

日本語マニュアル文における名詞間の接続情報を用いた ハイパーテキスト化のための索引語の抽出

○中川 裕志, 森 辰則, 松崎 知美

横浜国立大学工学部

情報化社会においてますます重要度が高くなっているマニュアルは決して読みやすいものではない。この状態の有力な解決策のひとつがハイパーテキスト化である。ハイパーテキストにおいてリンクを張る語はいわゆる索引語であり、情報検索におけるキーワードとは異なる。本論文では、基本単語が複合語を構成する割合を利用した新しい索引候補語抽出方法を提案する。これは、マニュアルにおける索引語の多くが複合語であるという事実によるものである。提案する方法では、さらに候補の中から索引語となる複合語を選択する方法も提案している。これらの方法は適合率と再現率から求めた Rijsbergen の E によって評価し、準最適化したものである。実験では5種のマニュアルで再現率 0.712、適合率 0.650 を得ることができた。

Index Words Extraction for Hypertextization by Noun-to-noun Connections of Japanese Manual Sentences

○Hiroshi Nakagawa, Tatsunori Mori, Tomomi Matsuzaki

Engineering, Yokohama National University

In this paper, first we propose a new method to extract the candidates of index words which is based on word's power to generate compound nouns, because majority of index words in manuals are compound nouns which include special nouns that express the main and/or key concept of the apparatus described in the manual. Second we propose the method to select index words from among the candidate which optimize Rijsbergen's E calculated from the recall and precision. Our result shows 0.712 of recall and 0.650 of precision on 5 Japanese manuals.

1 はじめに

マニュアルを読みやすいものにする、あるいはハイパーテキスト化などにおいて最大の問題はリンクを張る用語の抽出である。リンクを張るべき語は書物の巻末にある索引語にほぼ相当する。従来は索引語の抽出は人手で行なっていた。また、索引のないマニュアルも数多くみかける。さらに問題なのは索引語の選択基準が明確でないことである。このような状況から索引語を抽出する客観的基準を明らかにし、その基準に基づく自動抽出システムを開発することが、上記のハイパーテキスト化においては必須であるといえる。

ここで考えている索引語の性質は本やマニュアルの巻末にある索引語に近い。ただし、ハイパーテキスト化を目的とする場合には対象とする用語は次の性質を持つ。すなわち、リンクを張るためには同じ用語が定義される部分と、参照される部分に分かれて複数現れなければならない。この点が索引語と若干異なる。情報検索におけるキーワードと索引語との差は表 1 のようにまとめられる。

2 索引語候補となる複合名詞の重要度

マニュアルでは、索引語はしばしば、マニュアルの主要な概念を示す名詞を含む複合名詞である。多くの研究がここ十年間を通して、キーワードもしくは索引語となるような複合名詞の抽出の方法を提案してきた。日本語マニュアル文においては非常に多くの複合名詞が用いられるが、それら全てが必ずしも重要な索引語ではない。また逆に単名詞(「システム」や「実現」のようにそれ以上分解できない名詞)でも有力な索引語となることもある。

表 1: 索引語とキーワードの差異

	索引語	キーワード
検索対象の文書ベース	単一文書	膨大な文書群
文書あたりの数	数 10 から 数 100	数個から 数 10
特徴付けるもの	文書の一部分	単一の文書

そこで問題は与えられた複合名詞あるいは単名詞がどれだけ重要かを表す尺度を見つけることとなる。この目的のために、まず単名詞に対して前方、後方接続数を提案する。

定義 1 与えられたマニュアルにおいて、単名詞 N の前方接続数は、 $Pre(N)$ と表し、単名詞 N の直前について、 N と複合名詞を作る名詞の数とする。単名詞 N の後方接続数は、 $Post(N)$ と表し、単名詞 N の直後について、 N と複合名詞を作る名詞の数¹とする。

この定義に関する最も重要な点は、高い $Pre(N)$ 、 $Post(N)$ の値を持つ語と与えられたマニュアル中の主要な概念の間の関係である。このために、本研究では以下の基本的な仮定を提案する。

