自然言語処理 108－22
(1995. 7. 21)

WWW用機械翻訳システム：W3-PENSÉE

村田稔樹† 山本秀樹† 永田淳次†
† 沖電気工業(株) 研究開発本部 関西総合研究所
〒540 大阪市中央区城見1-2-27 クリスタルタワー25階
Tel:(06)949-5101,Fax:(06)949-5108, Email: *mura@kansai.oki.co.jp*

あらまし 近年、インターネットに代表されるコンピュータネットワークの普及により、世界各地から情報が発信されている。その情報はほとんどが英語であり、日本語が母国語である日本ではその情報の扱いに不便を感じることが多い。そこで機械翻訳システムを用いることになるが、現在の機械翻訳システムの翻訳品質は、人間の行なう翻訳には及ばない。そこで必要となるのが翻訳システムの使用方法の研究である。我々は、World Wide Web(WWW)の情報の翻訳を対象に、ユーザモデルを検討した結果得られた知見をもとに、通信路方式、事前／蓄積翻訳機能、タグレイアウト保存翻訳を特徴とするWWW用機械翻訳システムW3-PENSÉEを開発したので報告する。

和文キーワード World Wide Web, WWW出版、 情報検索、 機械翻訳システム、 インターネット、 ユーザモデル

WWW machine translation system : W3-PENSÉE

Toshiki Murata† Hideki Yamamoto† Junji Nagata†
† Kansai Lab., Research & Development Group, Oki Electric Industry Co., Ltd.

Abstract Though the Internet population has been steadily growing, for the general public, using the Internet is not an easy endeavor because the majority of information is in English. This paper describes the machine translation system for WWW users, called W3-PENSÉE , that translates WWW data on Internet. We paid special attention to make this machine translation system as user-friendly and suitable for Internet surfing. The system separates the sentences from the HTML tags used in the original WWW texts, translates the sentences and returns each tag to the original point so the HTML link information remains unchanged. The translated data are then sent to a cache in the translation system in advance while the original data is sent to the client's viewer. When the user clicks the translation button, the system sends the translated data stored in a cache to the user's viewer so as to make the user feel as if they are seeing the realtime translation.

英文 **key words** World Wide Web, WWW publishing, information retrieval, machine translation system, The Internet, user model

1 はじめに

近年、インターネットに代表されるコンピュータネットワークの普及により、世界各地から情報が発信されている。その情報はほとんどが英語であり、日本語が母国語である日本ではその情報の扱いに不便を感じることが多い。膨大な情報から必要な情報を得るには、斜め読みのようにさっと読めることが必要となるが、やはり母国語でない言語では多くの場合困難である。このような状況において、コンピュータで翻訳を行なう機械翻訳システムを用いることは、自然な流れだといえる。

しかし、現在の機械翻訳システムの翻訳品質は、人間の行なう翻訳には及ばない。特に長文になると訳質が低下することが多い。

そこで必要となるのが翻訳システムの使用方法の研究である。これまでは、いかに翻訳品質をあげるかということや、また、翻訳を目的とする人をいかに支援するかということに注力して研究が行なわれてきたが、ユーザの目的や状況を考慮して機械翻訳システムそのものの使い方を研究する必要がある。

我々は、インターネット上の電子メディアとして注目を集めている World Wide Web(WWW) の情報の翻訳を対象に、ユーザモデルを検討し、そこで得られた知見を踏まえた WWW 用機械翻訳システムを開発したので報告する。

2 ユーザモデル

本章では、WWW[1]ブラウザのユーザが機械翻訳システムを使用する場合のユーザモデルを考察する。

機械翻訳システムを導入するといっても、さまざまなレベル・形態が考えられるが、本論文では、WWW ブラウザに表示されるものがすでに翻訳が行なわれており、あたかもすべて自国語でネットサーフィンしているかのように思えるものを想定する。

2.1 環境

ユーザはすでに WWW ブラウザを持っており、それを用いて Internet にアクセスしている。そこに機械翻訳システムを導入する場合、今までユーザが使ってきた WWW ブラウザやサーバはそのまま使えるのが望ましい。しかし、WWW ブラウザはさまざまなコンピュータ用に各種存在する。また、WWW サーバも各種存在する。WWW ブラウザやサーバを変更 (改良) して機械翻訳を連結するとなると、膨大な数のソフトを修正する必要があり、その手間は計り知れない。よって、WWW ブラウザにもサーバにも手を入れずに翻訳機能を実現することが望まれる。

