特別論説 情報処理最前線

インターネットと多言語情報処理

Multilingual Information Processing on the Internet by Gen-ichiro KIKUI (NTT Information and Communication Systems Laboratories).

菊井玄一郎[1]
1 NTT情報通信研究所

1. はじめに

インターネットの目覚しい発展によって，世界中の計算機で公開されている膨大な量の文書にアクセスできるようになってきた．インターネット上の文書の大きな特徴はそれらがさまざまな言語で書かれているということである．WWWブラウザを使えば英語を始めとして西欧諸語からアジア系言語まで多様な言語の文書にアクセスできることが分かるだろう．これらの文書を情報源や語学教材として活用できればインターネットの有効性が高まることはいうまでもない．

しかし，これは必ずしも容易ではない．言語によってはブラウザにテキストを正しく表示することすら困難であるし（いわゆる「文字化け」の問題），これができたとしても，多くの人にとって外国語の文書を理解するのは簡単なことではない．また，多言語文書の中から自分の要求に合った文書を検索するのには多大な労力が必要だ．

これらの問題は，テキスト通信，自然言語処理，情報検索などの各分野で認識され，現在活発な検討が行われている．本稿ではとくにWWWに関するこれらの動向を以下の4つの話題に分けて解説する．前半2つが基礎技術，後半2つが応用技術である．

まず最初は，多言語文書の符号化と復号に関する話題だ．これらはネットワーク上の任意の文書を「文字化け」なく転送するための基礎として重要である．多言語情報処理という言葉でこれをイメージする読者も多いと思う．

2番目の話題はテキスト言語識別についてである．ネットワークから受け取った文書が何語かを識別することは，ハイフネーション，検索，自動翻訳など多くのテキスト処理の前提となる基礎的な処理である．

3番目の話題として，多言語テキストの情報検索技術を取り上げ，とくに，母国語で外国語の文書が検索できる「クロスリンガルな検索処理」を中心に説明する．

最後の話題は外国語文書の内容理解の支援である．この技術の究極である自動翻訳や，電子辞書を引いて原文に訳語の候補や注釈をつける自動辞書引き技術などを紹介する．

2. 多言語テキストの符号化と復号

計算機ネットワーク内で，テキストは符号（数字）の列に変換（符号化）されて転送される．符号列を受け取った側が元のテキストを正しく復元するためには，符号化とちょうど逆の変換（復号）を行わなければならない．WWWブラウザで「文字化け」が生じる1つの大きな原因は，受け取った符号列に対して，その符号化の方法（符号化法）に対応していない復号が行われてしまうことにある．なぜこんなことになるのかというと，インターネットでは歴史的な事情から，国や計算機などに応じてさまざまな符号化法が用いられており，受け取った符号列だけではどの符号化法で符号化されているのか自動的には決定できないからである．

そういうわけで，多くのWWWブラウザには「符号化法を選択するスイッチ」がついており，アクセスしたWWWページで文字化けが生じたらこのスイッチを人手で切り替えて正しい表示にできるようになっている．しかし，ページごとにスイッチを切り替えるのは煩雑であるし，選択を行う基準が「文字化け」の有無というのもうまく

ない．母国語や英語くらいならともかく，未知の言語のテキストの文字が化けているかどうか容易には判別できないからである．

本章ではこの問題に対するいくつかの解決策を以下に紹介する．

2.1 国際化文字符号化法

最も根本的な解決法は世界中の言語のテキストが正しく符号化できるような符号化法（国際化文字符号化法と呼ぼう）を1つ決めて，これ以外は使わないようにすることである．このような符号化法としてISO 2022とISO 10646の2つを簡単に紹介する．なお，詳しくは文献1)～4)などの解説や関係する規格書を参照してほしい．

2.1.1 ISO 2022

ISO 2022[5]を考える上で重要なのは文字集合という概念である．文字集合とは，主として国家などがその地域内で使われる文字を集めて規格化したものである☆．たとえば，日本では，漢字，かな，記号などからなるJIS X 0208が制定されている．各文字集合には固有の識別番号が与えられており，さらに，文字集合の各々の文字にはその文字集合内で固有の識別番号が付与されている．したがって，文字集合の識別番号と文字集合内における識別番号を組み合わせればどのような文字でも一意に特定できる．

