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

小島浩之（東京大学経済学部資料室） kojima@e.u-tokyo.ac.jp
前田朗　（東京大学経済学部図書館） maeda@lib.u-tokyo.ac.jp

インターネット上の情報資源に対するメタデータ作成を容易にするため、文章中からキーワードを抽出するシステムを構築した。東京大学情報基盤センター・中川研究室で研究開発されていたシステムを、主に実用面から再設計・再構築したものである。その結果、１）Webサービス「言選Web」、２）Perlモジュール"TermExtract"、３）Internet Explorerとの連携で動作する"termex"の公開に至った。本論ではこのシステムの機能と事業展開について述べる。

はじめに

現在、さまざまな図書館、研究機関でインターネット上の情報資源へのポータルサイト構築が進み、多くがDublin Coreの定義に準拠したメタデータを採用している。実際にメタデータの作成を体験すると、重要なアクセスポイントの一つとなるキーワード付与は意外と難しく、手間をとられることが分かる。また、個々の作業者におけるキーワードの採用基準にばらつきがでることも懸念された。それには、インターネット上の情報資源からキーワード候補を自動抽出するツールが有効である。

本論では、上記ツールの機能に加え、大学の研究成果を実用向けのサービスに展開するにあたり検討した事項に焦点を当てて述べていくことにする。

１．システム作成の経緯

東京大学経済学部資料室では、政府機関及び地方公共団体等の統計類、白書類を収集しているが、媒体が紙からインターネットに移っていく傾向にある。現在、今後ともそれらの情報を資料室で提供し続けるために、独自のメタデータベースの構築を進めている。

このメタデータベースの入力支援のため、インターネット上の情報資源からキーワード候補を抽出するツールを作成することにした。

この目的に沿うシステムとして、東京大学情報基盤センター・中川研究室と横浜国立大学・森研究室が研究し、日本語Perl（JPerl）にて実装し配布している「専門用語自動抽出システム」[1][2][3]に注目した。これは、文章を形態素解析ソフトウエアにより解析した結果から、１）複数の単語からなる専門用語を抽出し、２）その文章中における重要度の計算を行うものである。このシステムをメタデータ入力支援に流用するには、多少のカスタマイズが必要である。カスタマイズを容易にできないシステム設計に汎用システムとしての問題点がある。

そこで上記システムを開発した中川教授の協力のもと、「専門用語自動抽出システム」の構成を修正、学習機能の付加や高速化も行い、全面的（げんせん）に作り直した。これが本論で紹介する「言選Web」を中心としたシステムである。

本システムは元にしたシステムに比べて、完全なモジュール化（他システムに組み込み可）をはじめ、移植性・拡張性・メンテナンスの容易さの向上及び多機能化を図っている。また、パッケージとしての展開を考え、１）Webによる専門用語自動抽出サービス「言選Web」、２）専門用語抽出のためのPerlモジュール"TermExtract"、３）Internet Explorer(IE)と連携して専門用語を抽出する"termex"の3つのシステム構成とした。

作成したシステムはメタデータの入力支援のみならず、研究やビジネスでの利用も期待できることから、東京大学情報基盤センターで一般向けにサービスの提供とソフトウエアの配布を行なっている（http://gensen.dl.itc.u-tokyo.ac.jp）。

また、東京大学経済学部図書館では、実際に本システムを使い、前述のメタデータの入力作業に入っている。

２．専門用語自動抽出の基本原理

今回作成したシステムでは、専門用語を単語もしくは複合語から生成する。複合語を構成する最小単位の名詞を特に「単名詞」と呼ぶ。

日本語の場合、「茶筅」などの形態素解析ソフトウエアでまず文章を形態素（品詞）に分割

する。次に分割した形態素のうち原則として単名詞どうしを連結する形で、専門用語を生成する。英文の場合は品詞タグ付けソフト(POS Tagger)を利用し単名詞を連結するか、特定のストップワードにより文章を分割することで専門用語を切り出す。

このように抽出した専門用語に対し、重要度のランク付けをする。本システムでは専門用語を構成する単名詞が他の単名詞と連接して複合語をなすことが多いほど、重要な概念を示すと考える。

簡単な例で、「情報科学技術」を考える。この語は次のとおり3つの単名詞に分割ができる。ここで、それぞれの単名詞が他の単名詞の前と後にどれだけ結びつくかという統計量が分っているとする（下図の数値は仮の値）。

単名詞	前への連接	後への連接
情報	1	2
科学	2	3
技術	1	1

複合語全体の重要度はこれらの6つ（単名詞数ｘ2）の数値の平均から求める。本システムでは相乗平均でとっている。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。