[1]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り [1]、WWW ブラウザには文字・音声・画像・動画などのマルチメディアデータが表示・出力される。また、埋め込まれたリンクをたどることによって、他のデータを次々と呼びだして情報を得ることができる。

2.2 操作性

ユーザの目的は、WWW ブラウザで情報を得ることであり、翻訳することが目的ではない。翻訳することによって、情報獲得を容易にすることである。これまでの翻訳システムの利用方法は、翻訳が目的である状況のみを想定してきたが、それとは異なる。翻訳が目的ではないので、翻訳されていること自体が意識されないというのが理想である。以上のことから、WWW ブラウザ上に表示されたデータを翻訳するための操作は、必要ないかまたは非常に簡易でなければならない。

2.3 シーケンス

WWW ブラウザで情報を得ようとするユーザの操作シーケンスは以下のように考えられる。

1. ある情報を得ようとして、ある WWW サーバのデータをブラウザに表示
2. 得たい情報があるかないかを大雑把に判断
3. なければ、リンクをたどったり、他のページを探す (→ 2. へ)
4. あれば、熟読

上記 2. では情報があるかないかを調べているため、大雑把に読むことになる。また、得たい情報にいきつくまでは 2. と 3. を繰り返すため、読もうとするたびに時間のかかる翻訳が終了するのを待つとすると、操作効率が非常に悪くなる。したがって、翻訳処理は速くなければならない。

2.4 翻訳対象

翻訳対象は文書であるが、WWW 上のデータで文書の形態[2]をしているものは、多くの場合、HTML (Hypertext Markup Language)[2] で書かれたものか、プレーンテキストである。WWW では、マルチメディア文書や視覚に訴える文書が多い。イメージなどのマルチメディア情報とのリンクは HTML のタグを用いており、ハイパーテキストとしてのジャンプ機能も同じタグを用いている。ユーザは文書の中の文字列を読みとるだけでなく、イメージ情報、別文書へのリンクのある場所、強調されている部分、レイアウトから文書を理解していると考えられる。そのため、HTML 文書の場合は、HTML のタグを保存したまま翻訳する必要がある。また、プレーンテキストの場合はインデントや空行なども保存したまま翻訳する必要がある。

一方、WWW の文書は、ハイパーリンク先を示すためにメニューや箇条書きが多用される。そのため短い文や名詞句をうまく訳すことが要求されるが、現在の機械翻訳でもかなり良質の訳が得られる。したがって、WWW 文書は現在の機械翻訳のレベルに適した翻訳対象といえる。

[2]Content-Type: text/*

2.5 要求事項

以上から得られた要求事項をまとめると以下のようになる。

	要求事項
環境	使いなれたブラウザやサーバを用いたいので、ブラウザやサーバに依存しない形態にする
操作性	情報を得るためなので、翻訳の操作を意識させない
シーケンス	情報を得るまで繰り返しアクセスするので、高速に動作する
翻訳対象	レイアウトを重視した文書なので、タグやレイアウトを保存する

表 1: ユーザモデルから得られた要求事項

次章ではこれら要求項目を満たすための機能とその実現方法を述べる。

3 WWW 用機械翻訳システムに必要な機能と実現方法

3.1 通信路上で翻訳を行なう

WWW ブラウザやサーバに全く依存せずに翻訳するためには、ブラウザとサーバの通信路上で翻訳を行なえばよい。そこで、本システムでは、ブラウザから見ればサーバに見え、サーバから見ればブラウザに見えるような、ブラウザとサーバの間の通信路上に位置する一種の透過なサーバとして実現している。ブラウザ、サーバ間では、HyperText Transfer Protocol (HTTP) [3] というプロトコルを用いて情報を交換しているので、この HTTP のレベルで翻訳を行なえばよいことになる [4]。

