

解説



自然言語処理技術の最近の動向

自然言語処理技術の最近の動向^{†*}長 尾 確^{††}

1. はじめに

近年、自然言語処理の分野において、新しい研究の流れが起こっている。本特集では、この分野の若手研究者の斬新な視点からみた最近の重要な研究動向を解説する。また、本特集には、今後の研究の方向について洞察するというねらいも込められている。

本特集は、自然言語処理関係の研究者・技術者・学生を中心とした読者を対象として企画されている。また、読者は、構文解析などの自然言語処理の基本的手法に関する知識を有することを前提としている。つまり、本特集は一般的な入門編というより、上級編を目指したものである^{**}。

本学会誌の自然言語処理に関する特集には、第30巻第10号(1989年10月)の特集「自然言語理解」(天野真家編)がある。その特集は、形態素解析から始まり談話理解、さらに応用システムまでの話題を網羅したものであった。約3年隔てた現在において、いったいどんな変化が起こったと言えるだろうか。確かに応用システムに関しては、革新的なものはいまだに出現していない。しかし、研究分野では、前特集では述べられていなかった新しい動きが確実に起こっているのである。

たとえば、曖昧さの解消を、制約(constraint)や選好(preference)としての知識を必要に応じて用いる動的な問題として捉える流れや、文法的に不適格な文(ill-formed sentence)などを受け付ける頑健な処理を目指すアプローチ、また、超並列

処理の自然言語への応用、また、文(sentence)を越えた、意味的なつながりのある文章(text)の生成、そして、構文解析(parsing)と文生成(generation)を統一的な枠組で処理する、というような新しい研究成果が次々と発表されるようになったのである。

それは、自然言語処理の基礎技術が整備され、次なるステップに歩み出していることを意味している。しかも、そのような基礎技術を踏まえて、何が本質的な問題で、また何がこれから重要な方向なのかを見据えた上でのステップアップなのである。たとえば、いくら構文解析技術が進歩しても曖昧さの解消は、いまだに困難な問題として残されている。曖昧さの取扱いは自然言語処理において本質的な問題なのである。また、省略などを含む不完全な文の処理も曖昧さと並んで本質的である。この問題は以前はスコープの外にあったが、ようやく本格的に取り組まれるようになってきた。また、自然言語の意味が高度の文脈依存性をもつことはこれまでも多く議論されてきたところであるが、それを具体的にどのように扱うかに関しては、あまり提案されてこなかった。最近の超並列処理の流れは文脈依存のアポステオリな情報を扱う枠組として大いに期待されるものである。それは、統語論と意味論、語用論の自然な融合を目指しているからである。また、深い意味処理を必要とする文章生成の最近の流れでは、生成のためにどれだけの意味情報が必要なのか議論されてきている。ゴールのはっきりしていなかったこれまでの談話理解研究に比べるとかなり生産的であると言える。そして、人間の言語処理能力をよくよく考えてみると、言語理解と産出(深い意味の生成)がまったく別のモジュールで処理されているとは考えがたい。そこで、統一的な枠組で自然言語処理のさまざまな問題を扱おうという

[†] Recent Technical Trends in Natural Language Processing by
Katashi NAGAO (Sony Computer Science Laboratory Inc.).

^{††} (株)ソニーコンピュータサイエンス研究所

★ 本解説は、通常「編集にあたって」として述べられるものを拡張したものである。本特集の編集意図や各解説の概要については、本解説の1.と2.を参照のこと。

★★ もっとも、本学会誌は、そのへんぬりがなく、連載講座「自然言語処理入門」を用意しているので、専門家でない方もご心配なく。

考えが起こってきた。つまり「自然言語処理における統合」が一つの重要な研究テーマとなったのである*。中でも、自然言語文の解析と生成の統合は、最近特に注目され、重要な成果が得られている。

本稿では、2.で、本特集に含まれる解説の簡単な紹介を行い、3.で、本特集に含まれなかった自然言語処理のその他の話題について簡単に述べる。

2. 本特集の構成と概要

本特集には、本稿以外に、5編の解説記事が含まれている。これらの解説は、オムニバス形式で編集されている。これは、おのおのの執筆者が、互いに密な関連性をもつ諸問題に関して、それぞれ独自の視点で、問題を整理し、研究動向をまとめているからである。

