2004－DD－46　(3)
2004／9／24

Bigramの反復度を用いた技術用語抽出

中瀬 健太　　梅村 恭司
豊橋技術科学大学 情報工学系

自然言語処理において、文書中から技術用語などの重要語句を抽出する技術は、多くの応用のある問題である。分かち書きがなされない日本語や中国語においては、まず語境界を認定する処理が必要である。一般的には、単語辞書情報を用いた形態素解析が行われるが、辞書を元にした手法では未知語に対する柔軟性に関する問題が生じる。また、専門分野における技術用語の多くは複合語からなるため、語の単位の認定には困難さが存在する。本稿では、反復度という統計量を利用し、辞書情報を用いないキーワード抽出方式を提案する。本手法の特徴として、長い n-gram の統計値を用いずに Bigram のみに関する反復度を利用することで、スケーラブルかつ高速な用語抽出方式を実現する。

Technical terms extraction using adaptation of bigram

Kenta Nakase　　Kyoji Umemura
Toyohashi University of Technology, Information & Computer Science Department

In natural language processing, extracting important terms like technical terms is very important. Japanese and Chinese have no explicit word boundaries, so morphological analysis is needed as preprocessing. Though dictionary based morphological analysis is widely used, it is vulnerable to unknown word problem. Moreover, determining the boundary of a term is a difficult problem because most of technical terms are consist of multiword. We present a keyword extraction method using adaptation without dictionary information. We realize a scalable and fast extraction method using adaptation of bigram, not using statistics of long n-gram.

１．はじめに

自然言語処理において、文書中から技術用語などの重要語句を抽出する技術は、多くの応用のある問題である。しかし、新たな用語は常に生まれ続けているため、人手で技術用語を抽出することは難しく、また非常にコストがかかる作業である。さらに、専門分野における技術用語の多くは複合語からなり、語の単位の認定に関しても困難さが存在する。

日本語や中国語などは単語の境界が明確でないため、用語の抽出に際しては、まず単語の分割処理が必要である。一般的には、辞書情報を用いた形態素解析を行い、得られた単語の品詞情報などが利用される。この手法の問題点として、辞書情報の維持に高いコストがかかることと、辞書に登録されていない未知の語に対する汎用性が低下することなどが挙げられる。インターネットの利用などに伴い、情報の量が急増すると共に、新しい単語は日々生まれ続けており、未知語に対する汎用性の確保は重要な問題である。

そこで、辞書を始めとする言語知識を利用しない用語抽出アルゴリズムが提案されている。例えば、武田は統計的な情報のみを利用して用語抽出を行う方式を提案している[1][2]。武田は、文字列の反復度という統計量に注目して用語の抽出を行い、抽出した用語を情報検索エンジンのインデックス語として使用して、検索性能を向上させている。しかし武田の手法はドキュメントの全部分文字列の統計量を利用しており、用語抽出に際してはドキュメントの量と比較して非常に大きなメモリが必要となる。このため、抽出対象となる文書の大きさに制限が生じている。近年、コンピュータの利用増加に伴ってデジタルデータが急激に増大しており、インターネットにおける Web データや特許情報は GB オーダのデータサイズを有している。このようなデータを扱うためには、スケーラビリティの高い抽出方式が必要不可欠である。

本稿では、言語知識を用いず、かつスケーラビリティの高い用語抽出方式を提案する。これは、武田の方式を元にし、利用する情報を Bigram の統計量のみに制限することで実現する。コーパス中におけるユニークな Bigram の総数はコーパスのサイズ増加に対する影響が少ないため、巨大なコーパスに対しても適用が可能となる。

２．関連研究

コーパスからの用語抽出に関する研究としては様々なものが提案されている[4][5][6][7]。例えば文献[4]では、ある語の左右に連接する語の種類及び頻度に注目して専門用語の認定を行っている。これらの手法は、第一段階として形態素解析による語分割を行っている。我々は、明示的には形態素解析を行わず、反復度という統計量に注目し、統計的な情報のみを利用した用語抽出方式を検討する。

