自然言語処理技術とは何か

田中穂積
東京工業大学

1. はじめに

18世紀の産業革命に始まる大量生産技術の進歩は，労力の軽減にともなう余暇とさまざまな物質的恩恵をもたらし，以前とは比較にならないほど便利で豊かな物理空間を生み出してきた。一方，50年ほど前に誕生したコンピュータ技術の急激な進歩は，大量の電子的な情報を高速に扱うことを可能にし，主として紙と印刷技術に支えられたこれまでの情報の蓄積法，情報の伝達法，情報の処理方法を一変させた。最近のパーソナルコンピュータの普及、ワープロの普及、インターネットに代表される情報ネットワークの広域化は、我々を取り巻く情報環境、情報空間を質・量ともに大きく変えつつある。急激な変貌を遂げつつあるこの情報空間を、われわれにとって快適なものにする技術として、自然言語処理技術は極めて重要な役割を果たすものと期待されている。自然言語で書かれた文書が、情報空間を満たすもっとも重要なメディアの一つだからである。以下では、自然言語処理技術の現状について概説してみたい。

2. 自然言語処理技術の目指すもの

自然言語で記述された文書は、ワープロの普及とともに電子化が急速に進み、大量の電子化文書が家庭に、オフィスに、ネットワーク上に溢れている。このような状況では、大量の文書を処理し、文書中に含まれている情報の正確な検索が重要となる。それには、検索意図の理解、検索意図に沿う文書を取り出すこと、その要約技術が必要になる。最近の自然言語処理技術は、情報検索の基礎技術としてその重要性がとみに増している。

一方、地球規模の情報ネットワークの構築とともに、アクセス可能（取りだし可能）な電子化文書の多言語化が進んでいる。機械翻訳技術は文書の多言語化の問題に対処するためのキーテクノロジーである。自然言語処理技術なくして機械翻訳はありえないわけであるから、現在の機械翻訳システムの未熟さは、自然言語処理技術の未成熟さとおおいに関係している。

最近のパーソナルコンピュータの性能向上とその爆発的な普及により、誰でも使える機能（ヒューマン・インタフェース機能；対話機能）が求められるようになってきた。自然言語や、音声や映像といっ

たマルチモーダルな媒体を駆使した対話機能の充実が緊急の研究課題になっている。

自然言語処理技術者の夢は、あたかも人と話しをするようにコンピュータに話しかけたり、指令を出したり、コンピュータから言葉で答をもらったり、ある言葉を別の言葉に翻訳したり、話しをすればそれを文書にしてくれたり、意図した文書を探し出してくれたり、それを要約してくれるコンピュータの実現にあるといえる。

3. 理論と自然言語処理技術

ここで、人文学と自然言語処理技術との関連について述べておきたい。人文学の一つの柱として、人間の使う言葉の仕組みを研究する学問分野がある。言語学、心理学、哲学などがそれである。後述するように、これらの学問の理論は、部分的には自然言語処理技術に取り込まれて役に立っているが、到底十分とはいえない。言語学で言えば、それが形式化されず、人間の直観に訴える理論に留まっている限り、その理論をコンピュータで使いこなすことができない。

たとえば言語学者が使う「新しい情報」、「古い情報」について考えてみたい。言語学者は、助詞の「は」と「が」の機能の違いの一部を、これらの用語で説明することがある。「が」は新しい情報マークし、「は」は古い情報をマークするというのである。この理論によれば、助詞の「は」と「が」が英語の"the"と"a"の使いわけとも関係しているので、この理論を以前に検討したことがある。

自然言語処理の立場からのこの理論のもっとも単純な解釈の一つは、助詞「が」が付いた名詞句が現れたら、それを新しい情報としてデータベースに登録する。助詞「は」が付いた名詞句が現れたらデータベースにすでに登録されているものとして、そこを検索しにいく、というものである。もし助詞「は」が付いた名詞句がデータベースに存在すれば、それは古い情報であり、その名詞句の翻訳結果に"the"を冠する。もちろん助詞「は」と「が」には、情報の新旧のマーク以外の機能もあるが、以下では二つの助詞の機能を情報の新旧にだけ限定して考察を進める。次の文を考えてみよう。

> 「中古車を買った。お金は予想したほどではなかった。」

名詞句「お金」に付いた「は」は、情報の新旧という観点からどう解釈したらよいだろうか。前文までに現れる名詞句は「中古車」だけであり、名詞句「お金」はデータベースにない。それにもかかわらず、「お金」には助詞「は」が付いている。これは「古い情報」なのだろうか。答は「しかり」であり、情報の新旧という観点から次のように説明できる。

「中古車」の売買には「お金」が介在する。ということは、前文を聞いた段階で、われわれは「お金」の存在まで推論していると考えるのである。前文に現れる名詞句（「中古車」）を新しい情報としてデータベースに登録すること以外に、その名詞句を聞いて間接的に連想しうる名詞句（「お金」）をもデータベースに登録しておく。連想を働かせて登録したものを後から言及する場合にも、それを助詞「は」でマークすると考えるのである。

