講　　演

自然言語処理応用システムの新たな展開に向けて

－自然言語処理に関する講演会より－

東京工業大学工学部情報工学科
教授　田　中　穂　積

はじめに

東京工業大学の田中です。本日は「自然言語処理応用システムの新たな展開に向けて」ということでお話をさせていただきたいと思います。私自身は1971年から1972年にかけて，アメリカのスタンフォード大学に行く機会がありました。そのとき，MITのウィノラートという人がつくったシステムが非常に話題になっていました。自然言語処理というのはどんなものかなと思いつつ，MITにいる友達の所に遊びに行ったときに，テクニカルレポートの厚いドクター論文を手にして，こういうおもしろいことがあるのかと思いました。

長尾先生もちょうどそのときにスタンフォード大学に来られて，日本に帰ったらこれから自然言語を大いにやりたいとおっしゃったのを覚えています。私も帰ってから上司に自然言語処理のようなことをやってみたいということで，この分野に入ったわけです。実際には，1970年代の半ばくらいから始めて，もう20年近くやっているわけです。

自然言語処理応用システムへの期待

本日の話は自然言語処理応用システムということですので，そのあたりの話をまず最初にしたいと思います。実は1972年から科学技術庁で，5年に1度デルファイ法を用いた科学技術予測調査を行っています。情報エレクトロニクス分野全体で約160くらいの設問があるわけですが，その中から，自然言語処理に大いに関係があると思われるものをピックアップしてみましたところ7つありました。

デルファイ法の特徴は1度きりの予測調査ではなくて，予測の結果を回答者に1回フィードバックして，再度修正してもらうというものです。設問としてはまず，「話者を特定できる学習機能を内蔵した音声認識用ワンチップ集積回路が実現される」；音声認識も自然言語と大いに関係があるということでピックアップしてみました。実現予測時期というのは，大勢の人の平均を取った時期で，2003年です。重要度は0から100までの尺度で測られていますが，この場合64です。50を超えれば，重要なテーマの部類とみなしていいかと思います。

ちなみに非常に重要なテーマでトップ近くに上がっていたのは，チップに関するナノテクノロジーの話だったと思います。そういう非常に目標が具体的で分りやすいものについては，90を越す重要度がつけられたものもあります。総じて自然言語に関する話題についての重要度は，平均よりもかなり高い重要度をつけています。予測実現時期が2003年ということですが，自分だったらこれくらいの時期なのになということ

はあると思います。今世紀中には「実現」されないということだと思います。次の設問では「電子化された文書通信，電子化新聞等が家庭に普及する」：これは重要度がちょっと落ちていますが55です。実現予測時期は2004年ということで，2000年に入ってしばらくして「普及」する。

ここでちょっと注意していただきたいのは，質問の聞き方が4つに大別されていることです。1つは，原理的なものが「解明」される，それから，「実現」される，「実用化」される，「普及」するという4つの質問カテゴリに分かれています。第三の設問は，「図面を含む文書が理解できるシステム，例えば特許公報の要約が実用化される」：これは2000年に入ってすぐ「実用化」というのではなくて，2007年くらいです。重要度は68。

第四の「日本語の文章を音声入力（不特定話者による連続単語発声）することにより漢字混じり文に変換する汎用的な音声タイプライターが普及する」は，2008年に「普及」する。重要度は61です。一般に，自然言語関連の技術で「実現」されるとか「実用化」されるといった類いの問題については，2000年を超えてからになるだろうという調査結果になっています。

第五の「音声入出力によるポータブル型自動通訳機，簡単な日常会話を双方に通訳する」という設問では「商品化される」時期を聞いています。皆がこういったものを使える時期ということで，しかもポータブル型で簡単な日常会話ということですから，おそらく連続音声でなければならないということもあるのでしょう。そういう意味で，これも実現予測時期が2007年，重要度は64となっています。

第六は「図書，資料の要約抄録を自動的に行う装置（要求により任意の圧縮比で出力ができる）が開発される」時期は，予測では2010年になっています。重要度は60です。比較的重要度が低かったのは，最後の「行書程度の手書きの日本語文が読める装置がオフィスで普及する」。これは，実現予測時期が2005年，重要度48。この調査は，1992年11月に結果が出ていますが，実際このアンケートはこの2年半くらい前から準備されて始まっているということで，1990年くらいの皆さんの予測だと考えてよいかと思います。

