コロンブスの卵

梅村祥之

Columbus's Egg

Yoshiyuki Umemura

1．電波時計を購入

わかってしまえば何でもないことなのに，大変大きな効果を出した手法は「コロンブスの卵」と言われる。

最近，腕時計で電波時計というものが発売されたので買ってみた。これは，福島県にある電波送信所から送信される標準電波を，時計に内蔵されたアンテナで受けて，時刻を修正することにより正確な時刻にするものである。

普通のクオーツ時計は月差±15秒～20秒のため，年に最大4分遅れる恐れがある。会議に遅刻するのはまずいので，遅れを1分以内にしようとすると3ヶ月に1回時刻合せが必要であり，面倒であった。電波時計は遅れを気にする必要がなく，効果が大きい。(電波技術の門外漢のため，「何でもないこと」と思ったが，腕時計内蔵の小さなアンテナで電波を受信するのは高度な技術とのお叱りを受けるかも知れない)

2．自然言語処理，音声処理の新たな方法論

私の専門分野の言語処理／音声処理で，技術の発展に大きな効果を及ぼし，現在主流となっている「コーパスベースのアプローチ」という方法論も，ある意味「コロンブスの卵」的な感じがする。

自然言語処理というのは，文章を計算機で解析して単語の切れ目を見つけ，構文を解析し，文脈中での単語の意味を選ぶといった技術であり，応用としては仮名漢字変換，機械翻訳，情報検索などが代表である。また，音声処理には，音声認識と音声合成があり，パソコンソフトやカーナビで実用化されている。

それでは，言語処理，音声処理双方の技術を急速に向上させた方法論である「コーパスベースのアプローチ」とは何であろうか。言語処理を例にとると，大量の文章に，人手で「タグ」を付けてデータベースとして蓄積する。

「タグ」とは，単語の切目がどこで，文脈中の単語の意味が何で，品詞，活用形，構文構造がどうなのかなどを表す情報である。こうしてできたデータベースを「コーパス」という。

コーパスがあればしめたもので，漢字の読み分け処理として「今日」を「キョウ」と読むか「コンニチ」と読むか例にとると，コーパスの中から「今日」の含まれる文を拾い上げ，どういったタグの組合わせの時に「キョウ」で，どういった時に「コンニチ」かを計算機で集計して，判別率の高い規則を採用することになる(これを機械学習という)。

要するに，タグの付いた大量の実例データを基に，機械処理で集計して解析規則を見つけるという物量作戦の方法論であり，それ自体，特に種も仕掛けもない。

ところが，1990年代半ばまでこのアプローチが行われず，1990年代後半から主流になり，大成功を収めている。何でもない方法を使って成功を収めたという意味で「コロンブスの卵」的な感じを受けている。

1990年代半ばに何が変わったかというと，実は，1995年に代表的な自然言語のコーパスが発行されている。このコーパスの場合，新聞や雑誌から抜き出した20万文についてタグ情報を付けて収録している。(この開発プロジェクトはメーカー9社から出向して7年がかりの国家プロジェクトとして行われた)90年代半ばからコーパスベースの研究が盛んになり，実用に供する解析技術が急速に発展した。コーパスは，解析規則を機械学習するための教材として使われるだけでなく，解析精度を評価するためのテストデータとしても使われる。実に，コーパス以前と以後の研究論文を見比べると，コーパス以前は，解析技術の評価をするのにせいぜい数百文で調べていたのが，コーパス以後は20万文で評価するようになり，実用に耐えるものになった。

3．自然言語処理，音声処理における分析手法

次に音声・言語処理における分析手法(パターン認識手法)を見てみる。

分析は，**図1**に示すように各階層に分かれて行わ

れる。各々の処理は大体において**図2**に示すような次の構造を持つ。

・隣り合う入力データを大きなまとまりにまとめる

・その際に，辞書を引いて，まとまりやすさの程度に関する情報を得る。(まとまりやすさの要素には，周りの状態に関する情報も含む)