これで、助詞「は」と「が」の機能と情報の新旧に関する説明がやや厳密になった気がするが、それでも十分ではない。コンピュータは、知識を用いて連想を行なうことができるが、現実には、連想の連鎖をどこかで断ち切らなければ、データベースに登録する情報が溢れてしまう。「お金」から「金融」や「経済」が連想されたとしても、これらはおそらく後から言及されることはないだろうから、データベースに登録するのは止めにしたい。このように何をデータベースに登録するかを見究めることが、自然言語処理技術では大きな問題になるのである。

われわれは連想という用語を使って問題を解決した積りになっていても、自然言語処理技術の立場からは、どこまで連想を進めるべきかを定義しておかなければならないことが分かる。これは、認知科学的にも興味ある研究課題であるが、現在の言語学で

は、このレベルの詳細さの厳密性を追求していない。このことが、自然言語処理で利用可能な言語理論の範囲を狭めていると言える。

自然言語処理技術に取り込まれている言語理論もある。たとえば主辞駆動型句構造文法（HPSG：Head-Driven Phrase Structure Grammar)である。HPSGで使われている素性構造（文法情報の記述形式）は、細かい文法的な制約の記述が可能なため、自然言語処理の立場からも拡張が試みられている。

大量の例文（コーパス）を集めて、そこに含まれる統計的な情報を利用した自然言語処理技術の新しい流れがある。これはコーパスベースの自然言語処理技術とよばれている。大量のコーパスを用いることは、ボトムアップな研究のアプローチを強いる。これまでの言語学者は、ともすればトップダウンに理論を構築し、それを検証するために比較的少数の例文（それも自己の理論に都合のよい例文だけ）を取り上げることもなかったとはいえない。大量のコーパスが容易に利用可能になれば、網羅性の点で言語理論の構築法にも影響を与えることになるだろう。

4. 自然言語処理技術の要素技術

自然言語処理の要素技術を大別すると、1）形態素解析、2）構文解析（統語解析）、3）意味解析、4）文脈解析、5）文章生成がある。1）と2）は比較的研究が進んでいるのでやや詳しく説明し、3）から5）はまとめて説明する。

4.1 形態素解析

形態素は文を構成する最小の言語単位である。日本語など、単語と単語との間に空白を置かない言語では、形態素解析は特に重要である。後続する自然言語処理では、辞書に含まれている情報を利用するために、文を構成する形態素を切り出し認識しておくことが必要になるからである。これをわかち書きとよぶことがある。

たとえば「棒に当たる」は「棒 に 当た る」とわかち書きされる。形態素解析では、わかち書きして抽出した連続した文字列が形態素であることを保証するとともに、互いに隣接する形態素が接続可能かどうかを調べる。隣接した形態素間の接続可能性は2次元の表として表すことことがあり、これを接続表とよぶ。「当たる」の場合でいえば、「当た」が五段動詞の語幹であり、「る」は五段活用、上一段、下一段動詞の語尾になりうるが、接続表を用いれば、五段活用の語尾のみが五段活用の語幹に後続可能であるとして形態素を認定することができる。接続表の他に、字種の変わり目をわかち書きの情報として利用することもある。新聞記事など、わかち書き対象文書を限定すれば、98パーセント程度のわかち書きの精度をもつ日本語形態素解析システムが公開されている。

形態素解析の難しさは、接続表を用いただけでは、わかち書きが一意に決まらないことである。「くるまでまつ」には、少なくとも「くるまでまつ（来るまで待つ)」「くるまでまつ（車で待つ)」というわかち書きが考えられる。いずれも意味的に正しい形態素解析結果であるが、実際には、接続表には意味的な制約が書かれているわけではないので、意味的に異常な多数の形態素解析結果が得られてしまう。その数は、文の長さが長くなればなるほど急激に増える。多数の解析結果から妥当なものを選び出すために、単語間の接続可能性の度合を統計的な尺度でスコア付けする技術が研究されている。形態素解析の結果は、使用語彙調査などにも威力を発揮するだろう。

本節の冒頭で、わかち書きのない日本語文の場合に、形態素解析が重要であることを指摘したが、いかなる言語であっても、音声認識の場合には単語や形態素の認識が必要になる。音声認識の研究分野では、これを自動セグメンテーションとよんでいる。

4.2 構文解析

構文解析とは、文を構成する語の品詞の並びが、日本語なら日本語の文法にかなっているかどうかを判定するとともに、解析結果として構文構造（句構

造を表す樹状構造（木構造）や、係り受け関係を表す構造など）を抽出する。構文解析を統語解析とよぶこともある。ある言語の品詞の並びを表現する形式として、文脈自由文法(CFG ; Context-Free Grammar)とよばれる形式がある。ここで極めて単純な日本語文法のCFGによる記述を考えよう。

