キーワード自動抽出システム「言選 web」

前田朗（まえだあきら）

はじめに　～「言選 Web」へようこそ～

文章中の重要なキーワードをあらかじめ示してくれれば、概要をすぐにつかめるのにと思ったことはないであろうか。また、複数の文章を情報工学的に比較したいと思ったことは？「言選 Web」は文章からその概要をつかめるキーワードを取り出す Web 上のサービスである。これは、日本語のみならず、中文、西ヨーロッパの各言語（英語など）にも対応している。

「言選 Web」自体は Web アプリケーションであるが、そのエンジン部分や、Windows 用プログラムをフリーソフトとして配布している。Perl モジュール"TermExtract"、Windows アプリケーション"termex"、テキストマイニングツール"termmi"からなる、まさに「言選 Web」ファミリーともいうべき一群のソフトウェアである。

本稿では「言選 Web」をはじめとするこれらの専門用語自動抽出システムについて、基礎理論から活用法までを紹介する。

第一章と第二章は、基礎理論である。第三章と第四章では、専門用語自動抽出システムを実際に使いこなすための情報を提供する。もし、「言選 Web」の活用法だけを知りたければ、第三章からお読みいただくこともできる。

読後は「言選 Web」(http://gensen.dl.itc.u-tokyo.ac.jp/)にアクセスし、その有効性をぜひ確かめていただきたい。

第一章　用語抽出手法あれこれ

「言選 Web」では、１）文章から用語を抽出し、２）重要性の高い順に並べかえる、という２ステップの処理を行っている。ここでは、最初のステップである用語抽出をみていくことにしよう。

「言選 Web」における用語の抽出手法は大きく２種類にわかれる。ひとつは文章をいったん形態素（語の最小単位）まで分割して、その上で断片をつなぐように用語を組み立てる手法。そして、もうひとつは用語になりえない単語（もしくは文字）を消去し、残ったものを取り出す手法である。

形態素解析ソフトと用語の「まとめあげ」

まずは、形態素から用語にまとめあげる方法を説明しよう。形態素は前述のように語の最小単位のことである。日本語では文章を形態素に分割するソフトとして、茶筅、和布蕪がある。「わかち書きパッチ」を当てた案山子は文章から単語への分割を高速に行えるが、出力が形態素とは限らない。

試しに「漢字文献情報処理研究」という語を形態素解析ソフトにかけてみよう。和布蕪では次の４つの形態素に分解される。

漢字（名詞,一般）、文献（名詞,一般）、情報処理（名詞,一般）、研究（名詞,サ変接続）

キーワードとして「漢字文献情報処理研究」が一語で出て欲しいのに、これでは用語の単位として小さすぎる。形態素解析ソフトは、複合語に対応したものではないといえる。

そこで「言選 Web」では個々の形態素を用語レベルにまとめあげる。その基本ルールは、名詞の形態素（単名詞）が連続した場合に、それをまとめて複合名詞とみなすことである。

上記の例では、どの形態素も全て名詞である。よって全ての形態素を連結でき、「漢字文献情報処

理研究」と、一語にまとめあげることができる。

英文の場合は、既に単語の区切りが行われているので、品詞のタグ付けがされていればよい。この品詞のタグ付けを行うのが、POS Tagger という種類のソフトである。英語の場合は、Brill's Tagger という POS Tagger をフリーで入手できる。まとめあげのルールは日本語と比べて複雑になるが、基本的な考えは同じである。

中文では"ICTCLAS"を使う

中文では、中国科技院の Windows ソフト"ICTCLAS"で、品詞分割と品詞タグ付けを行える。「**计算所汉语词法分析系统**」とあるが、これは形態素解析ソフトと同義であろう。なお、Web 版もあり、以下の URL にて試すことができる。

http://mtgroup.ict.ac.cn/~zhp/ICTCLAS.htm

実際に、「ICTCLAS 的介绍及说明」を処理した結果は次のとおり（/nx などは品詞情報）である。

ICTCLAS/nx　的/u　介绍/vn　及/c　说明/v

さて、中文の場合は各形態素を次のとおりまとめあげる。例外の少ないシンプルなルールであるが、人民日報で試したところ、人手で指定した用語の 50%強を取り出すことができた。