仮定 1 高い $Pre(N)$ 、 $Post(N)$ の値を持つ単名詞は与えられたマニュアル中で用いられた主要な概念を表す。

これは今のところ単に仮定であるが、本論文の残り部分で、この仮定に基づいて抽出した索引語が適合率、再現率において良い値を持つことを示すことによって、この仮定を実証してゆく。またこの仮定によれば、単名詞でも主要な概念を表すものは高い $Pre(N)$ 、 $Post(N)$ を持つことになる。

ここまでで成分となる単名詞の重要度を $Pre(N)$ 、 $Post(N)$ として定義した。次になすべきことは、処理対象となるマニュアルに現れる複合名詞 N の重要性の尺度を定義することである。この尺度を $Imp(N)$ と呼ぶことにする。複合名詞 N が単名詞 N_1, N_2, \dots, N_k が継ったものだとする。すると N の使用された文脈を無視すれば $Imp(N)$ は、 $Pre(N_1), Post(N_1), \dots, Pre(N_k), Post(N_k)$ の関数として定義される。この関数には無限に多くの可能性があるため、全てを比較検討することはできない。そこでここでは、代表的な 4 種の関数として積、相乗平均、和、相加平均について比較検討することにした。これらは次式で表される。

定義 2 複合名詞 $N_1 N_2 N_3 \dots N_k$ (ここで N_i は単名詞もしくは「単名詞+」とする) の重要度、 $Imp_m(N_1 N_2 \dots N_k)$ ただし $m = 1, \dots, 4$ の各々を以下の式で定義する。

¹ここでいう「名詞の数」は頻度ではなく、同じ名詞は 1 回しかカウントしない。

$$\begin{aligned} & Imp_1(N_1 N_2 \cdots N_k) \\ &= \prod_{i=1}^k ((Pre(N_i) + 1) \cdot (Post(N_i) + 1)) \quad (1) \end{aligned}$$

$$\begin{aligned} & Imp_2(N_1 N_2 \cdots N_k) \\ &= \left(\prod_{i=1}^k ((Pre(N_i) + 1) \cdot (Post(N_i) + 1)) \right)^{\frac{1}{2k}} \quad (2) \end{aligned}$$

$$\begin{aligned} & Imp_3(N_1 N_2 \cdots N_k) \\ &= \sum_{i=1}^k (Pre(N_i) + Post(N_i)) \quad (3) \end{aligned}$$

$$\begin{aligned} & Imp_4(N_1 N_2 \cdots N_k) \\ &= \frac{1}{2k} \cdot \sum_{i=1}^k (Pre(N_i) + Post(N_i)) \quad (4) \end{aligned}$$

Imp_1, Imp_2 で $Pre(N), Post(N)$ に +1 しているのは、 $Pre(N), Post(N)$ がひとつでも 0 になると Imp が 0 になってしまうのを避けるためである。 Imp_1 と Imp_3 は各 $Pre(N), Post(N)$ の積あるいは和である。よって複合名詞を構成する単名詞の数が増えるとそれに相乗的あるいは相加的に大きな値を持つようになる。つまり、長い複合名詞ほど高い値を持つ。 Imp_1 は積なので Imp_3 より長い複合名詞が重要視されることになる。一方、 Imp_2 は Imp_1 の相乗平均、 Imp_4 は Imp_3 の相加平均をとったものである。よってこれらは複合名詞の長さには依存しない値を持つ。すなわち、長い複合名詞でもそれを単名詞換算した値になる。ただし、相乗平均は相加平均より小さな値になる。これをまとめると、 Imp の値が複合名詞を構成する単名詞の個数に対する依存度は、 $Imp_1 > Imp_3 > Imp_4 > Imp_2$ になる。当然のことながら Imp の定義の仕方にはこの定義の他にもたくさんのやり方がある。よりよい Imp はマニュアルの種類や内容に依存しているかもしれない。この他に複合語を構成する各単名詞 N の $Pre(N), Post(N)$ に重みを付ける方法についても検討した。これは、 Imp_1, Imp_2 の変形として、複合語を構成する最初あるいは最後の単名詞の $Pre(N), Post(N)$ を 1.5 乗する方法と、 Imp_3, Imp_4 の場合は 1.5 倍するを試した。しか

