自然言語処理 113－16
（1996．5．29）

日本語マニュアル文における名詞間の連接情報を用いた重要語の抽出

○松崎 知美, 雨宮 秀文, 森 辰則, 中川 裕志

横浜国立大学工学部

ある文献からその索引語となるべき重要語句を見つけ出し、索引を作ることは、その文献を読む人の大きな手助けとなると思われる。本研究では名詞間の連接情報を用いて重要語抽出を試みた。まずマニュアル文から名詞句を取り出し、ある名詞に対して、前方に連接される名詞、後方に連接される名詞の数を調べた。この連接される名詞の種類が、各名詞の複合語を合成する力である。複合語合成力の強い言葉の組合せが重要な複合語であるというアイデアにより名詞の重要度を計算した。重要度の順にソートした名詞群からある特性を用いて重要語を抽出することに成功した。

Key Word Extraction used Information of Noun-to-noun Connection of Japanese Manual Sentence

○ Tomomi Matsuzaki, Hidefumi Amamiya, Shin Wake, Tatsunori Mori, Hiroshi Nakagawa

Engineering, Yokohama National University

If manuals which are hard to read are hyper-textized when they encounter a word in reading the manual and couldn't remember meaning of it, users can get to access the meaning of that word just by clicking the word. This kind of words are characaterized as index words. Then index words extraction mechanism should be different from the traditional keywords extraction both in its concept and method. In this report, we propose a new method to extract index words which is based on word's power to generate compound nouns, because majority of index words in manuals are compoud nouns which include special nouns that express the main and/or key concept of the apparatus described in the manual.

1 はじめに

最近、WINDOWSのHELP機能やHTMLで記述された文書ファイルなどで見られるように、ブラウザ上で語句をクリックする事によってテキストからテキストへのハイパーリンクをたどる事が出来るようなツールが増えてきた。このようなツールの利点としてユーザが必要な知識、例えば理解できない語句の意味といったような情報をマウスの簡単な操作によって得られる事があげられる。この利点をマニュアルに生かすと、そのマニュアル自体が非常にユーザに理解しやすい便利なものになる。しかし、このようなツールを用いてマニュアルを作る際には、上で述べたようなユーザが情報を必要とする語句をあらかじめマニュアルライターが選ばなくてはならない。しかし、そのような語句を客観的に抽出する事は難しく、また、マニュアル自体の長さが長いものになるとその抽出作業は容易ではない。

そこで重要語の自動抽出が必要となる。従来の重要語抽出の研究では、主に文献の検索・分類・抄録作成用の索引データベースの作成のためのキーワード抽出を目的とし[荒木81, 細野83, 竹内86, 梅田86, 木本87, 林94]、抽出したキーワードを元にした大規模な文献検索システムのための研究[有田94]もなされている。

本研究も日本語のマニュアル(取り扱い説明書)を対象とした重要語抽出法について行なわれたが、従来の研究のように文献検索システムを目的とした重要語抽出ではない。本稿では上記のハイパーテキスト化されたマニュアルにおいて、ユーザがその定義やその語に対するより詳細な情報を必要とするであろう語句(すなわちインデクスの候補になる語であり、かつ上記のハイパーリンクを張るべきと思われる語句)を対象マニュアルの重要語であると考え、その自動抽出法を述べる。