- 名詞に類する語(ng n nr ns nt nz nx vn an i j) *以後「名詞」
  → 名詞、形容詞、助詞、后接成分、連詞（和、与）に結合する。複合語の先頭になる。
- 形容詞(ag, a)
  → 形容詞、助詞、后接成分、連詞（和、与）に結合する。複合語の先頭になる
- 助詞(u), 后接成分(k)
  → 名詞、形容詞に結合する
- 連詞(c)
  → 和,与の場合のみ。名詞に結合する。
- 区別詞(b)
  → 名詞, 助詞、連詞（和、与）に結合する。複合語の先頭になる

このルールで「ICTCLAS/nx　的/u　介绍/vn　及/c　说明/v」を試していただきたい。「ICTCLAS 的介绍」と用語抽出されるはずである。

カタカナと漢字熟語のみ抽出すると

日本語の場合、カタカナと漢字熟語の並びを抽出するだけでキーワード候補になりえてしまう。これは、キーワードの多くは名詞であることから、名詞候補を取り出したとも考えられる。

たしかに、カタカナは外来の名詞に使われることが多い。また、漢字熟語は文章中の使われ方によらず、熟語部分だけ切り出すと名詞としても扱えてしまう。次の例を見てみよう。

『日本語の場合、カタカナと漢字熟語を抽出するだけでキーワードになり得る。』
↓
（カタカナと漢字熟語のみ抽出）
↓
『日本語　場合　カタカナ　漢字熟語　抽出　キーワード』

特に「抽出する」が「抽出」の部分だけみると名詞扱いできることを確認いただきたい。上記の例では「場合」がノイズになるが、それ以外はキーワードらしき用語を抽出できていると思う。

文章を特定の語で分けてみる

情報検索システムでは、キーワードの自動切り出しを行う際に、キーワードとしてふさわしくない語を、あらかじめストップワードとして指定する。もし、キーワードになりえない語を文章からはずしていけば、残った語はキーワードとなる可能性が高いといえないであろうか。もちろん、ノイズは多いが、それは後述の重要度順の用語並べ替えにより、判別できる。

例えば、"I have a chinese journal"という英文であるが、情報検索システムで使われるストップワードを除いていくと、"chinese journal"だけが残る。シンプルだが有効に働くことがわかるかと思う。

中文に関しては、「言選 Web」オリジナルのストップワードリストを作成した。「動詞」となりえる形態素を全てストップワードにするなど機械的

な方法も試したが、人手により選定したストップワードがより効果的であった。中国語は品詞種別や活用、時制が見た目には区別できないことが多い。実際の文中では[述語][目的語]という並びでも、抽出されたものが十分キーワードとして通用する。つまり英文の場合"write a letter"という語がキーワードとしては不適格で"writing a letter"となればキーワードとして適格となるが、中文ではどちらも“写信”ということである。文章の内容が手紙を書くことに関してであれば文章中の品詞に左右されず、"写信"はキーワードとして適格といえる[2]。

中文では、文章中の品詞のキーワード抽出に与える影響が、日本語や西欧言語に比べて弱い。逆に言えばストップワードによる用語抽出の特性を生かせる言語であるといえる。

第二章　大事な用語から並べてみよう

オーソドックスな TF-IDF 法

重要なキーワードをランク付けするオーソドックスな手法が、TF-IDF 法である。TF は Term Frequency、IDF は Inverted Document Frequency の略である。これは、ある文献中に多く出てくる用語は重要だが、一般的な語は除外するという考えによる。文献中に多く出てくる語は出現回数を数えれば求められるが、一般的な用語を低くランクづけるのはどのように行っているのであろうか。

たとえば用語「漢字情報」が全 8 つの文献中、3 つの文献に使われていたとする。この情報を元に確率の考えを使えば、「漢字情報」がどのくらい一般的な用語かを示せる。今回の例の場合は、3/8 である。この確率が低い、つまり小さい値ほど一般的な用語と判断できる。しかし、用語の出現頻度では大きい値ほど重要性が高いため、このままでは両者を掛け合わせることができない。

