未踏テキスト情報中のキーワードの抽出システム開発

Keyword Extraction from Brandnew Japanese Text Data

梅村 恭司
Kyoji UMEMURA

豊橋技術科学大学 情報工学系（〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1 E-mail: umemura@tutics.tut.ac.jp ）

ABSTRACT. If keywords in documents are found are automatically, the keywords make it easier to handle the documents. We have developed a method to extract keywords automatically. This method uses statistical measure called adaptation. This report describes the whole algorithm and shows the result of outputs.

１．背景

最新の技術情報の速報やニュースなどを整理, 検索する際には, 記事の内容を特定できるキーワードの付与が行なわれている. キーワードを付与する作業を自動化できれば, キーワードが付与されていない文書の操作も容易になる. これまで検討されているキーワード自動抽出は, 辞書を用いて形態素解析を行い, その後, 品詞情報と頻度情報をもとにキーワードを弁別する手法がある. しかし, 辞書を用いる手法は日々新しい単語が生まれるインターネット時代の情報処理としては問題がある. それは, 処理の自動化が必要な最新の文章からキーワードを辞書に登録し続ける必要があるため生産性が悪いことと, 辞書に登録されていない全く未知の用語に対する汎用性がないことが挙げられる. そこで本プロジェクトでは, 辞書を用いないという条件のもとで, 文章からキーワードを自動抽出する手法を開発する. 人間による整理を行っていないテキスト情報は未踏のテキストと呼ぶこともできる. このシステムは未踏のテキストからキーワードを取り出す機能があるといってもよい.

本システムは, 文字列の頻度に加えて出現集中を示す統計量を用いることを特徴とする. また, システム構築に関しては, すべての部分文字列について, 出現集中を求める必要があるため, 単純に求めると計算量が膨大になるが, それを実現する手法にも工夫がある.
キーワードの抽出を辞書を行わないで実行するというアイディアは新しいものであり, 辞書を使うことが当然のこととされるキーワード抽出の現状とは一線を画するシステムである. また. そのアイディアをシステムとして実現する上で, 計算量の問題があり, それを解決したのがプロジェクトを実行した成果といえる.

２．目的

本プロジェクトでは、キーワード付与が行われていない文章の操作を容易にするために, キーワードを付与する作業を自動化することを目的とした. そこで, 辞書を用いないという条件の下で, 文章からキーワードを自動抽出する手法を開発した. 提案する手法は, 文字列の頻度に加えて, 出現集中を示す統計量を用いることを特徴としている.

３．出現頻度と出現集中

文字列の出現頻度は, 統計的に言語を処理するときの基本の統計量である. 情報検索においても, 「ある単語がドキュメントに現れる確率」に関する情報量で重みをつけると性能が向上することが知られている. 出現確率の P(1回出現) を推定するのには以下の式が使われる.

$$\hat{P}(\text{1回出現}) = \frac{DF(x)}{N} \qquad \textbf{(1)}$$

出現集中は adaptation[2]として知られる統計量であり, 「ある単語が一つのドキュメントに現れたという条件で, 同じ単語がもう一度出現する P(2回出現 | 1回出現) の推定値である. この確率を推定するために, 対象の文字列 x に関して, 「その文字列 x を含むドキュメントの数: $DF(x)$」と「その文字列 x を 2 回以上含むドキュメントの数: $DF_2(x)$」を数え上げる. そして, ベイズの規則を考慮した次式より推定する. ここで N は全ドキュメント数である.

$$\begin{aligned}\hat{P}(\text{2回以上} \mid \text{1回出現}) &= \frac{\hat{P}(\text{2回出現} \wedge \text{1回出現})}{\hat{P}(\text{1回出現})} \\ &= \frac{\hat{P}(\text{2回出現})}{\hat{P}(\text{1回出現})} \\ &= \frac{DF_2(x)/N}{DF(x)/N} \\ &= \frac{DF_2(x)}{DF(x)} \qquad \textbf{(2)}\end{aligned}$$

