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

原 正巳[†] 中島浩之[†] 木谷 強[†]

従来のキーワード抽出における単語の重要度を決定する手法は，頻度情報や位置情報など個々の単語に閉じた情報を利用していたため，高い抽出精度が得られなかった．本稿では特許明細書を対象に，テキストの表層情報を利用して実用的な処理速度を維持すると同時に，特定範囲内での単語の出現の有無を単語の重要度に反映させることで，キーワードを高精度で抽出する手法について述べる．まず，特許明細書に特有なフォーマット情報を利用してキーワードの抽出範囲を限定し，不要語の混入を回避した．次に，各抽出範囲ごとに出現する語のみに付与する重要度（範囲内重要度）を新規に導入し，抽出精度の向上を図った．また，テキストの内容を把握できるキーワードを獲得するために，文字列の包含関係に着目して，語の意味を具体的に表す語長の長い語を優先して抽出した．プロトタイプを作成し評価した結果，本手法が抽出キーワードの適合率と再現率の向上に有効であることを確認した．

Keyword Extraction Using a Text Format and Word Importance in a Specific Field

MASAMI HARA,[†] HIROYUKI NAKAJIMA[†] and TSUYOSHI KITANI[†]

Existing keyword extraction methods use only word-specific information such as word frequency and word location in a text in order to decide the importance of the keyword. Since they do not consider relationships among individual keywords, the extraction quality is not satisfactory to users. Our method proposed in this paper using Japanese patents also processes only surface information of the text to extract keywords. The simple mechanism performs keyword extraction fast enough to be used as a practical system. In spite of the simplicity of our method, a high quality of keywords can be obtained by choosing only a few crucial fields from entire patents and by considering word importance in a specific field in the text, based on a supposition that keywords should relate to each other in its context. To help users quickly understand the text with keywords, compound words including a few primitive words are chosen as keywords, since longer words usually have more concrete meaning than a primitive word. Moreover, the text is segmented by a simple algorithm for fast keyword extraction in our prototype system. According to the system evaluation, the proposed method has proved to be effective in improving both recall and precision of the extraction.

1. はじめに

近年，インターネットやパソコン通信などに代表される情報ネットワークの普及や，CD-ROM/MOなど大容量メディアの低価格化により，電子化されたテキストが大量に流通するようになってきている．それにともない，大量の情報から重要な情報のみを選択する必要が生じてきている．このような状況を背景として情報検索システムが注目されている．しかし，現状の検索システムは検索結果を十分に絞り込めず，膨大な検索結果から重要な情報を選択する作業はユーザの負担となっている．この負担を軽減するためには，全文を読むことなく内容を把握する手段が必要となる．

膨大な量のテキストを効率良く参照するための一手段として，キーワードの利用が考えられる．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有効である．キーワード抽出の研究は従来から広く行われており[1),2)]，実際に利用されているキーワード抽出システムもある[3)]．

本検討は特許明細書を対象とし，テキストの概要把握のために利用できるキーワードを抽出する試みである．特許明細書は，平成5年に公開公報が[4)]，平成6年には公告公報がCD-ROMで配布されるようになり，出願件数も年々増加の傾向にある．キーワードは大量の検索結果から必要な特許明細書を効率良く選択する

† NTTデータ通信株式会社 情報科学研究所
Laboratory for Information Technology, NTT DATA CORPORATION

ために役立つ．本稿では，特許明細書の定型フォーマット情報と，特定範囲での単語の出現の有無および文字列の包含関係を利用して語に重要度を付与し，複雑な解析をせずに実用的な速度で内容把握のためのキーワードを抽出する手法について報告する．また，プロトタイプの評価結果についても報告する．

2. 従来の手法

キーワード抽出の基本的な流れは，テキスト中の語から不要語を削除して，残された語に重要度を付与して，重要度の高い順にキーワードとするものである[1),2)]．不要語を削除した残りの語に対する重要度判定には，シソーラスの持つ語の意味分類やテキスト中の語の品詞情報や係り受け関係など，言語の持つ情報に着目する手法がある．また一方で，テキスト特有のフォーマットやテキスト内の語の出現頻度や出現位置など，表層的な情報を利用する手法が報告されている．

言語的な情報を利用する手法として，鈴木らは，シソーラスに基づいた意味分類によって段落間の関連を表す結束チャート[5)]を用いて，キーワードの重要度を決定する手法を提案している[6)]．まず，結束チャートを利用して，各段落の主題となる意味分類を推定し，その分類を基に各段落のキーワード候補を求める．そして，多くの段落に存在するキーワード候補をキーワードとしている．この手法は，段落の意味分類からテキストの意味分類を類推してキーワード抽出に利用することから，内容的に関連の深い語群がキーワードとして得られる可能性が高い．しかし，段落の意味分類を正確に行うためには，対象テキストに特化したシソーラス辞書を構築する必要がある．また，シソーラス辞書に登録されていない語が存在すると分類精度が低下する可能性がある．永田らは，従来のキーワード抽出法では主題を特定することができず，不要な語が多く抽出されたり，テキスト中に存在しない語は抽出できないことを指摘し，語の意味レベルに近い処理の導入を試みている[7)]．この報告では，抽出対象となるテキスト群のキーとなる概念を登録した辞書を利用して，テキスト中の文字列の組から概念を抽出し，その概念から索引規則辞書を利用してキーワードを生成する方法を提案している．新聞記事を対象とした実験での有効性が示されているが，実現にはキー概念辞書や索引辞書の作成が必要である．また，対象とする分野も限定される．

一般に，言語情報を利用するキーワード抽出は，字面にとらわれず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る．従来以上の精度を得るために要求されるハードウェアの能力も多大となり，大量のテキストを実用的な速度で処理することは難しい．さらに，意味の定義や概念辞書，シソーラス辞書の作成など，実現に必要とされる環境が完備されておらず，適用分野の大幅な限定が前提となるという問題がある．

