自然言語処理 120－13
（1997. 7. 25）

文字種切り出しと複合語分解によるキーワード抽出

沖電気工業 研究開発本部 関西総合研究所

下畑 光夫、杉尾 俊之

simohata@kansai.oki.co.jp, sugio@kansai.oki.co.jp

あらまし

テキストからキーワードを抽出する場合、最初に形態素解析を施して語を分割することが多い。しかし、形態素解析では単語辞書が必要である上、未知語の抽出精度が低くなってしまう。

本論文では、文字種によりテキストを分割して語を抽出し、次に抽出語を利用して語を分解してキーワードを獲得する方法について述べる。本手法は、文字種を基にしているため単語辞書を必要としないという特長がある。また、検索の時には検索キーワードにも同一の分解処理を施すため、検索範囲が複合語か単位語かを問わず拡大される。

キーワード　情報検索，キーワード抽出，文字種類，複合語

Keyword extraction using character type and decomposition of compound word

Kansai Lab., R & D Group, Oki Electric Industry Co.,Ltd.

Mitsuo Shimohata, Toshiyuki Sugio

Abstract

Although morphological analysis is frequently used for extracting keywords, morphological analysis needs words dictionary and does not work well when unknown word exists.

In this paper, we propose the method of extracting keywords, by dividing text based on character type and decompositing compound words using extracted words.This method has advantage of unnecessity for words dictionary. Because search keywords are also decomposited same as extracted keywords, searching can be extented to elementary word in compound word.

key words　information retrieval , keyword extraction , character type , conpound word

1 はじめに

最近では、インターネットに代表されるように大規模なデータが容易に蓄積できるようになってきている。所望のデータの検索に役立てるため、個々のデータに検索用のキーワードを付与することが多いが、大量のデータに人手でキーワードを付与することは困難であり、キーワードの自動抽出に対する期待が大きい。

キーワード抽出では、テキストに形態素解析を適用し、語に分割することが一般的である。[1][2][3][4] 形態素解析では単語辞書が必要であるが、対象テキストに適した辞書を作成することは容易ではない。また、内容が更新されるデータベースでは未知語の出現は避けられないが、未知語が出現した場合、形態素解析の精度が低下してしまう。検索においては新出語が検索キーワードとなることが多いことを考えると未知語の対処が重要となってくる。

森の研究 [5] や中渡瀬の研究 [6] では、N グラム統計を用い、字面処理でキーワードを獲得している。N グラム統計は単語辞書を必要としないが、キーワード獲得に必要な計算量が大きい。また、山本の研究 [7] では、単語辞書ではなく形態素区切りされたタグ付コーパスを用いて、文字のトライグラムにおける区切りパターンの確率から文字区切りを行っている。この方式でも、学習コーパスに出現しなかった語については抽出精度が低下してしまう。

本論文では検索キーワードを抽出することを主眼とし、文字種を利用して語を抽出し、さらに単位語に分割する方法について述べる。基本的に、ひらがなを区切り文字としてテキストを分割するため、計算量が少なくてよい。また、複合語の分解も抽出語を用いるため、キーワードの抽出・分解を通じて単語辞書を必要としない。検索時には、検索キーワードを同様に分解して使用するため、検索キーワードが複合語内で使われている場合や、逆に検索キーワードが複合語である場合でもヒットすることができる。

2 文字種による文字列切り出し

検索システムの多くは、検索キーワードを and や or で結合することでクエリーを与えているが、検索キーワードとして入力される語にひらがなはあまり使われない。ひらがなは、形容詞、動詞の活用部分や助詞、接続詞として用いられることが多く、名詞の一部 (または全体) として出現することは少ない。一方、検索キーワードには名詞が多く使われる上、他の品詞の言葉を入力する際も不確定要素である活用部分までを含めて入力されることは少なく、こういった相反する性質が原因であると考えられる。そこで、ひらがなはキーワードを切り出す区切り文字として有効であると考えられる。本手法では、文字をキーワード文字と非キーワード文字の 2 つに大別する。漢字、カタカナ、英字、数字はキーワード文字と、ひらがなは非キーワード文字とする。なお、記号は適宜キーワード文字、非キーワード文字に選別する。“-”や“&”や“%”などをキーワード文字としている。

