自 然 言 語 処 理 97−15
（1993. 9. 17）

複合語キーワードの自動抽出法

小川 泰嗣　望主雅子　別所 礼子
{ogawa,masako,ayako}@ipe.rdc.ricoh.co.jp
（株）リコー　情報通信研究所

本稿では、複合語キーワード抽出法を報告する。従来方式では、(1) 単語単位にキーワード性を判定し、キーワードと判定された単語の連続部分を複合語としていたため、複合語の範囲を正しく認定できなかった。(2) キーワード性を品詞に基づいて判定していたため、キーワードの抽出精度が低かった。これらの問題点を解決するため、(1) 抽出する複合語の記述に正規表現を導入し、この正規表現にマッチする単語の連続部分を複合語と判定する。(2) 品詞を補って単語の構文的・意味的性質を記述するキーワード素性を導入し、前後の単語の性質を考慮してキーワード性を評価する。その結果、複合語の範囲を正しく認定し、キーワード抽出精度を向上させることができた。

A Compound Keyword Assingment Method for Japanese Texts

Yasushi OGAWA, Masako MOCHINUSHI and Ayako BESSHO
Information & Communication R&D Center, RICOH Co., Ltd.

We report a new compound keyword assingment method for Japanese texts. Conventional methods judge whether an individual word to be suitable as a keyword based on its part of speech. Thus, they fail to determine the right boundaries of compound words, and extracted compounds contain many unsuitable words. To solve these problems, 1) we extract word sequences, which match a *regular expression* representing keyword specification, as compound keywords. 2) we introduce *keyword feature*, which describes word's more specific information necessary for keyword extraction, and evaluate compound keywords taking into account the proceeding/following words. Thus, our method extracts compound keywords with right boundaries much more precisely than convetional methods.

1 はじめに

キーワード自動抽出はテキストデータの管理に欠かせない技術として古くから研究されている [7][12]。

キーワード抽出法には、抽出するキーワードをあらかじめ限定する統制語方式と限定しない自由語方式がある。統制語方式では、統制語辞書（シソーラス）を作成しておき、登録テキストにシソーラスに含まれる語が存在するか否かを照合する。この方式は抽出処理は簡単であるが、シソーラスの作成・維持が問題である。自由語方式では、形態素解析などにより登録テキストを単語に分割した後、重要な単語をキーワードとして選択する。この方式は同一対象物に対するキーワードが統一されないという問題があるが、シソーラス（の管理）が不要であるため、自由語方式がより多く利用される傾向にある [9]。われわれも自由語方式を採用した。

自由語方式におけるキーワードの単位には、短単位語と長単位語（短単位語を構成要素とする複合語）がある。通常の検索システムでは、キーワード単位を不可分なものとして扱い、照合処理を行なう。したがって、長単位語をキーワード単位とする場合、表層文字列が部分的に一致し意味上も関連が強い単語も異なるものとして扱われるため、検索精度が落ちる[1]。一方、短単位語をキーワード単位とすれば、複合語の部分一致検索を容易に実現できる [6][18]。しかし、単純に短単位語を独立のキーワードとすると複合語の構成情報が失われるので検索精度が低くなるため、複合語の構成情報を用いることで高い精度を実現する検索システムをわれわれは提案している [13][14]。

本稿では、この検索システムのための、短単位語を処理単位とながら複合語を抽出する方法を提案する。（以下では、このような処理を『複合語キーワード抽出』と呼ぶ。また、明示しない限り『単語』は短単位語を意味する。）これまでにも複合語キーワード抽出法は提案されているが [4][18]、これらの方式では、単語ごとにキーワード性を判定し、キーワードと判定された単語の連続部分を複合語として生成するため、複合語の範囲が正しく判定できなかった。これはつぎのような理由による。(1) 複合語の構成単語がキーワードとしてふさわしくない場合でも、複合語としてはキーワードにふさわしい場合がある。しかし、単語ごとにキーワード性を判定するのでは、このような複合語をキーワードとして抽出することができない。(2) 各単語はキーワードにふさわしい場合でも、必ずしもその連続部分が 1 つの複合語を構成しない場合がある。しかし、連続部分を無条件に複合語とするのでは、適切な複合語キーワードを抽出することができない。

