特別論説

情報処理最前線

情報ハイウェイ時代のテキスト情報への知的アクセス†

野村浩郷†† 井佐原均†††
徳永健伸†††† 中村貞吾††

1. はじめに

インターネットということばで総称的に呼ばれているコンピュータネットワークの利用の普及が最近極めて急速に展開され，パソコン通信の一般化も手伝って，その利用者や利用場所は職場のみならず家庭にまで広がってきた．

インターネットへアクセスするためのインタフェースソフトウェアは，パソコンも含めていずれのコンピュータにも用意されるようになり，かつコンピュータの使用法に精通していない人でも容易に使えるようになっているため，今や，インターネットはだれでも使える状況になってきている．

現在のインターネットでは，コンピュータ，ネットワーク，および通信回線の容量・速度・料金の各種制約などから，実際には，インターネットが本来提供し得るサービス機能のほんの一部しか活用されていないといってよい．しかし，ハードウェア技術およびソフトウェア技術の進歩は驚くほどに急速であるので事態は早期に改善され，その結果としてインターネットの利用形態が一層発展し，かつ利用者数，利用対象，利用回数，および利用時間が激増するものと思われる．

さらに，現在，全世界的に開発が取り組まれている情報ハイウェイが21世紀はじめに実現されると，すべての職場および家庭が高速なネットワークで結ばれることになり，情報ハイウェイを通じてのコンピュータネットワークへのアクセスは生活のかなりの部分を占めることになる．

このような状況下では，コンピュータネットワークへの情報アクセスを知的に支援する知的情報アクセスの機能が不可欠となる．それなくしては，情報の洪水に流されたり埋もれたりしてしまうことになり，その結果として情報パニックや情報阻害に陥ることになる．そして，日常生活に支障をきたすことになる．

知的情報アクセスは，情報の意味を理解する機能を核とする技術により実現される．そのような技術を開発するためには，たとえば，現在の自然言語処理を高度な意味理解の機能を持つものに飛躍させなければならない．そのためには，高度な自然言語処理システムを設計するときの基礎となる言語データを整備する必要がある．一方，高度な自然言語処理を手助けするためには，ネットワーク上の情報の記述の仕方を工夫する必要もある．これらのための1つの方策として，文章に自然言語処理のための言語情報を付加することが考えられる．言語情報を付加した文章は，タグドテキストコーパスと呼ばれる．

本稿では，情報ハイウェイへの知的な情報アクセスを実現するための数々の課題の中から，テキストを対象とする自然言語処理に焦点を当てて，情報検索・情報抽出を中心とする「知的情報アクセス」，およびそれを実現しかつ将来の高度な自然言語処理の基礎データともなる「タグドテキストコーパス」について述べ，それらに関連するいくつかの問題にも触れる．

2. 知能ネット

知的情報アクセスの機能を備えたコンピュータネットワークは，IntelliNet（知能ネット）[1),2)]と呼ばれる．これは，Intelligent Network から作られた造語である．

† Intelligent Information Access to Text on Information Highway by Hirosato NOMURA (Kyushu Institute of Technology), Hitoshi ISAHARA (Communication Research Laboratories), Tatenobu TOKUNAGA (Tokyo Institute of Technology) and Teigo NAKAMURA (Kyushu Institute of Technology).

†† 九州工業大学情報工学部
††† 郵政省通信総合研究所関西先端研究センター
†††† 東京工業大学工学部

IntelliNetは，さらに，人にやさしい対話インタフェースをも提供する．テキスト，音，および画像を3要素とするマルチメディアにより対話し，人とコンピュータとの共同作業によって情報アクセスの目的を効率的に達成する．

IntelliNetは，知識としての情報を提供するという意味でも知的である．したがって，IntelliNetは，知的インタフェースと，知的アクセス手段と，そして知識の実体とからなる統合的な知能・知識システムである．

情報ハイウェイは，まさしくIntelliNetでなければならない．IntelliNetが持つ知的情報アクセスの機能の最も重要な核の1つは，情報の意味を理解する機能である．これは，ネットワーク上の情報に対しても，インタフェース上での対話に対しても必要な機能である．さらに，ネットワークアクセスのそれぞれの目的に応じて，それらの情報を整理・加工する機能も持つ．そして，必要ならば関連情報を電子図書館[3]などから探索して入手する機能を持ち，また，アクセスした情報から新しい重要な情報を発見する機能も持つ．