全体の抽出手順を図 1に示す。まず、テキストをキーワード文字から構成される文字列と非キーワード文字から構成される文字列に分割する。次に、分割された文字列を 2 文字以上のキーワード文字列、1 文字のキーワード文字列、非キーワード文字列の 3 つの場合に分けてキーワード抽出を行なう。2 文字以上のキーワード文字列には、接頭語・接尾語が接続した語や、助詞が省略されて結合された語が多く存在するために、それらの語を基本的な語 (単位語) へ分解する。1 文字のキーワード文字列は分解する必要はないが、複数文字列と性質が異なるため別の方式で抽出を行なう。また、非キーワード文字でもキーワードとして重要な語も当然存在するため、別方式で抽出する。以下の節で各 3 通りの処理について述べる。

なお、本論文では、BMIR-J1 [1] [8] に用意

[1]株式会社 日本経済新聞の協力によって、社団法人情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが、1993 年 9 月 1 日から 12 月 31 日の日本経済新聞記事を基に構築した情報検索評価用データベース（テスト版）を利用

されている600記事の本文を用いて実験・評価を行っている。タイトルと記事にあらかじめ付与されているキーワードは処理対象から除外している。

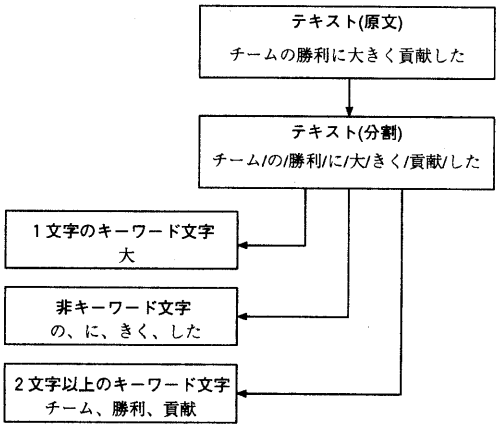

図 1: テキスト分割と処理の分岐

2.1 複数文字語の分解処理

テキスト中から切り出した2文字以上のキーワード文字列には、多くの複合語が存在する。複合語の生起規則には以下の2種類がある。

1. 単位語に接頭語または接尾語が付属した語 (例：新＋開発、経済＋力)
2. 単位語が複数連結して構成される語 (例：産業＋構造、情報＋検索)

ここでいう接頭語・接尾語とは漢字1文字から成り、単位語に接続することで単位語を修飾する語と定義する。図2に接頭語・接尾語の一例を示す。2種類の複合パターンを同時に含む語も存在する。(例：日本＋的＋経営)

複合語は、単位語に接頭・接尾語または他の単位語が付属した形をとるため、そのままの形では元の単位語の情報が複合語内に埋没してしまうという問題が生じる。そのため、検索キーワードと付与キーワードのマッチング処理を行なう際に、部分一致を用いないと検索漏れが生じる。また、抽出語について統計的処理を施す場合でも、ある単位語に関連する統計量 (頻度など) が複数の複合語に分散して

接頭語
現、初、同、前、全、再、他、両、各
接尾語
及、用、社、間、氏、化、用、力、内型、製、初、向、的

図 2: 接頭語・接尾語の例

しまい、統計量の質が低下してしまう。通常は、単位語辞書をあらかじめ用意し、複合語を分解することが多い。森脇の研究 [9] では、長い複合語どうしの共通文字列や差分文字列を用いて単位語を生成している。

本手法では文字種を利用して語を抽出しているが、抽出された語の中に単位語となるべき語が既に存在することが多い。そこで、既抽出語を単位語として利用する。(ただし1文字語は除く) したがって、複合語分解に用いる単位語はテキストから生成されることになり、単位語辞書をあらかじめ用意する必要がない。

既抽出語それぞれについて以下の複合語の分解手順を行ない、複合語の分解を行なう。

1. $init = 1$
2. 調べる語を $W(1,n)$(n は語長) とする。$k = 1$
3. $init = 0$ かつ $W(1,n)$ が抽出語として存在していれば true を返す。
4. $W(1,n-k)$ が抽出語として存在しているかチェック。抽出語であれば $init = 0$, 文字列 $W(n-k+1,n)$ を調べる語として2へ。
5. $W(1,n-k)$ の先頭文字が接頭語または、末尾文字が接尾語と一致するならば、$init = 0$, 接頭語・接尾語を除いた文字列を調べる語として2へ。
6. k を1増加。k が $n-1$ ならば false を返す。そうでなければ3へ。

上記の手順により、処理対象となる語が既抽出語の結合語であれば、その語は単位語に分解される。一部でも未抽出部分がある場合は、分解しない。したがって、「日本経済」という語は、「日本」と「経済」という2語が既

抽出語として存在していれば、この2語に分解されるが、どちらかの語が未抽出語であれば、分解されずに保持される。BMIR-J1では、22698個の複数文字から成るキーワード文字列が抽出されたが、そのうち8348個の文字列が複合語とされ、分解された。