[1]シソーラスによって部分一致検索を実現することも可能だが、日本語には複合語が多いため、全ての同義語・類義語をシソーラスに登録することは非現実的である。

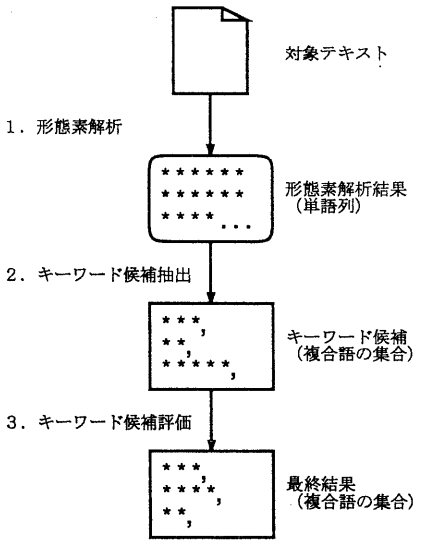

図 1: キーワード抽出の流れ

これらの問題点を解決するものとして、複合語キーワードを単語の文法的性質のパターンと捉え、テキストの解析結果からキーワードパターンに照合する単語の連続部分を抽出するパターンマッチング方式がある [5]。われわれはすでに簡単な遷移表を用いたアルゴリズムによるパターンマッチング方式 [2] が日本語に対して有効であることを検証している [13]。

本稿では、この方式を発展させ、キーワードパターンの記述に正規表現 [1] を用いて、より正確に記述できるようにした。本稿で提案するキーワード抽出法の処理フローは、これまでのものと同じである（図 1）。まず、入力テキストを形態素解析する。ここでは、単語辞書を使用し、単語に分割するとともにその品詞を決定する。つぎに、キーワードとなり得る複合語をキーワード候補として抽出する。最後に、各キーワード候補を評価し、不要な単語を削除して最終結果を生成する。以下、これらステップについて詳しく述べる。

2 形態素解析

キーワード抽出処理は単語を単位として行なわれる。したがって、キーワード抽出の第一ステップでは、入力テキストのべた書きの日本語を単語（形態素）に区切り、品詞判定する形態素解析を行なう。

2.1 アルゴリズム

形態素解析のアルゴリズムにコスト最小法を採用した[10]。すなわち、単語列のコストを単語列の構成単語のコストと隣接単語間の接続のコストの総和と定義し、候補単語列からコストを最小とするものを選択し、解析結果とする。

形態素解析の精度向上のためには、単語の接続の例外的現象や後続単語による構文的機能の変化に対応できなければならない。しかし、これらの現象を品詞のみで表現すると品詞数の増大を招き、単語辞書・接続表などの管理が面倒になる。そこで、本形態素解析処理系では、品詞からみた例外を『素性』で表現し、単語辞書・接続表の記述を簡単にした[8][11]。

2.2 品詞

キーワード抽出の視点からは、品詞は文法論で定義されるものとは必ずしも一致せず、独自の体系が要求される[12]。本形態素解析系で使用される品詞体系は、解析精度向上およびキーワード抽出以外のアプリケーションでの使用も想定して細分化されているが、その体系とキーワード抽出で要求されるものは必ずしも一致していない。そこで、品詞をグループ化し、後続処理にあったキーワード抽出用品詞にマッピングする（この操作を「品詞マッピング」と呼ぶ）。その結果、(1) データ作成者にとって、後述のキーワードパターンの記述が簡単になる、(2) キーワードパターンが簡潔になるため、キーワード候補抽出で使用するオートマトンが小型化され、処理を効率化できる。