表層的な情報を利用する手法では，伊藤らが，単語の頻度と複合語の知識を利用して，キーワードを抽出する手法を報告している[8)]．これは，テキスト中の名詞相当単語と，それらの単語の隣接語からなるすべての複合語のそれぞれに対し，出現頻度を利用して別々に重要度を付与し，それらを合わせて重要度順にキーワードとするものである．さらに伊藤らは，人間のキーワード抽出基準として出現頻度や語長など10種類の特徴量を仮定して，実際に人間が評価したキーワードを基に特徴量を学習することで，キーワードの抽出精度を向上できることを報告している．木本は，既存のキーワード抽出システム[9),10)]に対して，新聞記事を対象に記事の属する分野ごとの特性を利用することにより，キーワード抽出の精度向上を報告している[11)]．この報告では，テキストの論理展開の特徴を，テキストで重要な事項が記述されている位置と出現語の繰返し数という2つのパラメータでとらえ，テキストの属する分野の違いをそれらのパラメータに反映させてキーワードを抽出している．

表層的な情報を利用する手法は，内容あるいはその一部を的確に表す語はテキスト内部に必ず存在するという前提に基づいており，テキスト内に存在しない語は基本的に抽出できない．しかし，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を抽出できるという利点を持つ．この手法の多くは，キーワードの抽出箇所をテキスト全文として，抽出した単語の頻度情報を統計的に扱うことでキーワードの候補となる語に重要度を付与しているが，テキストの主題とは関連のない文から不要な語を抽出する問題や，テキストが長くなるにつれて処理に時間がかかるという問題があった．抽出範囲を限定する場合でも，テキストのフォーマット上の特徴を十分に利用してはおらず，文字数や段落数など経験に基づいて機械的に限定することが多かった．また，従来は頻度情報や位置情報など個々の語が持つ情報のみを利用してキーワード候補の重要度を決定しており，抽出語の関連性を考慮することが少なく，高い精度を得られないという問題があった．

本稿では，対象テキストを電子化データが容易に入手可能な特許明細書として，テキストの表層情報を

利用して実用的な処理速度を維持しつつ，高い精度でキーワードを抽出する手法を提案する[12]．第一に，テキストが持つ特有なフォーマットを利用して，キーワードの抽出に適した抽出範囲を実験的に決定し，処理対象を限定することを提案する．この手法により，テキスト全文を利用する従来の手法と比較して，抽出キーワードへの不要語の混入を低減でき，その結果抽出精度を向上することができるとともに，処理時間の削減にもつながる．また，抽出範囲への信頼性の点でも，文字数や段落数を利用する手法と比較して高くなる．第二に，特定の範囲内で出現する語にのみ付与する範囲重要度を定義する．記述内容の限定された範囲を処理対象とする範囲重要度の導入により，範囲内の語どうしの関連を考慮した重要度の付与が可能となる．

本検討では，キーワードをテキストの内容把握に利用するための手段と考え，文字列の包含関係を考慮して，基本単語ではなく語の意味を具体的に表す複合語を優先してキーワードとした．また，わかち書きの簡略化により，実用的な速度でのキーワードの抽出を試みた．本稿では，プロトタイプを作成・評価し，本手法の精度と速度の両面での有効性を示す．

本手法は，キーワードの抽出箇所をテキストの重要箇所に限定することで抽出精度の向上を図る点では，木本らの研究に近い．しかし，木本らの研究は新聞記事を対象に，テキストの重要位置を先頭からの文字数によって決定している．新聞記事はフォーマットが厳密には定められておらず，決定した重要度の信頼性にばらつきが生じる可能性がある．それに対して，本手法では一定のフォーマットを持つ特許明細書を対象に，テキスト内の見出しを利用して重要箇所を選択する．見出しの内容はテキストごとに一意であるため，重要な箇所を高い信頼性で判定することができる．また，木本らの手法では統制キーワード方式を利用しており，キーワード辞書を別途作成する必要があるが，本手法ではキーワード辞書を用意する必要はない．

3. 特許明細書からのキーワード抽出

本検討は，特許明細書のフォーマットおよび語の範囲内重要度と包含関係などを利用してキーワードを抽出する手法である．図1に本手法の処理フローを示す．

まず，特許明細書からキーワードの抽出対象を抜粋する（処理1）．この処理には，特許明細書に特有なフォーマット情報を利用する．次に，文字種の変化点に着目して，キーワード抽出対象内の文をわかち書きする（処理2）．次に，特許明細書で一般的に利用され，キーワードとなり得ない語を登録した不要語辞書と，不要語の持つ特徴を記述した不要語判定ルールとを用いて，わかち書きした語から不要語を削除する（処理3）．次に，残ったキーワード候補に対して，語の出現頻度および範囲内重要度を利用して重要度を付与し，さらに，文字列が包含関係にある語の組合せに対して，そのうちの最長の語に併合する処理によって重要度の補正を行う．そして，付与された重要度の上位からキーワードとして出力する（処理4）．次章および次々章で個々の処理の特徴を述べる．

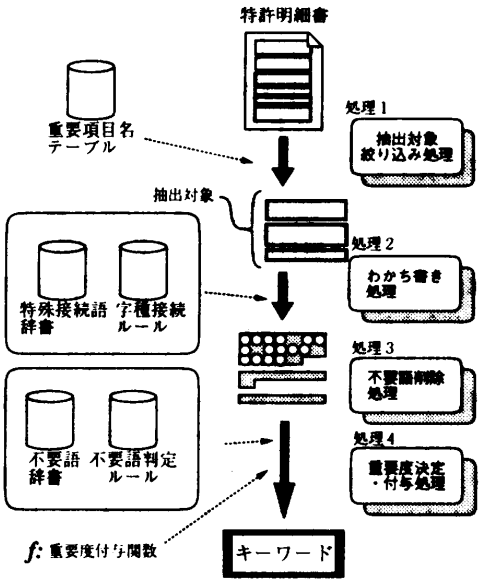

図1 キーワード抽出処理フロー
Fig. 1 Process flow of our keyword extraction.

