情 報 学 基 礎 39－6
（1995．9．14）

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

中渡瀬 秀一
木本 晴夫
{nakawtse, kimoto}@syrinx.dq.isl.ntt.jp
NTT情報通信研究所
〒238 神奈川県横須賀市武1-2356 ＮＴＴ横須賀研究開発センタ309Ａ
0468-59-3612

本論文では字面処理によって，テキストから重要語（キーワード）を自動抽出する方法について述べる．日本語の場合まず文章から単語を得るために，形態素解析が必要であるが，形態素解析には未知語や曖昧性の解消などの問題があり，これを解決するために，従来は複雑な規則や人間がメンテナンスしなければならない辞書が必要であった．本手法はNグラムの頻度情報を用いた完全な字面処理になっている．その手順では(1)まずNグラム頻度情報を使って重要な文字列を抽出し，(2)次にその中から無意味な文字列を排除する．実験ではこの手法が未知語や複合語の範囲を正しく識別し，抽出精度を向上させることを確認した．

Automatic Extraction of Keywords from Text Files Using n-gram Statistics

Hidekazu NAKAWATASE
Haruo KIMOTO
NTT Information and Communication Systems Laboratories
1-2356, Take, Yokosuka-shi, Kanagawa, 238-03 JAPAN
(telephone) +81-468-59-3612
(e-mail) {nakawtse, kimoto}@syrinx.dq.isl.ntt.jp

abstract

This paper describes a new method to extract free keywords automatically from a Japanese text. Morphological analysis is necessary to recognize words from a text for extraction of keywords. There exist, however, problems of unknown words recognition and ambiguity of compound words recognition, so dictionaries and complex heuristics are necessary to resolve them. Our method is based on the n-gram method and consists of 2 steps: (1) Evaluation of major strings using the n-gram statistics, and (2) Exclusion of nonsense strings. It was found that our method extracts keywords that is unknown word more precisely than conventional methods.

1．はじめに

今日，インターネットなどから参照可能なテキストは莫大な量に達しており[1]，この中から所望のものを効率よく入手することは困難になってきている．現在使用されている検索ツールとしてはキーワードで検索するテキストDB，WAISなどのサーチエンジン，そしてYahooやイエローページ（WWW用）などのシソーラスがある．これらを構築するにはキーワードや統制語（シソーラスの項目）を作成せねばならない．そこでこれらを機械的に高品質に作成する技術が必要である．

従来のキーワード抽出法では，まず文の分かち書きを，字種や区切り記号に着目したルール，分かち書き用の辞書を用いて語を品詞単位に分割する．次に，接頭語，接尾語を登録した辞書との照合により，分かち書きされた語から接頭語，接尾語を取り去り，さらに，複合語の分割を，最小単位の単語を登録した語彙辞書を利用して分割する．次に重要度評価を行い不要な語を除去しキーワードを絞り込む．このために様々なヒューリスティックが考案されてきた[9][10][11][12]．ヒューリスティックとしては単語の出現頻度や出現位置，特定のパタン，素性の導入などが提案されている．これらは主に重要度評価の方法について検討を行ってきた．しかしキーワード抽出法には以下のような問題もある．

1．1　語句の抽出誤り

従来のいずれの手法においても、まず形態素解析によって単語を抽出するというステップがある．そしてこのステップで得られた単語からキーワードが決定されるので，対象となるテキストから単語の抽出漏れや語の区切りの誤りがあると，たとえ重要度評価が優れていてもキーワードの精度を向上させることができない．

[1]例えばインターネットのニュースグループには１９９５年８月３日～１７日間に約１７０００件の(ｆｊ)記事（約４２Ｍバイト）が投稿されている．（saitoh@ics.es.osaka-u.ac.jp氏による）