品詞マッピングはつぎのような形式で記述する。

```
接辞名詞 付属名詞
修飾名詞 付属名詞
一般名詞
……
```

キーワード抽出品詞名を省略した場合、形態素解析品詞名がそのままキーワード抽出品詞名になる。

2.3 キーワード素性

前述の品詞マッピングにより品詞をキーワード抽出用に体系化し直しても、キーワード抽出には不十分である。それは、品詞は単語の形態素レベルの文法的性質を記述しているにすぎないため、同一品詞に属していても、単語によってキーワードとなり得るか否かは異なるためである。従来、不要語リストを設けることでこの問題に対処してきていた[7][12]。しかし、不要語リストは単語をキーワードに含めるか否かの2通りのわけ方しか想定してないが、実際には単語の構文レベルの性質や意味的な特徴に応じて、前後の単語を含めてキーワードとするかを判断しなければならない。そこで、われわれは不要語リストの代わりに『キーワード素性』を導入し、キーワード抽出の精度向上をはかっている[3][13][14]。

キーワード素性とは、品詞を補い、キーワード抽出の観点からのみた単語の構文的・意味的性質を記述するものである。キーワード素性は単語の性質に応じて用意され、その性質を有する単語それぞれに付与される。また、一つの単語に複数の異なるキーワード素性を付与することができる。

本システムでは、キーワード素性についても品詞と同様のマッピング処理を導入した。その結果、(1)（本来の目的通り）単語の性質に応じて素性付与すればよいので、データ作成作業が容易になる、(2) 形態素解析以降で同一処理を行なう素性が一つにまとめられるので、処理が効率化される。素性マッピングの記述法は品詞マッピングと同じである。

2.4 辞書管理

形態素解析辞書には、エディット可能なテキスト形式のソース辞書と処理実行時に参照される実行辞書の2種類がある。単語の登録・変更などはソース辞書で行ない、それをコンパイルして実行辞書を作成することで、単語の登録・変更を処理に反映させる。

キーワード素性が使用されるのは後述するキーワード候補評価の時点であるが、その際に単語のキーワード素性を知るために改めて辞書引きをしなくてもすむように、キーワード素性情報を品詞情報とともに形態素解析（実行）辞書に格納する。この場合、キーワード素性はソース辞書に記述しておかなければならないが、キーワード素性はアプリケーションに応じて設定されるものであり、頻繁に変更される。一方、形態素

解析の情報は変更は少ない。そこで、ソース辞書を形態素解析用とキーワード素性用の2つに分割し、キーワード素性を簡単に変更できるようにした。

3 キーワード候補抽出

形態素解析の結果得られる単語列から、キーワードパターンと照合する単語の連続部分をキーワード候補として抽出する。

3.1 キーワードパターンの記述

キーワードパターンは、キーワード候補として抽出する単語列を品詞で記述したものである。従来の単語単位のキーワード性の判定とは異なり、キーワードパターンによって単語列として複合語キーワードの持つべき性質を記述することができる。本システムでは、キーワードパターンの記述に正規表現[1]を使用した。これまでの遷移表による記述[13]と比較して正規表現は記述力が格段に大きいため、複合語の範囲を正確に認定することが可能となる。

各単語は <XX> のように記述する。ここで、XX は品詞マッピングしたキーワード抽出品詞名である。正規表現のためにつぎの記号を使用できる。

- *: 0回以上の繰り返し
- |: または
- (: 開き括弧
-): 閉じ括弧

例えば、<名詞><名詞>*<接尾辞> は名詞が1個以上連続し、末尾が接尾辞であるキーワードパターンを表す。

さらに、キーワードパターンの記述を簡潔化するため、マクロ機能も導入した。キーワードパターン定義においてマクロとして定義された識別子は、置換トークンで展開される。置換トークンには、単語記述あるいは識別子の正規表現を用いることができる。

3.2 アルゴリズム

つぎにキーワード候補の抽出アルゴリズムを説明する。テキストの形態素結果として得られる単語列の先頭から順にキーワードパターンと照合し、一致部分が

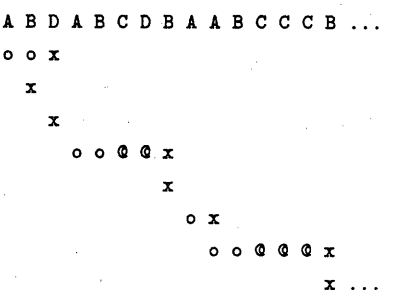

図2: キーワード候補抽出

発見された場合、その単語列を候補とする。照合には決定性有限状態オートマトン(DFA)[1]を使用する。なお、照合開始位置が同じで、キーワードパターンと照合する複数の単語列がある場合、最長のものをキーワード候補として抽出する。以下に処理手順を示す。