2 理論的背景

2.1 重要語抽出の流れ

本研究の重要語抽出の流れは図1の通りである。この内容について細かく以下に説明する。

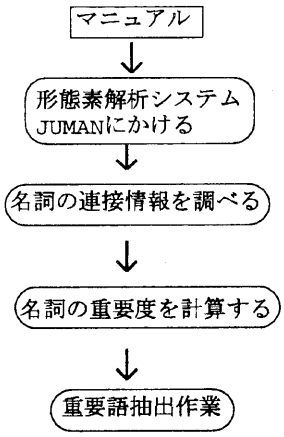

図1: 重要語抽出の流れ図

2.2 名詞の連接情報

重要語となる名詞句を探すに当たり、本研究では名詞の造語力というものに着目した。つまり、重要語の一部になるような名詞はマニュアルの記述するシステムや機械において重要な概念を表す言葉であり、いろいろな名詞と連接してたくさんの複合語を作る、という考え方である。本研究では対象とするマニュアル中の名詞と名詞の連接に着目し、一つのマニュアル中のすべての名詞について、前方および後方に連接される名詞の種類を数えた。前方に連接される名詞の種類数をその名詞の前方連接数、後方に連接される名詞の種類数をその名詞の後方連接数とした。(図2参照。)また、連接を扱う際には、「(名詞)の(名詞)」のように、名詞と名詞が「の」でつながっているものも名詞どうしの連接とした。マニュアル文においては「私の…」、「昨日の…」といった語が出てくる事がないのでこのような扱いが可能になる。

具体的には、

1 マニュアルを、形態素解析システムJUMANに通し、マニュアル文を形態素に分割した。

2 名詞、未定義語、および助詞「の」を取り出す。

3 全ての名詞について、前方や後方に連接される名詞の種類を数える。

という作業を計算機で自動的に行なった。ここで、形態素解析システムJUMANの辞書に登録されている名詞が、名詞の単位となる。このため、特別な語、複雑な複合語をこの辞書に登録することは避け、なるべ

く基本的な名詞が名詞の単位となることを目標とした。

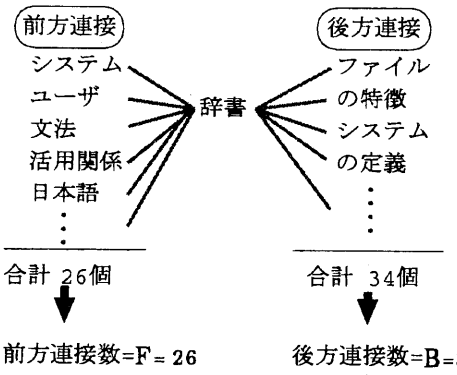

図 2: 連接数のカウント

2.3 重要度計算

上のような考え方を元に、「造語力の高い名詞からなる名詞句(複合語)は重要度が高い」ものとして、マニュアル中の名詞句の重要度を計算した。この際、例えば、「形態素解析辞書」という名詞句の一部である「形態素」や「形態素解析」といった名詞句も、本文中に出てくれば重要語となる可能性は十分にあるわけである。更に、「の」でつながった名詞句、「A の B の C」のようなものは、「A の B」、「B の C」、「A」、「B」、「C」の全てについて、重要語とする可能性がある。そこで以下のようにして複合語(名詞句)の重要度の計算を行なった。すなわち、名詞句 N を n 個の名詞からなる複合語「名詞 1, 名詞 2, …, 名詞 n」とする。ただし名詞 i($1 \leq i \leq n$) は、名詞+「の」でもよい。名詞 i の前方連接数を F_i、後方連接数を B_i とするとき、この名詞句 N の重要度 J は、相乗平均の考え方を用いて、次式で計算した。相乗平均の他に相和平均を用いることも考えられるが、相和平均は一つでも大きい値を含むと結果として大きい値を出す傾向が相乗平均に比べて強い。(例えば 90 と 10 の相乗平均は 10、相和平均は 50 である。) また、日本語は head final の性質がある(後ろに主要な意味を表す語が来る)ので、複合語中の一番後ろの名詞の前方連接数に重みを付けることも試したが、あまりよい実験結果が得られなかったので、重みを付けないほうが良いとの結論に達した。

名詞句	重要度 J
辞書	19.90
形態素辞書	17.18
形態素	14.83
形態素辞書ファイル	13.52
形態素の連接	13.25
辞書ファイル	12.90
活用辞書	12.20
形態素コスト	12.18
形態素辞書の記述	12.10