ISO 2022は基本的にこのような考え方で作られた文字符号化の方法である．なお，実際のテキストでは1つの文字集合内の文字が連続することが多いので，文字集合の識別番号は文字集合が切り替わったときだけ提示することにして無駄を省いている．**図-1**に日本語と中国語との混在しているテキストの例を示す．

現在ISO 2022に準拠した多言語WWWブラウザとしてMosaicL10n[6]，i18n Arena[4]が開発されている．また，テキストのみしか表示できないが，ISO 2022用のテキストエディタmule[4]上で動作するWWWブラウザ(w3-mode)も利用できる．このほかISO 2022をOSレベルでサポートした計算機環境も開発されている[7]．

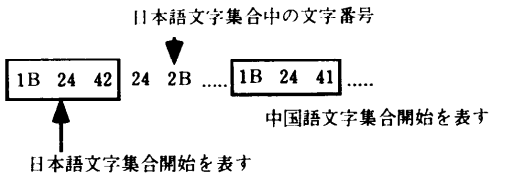


図-1 ISO 2022による符号の例

☆ ここでいう「文字」はワープロなどでおなじみの「フォント」とは違う概念であることに注意してほしい．明朝体の「あ」とゴシック体の「あ」は，同じ文字であるがフォントは異なる．

2.1.2 ISO 10646とUNICODE

一方，近年新たに提案されたISO 10646[8]は世界中の文字を含む単一の文字集合を新たに定義して，これを使うことが特徴である．したがって，ISO 2022のように「文字集合の切替え」は行わない．文字集合は4バイト（UCS-4）または2バイト（UCS-2）のアドレス空間で定義され☆☆，前者は後者を包含している．

最近話題のUNICODE(2.0)はこのISO 10646とほぼ同じものである☆3．UNICODEはいくつかのOS(Newton,WindowsNTなど)で採用され始めている．また，HTMLの文字集合としてISO 10646を採用する方向で議論が進んでいる[9]．

2.2 文字符号化法の通知

1つの符号化法を世界中で使うというのは理想だが実現には時間がかかる．複数の符号化法の混在する現状を前提とした上で，これに対処する1つの方法は，テキストと一緒にそのテキストの符号化法を通知する方法である．これは，一部のメイルシステムなどで実現されている．たとえばインターネットメイルのヘッダ部分をみるとFrom:やTo:などとならんで

Content-Type: text/plain; charset=iso-2022-jp

のような記述が見つかることがある．後半のcharset以降の部分がメイル本文の符号化法を表した部分である．WWW（HTTP）でも同様のことが可能である．

なお，これを実現するためには，ヘッダ情報を付加する処理や文書をネットワークに送信するサーバなどが当該文書の符号化法を知っていなければならない．これを確実に行うには，テキストを最初に符号化する処理（エディタやオーサリングツールなどの入力ツール）とヘッダ情報を付加する処理とが連携して，文書の符号化法に関する情報を正しく伝達することが必要である．

2.3 文字符号化法の自動識別

今まで述べた符号化法の識別はテキスト発信側

☆☆ 現在実質的に定義されているのはUCS-2のみである．

☆3 UNICODE1.0は2バイトの文字集合に限定されており，これはほぼUCS-2に対応している．

の協力を必要とするものであった．これが期待できないときは文字符号化法を自動推定するしかない．これは誤りを犯す可能性がある点で，いわば「最後の手段」であるが，現状ではこれに頼るしかない場合が多い．たとえば，多くのWWWブラウザで「日本語モード」を選ぶと，受け取った符号列の符号化法をSJIS，EUC(UJIS)，JISの3つの中から自動的に推定して復号・表示する．入力がこれら3つの符号化法のどれかであることがあらかじめ分かっている場合には比較的簡単に符号化法を推定することができるのだ[1]．

しかし，ネットワーク上の任意の文書を対象とした場合にはこれが成り立たない．この場合，確実にできることは符号化法の可能性を絞り込むことまでだ[4]．100％確実でなくてもよければ言語識別を応用して絞りこまれた中から1つのものを推定する手法[10]が利用できる（次章参照）．