1. 照合開始位置 *start* をテキストの先頭にセットする。
2. キーワードパターンと *start* 以降の単語列をDFAで照合する。
3. 照合に失敗した場合、*start* を一つ進めステップ(2)に戻る。*start* を進めることができなければ、キーワード候補抽出を終了する。
4. 照合に成功した場合、最長の単語列をキーワード候補とする。
5. *start* をキーワード候補の末尾単語のつぎの位置に進めステップ(2)に戻る。

例えば、大文字 A, B, ... で単語を表すこととし、キーワードパターンを ABCC*D* とする。小文字 x, o, @ でDFAの初期状態、中間状態、最終状態を表すと、o ··· @ が照合に成功した単語列に相当する。このとき、単語列 ABDABCDBAABCCCB... は図2のように処理される。この例で抽出されるキーワード候補は ABCD と ABCCC の2つである。

3.3 括弧処理

テキストにおける括弧の意味はその種類に応じて異なるため、

1. 括弧に囲まれた部分をキーワード抽出の対象にするか、

2. 括弧の前後を連続するものとしてして処理するか

を、括弧の種類ごとに実行時に指定できるようにした。両者を組み合わせることも可能である。

例えば、括弧を含む単語列 ABC(DEFG)HIJ... はつぎのように処理される。

- 通常の処理
  ABC, HIJ...
- 括弧の前後を連続する場合の処理
  ABCHIJ...
- 括弧のなかも処理
  ABC, DBFG, HIJ...
- 括弧の前後を連続させ、括弧のなかも処理
  ABCHIJ..., DBFG

4 キーワード候補評価

キーワード候補として抽出された複数の単語列を評価し、キーワードに相応しくないものを削除する。

4.1 アルゴリズム

キーワードパターンの記述には品詞しか用いていないので、キーワード候補にはキーワードから排除すべき単語が多数含まれている。そこで、キーワード抽出の観点からのみた単語の構文的・意味的性質を記述するキーワード素性を用い、キーワード候補に含まれる不要な単語を、前後の単語の性質を考慮しつつ削除する[2]。以下に処理手順を示す。

1. 数詞・助数詞の処理（後述）

2. 接頭辞・接尾辞の処理

   従来は、接頭辞、接尾辞はキーワードにするか否かを品詞レベルで一律に判定しているものが多かった。しかし、これらは語の性質によってキーワードを構成しうるかどうかが異なる。そこで、後述するようにわれわれは接頭辞、接尾辞にもキーワード素性を付与し、キーワード候補の削除、選択を行なっている。

3. 名詞の処理

   サ変名詞および複合語語基を有する一般名詞・固有名詞はキーワード候補の構成単語数が 1 の場合に削除する。

4. 重複語の削除

4.2 数詞・助数詞の処理

数詞をキーワードとするか否かはその数詞の意味によって異なるが、その意味は後続の単語によって決定される。そこで、数詞の連続部分は 1 かたまりとして扱い、後続単語の性質に応じてキーワードとするか否かを判断する [3][14]。

後続語が助数詞の場合、数詞＋助数詞は数量を表しているため、その数量をキーワードに含めるか否かを「助数特殊」素性によって識別できるようにした[3]。すなわち、後続語が助数特殊を有する助数詞の場合、数詞と助数詞をともにキーワードに残し、後続語が素性なしの助数詞の場合、数詞および助数詞をキーワードから削除する。また、後続語が助数以外の場合、数詞は削除しない。