し、より詳細な検討は今後と課題とし、本論文では上記 4 種の Imp を比較する。

以下に Imp_1, \dots, Imp_4 とその変形にもとづいたマニュアルからの索引語候補の抽出方法についてを述べる。

2.1 形態素解析

本研究の目標は日本語マニュアル文から索引語を抽出することである。ところが日本語は英語のスペースのような単語の区切りを持たないため、まず形態素解析して単語に分割することが必要となる。このために、本研究では形態素解析システム JUMAN [1] を用いた。索引語抽出の第一ステップとして、JUMAN を用いて与えられたマニュアルの全ての文を単語に分割した。JUMAN は入力 of 日本語を形態素解析するため約 10 万語の一般的な単語を用いており、その出力は分離された単語とその品詞情報の対の列である。

2.2 複合名詞の抽出

次に索引語の候補として、JUMAN の出力から単名詞もしくは「単名詞+の」の連続を全て拾い出し、それぞれについて $Imp_m, (m = 1, \dots, 4)$ を計算した。最後に各 Imp に対して、上記の索引語候補を Imp 値の大きい順にソートした。仮定 1 によって、 Imp の値を利用して重要な索引語を選ぶことが出来る。

なお、本研究では以下の 5 種のマニュアルを利用して実験を行なった。

3 索引語選択

3.1 正解索引語の選定法

前の章で述べたように、一つのマニュアル中に現れた全ての複合名詞句および単独名詞は Imp によって順序付けられる。しかし、これら全てが必ずしも重要な索引語ではないから、 Imp の値でランクづけされたこれらの複合名詞句から重要な索引語を選び出さなければならない。この目的のために、 Imp の値とマニュアル読者が選んだ正解索引語の間の関係を知ることが必要である。それぞれのマニュアルについて、これを読んだ 3 から 4 人の被験者によって、次の方法で正解索引語を選択させた。

表 2: 本研究で用いたマニュアル

マニュアル	総文数	大きさ (kB)
コンピューターソフトウェア (JUMAN)	436	31
〃 (SAX)	433	28
〃 (「たまご」)	628	30
家庭用ビデオデッキ (三菱電機 HV-F93)	1461	69
家庭用ゲーム機 (SONY PlayStation)	131	7

- 各マニュアルに現れる全ての複合名詞句と独立に現れた単名詞を被験者に提示する。ただし、提示の順序は出現頻度の大きい順とした。
- その語の定義がマニュアルを読む際にハイパーリンクされる必要があると思われるものを選択させる。
- 全員が一致して選択したものは正解索引語とする。
- 判断が分かれたものは、それが使われた文を読み被験者全員で索引語とするかどうかを最終判断をする。

このような方法で選ばれた正解索引語のマニュアル毎の数は次の通りである。また、2.2で述べた方法で求めた索引語の候補となる名詞の連続の個数も併せて示す。

4 窓方式

次に *Imp* と正解索引語の関係性を分析するが、処理の容易さのために *Imp* の大きい順に並べた複合名詞句上を一定幅の窓を移動させる方法を用いる。例えば、JUMAN のマニュアルで窓の幅

表 3: 正解索引語数と索引語候補数

マニュアル	正解索引語数	索引語候補数
JUMAN	104	882
SAX	216	797
たまご	102	786
ビデオ	264	1216
ゲーム機	41	150

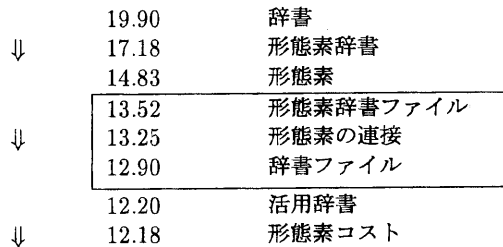


図 1: 窓方式：窓の幅 3

3で上から3語移動したとき窓を図1に示す。ここでは *Imp*₂ を用いている。

窓の位置に対応する *Imp* の値は、窓の範囲内の複合名詞句の中で最も高い *Imp* の値とした。

さてこの窓を動かして以下の性質を調べる。すなわち *Imp* と正しい索引語の関係性を示すために、x軸に *Imp* の値、y軸に(窓の中の正解索引語の数) / (窓の幅) をとったグラフを用いる。図2はJUMAN のマニュアルで *Imp*₁ を用いた場合の窓の幅 30 の例である。