実はこの調査とまったく同じ調査を，ドイツで行っています。ドイツでは今まで科学技術予測については非常に懐疑的で，そういうことをやっても無駄だ，政治的に利用される可能性もあるということで，こういった調査は今まであまり行われていなかったらしいのです。日本でこういう調査をやっているということで，日本の科学技術予測調査項目をそのままドイツ語に翻訳して調査を行い，その結果が，去年の暮れに公表されています。

それを見ると音声認識については日本よりも少し実現予測時期が早く，重要度が高い。また電子化新聞といったものについては，あまり重要ではないとドイツは見ている。また「図面を含む文章が理解できるシステムが実用化される」については，ドイツのほうが重要度が低い。「ドイツ語の文章を音声入力云々」という音声タイプライター関連では，重要度はドイツのほうが日本よりも高く，実現時期も日本よりも早く実現すると予測しています。

また自動通訳機については，これも音声がらみの話で，ドイツと日本とで，重要度は64，62とそう違いはないのですが，ドイツの予測実現時期は2005年で，少し早まっています。日本語の調査項目をドイツ語に直して行ったわけで，調査したという時期的な問題もあると思いますが，音声と翻訳，あるいは自然言語処理についての関心は日本よりも高い。また予測実現時期も早まっているということが，わかると思います。

自然言語処理応用システムの変遷

さて自然言語処理応用システムとして，これまでどういうものがあったか，代表的なものを挙げれば，日本語ワードプロセッサ，音声あるいは音声抜きの対話システム，機械翻訳システム，文章推敲システム，情報検索システム，文書要約システム，その他にもいろいろ考えられると思います。

日本は，日本語ワードプロセッサ，機械翻訳システムについては，日本電子工業振興協会でも委員会をつくって非常に熱心に対応してきたわけです。日本での技術的また研究面での実力は世界のトップレベルにあると言ってよいと思います。文章推敲システムも利用され始めている。情報検索システムは，以前からいろいろあり，日本ではこのあたりについても非常によく研究がなされてきましたし，実用的なシステムも一部で生まれ，商品化されたものも出ています。

しかし自然言語，言葉に関する応用システムは，人間にはかなわないところもまだ多々あるわけです。そういう意味で，先ほど科学技術庁のデルファイの予測時期で見たように，現在普及している日本語ワードプロセッサは例外的で，実際にそれがものになったり，世の中に広く「普及」するのは2000年を超えてからと見られているのだと思います。

最近の動向を見ますと，個々のシステムを，もっと大きなシステムの中の一部，あるいはコアとして位置づけるという動きに変わりつつあるのではないか。これまでの自然言語処理応用システムを統合化した，例えばオフィスにおける文書処理支援システムといったものが非常に重要になってきているのではないか。その背景には，膨大な文書がオフィスの中にあふれている。そういう現実があると思います。また国際社会での日本の地位が高まるにつれて，いろいろな言語の文書／書類を作成しなければならないということもあります。

情報検索も最近注目されています。例えばIBMも，以前は音声や機械翻訳の分野にも手を出していました。検索するためには，言葉の意味といった概念の体系を必要とするのは明らかです。そうした観点から，これまでの自然言語処理の研究を応用した情報検索の研究に力を入れ始めています。これについては，『自然言語処理技術の動向に関する調査報告書』の第2章の後ろの部分に，情報検索についてサーベイが載っていますので，参考にされるとよいと思います。最近の動き等が手際よくまとめられています。手前みそですが，委員の皆さんなかなかよく勉強して広範なサーベイを行い，うまくまとまっていると思います。

最近注目すべき動向としては自然言語だけでなく他のメディア，画像(動画像)，図面，音声といったものを統合化したマルチメディア，というよりもむしろいろいろなモードで，人間がシステムとコミュニケートできるマルチモーダルなインタフェースが重要になってきています。

