WWWブラウジングのための言語処理とヒューマン インタフェース

Natural Language Processing and Human Interface for WWW Browsing

平川 秀樹 H. Hirakawa　住田 一男 K. Sumita　武田 公人 K. Takeda

WWW (World Wide Web) により，インターネットの利用が爆発的に広まり，現在ではだれもが世界中の情報に容易にアクセスできるようになっている。この利便性の反面，情報が多すぎるため欲しい情報を得るコストが非常に高くなってしまう“情報洪水”の問題や，インターネット上の情報のほとんどが英語で書かれているために生ずる“言語障壁”の問題が顕在化している。このようなWWWブラウジングにおける問題への対処として，WWWページの言語解析とユーザの嗜(し)好を表すプロファイル(検索条件)により，好みにあった文書だけを選択・収集・統合提示する情報フィルタFreshEye$_{TM}$，および業界最大規模の23万語の辞書をもち，英文WWWページを日本語に翻訳・表示し，概要のすばやい理解を支援する機械翻訳システムASTRANSAC for Internet$_{TM}$を実現した。

With the recent rapid penetration of the Internet through the World Wide Web (WWW), people can now easily obtain information from around the world by WWW browsing. Two problems have emerged from this popularization of the Internet: the so-called “information flood,” and the “language barrier.”

To overcome each of these two problems, we have developed a software agent named FreshEye$_{TM}$ featuring a personalized information filtering function, and an English-to-Japanese machine translation system named ASTRANSAC for Internet$_{TM}$ featuring a 230,000-word dictionary. These products enable users to obtain the kind of information they need from WWW pages, and to read English WWW pages in machine-translated Japanese.

1 まえがき

インターネットの世界人口は1億人，わが国では500万人ともいわれ，その数は急速に増加しつつある。インターネットの普及の一つの転換点は，WWW (World Wide Web) の登場であり，文書とハイパーリンクによるネットワークインタフェースが幅広い支持を受けた。現状では，WWWは，パソコンやイントラシステムにおけるインタフェースのプラットフォームとしての地位を築きつつある。さらに，モバイル端末などでの利用も広がりつつある。この先には，ほとんどの情報はサイバースペースに置かれるようになるネットワーク社会がつながっている。

サイバースペース上に記憶され，流通する情報において，人々が日常的に情報蓄積・伝達の手段として利用する言語(自然言語)の役割はきわめて大きい。そこでは，言語メディアを利用したヒューマン インタフェース(HI)技術や言語メディア情報のコンテンツを理解する技術が不可欠のものとなる。

ここでは，ネットワーク社会への入り口であるWWWブラウジングを取り上げ，そこに見られる問題(“情報洪水”，“言語障壁”)と自然言語処理技術・HI技術を応用したWWWブラウジング用ソフトウェア(FreshEye$_{TM}$，ASTRANSAC for Internet$_{TM}$)について述べる。

2 WWWブラウジングと言語処理

2.1 情報洪水問題と情報フィルタリング

インターネットを効率的に利用する仕組みであるWWWブラウザの普及によって，世界中の最新データがリアルタイムに直接取り出せるようになり，情報の伝達の時間遅れがきわめて短くなった。ところが，個人，大学，企業などから発信されているページは，膨大な数となり，いかに必要な情報にたどりつくかが深刻な問題となっている。すなわち，提供される情報の量が多すぎて，ほんとうに必要な情報を発見することが困難になってしまうという“情報洪水の問題”である。

情報洪水に対する一つの解決手段として，個人の嗜好や興味に基づいて，欲しい情報を的確に収集し，わかりやすく提示してくれる情報フィルタリング技術は，重要なキー技術である。当社は，ニュースソースを対象とした情報フィルタリング技術を開発したが，これに基づく情報提供サービスが1996年6月から運営されている。現在，この技術は，インターネットの文書を対象として一般のインターネットユーザに利用可能なエージェント ソフトウェア(FreshEye$_{TM}$)に進展している。

2.2 言語障壁問題と機械翻訳システム

WWWブラウジングの第二の問題は，“言語障壁”である。

インターネットの公用語は，英語であると言われているように，インターネット上の情報の多くは英語で書かれている。そもそも英語を習得していないユーザが，英文のWWWページの情報を活用したり，楽しんだりすることは困難であるが，英語が読めるユーザにとっても，母国語でない英文WWWページをあれこれブラウジングする作業は，かなりの労力を要する作業である。

このようなユーザを支援するために，機械翻訳システムをWWWブラウザと連動して動作させることにより，英文WWWページを日本語で表示するシステム（機械翻訳ブラウザ）を開発した。

