自然言語処理技術と情報検索*

田中穂積
東京工業大学大学院情報理工学研究科

要約: ネットワーク技術の進展と共に，ネットワーク上に存在する情報を自在に検索する技術が重要になってきている．ネットワーク上に存在する情報の中で特に重要なものは，自然言語で書かれた文書である．自然言語をコンピュータで扱うための自然言語処理技術は，長期を要する技術であるため，これまで，情報検索の分野では未熟な技術として考えられ，自然言語処理技術をむしろ避けた情報検索システムが構築されている．しかし近年の自然言語処理技術の研究の進展により，情報検索に利用可能な技術も生まれてきている．またネットワーク上の情報検索では，専門家ではなく一般のユーザがこれを利用する機会も増えると思われる．自然言語を用いたユーザフレンドリーなインタフェースを持ち知的な検索を行なってくれる情報検索システムの構築が課題になっている．そこで，自然言語処理技術の現状を概観し，問題点を述べ，情報検索に利用可能な技術として現在どのようなものがあるかを述べる．

1　はじめに

大量のコンピュータパワー，大量の知識，そして高速の情報ネットワークをインテグレートした知的なシステム，音声，映像などさまざまな情報メディアを自在に使いこなすことができるシステム，これが将来の情報処理システムのイメージであろう．高速の情報ネットワークが各家庭にまで張りめぐらされ，家庭にいながらにして，世界に分散しているさまざまな情報にアクセス可能な時代が到来しようとしている．各家庭にまで張り巡らされた高速の情報ネットワークインフラの整備とともに，ネットワーク上に存在する情報を自在に検索する技術の重要性が増しているのはそのためである．

大量の電子化された情報の中で特に重要なものは，自然言語で書かれた文書である．情報検索に関連して，自然言語処理技術がこれから重要な役割を果たす見られているのはそのためである．ところが，情報検索に自然言語処理技術を応用することには問題があるとして，これまでむしろ避けられてきたと言って良い．特に日本語には，語と語の間に区切りがないため，構文解析以前の問題として形態素解析が大きな障害になっていた．形態素解析が不十分であるとして，自然言語処理技術を，日本語文書の検索に応用することが差し控えられていたのである．そのため，文字列単位の全文検索といったいわば力ずくの手法に頼ることが情報検索の分野でこれまでよく行なわれてきた．

全文検索もそれなりの工夫がこらされ，文書が頻繁に更新されたり追加されることがなければ，確かに確実で有効な方法である．しかし，日々更新され追加される大量の文書の場合には，全文検索に頼る情報検索には限界がある．全文検索用として，あらかじめ用意するテーブルの使用記憶量が，検索すべき文書の記憶量を越えるといった現象も現実に生じている．検索すべき文書の量が少なければ，これはさほど大きな問題にならないかも知れないが，検索すべき文書の量が膨大になると問題になる．そこで，最近もう

*Natural Language Processing and Information Retrieval
Hozumi Tanaka
Tokyo Institute of Technology

一度自然言語処理技術を見直し，それを応用した高度で自然な情報検索システムを開発しようとする動きがある．

一方自然言語処理技術に関していえば，解決が困難な問題が山積しているとはいえ，着実に進歩してきている．これまで困難であるとされていた問題も解決されるようになってきた．自然言語処理技術をここで見直しておくことは，情報検索の技術者にとっても有意義であると筆者は考えている．情報検索に関連して，最近の自然言語処理技術の中で次の4つが注目される．