そこで、確率の逆数を用いる。先の例では確率が 3/8 だが、これを反転させて 8/3 にするということである。しかし、このままでも十分ではなく、100 万件中 1 件も、100 万件中 2 件も、一般的ではない用語として大差がない。しかし、得られた数値としては 2 倍になってしまう。その調整として対数の計算を行う。また、全ての文献に含まれる用語は、対数の計算を行うと 0 になる(*log(1)=0*)ため、その調整のために 1 を加える。

最後に式の形でまとめると以下のとおりである。後述の termmi ではTF-IDF 法もサポートしている。

TF-IDF の重要度　=
用語の出現頻度 x (log (総文献数　/ 該当の用語を含む文献数) + 1)

FLR が「言選 Web」の基本

次に「言選 Web」がメインで用いている FLR を紹介する[1]。用語は単名詞そのものか複数の単名詞を組み合わせて作られる。この理論では、他の単名詞と連結して複合語をなすことが多い単名詞ほど、文書中で重要な概念を示すと考える。

簡単な例で、「漢字文献情報処理」を考える。この語は次のとおり 5 つの単名詞に分割できる。この際、それぞれの単語が他とどれだけ結びつくかを文章中から統計をとり、次のとおりわかったものとする。

単語	前の語に連接	後の語に連接
漢字	2	3
文献	3	4
情報	4	5
処理	2	0
研究	0	3

用語の重要度はこれらの 10（単名詞 x 2）の数値の平均から求める。平均値は、相乗平均が相和平均よりもより効果的なため、「言選 Web」では相乗平均を用いている。なお、相乗平均をとる際に、連接した回数が 0 回の単名詞に対応するため、実際には各回数に 1 を加算した値を用いている。

こうして得られた単名詞の連接情報に、用語の出現頻度をかけたものが「言選 Web」の重要度である。出現頻度(Frequency)に左(Left)と右(Right)の語の連接情報を組み合わせて使うため、これを FLR と呼ぶ。

中文は「文字」も使える

漢字は一字一字が意味を持つ「表意文字」とし

て使われることが多い。日本語の場合は形態素、英文の場合は単語を、用語の最小単位とし重要度の計算に用いた。しかし、中文の場合は、用語の最小単位として形態素と文字の２つが考えられる。

例えば、「漢字文献」を次のように重要度計算できるのではないであろうか。(数値は仮の値)

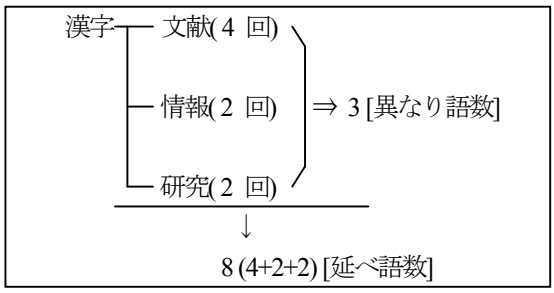
文字	前の字に連接	後の字に連接
漢	6	7
字	3	6
文	2	5
献	4	3

中文版「言選 Web」では実際に、文字と形態素、それぞれに着目した重要度計算のモードを用意している（停止语方式版と ICTCLAS 版）。それぞれの効果の違いはこれからの研究課題である。

「多様性」（パープレキシティ）でランクづける

単名詞が他の単名詞に連接した回数のカウントにはいくつかの方法がある。直感的にわかりやすいところでは語の延べ数や種類数である。これに換えて情報理論的な回数、すなわちパープレキシティを使うのが、東京大学情報基盤センター中川研究室における最近の研究理論である。

例えば、次の例で考えてみる。出現回数（延べ語数）だと 8 回 (4+2+2)、種類数（異なり語数）だと 3 回（「文献」「情報」「研究」の 3 種）である。

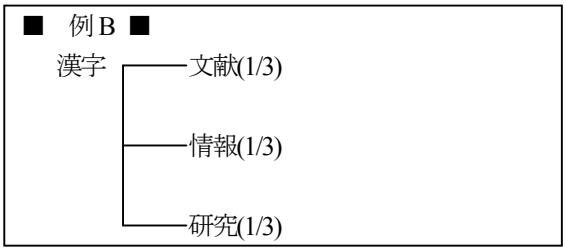


この回数のカウントでパープレキシティを使うと、たとえ連結する語の種類数が多くとも、特定の語とばかり結びつけば、カウントが少なくなる。語の連接が多くの語に分散していれば、カウントが多くなる性質を持つ。

