言語解析と語彙知識獲得のための支援環境

松本 裕治
奈良先端科学技術大学院大学
情報科学研究科

徳永 健伸
東京工業大学
大学院情報工学研究科

奥村 学
北陸先端科学技術大学院大学
情報科学研究科

大林 正晴　杉浦 芳樹
株式会社 管理工学研究所

1 はじめに

言語処理の様々な応用において、高品質の辞書の構築が重要な問題となっている。特に、分野や利用目的による言語使用の違いを考慮すると、コーパスなどの客観的な言語データを援用して辞書構築を行うことが好ましい。また、質の高い解析済コーパスを蓄積することも、辞書構築と同様に重要な問題である。本報告では、大規模コーパスを解析するための言語解析システム、および、解析された言語データを利用して、効率よい辞書編纂作業を実現するための支援環境である「計算機による辞書編纂環境 (Computational Lexicographer's Workbench)」を紹介する。次節で支援環境の全体像について述べ、以後の節でそれぞれの機能について説明する。

2 計算機による辞書編纂環境

本支援環境の全体像を図1に示す。本システムの目的は、大規模な日本語コーパスの解析結果や既存の辞書の情報を参照し、言語情報処理用の辞書の拡張や編集を効率よく行うことのできる環境の構築である。そのために、統計モデルに基づく言語解析システムや類義語のクラスタリングなどの基本的な処理システムを構築し、コーパスの解析を支援するとともに、解析されたコーパスデータを利用して解析システムの統計パラメータの学習を行うことによって解析システムそのものの精度向上を図っている。そして、解析されたデータを利用した言語知識抽出を支援することにより、辞書編集者の主観をできるだけ排除した辞書編纂環境を目指している。図1に示されるように本システムは4つの基本的なコンポーネントからなっており、それぞれは次のように要約できる。

- **言語解析部**では、大規模な日本語コーパスの解析を支援し、形態素、文節および係り受け関係の解析が行われたデータをSGMLによってタグ付けされた解析済みコーパスとして蓄積することを支援する。また、係り受けなどの語の共起関係を利用して語のクラスタリングを行う。
- **検索部**では、未解析あるいは解析済みコーパスに対して、正規表現や係り受け関係を利用した高機能な検索機能を提供する。また、動詞の用例を中心に検索して蓄積された用例データから動詞の格フレームの獲得を支援する。
- **辞書編集部**では、既存の辞書や部分的に構築された辞書の参照や編集、および、辞書内容の整合性のチェックなどの支援を行う。また、辞書項目の柔軟な参照のためのハイパーリンクの自動作成機能を提供する。
- **相互変換部**は、異なる辞書、あるいは、品詞体系などの異なる文法によってタグ付けされた言語データなどの相互変換を支援する。

これらの項目についての詳細を以下の節で順に説明する。

3 言語解析部

大規模な日本語コーパスを解析するための形態素および係り受け解析システムと、解析結果の誤りを修正するためのインタフェースを提供する。解析システムは統計モデルに基づいており、蓄積された解析結果を用いて統計的パラメータの学習を行うことにより解析システム自体の精度向上が可能である。

形態素解析システムとしては、任意長の文脈情報を考慮した連接規則の記述が可能であるように拡張した「茶筌」[1, 2]を利用している。同様に、統語解析についても、統計情報を利用した係り受け解析システム「茶掛」[3]を用いている。両者とも、蓄積された解析済みデータから統計的パラメータを学習することができるので、解析精度の向上を自動的に図ること、および学習したシステムにより解析済みデータを再解析することにより、人手による修正誤りや修正忘れの発見を支援することが可能になる。