IntelliNetは情報の意味を理解する機能を核とするが，マルチメディアの3要素であるテキストと音と画像とについては，従来から個別にそれぞれの意味理解の研究が進められてきた．特に，機械翻訳などを主たる応用とする自然言語処理では意味理解は最も重要な研究課題の1つであり，長年に渡って精力的な研究が推進されてきた．しかし，残念ながら，現在の意味理解の技術は，IntelliNetを実現するのに必要な技術レベルには程遠い状況にあるといわざるを得ない．

知的情報アクセスの当面の重要な技術課題の1つは，テキストを対象とした情報検索，情報抽出，情報発見，情報加工，および情報整理などを要素とする知的情報アクセスの統合技術の開発である．ここでいうテキストとは，電子メール，ニュース，通知，各種資料などの言語表現されたものを指す．したがって，それらの要素技術は，主としていわゆる自然言語処理に属する技術である．

現在までの自然言語処理の研究・開発は日本語ワードプロセッサや機械翻訳システムなどの実用化で大きな成果をあげてきたが，将来のIntelliNetにむけての知的情報アクセスの技術を開発するには，基本に立ち返って，基礎となる言語データを構築・整備し，いくつかの核技術とそれらによる要素技術，および目的に応じたそれらの統合技術を開発する必要がある．

3. 知的情報アクセス

現在のインターネットや将来の情報ハイウェイは，電子メールなどのコミュニカティブ情報，電子図書館などに蓄積されるストアド情報，および広報やニュースなどのデリバリ情報を提供する．それぞれへのアクセスの仕方は若干異なるが，知的情報アクセスという観点からは共通するところも多い．

3.1 情報整理

現在のインターネットにおいてさえも，あまりにも多くの電子メールが届くため，それらに十分には対応できないことはしばしばである．また，業務に必要な事項について情報検索を行うと情報が全世界から大量に収集されるため，混乱を起こすこともしばしばである．さらに，インターネットにはだれでも自由にアクセスできるFTPやWWW (World Wide Web)のサイトがたくさんあり，そこからいわゆるフリーソフトウェアやフリーデータが自由に入手できるため，入手した情報の活用準備に莫大な時間をとられることもしばしばである．したがって，大量に到来する情報を効率よくファイリングする手段が不可欠である．このファイリングが情報整理の最初のステップである．

電子メールをファイリングするためには，まず，個々の電子メールにインデクスを付ける必要がある．これにより，インデクスに基づくファイリングが可能となる．現在の電子メールに対しては，その発信者名，受信日付，および電子メールに記載されているSubjectの内容を活用することができる．

ファイリングは，電子メールだけではなく，コンピュータネットワークにより提供される種々の情報に対して必要である．そのため，インターネット上で情報を探索し収集するYahoo，WWW Robot，およびSoftware Agent[4]と呼ばれるソフトウェアシステムなどのようなものに簡易な自然言語処理の機能を入れて実現しようとする試みも出はじめている．

この程度のファイリングでも多いに役に立つが，知的情報アクセスという観点からすると，情報の内容に立ち入ったインデクシングが必要となる．そのためには，情報の内容を理解する機能が必要となる．これは，まさしく，自然言語処理の意味理解の問題であるが，現在の自然言語処理の技術では実現は難しい．そのため，表層的な言語表現の形を手掛かりとして，内容からのインデクシングを行うことになる．

表層的な言語表現の形を手掛かりとする方法は，いわゆるキーワード検出による方法である．このとき，文書が定型的な構成を持っている場合には，それも活用できる．

3.2 情報検索

情報検索とは，知りたい情報をストアド情報から見つけ出し，入手することである．情報検索の例としては，文献，特許，ニュースなどに対する検索がある．

文献検索や特許検索は，たとえば，キーワード付きの文献リストや特許リストの中から，あらかじめ設定されているキーワードの組合せに適合するものを見つけ出し，その一覧を作成する．ニュースの検索では，たとえば，そのタイトルに対して，検索を行う．このような方式による情報検索は従来から多く使われている．

これらに対して，テキスト検索と呼ばれているものは，テキストの内容に対して検索を行い，要求条件を満たすテキストを選び出す．最も単純なテキスト検索は，テキストの中からあらかじめ設定されている文字系列と同じあるいは類似の文字系列を発見できたときにそのテキストを検索結果として選び出す．文献検索，特許検索，およびニュースの検索などのいずれにも応用できる．たとえば，新聞記事のテキスト検索システムはすでに開発されている．