「パープレキシティ」が初耳でも「エントロピー」ならご存知のかたも多いかと思う。情報理論でいう「エントロピー」は、「情報を平均して何バイトで示せるか」、つまり情報の多様性を示す指標である。パープレキシティはエントロピーを 2 のべき乗した数値のことであり、同じく情報の多様性を示している。

パープレキシティは確率の考え方を使う。例えば先のケースを割合で示してみると次（例 A）になる。

この確率の分布が平均しているほど、パープレキシティが増大する。例えば、以下の例 B は例 A よりも多様性があるということである。



文章をどんどん「学習」させる

FLR における用語の重要度の計算は、単名詞がどれだけ連接したかの統計情報を必要とする。この統計データが正しければ正しいほど、計算の精度は上がるといえる。

文章を読み込むたびに、この統計データを蓄積し、次回以降の FLR の計算に生かす機能を与えれば、統計データひいては LR の精度が良くなっていくのではなかろうか。その考えから「言選 Web」のエンジン部分、"TermExtract"では学習機能をオプションとして実装している。

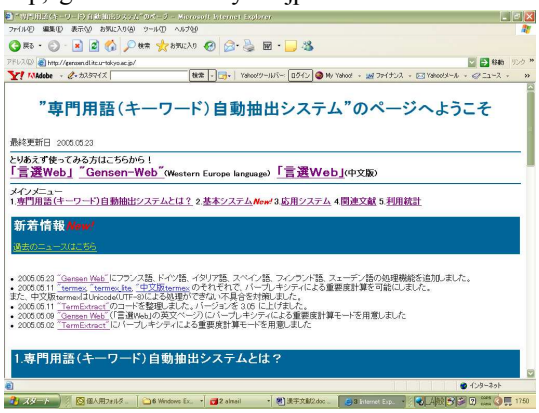
ただし、雑多な分野の文章を学習させると、あまりに一般的な用語の重要度が高くなるという弊害もある。あくまで特定の分野に限って「学習」させていただきたい。なお、ユーザが不特定多数の「言選 Web」では採用を見送っている。

第三章　"言選 Web"を使ってみよう

まずは"専門用語自動抽出のページ"にアクセス

まずは"専門用語自動抽出のページ"(以下のURL)にアクセスしてみよう。

http://gensen.dl.itc.u-tokyo.ac.jp



このページでは、「言選 Web」をはじめとする専門用語(キーワード)抽出システム全体について、案内を行っている。はじめての方は、このページから必要な情報やソフトを探して欲しい。

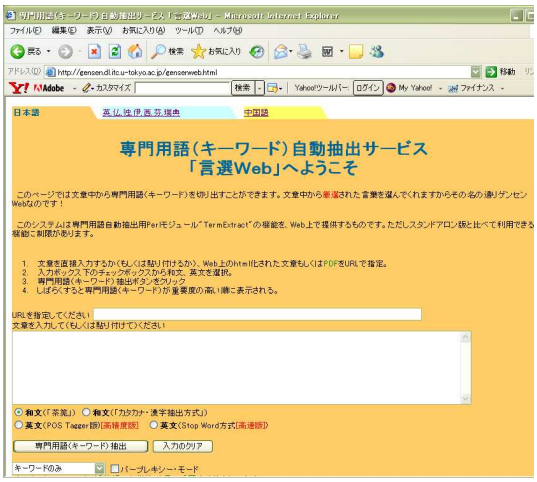
また。新機能の追加や、バグレポートの掲示も頻繁に行っている。言選 Web を既に使ったことのあるかたも、たまにアクセスしていただきたい。

