出現頻度と連接頻度に基づく専門用語抽出

本論文では，専門用語を専門分野コーパスから自動抽出する方法の提案と実験的評価を報告する。本論文では名詞 (単名詞と複合名詞) を対象として専門用語抽出について検討する。基本的アイデアは、単名詞のバイグラムから得られる単名詞の統計量を利用するという点である。より具体的に言えば、ある単名詞が複合名詞を形成するために連接する名詞の頻度を用いる。この頻度を利用した数種類の複合名詞スコア付け法を提案する。NTCIR1 TMREC テストコレクションによって提案方法を実験的に評価した。この結果、スコアの上位の 1,400 用語候補以内、ならびに、12,000 用語候補以上においては、単名詞バイグラムの統計に基づく提案手法が優れていることがわかった。

キーワード: 用語抽出，専門用語，単名詞，複合名詞

Term Extraction Based on Occurrence and Concatenation Frequency

In this paper, we propose a new idea of automatically recognizing domain specific terms from monolingual corpus. The majority of domain specific terms are compound nouns that we aim at extracting. Our idea is based on single-noun statistics calculated with single-noun bigrams. Namely we focus on how many nouns adjoin the noun in question to form compound nouns. In addition, we combine this measure and frequency of each compound nouns and single-nouns, which we call FLR method. We experimentally evaluate these methods on NTCIR1 TMREC test collection. As the results, when we take into account less than 1,400 or more than 12,000 highest term candidates, FLR method performs best.

KeyWords: *Term recognition, Domain specific terms, Basic Nouns, Compound Nouns*

1 はじめに

自動用語抽出は専門分野のコーパスから専門用語を自動的に抽出する技術として位置付けられる。従来、専門用語の抽出は専門家の人手によらねばならず、大変な人手と時間がかかるため up-to-date な用語辞書が作れないという問題があった。それを自動化することは意義深いことである。専門用語の多くは複合語、とりわけ複合名詞であることが多い。よって、本論文では名詞 (単名詞と複合名詞) を対象として専門用語抽出について検討する。筆者らが専門分野の技術マニュアル文書を解析した経験では多数を占める複合名詞の専門用語は少数の基本的かつこれ以上分割不可能な名詞 (これを以後、単名詞と呼ぶ) を組み合わせて形成されている。この状況では当然、複合名詞とその要素である単名詞の関係に着目することになる。

専門用語のもうひとつの重要な性質として (Kageura and Umino 1996) によれば、ターム性があげられる。ターム性とは、ある言語的単位の持つ分野固有の概念への関連性の強さである。当然、ターム性は専門文書を書いた専門家の概念に直結していると考えられる。したがって、ターム性をできるだけ直接的に反映する用語抽出法が望まれる。

これらの状況を考慮すると、以下のような理由により複合名詞の構造はターム性と深く関係してくることが分かる。第一に、ターム性は通常 tf×idf のような統計量で近似されるが、tf×idf といえども表層表現のコーパスでの現われ方を利用した近似表現に過ぎない。やはり書き手の持っている概念を直接には表していない。第二に、単名詞 N が対象分野の重要な概念を表しているなら、書き手は N を頻繁に単独で使うのみならず、新規な概念を表す表現として N を含む複合名詞を作りだすことも多い。

このような理由により、複合名詞と単名詞の関係を利用する用語抽出法の検討が重要であることが理解できる。この方向での初期の研究に (Enguehard and Pantera 1995) があり、英語、フランス語のコーパスから用語抽出を試みているが、テストコレクションを用いた精密な評価は報告されていない。中川ら (Nakagawa and Mori 1998) は、この関係についてのより形式的な扱いを試みている。そこでは、単名詞の前あるいは後に連接して複合名詞を形成する単名詞の種類数を使った複合名詞の重要度スコア付けを提案していた。この考え方自体は (Fung 1995) が非並行 2 言語コーパスから対訳を抽出するとき用いた context heterogeneity にも共通する。その後、中川らはこのスコア付け方法による用語抽出システムによって NTCIR1 の TMREC(用語抽出) タスクに参加し良好な結果を出している。彼らの方法はある単名詞に連接して複合名詞を構成する単名詞の統計的分布を利用する方法の一実現例である。しかし、彼らの方法では頻度情報を利用していない。上記のように複合名詞とそれを構成する単名詞の関係がターム性を捉えるときに重要な要因であるとしても、(Nakagawa and Mori 1998) が焦点を当てた単名詞に連接する単名詞の種類数だけではなく、彼らが無視したある単名詞に連接する単名詞の頻度の点からも用語抽出の性能を解析してみる必要があると考える。本論文ではこの点を中心に論じ、また複合名詞が独立に、すなわち他の複合名詞の一部としてではない形で、出現する場合の頻度も考慮した場合の用語抽出について論ずる。さらに、有力な用語抽出法である C-value による方法 (Frantzi and Ananiadou 1996) や語頻度 (tf) に基づく方法との比較を通じて、提案する方法により抽出される用語の性質などを調べる。

以下、2 節では用語抽出技術の背景、3 節では単名詞の連接統計情報を一般化した枠組、4 節では NTCIR1 TMREC のテストコレクションを用いての実験と評価について述べる。

2　用語抽出技術の背景

単言語コーパスからの用語抽出には三つのフェーズがある。第一フェーズは、用語の候補の抽出である。第二フェーズは第一フェーズで抽出された候補に対する用語としての適切さを表すスコア付けないし順位付けである。この後に順位付けられた用語候補集合の中から適切な数

の候補を用語として認定するという第三のフェーズがある。しかし、第三フェーズは認定したい用語数の設定など外部的要因に依存するところもあるので、本論文ではその技術的詳細に立ち入らないことにする。