表 1: 重要度 J の特に高かった名詞句

$$J = [\prod_{i=1}^{n} (F_i + 1) \times (B_i + 1)]^{1/2n}$$

実際に日本語形態素解析システム JUMAN のマニュアルについて、出現する名詞の全てについてこの計算を行なった結果、特に重要度の高かったものとその具体的な値を表 2.3に示す。

3 この重要度と従来の出現頻度によるキーワード抽出の比較

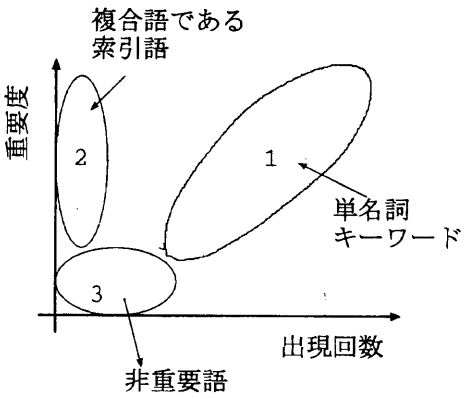

図 3: 重要度と出現頻度のプロットの一般的傾向

従来の文献検索のための重要語抽出の研究では、名詞の出現回数というものが重視されてきた。名詞の出現回数と、本研究で定義し、計算した名詞の重要度というものが、どのような関係に当たるのかを以下に述べる。いくつかのマニュアルについて、出てくる名詞の全てについて、重要度と出現回数を求めた。その結果を、横軸に出現回数、縦軸に重要度を取ってプロットすると、図3のような傾向で名詞が分布する事が分かった。まず、重要度も出現頻度も低い語(図中3の領域)は、非重要語である。図4は、実際に日本語形態素解析システムJUMANのマニュアルについて、このデータを取ったものだが、例えば、この3に当たる領域には、「ファイル」、「解析」、「文法」、「存在」、「情報」、「規則」、「構造」、「実行」、「システム」、「数」、「指定」といった、日本語において一般的な短い名詞が分布していた。図3の1の右上がりの領域にも名詞が分布する。この領域は、出現回数も重要度もある程度以上のものである事になる。出現回数がある程度以上の名詞というのは、重要度の高い複合名詞の一部となるような単名詞である事が多い。具体例としては、図4においては、「辞書」、「形態素」、「連接」、「コスト」、「活用」、「品詞」、「語」、「定義」、「記述」といった名詞であった。これに対して、図3の2の領域、すなわち、出現回数がさほど大きくなくても、重要度が高い領域に、高重要度の複合名詞句が分布する。図4で見ると、「形態素辞書」、「形態素辞書ファイル」、「活用辞書」、「形態素コスト」、「形態素解析」、「連接規則」などである。

このような散布図の具体例としてもう一つ、図5に、ある市販の家庭用ビデオデッキについての結果を掲げる。

ここで、問題になるのは、このようにして重要度を求めた全ての名詞から、実際に重要語としてどの名詞を抽出するかが問題になる。それを次節に述べる。

4 重要語の抽出法

前節までに求めた重要度の値は、マニュアルの規模や性質に依存するので、絶対的な値とはいえない。そこで、あるマニュアルについて、全ての名詞の重要度を計算する事は出来るが、どの名詞を重要語として抽出するかという事は難しい問題である。例えば、ある重要度以上の語を抽出するとしても、その境界値を、一つ一つのマニュアルについて人間が判断するのではなく、自動的に決定したい。そこで本研究では、名詞句を重要度の順にソートした後、その上で窓を動かし、傾向を調べた。