4.3 接頭辞の処理

識別力を増やす語にキーワード素性を付与し、キーワード素性の付与されていない接頭辞は削除する[4]。

- キーワード素性が付与されていない場合

  キーワード候補から削除する。

- 接頭特殊 1 を有する場合

  接頭特殊 1 は形容詞的な機能を持つ接頭辞である。直後に修飾関係にある名詞句が続くのでキーワード候補とする。

  遠：遠赤外→○

- 接頭特殊 2 を有する場合

  接頭特殊 2 は副詞的な機能を持つ接頭辞である。直後には相体言 (形容動詞、修飾名詞など) の機

[2]格・出現頻度・出現位置などの情報をもとに単語（複合語全体）のキーワード性を詳細に判定する方式は多数提案されている [9][17][16]。しかし、本方式ではこれらの情報を使用していない。

[3]キーワードに含める数量が何であるかは対象テキストの分野に依存するため、分野ごとにキーワード素性を用意する必要がある。以前の発表 [3][14] では、情報処理分野を対象に「情報処理分野助数」素性を用意したが、本システムではこの素性を「助数特殊」にマッピングして使用する。

[4]以前から接頭辞のために「修飾性接頭語」というキーワード素性を用意していたが、今回はこれを 3 つに細分化した。

能を持つ語が続き、両方が結合して形容詞句をつくる。形容詞句が名詞と結合している場合だけキーワードとする。つまり、後続単語が1単語の場合はキーワード候補から削除し、2単語以上の場合はキーワード候補とする。

超：超小型→×、
超小型プロセッサ→○

- 接頭特殊3を有する場合

接頭特殊3は指示的な機能を持つ接頭辞である。既出の単語を指示しているので、接頭辞をキーワード候補から削除する。また複合語語基（その分野でよく出現する名詞）が結合している場合は両方をキーワード候補から削除する。

同：同社横浜工場→横浜工場、
同横浜工場→横浜工場

4.4 接尾辞の処理

複合語の主要部になるので、多くはキーワード候補とするが、名詞句以外になる接尾辞には素性を付与し、異なる処理を行なう。

- キーワード素性を付与されていない場合

キーワード候補から削除しない。

- 接尾特殊1を有する場合

接尾特殊1は副詞句をつくる機能を持つ接尾辞である。1単語と結合して副詞句をなす場合は両方を削除するが、2単語と結合している場合は、接尾辞だけを削除する。副詞句に名詞句が続く場合は全体をキーワード候補とする。

上：製本ライン上→製本ライン、
ライン上→×、
ライン上事故→○

- 接尾特殊2を有する場合

接尾特殊2は形容動詞をつくる機能を持つ接尾辞である。後ろに修飾する名詞句が続く場合はキーワード候補とするが、名詞句が続かない場合はキーワード候補としない。

的：全国的→×、
全国的オンライン→○

- 接尾人名を有する場合

人名を表す接尾辞である。接尾辞はキーワード候補から削除する。

氏：米山氏→米山

5 評価

5.1 評価対象データ

われわれが提案したキーワード抽出法を200件の新聞記事を用いて評価した。評価では、記事をエレクトロニクス関係の50件とそれ以外の150件に分けた。エレクトロニクス関係を別扱いにしたのは、エレクトロニクス関係での利用を前提にキーワード素性を付与したので、その影響を確認するためである。例えば、助数特殊はエレクトロニクス関係の数量を落さないように設定した。

なお、記事の長さは平均で650文字（1.3 KB）である。

5.2 キーワード抽出精度

キーワード抽出の精度をつぎのように定義される再現率 R・適合率 P [15] によって評価した。

$$R = \frac{\text{抽出結果に含まれる正解キーワード数}}{\text{正解キーワード数}}$$

$$P = \frac{\text{抽出結果に含まれる正解キーワード数}}{\text{抽出キーワード数}}$$

ここで、正解キーワードとは人がその文書から抽出されるべきだと判断したキーワードのことである。再現率 R は正解キーワードのうちどれだけが実際に抽出されたか、適合率 P は抽出キーワードのうちどれだけが正解であったかを表すものである。