今まではディスプレイ上の文字と対話していたわけですが，動画像技術を使って，ディスプレイ上の人間の顔や目や口が動く。応答の内容に応じていろいろな表情をして，音声で応答する。当惑したときはそのような顔つき目つきをするといったシステムもソニーの研究所で開発されているわけです。このように単に自然言語だけでなく，それ以外のメディアを統合化したマルチモーダルなインタフェースの研究にも，これから重点を置いていかなくてはいけない。

もう1つの問題は，既存の不完全なシステムと，われわれはどうつき合っていくかということです。これはヒューマンインタフェースの問題と関連するわけですが，きちっとやっていかなければならない。自然言語処理応用システム自身の評価の問題も，これから十分に行っていかないと，技術が着実な進歩を遂げにくくなる

ので重要です。

すでに述べましたが自然言語の技術にはいろいろ困難な問題もあるわけです。このへんで自然言語処理技術の見直し，特に解析技術の見直しをしておく必要があるのではないかと個人的には思っています。非常にコンベンショナルな分け方ですが，自然言語処理技術を形態素解析技術，統語解析技術，意味・文脈・談話解析技術の3つに分けました。実際にはこの3つは複雑に絡み合っていて，これらをどう統合化するかという研究も重要ではないかと思います。これが，多様な解析レベルの知識の統合化という問題です。これについて幸いにして，いくつかおもしろい技術も現れ始めているという気がします。

分かち書きに関連する形態素解析技術については，一見完成した技術と見られがちですが，実際には問題があります。よく学会の発表などで，99%以上の精度で分かち書きができたという報告があります。ところがそれだけの高い精度を出すために，実際には辞書に複合名詞のようなものもどんどん登録して精度を上げている場合もあるわけです。分野にチューンナップした辞書を使うことによって，その精度がはじめて得られたと言えるわけです。

しかし実際には複合名詞のようなものを辞書に全部登録するわけにはいかないので，それをどう分割して辞書にある形態素の並びにするかということが，まだ大きな問題として残っているわけです。情報処理学会などの研究会の発表でも，一時期統語解析についての研究発表が非常に多かったのですが，数年前にまた形態素解析技術についての研究が息を吹きかえしたという経緯があります。このあたりの技術は，ワードプロセッサの技術と非常に密接に関連があるわけで，もう少し研究しなければなりません。

統語解析技術についてはかなりよいアルゴリズムが開発されてきて，そのツールも整備されてきているわけです。1つの問題は，文法をどのように書けばよいのかということがあります。計算機上で使える文法の理論です。人間にわかる文法理論ではなく，計算機の上に乗る計算機が分かる文法理論でなくてはいけない。言語学者は必ずしも計算機の上に乗る言語理論を目指して文法理論なり言語理論をつくっているわけではないので，われわれ自然言語処理を研究している人間にとって使える文法理論については，それなりの新しい考え方も必要なのではないかという気がします。人間にわかる文法理論だけではいけないということです。

3番目の意味・文脈・談話解析技術については，いろいろ難しい問題があります。談話，文脈解析になりますと，1つの文だけで閉じた技術であってはならない。1文の中での関係ではなくて，文と文との間の関係といったものも考慮して解析しなくてはいけない。

従来，ともすれば1，2の技術が中心で，3の技術は，非常に局所的な構造を対象にしていた。1つの文章の中でいうと名詞句の中での構造，あるいは単文の中での構造です。しかし現実の文章はもっと複雑で，いくつかの文が1つの文章になってしまっていたり，かなり複雑な構造を取る。そうすると全体構造というものを，どのように部分構造から把握していけばいいか。部分構造をきちっと解析して，それを積み重ねていけば全体構造が把握できるかということも問題になってくるわけです。このあたりについての研究が今後必要だと思います。

また音声対話というものを，自然言語処理の研究者も1つのチャレンジとして受け止めなければいけない時期に来ているのではないか。音声対話を調べてみると，未知語がよく出てきます。音声認識システムの限界で誤った語を認識した結果，未知語が出てくるということもあるでしょうが，音声対話システムで使われているボキャブラリは，せいぜい数百から2～3,000どまりです。そのため未知語の問題がどうしても避けられない。未知語が現れたらすぐ止まって

