# 構文情報を用いたキーワード抽出

3S−5

野美山　浩[1]　諸橋　正幸[1]　細野　公男[2]　原田　隆史[2]　関根　さゆり[3]　梅田　栄廣[4]
[1] 日本アイ・ビー・エム　東京基礎研究所　[2] 慶應義塾大学　文学部図書館・情報学科　[3] 日立製作所　[4] 日本電気

## 1　はじめに

従来のキーワード抽出手法は、文書を解析することによって自立語を切り出し、それらを独立に扱うことで効率的な検索を実現してきた。しかし、このような手法では、意味的な観点が陽に考慮されないため、重要でない語もキーワードとして抽出されてしまう。その結果、これらのキーワードを用いて検索する際に、重要でない文書も検索されてしまう問題が生じる。

本稿では、検索対象を構文解析することによって得られる語の係り受け関係、否定、時制といった構文情報を用いて、キーワードであるかどうかを評価し、重要なキーワードを選別する手法を提案する。構文解析は、会話型パーザJAWB[1]を用い、正解を教えることで解析を行なった。解析結果は、文節間の係り受け関係を表す木構造として得られる。また、構文的情報に加え、語の意味分類として、分類語彙表[3]を用いた。

同様な研究に木本[2]の研究がある。我々の手法は、構文解析を行なうことで、係り受け関係や、より一般的な構文属性も考慮した。また、その重み付けを事例ベースから計算する手法を用いた。

## 2　構文的観点から見たキーワード出現特長

まず最初に、構文的観点から見たキーワードの出現特長の調査を行なった。調査は、JICSTの抄録中、“コンピュータ”をキーワードとして含む抄録を対象とした(表5中の訓練標本)。まず、実際の文章を構文解析し、その解析結果に対し、キーワードである語を人手で選択し、それを元にキーワードの出現特長を調査した。

以下に構文的観点毎にキーワードの出現特長を調査した結果を示す。なお、本稿において、W(x/y)は、Wの出現数yの内、x個がキーワードであることを表す。

### 2.1　語＋“の”＋名詞とキーワード

特定の名詞の“の”格について、比較的高い確率でキーワードとなった語の一部を表1に挙げる。

表1: 語＋の＋名詞

| 意味分類 | 具体例 |
|---|---|
| 概念 | “概念”(14/14),“概要”(18/20) |
| 特徴 | “特徴”(12/12),“特長”(4/6) |
| 手法 | “手法”(6/9),“技法”(5/5),“法”(23/33) |
| 使用 | “利用”(17/21),“使用”(5/7),“応用”(10/13) |
| 開発 | “構築”(7/7),“改善”(8/14),“開発”(35/46),“更新”(4/5) |
| 例示 | “事例”(6/9),“例”(14/18) |
| 動向 | “動向”(18/21),“傾向”(3/4) |

### 2.2　動詞とその格

動詞とその格となる名詞において、キーワードの出現を調べたものを表2に示す。格は、表層の格を用い、受け身、使役等の考慮はしていない。

### 2.3　「時」を表す語

時を表す語とそれが修飾する部分木中ある語がキーワードかどうかを調べた結果を表3に示す。

### 2.4　否定形

文が否定されている場合、その文の中にはキーワードは含まれない可能性が高いと推測される。否定形の述語の下位ノードにキーワードが含まれているかどうかを調査した結果、否定の形をとる述語が、全部で51あった。その否定形の述語の下の部分木に存在する文節の数(296)のうち、キーワードを含む文節は22であった。

表2: 語＋助詞＋動詞とキーワード

| 助詞 | 語がキーワードである<br>ことが多い動詞 | 語がキーワードでない<br>ことが多い動詞 |
|---|---|---|
| を | “通す”(10/13),“決定する”(12/20)<br>“検討する”(71/137),“介す”(8/12)<br>“分析する”(5/15),“改善する”(10/21)<br>“測定する”(21/41),“計測する”(4/4)<br>“試みる”(18/32),“提示する”(19/25)<br>“予測する”(18/25) | “とりまとめる”(0/4)<br>“強化する”(0/13),“変化する”(0/8)<br>“分ける”(2/6),“分離する”(0/4)<br>“引き下げる”(0/1)<br>“減少する”(0/5),“節減する”(0/2)<br>“短縮する”(0/2) |
| が | “開発する”(10/22) | “存在する”(0/11),“減少する”(0/6)<br>“減る”(0/2),“無い”(0/2)<br>“必要な”(0/3),“可能な”(2/33)<br>分類コード 3.13(繁閑, 敢不適)(0/5)<br>分類コード 3.19(多少, 大小)(3/39)<br>分類コード 3.37(経済概念)(0/3) |
| は | “図る”(4/4)<br>“提供する”(5/8),“利用する”(1/1)<br>分類コード 2.37(所有, 取得)(13/27) | “示す”(1/13)<br>“見える”(0/7),“大きい”(0/5)<br>分類コード 3.50(明暗, 音)(0/2) |

表3:「時」を表す語とキーワード

| 具体例 |
|---|
| “最近”(19/53),“現在”(15/48),“現代”(2/2),“最新”(11/13)<br>“始めに”(6/19),“初期”(1/5)<br>“〜後”(7/14),“その後”(2/23)<br>“〜以降”(3/9),“〜以来”(3/11)<br>“将来”(8/27),“今後”(10/30),“次代”(1/1),<br>“次の”( 0/ 3)<br>“過去”(4/8),“以前”(0/13),“従来”(4/37),“〜前”(1/9),<br>“〜年間”(5/39),<br>“〜年代”(1/2),“〜時代”(3/13),“当時”(0/1) |