３．従来法

本稿で提案する手法も、武田の手法と同様に反復度に基づいている。反復度は、Church により提案された文字列の集中出現を表す統計量である[3]。反復度は、ある文字列がドキュメント中に出現した際、同じ文字列が再び出現する確率である。反復度βは次式で推定できる。

$$\hat{\beta} = \frac{df_2(x)}{df(x)}$$

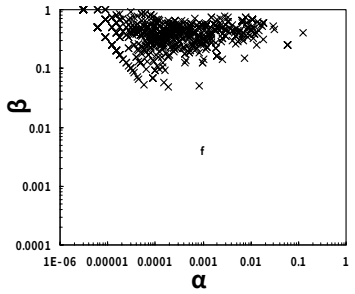
df(x) : 文字列 *x* を含むドキュメントの数
$df_2(x)$: 文字列 *x* を 2 回以上含むドキュメントの数

また、*df(x)*を用いて文字列の出現確率αを定義する。αは下式で推定される。

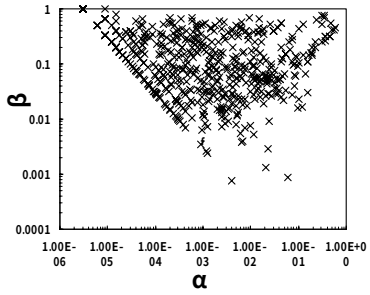
$$\hat{\alpha} = \frac{df(x)}{N}$$

N : ドキュメントの総数

Church は英単語の反復度について、反復度が内容語に対して高い値を持ち、機能語に対しては低い値を持つと報告している。武田はこれを日本語と中国語について検証し、技術用語が高い反復度を持つことを報告している。図 1は横軸に出現確率α、縦軸に反復度βの値を取り、NTCIR1 コレクションの論文アブストラクト情報に付与された著者キーワードの分布と、ランダムに切り出した文字列の分布を比較したものである。図に示されるように、技術用語はランダムに切り出した文字列と比較して高いβを持つ。



(a) 著者キーワード



(b) ランダムに切り出した文字列

図1 著者キーワードとランダムに切り出した文字列の分布

また、反復度は語の境界において大きく減少するという特徴がある。例を図 2に示す。

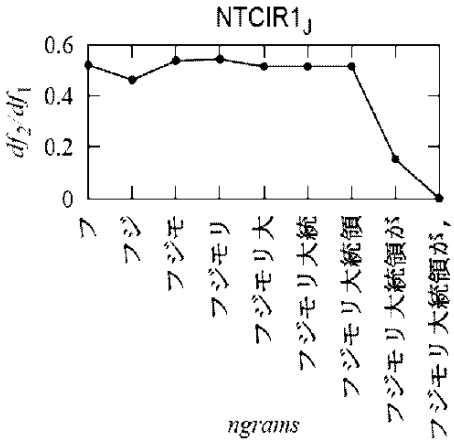


図2 語境界における反復度の推移[1]

文字列の出現頻度が比較的緩やかに減少するのに対し、反復度は語の境界で急激に減少する。この変化を捉えることで辞書を使わずに語の境界を認定し、さらに、技術用語のβ値の高さを利用して用語抽出を行っている。