3. テキストの言語識別

ネットワーク上でアクセスした文書の言語を知る最も確実な方法は，文書の発信元にこれを通知してもらうことである．それが不可能な場合には先の符号化法と同様，文書の受け手が何らかの手段で言語を推定しなければならない．

3.1 言語名の通知

WWWでは文書を送信する際のプロトコル（HTTP）の一部にその文書の言語を通知する手段が含まれている．しかし，現在これはほとんど使用されていない．また，HTMLのタグを使って文書の任意の部分が何語で書かれているかを記述できるようにしようという提案もある[9]．このような手段を文書送信側が用いるようになれば，WWW文書に関する限り，受信側は簡単に文書の言語を知ることができる．

しかし，この場合は送信側で文書の言語が既知でなければならないことに注意が必要である．文書作成者が言語名を人手で入力するのが確実であるが，文書，あるいは，文書の部分ごとにこれを入力するのは繁雑であろう．これを避けようとすると，結局，自動識別に頼ることになる．

3.2 言語の自動識別

3.2.1 統計的手法による言語識別

ここでは西欧系言語と東アジア系の言語に絞って解説するが，同様の手法はほかの言語にも適用可能である．

(1) 西欧系言語の自動識別

テキストの言語を自動識別する処理は多言語社会である西欧で活発に研究されている(文献11)，12)など)．与えられたテキストが西欧言語のうちの何語かを精度よく識別する方法の基本的な考え方は，「単語(や文字)の頻度分布の言語による違いを利用する」というものである．

簡単にいうと，a, the, atなどが多く出現すれば英語と判断し，der, fürなどが多く出現すればドイツ語と判断するわけだ．単語そのものを使わず，なるべく言語間の違いが大きく表れて，かつ，短い文書でも多く出現するようなものを選ぶと精度が向上する．1つの解は単語を語尾n文字で同値類に分け☆4，この同値類の頻度を用いる方法だ[10]．また，一文書内での複数言語の混在に対処するためには文書をパラグラフ単位に区切ってこれら各々に対して識別を行えばよい．

なお，各言語特有のアクセント文字（たとえばç）が出現したらその言語(フランス語)に決めるという方法もあるが，このようなアクセント文字が現れない文書も多く精度が悪い．

(2) CJK言語の識別

中国語，日本語，韓国語をまとめてCJK言語と呼ぼう．これらの言語の特徴は，単語と単語の間が明確に区切られていないこと，文字の数が多いことである．とくに，漢字はそれ自体で意味をもつことが多く，荒っぽくいって1つの単語のようなものと考えられる．このような言語では文字の頻度統計を用いた自動識別が有効である[10]．

(3) 西欧系言語とCJK言語との識別

西欧系言語とCJK言語との間では利用される文字が大きく違うので，テキストが国際化文字符号系で符号化されていればこれらの間の識別は容易である．問題は現状のように与えられたテキストの符号系が分からない場合である．これに関しては次に述べる．

3.2.2 符号系と言語の同時識別

現状ではネットワークから得られたテキストの符号系も言語も分からないという場合が多い．このような場合に対処するため復号処理と言語識別とを組み合わせた方法が提案されている[10]．

☆4 たとえばtheはtheという同値類，spazierungはX-rungという同値類に置き換える．

基本的な考え方は次のとおりだ．まず符号列に対して可能な復号をすべて試み，次にエラーなしに得られたすべての復号結果（文字列）に対して統計的な言語識別処理を行い言語とその尤度を求める．最後に，各文字列の中で識別尤度の最も高いものを復号結果として出力する．なお，復号結果中に複数の文字集合が混在している場合は，文字集合ごとに文書を分割してから識別処理を行う．

以上の方法によって，西欧系言語あるいはCJK言語（これらの混在も許す）で書かれた文書であれば，符号系が未知であっても復号と言語識別を実現している．

3.3 その他の手がかりによる言語識別

統計的手法の欠点は学習データが必要なことと，識別対象の文字列がある程度長くなければならないことである．この欠点を補う方法として以下が考えられる．

(1) 文字集合

テキストの文字集合と言語の間にはある程度の依存関係がある．たとえば，日本語文字集合が使われている文書は日本語である可能性が高い．しかし，日本語の漢字を用いて中国語テキストを表現する例がある☆5など，文字集合と言語との結びつきは緩いことに注意が必要だ☆6．

