人工知能学会研究会資料
SIG-F/H/K-9101-1(12/5)
<講　　演>

人工知能の研究と自然言語処理

田中穂積
（東京工業大学工学部）

言葉・思考・知能

思考と言葉（自然言語）とは不可分の関係にある。我々は言葉を使って思考しているとも言われている。確かに言葉を抜きにした思考を想像することは難しい。思考の過程で呟く癖のある人を時折見かける。ぶつぶつ呟く癖はないと抗弁しても、君は口に出さなくとも、心の中で呟きながら思考を進めているのだと言われると、確かにそのようにも思えてくる。

もともと言葉は、我々が思考した結果を「相手」に伝達する手段として生まれ、発達してきた。しかし、この「相手」という所を「自分自身」という言葉で置き換えれば、思考を進めるということは、言葉を用いて自分自身に思考を伝え、それにより思考を推敲することだ、と言い換えることができる。一方、思考は知能と不可分の関係にあるのは明かだろう。思考の裏付けがあってこそ、我々に知能があるといえるのだろう。したがって、言葉と知能は不可分の関係にある。

この結論を冒頭でいきなり述べたとしても、それは何の疑問もなく自然に受け入れられたに違いない。それほどこれは当り前の結論であると思われる．にもかかわらずくどくどと前置きを述べたのは、言葉の本質を探ることが、知能の本質を探ることでもあるということを、「思考」というキーワードを通して説明したかったからである。

これまで多くの研究者が、コンピュータによる自然言語理解など、自然言語処理に関する研究課題は、人工知能の応用に属すと見なしてきた。これは一般には、人工知能の本質に関する研究が別途あり、そこから得られた成果が理論であって、その理論を応用することが自然言語処理の研究であると考えられていることによる。理論と応用という言い回しからも分かるように、一義的には、理論が応用に先行すると一般には見られている。人工知能の応用という言い回しには、暗黙の内にこうした意味合いが含まれている。自然言語処理の研究にそうした意味合いが含まれていることは部分的には認めて良いだろう。しかし、自然言語処理の研究は応用にのみ留まるものではない。

冒頭での議論からも明らかなように、自然言語処理に関する研究は、好むと好まないとに関わらず、知能の本質と直面しなければならない。自然言語処理に関する研究には、知能の本質に関わる問題が数多く含まれている。自然言語処理の研究が長期を要す研究課題であるのはそのためだろう。我々はいかなる機構を用いて言葉を理解したり、言葉を作りだしたりしているのだろうか。未解明の問題は山積しているといってよい。問題が山積しているからといって、悲観することはない。むしろこのことは、自然言語処理が、人工知能に関する重要な研究課題の宝庫であるということを意味しているからである。実際、人

工知能に関する新しい理論が、自然言語処理の研究を通じて生まれると筆者は確信している。ただしここで誤解を招かぬように一言付け加えるとすれば、困難な問題が山積しているからといって、現在の自然言語処理の研究が極めて困難な状況にあり、工学的に役に立つレベルに程遠いと言うことではない。現在の自然言語処理の研究のレベルでも、十分工学的に役立つことは可能であるということを指摘しておきたい。たとえば現在のワープロは、入力された仮名文の意味を必ずしも理解して、かな漢字変換を行っているわけではない。現在の自然言語処理の研究成果の極く一部を使っているに過ぎない。それにも関わらず現実に役立っている。この文章作成もその恩恵を蒙っている。

以下では、自然言語処理の研究で、今後何をなすべきか、気が付くままにそのいくつかを取り上げて説明してみたい。

言葉と知識

人間の知能を省みることなく全く無縁なまま、人工知能の研究が発展し続けることは可能だろうか。飛行機の実現が、鳥の飛行のメカニズムを解明せずとも可能であったのと同様なことが、人工知能の研究にも当てはまるのだろうか。ライト兄弟が飛行機を空に飛ばすことに成功してから、飛行のメカニズムの研究が始まったという事実を見よ、と主張することもできるかもしれない。この場合には、人工知能が実現してから、人工知能の研究が始まることになるのである。少し考えてみれば分かることだが、これはいささか奇妙なことである。