3 WWW情報フィルタ

WWW上に公開されているホームページを対象に，ユーザに取って代わって，ユーザの関心に沿った情報を収集するエージェントソフトウェア（FreshEye$_{TM}$）を開発した。このソフトウェアは，次の特長をもっている。

(1) 英語，日本語両言語に対応した情報の選択，順位づけ

(2) 新聞記事ページに対する記事単位の情報選択

(3) 学習機能によるユーザの好みへの適応

Windows®(注1) 95上で動作し，ブラウザと連携して動作する。

3.1 システムの構成

図1にシステムの構成を示す。システムはユーザが指定した時間間隔で起動し，複数のページに自動的にアクセスして必要な情報を取得する。主な処理モジュールの処理内容を以下に述べる。

(1) ページの取得　ユーザが設定した検索サーバや新聞社が公開している新聞記事のページにアクセスし，複数のページを取得する。例えば，検索サーバへのアクセスを設定している場合，システムはその検索サーバに検索を依頼し，前回の処理時に得られた結果と比較することにより，新規に登録されたページの検出が可能である。

(2) 書式解析　新聞記事のページでは，1ページ内に複数の記事が含まれる場合がある。これらの記事を分割することにより，記事単位の情報選択を可能にした。

(3) 言語解析　日本語文に対する単語切り（形態素解析），英語単語に対する語幹抽出などを行う。WWW上の文書は日本語，英語混在の文書も多い。このような日本語英語が混在する文書に対応している。

(4) フィルタリング処理　フィルタリング処理では，ユーザの関心や興味を記述したプロファイル（検索条件）との類似度を算出し，類似度の順に情報をランキングする。類似度の算出ではベクトル空間法をベースに語の頻度情報，見出しや段落情報，係受け情報などを解析しランキングを行う(1)。

(5) ページ生成　フィルタリング処理で選択されたページについて抄録を生成する。ここでは，プロファイル中の単語を含む部分テキストを抽出する方式を採用した。こららの抄録を組み合わせてブラウザが表示できるHTML（HyperText Markup Language）文書を作成する。

(6) 学習　Webブラウザで，任意の文書を参照時にその情報に関心があるかないかをフィードバックする学習機能を設けた。システムは，指定された記事からキーワードを自動抽出し，プロファイルを更新する。

(7) トピックス管理　アクセスするページや検索サーバ，フィルタリングのためのプロファイルなどを設定するインタフェースである。これらの情報を，ユーザは階層的に話題別に管理することが可能である。

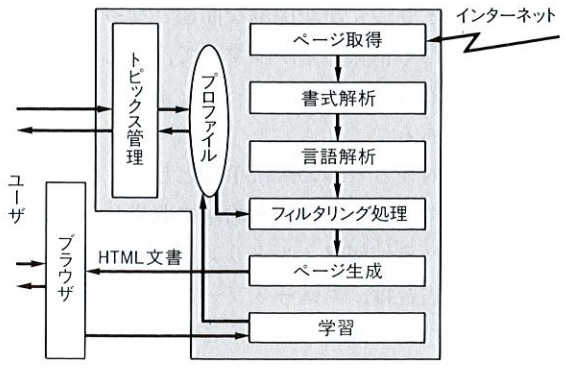

図1．FreshEye$_{TM}$のシステム構成　インターネットからWWW文書を取り込み，解析結果をユーザの好みを表現するプロファイルと照合，選択して表示する。

System configuration of FreshEye$_{TM}$

3.2 動作例

ユーザが設定した時間間隔（毎日1回，週2回など）で，システムは自動的に情報収集を始め，情報収集を終えた時点でフィルタリング結果を整理しHTML文書を作成する。図2に，作成されたHTML文書をブラウザで表示したようすを示す。図示したように，選択されたページや記事は話題別に一覧できるようになっている。HTML文書はパソコン上のファイルとして作成されるため，ネットワーク経由でページを表示する場合に必要となる待ち時間なしに，即座にブラウザに表示することが可能である。

