言語データベース統合管理システム

5B-6

小倉 健太郎　　橋本 一男　　森元 逞

ATR自動翻訳電話研究所

1.はじめに

自然言語処理研究にとって、言語データは言語現象解明のための基礎データとして、また、辞書作成の基礎データとして重要である。複雑な言語現象を解明するための言語データとしては、いろいろな側面から言語現象を眺めたデータである必要がある。また、言語データから統計的なデータを抽出し、これを言語処理や音声認識へ応用する立場からは、言語データは大量なものである必要もある。本稿では、ホストコンピュータのRDBとワークステーション上のオブジェクト指向表現を組み合せることにより、言語データベース利用者が使いやすく、大量で多様な言語情報を扱えるようにした言語データベース統合管理システムについて述べる。

2.言語データ

言語データベース統合管理システムで扱うべき言語データについて述べる。言語データベースの利用価値を高めるためには、単にテキストをそのまま格納しておくだけでは不十分である[1][2]。言語の3大要素、構文、意味、文脈を明らかにするためには、単語に関する情報、単語(概念)間の関係についての情報が必要である。また、翻訳を考える場合、言語間の対応の情報も重要になってくる。

言語データの内容は以下の通りである。

(1)文章の構成(全体部分関係)
　(a)会話、(b)発話、(c)文、(d)文節、(e)単語
(2)単語に関する情報
　(a)会話(文章)に現れた単語、(b)その文脈での読み、
　(c)正規表現、(d)品詞、(e)活用型、(f)活用形、
　(g)音便
(3)単語間の関係
　(a)係り受け関係、(b)格関係、(c)接続関係[3]
(4)言語間の対応[4]
　(a)文、(b)格要素、(c)文節、(d)単語、(e)その他

この言語データは、一般性を持たせるため特殊な言語理論には依存していない。品詞体系は学校文法に準拠しており、意味関係も一般なものを網羅したものにしている。品詞体系および意味関係の体系には、階層構造を持たせて、上位の概念を利用した検索ができるようにしている。

言語データは各種情報毎に収集支援システム[5]を使って収集している。文の表現など言語表現を網羅し、言語の統計的なデータに実際上の有意性を持たせるため、言語データとしては、最低100万語集める必要があると考えて収集を進めている。

3.言語データベース統合管理システムへの要求条件

言語データベースを効果的に利用し、管理するためのシステムに要求される条件を以下に示す。

(1)マン・マシンインタフェースについての要求条件
　(A)検索の質に関する要求
　　(a)複雑なデータ、(b)"is-a"階層の利用
　　(c)各種の情報の組み合わせ利用
　　(d)多言語　　(e)統計データ
　(B)ユーザへの便宜
　　(a)データ操作のしやすさ(メニューなど)
　　(b)データ内容、検索結果の見やすい表示
　　(c)言語データ検索条件の容易な記述
　(C)多ユーザ支援
　　(a)初心者から熟練者まで
　　(b)ユーザ個別の要求を充足
(2)データの保守管理についての要求条件
　　(a)大量データ　　(b)データの修正が容易
　　(c)機密保持　　(d)一貫性の保証

また、言語データベースには大きく分けて二つの利用の仕方がある。一つはあるテキストをいろいろな側面から詳しく分析するミクロ分析であり、もう一つは言語現象を全体的な傾向から分析するマクロ分析である。

ミクロ分析は、一般的には分析するデータ量は少ないが、各種の情報を組み合わせた複雑な言語データの検索が必要になる。"の"で結び付けられた単語が、どのような意味関係で結び付けられている時に、英語でどのように表現されるかを検索するなどがミクロ分析の例である。

またマクロ分析には、検索自体は比較的単純であるが、大量の言語データの検索が必要になる。マクロ分析は、言語データの統計的な分析であり、例えば、電話会話のデータに対して、品詞や意味関係毎に頻度を求めることなどが考えられる。

4.システム構成

言語データベース統合管理システムへの要求条件を満足するためのシステム構成について述べる。図1に言語データベース統合管理システムの構成を示す。システムは、言語データ収集システムにより得られたデータを格納管理する部分と、言語データを利用するためのマン・マシンインタフェース部[6]からなる。大量の言語データを効率良く格納するため、ホストコンピュータに、拡張したRDB(Relational Database)を用いて言語データを格納している。複雑で、多様な検索を容易にするためワークステーション上では、言語データはオブジェクト指向の表現になっている。

このシステムは、ミクロ分析とマクロ分析用に、2つのモードを用意する。ミクロ分析モードでは、言語データは予めホストコンピュータからワークステーション上にロードされており、ワー

An Integrated Linguistic Database Management System
Kentaro OGURA, Kazuo HASHIMOTO, Tuyoshi MORIMOTO
ATR Interpreting Telephony Research Laboratories