４．提案法

４．１ 概要

我々は武田の手法と同様に反復度を利用して技術用語抽出を行うが、利用する情報を Bigram の統計量のみに制限することでスケーラビリティを向上する。

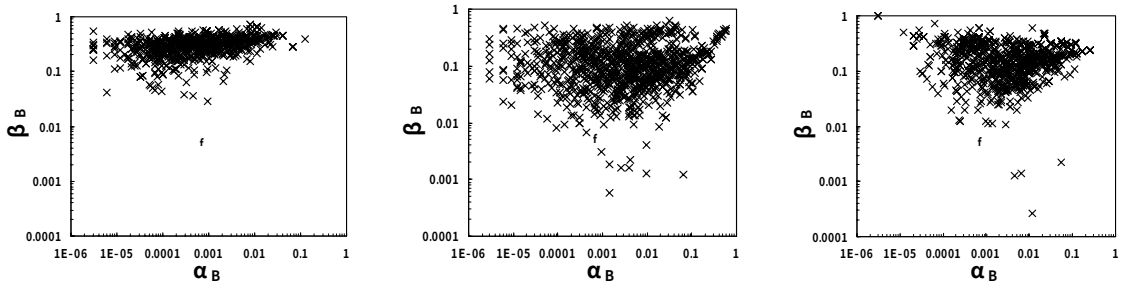
技術用語を構成する Bigram の反復度について調査を行った結果、技術用語そのものだけでなく、技術用語を構成するすべての Bigram が高い値を持つことが観察された。よって、語の反復度の値を、語を構成する Bigram の反復度の最小値を代替として用いることを考える。また、出現確率に関しては、語を構成する Bigram のうち、出現がまれな Bigram の特徴を捉えるために、同様に Bigram の出現確率の最小値を利用する。

このような考えに基づき、Bigram の統計量のみを利用したα_Bとβ_Bを以下のように定義する。これらの統計量を従来法におけるαとβといった特徴量の代替として利用する。

$$\alpha_B(w) = \min_{b_i \in BI(w)} \left\{ \frac{df(b_i)}{N} \right\}$$

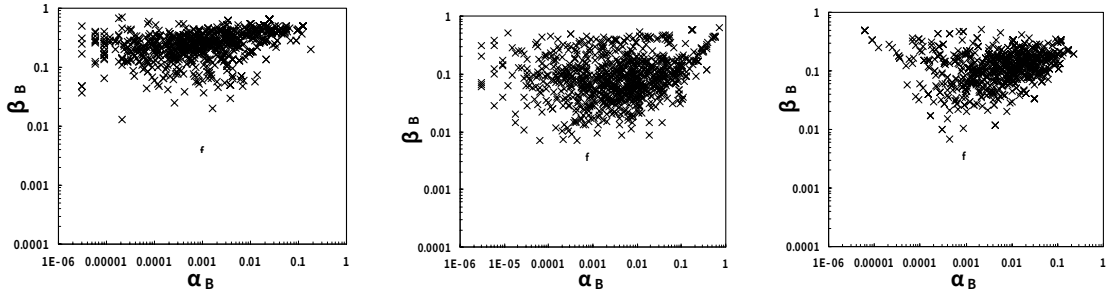
$$\beta_B(w) = \min_{b_i \in BI(w)} \left\{ \frac{df_2(b_i)}{df(b_i)} \right\}$$

BI(w) : 文字列 *w* を構成する Bigram の集合
（例）: *w* : 技術用語 → *BI(w)* : {技術、術用、用語}



(a) 著者キーワード　(b) ランダムに切り出した文字列　(c) 著者キーワードの境界を含む Bigram

図 3 文字列の分布（NTCIR1 論文アブストラクト）



(a) 技術用語　(b) ランダムに切り出した文字列　(c) 技術用語の境界を含む Bigram

図 4 文字列の分布(TMREC 用語抽出タスクコーパス)

ここで α_B と β_B はそれぞれ、文字列 w を構成する Bigram の集合の出現確率および反復度の最小値である。このように定義した α_B と β_B をそれぞれ横軸、縦軸にとって描いたキーワードとランダム文字列の分布を図 3、図 4 に示す。

Bigram の統計量のみしか使用していないにも関わらず、図 1と同様に、キーワードの β_B は高い値を示し、ランダムな文字列と比較して技術用語の分布は偏っていることがわかる。よって、α_B や β_B の値は α や β の値の近似とまではいえないが、α_B と β_B の情報を利用して技術用語と非技術用語を弁別できると考えられる。

