

キーワード構成の分析とその応用

安藤 一秋, 李 泰憲, 獅々堀 正幹, 青江 順一

徳島大学工学部 知能情報工学科

{ando, aoe}@is.tokushima-u.ac.jp

Context に基づいたキーワード抽出の実現の第一歩として, 複合語生成規則とキーワードの構成概念を利用した複合語キーワード生成を試みる. まず, 人間が文書中の離れた文字列からもキーワードを生成すること, 代表的な単語を見るだけでキーワードとなる分野や上位語を認知できる能力をもつことに着目し, 原文中に出現しないキーワードの分析を行った. そして, 分析結果から, 複合語の係り受け規則を利用した生成規則と概念間の距離, キーワードの構成概念およびその強度を考慮した複合語キーワード生成規則を構築した.

Analysis of Japanese Compound Keywords and Its Application

Kazuaki Ando, Taehun Lee, Masami Shishibori, and Jun-ichi Aoe

Department of Information Science and Intelligent Systems, Faculty of Engineering,
Tokushima University

{ando, aoe}@is.tokushima-u.ac.jp

This paper proposes rules for generating keywords using productions for compound words and general concepts consisting of keywords in order to achieve a compound keyword extraction method based on Context. We focus on the following points to define the rules: 1) Human can easily create compound keywords from separated words in a text; 2) Humans can usually recognize words which means a superordinate concept and fields like natural language processing or information retrieval as keywords by finding specific words without reading the whole text. By analyzing a variety of Japanese compound keywords that do not appear in texts, this paper defines production rules describing semantic dependent relationships between separated words, concepts consisting of the keywords, strength and distance between the concepts.

1. はじめに

近年、ハードウェア技術の発展により、全文検索法を採用した DBMS (DataBase Management System) が開発されるようになった[1][2]。しかし、キーワードを索引にもつ DBMS も安定性、容量の面で現在も盛んに研究開発されている[2]。キーワードは、情報検索分野だけでなく自然言語処理、例えば自動抄録や要約の実現[3]にも有用である。また、キーワードは、論文などの見だし語として内容を把握する手段にも利用される[1][4]。

これまでに多種多様なキーワード抽出法が提案されたが、一般的なキーワード抽出技術としては、不要語辞書を用いてキーワードには適さない語を削除した後、残った候補に重要度を付与して重要度の高い順に抽出する方法が採用される[4-9]。しかし、この手法は、文意を踏まえた上でキーワードを抽出していないため、不適切なキーワード候補が多数抽出される[2][4][6][7][9]。それにより、キーワードのノイズが増加すると文献検索時の適合率が減少するため、得られた文献集合から更に必要なものだけを取り出す作業が必要になる[2][9]。したがって、文意を考慮したキーワード抽出法の考案が重要な課題となる。

その解決法として、キーワード抽出に言語情報を利用する研究がある。内山ら[6]は、重要キーワードが文書中の主語と目的語の節に多く含まれていることに着目し、これらの文節や括弧などで強調された文節からキーワード候補を抽出し、キーワード候補の前後にある格助詞情報などから重要度を決定する手法を提案した。意味的情報を用いる手法として、鈴木ら[5]は、結束チャートを用いて文書中の意味分類の出現傾向を大段落ごとに抽出し、それらの中からキーワード候補を抽出する方法を提案した。この手法は、段落単位で意味分類を類推し、キーワードを抽出するため主題を考慮したキーワードを抽出できる可能性をもつ。

これまでに提案された言語情報を用いるキーワード抽出法[5][6]でも、原文中に出現する単語のみを対象に抽出を行う。しかし、我々人間は、原文中に

出現しない語でも文意から判断して必要な語はキーワードとして付与する[2][7][9]。そのため、キーワード抽出の評価の際には、原文中に存在しない語を削除して[4]評価が行われることが多い。しかし、原文中に存在しない語もキーワードに含めて利用する方が、再現率を高めることが可能である[7]。

このような問題に対し、永田ら[9]は、文書中から概念を抽出し、あらかじめキーワードを構成する基本概念(キー概念)とキーワードの関係を記述してある索引辞書を用いて、原文中に出現しないキーワードを生成する方法を提案した。しかし、概念を取得する範囲に関して議論されておらず、不要なキーワードが生成される可能性をもつ。また、1 キーワードを構成するために必要なキー概念の定義が明確でないため、規則生成時に作成者の主観が入りやすく、自動構築も難しくなる。