1) 新しい高速な自然言語処理アルゴリズムが開発されている，

2) 大量の電子化された文書 (コーパス) が利用可能になり，文書に含まれる統計情報を用いた自然言語処理技術の進展がある，

3) 自然言語処理で用いる辞書として大規模なものが入手可能になってきている，

4) 大規模な知識が開発されている．

1) に関しては，高速で高精度の日本語の形態素解析システムが開発され，それを誰でも容易に入手し，日本語文書の語切りを容易に行なうことが可能になったこと，また日本語の文法にしても，広範囲の日本語の文を解析可能な日本語文法や，構文解析システムが開発されていることをあげることができる．これには，3) で述べた大規模な電子化辞書 (たとえば日本電子化辞書研究所で開発したEDR辞書など) が利用可能なことも関係する．2) は大量のコーパスから抽出した統計情報を利用して，自然言語の解析結果に含まれる曖昧性を解消する手法として様々なものが開発されている．この技術は，少なくともわが国では現在もっともホットな研究課題である．4) は意味のレベルにまで踏み込んだ自然言語処理を行なうための基本的な知識であり，これまでとかく避けてきた意味理解に一歩踏み込んだ次世代の情報検索を可能にするものであるので今後重要になると思われる．以下では，これからの情報検索で重要になると思われる自然言語処理技術の現状と課題を述べる．

2 自然言語処理の概要

自然言語処理は通常次のようにして進む．

1) 文中の形態素を切り出し，切りだした形態素から単語を知りその辞書引きを行い，品詞を見て，文中の品詞がどのような並びをしているかを文法を用いて調べる．文中の形態素を認定する仕事を形態素解析とよぶ[1]．

2) 品詞の並び，あるいは形態素の文法カテゴリーの並びが文法にかなっているかどうかを調べる．これを構文解析とよぶ．

3) 同音異義語の曖昧性解消や，係り受けに関する曖昧性解消を行う．これらを意味解析とよぶが，意味解析にあたり前後の文脈を参照することも必要になる．文脈を参照する仕事には省略語の補強，代名詞の参照先の決定などが含まれるが，これらを文脈解析とよぶ．

4) 意味解析の結果を受け，文章を生成する．

上記した自然言語処理の順序は，多くの自然言語処理システムで取られている順序であるが，我々人間は，1)，2)，3) を同時に行なっているように思われる．このような自然言語処理のモデルをいかに構築するかを巡り，認知科学的な立場からの研究もある．

[1] 言語学では「文を構成する最小の言語単位」を形態素とよぶ．多くの場合単語と一致するが，活用する語「聞き」は，「聞」と「き」という形態素に分解される．前者の形態素の文法カテゴリーは，「か行五段活用動詞の語幹」，後者は「か行五段活用動詞の連用形語尾」である．これから「聞く」という単語を知り，辞書びきを行なう．

3　形態素解析

英語などの言語では単語と単語の間に空白があるので，形態素解析は大きな問題にならない．ところが日本語，韓国語，タイ語などには単語と単語の間に空白がなく，形態素解析は大きな問題になる．現在の情報検索は，文書中に含まれる単語をベースにして行われるので，この種の言語の文書を形態素解析し，文書中に含まれる単語を認定しておくことは，情報検索のための第一歩であるので重要である．

幸いなことに，日本語については，最近形態素解析のための高速で優れた方法，システムが開発されている．いずれも表層レベル情報を用いたものであり，深い意味理解に基づくものではない．文書が対象とする分野（たとえばニュースや科学技術分野など）を限定し，それに合わせた単語辞書を用いれば，99%以上の精度で形態素解析を行うシステムも現れている．しかし，システムにとって単語辞書の整備が十分でない新しい分野の場合には，90%を少し越える程度の形態素解析の精度しか得られないことも知られている．

松本らの開発したJUMANと呼ばれるシステムが良く使われている．これはフリーのソフトウェアシステムであり，多数の自然言語処理の研究者や応用システムの開発者によって使いこまれている．分野さえ限定し，分野に適した単語辞書を付加し整備しさえすれば，これまでとは比較にならない形態素解析の精度を得ることができるので，情報検索の技術者も，こうした既存の形態素解析システムを一度利用してみられるとよい．