さらに、重要度に文献中の用語の出現頻度を乗ずることで補正をかける方法を採用している。

３．専門用語自動抽出サービス「言選Web」

「言選Web」は専門用語を自動抽出するWebサービスである。幅広く研究やビジネスにも使用してもらいたいと考え、極力シンプルな機能だけを持たせることにした。高度な機能を利用したい場合は後述の"TermExtract"か"termex"の使用を推奨している。

使用法は、１）処理対象のURLを指定もしくは、テキストデータを入力（もしくは貼り付け）、２）和文か英文（英文は2種）のいずれかのチェックボックスを選択、３）「専門用語抽出」ボタンをクリックするだけである。これにより、文章から抽出した専門用語の一覧が、重要度の高い順にWebブラウザ上に表示される。Webブラウザの機能を利用し、結果をパソコンのファイルに保存することもできる。

「言選Web」には英文のページ(http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_eng.html)も用意し、英語圏からも利用できるようにしている。

「言選Web」にはURL指定で他のWebサイトのデータを処理する機能を付けた。このため、他のWebサイトに不正なアクセスをできないよう設計する必要があった。検討した事項は次のとおりである。

（１）リンクしたWebページ

「言選Web」は指定したURLにあるHTMLのみ処理を行う。ただし、フレームに限り、HTMLのリンクをたどった処理を行うこととした。Webサイトのトップページを指定し、サイト内の情報をまとめて処理できることは、便利に思える。しかし、ロボット規約の遵守義務まで考慮すると扱いが難しくなるため、Webブラウザの範囲内のアクセスに留めた。

（２）動的ホームページ

動的に変化するWebページへのアクセスは、アクセス先の負荷が大きい場合があるため禁止した。

（３）同時実行制御

専門用語自動抽出処理が同時に複数実行されないように設定した。これは「言選Web」を踏み台にしたDoS攻撃を防ぐためである。

（４）制限時間

アクセス先に負荷のかかる要求を行っても、一定の時間で要求を中断するようにした。

４．Perlモジュール“ TermExtract”

専門用語を自動抽出するためのPerlモジュールである。元にしたシステムでは、他のシステムへの組み込みが難しかった。このモジュール化により様々な用途に使えることを目指した。実際にこのモジュールを使い、前述の「言選Web」と後述する"termex"の2つのシステムの開発に成功している。これらに加え、テキストマイニング機能をより充実させたシステムの開発も検討している。

このモジュールは広く配布することを考え、次の用件を満たすよう作成した。

（1）標準化

CPAN（総合 Perl アーカイブネットワーク）で標準のインストーラーを用意した。さらに、Perl の作者が推奨しているとおり、-w コマンドラインオプション、use strict モジュール、の両制約の中で動作するようにした。将来的には CPAN への登録も考えている。

（2）移植性

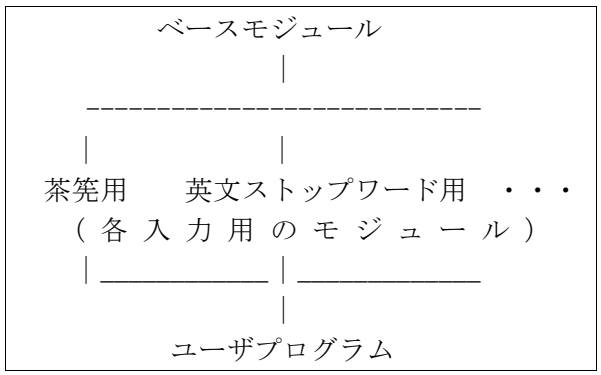
UNIX 環境、Windows 環境によらず動作する。また日本語パッチを当てた JPerl とオリジナルの Perl のいずれでも動作するようにした。

（3）オープン性

プログラム中にモジュールの仕様を書き込んだ上、Web ページでも公開した。プログラムのメンテナンスと第 3 者によるプログラムの改造を容易にするだけではなく、将来的に他のプログラミング言語（Java 等）への移植も考えてのことである。

（4）拡張性

モジュールは 2 階層の構成とした。重要度計算を行うモジュールを上位とし、形態素解析結果から専門用語を生成するモジュール、ストップワードによる専門用語を切り出すモジュールを下位におく設計である。ユーザプログラムは下位のモジュールを呼び出すことになる。（下図参照）



この構成により、例えば、ドイツ語用の処理を行いたければ、下位のモジュールのみを新規に作成すればよいことになり、拡張性が高くなる。実際に、日本語形態素解析ソフト「和布蕪」や、英文のストップワードによる処理といった新規の案件にも容易に対応することができた。