これは形態素解析では辞書に登録されているされている語にマッチした文字列を語と見なすため，まだ辞書に登録されていない新語を抽出することができないことに起因する．図１は形態素解析の誤り例である．これらは辞書に［母一人子一人］［松商］［メモリアル］［全人代］等が登録されていないためにこのように解釈されてしまう．従来の複合語の処理では最小単位の連続した部分が候補となっていたので，複合語全体としては連続する名詞全体を候補にすればよいが，部分の語を候補にするときの誤りの原因となる．［全人代］のような略語は新聞では頻繁に使われるが最小単位としては個別の文字に分割されてしまうため不要な部分語を生成する可能性がある．［万里の長城］のような場合，名詞接続助詞で接続しているので普通は候補にならない．

母｜一人子｜一人	×
母一人｜子一人	○
松｜商学｜園	×
松商｜学園	○
メモ｜リアル｜チケット	×
メモリアル｜チケット	○
全｜人｜代　（全国人民代表会議）	×
全人代	○
敷島製｜パン	×
敷島製パン	○
北｜の｜政所	×
北の政所	○
万里｜の｜長城	×
万里の長城	○

図１　形態素解析の誤り例

このように辞書への登録漏れは形態素解析の精度を低下させてしまう．これを防ぐために人間が辞書を絶え間なくメンテナンスして，新語を登録してゆく必要がある．しかし日々増加してゆく新語や専門用語を人手で登録してゆくのには限界があるため，自動的に語を獲得する方法が必要である．

そこで本論文ではキーワード（重要語）の抽出方法として，Ｎグラム統計（あるｎ個の文字も組み合わせがどのような頻度で生じるか調べたも

の）を用いた、テキストからの重要語自動抽出方法について述べ，この方法を新聞記事に適用した結果を報告する．

2．Nグラム統計による語（キーワード候補）の抽出

Nグラム統計を用いた完全な字面処理による定型表現の自動抽出の試みとしては，[2][3][4][5][6]がありそれらでは文字列の接続の仕方に関するヒューリスティックを用いて単語や定型表現の自動抽出を行い，その有効性を示している．これに対して，本論文で提案する手法では文字列の出現頻度を正規化する方法となっている．

2．1　正規化頻度

本手法のアイデアはすでに[1]で説明してあるのでここでは要点をまとめておく．

［正規化頻度の要点］

1：*テキスト中で出現頻度の高い文字列は重要*

2：*文字列長の違いにより出現頻度分布に差があるためこれを正規化して比較*

3：*無意味な文字列が正規化頻度の比較で除去できる*

2，3については以下に補足しておく．

・2について

長さNの文字列の中にはN－n＋1個のnグラムが存在するが，文字種がm種の場合，可能なnグラムの種類はm^nであって，n＋1グラムとnグラムの可能な種類の比はmであり，あるnグラムの出現頻度が同じであってもその重要度は異なる．そこで出現頻度に重み付けをし，正規化する．文字種がmでそれらが均等に出現する場合には，n＋1グラムとnグラムの重みの比はmであるが，実際の言語ではそのようなことはないため，著者はnについて単調増加する対象文章に依存する数列として以下のものを検討し，簡単なキーワードの再現実験により有効性を示した．

nグラムの重み係数（Mn）＝

X1＋X2＋．．．＋Xn

Xn：対象文章に出現するnグラムの種類

これにより文字列Aの正規化頻度N（A）は

N(A)＝　Aの出現頻度・(X1+X2+…+Xn)

（ただし頻度は0オリジンでカウント）

となる．

・3について

図2は正規化頻度の計算結果例であるが，ある単語の部分文字列はその単語より低い正規化頻度となっている．このため重要語候補は正規化頻度の高い文字列から順にとり，同時に正規化頻度上位の文字列の下位にある部分文字列を排除することによって候補から無意味な文字列を排除できると考えられる．

[正規化頻度]	[文字列]
167724	オブジェクト指向データベース
142668	ブジェクト指向データベー
142668	ジェクト指向データベース
142668	オブジェクト指向データベ
129696	ブジェクト指向データベ
129696	ジェクト指向データベー
129696	オブジェクト指向データ
129696	ェクト指向データベース
122700	ブジェクト指
122700	ジェクト指向