このグラフを見ると、*Imp* の値が小さくなるにつれて、窓内の正解索引語の数も小さくなるのが分かる。この傾向は他の *Imp* でも同様である。このことから、適当な *Imp* の値をしきい値として用いることが出来るのが分かる。つまり、しきい値より高い *Imp* の値に対応する窓の中の候補のうち *Imp* の値が最大の、単名詞あるいは複合名詞句のみを索引語として選ぶのである。しかし、*Imp* の値の分布は個々のマニュアルに依存する。例えば我々が分析した5つのマニュアルにおいて *Imp*₂ の最大値は、高いもので21.45、低いもので4.47であり、その平均値は高いもので3.14、低いもので1.63である。他の *Imp* では、複合名詞の長さ依存性がより大きいので、この分布はもっと

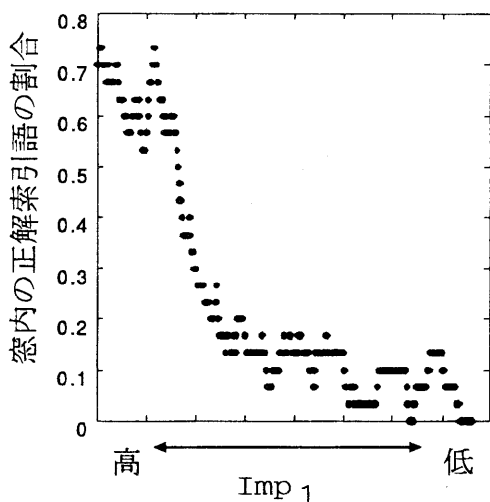


図 2: JUMAN のマニュアルにおける正解重要語の割合 (窓の幅: 30)

広がっている。それゆえいろいろなマニュアルに一つの固定したしきい値を適用することはできない。Imp のしきい値を決定するためにマニュアルに依存しない方法が必要である。この目的で、(窓の中の複合名詞句の数) / (窓の幅) の割合 (これを複合語率と呼ぶ。)に着目した。当然のことながら、(窓の幅) = (窓の中の複合名詞句の数) + (窓の中の単名詞の数) である。複合語率に着目した理由は重要な索引語の多くが複合名詞であり、単名詞の多くが重要な索引語でない傾向があるからである。いくつか重要な索引語となる単名詞もあるが、Imp の値が高い領域では、ある幅の窓の中でそれらの多くが複合名詞であることが期待される。ところで、個別のマニュアルにおいて単名詞の数と複合名詞句の数とともにマニュアルの長さとともに増えていくという傾向が予測されるから、複合語率はマニュアルにあまり依存しないと予想される。当然のことながらこれらはマニュアルの長さにも必ずしも比例するものではない。しかし、少なくとも複合語率はマニュアルの長さからの影響が小さいから、個々のマニュアルへの依存度が低い。複合語率を用いるのが正しいことを実証するために、(窓の中の正解索引語の数) と (複合語率) の相関係数を計算してみると各マニュアルで全ての Imp で

0.6 から 0.9 程度となり、充分高い相関があることがわかった。言い替えば、Imp が高い名詞句の多くが複合名詞句であるという予想が経験的に証明されたことになる。よって、まず窓内の複合語率にあるしきい値を決めておく。そして、窓を動かして、そのしきい値より大きい複合語率を持つ窓の中の候補となる単名詞あるいは複合名詞句 (つまり窓内の最も Imp の高い候補) を索引語として選択すればよい。

4.1 E によるしきい値の決定

窓方式において複合語率によって索引語選択のしきい値を決めるためには窓の幅も決めなければならない。以下でこの窓方式の 2 つのパラメータを決定する方法を提案する。普通、キーワード抽出法は次式に示す再現率と適合率によって評価される。

定義 3 再現率・適合率の定義

$$\text{再現率} = \frac{\#((\text{抽出索引語}) \cap (\text{正解索引語}))}{\#(\text{正解索引語})} \quad (5)$$

$$\text{適合率} = \frac{\#((\text{抽出索引語}) \cap (\text{正解索引語}))}{\#(\text{抽出索引語})} \quad (6)$$