なお，このソフトウェアの試用版は，次のページからダウンロード可能である（97年9月現在）。

http : //www.toshiba.co.jp/tech/software/fresheye

http : //softpark.jplaza.com/MISC

(注1) Windows95は，Microsoft社の商標。

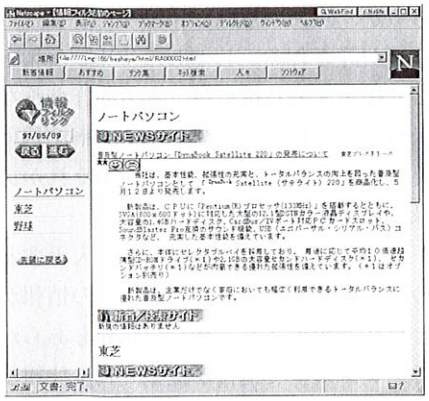

図2．フィルタリング結果の表示　ユーザの好みに合致する記事が選択され，一つの画面にトピックごとにまとめられて表示される。
Example of display of selected articles

4 機械翻訳ブラウザ

機械翻訳ブラウザとは，WWW ブラウザに言語翻訳機能を追加付与し，母国語外言語で記述された WWW ページを母国語に機械翻訳して表示できるようにしたものである。図 3 は，当社製品である ASTRANSAC for Internet$_{TM}$の表示画面例であり，画面下側にオリジナルの英文 WWW ページ，画面上側にその翻訳結果である日本語の WWW ページが表示されている。

4.1 機械翻訳ブラウザの実現方式

WWW ブラウザに機械翻訳機能を組み込む方式には，次のものがある。

(1) クリップボード翻訳方式　ユーザのテキストドラッグ操作により Windows® のクリップボードに登録された文字列を翻訳エンジンにより機械翻訳し，結果を専用ウィンドウに表示する方式である。これは，英文をベースに読み，その一部を翻訳したい場合などの使用に適している。

(2) ページ翻訳方式　WWW ブラウザにいったん英語を表示し，このページデータ（HTML データ）を解析し，文字部分を抽出・機械翻訳して，日本語のページデータ（HTML データ）を生成し，これを WWW ブラウザに表示する方式である。この方法では，ページ全体のレイアウトを保存しながら，ページ全体を翻訳表示できる。

(3) 擬似プロキシ翻訳方式　図 4 に示すように，WWW ブラウザとプロキシサーバ（インターネットと WWW ブラウザのデータの授受を仲介する）の間に擬似プロキシを配置し，プロキシサーバから送られてくる HTML データストリームから，文字データ部分を解析抽出し，母国語に機械翻訳するという方式である。特長は，英語画面をいっさい表示することなくブラウジングを行うことができる点である。ASTRANSAC for Internet$_{TM}$では，これら 3 方式すべてを実現しており，あらゆる要望にこたえられるようにしている。

4.2 インタフェースの特長

種々のユーザの要望に対応するため，多彩な表示形態・翻訳形態，カスタマイズ項目を用意している[2]。

(1) 画面表示方式　WWW ブラウザの画面を上下に二分割し，原文・訳文を同時表示する画面分割方式と，原文ページ・訳文ページを二つのブラウザ画面に表示する 2 画面表示方式，一つの WWW ブラウザ画面内に訳文・原文をパラグラフ単位に交互に示す交互表示方

図3．機械翻訳ブラウザの画面表示　元の英語ページとその翻訳結果の日本語ページが上下の 2 画面に分割表示されている。
Example of machine translation WWW browser display

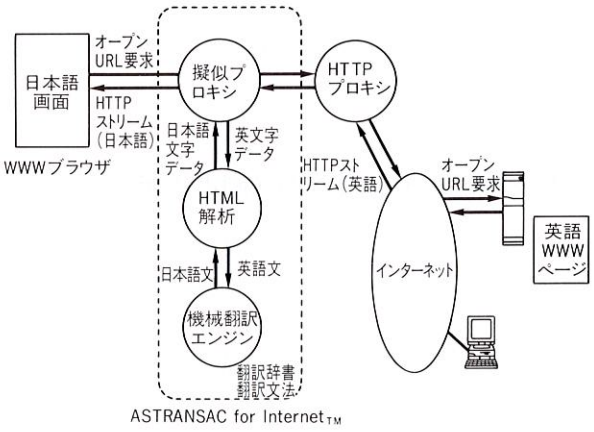

図4．擬似プロキシ方式の構成　擬似プロキシにより，テキスト部とデータ部を並行に分離し，英語ページを読み込みながら日本語に翻訳する。
System configuration of pseudo proxy method

式の3方式を提供する。

(2) 翻訳情報出力モード　文全体を翻訳し訳文を出力する標準翻訳モード，原文中の高校レベル以上の単語にだけ訳語を付与表示する単語翻訳モード，WWWページ中の単語とその訳語のリストを別画面表示する単語帳モードの三つの翻訳情報出力モードを提供する。

