自然言語処理技術を応用した電子化辞書の作成環境の構築

松本　裕治[†]　徳永　健伸[††]
奥村　学[†††]　杉浦　芳樹[††††]

日本語処理に向いた十分な語彙数と正確な記述内容を持ち、日本語処理に必要な文型、意味、用例などの情報を体系的に含んだ辞書が期待されている。しかし辞書の編纂には専門的な言語学的分析が必要であり、多くの人手と膨大な時間が必要とされる。従って、効率よく高品質な辞書を開発するための統合環境の開発を本プロジェクトでは目的としている。形態素解析、係り受け解析、クラスタリングなどの自然言語処理技術を用いて、大量の日本語コーパスや既存の辞書から、一次辞書データを作成する。さらにネットワークを介しての複数のユーザによる協調編集を可能にし、文型検索、類似検索、ハイパーリンクなどの高度な検索機能を実現する。

1　はじめに

パソコンの普及に伴ってインターネットの利用者も急増し、現在の技術よりも適用範囲が広く高度な日本語処理技術が求められている。それによって、日本語でのインタフェイスや検索などの分野で、日本語処理技術のいっそうの向上が必要となった。高品質の辞書は日本語処理技術の前進に不可欠であると思われる。必要とされる辞書とは、当然のことながら十分な語彙数を持ち、文型や意味情報などが正確に、しかも体系的に含まれたものである。

電子化された辞書では、『岩波国語辞典』や『三省堂大辞林』などがCD-ROMなどの形で出版されている。しかしこれらは印刷された書籍の代わりに利用されるレベルに留まっているのが現状で、日本語処理への応用に直接利用することができない。

一方、文型などの情報が体系的に含まれており、日本語処理への応用も可能なものにIPAL辞書[1][2][3]がある。これはIPAによって開発された動詞、形容詞、名詞の辞書で、利用分野を特定しない、汎用的な辞書の提案を目的としているが、辞書そのものを編纂するにあたっての語彙量が少ない。このように既存の辞書には一長一短がある。

辞書を開発する環境もまた問題を抱えている。開発に携わる技術者が情報交換を行うための共通の基盤がなかったり、開発された各種システムの動作が統合されていず、ばらばらだったりしているのが現状である。

本プロジェクトは辞書の作成環境を整備し、その利用技術の提供を目的にしている（図1参照）。

† 奈良先端科学技術大学院大学 情報科学研究科
†† 東京工業大学 大学院情報理工学研究科
††† 北陸先端科学技術大学院大学 情報科学研究科
†††† 株式会社 管理工学研究所

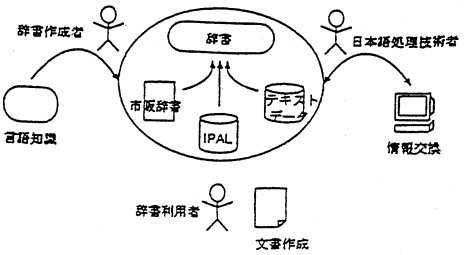

図 1　システム概要

2　既存の日本語処理システムの統合

辞書を作成するにあたっては、見出し語、品詞、文型、用法などの情報が必要であり、これを大規模なテキストコーパスから抽出することを考える。これには既存の日本語処理システムを応用・拡張して利用する。

2.1　形態素解析

形態素解析には高速・高精度のシステムが要求される。また日本語の形態素の分類には多くのバリエーションがあり、特定の形態素分類のみを対象として動くシステムでは、利用範囲が限定されてしまう。本システムでは、奈良先端科学技術大学院大学で開発された日本語形態素解析システム「茶筌」[4] をベースに、異なる形態素文法の辞書間の変換をサポートする機能を開発する。また精度を上げるために正しい解析結果を蓄積し、品詞や読みを含む形態素情報や、形態素の接続規則などのコスト値を学習する技術を開発する。

2.2　係り受け解析

日本語処理の基本的な解析機能として、また言語知識学習を行う基本的なデータ構築のためのツールとして、文節間の係り受けを解析することが重要である。さらにそれらのデータを利用しやすい形で蓄積する必要がある。そこで形態素から統計モデルに基づく文節単位の係り受け解析 [5] を行い、解析結果の構築と蓄積を支援する。また作成された係り受け情報により、文節の中に含まれている一つ一つの用言ごとに、それを連用修飾している文節内の名詞と格助詞の組み合わせを抽出し、用例データや格パターン (文型) データを獲得する。さらに修飾関係と品詞の組み合わせをもとに、連接する格助詞や活用などの情報を考慮して用法を決定する。

2.3　類義語抽出

辞書を編集する上で、語の類似関係は重要な情報である。とくに格フレームなどの編集を考える際、語の共起の傾向に基づく類似性が重要となる。このため語を共起情報に基づいてあらかじめクラスタリングし、それを辞書編集者に提示、あるいは検索させることによって辞書編集を支援することが考えられる。クラスタリングはこのための基礎的な技術である。これまでにクラスタリングのための様々なアルゴリズムが提案されているが、十分な精度で大規模な事例に適用できるものはほとんどない。しかし東京工業大学と日立基礎研究所で共同開発されたベイズ理論に基づく階層クラスタリング・アルゴリズム（ＨＢＣ）[6] は、これらの問題を解決する可能性を持っている。現在のＨＢＣでは、大規模な事例に適用するためには計算コス

トが高すぎるので、精度のよい近似手法を開発する。具体的な類義語抽出は、名詞、動詞そして名詞と動詞の組み合わせ（格パターン）の頻度を求め、頻度行列表を作成する。そしてその頻度行列表をもとに類似度を計算する。類似度の一番高いものをペアにしてゆき、次々に同等の処理を繰り返してペアを作っていく。これによりペアの階層構造ができる。このペアの近いもの、つまり類似度が高いものを類義語と見なして抽出する。

