日本語ワードプロセッサから始まる東芝の自然言語処理技術 ── 歴史と展望

Toshiba Natural Language Processing Technologies Starting from Japanese-Language Word Processors: History and Prospects

住田　一男

■ SUMITA Kazuo

自然言語は，人どうしが意図を伝え，知識を残すためのもっとも自然で基本的な手段である。東芝は，日本語ワードプロセッサの実現以来，自然言語処理技術の深耕とともに，その応用製品とサービスの開発を行ってきたが，2008年その日本語ワードプロセッサがIEEE（電気電子技術者協会）マイルストーンに認定された。

今後は，オフィスでの業務効率化のための知的な新ソリューションやデジタルメディア機器における知的な新機能の実現，音声翻訳をはじめとした新商品創出に向けて，多言語での自然言語処理技術の研究開発を推進する。

Natural language is the fundamental means by which we convey our intentions to others and record our knowledge. Since its development of the first Japanese-language word processor, Toshiba has been further advancing natural language processing technologies and developing applications and services using them. In 2008, the Japanese-language word processor was recognized as an IEEE (Institute of Electrical and Electronics Engineers, Inc.) Milestone.

We will continue to promote the research and development of natural language processing for multiple languages in order to provide new intelligent solutions that improve the efficiency of office work, to realize new intelligent functions for digital media products, and to create innovative products such as speech-to-speech translators.

日本語ワードプロセッサから始まる東芝の自然言語処理技術

日本語ワードプロセッサ（WP）が2008年，IEEEマイルストーンに認定された。ここでは，このIEEEマイルストーン受賞を記念して，東芝の自然言語処理技術のこれまでの歴史を振り返る。

当社が1978年にわが国初の日本語WP“JW-10”の製品化に成功する以前，日本語文書を活字形式で作成するには専門家のタイピストに依頼しなければならず，多大な手間とコストが必要であった。日本語WPが発表されてから30年が経ち，携帯メールや電子メール，インターネットなどにより，誰もが自由に情報を発信しアクセスできる環境があたりまえとなっている。文章が自由に入力できなければ，パーソナルコンピュータ（PC）や携帯電話，インターネットもこれほどまで普及することはなかっただろう。

1985年に当社がポータブルWP Rupo™を商品化した以降，他社からもポータブルWPが商品化されるようになったことで，オフィスにとどまらず多くの家庭へと日本語WPの普及が進んだ。しかし，ノートPCをはじめとするPCの普及が進み，また日本語WPの機能がパッケージソフトウェアとして提供されるようになった。その結果，専用WPはPCで代替可能となり，2000年前後には当社を含めて多くのメーカーが専用WPの