より高度なテキスト検索については，たとえば，米国の TREC（Text Retrieval Conference）[5] が研究のコンペティションを実施してきた．これは，各研究で作成したテキスト検索システムの能力を，共通のテキストを材料として競うものである．ここに参加したシステムが採用した検索方式は，テキストの中での語の出現傾向モデルによるもの，確率・統計モデルによるもの，シソーラスの情報を活用するもの，および文の言語モデルを使う解析によるものなどがある．

将来の情報ハイウェイでは，もっと多様で柔軟な知的情報検索が要求される．これは，知的情報アクセスの第1の側面であり，知りたいことを知りたいときに平易な手段で検索でき，その結果として知りたいことのみを適切な形で入手できるようにすることである．そのためには，情報の内容に立ち入って検索ができなければならない．これには，意味理解の機能が必要である．

知的情報検索では，検索システムに知りたいことの内容を適切に伝えられなければならない．ところが，知りたいことの内容が明確に定義できていないことも多い．そのような場合には，インタフェースシステムを通して検索システムと対話することにより，知りたいことを明確化することになる．すなわち，そのような対話機能を持つナビゲーション機能が必要である．ここでも意味理解の機能が必要である．

情報ハイウェイのストアド情報は，物理的には全世界に分散されて格納される．知的情報検索システムは，検索対象の範囲が指定されている場合にはその範囲で，そうでない場合にはそれらの全てのストアド情報にアクセスしなければならない．

3.3 情報抽出

情報抽出とは，デリバリ情報あるいはストアド情報から知りたい情報を見つけ出すことである．たとえば，ニュースから特定の話題の特定の項目に関する情報を収集する．

情報抽出に関しては，米国の DARPA の研究プロジェクトである TIPSTER[6] の一貫として MUC（Message Understanding Conference）[7] が研究のコンペティションを実施してきた．これは，共通のテキストを材料として，各研究で作成した情報抽出システムの能力を競うものである．

ここに参加したシステムが採用した処理方式は，主としてテンプレート方式である．たとえば，企業間の国際的な提携に関するニュース記事に対しては，それぞれの企業名，所属国名，企業提携の内容などをスロットとする文のテンプレートをいくつも用意しておく．そして，自然言語処理のいわゆる構文解析や意味解析の簡易な処理を経て，その結果に基づきニュース記事の文とそれらのテンプレートとのパターンマッチングを行う．

その結果として得られたそれぞれのスロットの内容を抽出できた情報とする．

企業名や国名などの言語表現には正式名称の他に通称や略称が使われることも多く，またそれらの出現個所や出現順序にバリエーションもあるので，通称や略称の辞書を用意したり，テンプレートのパターンマッチング法を柔軟にするなどいろいろな工夫が必要である．

1993年に行われた第5回目のConferenceでは，企業間の国際的な提携に関する短い新聞記事が共通の材料として使われた．情報抽出されるべき項目とその内容として人手によりあらかじめ作成されていた「正解」と比較して，最も成績がよかったシステムは約60%の精度で情報抽出できたと報告されている．

情報ハイウェイでは，大量のデリバリ情報が提供されることになるので，知的情報抽出が重要となる．これは，知的情報アクセスの第二の側面であり，必要な情報をもらさずキャッチするためには，常に情報抽出を行っていなければならない．そのとき，上記のようなテンプレートの簡易なパターンマッチングの方法だけでは明らかに限界がある．したがって，より一層の高度な意味処理の開発が必要であり，それを使う知的情報抽出の技術の開発が必要である．

3.4 情報加工

情報検索や情報抽出により入手した情報は，情報の単なる羅列であり，そのままの形では役にたたない．これを目的にそって役に立つ形に整形するのが，情報加工である．

情報加工は，入手した情報の全体を見渡してそれをレポートにまとめること，その要約を作成すること，ファイリングすること，その内容から情報発見をすること，そして必要ならば電子図書館などにアクセスして情報の関連付けや補足をすることなどを行わなければならない．さらに，入手した情報には多くの重複がある場合があり，重複の除去も行わなければならない．

文書の要約作成システムの研究は従来からいくつかある．キーワードを含む文を残して他の文を削除したり，文の要素の中で主要な要素を助詞の種類により判別しそれにより文を短くしたり，文と文との連節の仕方を言語表現上の特性から判定し主となる文のみを残すなどの方法がある．