本稿では、Context に基づいたキーワード抽出の実現の第一歩として、複合語生成規則とキーワードの構成概念を利用した複合語キーワード生成を試みる。まず、人間が文書中の離れた文字列からもキーワードを生成すること、代表的な単語を見るだけでキーワードとなる分野や上位語を認知できる能力をもつことに着目し、原文中に出現しないキーワードの分析を行った。そして、分析結果から、「曖昧性を解消する→曖昧性解消」のような複合語の係り受け規則[10]を利用した生成規則と概念間の距離、キーワードの構成概念およびその強度を考慮した複合語キーワード生成規則を構築した。構成概念は、キーワードを構成する各基本単語単位に定義されるため、生成規則の自動構築が可能になる。また、概念間の距離を導入することで、無用なキーワード生成を抑制できる。

2. キーワード構成の分析

2.1 対象データ

学術情報センターの情報検索システム評価用テストコレクション[11]から自然・音声言語処理に関する

表1 分析に用いたキーワードと抄録情報

データファイル数	65
キーワード情報	
平均キーワード数	4.0
最大キーワード数	6
最小キーワード数	2
キーワード総数	263
複合語キーワード数	217
基本単語キーワード数	46
平均構成単語数	2.11
最大構成単語数	6
最小構成単語数	1
抄録情報	
総文数	363
平均文数	5.6
最大文数	15
最小文数	2
総サイズ(KB)	43.2

データファイル[†](65 件)を抜粋し、キーワードの構成分析を行った。

表 1 に、対象ファイルに含まれるキーワードと抄録情報を示す。表中の平均、最大、最小キーワード数は、1データファイルに対する値であり、平均、最大、最小構成単語数は、各キーワードを構成する基本単語数を示す。また、総文数はタイトルも含んだ数である。学術情報センターに提出するデータシートには6 個以内のキーワードが字数と語数の制限なしに自由に記入できる。分析の結果、一般著者は平均4 個のキーワードを付与することがわかった。また、全体キーワードの約 83%が複合語キーワードであり、それらは、平均2 個の基本単語で構成されていた。人間は、有意なキーワードを構成するために複合語を使用することが検証できる。

次に、このキーワード集合を用いて、原文中に出現しないキーワードの分析を行った。ここで、「原文中に出現しないキーワード」とは、著者が付与したキーワードと完全一致する単語列が存在しないことを意味する。データファイルには、本文が含まれていないため、著者が付与したキーワードの中で、タイトルと抄録に出現しないものを対象にした。表2は、キ

[†] テストコレクションの各データファイルには、全国大会、研究会などで発表された論文のタイトル、著者名、抄録、主催学会名、著者が付与したキーワードのリストなどが含まれる。

表2 キーワードの出現分布

キーワード総数	263
タイトルと抄録に存在する数	160
タイトルだけに存在する数	12
複合語キーワード数	11
短単位キーワード数	1
抄録だけに存在する数	72
複合語キーワード数	56
短単位キーワード数	16
タイトルと抄録の両方に存在する数	76
複合語キーワード数	59
短単位キーワード数	17
タイトルと抄録の両方に存在しない数	103
複合語キーワード数	91
短単位語キーワード数	12

ーワードの出現状況を示す。総キーワードの内、約60%がタイトルまたは抄録中に存在した。特に、タイトルに存在したキーワードは、全体の約55%を占めることから、キーワードの尤度付けに利用される「タイトルに含まれる語は重要キーワードになりやすい」[6]というヒューリスティックスの正当性が検証できる。また、タイトルは、短い文字列の中に凝縮された内容が詰め込まれるため、複合語の使用率も高い。そのような本文の主題を凝縮した複合語がキーワードとして採用されやすい。

総キーワード中の残りの約40%は、タイトルと抄録の両方に出現しなかった。次節では、これらのキーワード、特に複合語キーワードに対して言語情報を踏まえた分析を行い、生成規則定義の準備とする。

2.2 キーワードの分析

本節では、タイトルと抄録の両方に出現しない複合語キーワードを、キーワードの構成情報が一文中に存在するもの、構成情報が複数の文に分離して存在するものおよび一部だけ存在するもの、そして、構成情報が原文中に全く存在しないものに分割して分析する。ここで、キーワードの構成情報とは、表層的な情報(表記や文法)を意味する。