機能面で手を加えたのは、いくつもの重要度計算のオプションを選べるようにしたことである。これにより、ユーザが扱う文書の特性にあったカスタマイズを行うことができる。

また、インターネット上の情報資源を対象にテストを繰り返し、ノイズの少ない結果が出力されるよう調整した上、「茶筅」の複数バージョンへの対応（バージョンにより未知語の扱いが異なる）も行った。

中川教授の教示により追加した機能が、学習機能とストップワードによる英文処理である。

学習機能は、文章量が少ない場合でも、それまでに処理した文章の情報を生かして重要度計算を行うことにより、重要度計算の精度を高める機能である。学習機能を使用する設定にすると、それまでに処理した単語の連接情報が蓄積されていく。この情報は専門用語の重要度計算に用いるものであるが、言語学的にもこの統計情報は意味のあるものではないかと考えている。後述の"termex"にはこの統計情報を確認するプログラム、get_stat.pl も用意した。

また、ストップワードによる英文処理は、特定のストップワードで英文を分割することにより、専門用語を抽出するものである。POS Tagger を使う方式に比べて高速に動作する。現在、東京大学情報基盤センター・中川研究室では、このストップワード方式を使い、多言語化の研究を進めている

５．Windows プログラム“ termex”

Windows環境で、専門用語を抽出するためのシステムである。「言選Web」に比べて、次のメリットがある。

（1）他の Web ページへのアクセス制約がない。

（2）学習機能を利用できる。

（3）複数の Web ページを蓄積し結果を出せる。

（4）Perl スクリプト中のコメント行を解除することで、高度な設定ができる。

使い方には、IE との連携による方法と、Windows 上のテキストファイルをそのまま処理する方法がある。具体的には次のとおりである。

（1）IE との連携

１）IE で処理対象の Web ページを開き、マウス右ボタンのメニューから、「ソースの表示」を選ぶ。コマンドプロンプト（MS-DOS プロンプト）のウインドウが起動し、しばらくすると閉じる。
２）上記の１）の操作を Web サイト内の必要なページに対して行う。
３）重要度集計用のアイコンをダブルクリックする。Windows の「メモ帳」が起動し、専門用語が重要度の高い順に表示される。

（２）Windows のテキストファイル
１）Windows のテキストファイルを、重要度計算用のアイコンにドラッグ＆ドロップする。
２）Windows の「メモ帳」が起動し、専門用語リストが重要度の高い順に表示される。

なお、IE との連携では、IE の標準のソースエディタを別エディタに切り替えるフリーソフトの活用を考えついた。本来であれば Active X などの技術を利用すべきところかもしれないが、簡便な方法として面白いものであると考えている。

６．検索エンジン対策

作成したシステムを研究・ビジネス向けに一般広報するにあたり、図書館の Web ページからのリンクによるナビゲートだけでは難しいと考えた。専門用語の自動抽出は図書館のサービスとしては特異なため、研究者が目的を持って図書館のサイトを探すとは考えにくい。

そこで、多くのユーザ数を誇る２つの検索エンジン Google と Yahoo Japan について対応を行い、検索エンジンから研究者に見つけてもらうことにした。

まずは、「キーワード 自動抽出」というフレーズで Google の上位になることを目指し、2003 年 7 月時点で最上位に表示されるようになった。また、Yahoo Japan についてはディレクトリの登録申請を行い、2003 年 8 月に採用されている。

おわりに

図書系の研究部門の研究成果が実用的なサービスとなること、そのために図書館職員が実用的な観点から企画を考え進めていったことは、意義のあることではないかと考えている。

今後、図書系の事業として、このような研究支援のシステムの開発・運用も有望となるのではなかろうか。

謝辞

本システムの開発にあたり、多大な助言をいただいた東京大学情報基盤センター・中川教授に感謝いたします。

参考文献

[1] Hiroshi Nakagawa : ″Automatic Term Recognition based on Statistics of Compound Nouns″, Terminology, Vol.6 No.2 pp.195-210, 2000.

[2] Hiroshi Nakagawa, Tataunori Mori : ″A Simple but Powerful Automatic Term Extraction Method″, Computerm2 : 2nd International Workshop on Computational Terminology, COLING-2002 WORKSHOP, pp. 29-35, Taipei, August 31, 2002.

[3] 中川裕志、森辰則、湯本紘彰 : ″出現頻度と連接頻頻度に基づく専門用語抽出″, 自然言語処理, Vol.10 No.1, pp.27-45, 2003.