筆者の考えでは、人工知能の研究を、飛行機の研究になぞらえることには問題がある。知能をどう人工的に組み立てるかという問題は、飛行機を組み立てること以上に複雑で微妙な問題である。研究を進める上で鳥と飛行機を切り離すことができたようには、人工知能を人間の知能と切り離すことができない。両者の決定的な違いは、飛行機が飛ぶということは、物理的現象の一種であるのに対して、人工知能が動くということは精神的な現象を模擬しているということである。それが精神的な現象である限り、人間の知能と切り離すことができないのである。

人工知能には知識が必要であると言われる。知識獲得や学習の問題が重要なのはそのためである。数学的な問題を解くためには、数学に関する知識がなくてはならない。同様に、言葉を理解するためには、言葉を理解するための知識が必要になる。そこには文法的な知識も必要になろう。辞書的な知識も必要になろう。それだけも十分とはいえず、我々が日常何気なく使っている常識なども必要になろう。

たとえば、

Bring me a jack.

It will be in the toolbox on the table.

という文では、現場にいて（居あわせていて）状況が分かっている人を除き、jack の意味

は、後文の table ではなく tool box を聞いてから決まる。常識的に tool box の中には道具としての jack（ジャッキ）があり、トランプカードの jack（ジャック）があるとは考えられないからである。一般に常識は、言語的な知識というより、日常生活から学んだ文化的、慣習的な知識であり、それがどのようなものであるかについて、定説といえるものはまだない。言語理解システムの研究では、このような常識が必要に応じて問題毎に個別に用意されることが多く、残念ながら常識の一品料理的な色彩に留まっている。常識に関する体系的な研究が望まれる。しかし、常識を含む知識に関しては、以下に述べる問題がある。

知識は、飛行機と違って、客観的に外から観察できる代物ではない。知識の外形、構造を直接知ることができない。ここに知識に関する研究の難しさ、もどかしさ、問題がある。確かに我々は前述した知識を持っていると断言できる。しかしそれを手に取り観察することができない。唯一可能なことは、自分の持っている知識を内省することである。内省を抜きにして知識を論ずることは、滑稽であるとすら思える。したがって、自分自身のもつ知識についてもっと深く内省し、それの結果を人工知能の持つ知識に反映させることが必要であろう。ところが、内省により言葉を理解するために使われている知識を全て洗い出すことが難しい。言葉の理解の過程で、無意識の内に使っている知識が数多くあると思われるからである。

自然言語理解システムを作ってみれば、この問題はすぐに顕在化する。なぜなら、人間が当り前であると思っていたことが、自然言語理解システムとして実現されていないことがあまりに多く、それが自然言語理解システムの応答からすぐ分かるからである。自然言語理解システムは、問題の所在を明らかにしてくれるという意味で有用であるという消極的な見方もできる。認知科学の研究者が自然言語理解システムに興味を持つ理由の一つは、この辺りにあるのかも知れない。もう少し積極的な意義をあげれば、自然言語理解システムに組み込まれた知識は、外から眺めることができる。それだけでなく、操作も可能である。これは、知識の研究にとって重要なことである。知識のシミュレーションが可能になるからである。この時問題になるのは、知識をどのような形式で表現すべきかということだろう。知識表現形式の問題は、人工知能の重要な問題である。