(2) ドメイン

文書のURLから発信元の国（地域）を割りだし，その国における主要な言語を文書の言語とする方法である．たとえば，フランス発信のテキストをフランス語と推定するわけだが，これもあくまで可能性が高いというだけである．

(3) アンカー（リンク）

HTMLページでEnglish version is here.といったリンクをクリックすると，ほとんどの場合英語のページが現れる．このように，リンクに付与された文字列（注釈）が飛び先のページの言語名を表すことがよくある．したがって，リンク近傍のテキストをうまく解析することによって飛び先のページの言語を推定することができる．

☆5 ネットニュースfj.soc.chineseで利用されている．
☆6 筆者は，JIS X 0208のキリル文字を使ってロシア語を表現したhtml文書に遭遇したことがある．

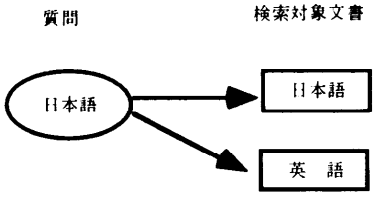
4. 多言語検索支援

今まで述べてきたような基礎技術が確立すれば，これを土台として，多言語文書の利用を支援するためのさまざまな情報処理が可能になる．本章では，インターネットということでまず必要となる多言語情報検索について述べる．

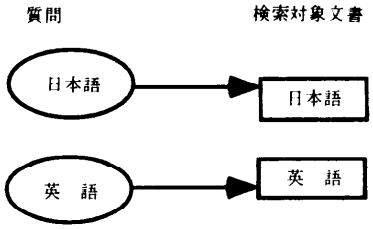
4.1 多言語テキストの内容検索

多言語テキストの内容検索とは，さまざまな言語で書かれた大量の文書の中から利用者の要求に合った内容の文書を検索することをいう．最も分かりやすいのは，日本語で「電話」と入力しただけで，電話に関する文書ならば言語を問わず検索するものである．このように検索要求の言語とは別の言語の文書でも検索できるようなシステムを「クロスリンガル検索システム」と呼ぼう(図-2(a))．また，入力の言語と同一言語の文書しか検索できないが，さまざまな言語の検索要求を受けつけて入力された言語の検索結果を返すようなシステム(図-2(b))も広義の多言語検索に含まれる．

さて，以下では情報探索支援ツールを「ディレクトリ型」と「質問型」とに分けて説明する．ここで，ディレクトリ型（たとえばYahoo[13]）とは情報資源をトピックごとに分類した階層型ディレクトリを提示するもので，利用者はこのトピック



(a) クロスリンガル検索



(b) クロスリンガルでない多言語検索

図-2 多言語検索の分類

の階層をたどることによって望む情報に接近する．一方，質問型（たとえばAltaVIsta[14]）とは全文検索などのテキスト検索エンジンを用いて，WWW文書の中から利用者の与えた検索要求（キーワードや語句）と関連性の高い情報資源のリストを提示するものである．

4.1.1 ディレクトリ型検索

まず，ディレクトリ型検索からみていこう．1）分類名が複数の言語で記述されているもの，あるいは，2）複数の言語の文書が分類されているものはクロスリンガル検索システムといってよいだろう．この点から既存のディレクトリ検索のほとんどは，検索言語が1つに固定されたクロスリンガル検索システムである．つまり，利用者は特定の言語（たとえば日本語）で階層構造ディレクトリをたどることによって複数言語の文書に到達できる．

4.1.2 質問型検索

まず，単言語の場合を考えよう．質問型検索の基本は「検索要求に現れる語句と同じ語句を含むような文書を探して出力する」というものである．多くのシステムではこれを行う前に，検索要求や対象文書の屈折変化系（たとえば英語における複数形）を原形に戻したり，表記のゆれ（日本語における「コンピュータ」と「コンピューター」）を吸収したりといった「単語の正規化」の処理を行うことによって，検索精度（再現率）を高めている．また，日本語などの単語境界が明示されない言語で単語単位のマッチングを行う場合にはテキストを単語単位に切り離す処理を行う場合がある．