上記200件の記事ごとに2人が独立にキーワード付けをおこない、正解キーワードを設定した。この正解キーワードを用いて記事ごとに再現率・適合率を算出し、それらの平均値をキーワード抽出精度の指標に用いた。

キーワード素性を用いた候補評価の有効性を検証するため、候補評価の前後で再現率・適合率を比較した。その結果を表1に示す。ここで、向上率とは評価前後の再現率・適合率の変化の割合を示している。両者を比較すると、再現率はわずかに減少しているが、適

表 1: キーワード候補評価の有効性
（完全一致による再現率・適合率の比較）

		評価前	評価後	向上率
エレクトロニクス	R	0.72	0.70	-0.03
	P	0.16	0.24	+0.52
その他	R	0.70	0.63	-0.09
	P	0.15	0.23	+0.53
全体	R	0.70	0.64	-0.09
	P	0.15	0.23	+0.53

表 2: キーワード候補評価の有効性
（部分一致による再現率・適合率の比較）

		評価前	評価後	向上率
エレクトロニクス	R	0.97	0.94	-0.03
	P	0.38	0.55	+0.44
その他	R	0.96	0.92	-0.05
	P	0.38	0.55	+0.46
全体	R	0.96	0.92	-0.04
	P	0.38	0.55	+0.45

合率は大きく改善されていることがわかる。再現率が減少しているのは、キーワード候補評価によって削除される複合語のなかに正解キーワードが含まれていることを示しているが、向上率の絶対値が小さいことからその割合は非常に少ないことがわかる。一方、適合率が向上しているのは、キーワード候補のなかに含まれていた不要なキーワードが削除されていることを示している。すなわち、候補評価によってキーワードにふさわしくない複合語のみが効率的に除去されていることが確認できた。

つぎに、記事分野の影響を検討する。この表からわかるように、その他と比較してエレクトロニクスでは再現率・適合率とも数 % 高い。これは、キーワード素性をエレクトロニクスに合わせて調整したことがそのまま反映されているためと考えられる。

再現率 64 %・適合率 23 % という値はやや低いと考えられるかもしれない[5]。表 1 は、システムが抽出したキーワードと正解キーワードが完全に一致する場合に、両者が一致すると判断していた（完全一致）。しかし、われわれが開発している検索アルゴリズム [13][14] は部分一致検索を実現しているので、再現率・適合率の計算時のキーワードの一致を、システムが抽出したキーワードと正解キーワードに共通の構成単語があるか否かから判断する方がよい（部分一致）。部分一致による再現率・適合率を表 2 に示す。予想されたように、再現率・適合率ともに完全一致の場合よりも大きくなっている。評価前後の変化、記事分野の影響は完全一致の場合と同様の傾向が確認できる。

[5] ここでの実験では、キーワードの格・出現頻度・出現位置によるキーワード候補の評価のテクニックは採用していない。これらの技法を用いることで、再現率・適合率を向上させることはもちろん可能である。

表 3: 抽出キーワード数の比較

	評価前	評価後
複合語数	63.7	36.2
短単位語数	112.5	83.0
複合語当たりの短単位語数	1.77	2.29

5.3 抽出キーワード数

抽出キーワードの個数を記事ごとに評価した。複合語の個数および複合語の構成単語の個数を計測した結果をまとめたのが表 3 である。

ここでもキーワード候補評価を行なう前後を比較した。表 3 からわかるように、キーワード候補評価により抽出キーワード数は候補の 57 % に削減されている。前に示したデータで、再現率の減少がわずかであること（完全／部分一致のいづれの場合でも 3 % 減）を考えるとキーワード候補評価が有効に作用していることがわかる。

また、複合語の構成単語数も比較した。構成単語数は上昇しているが、これはキーワード候補評価によって識別性の低い構成単語数 1 のキーワード候補が削除されるためと考えられる。

6 おわりに

われわれは日本語テキストを対象とした複合語キーワード抽出法について報告した。本稿で提案したアルゴリズムは、(1) 形態素解析による単語の認定、(2) 品詞に基づくキーワード候補の抽出、(3) キーワード

