

自然言語実験支援環境

LINGUIST

小野寺 浩^{*} 佐々木 博司^{*} 佐野 洋^{**} 田中 裕一^{**}^{*}富士通エフ・アイ・ビー(株)^{**}(財)新世代コンピュータ技術開発機構

はじめに

分析の対象はきわめて身近であるにもかかわらず、計算機援用の自然言語研究の支援環境は高度に複雑なシステムである。言語をはっきりと意識して客体化することは一般に難しいからであろう。

言語研究者は文法の枠組の構築や辞書の記述の妥当性など言語学的側面にその探究の対象がある。特定の分析に従い、文法についてのさまざまなアイデアを評価し実証するのである。計算機援用の立場から考えると、そうした要請に応じるためには、各種の処理が協調的に行われる必要がある。例えば、文法の実行制御過程が視覚的に得られたり、任意のデータフローに使用者がアクセスできると共にそれらを随時操作できることが望ましい。辞書の記述については、辞書項目やその記述内容が実験時にすぐさま変更が可能であることも必要である。使用者への操作使用のための負担が少ないほど、言語についての分析を深めることに傾注できる。限られた時間を有効に使うことができるといえよう。

計算機援用の言語研究では、一般に、他のソフトウェア開発にくらべて、扱うデータに相関する情報が多い。その規模が大きくなるに従い、言語分析のアイデアなりをすぐさま既存の構築のモデルに反映して検証なり評価することが難しくなる。例えば、文分析において、ある品詞を再区分したい場合を考えると、関連する辞書項目の変更や、分析のための文法中の関係するカテゴリーの再構成などの作業が挙がる。様々な言語データすべてに亘り、一貫した変更を使用者に求めることは無理であろう。

言語分析の開発研究を進める上でどのような支援環境を整備するかは、使用者のシステム使用の目的の意図に密接に結びついており、システムが採る形態を決める上で重要な問題であるといえる。我々は、言語研究という要請をベースに、特定言語(日本語)に傾注した文法や辞書の研究開発を目的とする言語研究支援システムを開発した。このシステムは、文法や辞書の開発の支援を行い、解析と生成エンジンとその動作実験環境を備えている。

1 システム概要

システムの特徴を挙げる。

1.1 基本枠組

- ・属性と値の対のリストを基本データとする。
- ・文法記述形式は句構造をベースとする論理文法の枠組である。

このデータ構造には辞書内容や文法属性を素性と値の対で表現できる特徴がある。文法については理論文法の枠組を比較的素直に実現できる特徴を持つ。

1.2 インタフェース

- ・マルチウィンドウシステム上に本システムを構築している。
- ・使用者への負担の軽減のためメニュー操作を基本としている。

実システムは逐次型推論マシンPSI-II上に実現している。

2 支援ツール

開発支援環境は、上記の特徴を満足し、しかも使用者に操作負担を強要しないことを考慮している。分析のためのエンジンには、Prologを用いた。文の解析にはボトムアップパーザを用い、一方、文の生成はトップダウンパーザを用いている。実効メカニズムはあくまで透明性を失わず、単一化の記述も容易である。

この基本実行素性をベースに、基本データ構造を扱う辞書ハンドラ、入出力管理や記憶管理機能などの支援ツールを設けた。文法の実行過程の視覚化にはビジュアルトレーサを用意し、同時にデバッグ機能も付与した。分析結果は専用のインスペクタを使うことでデータへのアクセスと変更操作が可能となっている。すべての分析対象と結果は、記憶管理機構が管理しており、各種の入出力デバイスを通じて、使用者は編集なり修正を行える。

文法支援環境については別稿に譲り、本稿では、辞書操作系、言語構造インスペクタと記憶管理系について述べる。

2.1 辞書操作系

2.1.1 構成

辞書操作系は辞書データの作成や辞書属性の参照のほか属性の自動変更や使用者による辞書の構築、ならびに修正変更操作を担う。一般に、当該分野では、(1)辞書の単語が拡大した場合、それらの属性を一括管理することが困難である(2)辞書の既性と関連する文法の規則による分析と実験評価を通じて、辞書記述の妥当性を検証するまでに時間を要する、などの問題点を抱える。本操作系では、特に上記の問題点の解消を目指している。

2.1.2 機能

辞書操作においては辞書テンプレートに対する操作とデータに対する操作の二つの編集モードがある。このモードは任意に切換えが可能であって、モード変更の際して編集内容は自動的に保存と維持がなされる。

データ編集 辞書内容は部分修正が可能である。実験時におけるメモリ上の辞書エントリは、随時、辞書操作系からエントリ単位で追加・削除することが可能である。これによって、辞書データベースの構築と平行して辞書記述データの評価実験がインタラクティブに行える。一方、エントリ編集の省力化も図られている。使用者へ

