言語・画像のデータ依存情報処理

6

長尾　真●国立国会図書館長

言語研究方法論の変遷

20 世紀前半の言語学は Bloomfield に代表される構造言語学と称される時代で，言語データを集め，その中にひそむ構造を明らかにするという研究態度であった．そして音韻論や形態論のレベルでは大きな成果をあげた．これは人手で取り扱う範囲の言語データ量でも，安定した規則性を取り出すことができたからである．これに対して構文論の世界においては，現象が複雑多様であり，人手で集められる程度の言語データでは安定した文法規則を取り出すことができず，大まかな規則性について記述ができるという程度であった．

こういった構造言語学の行きづまりを打破したのは Chomsky であった．彼は文法機能は人間の脳にそなわったものであり，人間はそれを使って文を発話しているとし，その文法機能を仮説的に句構造文法で与えることを考えた．これが生成文法理論である．構造言語学がボトムアップの方法であるとすれば，これはトップダウンの方法である．この方法は画期的なものであったが，現実に存在する文の集合を過不足なく覆う文法を作ることはできずに終わってしまったといってよいだろう．

事例言語学

言語学は文法を明確に打ち立てることに多くの努力を割いてきた．しかし言語を句構造文法といったある種のモデルの世界で完全に書ききることはほとんど不可能である．各種の微妙な表現を文法規則で処理しようとするとその表現(句)自体を規則として採用しなければならなくなる．

言語の解析や言語データの検索をこのような句単位で行えばかなり精確なことができ，さらにシソーラスを用いて類似の表現を同じものと見なして処理することによって事例の数を限定することができる．このように事例を用いて言語処理を行うのを事例言語学と称してもよいだろう．比較的少数の文法規則を用いて深い解析を行うことによって言語を説明する方法に対して，大量の事例によって浅い解析で言語をくまなく覆うという方法である．用例主導機械翻訳はこの考え方によっている．

幸いなことに 2009 年 6 月の著作権法改正によって，コンピュータを用いて言語データの解析をし，言語の性質・構造を抽出したり，比較・分類や統計的性質を明らかにしたりするためには，著作権者の許諾を得ずに本や資料をディジタル化したり，そうなっているものを集めて使用することができることとなった．この改正によって事例言語学の研究がより本格的にできる環境がととのったと言える．

事例言語学の諸要素

大量の言語データの処理としてはまず出現する単語を網羅的に取り出し，一般用語と各専門分野の用語に分離する．次にこれらをシソーラスとオントロジーのシステムに整理することが必要となる．一般語については既存のシソーラスとオントロジーを利用し，その改良という方向で処理できるだろう．専門用語のシソーラスとオントロジーについては専門用語の現れる文を精密に解析し，そこから同義語，上位下位概念，部分全体概念等の意味チェックを行いながら構築していくことが必要となり，かなり難しい処理となる．

言語解析のための文法の基本は格構造表現をとるのが良いと考えられる．動詞を中心として，その動詞の主格，目的格や補格などにどのような語句が入るかを大量の言語データを自動解析することによって明らかにすることが行われている．つまり主格に来る単語，目的格に来る単語などを動詞とともに記憶しておき，文の解析のときにその特定の動詞の主格集合や目的格集合の中に現れる語(あるいはそれと同義の語)があれば，正しく解析されたとして格構造表現を作り出す．このような解析が言語

データの量に対応してどの程度の精度になるかの実験例を図-1に示す．対訳テキストを解析することによって日本語の句に対する相手言語の句を抽出し，これを用いて機械翻訳を行う研究も進められていて，質のよい翻訳結果が得られるようになってきている．

人間が言葉の意味を知るためには辞書を引き，そこに書かれている説明で分かったとしている．これは単語の意味を表すのに言い換えをしていることである．言葉の種々の意味，あるいはニュアンスをあますところなくすくい上げるために，種々の言い換え表現，あるいはその語の用例によって人にその語の使い方を教えているといってよい．そう考えると言語処理においてもその語の使われる句や文脈を集めることによって，その総体がその語の意味を表現していることになる．