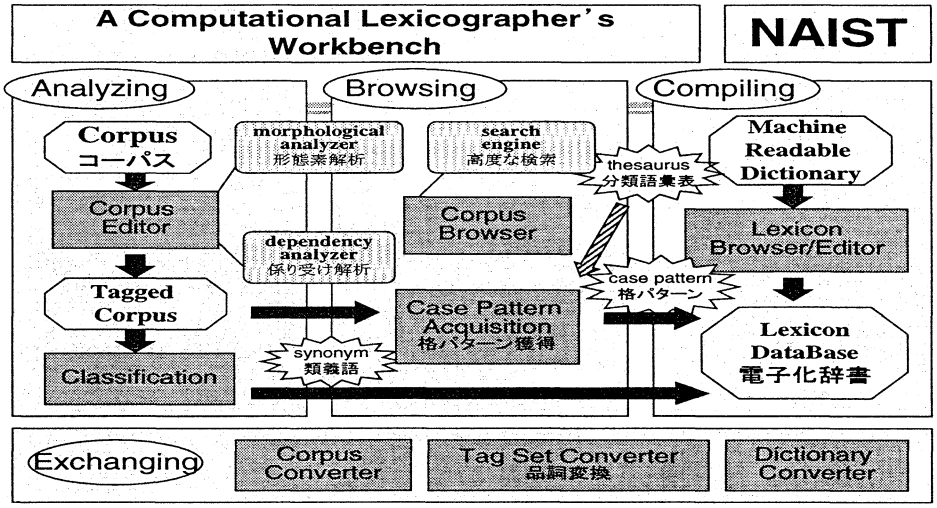

図 1: 計算機による辞書編纂環境の概念図

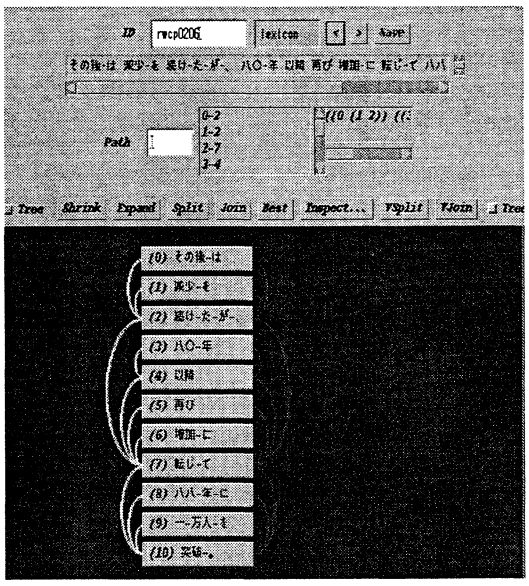

図 2: 係り受け解析インタフェース

形態素および統語解析については、どのような品詞体系や文法にしたがうかということも重要な事項である。本プロジェクトでは、RWCP コーパスで採用されている IPA 品詞体系 [4] に基づいた品詞分類を用いている。また、統語解析については、文節間の係り受けまでの解析に留めている。これは、本プロジェクトの目標である辞書の語彙記述に関する情報には文節を構成する単語間の係り受け関係が有効であると考えられることと、それ以上の解析には特定の文法に依存した考え方を導入する必要があり人手による解析結果の修正にも専門の知識を必要とするなどの困難を伴うからである。

形態素解析、係り受け解析それぞれについて、解析結果の表示と修正を支援する環境を構築している。前者は「美茶」[5] と呼ばれ、形態素解析で生じる曖昧性をグラフ状に表示し、マウス操作によって正解の選択や解析誤りの修正を行うことができる。

係り受け解析結果の表示と修正のための GUI の画面を図 2に示す。文節が一つの箱によって示され、文節間の係り受けが弧によって描かれている。図中、箱の右側の弧が解析システムによる結果であり、利用者による修正は箱の左側で行われる。文節への分割誤りの修正や文節中の品詞の修正などもマウス操作で行うことができる。修正結果の品詞、文節、文節間の係り受け関係は、SGML によるタグ付きデータとして保存される。

辞書を編集する上で、語の類似関係は重要な情報である。係り受け解析データから得られる動詞と名詞の共起関係から名詞の類似度を求め、階層的なクラスタリングを行う機能も提供する [6]。

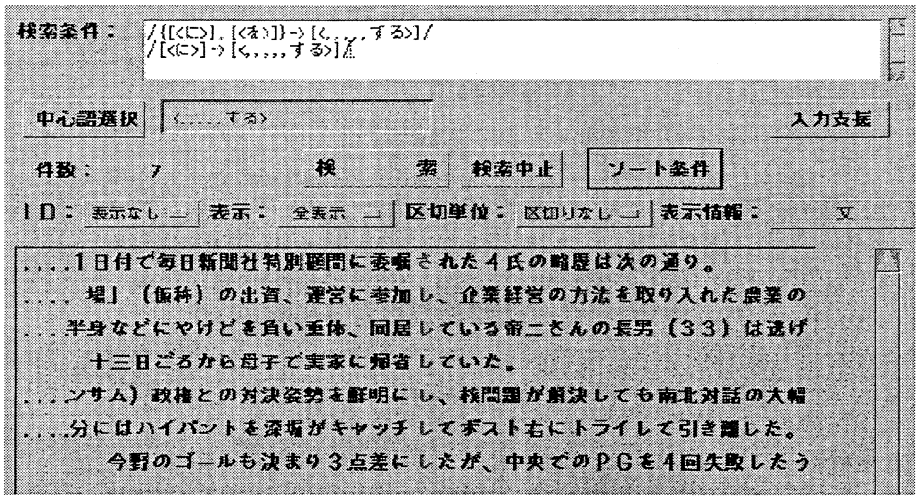

図 3: 検索 GUI

4 検索部

検索部では、係り受け解析済みコーパスに対して、語彙、品詞、係り受け関係によって記述された正規表現を用いて用例の検索を行う。検索結果は、検索対象語を中心に配置した KWIC(Key Word In Context) 形式によって表示される。図 3に検索 GUI の実行画面を示す。

