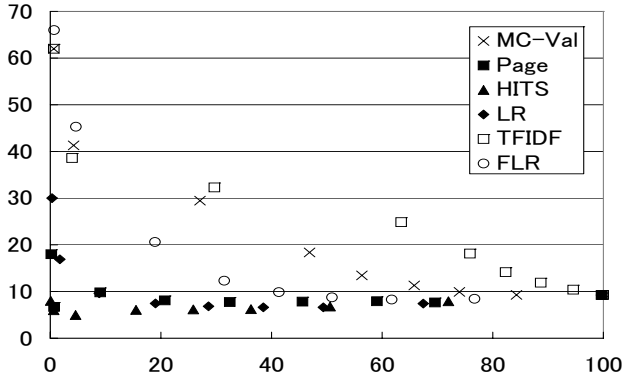

図 4　TechWeb での各尺度の精度比較

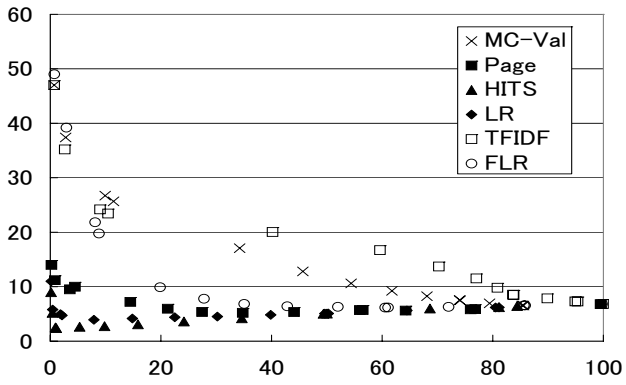

図 5　FOLDOC での各尺度の精度比較

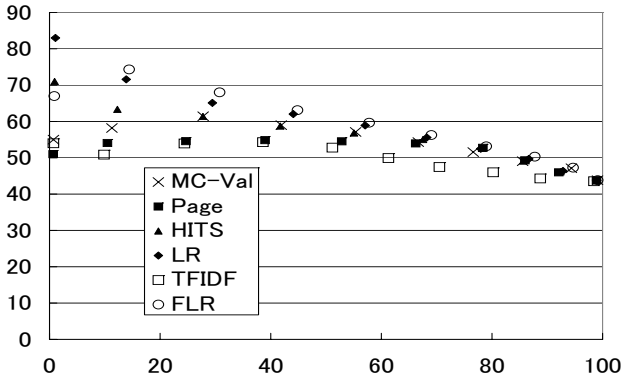

図 6　TMREC での各尺度の精度比較

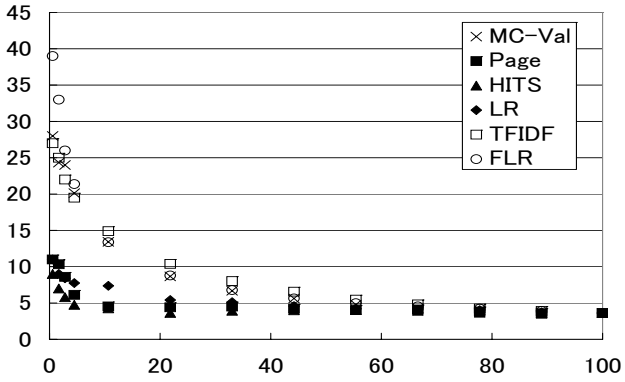

図 7　TMREC 実験結果（辞典掲載語に限定）

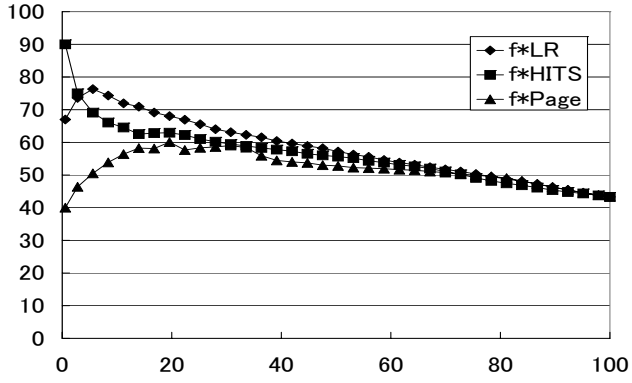

図 8　TMREC 実験結果（FLR，FHITS，FPageRank）

のであり，巨大な索引とでもいうべき用語集合に近い。

TMREC の正解集合には，認知科学辞典 [6] に掲載されている 671 語にマークがつけられている[*5]。

TMREC のコーパスはそのまま用い，正解集合をこの 671 語に限定した場合の実験結果を示す（図 7)。

グラフの形状が少し異なってはいるが，尺度間の優劣関係は辞書テストコレクションと同様の傾向を示してい

[*5] うち本実験でコーパスから候補語として抽出されていたのは 546 語

る。このことから，TMREC テストコレクションと Web 辞書テストコレクションの本質的な差は正解集合の選択手法であると考えられる。

4.2 グラフ着目範囲と精度

LR の尺度値と出現頻度の積を用いる FLR は，TMREC はもちろん Web 辞書においても Recall の小さい領域においては高い Precision を示しているが, LR・HITS・PageRank 単体では，Web 辞書において低い精度しか得られない。