2.1 候補抽出

西欧の言語と異なって空白のような明確な語境界がない日本語や中国語では、情報検索に使う索引語として文字 N-gram も考えられる (Fujii and Croft 1993; Lam,Wong and Wong 1997)。しかし、専門用語という観点に立てばやはり人間に理解できる言語単位でなければならず、結果として単語を候補にせざるをえない。また、NTCIR1 TMREC で使用されたテストコレクションでも単語を対象にしている。さて、単語も詳細に見ると単名詞と複合語に分かれる。関連する過去の研究では単語よりは複雑な構造である連語 (Collocation) や名詞句の抽出を目標にする研究 (Smadja and McKeown 1990; Smadja 1993; Frantzi and Ananiadou 1996; Hisamitsu and Nitta 1996; Shimohata,Sugio and Nagata 1997) が多い。連語や複合語のような言語単位を対象にする場合には、それらはより基本的な構造から構成されることを仮定しなければならない。ここでは、単名詞を最も基本的な要素とする。用語候補が単名詞のどのような文法的構造によって構成されるかという問題も多く研究されてきた (Ananiadou 1994)。どのような構造を抽出するにせよ、まずコーパスの各文から形態素解析によって単語を切り出す必要がある。形態素解析の結果としては各単語に品詞タグが付けられる。よって、複合名詞を抽出するなら、連続する名詞を抽出すればよい。これまでの研究では、名詞句、複合名詞 (Hisamitsu and Nitta 1996; Hisamitsu,Niwa and Tsujii 2000; Nakagawa and Mori 1998)、連語 (Smadja and McKeown 1990; Daille,Gaussier and Lange 1994; Frantzi and Ananiadou 1996; Shimohata et al. 1997) などを抽出することが試みられた。

2.2 スコア付け

前節で述べた用語候補抽出の後、用語候補に用語としての重要度を反映するスコア付けを行う。当然ながら、用語としての重要度はターム性を直接反映すると考えてよく、それゆえにスコアはターム性を反映したものが望ましい。しかし、ターム性というのは前にも述べたように直接計算することが難しい。このため、tf×idf のような用語候補のコーパスでの頻度統計で近似することがひとつの方法である。一方、(Kageura and Umino 1996) は用語の持つべきもうひとつの重要な性質、ユニット性を提案している。ユニット性とは、ある言語単位 (例えば、連語、複合語など) がコーパス中で安定して使用される度合いを表す。これを利用するスコアも用語の重要度を表す有力な方法である。例えば、Ananiadou らが (Frantzi and Ananiadou 1996, 1999) で提案している C-value は入れ子構造を持つコロケーションからユニット性の高い要素に高いスコアを付ける有力な方法である。(Hisamitsu et al. 2000) は、注目する用語と共起する単

語の分布が全単語分布に比べてどのくらい偏っているかをもってターム性を計ろうとしている。(Kageura and Aizawa 2000) は日英 2 言語コーパスを用い、日本語の用語の対訳が英語のコーパスの対応する部分にも共起することがターム性を表わすというアイデアに基づいた用語抽出法を提案している。同様の考えは (Daille et al. 1994) にも見られる。これらの研究は、用語の現れ方や使用統計に基礎をおくものである。一方、(Nakagawa and Mori 1998) は、単名詞と複合語の関係という用語の構造に着目してターム性を表わそうとしている。本論文の次節以降で我々は、ターム性を直接的に捉えようとする (Nakagawa and Mori 1998) に対して、連接する単名詞の種類数だけではなく、頻度も考慮した場合を提案し実験的比較を行った。

3　単名詞の連接統計情報の一般化

3.1　単名詞の連接

2 節の用語抽出技術の背景で述べた多くの研究では実質的に用語の対象にしているのは名詞である。実際、専門用語の辞典に収録されている用語も大多数は名詞である。例えば、(平山, 氏家 1996; 長尾 1990; 青木 1993) などでは収録されているのはほとんどが名詞である。そこで本研究では対象とする用語を単名詞と、その単名詞のみで構成される複合名詞とした。実際、用語の大多数は (平山, 氏家 1996; 長尾 1990; 青木 1993) に見られるように複合名詞である。しかし、これらの複合名詞の要素となる単名詞はあまり多数にのぼるわけではない。この考え方から、単名詞に連接して複合名詞を構成する単名詞の異なり数に着目するというアイデア (Nakagawa and Mori 1998) が生まれる。しかし、連接する単名詞の異なり数だけではなく、頻度など他の要素も考慮することは重要である。連接する単名詞のどのような性質に着目したときに性能の良いスコアになるかを調べるのが本論文の課題のひとつである。

まず、特定のコーパスを想定したとき、単名詞 N が連接する状況すなわち単名詞バイグラムを一般的に図 1 のように表わす。

$$
\begin{array}{ll}
[LN_1 \quad N](\#L_1) & [N \quad RN_1](\#R_1) \\
[LN_2 \quad N](\#L_2) & [N \quad RN_2](\#R_2) \\
: & : \\
[LN_n \quad N](\#L_n) & [N \quad RN_m](\#R_m)
\end{array}
$$

図 1: 単名詞 N を含む単名詞バイグラムと左右連接単名詞の頻度

図 1 において、LN_i $(i = 1, ..., n)$ は、単名詞バイグラム $[LN_i \quad N]$ において N の左方に連接する単名詞 (n 種類) を表わし、単名詞バイグラム $[N \quad RN_i]$ において RN_i $(i = 1, ..., m)$ は N の右方に連接する単名詞 (m 種類) を表わす。また、() 内の $\#L_i$ $(i = 1, ..., n)$ は N の左方に連接する単名詞 LN_i の頻度を表わし、$\#R_i$ $(i = 1, ..., m)$ は N の右方に連接する単名詞 RN_i の頻度を表わす。もちろん、単名詞バイグラム $[LN_i \quad N]$ や $[N \quad RN_j]$ はより長い複合名詞の

一部分であってもよい。以下に “トライグラム” という単名詞を含む単語バイグラムがコーパスから得られた場合、そこから連接頻度を求める簡単な作例を示す。

例 1: 単名詞 “トライグラム” を含む単語バイグラムの抽出例

トライグラム 統計、トライグラム、単語 トライグラム、クラス トライグラム、単語 トライグラム、トライグラム、トライグラム 抽出、単語 トライグラム 統計、トライグラム、文字 トライグラム

この例を図 1 に示す形式で表記すると図 2 のようになる。

[単語 トライグラム](3)　　[トライグラム 統計](2)
[クラス トライグラム](1)　[トライグラム 抽出](1)
[文字 トライグラム](1)