2.2.1 キーワードの構成情報が一文中に存在

キーワードの構成情報、特に、キーワードを構成する基本単語が何らかの形で一文中に存在する場合

である。約 37%のキーワードがこの分類に該当する。

(A) 係り受け関係の語から複合語生成

このタイプの基本形としては、

「曖昧性を解消する→曖昧性解消」

「知識を獲得する→知識獲得」

などが挙げられ、複合語の自動分割に用いられる係り受け規則[10]で抽出可能である。

また、その変形として、

「音声の認識および合成→音声認識、音声合成」のように並列関係にある語から複数のキーワードが生成されるものや

「対話音声を合成する→音声合成」

のように複合語が構成される際に要素の一部が削除されるものも存在した。

(B) 複合語から複合語への短縮

・複合語の構成要素が短縮語へ変換される

「学習方法→学習法」

・複合語の構成要素が削除される

「比喩文生成システム→比喩生成」

・複合語と複合語の合成

「日本語処理+テキスト処理

→日本語テキスト処理」

(C) 複合語間に助詞を含む

単なる複合語として表現されるキーワードだけでなく、複合語と複合語の間に助詞が存在するキーワードも存在した。例えば、

「格要素の貢献度」

抄録中には、「貢献度の高い格要素」という表現が存在し、この表現から派生したものと考えられる。これは、人間が生成するキーワードの特徴的なパターンであり、単なる名詞だけで構成されるキーワードより、助詞を含むことで、更に意味を限定できると考えられる。

(D) 一般語から英語の短縮語への変換

このタイプに属するものとしては、

「文脈自由文法→CFG」

「隠れマルコフモデル→HMM」

「人工知能→AI」

などのキーワードがあった。

また、このパターンの逆、短縮語から一般語への変換により生成されるキーワードも存在した。

「HMM→隠れマルコフモデル」

研究目的に合致するタイプは、短縮語から一般語へ変換されるこちらのタイプである。

2. 2. 2 キーワードの構成情報が複数の文に分離、原文中に一部だけ存在

キーワードの構成情報が、複数の文に分離して存在するもの、タイトルおよび抄録中に構成情報の一部しか存在しないもの。約 34%のキーワードがこの分類に属す。

(E) 代名詞の意味を考慮

文脈を捉えることで、代名詞の指す語を考慮して複合語を生成するパターン。

例えば、

「... 言語... それを習得する→言語習得」

のように代名詞「それ」が指す語「言語」と結びついてキーワードとなる。

(F) 複合語の変形

このパターンは、複合語の構成要素の一方が同義語または上位下位語へ変換されるパターンである。

「知識取得→知識獲得」

(G) 分野名および上位語

キーワードの構成要素の一部と抄録全体または一部に存在する概念の組み合わせや文脈から連想される分野などが組み合わせられてキーワードになる。

「日本語意味解析、係り受け処理→日本語処理」

このパターンは、原文中に存在する複合語の上位概念を表す複合語がキーワードとなると考えられる。抽出には、概念レベルの規則や文脈意味解析を利用した推論などが必要となる。

2. 2. 3 構成要素が全く存在しない

キーワードの構成情報が全く出現しないものは約 29%存在した。これは、キーワードを構成する基本単語が全く出現しないことを意味する。

(H) 分野名および上位語

抄録全体または一部に存在する概念の組み合わせにより、連想される分野および上位語がキーワー

ドになる。このタイプには、

「ユーザインターフェイス」

「自然言語処理」

などのような分野名や上位語表現が多い。

これも(G)と同分類と考えられるが、(G)の場合、原文中に構成要素の表層情報が存在する。しかし、両方ともキーワードを抽出するためには、意味情報の利用が必須である。

(I) システム名や造語

提案するシステム名やその論文中だけにしか使用されない著者定義の造語。

以上の分析は、著者が付与したキーワードとタイトルおよび抄録との関係の分析である。データファイル中のキーワードは、著者が本論に対して付与したものであり、抄録に対して定義されたものではない。しかし、タイトルと抄録は、その文書の概要をとりまとめて、短く表現したり書き抜いたものである。したがって、多少、情報量の差異は生じるが、本論に対する分析と同等であると考えて問題ない。

3. キーワード生成規則

2章での分析結果に基づき、原文中に出現しないキーワードの生成規則を定義する。本研究の最終目的は、Context に基づいたキーワード抽出であるため、基本単語ではなく語の意味を具体的に表す複合語のみを扱う。したがって、分析に用いたキーワード集合内に含まれる基本単語に対する生成規則の定義は行わない。