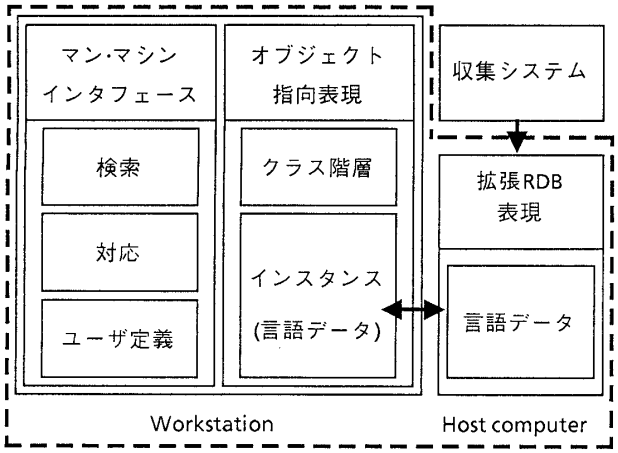

図1. 言語データベース統合管理システムの構成

クステーションの中で実際の検索が行われる。マクロ分析モードでは、実際に言語データの検索要求があった時点で、ホストコンピュータの中で検索が実行され、結果がワークステーションに送られる。検索のコマンドは、形式的には、モードに関係なく同様の検索記述言語が使える。

データ格納に既存のデータベースシステムを改良したものを使い、主にRDBとマン・マシンインタフェースのつなぎと、マン・マシンインタフェースを開発するという方式により、上記の要求を満たすシステムを最初から開発するのに比べて効率的な開発が期待できる。

4.1 言語データのオブジェクト指向表現

複雑な言語構造を直接的に表現するために、ワークステーション上はオブジェクト指向表現にしている。(図2)

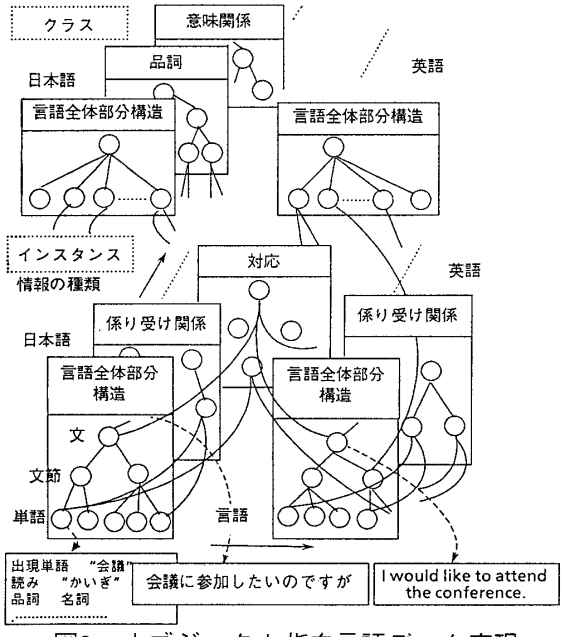

図2. オブジェクト指向言語データ表現

言語データは二種類のデータからなる。一つは、同じ概念の集合を表すクラスであり、これは階層構造を持つ。もう一つは実際の言語データを表すインスタンスである。

4.2 データ格納

収集システムを使って集めた言語データはホストコンピュータのRDBに格納する。全体部分のような関係を効率良く扱えるように、値として多値を許すように拡張する。会話、発話、文、文節、単語、意味関係、構文関係、各レベルの対応関係を関係表(Relation)とし、実際のデータを組(Tuple)とし、発話者、品詞、意味関係名などを属性(Attribute)で表す。RDB表現とオブジェクト指向表現との対応を考えると、関係表が最下位のクラス、組がインスタンス、属性がインスタンス変数に対応している。言語データの修正、機密保持、一貫性、統計処理については既存のRDBに依存することによりシステム開発の手間を大幅に省くことができる。

4.3 マン・マシンインタフェース

言語データの全体・部分関係を直接的に表現でき、上位下位の階層関係を利用でき、容易に各種の情報が組み合わせて利用できるように、ワークステーション上のデータは、オブジェクト指向の表現にしている。また、オブジェクト指向表現を採用することにより、ユーザの言語データへの操作を容易にするとともに、システム開発も容易になる。

マン・マシンインタフェースとしては、検索、対応、ユーザ定義機能を用意している。対応は、日本語と英語を対応をとりながら表示する機能、ユーザ定義はユーザが既存の品詞体系、意味体系など言語データの体系に満足できない場合に、体系の修正や追加をユーザごとに行えるようにするための機能である。これにより、ある特定の言語理論に基づいて言語データを利用したいというようなユーザもこの言語データを利用できる。マクロ分析モードでは、統計処理ための機能を用意する。

5.おわりに

言語データベース統合管理システムを利用することにより、多様な言語情報を組み合わせることによって、いろいろな側面から、言語現象のミクロ分析およびマクロ分析が容易に行える。また、言語データの維持・修正などの管理が容易になる。この言語データベース統合管理システムは、データ格納にVAX8810/ULTRIX、ユーザインタフェースにシンボリックス・リスプマシンを使用する。現在、ユーザインタフェース部のプロトタイプが完成したところである。RDB上の格納およびワークステーションへの転送機能については、現在開発を進めている。

謝辞 日頃御指導頂く当研究所樽松明社長に感謝致します。

＜参考文献＞
[1]森元・小倉・飯田、自動翻訳電話研究用言語データベースの収集について、情処学会第36回全国大会4U-5、1988
[2]篠崎・小倉・森元、言語データベースの品質管理、情処学会第36回全国大会4U-3、1988
[3]井ノ上・小倉・森元、係り受け意味関係の問題点とその考察、信学会NLC87-25、1988
[4]小倉、言語対比データの構築について、信学会創立70周年記念総合全国大会1642、1987
[5]小倉・篠崎・森元、言語データベース収集支援システム、情処学会第36回全国大会4U-4、1988
[6]橋本・小倉・森元、言語データベース統合管理システムのマンマシンインタフェース、情処学会第37回全国大会、1988