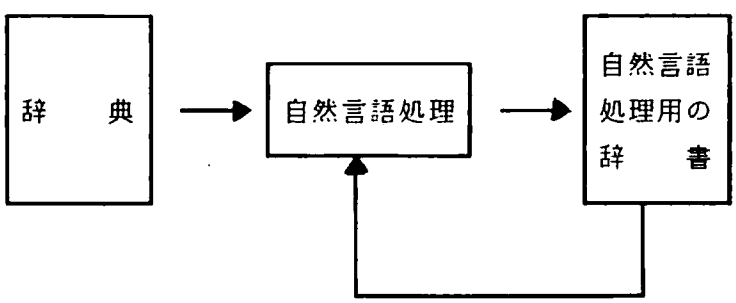
自然言語処理用の辞書

人工知能が人間並に賢くなるためには、人工知能が人間並の量の知識を持たねばないだろう。言葉の理解に大量の知識が必要になることは、これまでの議論から明かだろう。言葉の学習について、６０年代、７０年代と異なり、白紙の状態から学習可能な学習理論やアルゴリズムの存在について、最近では疑問が提出されている。学習は大量の知識を前提にして初めて可能になるというのである。新しい知識は何もないところからは生まれない。短絡的に表現すれば、少量の知識ベースからは役に立つ知識を学習することは困難である。意味ある知識の学習は、大規模知識ベースの存在を前提に、はじめて可能になるというのである。エキスパートシステムの性能を一層向上させるためにも、高度な知識ベースの構

築が必要であるとされ、米国のＭＣＣでＣＹＣと呼ばれる大規模知識ベース構築が始まり、その一部が公開されている。

米国と異なりわが国では、機械翻訳の研究と開発が盛んに行われている。そこで使われる辞書は、これまでの人工知能の分野での自然言語処理の研究で使われきた辞書より大きい。機械翻訳では多種多様な文を処理する必要があるからである。機械翻訳システム開発の経験を踏まえて、わが国では機械翻訳の研究者が大規模知識ベースの必要性を唱えている。これは電子化された大規模な辞書の開発を目指すものであるが、辞書の品質向上が機械翻訳システムの質の向上に大きな役割を果たすことを彼らが実感したからに他ならない。実際、機械翻訳結果に含まれる誤りを分析すると、単語の記述内容の不備が原因であることが多い。そのため辞書の整備に多大の労力が費やされることになった。通産省が推進している電子化辞書計画は、このことも考慮した大規模辞書開発計画であり、その公開も真近かと言われている。公開された辞書は、使用者側が使いこなしてよりよい辞書に育て上げる努力が必要だろう。

こうした大規模知識ベースを如何にして構築すべきか、という問題を次に考えてみたい。第一の方法は、我々の持つ知識を記述した既存の百科辞典や言葉の辞典から、機械に分かる形式の知識を機械的に抽出することである。それには自然言語処理技術が利用できる。これらの辞典の記述は、自然言語で書かれているからである。この方法がどれほど有効かは、現在の自然言語処理の研究レベルによる。自然言語による辞典の記述は、明瞭さ、簡潔さの点で、他の自然言語の文より優れている。辞典の記述は、現在の自然言語処理技術である程度カバーできる範囲にあるともいえる。いわば、自然言語処理用知識のブーツストラップ構築の研究が、わが国でもっと活発化することが望まれる。



第二の方法は、労力を要すが、これまでの辞書作成と同様に、人間が丹念に知識を記述して行くことである。その過程で、部分的には計算機の助けを借りることはあっても、人間が個々の知識を丹念に記述して行く。ＣＹＣ計画でもＥＤＲ計画でも、基本的にはこの方法を採用している。これらの計画の推進者は、第一の方法は研究としては成り立つものの、現在の自然言語処理の技術が未熟で現実的ではないと主張している。現時点で自然言語処理技術を導入しても、最終的に人間を頼りにせざるをえず、労力がかえって増えることも考えられる。ＣＹＣ計画でもＥＤＲ計画でも、この点を考慮して、第二の方法を取っているものと思われる。それはそれで確実さを重視した一つの見識であろう。