(A)から(C)のパターンに対しては、複合語の生成規則として、宮崎の係り受け規則[10]が利用できる。(D)に関しては、短縮語辞書を用いることで実現できる。(E)に関しては、文脈解析など深い解析を必要とするため今後の課題とする。(F)に関しては、複合語生成規則の拡張およびシンソーラス辞書を利用することで抽出可能である。(G)と(H)に関しては、概念および距離を考慮した生成規則により抽出を行う。(I)に関しても今後の課題とし、今回は対象としない。規則の記述には、我々が提案した多属性情報照合エンジン[12]を拡張したものをを用いた。

係り受けの型	条件	例
AをBする	対象	原因究明
Aを対象にBする	対象	記者会見
AでBする	道具/手段/材料	自費出版
Aを目的にB	目的	改修工事
AするためのB	目的	発声器官
AのためにBする	原因/理由/結果	予防接種
AによりBする	原因/理由/結果	引責辞職
AすべきB	義務	作業量

図1 複合語の係り受け規則例

3.1 複合語の係り受けに基づく規則

人間は文書中の離れた文字列からもキーワードを生成する能力をもつ。この能力を機械的に実現するため、宮崎の一般語の係り受け規則[10]に基づき34種の複合キーワード生成規則を構築した。図1に係り受け規則の例を示す。また、係り受け規則と共に2章の分析により生成された17種の規則も同時に構築した。以下に規則例を示す。

規則例1 A+B+C+D → AC

A={ (品詞, 普通名詞) }

B={ (品詞, 格助詞「を」) }

C={ (品詞, サ変/語幹) }

D={ (サ変/活用形) }

例. 音声/を/合成/する→音声合成

規則例2 A+B+C+D+E → BD

A={ (品詞, サ変名詞) }

B={ (品詞, 普通名詞) }

C={ (品詞, 格助詞「を」) }

D={ (品詞, サ変/語幹) }

E={ (サ変/活用形) }

例. 対話/音声/を/合成/する→音声合成

3.2 概念を利用した規則

(G)や(H)のパターンのように、人間は文書中に出現する概念語を組み合わせることでキーワードを構成する。これを機械的に模倣するために、概念語からなるキーワード生成規則を構築する。

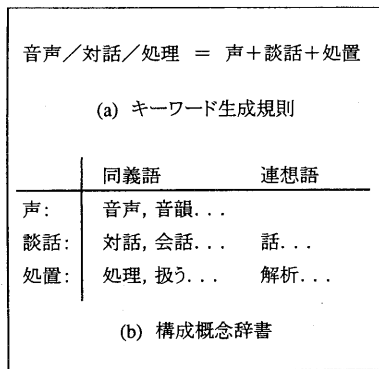


図2 キーワード生成規則と構成概念

永田ら[7]は、キーワードの構成要素となる基本概念の組み合わせを索引辞書として定義し、文書中から概念を抽出することでキーワードを生成する方法を提案した。しかし、1 キーワードを構成するために必要なキー概念の数が明確でないため、規則の生成時に作成者の主観が入りやすく、自動構築も難しい。また、概念の取得範囲なども議論されていない。

そこで我々は、複合語キーワードを構成する各基本単語に対して、それぞれ構成概念を与えることにした。つまり、複合語キーワード生成には、キーワードを構成する要素数と等しい構成概念が必要となる。これにより、複合語キーワード生成規則の自動構築が可能になる。また、構成概念を文書中から抽出するための手掛かりとして、構成概念と同一概念をもつ同義語と構成概念から連想できる語、連想語の情報を利用する。図2に概要図を示す。