図2　正規化頻度の計算例の一部．

具体的には本手法は次の手順に従う．

1）対象となるテキストの中に出現する任意の文字列（実際は必要な長さまで，実験では20文字）を記録し，その出現頻度を調べる．

2）文字列の長さ別の種類数を数え，nグラムの重み係数を得る．

3）1で得られた頻度を重み係数で正規化する．

4）正規化頻度の上位の文字列の部分になっているそれより下位の文字列は排除する．

こうして上記の4ステップによって本手法では正規化頻度という重要度評価のされたキーワード候補を得る．

2．2　正規化頻度の性質

以下では正規化頻度について確率的視点からの解釈について述べる．まず正規化頻度の持つ性質をいくつか紹介する．

［定義：正規化頻度］
文字列A、文字ａ（文字列は大文字、文字は小文字で表す）、そしてそれぞれのテキスト中における出現頻度＃A、＃ａと表し、またAの文字列長［A］、文字列長がαである文字列の種類を@αと表す．この時，正規化頻度は

$$N(A) = \#A \cdot \sum_{k=1}^{[A]} @k$$

と定義される．

［性質１：文字列長依存性］
任意のA，Bに対して
［A］<［B］、＃A＝＃B
ならば
N（A）＜N（B）

証明：定義より明らか

［性質２：頻度依存性］
任意のA，Bに対して
［A］=［B］、＃A＜＃B
ならば
N（A）＜N（B）

証明：定義より明らか

［性質３：接続確率依存性］
任意のA，ａに対して、

$$\frac{\#Aa}{\#A} > \frac{\sum_{k=1}^{[A]} @k}{\sum_{k=1}^{[A]+1} @k} \quad (1)$$

ならば
N（Aａ）＞N（A）
（aAについても同様）

証明：定義より明らか

ここで１の左辺は文字列Aの左右どちらかにａが接続する確率を表す．例えばA＝’私’、ａ＝’は’、テキストに出現するすべてのAｘ，ｙAの形の文字列が以下の様であったとする．

を私
の私
が私
私
私は
私は
私は
私が
私が
私の
私を

図３　ｙ私，私ｘの形の文字列

Aはここにあげた文字列にしか含まれないので
＃A＝１１、また＃Aａ＝３
よって［私］の右に［は］が接続する確率は3/11である．

一方、(1)の右辺を考えるために単純に@[A]＝１とする．この時Aｘ，ｙAの形の文字列はｘ，ｙにとりうる文字のの種類だけ（＝@([A]＋１)）存在するため，@[A]／@([A]＋１)はAにある１文字が接続する平均的な確率にほぼ近い．さらに@[A]／(@([A]＋１)＋@[A])は接続する文字として空文字を含めた場合と見なすことができる．このように考えると（１）式の右辺は長さが[A]である文字列にある１文字が接続する平均的確率（種類ベース）にほぼ近いと考えられる．したがって性質３は

文字列Aに文字ａが平均以上の確率で接続するならN（Aａ）＞N（A）である．

と考えられる．そして単語の中の部分文字列の正規化頻度が単語のそれよりも低いのは単語を構成する文字どうしがが高い確率で接続されているからだと思われる．実際の計算例では図４の様にな

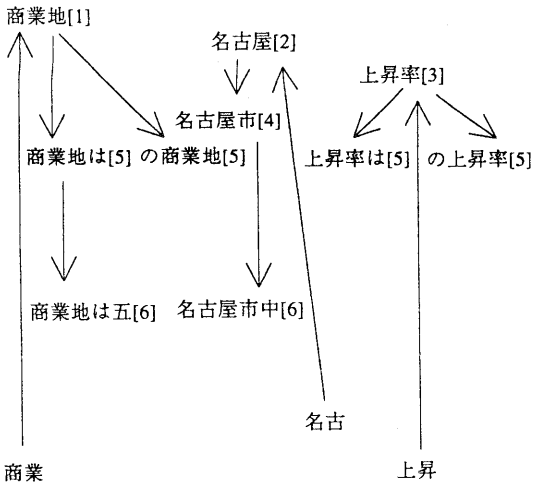

図４　正規化頻度の大小関係

っている（矢印は文字列どうしのの包含関係を表す．括弧内は正規化頻度による順位）．図５は計算された実際の値である．

[文字列]	[正規化頻度]	[順位]
商業地	29653	1
名古屋	25091	2
上昇率	20529	3
名古屋市	13704	4
上昇率は	10278	5
商業地は	10278	5
の上昇率	10278	5
の商業地	10278	5
名古屋市中	9112	6
商業地は五	9112	6