	19.90	辞書
	17.18	形態素辞書
	14.83	形態素
	13.52	形態素辞書ファイル
⇓	13.25	形態素の連接
	12.90	辞書ファイル
	12.20	活用辞書
	12.18	形態素コスト
	12.10	形態素辞書の連接
	11.83	連接
⇓	11.22	辞書システム
	10.93	辞書の記述
	10.88	連接コスト
	10.85	文法辞書
	10.44	形態素解析
	10.34	形態素辞書の指定
	10.26	連接規則辞書
	10.17	システム辞書ファイル
	10.00	コスト
	9.63	活用辞書の記述
	9.43	形態素の記述
	9.37	形態素文法
	8.97	連接規則辞書の記述

図 6: ソートした名詞の上で窓を動かす

窓を動かすとは図6のように、重要度順にソートした名詞を、上から順にある幅の窓中の名詞について傾向を見るものである。ただし、マニュアル中の出現回数が1回である名詞は、電子化したマニュアルにハイパーリンクを張るということが不可能なのであらかじめ除去した。ここで調べたのは、窓内の名詞のうち正解重要語の占める割合と、複数の名詞からなる名詞(複合語)の占める割合である。そこでまず、正解重要語の決定法として本研究で用いた方法を以下に述べる。

4.1 正解重要語の調査

正解重要語の決定は、数人の被験者に、対象としたマニュアルのコピーを読ませ、インデクスとなるべき名詞句を選ばせた。まず、重要語の目安としては以下のとおりであると説明した。

- 当マニュアルが電子化されハイパーリンクが張れるものと考え、マニュアルを読む過程でその後に

対する詳細な説明のリンクを張って欲しいと思われる語。

これは本研究の重要語抽出の目的を説明したものである。選び方は以下のようにした。

- 名詞と名詞が隣合っているものは、その間でマークを切らない。例えば「形態素辞書ファイル」とあった時、「形態素辞書」や「辞書ファイル」のみにマークをしない。
- 名詞間をつなぐ助詞は「の」のみとし、それ以外の助詞はその前後で名詞句が切れるものとする。
- マークする名詞句の長さは自由である。ただし、名詞句を切るのは助詞の「の」のところに限る。例えば「形態素辞書のファイルの定義」という名詞句があった場合、「形態素辞書のファイル」だけや「ファイルの定義」だけにマークするのは可能である。
- 括弧内の単語はないものとして扱う。
- 図や表の中の名詞句はマークの対象としない。

このようにして、1つのマニュアルについて3から4人の被験者に、1ページにつき3つ以上の名詞句を選ばせた。その結果の全てを合わせたものを重要語とした。

このようにして決定された正解重要語が重要度の順にソートした名詞の上で窓を動かした時、窓内にどれだけ含まれるか、その割合の変化を示したグラフが図7である。用いたマニュアルは、日本語構文解析システムSAXのマニュアルである。

重要度がある程度低くなったところから人手で選ばれなかった語ばかりになることが分かる。この選ばれなかった語ばかりになるところの特徴とは何であろうか。それを調べるために、同じマニュアルについて、窓の中の複合名詞からなる名詞の割合を求めた。その結果を図8に示す。

これら2つの相関性を見るために相関係数を計算したところ0.92であった。つまり両者には正の非常に強い相関がある。

そこで、窓を動かしていった際、単独の名詞からなる名詞句の割合が一定の割合を越える部分の抽出を止めるといった方式で重要語の抽出を行なうことにした。その際問題になるのは窓の幅と、抽出を止める単独名詞の割合の最適値を求めるということである。そこで、重要語抽出の評価法について以下に述べる。

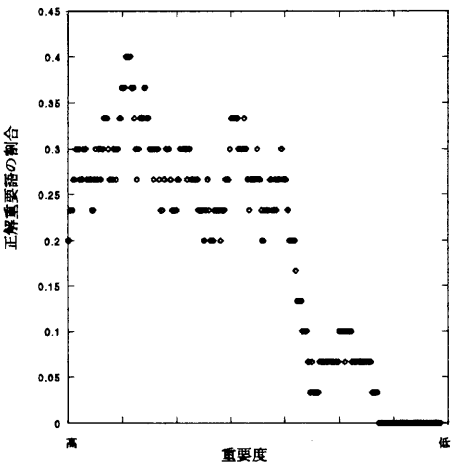

図 7: SAXのマニュアルにおける正解重要語の割合 (窓の幅： 10)

