情 報 メ デ ィ ア 21－3
（1995．5．12）

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

村田稔樹† 山本秀樹† 永田淳次†
† 沖電気工業(株) 関西総合研究所

世界的な規模のWWWサーバを人類共通の資産として広く活用することを考えると、言語の違いが情報発信、情報入手の障害になっている。そこで、本論文では、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について述べる。本システムの特徴は以下の通りである。

- WWWのデータの中のタグ情報を保持しながら翻訳するために、同じレイアウトの翻訳結果を得ることができる
- 翻訳結果のリンクから次の文書を検索できる。その際必要ならば翻訳する。
- 複数の人が同時に使用可能
- 事前蓄積翻訳方式により見た目の翻訳時間の短縮をはかっている
- ユーザは端末側のブラウザを一切変更せずに翻訳機能を利用できるようになる

本方式を利用することで、世界的な規模のWWWサーバを人類共通の資産として広く活用できるようになる。

The machine translation system for WWW users that enables to use the Internet without feeling language gap

Toshiki Murata† Hideki Yamamoto† Junji Nagata†
† Kansai Laboratory, Oki Electric Industry Co., Ltd.

Recently, several varieties of documents are distributed via electronic mail and WWW publishing, and to use the information effectively, machine translation systems are needed. However, conventional machine translation researches focus on translation algorithms and language modeling. This paper describes the machine translation system for WWW users, called W3-PENSÉE , that translates WWW data on the Internet. We proposed three types of W3-PENSÉE and pre-stored translation method that means W3-PENSÉE starts translating before a user orders and reusing the results translated previously by others. This system improves the effectiveness of the Internet browsing.

1 はじめに

ワークステーションやパソコンの普及によって様々な文書が、電子メールなどの電子メディアを通じて配布されるようになってきている。またインターネットに代表される計算機ネットワークの普及により世界各地の情報の入手や、世界中への発信が容易になってきた。これらのネットワーク上では、英語で情報発信されることが多く、そのことが英語圏以外の人々にとっては情報入手や発信の障害になっている。このような言語障壁の緩和のために計算機を使用する研究として、機械翻訳システムの研究や言語教育支援システムの研究が盛んに行なわれている。

これまでの機械翻訳システムの研究は、翻訳品質を向上させるための翻訳方式に関するものや、自然言語の現象のモデル化に関するものが中心であり、情報交換に用いられる電子メディアに対してどのように機械翻訳を用いるかといった視点からの研究は少ない。文献[1, 2]は、電子メールによる機械翻訳の利用技術の研究を行なっているが、この研究では電子メールは機械翻訳の起動方法として使用されているだけであり、電子メール上の情報の翻訳を目指した研究ではない。機械翻訳の研究においては、翻訳対象の性質を考慮した研究が重要である。

近年、インターネット上の電子メディアとしては、World Wide Web(WWW)が注目を集めている[3]。WWWは、分散型のマルチメディアハイパーテキストシステムである。このハイパーテキストには、他のハイパーテキストシステムや音声・画像・動画などのマルチメディアデータを指すリンクを埋め込むことができる。リンクは同一計算機(WWWサーバ)内のデータだけでなくネットワーク上に分散している他の計算機上のデータを指すことが可能である。他の計算機のデータに対するアクセスは、WWWサーバ間のプロトコルであるHTTP [4]だけでなくftp、Gopher, WAISといった他のインターネット標準プロトコルを使用できる。

WWWは、(1) WWWサーバに情報をおくことで、潜在的に全世界のインターネットユーザに向けて情報発信が可能であること、(2) Mosaic[1]をはじめとする端末側のソフトが、各種OS用に開発されていること、から利用者が増加の一途をたどっている。しかしながら現状のWWWサーバとブラウザだけでは、膨大な情報の中から必要な情報を見つけるための支援は行なわれていない。例えば、情報発信側は自分の日常使用する言語で情報発信できればより迅速に情報提供可能であるが、他の言語への翻訳を行なった後に情報を提供しようとすると時間的遅れや余分な労力が発生する。また、情報検索者はあるWWWサーバに必要な情報があることがわかったとしても、その情報が自分の母国語で以外の言語で記述されていると理解に時間がかかってしまう。