4. 処理の特徴

4.1 抽出対象の絞り込み

本手法では，特許明細書全体をキーワードの抽出対象とはせず，明細書に特有なフォーマットを利用して抽出対象箇所の絞り込みを行っている．

特許明細書は，“【”と“】”とで囲まれた項目名により，内容ごとに項目に分けて記載されている．項目名には，たとえば【特許請求の範囲】，【従来の問題点】，【発明の解決すべき課題】，【実施例】などがある．項目名および項目名を見出しとする項目の記載内容は，発明について項目ごとに異なる観点から記述するよう特許庁により定められており，ほぼ一定の品質が保たれている．そこで今回は，キーワードの含まれる割合を項目ごとに実験的に調査し，キーワードの抽出精度の高い項目の組合せを選択してキーワード抽出対象項目としている．

キーワードの抽出範囲を限定することで，キーワードに不要な語が混入することを回避し，抽出精度を向

上させることができる．また，特許明細書の全文を処理対象とする場合に比べて，高速な処理が可能となる．

“【発明の名称】”，“【特許請求の範囲】”などの項目からキーワードを抽出する項目を選択するために，まず正解出現率を定義する．

$$\text{正解出現率} = \frac{\text{項目内正解キーワード数}}{\text{項目内キーワード候補数}^{\star}} \quad (1)$$

正解出現率が高いということは，項目内で正解キーワードが含まれる割合が高いことを意味しており，キーワード抽出項目として適しているということができる．次に考慮すべきことは，正解キーワードは，複数項目内で存在しているということである．キーワード抽出項目を選択する場合には，キーワードが各項目間で極力重複しないように選択することが望ましいと考えられる．

そこで，正解出現率を利用し，かつ重複しないようにキーワードを抽出できる項目を選択するために，以下に述べる方法を実行した．まず特定の特許において，正解出現率が最大となる項目を求めて重要項目とする．次に重要項目に含まれる正解キーワードを除いた残りの正解キーワードを利用して，重要項目以外の項目から正解出現率が最大となる項目を選択し，重要項目に加える．この処理を繰り返し行い，処理対象となる正解キーワードまたは項目がなくなった時点で処理を終了する．得られた重要項目の集合が，その特許のキーワード抽出対象項目となる．以下にアルゴリズムを示す．

```
while（A，C どちらかが空集合になるまで）
do begin
    A' := 正解出現率が最大となる項目 A_i (∈ A)
    W := W + A'
    A := A − A'
    C := C − (A'内でヒットした正解キーワード)
end
```

ただし，

$A = \{A_1, A_2, .., A_n\}$: 項目集合

C: 正解キーワード語集合

W: 重要項目集合

この方法によって得られた重要項目は，項目間で重複が少なく正解キーワードを含む重要な項目である．そこで，予備実験用データである特許明細書150件において，実際に出現した各項目がどの程度の割合で重要項目として選択されているかを求めて，項目重要度と定義した．項目 W_i の項目重要度を式(2)に示す．

$$W_i \text{の項目重要度} = \frac{\text{全特許での項目 } W_i \text{ の選択数}}{\text{全特許での項目 } W_i \text{ の出現数}} \quad (2)$$

項目重要度が高いということは，その項目が特許明細書に出現したときに，重要な項目である可能性が高いことを意味する．

4.2 わかち書き処理

抽出キーワードがテキストの内容を具体的に表すためには，複合語として抽出されることが望ましく，複合語単位でのわかち書きに関する研究が多く発表されている[13)~15)]．

特許明細書は，専門用語を始めとする複合語が多く出現し，かつ一文が通常のテキストと比較して非常に長いという特徴を持つ．この特徴は，従来の形態素解析では処理精度および速度の点で不利となる．そこで本検討では，稲垣ら[16)]の研究を参考にして，字種の並び順に着目して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った[12)]．これにより，複合語の不必要な分割をおさえることができる．

4.3 キーワード候補の獲得

本検討では，一般的な方法である不要語辞書を利用したキーワード候補からの不要語削除[1),2)]のほか，不要語判定ルールを作成して不要語を削除している．

不要語辞書には，無作為に抽出した特許明細書約250件から得た不要語707語を登録した．これらは，見出しや事例などで特許明細書に特有な利用をされる語（例：公開，公報，具体例，効果，ブロック図，など）や，一般的に利用されるがその語自体は特許明細書の論旨を支えるが，その内容を端的に特徴づけるものとはいえない語（例：可能，一致，一例，下記，解決，など）であり，キーワードとはなり得ないと考えられる語である．不要語辞書内の語と一致する語は，不要語としてわかち書き結果から削除する．

一方，接辞語や記号などは他の文字種と多様に接続するため，不要語辞書への登録が困難である．そこで，これらの語を削除するため，以下に示す不要語判定ルールを用いている．

[不要語判定ルール]

(1) ひらがな，記号，数字のみで構成された語
例：“そして”，“3 − 5”など

(2) 1文字からなる漢字
例：“即”，“等”など

(3) 接頭表現や接尾表現，末尾の数字列
例：“当該装置”の当該，“発明図3”の“3”など

☆ 項目内の文をわかち書きした結果から不要語を削除して，残された語の種類の数．

(4) 接頭表現や接尾表現，末尾の数字列を削除した後に，不要語辞書に属する語や上記ルール(1)，(2)，(3)に該当する語

このうちルール(3)については，特許明細書では利用される語が慣習的に限定されているために，250件の不要語登録用データからの登録語は100語程度に限定される．さらに不要語登録用データを増加しても，登録語が大きく増加することはないと考えられる．