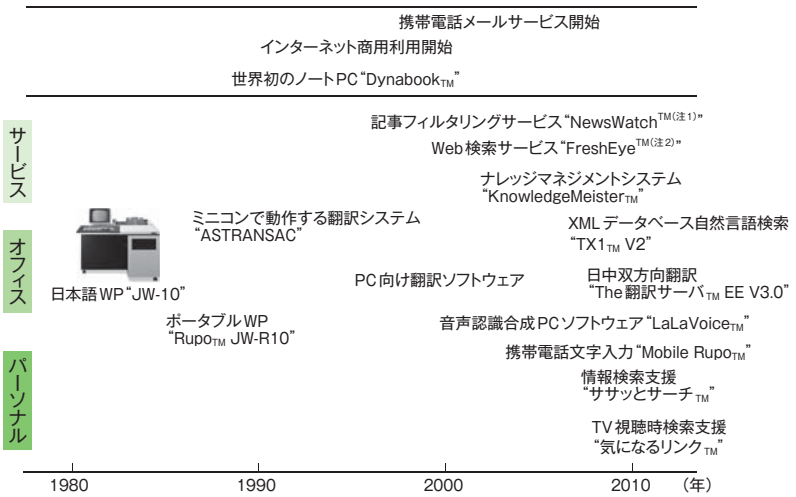


図1. 自然言語処理の応用 ― 東芝は，“かな漢字変換”の技術を通じて深耕した自然言語処理技術を様々な応用に適用してきた。

Applications of natural language processing

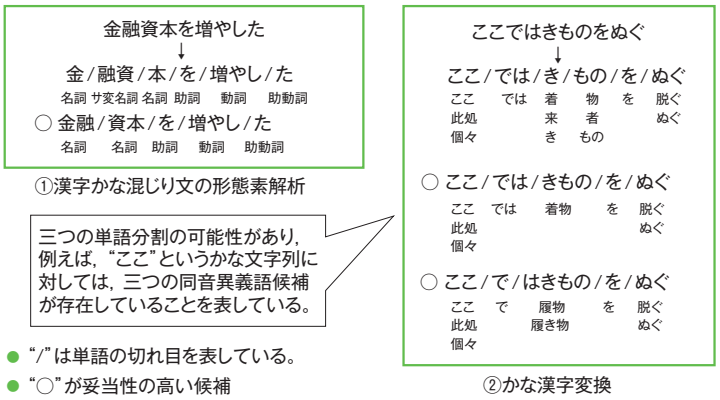
（注1），（注2） NewsWatch，FreshEyeは，（株）ニューズウォッチの商標。

自然言語処理の要素技術としての形態素解析と“かな漢字変換”

われわれが普段使うことば（言語）は，コンピュータを動かすために定義したプログラム言語とは異なる性質を持っている。このため，情報科学の領域では，“言語”のことをプログラム言語と区別するために特に自然言語と呼んでいる。自然言語を取り扱う技術全般を自然言語処理と呼ぶ。

自然言語処理の主な要素技術としては，形態素解析，構文解析，意味解析などが挙げられる。また，機械翻訳や情報検索，テキストマイニング，音声対話などの応用を実現するためにはそれぞれに固有な処理技術も必要となる。

形態素解析は，文中の語を認識し，それらの語の品詞などを特定する処理であり，自然言語処理におけるもっとも基本的な要素技術である。英語では語の間に空白があるが，日本語では各語の間に明示的な区切り記号が存在しない。このため，日本語文の形態素解析の主な目的は，入力文を辞書中の語に分割することである。



形態素解析結果の例

日本語WPにおける“かな漢字変換”技術もこの形態素解析に相当する。形態素解析の結果の例を**図**に示す。ひらがなの入力文であれ，漢字かな混じりの入力文であれ，語の区切りは複数候補が存在する。このため，形態素解析では，辞書や文法に基づいて，これらの複数候補から正しい語の系列を求める。更にかな漢字変換では，ひらがな入力のため，漢字かな混じり文に比べてあいまい性が増大する。そこで，このあいまい性を解消するための技術が必要になる。

開発から撤退することになった。

現在，日本語WPの機能は携帯電話でも実現され，基本的な機能となっている。更に，携帯電話では少ないキー入力で文章を入力したいというニーズから，Mobile Rupo$_{TM}$では，ユーザーの入力履歴からどのような単語が入力されるかを予測する入力予測機能が提供されている。

日本語WPにおいて入力のひらがな文字列を漢字かな混じり文に変換するソフトウェアは，“かな漢字変換”と呼ばれる。当社は，かな漢字変換に関して，品詞や文法の精緻（せいち）化や，“熱い”コーヒーと“厚い”本のように同音異義語を文脈によって適切な語を選択するAI（Artificial Intelligence）辞書などの開発により，高精度化を進めてきた。そして，その技術開発を通じて深耕した自然言語処理技術（**囲み記事参照**）を様々な応用に適用してきた。これらの応用例を**図１**に示す。

インターネット上のサービスへの応用（1995年～）

インターネットの商用利用が1990年に始まり，World Wide Web（WWW）上で動作し，文字と画像を一つの画面に自動的に配置して表示することが可能なMosaic（1993年）やNetscape（1994年）などのWebブラウザが発表されると，WWWによる情報提供が一般的になっていった。以降，WWW上の情報量は爆発的に増えた。2000年には6.2 E（エクサ：10^{18}）バイトと推定されているが，2011年には1.8 Z（ゼタ：10^{21}）バイトにも達すると予想されている。