しまう解析技術では困る。タフな解析技術でなければいけない。

また今日私のしゃべっている内容を原稿に起こしてみると分ることですが，おそらく助詞が脱落していたり，誤っていたり，アーとかエーとかいう無意味語が入ったりしています。実際の音声対話にはこうしたいろいろな現象があります。それに耐えてきちっと解析する技術を開発しておく必要があります。機械翻訳についても，似たようなことが言えると思います。書いた言葉であっても，こういった問題は必ず出てくるわけです。したがってタフな解析技術が，重要なものとして浮かび上がってくる。

知識については，あまり細かいことを言ってもできることは限られています。知識が必要だということは，昔からいろいろな人が言っています。知識には深い知識と浅い知識があると思います。浅い知識の利用というのは，意味まで立ち入らないレベルの文法的な現象を分析した結果を知識として蓄えておき解析に役立てる。ヒューリスティックスと呼ばれているものも，このようなレベルの知識がほとんどです。一方，深い知識も必要になる。この2つはパラレルに研究，あるいは開発していかなければならない。

現実にどちらが難しいかというと，深い知識の利用が難しいわけです。そういったことを十分認識しながら知識を利用していく。また対話などでよく言われているのは，意図やプランあるいはユーザモデルといったものをどうシステムの側が獲得していくかという問題もあると思います。システムの使用者が初心者か熟練者かどうかをシステムが推定し，それに応じた応答を行うのが，ユーザモデルの問題です。こうした研究も近年進展している。

自然言語応用の立場からのプラス要因の1つはメモリが安くなり，大規模なメモリが，非常に安価に使えるようになってきたことが挙げられます。これをどのように使っていくか。またCPUの価格の低下とともに超並列の研究も進められています。数百だと超という言葉は使いたくないのですが，数百でも超という言葉を使う人もいるようです。それでも以前は，とてもそんな数のCPUを同時に使うことは考えられなかった。しかしそれが現実のものになってきている。こういう資源の活用もこれから考えていかなければいけない。

ちなみに私どものところでは頑健な解析技術ということで，2つくらい未知語が含まれていて単語の脱落もある文を，どう解析するかという研究をしています。文の長さが単語の数でせいぜい20とか30くらいの文を，既存の方法を使って，しかも意味，文脈解析などをしないレベルの解析を行っただけでも，今のSPARCワークステーションで1日走らせても結果が出てこない。それくらいの計算量になっているようです。そういうことで超並列に計算を行うということも，これから非常に重要になってくるし，それが現実のものになりつつある。

コーパスの利用も，これから重要になってきます。コーパスというのは，電子化された文を大量に集めたものをいいます。報告書の総論のところでも触れていますが，アメリカではこのあたりについては非常によくやっています。いろいろなコンソーシアムをつくって皆で利用し合って，大量のコーパスを整備している。これらのコーパスは，用例ベースの機械翻訳だとか統計情報を使って解析結果の曖昧性の解消を行う研究などに使われています。

コーパスは，わが国でもきちっとしたものを整理しておく必要があります。それが将来の新しい自然言語応用システム展開のベースになるものだと考えられます。このあたりについての努力が，今後なおいっそう必要だという気がします。

電子化辞書研究所もいろいろなコーパスを集めておられるし，各メーカともいろいろなコーパスは持っているようです。数千万例文のコーパスが必要だという話もあります。現実には量

的な問題の克服には，著作権の問題があってなかなかうまくいかないようです。

自然言語応用システムに対する別のプラス要因ですが，大規模辞書が利用可能になっている。電子化辞書研究所で開発したEDR辞書の評価版が公開されました。私どもではこのEDR辞書を使ってみましたが，われわれは，十分に使えると判断しています。こうした大規模な辞書にはバグがつきものです。これからそれを皆で使って直してより良い辞書にしていくことが必要だと思います。

米国の場合ですが，ペンシルバニア大が開発したコーパス（ペントリーバンク）を使って仕事をした場合には，必ず謝辞を書いて感謝の意を表明する。実際に著者に聞いてみると，実はバグが多く使うのに苦労したと言うのですが，全体の方向としては使ってよかったと見ているわけです。そしてバグレポートを送ると，すぐに直したバージョンが返されフィードバックのサイクルが早い。使った人が集まってシンポジウムを行いながら意見交換をする場を作っていくことも彼らはうまいと思います。