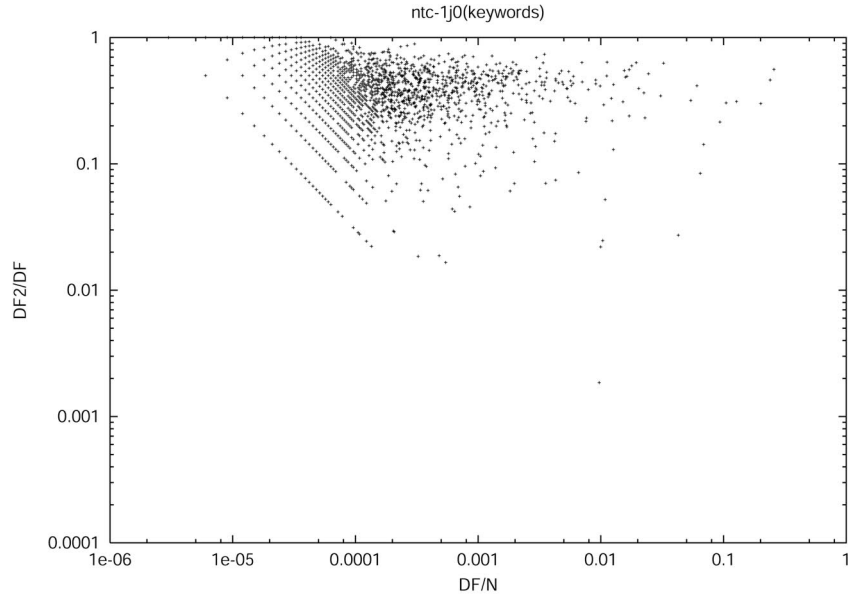


図1) 論文情報におけるキーワードの分布

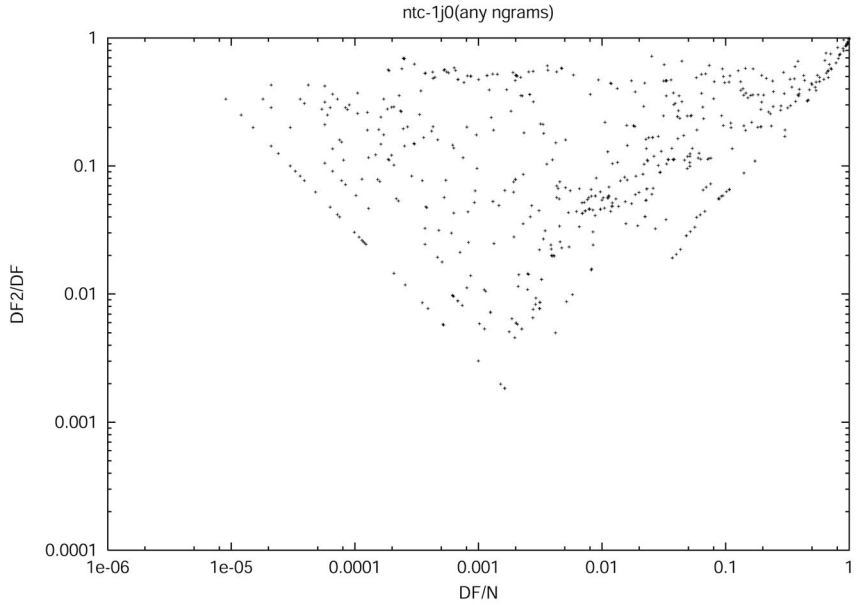


図2) 論文情報における部分文字列の分布

文字列の出現がポアソン分布に従うとすれば, P(2回出現|1回出現) は P(1回出現) の値は等しくなるはずであるが, 実際のテキストの文字列の分布はランダムではない. このため, P(2回出現|1回出現) は P(1回出現) は等しくならない. 文献 2 によれば英語のキーワードの P(2回出現|1回出現) は, P(1回出現) には依存しない一定の値となることが報告されている. 我々はこれに注目し, キーワードの分析に P(2回出現|1回出現) を利用することを着想した. このため, すべての部分文字列について, P(2回出現|1回出現) と P(1回出現) を分析することを実行した. すなわち, 文字列の出現がポアソン分布に従うと仮定すると $DF(x)/N$ と $DF_2(x)/DF(x)$ は同じ値になるが, 実際のコーパスでは $DF_2(x)/DF(x)$ の値が大きく, キーワードと認める文字列なら $DF(x)/N$ に比べその差は特に大きくなることが観測できる. 以下に詳しく分析を行う. なお $DF(x)/N$, $DF_2(x)/DF(x)$ において, 引数の文字列 x が明らかな場合は, これ以後, 単に DF, DF_2 と書くことにする.

４．日本語における adaptation の分布》

英語で報告されている事象が日本語でも成立しているかどうか確認するために, 論文情報について分析した. 論文情報には, 日本語の論文のアブストラクト情報 33 万件があり, それぞれのアブストラクトに著者が記述したキーワードがある. そのキーワード 10000 個について, DF/N を横軸に DF_2/DF を縦軸にプロットしたものを図 1に示す.

英語で観測されたのと同様に, DF/N と関係なく DF_2/DF が一定の値を示す傾向が観測できる. 同じ論文情報について, ランダムに文字列を取り出して表示した

表1) 各部分文字列における統計値

DF	DF_2	DF/N	DF_2/DF	x
124696	79894	3.75×10^{-1}	6.41×10^{-1}	ロ
3672	2413	1.10×10^{-2}	6.57×10^{-1}	ロボ
3320	2237	9.97×10^{-3}	6.74×10^{-1}	ロボッ
3319	2237	9.97×10^{-3}	6.74×10^{-1}	ロボット
577	96	1.73×10^{-3}	1.66×10^{-1}	ロボットに
30	1	9.01×10^{-5}	3.33×10^{-2}	ロボットにつ
30	1	9.01×10^{-5}	3.33×10^{-2}	ロボットについ
30	1	9.01×10^{-5}	3.33×10^{-2}	ロボットについて
23345	12250	7.01×10^{-2}	5.25×10^{-1}	ボ
4045	2551	1.22×10^{-2}	6.31×10^{-1}	ボッ
3370	2551	1.01×10^{-2}	6.68×10^{-1}	ボット
578	96	1.74×10^{-3}	1.66×10^{-1}	ボットに
30	1	9.01×10^{-5}	3.33×10^{-2}	ボットにつ
30	1	9.01×10^{-5}	3.33×10^{-2}	ボットについ
30	1	9.01×10^{-5}	3.33×10^{-2}	ボットについて

総ドキュメント数 $N = 332921$)

ものが図 2である．図 1に比べて， DF_2/DF の存在範囲が広いことが観測できる．そこで，キーワードの分布範囲にある部分文字列をうまく取り出すことによりキーワードを取り出すことが可能であることが示唆される．

なお，この図で， DF/N が小さいところでみられる右下向きの線は，DF_2 の値が整数であるために生じたものである．データのドキュメントの数がさらに増加すれば，この領域は DF/N が小さい方向に移動する．

また, 図 2の DF/N の大きいところでみられる右上がりの部分は, 非常に頻繁に出現する記号や助詞を構成する文字列である．

どのような文字列が adaptation の値が高いかを調べるために，具体的な単語を含む文字列について， DF/N と DF_2/DF の値を調査する．サンプルとして，「ロボットについて」という語の部分文字列 x の一例とそれに対応する DF ， DF_2 ， DF/N ， DF_2/DF を表 1に示す．ここで観測できることは二つある．一つは，キーワードを構成する文字列では，DF/N に比べて DF_2/DF が大きい．このことは, キーワードとなる語はドキュメント中に複数回出現することを表す．また，表 1においても容易に確認できる．もう一つは, 語の境界を越えると DF_2/DF が小さくなることである．これは，表 1において，部分文字列 x が「ロボット」から「ロボットに」と変化すると, それまではほぼ一定値の DF_2/DF が小さくなっていることから確認できる．一般に DF/N に比べ，DF_2/DF の方が値の存在する範囲が小さく, 単語の区切りが明確である．