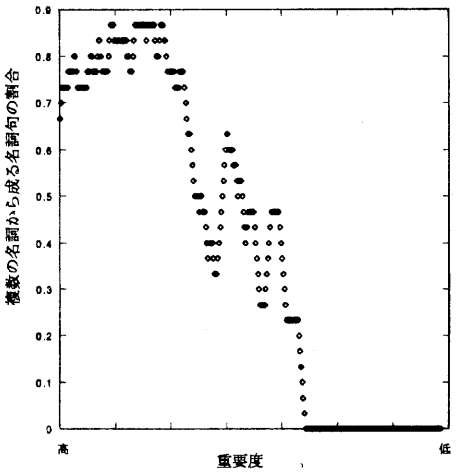

図 8: SAXのマニュアルにおける複数名詞から成る名詞句の割合 (窓の幅： 10)

4.2 抽出語の評価 (適合率と再現率)

従来のキーワード抽出においてはその評価方法として、適合率、再現率といった考え方がとられてきた。これについてまず、説明する。

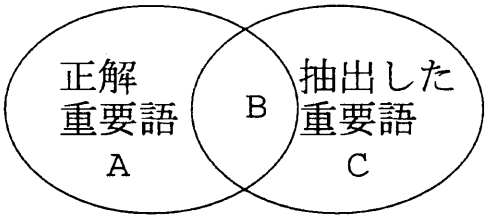

図 9: 適合率・再現率の説明

図 9において A の領域を正解重要語の集合、C の領域を抽出した重要語の集合、B の領域をその両者の一致部分と考える。適合率とは抽出した重要語が正解重要語である割合、すなわち図の B/C、再現率とは正解重要語のうち抽出されたものの割合、すなわち図の B/A を指す。

この両者は、抽出する名詞が多いほど再現率が上がるが、適合率は下がり、また、抽出する条件を厳しくすれば、適合率は上がるが、再現率は下がるという風な裏表の関係にある。これらの両方が高くなることが必要とされる。

4.3 抽出条件の最適化

前節の適合率、再現率が共に良くなるように抽出条件、すなわち窓の幅と抽出を止める単独名詞の割合を決定する必要がある。そこで窓の幅を 5 から 30 まで、抽出を止める単独名詞 (単純語と呼ぶ) の割合を 0.1 から 0.9 まで変化させて、適合率と再現率の積を求めた。当然、適合率と再現率の積が高くなることが望ましい。このような観点から適合率、再現率の積の変化を見た場合、大きく影響を及ぼすのは抽出を止める単純語の率であることが分かった。そこで、3 本のマニュアル、日本語形態素解析システム JUMAN(松本裕治ら. 日本語形態素解析システム JUMAN 使用説明書 version1.0. 1993.)、家庭用ビデオデッキ (三菱電気株式会社. HV-F93)、家庭用ゲーム機 (Sony Computer Entertaiment Inc. PlayStation) について抽出を止める単純語の率を横軸に、適合率と再現率の積を縦軸にとったグラフを図 10から 12に載せる。ここでいう、窓の幅とは、窓に含まれる名詞の個数である。

これらのグラフから、適合率と再現率の積が高いところで安定するのは抽出を止める単純語の率が 0.7 程度のところということが分かる。また、窓の幅は小さい方が良いことが分かった。そこで窓の幅 5、抽出を止める単純語の率 0.8 としてテストセットに適用してみた。

テストセットとしては、2 本のマニュアル、日本語構文解析システム SAX(松本裕治ら. 構文解析システム SAX 使用説明書 version2.0. 1993.) と仮名漢字変換システム「たまご」のマニュアルを用いた。これらについて以上に述べてきた方法で、重要語抽出を行なったところ、適合率、再現率は表 2のようになった。

