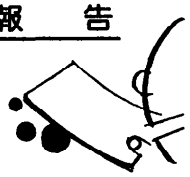
報　告



パネル討論会

自然言語処理の現状と課題

昭和57年後期第25回全国大会† 報告

パネリスト

〔I〕 基礎的な問題　田中穂積[1]，堂下修司[2]，吉田　将[3]，田町常夫[4]，村田賢一[5]，野村浩郷[6]　**司会**　淵　一博[7]

〔II〕 応用に関する問題　石原孝一郎[8]，林　達也[9]，千葉成美[10]，諸橋正幸[11]，森　健一[12]　**司会**　淵　一博

ディスカッサント　辻井潤一[13]

基礎的な研究課題と応用システムの諸問題とに分けて問題提起と討論を行った．まずはじめの基礎的な研究課題に関して，田中穂積氏は，自然言語処理を行うプログラミング言語の持つべき機能を検討し，Prolog が優れていると結論している．堂下修司氏は，自然言語処理の困難さを多角的に論じ，最後に自然言語理解を行うための理論的枠組みと，そこで用いられる推論形式と利用法とを考察している．吉田将氏は，自然言語処理に対して言語表現に密着した意味構造の形式化が必要であることを述べている．そして討論の会場で議論になったサブセット日本語に対する氏の肯定的見解を述べている．田町常夫氏は，機械翻訳システムの実現には，意味，知識の利用だけでなく，原文の情報が必要になることを指摘している．村田賢一氏は，サブセット日本語が具備すべき諸条件を考察している．これは，パネルでも大きな問題となったものである．最後の野村浩郷氏は，我が国と海外の機械翻訳システムの開発動向をまとめ，我が国が今後力を注ぐべき技術項目を抽出している．

次の応用システムの諸問題は，開発の現場におられる方々から，5つのテーマについての問題提起と討論を行った．石原孝一郎氏は，機械翻訳システムには困難な問題が含まれているので，人間の支援を受けたシステムの開発と，研究協力の課題とを述べている．林達也氏は機械翻訳と自然言語による QA システムでは，文脈レベルの処理の問題が将来必要になることを指摘し，日本語のコントロールについての考え方を述べている．会場ではこれに対する議論が活発に行われた．吉田将氏，村田賢一氏の小論を参考に読者自身もこの問題を考えてみられるとよいだろう．

千葉成美氏は，音声認識に固有の問題とその解決技術の現状をまとめ，連続音声の認識には自然言語処理技術との統合が不可欠であると結論している．

諸橋氏は自然言語による質問応答システムの問題として，氏の作成経験に基づいた問題点を実用化の観点から検討している．森健一氏は，自然言語処理では数十万語以上のデータに耐えうるタフなアルゴリズムの必要性とともに，自然言語処理システム開発ツールとしてどのようなものが必要かを論じている．

フロアとの質疑応答の詳細は今回の報告に盛り込むことができなかった．そこで最後に辻井潤一氏からパネル討論に対するコメントをいただいた．

† 日時　昭和57年10月19日（火），9：30～17：00，20日（水），9：00～17：00，21日（木），9：00～16：30
場所　九大工学部・理学部
1) 電総研，2) 京大，3) 九大，4) 九大，5) IPA，6) 武蔵野通研，7) ICOT，8) 日立，9) 富士通研，10) 日電，11) 日本 IBM，12) 東芝，13) 京大

自然言語処理の今昔

淵　一博

自然言語処理に対する最近の関心の高まりは，一昔前に比べると驚くべきものがある．自然言語は，コンピュータにとって夢であるが，とうてい現実にはならないというのが大方の観測であったようだ．

たとえば機械翻訳（MT）がある．20年ほど前に MT のブームがあった．しかし困難性が認識され，その種の研究は計算言語学（CL）に衣がえした．CL はずっと地味な存在だった．ところがこの数年 MT の復活のきざしが出てきた．限定された対象ではあるが，京都大学（長尾研）で開発された科学技術文献の表題翻

訳システムは，文献データベースと結合され，筑波の工技院センターで実用化されている．いくつかの計算機メーカで英和，和英の MT 試作が始められているのは現実的価値を認めてのことであろう．

一方，日本語ワードプロセッサはいままさにブームのなかにある．これは漢字かなテキスト処理のレベルにあるが，むしろそのことが現実の展開を可能にしている．マイクロコンピュータを中心とするハードウェア技術（プリンタ，ディスプレイ，プロッピ等を含む）がそれを支えている．ソフトウェア技術の進歩もある．しかし自然言語処理の水準で言えばあまり欲ばっていないところに良さがある．

ワードプロセッシングにしても今後もっと欲ばりたくなるかもしれない．しかしそこには大きな落し穴がある．ギャップをとびこすのはそう簡単ではない．MT にしても，その本質的な困難性が解決されたわけではない．そのギャップの手前で努力する余地があり，またその現実的意義もある．それとともに，そのギャップをこすための自然言語処理の研究もこの 10 年厚みを増している．ワードプロセッシング的技術はそういう研究の大きな助けになるツールでもある．ギャップをこすのは一足ではないが，一昔より楽観してもよさそうである．

本格的な自然言語処理（文法的処理，意味的処理）をとり入れるには，さらにいっそう，言語の本質にせまる研究が必要である．それを促進するには良いツールの整備がまた大事なことである．そういうツールとして，マイクロコンピュータの意義は大きいが，将来，現行のものの改良だけでなく，自然言語処理にもっと適したプロセッサを，自然言語処理の研究の方から示唆する可能性はないだろうか．その可能性は大いにあると見たい．（そのステップの一つは Prolog マシンのようなものであろう．）自然言語処理専用マシンというのは，ところで「専用」なのであろうか．高機能の記号処理を高効率で実行するマシンということから，それはむしろ，汎用的なマシン（新型コンピュータ）ということになるだろう．

[I] 基礎的な研究課題

自然言語処理技術とプログラミング言語

田中 穂積

自然言語処理にどのような機能をもつプログラミング言語が適当かという問題について，non-determinism の観点から考察してみたい．

自然言語処理では，構文解析，意味解析，さらには談話解析が必要になる．ここでは，構文解析と意味解析のなかから，幾つかの例をあげ，non-deterministic programming の有効性を述べる．

自然言語の構文解析で用いる文法規則が非決定性を有することは良く知られている．たとえば日本語で，(1) $NP \rightarrow$ DET ADJ N (その赤い箱)，(2) $NP \rightarrow S$ NP (彼が見た人) などの文法規則は，矢印の向きに従って左から右に，トップダウンに処理が進む場合には，文法規則の適用に非決定性が含まれる．depth-first なら，(1) の適用に失敗すれば，(2) の規則が適用される．ボトムアップの場合でも，$PP \rightarrow NP$ P (出勤が)，$S \rightarrow NP$ Aux (出勤です) の規則があれば，これらの適用にも非決定性が含まれる．さらに $N \rightarrow N$ の N，という規則により，「N の N の N」という文は「(N の N) の N」とも「N の (N の N)」とも解析できる．これは再帰的な規則に含まれる非決定性の 1 つである．

意味解析ではフレームという意味表現の枠組みがよく使われる．フレームはいくつかのスロットから構成され，そこには，スロットを満たす時に必要なさまざまな制約条件が記述されている．格文法で用いるフレームでは，動作主格というスロットには有生という意味マーカをもつ名詞句がフィラになりうる等々である．意味解析は，これらのスロットを順次調べて，スロットに付与された制約条件を名詞が満たすかどうか調べるのである．この解析過程は，非決定的な解析過程として容易に実現されよう．

一般に，言語学的には，1 つの構文解析木には，1 つの意味が対応していると考えられている．構文解析木の姿が異なれば，通常それらの意味も異なると考えられているのである．ところが機械処理の立場からは，構文解析木が 1 つでも，これから異なる意味解析結果が得られることが望ましいことがある．たとえば，「猫を買い与える女は美しい」という文は，[$_S$[$_{PP}$[$_S$ 猫を買い与える] 女は] [$_{pred}$[$_{adj}$ 美しい]]] と構文解析できる．ここで，「女」が「買い与える」とどのような格関係を有しているかによって，2 通りの意味解析が可能である（「女が誰かに猫を買い与える」，「誰かが女に猫を買い与える」）ことに注意してほしい．連体修飾される名詞が，連体修飾する用言と，どのような格関係を有するかまでを構文解析で決めることは一般にしないし，それらを木構造の形状の相違