（1）文字列の分割)

図 1において，キーワードが集中している領域の文字列をキーワードとして選ぶという明白な方法があるが，これでは，キーワードとして正しい語が選ばれない．それは, キーワードを構成する文字の部分文字列の出現パターンは, キーワードの出現パターンとほぼ同じであるため, キーワードの部分文字列かキーワードかを分別することができないからである．このことは，表 1において，「ロボット」の部分文字列である「ロボッ」や「ボット」が，ロボットの DF/N ， DF_2/DF と非常に近い値をとっていることより確認できる．

これを解決するために, 与えられている文字列を分割したとき, それぞれの文字列が単語らしい確率の積が最大になるように分割を行って, キーワードの境界を求めるという方針にした．

ここで問題となるのは単語らしいという尺度を辞書を用いずに求めることである．ここで，DF_2/DF の値をもとに，文字列の単語らしさを推定するにした．

DF_2/DF は, もともとキーワードらしさを推定する尺度と解釈できるが, これを単語らしさと置き換えて処理を試みる．この尺度については改良の余地があることは明らかであるが, システムの目的がキーワードの境界が正しく求まることであるので, キーワードらしさで単語境界を求めることも一つの方法と考えられる．

ポテンシャルの推定について, DF_2/DF の対数の値を使うことを基本とするが，図 1から考えて，いくつかの変更を行った．その変更は，図 3のフローチャートで示す．

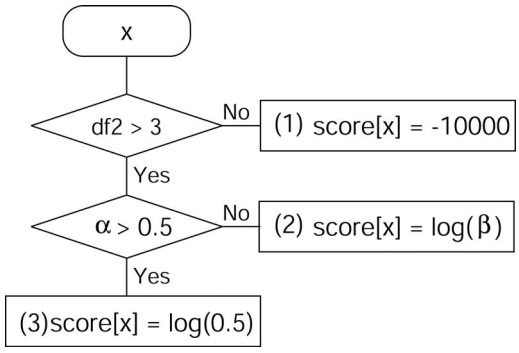


図3 ポテンシャルの計算法

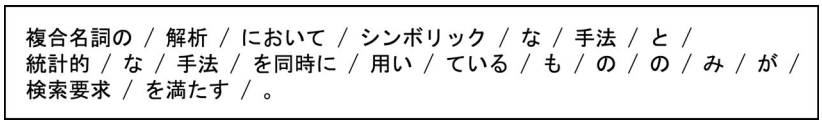


図4 本手法による分割

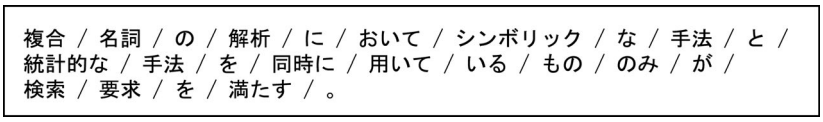


図5 形態素解析システム Chasen による分割

まず，DF_2 があまりに少ない文字列 x は，単語とみなさずポテンシャルを低く定義している（図 3(1)）．この理由は，少しの偶然のために DF_2 の値が変化すると，それが大きく DF_2/DF に影響し，ポテンシャルとして信用ができないからである．これは，図 1，図 2の DF/N の小さい領域に不自然な線が存在していることからも分かる．

次に，出現確率 DF/N が 0.5 を超えるところでは，単語らしさの推定値を一律 log(0.5) に制限した．これは図 2の $DF/N > 0.5$ の領域で，すべての部分文字列の出現集中の確率がキーワードの出現集中を上回っていることに対処するものである．文章において助詞などは，多数出現し，とりたててキーワードとは認められないという考え方をした．すべての文字列分割において，以上のようにポテンシャルを求め，各文字列のポテンシャルの合計が最大となる分割を求める．

ここで，分析対象の文字列の長さを N とすると，文字列のすべての分割の方法は $2^{(N-1)}$ 個あり，すべての分割について，単純な方法でポテンシャルの合計を求めると計算量が増えすぎて，実際に使用することができない．これを解決する方法はダイナミックプログラミングを用いる方法であり vitabi サーチとして知られている．ここで，単語らしさを決めるためにいくつかの閾値を用いたが，これはすべての単語を網羅しなければならない辞書と異なり，サンプルのキーワードから値を求めることができる．例えば，DF_2/DF の飽和の値 0.5 はドキュメントの長さに影響を受けるが，サンプルのキーワードから決定できる．具体的には，検査対象のドキュメントで図 1の DF/N と DF_2/DF の範囲を求め，0.5 や 0.1 などの値を決定する．この操作は，新しい文書でも実行できる操作である．また，DF_2 の制限は，統計的に安定させるためのものなのでコーパスの種類に依存しない．

この方法で分割した例を図 4，図 5に示す．DF/N が大きいときに DF_2/DF にキーワードの情報が含まれないことの当然の帰結として，Chasen[3]に比べ，DF/N の大きい助詞・助動詞の分割は不自然だが，キーワードは正確に取れていることが観測できる．