形態素解析システムの多くは，隣接する形態素と形態素との間の接続可能性に関する制約情報を用いて形態素の切り出しを行う．接続が不可能な隣接する形態素は，たとえ辞書にこれらの形態素が登録されていたとしても，妥当なものではないので切り出されることなく排除される．たとえば動詞の語幹「聞」に，動詞の語幹「来」が接続することはないと接続表に書いてあれば，「聞きます」は，「聞」「き」「ます」のような形態素の分割は許されるが，「聞」「き（来）」「ます」のような分割は許されない．接続可能性の制約は表の形にして持っていたり，例外的な接続可能性を精密に検査するために，検査手続きを起動可能なようになっていることが多い．

一般に，形態素と形態素の間に空白のない言語の文の形態素解析を行うと，非常に多数の形態素解析結果（形態素の系列）が得られる．文の長さにもよるが，解析結果の数が天文学的な数字にのぼることもまれではない．そこで，どの形態素解析結果がもっともらしいかを決めるために，形態素の系列の統計的なもっともらしさを計算する必要がある．この計算を行うために，形態素間のバイグラムやトライグラムの確率を用いることがよく行われる．

こうした形態素解析システムでは形態素の認定を辞書を用いて行う．そのため，辞書に登録されていない語が文書に含まれる場合が問題になる．これを未定義語の問題とよんでいる．人名や地名の数は比較的固定しているため，そのほとんどすべてをあらかじめ辞書に登録しておくことも，労力とお金の問題を厭わなければ，それなりに可能になってきた．しかし，日々新造語が生まれており，これにどう対処したら良いかが問題である．複数の名詞が並んだ複合語の形態素解析も問題になる．たとえば「歩行者通路」は「歩行」「者」「通路」とわかち書きすることは妥当であろうが，「歩」「行者」「通路」とわかち書きするのは妥当ではない．こうした問題を避けるために，「歩行者通路」を一つの形態素として辞書にそのまま登録することがよく行われている．漢語の造語能力は無限にあるので，こうしたその場しのぎの方法に限界があるのは明らかであろう．

4　構文解析

構文解析技術は1980年代に入り特に進歩が著しい．数十語からなる文の構文解析に要する時間が，1秒以内と高速な解析アルゴリズムが開発されている．良く使われる構文解析アルゴリズムは，文法は一般の文脈自由文法を用いる．チャート法，一般化LR(GLR)法はいずれも文脈自由文法をベースにしている．前

者は，解析すべき文の長さをnとしたとき，nの3乗のオーダーの解析時間を要す標準的なアルゴリズムである．一方GLR法は，文脈自由文法規則の右辺の長さ(記号数)の最大をmとしたとき，最悪の場合，nのm+1乗のオーダの解析時間を要す．しかし経験的にGLR法はチャート法より解析時間が少ないことが知られている．

一般に，解析すべき文の長さが長くなればなるほど，構文解析結果の数も天文学的な数字になる．形態素解析と同様に，妥当な解析結果を得るために，解析結果に統計的なスコアを与える方法がいくつか提案されている．

一つは確率文脈自由文法を用いる方法である．各規則に確率をあらかじめ割り振っておき，それを構文解析結果のスコアの計算に利用するものである．この方法によれば，構造が異なる構文解析結果に等しいスコアを付与することがあり問題がある．

この問題を解決するために，文脈に依存した確率を計算する方法が開発されている．これに関連して，GLR法で用いるLR表中の各アクションに確率を付与するBriscoeらの方法が注目される．GLR法で用いるパーザの状態が左文脈を，先読み語が右文脈を与えると考えるのである．残念ながらBriscoeらの与えた方法には理論的な欠陥がある．筆者らはそれを解決した新しい確率LR法を開発しており，実験を準備中である．

構文解析で問題となるのは，解析のアルゴリズムではなく，むしろ構文解析に用いる文法規則の開発にある．言語学者は必ずしもこうした文法規則を開発することに興味を示さない．自然言語処理の研究者が文法を開発しなくてはならないのである．現実の文書に現れる文の解析が可能な文法を開発することは一般に容易ではない．そこで，文法規則をコーパスから自動的に獲得する研究が行なわれている．