素性に基づくキーワード候補の評価の 3 つステップから構成される。(2) のキーワード候補抽出では、抽出する複合語の記述に正規表現を導入し、複合語の範囲を正しく認定できるようにした。(3) のキーワード候補評価では、品詞を補って単語の構文的・意味的性質を記述するキーワード素性を導入、前後の単語の性質を考慮してキーワードに相応しくない単語を削除する。その結果、複合語キーワードを精度良く抽出することが可能となった。さらに、キーワード素性の管理を品詞などの管理とは独立にすることで、データ管理の簡素化を達成できた。

今後は、キーワード抽出精度とくに適合率の向上、キーワード素性付与の自動化などを行ないたい。

謝辞

形態素解析系の開発を行なった小松順子、小島裕一、伊藤篤、今郷詔さんに感謝いたします。

参考文献

[1] A. V. Aho, R. Sethi, and J. D. Ullman. *Compilers: Principles, Techniques, and Tools.* Addison-Wesley, 1986.

[2] P. Anwyl, M. Kanaya, and T. Morita. Automatic keyword assignment in English documents. 第 *41* 回情報処理学会全国大会予稿集, pp. 3- 87–88, 1990.

[3] 別所 礼子, 広瀬 雅子, 小川 泰嗣, 西村 美苗. テキストデータベースのためのキーワード抽出法. 第 *45* 回情報処理学会全国大会予稿集, 1992.

[4] 会森 清, 依田 透, 嵩原 哲. 日本語キーワード抽出システムの開発および今後の課題. ドクメンテーション・シンポジウム予稿集, pp. 15–19, 1988.

[5] 細野 公男, 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. 情報処理学会自然言語処理研究会, Vol. 39, No. 1, pp. 1–8, 1983.

[6] 今郷 詔, 望主 雅子. 短単位キーワードを用いた文書ファイリングシステム. 第 *43* 回情報処理学会全国大会予稿集, pp. 4- 215–216, 1991.

[7] 石川 徹也. 日本語テキストを対象とした自動索引システムの課題:総論. 情報の科学と技術, Vol. 42, No. 11, pp. 994–1002, 1992.

[8] 伊藤 篤, 望主 雅子, 小島 裕一. 日本語形態素解析における素性を用いた解析方式. 第 *43* 回情報処理学会全国大会予稿集, pp. 3- 113–114, 1991.

[9] 木本 晴夫. 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 *D-1*, Vol. J74-D-1, pp. 556–566, 1991.

[10] 小松 順子. コスト最小法に基づく逐次確定型・形態素解析. 第 *47* 回情報処理学会全国大会予稿集, 1993.

[11] 望主 雅子, 伊藤 篤. 日本語形態素解析における素性の導入. 第 *43* 回情報処理学会全国大会予稿集, pp. 3- 111–112, 1991.

[12] 諸橋 正幸. 自動索引付け研究の動向. 情報処理, Vol. 25, No. 9, pp. 918–925, 1984.

[13] Y. Ogawa, A. Bessho, and M. Hirose. Simple word strings as compound keywords: An indexing and ranking method for Japanese texts. In *Proc. of 16th Int. Conf. on Research and Development in Information Retrieval*, pp. 227–236, 1993.

[14] 小川 泰嗣, 別所 礼子, 岩崎 雅二郎, 西村 美苗, 広瀬 雅子. 短単位キーワードに基づくテキストデータベースシステム. 情報処理学会データベースシステム研究会, Vol. 70, No. 5, pp. 1–8, 1992.

[15] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[16] 妹尾 宏. 文字情報を主体としたデータ放送サービスにおけるインデキシングの検討. 情報の科学と技術, Vol. 42, No. 11, pp. 1023–1032, 1992.

[17] 内山 恵三, 中村 正規. 重要キーワードとその活用方法. 電子情報通信学会データ工学研究会資料, Vol. DE91-23, pp. 83–93, 1991.

[18] 山口 義一, 杉山 時之. 自然語による索引語自動抽出システムの概要とその索引語の分析. 科学技術文献サービス, No. 85, pp. 31–40, 1988.