５．システム実現上の技術的問題点と解決方法》

分割を行うときに，$DF(x)$ と $DF_2(x)$ の値を用いる．この $DF(x)$，$DF_2(x)$ の引数 x は，すべての部分文字列を取りうる．ここで，分析対象の文字列の長さ N の全部分文字列の総数は，$N(N-1)/2$ となる．今回，分析の対象としているテキストは 100M($=10^8$)バイト以上ある．この文字列の部分文字列の $DF(x)$ と $DF_2(x)$ を表にしようとすると，10^{16} 程度の項目を持つ表を作ることになり，作成することも難しいが，それを保持しておくのも実際的な記憶容量ではない．

表を使用しないで，検査対象の文字列ごとにドキュメントを先頭から調べていくことをすると，処理時間が問題となる．与えられた文字列が，それぞれのドキュメントに含まれている数を数え，それが 1 以上であるドキュメントの数と 2 以上であるドキュメントの数を数えるということで，計算時間を無視すれば，必要な情報が集まる．この方法でも理論的には計算はでき，プログラムを実現するのは簡単である．しかしながら，一つの部分文字列の値を計算するたびに100Mバイトを検査することになる．100 文字程度の文章の部分文字列をすべて分析するのには，およそ 5000 文字に対する分析が必要となる．一つの文字の計算に 1 分かかるとすると，5000 分ほどの計算になり，これはおよそ 100 時間，つまり 4 日の計算量になる．これでは，計算時間が実用的なものといえない．

そこで，我々は，Suffix Array[4]として知られるデータ構造を利用した．この Suffix Array は，テキストの 5 倍のメモリ空間を要するが，すべての部分文字列の位置が $\log(M)$（M :テキストの大きさ）で特定できるデータ構造である．Suffix Array は，分析対象となるテキスト中のすべての場所から最後までの部分文字列を，辞書順に並べた情報を保持している．この文字列は，テキストの長さと同じだけの数がある．任意の文字列について，Suffix Array をニ分探索すると，それが出現した場所から始まる文字列を特定できる．つまり，任意の文字列の出現場所を $\log(M)$ の計算量で特定できる．特定された場所について，それが属するドキュメントを特定し，ドキュメントの重複を検査して $DF(x)$ と $DF_2(x)$ を求める．

対象となる文字列を普通の形で表現すると，$M(M-1)/2$ だけ必要なメモリの領域が必要となりメモリ量が問題となる．Suffix Array で巧妙なのは，この 部分文字列を，もとのテキストの始まる場所という一つの整数で表現するというところである．この表現形式のため，どんなに長い文字列でも同一のメモリ量で表現できる．このため，必要なテーブルのメモリ容量は，分析対象のテキストに比例するもので済み，100M バイト以上の文字列でも分析ができることになる．

Suffix Array の例を表 2に示す．Suffix Array は整数の配列であるが，それで表現される文字列を並べて示している．この配列を利用すると，二分探索することで文字列 bc の出現は 2 回であり，その場所が 3 と 6 であるということを求めることができる．

表2　Suffix Array の例

全体	ababcabcd
suffix	文字列
0	ababcabcd
2	abcabcd
5	abcd
1	bcabcabcd
3	bcabcd
6	bcd
4	cabcd
7	cd
8	d

```
char * text;
int * suffix;

int suffix_index_compare(int *a, int *b)
{
      if(*a == *b) return(0);
      return strcmp(text + *a, text + *b);
}

main()
{
      int i;
      int n;
        .....
      n = strlen(text);
      for(i=0;i<n;i++) { suffix[i] = i; }
      qsort(suffix, n, sizeof(int),
  &suffix_index_compare);
        .....
}
```

図6　Suffix Array を生成する方法

Suffix Arrayは図6のようなCプログラムで作成できる．メモリ上に検査対象となる文字列を text という変数から指し，それと同じ数の整数の列を suffix という変数から指してあるとする．そこで，この整数の列を，その内容の整数の場所から始まる文字列の辞書順でソートすると，Suffix Array ができる．ここで qsort は C のライブラリのソート関数である．

これで，通常のコンピュータで分析の実行ができるが，さらに高速化できる工夫がある．この方法では，検査しようとする文字列 x の総頻度が数百以下の場合は高速に $DF(x)$ と $DF_2(x)$ が求まるが，総頻度が大きくなるとその回数だけ検査が起きるため速度が低下することが観測できる．総頻度は，調査対象の全文字列の長さよりも小さい値であるけれども，数万になることがある．このような数について場所の検査をすることが速度低下の要因となる．

そこで，実際の分析プログラムでは，頻度が大きいときには，一度求まった文字列の頻度情報は記憶しておき，再計算を避けるようになっている．この工夫のため，最初はスロースタートで分析が始まるが，ある程度の情報が蓄えられる量になると，分析速度が向上するようになる．

（1）キーワード候補の選別》

部分文字列 x の出現確率 $DF(x)/N$ が大きい場合は，どのドキュメントにも現れる文字列を示すため，文書を識別する能力が低く，逆に，DF/N が小さい場合，例えば一度しか出現しないような文字列は，他の文書との関連を示す能力がないことになる．そのため，キーワードはドキュメントを特定できる単語という性質を持つことから，DF/N がある一定の範囲内に含まれる．さらに，単語がドキュメントの内容に係わるものかを推定するために，DF_2/DF の値を考慮して求めた単語らしさのポテンシャルを再び用いる．DF/N と DF_2/DF の値の範囲については，コーパスを利用して学習を行い，抽出する範囲の候補を選別する．これらの条件を満たしたもののうち，文字列 x の長さ $len[x]$ が 1 より大きいもののみをキーワードと考えている．これは，キーワードとなる語は一語より長いという経験則に基づいている．具体的には，図 7の条件をすべて満たしたものをキーワードとして抽出している．