不要語判定ルールを用いることで，不要語辞書への登録語数を最小限におさえ，辞書の複雑化や巨大化が回避できる．以下，不要語削除の結果残った語をキーワード候補と呼ぶこととする．

5. キーワード候補の重要度決定

本手法では，キーワード候補に重要度を付与するために，**範囲内重要度の付与**と**最長語併合による重要度の補正**を行っている．それぞれを以下に説明する．

5.1 範囲内重要度

5.1.1 範囲内重要度の定義

テキストの内容を踏まえたキーワード抽出を行うためには，出現する語どうしの関係を考慮する必要がある．本検討では語の関連性に関して，前章で述べた抽出範囲の限定を踏まえて以下の仮定をした．

仮定1. 特定範囲の多くで同時に出現する語どうしは関連性が高い

仮定2. 特定範囲に出現する語は，その範囲に出現する全語数が少ないほど重要性が高い．

以上の仮定から，範囲内における単語の重要度を式(3)に定義した．

段落や章など限定された範囲 A_i において，キーワード候補が N_i 種類存在するとき，範囲 A_i に存在するキーワード候補の範囲 A_i 内での重要度を $1/N_i$ と定める．このとき，キーワード抽出の対象となる範囲全体での重要度の平均を範囲内重要度と定義する．

M個の範囲 $A_1,..A_i,..,A_M$ 内にキーワード候補がそれぞれ $N_1,..,N_i..,N_M$ 種類存在するとき，あるキーワード候補 K の範囲内重要度 $C_r(K)$ を式(3)に定義する．

$$C_r(K) = \frac{1}{M}\sum_{i}^{M}\frac{\alpha_i}{N_i} \qquad (3)$$

ただし，

$$\alpha_i = \begin{cases} 1 & (A_i \text{ に K が存在するとき}) \\ 0 & (A_i \text{ に K が存在しないとき}) \end{cases}$$

である．

5.1.2 範囲内重要度の意味

段落，章など限定された範囲には，特定のテーマが存在する．各範囲におけるテーマは，その範囲内の語単独の意味の総和ではなく，語の意味が相互に関連することで表現されていると考えられる．本手法で提案する範囲内重要度をキーワード候補の重要度に反映させることで，特定範囲に存在する内容的に関連性の深い語の重要度を同時に向上させる効果が期待できる．

5.1.3 同一文内を1つの範囲と考えたときの例

例として，次の2文からなる文章を考える．

「キーワード抽出法には，構文や意味などの言語情報を利用するキーワード抽出法や言語情報を利用せずに出現頻度などの表層情報を利用するキーワード抽出法がある．本稿では，表層情報を利用したキーワード抽出法について述べる．」

この文章から，6種類のキーワード候補を得る．

第一文： キーワード抽出法，構文，意味，言語情報，出現頻度，表層情報

第二文： キーワード抽出法，表層情報

このとき，範囲内重要度の対象となる範囲の数は文の数(2)となり，また第一文での語の種類は6種類，第二文では2種類となる．したがって「キーワード抽出法」および「言語情報」の範囲内重要度は，式(3)からそれぞれ

キーワード抽出法：

$$\frac{1}{2}\cdot\left(\frac{1}{6}+\frac{1}{2}\right)=\frac{1}{3}$$

言語情報：

$$\frac{1}{2}\cdot\left(\frac{1}{6}+\frac{0}{2}\right)=\frac{1}{12}$$

となる．多くの箇所で出現している「キーワード抽出法」は内容に関連が深い可能性が高いと考えられるが，高い範囲内重要度を獲得しており直観と合致する．一方，1つの範囲内に多くの語があると，個々の語の重要性は分散されて低くなると考えられる．範囲内重要度の定義から，「キーワード抽出法」が個々の範囲で獲得した重要度は，対象とする範囲内の語が多いほど減少する傾向にあり，この点でも直観に合致する．

5.2 最長語併合処理の利用

5.2.1 最長語併合の定義

最長語併合とは，語bが語a全体の文字列を包含する語のうち最長の語であるとき，語bを両方の語の代表としてキーワード候補とすることである[8),17)]．

5.2.2 最長語併合の意味

語長が長い語はテキストの内容を表していることが多い[8)]．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く，特に特許明細書では，語の一部のみを利用してその語に関連した内容を表すことが多い．したがって，単純な出現頻度や範囲内重要度などの数値情報のみでは，重要語に高い重要度を付与できないという問題がある．

この問題に対処するために，キーワード候補どうしで文字列の包含関係をチェックし，他の候補を包含する語のうちで最長の語のみをキーワード候補とし，高い重要度を付与する方法[8),17)]を採用した．特定のキーワード候補が他の多くのキーワード候補を包含するということは，その候補に関連する内容の記述が多く，その語の表す内容の重要性が高いと考えられる．

5.2.3 最長語併合の例

最長語併合の例を以下に示す．

例1：「言語」「言語処理装置」「自然言語処理装置」
⇒「自然言語処理装置」

例2：「自然言語」「言語処理」「音声言語処理装置」
⇒「自然言語」「音声言語処理装置」

「自然言語」と「言語処理」のような部分的な一致は最長語併合の対象としない．

5.3 重要度決定

5.3.1 範囲内重要度の利用

本検討では，範囲内重要度の適用範囲を特許明細書のフォーマットとして定められている項目に限定した．項目より広い範囲では語の意味的な関連性は低いと考え，処理の対象としない．これは項目が特定の内容に即して記述される最大の単位であるという仮定による．