更に、同義語と連想語の概念強度の違いを考慮し、同義語のみで生成されたキーワード(α)と同義語と連想語から生成されたキーワード(β)、連想語のみで生成されたキーワード(γ)の優先度を以下のように定義する。

$$\alpha > \beta > \gamma$$

β に関しては、同義語を含む割合も考慮し優先度を付与する。例えば、

$$\text{同義+連想+同義} > \text{同義+連想+連想}$$

なる優先度になる。

抽出されたキーワードの尤度付けには、上記の関係と概念間の文書上での距離を用いる。ここでの距

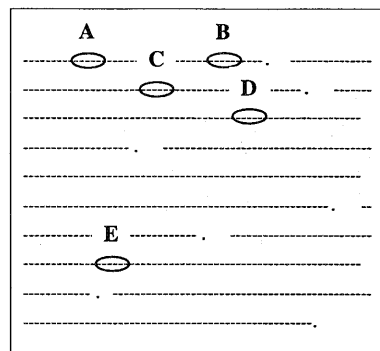


図3 概念間の距離

離とは概念を取得する範囲を意味する。距離の概念を導入することで、離れすぎた概念間の規則発火を抑制できる。図3の例では、ABの関係は距離0、ACの関係は距離1、ADの関係は距離2、AEの関係は距離5と定義される。例えば、距離0 > 距離1 距離2 > 距離5なる関係が存在する。

3.3 その他の構成規則

(D)に関しては、分析に用いた65ファイル中に存在する用語に対してのみ辞書を作成した。本研究の最終目的は、Contextに基づいたキーワードの抽出であるため、「文脈自由文法→CFG」なるタイプより、「CFG→文脈自由文法」を優先する。

4. 実験

4.1 実験データ

規則の抽出に利用した抄録65件を用いて、プロトタイプシステムの間接評価を行った。実験に使用した65件の抄録は、学術情報センターの情報検索システム評価用テストコレクション[11]から抜粋した自然・音声言語処理に関するデータファイルである。表3に実験データの詳細を示す。形態素解析には、本研究室で開発したコスト最小法に基づくエンジンを用いた。解析辞書は約25万語の語彙をもつ。

4.2 再現率と適合率による評価

プロトタイプシステムの評価は、規則の抽出に利用した抄録に付与されているキーワードに対する再現

表3 データファイル情報

データファイル数	65
総形態素数	12351
平均形態素数	190.0
タイトルの最小形態素数	5
タイトルの最大形態素数	22
タイトルの平均形態素数	12.1
抄録の最小形態素数	232
抄録の最大形態素数	63
抄録の平均形態素数	168.1
一文あたりの最小形態素数	5
一文あたりの最大形態素数	120
一文あたりの平均形態素数	34.3

率と適合率によって行った。以下に評価式を示す。

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

$$\text{適合率} = \frac{\text{抽出結果に含まれる正解キーワード数}}{\text{抽出されたキーワード数}}$$

ここで、正解キーワードとは著者が付与したキーワードの内、タイトルおよび抄録中出现しなかったキーワードの数である。また、再現率は、抽出漏れを、適合率はノイズの割合を示唆する指標である。

実験結果を表4に示す。表中の括弧内の数は、その規則によって抽出されたキーワード数を示す。再現率に関しては、多少低い値ではあるが68.1%とまずまずの値が得られた。抽出できなかった29のキーワードには、(I)のような原文中にキーワードの構成情報となる手掛かりが全く存在しないものが約30%を占め、(E)のように文脈理解が必要なもの、「学習方法→学習法」のように形態素の変形を要するものが多く存在した。また、構成情報の一部は存在するが抽出に高度な推論および知識を要するものが約48%存在した。本実験では、「自然言語処理」や「音声言語処理」などの大分類名や「形態素解析」や「ユーザインターフェイス」などの中分類名および頻出キーワードに対してのみ概念を利用した規則を構築した。約48%のキーワードは、頻出キーワードではなく、また、構成情報が一部しか存在しないため、意味ネットワークなどを用いて推論する必要がある。

適合率に関しては、かなり低い値になった。これは、1抄録あたりの平均抽出キーワード数に関しては

表4 実験結果

総キーワード数	91
抽出された正解キーワード数	62
係り受けに基づく規則	25 (439)
概念を利用した規則	31 (287)
その他の規則	6 (12)
抽出できなかった数	29
再現率(%)	68.1
適合率(%)	8.4

表5 抽出キーワードの妥当性判定

	A	B
係り受けに基づく規則	230	209
概念を利用した規則	239	48
その他の規則	12	0
妥当性(%)	65.2	

11.4と抽出数が抑制できているが、正解キーワードの数は1抄録あたり1.4個とかなり少ないこと、更に、本方式は、初期段階のプロトタイプバージョンであり、キーワードの尤度付けを行うための各種手法や情報(出現頻度、不要語)などを全く使用していないことが原因と考えられる。したがって、本方式に尤度付けを行う技術を導入することにより、適合率を向上させることが可能である。

