自 然 言 語 処 理 106－1
（1995．3．9）

単語共起と語の部分一致を利用したキーワード抽出法の検討

原 正巳　中島 浩之　木谷 強

NTT データ通信 (株)
技術開発本部 情報科学研究所

本方式は、記述項目と内容が定められている定型フォーマットのテキストを対象に、単語の共起情報や語の部分一致などの情報を利用して、内容把握のためのキーワードを抽出する手法である。提案する方式では、まず、キーワード抽出の対象とする項目をテキストから抜粋し、字種の接続関係を利用して、複合語の分割を極力避けながらわかち書きを行なう。次に、わかち書き結果から不要語を削除して得たキーワード候補に対して、出現頻度および他の語との共起情報、語の部分一致などの情報を利用して重要度の順位づけを行なう。本検討では、記述項目が統一されている特許明細書を処理対象として、プロトタイプを試作し評価を実施した。評価の結果、本手法により特に出現頻度の低い語に高い重要度を付与できることが明らかになった。また、十分高速なキーワード抽出が期待できることがわかった。

Keyword Extraction Using Word Co-occurrences and Partial Word Matching

Masami Hara, Hiroyuki Nakajima, Tsuyoshi Kitani

Research and Development Headquarters
NTT DATA COMMUNICATIONS SYSTEMS CORP.

This paper describes a method for extracting keywords from Japanese text in which fields of content and the text structure are predefined. The proposed method uses information about word co-occurrences and partial word matching to extract keywords which are used to help users quickly understand the text. The method first identifies fields to be processed in the text. Second, it divides the sentences into words referring to a table which defines whether or not a word boundary must be inserted between adjacent characters. The word separation is based on the transition of character kinds, which works to keep compound words unseparated often comprising of a few Kanji characters. Third, possible keywords are selected by removing ones inappropriate as keywords. Finally, they are ranked in the likely order based on the information about word frequency, word co-occurrences, and partial word matching obtained from the text. A prototype system is developed and evaluated to process patents whose fields of content and the structure are predefined. According to the evaluation results, this method is proved to be effective particularly in giving a high priority to important words appearing infrequently in the text. The results also prove that the system is expected to extract keywords fast enough to be used as a practical system.

1 はじめに

近年、CD-ROMやネットワーク上のテキストの増加に伴い、電子化されたテキストが大量に流通している。このような状況を背景として、全文を読むことなくテキストの内容を把握する必要性が高まっている。キーワードは、テキストの内容を簡潔に表現するものであり、テキストの内容把握に役立つ情報と考えられる。

本稿では、定型的なフォーマットを持つテキストを対象として、実用的な速度で高精度なキーワード抽出を行なう手法を提案する。また、特許明細書を対象に実施したプロトタイプの評価結果を報告する。

2 従来のキーワード抽出

キーワードの抽出手法は2つに大別できる。1つは語句の持つ概念や品詞、係り受け関係など言語の持つ情報に着目してキーワードを抽出する手法である。この流れの研究としては、シソーラスを利用して、テキスト中の文字列から共通する概念を抽出し、その概念を基にキーワードを生成する方法[1]や、出現する個々の語の意味分類から、テキストの主題となる意味分類を決定し、その分類情報を基にキーワードを抽出する方法[2]などがある。言語情報を利用したキーワード抽出は、字面にとらわれない、内容を踏まえた抽出が可能となるが、現状では意味や文脈まで考慮した解析は技術的に困難であり、大量のテキストを実用的な速度で処理することも難しい。さらに、意味の定義や概念辞書の作成など、実現に必要な環境が完全には整備されていないという問題があり、実用レベルに達していない。

言語の持つ特徴に着目する方法とは別に、テキストの持つ表層的な情報を利用する方法がある。この方法は、フォーマットや表現などのテキストの持つ特徴と、語の出現頻度や語の出現位置を利用してキーワードを抽出するものである。この立場を採る研究としては、テキスト中の単語と、それらの単語の組合せからなる複合語に関して、出現頻度を基にそれぞれ別の算出法で重要度を算出し、それらを併合して順序付けを行なう方法[3]や、キーワード候補とする単語列を品詞の並びであらかじめ記述し、テキスト中でそのパターンとマッチしたものをキーワードとする[4]手法などがある。また、分野に依存したテキストの表現傾向を利用する手法もある[5]。テキストの持つ特性を利用する方法は、テキストに現れない語をキーワードとして生成することは困難であるが、複雑な言語解析を必要としないため、言語情報を利用する手法に比べて高速にキーワード抽出を実施でき、これまでにいくつかのキーワード自動抽出パッケージも市販されている。しかしこの方法では、テキストの内容把握に利用できるほどの高い品質ではキーワードを抽出できない問題があった。本報告では、内容把握のためのキーワードを表層的な情報を用いて自動抽出する手法を提案する。