キーワード候補 K_i の範囲内重要度を利用した重要度 $I(K_i)$ を式 (4) のように，K_i の出現確率 $C_f(K_i)$ と範囲内重要度 $C_r(K_i)$ の和で定義する．

$$I(K_i) = \alpha C_f(K_i) + \beta C_r(K_i) \qquad (4)$$

$$C_f(K_i) = \frac{Freq(K_i)}{\sum_i Freq(K_i)}$$

ただし，

$Freq(K_i)$: K_i の出現頻度

$C_r(K_i)$: K_i の範囲内重要度

α, β: 定数 ($\alpha + \beta = 1.0;\ \alpha, \beta \geq 0$)

5.3.2 最長語併合の利用

最長語併合では，併合された語の重要度を最長語の重要度にどのように反映させるかが課題となる．出現頻度を単純に加算する方法[17)]も検討されているが，語長が異なるということは，語の意味の具体性に差異があることを意味しており，語長を考慮したウェイトづけが必要だと考えられる．本検討では，最長キーワード候補とそこに包含される語の重要度のウェイト比を語の長さの比と単純化して，最長語併合を行うこととした．

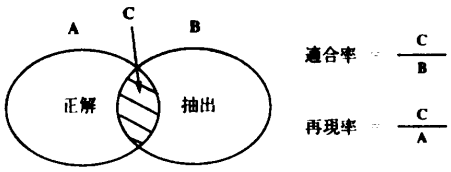

図2 適合率と再現率
Fig. 2 Precision and recall.

最長キーワード候補 $K_{\max}$ の語長を N，$K_{\max}$ の部分文字列であるキーワード候補 K_i の語長を M_i 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，式 (5) のように各単語が出現頻度の全体に占める割合 $C_f(K_i)$ と範囲内重要度 $C_r(K_i)$ の和と定義する．

$$I'(K_{\max}) = I(K_{\max}) + \sum_i \frac{M_i}{N} I(K_i) \qquad (5)$$

式 (5) で得られた重要度を用いてキーワード候補の重要度順位を決定する．ただし，同一重要度の候補が出現した場合は，語長の長い候補を上位に置くこととする．

6. 実験・評価方法

6.1 適合率・再現率

キーワード抽出の評価には，図2に示される適合率と再現率の2種類の基準が一般的に用いられる[1)]．

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のうち，抽出できた正解キーワードの割合を示す．それぞれ，抽出したキーワードのノイズの少なさと漏れの少なさを表す指標である．

適合率と再現率を求めるためには正解キーワードが必要となる．従来のキーワード抽出の評価では，正解キーワードは評価者あるいは非専門家によって付与されていることが多く，正解/非正解の判断基準が曖昧であった．今回の評価では，特許明細書に対して専門家が付与するPATOLISキーワード☆を基に作成した．PATOLISキーワードは，専門家が一定の作成基準に従って作成しているため，品質が安定していると考えられる．

PATOLISキーワードは，キーワード検索で利用するために作成されたキーワードである．一方，本手法のキーワードは特許明細書の内容を把握するための

☆ (財) 日本特許情報機構で付与されたキーワード．PATOLISは(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．

ものであり，極力単語分割をせずに抽出している．したがって，本手法のキーワード（平均語長 5.6 文字）は，PATOLIS キーワードから作成した正解キーワード（同 4.9 文字）に比べて語長の長い語の割合が大きい．この結果，両者のキーワードは完全には一致しにくい．今回の評価では，両者が完全一致する場合のほかに，抽出キーワード文字列が正解キーワード文字列を完全に含む場合も一致と見なして評価を行った．抽出キーワード文字列が正解キーワード文字列を完全に含む場合には，抽出キーワードは正解キーワードの意味をより詳細かつ具体化していると考えられる．本検討の目標は，テキストの内容把握のためのキーワードを抽出することである．そこで，抽出キーワード文字列が正解キーワード文字列よりも具体性を増す場合には，抽出キーワードの誤りや品質の低下とは意味しないと考え，語長の長い語を正解と見なすこととした．

なお，今回の評価では既存キーワード抽出手法との比較ではなく，提案する手法の有無による精度の変化について評価を実施している．また，長いキーワードを優先することが精度評価に影響しないように，比較した全手法に対して長いキーワードを優先する方法を採用している．このような評価を実施した場合，1 つの正解キーワードが複数の抽出キーワードと一致してしまい，特に適合率が過大に評価される問題点がある．そこで今回の評価では，適合率が過大に評価されることを避けるため，適合率の計算の際に同一の正解キーワードに n 語の抽出キーワードがヒットした場合，1 度だけカウントして残り $n-1$ 語は抽出キーワード数から除く補正を行って，適合率の過大評価を回避している．

以下に例を示す（下線が正解キーワードと一致した抽出キーワード）．

正解： 自然言語，機械翻訳

補正前： 自然言語処理，自然言語理解技術，
文字認識
→ 適合率 $2/3 = 66.6\%$，
再現率 $1/2 = 50.0\%$

補正後： 自然言語処理，自然言語理解技術，
文字認識
→ 適合率 $1/2 = 50.0\%$，
再現率 $1/2 = 50.0\%$

6.2 実験用データ

6.2.1 実験用テキスト

実験用テキストには平成 6 年度の公開公報を用いた．特許明細書には以下の特徴があるため，本実験データとして採用した．

表 1 実験データ
Table 1 Data used for the experiments.