として反映させることもできない．そこで，構文解析結果（木）がたとえ1つであっても，このような埋め込み文の場合には，異なる意味解析の結果が得られる機構が必要となる．

以上は，non-deterministic programming によって容易に実現できる．したがって，non-deterministic programming 言語の1つである Prolog は，自然言語の解析に適当なものの1つであると考えられる．non-determinism を逐次処理で実現する場合には，バックトラックの考え方が必要になる．かつて自動バックトラックの害が強調されたことがある．筆者は，基本が自動バックトラックであることはむしろ良いことだと考える．基本的には自動バックトラックで，必要が生じたらそれに制約を付加し，知的なバックトラック機構を開発する研究を積み上げればよいのだから……．

ロジックと自然言語処理技術

堂下 修司

自然言語の研究は，近年急速に発展したが，それが十数年前に比べて，自然言語に関する理論の進歩や人工知能的手法の進歩によるものか，または主として，計算機 tool の高速化・大容量化に伴うものなのかは意見の分れるところである．ここでは，自然言語処理について，論理的側面と実際的側面の関係を中心に述べる．

形式言語理論（オートマン理論），生成文法，変形文法，格文法，λ 範疇文法，モンテギュ文法などの理論は，一定範囲の言語についての論理的構造を与えてくれる．自然言語の処理は，プログラミング言語の上で行われるが，記述性の良い明解な自然言語処理系を構成するためには，自然言語のモデルとそれを扱うプログラミング言語とが整合していることが好ましい．文脈自由文法と Prolog，λ 範疇言語と Lisp などはよい例であろう．このことは自然言語とプログラミング言語の相違性を明らかにするためにも有用である．しかし一方，そのようなモデルだけで自然言語が自然な形で処理されないことは，種々の実例が示している．自然な処理のためには数万語の単語辞書，数十万件の用例集など言語の個別的知識とともに上述のような言語系と，その言語の文章が述べている対象世界に関する知識やその上に展開される推論との関係を明らかにしなくてはならない．現在の多くの本格的な言語理解システムや機械翻訳システムは，記号列としての言語系のなかで閉じているが，言葉を他の言葉で表現し，説明するだけでは意味を与えたことにならない．プログラミング言語と同様に，言語の意味はインタプリタを介して対象世界の上に定めなければならない．しかし，自然言語系においては，いわゆる semantic gap はプログラミング言語に比べて格段に大きく，言語世界と対象世界の統合は容易ではない．多分，この結合はコモンセンス的世界に対しては不可能に近く，特定の領域に限定したエキスパート的世界に対して試みられるであろう．

つぎに，自然言語の処理系について考えよう．自然言語は人工知能の研究対象として重要な分野であり，逆に人工知能的手法は自然言語処理に欠かせない．一口で言えば，人工知能系は，推論系であるが，それはまた「人工的に規定された有限記述の処理系が，多くの例外・あいまい性・誤りを含み部分的にしか明示化されていない自然的な対象（自然言語や各種のパターンなど）や，明確に規定されてはいるが，本質的に非決定性の処理が必要な対象（証明・演繹システムなど）に関する処理」を行うことである．現時点では自然言語そのものに関する人工知能的処理は，限定されている．今後，言語系に関するより高度の意味的処理および対象世界に関する推論系との結合のためには，自然言語処理にもっと人工知能的手法を取り入れる必要があろう．

もう一つ，自然言語の処理を複雑にしているのは，同じ対象に対しても，その理解者（話し手・聞き手）により，捉え方・言語的表現・評価等が異なるということである．対象世界・言語系・推論系はそれぞれ共通的・一般的意味をもってはいるが，それは部分的であり，大きな自由度を残している．その自由度に対して，個々の理解者は異なった view をもち，個別的意味を構築している．

自然言語の理解においては，自然系と人工系の関係にも留意する必要がある．機械が自然系を自然なまま扱うことは理論的に不可能であり，なんらかの形で有限記述化しなければならない．自然系と人工系の間には無限のギャップがあり，有効な機械処理系を構成するには，人工系を段階的に拡張し，本質的な自然性を保持した，ゆるやかな枠組をもった「擬似自然系」を設定する必要がある．この擬似自然系と真の自然系の間には，なお越え難いギャップが残るため，自然系の理解には，人の高度の介入が不可避であろう．図-1 は，このような観点から，自然言語理解系を図式化し

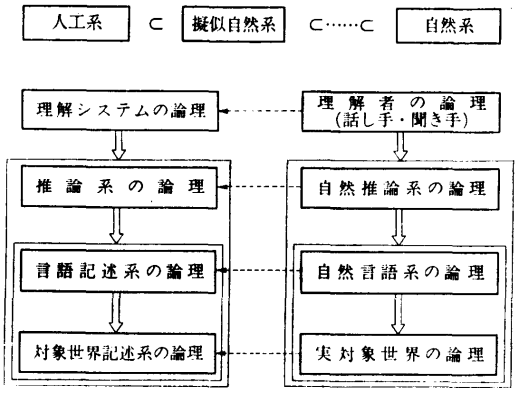


図-1 自然言語理解系の構成

たものである．図において「論理」という語を用いているが，自然言語における記述では，論理的道筋の正当さよりも，推論に用いられる前提や推論規則，その結論の当否，その意図・評価などに主たる関心があり，特に日常的文章ではその傾向が強く，推論法や手続きの記述を主とするプログラミング言語と異なる点に留意する必要がある．

このようなシステムに用いられる推論は多様なものとなるであろう．通常推論は推論形式に従って，類推型，帰納型，演繹型などに分けられるが，ここでは以下のように推論の目的に従って別の分け方を試みる．

(a)* 目標発見推論——多様な状況や知識の集まりから，一定の基準に従って自律的に発想し，目標を設定する．

(b)* 目標指向型推論——与えられた目標を達成するために，状況や知識を整理・統合し，新たに獲得し，それらに一定の枠組を与える．

(c)* 目標到達型推論——与えられた枠組の下で，与えられた目標に到達するための筋道を見出す．

(d) 計算型(手続き型)推論——過去の推論によって得られた筋道（すなわち手続き）をたどり，与えられた入力に対し，効率的に所期の結果を得る．

(e)* 経験探索型推論——過去の推論によって得られ，蓄積されている多種多様な知識を組合わせ，与えられた状況に対して，最も適当と思われるものを探し出す．

(f) 再生型推論——個々の状況に対しほぼ1対1に対応する過去の事例に基づき，与えられた状況に対する直接的な答を見出す．

人の場合，これらは一体となっており，まず最も単純で容易な下位の推論を試み，十分な答が得られないときに，より複雑な上位の推論を試みることにより，全体として効率的な推論系を構成していると考えられる．あるレベルの推論の結果はそれより下位の推論の知識として蓄積されており，上位の推論に際してはそれらの知識を十分に取り入れている．＊印を付した推論(a)，(b)，(c)および(e)は，非決定性をもつが，特に(a)や(b)では多重非決定性をもち，「非決定性による複雑さの爆発的増大」の危機にさらされている．それを実行可能な程度に下げるには，システムは，推論の一般的メカニズムをもつのみでなく，それを有効に制御するために，下位の既得の知識の全面的援用が必須である．機械に人間に近い推論を行わせるためには，これらの推論系が有機的に結合されること，長期間にわたり知識の集積を行うことが不可欠であるが，現時点ではいずれもまだ不十分である．

このように述べてくると，自然言語の処理に論理はあまり役立たないと思われるかもしれない．しかし，もし中核的な処理をカバーする論理的枠組をもたないで，言語を数万・数十万の個別的知識のみで網羅的に処理すれば，収拾のつかない，時間的・空間的に発散するシステムとなるであろう．言語の中核部分を論理的モデルの下に処理し，その核処理系の下に個別的言語知識を附加することによりパフォーマンスの向上を図ることは，見通しの良い，コンパクトな，拡張性に富んだシステムを構成するために必須のことと思われる．このことはまた，コンパクトで能力の高い擬似自然言語系の設定にも役立つものと思われる．