以下で、これらの解説について、簡単に紹介する。

2.1 自然言語処理における曖昧さとその解消

最初の解説では、自然言語処理の基本的な問題の一つである曖昧さの扱いについて述べている。

コンピュータによる自然言語処理の初期の段階から、曖昧さの問題は最も重要なものの一つとして認識されてきた。自然言語文を解析するための初期のアプローチでは、文法を詳細化したり、意味的な条件を加えたりする、いわば制約に基づく曖昧さの解消が行われていたが、その後、曖昧さの候補が現われてから文脈に応じて適切なものを選択するという、いわば選好に基づく曖昧さの解消が行われるようになった。そして、この問題の動的な性質が明確になってきたために、情報単調性と呼ばれる、動的問題に積極的に取り組む考えとそれに基づく枠組がいくつか提案されるようになった。たとえば、素性構造と呼ばれるデータ構造とその構造上の単一化という手法がある。これは、現在で最も一般的な自然言語解析の枠組であるが、さらに進んで、論理におけるアブダクション (abduction) などが曖昧さの解消を含めた自然言語の解釈の重要な枠組として位置付けられるようになった。

本解説では、自然言語処理の研究分野におい

て、曖昧さの問題がどのように扱われてきたかについて概観している。まず、構文的、意味的な制約を精密化していくことによって曖昧さを解消しようとする試みについて述べ、次に、曖昧さを選好と呼ばれるヒューリスティックスによって解消する手法について解説し、最後に、情報単調性に基づく、文脈によって変化するパラメトリックなデータ構造を用いて曖昧さを扱う枠組について解説している。

2.2 頑健な自然言語処理へのアプローチ

次の解説では、自然言語処理における頑健性 (robustness) に関連する問題と、それに対する種々のアプローチを紹介している。

現在では、ワープロや機械翻訳などの、自然言語処理技術を応用したシステムが開発され、広く使われるようになってきている。この種のシステムには、形態素解析技術・構文解析技術・意味解析技術などがその要素として利用されている。これらの技術は、基本的に、文法的に適格 (well-formed) な表現を対象として研究開発されてきたものであり、実用規模の辞書・文法も開発されてきている。

これに対して文法的に不適格 (ill-formed) な表現に対する処理 (頑健な処理) については、これまであまり重要視されていなかった。しかし、頑健な処理は、より柔軟な自然言語入力、より広範な機械翻訳システムの実現に有用であるのみならず、well-formed な入力を期待できない自然言語インタフェースなどでは必須の技術である。

近年の要素技術の蓄積により、頑健性への本格的な取り組みが可能になり、また、頑健な処理の重要性が再認識され始めてきた。頑健な自然言語処理のためには、次のような問題を扱う必要がある。(1) ミススペル、未知語などの語彙レベルの問題。(2) 構文レベルの文法上の誤りを含む文の処理と誤りの同定の問題、および、それらの誤りから回復するための技術の開発。(3) 意味を重視した言語理解、および、大規模な知識ベースの構築。(4) 慣用的な表現、語用論に関連する言語現象の問題。(5) 文脈に対する柔軟な処理、および、協調的な応答に関する問題。(6) 比喩や換喩の理解の問題。

本解説では、これらの問題のうち特に(2)と(3)に関する話題について述べている。

* 昨年、日本ソフトウェア科学会が、このテーマでシンポジウムを開催し、学会誌にその特集論文と解説を掲載している。その中で、横田・竹沢両氏の解説が非常によくまとまっている。

2.3 超並列自然言語処理

次の解説では、超並列的な自然言語処理の手法について述べている。

コネクションマシンのような超並列マシンが製品化されるに至って、近年では、超並列的な知能処理が興味を集めるようになってきている。また、自然言語処理においては、言語現象としてユニバーサルな統語論的制約のみを扱う時期は過ぎ、意味論、語用論を含む記憶モデルをも前提とする統合的な認知現象のテストベッドとしてみなされるようになってきている。

また、扱われる制約も、これまでの文脈自由文法に代表されるアプリオリなものに限られていたのが、文脈的制約のようにアポストオリな獲得が威力を発揮しそうな制約を必要とする言語現象も扱われるようになってきた。さらに、Frege/Montague 的な統語論、意味論の係わり合い方で前提とされてきた、全体の意味の構成要素からの単調的な組み上げ、つまりコンポジショナリティも、現在問題になっているような言語現象の処理にとっては十分な仮説になっていないことが指摘されている。

超並列的な認知理論による自然言語処理では、種々の性質の異なる制約が超並列的にダイナミックに変化しながら係わり合っていくことを前提にしており、アプリオリシティ、コンポジショナリティ、言語現象の記憶からの切り離しなどはもとも前提とされていない。

本解説では、このような話題に関してこれまで提唱されてきた超並列的な自然言語処理の手法について紹介するとともに、これまでの手法の問題点とその解決への課題・展望について述べている。

2.4 自然言語生成への招待

次の解説では、自然言語生成の最近の研究成果について述べている。

自然言語理解では、形態素解析・構文解析・意味解析・談話解析の4つの処理を想定するのが一般的であるが、自然言語生成では、(1)what-to-say の決定と、(2)how-to-say の決定の二つに処理を分けることが多い。what-to-say の決定(それを text planning と呼ぶ)では、発話時の状況などを参考に、システムがもつ情報の中から出力すべき情報内容を選択し、その構成に関する

