4G-3

自然言語研究開発支援システム

飯野香　　奥村明俊

日本電気（株）C＆C情報研究所

1．　はじめに

自然言語処理の研究開発では大量のデータの管理とシステム構築の処理を総合的におこなう必要がある。一般のソフトウエア開発と比べてフィードバックする情報が多いため様々な処理をより協調的におこなわなければならない。研究開発を進める上でどのような支援環境を整備するかは重要な問題である[1)]。

自然言語の研究開発において支援すべき対象は、
1）知識ベース（辞書，ルール），エンジン
2）上記の共有資産の管理
の2つに分けることができる。

本論文では自然言語を対象とした研究開発全体を一連の流れとして明確にし必要な機能を明らかにする。そして、1，2）を支援するために現在構築中である自然言語研究開発支援システムの概要を述べる。

2．　支援方針

最近、自然言語処理の研究開発をおこなうための環境やツール類が報告されている[1,2)]。これらの技術は日本語解析環境や処理系として優れた機能を提供しているが、自然言語処理システムを構築するためには他の作業の支援を含めたより包括的な環境が必要である。そこで、自然言語処理システム構築作業全体をモデル化し総合的な支援環境を考える。

システム構築作業は、図1に示すようにいくつかの状態を経ておこなわれると考えられる。まず構築するシステムの対象となる言語データを収集し、言語現象を分類する（A,B）。次に、辞書やルールをどの様に構成すべきかのモデル化をおこなう（C）。モデル化したものをシステムとして動作させるために辞書作成やルールのコーディング、デバック作業をおこなう（D,E）。システムの動作確認後、モデルの能力を評価する（F）。これらの一連の流れの中でフィードバックがおこなわれる。

各状態は、さらに細分化された状態の遷移から構成される。開発作業全体は階層化された状態遷移で表すことができる。そこで、各状態における支援機能と開発全体として必要な資産を管理するデータベース（LRDB：Language Research Data Base)の構築を考える。

3．　支援機能

各状態における支援機能の概要について述べる。

3．1．　言語データ収集機能　（A）

言語データの収集，検索，分類をおこなう。

- データ収集　　ネットワークを利用した収集
- テキスト検索
  - 表層パターンによる検索
  - 辞書，解析機能を用いた検索
- データの分類
  - 文の種類による分類（単文，複文，重文）
  - システムの処理の成功・失敗による分類

3．2．　言語現象分類支援　（B）

言語現象の分類を支援する。作表システムとグラフィックエディタを起動する。

3．3　モデル化　（C）

システムとして動作させるためにモデル化する。その結果を仕様書としてまとめるエディタを起動する。

- 仕様書作成支援エディタの主な機能
  - テンプレート利用による作成補助
  - 作表システム，グラフィックエディタ
  - 作成されたフローチャートの文章化
  - 内容チェック（項目もれ，字句誤り，等）

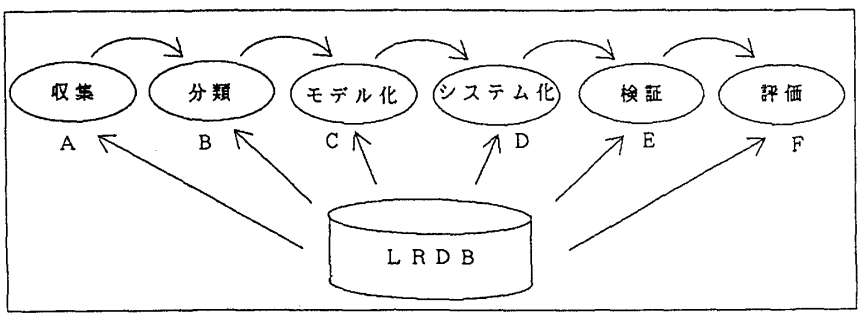

図1　開発概要

Support System of Natural Language Research and Development

Kaori IINO,　Akitoshi OKUMURA

C&C Information Technology Research Laboratories, NEC Corp.