機械翻訳をWWWサーバの情報検索に使用できるようになると、情報発信、情報入手の時間が短縮され、世界的な規模のWWWサーバを人類共通の資産として広く活用することが可能になる。そこで、本論文では、WWWサーバの情報検索に使用するWWW用機械翻訳システムについて述べる。本システムは、通常のWWWの情報検索操作の際に、自然に機械翻訳を使用できることを目標とする。以下ではWWWサーバと機械翻訳の方式についてのべる。次にWWWの情報検索操作について考察し、それをもとにWWW用機械翻訳システムに必要となる機能を明らかにする。そしてその要求を満たすため3つの実現方法を提案する。さらに3つの中で実際に開発したW3-PENSÉE [2]システムとその評価について言及する。

2 WWWサーバと機械翻訳の概要

2.1 WWWの概要

世界中のWWWサーバはそれぞれ一意のドメイン名をもつ。そのサーバにおかれた文書は、プロトコル名、ドメイン名、ホスト名およびパス名によって世界中で一意に決まる名前、URL (Uniform Resource Locator)を持つ。利用者がブラウザから

[1] Mosaic はイリノイ大学の商標である

[2] PENSÉE は移植性の高い翻訳エンジンと文法・辞書からなる機械翻訳システムである。PENSÉEは、沖電気工業(株)、大阪ガス(株)および(株)オージス総研の登録商標。

URLを指定すると、ブラウザは直接または代理サーバ(proxy サーバ)を通じて間接的にその文書を持つサーバに対して送信要求を送る。その文書を持ったサーバは文書に、文書の種別などのヘッダつけた文書をブラウザに送る。ブラウザはその文書の種別に従って表示・音声の再生・録画の再生などを行なう。

WWWサーバに置かれている文書は、通常HTML (Hypertext Markup Language)[5]と呼ばれる言語で記述される。HTMLは、ISO のSGML (Standard Generalized Markup Language) によって定義されている。SGML は、構造化文書の型とそれを記述するタグ(マークアップ)を定義するためのものである。タグの中には、文書構造を示すものや他の文書へのリンクを示すものがある。WWWのブラウザを使用しているときに、他の計算機上の文書へのリンクを指定すると、上記の機構によって、ブラウザはその文書を入手する。

2.2 機械翻訳の研究

機械翻訳の研究では、 SGMLのようなタグを含んだ文書の翻訳方式についていくつかの研究がなされている[6, 7, 2]。それらの研究では、基本的にタグの部分は翻訳せずに、訳文にもそのタグを残すという方式をとり、翻訳の前後で図や表の位置など、文書のレイアウトを同じ状態に保つ。この方式に基づきワークステーション上のDTP (Desktop Publishing) システムとリンクしたシステムがすでにいくつか実用化されている[6, 2]。

また、近年個人の利用を考慮したパソコン用の機械翻訳システムが数多く実用化されている。それらは、汎用ワードプロセッサソフトや専用エディタを用いて会話的に翻訳する場合や、ファイル単位に一括して翻訳するために使用されている。会話的に使用する場合には、機械翻訳に必要な前編集や後編集を支援するための機能が使用できるようになっている。

別の研究として、機械翻訳をサーバにおき、その操作を電子メールを介して行なうシステムがある[1, 2]。これらのシステムはユーザ側の計算機が電子メール機能さえ持てば機械翻訳を使用できるという利点があり、多く利用されている。しかしながら、この種のシステムにおける電子メールの役割は従来GUIなどを使用して行っていた機械翻訳の起動を電子メールに変えただけであり、電子メールでの人間の情報伝達の支援を目指したシステムはなっていない。