さて，このような処理を多言語環境で実現するためには，まず，検索対象や検索要求の言語に応じた単語の正規化処理を適用することが必要となる．これができれば，クロスリンガルでない多言語検索（図-2(b)）は一応可能となる．日本語と英語を対象としたシステムの例として文献15），16）があげられる．

質問型のクロスリンガルな検索処理はさらにいくつかの処理が必要である．これは最近注目を集めている研究分野なので，少し詳しく説明する☆7．

基本的な方法は，和英辞書のような対訳辞書を用意しておき，入力された検索要求の各単語に対してこの対訳辞書を引いて別の言語の質問を作るというものである[18]．

たとえば，「カリフォルニアの地震」という日本語の検索入力に対しては，california earthquakeという英語質問を作り，後者を使って英語の文献を検索する．

一般に1つの単語に対して訳語の候補は複数あるので，これらをどのように使うかが問題となる．1つの方法は各単語の訳語候補のすべての組合せで単語列を作り，これの選言結合（OR）によって検索を行うというものである（EMIR[19]）．たとえば，入力が2単語からできていて，一方が2つ他方が3つの訳語をもつとき，6個の訳語単語列を作り，これらのOR条件で検索するわけである．単語列の中にはおかしなものも含まれるが，それらは実際の文書にはまず存在しないので悪影響はないというのが根底にある考え方である．また，日本語の検索入力を英語に変換する文献20）の例では，日本語単語の意味属性に着目した翻訳辞書および翻訳規則（語彙数，規則数ともに約2300）を作成し，これを用いた一種の翻訳システムを作ることによって訳語の多義を解消している．

一方，対訳辞書を手作業で作るのではなく，あらかじめ作成された対訳コーパス（対訳文書の集合）を訓練データとして対訳関係を「学習」し，これを使って質問文を生成する手法も研究されている[21]．

いずれにせよ，これらの方法は，ある単語に対する訳語をその単語に対する「同義語」と考えて後者と前者の両方で検索を行うという点でシソーラスで同義語を引いて検索する方法と類似している☆8．この点に着目して，シソーラス利用と似た効果をもつLSI（Latent Semantic Indexing[23]）の手法をクロスリンガルな検索に用いる手法も提案されている[24]．

4.2 言語を指定した検索

インターネット上にあるさまざまな言語の文書の中から利用者の望む言語の文書のみを検索する機能も重要である．3章で述べたような，各テキ

☆7 研究論文や国際会議などの情報は文献17）を参照してほしい．

☆8 初期の研究として文書中の単語を言語間で共通な概念記号に変換して検索する手法がある[22]．

ストに対して言語のラベルを与える仕組みを使えば，これは容易に実現することができる[18]．

4.3 検索出力の表示

現在の検索精度ではシステムの提示した検索結果の中からほしい情報を取捨選択するというインタラクションが必ず介在する．しかし，情報の取捨選択のためだけに外国語テキストを読むのは苦痛であろう．したがって，クロスリンガルな検索システムでは検索結果の一覧を利用者の言語で表示することが有効である．WWWページの検索で検索されたHTML文書のタイトルを利用者の言語に翻訳して提示するシステムが提案されている[18]．もちろん，利用者が翻訳機能つきのブラウザ（次章参照）を利用している場合にはこれを用いて検索結果を翻訳することもできる．

5. 外国語情報の理解を支援する技術

計算機による外国語理解支援は単に外国語で書かれた情報の理解に役立つだけでなく，外国語学習☆9のための有効なツールとなることが期待できる．

さて，WWWを対象とした支援システムはネットワーク内における位置によって，個々の利用者のWWWブラウザに組み込まれた「ブラウザ型」と中継サーバ（proxy server）に組み込んで複数人で使う「Proxy型」とに分けられる[25),26]．大雑把にいって支援システムが高価であったり，高い計算能力を要求したりする場合は後者が適しており，そうでなければ前者が適している．そういうわけで技術の普及とともに後者から前者の形態に移っていくようである．

(1) オンライン辞書引き

まず，オンライン辞書引き処理をあげよう．これは与えられたテキストの各単語に対して訳語辞書（たとえば英和辞書）を引いて，その結果を利用者に表示するものである．EtoJ_Proxy[27]はProxy型のシステムで，元のページの各英単語をその日本語訳に遂語的に置き換えたページを出力する．辞書は「ライフサイエンス」の分野に依存した日英対訳辞書である．専門用語の遂語訳を行うシステムの利点として「難解な英語専門用語を日本語に置き換えることによって視認性が向上し，ブラウジングの効率化に役立ち，かつ，翻訳に比べて高速である」点があげられている[27]．