4.3 抽出キーワードの妥当性に対する評価

再現率と適合率に対する評価だけでなく、システムにより抽出されたキーワードの妥当性も評価した。評価は、3人の被験者(博士後期課程院生)に抽出キーワードが抄録に対するキーワードとして成立するかどうかをA(キーワードに使用可能)B(不可能)ランクで判定してもらった。表5に、2人以上が指定したランクに集約した結果を示す。

判定の結果、約65%がキーワードとして適切であるという評価を得た。しかし、係り受けに基づく規則により抽出された約47%のキーワードがBにランク付けされた。Bランクに属する約60%は、複合語としても成立しない語、例えば、「仮定場合」や「人名関」であった。これは、適合率の考察でも述べたように、不要語に関係するものが大部分を占めた。また、規

則に品詞情報しか利用していないことも挙げられる。不要語辞書などを導入することで、妥当性の更なる向上が期待できる。

概念を利用した規則に関しては、約 83%が A ランクという結果を得た。更に、その内の約 95%が同一文内の構成概念によって生成されており、概念間の距離を尤度付けに利用する有効性が確認できた。また、同一文内の同義語から生成されたキーワードは、その内の 50%占め、同義語と連想語または連想語のみで生成されるキーワードより、キーワードとして成立しやすいことが確認できた。

5. まとめ

従来のキーワード抽出技術は、出現頻度や原文中に出現する語をキーワードとして採用するものが多いが、主題に関係ないキーワードを多く抽出するためノイズが増加する。そこで、本稿では、Context に基づいたキーワード抽出の実現の第一歩として、原文中に出現しない単語で構成されるキーワードの分析を行い、その生成規則を構築した。提案手法は、まだ初期段階であるが、実験により有効性を確認できた。また、提案手法は、原文中に存在しないキーワードを生成できるので、従来手法と融合させるだけでも、キーワード抽出の再現率を向上できると考えられる。

今後は、不要語辞書の作成や係り受けの条件および文末表現などを考慮した規則の拡張、重み付けによるキーワード尤度の決定法など考案し、章単位でキーワード抽出の範囲を拡大し、評価を行う予定である。

参考文献

- [1] 伊藤哲, 丹羽寿男, 萱嶋一弘, 丸野進, 木泰治, “利用目的に応じて最適可能なキーワード抽出手法”, 信学技法, NLC93-53, p.41-p.46, 1993.
- [2] 木本晴夫, “統合型大規模テキストデータベースへの自動索引とその評価”, 情報処理学会 DBS 研報, 90-9, pp.73-81, 1992.
- [3] 窪田健一, 山下浩一, 吉田敬一, “要約文生成の

ための単語抽出法”, 情報処理学会 NL 研報, 128-20, pp.143-150, 1998.

[4] 原正巳, 中島浩之, 木谷強, “テキストのフォーマットと単語の範囲内重要度を利用したキーワード抽出”, 情報処理学会論文誌, Vol.38, No.2, pp.299-pp.309, 1997.

[5] 鈴木斎, 増山繁, 内藤昭三, “語の意味分類の出現傾向を考慮したキーワード抽出の試み”, 情報処理学会 NL 研報, 98-10, pp.73-80, 1993.

[6] 内山恵三, 中村正規, “重要キーワード抽出方式とその活用方法”, 情報処理学会 DBS 研報, 84-19, pp.151-161, 1991.

[7] 永田昌明, 木本晴夫, “重要概念抽出に基づく新聞記事からのキーワード生成”, 第 37 回情報処理学会全国大会論文集, pp.1030-1031, 1988.

[8] 小川靖嗣, 持主雅子, 別所礼子, “複合語キーワードの自動抽出法”, 情報処理学会 NL 研報, 97-15, pp.103-110, 1993.

[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, Vol.J74-D-I, No.8, pp.556-566, 1991.

[10] 宮崎正弘, “係り受け解析を用いた複合語の自動分割”, 情報処理学会論文誌, Vol.25, No.6, pp.970-979, 1984.

[11] NACSIS Test Collection for IR Systems, 学術情報センター, 1999.

[12] 安藤一秋, 辻孝子, 獅々堀正幹, 青江順一, “日本語定型表現のパターン記述規則と効率的な照合アルゴリズム”, 電子情報通信学会論文誌, Vol.J80-DII, No.7, pp.1860-1869, 1997.