3 WWW用機械翻訳の機能

3.1 WWW用機械翻訳に要求される機能

WWWサーバを検索しているユーザは、画面を眺めてみて、必要な情報がその文書にあるかそれともその文書の中のリンクが指す文書にあるのかを判断しているといえる。もし、その文書に必要な情報がある場合は、その文書を丁寧に読むがそうでない場合は、リンクをたどって次の文書を表示させるか、たどってきたリンクを後戻りし別のリンクをたどるか、あるいは全く別の文書のURLを入力するかである。ユーザが画面を見る際には、単に文字情報を読むだけでなく、画面にあらわれた図、文字の大きさやフォントが異なる部分、および他の文書へのリンクがある部分などを見ることによって文書を判断していると考えられる。

このような一連の操作の中で使用されるWWW機械翻訳に必要な機能について検討する。ユーザが指定した文書が外国語(例えば英語)の場合は、ユーザから翻訳要求があればそれをユーザの母国語(例えば日本語)に翻訳できることが必要である。その際、原文の持つ、図、文字の大きさやフォントの違い、および他の文書へのリンクといった文書構造を翻訳結果に忠実に保存することが重要である。

次に、ユーザが翻訳結果のリンクをたどることを考えると、翻訳結果のリンク先は、原文の対応するリンク先をたどったときと同一の文書にいくべきである。さらに、リンク先の文書がもし外国語の場合はそれをまた翻訳する必要がある。ユーザがリンク元の文書を翻訳した場合は、リンク先の文書もまた翻訳すると考えられる。操作を簡単にするには、翻訳結果中のリンク先の文書が、外国語の場合は自動

的に翻訳を起動するのが望ましい。

翻訳速度について考えると、ユーザが読みたいと思う文書はできるだけ速く翻訳できることが必要である。WWW用機械翻訳の場合、翻訳対象であるWWWサーバの文書はそう頻繁には変更されないと考えられる。また業務でWWWサーバを検索することを考えると複数の人が同じ文書の翻訳結果を要求することが予想される。いったん翻訳した結果を複数のユーザで共有できる機能があると翻訳速度に対する要求は満たすことができる。

上記の検討から、WWW用機械翻訳には、以下の機能が要求される。

3.2 タグ文書の翻訳機能

タグを含んだ文書を翻訳する機能である。本機能はタグの部分を翻訳せずに、それ以外の部分を翻訳し、翻訳結果に原文と同じタグを埋め込む。その結果翻訳の前後で文書のレイアウトを保存する。

3.3 言語自動判別機能

WWWのデータの記述言語(例えば英語か日本語かなど)を判別する機能である。システムは文書を翻訳すべきと判別すると翻訳を実行する。

3.4 事前翻訳機能

ユーザが読む必要であることを判断して翻訳操作をした後は、システムはできるだけ早く翻訳結果を出力できるのが望ましい。そこで、原文を表示した直後に翻訳を起動しておき、ユーザが読む必要があると判断すればすぐに翻訳中の結果を出力みせるようにする。すなわち、システムに対する翻訳を指定は、通常の機械翻訳のように翻訳機能の起動を意味するのではなく、翻訳結果の表示要求を意味するようにする。これは通常のWWWの検索と親和性が高い処理だといえる。

3.5 蓄積翻訳機能

WWWサーバのデータは頻繁に変更されるものではないので、一度翻訳した翻訳結果を蓄積しておき、翻訳の要求があったときは原文が修正されている場合だけ翻訳し、そうでない場合は翻訳しないようにする機能である。

4 実現方式

本章では、上記の機能を満たすWWW用機械翻訳システムの3つの実現方式について提案する。WWW用機械翻訳システムは、図1に示すように、基本的には機械翻訳システムをおく場所によって3つの型に分けられる。