ただし #X は 集合 X に含まれる要素数。

そこで、ここでの索引語抽出もこれらによって評価する。そして、再現率と適合率を最適化することによって窓方式のパラメータを決定する。再現率と適合率を同時に最適化するにはいろいろな方法がある。ここでは、Rijsbergen によって定義された E の値を用いる。Rijsbergen の E は次式で表される。

$$E = 1 - \frac{(1 + b^2)PR}{b^2P + R} \quad (7)$$

ここで P は適合率を、R は再現率を、b は再現率にどれだけ重みを置くかを表す。ハイパーテキスト化という観点からどのような重みにすべきかを考察する必要はある。例えば、初心者のためにできるだけ多くの用語に索引をつけたいなら再現率を重視すべきであろう。ただし、どのような重みがよいハイパーテキスト化に継るかは別の課題で

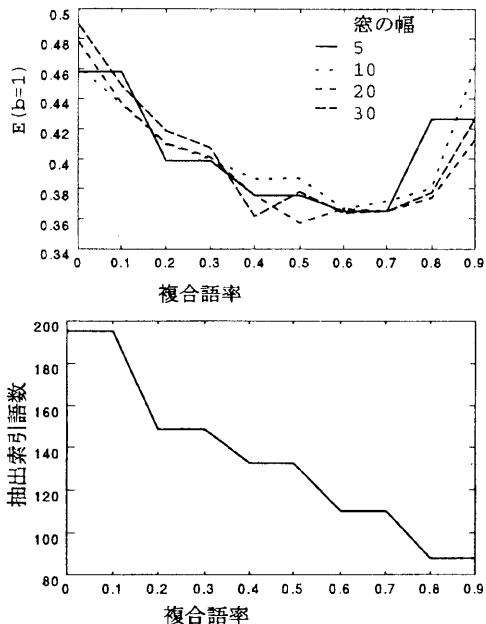


図 3: JUMAN のマニュアルの E と抽出索引語数 (窓幅 = 5)

ある。したがって、ここでは両者を同じ重みに考え、 $b=1$ とした。もっとも、以下に述べる方法論は b の値を変えて計算しても使える。

まず、これまでに述べた窓方式における 2 つのパラメータである窓幅と複合語率のしきい値の種々の値の組み合わせに対して選ばれた複合名詞句について、適合率と再現率を評価した。適合率と再現率を計算するために、3.1 で述べた正解索引語を用いた。与えられた 2 つの窓方式のパラメータについて Imp の値によってランクづけされた複合名詞句から索引語を選んだ。

対象としている 5 種のマニュアルについて、 Imp_1, \dots, Imp_4 およびその変形の各々を用いて、 E を計算してみた。索引語選択のパラメータとして、窓の幅は 5, 10, 20, 30 の場合について複合語率のしきい値を 0 から 1 まで 0.1 きざみで動かして計算した。一例として JUMAN において Imp_1 を用いた場合を図 3 に示す。複合語率のしきい値に対応して選択された索引語の数も下の図に平行して示す。

上で述べた組合せ全ての図を示すと量が多過ぎるので、 Imp_1, \dots, Imp_4 の各々で、5, 10, 20, 30 各々

の窓幅に関して、1) 0 と 1 の間の全複合語率に対する E の平均値 avE , 2) 複合語率を 0.1 から 1 まで 0.1 刻みで変化させたときの E の最適値 (最小値) $minE$, E が最適値の場合の再現率、適合率を 4 つのマニュアルで平均したもの (表 4 中 $av\bar{E}$ などと表す。) を比較した。この結果から 5 種のマニュアルに共通に適用できる準最適な窓幅と複合語率のしきい値を選ぶとき次の点に考慮する必要がある。つまり、未知のマニュアルにこの方法を適用する場合には正解索引語は予めわかっていないし、最適な E も分からない。そこで、複合語率しきい値が最適な E を求めるものから少々ずれていても比較的良好な E が確保できる方法が望ましい。つまり、選択方針は、ある複合語率しきい値で最小の E を出すよりも、複合語率しきい値平均 avE がよい窓幅を選ぶほうがよい。そこで我々は avE に着目し、これが最小になる窓幅を選んだ。