表 2: テストセットへの適用結果

マニュアル	適合率	再現率
構文解析システム SAX	34.6 %	71.2 %
仮名漢字変換システム「たまご」	26.5 %	66.0 %

これらの値は、サンプルセットでとった値に劣らないものである。ゆえに、前節で求めた抽出条件が適当であったといえるだろう。

5 まとめ

文献検索ではなく、マニュアルを読む際にリンクが張ってあって説明がなされるとユーザの助けになる語を重要語として重要語抽出を行なった。名詞の造語力に着目し、「造語力の高い名詞からなる複合語は重要度が高い」ものとして、名詞の重要度を定義し計算することに成功した。この重要度と従来良く用いられた出現回数について比較し、検討を行なった。更に重要度をつけた名詞から、重要語を抽出する方法として窓を用い、窓の中の複合語の割合により抽出するものを決定する方法を考えた。このようにして、重要語を抽出するシステムとしての性能の評価として、従来のキーワード抽出に用いられていた、適合率、再現率の考え方を用いた。このため、被験者による正解重要語の抽出も行なった。窓を用いた抽出の細かい条件は、3 本のマニュアルをサンプルセットとして決定し、2 本のマニュアルについて適用した結果、望ましい値が

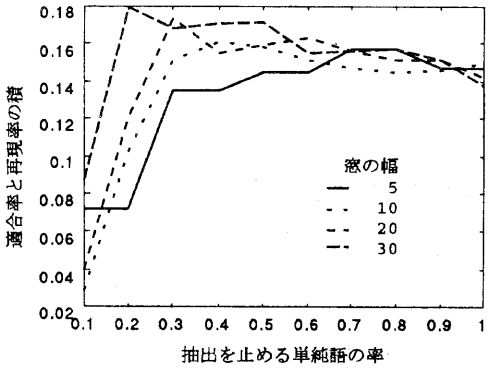

図 10: 日本語形態素解析システム JUMAN

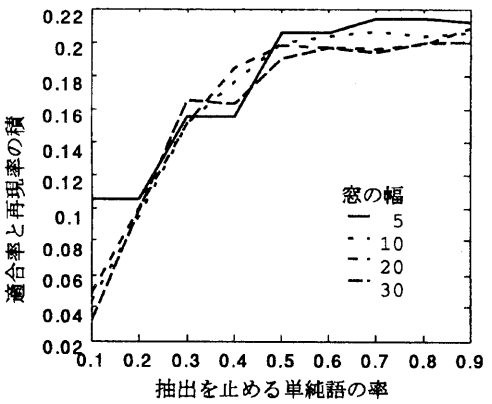

図 11: 家庭用ビデオデッキ

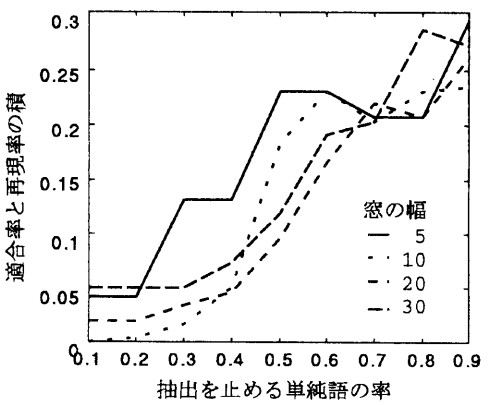

図 12: 家庭用ゲーム機

得られることが分かった。

今後は、この重要語を用いて自動ハイパーテキスト化を行なうことが課題となる。また、本研究で用いたような名詞の連接情報を用いて、マニュアル独自のシソーラスの自動作成についての研究を予定している。

6 謝辞

本研究の初期において多大な御尽力をいただいた現 NEC の和気真氏に感謝致します。

参考文献

[荒木 81] 荒木啓介 金子明夫 高野文男 日夏健一. 日本語論文タイトルからのキーワード自動抽出. 情報処理学会研究報告 NL-26-3, 1981.

