# 文字列に対する重要度付けによるキーワードの抽出

3Y－2

塚田政嘉，　黒川恭一
防衛大学校情報工学教室

## 1　はじめに

近年、電子メディアの普及に伴い大量の情報を容易に入手する事が可能になった反面、それらの膨大な情報の中から必要とする情報を効率よく見つけ出すことが困難になりつつある。そのような中で、原情報で述べられていることを的確に表している語をもとにして、入手する情報を絞り込む方法の１つとしてキーワード検索がある [1]。個々の原情報からその内容を的確に表現しているキーワードを自動的に抽出する方法はこれまでにも種々提案されている [2][3][4][5]。本稿ではキーワードを抽出するための文章として科学技術論文に限定し、カタカナ及びカタカナを含む語に注目することによりキーワードを自動的に抽出する方法を提案する。

## 2　重要度付け

キーワードの重要度付けとしては、基本的に文章中に出現する語の表層上の特徴 (出現位置、出現頻度、それらが用いられた表現等) をもとに付与する方法が文献 [6] で提案されている。そこで本稿でも、この方法を用いることとした。

### 2.1　候補語の限定

キーワードの候補となるのは、カタカナのみで表記された語もしくはカタカナと漢字の組み合わせで表記された語 (以後キーワード候補語と呼ぶ) とした。

これは、科学技術論文においては用語がカタカナもしくはカタカナと漢字の組み合わせで表記されていることが多く、さらにそれらの語が論文の内容を表している度合いが強いためである。そのことを確かめるために、カタカナ及びカタカナと漢字の組み合わせで表記された語が、論文に記載されたキーワード全体に占める割合を電子情報通信学会の論文誌を対象に調べてみた。電子情報通信学会の 1997 年 11 月号の論文誌 A,B-I,D-I において掲載された論文 48 件では、カタカナ及びカタカナと漢字の組み合わせで表記された語はキーワード全体の 213 個に対して 103 個 (全体の 48.4 %) であった。このことから、科学技術論文におけるキーワードの自動抽出において、カタカナ及びカタカナと漢字の組み合わせで表記された語に注目することは有効であると考える。

ただし、カタカナだけで表記された語に関しては「システム」や「コンピュータ」等の広い意味を持つ語もキーワードとして抽出してしまうので、この様な広い意味を有する語に関してはあらかじめキーワードとしての禁止語リストに登録することによりキーワードとして抽出しないようにした。

### 2.2　重要度付けのルール

キーワード候補語に対して、以下に示す重要度付けのルールに従って点数を付与するものとした。

(1) タイトル：タイトルに含まれる場合は２点。
(2) 助詞：論文のあらましと結論においてキーワード候補語の後に「は」「が」「を」がある場合に１点、本文中では 0.5 点。
(3) カタカナの前もしくは後に漢字が１字しかない場合はその漢字を削除してキーワード候補語とする。
(4) カタカナ＋漢字の順の語で、カタカナ全てと漢字の最初の２文字が一致している場合は同じ意味合いを持つ語とする。
(5) 論文中に同一のキーワード候補語が複数回出現する場合にはそれらの合計点をそのキーワード候補語の点数とする。

## 3　キーワードの抽出

### 3.1　抽出方法

キーワードの抽出は、キーワード候補語に上記のルールに従って付与された点数が、設定したしきい値を超えた場合に行うものとした。このしきい値の設定は、論文の長さによらず、それぞれの論文に含まれたキーワード候補語全ての点数の合計 (以後トータル点数と呼ぶ) にもとづいて決定した。これは、カタカナ及びカタカナと漢字の組み合わせで表記された語が出現する頻度が論文の長さに依存するとは言いきれないため

A Method of selecting keywords by attaching important value
Masayoshi Tsukada and Takakazu Kurokawa
National Defense Academy, 1-10-20 Hashirimizu, Yokosuka 239, Japan