要約を作成するにも，情報の意味理解の機能が不可欠である．また，目的にそった要約を作成するということも必要である．

情報加工の他の項目についての研究は少ない．たとえば，情報発見についての試みはほとんどないといってよい．情報アクセスの最終段階の処理として情報加工は欠かすことのできない処理であるので，今後研究を推進しなければならない．

4. タグドテキストコーパス

タグドテキストコーパスとは，情報を付加した文章の集まりである．付加する情報のことをタグと呼び，文章のことをテキストと呼ぶ．そして，それらを集めたものをコーパスと呼ぶ．

4.1 タグの種類

タグの種類は，大別して，文書構成に関するもの，文章構成に関するもの，および文字属性に関するものがある．タグの種類は，メディアや言語によっても異なる．ここでは，このような違いについては言及しないことにする．

文書構成に関するタグとしては，マークアップ言語とも呼ばれているものがある．その例として，SGML（Standard Generalized Markup Language）[8]やHTML（Hypertext Markup Language）[9]がある．SGMLは，1986年にISO規格として制定されており，欧米ではかなり使われている．日本においても，JIS規格となっている[10]．HTMLは，インターネット上のWWWサーバ用の文書作成に用いられている．さらに，UNIX上のいわゆるTexも文書構成に関するタグ付けの例ともみなされる．Texは，大学や研究機関などでの論文作成などによく使われている．

マークアップ言語は，文書の表題，著者名，および章節構成とそれらの表題などを規定する．また，文章と図表との関係を規定し，かつ文書の表示・印刷イメージのレイアウトなどをも規定する．これらは世界共通のものであるが，通常使われているワープロはそれぞれ独自の内部的な方式により，同様な機能を実現している．

文章構成に関するタグは文法的なタグであり，文やその要素に対する文法的な情報を与える．現在はまだ文書作成には使われていない．将来的にも，必ずしも文書作成者が文書作成時に与えるものではない．デリバリ情報として発信するときや

ストアド情報として格納するときに，コンピュータネットワークが与えるものである．

文章構成に関するタグは，語の表記，語の品詞，語の活用，複合語の内部構成，構文構成，意味構成，文脈構成，およびそれらに対する意味的な属性などである．すなわち，通常の自然言語処理で使われているあるいは使われようとしている種類のものと同じである．

このような文法的なタグは，自然言語処理の結果として生成されるものである．現在の自然言語処理技術では，構文解析あるいは極めて初歩的な意味解析のレベルまでしかできない．また，処理結果に曖昧さが多いため，人手による最終仕上げの処理が必要である．

文字属性に関するタグは，ワープロで通常使われているような文字のフォント，サイズ，スタイル，および言語の種別などに関するものである．ワープロではディスプレイ上での操作により指定し，ワープロの内部にそれらの指定が記述される．Texでは，テキストの中に制御文としてテキストと同じレベルで記述する．

文字属性に関するタグの情報は，現在の自然言語処理ではほとんど使われていない．しかし，将来の知的情報アクセスにおいては，この情報は処理上の大きな手かがりを与える可能性がある．

4.2 知的情報アクセスでの活用

情報検索や情報抽出では，処理対象であるテキストに豊富なタグが付いていると精度の高い処理が可能となる．特に，意味理解の機能を使う将来の知的情報アクセスでは，意味理解の処理の結果であるべき情報の全部ないしは一部が文章構成タグとしてあらかじめ付加されていることになる．これにより，自然言語処理で発生する曖昧さの数が大幅に減少でき，その結果自然言語処理の負荷が大幅に軽減される．

自然言語処理の最大の難関は，曖昧さの爆発にある．タグはこの曖昧さの爆発を前もってかなり抑える効果があり，効果的である．しかし，前にも述べたように，タグは自然言語処理の結果として付けられるべきものであり，これを前もって付けるということには，ジレンマがある．このジレンマの解決こそが将来の知的情報アクセスの実現の鍵である．

ストアド情報およびデリバリ情報のいずれに対しても情報検索および情報抽出が多く行われる．したがって，アクセスされるたびに自然言語処理を行っていたのでは無駄が多い．情報にタグを付けておくと，そのような無駄が回避される．

