NEWS RELEASE

NTT
日本電信電話株式会社

平成11年9月24日

日本電信電話株式会社

# CD-ROM版日本語語彙大系と文章解析プログラムを完成

ＮＴＴは、日英翻訳システムの研究成果として、大規模で信頼性の高い日本語の意味辞書「日本語語彙大系CD-ROM版」を完成し、さらにこの信頼性の高い意味辞書の体系を用いて任意の日本語の文章を解析する形態素解析プログラムを開発しました。本プログラムは、研究目的の利用に関しては登録者に無償で提供します。

大規模な日本語単語の意味の違いを体系化し、これら意味体系を検索機能や日本語解析機能とともに提供するのは、世界で初めてです。

## ＜開発の背景＞

人は対話や読書などで自然に言葉を覚えていきますが、コンピュータが言葉を覚えるには、単語を計算機上に逐一収録していかなければなりません。また、人はそれらの語句がどのような状況で使用できるか経験的に学習していきますが、コンピュータでは、単純に収録語数が増えると、「平野（ひらの、へいや）」のような同形語が増えて意味を判断するのに曖昧さが生じてきます。また「米国産」のように、意味をとらえないと米国/産を米/国産のように分割を誤るという問題も出てきます。

このような曖昧さを解消し、コンピュータによる日本語の解析精度を向上させるために、ＮＴＴコミュニケーション科学基礎研究所ではコンピュータが直接扱うことができる言語知識ベースの構築とそれを用いた日本語解析プログラムの研究を進めてきました。

今回完成したのは、（１）日本語語彙大系(CD-ROM版(*1,*2))、および（２）CD-ROM版と共通の意味体系に基づく日本語形態素解析プログラム ALTJAWS（アルトジョーズ）（*3）です。

（１）は、'97年9月に岩波書店から出版された「日本語語彙大系（全五巻）(*4)」に、さまざまな検索機能を組み込んで実現した電子化辞書で、（２）は、日本語の文章を入力すると、単語単位に区切り、その品詞を決定するとともに、（１）の体系に基づく意味情報がコードで出力されるプログラムです。

**＜技術のポイント＞**

今回完成した日本語語彙大系CD-ROM版は以下の3体系を１枚のCD-ROMに収録しています。

・意味体系：日本語の一般名詞、固有名詞、用言の意味の用法をツリー構造で3,000カテゴリ、最も深いレベルで12段に体系化（図1）
・単語体系：30万語の単語を意味体系（3,000カテゴリ）によって定義。
・構文体系：6,000の用言に対して、主語・目的語にくる名詞を意味体系で定義し、対応する英語構文を収録。

今回の辞書は、単語レベルでは約30万語に対して、名詞を中心に各々の語に意味カテゴリーを付与しています。図2の例は「ホテル」に対して「宿泊施設」「企業」の意味カテゴリーが付与されている状況を示しています。

文のレベルでは日本語の用言（動詞や形容詞）の主語や目的語にどのような意味をもつ名詞が使われるかという構文のパターン分けを行い、14,000パターンの辞書を構築しています。例えば「取る」という動詞に対して、「ホテルを取る」は"reserve"、「使用料を取る」は"charge"というように、目的語になりうる名詞を、意味カテゴリーを用いて表現することで、構文の意味の違いが分かるようになっています（図3）。

「日本語語彙大系CD-ROM版」の仕様は、電子出版の共通規約であるEPWING規約*5に準拠しています。読みの情報から検索できるだけでなく、以下のように辞書体系間を移動したり、複合した検索をすることが可能となっています。

---

・単語の類語や関連した語を検索する場合：→図4
・用言（動詞や形容詞）の訳語を調べる場合：→図5
・日本語を学習するときに、（１）「～とあう」（２）「～をあう」のような、動詞と助詞の組み合わせが自然かどうか調べる場合：→図6

---

日本語形態素解析プログラムでは、これら実用的な語彙規模で体系化した辞書のうち、意味体系と単語体系を実装することで、固有名詞などが多く出てくる新聞記事においても、文章を精度よく単語に分割することができるようになりました。図7は新聞記事などでよく出てくる形式の複合語を解析した例です。

---

STEP1: 日本文が入力される

STEP2: 辞書から形態素(単語や接辞)を検索し、候補を抽出する。

STEP3: 3-1 文法的に接続可能な候補をつないで、候補列を作る。

3-2 単語の品詞や意味カテゴリなどを用いて意味的な係り受けの有無をチェックする。

3-3 分割数が少なく、係り受け数の多い候補を最終候補とする。

STEP4: 形態素の単位と品詞情報と意味カテゴリを出力する。

---

このように、単語間の接続条件や係り受けを自動的に調べることで、正しいと想定される候補を抽出することができます。また、図8は、「私は休暇を取る許可を取り、妻はホテルを取った。」の文章をALTJAWSで解析した結果です。

この意味体系に基づく辞書は、日本語の意味理解の研究に大きな前進をもたらすと同時に、日本語と英語の表現の違いに関する分析など言語学の研究、日本語を学ぶ人々にとっても貴重なデータとなるものとして期待されます。

今回のCD-ROMは、情報処理だけでなく、国語学、言語学の分野の研究者、日本語に関心のある一般の方々、日本語に興味のある方々が、研究や日常の仕事などで利用していくことができます。

一方、形態素解析プログラムは、さらに情報処理の分野において文書の内容の検索、文書へのキーワード自動付与などテキストを扱う研究、電子メールの自動分類やフィルタリングなど広範な情報流通に関する研究などの実験への適用が可能です。本プログラムは、大学・研究機関での研究目的に限り、 無償で利用可能です。詳しくはALT（アルト）資産管理事務係(*6)へ お問い合わせ下さい。

**＜今後の展開＞**

言葉はつねに変化していくものであるため、NTTでは、日本語語彙大

系が新しい言葉や概念にも対応できるように拡充と検証を進めていく予定です。また、研究用に提供する形態素解析プログラムは、機械翻訳の最初のステップに当たります。こちらに関してもNTTでは今後、構文解析機能等のパッケージ化も進めていきます。

*1: http://www.kecl.ntt.co.jp/icl/mtg/resources/GoiTaikei/
*2: http://www.iwanami.co.jp/hotnews/GTCD/ (作成準備中)
*3: http://www.kecl.ntt.co.jp/icl/mtg/resources/altjaws2-j.html
*4: http://www.iwanami.co.jp/hotnews/GoiTaikei/
*5: http://www.epwing.or.jp/
*6: 〒244-0805 神奈川県横浜市戸塚区川上町90-6　東戸塚ウエストビル8階
NTTアドバンステクノロジ株式会社 ALT資産管理事務係
TEL：045-826-6185 (平日9:00-17:00)　FAX：045-820-1532
E-mail: alt@totsuka.ntt-at.co.jp

---

図1 意味の体系
図2 名詞の意味体系と単語(ホテル)の対応関係例
図3 日本語語彙大系を使った訳語選択例
図4 「ホテル」の意味カテゴリと同一カテゴリの語の検索
図5 構文体型の検索例「運動する」
図6 フレーズ（○○と．．．あう）からの検索例
図7 複合語の形態素解析ステップと処理例
図8 日本語形態素解析プログラム(ALTJAWS)の出力例

＜本件に関するお問い合わせ先＞
ＮＴＴ先端技術総合研究所
企画部　真鍋、活田、佐々木
Tel: (046) 240 5152, Fax (046) 270 2365
E-mail：st-josen@tamail.rdc.ntt.co.jp

NTT NEWS RELEASE