国語辞典などには、各単語にいくつかの語義が記述されている。たとえば「さくら」という単語には、「落葉高木としての桜」、「桜色」、「馬の肉」といった語義が書かれている。知識としては、単語としての「さくら」ではなく、語義をベースにする必要がある。「さくら」は多義語であり、自然言語の意味処理では、「さくらを食べた」という文の「さくら」が「馬の肉」であることを知る必要がある。ＥＤＲの辞書では、大略この語義相当のものを概念と呼んでいる。概念を集め、概念相互の間に関係を付けたものを概念辞書と呼んでいる。しかし、これらの概念相互間に関係付けする作業は困難なものになる。関係と言ってもそれはどのようなものかをまず決める必要がある。上位下位関係、全体部分関係、同義関係、類義関係、反義関係などは比較的容易に思い浮かぶが、それ以外にも概念間の関係として様々なのものが考えられよう。それらをフィクスすることは容易でない。いわゆる深層格関係にしても、合意の得られた深層格関係のセットを確定することが難しい。そこで、さし当り合意の得られた関係についてのみ作業を進めることになる。国語辞典などでは、同義、類義、反義関係は記述されているものの、それ以外の語義間の関係は、語釈文を介して読み手に自然に伝わるようになっている。

このように語義単位に分割した辞書記述の問題点は、語義相互間の関係を設定することが難しく、その作業はたぶんに推測と言語的直観に頼らざるをえないことである。先の「さくら」の例で挙げた三つの語義も、それぞれが全く独立なものではなく、相互に関連し合っている。肉の色が「桜色」に近いことが、「馬の肉」という語義を成り立たせていると思われる。もっと中核的な語義を設定して、そこから隣接関係にある語義を導き出すことができると都合がよい。このような立場に立つものとして、語源を重視するものがある。語源は歴史的のものであり、それだけで十分であるとは言えない。最近、基本的な動詞の意味を、少数の中核的な語義から派生すべきであるとする研究がある。彼らはこれをmonosemyとよんでいる。この考え方は、比喩や隠喩理解に自然な解釈を与える可能性があり注目に値する。筆者にはSchankの考え方も、基本的にはこれと似ているように思われる。しかし、この中核的な語義をどう設定するか、今後の研究に待たねばならない。

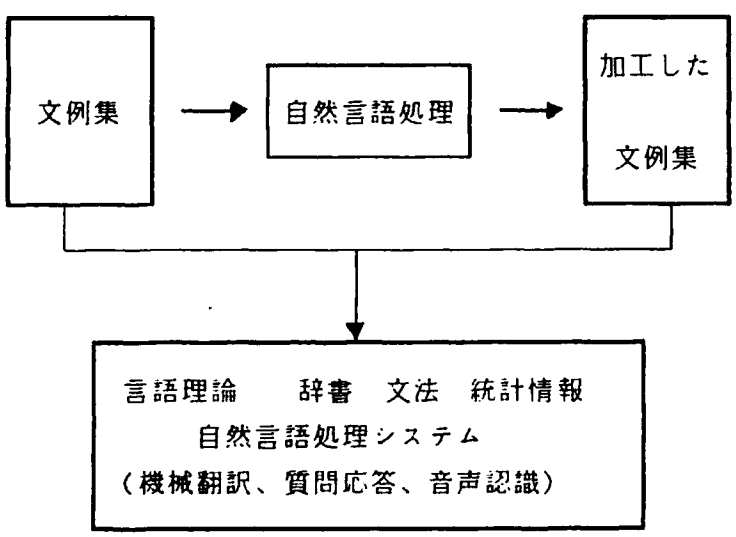
文例データベース

自然言語処理の研究を進めるにあたり、様々な文例を集めておくことは重要である。我々が想起できる文例には限りがある。言語学の研究で、理論をよく説明する文例は作れても、理論の反例となる例文を作り出すことが難しいと言われる。必要な文例が直ちに取り出せる文例データベースを作成しておくことは、今後の自然言語処理の研究の進展に大きな寄与をなすと思われる。実際、辞書の作成では、その語が用いられている文例、言い換えると用例を集めることから出発する。したがって、いかに多くの用例を集めるかが良い辞書を作ることに直結している。文例データベースがあり、それが様々な角度から検索可能になれば、辞書作成者にとっても大きなメリットになるだろう。それはまた、自然言語処理用の知識ベース構築にも役立つことになる。そのためにも、なるべく加工しない生のままの文例集をまず初めに作るべきだろう。生の文例集であっても、直接役立つ情報が含