特許明細書	件数	分野	公開年度
不要語登録用データ	250 件	全分野	平成 4 年度
予備実験用データ	150 件	計算・計数分野	平成 4 年度
実験用テキスト	150 件	計算・計数分野	平成 6 年度
合計	550 件		

- CD-ROM で公開されており，電子化データの入手が容易である．
- 記述項目や項目内の記述内容が統一されている．
- 評価用に正解キーワードが入手できる．

実験用テキストとして，特許明細書 150 件を CD-ROM から抽出して用いた．対象とする分野は計算・計数分野とした．実験用テキストおよび不要語登録用データ，予備実験用データはすべて異なる特許明細書を利用している．**表 1** に，利用した各データの内訳を示す．

6.2.2 正解キーワード

PATOLIS キーワードは品質が安定しているが，検索を主な目的として作成されたため，一般的な語や内容把握のキーワードとして利用するには不要なキーワードが含まれている．これらの語は，本検討が対象とする内容把握のためのキーワードの正解としては不適当であると考えられる．本検討では，評価対象となる計算・計量分野の特許 150 件の PATOLIS キーワードに対して，以下に述べる修正・削除を行い，正解キーワードを作成した．

まず，他の正解キーワードの部分文字列となっている正解キーワードについては，語長の長い正解キーワードがあれば内容を把握できると判断して削除した．この処理により，抽出キーワードが複数の類似した正解キーワードと一致することを回避できる．

次に，語の意味が失われるなどの分割誤り（例：KWIC→KW, IC）を修正した．また，個々の特許明細書中でのみ意味を持つ記号，図/式番号，変数（例：X，図 3，変数 B，など）を削除した．また，いずれの特許明細書においても内容把握には重要と考えられないキーワード [非重要語]（例：判定，設定，など）を削除した．この処理により，評価精度の信頼性を向上する．

さらに，PATOLIS キーワードには本文中に登場しない語（創出キーワード）が存在するが，本文中の語に置換えが可能な語は本文中の語で置き換え，逆に置換えが不可能な場合は削除した．この処理により，正解の同義語が抽出されているにもかかわらず，正解キーワードと一致しないという問題を回避する．

表2 修正・削除語数
Table 2 Number of modified words in PATOLIS keywords.

PATOLIS キーワード		5326 語
修正・削除内訳	部分文字列	994 語
	図，式番号，変数	111 語
	非重要語	1832 語
	書換え	32 語
修正後の正解キーワード (1 件平均)		2357 語 15.7 語

PATOLIS のキーワード数および修正・削除後の正解キーワード数を表2に示す．

7. 予備実験

キーワードの抽出範囲とする項目の選定および重要度付与式 (4) の係数 α, β を予備実験により決定した．利用したデータは，平成4年度の公開公報である．

7.1 キーワード抽出範囲の指定

4.1 節で述べた方法により各項目の項目重要度を求めたところ，表3を得た．この表の上位から順にキーワードの抽出範囲を拡大して，それぞれ適合率と再現率を求めたところ，【実施例】を含めると適合率の上昇と比較して再現率が大きく低下し，結果として全体の精度が低下することが分かった．そこで今回の実験では，項目重要度の上位7項目である【発明の名称】，【構成】，【目的】，【効果】，【産業上の利用分野】，【符号の説明】，【特許請求の範囲】の7項目をキーワード抽出範囲として用いることとした．

7.2 重要度付与係数の決定

重要度決定の式 (4) における定数 α, β を決定するため，特許明細書150件を無作為に選択し，前述の予備実験で決定した抽出項目7項目に対して，$\alpha+\beta=1.0$ となるように α を 0.1 刻みで変化させたときの適合率と再現率の変化を調査した．

実験の結果，

$$\alpha = 0.2, \quad \beta = 0.8$$

のとき，適合率–再現率曲線が最も高い精度をとることが分かった．そこで，本検討ではこれらの値を利用して，実験・評価を実施した☆．

8. 評価結果と考察

予備実験の結果に基づき，適合率・再現率による評価を実施した．

まず，範囲限定の効果を確認するために，抽出項目限定 + 最長語併合と，全項目抽出 + 最長語併合，全項目抽出 + 出現頻度順の3通りの方法で重要度を付与した際の適合率・再現率のグラフを図3に示す．比較データとして全項目抽出 + 出現頻度順をあげたのは，最も単純な順位づけである全項目抽出 + 出現頻度順との比較を通じて，本手法の有効性を明確にするためである．各プロットは，右下から左上に抽出語数を5から40まで5刻みに変化させたときの適合率と再現率を表す．ただし，必要なキーワード数を抽出できない明細書については評価データからは除外している．

図3から，抽出項目の限定がキーワード抽出の精度向上に有効に作用することが分かる．全項目抽出 + 出現頻度順と抽出項目限定 + 最長語併合とを比較すると，適合率と再現率ともに平均10%強の向上が見られる．

全項目抽出の場合に出現頻度順と最長語併合を比較すると，再現率と比較して，適合率の向上が著しい．適合率については，最長語併合は出現頻度順での重要度付与よりも語長の長い複合語を抽出しやすいため，正解キーワードを含む場合も一致と見なす今回の評価方法では，抽出キーワードと正解キーワードとが一致しやすくなったためだと考えられる．また最長語併合では，類似した語が上位に集中する傾向があり，その結果として適合率が向上している可能性もある．このことは，再現率の向上がほとんど見られないことからも予想できる．すなわち，最長語併合によってヒットした抽出キーワードの大部分は，同一正解キーワードを含むものであり，一方で併合する語がないために下位に押しやられた抽出キーワードが存在する．それらが再現率への効果を相殺している可能性があり，この結果再現率が向上しないと考えられる．

全項目抽出 + 最長語併合と抽出項目限定 + 最長語併合を比較すると，抽出項目限定は適合率・再現率ともに向上しており，抽出キーワードから不要語を適切に削除できているといえる．また，特に再現率の向上については，同一の正解キーワードを含みやすい最長語併合の抽出結果に対して，異なる正解キーワードが抽出できていることを意味しており，このことも抽出範囲限定の有効性を示している．

図4は，抽出項目限定 + 最長語併合に範囲内重要度を併用した精度を表す．15語抽出時の再現率を比較した場合，出現頻度順では45.5%であるのに対して，範囲内重要度を併用した場合には62.5%となり，1.37倍の数の正解キーワードを得ることができている．範囲内重要度を併用しない場合との比較でも，適合率と

☆ 本検討で抽出対象項目とした【符号の説明】は，特許明細書内に現れた名詞の列挙である．個々の語どうしには意味の関連性はないと考えられるため，範囲内重要度による重要度付与は行わない．

表3 項目名と項目重要度（%）
Table 3 Field name and their importance (%).