同じくProxy型の辞書引きシステムの[25]は，WWWページの各単語を訳語に置き換えるのではなく，リファレンス（辞書）項目を参照するリンクを埋め込んで出力する．利用者は出力されたページ中の各単語をクリックすることによってその単語に対する辞書項目にアクセスできる．

(2) 自動翻訳

外国語情報の理解を支援する技術の究極は（質のよい）自動翻訳システムであろう．従来，翻訳システムはなかなか一般の人々からは遠い存在だったが，WWW用のProxy型翻訳システム[28]が発表されて以降，WWW用システムの開発が続き，現在ではブラウザ型の低価格な英日自動翻訳システムが多数市販されている．

(3) 対訳アライメント

WWW上には対訳，あるいは，抄訳関係にあるテキストがいくつか存在する．対訳関係にある段落や文を対応づけるアライメント技術（文献29）など），そして，この対応関係を分かりやすく利用者にみせる技術も重要である．こうすることによって利用者は一方の言語を頼りに他方の言語のテキストを読解することができる．また，対応づけられたテキストコーパスは外国語の文書を作成する際の大きな助けになる．

読解支援技術は単純な「辞書引き」から「翻訳」までいくつかのレベルがあり，それぞれ，文書の質や利用形態によって向き不向きがある．たとえば契約書などはかなりよい自動翻訳が利用できたとしても原文にあたって確認することが必要であろう．したがって，これらさまざまなレベルの支援機能を多層的に組み合わせて自由に使えるものが望ましい．

6. さいごに

本稿ではインターネット上の外国語文書の利用に関わる情報処理技術について述べた．なお，ここで触れることのできなかったが，多言語文書の発信を支援する技術も重要であることをあげておく．

現在，インターネットは爆発的な勢いで世界各

☆9 CALL: Computer Aided Language Learningと呼ばれる．

地へ浸透しており，利用者の裾野も拡大している．これに伴ってネットワーク上にはますますさまざまな言語の文書があふれるようになるだろう．また今後はインターネットだけでなく，企業活動の国際化などにともなって企業内ネットなどの比較的閉じた情報システムでも多言語文書を扱う機会が増えると思われる．

このような状況を考えると多言語情報処理は，今後，世界中の人々の情報共有のためにますます重要な役割を果たすものと思われる．本稿によってこのテーマに興味をもっていただければさいわいである．

謝辞 本稿に対して貴重なコメントをいただいた，本誌編集委員，および，NTT 情報通信研究所，林良彦，加藤恒昭，奥雅博の各氏に感謝する．もちろん内容の誤りは筆者一人の責任である．