5　意味・文脈解析

意味解析・文脈解析には知識が必要になる．特に概念間の関係については，「人間は液体を飲む」という一般化した常識を用いて，「花子は酒を飲む」，「太郎は水を飲む」など，人間や液体の下位概念を用いた無数の文の意味的妥当性を計算したり意味的曖昧性の解消をはかることができるので重要である．曖昧性の解消も，「人間は液体を飲む」という知識から，「太郎はタバコを飲む」という文の「飲む」は，「液体を飲む」という意味の「飲む」ではないことが分かる．

最近8万から数十万の数の概念間の関係を体系化した知識として，国立国語研究所の分類語彙表，電子化辞書研究所の概念辞書，プリンストン大学の開発したWordNetが利用可能になってきた．これらは意味解析だけでなく，類義語を利用した情報検索にも役立つ知識である．情報検索の研究者も，これらの知識を組み込んだ知的検索システムの構築を試みても良い時期にきているといえる．少し考えてみれば分かることだが，それでも計算機が持っている知識の量・質とも十分でない．知識をどう表現するかという技術も十分でない．計算機による文書の深い意味理解をめざすためには解決すべき沢山の問題がある．

統計的な知識を用いて曖昧性を解消する研究も盛んに行われている．たとえば“crane”という単語には「起重機」という意味と「鶴」という意味がある．それぞれの意味毎に，“crane”という語を含む文を大量に集めておき（これをコーパスとよぶ），“crane”という語の前後にどのような語が現れやすいかに関する統計データをとり，この統計データを用いて多義語の曖昧性を解消するのである．

大量のコーパスを集め，そこから抽出した統計的な知識を利用した自然言語処理技術は，これ以外にもさまざまな技法が提案されている．この技法は，コーパスを集めさえすればよいので，大規模な知識ベースを設計する必要がない．自然言語の深い意味解析の安定した技術が開発されるまでのつなぎとしてコーパスベースによる自然言語処理は，速効性があり有効な方法であるといえる．ただ，大量のコーパスの収集がわが国では個別に行われているのが問題で，こうした現状を今後改める必要がある．

6 対話理解

高速な情報ネットワークが各家庭にまで入り込み，家庭に居ながらにして，ネットワーク上のさまざまな情報を検索することが可能になるためには，さらに高度で自然な，人間と機械との間のインタフェース機能が必要になると思われる．インタフェースとしてさまざまなものが考えられるが，もっとも自然で望ましい対話は，我々が日頃使用している自然言語を用いた対話であろう．これには音声による対話も含まれている．検索結果は情報検索者の意図するものでなければ意味がないので，情報検索者は，自然言語を用いて対話を行い，検索の意図をシステムに伝えることになる．このような場面では，システム側に対話の意味理解が求められることになる．特に，家庭にいる一般のユーザが気軽にネットワーク上の情報を検索するためには，自然言語を中心にしたより高度で自然なインタフェース機能の充実が望まれる．端末からの対話が不得手のユーザには音声による対話機能の実現が望まれている．しかし自然言語や音声による対話機能は問題が多い．今後さらなる研究が求められている．

7 機械翻訳

高速情報ネットワークが世界的な規模に拡大するにつれて，ネットワーク上には世界各国の情報，知識が分散して存在することになる．これらが自然言語で書かれた文書である場合，そのほとんどは，その国の言葉で記述されていると仮定してよいだろう．各国の言語で記述された文書や，話し言葉を含む映像などにアクセスするためには，翻訳技術が必要になる．世界的に分散したこれらの知識や情報の検索を可能にし，情報のグローバリゼーションを達成するためには，翻訳技術がきわめて重要な役割を果たす．これは機械翻訳技術に頼ることになるのが，これも自然言語処理技術の一分野であることは，明らかだろう．

機械翻訳の方式は，トランスファー方式と中間言語方式の二つに大別することができる．中間言語方式は，ソース言語とターゲット言語の双方に共通な，中間言語とよばれるレベルを設定する（日英翻訳の場合には，日本語がソース言語で英語がターゲット言語になる），このレベルはソース文の意味を理解したレベルであると考えてもよい．多言語間翻訳をめざす場合には，中間言語は言語によらない普遍的なレベルとされているので，この構造からさまざまなターゲット文を生成することができる．こうしてソース言語とターゲット言語の対を陽に考慮することなく翻訳システムを構築することができるので，多言語間翻訳に都合の良い方式であるとされている．