3 単語共起と最長語併合を利用したキーワード抽出

本手法は、章や段落ごとに特定の見出しが付与された定型フォーマットのテキストを対象として、テキストの持つ表層上の特性を利用するキーワード抽出である。テキストから抽出したキーワード候補に対して、単語の共起や語の含有関係、出現頻度情報などを利用して重要度を付与し、重要度の高い語をキーワードとする。

図1に本手法による処理フローを示す。まず、テキストからキーワード抽出の対象とする見出しを選択し、その見出しに属する文を抜粋する(処理1)。次に、文字種類の変化情報を利用して、選択した見出し内の文のわかち書きを行なう(処理2)。さらに、一般的な語を登録した不要語辞書と不要語の特徴を記述した不要語判定ルールを用いて、わかち書きした語から不要語を削除する(処理3)。最後に、残ったキーワード候補に対して単語の共起度および部分一致する単語を長い単語に併合する処理により重要度を付与し、上位からキーワードとして出力する(処理4)。次章では、それぞれの処理の詳細について述べる。

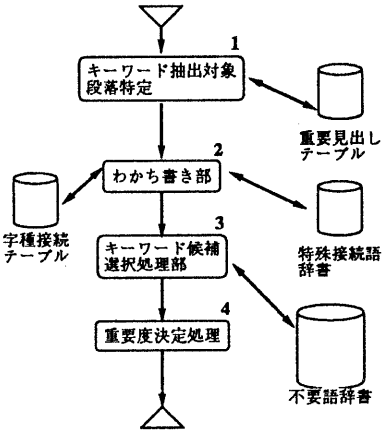

図1: 本手法によるキーワード抽出フロー

4 処理の特徴

4.1 わかち書き処理

従来のキーワード抽出の多くは、わかち書き処理として形態素解析辞書を用いた形態素解析を行なっているが、形態素解析の失敗による単語分割誤りは避けられない。また、わかち書きの結果が単語単位に分割されてしまい、内容把握に必要な複合語の獲得には不向きである。そこで本検討では、直前の字種と後続文字の字種の

接続可能性を基にわかち書きを行なうことで、語の分割を最小限に押え、複合語を得ることを試みた。表1は字種接続テーブルの一部を示しており、文字列は接続不可(×)の条件で分割される。

表 1: 字種接続テーブル

A＼B	漢字	平仮名	片仮名	記号	...
漢字	○	×	○	×	...
平仮名	×	○	×	×	...
片仮名	○	×	○	×	...
記号	×	×	○	×	...
:	:	:	:	:	:

○: 接続可、×: 接続不可

今回試作したプロトタイプでは、字種接続テーブルを参照する前に、漢字のみに書き換え可能な語は書き換えを行ない(例:「書き換え」→ 「書換」)、不可能な語(例:「ひん度」)は特殊接続語辞書を参照して単語分割の誤りを防ぐ。

4.2 キーワード候補の獲得

キーワード候補選択処理では、わかち書き処理で得られた語から、特許で一般的に使用される語や接辞語、記号など、キーワードに適さない文字列を削除してキーワード候補を得る。不要語削除のために、本検討では不要語辞書および不要語判定ルールを用いている。

不要語辞書は、特許明細書で一般的に使用される語を登録する辞書であり、無作為抽出した特許明細書約250件から得た707語を登録している。わかち書き結果のうち、不要語辞書内の語と一致する語は削除される。