この結果、は Imp_1 では 5, Imp_2 では 5, Imp_3 では 30, Imp_4 では 5, となった。また、複合語の最初と最後の単名詞に 1.5 の重みをかけた変形 Imp でも最適な窓幅は全て 5 であった。これら最適の窓幅でかつ最適の複合語率しきい値の場合を一覧表にして表 4 に示す。なお、変形 Imp については一番良い結果を得た Imp_1 の変形 (最初の単名詞を 1.5 乗する Imp_{S1} と最後の単名詞を 1.5 乗する Imp_{E1}) のみ記す。ちなみに、最適な E を与える複合語率のしきい値も示してある。また一例として Imp_1 の場合のマニュアル毎の値 (再現率、適合率、抽出索引語数も加えた。) を表 5 に示す。

我々は avE を選択の基準とするので、全しきい値に対して平均的に見て最も良い E すなわち avE を与えるのは Imp_1 (積) である。よってここでは Imp_1 を選ぶ。また Imp_1 を用いた場合の複合語率のしきい値の準最適値としては 0.2 になる。これらの準最適なしきい値を用いた場合の再現率、適合率は表 6 のようになる。

また、ふたつのパラメータの全て値の組み合わせにおいて最適な E となる場合の再現率、適合率を表 7 に示す。

これらの結果を見ると複合語を構成する各 $Pre(N), Post(N)$ の積である Imp_1 を用い、窓の幅 = 5, 複合語率のしきい値 = 0.2 として得た適合率・再現率は、最適な場合の適合率・再現率とかなり近いことが分かる。このことより、本研究の方法は必ずしも最適ではないが、マニュアルの索引語

表 4: $Imp_1 \dots Imp_4$ と変形 Imp_1 に対する窓幅毎のデータのマニュアル平均値

Imp	窓幅	複合語率 しきい値	$av\bar{E}$
	$min\bar{E}$	再現率	適合率
Imp_1	5	0.24	0.373
	0.315	0.705	0.675
Imp_2	5	0.28	0.426
	0.348	0.698	0.636
Imp_3	30	0.30	0.376
	0.320	0.742	0.634
Imp_4	5	0.16	0.428
	0.339	0.751	0.594
$ImpS_1$	5	0.24	0.396
	0.332	0.698	0.651
$ImpE_1$	5	0.24	0.391
	0.331	0.705	0.647

表 5: Imp_1 に対するマニュアル毎のデータ

マニュアル		avE	$minE$
最適 複合語率	再現率	適合率	抽出語数
JUMAN		0.405	0.365
0.6	0.654	0.618	110
SAX		0.388	0.311
0.0	0.676	0.702	208
たまご		0.343	0.323
0.4	0.667	0.687	99
ビデオ		0.367	0.330
0.2	0.822	0.565	384
ゲーム機		0.361	0.247
0.0	0.707	0.806	36
マニュアル平均		0.373	0.315
0.24	0.705	0.675	-

を選ぶのに十分有効であることが証明された。

5 他研究との比較

この節では、以上述べてきた Imp とりわけ Imp_1 を用いた索引語抽出法と、他の索引語抽出法との比較を行なう。ただし、この論文で述べているようなハイパーテキスト化を目的にした索引語抽出法そのものの研究はほとんどないので、従来の文献検索におけるキーワード抽出法のうち類似の方法と比較することにする。

伝統的なキーワード抽出法では大抵与えられた文書での単語の出現頻度に依っている。そこで形態素解析の結果として得られた単語のうち出現頻度の高いものを索引語として選択する方法を表 2 の 5 つのマニュアルに適用してみた。最適な $E(b=1)$ の値と、その場合の再現率、適合率、抽出索引語数は表 8 の通りである。

自動的な索引語抽出において問題になるのは出現頻度のしきい値の選び方である。そこでここでは最適な E の得られる場合の (抽出索引語数)/(全索引語候補数) に着目し、このマニュアル平均をしきい値としてみた。具体的にはマニュアル平均の