一方，それほど大きな辞書ではないのですが，情報処理技術振興財団の開発したIPALという辞書があります。これは言語学者の卵も含めて，ていねいに作った比較的小規模な辞書です。これも非常に安価に手に入ります。こういったものを大いに使っていける状況になってきています。

ここで，コーパスを利用したわれわれの研究の一例を話します。実際に使ったコーパスは，愛知淑徳大学の田中康仁先生がつくられたもので，４文字漢字という漢字の並びの16万ほどのコーパスです。それを利用させていただいて，複合名詞の解析をしました。

複合名詞の解析は，なかなか難しいのです。１つの理由は，日本語は語と語の間に明示的な区切りがない。複合名詞は文と同じような構造を持つこともある。漢字からなる名詞の並びですから文法的な手がかりが少ない。数が多いので辞書に登録するのは困難だという問題があって，なかなかやっかいな問題です。形態素解析が成功したからといって，それで終わりかというとそうでもない。

例えば「歩行者通路」というのは，歩行／者／通路，歩／行者／通路のように少なくとも２通りの切り方があるわけです。それだけではなくて，一方の切り方が正しいとしても，どういう構造にするかという問題が残る。歩行／者／通路という分け方が正しいとしても，[[歩行／者]通路]とする構造が正しいか，[歩行／[者／通路]]とする構造が正しいかを決めなければなりません。構造まで含めて解析するのが形態素解析の最終的な目標なわけです。そこをうまくやっておかないと機械翻訳の場合には，翻訳結果がおかしくなる。また，読みを音声で出すときも，どこにアクセントを置くかが構造により変わってくるので，音声合成をする場合にも問題がある。

構造まで含めた形態素解析については，IBMの藤崎氏の研究がありますが，だいたい76%くらいはうまくいく。それも文字数にして平均4.2くらいの複合名詞についての研究です。

われわれは田中康仁先生の作られた４文字漢字のコーパスを用いて，どのような意味カテゴリーの並びが，１つの構造としてまとまりやすいかという統計データを抽出しました。このような統計データを主として用いて，複合名詞を解析したところ，4.2文字ほどの複合名詞の解析では，90%以上の精度で，正しい構造が得られることが分りました。

EDRのコーパスからは，私共のところの徳永先生が中心になり文法を抽出するという研究をしています。現在使っているEDRコーパスは例文数が３万4,739で，人が括弧による構造を付加した括弧つきコーパスになっています。文法規則を，これから自動的に抽出する実験を進めていますが，明るい見通しを得ています。将来

は，こうしたコーパスを大量に集めることで，大規模な文法規則を自動的に獲得することも夢ではないと感じています。括弧つきコーパスも，私がここでお話しした以外の使い道もいろいろとあると思います。ぜひ今後全体を公開していただいて，皆さんが使い合って，よりよいコーパスに整備していただきたいと思っている次第です。

こうした努力の結果作成された大規模コーパスは自然言語処理応用システムをつくる場合の体力のようなものになるので重要だと思います。

一方，言語データの分析を十分に行わなければならない。長尾先生はこれをヒューリスティックスと言っていますが，実際に言語データはどういうものかということを分析しなければいけない。言語学者によるトップダウン的なアプローチではなく、ボトムアップ的な立場での言語の研究の重要性を指摘したいわけです。

アメリカではワードネットという概念間の関係についての知識ベースが開発されています。7万くらいの語義で，しかも関係も上位下位だけでなく同義，反義，全体部関係についての意味ネットワークが，フリーのソフトウェアとして誰でも使えるような状況にあります。これと似たものでわが国でよく使われているのは分類語彙表です。これは3万2,600語が798，1,000くらいの意味分類項目に分けられています。

問題は，われわれの使っているこの分類語彙表が非常に荒いということです。ある単語の語義を全部すべてではなくて，代表的な語義だけを考えて分類していくという考え方で作られています。EDRの概念辞書は，40万概念を上位／下位関係を基に配列した労作ですが，これとワードネットはちょっと関係の数が多過ぎます。ですから分類語彙表とワードネットの中間，5,000くらいの体系ができないものかということも検討しています。