また，これらの情報は一つの国のことばで発信されているわけではなく，様々な国のことばで発信されている。このことを示す例として，インターネットユーザーの国別人口を**図２**に，国別特許出願件数を**図３**に示す。これまで，英語と日本語の占める割合が多かったが，

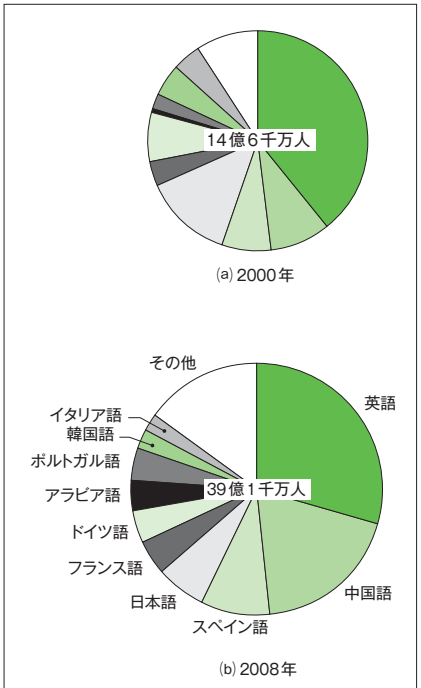


図２．インターネットユーザーの国別人口―ユーザー数が全体的に増加するとともに，使われていることばの分散化が進んでいる。特に中国の比率が高くなっている。

Population of Internet users by language

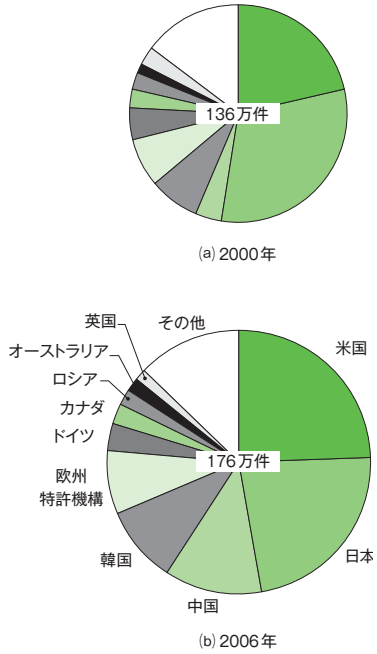


図3. 1年当たりの国別特許出願件数 — 米国と日本の出願件数が占める比率は依然として大きいが，日本の出願特許の比率が減少し，中国の出願特許の比率が大きくなっている。

Annual number of patent applications by country

近年，中国語をはじめとして様々なことばの比率が増しつつある。

このような情報の爆発と多言語化の状況を背景に，当社は自然言語処理に基づく応用技術の開発とインターネットの分野への応用を図ってきた。

■情報フィルタリングと情報検索

当社は，情報フィルタリングシステムを開発し，(株) ニューズウォッチを米国ベンチャーなどの複数社と共同で設立し，新聞記事を対象とした情報フィルタリングサービスを1996年より開始した。新聞記事の配信を新聞社からネットワークを介して受け，利用者の関心に合わせたトピックごとに分類し，電子メールやWebで毎朝利用者に送り届けるサービスである。自然言語処理に基づく情報フィルタリング技術によって高精度に新聞記事をトピック別に分類するとともに，100を超える新聞メディアとのネットワークを生かした情報サービスを継続している。また，企業サイトに特化した検索サービスを提供している。このサービスは，Web検索エンジンとして当社で開発し，1998年にサービスを開始したFreshEye™に基づく検索サービスである。

これらサービスの立ち上げ以降も，検索技術の精度向上に努めている。例えば，技術ベンチマークを目的とした国際ワークショップNTCIR (NII Test Collection for IR Systems) で上位の成績を収めている[1]。

