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）、前田朗（東京大学経済学部図書館）

現在、さまざまな図書館、研究機関でインターネット上の学術情報資源へのポータルサイト構築が進み、多くが Dublincore 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以外と難しく、手間をとられることが解る。そこで採録者の負担を軽減するために準備したのが本システムである。

本システムは、単なる文章の単語分割ではない。一般に文章中では複数の単語の組み合わせで複雑な概念を表す場合が多く、文章の内容が専門的な事項に特化すればその傾向はさらに顕著なものとなる。したがって文章中からキーワードを抽出する場合、単語分割機能だけでは意味を成さない。そこで、このシステムでは、（1）形態素解析プログラムによる単語分割、（2）複合語の作成、（3）文章中における重要度の計算、という3つのステップを踏むことで、複合語により複雑な概念を表すことが多い専門用語をキーワードとして文章中から抽出することに成功した。

専門用語の自動抽出にあたっては、東京大学情報基盤センター中川裕志教授、横浜国立大学環境情報研究院森辰則助教授の「専門用語自動抽出システム」を元に、中川教授の教示を受けつつシステムを再設計、再構築した。その結果、単にメタデータの入力支援用のカスタマイズにとどまらず、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 と連携して専門用語を抽出する"termex"の3つのシステムを作成し公開するに至った（http://gensen.dl.itc.u-tokyo.ac.jp/）。また、学習機能の付与や、オリジナル版 Perl への対応、英文の高速版の作成をはじめ新規に実現した機能も多い。このように本システムは研究部門と実務部門（図書館職員）の連携により、当初の業務用ツールの構想が新サービスの域にまで発展したものである。

本システムを利用することで、テキストドキュメント中から迅速に専門用語を抽出することができる。加えてオプションの学習機能を使うことにより、特定分野の専門用語抽出についてされに精度を上げることも可能であり、言語学的な研究にも利用可能だと考えている。