有用な自然言語処理システムを実現するためには，対象分野を限定し，一般用単語辞書とともに，その分野に関する十分な専門用語辞書を用意し，その分野に固有の知識に基づく推論と組合せて十分な意味的処理を行うことが当面の課題であろう．

意味構造の形式と体系

吉田 将

“意味とは何か，意味の捉え方としてどのようなものがあるのか，それらはどのように形式化されるのか，それらの形式化はどのような立場から体系付けできるのか，等について意見を述べよ．また，それぞれの形式化が自然言語処理の問題においてどのような利害得失を持つのか，今後の研究方向はどうあるべきなのか，等について意見を述べよ”，ということが，私に対して与えられた課題だろうと思う．

これに正面から答えるためには以下のキーワードを

解説し体系化すれば良いであろう．

セマンティック・マーカ，セマンティック・ネットワーク，概念依存表現，……，生成文法，モンテギュ文法，スクリプト，談話構造

しかし，以下で自然言語処理に意味を導入する場合の問題点について，断片的な主張をしたい．

次のような言葉をしばしば聞く．“自然言語処理は次の段階に分けることができる．形態素解析，構文解析，意味解析，……，構造変換，文合成．各段階における処理がそれぞれあいまいさを残している．このあいまいさは構文解析，意味解析へと段階が進むにつれて減らすことができる．すなわち，形態素解析におけるあいまいさは構文解析や意味解析の段階において減らすことができるし，構文解析におけるあいまいさは意味解析の段階において減らすことができる．……．”

“意味を考慮に入れればあいまいさを減らすことができる”という主張は本当だろうか．

“すももももももものうち．”という文の解釈として，“李も桃も桃のうち．”は正しいけれども，“酢も藻も桃も桃のうち．”は酢も藻も桃のうちではないので正しくない．したがって，後者のような解釈は採用しない．

この判断は一見，もっともなように思える．しかし，このような判断を下す過程において，われわれ人間は多くの推論過程を見過し，また多くのあいまいさの存在を見過している．

国語辞典によると，桃と李は次のように説明されている．

もも(桃)①……落葉樹，実を食べる．②桃色．

すもも(李)……白い花を開く落葉高木．

もも(桃)は“樹”，“色”などの意味をもっている．つまり，“桃のうち”の桃は樹であるかもしれないし桃色であるかもしれないことになる．したがって，“酢”も“藻”も樹ではなくまた桃色でもないことを調べなければ，“酢も藻も桃のうちではない”という判断を下すことができないことになる．しかし，実際には，“AもBももものうち”という言い方をする場合，“もも”は桃色の意味ではあり得ない．この判断は意味に基づいてではなく言語表現に基づいて行われたものである．つまり，意味を導入しただけではあいまいさを減らすことがむずかしく，言語表現が自然かどうかということをも含めて，はじめてあいまいさを減らすことができるというわけである．詳しく意味を識別すればするほど，あいまいさの組合せ数が増大する．この増大を食い止める手段として意味による言語表現の違いを利用すべきだというわけである．“意味”と“言葉”とは独立なので当然のことである．

やや結論を急ぐが，

① 意味の違いを識別すればするほど，あいまいさの組合せが指数関数的に増大する．したがって，

② 言語表現の違いを使って意味の違いを識別する方法の研究にもっと力を注ぐ必要がある．

③ 意味の導入の詳細さの程度に見合った，あいまいさを解消（減少）するための評価モデルを与えなければならない．

④ 構文処理までの段階においてできるだけあいまいさを落しておくべきである．そのための有効な手段を考える必要がある．

私はまだこれらのことについて答えをもっていない．ただ次の点を主張したい．現在の人工知能的自然言語理解のやり方は言語表現と意味との関係を単純化し過ぎてしまっている．自然言語処理の立場からは，もっと言語表現と密着した意味の表現が必要である．それを実現するための膨大な仕事量を覚悟すべきである．

人工知能の研究は知識の表現を重要なテーマとしている．そこでは，知識を表現するための道具立て——表現の形式と枠組——の研究が主体となっており，その枠組の中に記述する内容を，各自が興味をもっている各々の限られた世界の限られた“ものごと”の人工的な意味に限っている．しかし，自然言語処理の立場からは個々の言葉の意味をどう表現するかを問題にしなければならない．個々の言葉の意味を表現する作業は膨大ではあるけれども，その作業を達成することは可能であると私は思っている．われわれはそう信じて研究の力点を“語の意味そのものの記述”に置いてきた．そこがアメリカを中心とした人工知能的知識表現の研究とは異なっていた．しかし，われわれが求めた意味構造の形式的側面は結果的には彼らの形式とほぼ同じであった．

話題は私に対して与えられたテーマからそれるが，パネル討論の会場において議論になったサブセット日本語のことに触れたい．種々の意味において，制限日本語は必要であり，かつそれの実現は望ましいことであると私は思っている．要は人間にとっても機械にとっても，

■ あいまいさが少ない（できるならばない），

■ 平易な文である，

■ 記述力が充分である,
■ 自然な表現ができる(言語として不自然でない),
■ 他の言語へ翻訳しやすい,

という性質をもった日本語を開発せよということである.

このような日本語を実現する手段として語彙,構文,文体に制限を置こうとするものである.このことは新しい可能性の追求であり,決してわれわれの自由を奪うものではない.自然言語においては漠然さはつねにつきまとっている.しかし,情報伝達の道具として使用する自然言語においては,あいまいさはなくさなければならない.機械にも外国人にも情報を伝えなければならない時代である.もはや,日本語は日本人だけのものではなくなってきている.情報化時代にふさわしい日本語の開発・国際化が必要である.(なお,本文はゆるく制限した文法に従って書いた.)

機械翻訳システムの諸問題

田町 常夫

機械翻訳はその歴史も古く,自然言語研究の動機ともなったものであるが,問題の困難さのために一度はあきらめられる時代もあった.その後言語構造に関する基礎的な取組みが次第に進み,一方処理機械のハード,ソフト面での進歩もあって,今日では機械翻訳のための道具立てはかなり進んでいるといえる.しかし一般に考えられる機械翻訳の疑問点は依然として残っており,翻訳システムとしての実現には問題が多い.

システムとしては入出力,辞書,アルゴリズムそれぞれの問題があるが,入力対象,出力に対する要求,操作条件等に依存して考え方がかなり違ってくる.全般的に翻訳の質を高めることがこれからの問題であろうが,文の種類によってはあまり意味を考えなくてもよい場合もあるし,また意味の深層までさかのぼってもあいまいさの処理に役に立たない場合もある.そうかといってあいまいさのなかから無理にどれかを選択することは客観性を欠くことにもなりかねない.ここで言いたいことは,一口に翻訳と言っても目的,対象によってかなり処理方法が違ってくることになるであろうから,まず目的を明確にとらえる必要があるということである.

つぎに翻訳の質の問題に関連して,機械翻訳では意味や知識がどのように必要なのかという問題がある.機械翻訳の言語処理上の特色は入力言語から出力言語への情報の受渡しの過程にある.したがって入力言語の分析の段階では他の言語処理と同様,最低限,概念間の依存関係を知るための概念の内包的意味が必要となるが,そのほかに意味構造の変換のために少なくも訳文の側の概念の依存関係を得るための推論が一般に必要である.

この変換の過程は,今日ではトランスファ方式や中間言語を介する方式によって実現しようとしているといえる.

意味構造論の立場からは,意味的異常性の検出が構文のあいまいさの減少に役立つことになる.また同義性の問題は定義の仕方によって文の標準化に有効な役割をもつことになろうが,翻訳の場合にはこれが訳文のあいまいさをふやす条件となって悪い方向に作用する.中国言語を介する翻訳ではこのことが問題であり,適訳を得るためにはどうしても原文の情報が必要になる.原文の構造に含まれるこまかなニュアンスは語用論の問題も多くて,これを意味構造に反映することは難しいが,かりにそれができたとしてもこれを多国語間に通ずる完全な中間言語として用いることはまず無理である.ここでも意味構造のレベルでのトランスファが必要になると考えられる.現在のトランスファ方式は,こまかなニュアンスをすべて原文の構造で代表させて変換を行っていると見ることができ,その意味では妥当な考え方であるといえる.