「言選 Web」は準備不用

「言選 Web」は文章中から専門用語（キーワード）を自動抽出する Web 上のサービスである。使い方は次のとおりいたって簡単である。

1）URL 入力欄に解析を行う Web ページの URL を入れるか、テキストボックスに解析対象の文書を貼り付ける
2）専門用語（キーワード）自動抽出」ボタンをクリックする、
3）文章中の用語が重要な順に表示される。

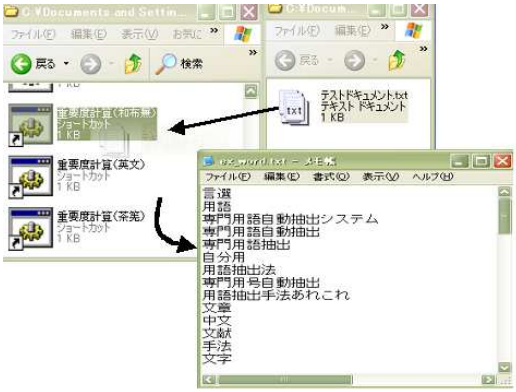
「言選 Web」には和文版のほかに、中文、西ヨーロッパ言語版も用意している。これは上部のタブメニューで切り替えを行える。インターネットが利用できる環境なら、準備は不要である。「言選 Web」の機能をぜひお試しいただきたい。



termex なら自分専用

「言選 Web」は不特定多数の人にサービスするため、いくつかの制約を設けてある。この制約なしで使いたいという要望にこたえるのが、Windows 用ソフト"termex"である。これには、通常の和文版と中文版、また和文の簡易版である"termex lite"がある。

使い方は処理対象のテキストファイルを termex の実行用アイコンにドラッグするだけでよい。Windows の「メモ帳」が起動し、用語抽出の結果が表示される。



Web ページを主に扱うなら、Internet Explorer との連携機能を試してみよう。Web サイト内の複数のページを蓄積した上で、用語を抽出できる。

また、"termex"は重要度計算パラメータの設定や、学習機能を利用できるなど、個人ユースに特化してある。

termmi で文章をまとめて解析

「言選 Web」のキーワード抽出機能をテキストマイニングに応用したのが、"termmi"である。テキストマイニングとは、大量のテキストデータから隠れた知識を発見するための手法の総称である。termmi はその中でも「手軽」「シンプル」「無料」であることに特色がある。

Windows のフォルダの中にテキストファイルを複数いれ、"termmi"のアイコンにドラッグすることでフォルダ内の全テキストファイルの解析が行われる。全テキストで共通な用語、個々のテキストのユニークな用語、個々のテキストの用語抽出結果、形態素解析結果を見ることができる。

また、別に用意したベクトル空間法による類似度計算スクリプト(vector_space.pl)を使うことで、文献群の中で特徴的な文献を数値的に判断できる。

termmi は、統計的な手法を駆使する学術的なテキストマイニング研究とは異なる方向性である、しかし、手軽でシンプルなだけに一般の利用にはむしろよいのではないかと考えている。

第四章　"言選 Web"を使いこなす

「言選 Web」 (Web によるサービス)

いくつかの制限を理解しよう

「言選 Web」は不特定多数のユーザが利用する。そのため、やむなく機能を制限している部分がある。不便を感じるかたもいるかもいるかと思うが、どうかご理解いただきたい。

A.一定時間過ぎると処理を中断
B.データ量が大きい場合は、処理を中断。
C 動的 Web ページへのアクセスを禁止
D.同時アクセスの禁止

目的にあった用語抽出法を選ぼう

「言選 Web」の利用法は簡単だが、オプションがいくつかあり、どれを選んだらよいか悩むところかと思う。以下は開発者からのお薦めである。

A. 日本語であれば、まずは「茶筅」版を選び、よい結果が出なければ、「漢字・カタカナ抽出方式」を試す。
B.英語の場合は、"POS Tagger version" がデフォルトだが、多量の文章では「制限時間オーバー」を起こしがちである。大量のデータの場合は、"Stop　Word　version"を選びたい。
C.中文の場合は、ストップワード版とは別 WEb ページになるが ICTCLAS 版もある。通常のストップワードで満足できない場合は、少し手数がかかるが、ICTCLAS 版もお試しいただきたい。

