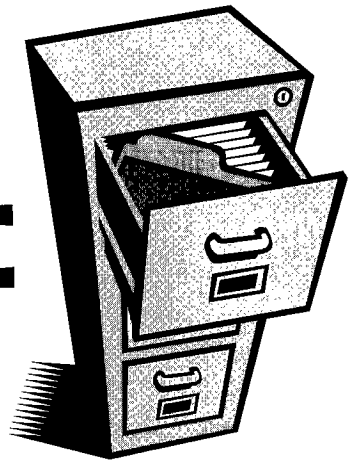
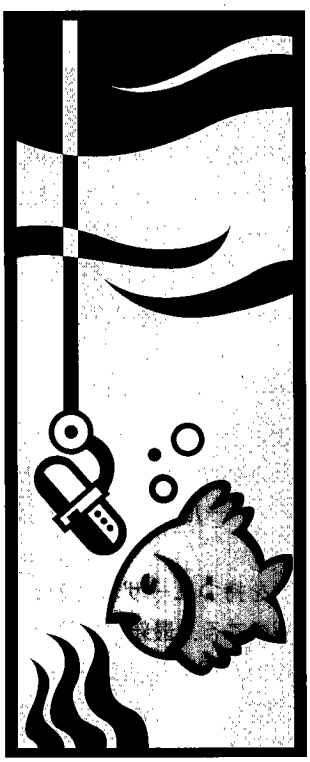
特集

使いやすくなった自然言語処理のフリーソフト



－知っておきたいツールの中身－



編集にあたって

使いやすくなった自然言語処理のフリーソフト
－知っておきたいツールの中身－

那須川 哲哉
日本アイ・ビー・エム（株）
東京基礎研究所
nasukawa@jp.ibm.com

久光 徹
（株）日立製作所
中央研究所
hisamitu@harl.hitachi.co.jp

李 航
NEC
情報通信メディア研究本部
h-li@cp.jp.nec.com

情報処理技術の発展が急速な勢いで文章を電子化している．書籍や各種報告書など従来紙に記述されてきた文章はもちろん，電子メールやチャットの普及により，従来口頭で伝達されていた内容までが電子データの形で計算機上に蓄積されつつある．その結果，読まなければならない膨大な文書の山に辟易している読者も多いのではないだろうか．

しかしこの流れは情報処理にかかわる者にとっては大きなチャンスである．この文書の山を何とかしたいという需要はビジネスチャンスであるし，電子化された膨大な情報の分析により学術的にもビジネス的にも有用な知見が得られる可能性がある．

ところが文書データの処理は一筋縄ではいかない．日本語の場合，文章が分かち書きされないため，単語を切り出すだけでも大変であるし，文字コードの問題もある．さらに，膨大な文書の中から必要な文書のみを取り出したり，編集加工したり，文中の単語と単語がどのような関係にあるかを調べたりといった処理は非常に複雑である．

それに対し，近年，多種多様な文書データに対応できる質の高い自然言語処理ツールがフリーソフトとして公開され，専門家でなくとも自然言語で書かれた文書データを簡単に処理できるようになってきた．

たとえば，日本語形態素解析ツールを用いれば，日本語の文章を意味のある単位に分割（「日本語」「の」「文章」「を」…というように）し，各要素に品詞を付けることができる．文書中の名詞一覧を作成するといった処理が簡単に実現できるわけである．さらに構文解析ツールを用いれば，言葉と言葉の（係り受けなどの）関係を含む文の構造を調べることができ，たとえば，文書中から，ある言葉を修飾しているさまざまな形容詞を取り出すことが可能になる．さらには，特定の言葉を含む文書を探し出すための検索ツールや，多言語の文書を自由自在に編集できるツールまでもフリーで公開されている．

これらのソフトは，日々増え続けるデータの処理に欠かせないインフラとなっており，情報処理システムの成功例といえる．本特集では普及度の高いいくつかの自然言語処理ツールを取り上げ，開発者たちから基本的な使い方と利用事例に加え，論文やマニュアルなどではなかなか知ることのできない開発の経緯，設計思想，開発のノウハウや苦労話，ツールを上手に使うコツなどを紹介してもらう．

第1編では，自然言語処理およびその関連分野で開発されているフリーソフトウェアを網羅的に紹介する．具体的には，形態素解析ツール，構文解析ツール，全文検索ツールなどのフリーソフトウェアの特徴，入手先，コア技術などについて述べる．

第2編では，現在広く利用されている形態素解析システム「茶筌」の開発経緯について紹介する．「茶筌」の現在のシステムに至る経緯の説明からなぜ「茶筌」が現在のようなシステムになったかを理解することができる．

第3編および第4編では，構文解析ツールを紹介する．第3編の「KNP」は日本語の，第4編の「Apple Pie Parser」は英語の構文解析システムである．各々で構文解析の難しさや，それを克服して精度を向上させた工夫が紹介されている．この2つのツールは対象言語だけでなく解析のアプローチなども異なっている．第4編ではインストール方法や「Apple Pie Parser」以外の英語向けのフリーソフトウェアも紹介されている．

第5編では，文書検索ツールとして，個人やウェブサイトで広く利用されている，キーワードによる全文検索システム「Namazu」について紹介する．その生い立ち・技術的背景から，インストール方法，利用例まで丁寧に解説する．

第6編では，多言語機能を持つユーザ環境である，最近のGNU Emacsについて解説する．最新の多言語入力環境の紹介とともに，システム記述言語であるEmacs Lispの利用により，GNU Emacs自体，自然言語処理プログラムの優れた開発環境であることを示す．

（平成12年10月2日）