■機械翻訳

WWW普及に伴って，WWWを介してほかの国のことばで書かれた情報へアクセスする機会が増えてきた。当社は，自然言語処理の応用の一つとして，機械翻訳の技術開発を行ってきた。製品としては，ミニコンピュータ上で動作するシステムから始まり，1995年には，PC向けのパッケージソフトウェアを製品化した。現在，パッケージソフトウェア“The翻訳$_{TM}$”やサーバで動作する“The翻訳サーバ$_{TM}$”が，東芝ソリューション(株) から提供されている。

The翻訳サーバ$_{TM}$は，サーバFreshEye™やその他のWWWのポータルにおいて翻訳サービスで利用されるとともに，特許庁の電子図書館における海外の利用者に対する特許情報サービスでも利用されている。また，言語の多様性拡大という動向に合わせ，中国語翻訳の技術開発を行っている(この特集のp.10－13参照)。

オフィスやイントラネットへの応用 (2000年～)

WPやPCが普及するとともに，ネットワーク環境が整備されると，オフィスで電子化された文書情報がはんらんし，それらの情報を有効に共有し活用する必要性が増してきた。1990年代後半には，組織のフラット化など組織構造の変化や，経験を積んだ社員が定年を迎え企業を去ることにより，知識の継承の重要性も意識されるようになった。企業内の個々人が持つ知識の共有と活用，創造を活性化することによって経営を進める仕組みや組織活動は，ナレッジマネジメントと呼ばれる。

■ナレッジマネジメント支援

当社は，情報検索やテキストマイニングなどの自然言語処理技術によって実現したナレッジマネジメントシステムを開発し，1990年代後半から，社内での実践を進めてきた。そして，東芝ソリューション(株) がその社内実践システムを製品化し，現在，ナレッジマネジメント支援ソフトウェア“KnowledgeMeister$_{TM}$”を提供している。

KnowledgeMeister$_{TM}$は，情報収集，知識共有，知識分析，知識蓄積，及び文書管理の製品群から成り，自然言語処理技術が生かされている。

例えば，知識共有では，指定したキーワードを含む文書を検索する全文検索機能と，思い浮かんだフレーズを，そのまま質問文として使用できる自然言語検索機能を提供している。また，検索された文書から質問文に近い文を提示する機能や，質問文と検索結果から検索絞り込みのための関連キーワードを提示する機能，蓄積した情報を意味内容ごとに自動で振り分ける文書分類機能など，自然言語処理に基づく先進的な機能を提供している。

また，知識分析は，コールセンターなどで蓄積した顧客からの問合せ内容の全体像をクラスタリング分析とテキストマイニング分析によって把握し，問題点分析や製品開発に生かすヒントを得るために利用されるエンジンである。テキストマイニングでは，文章中から原因や結果などの因果関係を抽出することが可能である。現在，WWW上の掲示板において交わされる特定の製品やサービス，企業などに対する批判を監視し，適切に対処することを目的として，風評を対象としたテキストマイニングの研究開発に取り組んでいる(同，p.18－21参照)。

■情報の構造化と構造化データに対する検索

社内文書の管理において，作成日付や作成した組織名などの整合性や首尾一貫性が重要となる。社内文書はWPソフトウェアで作られるため，地の文に埋もれてしまい，文書の改訂などにより，このような情報の整合が取れなくなってしまう場合が多かった。このため，このような情報を抽出し，構造化する技術を開発した（同，p.22－25参照）。

構造化された文書は，XML（Extensible Markup Language）データとして表現され，XMLデータベース“TX1TM”により，効率的に管理することが可能となる。

■自然言語処理ソリューション

TX1TMでは，これまで全文検索しかサポートしていなかったが，自然言語検索を実現した（同，p.14－17参照）。これにより，思い浮かんだフレーズを質問文として，そのフレーズに関連する情報を高速に検索することが，XMLデータベースで可能になった。

また，言いかえに基づく自然言語検索や文書チェックなど，自然言語処理に基づくソリューション開発に取り組んでいる（同，p.30－34参照）。

デジタルメディア機器への応用（2005年～）

ブロードバンド（高速大容量通信）ネットワークが家庭にまで引かれるようになり，テレビ（TV）やビデオレコーダなどのデジタルメディア機器もこのネットワークにつながることが一般的になってきた。デジタルメディア機器がネットワークにつながることで，利用者は様々な情報にいつでもアクセスできるようになった反面，各利用者にとってほんとうに必要な情報を適切に絞り込み，送り届けることが必要となってきた。PCと違いTVやビデオレコーダはキーボードのような入力手段を持たないため，情報の持つ内容に従って振り分ける技術や，簡単に検索できる機能が求められる。