最　後　に

最後にまとめに入らせていただきます。われわれは望まれるヒューマンマシンインタフェースとして自然言語に注目しているわけです。1つは初心者がすぐ使えるインタフェースとして（媒体として）自然言語は非常に効果的だからです。一方熟練者でも使いやすいインタフェースの研究も必要です。この2つの要素をどのように調和させるか。そのためにはユーザモデルの研究が必要だという気がします。おそらくこれからコンピュータは超並列，高信頼な方向に向って進化します。大規模な知識ベースと知識システムも片方で構築される。それらを結ぶ情報ネットワークが整備されるでしょう。そして，これらをインテグレートした知的なシステムが開発される。そのときシステムと人間との間のインタフェースが，最も重要なものとなると思います。これは，ある意味で，ソフトウェアの問題です。マルチメディアと情報スーパーハイウェイの問題が，マスコミをにぎわしています。私見になりますが，私は，こうしたハードとインフラの問題は，時間とお金で，ある程度解決ができるような気がしています。マルチメディア関連の産業が，大きく育つかどうかの決め手は，私はソフトウェアにあると思います。その中のコアが，システムと人間との間のインタフェースに関するソフト，特に自然言語に関する優れたソフトウェアであるという気がします。以上で私の講演を終わらせていただきたいと思います。

質 疑 応 答

質問　おもしろく聞かせていただきました。デルファイ予測の調査についてですが，どういう分野の人を対象としたのか，割合や人数も教えていただきたい。また先ほどソニーの研究所

のデモンストレーションのお話がありましたが，これは87年くらいに北大（現慶応大）の安西先生が，札幌のレストランの案内で実験モデルをつくったと思います。また実際の実用モデルとしては，92年くらいにトスバーグというハンバーグの注文システムを東芝で作っています。ソニーのものは，どういう用途のものなのでしょうか。

また，EDRの辞書に関しては，ああいう種類のものを共有・調整してよりよいものに変えていけばいいということでしたが，EDR内部の人の話を聞いても，概念辞書等の記述情報というのは担当者しか内容が分からないということがあるので，今後はガイドライン的なツールを使っていくという方向にもっていかなければならないと思います。そのへん，実態はどういう方向にいっているのでしょうか。

田中　答えやすいほうからお話しします。最後については，たぶんこの中にEDRの方がおられたら，その方にお答えしていただくのがいいかと思います。今まではなかなか全体としてまとまらなかったので，オープンにできませんでしたが，これからはオープンにしながら，内容を洗練していくとか考え方を整理していくという方向に向うと思います。今まではとにかくつくるのに精一杯で，なかなか他のことが考えられなかったという事情があったのではなかったかと推察しています。

最初のデルファイ予測についてです。デルファイについては，「あなたの専門度はどのくらいですか」といった項目もあって，そういうものも込みで予測結果の判断をしなくてはいけないわけです。回答者は各項目につき70名から100名ほどが回収され，そのデータを元にして，デルファイ予測が行われました。

ドイツ側は，もう少し少なく，30名とか20名くらいの回答数のものもあります。しかし，両国とも実現予測時期がだいたい似ているのがおもしろいところです。調査した時点が違うので，少し技術が進んだ分，ドイツ側には音声についての強気な見方がありますが，それ以外はほとんど同じです。160項目中トップ10に現れたものでドイツと日本で重複しているのは，5項目もあります。トップ10の中で注目されるのは，素子も含めて7割から8割は，コンピュータに関する技術です。

ソニーのシステムの原型は，以前文部省の特別推進研究で現在中京大学の戸田先生が北大におられたときに似た研究をしています。ネネという名前だったと思います。ただそのころの技術よりも今の技術がかなり進んで，画像処理技術，処理スピードが進んでいます。これらを組み合わせてみると，ソニーのシステムは，あのころ見たシステムよりもずっと臨場感があっておもしろいシステムになっている。画面上の応答が，図や音声だけではなくて，人間の顔のさまざまな表情の変化として見られるので，対話者は冷たい機械との対話とは異なる印象を持つことができます。図，音，動画を応答に合せて同期させた，マルチモーダルなインタフェースの実験であると考えられます。原理的なものは，あの時代に戸田先生たちが考えられていたと思います。大筋のところはそのように考えていいのではないでしょうか。