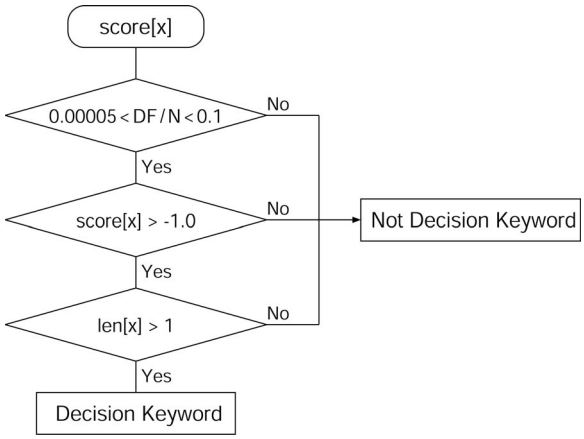


図7 キーワードの検出

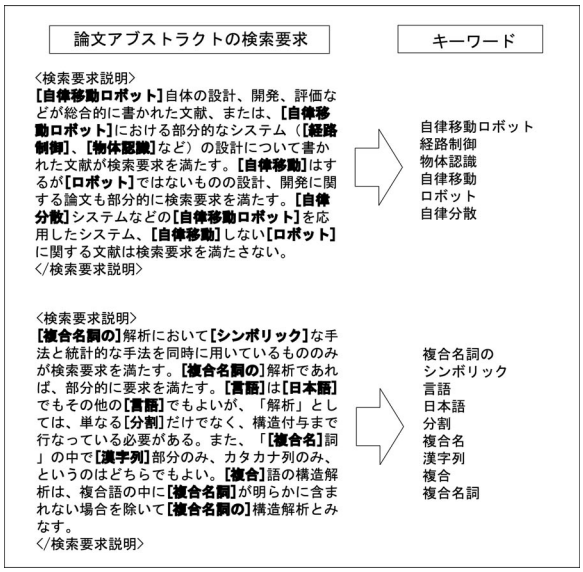


図8 検索要求から抽出されたキーワード

6．キーワード抽出システムの評価》

論文アブストラクト 33 万件[1]に対する検索要求 30 件に関して，本手法を用いてキーワード抽出を行った．その一例を図 8に示す．

また，実際に 30 件の検索要求すべてから抽出したキ

正しい抽出語				誤った抽出語
DHP	LFG	PAC学習	アルゴリズム	Ag
インターネット	エージェント	オブジェクト	カタカナ	al
キーワード	クラスタリング	グラフ	コロケーション	iアルゴリズム
シソーラス	シナリオ	シンボリック	スキーマ	lient
ツール	テキスト	データベース	データマイニング	roba
トラヒック	ニューラルネットワーク	ニューロン	ネットワーク	、（
バス	プログラム	マイニング	マッチング問題	「自律
ユーザ	ラティス	リダクション	ロボット	しば
位置計測	異表記	英語	英単語	イ論
音声	音声認識	画像	階層	ラルネットワーク
慣用表現	漢字列	機械	機械翻訳	レンド
学習	機能	記述	係り受け解析	人工ニュー
経路制御	計算	検索	検知	複合名
言語	故障	故障診断	語彙機能文法	複合名詞の
高速	字幕	次元	自然言語	文字の
構文解析	自動索引	自律	自律移動	報資源
自律移動ロボット	自律分散	辞書	手話	
性能	素子	対話	対話システム	
知識獲得	知的	抽出	同時送信	
内容検索	日本語	日本語文	入力	
認識	認知	複合	複合名詞	
物体	物体認識	分割	分析	
文書	文法	並列	翻訳	
要素	類推	連想記憶	連想検索	

図9 抽出されたキーワードのリスト

① 339501件の論文アブストラクト使用

我々は過去数年来、窒素の放電光による**[オゾン生成]**の可能性について研究を続けてきた。昨年の前報でも**[二重]**[**ガラス**]管と**[スパイラル]**[**電極**]を用いた。窒素の放電光の**[エネルギー]**で**[オゾン]**を**[生成]**しようとする**[放電管]**を製作し、その結果について報告を行なった。しかし、この様な**[放電管]**では**[オゾン]**の原料として用いている酸素ガスに**[沿面放電]**が発生し、**[オゾン生成]**に大きな影響を与えていることが判明した。即ちこれまで行なってきた測定では窒素の放電光の効果が正しく測定できていないことが判明した。そこで前報で報告した結果の再評価を行なうと供に、新たに、正しい結果の解釈を与える。

② 33949件(=①/10)の論文アブストラクト使用

我々は過去数年来、**[窒素の放電光]**による**[オゾン生成]**の可能性について研究を続けてきた。昨年の前報でも**[二重]**[**ガラス**]管と**[スパイラ]**[**ル電極**]を用いた。**[窒素の放電光]**の**[エネルギー]**で**[オゾン]**を**[生成]**しようとする**[放電管]**を製作し、その結果について報告を行なった。しかし、この様な**[放電管]**では**[オゾン]**の原料として用いている**[酸素]**[**ガス**]に**[沿面放電]**が発生し、**[オゾン生成に]**大きな影響を与えていることが判明した。即ちこれまで行なってきた**[測定]**では**[窒素の放電光]**の効果が正しく**[測定]**できていないことが判明した。そこで前報で報告した結果の再評価を行なうと供に、新たに、正しい結果の解釈を与える。

③ 3394件(=①/100)の論文アブストラクト使用

我々は過去数年来、**[窒素の放電]**[**光による**][**オゾン生成**]の可能性について研究を続けてきた。昨年の前報でも**[二重]**[**ガラス**]管とス**[パイ]**[**ラル**][**電極**]を用いた。**[窒素の放]**[**電光の**][**エネルギー**]で**[オゾン]**を**[生成]**しようとする放電管を製作し、その結果について報告を行なった。しかし、この様な放電管では**[オゾン]**の原料として用いている**[酸素]**ガスに**[沿面]**放電が発生し、**[オゾン生成]**に大きな影響を与えていることが判明した。即ちこれまで行なってきた測定では**[窒素の放電光]**の効果が正しく測定できていないことが判明した。そこで前報で報告した結果の再評価を行なうと供に、新たに、正しい結果の解釈を与える。