サブセット日本語

村田 賢一

0 限られた語彙で平易な文章を作るという運動は,元来,教育的ないしは社会的配慮から生まれてきたもので,相当の実績もある(例えば Basic English, VOA 等).このように語彙その他に人工的制限をつけたものを,ここでは「サブセット英語」等々ということにする.またその総称には「サブセット自然言語」を用いることにする.

1 電子計算機に自然言語を「理解」させようという大変困難な課題を少しでも容易にするために,サブセット自然言語を採用してはどうかということは,誰しもが思いつくことではある.

しかし,ちょっと考えてみると,問題は決して単純ではない.まず既存の Basic English のようなものを使うことを考えると,(日本語でどのようなものが存在するかはさておき)実用されているというのは大変な強みではあるけれども,電子計算機側の困難さをどの程度軽減してくれるかがはなはだ疑問なのである.

幼児の絵本のように内容も単純なものならば良いが，大人が普通に（英語の例で）5〜6千語程度[1]の語彙を用いて表現している内容を1,000語程度の制限語彙で表現しようとすると，どうしても語ないしは語の組合せ（熟語）のもつ多義性を活用せざるを得ず，計算機の方はこのような文を理解するために，文脈などの高度に知的な情報と，熟語等に関する豊富な知識を用いて一義化しなければならない．これは電子計算機にとり大変な負担である．一例をあげれば，turn in という熟語には十数個の互いに全く異なった意味があるが，これがどのような環境でどの意味に使われるかを計算機に教えこむことはそれ自体大変な仕事である．

2 所期の目的を達成するには，結局，計算機用のサブセットを新規に設計しなければならない，ということになろう．そこで，設計に際して考慮すべき諸条件について考察を加えたい．特に設計にまつわるトレード・オフ関係および実用化のための具体的問題に触れたい．

3 電子計算機による自然言語理解において最も困難な作業の1つは，無数の可能な解釈のうちから正しい解釈を選択する過程であろう．そもそも可能な解釈が無数に表われるという現象自体回避したいところである．そのためにはサブセット内部に一義化のための強力な機能があることが望ましい．なお，語彙が多いこと自体は電子計算機にとり致命的な問題ではないであろうから，語彙の数量的制限にはあまりこだわる必要はない．必要なのはむしろ用法的制限である．以上を踏まえて，サブセットが具備すべき諸条件を列挙してみると，大略次のようになろう[2]．

① 文章レベルにおける多義性が存在しないこと．

② 文レベルにおける多義性がほとんど存在しないこと．

③ 語彙レベルにおける多義性，多品詞性が最少限であること．

④ 上記において多義性が存在する場合も，環境によって意味が一義化される過程が，現在の情報工学技術で処理可能な程度に，単純かつ明快に与えられていること．

具体的に言えば，Basic English 等の行き方とは対照的に，語彙を増やし，その代りに極力1語1意味になるよう豊富な語彙を使い分ける，ということである．もちろん完全に1語1意味というのは非現実的だから，④の条件が満足される場合には多義性を許すことにする．

4 しかし，この方針をあまりにも徹底して貫くと，極度に人工的なサブセットが出現し，自然言語とは言い得なくなる恐れがある．そこで，ここは極めて難しいところであるが，第5の条件

⑤ 自然言語のサブセットとして，その使用者に受け入れられ得るものであること．

を導入し，実際の設計に当たっては電子計算機の側の都合である①〜④と，人間の側の都合である条件⑤の間のトレード・オフを重視する必要がある．

5 最後に，実用化のための具体的問題について触れたい．言語のフォーマルな定義は，普通次の3項目を与えることであるとされている．

ⓐ 語彙項目リスト

ⓑ 統語規則

ⓒ 意味解釈規則

しかし，実用化のためにはもう1つの重要な側面が残されている．それは，「良い文章を書く技術を訓練により徹底させること．」である．「良い文章」という価値判断は立場により異なるが，電子計算機側の都合からいえば，それは

㋑ 共有知識（特に当事者だけに共有のもの）に依存する表現や省略表現がなく，明示的で一義的な文章．

ということになろう．他方，人間側の都合から言えば

㋺ 必要十分な情報だけを与える文章．

ということになろうか．㋑と㋺は相反する条件であるから，再び適切なトレード・オフが必要になる．

1）日本語の場合には，もっと多い．

機械翻訳システムの開発動向

野村 浩郷

ここ1〜2年間の日本における機械翻訳の研究・開発の状況には目をみはるものがある．取り組み機関も大学研究所およびメーカに及んでいる．科学技術庁の文献抄録翻訳プロジェクトのように国家的プロジェクトもある．海外では，EC の EUROTRA プロジェクトが正式に承認され，1983年初頭から，EC 内の公文書を7カ国語に翻訳するためのパイロットシステムが5カ年計画で開発される．注目すべきは言語学，心理学，および計算機科学の分野の研究・技術者の協調が急速に進んでいることである．産業的にみるならば，翻訳は，一千億産業と五千億産業ともいわれている．

機械翻訳システムの研究・開発がこのように盛んに

なったのは，国際化社会における需要の大きさもさることながら，日本語を含む翻訳処理技術が将来の高度に知的な情報処理技術の根幹技術に関連するとともに辞書などの基本データの着実な蓄積をもたらすからである．高度な翻訳はまた頭脳の高度な知能機構の解明にも関連している．にもかかわらず，研究の過程で得られる個別的技術は，高度な OA システムやマンマシンインタフェースとして段階的に実現可能である．

国内の機械翻訳実験システムの数は少なくとも 17 あり，欧州全体のそれに匹敵するかそれ以上である．その内訳は，日英翻訳が 7，英日翻訳が 10 である．論文標題の英日翻訳システムは実用化されている．実験システムの数は近いうちに 20 を越えるであろう．翻訳対象は計算機などのマニュアルが圧倒的に多い．文法や語彙を制限するコントロールド言語のみを扱うことを目的としたシステムは少なく，現在は，研究の途中段階としてそうなっているものが多い．

現在までのシステムの大部分は研究室などで作成された実験システムであり，真の意味での実用化までに解決されなければならない課題は多い．研究方法を分類すると，(A)高度な意味処理技術の確立を目指す知能処理の研究を主とするもの，(B)確立しつつある構文解析技術の応用を主とするもの，および(C)人間の研究・作業環境の改善の一貫として可能な範囲で翻訳サポート機能を入れていくものに大別できる．研究としてはAが，実用化としてはBが多い．Cは辞書さえあればすぐにでも初歩的機能が実用化でき，しかもそれ自体有用なものであるが，本格的な取組みは全くないといってよい．

機械翻訳システムの開発に必要な技術は，(1)言語解析技術，(2)言語構造変換技術，(3)言語生成技術，(4)文法や意味や知識の表現技術，(5)システム作成や実験や運用を行うためのマンマシンインタフェース技術である．さらに，これらを動かすための計算機システム，ファイルシステム，端末装置，OCR 装置などがサポートされなければならない．また，プログラム記述のためのプログラミング言語が必要である．現在までの多くの機械翻訳システムは，アセンブラ，FORTRAN, PL/1, Lisp, Prolog, C などで記述されている．

言語解析技術は，句構造に基づく構文解析，ATN に基づく構文解析，格解析，文形マッチなどに大別される．構文解析だけでは不十分であるので意味カテゴリのチェックなども行われる．解析技術として重要なものは意味解析技術であり，これは現在の研究の中心である．意味理解技術は知識を使って意味を理解することであるが，これは今後の課題である．その他に文脈処理も今後の問題である．言語構造変換技術は，木構造変換とフレーム構造変換などに大別できる．言語生成技術は，句構造からの変形生成やフレーム構造の線形化などに大別できる．言語構造変換および言語生成ともに，あまり研究が進んでいない．