一方、接辞語や記号などは不要語判定ルールを用いて削除している。不要語判定ルールは次のように定義している。

1. ひらがなや記号、数字のみで構成された語(“そして”、“3－5”など)
2. 1文字からなる漢字(“即”、“現”など)
3. 接頭表現や接尾表現、末尾の数字列(“当該装置”の当該、“発明図3”の“3”など)
4. 接頭表現や接尾表現、末尾の数字列を削除した後で不要語辞書に属する語や上記1,2,3に該当する語

不要語判定ルールにより、不要語辞書への登録語数を最小限に押えることが可能である。

4.3 重要度の決定

本手法では、キーワード候補に重要度を付与するために、**単語共起**と**最長語併合処理**を利用している。それぞれについて以下に説明する。

4.3.1 単語共起度

1. 単語共起度の定義

あるキーワード候補の単語共起度とは、そのキーワード候補が文や見出しなど特定の範囲内で他の特定のキーワード候補と共に出現する回数を、その範囲内のキーワード候補数で除し、それをテキスト全体で合計したものである。

例えば、「A は B のために、C を利用する方法である。」という文では、キーワード候補 A は B,C とそれぞれ1回づつ共起しており、キーワード候補語数が3であることから、候補 A の単語共起度は 1/3+1/3=2/3 となる。B,C についても同様に他のキーワード候補との単語共起度は 2/3 となる。

2. 単語共起度の意味

文や段落、章など特定の範囲が持つ意味は、範囲内の語単独ではなく、語の組合せによって表現されている。例えば、前述の例文ではキーワード候補 A,B,C の組合せが、この文の意味を示していると考えられる。

ある語の対が多くの範囲で共起するということは、その語の対に関連する内容がテキストで強調されていると考えられる。そこで本検討では、単語共起度を重要度付与パラメータの一つとして利用することとした。

単語共起度を利用すると、内容的に関連性の深い語を対で抽出できる可能性が高くなるという利点があるが、一方で、頻出語や不要語の重要度を同時に上げてしまう問題がある。今回の実験では、対象テキストである特許明細書に特化した不要語辞書を充実することで不要な語を可能な限り除去して、この問題に対処している。

4.3.2 最長語併合処理の利用

1. 最長語併合によるキーワード候補の決定

最長語併合とは、語 B が語 A 全体を含む語のうち最長の語である時、語 B を代表のキーワード候補とすることと定義する。例えば、「言語」「言語処理装置」「自然言語処理装置」というキーワード候補が存在する時、「言語」及び「言語処理装置」を、これらを含む最長語である「自然言語処理装置」に併合する。その結果、キーワード候補は「自然言語処理装置」一語になる。

2. 最長語併合の意味

テキストの内容を表すキーワードは語長が長いことが多い。これは、語長が長いほど語の意味が具体化されるためである。しかし、一般的に語長の長い語は出現頻度が低いことが多いため、単語共起度のみを利用して重要度を付与する方法は、重要語の出現頻度が少ない場合に高い重要度を付与できない。

そこで、キーワード候補にテキスト内で部分一致する語の出現頻度を加算して、キーワード候補の重要度を向上させることとした。キーワード候補に内包される語が多いということは、その候補が重要な語であり、その内

容に関する記述が多いと考えられるためである。

4.3.3 重要度決定

1. 単語共起度の利用

今回の検討では、単語共起度の適用範囲を文と見出しの2つとした。見出しはテキスト内で表題のついた最小単位である。見出しより広い範囲では、共起による意味的な関連性は少ないと考えたため採用しなかった。

キーワード候補Kの単語共起度を利用した重要度$I(K)$を式1に定義する。

$$I(K) = \alpha Freq(K) + \beta Cp(K) + \gamma Cs(K) \quad \cdots\cdots (1)$$

Freq(K): Kの出現頻度
Cp(K): Kの見出し内単語共起度
Cs(K): Kの文内単語共起度
α, β, γ: 定数 $(\alpha, \beta, \gamma \in (0, 9])$