下準備が必要な中文 ICTCLAS 版

「言選 Web」はサーバ内部に形態素解析ソフトを組み込んでいる。そのため、[形態素解析]→[用語まとめあげ]→[重要度ランク付け]の一連の処理をユーザが意識することはない。

しかし、中文の形態素解析ソフト ICTCLAS だけは「言選 Web」サーバ内に組み込めなかった。そのため、事前に中文を ICTCLAS で処理し、その結果(ICTCLAS タグ付け済みのテキスト)を「言選 Web」のテキストボックスに貼り付けて使う必要がある。

パープレキシティモードはどこが違う？

パープレキシティモードは「情報の多様性」をもとに用語の重要度ランクづけを行う。抽出される用語自体は同じだが、その並び順や重要度の値が異なるということである。

東京大学情報基盤センター中川研究室による最

近の研究では、パープレキシティモードが通常のFLR よりも優れた結果を示した。「言選 Web」ではオプション扱いであるが、かなり優れた性能を示すはずである。ぜひ、お試しいただきたい[3]。

TermExtract (Perl モジュール)

サンプルスクリプトから自作スクリプトへ

"TermExtract"は「言選 Web」のエンジン部分をなす Perl モジュールである。そのまま使えるサンプルスクリプトも付属している。

このサンプルスクリプトを参考に、自分専用のスクリプトを作成してみよう。自作の Perl スクリプトの中に「専門用語（キーワード）自動抽出機能」を組み込むことすら容易に実現できるはずである。

自分用の追加モジュールを作る

TermExtract では、「茶筅」や英文テキストファイルなど、さまざまな入力形式のデータに容易に対応させるため、入力のデータ形式依存部分と重要度計算部分でモジュールを切り離して作っている。

このデータ形式依存の部分と、重要度計算部分(Calc_Imp.pm)のデータ入出力仕様さえご理解いただければ、TermExtract の追加モジュールを自作できる。例えば、「アラビア語対応モジュール」など現在は、"TermExtract"がサポートしていない言語への対応も可能であろう。また、第一章で述べた用語切り出し方法を自分の好みに変更することもできる。詳細な仕様を Web でも公開しているので、Perl を使い慣れたかたなら、自在に改良ができると思う。

termex (Windows 専門用語抽出)

各種パラメータを変更してみよう

termex の Perl スクリプトをエディタ（Windows の「メモ帳」など）で開き、パラメータを変更できる。その説明はスクリプト中のコメント欄に詳しいが、以下にまとめてみた。

A. 重要度計算で、連接情報の重要度計算のモードを選択できる。可能なモードは、連接語の"延べ数"、"異なり数"、"パープレキシティ","連接情報を使わない"のいずれかである。

B. 連接情報と組み合わせる頻度を Frequency, と TF(Term Frequency)のいずれかから選ぶことができる。TF は用語が他の用語の一部に含まれていた場合もカウントするが、Frequency はカウントとしない。例えば「情報システムと情報」の場合、TF, Frequency のカウントは次のとおりである。

「情報システムと情報」
TF →
「情報」2 回, 「情報システム」1 回
Frequency →
「情報」1 回, 「情報システム」1 回

上記 A の"連接情報を使わない "設定がなされていれば、頻度情報のみ出力できる。

C. 重要度計算で、「ドキュメント中の用語の頻度」と「連接語の重要度」のどちらに比重をおくかを設定する(デフォルト値は１)。値が大きいほど「ドキュメント中の用語の頻度」の比重が高まる。

D. オプションの学習機能を使用できる。デフォルトは、"使用しない"である。

学習機能をうまく使うには

学習機能は雑多なテーマの文献を学習させてしまうと、一般的すぎる語が上位にきてしまう。常に同じテーマの文献に限って使用していただきたい。

また、学習機能を使い続けると、重要度中の頻度と連接情報の重みが変わっていってしまう。そのため、「重み付けの割合を頻度側に戻す」などの手当てが必要になる。

termmi (テキストマイニングツール)

どの文献が似ているかを調べてみる

"termmi"には、「ベクトル空間法」による類似度判定用スクリプト vector_space.pl が付属している。「ベクトル空間法」では TF-IDF の重要度を使う方法がオーソドックスであるが、termmi では