これは, 語基グラフにおける情報を利用した結果, 特定の語基（ノード）が高い重要度を獲得し, 用語の重要度に影響を与え過ぎたためである。LR, HITS, PageRank と，語基グラフの着目範囲が広がるほど顕著にその傾向が現れ，頻度情報との積を取った FLR，FHITS，FPageRank は順に精度が低下する傾向を示した（図 8)。

このことは，Web コミュニティの解析から得られる「グラフのより広範囲な情報を用いた手法が有利である」という直感が必ずしも成立しないことを示唆している。

4.3 考察

本稿では性質の異なるテストコレクションを用いて，様々な用語抽出尺度の精度を比較した。

辞書テストコレクションと認知科学辞典掲載語に限定した TMREC テストコレクションはいずれも同様の傾向を示し，手を加えない状態での TMREC テストコレクションとは全く異なる傾向を示した。この辞書テストコレクションの正解集合は辞書の見出し語であるから，言語や文書の分野にかかわらず，辞書の見出し語は同様の性質を持っていると考えられる。

辞書の見出し語と TMREC の正解集合の間には様々な違いが存在している。語基 1 つのみから成る単語基の用語と，複数の語基から成る用語とで割合が大きく異なっている（表 3)。e-Words と TMREC については用語の字種（漢字・カタカナ・アルファベット）毎の結果も示す。

TMREC では単語基よりも複数語基のほうが候補数に対する正解語数の割合が高いが，辞書コーパスでは単語基のほうが候補数に対する正解語数の割合が高い。また，TMREC では正解語数の 6 割を占める漢字複合語が，e-Words では 1 割程度でしかない。

FLR や MC-Value といった手法は複合語の性質を利用しているため複数語基の用語抽出に強いが，副次的効果として単語基の抽出に弱いと考えられる。このため，複合語の正解語の割合が高いテストコレクション（TMREC）では高い精度を示すが，そうでない場合は単語基用語の抽出に強い TF・IDF が優位となり，テストコレクションによって傾向が異なる要因となっている。

用語抽出においては，目的とする正解集合に応じた抽出尺度を選択しなければ適切な結果を得られない，あるいは非常に悪い結果しか得られないと考えられる。

知的ＣＡＤのためのフィジカル・フィーチャー・データベース**における、**挙動表現**と**推論**の枠組について述べる。**フィジカル・フィーチャー**とは**物理現象**と関連する**属性**との**記述**である。**知的ＣＡＤシステム**は**フィジカル・フィーチャー**の**知識**を利用して**モデル**の**構築支援**、**一貫性管理**、**自動生成**を行なう。現在の**フィジカル・フィーチャー**の表現は**因果関係**が中心であるが、機械の**モデリング**のためには**空間情報**や**場**からの作用の**時間変化**なども¡ **記述**する必要がある。このため、今後**フィジカル・フィーチャー・データベース**を**因果関係**、**空間**、**時間**など複数の**オントロジー**の組合せに拡張する方針について述べる。

表 2　TMREC コーパスと用語候補

	単語基	複数語基
e-Words	1886/13408	1276/16789
e-Words ／漢字	56/6079	225/7971
e-Words ／カタカナ	481/2211	617/4324
e-Words ／ Alphabet	1349/5118	434/4494
アスキー	1888/12660	1534/15432
TMREC ／漢字	979/4592	4567/9046
TMREC ／カタカナ	505/1064	947/1698
TMREC ／ Alphabet	386/832	357/637
TMREC	1870/6488	5871/11383
TechWeb	5113/21099	4686/84527
FOLDOC	3886/27377	2847/71327

表 3　単語基と複数語基についての正解語数／候補数

5 おわりに

専門用語抽出アルゴリズムと正解集合の関係について考察した。その結果，用語抽出アルゴリズムの効果は，正解集合の大きさ，および正解集合における単語基と複合語の比率に依存することがわかった。

参考文献

[1] 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.

[2] K. T. Franzi and S. Ananiadou. “extracting nested collocations”. *COLING*, pp. 41–46, 1996.

[3] 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.

[4] Jon M. Kleinberg. “Authoritative Sources in a Hyperlinked Environment”. *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632, 1999.

[5] 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.

[6] 日本認知科学会. 認知科学辞典. 共立出版, 2002.