図５　正規化頻度計算例

[2]再現率＝ 抽出結果に含まれる正解キーワード数 / 正解キーワード数

[3]適合率＝ 抽出結果に含まれる正解キーワード数 / 抽出キーワード数

3　評価

本手法の有効性を確かめるために、新聞記事を用いてキーワード抽出実験を行った．ここでは評価尺度として再現率[2]、適合率[3]を使用した．再現率は正解キーワードがどれだけ抽出されたかを，適合率は抽出キーワードのうちどれだけが正解かを示すものである．

3．1　実験に使用したデータ

用意した新聞記事は２４０件で，記事の長さの平均は６６３文字である．これらにはあらかじめ５人の専門家によってキーワードを付与（記事中の語を使用）しておいた．そしてこれらのキーワードを正解データとして実験結果と比較した．付与されたキーワードの総数は６９７語である．本手法ではテキスト中に１回しか出現しない語は抽出されない．このような語は上のうち２３０語であった．したがって再現率は

（６９７－２３０）／６９７≒６７％

である．

3．2　適合率

キーワード候補としては正規化頻度の高い順に任意の個数の語を抽出すればよいのだが，ここでは正解キーワードの正規化頻度より高い値を持つ語をすべて候補とした．ただし本手法では正解キーワードを包含する文字列の正規化頻度が正解キーワードのそれより高い場合，正解キーワードとなる語が抽出されないので，その場合には包含する文字列で正規化頻度の最大のもので代用した．表１に適合率の計算結果を示す．