たとえば文の解析や機械翻訳のときに意味のチェックを行い複数の可能性から最も妥当なものを選ぶという場合などである．この場合は，たとえば文中の近くの単語A，B，Cの意味関係をチェックすることになる．単語Aが現れる句(事例)集合の中に単語Bの現れる句があるか(つまりAとBの共起する句)，その出現頻度はどうかを調べる．AとCとの共起についても同様に調べ，同一文の中でAとCよりもAとBとの関係が密であるといったことから妥当な意味関係を持った解析ができることになる．

情報検索の分野においてもいろんなことが可能となってきている．クラスタリングの技術を用いてテキストの自動分類をしたり，類似テキストあるいは関連テキストを検出して相互にリンクを付けたりすることも可能となり，連想検索への道が開かれる．Google検索などではよほど考えたキーワード群で検索しても何万，何十万というテキストが出てきて，ランキングの上の方に自分の本当にほしい情報が出てくることは稀である．したがって質問を文で与え，その文と類似する文を含むテキストだけを検索結果として取り出すといった試みも行われ始めている．こういった方向で処理の精度をあげてゆくことによって，ほしい情報を含んだテキストだけを取り出すだけでなく，うまく質問をし，うまく処理すれば質問に対する解答を含んだ文だけを取り出すことも可能となるだろう．情報検索から事実検索あるいは知識検索へという方向である．

Web情報には種々のものがあるが，検索して高いランキングで出てきたものがかならずしも正しいものとは限らない．したがって得られた情報がどの程度の信頼性を持つものか，その情報を否定するような情報が存在しないかといったことや，検索出力された多くの情報がどのような種類のものであるかを大まかなクラスタリングによって分類し，意見分布を調べてみるといったことが大切である．図-2にその例を示す．

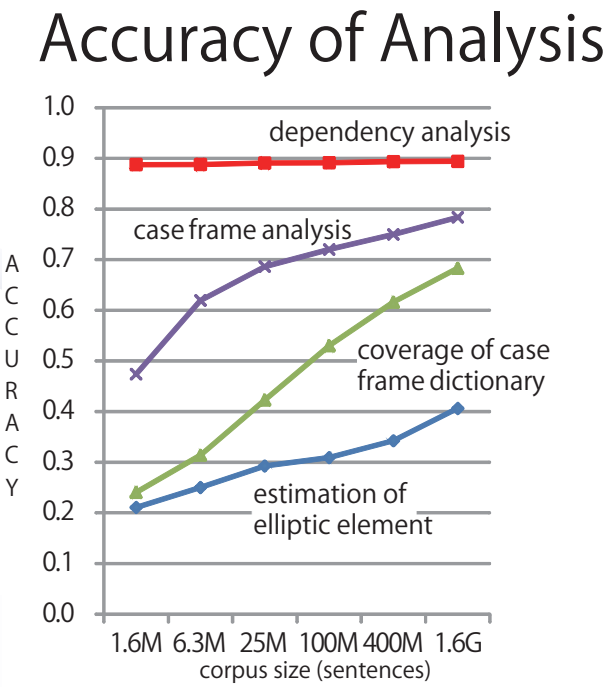

図-1　コーパス量による精度の向上(黒橋禎夫氏(京都大学)による)

情報の信頼性については，その情報の発信者がどのような機関，人であるか，発信者の氏名，アドレス，連絡のためのメールアドレスや電話番号がのっているかといったこと，また情報内容の表現が妥当な文章表現になっているかなど種々のチェックをすることによって推定することができるだろう．これと検索出力情報の意見分布を見ることによって情報検索がより信頼性のあるものとなってゆくわけで，いわば第2世代の検索方式といってもよいだろう．これらのほかに自動抄録，知識データベース構築，そのほか種々の研究テーマが山積しており，これらは将来の電子図書館の建設のためにも必須の技術である．

画像処理研究

画像処理の研究も大量画像の取扱いの時代にはいってきた．写真の中の人の数をかぞえるといったことも実用に近づいているが，これも膨大な人の顔の画像を記憶しておいてうまく相関をとることによって検出するという手法が使われている．なめらかな曲面がどのように湾曲しているか，凹凸はどのようになっているかといったことについても，人間は現在の画像処理で行われているように3次元曲面の方程式や反射についての物理学的法則に基づいて頭脳の中で計算しているのではなく，多くの