語に関する情報は、⟨表層形, 品詞, 品詞細分類, 活用型, 活用形, 基本形⟩という形で指定され、文節は鉤括弧 ([ ])、係り受けは矢印 (->) によって記述される。日本語は語順が比較的自由な言語であるため、一つの文節に複数の文節が係る場合には、係り側の文節を集合として記述し、語順に依存しない検索を行うこともできる。さらに、以上の記述を含んだ正規表現による検索も実現する予定である。図 3では、「に」を持つ文節が「する」という動詞を修飾する例文を検索した結果を示している。「する」の基本形が指定されているだけであり、基本形以外の活用形も照合している。また、「する」が中心語として指定されているので、検索結果は「する」を中心にして配置されている。図ではわかりにくいが、照合した「に」および「する」には利用者が指定した色が付けられており、その場所を容易に認識できる。

さらに、このように特定の動詞を中心に検索した結果からその動詞の用例の集合を求め、シソーラスなどの名詞の階層構造を利用して、動詞の格フレームの獲得を行う過程を支援するシステムを構築している [7]。これにより多義性を考慮した用例の分類と、それぞれの用法の格フレーム記述を行うことができる。

5 辞書編集部

辞書の語彙項目の記述にあたっては、それぞれの語彙項目にどのような情報が含まれるべきか、すなわち、語彙項目のフォーマットを定義する必要がある。また、記述された内容に矛盾、すなわちフォーマットの誤りが生じないかを確認する必要がある。また、語の間の関係は単純ではなく、お互いに関連し合うことが多い。こうした関係を表現するには、関係モデルよりハイパーテキスト構造が適している。関連する語同士にリンクを張り、それを用いることによって柔軟な辞書の参照を実現できる。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きわめて繁雑な作業となるため、これを自動化する必要がある。リンクテーブルの構築のために、自動的ハイパーテキスト化を行う [8]。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。まとめると、辞書編集部提供するのは、以下の諸機能である。

- 品詞ごとの語彙項目のフォーマットの定義
- 既存の機械可読辞書から定義された辞書項目フォーマットへの変換の支援機能
- 語彙項目の記述内容の矛盾のチェック機能
- 辞書の自動ハイパーリンク機能
- 語彙項目の内容を編集するためのグラフィックインタフェース

6 相互変換部

辞書作成作業では、編集者個人の作業だけでなく、他の技術者や編集者と情報交換を行うことが重要である。また、他の既存の辞書からの情報を変換したいという要求もある。しかし外部辞書の内容の利用を行う場合、同じ辞書同士では生じなかった体系間の変換の問題が生じる。この変換を柔軟に行うために、品詞変換機能を提供する。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も提供する。この機能は編集部とも共通の機能である。また、より汎用的な交換手段として、SGML や HTML 形式などの変換機能を提供する。

品詞体系の変換については、異なる品詞体系によってタグ付けされた同一テキストを比較することによって、自動あるいは半自動的に体系間の変換規則を抽出する支援環境を提供している。

7 おわりに

日本語コーパスの形態素解析と係り受け解析を支援し、解析済みコーパスを作成しつつ解析システム自体の精度向上を図る環境の提案、および、解析済みコーパスの高度な検索機能と、それを基礎データとして辞書の語彙項目の記述を支援する様々な機能をもつ辞書編集支援環境を紹介した。

辞書作成を効率的に支援し、かつ、均質で精度の高い辞書を作成するには、本稿でのべたような多くの機能が必要である。それぞれの機能のほとんどは、独立したシステムとしても有用であり、個別利用が可能であるように独立に開発してきた。それぞれのシステムの利用環境の整備とともに、全体的なシステムとしての結合も進める予定である。

本プロジェクトで構築したシステムは、整備が進み次第、順次フリーソフトウェアとして公開して行く予定である。また、今後は、利用者からのフィードバックを吸収してシステムの改良を続ける体制を整えて行きたい。

謝辞

本研究は「IPA 創造的ソフトウェア育成技術」の一環として平成 8 年度および平成 9 年度に行った成果である。情報処理振興事業協会からの研究支援に感謝する。また、本システムの設計と開発にあたって協力いただいた富士通 (株) 橋本三奈子女史、(株) シー・ラボの山田和久氏をはじめとする関係者各位に感謝する。

参考文献

[1] 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, NAIST Technical Report, NAIST-IS-TR97007, 1997.

[2] 北内啓、山下達雄、松本裕治: 日本語形態素解析システムへの可変長連接規則の実装, 言語処理学会第三回年次大会論文集, pp.437-440, 1997.

[3] 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, 97-NL-117, pp.83–90, 1997.

[4] テキストデータベース報告書 (平成 8 年度)、技術研究組合 新情報処理開発機構、1997.

[5] 山下達雄、松本裕治: 形態素解析視覚化システム ViJUMAN version 1.0 使用説明書, NAIST Technical Report, NAIST-IS-TR96005, 1996.

[6] Iwayama, M. and Tokunaga, T.: Hierarchical Bayesian clustering for automatic text classification, Proceedings of IJCAI '95, pp. 1322-1327, 1995.

[7] 宇津呂 武仁, 中塚 幸毅, 松本 裕治: コーパスを用いた動詞格フレーム辞書構築のための支援環境, 自然言語処理シンポジウム「実用的な自然言語処理に向けて」論文集, 1997.

[8] 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, A4-1, pp.77-80, 1996.