[細野 83] 細野公雄 後藤智範 諸橋正幸. パターン・マッチングによる重要語の自動抽出. 情報処理学会研究報告 NL-39-1, 1983.

[竹内 86] 竹内晴彦 岩坪秀一 西野博二. 多変量解析によるキーワードの自動抽出と文献の自動分類. 情報処理学会研究報告 NL-54-2, 1986.

[梅田 86] 梅田茂樹 諸橋正幸 細野公雄 原田隆史 後藤智範. 漢字クラスターによる日本語文献の重要語抽出. 情報処理学会研究報告 NL-58-5, 1986.

[木本 87] 木本春夫. キーワード自動抽出と重用度評価. 情報処理学会研究報告 NL-64-1, 1987.

[林 94] 林淑隆 獅々堀正幹 伊与田敦 津田一彦 青江順一. 複合語キーワードの効率的抽出法. 情報処理学会研究報告 NL-104-9, 1994.

[有田 94] 有田健 津田和彦 入口浩一 青江順一. 複数キーワードによる検索の一高速化の手法. 情報処理学会研究報告 NL-104-7, 1994.

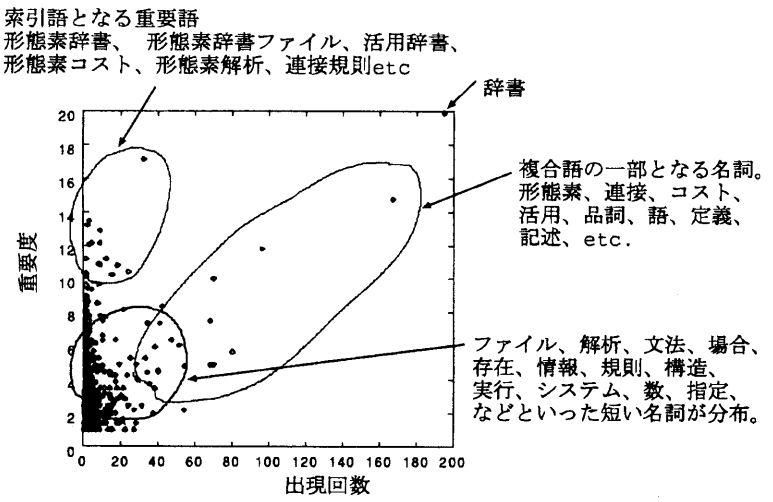

図 4: 形態素解析システム JUMAN のマニュアル

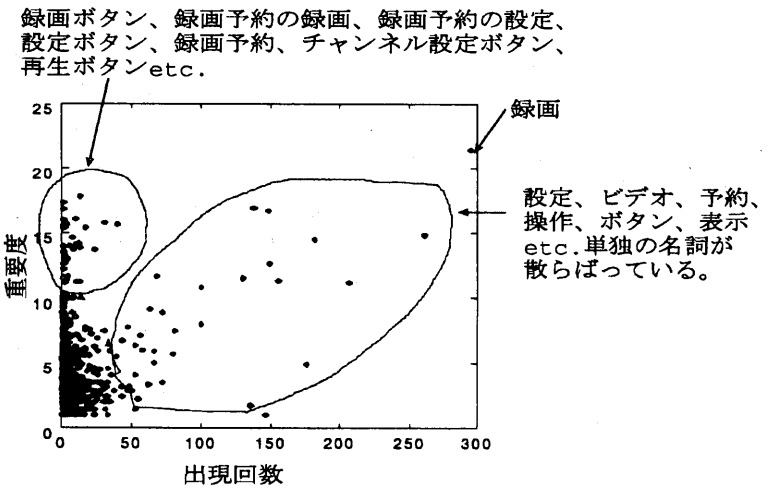

図 5: 家庭用ビデオデッキのマニュアル