４．２ 用語抽出処理

本稿で提案する、α_B と β_B を利用した用語抽出アルゴリズムについて述べる。

日本語は分かち書きされていないため、用語抽出に際してはまず、単語の境界を認定する処理が必要である。本手法では辞書等を用いず、統計的な手法により語の認定を行う。なお、厳密な形態素の認定を行うことを目的とはせず、技術用語における境界情報に注目した語分割処理を α_B と β_B の情報により行う。さらに、反復度の高い文字列シーケンスを用語の候補として扱う。

反復度は語の境界において大きく減少するという特性を持つ。武田は任意長文字列についての反復度の減少を報告しているが、Bigram の統計量のみに注目した場合についても同様に、語の境界において反復度の減少が観測される。キーワードの境界を含む Bigram の分布を図 3、図 4 に示す。分布領域がランダム文字列と同様に大きく広がり、キーワードの分布と比較して偏りが生じている。よって、α_B や β_B の情報によってキーワード境界の認定も行うことが可能であると考えられる。具体的な例として、「非線形振動を」という文字列に対する α_B、β_B の値を表 1に示す。

表 1 「非線形振動を」に対する α_B 及び β_B

Bigram	α_B	β_B
非線	0.0242	0.473
線形	0.0375	0.460
形振	0.0005	0.293
振動	0.0309	0.583
動を	0.0478	0.166

表に示されるように「非線形振動」という技術用語を構成する Bigram はすべて高い β_B を持っていることが分かる。さらに、「動を」という技術用語の境界を含む Bigram において、β_B の値が大きく減少していることが分かる。この減少を観察することで、用語の境界を捉えることができると考えた。さらに、得られた用語候補から、出現確率や反復度の値を元にして特徴的な語を選択する。ここでは、高い反復度の値を持つ

文字列シーケンスの中で、出現確率の値が低い用語を選択する。以上のような、語分割処理と用語選択処理を行う。

まず、語分割処理では大まかな語境界の認定を行う。先ほど示したように、用語を構成する Bigram はすべて高い反復度の値を持つため、反復度 β_B が、あるしきい値 K よりも低くなった Bigram の位置において文字列を分割する処理を行う。しきい値は、正解データが付与されたドキュメントを学習コーパスとして用い、もっとも高い精度が得られる値に決定する。

次に、得られた文字列のリストから用語を選択する処理を行う。用語の判定には α_B の値及び、Bigram の平均反復度 β_A の値を用いる。多くの技術用語は複合語からなり、語を構成する Bigram の反復度には揺れが存在する。全体的な反復度の高さを捕らえるために、スコアには β_A を用いる。β_A は以下の式で定義される。

$$\beta_A(w) = \underset{b_i \in BI(w)}{average}\left\{\frac{df_2(b_i)}{df(b_i)}\right\}$$

この値を用い、用語らしさとして以下のようなスコアを定義する。

$$score(w) = \log\left(\frac{1}{\alpha_B(w)}\right)\beta_A(w)$$

語分割で得られた文字列のリストについて、技術用語とその他の文字列に対するスコア値の分布を図 5、図 6に示す。ここでの技術用語とは、コーパスにおいて正解キーワードとなっているものを指す。それぞれのリストからサンプリングを行い、スコアを付与した。図に示されるように、技術用語は高いスコアを持ち、それ以外の文字列は低いスコアを持つ。このスコアがあるしきい値よりも高い語を用語として抽出する。しきい値の決定に際しては、文献[7]で用いられている方式と同様に線形判別分析によって行う。しきい値φの決定は下式に基づく。ここで、μ_0、μ_1 はそれぞれ、技術用語及びその他の文字列に対するスコア分布の平均値であり、同様に σ_0、σ_1 は分散値である。