文法や意味や知識の表現技術は論理式，プロダクションシステム，セマンティックネットワーク，フレームなどの技術であるが，今後に残された問題である．

以上の他に(6)辞書，(7)言語データ，(8)知識について考えねばならない．前述の諸技術がいくら進んでもこれらがないと言語処理は不可能である．しかし，逆に，これらのデータがそろっていると前記諸技術は未熟でも何がしかの処理は可能である．日本においては，特に，これらへの取組みが不足している．辞書は解析用辞書，語彙変換用辞書，生成用辞書が必要であるが，いずれも満足なものは作られてない．言語データとは，文法や格フレーム，イディオムなどである．構造変換用辞書も含まれる．辞書は人間用の辞書を基本にしてある程度のものが作れるが，言語データは言語分析などにより集めなければならない．意味カテゴリやシソーラスなども作らねばならない．語と語の連接条件なども集めなければならない．これらは言語的知識ともみれるが，一般常識や専門知識に相当するものが，いわゆる知識である．ターミノロジもこれに含まれる．外国には LEXIS や EURODICAUTOM などの大規模なターミノロジバンクがあるが，日本にはない．ターミノロジバンクは変換のときに，専門用語の訳語を定めるのに用いられる．このときの専門用語とは辞書にない語ということになる．しかし，本来の専門用語のもつ役割は異なるであろう．意味を理解するときに参照する場合には，用語は知識として扱わなければならない．これに比べて，一般常識は，辞書にある言葉であるが，辞書に記述されていない知識とでも考えることができる．これらの区別を明確にしようとすることはほとんど意味がない．いずれにせよ，辞書や言語データや知識は知識ベースとして発展させられるべきであろう．

現在までの機械翻訳実験システムは翻訳能力を評価するところまでに至ってない．すなわち，実験室用のサンプル文についての翻訳が主に行われており，それらはシステム作成時に参考にした文であるものが多

い．そのような文の数が十分多いときにはある程度の評価はできるが，現在のところ少ないようである．日本における機械翻訳システムの技術的進歩は，ここ当面，辞書作りと言語データ収集の度合いに比例するであろう．

[II] 応用システムの諸問題

英日機械翻訳システム

石原孝一郎

1. 機械翻訳実験システム

多くの英日翻訳実験システムの翻訳過程は大略次の通りである．

（1） 辞書検索：英文を入力し，単語やイディオムを調べ，その品詞などを知る．

（2） 構文解析：単語から句を切り出し，単語や句に対して，主語，述語，目的語など構文上の役割りを決める．必要ならさらに節を切り出し，単語，句，節の構文役割りや修飾関係を決定する．

（3） 文型変換：文の構造に注目して，できあがった英文解析ツリーを日本語に適した語順にならべかえる．

（4） 日本文生成：単語や句に対訳をあてはめ，助詞を補ったり，語尾変化をさせて日本文を作成する．

以上の手順は，通常人間も行う常識的な手法である．解析の過程では，（たとえば名詞＋コンマ＋名詞という並びがあれば，後の名詞は同格で並置されているといったような）文法規則や（冠詞の後に動詞はこないというような）禁止規則を記憶していてこれらを活用して文の解析を行っている．このような構文解析が現在は主となっているが，意味解析も部分的に取り入れられつつある．たとえば，次の２文は，構文上は

① I bonght a car with 4 wheels.

② I bonght a car with 4 dollars.

まったく同じ形をしているが，前置詞句の役割りや修飾先が異なっている．これを区別するには，wheel が car の部品であることと，dollar が金額であって buy の手段であることを知る必要がある．われわれのシステムでも名詞をいくつかの種類に分類し，前置詞との結びつきにおいて構文役割りや訳語の決定に用いている．前置詞句は構文役割りも多様であり，訳しわけも難しいので，力を入れて解析しているが，現状の構文解析主体の解析法ではいずれも限界がくるものと思われる．

2. 機械翻訳の課題

上述したごとく構文解析には限界があるので，単語や句のもつ意味情報を手がかりとする意味解析に多くの研究者が取り組んでいる段階である．これ自身，挑戦的な課題であるが，さらに一つの文の中だけで意味が決定できないことが多い．次の文をみてみよう．

③ I got at it.

この文単語ではほとんどが意味をなさない．仮に前後の関係から is が book（本）を受けていることが分ったとしてもなお，got at の訳として「手が届く」「発見する」「理解する」などが考えられる．これを区別するには，前後の文脈でなにが問題となっているかを理解しなくてはならない．また話題に関する広い背景知識が必要となることもある．このような人工知能の領域にまで深入りするなら，ここ10年ぐらいの間に有効な解を見出すことは恐らく不可能であろう．ただ機械翻訳の有利な点は，翻訳出力を人間が解釈するので，多少のあいまいさやまずい点があっても，人間がこれを是正してくれる可能性があることである．

したがって当面の機械翻訳システムは完全な自動翻訳をねらうよりも，人間よりも多くの文法規則や語意を正確に記憶でき，校正や清書機能のすぐれた機械と，パターン認識に秀で，莫大な知識を駆使できる人間との長所をうまく組合せたトータルシステムを開発していくのが課題である．

3. 今後の展開

このように多くの課題を抱える機械翻訳をさらに発展させるために，関連する研究グループ間での良い意味での競争とともに，ある面での協力が必要となろう．望ましい協力分野を２，３述べてみたい．

（1） 意味体系の基礎と辞書構築：現在の構文解析技術が既存の文法理論に立脚しているのと同様に，今後の意味解析にもある程度統一された意味コード体系が必要となろう．そのような体系に基づいた辞書の構築と翻訳ソフトウェアの開発を分離することにより一層の技術の発展が期待できる．

（2） 翻訳の評価基準：何事であれ，結果を客観的，定量的に評価することによって進歩改良がもたらされる．翻訳結果を評価するのが大変難しいことは，われわれの行っている４段階評価で，人によって２段階も異なることがあることからも理解できる．人工知能の分野で成功しているチェスをするプログラムは少なくとも相対評価が非常に明確であることも，成功の一因と考えられる．翻訳においても，なんとか共通

性，普遍性のある評価基準を整備していくことが望まれる．

（3） 科学技術分野における共通語，共通文法：ニュアンスや行間の意味があまり問題とならない科学技術分野は，当面の機械翻訳の有力対象と考えられている．せめてこの分野において英語にせよ，日本語にせよ，機械処理を念頭においた擬似自然語または制限文法があれば望ましい．しかしこれには今後に多くの課題が残されている．

日英機械翻訳システム

林　達也

われわれは，自然言語処理として機械翻訳，DB のフロントエンド等の研究を行っているので，機械翻訳それ自体の話と同時に，他の自然言語処理と対比させた形で問題点を見てみることにし，最後に全般的課題について考えてみたい．

1. 機械翻訳システムの現状

現在開発が進められている機械翻訳システムの現状を見てみると，これらのシステムは適用対象によって二つのグループに分けることができる．第一のグループは，科学技術文献の表題，抄録，社内連絡文等のどちらかといえば翻訳家が乗り出すまでもない分野を対象とするもので，量的に短く文型が限定されているかあるいは訳文の品質が悪くても情報伝達の迅速性の方が重要視されるものである．例に挙げたシステムは日英相互翻訳であるが，一般に英日による情報収集のみに専念していると海外との摩擦がつのる一方であるから，日英による情報提供も今後もっと考えていかなければならないだろう．

第二のグループは，マニュアル，科学技術文献，行政文書等のこれまで専門の翻訳家が念入りに作業を行っている分野を対象とするもので，これに機械で支援しようというものである．

全体としてみると，第一のグループは Machine Translation，第二のグループは Machine Aided Translation を目指しているといえよう．

2. 問　題　点

一般に日本語が絡む場合の問題点について考えて見る．図-1 において，日本語はよくいわれるように暗示的な言語であり，当事者間に状況が一度設定されるとその場の効果を最大に活用して，必要最少限のことしか明示的に表現しない．

そのため主語や目的語はよく省略されるし，「僕はうなぎだ」といった類の言い方もなされる．それでもコミュニケーションが成立するのは，「場の効果」という氷山の水面下の部分が大いに物を言っているからである．

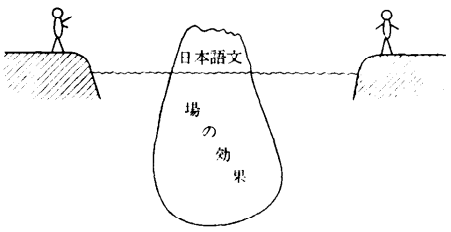


図-1　日本語によるコミュニケーション

機械処理の場合，水面下の部分を把握するのは一般に困難（翻訳の場合でいえば構文や意味レベルを跳び越えていきなり文脈知識レベルの技術が必要とされる）であり，水面下の部分を浮上させるなんらかの措置が当面必要となる．