1) S→PP S　　3) V →Vs Ve
2) S→V　　4) PP→N P

N、P、PP、S、Vは、それぞれ名詞、助詞、連用修飾句、文、動詞を表す。CFGを用いた構文解析法には大別してトップダウン法とボトムアップ法がある。再び「棒に当たる」を取り上げる。

トップダウン法による構文解析は次のように進む。規則1）を適用して記号SをPPとSに分解する。左から右に解析が進むとして、つぎはPPを分解することになるが、今度は規則4）を適用してPPをNとPに分解する。Nはこれ以上分解できないので、解析対象文の先頭の「棒」の品詞を調べるとNであることが分かるので整合する。つぎはPであるが、これも「棒」の次の語「に」の品詞を調べると助詞Pであることが分かる。こうして「棒 に」が規則4）によりPPにまとめられる。規則4）の適用は成功し、規則1）で分解したPPの内部構造が求まったことになる。そこで、次はPPの右のSの解析に進む。このSに規則1）を適用してSをPPとSに分解する。ところが残りの「当たる」はPPにまとめることができない。（これはなぜか？読者の宿題として残す。）そこで、今度は規則1）の代わりに規則2）を用いてSをVに分解する。以下詳しくトレースしないが、残りの「当たる」はVsとVeの並びであることが分かり、Vとしてまとめる。以上の処理により、解析対象文から下記の構文構造を抽出することができる。

[S [PP [N 棒] [P に]]
　[S [V [Vs 当た] [Ve る]]]]

トップダウン法では、解析対象文がS（トップ）であると仮定して解析を始め、規則の適用によりSをより細かい記号に分解することを繰り返してゆく。分解が不可能なレベル（ボトム）に到達して初めて解析対象文を構成する語の品詞との整合性を調べる。

これとは逆に、ボトムのレベルの語の品詞の並びを、より上位の記号にまとめることを繰り返し、文の末尾にきた時、文全体がS（トップ）にまとめあげることができれば、解析成功とする方法がある。これをボトムアップ法とよぶ。

構文解析でも、実際に文法の体系が複雑になればなるほど、多数の構文解析結果が得られる。これを構造的な曖昧性とよぶ。構造的な曖昧性の解消には、自然言語処理技術者の苦闘の歴史が刻み込まれている。最近、統計的な情報を用いて構造的な曖昧性を解消する技術が研究されている。

4.3　意味解析、文脈解析、文章生成技術

意味解析と文脈解析は判然と区別ができるわけではない。意味解析の典型は、格文法で言う深層格構造を文から抽出することである。たとえば「犬が棒に当たる」から深層格構造として［出来事＝当たる、動作主＝犬、目標格＝棒］を抽出する。文に代名詞や省略が含まれているときには、解析対象文の前後（文脈）を調べる必要がある。意味解析でも文脈解析が必要になることがある。たとえば代名詞を含む"A taxi is comming over there. Let's take it"という文の翻訳では、itが何を指すか（何を意味しているか）を決めない限り、takeをどう訳すべきかが決まらない。takeに多義性があるためである。Itが前文に現れるtaxiを指していることが分からなければ、これを「乗る」と訳すことができない。語義の曖昧性解消は困難ではあるが意味解析の重要な研究課題である。

2で述べた情報の新旧は文脈解析技術と関係しているが、厳密な理論化がなされていないことはすでに指摘した。対話文では、言外に述べられていることが実は重要であることが知られている。発話行為とよばれる理論である。この理論も自然言語処理の立場からは十分な厳密性を持っているとは言えない。発話行為論の今後の研究の進展に期待したい。文章

生成は、意味構造や概念に近いレベルの深い構造から文章を作り出す技術である。これは、曖昧性が解消された構造を出発点とすることができるという利点はあるが、その構造からさまざまな文を作り出すことができるという文生成に関する曖昧性が問題になる。生成した文の良否を判定する評価基準が必ずしも明確ではないことも問題である。省略や代名詞化を行なう文章生成技術も確立されているとは言えない。高度な文章生成技術の研究は緒についた段階にあるといえる。

5. おわりに

自然言語処理技術には長期を要する研究が必要である。言語学からはより厳密な理論が求められている。自然言語処理の立場から、言語学に新たな知見をもたらす可能性もあることを指摘しておきたい。2. で述べた言語理論HPSGの素性構造は、自然言語処理の側からの知見を言語理論に取り込んだとみることもできる。今後わが国で、両者の交流が盛んになることを期待したい。そのための学会（言語処理学会Tel. 06 (6873) 2301）も設立されている。

これまで言語学は意味の問題を避けてきたように思われる。筆者は、この困難な問題の解決に取り組む言語学者の数が増えること、書き言葉だけでなく、話し言葉の研究がもっとなされるべきことを指摘して結びとしたい。これらに興味のある読者は、電子情報通信学会から出版された「自然言語処理—基礎と応用」の一読をお勧めする。

*　　*　　*