2. 最長語併合の利用

単語共起度により付与された重要度に対して、最長語併合を利用して各キーワード候補の最終的な重要度を決定する。最長語K_{max}の単語共起度の重要度を$I(K_{max})$、K_{max}に最長語併合の対象となる語群K_{sub}の重要度を$I(K_{sub})$としたとき、最長語併合を利用したK_{max}の補正重要度$I'(K_{max})$は、次のように定義される。

$$I'(K_{max}) = I(K_{max}) + \tau \sum I(K_{sub}) \quad \cdots\cdots\cdots (2)$$

τは1未満の正数である。τをα, β, γよりも小さい値とする理由は、併合される語は語長が短く、キーワード候補より重要度が低いと考えられるためである。

上記1,2の重要度付与の結果、同一重要度の候補が出現した場合は、語長の長い候補を優先して抽出する。

5 実験・評価方法

5.1 評価基準

5.1.1 適合率・再現率

キーワード抽出の評価には、図2に示されるような適合率と再現率の2種類の基準を用いた評価が一般的である[6]。図2に示すように、適合率は抽出したキーワードが正解キーワードと一致する割合を示し、再現率は全正解キーワードに対する抽出した正解キーワードの割合を示す。それぞれ、出力のノイズの少なさと漏れの少なさを表す指標である。

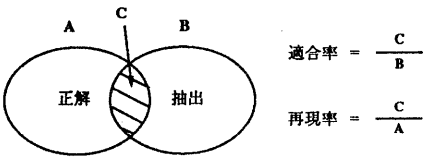

図 2: 適合率と再現率

適合率・再現率を求めるためには正解キーワードが必要である。正解キーワードはPATOLISキーワード(後述)を基に作成している。PATOLISキーワードには単語のキーワードが数多く含まれている。一方、本手法で抽出されるキーワードは、長めの語を優先して抽出しているため、両者が完全一致する確率は低い。そこで今回の評価では、完全一致の他に抽出キーワードがPATOLIS正解キーワードを含む場合も一致とみなして適合率・再現率を求めることとした。

5.1.2 必要抽出数

本評価では適合率・再現率とは別の評価尺度として、必要抽出数を定義した。必要抽出数とは、**再現率n割**$(n=3 \sim 8)$**を達成するためには何語のキーワードを抽出する必要があるかを表すものである。**同程度の正解が含まれている場合、抽出キーワード数が少ないほど有効な抽出法であるといえる。

なお、本評価では、キーワードの抽出数を最大にしても正解のn割に達しない特許明細書は評価対象から除外した。これは、PATOLISキーワードでは単語分割されたキーワードが含まれるために語数が多く、最長語併合で語数を絞り込む本手法ではn割の正解を得るのに十分なキーワード数が獲得できない場合があるためである。

5.2 実験用データ

5.2.1 実験用テキスト

実験用テキストには平成6年度の特許公開公報を用いた。特許データを実験用テキストに選んだ理由は以下の通りである。

- CD-ROMで公開されており、電子化データを容易に入手できる。
- 見出しの表題や記述内容が統一されている。
- 比較評価の基準となる正解キーワードを入手できる。

実験用テキストは、不要語辞書作成データや予備実験データとは異なる特許として、CD-ROMから抽出した計算・計数分野に属する特許150件を用いた。

5.2.2 正解キーワード

特許明細書のキーワードとして一般に利用されているものに、(財)日本特許情報機構(JAPIO)から提供されるキーワード(以降PATOLIS[1]キーワード)がある。本研究では、計算・計量分野の特許150件に対するPATOLISキーワードを修正・削除することで正解キーワードを作成した。

[1](財)日本特許情報機構が提供する特許情報オンライン検索システム

まず、PATOLIS キーワードの中には本文中に登場しない語が存在するが、本文中の語に置き替えが可能な語は、本文中の語で置き換えた。また、語の意味が失われるなどの分割誤りがあるもの (KWIC→KW, IC となっているもの等) の修正を行なった。

次に、文中でのみ意味を持つ記号 (図／式番号、変数)、いずれの特許明細書においても内容把握に重要とは考えられない語 (非重要語) を削除した。さらに、他の語の部分列となっている語についても、長い方の語があれば内容を把握できると判断し、削除した。削除した語数を表2に示す。

表 2: 削除語数

PATOLIS キーワード	5713 語
図、式番号、変数	227 語
非重要語	1339 語
部分列	754 語
分割の修正	7 語
正解キーワード	3386 語