図10 テキストサイズの変化に伴うキーワードの振る舞いの変化

ーワードを図 9に示す．ここでは，語を正しく抽出したものと, 単語の切り出しを誤って抽出したものに分類している．図 9より，抽出した 113 件のキーワードのうち 97 件は単語を捉えていることから, 全体の 85.8%は正確に語の境界を抽出できていることが分かる．

キーワードを正しくとらえているかということを, 定量的に評価することは困難であるため，ここでは図 8や図 9によって，実際に抽出したキーワードを示した．

辞書を全く使用していないシステムにもかかわらず，キーワードの境界の特定と, キーワードの選び出しができている．故に，図 1及び図 2の分布の差を用いて，キーワード抽出を行えることが分かる．以下に，本システムの振る舞いをより明らかにするため, テキストサイズ及びテキストの種類を変化させ，さらに検討を行う．

（1）テキストのサイズによる振る舞いの変化》

本システムは，部分文字列の DF/N ， DF_2/DF の確率の推定値によりキーワード抽出を行っている．この確率の推定値はテキストのサイズが大きいほど正確に求められるため, テキストのサイズが大きいほど望ましい．ここでは, テキストサイズの変化に伴うシステムの振る舞いの変化を示す．なお，この実行においては，パラメータの再分析を行わず, 同じパラメータを用いて実行を行った．

図 10に 339501 件のテキストを用いて得られたキーワードと，約 10 分の 1(33949)件，約 100 分の 1(3394)件と変化させ，得られたキーワードを示す．これより，テキストサイズが小さくなるに従い, 単語の区切りをとらえることができなくなり，語に助詞が付加されたり，語の途中で切れていることが読み取れる．また，339501 件すべてを用いた場合は得られなかったキーワードとしては成り立たないような「測定」，「光による」などの語も抽出されていることが分かる．これは，テキストサイズが小さくなることにより, 確率を正確に推定できなくなったためである．

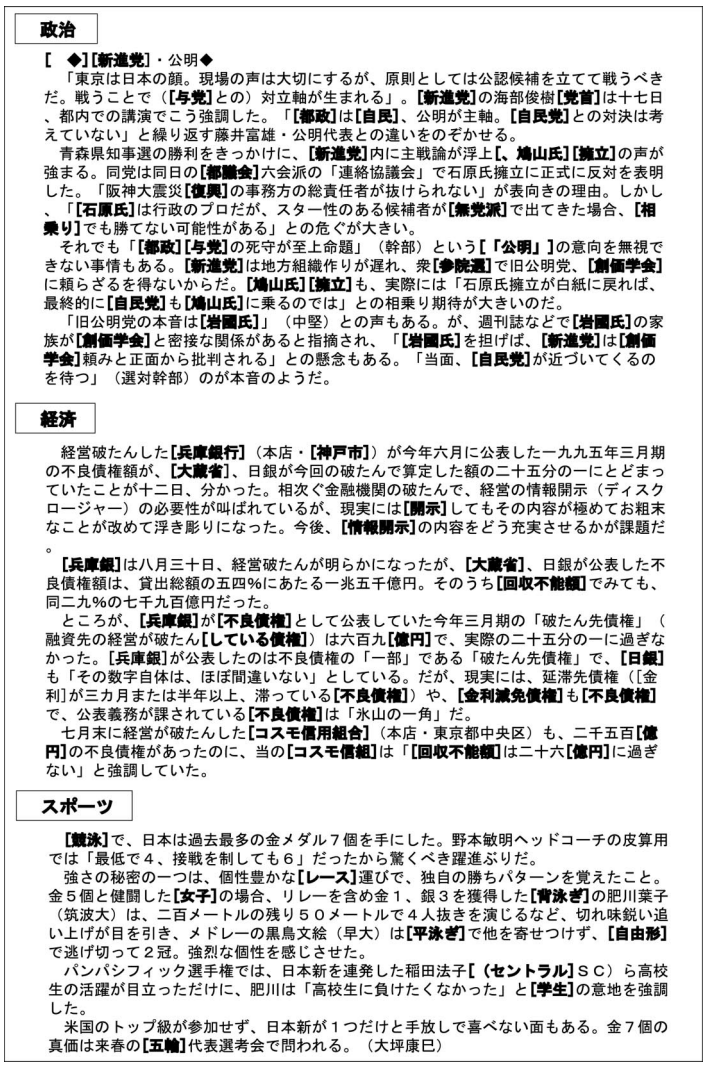
（2）テキストの種類による振る舞いの変化》

本アルゴリズムを, これまで検討してきた論文アブストラクトとは異なる, 新聞記事に対して用いた結果を図 11に示す．これより，対象とするデータが論文アブストラクトから新聞記事へと変化しても, 本システムによってキーワード抽出が行えることが分かる．

また，新聞記事は政治，経済，社会，スポーツ，国際など様々な分野に分類されている．そのため，新聞記事に対しキーワード抽出を行うことにより, 本システムの抽出能力が分野ごとに異なっていることが判明した. 図 11には，テキストサイズが同程度であった政治，経済，スポーツの記事を示した．これらを比較すると，政治や経済の記事においては, 党名や人名などの固有名詞も比較的とらえているが，スポーツの記事においては，人名などの固有名詞をとらえることができていないことが分かる．このことは，政治や経済の記事では，出現する人名がある程度固定されているのに対し, スポーツ記事では, 競技や大会ごとに出場選手名が一回しか出現しないことが多いため, 出現集中が単語らしさを表せていないためである．また，図 11のスポー記事において，キーワードとして抽出したい「金メダル」などの語をとらえられないことも，スポーツ記事という性質上，各記事の情報が大きく異なり，「金メダル」に関する記事が少なかったため，本システムでは，この語をキーワードとしてとらえることができなかったと考えられる．