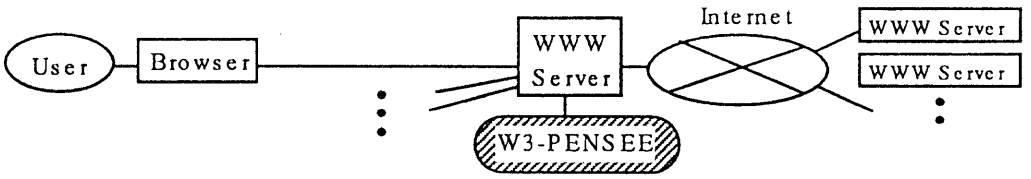
WWWサーバ型: WWWサーバ型(図1(1))は、WWWサーバ側に機械翻訳システムをおき、WWWサーバの外部コマンド実行機能(Common Gateway Interface:CGI)として翻訳を実行する。いったん翻訳機能を使用するとその後のWWWサーバのアクセスはすべてこの翻訳システムを経由するように、文書中のリンクをすべて書き換える必要がある。翻訳指定などの操作ボタンは、WWWサーバ型の翻訳システムからブラウザに送信する文書の先頭に追加することによって実現する。翻訳操作ボタンは、翻訳結果の文書へのリンクとして実現する。

通信路型: 通信路型(図1(2))は、WWWサーバとブラウザとの通信路に機械翻訳システムをおくシステムである。ブラウザとWWWサーバの間のすべてのデータはWWW用機械翻訳システムを中継することになる。したがってWWWサーバ型と異なりリンクの書き換えは必要なくなる。詳細は第5章で述べる。

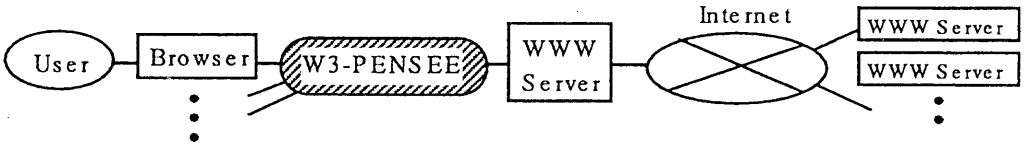
ブラウザ型: ブラウザ型のシステムはブラウザ側の計算機に機械翻訳システムをおく。ブラウザのメニューの中に機械翻訳の起動メニューができるようにブラウザと機械翻訳は密にリンクされる。(図1(3))

この3つのシステムの中では、ブラウザに依存しないことおよび機械翻訳の使用を意識させないことを考慮すると、通信路型のシステムが優れている。以下では、機械翻訳システムPENSÉEを用いて、Sun[3]ワークステーション上に開発した通信路型のシステムの実現方法について述べる。