FLR による重要度を利用している。なお、オプションとしてオーソドックスな TF-IDF の処理モードも用意してある。

termmi では文献群全体の処理結果と個々の文献を比較する。もし特定の文献との比較をしたければ OUTPUT フォルダ中の任意のファイルで SUMMARY フォルダ中の total.txt を上書きし、その上で vector_space.pl を実行いただきたい。

FLR の重要度と、ベクトル空間法の組み合わせは、研究で実証されていない機能である。読者のかたにも評価いただければ幸いである。

どの形態素が隣り合うかを確かめる

"termmi"では、どの単名詞が連接しているかの情報を、学習用データベースに保存している。これは､次回の処理を行うまでPC 上に残っている。この学習用データベースは、付属スクリプト get_stat.pl を使うことで内容を見ることができる。単語の連接は単語バイグラムモデルにも使われる統計データである。termmi の場合は「キーワード関係のみ」の単語の連接に限定されるが、研究目的にもご活用いただけるかと思う。

おわりに　～さらなる展開に向けて～

「言選 Web」は平成 15 年 4 月 23 日から一般に公開を始め、月間アクセス件数を数百から多いときには数千まで伸ばしてきた。インターネット・エクスプローラーのお気に入りに登録された件数は、3,000 に達しようとしている。インターネット上での評判もおおむね好評といえそうである。

「言選 Web」の機能はキーワードを抽出するだけというシンプルなものである。だからこそ、その応用範囲は広いといえる。開発側としては、メタデータベースにおけるキーワードの選定、言語学の研究、Web サイトの解析などの利用を想定しているが、それ以外の用途にも活用いただければ喜ばしい限りである。

「言選 Web」は発展していくサービスである。公開から 2 年あまりがたった今も、新たな研究理論の導入や、テキストマイニングへの応用、多言語対応など、さまざまな手当てを行ってきた。今後とも、研究と実用サービスへの橋渡しとして、改良を続けていきたい。

謝辞

東京大学情報基盤センター図書館電子化部門の中川裕志教授には、当システムの構築にあたり、専門用語抽出理論の実装、システムの仕様の検討などにおいて多大なる助力をいただきました。

また、東京大学経済学部資料室の小島浩之助手のメタデータベース（東京大学経済学部サブジェクトゲートウエイ Engel）での「キーワード半自動付与」構想がなければ「言選 Web」は生まれなかったといえます。さらに中文版「言選 Web」においても、小島浩之助手を抜きにしては語ることができません。

お二人に深く感謝いたします。

参考文献

[1] Hiroshi Nakagawa, Tatsunori Mori: "Automatic Term Recognition based on Statistics of Compound Nouns and their Components", Terminology, Vol.9 No.2, pp. 201-209, 2003

[2] Hiroshi Nakagawa, Hiroyuki Kojima, Akira Maeda, "Chinese Term Extraction from Web Pages Based on Compound word Productivity", 42nd Annual Meeting of the Association for Computational Linguistics (ACL2004), Third SIGHAN Workshop on Chinese Language Processing, pp.79-85, Barcelona, Spain, July, 2004

[3] 森山聡, 吉田稔, 中川裕志 “複合語のパープレキシティに基づく重要語抽出法の研究 “, 言語処理学会第 11 回年次大会発表

[4] 前田朗、小島浩之, 中川裕志 “「言選Ｗｅｂ」の世界,” 図書館の窓, vol.43 No.3 pp.61-65 http://www.lib.u-tokyo.ac.jp/koho/kanpo/vol43/vol43-3.pdf

[5] 小島浩之、前田朗、 "キーワード（専門用語）自動抽出システムの構想とその展開", 第 51 回日本図書館情報学会研究発表要綱, pp.17-20, 2003

[6] Xiayan Zhang et al "Chinese Named Entity Recognition with Hybrid Statistical Model". Web technologies Research and Development-APWeb 2005: 7thAsia-Pacific Web Conference Shanghai, China, March 29-Aprill, 2005: Proceedings. PP.901-912, 2005

[7] 新田義彦 “正規表現とテキスト・マイニング”, 明石書店, 2003