(3) 翻訳ハイパーリンク　原文と訳文の対応関係を保持し，訳文画面（もしくは原文画面）に埋め込まれたマークをクリックすることにより，対応する原文画面（もしくは訳文画面）を呼び出すことができる。

以上のほか，パソコン上のファイルを翻訳するファイル翻訳機能，ユーザの好みにより訳出方法を選択切換えできる翻訳環境指定機能（「ですます調／である調切換え」など37項目）などを提供している。

4.3 機械翻訳エンジン

従来の機械翻訳システムは，マニュアル文書の翻訳など，いわゆる産業翻訳における翻訳者の業務の支援を目標としてきた。最終目標として高品質翻訳を提供する情報発信のための翻訳システムである。これに対し，WWWブラウジングにおける翻訳システムは，文書の内容概要を把握する情報獲得のための翻訳システムである。

情報獲得のための翻訳では，①非常に多種多様な分野の文書が翻訳対象となる，②ユーザ自身による翻訳知識の追加（ユーザ辞書登録など）は期待できない，③リアルタイムで使用されるため非常に高速な翻訳処理がパソコン上で要求されるなどの条件での使用を考慮する必要がある。

①，②に関しては，広範な分野の文書に対する翻訳品質の向上がキーである。このため，広範なインターネット文書の用語をカバーするよう，従来，約8万語であった翻訳用辞書を，現在は約23万語にまで拡張しており，インターネット向けソフトウェアとして最大規模の辞書をもっている。また，③に関しては，高速翻訳を実現するため，翻訳ルールをコンパイルする方式を取り入れ，従来のエンジンに対して3倍以上の高速化を実現した。

5 自然言語処理の今後の課題と応用

情報洪水に対処するうえでは，自然言語処理の応用の一つである抄録・要約技術が有効である。現状のWWWページ検索エンジンにおいて，WWW文書の所在を示すURL（Uniform Resource Locator）の一覧表の代わりに，抄録または要約された情報を提示するだけで，大幅にHIは改善されることが期待できる。

ここでは，膨大な情報を効率的に伝達する手段に言語処理を導入するわけであるが，もっとも重要なことは伝達情報をいかに正確に伝えられるかである。先に述べたブラウザ翻訳システムの出力訳文は，斜め読みに使用するというレベルであるが，訳文の意味が不確実のときには，原文を参照することが許される。しかし，検索した情報を抄録または要約して提示する技術には，誤りが許されない。意味が正確に伝わらなければ，情報伝達の効率が低下ばかりでなく，重要な情報を見落としてしまうおそれがあり，精度の高い抄録・要約技術が要求される。

また，キーワード検索によって情報を入手するこれまでのプル型の利用形態は，ワールドワイドの情報が検索のたびに対象となり効率が悪く，欲しい情報をあらかじめ指定することでプッシュ型で情報が提供される仕組みが注目されている。このプッシュ型の情報提供には，言語処理を応用した精度の高い文書の類似性抽出，情報フィルタリング技術が要求される。

6 あとがき

ここで述べたソフトウェアエージェントFreshEye$_{TM}$の試用版は，インターネットで公開され，すでに500本以上がダウンロードされた。また，ASTRANSAC for Internet$_{TM}$は，パッケージソフトウェアや当社パソコンバンドルの形態で商品化されており，数万規模のユーザをもっている。今後は，テキストコーパス（大量文書）からの辞書自動作成[3]など，コーパス利用技術により，基本となる言語処理能力の飛躍的向上を図るとともに，インターネットを利用した翻訳サービスや情報提供サービスなど，新規分野への技術応用を行っていく。

文　献

(1) 住田一男，他：情報フィルタリング技術，東芝レビュー，51，1，pp.42-44（1996）

(2) 伊藤悦男，他：インターネット翻訳におけるユーザインターフェース，情報処理学会第53回全国大会（1996）

(3) 熊野　明，他：対訳特許文書からの機械翻訳辞書自動作成，東芝レビュー，50，1，pp.47-50（1995）

平川　秀樹　Hideki Hirakawa

研究開発センター 情報・通信システム研究所主任研究員。自然言語処理システムの研究開発に従事。ACL，情報処理学会，AI学会，言語処理学会会員。

Communication & Information Systems Research Labs.

住田　一男　Kazuo Sumita

研究開発センター 情報・通信システム研究所主任研究員。自然言語処理システムの研究開発に従事。情報処理学会，AI学会，電子情報通信学会会員。

Communication & Information Systems Research Labs.

武田　公人　Kimito Takeda

東京システムセンター システム開発部開発担当主査。自然言語処理システムの研究開発に従事。情報処理学会会員。

Tokyo System Center