6 予備実験

キーワードの抽出範囲となる見出しの選定、および重要度付与パラメータの重みを予備実験により決定した。

6.1 キーワード抽出範囲の指定

特許明細書には“【発明の名称】”、“【特許請求の範囲】”等の見出しが定められており、発明内容について見出しごとに異なる観点から記述される。観点の違いとキーワードの関係を確かめるため、それぞれの見出しにおける

$$\text{登場頻度} = \frac{\text{見出し内正解キーワード数}}{\text{見出し内キーワード候補数}} \quad \cdots\cdots (3)$$

を調査したところ、最大の見出しと最小の見出しでは4倍近い開きがあることが明らかになった。登場頻度の違いはキーワード抽出精度に影響を与えると予想されるため、適切な見出しを選んでキーワード抽出に利用する必要がある。そこで明細書の見出しのうち、登場頻度の高い10見出しを選び、その組合せによる適合率・再現率の差異を調査した。

評価値として適合率と再現率の積を用い、キーワード抽出数10,15,20,30語とした際の評価値を比較したところ、どの抽出数においても、登場頻度の高い見出しのみを用いる組合せにおいて最も良い精度が得られた。本実験では、その中でも最も良い精度を得られた組合せである【発明の名称】、【構成】、【目的】、【符号の説明】、【産業上の応用分野】の5見出しをキーワード抽出範囲として用いることとした。

6.2 重要度付与パラメータの決定

式1、式2における定数 α,β,γ,τ 決定のため、特許明細書50件を無作為抽出し、前記の予備実験で決定した見出しに対して、α,β,γ を区間 $[0,9]$ で独立に1刻みで、また τ を区間 $(0,1]$ で0.1刻みで変化させた時の適合率と再現率の積を調査した。

実験の結果、適合率と再現率の積を極大にする値

$$\alpha = 7, \beta = 1, \gamma = 3, \tau = 0.5$$

が得られた。これらの値を使って、本実験を実施した。

7 評価結果・考察

予備実験で得られたキーワード抽出対象とする見出しと、重要度付与パラメータを利用して、適合率・再現率と必要抽出数の実験・評価を実施した。

7.1 適合率・再現率

本手法 ($\alpha,\beta,\gamma,\tau \neq 0$) と、単語共起度のみ ($\tau = 0$)、最長語併合のみ ($\beta,\gamma = 0$)、出現頻度順 ($\beta,\gamma,\tau = 0$) の4手法における適合率・再現率のグラフを図7.1に示す。各プロットは、右下から左上に 10,15,20,30,40, 可能な限りの抽出語数に対する適合率と再現率を表す。

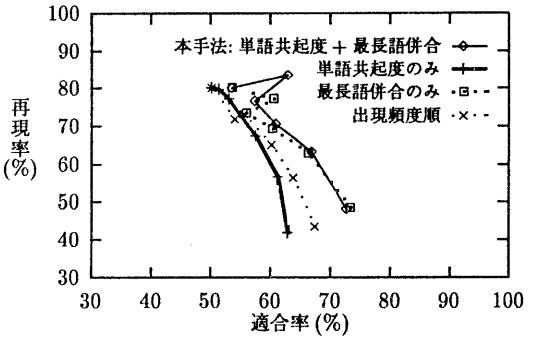

図 3: 適合率・再現率曲線

単語共起度のみと出現頻度順を比較すると、単語共起度のみを利用した場合、適合率には劣るが、再現率ではほぼ同等の精度を維持している。適合率の低下は、単語共起度の利用により、正解と重複して一致していた抽出キーワードが減少し、見かけ上正解との一致が減少したためである。単語共起度と出現頻度順の再現率を比較すると、抽出10語では42%対43%であるのに対して、抽出30語では77%対73%となり、抽出語数が増加すると、単語共起度と出現頻度順の再現率比が逆転する。この結果から、単語共起度は特に出現頻度の少ないキーワードについて、その重要度順位を挙げることに有効であることがわかる。

最長語併合のみと出現頻度順の比較では、最長語併合のみの抽出精度は出現頻度の抽出精度を大きく上回って

いる。最長語併合では長めの複合語がキーワードとなりやすいが、正解キーワードを含む場合も一致とみなす今回の評価では抽出キーワードが長く、正解と一致しやすいことが精度の高い原因と考えられる。最長語併合に単語共起度を併用することで、特に抽出語数が多い部分で適合率・再現率ともに若干の向上が見られる。