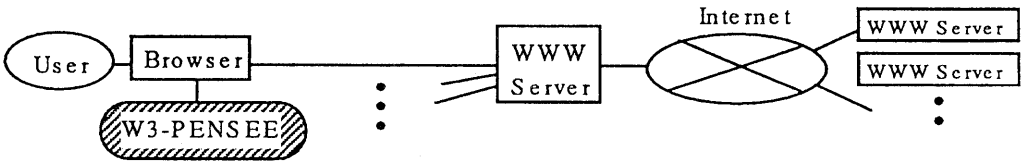
[3]Sunは、Sun Microsystems, Inc.の登録商標



(1): WWW サーバ側に機械翻訳をおいたシステム



(2): 通信路に機械翻訳をおいたシステム



(3): ブラウザ側に機械翻訳をおいたシステム

図 1: WWW 用機械翻訳:W3-PENSÉE の 3 つの実現方式

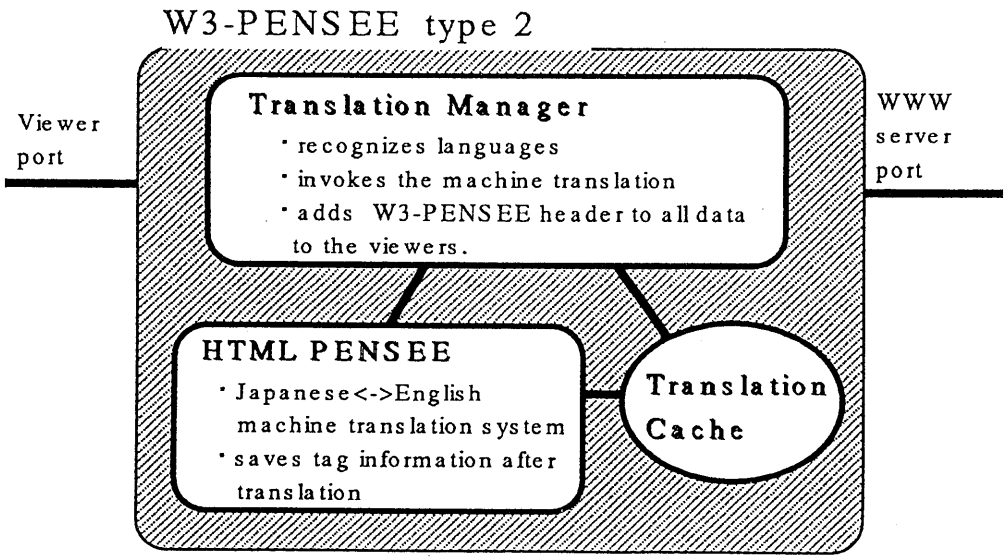


図 2: 通信路型 W3-PENSÉE の構成

5 通信路型システムについて

通信路型 W3-PENSÉE の構成を図2に示す。図3および以下に処理の流れを示す。

1. ユーザがブラウザを用いて、W3-PENSÉE に対して URL を要求する。
2. W3-PENSÉE は、要求されたデータを HTTP プロトコルを用いて WWW サーバから転送する。
3. 転送したデータの言語を判別し、必要ならば翻訳機能を実行し翻訳結果をディスク (図2の Translation Cache) に格納する。
4. 3 と並行して、指定されたデータに、翻訳結果表示指定のボタンを付与し、ブラウザに転送する。
5. ユーザがブラウザを用いて他の文書を指定した場合は、その文書に対して同様の処理を行なう。また、ユーザが翻訳結果表示指定を指定した場合は、翻訳結果をブラウザに転送する。

図4(1) に英語の WWW サーバのデータ例を、図4(2) にその翻訳結果を示す。(1) の文書の先頭部分は、本システムが付加したヘッダである。「翻訳」は、翻訳結果へのリンクになっている。(2) の文書の「原文」は、原文へのリンクになっている。

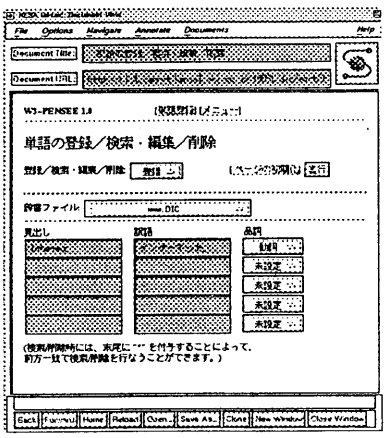


図 5: 単語登録画面

開発したシステムでは、翻訳機能だけでなく、ユーザ辞書の登録 / 検索・編集 / 削除機能も実装した (図5)。ユーザは、『単語登録』ボタンを押すことでユーザ辞書の登録 / 検索・編集 / 削除機能をブラウザ上で使用できる。ブラウザ上から登録した辞書を使って再度翻訳を実行するための『再翻訳』ボタンも実装している。

6 考察

WWW サーバの検索操作を考慮した WWW 用機械翻訳システムを UNIX [4]ワークステーション上に開発した。

6.1 WWW 用機械翻訳の利用形態について

従来の機械翻訳システムは、翻訳システムを使用する人と、その翻訳結果を利用する人が異なる場合が多かったと考えられる。そのため、より翻訳精度をあげるための前編集支援や後編集支援に力が注がれてきた。一方、WWW 用機械翻訳は、翻訳システムを起動する人が翻訳結果をその場で理解し検索を進めるといった利用形態をとる。すなわち、これまでの前編集支援や後編集支援を使用しない新しい利用形態の機械翻訳システムを開発したといえる。