図 2: 単名詞 “トライグラム” を含む単語バイグラムと左右連接単名詞の頻度の例

3.2 単名詞バイグラムを用いた単名詞のスコア付け

3.2.1 連接種類数 $\#LDN(N), \#RDN(N)$

図 1 において単名詞バイグラムで単名詞 N の左方にくる単名詞の種類の異なり数、すなわち n を以後 $\#LDN(N)$ と書く。同様に、単名詞バイグラムで単名詞 N の右方にくる単名詞の種類の異なり数、すなわち m を以後 $\#RDN(N)$ と書く。図 2 の例では、$\#LDN$(トライグラム) $= 3$、$\#RDN$(トライグラム) $= 2$ である。(Nakagawa and Mori 1998) では、この $\#LDN(N), \#RDN(N)$ を単名詞 N のスコアにしている。$\#LDN(N), \#RDN(N)$ は頻度に影響されないので、コーパスが出現する複合名詞の用語をカバーする程度に大きくなれば、もはや一定の値になる。$\#LDN(N), \#RDN(N)$ は N が固有の分野においてどれほどたくさんの概念 (複合名詞で表される) を作るときに使われるかを表す。つまり、分野における基礎概念である度合を表す。よって、N の持つ概念としての重要さを直接表現しているので、ターム性の重要な一面を計っているといえよう。

3.2.2 連接頻度 $\#LN(N), \#RN(N)$

単名詞バイグラムを特徴付ける要因には、連接単名詞の異なり数の他に頻度情報 $\#L_i, \#R_j$ がある。この二つの要因を組み合わせ方としては種々の方法が考えられるが、簡単なのは異なり単名詞毎の頻度の総和をとる方法であり、次式で表わされる。ただし、記法は図 1 の記号を用いる。

$$\#LN(N) = \sum_{i=1}^{n} (\#L_i) \tag{1}$$

$$\#RN(N) = \sum_{i=1}^{m}(\#R_i) \tag{2}$$

$\#LN(N), \#RN(N)$ は、それぞれ N の左方、右方に連接して複合名詞を形成する全単名詞の頻度である。図 2 の例だと、$\#LN(\text{トライグラム}) = 5$、$\#RN(\text{トライグラム}) = 3$ である。

3.3　複合名詞のスコア付け

以上のような方法で単名詞の左右に連接する単語の種類数あるいは頻度を用いたスコアを定義した。これら左右のスコアを組み合わせて単名詞そのもののスコアを定義する必要がある。一方、我々が注目している用語は単名詞だけではなく、複数の単名詞から生成される複合名詞も含まれる。先に述べたように専門用語ではむしろ複合名詞が多いので、複合名詞のスコアを定義することも必要である。複合名詞のスコア付けには、ふたつの考え方がある。第一の考え方は、複合名詞のスコアはその構成単名詞数すなわち長さに依存するというものである。この考え方に従えば、長い複合名詞ほど高いスコアがつくことが自然である。第二の考え方は、スコアは複合名詞の長さに依存しないというものである。この考え方に従えば、長さに対して依存しないような正規化が必要になる。専門用語に複合名詞が多いことは認めるにしても、長い程、あるいは逆に短い程、重要であるという根拠は今のところない。よって、我々は第二の考え方を採る。

まず、前節までで導入した 2 つの単名詞のスコア関数を抽象化し、単名詞 N の左方のスコア関数を $FL(N)$、右方のスコア関数を $FR(N)$ と書くことにする。単名詞 $N_1, N_2, ..., N_L$ がこの順で連接した複合名詞を CN とする。CN のスコアとして前節で定義した各単名詞の左右のスコアの平均をとれば、我々の採った第二の考えに沿った、CN の長さに依存しないスコアを定義できる。ここでは、相乗平均を採用する。ただし、CN の構成要素の単名詞のスコアが一つでも 0 になると CN のスコアが 0 になってしまうので、これを避けるために次式で CN のスコア $LR(CN)$ を定義する。

$$LR(CN) = (\prod_{i=1}^{L}(FL(N_i)+1)(FR(N_i)+1))^{\frac{1}{2L}} \tag{3}$$

例えば、図 2 の場合、連接頻度をスコアとすれば、$LR(\text{トライグラム}) = \sqrt{(5+1)(3+1)} \simeq 4.90$ である。式 (3) によりれば、複合名詞と同時に単名詞のスコア付けもできている。(3) で CN の長さ L の逆数のべき乗となっているので、$LR(CN)$ は CN の長さに依存しない。したがって、単名詞も複合名詞も同じ基準でそのスコアを比較できる。なお、ここで定義した相乗平均の他に相加平均を用いる方法もあるが、以下では予備実験において若干性能の良かった相乗平均のみについて議論する。

3.4 候補語の出現頻度を考慮した重み付け

これまでに述べてきたのは、連接種類数にせよ、連接頻度にせよ、(3) の $LR(CN)$ に関しては、抽出された用語候補集合内での統計的性質についての議論であった。一方で、用語候補が純粋にコーパス中で出現した頻度という別種の情報が存在する。つまり、前者が用語候補集合における構造の情報、後者が、コーパスにおける個別用語候補の統計的性質であり、両者は別種の情報であるといえる。したがって、この両者を組み合わせることによってスコア付け方法の性能改善が期待できる。そこで、用語候補である単名詞あるいは複合名詞が単独で出現した頻度を考慮すべく、(3) を補正して、次のように $FLR(CN)$ を定義する。

$$FLR(CN) = f(CN) \times LR(CN) \tag{4}$$

$f(CN)$ は候補語 CN が単独で出現した頻度である。ここで単独で出現した用語というのは、他の複合名詞に包含されることなく出現した用語のことを指す。例えば、例 1(図 2) の場合、“トライグラム” は単独で 3 回出現しているので、連接頻度をスコアとすれば、FLR(トライグラム) $= 3 \times \sqrt{(5+1)(3+1)} \simeq 14.70$ となる。

3.5 MC-value

比較のために、単名詞バイグラムによらない用語スコア付けとして C-value (Frantzi and Ananiadou 1996) を考える。C-value は次式で定義される。

$$\text{C-value}(CN) = (length(CN) - 1) \times (n(CN) - \frac{t(CN)}{c(CN)}) \tag{5}$$

ここで、

CN:	複合名詞[1]
$length(CN)$:	CN の長さ (構成単名詞数)
$n(CN)$:	コーパスにおける CN の出現回数
$t(CN)$:	CN を含むより長い複合名詞の出現回数
$c(CN)$:	CN を含むより長い複合名詞の異なり数

である。

ところがこの式では、$length(CN) = 1$　すなわち CN が単名詞の場合 C-value が 0 になってしまい、適切なスコアにならない。C-value 以前の類似の方法の (Kita 1994) では、複合語を認識するための計算コストを用語の重要度評価に用いていた。C-value においても、このような背景から、一度複合名詞が切り出された後は、その構成要素の名詞数に比例する認識コストが重要度になる。ただし、複合名詞全体がすでに認識されている場合、名詞を順に認識していけば、最後の名詞を認識する手間は必要なくなる。したがって、(5) では $(length(CN) - 1)$ となる。しかしながら、人間が言葉を認識する上では全ての構成要素の単名詞を認識している

[1] (Frantzi and Ananiadou 1996) では nested collocation と呼ばれる.