なお、この複合語分解では、カタカナまたは英文字だけからなる文字列の途中での分解は行なっていない。複合語から単位語への分解処理は、分解された単位語を重ねた意味と元の複合語の意味がほとんど同一である場合に効果があるが、カタカナや英文字から成る語では分解された語の意味を重ねた以上の意味を持つことが多く、複合語分解処理が適切に作用しないためである。したがってソフトウェア ⇒ ソフト + ウェアやforget ⇒ for + getといった分解は、各単位語が既抽出語として存在していたとしても行なわない。

2.2 1文字のキーワード文字列処理

1文字語は分解する必要がないために前節のような分解処理は不要であるが、特有の性質を持つため、キーワードとして有効な1文字語を選別する必要がある。キーワード文字1文字の前後にひらがなが接続する文字列は以下の2つの場合が考えられる。

1. 形容詞・形容動詞や動詞の語幹と活用語尾 (例: 大きい、走る)
2. 漢字、英字などの1文字の名詞に助詞が接続したもの (例: 鍵が、国の)

1に該当する語は活用するために表記に揺れが生じる。また、このような語は検索の中心概念となることが少ないと考えられるため、検索キーワードとしては有用ではないと考えられる。しかし、名詞はキーワードとして重要な役割を果たすことが十分考えられるため、2に該当する語は、キーワードとして抽出しなければならない。1と2の場合の判別規則として、キーワード文字に後接するひらがな文字列を用いる。後接するひらがな文字列が名詞に接続する助詞から始まる場合は、名詞であると判定し抽出する。また、ひらがなでなく記号(句読点や括弧)で終っている場合も名詞と判定する。図3に名詞の判定に用いた助詞を示す。

格助詞	が, の, を, に, へ, と, より, から, で
係助詞	は, も, こそ
副助詞	や, か, さえ, でも, しか, まで, など

図 3: 名詞の判定に用いた助詞

BMIR-J1に1文字語抽出を行った結果と人による名詞判定との適合率・再現率を図4に示す。評価は総抽出数ではなく総抽出種類数で行なった。したがって、表記上等しい文字列は頻度に関わらず1つとして扱っている。実験では、全種類の漢字について評価した場合と、検索に有用でない漢字を除いて評価した場合について行なった。検索に有用でないとは、「私」「仮」などの代名詞や接続詞に用いられる漢字や、「化」「経」などの1文字で名詞になる可能性が少ない漢字を指す。結果を見ると、有効な1文字漢字を選択することで再現率が大幅に向上している。

	全漢字	選択漢字
総抽出種類数	704	1802
自動抽出数	1340	959
正解数	1190	1039
正解自動抽出数	971	855
適合率	81.6%	82.3%
再現率	72.5%	89.2%

図 4: 1文字語の抽出精度 (全漢字 / 選択漢字)

2.3 非キーワード文字列の処理

非キーワード文字は区切り文字として使用されるため、抽出されるキーワードに非キーワード文字は含まれない。しかし、非キーワード文字を含む語をキーワードとしたい場合も十分考えられる。そのような場合に対処するため、非キーワード文字から成るキーワードをあらかじめ登録し、登録した語がテキスト中に現れた時にキーワードとして抽出している。

この処理では、表記上一致する文字列が存在するかどうかで判定しているため、偶然に

表記が一致するために抽出される可能性がある。例えば、「のみ」というキーワードを登録した場合に、名詞としての「のみ」だけではなく、副助詞としての「のみ」も抽出してしまう。そのような場合には、不適切なキーワード付与を行なうという問題が生じる。非キーワード文字列部分について、Nグラム統計によるキーワード抽出の適用することで自動抽出も可能であろう。

3 検索処理

2.1で述べたようにテキストから抽出した複合語は単位語に分割されるため、検索キーワードに単位語が入力された場合に、キーワードのマッチングを完全一致にしておいてもデータを収集することができる。

逆に検索キーワードに複合語が入力されることも考えられるため、検索キーワードにも複合語分割処理を施している。単位語を分割するアルゴリズムはテキストデータの分割方法と同一である。検索キーワードを単位語に分解し、単位語をandで結合したものを新たなクエリーとして検索する。(図5) したがって、検索キーワードが複合語であっても分解し、
データ中で独立した語または別の複合語を構成する単位語となっているデータでもヒットすることができる。