6.2 ブラウザ依存性について

本システムは、HTTP を用いてサーバからデータを受けとり必要ならば翻訳を行ない HTTP でブラウザに翻訳結果を送信する。従ってブラウザに全く依存しない方式である。

しかしながら、実際に各種ブラウザを評価したところでは、 HTML2.0[8] の仕様を完全に満たしていないブラウザも見られた。ブラウザのバージョンアップは盛んに行なわれており、このような不具合を持つブラウザはなくなっていくと考えられる。

6.3 HTML、HTTP の仕様変更に対する対応性について

開発したシステムは、HTML2.0、HTTP1.0 に準拠している。HTTP の仕様が上位互換性を維持

[4]UNIX は、X/Open Company Limited がライセンスしている登録商標。

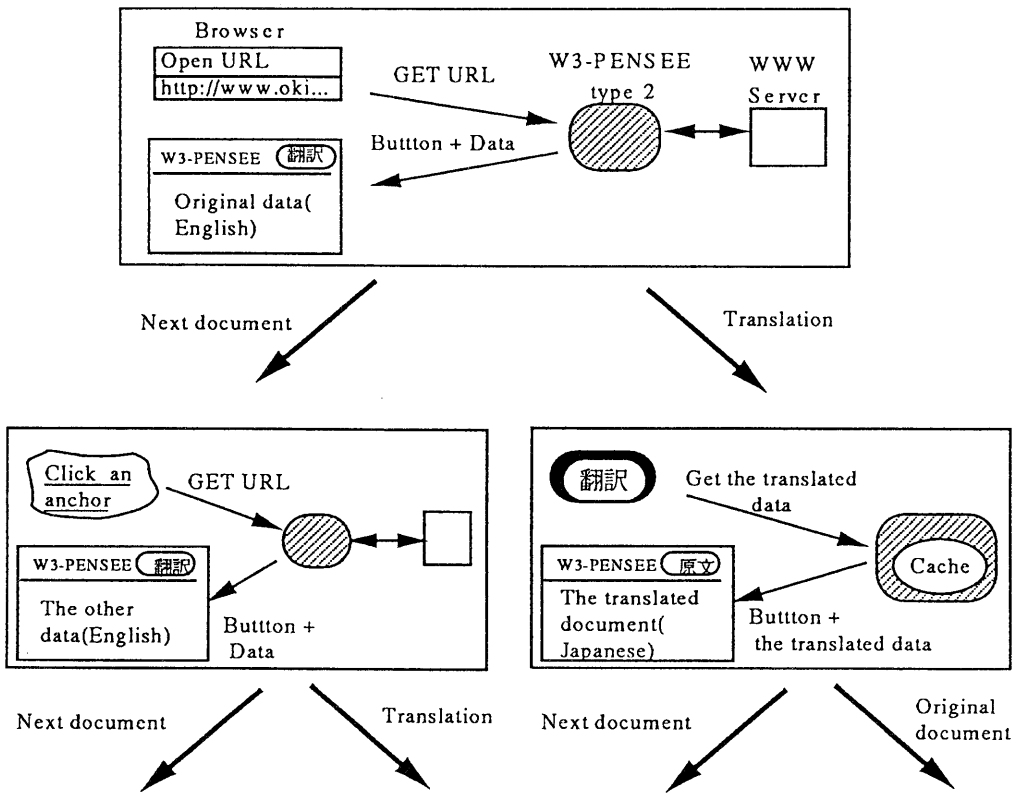
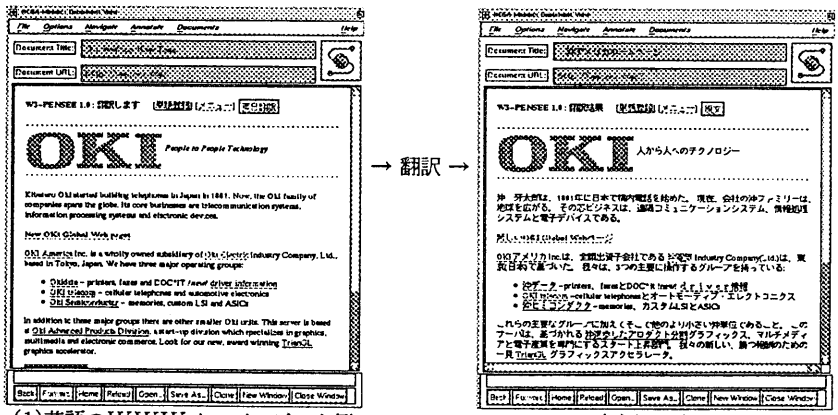