$$\phi = \frac{\mu_1\sigma_0 + \mu_0\sigma_1}{\sigma_0 + \sigma_1}$$

得られたしきい値φよりも高いスコアを持つ語を用語として抽出する。

以上のプロセスによって抽出された語の例を表 2に示す。

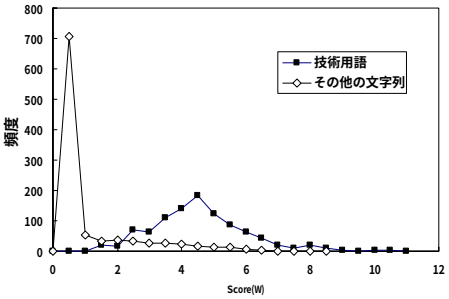


図5　スコア分布(NTCIR コーパス 1000 サンプル)

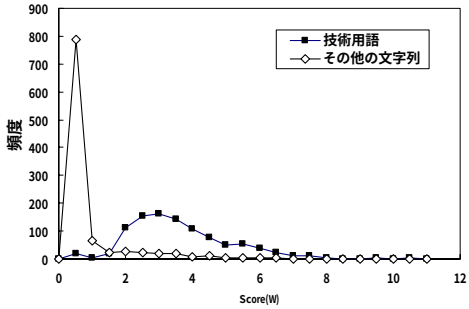


図6　スコア分布(TMREC コーパス 1000 サンプル)

表 2 抽出結果例

文書	抽出用語
パラメータ励振回路における位相角の変化について。可飽和鉄心を含むパラメータ励振回路方程式に対してＭＡＰＰＩＮＧ法を適用して不変閉曲線を描く。	パラメータ励起回路 位相角 過飽和 鉄心 不変閉曲線
リンクを辿る検索方式としてファジィ検索方式を採用することでリンクに適切さが計算され、柔軟かつ効率的な文書間ナビゲーションが可能となる。	リンク 検索方式 ファジィ検索方式 文書 ナビゲーション

５．実験

NTCIR1 論文アブストラクト、及び NTCIR1 の用語抽出 (TMREC) タスクのコーパスを用いて抽出実験を行った。実験に使用したデータの詳細を表 3に示す。

表3 実験データ

コレクション	ドキュメント数（件）	データサイズ（MB）
論文アブストラクト	332918	193
TMREC	1785	1.5

５．１　抽出精度評価

論文アブストラクトについては、すべてのドキュメントに対して著者キーワードが付与されており、このキーワードを正解データとして実験を行う。また、

TMREC コーパスについては、用語抽出タスクの正解データを用いた。

まず、抽出にあたって Bigram の反復度や出現頻度の情報が必要となるため、これらを論文アブストラクト 33 万件の情報を用いて計算した。次に、それぞれのテストコレクションから 100 件を学習コーパスとして抽出し、語分割のしきい値 K とスコアしきい値φを決定する。

表4 抽出パラメータ

コレクション	K	φ
論文アブストラクト	0.26	2.97
TMREC	0.24	1.41

さらに、コレクションから学習コーパスを除いてそれぞれ 100 件ずつをテストコーパスとして抽出し、抽出実験を行った。実験結果を表 5、表 6に示す。なお、F-measure は次式で定義される。

$$F-measure = \frac{2\times 精度 \times 再現率}{精度 + 再現率}$$

評価は精度、再現率、F-measure について行い、武田の手法を従来法として比較する。

表5 NTCIR1 コーパスを用いた実験結果

	F-measure	精度	再現率
従来法	0.105	0.074	0.183
提案法	0.156	0.103	0.326

表6 TMREC コーパスを用いた実験結果

	F-measure	精度	再現率
従来法	0.286	0.403	0.222
提案法	0.403	0.350	0.475

実験結果より、利用する情報を Bigram の統計量のみに制限したにもかかわらず、全部分文字列の統計量を利用した従来法よりも高い抽出精度が得られていることが分かる。また、スコアしきい値を変化させた際の F-measure の変化を図 7に示す。