そこで特に日本語が絡む場合の自然言語処理のポイントとしては，次があげられる．

- 自然言語でなく自然な言語
- 意訳でなく直訳
- 言語間マッピング（自然言語対自然言語，自然言語対機械語）の枠組

3. テーマ別課題

上述の問題点を背景に機械翻訳，OB フロントエンド，といった自然言語処理のテーマ別に技術的課題を整理してみると図-2 のようになるであろう．

機械翻訳は一般に次元（用語，文型の種類）が大きくマッピングの対象が多種多様でそれらを充分に把握することはとても困難である．いきおい直訳にとどまらざるを得ないが，かといって日本語の水面下の部分を浮上させないとそれすらおぼつかないと言える．文脈（場の効果）の利用は最少限にとどめる必要があろう．

DB の場合は機械語へのマッピングを行うためある意味で意訳ができなければならない．しかし幸いなこ

	次元	マッピング	変換レベル	コントロール度	文脈利用
翻訳	多大	△多大	(直訳)	小	(小)
DB	小	◎データ	意訳	中	(多大)

図-2　テーマ別課題

とにマッピングの対象はデータのみといってよく次元も小さい．そしてデータセマンティクスはあらかじめ自然言語と柔軟に対応付けられた形で保持しておくことが可能であり，自然語文は離散的である．そこでこの場合の課題は，翻訳とは逆に文脈利用を最大限許すことである．すなわち，「僕はうなぎだ」式の表現も認めさせるようにすることがエンドユーザ向として重要であると考えている．

4. 全般的課題

以上主に技術面での問題点や課題について述べたが，最後に今後の周辺環境面での課題について触れると次の通りとなる．

（1） 知識の獲得収集の効率化

- 辞書，シソーラス，コンコーダンス，テキストベース等各種言語データの整備
- 流通の自由化

（2） 日本語のコントロール

- テクニカルライティングの確立
- 翻訳家との協調
- 言語センタ

（3） 国 際 協 力

- 対欧米
- 対アジア
- 日本語の輸出

（1）に関しては，自然言語処理にはいろいろな意味での知識が関係しているので，その基礎となる各種言語データの整備を官民一体で協力して行い，かつまたその流通自由化を促進する必要があろう．

（2）に関しては，日本語の中で閉じている間は以心伝心で済むので問題にならない．また人手による翻訳の場合も，日本語に堪能であればまずは原文を把握する上で問題にはならないだろう．しかし高度に専門的分野になるとそれも難しくなるはずである．また今後，社会が計算機と共存していくこと，さらには（3）とも考えると，日本語の適切なコントロールによる多義性の解消が課題となろう．

（3）に関しては，欧米の翻訳プロジェクト等にドッキングを計ることはわれわれにとり直接のメリットがあるが，それだけでなくアジアの国々に対して情報提供を計ることも今後重要となろう．マレーシアのマハディール首相も「Learn to the East」と言っている今日，アジアへの協力も考えていく必要がある．

また日本語の上には科学技術上，経済上，文化上価値の高い情報が多量に乗っているわけであり，直接日本語に接したいという海外からの要望も増えてくるのではないかと考えられる．それに対処するためには，かつて英国が英語を輸出したように，日本も世界で6番目に多く使用されているにもかかわらず日本人以外は使っていない日本語の輸出を考えてもよいのではないか．

日本語の国際化のためには水面下の部分を浮上させ「場の効果」の活用を抑制せざるを得ず，この点機械処理からの要請ともマッチしよう．

上記の課題全般にわたって中心的役割を果す言語センタ的機関が望まれるところで，そこでインタナショナル（コントロールドというとアレルギを生じる向きもありそうなので）ジャパニーズをまとめ上げて政府刊行物あたりから適用してみることを提案したい．

音声認識の諸問題

千葉 成美

音声認識の問題は，音響処理や言語処理，ハードウェアなど多岐にわたるが，ここでは言語処理の側面からみた問題点について考察したい．

自然言語処理というと，普通，機械翻訳や質問回答システムなどに関連した文字列の処理を指す場合が多い．しかし，言語の成り立ちを考えると，書き言葉よりも話し言葉の方がはるかに先行している．したがって，話し言葉を直接処理する音声認識は，本来最も基本的な自然言語処理であるといえる．

文字列の処理は，文字という有限のシンボルの集合を相手にした，抽象化された世界での処理である．これに対して，音声認識は，音声現象という，アナログ的でかつ変動の激しい事象を相手にしており，文字列の処理よりもはるかに難しい面をもっている．

音声認識技術の究極の目標は，不特定多数の話者が発声した任意の語彙を含む文を連続的に発声した音声を完全に認識することである．音声信号は，音素と呼ばれる最小単位から構成されており，その種類は言語により異なるが数十種である．したがって連続的な音声を音素の系列に分解して，それぞれを正しく認識することができれば，あとはディスクリートなシンボルの世界の処理であり，従来の音声学や言語学の知識を利用することにより，容易に単語や文として認識することができる．しかし，連続音声中の音素は，その前後の音素環境により物理的な性質が大きく変化するため，それぞれを正しく確定的に認識することは不可能である．したがって，人間が通常行っているように，

シンタックスやセマンティックスの情報を援用した高度の言語処理により，意味内容を理解することによって，初めて全体を正しく認識することができるものと考えられている．このようなアプローチは音声理解と呼ばれ，これまで1,000程度の語彙からなる文の認識，理解はほぼ達成されている．しかし，このために数 MIPS のコンピュータを使用しても，一つの文に対して数分あるいはそれ以上の処理を必要としており，数千語以上の大語彙を対象にした本格的な連続音声の認識が実現するまでには，今後かなりの年月を要するものと考えられている．

そこで，認識対象の語数や話者，発声方法などに制限を加え，その範囲内で実用可能なレベルの認識性能を実現しようという試みが古くから行われてきた．**図-1** は，このような認識対象による音声認識技術の分類の一例を示している．

数百語程度までの限定語音声を対象にする場合には，単語全体の音声パターンを認識の基本単位とすることにより，音素認識の問題を避けることができる．このため，パターンマッチング，特に DP マッチング方式（動的計画法により最適時間正規化とマッチングを同時に行う方式）や，識別関数方式などにより，高い認識率が得られており，産業用の分野を中心に実用化が進んでいる．

最近になって，DP マッチングをさらに発展させて連続的に発声された単語の認識を可能とした2段 DP マッチングに，有限オートマトンによるシンタックス制御を取り入れた方式が開発され，限定文音声の実時間認識が可能になった[1]．この方式では，オートマトンの記述が単語レベルで行われる．このため，前述の音声理解システムにおいて行われたような，音素レベルのネットワーク表現と異なり，専門知識がなくてもシンタックス構造を記述できる特長があり，実用上の大きな利点となるものである．

このような限定文音声認識は，航空管制官訓練システムや，OA システムなどのほか，知能ロボットの制御用などへの応用が考えられる．

任意語音声の認識において，日本語の場合は，カナ文字に対応した単音節を区切って発音したものを認識すれば，その組合せによって任意の日本語を表すことができる．日本語単音節は約100種あり，原理的には限定語音声認識のように，パターンマッチングにより認識することができるが，子音部分だけを抽出してマッチングを行うことにより，子音認識の精度を上げる方法もよく行われる．

単音節の形での音声の発声は，やや不自然であると同時に，入力速度の面でもかなり不利となる．そこでキーボードなどと比較した日本語テキスト入力の評価実験を行った[2]．その結果，カナキーボードになれていない被験者の場合，単音節音声による入力は他の入力手段と同等かそれ以上の入力速度をもつことが実証され，また，主観評価でも最も使いやすいという結果が得られた．現在単音節認識装置と音声ワードプロセッサの実用化の試みがある．

単音節認識では，完全な音素識別が必要となるため，特定話者の場合でも認識率は前後にとどまっている．このため，単語辞書を利用した言語処理により，認識誤りを訂正する技術が重要であり，今後の進展が期待される．