７．他の研究との比較》

キーワードを取り出すシステムに関する研究は多いが, 全く辞書を使わずにキーワードをテキストより取り出すという研究は少数である．森ら[5]は統計処理により未知語の抽出を行っている．また，この報告でも Suffix Array を使用している．しかし，ここで使用されている統計量は，コーパス中の頻度であり，これは，ほぼ，出現確率と同様な統計量である．森らは出現集中の情報は利用していない．表 1で示したように，出現集中は，出現確率とは異なった情報であるため, 利用している情報が森らの手法とは異なる．我々の方法は，逆に，キーワードの境界を特定するためには出現確率を利用していない. 辞書を使用しないで単語を切り出すという目的は同一であるが，その手法は異なるものである．



政治

[　◆][新進党]・公明◆

「東京は日本の顔。現場の声は大切にするが、原則としては公認候補を立てて戦うべきだ。戦うことで（**[与党]**との）対立軸が生まれる」。**[新進党]**の海部俊樹**[党首]**は十七日、都内での講演でこう強調した。「**[都政]**は**[自民]**、公明が主軸。**[自民党]**との対決は考えていない」と繰り返す藤井富雄・公明代表との違いをのぞかせる。

青森県知事選の勝利をきっかけに、**[新進党]**内に主戦論が浮上**[、鳩山氏][擁立]**の声が強まる。同党は同日の**[都議会]**六会派の「連絡協議会」で石原氏擁立に正式に反対を表明した。「阪神大震災**[復興]**の事務方の総責任者が抜けられない」が表向きの理由。しかし、「**[石原氏]**は行政のプロだが、スター性のある候補者が**[無党派]**で出てきた場合、**[相乗り]**でも勝てない可能性がある」との危ぐが大きい。

それでも「**[都政][与党]**の死守が至上命題」（幹部）という**[「公明」]**の意向を無視できない事情もある。**[新進党]**は地方組織作りが遅れ、衆**[参院選]**で旧公明党、**[創価学会]**に頼らざるを得ないからだ。**[鳩山氏][擁立]**も、実際には「石原氏擁立が白紙に戻れば、最終的に**[自民党]**も**[鳩山氏]**に乗るのでは」との相乗り期待が大きいのだ。

「旧公明党の本音は**[岩國氏]**」（中堅）との声もある。が、週刊誌などで**[岩國氏]**の家族が**[創価学会]**と密接な関係があると指摘され、「**[岩國氏]**を担げば、**[新進党]**は**[創価学会]**頼みと正面から批判される」との懸念もある。「当面、**[自民党]**が近づいてくるのを待つ」（選対幹部）のが本音のようだ。

経済

経営破たんした**[兵庫銀行]**（本店・**[神戸市]**）が今年六月に公表した一九九五年三月期の不良債権額が、**[大蔵省]**、日銀が今回の破たんで算定した額の二十五分の一にとどまっていたことが十二日、分かった。相次ぐ金融機関の破たんで、経営の情報開示（ディスクロージャー）の必要性が叫ばれているが、現実には**[開示]**してもその内容が極めてお粗末なことが改めて浮き彫りになった。今後、**[情報開示]**の内容をどう充実させるかが課題だ。

[兵庫銀]は八月三十日、経営破たんが明らかになったが、**[大蔵省]**、日銀が公表した不良債権額は、貸出総額の五四％にあたる一兆五千億円。そのうち**[回収不能額]**でみても、同二九％の七千九百億円だった。

ところが、**[兵庫銀]**が**[不良債権]**として公表していた今年三月期の「破たん先債権」（融資先の経営が破たん**[している債権]**）は六百九**[億円]**で、実際の二十五分の一に過ぎなかった。**[兵庫銀]**が公表したのは不良債権の「一部」である「破たん先債権」で、**[日銀]**も「その数字自体は、ほぼ間違いない」としている。だが、現実には、延滞先債権（[金利]が三カ月または半年以上、滞っている**[不良債権]**）や、**[金利減免債権]**も**[不良債権]**で、公表義務が課されている**[不良債権]**は「氷山の一角」だ。

七月末に経営が破たんした**[コスモ信用組合]**（本店・東京都中央区）も、二千五百**[億円]**の不良債権があったのに、当の**[コスモ信組]**は「**[回収不能額]**は二十六**[億円]**に過ぎない」と強調していた。

スポーツ

[競泳]で、日本は過去最多の金メダル７個を手にした。野本敏明ヘッドコーチの皮算用では「最低で４、接戦を制しても６」だったから驚くべき躍進ぶりだ。

強さの秘密の一つは、個性豊かな**[レース]**運びで、独自の勝ちパターンを覚えたこと。金５個と健闘した**[女子]**の場合、リレーを含め金１、銀３を獲得した**[背泳ぎ]**の肥川葉子（筑波大）は、二百メートルの残り５０メートルで４人抜きを演じるなど、切れ味鋭い追い上げが目を引き、メドレーの黒鳥文絵（早大）は**[平泳ぎ]**で他を寄せつけず、**[自由形]**で逃げ切って２冠。強烈な個性を感じさせた。

パンパシフィック選手権では、日本新を連発した稲田法子**[（セントラル]**ＳＣ）ら高校生の活躍が目立っただけに、肥川は「高校生に負けたくなかった」と**[学生]**の意地を強調した。

