

日本語処理に関係する機構を領野分解してみたいと思います。日本語処理の主な処理は次の通りです。

- (1) 分かち書き化
- (2) 修飾・被修飾関係の解析
- (3) 格の解析
- (4) 欠落後補てん
- (5) 指示代名詞解決
- (6) 意味理解・・・イメージ把握
- (7) 文章の理解
- (8) データマイニング (文法規則の抽出、汎化・特殊化)

これに必要な知識ベースとしては次のものがあります。

- (1) 単語辞書
- (2) 熟語辞書
- (3) 修飾・被修飾関係解析辞書
- (4) 事象や概念のモデル・・・実例文などの参照モデル
- (5) イメージ要素群・・・意味理解の基盤としてのイメージ

領野としてはこれらを全部一緒くたにするのが最も簡単です。ですが、メモリが足りない、並行処理に限度がある、・・・と現実的ではありません。特に、意味理解では文章を沢山解析した内容をイメージも含めて持つ必要がありますから、ここだけは別コンピュータで処理し、データを保管すべきでしょう。

辞書ごとにOODBでデータを持つことがもっとも考えられる場合です。辞書コンピュータの中で、データをオブジェクトとして持ち、テキストデータをオブジェクト形式にするためのオーバーヘッドをなくすのです。記号での文章の意味解析結果もイメージもオブジェクトとして、コンピュータ間を行き来するようにする。

データマイニングは知識ベースのあらゆる階層 (辞書、モデル、文章、イメージ) でデータをアクセスして、ブロードキャストで複数の専門解析コンピュータに送って、規則を発見する。その結果はデータに名前付けがなされていますので、その名前で解析結果を反映させていくことができます。

そんなのが領野分割のありようです。