と考えられる。そこで、我々は (Frantzi and Ananiadou 1996) の定義を次のように変更した。また、変更した定義を以後、Modified C-value 略して MC-value と呼ぶ。

$$\text{MC-value}(CN) = length(CN) \times (n(CN) - \frac{t(CN)}{c(CN)}) \tag{6}$$

例 1(図 2) の場合、MC-value(トライグラム) $= (7 - 7/5) = 5.6$ である。

4　実験および評価

4.1　実験環境および方法

本節では、まず実験の主な環境となるテストコレクションについて述べる。我々が用いたのは NTCIR-1 の TMREC タスクで利用されたテストコレクションである (Kageura 1999)。1999 年に行われた NTCIR-1 のタスクのひとつであった TMREC では、日本語のコーパスを配布して用語抽出を行う課題が行われた。主催者側が人手で準備した用語に対して参加システムが抽出した用語の一致する度合いを評価した。ただし、これらは何らかの客観的定量的基準に基づいて人手で選択されたものではなく、抽出者の直観によるものである。翻って、ある学問分野における正しい用語とは多くの専門家の時間をかけた合意の産物であり、簡単に定義できない。さりとて、この問題に深入りしても当面大きな成果が得られる保証もないので、上記の評価方法を用いる。なお、以下では NTCIR1 で準備された用語を簡単のため正解用語と呼ぶことにする。さて、日本語コーパスは、NACSIS 学術会議データベースから収集された 1,870 の抄録からなる。対象の分野は、情報処理である。主催者側で準備した正解用語は 8,834 語であり、単名詞と複合名詞が多く含まれる。参加システム側で形態素解析を行うタスクと、主催者側で予め行って形態素解析済みコーパスを配布して利用するタスクがあった。我々は、形態素解析済みで品詞タグ付きのコーパスを利用した。

我々は、この品詞タグ付きのコーパスから用語候補として連続する名詞を抽出した。ただし、“的” と “性” で終了する形容詞は分野固有の複合語の用語に含まれることが多いと考え、例外として単名詞扱いしている。この結果、用語候補数は 16,708 になった。これらを 3 節に述べた諸方法でスコア付けし、スコアの降順に整列した。こうして作られた用語候補を上位から PN 個取り出した場合について、NTCIR-1 TMREC テストコレクションとして供給された正解用語と比較し、抽出正解用語数、適合率、再現率、F-値を計算し評価する。これらは次式で定義される。

$$\text{抽出正解用語数}\,(PN) = \text{上位}\,PN\,\text{候補中の正解用語数} \tag{7}$$

$$\text{適合率}\,(PN) = \frac{\text{抽出正解用語数}\,(PN)}{PN} \tag{8}$$

$$\text{再現率}\,(PN) = \frac{\text{抽出正解用語数}\,(PN)}{\text{NTCIR-1 TMREC テストコレクション中の全正解用語数}} \tag{9}$$

$$\text{F-値}\,(PN) = \frac{2 \times \text{再現率}\,(PN) \times \text{適合率}\,(PN)}{\text{再現率}\,(PN) + \text{適合率}\,(PN)} \tag{10}$$

4.2 各方法の比較実験および考察

以下の各々の手法によりスコア付けをし順位を求めた場合について、PN が 3,000 語までの抽出結果を示し、考察を行なう。

1. 連接種類数 $\#LDN(N), \#RDN(N)$ を用いた LR 法 (以下、「連接種類 LR 法」と呼ぶ。)
2. 連接頻度 $\#LN(N), \#RN(N)$ を用いた LR 法 (以下、「連接頻度 LR 法」と呼ぶ。)
3. LR(連接頻度) 法に候補語の単独出現数を考慮した FLR 法
4. MC-value 法
5. 単名詞、複合名詞の単独での出現頻度をスコアとする語頻度法

NTCIR1 のように専門用語が高い密度で現われるコーパスでは語頻度法が有効に機能すると考えられるので、比較対象の一つに加えた。もしも、語頻度法のような簡単な方法が高い精度を示すなら、ここまで検討してきた 1 から 4 のような複雑な方法は必要がないからである。