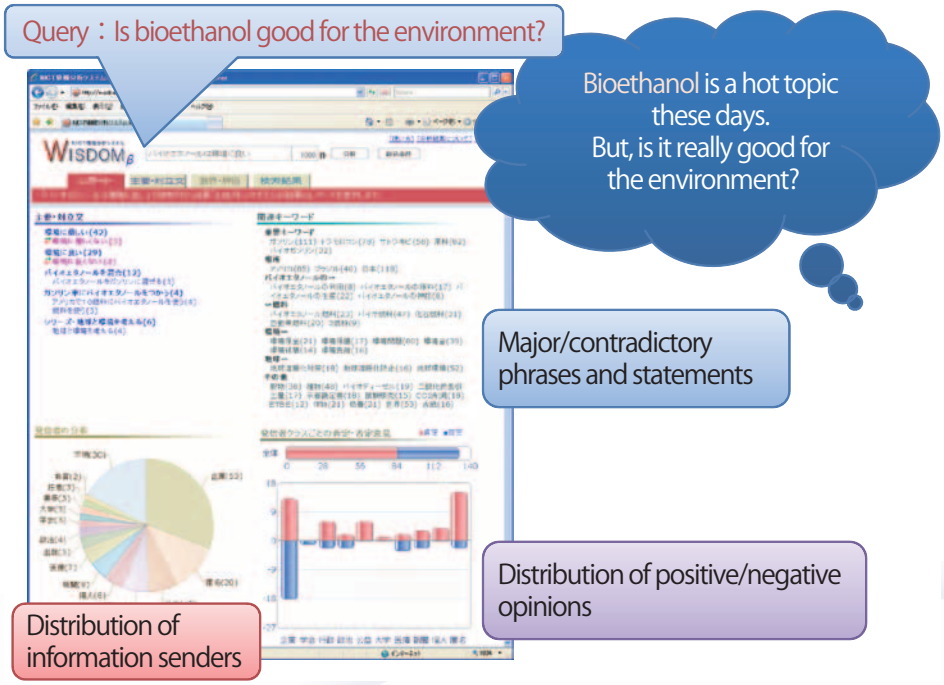

図 -2　検索出力された情報の分類や意見分布の表示（黒橋禎夫氏（京都大学）と NICT との共同研究による）

曲面と光の当たり具合についての膨大な画像を記憶していて，これを超高速に相関計算することによって曲面形状の推定をする方法をとっていると考えてよいのではないだろうか．その場合の計測の精度は数理的アプローチにくらべて格段に低いが，人間は試行錯誤的にやることによって，その精度を向上させて所期の目的を達成しているので，画像処理においてもそのようなプロセスを導入することによって目的を達成することができるようになるだろう．3 次元画像・映像の問題はこのような方法論よりもコンピュータグラフィクスとの関係で考えるのが適切と思われるが，紙面の制約もあって論じない．

データ依存情報処理技術

言語の場合も画像の場合も，文法，規則や物理法則を用いる方法では，そこからはみ出るデータが現れたときのシステムの改善が難しく，最初に戻ってモデルを作り直さなければならない．これに対してすでに持っているデータにならって処理を進める事例ベースのシステムにおいては，新しい状況が生起し処理に失敗したときには，その新しい状況を既存のデータに追加することによってそれまでのシステムの処理にほとんど影響を与えずに新しい現象にも対処できるようになるという利点がある．つまり新しい状況に対する適応能力が高いのである．

言語だけでなく画像などにおいても論理的に扱える部分はあまり大きくない．情報処理の他の多くの問題においてもそうであり，そういった対象に対してはここに述べたような処理方法が 1 つの有力な方法となる．英語でも case-based reasoning という言葉が人工知能研究の分野でも使われていたし，data-intensive research という言葉も見られるようになってきた．これらはいずれも法則的に扱えない微妙な状況に対処するのに適切な手法であって，これから参照できるデータ量が巨大になればなるほど注目をあびるようになるだろう．そのときに巨大なデータをどのように構造化して処理を効率化してゆくかもまた面白い課題となる．

（平成 21 年 11 月 30 日受付）

長尾　真（名誉会員）

mngo@ndl.go.jp

1936 年生まれ．工学博士．専門は，自然言語処理，画像処理，パターン認識，電子図書館．京都大学工学部電子工学科卒業，同大総長（第 23 代），（独）情報通信研究機構理事長を経て，2007 年から国立国会図書館長．