究極的な目標である任意語連続音声の認識では，従

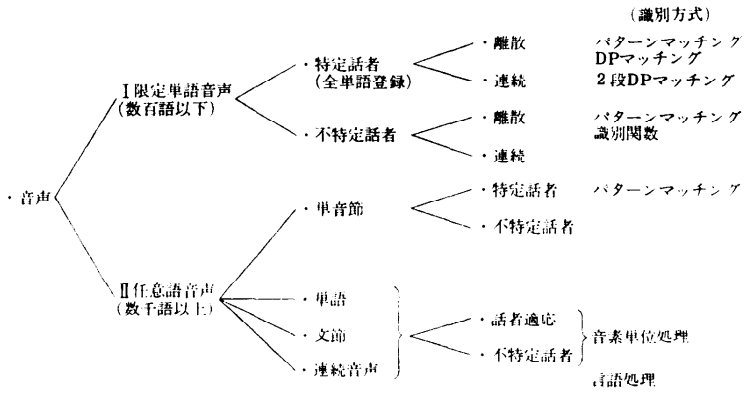


図-1　音声認識技術の分類

来の音声理解システムにおけるたかだか1,000語の限定タスクのため言語処理とは桁違いに大規模で効率的な自然言語処理技術が必要となる．このような本格的な自然言語処理技術の開発は最近急激に立上りつつある機械翻訳の研究の一環として，強力に進められている．音声認識の今後の進展は，かなりの程度までその成果にかかっている．

文 献

1) Sakoe, S.: A Generalized Two-Level DP-Matching Algorithm for Continuous Speech Recognition, Trans. IECE Jpn., Vol. E 65, No. 11, pp. 649-656 (Nov. 1982).
2) 吉田：日本語ワードプロセッサに対する各種入力手段の比較評価，情報処理学会昭和57年度前期全国大会論文集 2G-3, pp. 1019-1020.

自然言語による質問応答システム

諸橋 正幸

1. 必 要 性

計算機利用者の増大，通信技術の発達などを背景として，従来よりはるかに多様な情報が多くの異なる検索要求をもった利用者によって使用できるようにすることが最近のデータベース検索システムの主要課題となっている．このようなシステムを実現するには，従来の設計思想である，大規模データベースに対する特定検索要求への高速応答に重点を置いた検索機構ではなく，多種の検索要求（ときには，データベース設計者があらかじめ予期していなかったような検索要求）に対して適切な応答をする知的な検索機構が必要である．一方，検索言語についても従来の使用環境と異なった状況を考える必要がある．すなわち今後の傾向としては，利用者数が多いかわりに各利用者の検索回数は逆に少なくなる（いわゆる，たまさかな利用者が増大する）であろう．こうした利用者に対し，毎日端末の前に坐って検索に明け暮れるオペレータと同様の検索言語習得訓練を行うことは非現実的であり，したがって訓練をほとんど必要としない検索言語あるいは手段が必要となってくる．こうした条件を満たす検索システムとして「自然言語による質問応答システム」は理想的である．しかしながら，この種のシステムを実現するために解決すべき技術的問題，実用化のための経済上の，あるいは使用環境の問題などは決して少なくない．

2. 技術的背景

複雑多岐な検索要求に対して柔軟に対応できるデータベース・スキーマとして初期の質問応答システム（LUNAR, SHRDLU, REQUST QBE REQUEST など）が採った方式は，自然言語の構造（チョムスキーの生成変形文法やフィルモアの格文法に則ったもの），あるいは人間の思考過程のシミュレーションや述語論理に基づく推論機構による情報の格納と検索のスキーマであり，これらの研究は現在でも知識工学やエキスパート・システムの分野での主要な研究方法の一つとなっている．このアプローチによれば高度に複雑な検索要求に対して大きな効果を発揮する．また，質問文に必要とする検索条件のすべてが提示されていないような場合にも欠けている条件を的確に把握し，質問者にフィードバックすることも可能であろう．その反面，蓄積すべき情報の内部表現が複雑で，言語理論，対象とする応用分野の知識の両面にわたる深い造詣をもつ人々の関与が不可欠である．しかも，大規模データベースを矛盾なく構築するためには，彼らの長時間にわたる多大な努力が必要である．また適用分野に依存した表現形式の工夫が多く導入される傾向が強く，ある分野の情報に関して構築されたデータベース・スキーマがそのまま他の分野に適用できる可能性が少ないなどの欠点がある．

それに対して，コッドの提唱した関係形式データベースは純粋にデータベース・スキーマという観点から評価すると，利用者の立場から見たデータベース・モデルとして非常に洗練された枠組を与えた点で特徴的なものである．すなわち，計算機内部での検索手順とは独立に検索要求の表現手段が与えられる．QBE はその表現手段をそのまま画面上に視覚化させて成功した例である．この特長は，それまでのデータベース・モデルのような検索者自身が検索手順を明示的に与えるモデルに比べて，検索者がデータベース設計者の意図する検索手順に必ずしも則る必要がない点で大きな利点となる．いいかえれば，ある一つの情報を得るために発する質問がいろいろな表現をとることを許すことにほかならない．この利点は自然言語による質問を許すようなシステムを考える上で重要な機能である．関係形式モデルを利用して自然言語による質問応答システムを作ることは，コッド自身手がけており，その後「やちまた」でも試みられた．「やちまた」では，関係形式モデルが汎用のデータベース・スキーマであることを利用して，種々の適用分野のための質問応答システムが簡単に（従来のデータベース設計者が必要とする知識とほぼ同等の知識で）作ることができ

るような生成機能を設けることに成功し，汎用のシステムとしての可能性を初めて示した．

自然言語によるシステムを考える上でのもう一つの大きな要因は質問文に現れる表現とデータ・モデルとの関連をいかにして実現するかという点である．REQUEST はチョムスキーの言語モデルをかなり忠実に実現したが深層構造に直接対応するデータベース・モデルが存在しないために，独自の方法で必要な情報を取り出す．「やちまた」は格文法の枠組に近い言語モデル（名詞句データ模型）を作り，それを関係形式データ・モデルに対応させる．この言語モデルでは格にあたる部分の表現が非常に柔軟で関係形式の枠組に合いさえすれば品詞がなにであってもこだわらない．したがって言語モデル自身の一つのデータベース・スキーマに近い概念であり，名称もデータ模型となっている．言語モデルをデータ・モデルと非常に近い形にしたために，「やちまた」ではデータベースのさしかえがシステムに与える影響が少なく，汎用システムとして機能する理由はここにある．汎用システムとして自然言語による質問応答を考える上で重要なことは，質問文で表現される文の言語モデルとデータベース検索のためのデータ・モデルをいかにうまく結合させるかにある．

3. 実用化の条件

自然言語による質問応答システムが実用のものとして使えるようになるためには，いくつかの条件を満足する必要があろう．

（1） 処理時間の問題，メモリの問題

従来の検索言語による質問応答に比べ，処理時間がかかることは否めないが，1節で述べた環境の変化からこうしたシステムが次第に必要となってくることは明らかであり，オンライン検索に十分な応答速度が得られれば問題はなくなるであろう．ハードウェアの技術的発展の度合を考えてこの問題は将来にわたってまで問題となることはなかろう．メモリについても同様である．

（2） 複雑な質問になると質問文が簡単に作れなくなる

これは検索要求自身がもつ本質的な問題であり検索言語が変わっても大きく変わる問題ではない．ただし形式言語の場合は思考の枠が限定されるために，複雑な検索要求は一つの検索文で実現できない等のことに比較的早く気づくために対処の方法がわかりやすい傾向はあるかもしれない．自然言語による検索においても検索要求を論理的に表現する必要があることは検索者自身心得ていてもらいたいものである．

（3） 形式言語による検索に比べ質問文が長くなる

これは，われわれが実験した際に被験者からしばしば耳にした苦情である．たしかに質問文をカナ鍵盤から入力することは少し長い質問になると苦情を伴うものである．ただし，日本語ワード・プロセッサが普及しはじめたことにより，カナ鍵盤に対するアレルギが解消されれば状況は好転するであろうし，音声入力が実用化されればこのシステムは最大のアプリケーションの一つとなろう．