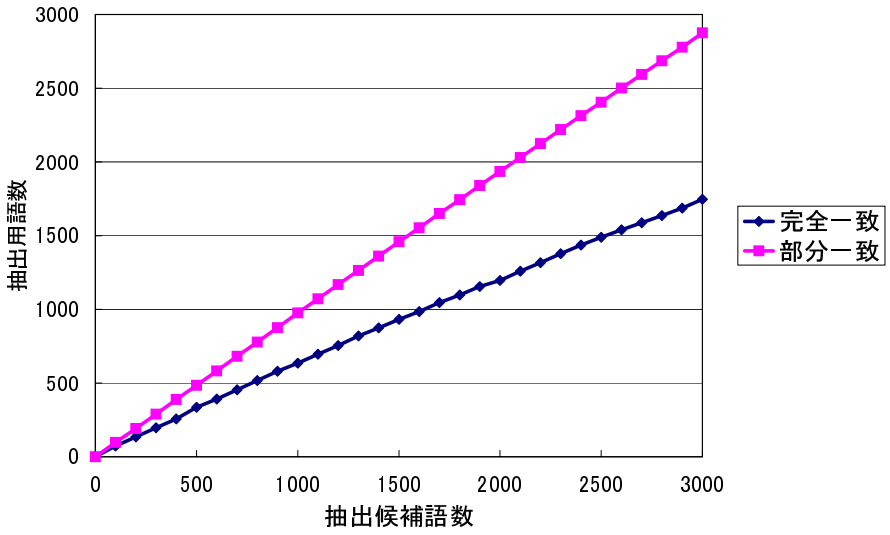

図 3: 連接種類 LR 法で抽出した候補語上位 3,000 語における完全一致数と部分一致数

まず、図 3 に連接種類 LR 法によって抽出された候補語 3,000 語までの場合について、正解用語との完全一致用語数と、正解用語を含んだより長い候補語も数えた、部分一致用語数を示す。例えば、正解用語に「エキスパートシステム」という用語があって、候補語に「エキスパートシステム構築支援」というような用語が抽出された場合、これは部分一致用語数として数えられる。正解用語を含んだより長い候補語も正解とすると 3,000 語まではかなりの部分をカバーしていることがわかる。そこで、この連接種類 LR 法を基準として、語頻度法、連接頻度 LR 法、MC-value 法、および、FLR 法を比較する。図 4 に完全一致用語数の変化を示し、図 5 に部分一致用語数の変化を示す。いずれも、各手法の一致用語数から、基準となる連接種類 LR

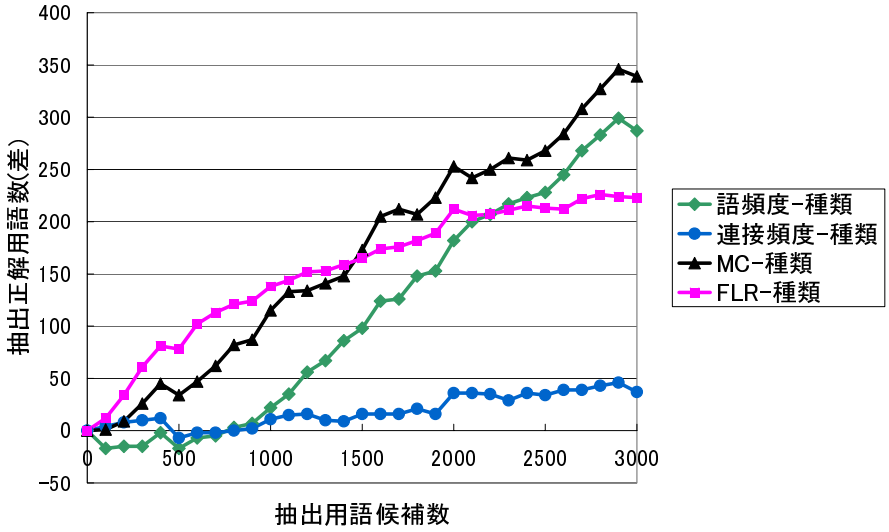

図 4: 語頻度法, 連接頻度 LR 法,MC-value 法,FLR 法における完全一致数の変化 (連接種類 LR 法との差をプロット)

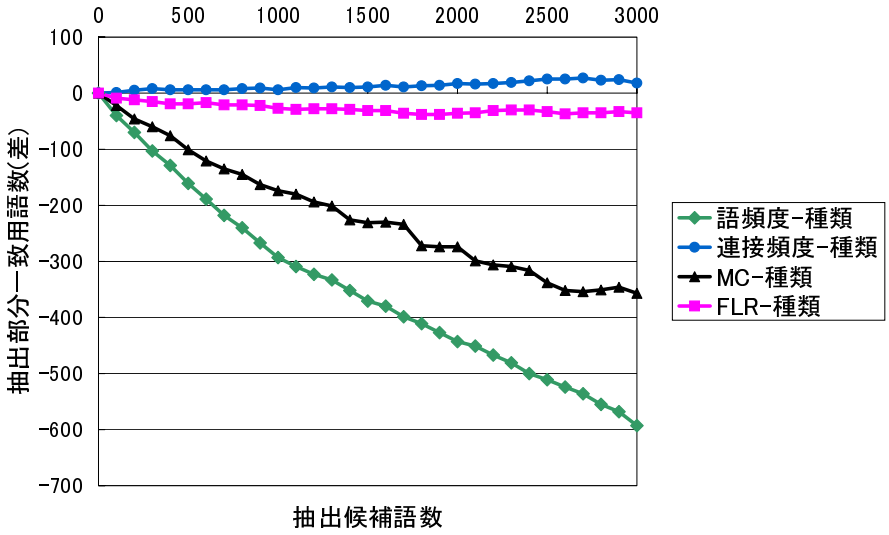

図 5: 語頻度法, 連接頻度 LR 法,MC-value 法,FLR 法における部分一致数の変化 (連接種類 LR 法との差をプロット)

法の一致用語数を減じた数を記している。例えば，図中「$FLR-$ 種類」と示されているプロットは、FLR 法により求めた一致用語数と連接種類 LR 法により求めたものの差の変化を示すものである。

まず、完全一致用語数では「連接頻度 – 種類」のプロットが 0 よりもほぼ上にあることから、連接種類数を手掛かりとするよりも連接頻度を用いる手法のほうが若干優れていることがわかる。一方、FLR 法、MC-value 法、語頻度法は、いずれも、連接頻度 LR 法、連接種類 LR 法を上回る結果となった。さらに、1,400 語までは FLR 法が最も優れた結果を示し、それ以降は MC-value がこれを上回った。また、部分一致用語数では連接頻度 LR 法が最も優れた結果を示した。しかしながら、連接種類 LR 法と FLR 法は、共に大差はないが、語頻度法ならびに MC-value 法はこれらを大きく下回る結果となった。これらを見てもわかるように、我々の提案する手法では完全に間違った候補語は抽出されにくいのに対して、語頻度法や MC-value 法は正解用語とまったく関係のない候補語も抽出される傾向にあるといえる。

さらに、候補語 3,000 語、6,000 語、9,000 語、12,000 語、15,000 語の各々について、抽出正解用語数、再現率、適合率、F-値を求めた。表 1 に抽出正解用語数を、表 2 に再現率、適合率、F-値を示す。

表 1: 各方法により抽出された完全一致用語数

PN	連接種類 LR	連接頻度 LR	FLR	語頻度	MC-value
3,000	1746	1784	1970	2034	2111
6,000	3270	3286	3456	3740	3671
9,000	4713	4744	4866	4834	4930
12,000	5974	6009	6090	5914	6046
15,000	7036	7042	7081	6955	7068