3．4　システム化　（D）

ルール，辞書，エンジンを作成しシステムとする。

- ルール，辞書，エンジン作成支援エディタ
  - チェック機能：エラー検知，ガイド
  - 参照機能：関連情報の参照
  - 作成補助：テンプレート利用による作成補助
    関連ファイルの自動作成
    ルールの文章化

3．5　検証　（E）

システムの動作検証をするため、検証に関与するファイルを準備しデバッガを起動する。

- 入力ファイルの準備，編集
- 部分辞書，併合辞書の作成
- デバッガ
  - 対象の指定　（辞書，ルール，テーブル，入力ファイル）
  - 動作モジュールの指定（解析，生成，等）
  - 出力内容指定（構造，変数）
  - 出力形式の指定（図，色，音）
  - 動作制御（進行，停止，後退）
- 実行結果の出力
  - 対象，内容，形式の指定

3．6　評価　（F）

システムの動作を確認した後、大量データに対してテストし統計的な処理をおこなう。他システムの出力結果や正解データが存在する場合、それらの比較結果を出力する。

- 成功例収集，失敗例収集，統計処理
- 同機能他システム，正解データとの比較

3．7　ＬＲＤＢ（Language Research Data Base）

ＬＲＤＢはどの状態からでも参照できるデータベースである。参照機能の他に以下の機能がある。

- 変更点と改版履歴の保存
- バックアップ作成
- 相互的に関連するファイルのバージョンチェック

対象となるファイルは以下のものである。

- 辞書，ルール，テーブル，エンジン
- テキストデータ
- 説明書，仕様書
- 一般データベース
  - 一般辞書，論文，特許，新聞，参考文献，等

4．　システム構成

以上の議論に基づいて、支援システムをワークステーション上に現在構築中である。

システムの構成は図２のようになる。

資源ファイル及び環境ファイルはシステムの参照すべき情報の設定及びシステムの動きの設定を記述するファイルである。資源ファイルはシステム全体でひとつだけ存在するもの（辞書など）の存在場所を書いたファイルで原則として全ユーザに共通である。環境ファイルはユーザごとの設定を指定するためのファイルである。ここではルールや入力テキスト等のファイル指定の他、どの情報を詳細に見たいか等のシステム動作設定を行う。このファイルによりユーザごとにシステムをカスタマイズすることが可能となる。

各プロセスは図１の各状態に対応している。ただし、マルチウィンドウ環境により、この状態の流れは特に意識せずに並列に動かすことができる。また同様に独立したシステムとしてのみでなく、既存の環境に付け加える形でこの支援環境を使うことができる。

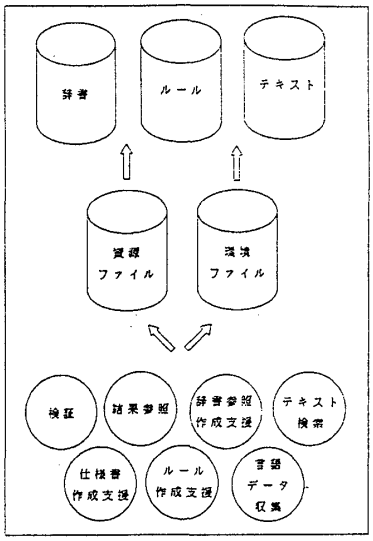

図２　システム構成図

5．　おわりに

総合的な支援環境を整備するため自然言語処理システムの構築手順を分析し、必要な支援機能を抽出した。それに基づいて自然言語処理研究開発支援システムを構築中である。本システムにより、効率的なシステム開発が可能になることが期待できる。

今後は共有資産の管理法に関して検討を進め、ＬＲＤＢを管理支援システムとしてより強力なものとする予定である。また、ユーザカスタマイズ機能をさらに強化することでユーザインタフェースを改善していきたいと考えている。

システム作成に御協力頂いている日本電気技術情報システム開発（株）の奥井伸司氏と落合尚良氏に感謝の意を表する。

参考文献

[1]　渡辺日出雄，丸山宏
「対話的日本語解析環境：ＪＡＷＢ」
情処第38回全国大会 pp.396-397

[2]　杉村領一，他
「汎用日本語処理系ＬＴＢの構成」
情処第37回全国大会 pp.1072-1073