（4） 応答能力の問題

現段階ではやはり，どんな質問にも答えられるシステムは難しいが，検索者に提供情報の種類，性質を知らせておけば避けられる問題である．

自然言語処理システム開発ツール

森 健一

1. タフなアルゴリズム

日本語を含む自然言語処理システムの研究開発が盛んになってきており，いくつかの国家的プロジェクトも並行して実施されるようになってきた．

自然言語処理においては，言語処理のアルゴリズムが，単に実験的な数千語の世界で適当な性能を示したからといって，現実の世界で本当に役に立つかというと，何の保証もない．そのアルゴリズムは本質的に多様で，生きた世界で通用する「タフさ」をもっていなければならない．パターン認識の研究分野で，多くの研究室的論文が，このタフさに欠けているために，単に論文のための論文にしかすぎなかったことからも，この点は強調しすぎることはない．アルゴリズムがタフかどうかは，実験室的な規模でなく，数十万語の実データでテストすれば容易に判定できる問題である．実験のたびに数十万語のデータを用いることは困難であるとすると，少なくとも，どのような場合にアルゴリズムの弱点が顕著になるかの感覚は実験者が自覚している必要がある．このためには実験の過程で，一度は数十万語以上のデータでテストしてみる必要があろう．

2. 自然言語処理システムのための開発ツール

自然言語処理システムを研究開発するうえで必要となる開発ツールは，3つに大別される．

（1） 言語処理シミュレーション言語

当然言語処理のアルゴリズムを研究開発するためのツールとしては，言語処理の各プロセスを記述するための便利なシミュレーション言語が必要となる．この言語は，対象とする自然言語の構文解析，意味処理，文脈処理，文生成処理のすべての過程を通して統一的に記述できることが望ましい．さらに，アルゴリズム部と，辞書部とが分離された形で記述されることが必要である．

（2） 辞書開発用ツール

自然言語処理には少なくとも数万語以上の単語辞書を作成する必要がある．この辞書の単語数は応用が広がれば，数百万語にも発展していくことが予測される．その開発と管理には最初から慎重に計画された辞書開発用ツールの用意が必要となる．

（3） テスト用自然言語データベース

辞書作成と似たようなものに，自然言語の用例を豊富に含んだテスト用のデータベース作成がある．このための文章は特定の分野に限定されることなく，自然言語処理のアルゴリズムのタフさをテストするのに十分な量をもつ必要がある．研究者はとかく小さな辞書を使い，小さなデータベースで実験しがちである．上記(2)(3)のを作成するには，大変な時間と労力を必要とする．国家的なプロジェクト分は，その費用の相当分をさいて，辞書と自然言語データベースを整備することが必要だと思う．

パネル討論に対するコメント

辻井 潤一

機械翻訳がさかんに研究された1960年代と現在とを比べてみると，言語処理をとりまく周辺技術の状況は，ハードウェア・ソフトウェアの両面から大きく変化してきている．また，言語処理そのものの技術を考えてみても，各種の構文解析手法の定式化・格文法やモンテギュ文法等の意味処理技法の深化など，数多くの成果がこの20年の間にあげられてきた．自然言語処理が，単なるアカデミックな興味だけでなく，応用システムを含めた，広範な興味をひき起しつつある所以である．

このような状況は，今回の各パネラの意見にも反映されていたが，特に私には次の3点が印象深かった．

（1） 形態素解析・構文解析などにおいては，これまでにすでにさまざまな技法・アイデアが提案され，各技法の有効性とその欠点が定性的にはかなり明瞭になってきたこと．

（2） しかしながら，人間の高度な知的機能と関連した，いわゆる意味の問題には，依然として明解な手法は確立されておらず，むしろ，研究の進展に伴って，人間と機械の間のギャップがますます明瞭になってきたこと．

（3） これまでに開発されてきた技法を組み合せて，大規模な応用システムを開発しようとする動きが，各種の研究機関・メーカの間でかなりはっきりとでてきたことである．

(1)・(2)は，「現在の技術で実現可能なことと実現不可能なことがかなりはっきりとしてきた」とまとめて言ってもよいであろう．その結果，実現可能な範囲でシステム化してゆこうというのが，(3)の動きである．

このような現象は，自然言語処理の隣接分野である人工知能の分野にもみられる．そこでは，これまでの技法の蓄積（例えば，発見的探索技法やプロダクション・システム）の範囲内で，現実的な応用システムを開発してゆこうとする知識工学の動きが活発である．その一方で人工の知能と自然知能（人間の知能）とのギャップの大きさは，言語学や心理学との連携を強めて，より基礎的・サイエンス的に自然知能を探究してゆこうとする認知科学の流れを作り出している．

過去の蓄積を工学的に応用してゆこうとする立場に立つか，あるいは，人間と機械のギャップに目を向けて，これを少しでも埋めようとする立場に立つかは，各研究者の趣味の問題でもあり，どちらの立場が正しいといえるものではない．一つの研究が同時に2つの問題に対するアプローチとなっていることも多い．しかしながら，あえて言えば，私自身としては，現在の時点で一度大規模な自然言語応用システムを開発してみることが，今後の自然言語研究にとって不可欠なのではないかと感じている．

一つには，自然言語の現象自身が非常に多様であり，(1)の，すでに定式化された各手法も，どの程度現実の言語現象をカバーできるものなのかが明確になっていない，ということがある．これまでに定式化されてきた技術やアイデアは，研究室という温室の中で育てられてきたものであり，これが本当に現実の現象に対してどこまで有効かは，ある程度の目安はあるにしても，研究者自身よく分っていない．多くの過度に洗練された人工知能分野の手法が，知識工学という試錬の中で淘汰され，現実的なプログラミング手法として再定式化されつつあるのと同じことを，自然言語処

理の各手法に対しても行ってみる必要がある．

自然言語処理は，質の技術であると同時に，量の技術でもある．これまで数多くの機械翻訳システムが作られ，デモンストレーションの段階までゆきながら，現実の応用段階にまで至らなかったのは，それを量的に拡大しようとしたときに，本質的な困難に遭遇してきたからである．例えば，大規模な文法や辞書を多数で開発するといった問題をどうシステム的にサポートできるか，といった問題は，知識工学システムが大規模化した際に生じる問題とほとんど同じか，それ以上にむずかしい問題をはらんでいる．これまでの温室育ちの技法やアイデアは，この側面からもう一度再定式化・再検討されなければならない．

現時点で大規模な応用システムを開発してみることのもう一つの意味は，人間と機械とのセマンティック・ギャップを埋めようとする研究にとっても重要である．Winograd が対話システム GUS を開発したとき，その感想として，「現実の会話では，設計者が考えてもみなかった応答があらわれる．GUS が行っている会話は，自然らしく見えるものではあるが，現実の人間の会話ではない」と書いている．われわれ人間は，頭の中で種々の状況を考え，さまざまな文とその文脈を考えることができる．初期の自然言語理解の研究は，このような内省の結果得られたさまざまな文や文章をもとにして，そこにあらわれる問題を解決するための技法やアイデアを作りあげてきた．そのような文や文章が，現実の言語現象の中でどの程度一般的にあらわれるものなのかについては，ほとんど問題にされなかった．研究のある時期，このような思考実験を行うことはもちろん重要であったし，不可欠であった．しかし，そればかりを続けてゆくことの危険も非常に大きくなってきている．Winograd の述懐のように，われわれが内省によって作り出す現象以外のものが現実の言語現象の中に数多くあらわれること，逆に，われわれが内省によって作り出した難問は，現実の言語現象の中では，ごく少数の例外であるかもしれないこと，等々である．大規模な応用システムを開発してみることは，なにがまず解決されねばならない問題なのかを明確にしてくれよう．

あとがき

パネル討論会は自然言語処理技術に関する基礎的な研究課題について討論を行い，つぎにそれらの応用システムについての諸問題の討論を行った．自然言語処理のすべての問題が取り上げられたわけではないが，主要な問題の幾つかが取り上げられ討論された．全体的に，機械翻訳に関する話題が多いのは，これが最近の自然言語処理技術の1つの集約点に位置しており，その実現可能性に向けて，多くの研究者，技術者の努力が我が国で傾注されていることの反映でもあろう．