当社は，時事性の高い話題となっているキーワードの抽出とその推移を表示し，容易な話題の把握とワンクリックでの関連Webページ表示を実現した“ホットワードリンクTM”[2]や，視聴中のTV番組で気になるシーンに関連キーワードを表示し，ワンクリックでの簡単なWeb検索を可能にする“気になるリンクTM”を開発し，TV番組の視聴支援に適用した[3]。また，電子プログラムガイドの内容や視聴者の視聴履歴から，利用者が関心を持つ番組を推定し，利用者に適切に推薦する技術の研究開発も行っている[4]。

また，Web上のブログなどでの評判情報を，店に並んでいるリアルな商品と関連付けて携帯電話でアクセスする“ユビdeコミミハサンダーTM”は，語と語の間の意味的な関係を蓄積したオントロジー(注3)を用いて実現している[5]。このように携帯電話などのモバイル機器の自然言語処理による付加価値向上にも取り組んでいる。

今後の展開に向けて

■コミュニケーションとインタラクション

当社は，これまで自然言語処理を文書入力や作成，及びインターネットやイントラネットでの情報アクセスと知識共有における応用を進めるとともに，デジタルメディア機器の高付加価値化に向けた取組みを進めてきた。自然言語は，人にとってもっとも基本的で自然なコミュニケーションやインタラクションのための手段である。人どうしのやり取りと同じように，機械に対して問合せや指示ができるようになれば，新たな機器であっても操作方法を覚える必要がなくなる。

実用的な音声翻訳（同，p.26－29参照）や音声対話の実現にチャレンジしている。音声認識の精度や，口語の翻訳，対話理解の技術など，実用レベルにするには多くの課題が残されている。現在，ほう芽的な製品やサービスも存在するが，まだまだ実用的なレベルになっているとは言いがたい。利用者のニーズに応えられる精度と性能を実現する必要がある。

(注3)　特定分野の概念をまとめた語い体系。

■多言語処理への対応

経済のグローバル化の進展に伴って，日本語と英語だけにとどまらず，様々な言語の情報を取り扱う重要性が増しつつある。また，デジタルメディア機器に対する自然言語処理に基づく付加価値向上に関しても，多くのマーケットに展開するためには，それらの国々の言語に対応していくことが求められる。

当社は，中国北京の研究開発センターや英国のケンブリッジ研究所に研究拠点を設け，日中英の三拠点体制でこれら多言語での自然言語処理技術の研究開発を進めていくとともに，製品応用を進め新たな価値創出に努めていく。

文　献

(1) Sakai, T., et al. "Toshiba BRIDJE at NTCIR-6 CLIR : The Head/Lead Method and Graded Relevance Feedback". Proc. of NTCIR-6. Tokyo, 2007-05, NII. p.36－43.

(2) 岡本昌之，ほか．気になるキーワードから時事の話題を検索できるホットワードリンクTM．東芝レビュー．62, 12, 2007, p.58－61.

(3) 山崎智弘，ほか．番組視聴中に気になったシーンを簡単に検索できるキーワード抽出技術．東芝レビュー．63, 4, 2008, p.26－29.

(4) 折原良平，ほか．コンテンツの高精度推薦技術によるデジタル機器の価値向上．東芝レビュー．61, 12, 2006, p.13－17.

(5) 川村隆浩，ほか．ネットとリアルを結び付けるオントロジー技術“ユビdeコミミハサンダーTM”．東芝レビュー．61, 10, 2006, p.62－65.

住田　一男
SUMITA Kazuo, D.Eng.
研究開発センター　知識メディアラボラトリー研究主幹，工博。自然言語処理の研究・開発に従事。
Knowledge Media Lab.