・いろいろな組合わせで，文全体としてのまとまりの程度が最大となるものを探す (その際，高速化のためにダイナミックプログラミングなどの手法を使う) すると，この手法を可能にする味噌は，要素 (例えば単語) の生起確率と要素間の結合のしやすさ (例えば名詞の後に助詞の「は」が来やすい) の情報を持った辞書の存在である。

辞書の基本部分は，言語学の成果である国語辞書から得られ，生起確率や結合確率は先のコーパスベースの機械学習を利用して構築される。

4．おわりに

インターネット接続で，これまでと同じ電話線を使ってISDNの20倍の高速通信を実現したADSLなど，コロンブスの卵はいろいろと転がっている。柔軟な発想でブレークスルーを目指したいと思う。

(2001年7月11日原稿受付)

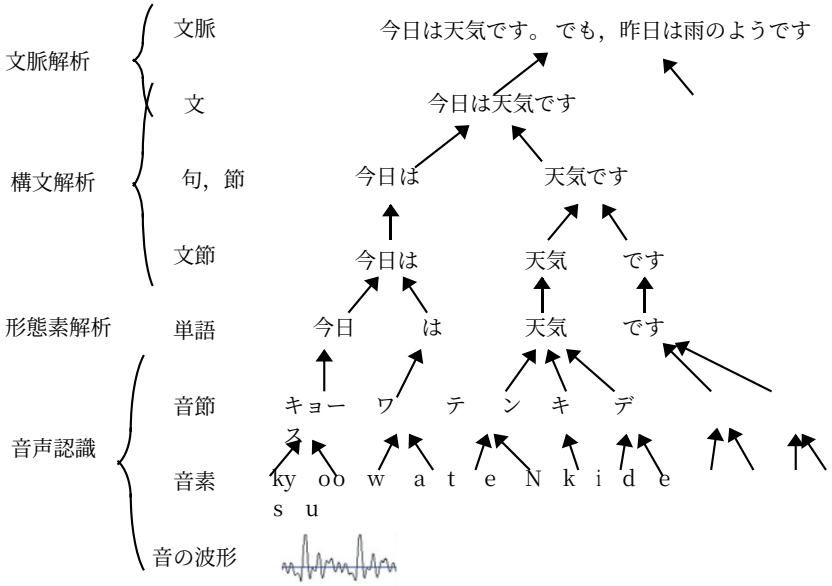

図1　音声・言語処理における解析の過程を示す。
各階層で，要素から塊 (例えば音節から単語) へまとめ上げの処理が行われる

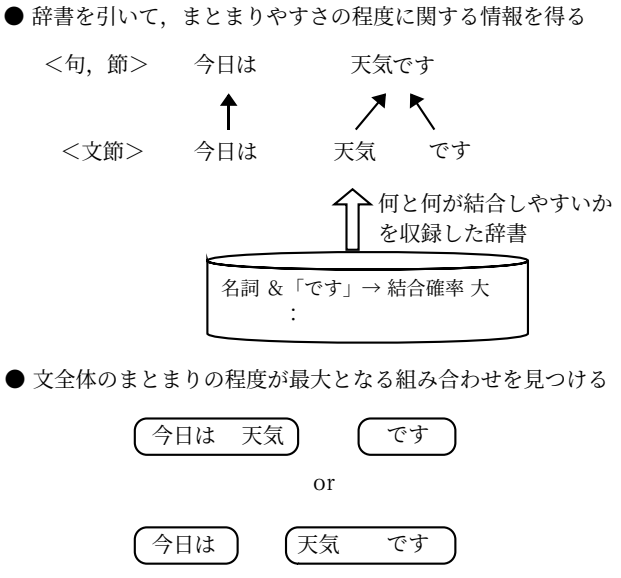

図2　要素から塊にまとめ上げる手法 (構文解析の段階を例にとる)