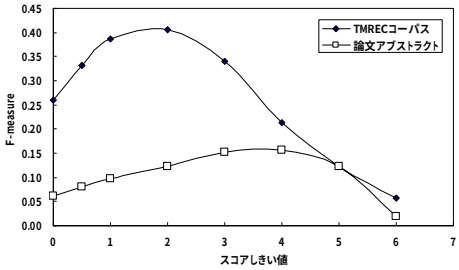


図 7　スコアしきい値変化に対する F-measure 推移

図より、しきい値φの値がおおよそ適当な値に決定されている事が分かる。

５．２　抽出速度評価

従来法ではキーワード抽出に際して、全部分文字列の統計量を利用する。梅村ら[8]の方法により、高速に全部分文字列に対する統計量を取得可能であるが、SuffixArray を元にしているため、大量のテキストに対して、インデックスの構築に時間がかかるという問題がある。提案法では、Bigram の統計量のみを利用しているため、インデックスの構築に関しても容易に行う事ができる。また、ドキュメントの追加・削除に対するインデックスの更新処理も容易である。従来法とのインデックス構築時間の比較を図 8、図 9に示す。

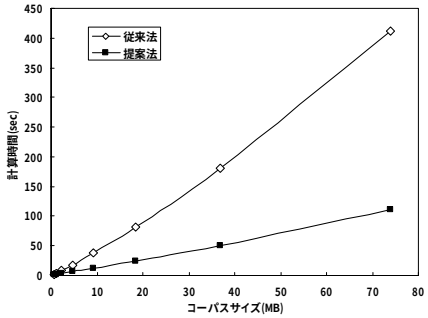


図8　インデックス構築時間

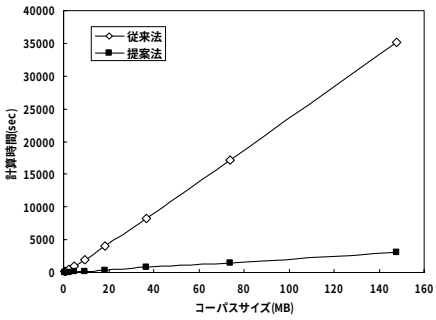


図9　用語抽出時間

従来法と比較して、高速に実行されていることが分かる。

６．考察

提案法では、従来法に比べて情報を制限したにもかかわらず、従来法よりも高い抽出精度を得ることができた。この理由として、まず従来法で利用されている経験的なしきい値を減らし、さらにこれらを学習コーパスの値で自動的に決定する処理を導入したことによる効果が考えられる。また、従来法では分割しがちであった複合語を多く抽出している点も挙げられる。さらに、利用する情報を Bigram 統計量のみに制限していても、多くの技術用語を構成する Bigram の反復度

がすべて高い値を持つことから、Bigram の統計量のみでも十分な用語認定が行えたと推定される。

しかし、他の形態素解析を行う用語抽出法と比較すると、得られた F-measure は低い値となっている。[5]では、TMREC コーパスを用い、用語抽出方式の評価を行っており、精度 0.46、再現率 0.83、F-measure0.59 という値が得られている。本稿で提案した手法では、現在 F-measure で 0.40 程度であり、これらの手法に及んでいない。この主な原因としては、語分割のプロセスにおける誤りの発生が挙げられる。

提案手法では Bigram の統計量のみを利用して語分割及び用語抽出処理を行っている。本研究における目的は用語抽出であり、語分割プロセスにおいて正確な形態素を認定することを目的とはしていない。しかし、用語を過度に分割している例や、分割すべき点で分割されない例が見られており、語分割プロセスの精度向上が必要である。統計的な情報のみを用いた形態素解析手法として、[9]][10][11]などがある。[9]では、可変長 n グラムモデルによる日本語単語分割手法が提案されている。また、[10]では、単語分割問題を分類問題として定式化し、アダブーストを利用した語分割手法を提案している。これらの手法は辞書情報に依存しないため、未知語に対する汎用性が高いと考えられる。今後これらのような統計的形態素解析技術の利用を検討したいと考える。