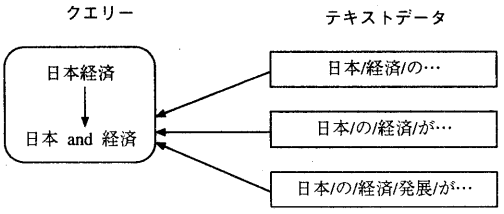

図5: 検索語とテキストデータのマッチング

4 評価

本手法によるキーワード抽出・検索をBMIR-J1に用意されている検索例と比較した実験を行なった。実験では、事前には単位語辞書を保持せず、抽出語のみを用いて複合語の分解を行なった。クエリーとして、一つの単位語を与えた。図6に検索キーワードと適合率・再現率を示す。図中、分子は抽出正解数を、分母は抽出数(再現率では正解数)を表している。また、図7に各検索キーワードを含む複合語の中で検索キーワードに分解された複合語を示す。図中の複合語において、"－"の箇所で分割されている。

検索要求は検索キーワードをデータに含んでいれば正解であるため、適合率はどの場合も100%であった。しかし、検索キーワードが複合語内しか現れず、かつ分解できなかった場合があるため、再現率は100%に達しない場合があった。例えば、「所得税減税」という複合語は「所得税」が独立語として存在しなかったために分解されず、3件のデータが検索から漏れた。しかし、検索キーワードが複合語にしか現れていないにも関わらず、単位語に分解したために収集できたデータも2件あった。「液晶－表示－装置」と「設備投資－減税」の2語が分解されたため、独立語として検索キーワードが存在しないデータを収集することができた。

クエリー	適合率	再現率
任天堂	5/5	5/5
農薬	3/3	3/4
液晶	6/6	6/7
減税	5/5	5/12

図6: 検索結果の適合率・再現率

5 課題・結論

本論文では、文字種を利用してキーワードを抽出し、さらに複合語を単位語に分解する方法について述べた。本手法では、単語辞書などの対象依存が大きい情報を必要とせず、接頭語・接尾語の情報や助詞の情報といった一般性の高い情報を基に語の抽出・分解を行なうことができる。

また、複合語の分割性能の向上のために単位語辞書を、また不要な語を排除するために不要語辞書を作成すると効果が高い。本手法

農薬
農薬－部門, 農薬－原料, 農薬－メーカー 農薬－会社, 農薬－事業, 農薬－需要 農薬－業界, 農薬－市場,
液晶
フィルム－液晶, 液晶－モニター 液晶－事業, 液晶－表示－装置, カラー－液晶 液晶－分野, カラー－液晶－パソコン
減税
減税－規模, 投資－減税, 減税－対象品目, 政策－減税 減税－措置, 住宅－減税, 設備投資－減税

図 7: 分解された複合語

では、このような辞書を作成する場合に対象文書から複合語や不要語の候補が抽出されているため、作成の労力が比較的少なくてよいという利点がある。

本手法では、複合語を単位語に分割し検索していることに起因する不要なデータのヒットや非キーワード文字からなるキーワードを表記的に抽出することといった適合率を悪化させる要因がある。今後の課題には、適合率の向上が挙げられる。方策の一つとして、語の連結情報を利用することが考えられる。テキスト内でキーワードが文内で共起したのか、複合語内で共起したのかといった情報を持つことで検索精度の向上を期待できる。

また他の課題として、キーワードを統計的尺度 (tf・idf など) を用い、キーワードの重み付けや企業名によく見られるような固有名詞となっている複合語を分割せずに抽出することなどを考えている。

参考文献

[1] 原田隆史, 細野公男他: 抄録からのキーワード自動抽出, 情報処理学会情報学基礎研究会,31-8,p55-61

[2] 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,VolJ74-D1,p556-566,1991

[3] 水野聡, 島田静男, 中牟田純, 近藤邦雄, 佐藤尚: 日本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, Vol.91,No.6,p41-45,1992

[4] 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,1993

[5] 森信介、長尾真:n グラム統計によるコーパスからの未知語抽出、電子情報通信学会論文誌、 Vol,No,p.7-12,1995

[6] 中渡瀬秀一, 木本晴夫: 統計的手法によるテキストからの重要語抽出メカニズム、電子情報処理学会情報学基礎研究会,39-6,p41-47,1995

[7] 山本幹雄, 増山正和, 品詞・区切り情報を含む拡張文字の連鎖を用いた日本語形態素解析: 言語処理学会 p421-424,1997

[8] 芥子ほか: 情報検索システム評価用ベンチマーク Ver1.0(BMIR-J1) について, 情報処理学会データベースシステム研究会, 106- 19,1996

[9] 森脇敏, 河部恒, 辻井潤一: 辞書を使わない日本語専門用語の自動分割, 言語処理学会 第 2 回全国大会,p.273-276