デリバリ情報は情報の発生から受信者への到達までの時間が短いので，豊富なタグを付加するための時間的な余裕がない．これに比べて，ストアド情報は恒常的に蓄積される情報であるので，豊富なタグ付けを行える．

ストアド情報にしろデリバリ情報にしろ，上で述べたすべての種類のタグが付けられていないと役に立たないというわけではない．極めて初歩的なタグが付けられているだけでも，情報検索・情報抽出の速度や精度は大幅に改善される．

ワープロでは，タグはテキストの裏に隠された内部的な記述として保持される．これに比べて，たとえばTexでは，タグはテキストに混在する制御文としての表層的な記述として保持される．これらの違いは，ワープロ発展の歴史的な経緯の違いによるものである．これらの記述に関して標準化や互換性を図ることが重要である．また，自動変換の機能も必要である．いずれの方式にも利点があるので，統合化された環境における選択として位置付けられることが望ましい．それによりタグの情報が知的情報アクセスで活用できるようになる．

タグが付けられたストアド情報は，一種の知識ベースとも見なされる．日々情報が追加・修正される知識ベースでは，このような形での構成がとても効果的となる場合が多い．従来のデータベースや知識ベースでは，格納情報を特別な形式で構成している．そのとき，言語表現されていた情報のすべてがエンコードされるとは限らない．エンコードされる情報は，そのときの関心事に基づいて抽出される．たとえば，法律知識ベースでは，その時点までの判例に基づいて情報がエンコードされる．ところが，それまでとは異なる新しい観点からの判例がでてきたときには，その新しい観点も加えたより広い観点からの法律知識ベースの再構築が必要となる．このような場合に，タグドテキストコーパスの形式による知識ベースは威力を発揮する．すなわち，タグドテキストコーパスの形式での情報格納は，情報アクセスだけではなく，格納情報に対する新しい情報の融合および古

い情報の更改などのために有効である．このような情報の表現の仕方は，「意味の表現」[19]や「理解の表現」[20]の1つの方法を与えると見なすことができる．

情報検索や情報抽出ではそれぞれの処理の目的があり，それによってテキストに付けられているタグの使い方も異なってくる．したがって，情報検索や情報抽出で使われるタグは，情報の解釈の仕方を完全に規定するためのものではなく，それを支援するためのものでなければならない．タグにより情報アクセスの目的が歪められたりすることがないように，タグのセットを階層的な構成にし，それぞれのタグの役割とそれぞれのタグの間の関係を明確にしておく必要がある．

4.3 自然言語処理の新しい動向

将来の高度な自然言語処理技術を開発するために，基本に立ち返った研究活動が開始されている．そのような活動の中の主なものの1つは，自然言語処理のための基礎データを得ることを目的とした言語データの作成に関するものである．ここでいう言語データとは，上で述べたタグドテキストコーパスである．

自然言語処理のための言語データは，2つの目的のために使われる．1つは，自然言語処理システムが使う辞書や文法を開発するためであり，もう1つは，自然言語処理システムの診断や評価を行うためである．機械翻訳システムなどの従来の自然言語処理システムは，十分な量と内容を持つこのような言語データを用意することなしに作成されてきた．

現在では，自然言語処理の方式として，コーパスベーストな処理や統計情報に基づく処理の方式が盛んに研究されており，これらの処理のためにもタグドテキストコーパスは不可欠となってきている．

4.4 タグドテキストコーパス作成の動向

米国では，米国の計算言語学会（ACL）がタグドテキストコーパスを作成しつつある．これは，TEI（Text Encoding Initiative）[11]と呼ばれている．また，ペンシルバニア大学では，長年の独自の研究に基づき，独自のタグドテキストコーパスを作成している[12]．さらに，音声データについてのタグドコーパスの作成も進められている．

欧州では，従来から，辞書に関する長期的な研究・開発プロジェクトが推進されてきた．これらを母体として，最近は，さらに米国のACLの活動との協力も含めて，言語データの作成が進められている．

東南アジアのいくつかの国でも，欧米のこれらの活動に呼応して，言語データを作成するプロジェクトの計画が検討されている．

日本でもいくつかの活動が推進されている．たとえば，米国のTEIの言語データ作成に関連して，いわゆる文化系の研究者が中心となって連絡・協力をしている．その他にも，たとえば，エィ・ティー・アール音声翻訳通信研究所（ATR）では，対話文を集めたテキストについて，ATRコーパスと呼ぶタグドテキストコーパスを作成している[13]．電子化辞書研究所（EDR）では，新聞記事のテキストについて，EDRコーパスと呼ぶタグドテキストコーパスを作成している[14]．リアルワールドコンピューティング（RWC）[15]でも，タグドテキストコーパスの作成を進めつつある．