米国のトップ級が参加せず、日本新が１つだけと手放しで喜べない面もある。金７個の真価は来春の**[五輪]**代表選考会で問われる。（大坪康巳）

図11 テキスト種類によるキーワードの振る舞いの変化

８．波及効果》

今回の開発は，日本語の情報処理の基礎技術として重要な位置にある．英語では，単語の境界が明示されているが日本語では単語の境界が明示されていない．このため，形態素解析システムが広く使われる．しかし，形態素解析システムには日々生まれてくる新しい言葉に対応できないという重大な問題がある．ここで開発したシステムはこの形態素解析の問題を解決する作用があり，形態素解析システムを補うという位置にある．形態素解析システムが日本語の情報処理の基礎技術として重要な位置にあるのと同様に，辞書を用いないキーワードの抽出システムも日本語の情報処理の重要な基礎技術である．この技術は以下のような発展をしてビジネスにつなげることができる．

- 未踏テキスト中のイベントの検出システム
  このシステムは，「この時代，地域の重大事件は何か」という問いにシステムが回答できるものである．現状の検索システムでは，事件に関わる具体的なキーワードがないと検索ができないため，このようなシステムを自動で作成するのは難しい．現状では，重大事件を選びだす人間の作業が必要である．
  今回の成果を利用すれば，人間が指定するべきキーワードの候補を，システム側から提案することができる．
- 未踏テキストのハイパーテキストへの自動変換
  Web などで広く使われているハイパーテキストであるが，参照情報を選びだし，相互に関係のある状態にするには人間が文章の参照部分を取り出し，関連あるドキュメントを整理するという作業を行っている．あるいは，その作業をするほうが効果のある情報が，この作業のコストが高いために，参照情報が整備されないことになっている．
  キーワードは参照情報の候補とも言えるものであり，このキーワードを特定し，同様なキーワードをもつ文書を相互参照の対象とすれば，テキストの利用価値が向上する．

上記以外にも適用範囲はある．キーワード抽出の要素技術である，単語の分割については形態素解析システムを補うシステムである．このため，形態素解析を使用しているアプリケーションであるならば，この技術を適用できる分野であるといえる．

９．達成度》

当初の予定である目標は達成した．辞書を用いないキーワード抽出システムは動作しており，そのパラメータ

を決定するための分析情報の資料と、システムの動作を記述する資料と複数の種類のテキスト情報に関する処理結果の資料を作成できた.

このプロジェクトにおける最大の成果は実際的な速度で統計情報を求める分析システムと考える. 論文を作成するには原理的な分析を行うためのプロトタイププログラムで作業をして結果を整理すれば十分であるため, 速度に関しては注意を払うことが少ない. このため実用システムにつなげるには, 多くの作業が残るものになる. 今回の成果の分析システムは, 実用システムの一部として使用できるものになっている.

１０．今後の課題》

分析が本格化して, 出現集中による効果も明らかになったが, 「金メダル」など, ドキュメントに対する出現集中は観測されないが, ジャンルに対する集中が観測される単語の処理が必要であることが明らかになった. また, 時間に関しての出現集中もキーワードの選別に必要であることが明らかになった. そして, キーワードが単語となる比率を高めるために, 頻度情報も使う必要があることが明らかになった.

次にするべきことは二つある. 一つは, 今回, 提案したドキュメントへの出現集中だけでなく, 分野に関する出現集中や時間軸に関する出現集中を分析し, これをキーワード抽出の能力の向上につなげることである. もう一つは, 形態素解析システムを補う結果を, 実際に形態素解析システムの改良につなげていくことである.

１１．まとめ》

本報告では, 文字列の出現を統計的に解析し, キーワードを抽出する方法を示した. この方法は辞書を用いない方法である. 本手法は, 新しいドキュメントにキーワードを付与する人手の作業の軽減や, 最新情報のサーベイなど, 辞書の整備が間に合わない種類のテキストの処理に役立つ自動抽出法である.

参加企業及び機関》

豊橋技術科学大学 情報工学系
ソフトウェアシステム研究室

謝辞》

まず, 英語における出現集中について示唆をいただいた AT&T の Ken Church 博士との議論が, このプロジェクトをスタートさせることを可能とした. このプロジェクトは, 豊橋技術大学情報工学系 梅村研究室のメンバで行った. 梅村がプロトタイプを作成し, 全体の調整を行った. 山本英子君は, 分析を行うための環境整備の全般を行い, 特に高速化した分析システムの開発を行った. 武田善行君は, 山本英子君のシステム開発を助け, 分析プログラムのレビューを行うと同時に, 多くの言語情報を整形し, キーワードの情報を整備し, 分析作業を行った. そして図 1, 図 2のように, 分布をわかりやすく表示し必要なパラメータの特定に貢献した. 田中路子君は, テキストの大きさによる振る舞いの変化の分析や新聞記事のジャンルによる振る舞いの変化の分析を行うとともに, この資料の最終的な仕上げを行った.

このプロジェクトの実行にはプロジェクトリーダの東京大学松島克守教授には, 適用範囲と分野と応用の方向性について適切なアドバイスをいただいた. サポート組織として紹介いただいたネイチャーランドの斎藤昌義氏には, 単なる事務的なサポートの範囲をこえて, ビジネスに必要なスペックの提案などの有用なアドバイスもいただけた.

参考文献》

[1] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato and Souichiro Hidaka.: Overview of IR Tasks at the First NTCIR Workshop, NTCIR Workshop, vol.1, pp.11--44 (1999).

[2] Kenneth W. Church.: Empirical Estimates of Adaptation, Coling-2000, pp.180--186 (2000).

[3] Yuji Matsumoto, Akira Kitaushi, Tatsuo Yamashita, Yoshitaka Hirano, Osamu Imaichi and Tomoaki Imamura.: Japanese morphological analysis system chasen manual, NAIST,Technical Report NAISTIS-TR97007(1997).

[4] Udi Manber, Gene Myers: Suffix arrays: A new method for on-line string searches, SIAM Journal on Computing, Vol.22, No.5, pp.935—948(1993).

[5] 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, 情報処理学会論文誌, Vol.39, No.7, pp.2093-2100(1998).