ナと漢字の組み合わせで表記された語が出現する頻度が論文の長さに依存するとは言いきれないためである。

### 3.2 抽出精度の評価法

キーワードの抽出精度は、一般的に再現率と適合率の組み合わせによって評価されると文献 [3] において定義されている。本稿においても抽出精度の評価には、キーワードとして抽出した語が論文に記載されたキーワードと合致した数 (以下「ヒット数」と呼ぶ) をもとに、適合率と再現率で判定した。ここで、

$$\text{再現率} = \frac{\text{ヒット数}}{\text{抽出したキーワード数}} \quad (1)$$

$$\text{適合率} = \frac{\text{ヒット数}}{\text{論文記載のキーワード数}} \quad (2)$$

とした。

## 4 評価

今回提案した方式の評価を、以下のように行った。まず重要度付けの対象とした文献は、情報工学に関連した論文 20 件であり、各論文に記載されたカタカナ及びカタカナと漢字の組み合わせで表記されたキーワードの数が 2～4 のものとした。なお、これら 20 件の論文に記載されていたキーワードの数は全部で 89 個であり、そのうちカタカナ及びカタカナと漢字の組み合わせで表記されたキーワードの数は 56 個 (全体の 62.9 %) であった。

これらの科学技術論文においてトータル点数に対してしきい値を変化させたときの再現率と適合率を調べた。表 1 はしきい値をトータル点数の 1/5～1/15 まで変化させたときの再現率と適合率の平均を表したものである。表 1 からしきい値を高く設定するとキーワードとして抽出される語数が少なくなるために適合率は向上するが再現率は低くなり、一方しきい値を低くするとキーワードとして抽出される語数が増えるので再現率は向上するが適合率が低下することが分かる。このように再現率と適合率との間にトレードオフの関係が存在するため、今回調べた結果からは、しきい値をトータル点数の 1/11～1/12 付近に設定して自動的にキーワードを抽出することが最も適していると考える。

## 5 むすび

本稿では、科学技術論文からキーワードを抽出する場合においてカタカナに注目し、自動的にキーワードを抽出する方法を提案した。これまでに発表された 20 件の論文を対象にキーワードの抽出を行った結果、しきい値の値によって比較的高い再現率と適合率が得られることが分かった。近年、カタカナ表記された語が多用されていることが社会現象として報道されている。カタカナ表記された語の出現頻度はこれからも高まり、科学技術論文のみならず、様々な分野の文章においてその重要度が高まるものと考える。今後、調査する文献、文章数を増やし、より詳細な検討を進めて行く予定である。

## 参考文献


[1] 諸橋正幸: "自動索引付け研究の動向," 情報処理,25,9,pp.918-924,1984.

[2] 長尾,水谷,池田: "日本語文献における重要語の自動抽出," 情報処理,Vol.17,No.2,1976.

[3] 木本春夫: "日本語新聞記事からのキーワード自動抽出と重要度評価," 電子情報通信学会論文誌,Vol.J74-D-I,No.8,pp.556-566,1991.

[4] 高野,荒木,金子,日夏: "日本語論文タイトルからのキーワードの自動抽出システム(JAKAS)," 情報処理学会自然言語処理研究会資料,26-3,1981.

[5] 細野,後藤,諸橋他: "パターン・マッチングによる重要語の自動抽出," 情報処理学会自然言語処理研究会資料 39-1,1983.

[6] 渡辺日出雄: "文章内容を反映したキーワードの重要度付け," 第 52 回情処全大,5P-1,1996.


表 1 しきい値と再現率及び適合率の関係

| しきい値 | T/5 | T/6 | T/7 | T/8 | T/9 | T/10 | T/11 | T/12 | T/13 | T/14 | T/15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 再現率(%) | 28.6 | 32.1 | 35.7 | 37.5 | 41.1 | 48.2 | 55.4 | 57.1 | 58.9 | 64.3 | 66.1 |
| 適合率(%) | 66.7 | 60.0 | 62.5 | 60.0 | 60.0 | 55.3 | 55.4 | 55.2 | 53.2 | 50.0 | 50.0 |

T : トータル点数