しきい値は 40.7% である。このしきい値によって選択された準最適な索引語のマニュアル毎の E 、再現率、適合率は表 9 の通りである。

これらいずれをとってみても、前節で述べた Imp_1 を用いた方法より E は大幅に劣る。これは、単なる出現頻度を用いると重要だが出現頻度の低い複合語が選択できないことによる。

次に複合語の性質を利用してキーワード抽出を行なう他研究と比較してみる。本研究と最も類似しているのは亀田による重要キーワード抽出 [2] である。この方法は複合語内に現れる単名詞の頻度を考慮しているので、我々の提案する $Pre(N), Post(N)$ に近い性質を取り出す。また、複合語の構成単語数も考慮しているが、これは我々の Imp_1, Imp_3 が複合語の構成単語数が長くなるほど大きな値を持つのに近い。ただし、[2] ではこのふたつの要因が最適化された重みの線形和として扱っている。[3] も複合語を構成する個々の単名詞に適当な重みを与え、これらを加算を基本とする演算を適用して複合語の重要度を計算している点で、[2] や我々の方法に似ている。[3] の特徴は個々の単名詞の重みをその複合語内での位置によって計算している点である。このように [2, 3] と我々の方法は似ているが、我々は個々の単名詞

表 6: Imp_1 の窓方式のパラメーターを固定した際の 5 つのマニュアルの再現率・適合率

マニュアル	再現率	適合率	E
JUMAN	0.731	0.510	0.399
SAX	0.606	0.724	0.340
たまご	0.716	0.629	0.330
ビデオ	0.822	0.565	0.330
ゲーム機	0.683	0.824	0.253
マニュアル平均	0.712	0.650	0.330

窓の幅 = 5, 複合語率のしきい値 = 0.2

表 8: 出現頻度による索引語抽出法の最適値データ

マニュアル	最適な E	再現率	適合率	抽出索引語数
JUMAN	0.544	0.500	0.419	124
SAX	0.322	0.944	0.528	386
たまご	0.509	0.912	0.336	277
ビデオ	0.447	0.943	0.392	636
ゲーム機	0.322	1.000	0.512	80

表 9: 出現頻度による索引語抽出法の準最適値データ

マニュアル	準最適な E	再現率	適合率	抽出索引語数
JUMAN	0.590	0.913	0.265	359
SAX	0.370	0.787	0.525	324
たまご	0.537	0.961	0.306	320
ビデオ	0.468	0.765	0.408	495
ゲーム機	0.431	0.707	0.475	61

表 7: 5 つのマニュアルの最適な再現率・適合率

マニュアル				
窓の幅	複合語率	再現率	適合率	E
JUMAN				
20	0.5	0.692	0.600	0.357
SAX				
5	0.0	0.676	0.702	0.311
たまご				
20	0.5	0.667	0.708	0.313
ビデオ				
5	0.2	0.822	0.565	0.330
ゲーム機				
30	0.2	0.732	0.789	0.241

の重要度をいかに多くの複合語を構成するのに寄与するかという観点から特徴づけた点が発想としては異なるといえる。そして我々の Imp_1 は複合語を構成する単名詞の重要度の積であるから、他の方法に比べて複合語を構成する単名詞数が大きく寄与する。 Imp_1 が再現率、適合率において最も良い結果を得たわけだから、我々の提案する索引語抽出方法はここで比較した他の方法に劣らない可能性が高いといえよう。さらに我々は窓を用い、窓内の複合語率をしきい値に用いることにより、文書の長さ依存する度合いの低いしきい値を考案した。この点は我々独自のものである。

6 おわりに

本論文において、日本語マニュアル文から重要な索引語を抽出する新しい方法を提案した。抽出した索引語を活用した読みやすい電子化マニュアルは、進歩したソフトウェアシステムをも含む今日の複雑なシステムのユーザにとって必須である。

参考文献

- [1] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. 日本語形態素解析システム juman 使用解説書 version1.0, 1993.
- [2] 亀田雅之. 疑似キーワード相関法による重要キーワードと重要文の抽出. 言語処理学会第 2 回年次大会, pp. 97-100, 1996.
- [3] Ayako Bessyo Yasushi Ogawa and Masako Hirose. Simple word strings as compound keywords: An indexing and ranking method for japanese text. *SIGIR'93*, pp. 227-236, 1993.