2.4 ハイパーリンク

語の関連は単純ではなく、複雑なネットワーク状（網目状）である。こうした関連を表現するには、リレーショナルモデルよりハイパーテキストの方が適している。語同士のリンクを用いることによって柔軟な検索を実現できる。しかしその語義までを特定して、人手でリンクを張るのは膨大な作業となるため、これを自動化する必要がある。リンクテーブルの張り方に関しては、北陸先端科学技術大学院大学の「IPAL辞書の自動的ハイパーテキスト化」[7]がある。それは関連語や文型、意味素性を利用し、対応する語の語義の決定を行い、自動的にリンクを生成するものである。この技術を応用し、語のネットワークを構築する機能を開発する。

3 日本語処理を応用した検索

辞書編纂のためには単語の検索も必要であるが、国語辞典やすでにあるIPALなどの電子化辞書の情報を参照したり、あるいは抽出された情報の元となった文を、文例や用例として参照したいという要求がある。そのためには単純な正規表現による検索、そして文型や関連語も検索可能にする必要がある。また検索においては表記の揺れについても考慮しなければならない。そこで表記の揺れを応用した検索システム、文例を形態素情報や係り受け情報を利用し、文型で検索するシステムを開発する。文型検索は検索条件に格パターン、語、品詞、活用形などを、文中における相対的な位置関係（前、後など）も含めて指定し、検索することを可能にする。またこれらの検索技術では、一般の日本語文書の全文検索への応用も考慮する。

また単語を解析してゆく上でその語の用法、文型などの分類情報やその頻度などの情報を提供することは重要である。上記の検索を利用し、頻度計算や分類を行い、それらの結果を元に分布図などを利用してグラフィカルに表示することによって、言語知識獲得のサポートを行う。

4 協調編集

辞書の記述は大勢で協調して作業することになるので、まずその作業環境を整えなくてはならない。そのためには文例や辞書の作業単位を担当者ごとに分割し管理する機能、作業結果に不整合が発生しないようにするための排他制御機能を提供する。またUNIXシステムのみでなく、Windowsなどの環境も考慮する必要がある。具体的にはネットワークを使用した共同作業環境を開発し、イントラネット、インターネットでも利用できるように、WWWを中心としてJava, CGIによるインタフェイスを提供する。

また複数の作業者で辞書を記述してゆくと、記述内容に矛盾が生じる可能性があるが、その

対策として辞書不整合をレポートする機能を開発する。

5 相互運用性

辞書作成作業では、1つのチームという枠組みに捕らわれないで、他の日本語処理技術者と情報交換を行いたい、また他の既存の辞書と情報を交換したいという要求もある。しかし外部辞書を使って内容の変換を行う場合、同じ辞書同士では生じなかった問題も、そこでは起こってくると思われる。そこで、保持している情報の体系化、分類方法を洗い出し、必要な情報は変換できるシステムを作り出さなくてはならない。この変換を柔軟に行うために、品詞変換機能を、また変換方法をフォーマット定義という形で記述して、外部辞書との変換を柔軟に行う機能を提供する。より汎用的な交換手段として、国際的に規格化され、広く使用されているSGML、またWWWにより普及しているHTML形式などの変換機能を提供する。また辞書の文書化をサポートするためにTeX変換機能を作成する。これらの各種変換機能により、辞書情報交換の基盤を提供する。

6 おわりに

辞書の作成作業も日本語処理も、コンピュータを使用して行われるようになってきた。しかしまだかなりの部分を手作業に頼り、共同作業の環境も十分とは言えない。また工学系の技術者はUNIXを中心としたシステム、文系の研究者はDOS、Windowsを中心としたシステムを使用する傾向があり、利用OSの違いがいが、相互に意見を交換したり、共同で作業するための妨げになりがちである。本プロジェクトでは、現在ある日本語システムを応用、統合し、共同作業、情報交換基盤を整備し、WWW、Javaなどのインターネット技術を使用することにより、相互運用性の高い日本語処理環境の確立を目指す。

謝辞

なお，本プロジェクトは情報処理振興事業協会(IPA)の創造的ソフトウェア育成事業の支援によるものである。またIPAの橋本三奈子氏、桑畑和佳子氏から、辞書を作成するための環境について貴重なご意見を頂戴した。厚くお礼申し上げる。

参考文献

1) 情報処理振興事業協会. 計算機用日本語基本動詞辞書ＩＰＡＬ(Basic Verbs)辞書編・解説編. 1987.
2) 情報処理振興事業協会. 計算機用日本語基本形容詞辞書ＩＰＡＬ(Basic Adjectives)辞書編・解説編. 1990.
3) 情報処理振興事業協会. 計算機用日本語基本名詞辞書ＩＰＡＬ(Basic Nouns)辞書編・解説編. 1996.
4) 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明. 日本語形態素解析システム『茶筌』version 1.0 使用説明書. In *NAIST Technical Report*, 1997.
5) 藤尾正和, 松本裕治. 統計的手法を用いた係り受け解析. 情報処理学会研究報告, pp. 83–90, 1997.
6) H.Tanaka T.Tokunaga, M.Iwayama. Automatic thesaurus construction based on grammatical relations. In *Proc. IJCAI95*, pp. 1308–1313, 1995.
7) 梁慶昇, 奥村学. Ipal辞書の自動的ハイパーテキスト化. 言語処理学会 第2回年次大会, pp. 77–80, 1996.