表１　適合率

	適合率	適合率(部分一致)
最大	1.000	1.000
最小	0.013	0.013
平均	0.216	0.253

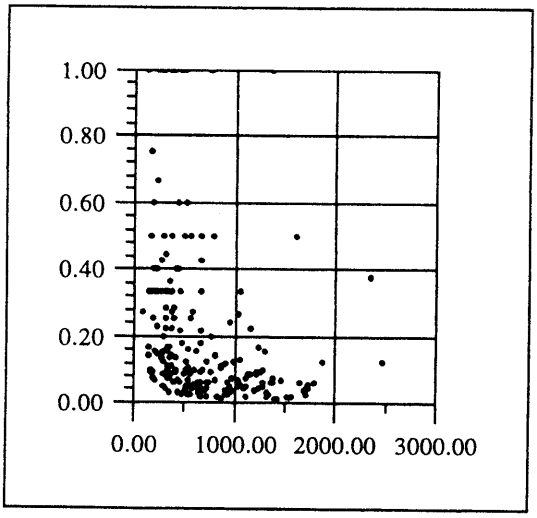

図６　記事長と適合率の関係
（縦軸：適合率，横軸：記事長）

これより適合率は１.３％～１００％の広い範囲で分布していることがわかる．全体の平均は約２２％である．

記事の長さと適合率の関係は図６のようになった．これからわかるように記事長が長くなると適合率の上限が低下していることがわかる．これは記事が長いほどそこに含まれる語が多いことに由来する．次に実験条件などが同一であるわけではないが参考までに従来の方法[10][12]による測定値と比較してみる．[10]は語の特徴によるヒュリスティックを利用した手法であり，これによると再現率＝６５％，適合率＝５０％と報告されている（改善前：再現率＝７０％，適合率＝１０％，フリータ－ム方式）．[12]は複合語の処理に独自のヒューリスティックを取り入れた方法である．この場合，再現率＝６４％，適合率＝２３％（改善前：再現率＝７０％，適合率＝１５％）となっている．後者は本手法の結果と近くなっている．後者の手法では正解キーワードと抽出キーワードとの部分一致も適合と見なす場合の評価もしている（適合率＝５５％）．これは検索アルゴリズム[13]が部分一致検索を実現しているためである．そこで本手法でもその場合の適合率を計算した（表１）．[12]では部分一致を適合と見なすことによって適合率が５５％にまで改善され，[10][12]ともに適合率は約５０％になったが，本手法では改善は５増％にとどまった．部分一致を含めることによって適合率が改善されるということは正解キーワードと抽出キーワードとの間の共通構成単語が多いということであるが，本手法の場合は別の原因があるということになる．

図７は適合率の低い文章の例である．これから抽出された抽出キーワードを図８に示す．

糖尿病　～（２６）～　池田義雄　年に一度は全身チェック　併発症の検査

わが国における糖尿病の圧倒的多数は、男女を問わず中高年者層によって占められています
中高年ということになると、その健康問題といわゆる成人病の存在を無視して語るわけにはまいりません
この成人病の代表疾患の一つが糖尿病であることについては、皆さん既によくご承知の通りです
高齢化社会が現実のものとなっている今日、無病息災はもはや幻影だといわれています
それ故、一病息災は無論のこと、二病でも三病でも、それをコントロールして息災に暮らすことが当たり前となり、そういう人の数は年々ふえてきております
糖尿病にプラスして、何か併発症（高血圧、高脂血症、痛風、肝障害など）を、かかえている人の場合、それらがどのようないきさつでみつかったのか興味がもたれます
最も多い例は、最初に糖尿病がみつかって、その後の精査と長年の経過の中で、時には意図的に行われた検査によって、時には偶然の機会に、その異常がみつかってきています
ともあれ、糖尿病プラスアルファをもつ人の数は大変多くなっているというのが昨今の印象です
このような事情ではありますが、中にはついつい見逃されていて、やがて重大な結果を招いている場合も少なくありません
．．．．
（以下略）

図７　適合率の低い記事の例
（適合率＝０．０１５１，１４２０文字，
正解キーワード＝糖尿病，病気）

[抽出キーワード]	[正規化頻度]
コントロール	20276
の	17752
糖尿病*	17168
りません	15775
血糖コントロール	13500
い	12997
ありま	12876
ありません	12411
です	11590
てい	11590
と	11095
が	11095
という	10730
ている	10730
に	10461

```
は                  9510
のような            9465
ます                9272
こと                9272
年に一度は全身チェック 8822
れてい              8584    . .  (以下略)
```

図８　抽出キーワード

図８の示すとおり，平仮名だけからなる文字列が多いのがわかる．これを除けば残りはほぼキーワード候補として妥当である．そこで抽出キーワードから平仮名だけによる文字列，さらに括弧など記号を含む文字列を削除した場合を計算した．その結果を表２に示す．

表２　適合率（かな文字列削除）

	適合率	適合率(部分一致)
最大	1.000	1.000
最小	0.019	0.019
平均	0.294	0.348

キーワードの中には平仮名だけで構成される語もあるので，これを削除すると再現率が低下する．正解キーワードの中でそのような語は［いじめ］，［おもちゃ］の２語であった．しかし適合率はこれによって約１０%改善され約３５%（部分一致）となった．これは[10][12]の５０%には及ばないが本手法では［全人代］，［政労協］のような未知語を正しく抽出できる（従来は[全][人][代]のように解析され，部分文字列が抽出されてしまう）という独自の効果を持つ．そのため既存の方法におけるヒュリスティックと本手法とを組み合わせると，さらに改善されることが期待できる．つまり本手法で抽出されなかった正解キーワードが既存の方法で抽出され（再現率改善），不要語の絞り込みが行われる（適合率改善）．