のデータエントリー一覧の表示機能やメニュー方式によるエントリーの選択機能を持つ。エントリー単位の複写機能や追加機能も持っている。

テンプレート編集 辞書テンプレートは辞書データの記述形式を規定する。言語分析では辞書記述形式を採ることにも含まれる。こうした試行の後に記述形式を変更する際に、テンプレートの編集機能が有用である。テンプレートに関連する既存の辞書データは自動的に新しい記述形式に変更される。

2.2 言語構造インスペクタ

2.2.1 構成

言語構造インスペクタは文の分析結果である構成素構造を視覚的に参照することができ、さらに、個々の構成の値を変更することが可能である。

2.2.2 機能

変数の記憶管理をおこなっており、未定義の変数の生成が可能である。一つのエントリー内の変数名の識別が行える。構成素構造の各種編集機能には、修正機能、削除機能に加え、挿入機能、入替え機能がある。新たに要素を追加する挿入機能では、属性の入力の簡便化のためにメニュー選択方式を採用している。

マウスオペレーション 固定した項目についてはメニューによる編集を基本としている。

階層表示 構成素構造は視覚に訴える階層表示形態で表示される。

2.3 記憶管理系

2.3.1 構成

研究開発の対象である文法や辞書データは、記述の妥当性を評価しなければならない。実際にテキストを解析して得られる構成素構造の妥当性を評価することや任意の構成素構造を生成して文の生成の実験を行うことなどが基本的な評価手順となろう。記憶管理系はこうした試行作業を効率よく行うために容易されている。

2.3.2 機能

実験に必要なテキストやその分析結果の構成素構造を、使用者が定義した入力媒体から読み込み、それぞれのブールに保持する。さらに、文解析ならびに文生成時に対象データをストリーム経由で取り出すことが可能である。
データ管理 解析と生成の結果は、それぞれ構成素構造とテキストに対応し、システムブールに出力される。ブール内のデータは削除、複写、ストリーム出力などの操作の対象ともなる。データ保存のために、テキストと構成素構造は定義した出力媒体に出力できる。

有機統合 言語構造インスペクタと連携することで、構成素構造を修正しながら文の生成の試行実験を行うことも可能である。

3 システム

図1に辞書操作系を起動している状態の画面を示す。画面はデータ編集モードである。辞書テンプレートに従ってデータ編集を行う。テンプレートも編集対象になる。テンプレートの編集が終了すると、そのテンプレートに関連する辞書データは自動的に変更される。

4 今後の課題

現在のパイロット版では、比較評価のための機能を欠く。分析データの差分検索は重要である。また、分析手順の半自動化機能も有用な手続きである。例えば、分析例の収束をパッチ処理的に行ったり、既存のデータとの比較を行ったりする機能が必要である。

[1]に指摘されているように、実験結果を記録にとどめ、文書化したり、あるいは機械処理可能な形態で蓄積するデータベース化機能の強化は言語知識ベースとしての面から不可欠な要素である。辞書や文法規則など、言語知識ベースに関するバージョン管理は漸進的システム構成には欠かせない。分析の履歴の管理なども実験支援環境の機能には不可欠であろう。

バックアップ機能は使用者の限られた記憶能力を補完するため、また、使用者が、最も関心のある言語分析対象への集中を促すために不可欠な要素である。

今後はこれらの不備を解消しつつ、支援環境自身をより使いやすい形態にして行くと共に、言語分析の成果を蓄積してゆく。

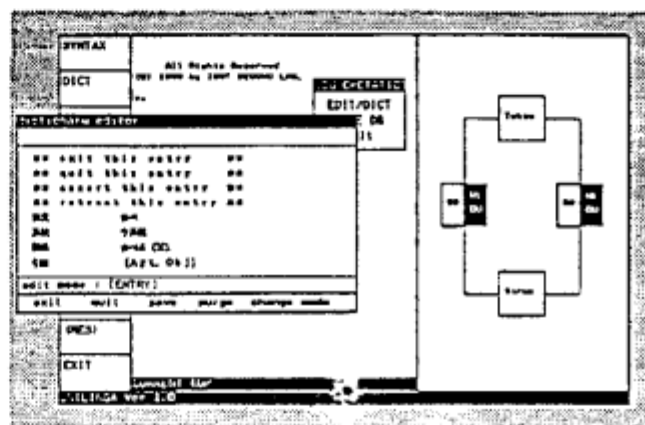


図 1

参考文献

- [1] 飯野香, 奥村明俊, 自然言語研究開発支援システム 情報処理学会第39回全国大会 pp714-715
- [2] Oliviero Stock: Copying with dynamic syntactic strategies: An experimental environment for an experimental parser. Third conference of the European chapter of the association for computational linguistics, pp234-240.