この結果を見ると、まず単名詞バイグラムによる方法の中では、$\#LN(N), \#RN(N)$ に候補語の独立出現数を補正した FLR のスコアが一番性能がよい。語頻度法が 6,000 語の場合には一番性能が良く、MC-value は抽出用語数が 3,000 語、および 9,000 語の場合には全ての方法の中で最も性能がよい。しかし、抽出用語数が増えるにつれて FLR との差は小さくなり、上位 12,000 語および 15,000 語を抽出した場合には FLR が最高の性能を示した。単純な語頻度法は 6,000 語付近で最高の性能を示すが、それ以外では FLR あるいは MC-value に劣ることが実験的に判明した。

さて、このような傾向から見てどの方法が優れているかについて考えてみる。専門用語辞書をみると、(平山, 氏家 1996; 長尾 1990; 青木 1993) では、各々10,000 語から 40,000 語を収録している。よって、15,000 語という多数の抽出で高い性能を示した FLR が有望な方法である。一方、インターネット上の情報通信用語辞典 e-Words(株式会社インセプト 2002) では、2002 年 5 月時点で約 3,200 語を収録している。この領域では MC-value が最も高い性能であった。目的

表 2: 各方法により抽出された完全一致用語における再現率、適合率、F-値

PN	連接種類 LR	連接頻度 LR	FLR	語頻度	MC-value
3,000	.197	.202	.223	.230	.239
	.582	.595	.657	.678	.704
	.295	.301	.333	.343	.356
6,000	.370	.372	.391	.423	.415
	.545	.548	.576	.623	.612
	.441	.443	.466	.504	.496
9,000	.533	.536	.550	.547	.557
	.524	.527	.540	.537	.548
	.529	.532	.545	.542	.553
12,000	.676	.680	.689	.669	.684
	.498	.501	.508	.493	.504
	.573	.577	.584	.567	.580
15,000	.796	.796	.800	.786	.799
	.469	.469	.472	.464	.471
	.590	.591	.594	.583	.593

表の各セルの内容は上段が再現率、中段が適合率、下段が F-値 を表わす。

とする抽出語数が決まれば、採用すべき方法が決まるようにも見えるが、実際は既に述べたように NTCIR1 で主催者が用意した用語の性質にも定量的根拠が薄いので早急な結論は出しにくい。いろいろな分野への適用を通じてどの方法が望ましいかが見えてくると考える。

4.3　抽出用語の性質

さて、これまでは抽出用語の質をそのまま候補語中の正解用語数で議論してきた。しかし、テストコレクションの正解が実用的にどのくらい有効な指標になっているかは議論の余地がある。そこで抽出用語に対する直接的な評価を以下に試みる。まず、用語の長さは抽出用語の品質に密接に関係するのでこれを調べる。語頻度法、連接頻度 LR 法、FLR 法、MC-value 法の 4 つの手法における上位から並べた正解用語の長さを図 6 に示す。ただし、長さは複合名詞を構成する単名詞数で表わした。なお、正解用語の平均語長は 2.56 である。図 6 を見ると、候補語上位 1,400 語付近までで MC-value 法は連接頻度 LR 法や FLR 法に比べて平均語長が短い傾向にある。すなわち MC-value 法では語長の短い語が高いスコアを得る傾向にある。ところが、上位 1,400 語までは FLR が最も多くの正解用語を抽出している。上位 1,400 語以降、MC-value は語長の長い語も抽出するようになるにつれて、より多くの正解用語を抽出するようになった。

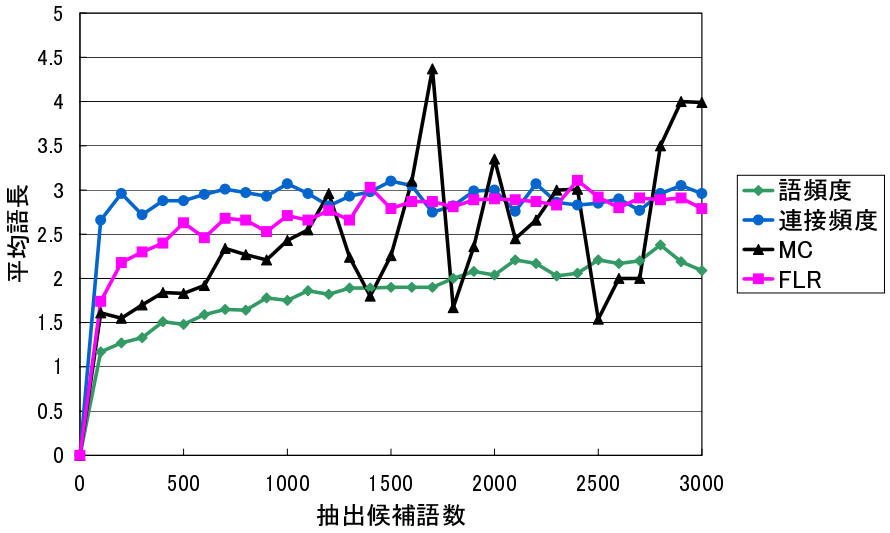

図 6: 各手法における 100 語毎の平均語長

連接頻度 LR、FLR の手法は 1,000 語付近までは FLR のほうが短い語を抽出しているが、それ以降は同程度の長さの語を抽出し、比較的安定している。語頻度法は安定して短い用語を抽出する傾向にある。この理由は後で述べる。