日本電子工業振興協会（JEIDA）では，平成5年度より，研究者・技術者が研究・開発のために共有する実際の言語データの構成法の検討を長期的な観点から進めている．また，JEIDAが発行する報告書などをテキストとして活用し，これによりJEIDAタグドテキストコーパスを試作しつつある[16),17)]．

情報処理学会（IPSJ）では，同様な目的のための学術的側面を検討するために，平成6年10月に，電子化テキストコーパス作成技術研究グループ（ETC）を2カ年計画で発足させ，研究を進めている．ETCでは，IPSJが発行する学会誌，論文誌，および研究報告などの記事や論文をテキストとして活用するIPSJタグドテキストコーパスの試作を検討している[18]．また，IPSJが検討を進めている論文などの電子投稿の際のテキスト構成法との関連についても検討をしようとしている．

これらの活動は個別的に行われており，現時点での連携は少ない．現在はそれぞれが試行錯誤の状況にあり，ある程度の経験を積むと標準的な構成法が定まってくるものと思われる．しかし，上記の最後の2つと東南アジアのいくつかの国の関連プロジェクトは，アジア言語のタグドテキストコーパス作成に関して，連携の可能性について検討をしようとしている．

タグドテキストコーパスを実際に作成するに当たっては，その材料となるテキストの使用に関して，著作権の問題がからんでくる．これは，従来からの自然言語処理の研究の推進においても難題とされてきた．現在は，いくつかの新聞社などの協力により，研究目的の使用に限って，新聞記事のCD ROMが有償で提供されている．

4.5 タグドテキストコーパスの構成

タグは，タグラベルとタグバリューの対からなる．すなわち，

＜ tag-label tag-value ＞

の形をしている．タグは，文章，章，節，段落，文，文の中の節や句，語などのすべてのテキストの要素に対して付けられる．

従来のタグの種類の設定は，欧米言語について行われてきた．したがって，文法的なタグの種類も欧米言語に対するものとなっている．アジア言語に対しては，アジア言語としての言語的特性を取り入れたタグの種類の設定も必要である．また，欧米言語に対するタグとの整合性をとることも重要であり，このような観点からの包括的なタグの種類の設定が必要である．

タグドテキストコーパスは，先に述べたタグの種類に応じていくつかのレベルに分けられる．ここでは，文章構成に関するタグについてのみ述べる．すなわち，文法的なタグについて述べる．

第1のレベルは，形態素レベルである．これは，単語に対する文法的なタグである．そのためには，まず，文を単語の系列に分割しなければならない．この処理は，単語への分割の仕方がたくさんあるため，曖昧さが爆発する困難な処理である．英語などでは，単語と単語の間にスペースを入れて書くので，この処理は不要である．日本語を含むアジア言語の一部では，単語と単語の間にデリミタとなる情報を入れないで書くため，文を単語の系列に分割する処理は必須である．

形態素レベルのタグは，単語の系列に分割した文のそれぞれの単語に対して，表記，品詞，活用形，原形，内部構成などのタグを付ける．内部構成とは，接頭語や接尾語をともなっている単語の場合には，接辞と語との構成を示す．また，複合語の場合には，それを通常の語へ分割したときの構成を示す．

たとえば，「超高速」という単語に対しては，

＜語 超高速＞
＜品詞 名詞＞
＜活用形 NONE>
＜原形 SELF>
＜構成 接頭語＋語＞
　＜接頭辞 超＞
　＜品詞 接頭語＞
　＜活用形 NONE>
　＜原形 SELF>
　＜構成 SELF>
　　＜語 高速＞
　　＜品詞 名詞＞
　　＜活用形 NONE>
　　＜原形 SELF>
　　＜構成 SELF>

のようなタグを付ける．ここで，NONEはこの項目に関知しなくてよいことを示し，またSELFは自分自身であることを示す．ただし，ここでは，実際のタグの種類の一部のみを示してある．また，説明の都合上，分かりやすく例示している．