いずれの手法でも、最大限可能なキーワードを抽出した場合の再現率は80%に収束しているが、これは、以下の4つの原因による。

原因1. PATOLISキーワードでは、予備実験で決定した見出しに含まれない箇所からキーワードが抽出されることがある。

原因2. 最長語併合処理のため、PATOLISキーワードに比べて抽出数が少ない。

原因3. PATOLISのキーワードでは認めている漢字かな混じり語は、本検討の方法では抽出できない。

原因4. PATOLISキーワードで認めているひらがなのみからなる語を、本検討では不要語としている。

再現率を100%に近付けるために、原因1,2に関してはキーワードの抽出範囲を広げる必要があるが、予備実験からわかるように、抽出範囲の拡大は全体の精度低下につながる恐れがある。本手法を用いた上でこれ以上精度を向上するには、テキストの意味や内容に関する解析が必要となる。原因3に関しては、特殊接続語辞書を充実することが有効である。

7.2 必要抽出数

表3は、出現頻度順、単語共起度のみ、最長語併合のみ、本手法の各手法でn割の再現率を得るために必要とするキーワード抽出数を示している。

表3: 必要抽出数

n(割)	4	5	6	7	8
出現頻度順(語)	11	14	19	23	27
単語共起度のみ(語)	11	15	17	21	23
最長語併合のみ(語)	9	12	15	19	22
本手法(語)	**9**	**11**	**15**	**18**	**20**
抽出可能データ(件)	148	145	133	117	89

この表より、単語共起度のみと最長語併合のみの場合は、出現頻度順よりも少ない抽出数で同一の再現率を得られることがわかる。さらに、両手法を併用した本手法によれば、より少ない語数で高い再現性が得られる。少ない抽出数での再現率向上は、同時に適合率の向上を示しており、適合率・再現率の評価では明確ではなかった本手法の有効性を示している。

8 おわりに

単語共起と最長語併合を利用したキーワード抽出の有効性について述べた。単語共起と最長語併合が、単独では出現頻度の低いキーワードに高い重要度を付与し、適合率・再現率の向上に役立つことを確認した。

今回のプロトタイプは、SPARCStation10上で特許明細書1件(平均約21KB)を約4.7秒で処理する。このうち、わかち書きのためのかな漢字混じり語書き替え処理はUNIXのシェルスクリプトで記述しており、実行に約2.6秒を占めている。キーワード抽出処理部はCommon Lispで作成しているが、使用言語を変更することにより高速化が可能であり、大量のテキストを処理するのに十分高速な処理速度を得ることが期待できる。

今回の手法は、特に出現頻度の少ないキーワード候補の重要度順位を上げる効果があるが、キーワード抽出数が少ないと出現頻度の少ない語は抽出されにくくなり、精度向上に反映しにくいという問題がある。今後の課題として、出現頻度の多い語と少ない語で単語共起度の影響を変えて精度向上の可能性を探ることや、抽出対象見出しと、重要度決定のために用いた定数の分野依存性を調査する必要がある。また、係り受けの深さを単語共起度に反映させたり、単語共起度の代わりにシソーラス辞書内での距離によって重要度付与を行なう場合の効果も検討したい。さらに、特許明細書以外の定型フォーマットのテキストに対する適用可能性についても調査することが課題である。

謝辞

本検討の機会を与えた下さった、安部孝二情報科学研究所長に感謝いたします。また、PATOLISキーワード取得に協力をいただいた総務部知的財産室各位にお礼申し上げます。

参考文献

1) 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", 情報処理学会第37回全国大会,(1988).

2) 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", 情報処理学会自然言語処理研究会98-10,(1993).

3) 伊藤哲, 丹羽寿男, 萱嶋一弘, 丸野進, 〆木泰治: "利用目的に応じて最適化可能なキーワード抽出手法", 電子情報通信学会NLC93-53,(1993).

4) 小川泰嗣, 望主雅子, 別所礼子: "複合語キーワードの自動抽出法", 情報処理学会自然言語処理研究会97-15,(1993).

5) 木本晴夫: "キーワード自動抽出における分野特性の利用", 電子情報通信学会春季全国大会D-303,(1989).

6) 伊藤哲郎: "情報検索", ソフトウェア講座19, 昭晃堂,(1986).