３．３　抽出される語について

次に正解キーワードと抽出キーワードを比較した例を紹介する．図９は評価で用いた記事とそこから抽出されたキーワードの一例である．この場合正解キーワードをすべて得るためには７番目の［ソ連軍］まで抽出しなければならない．この時の適合率は約７０%であるが，ここで不適合である抽出キーワードを見てみるとそれらには［撤退計画］［拒否］など記事の内容から見て重要である語が含まれており，一概に正解キーワードと一致しないものを不適合とするのには問題があることがわかる．

```
[正解キーワード]ソ連，パキスタン，アフガニスタン
[記事]
アフガンがソ連軍撤退案　パキスタンは拒否
【ワシントン二日共同】二日付のワシントン・ポスト紙が米国務省や外交筋などの話として報じたところによると、アフガニスタン政府はこのほど、ソ連軍の撤退計画案をパキスタン側に提示した
これに対しパキスタンは撤退終了までの期間が長すぎるとしてこの案を拒否したが、ソ連軍が一九七八年十二月アフガニスタンに侵攻して以来、アフガニスタン側が撤退計画を提示したのは初めて
アフガニスタン問題解決のためコルドベス国連事務次長は三国間を往復し、三月中旬に撤退計画の詳細を得て直ちに同計画をパキスタン側に示した

[正解キーワード]ソ連，パキスタン，アフガニスタン
[抽出キーワード]      [正規化頻度]
アフガニスタン          3909
パキスタン              2691
をパキスタン側に        1499
撤退計画                1380
パキスタンは            1102
示した                   970
ソ連軍                   970
ワシントン               897
提示した                 690
計画を                   485
二日                     288
拒否                     288
案を                     288
（以下１文字の語）
```

図９　記事と抽出キーワードの例

４　終わりに

本論文ではテキストからの重要語自動抽出方法について報告した．本手法では文字列の出現頻度の正規化に注目し，任意の文字列から重要な語の候補を抽出できるようにした．また確率的な解釈も示し，正規化頻度の性質を理解しやすいようにした．本手法は高い確率で接続されている文字の集まり，高い頻度で現れる文字列ほど正規化頻度が高いという性質による．そして実験の結果，これだけで単純なフリータ－ム方式を改善できることが確かめられた．また既存の方法にない抽出効

果があるため，既存法との併用でさらにキーワード抽出精度が改善されることが期待される．

参考文献

[1] 中渡瀬　秀一：統計的手法によるテキストからのキーワード抽出法，電子情報通信学会データ工学研究会，Vol. 95，No. 81，pp. 9-16(1995).

[2] 長尾　眞，森　信介：大規模日本語テキストのnグラム統計の作り方と語句の自動抽出，情報処理学会自然言語処理研究会，96-1，pp. 1-8(1993).

[3] 森　信介，長尾　眞：nグラム統計によるコーパスからの未知語抽出，電気情報通信学会言語理解とコミュニケーション研究会，95-8，pp. 7-12(1995).

[4] 新納　浩幸，伊佐原　均：疑似Nグラムを用いた助動詞的定型表現の自動抽出，情報処理学会論文誌，Vol. 36，No. 1，pp. 32-40(1995).

[5] 新納　浩幸：文字列と後続文字列との接続割合の変化を利用した定型的文末表現の自動抽出，情報処理学会自然言語処理研究会報告，104-6，pp. 39-46(1994).

[6] 池原　悟，白井　諭，河岡　司：N-gramを用いた連鎖型共起表現の自動抽出法，言語処理学会第1回年次大会発表論文集，pp. 313-316(1995).

[7] 北　研二，小倉　健太郎，森元　つよし，矢野　米雄：仕事量基準を用いたコーパスからの定型表現の自動抽出，情報処理学会論文誌，Vol. 34，No. 9，pp. 1937-1943 (1993).

[8] 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).

[9] 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，pp. 55-61 (1993).

[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(D-1)，Vol. J74-D-I，pp. 556-566(1991).

[11] 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，pp. 41-45(1992).

[12] 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，Vol. 97，No. 15，pp. 103-110(1993).

[13] 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).