もう1つの例として，動詞「示すことができるものとする」に対する略記は，

動詞「示す」連体形＋形式名詞「こと」＋
格助詞「が」＋動詞「できる」連体形＋
形式名詞「もの」＋格助詞「と」＋
動詞「する」終止形

となる．ただし，ここでは，便宜的に，全体を1つの動詞として示した．このように，定まった用法と定まった意味を持つ複合語はたくさんあり，それらを1つの単語として扱うなどの工夫も必要である．

第2のレベルは，構文レベルである．これは，文の係受構造および格構造を記述する．係受構造とは，文の要素が文の他の要素を修飾する構成に関するものである．これには，従属節と主節との間の係受構造，埋込み文である節とそれが連体修飾する名詞句との間の係受構造，および連体助詞「の」を持つ連体修飾句とそれが連体修飾する名詞句との間の係受構造とがある．格構造とは，文の中のそれぞれの節について，その節の述語がその節の中の他の要素に対して持つ意味的な関係の構成を表すものである．これらの要素は，いわゆる5W1Hを表すものに相当し，通常は，「が」，「を」，「で」，「に」などの助詞をともなう形で言

語表現される．

第3のレベルは，意味レベルである．これは，構文レベルで設定された各要素にそれぞれ意味タグを付ける．意味タグのラベルとしては，構文レベルのものも加えて，

Predicate	要素
PartOfSpeech	品詞または句節標識
Inflection	活用形
Form	内部構成
Attribute	意味属性
Function	文法機能

などがある．このようなタグを付けたタグドテキストコーパスの容量は大変大きくなり，かつプリティプリンティングすると多くのスペースを必要とする．したがって，誌面の都合により，実際の例をここには掲載できない．

第4のレベルは，文脈レベルである．これは，文の要素と文の他の要素に対する照応関係のタグを付ける．照応関係とは，たとえば，代名詞や指示詞とそれらが差し示す言語要素との間の関係である．文の要素の言語表現が省略されることもしばしばある．省略された要素とそれを補完する要素との間の関係を省略関係と呼ぶ．省略関係も，照応関係の一種である．照応関係は，異なる文にまたがる場合も多い．

これらの他にも，並列構成に関するものおよび意味的論理構成に関するものなど，いくつかの種類のレベルのものがある．意味的論理構成のタグは，先に述べた知識ベースを構成するために使われる．たとえば，法律知識ベースでは，法的判断の論理的な流れを記述するのに使われる．これは，文章の意味を記述する典型的な例である．

現在までのタグドテキストコーパスには，たとえば，形態素レベルの文法的なタグを付けたブラウンコーパス，文書構成と形態素レベルの文章構成に関するタグを付けたTEIコーパス[11]，および形態素レベルおよび構文レベルの文章構成に関するタグを付けたPenn Treebank[12]などがある．Penn Treebankでは，文法的なタグの中で品詞に関するものが48種類あり，実際にはさらに細分化したものを用いている．先にいくつかのレベル分けをして示したタグは，これらをもっと多様化しかつ詳細化するものである．

インターネットや情報ハイウェイは全世界を結ぶので，テキストはいろいろな言語で書かれていることになる．現在では，WWWサーバでの機械翻訳を行うシステムもあるが，そのような処理を行うためには，バイリンガルないしはマルチリンガルなタグドテキストコーパスが必要になってくる．そのためには，言語間の対応する言語要素の間の関係についてのタグも必要となってくる．

知的情報アクセスに使うための情報へのタグ付けは，ストアド情報に対しては精緻な知的情報アクセスのためにできるだけ多くのレベルのものについて，一方，デリバリ情報に対してはリアルタイムな知的情報アクセスのためにできるだけ簡潔なレベルのものについて行うことになる．タグを付けると情報のサイズが大きくなるが，これは知的情報アクセスの機能の発揮によって十分に補われる．

ここで重要なことは，自然言語処理のための基礎データとしてのタグドテキストコーパスと知的情報アクセスのためのそれとが，全体と部分という関係を持つように設計されることである．

5. むすび

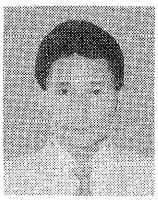
将来の情報ハイウェイにIntelliNetとしての機能を実現するために，特に，テキストを対象とする高度な自然言語処理の立場から，知的情報アクセスおよびそれを実現するための基礎となるタグドテキストコーパスについて述べた．これらは情報処理の新しい課題であり，それらへの取り組みは今まさにその緒についたばかりである．21世紀には情報ハイウェイへの情報アクセスが日常生活の重要な部分を占めることになる．したがって，これらの課題の解決は必須であり，今後重点的に取り組まなければならない．