まれている。ＫＷＩＣなどは、それから容易に作り出すことができる。いずれにしても、使用目的に応じた生データの加工は、自然言語処理技術の助けや、計算機技術の助けを借りて行うべきだろう。

最近では、文例を利用した自然言語処理、機械翻訳システムの実現が模索されている。人工知能の分野では、これらはコーパス・ベース、事例ベースのシステムなどと呼ばれることがある。人間の翻訳でも、翻訳事例を記憶しておくことの重要性はよく経験する。文例データベースはこれらの研究にも役立つだろう。機械翻訳の場合には、いくつかの言語の文が並置しており、そのおのおのが相互に対訳になっている文例データベースの構築が望まれる。文例集からは、自然言語処理に必要な統計的な情報を抽出することができる。最近の音声認識システムでは、n-グラムなど音素の並びに関する統計情報が重要な役割を果たしている。自然言語処理でそれに相当するものは、単語の並び、品詞の並びの統計情報であろう。興味深いことに、漢字仮名変換プログラムにより、文例集を音素の並びに変換し、それから音声認識用の統計情報を取り出し利用されている。

文例データベースにはどれくらいの分量の文例が必要だろうか。言語学者によれば、数百万例文は必要であると言う。それ以上なければ役立たないという人もいる。数百万例文以上になると、そのＫＷＩＣをどうするか、検索を高速化するためのデータベース構成法など、大規模故の問題を解決しなければならなくなるだろう。最近のＶＬＳＩ技術の進展を見れば、記憶容量の問題は、さして重要ではないだろう。



言葉と推論

言葉の理解には推論がつきものである。「文間を読む」という表現には、そのことがよく現れている。先に述べたjackの意味の決定も、常識を用いた推論が使われていた。これまで、知識の重要性を強調してきた。しかし、知識だけあっても問題は解けない。適切な知識を適切な時期に引き出して初めて問題が解ける。この適切な時期に適切な知識を引出

し、それを使って問題を解くことが推論の本質だろう。多くの推論では、固定化した知識があり、それを参照する。言葉の理解では、それまでに理解して得た知識も、それ以後の推論に使う。この時、文を聞いた段階で、曖昧性解消のための推論結果が必ずしも一つに絞られないことがある。後続文により、前文までに得た知識が次第にクリアなものになることがある。これは先のjackの例からも明らかなことだろう。言葉の意味理解には、こうした漸進的な推論が可能なメカニズムが必要になる。これを漸進的曖昧性解消と呼ぶことにする。jackの例は、文間レベルでの漸進的な曖昧性解消であった。漸進的な曖昧性解消は、文内のレベルでも生じる。「さくらを食べる」という文では、「さくら」は文内の後方に位置する単語「食べる」を聞いて初めて曖昧性が解消される。

漸進的曖昧性解消で忘れてならない問題がもう一つある。それは、複数個の曖昧性が残されたとき、できることならそれらに優先順位をつけることである。優先順位をどう計算するかは、優先意味論の研究から確率やニューロンネットワークを用いる研究まで様々な提案がある。今後の重要な研究課題であろう。筆者の好みを言えば、正統的で保守的かも知れないが、ニューロ的な手法ではなく、確率的な手法に未来があるように思われる。この種の漸進的曖昧性解消は、我々人間が日常的に行っていることの様に思われる。そのために、人工知能の分野で興味ある研究が幾つかなされている。仮説推論や単一化操作を拡張したり、弁別ネットワークを拡張して実現しようとする研究などがある。