### 2.5　相言

形容詞、形容動詞、副詞とキーワードとの関係に着目した分析を行った結果、キーワードを修飾しやすい傾向にある語が存在した。

“新しい”(18/33), “最適な”(12/15)

---

Automatic Indexing Using Syntactic Infomation

Hiroshi Nomiyama[1], Masayuki Morohashi[1],Kimio Hosono[2], Takashi Harada[2], Sayuri Sekine[3], Eiko Umeda[4]
[1] IBM Research, Tokyo Research Laboratory
[2] School of Library and Information Science, Keio University
[3] Hitachi Ltd.
[4] NEC Corporation

### 2.6 指示語

いくつかの指示語について、それが修飾する語がキーワードかどうかを調査した結果を以下に示す。

"この"(51/86),"その"(26/84)

### 2.7 その他の情報

その他キーワード抽出に有効であると判断された属性を以下に列挙する。

- 品詞　固有名詞 (202/285), 形容詞 (0/141)...
- 不要語　"進歩"(0/10),"影響"(0/20),"現在"(0/9) ...
- 題名中に出現する文節 (994/1457)
- 抄録の最初の文中に出現する文節 (864/1535)

## 3 処理系

我々は前節の調査に基づき、の 292 の規則を作成した。そのタイプ別の分類を表 4に示す。

表 4: 規則のタイプ別分類

| タイプ | 規則数 |
|---|---|
| 名詞＋"を"＋動詞 | 34 |
| 名詞＋"が"＋動詞 | 12 |
| 名詞＋"は"＋動詞 | 10 |
| 名詞＋"の"＋名詞 | 74 |
| 指示語 | 2 |
| 相言 | 4 |
| 否定 | 2 |
| 時制 | 60 |
| 格助詞相当語 | 1 |
| 不要語 | 76 |
| 品詞 | 13 |
| 複合語 | 1 |
| ノード、文の位置情報 | 3 |
| 文タイプ | 1 |
| 合計 | 292 |

これらの規則を用いて効率的な実行系を生成するために、文献 [4] の手法を用いた。この手法は、事例の集合から情報を得ることによって、規則系の最適化を図るものである。規則系の最適化は、２つの過程からなる。最初の過程では、規則を評価関数によって評価し、規則を組み合わせることによって、有効な組み合わせ規則を学習する。規則の組み合わせにより、１つだけでは、充分でない規則から、より有効な規則を生成することができる。２番目の過程では、このようにして得られた組み合わせ規則の集合、を実行時の効率性の観点から見て、効率的な順序で実行される順番を決定する。この手法によって、重み付けを主観的な観点からでなく、事例ベースによって正当化された客観的尺度から決定することができる。

表 5: 標本の内訳

| | 文数 | 文節数 | キーワードを含む文節 |
|---|---|---|---|
| 訓練標本 | 1535 | 12168 | 3860(31.7%) |
| 評価標本 | 717 | 5514 | 1675(30.4%) |

## 4 実験

前述の手法を用いて、人手で作成した規則 (293) を最適化することによって、1451 の組み合わせ規則が得られた。

得られた組合せ規則を評価標本に適用した結果を表 6に示す。

表 6: 実験結果

| 抽出結果 / 人手 | キーワード | キーワードでない | 合計 |
|---|---|---|---|
| キーワード | 1375 | 300 | 1675 |
| キーワードでない | 1199 | 2640 | 3839 |
| 合計 | 2574 | 2940 | 5514 |

実験の結果、キーワード自動抽出の再現率は 82.1%、適合率は 53.4%であった。

文の構造を反映した抽出による効果を調べるために、同一の語に着目して、それがどのように判断されたかを調査した。"システム" について調査した結果を表 7に示す。

表 7: "システム" に対するキーワード判定

| 抽出結果 / 人手 | キーワード | キーワードでない | 合計 |
|---|---|---|---|
| キーワード | 65 | 4 | 69 |
| キーワードでない | 18 | 5 | 23 |
| 合計 | 83 | 9 | 92 |

## 5 おわりに

構文情報を用いて、キーワード抽出を行なう手法を提案した。本手法によって、重要でないキーワードを除くことができる可能性を示した。今回の実験は、抄録中のすべての文を対象としたが、抄録中の文のすべてが文献中に述べられていることを述べている訳ではない。抄録中の文から、文献中で述べられていることを記述している文（主題文）を抽出することによって [5]、より高い精度でキーワード抽出を行なえる可能性がある。また、本手法を適用した結果、キーワードである確率が得られる。この確率を利用して、重み付きの検索システムを実現することも容易である。

## 参考文献


[1] Maruyama, H. et al., "Interactive Japanese Parser for Machine Translation," COLING'90, Vol. 2, pp. 257-262,1990.

[2] 木本, "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 D-I,Vol. J72-D-1, No. 8,pp.556-565,1991.

[3] 国立国語研究所, "分類語彙表," 秀英出版,1965.

[4] 野美山他, "事例ベースを用いた規則制御の最適化," 情報処理学会自然言語研究会,Vol. 3, pp.191-192, 1992.

[5] 梅田, "抄録からの主題文の自動抽出," 慶應義塾大学図書館・情報学科卒業論文,18801604,1991

[6] 関根, "抄録からのキーワード自動抽出の高度化," 慶應義塾大学図書館・情報学科卒業論文,18805236,1991