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

4B－1

松崎 知美　和気真　森辰則　中川裕志　*

横浜国立大学 工学部

1　はじめに

重要語の抽出を行ない、抽出した重要語を用いて、「知識オンデマンド」とでも呼べるようなシステムの開発ができないだろうか。例えば、何らかの作業の途中で、定義を忘れた言葉があれば、その言葉をクリックすれば、すぐ定義を調べられるようなシステムを、である。そのためには、３つ４つといった程度のキーワードではなく、多数の重要語が必要となる。多数の重要語にハイパーリンクを貼ることにより、必要な知識をいつでも取り出せるシステムの構築が可能になると思われる。文献検索において使用される、その文献を特徴付けるキーワード抽出の研究は、これまでにも盛んに行なわれてきた。しかし、我々が抽出したのは、このようなキーワードを含むかもしれないが、必ずしも一致するものではない。むしろ、上記のようなシステムに対する必要性から、索引語となるべき重要語に視点を絞り、重要語抽出の研究を行なった。以下に、このような重要語抽出の研究について報告する。

2　重要語の抽出方法

2.1　理論的背景

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示す言葉であり、いろいろな名詞と連接してたくさんの複合語を作るのではないか、という考え方である。本研究では、マニュアル文を研究対象とした。また、名詞と名詞の連接に着目し、一つのマニュアル中で、全ての名詞に対して、連接される名詞の種類を数えた。これは、本研究で用いたマニュアル文において重要語を探す場合、有効な考え方だと思われる。マニュアル文においては、「私の…」、「昨日の…」といった語が出て来ることはないので、このような扱いが可能となる。このようにして、全ての名詞の、前方に連接される名詞の種類と、後方に連接される名詞の種類をカウントした。連接される名詞の種類が多かった名詞を表1に示す。

名詞	解析	形態素	辞書	システム	連接	ファイル
前方連接する名詞の種類	5	10	21	7	13	13
後方連接する名詞の種類	8	19	17	4	9	4

表 1:「日本語形態素解析システム JUMAN の使用説明書」中で前方および後方に連接される名詞の種類が多かった名詞

2.2　計算法

上のような考え方を元に、「造語力の強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、重要度の計算を行なった。具体的な計算式を以下に示す。名詞句Ｎをｎ個の名詞からなる名詞1, 名詞2, …, 名詞ｎとする。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。

*Index Word Extraction based on Noun-to-Noun Connections in Japanese Manual Sentences

by Tomomi Matsuzaki, Shin Wake, Tatsunori Mori and Hiroshi Nakagawa,

Yokohama National University, Tokiwadai, Hodogaya-ku, Yokohama 240, Japan.

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 2: 重要度 J の特に高かった名詞句

相乗平均:$J =$
$[\prod_{i=1}^{n}\{($ 名詞 i の前方に連接された名詞の種類数 $+1)$
$\times($ 名詞 i の後方に連接された名詞の種類数 $+1)\}]^{1/2n}$

3 実験結果とその検討

実験に用いたのは、日本語形態素解析システム JUMAN の使用説明書である。重要度 J の値が特に大きかったものを表3に示す。相乗平均による重要度がどの程度有効であるかを調べるために、従来よく行なわれてきた出現回数による抽出との比較を行なった。すべての名詞句のマニュアル中における出現回数を計算し、個々の名詞句について、出現回数を横軸に、算出した重要度を縦軸にプロットした。その結果が図 1である。出現回数が多い名詞というのは、どうしても、複合語の一部となることが多いような短い名詞が多くなる。複雑な概念が盛り込まれた複合名詞はこれらの短い名詞に比べると、出現回数は少なくなってしまう。プロットされた具体的な名詞を見ていこう。重要度、出現回数ともに最大の名詞句は、「辞書」(グラフ中一番右上のプロット)であった。グラフの原点からこの「辞書」のプロットへ向かって、右上がりに並ぶ、出現回数 50 回以上の点(図中 2 の領域)は、「形態素」、「連接」、「コスト」、「活用」、「品詞」、「語」、「定義」、「記述」と、全て複合語の一部となるような短い名詞であった。このような名詞は、出現回数が多いからといって、このまま索引語とするには適さないであろう。これに対し、出現回数は 50 回以下でも、重要度 10 以上、つまりグラフの左上にプロットされた名詞句(図中 1 の領域)は、「形態素辞書」、「形態素辞書ファイル」、「辞書ファイル」、「活用辞書」、「システム辞書」、「文法辞書」、「形態素解析」、「連接規則辞書」などである。実際にこのシステムを利用している人を被験者としてマニュアル中から重要語を選んでもらったところ、これらの語が選ばれていた。このことからもこの計算法による重要度は意味のあるものであるといえるだろう。図中 3 の領域はどうだろうか。この辺りには、「ファイル」、「解析」、「文法」、「場合」、「存在」、「情報」、「規則」、「構造」、「実行」、「システム」、「数」、「指定」といった短い名詞が分布している。これらも、索引語にはむかない。以上まとめるとこの重要度計算は、複雑難解なマニュアルから、索引語とするのに望ましい難解な複合語、すなわち、そのマニュアル独特に頻出する概念からなる複合語を抽出する際に効果を期待できる。縦軸の重要度を求めたことによって、図 1のグラフの 1 の領域と 3 の領域を分離できた。なお、マニュアルによって変わってしまう、1 と 3 領域を分ける閾値の求め方が今後の課題である。

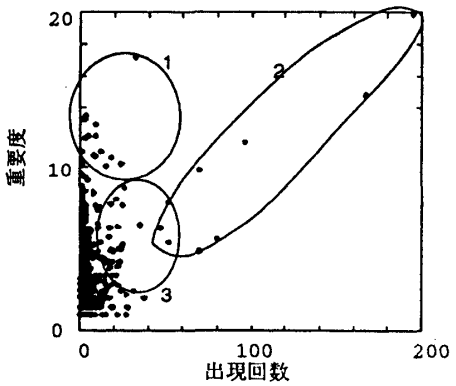

図 1: 出現回数と重要度の分布

最後に、JUMAN 使用説明書を分析することを快諾して頂いた奈良先端大の松本裕治教授に謝意を表します。

参考文献

[1] 日本語形態素解析システム JUMAN 使用説明書 version 1.0