図 3: 処理の流れ



(1)英語のWWWサーバのデータ例　　(2)翻訳結果

[a]

[a]NCSA Mosaic は、 National Center For Supurcomputing Applications によって開発された。

図 4: WWW 用機械翻訳システムの翻訳結果

する限り本システムは変更なく使用できる。

6.4 翻訳速度について

蓄積翻訳機能により、一旦キャッシュされた翻訳結果を見る場合は、翻訳時間は0のように見える。WWWを日常業務として検索される場合を考えると、これは、業務の効率の向上に大きく寄与する。

6.5 機械翻訳からの方式の独立性について

本システムは、機械翻訳と文単位でインタフェースをとっているためどんな機械翻訳でも接続可能である。すなわち、機械翻訳の方式に独立している。

また言語についても日英、英日だけでなく他言語間の翻訳システムも接続可能である。対応する言語の数が増えても方式を変更する必要はない。すなわち、言語の数に対する拡張性を持っている。

7 おわりに

本論文では、電子情報メディアを用いた情報伝達形態を考慮することによって、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を提案した。本システムの特徴は以下の通りである。

- ユーザのブラウザのソフトウェアを一切変更せずに翻訳機能を利用できるようになる
- WWWのデータの中のタグ情報を保持しながら翻訳するために、全く同じレイアウトの翻訳結果を得ることができる
- 翻訳結果のリンクから次の文書を検索できる。その際必要ならば翻訳を行なう
- 事前翻訳方式、および蓄積翻訳方式により翻訳結果を得るまでの応答時間の短縮をはかっている

本方式を利用することで、情報発信、情報入手の時間が短縮され、世界的な規模のWWWサーバを人類共通の資産として広く活用することが可能になる。

今後の課題は、タグを含んだ文書の翻訳の精度向上、複数のWWW用翻訳システム間でのユーザ辞書の共有化などがある。

謝辞

WWWデータの使用を御快諾いただいた OKI America, Inc に感謝します。

参考文献

[1] 西野文人, 中村直人：機械翻訳電子メール, 情報処理学会研究報告(自然言語処理), Vol. 75, No. 5 (1990).

[2] 永田淳次, 山本秀樹：実用性が高くユーザフレンドリな機械翻訳システム, 沖電気研究開発, Vol. 62, No. 2 (1995).

[3] Berners-Lee, T., Callian, R., Luotonen, A., Nielsen, H. F. and Secret, A.: The World-Wide Web, *Communications of the ACM*, Vol. 37, No. 8, pp. 76 – 82 (1994).

[4] Berners-Lee, T., Fielding, R. T. and Nielsen, H. F.: Hypertext Transfer Protocol – HTTP/1.0 (1994), (Internet Draft).

[5] Berners-Lee, T. and Connolly, D.: Hypertext Markup Language (HTML) A Representation of Textual Information and MetaInformation for Retrieval and Interchange (1993), (Internet Draft).

[6] 伊藤悦雄, 武田公人, 平川秀樹, 天野真家：DTP形式情報を保存する機械翻訳システム, 情報処理学会第42回全国大会講演論文集, pp. 2C–10 (1991).

[7] 石川, 檜山：タグ付文書の英日機械翻訳支援システム, in *CALS JAPAN*, p. 794 (1994).

[8] Berners-Lee, T. and Connolly, D.: Hyper-Text Markup Language Specification – 2.0 (1995), (Internet Draft).