種々の項目について決定する。また、how-to-say の決定 (realization と呼ばれる) では、主に構文情報、語彙情報を参照しながら構文構造や語の選択などに関する決定を行い、出力すべき情報内容を文字列として生成する。

最近の自然言語生成に関する研究では、結束性 (coherence) のある文章をいかにして生成するかという問題が中心的なテーマである。文章の結束性とは直観的に言って、その文章に含まれる各文の間に自然な意味的・概念的つながりが認められることを言う。つまり、結束性は上で述べた what-to-say の決定に関係する。結束性は、単文を越えた文章を生成するときの重要なポイントである。

この解説では、まず、自然言語生成では何が困難であるかを自然言語解析と対比しながら述べ、次に、文章の結束性に関する研究を中心とした自然言語生成に関する話題を解説している。

2.5 自然言語の構文解析と文生成の統合

本特集の最後の解説では、言語表現の理解と産出の統合を目指した、一般的な枠組による、自然言語文の解析と生成の統合について解説している。

認知科学的な立場から、言語の理解と産出が同一の文法、辞書、および処理手続きに基づいているという仮説が立てられている。このような仮説は、認知過程における多様な情報の流れを実現するための一般的な要請から導かれる。理解と産出を統合するためには、情報の流れをできるだけ固定しない処理メカニズムを設計する必要がある。そのようなメカニズムのベースには、制約に基づく言語理論が望ましい。また、これまでに提案されてきたアルゴリズムを自然発生 (emerge) させるような一般的な処理の戦略が必要である。

本解説では、言語理解と産出の統合の基礎となる文法理論について簡単に述べ、特に構文解析と文生成の統合のためのいくつかの処理メカニズムを紹介している。また、言語的制約の宣言的意味論としての言語能力と、処理の制御形態としての言語運用とを統合する方法について論じ、その認知的科学的意義を指摘している。

3. 自然言語処理のその他の話題

ここでは、本特集で特別に取り上げることでできなかった、その他の重要な話題について簡単に述べる。

3.1 談話理解

今回の特集では、自然言語の深い意味処理である談話理解に関する話題については特に触れなかった。文章生成の解説の中で、談話の深い意味関係を表わす修辞構造について述べているが、これは文章生成の前提としての意味構造に関する話題であり、自然言語理解の言うところの談話理解、つまりどのように談話の意味解釈を行うかといった話題とは異なる。これに関しては、本学会誌の以前の特集「自然言語理解」の片桐氏の解説²⁾と最近の話題を比べてみても、さほど大きな進展はみられない。

3.2 対話処理

談話理解をある特定の発話者あるいは執筆者の発話や文章を理解することだと考えると、対話の理解は複数の発話者のコミュニケーションを解釈することだと言える。このとき発話者の意図を聞き手がどのように認識するか、などの問題に焦点が当てられる。これをプラン認識の問題と言う。また、自分の意図を効果的に相手に伝達するために、相手の信念などを考慮して発話の戦略を立てる。これは文章生成とも関連するが、特にプラン生成の問題と呼ばれる。また、発話を物理的な行為と同様に考える発話行為という考えもある。これらの話題も、以前の特集にある飯田・相沢両氏の解説¹⁾から大きく進展しているわけではない。

これは、このような研究が、非常に狭いドメインを設定してしまっているため、分野全体に与えるインパクトが比較的少ないことにも関係があると思われる。問題の領域を広げるためには、談話や対話にとって本質的というより、自然言語文の解析や生成にとって重大な難問に挑まなければならないのである。

ただし、人間とコンピュータの協調的なコミュニケーションを探究する上での、自然言語対話の研究は、今後さらに重要性を増していくであろう。

3.3 音声処理

音声処理と言語処理の統合は、最近非常に盛り上がりを見せている話題である。音声認識のブレークスルーが言語処理との統合によってもたらされるであろうと期待している人は少なくないであろうし、言語処理、特に対話処理に興味のある人が音声による対話を対象としようとすることは

きわめて自然である。また、音声言語が、自然言語処理にとって非常に重要な文脈情報（たとえば、韻律情報など）を合わせもっていることは、今後の重要な研究の広がりを予感させる。しかしながら、今のところ、音声処理との統合による革新的な研究成果は得られていない。

3.4 非字義的意味の処理

比喩や換喩などのいわゆる非字義的 (non-literal) な意味の扱いは、これまであまり大きく取り上げられてはいなかった。これは字義どおりの解釈を行った後で常識的な推論によって対処できるはずだという甘い思惑があったからである。しかし、問題はそれほど単純ではない。もっとも、解釈に対して考えられる常識的な推論がどのようなものであるかを明確にするのは非常に困難であることは容易に推察できることではあるが。