次に具体的な抽出用語例を示そう。全てを示すことは紙面の関係でできないので、最上位の抽出用語を示して各スコア付けの特徴について考えてみる。

表 3 に各手法におけるスコアの最上位 15 候補を示す。この結果を見ると、明らかに連接種類 LR によるスコア付の上位候補は複合名詞が多い。一方、FLR、語頻度、MC-value の各手法によるスコア付けの上位候補には単名詞が多い。FLR では、出現頻度の高い単名詞を優遇する補正をしているし、MC-value でも単名詞の頻度がその単名詞を含む複合名詞の頻度を強く反映した構造になっているから、この結果は偶然ではない。MC-value の場合 “研究、論文、方法、手法” などという分野の用語でない名詞が多く抽出されているが、これも大量かつ多種類の複合名詞に含まれるであろうこと、および MC-value が多数かつ多種類の複合名詞に含まれる単名詞のスコアを高くつけることから得られる帰結である。一方、FLR 法では、連接頻度を用いることにより、これらの単純に頻度が高いだけの名詞をスコアを低くする効果がある点が有利である。さて、図 6 で見たように語頻度法が短い用語を抽出する傾向についてであるが、表 3 を見れば、「我々」「方法」のような一般に使用される単名詞を抽出している。このような一般的な単語は高い頻度で現われるということを示しており、同時に必ずしも専門用語としては重要でないことを考えれば、語頻度法では専門用語を選択的に抽出する能力には限界があると言わざるをえない。

表 3: スコアの最上位の 15 用語候補

連接種類 LR		FLR		語頻度		MC-value	
知識		知識		システム		学習者	
学習知識		システム		知識		問題解決	
学習		問題		研究	×	システム	
言語的知識		学習		本稿	×	知識	
知識システム		モデル		手法	×	研究	×
学習システム		情報		問題		本稿	×
問題知識	×	問題解決		論文	×	手法	×
学習問題		設計		方法	×	問題	
言語的		知識ベース		学習者		知識ベース	
システム		推論		情報		論文	×
問題		支援	×	モデル		方法	×
論理的知識		知識表現		我々	×	支援システム	
学習支援システム		エージェント		ユーザ		計算機	
設計知識		学習者モデル		機能	×	情報	
学習問題解決システム		構造	×	対象		モデル	

無印:正解用語,　×:不正解用語

4.4　NTCIR-1TMREC の結果との比較

ここまで述べてきたスコア付け方法の客観的評価を行うために、NTCIR-1 TMREC タスクで上位の成績を残したチームとの比較を行う。なお、NTCIR-1 には C-value によるスコア付けをするチームも参加しているが、NTCIR-1 の参加規定によりどのチームかは不明である。しかし、後で述べるように本論文で提案した C-value を修正した MC-value が良好な結果を示していることから、我々の C-value の修正法には若干の独自性が認められると考えられる。NTCIR-1 TMREC の上位 2 チームの手法を以後 T1,T2 と呼ぶ。

T1、T2、ならびに、本論文で性能の良かった FLR および MC-value の各スコア付け方法において、上位から 3,000 語までの範囲で 1,000 語毎に求めた適合率を表 4 ならびに図 7 に示す。また、同様に上位から 15,000 語までの範囲で 3,000 語毎に求めた適合率を図 8 に示す。

表 4 ならびに図 7 によれば、スコア付け 1001〜2000,2001〜3000 語の部分では MC-value が他を上回ったが、1〜1000 語部分での抽出精度は我々の提案した FLR によるスコア付けが、最も優れた結果を示した。また、図 8 に示すとおり、3,000 語以降については、候補語数が多くなるにつれて、手法 T1、T2 は適合率を落とすが、FLR 法と MC-value 法の抽出精度の下がり方はなだらかであった。このことは FLR 法や MC-value 法が安定して正解用語を抽出していることを示している。最終的に FLR は候補語上位 16000 語のうち、7412 語が正解用語であった。

表 4: NTCIR-1 TMREC 参加上位 2 チームと FLR、MC-value の比較 (1000 語毎の適合率)

PN	FLR	MC-value	T1	T2
1 から 1,000	.773	.754	.705	.744
1,001 から 2,000	.635	.707	.607	.584
2,001 から 3,000	.562	.640	.618	.518

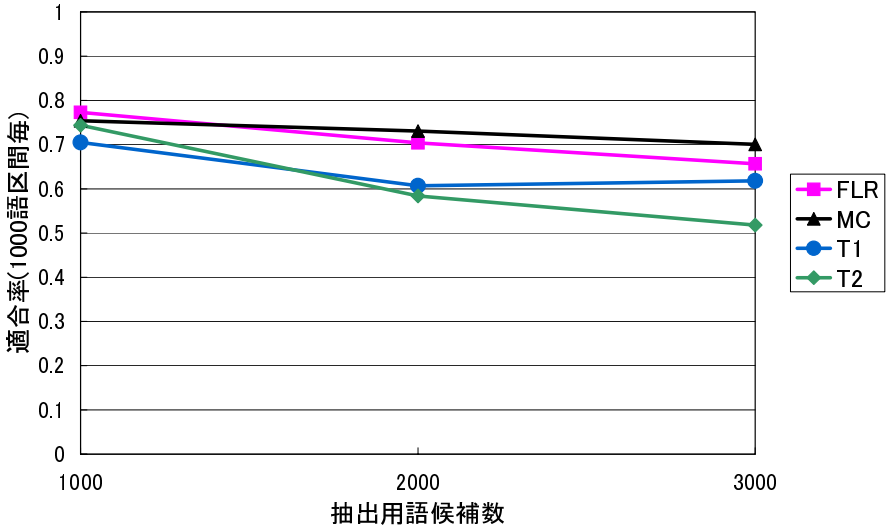

図 7: NTCIR1 TMREC タスク参加上位 2 チームと FLR、MC-value 方式の比較 (1000 語毎の適合率)

この結果は他の研究と比較しても高い結果といえるだろう。NTCIR の中のチームの候補語上位 16,000 語での抽出結果では、$T1$ の正解用語数 6536 語が最高である。また、最も多く正解用語を抽出したチームは $T2$ で、正解用語数 7944 語であるがこれは候補語上位 23270 語からマッチしたものであり、かなり低い適合率である。我々は、名詞の連続だけを取り出したが、正解用語の中には形容詞と名詞の連接や、助詞 “の” によってつながった用語もある。これらを広く抽出すれば再現率は高まるが、上位のスコアの抽出後においてすら非正解用語を多数抽出してしまい、あまり好ましくない。