ところで漸進的曖昧性解消は、早期意味解釈のモデルに必須の機能である。早期意味解釈とは、情報が得られた時点で直ちに必要にして十分な意味解消を行うもので、それにより組み合せ的な爆発を防ぎ効率のよい意味解析を行うことができる。早期意味解釈のモデルでは、必要にして十分な曖昧性解消を行うべきで、推論を必要以上に深く進めぬことが望ましい。推論を深く進めすぎて、後からその誤りに気付いたとき、その修正には多大なコストがかかるからである。しかし現実には、必要以上に推論を深く進めないということの正確な定義が難しい。人間でも以前の推論結果を訂正せざるを得なくなることもある。翻意の問題である。袋小路文は文法解析における翻意を要す文である。後述するが会話などでは、後から言い間違いを訂正するための発話もしばしばなされる。このような翻意を許すことは、原理的に、言葉の理解における推論が非単調でなければならないことを意味している。非単調な推論は人工知能の重要な研究課題の一つである。また翻意には、誤解や相互信念の問題も含まれている。いずれも人工知能の根幹にも触れる重要な研究課題である。

音声と発話

人間は音声を用いて言葉を相手に伝えることができる。これをここでは発話と呼ぶことにしよう。音響信号レベルではなく言語レベルでの発話に関する研究は、これまで一部の言語学者によりなされてきたが、十分であるとは言えなかった。しかし、わが国ではＡＴＲ研究所が音響から言語レベルに至るまでの発話に関する基礎研究を開始し、活発な研究が行われている。

発話は、書かれた言葉と異なる。発話の速記を読まされて、あまりの文法逸脱性に驚くことがある。聞いている時には、さして気にもならず、逸脱性にも気が付かないにもかかわらずである。おそらく書かれた言葉は、読み手に何を伝えるべきかを書き手の側で考える余裕が（話し手とに比べてない）ことによると思われる。話し言葉は、言いたいこと（意図）が脳裏にあっても、それを組立てる（計画する）時に、文法的な知識の参照を十分に行う時間が無いこと、話している最中に、話の意図が変わったり忘却することがあり、それのため話の先頭と末尾とで、意味的な首尾一貫性が取れなくなることがあることによると思われる。いずれにしても、これは、発話で用いる脳中の記憶が一過性のものであることと関連する。

発話に特徴的な現象を思い付くままにあげると、
・文法からの逸脱性の大きい発話が多い、
・言い回しに冗長性が多く含まれる表現が多い、
・アー、アノー、ナルホド、エート、などといったあいづち表現が随所に現れる、
・翻意の発話がある、
・ツマリなどといって別の言い回しをすることが多い、
・割り込み発話がある、
・イントネーションなど、感情的な情報をのせることができる、
などがある。

音声理解システムは、こうした発話に特徴的な問題を解決しなければならない。上記した問題の幾つかは、これまで自然言語処理の研究として本格的に取り組まれてこなかったことのように思われる。一部の問題は、自然言語処理システムの頑健性と関係する。音声理解との関連で、今後この種の研究が一層活発になることが期待される。

円滑な発話を行うためには、相手が自分の意図をどの程度理解しているか、それを絶えず監視して発話する必要がある。この監視は「聞き手のモデル」を話し手が構築することに相当する。聞き手のモデルの精度が悪いと、発話そのものが相手に苦痛を与えることになりかねない。特に誤解に基づく会話は苦痛である。その意味で、聞き手のモデル化に関する研究、狭くは誤解の研究も需要であろう。音声に基づく発話理解（音声理解）システムを構築するためには、人工知能のほとんどの基礎研究の成果を集大成する必要があると思われる。その意味でチャレンジングな研究課題であるといえよう。それは、新しく発足する研究会で議論されるべき重要な研究課題であろう。

最後に、ＤＮＡの２重らせん構造の発見をきっかけにして、遺伝子や生命の本質に関わる重要な事実が次々に発見されたり、様々な現象が解明されたのに匹敵する自然言語処理に関する理論が、本研究会から生まれることを期待したい。それは意外に単純なものかも知れない。