【発明の名称】	【目的】	【構成】	【効果】	【符号の説明】	【特許請求の範囲】	【利用分野】
100.0	97.2	90.0	83.3	69.7	59.3	54.4
【実施例】	【従来】	【作用】	【発明の効果】	【解決すべき課題】	【図面の説明】	【解決手段】
53.3	31.1	27.0	21.4	21.2	18.7	15.9

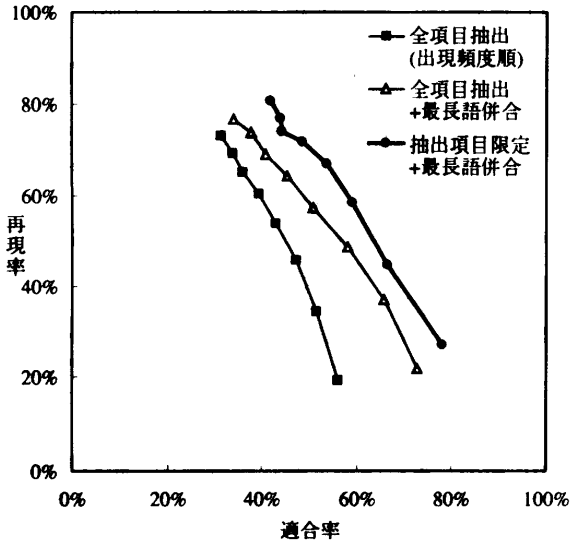

図3 キーワード抽出項目限定の効果
Fig. 3 Effect of selecting fields for keyword extraction.

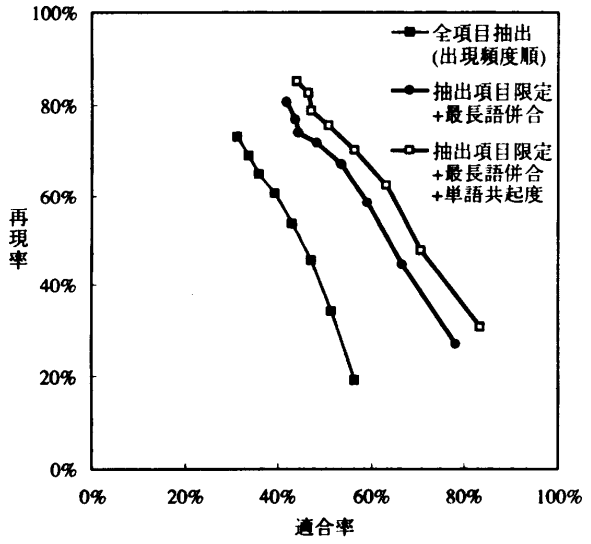

図4 範囲内重要度の効果
Fig. 4 Effect of using word importance in a specific field.

再現率が向上しており，範囲内重要度の有効性を示している．

今回の手法では，最大限可能なキーワードを抽出した場合でも約85%の再現率にとどまることを実験で確認している．これは，大きく2つの原因によると考えられる．

原因1. 正解キーワードには，予備実験で決定した項目とは異なる項目から抽出されたキーワードが存在する．

原因2. 正解キーワードのうち，本手法では不要語として削除された語が存在する．

本検討では，重要な箇所をあらかじめ限定して，そこからキーワードを抽出している．しかし，表3からも分かるように，選択した項目以外にもキーワードは存在している．限定した箇所から最大限可能なキーワードを抽出しても再現率が約85%にとどまったことは，【実施例】や【発明の効果】など，抽出項目の限定によって削除された項目にのみ含まれるキーワードが存在することを意味している（原因1）．これらのキーワードを抽出することは本手法では不可能であり，再現率向上のために，原因1に関してはキーワードの抽出項目を増やす必要がある．しかし，予備実験の結果によれば，抽出項目の不用意な拡大は正解キーワードの取得以上に，誤ったキーワード候補の混入を招き，全体的な精度の低下につながる恐れがある．ユーザが適合率と再現率のどちらを重視するかで，抽出項目を変更できるようにすることが好ましい．原因2に該当する例としては，たとえば“対応”や“書込”などの語があげられる．これらの語は，一般的に使用される語として本実験では不要語辞書に登録されているが，計算機や記憶媒体に関連する特許では正解キーワードとして採用されている．このように語の要・不要は，たとえ一般的な語であっても文脈によってその重要性が変化するという問題がある．この問題への対処として，単語のレベルで要・不要を判断するのは困難である語を不要語辞書に登録しないという方法も考えられるが，その場合多くの不要語がキーワード候補に混入するのは回避できない．この問題はキーワード抽出の研究で必ず直面する問題であり，今後の研究課題といえる．

処理速度に関しては，キーワード抽出項目のみを処理対象とすることで抽出の高速化を図っている．今回のプロトタイプをSUN SPARCStation20上で実行した結果，特許明細書1件（平均約21 KB）あたり平均約0.7秒で処理することができた．本プロトタイプではキーワード抽出処理部をC言語で記述しており，