参考文献

1) Nomura, H. : Information Extraction and Generation on Information Highway, Proc. of Second Natural Language Processing (1995).
2) Nomura, H.: Natural Languagc Proccssing on Information Highway, International Workshop on Language Engineering (1994).
3) 「ディジタル図書館」編集委員会：ディジタル図書館 (1994).
4) Ykoster, M. : Robots in the Web: Threat or Treat?
5) Harman, D. K. : Overview of the First Text Retrieval

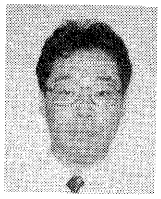
Conference (TREC-1) (1993).
6) Harman, D. K. : The DARPA TIPSTER Project, ACM SIGIR FORUM (1992).
7) Proc. of Fifth Message Understanding Conference (MUC-5), Morgan Kaufmann Publishers (1993).
8) Bryan, M. : SGML: An Author's Guide to the Standard Generalized Markup Language, Addison Wesley (1988).
9) Berners-Lee, T. and Connoly, D. : Hypertext Markup Language - 2.0, Internet-Draft, MIT (1995).
10) JIS X 4151, 文書記述言語 SGML, 日本規格協会 (1992).
11) Sperberg-McQueen, C. M. and Burnard, L. : Guidelines for Electronic Text Encoding and Interchange (1994).
12) Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A. : Building a Large Annotated Corpus of English: The Penn Treebank, ACL, Vol. 19, No.2 (1993).
13) エイ・ティー・アール音声翻訳通信研究所：ATR コーパス.
14) EDR コーパス：電子化辞書研究所.
15) 大津：リアルワールドコンピューティング, 人工知能学会誌, Vol.9, No. 3 (1994).
16) 自然言語処理技術の動向に関する調査報告書, 日本電子工業振興協会, 94-計-4 (1994).
17) 自然言語処理システムの動向に関する調査報告書, 日本電子工業振興協会, 95-計-3 (1995).
18) 「電子化テキストコーパス作成技術」研究グループの新設について, 情報処理, Vol.35, No.9 (1994).
19) 野村, 内藤:自然言語理解における意味表現, 情報処理, Vol.27, No.8 (1986).
20) 野村: 自然言語理解の構造-理解の表現-, 情報処理, Vol.30, No.10 (1989).

(平成 7 年 10 月 13 日受付)



野村 浩郷（正会員）

1944 年生. 1967 年大阪大学工学部卒業. 1969 年同大学院工学研究科修士課程修了. 同年日本電信電話公社電気通信研究所入所. 基礎研究所勤務. 1988 年九州工業大学情報工学部教授. 工学博士. 言語知能, 自然言語処理, 機械翻訳などの研究に従事. 日本言語学会, 計量国語学会, 電子情報通信学会, 人工知能学会, 米国 ACL 学会各会員.



井佐原 均（正会員）

1954 年生. 1978 年京都大学工学部卒業. 1980 年同大学院工学研究科修士課程修了. 同年通商産業省工業技術院電子技術総合研究所入所. 1995 年郵政省通信総合研究所関西先端研究センター知的機能研究室長. 工学博士. 自然言語処理, 機械翻訳の研究に従事. 人工知能学会, 言語処理学会, 認知科学会, ACL 各会員.

徳永 健伸（正会員）

1961 年生. 1983 年東京工業大学工学部情報工学科卒業. 1985 年同大学院理工学研究科修士課程修了. 同年（株）三菱総合研究所入社. 1986 年東京工業大学大学院博士課程入学. 現在, 同大学大学院情報理工学研究科助教授. 博士（工学）. 自然言語処理, 計算言語学, 情報検索の研究に従事. 認知科学会, 人工知能学会, 計量国語学会, ACL, ACM SIGIR 各会員.

中村 貞吾（正会員）

昭和 34 年生. 昭和 57 年九州大学工学部電子工学科卒業. 昭和 59 年同大学院修士課程修了. 昭和 62 年同大学院博士後期課程退学. 工学博士. 同年同大学工学部電子工学科助手. 平成 4 年九州工業大学情報工学部講師. 計算言語学, 自然言語処理に関する研究に従事. 電子情報通信学会, 人工知能学会各会員.