中間言語方式の問題は，中間言語の設計が難しいこと，たとえそれが可能であるにしても，ソース文を解析して中間言語のレベルの構造を抽出する技術が未熟であることである．中間言語方式は，意味理解に深く踏み込んだ理想に近い方式であり，CICC，EC，カーネギーメロン大学での試みの他に，NEC や富士通の試みもあるが，なお一層の研究が必要である．

中間言語方式ではなく現実的な方式としてトランスファー方式がある．この方式は，ソース言語の解析結果を，ソース言語に近い表現にとどめ，それをターゲット言語に近い表現にトランスファー（変換；移行）する．そしてそこから最終的なターゲット言語の文を生成する．この方式はソース言語とターゲット言語の対を考慮するレベルが設定されていることが特徴である．解析レベルをもっとも深いレベルに設定すれば中間言語方式に近くなり，解析を行わずに単語単位の翻訳を行うだけにすれば直接翻訳方式になる．

現在では直接翻訳方式の機械翻訳システムは皆無といってよいから，中間言語方式とトランスファー方式とを比較することになる．トランスファー方式の擁護者は，人間の翻訳ではソースとターゲット言語の対を絶えず考えながら翻訳していると主張する．理想的な意味での中間言語方式はまだ確立されていないので，商品化された機械翻訳システムは，解析レベルの差はあるとはいえ，トランスファー方式を採用しているといってよいだろう．ただし中間言語方式を標ぼうするシステムでは，トランスファー方式と比べて，より深い意味解析を行おうとする姿勢が強い．

8 要約と文章生成

検索結果の量が大量なら，検索結果を要約して示す要約システムも必要になるだろう．このことは，自然言語処理技術の中で比較的未熟であった意味処理技術の研究がこれからますます重要になることを意味している．要約結果は，複数の文の系列として出力する必要があるため，一文ではなく，複数の文の系列からなる文章の生成技術もこれから重要になってくるだろう．省略を含む文や適当な代名詞化をおこない，こなれた訳文を生成する技術の確立は今後の課題である．

9 おわりに

情報検索に関連させながら，自然言語処理技術の現状と問題を駆け足で見てきた．本文中でも指摘したように，自然言語処理を行うために開発された知識・技術のうち，形態素解析技術と概念の体系は，情報検索技術者が直接利用可能な段階に到達していると考えられるという指摘をした．自然言語処理技術の本質は意味解析・文脈解析にあるが，この技術はまだ未熟であることも指摘した．本稿により，情報検索の研究者・技術者が最近の自然言語処理技術に興味をもち，それを応用した新しい世代の情報検索システムの構築を試みることがあれば幸いである．

最近の自然言語処理技術を理解する早道は，文献を購読するより，実際の解析システムを動作させてみることだろう．松本らの開発した形態素解析システム JUMAN はフリーの辞書付きソフトウェアである．これに興味ある読者は http://cactus.aist-nara.ac.jp/lab/nlt/NLT.html を参照のこと．GLR 法は，文脈自由文法規則から LR 表を作成するところに第一のハードルがある．幸いにして筆者らが開発したシステムがフリーのソフトウェアとして，ICOT の後身である AITEC で公開されている．これは MSLR とよばれているが，AITEC のホームページ http://www.icot.or.jp を参照してアクセスされるとよい．機械翻訳システムとしては，安価なソフト（1万円以下）が多数販売されている．これらを購入して，機械翻訳システムの現在の実力を試されると良い．購入にあたり，アジア太平洋機械翻訳協会に問い合わせると良い (http://www.jeida.or.jp/aamt/)．対話理解に関連して音声認識の機能についてもふれておいた．これについても，最近安価な音声認識システムが発売されているので，購入されて性能を評価されることをお勧めする．