大量のテキスト処理に十分高速な処理速度を実現している．

9. おわりに

従来のキーワード抽出法では，単語に重要度を付与するために頻度情報や位置情報など個々の単語に閉じた情報を利用しており，テキストのフォーマットや単語どうしの関連性を考慮していなかったため，高い抽出精度を得られなかった．本稿では特許明細書を対象に，テキストの表層情報を利用して実用的な処理速度を維持しつつ，高い精度でキーワードを抽出する手法を提案した．まず，特許明細書に特有なフォーマット情報を利用して，キーワードの抽出範囲を限定することで，キーワードへの不要語の混入を回避した．次に，キーワードどうしの関連を特定範囲で同時に出現するかどうかと関係すると仮定して，抽出精度の向上を図った．また，キーワードを利用してテキストの内容を把握できるように文字列の包含関係を考慮して，語の意味を具体的に表す語長の長い語を優先して抽出した．さらに，わかち書きを簡略化して全体の処理を高速化した．プロトタイプを作成し評価した結果，本手法が適合率と再現率の向上に有効であることを確認した．また実用上，十分高速なキーワード抽出が可能であることを示した．

本手法では，具体性が高いと考えられる語長の長い語を優先してキーワードとしており，テキストの内容把握のための利用を想定している．しかし，キーワードは検索用のキーとして利用されることが多い．また，近年，ソフトウェアによる全文検索技術も一般的になってきたが，テキスト量が増加するとインデックス量も膨大になるという問題がある．テキスト全文をインデックス化する代わりに，本手法で得られたキーワードをインデックス化すれば，全文を検索対象とする場合に比較して検索の再現率低下は避けられないが，インデックスサイズを小さくできると同時に，検索の適合率を高めることができるという報告がある[18]．

今回の評価では，特許の特定分野に特化した不要語辞書を用意しなかったが，実際には分野によって不要語に差異があることが予想される．今後の検討では，特許明細書の分野別の特性を不要語辞書や抽出対象項目，および重要度付与式に反映させた場合の，精度向上への影響を調査する．一方，単語の分割誤りを救済し，語句をより高い精度で認定するために簡易な形態素解析を行い，その結果から複合語を合成する手法も検討したい．さらに，SGML により構造化されたテキストも多くなっているため，特許明細書以外の定型フォーマットを持つテキストに対しても，本手法の適用可能性を検討していく．

謝辞 本研究の機会を与えて下さった，安部孝二前情報科学研究所長に感謝いたします．

参考文献

1) 伊藤哲郎：情報検索（ソフトウェア講座 19），昭晃堂 (1986).
2) 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).
3) 情報科学技術協会：特集＝日本語テキストを対象とした自動索引システム，情報の科学と技術，Vol.42, No.11 (1992).
4) データベース振興センター：特許公報における CD-ROM 公開公報について，データベース白書 1993，pp.227–228 (1993).
5) 佐々木一朗，増山 繁，内藤昭三：結束チャートの自動生成と日本語文章の語彙的結束構造解析への応用，電子情報通信学会言語理解とコミュニケーション研究会 (1993).
6) 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 (1993).
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) 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，電子情報通信学会言語理解とコミュニケーション研究会 (1993).
9) 木本晴夫，斉藤 雅：自然言語処理技術のデーターベースへの利用，国際シンポジウム “Computer World '90” (1990).
10) 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-1, pp.556–566 (1991).
11) 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 (1989).
12) 原 正巳，中島浩之，木谷 強：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，情報処理学会自然言語処理研究会 (1995).
13) Ogawa, Y., Bessho, A. and Hirose, M.: Simple Word Stings as Compound Keywords: An Indexing and Ranking Method for Japanese Texts, *Proc. 16th Int. Conf. on Research and Development in Information Retrieval*, pp.227–236 (1993).
14) 別所礼子，広瀬雅子，小川泰嗣，西村早苗：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論文集，pp.4-219–4-220 (1992).
15) 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 (1993).
16) 稲垣博人，小橋史彦，中川 透：簡易文章構造解析による文構造の決定，第 42 回情報処理学会全国大会論文集，pp.3-104-3-105 (1991).
17) 水野 聡，島田静雄，中牟田純，近藤邦雄，佐藤 尚：日本語キーワードの自動抽出手法，情報処理学会自然言語処理研究会 (1992).
18) 木谷 強，高木 徹，木原 誠，関根道隆：フルテキストと抽出キーワードを利用した情報検索，情報処理学会情報学基礎研究会 (1996).
19) 原 正巳：出現度数と分野情報を利用したキーワード抽出法の検討，第 43 回情報処理学会全国大会論文集，pp.3-185-3-186 (1991).

(平成 8 年 5 月 30 日受付)
(平成 8 年 11 月 7 日採録)

原 正巳（正会員）

1989 年東京工業大学理学部情報科学科卒業．同年 NTT データ通信（株）入社．キーワード抽出，抄録自動作成，文書分類など自然言語処理技術の研究開発に従事．現在，同社情報科学研究所に勤務．言語処理学会会員．

中島 浩之

1992 年東京工業大学工学部情報工学科卒業．1994 年同大学院理工学研究科情報工学専攻修了．同年 NTT データ通信（株）入社．現在，同社情報科学研究所で自然言語処理技術の研究開発に従事．人工知能学会会員．

木谷 強（正会員）

1960 年生．1983 年慶應義塾大学工学部電気工学科卒業．同年日本電信電話公社入社．1991～1993 年カーネギーメロン大学 Center for Machine Translation 研究員．形態素解析，情報抽出，情報検索などの自然言語処理の研究に従事．現在，NTT データ通信（株）情報科学研究所主任技師．工学博士．言語処理学会，ACM 各会員．