用語抽出の観点からは、複合語の抽出が重要な問題である。複合語の抽出に関する研究はさまざまなものがあるが、辞書を使わない手法はあまり存在しない。本稿で提案した手法は、抽出精度の観点からは辞書を使った手法に及ばないが、言語の知識を直接的には利用しておらず、汎用性の観点から優れているといえる。提案法は言語知識を用いていないため、新語に対して柔軟に適応することが可能である。

また、Kageura らは、専門用語の性質としてターム性とユニット性を挙げている[12]。ターム性とはある言語的単位の持つ分野固有の概念への関連性の強さである。また、ユニット性は語同士の文法的・意味的なつながりの強さである。提案法では、反復度を利用した語分割及び用語抽出を行っている。反復度は重要語に対して非常に高い値を持つため、ターム性の尺度として有用であると考えられる。しかし、ユニット性の観点からは、特に Bigram の反復度のみの利用では、不十分であり、語同士の結合性を捕らえる他の尺度が必要だと考えられる。

７．おわりに

本稿では、Bigram の反復度を利用した技術用語抽出法を提案した。利用する情報を Bigram の統計量のみに制限したにも関わらず、技術用語抽出に関する実験では、全部分文字列の統計量を利用した従来法に比べて、同等以上の抽出精度を実現した。また、コーパス中に存在する Bigram の総数の増加は、コーパスサイズの増加に対して緩やかであり、巨大なコーパスに対しても適用可能である。よって、従来法に比べて理論的なスケーラビリティを大きく向上することができた。

今後の課題として、統計的形態素解析を利用した語分割などによる精度向上などが挙げられる。

参考文献

[1] 武田善行, 梅村恭司: キーワード抽出を実現する文書頻度分析, 計量国語学会, Vol.23, No.2, pp.65-90, 2001

[2] Yoshiyuki Takeda, Kyoji Umemura: Selecting Indexing Strings Using Adaptation, In Proceedings of the 25th Annual International ACM SIGIR Conference, pp.427-428, 2001

[3] Kenneth W. Church. Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to p/ 2 than p^2, In Coling-2000, pp.180-186, 2000

[4] Hiroshi Nakagawa, Tatsunori Mori: A Simple but Powerful Automatic Term Extraction Method, 2nd International Workshop on Computational Terminology, COLING-2002, pp.29-35, 2002

[5] Hiroshi Nakagawa: Automatic Term Recognition based on Statistics of Compound Nouns, A: Terminology, vol. 6, pp. 195-210, 2000

[6] Toru Hisamitsu, Yoshiki Niwa: A Measure of Term Representativeness Based on the Number of Co-occuring Salient Words, Coling-2002, 2002

[7] Masao Utiyama, Masaki Murata, Hitoshi Isahara : Using Author Keywords for Automatic Term Recognition, Terminology Vol. 6, No. 2, pp.313-326, 2000

[8] 梅村恭司, 真田亜希子 : 文字列を k 回以上含む文書数の計数アルゴリズム, 自然言語処理, pp. 43-70, 2002

[9] 小田 裕樹, 北 研二 : PPM* 言語モデルを用いた日本語単語分割, 情報処理学会論文誌, Vol. 41, No. 3, pp. 689-700, 2000.

[10] 新納浩幸, 決定リストを弱学習器としたアダブーストによる日本語単語分割, 自然言語処理, Vol.8, No.2, pp.3-18, 2001

[11] Rie Kubota Ando, Lillian Lee. : Mostly-Unsupervised Statistical Segmentation of Japanese Kanji sequences. Journal of Natural Language Engineering, vol.9, 2003.

[12] Kyo Kageura, Bin Umino : Methods of Automatic Term Recognition : A Review, Terminology, vol.3, no.2, pp.259-289, 1996