参 考 文 献

1) Lunde, K.（春遍雀来，鈴木武生訳）：日本語情報処理,ソフトバンク(1995).
2) 太田昌孝: 文字コードと国際化, bit Vol.27, No.6, pp.4-10, 共立出版(1995).
3) 柳田俊彦: 特集: 文字コード, 日経バイト 1996 年 6 月号, pp.186-205(1996).
4) 錦見美貴子, 高橋直人, 戸村 哲, 半田剣一, 桑理聖二, 向川信一, 吉田智子: マルチリンガル環境の実現 X-Window/Wnn/Mule/WWW ブラウザでの多国語環境, プレンティスホール出版(1996).
5) ISO 2022: 1986. International Standard — Information Processing — ISO 7-bit and 8-bit coded character sets — Code extension techniques(1986).（日本工業規格 JIS X 0202(-1991), 日本規格協会(1991).
6) Toshihiro Takada: Multilingual Information Exchange through the World-Wide Web, Proceedings of WWW'94(1994).
7) 片岡 裕, 片岡朋子, 上園一知, 大黒屋秀治郎, 大矢俊夫, 小原啓義: 全世界の文字と言語の完全混在処理環境: Internationalized Multilingual System-The Waseda I18N & ML System, ディジタル図書館 No.6, pp.22-31(1996).
(http://www.DL.ulis.ac.jp/DLjournal/)
8) ISO/IEC 10646-1:1993. International Standard — Information technology — Universal Multiple-Octet Coded Character Set (UCS)— Part 1: Architecture and Basic Multilingual Plane(1995).(日本工業規格 JIS X 0221-1995, 日本規格協会(1995)).
9) Yergeau, F., Nicol, G., Adams, G. and Duerst, M.: Internationalization of the Hypertext Markup Language, draft-ietf-html-i18n-05.txt, Internet Draft(1996).
10) Kikui, G.: Identifying the Coding System and Language of On-line Documents on the Internet, Proceedings of COLING'96(1996).
11) Peter, H.: Language Identification for the Automatic Grapheme-to-Phoneme Conversion of Foreign Words in a German Text-to-Speech System, Proceedings of Eurospeech 1989, pp.220-223(Sep.1989).
12) Cavnar, W. B. and Trenkle, J. M.: N-gram Based Text Categorization, Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, pp.161-169(1994).
13) http://www.yahoo.com/
14) http://www.altavista.digital.com/
15) http://www.fujitsu.co.jp/hypertext/wais/1/gaiyo.html
16) http://navi.ntt.jp/j/info/infobee.html
17) http://www.ee.umd.edu/medlab/mlir/mlir.html
18) 菊井玄一郎, 鷲崎誠司, 林 良彦, 砂場倫太郎: インターネット情報ナビゲーションにおける多言語機能, 自然言語処理の応用に関するシンポジウム論文集, 情報処理学会(1995).
19) Fluhr, C., Schmit, D., Ortet, P., Elkateb, F., Gurtner, K. and Semenova, V.: Distributed multilingual information retrieval, Proceedings of MULSAIC'96 (1996).
20) 三日市紀子, 石川徹也: 電子図書館における専門用語の課題－検索インターフェースとしての複合語生成・翻訳システム, 専門用語研究 No.9, pp.3-10, (1995).
21) Sheridan, P. and Ballerini, J. P.: Experiments in Multilingual Information Retrieval using the SPIDER System, Proceedings of SIGIR'96(1996).
22) Salton, G.: Automatic Processing of Foreign Language Documents, Journal of the American Society for Information Science, 21:187-194(1970).
23) Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R.: Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science, Vol.41, No.6, pp.391-407 (Sep.1990).
24) Duamis, S. T., Landauer, T. K. and Littman, M. L.: Automatic Cross-Linguistic Information Retrieval using Latent Sematic Indexing, In SIGIR'96-Workshop on Cross-Linguistic Information Retrieval, pp.16-23(1996). (http://superbook.bellcore.com/~std/LSI.papers.html)
25) 伊藤修一, 梅村恭司: WWW での辞書引き方法の比較検討, 情報処理学会研究報告 96-HI-64, pp.49-54, (1996).
26) 村田稔樹, 山本秀樹, 永田淳次: 言語の違いを意識しないインターネット利用を可能とする WWW 用機械翻訳システム, 情報処理学会研究報告 95-IM-21, pp.19-26(1995).
27) 鵜川義弘, 秋元 学, 藤田信之, 佐藤 豊: DeleGate Proxy を使った Web 情報の英和逐語訳サーバー EtoJ_Proxy —, Proceedings of Japan World Wide Web Conference '95(Nov.1995).
28) Yamamoto, H., Murata, T. and Nagata, J.: W3-PENSEE: WWW Machine Translation System that Supports the Comfortable Internet Surfing, Proceedings of International Symposium on Digital Libraries 1995,

pp.159-166(Aug.1995).
29) Haruno, M., Ikehara, S. and Yamazaki, T.: Learning Bilingual Collocations by Word-level Sorting, Proceedings of COLING'96(1996).

(平成8年10月2日受付)

菊井玄一郎（正会員）

1961年生．1986年京都大学工学部電気工学第二専攻修士課程修了．同年NTTに入社，情報通信処理研究所に配属される．1990年から1994年までATR自動翻訳電話研究所に出向（この間1993年ドイツ人工知能研究センタ（DFKI）滞在研究員），1994年よりNTT情報通信研究所．自動翻訳，文書訂正支援，多言語インターネット情報検索など主に自然言語処理関係の研究・開発に従事．