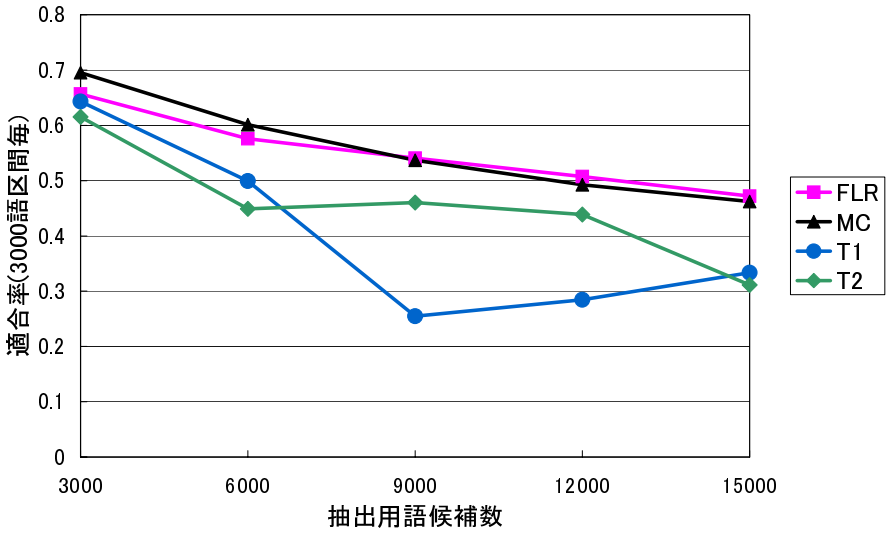

図 8: NTCIR1 TMREC タスク参加上位 2 チームと FLR、MC-value 方式の比較 (3000 語毎の適合率)

5　おわりに

本論文では、専門分野コーパスからの専門用語の抽出法について検討した。まず、用語抽出技術の背景を述べ、次に本論文の核心である単名詞 N に連接する単名詞の頻度の統計量を利用する N のスコア付け方法を提案した。これらスコア付け方法を複合名詞のスコア付けに拡張した。比較対象としては、既存の C-value を修正した MC-value 法ならびに語頻度法を検討した。これらのスコア付け法を NTCIR-1 TMREC タスクのテストコレクションに適用して結果を評価した。その結果、スコア上位の候補、および 12,000 語以上を抽出する場合においては我々の提案する FLR 法の性能が優れていることがわかった。一方、1,500〜10,000 語程度の専門語を抽出したいのであるなら、MC-value 法のほうが優れた結果を示すが、正解用語を含む長めの語でよいのであれば、FLR 法の出力は正解用語の大部分をカバーすることができることもわかった。

今後の課題としては、より多様な情報、例えば文脈情報を利用して用語抽出の性能の向上を計ることが重要である。しかし、一方で、専門分野の用語として真に欲しいのはどのような性質を持つ用語なのかを定式化するという根本的問題も考察していく必要があろう。このような考察は哲学的なものというよりは、実際のコーパスの統計処理を用いた実験的なものでなければ実用性に乏しい。その意味で、このような観点から設計した用語抽出タスクを企画することも望まれる時期にきているのではないだろうか。

参考文献

Ananiadou, S. (1994). "A Methodology for Automatic Term Recognition." In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, pp. 1034–1038.

Daille, B.,Gaussier, E.,and Lange, J. M. (1994). "Towards Automatic Extraction of Monolingual and Bilingual Terminology." In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, pp. 515–521.

Enguehard, C. and Pantera, L. (1995). "Automatic Natural Acquisition of a Terminology." *Journal of Quantative Linguistics*, **2** (1), 27–32.

Frantzi, K. and Ananiadou, S. (1996). "Extracting Nested Collocations." In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96)*, pp. 41–46.

Frantzi, K. and Ananiadou, S. (1999). "The C-value/NC-value method for ATR." *Journal of NLP*, **6** (3), 145–179.

Fujii, H. and Croft, W. B. (1993). "A Comparison of Indexing Techniques for Japanese Text Retrieval." In *Proceedings of SIGIR '93: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 237–246.

Fung, P. (1995). "Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus." In *Proceedings of the Third Workshop on Very Large Corpora*, pp. 173–183.

Hisamitsu, T. and Nitta, Y. (1996). "Analysis of Japanese compound nouns by direct text scanning." In *Proceedings of the 16th International Conference on Computationa Linguistics (COLING 96)*, pp. 550–555.

Hisamitsu, T.,Niwa, Y.,and Tsujii, J. (2000). "A Method of Measuring Term Representativeness." In *Proceedings of 18th International Conference on Computational Linguistics (COLING 2000)*, pp. 320–326.

Kageura, K. (1999). "TMREC Task: Overview and Evaluation." In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 411–440.

Kageura, K. and Aizawa, A. (2000). "Automatic Thesaurus Generation through Multiple Filtering." In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pp. 397–403.

Kageura, K. and Umino, B. (1996). "Methods of automatic term recognition:A review." *Terminology*, **3** (2), 259–289.

Kita, K. (1994). "A Comparative Study of Automatic Extraction of Collocations from Copora: Mutual Infomation v.s. Cost Criteria." *Journal of NLP*, **1** (1), 21–29.

Lam, W.,Wong, C.-Y.,and Wong, K.-F. (1997). "Performance Evaluation of Character-, Word- and N-gram-Based Indexing for Chinese Text Retrieval." In *Proceedings of the Second*

International Workshop on Information Retrieval With Asian Languages (IRAL 97), pp. 68–80.

Nakagawa, H. and Mori, T. (1998). "Nested Collocation and Compound Noun for Term Recognition." In *Proceedings of the First Workshop on Computational Terminology (COMPTERM 98)*, pp. 64–70.

Shimohata, S.,Sugio, T.,and Nagata, J. (1997). "Retrieving Collocations by Co-occurences and Word Order Constraints." In *Proceedings of the 35th ACL and 8th EACL*, pp. 476–481.

Smadja, F. (1993). "Retrieving collocations from text:Xtract." *Computational Linguistics*, **19** (1), 143–177.

Smadja, F. and McKeown, K. (1990). "Automatically Extracting and Representing Collocations for Language Generation." In *Proceedings of the 28th Annual Meeting of the Association of Computational Linguistics (ACL 90)*, pp. 252–259.

株式会社インセプト (2002). "情報・通信辞典 e-Words." http://www.e-words.ne.jp/.

長尾真（編） (1990). 岩波情報科学辞典. 岩波書店.

平山博, 氏家理央（編） (1996). 電子情報通信英和・和英辞典. 共立出版.

青木繁（編） (1993). 建築大辞典. 彰国社.