最近の話題で面白い点は、字義的な意味解釈とそれ以外の意味解釈がそれぞれ別々のモジュールで処理されるのではない、ということである。これは、もともと、処理すべき意味的な現象があらかじめ分類されているのではなくて、発話された文脈に応じて解釈が変化していくものであるという立場である。これは、曖昧さの扱いに関しても同様のことが言える。

3.5 大規模な辞書やコーパスの構築

大規模な辞書や文法、知識ベースの構築に関する話題は、以前より増して盛んに取り上げられている。COLING などの国際会議で非常に大きな割合を占める最近のテーマは、確率・統計的手法による大規模コーパス (例文集) の粗い解析 (tagging) とそれを用いた文の解析や、対訳例を用いた翻訳などである。これらは、とにかく大雑把でいいから処理をスケールアップしようとするアプローチである。これに関連して、大規模コーパスを世界規模で収集、解析していこうとするプロジェクトである TEI (Text Encoding Initiative) のような動きもある。これは著作権なども絡んで、なかなか思ったようにはかどらないのが現実のようである。しかし、これには日本もただ乗りするわけにはいかず、全面的に協力する必要がある。なぜなら、大量な言語データに基づく研究は今後の重要な鍵となるからである。また、日本電子化辞書研究所 (EDR) が構築した大規模な辞書がそろそろ利用可能になるかもしれない。このような

リソースは多くの人が使って、こなれていくことに意味があるので、できるだけ早く配布してほしい。大規模な辞書などは、その開発・保守にとつてもない労力を必要とするから、世界規模の共有財産にして、多くの組織の協力の下に維持していくのがいいのであろう。

3.6 応 用

今回の特集では、応用に関する話題が一切ないのであるが、これは、実用に耐え得るだけの新しい応用システムが依然として実現されていないことに起因している。たとえば、現在、機械翻訳システムが十分に実用化していると考えている人は少ないであろうし、ワープロに多少の機能を付加したに過ぎないもの以上の文書処理システムが存在しているわけではない。対話型機械翻訳や文書校正支援のシステムが注目されつつあるが、これらは今後に期待したいところであろう*。

4. おわりに

以上、多少独断的に自然言語処理の最近の動向について述べた。特筆すべきことは、本特集の執筆者を含めた日本人の若手研究者が非常に良い仕事をして、活躍している点である。人工知能のような新しい研究分野における日本人のオリジナリティは低いと思われがちであるが、少なくとも最近の自然言語処理の研究の流れにおいては、そうではない。いくつかの先駆的な研究が日本人研究者によってなされているのである。

先に述べたように、本特集は自然言語処理のアドバンスト・コースとして企画したものである。これから研究を始めようとする人たちにとっては、それほど有益ではないかもしれないが、あ

る程度の知識をもっている人たちならば、最近の研究動向を理解して、今後の方向性を読みとってくれるであろうことを期待したい。そして、一刻も早く第一線の研究者と肩を並べて欲しい。研究は人によって生み出され、人と人とのインタラクションによって大きくなっていくものなのであるから。本特集が、自然言語の研究者にとっての一助になってくれれば、企画したものとしては、うれしいかぎりである。

最後に、大変お忙しい中、筆者からの執筆依頼を快く引き受けてくださった解説者のみなさまと、本特集に先だって行われたソニー CSL におけるワークショップに出席してくださった方々に感謝いたします。

参 考 文 献

- 1) 飯田 仁, 相沢輝昭: 意図の理解, 情報処理, Vol. 30, No. 10, pp. 1216-1223 (1989)
- 2) 片桐恭弘: 文脈理解のモデル, 情報処理, Vol. 30, No. 10, pp. 1199-1206 (1989)
- 3) 橋田浩一, 竹沢寿幸: 自然言語処理における統合の諸相, コンピュータソフトウェア, Vol. 8, No. 6, pp. 511-524 (1991)

(平成4年4月14日受付)



長尾 隆 (正会員)

1962年生。1985年東京工業大学工学部卒業。1987年同大学総合理工学研究科修士課程修了。1987年4月より1991年3月まで、日本アイ・ビー・エム(株)東京基礎研究所において、自然言語処理、機械翻訳の研究に従事。1991年4月より、(株)ソニーコンピュータサイエンス研究所において、自然言語対話システムの研究に従事。訳書に「決定不能の論理パズル—ゲーデルの定理と様相論理」(R. スマリヤン著, 白揚社, 1990年)がある。日本ソフトウェア科学会, ACL 各会員。

* この次に本学会誌で自然言語処理の特集をやるときは、もしかしらば応用が中心の話題になるかもしれない。そのときには、自然言語処理が実用的には役に立たないと思っている人たちの、うならせるだけの製品が現われていることを期待したい。