3.2 簡易な操作と言語自動判別機能

WWW ブラウザのユーザにとって、画面上に表示されたボタンを押すことは非常に簡易な操作である。本システムでは WWW ブラウザの表示画面の最上部に翻訳するためのボタンを表示させ、そのボタンを押すことによって翻訳結果を表示することにした。ただし、ブラウザが要求した WWW のデータの型は、HTTP では Content-Type に記述されており、それが文書を示しているときのみ翻訳ボタンを付加する。

文書が母国語 (ユーザが要求している言語) で記述されていた場合は、実際には翻訳する必要はない。そこで、言語自動判別機能により翻訳対象言語かどうかを判別し、対象言語の時のみ翻訳ボタンを付加するようにした。この機能と後述の蓄積翻訳機能と組み合わせて、翻訳ボタンを押す操作をも省略できるようになった。

また、利用者の母国語で書かれた文書に対しては単に中継するだけとなり、無駄な負荷が発生しない。

3.3 リアルタイム翻訳を実現する方法

2.3 節で述べたように翻訳は速ければ速い方がよいが、従来の翻訳技術では、リアルタイム翻訳 (翻訳にほとんど時間がかからないこと) はまだ実現されていない。しかし、WWW 用機械翻訳に特化して工夫すれば、リアルタイム翻訳は可能になる。それを実現する仕組みが、事前翻訳機能、蓄積翻訳機能と、翻訳と表示の非同期処理である。

3.3.1 事前翻訳機能

WWW ブラウザで表示する文書には、多くの場合画像データが埋め込まれている。画像データはデータサイズが大きいため、インターネットを通じてすべてが送られてくるまで時間がかかる[3]。また、WWW ブラウザに表示されてから、ユーザが何かを WWW ブラウザに指示するまで、いくらかの時間が過ぎているはずである。これらの時間を有効活用する方法として事前翻訳機能を提案する。

WWW ブラウザに対して、ユーザが表示したい URL [4]を入力すると、ブラウザは本システムにその URL の指しているデータの要求を行なう。本システムは、そのデータを WWW サーバから得る。その後のデータの翻訳処理方法は 3 通り考えられる。

1. 得られたデータを翻訳し、翻訳結果を WWW ブラウザに送る
2. 得られたデータをそのまま WWW ブラウザに送り、次にそのデータの翻訳結果が要求されれば、翻訳を始め、翻訳結果をブラウザに送る
3. 得られたデータを WWW ブラウザに送りながら、翻訳も始める。次にブラウザから翻訳結果の要求がなされれば、翻訳結果をブラウザに送る。

1. は、翻訳結果がすべて送り終らなければ表示されないので、翻訳が終了するまで待たなければならない[5]。

2. は、1. と同様、結果を要求してから翻訳を始めるため、翻訳が終了するまで待たなければならない。

3. は、画像の表示時間やユーザが翻訳要求をするまでの間に翻訳が進むので、翻訳要求した時点で

[3]最近は、すべてが送られてこなくても、送られてきたものから徐々に表示される場合もある。しかし、ほとんどの場合、ユーザはその画像が送り終られるまで待っているようである。

[4]Uniform Resource Locator の略。文書を世界中で一意に決めるための名前。

[5]最近は、データを送り終らなくても順次表示できる WWW ブラウザもあるが、その場合でも実際の翻訳の速度はかかるため、待つ必要がある。

は翻訳が終了しており、要求されると同時に翻訳結果をブラウザに送ることができる。つまり、ユーザから見るとリアルタイム翻訳が行なわれたことになる。

そこで、我々は 3. を採用し、この機能を事前翻訳機能という。

3.3.2 蓄積翻訳機能

WWW での文書は一度作られるとあまり修正されない場合が多い。そのため、通信の負荷を下げるため、一度得たデータはキャッシュしておくことが多い。

翻訳でも同じように、一度翻訳した結果を保存しておく。つまり、前節で WWW ブラウザから要求されたデータを翻訳する際、翻訳結果を保存しておく。次に、また WWW ブラウザからある URL のデータが要求された場合、その URL に対応する翻訳結果が保存されていないかどうかをチェックする[6]。もしあれば、その翻訳結果をブラウザに送る。すると、ユーザから見るとまさにリアルタイム翻訳が行なわれたことになる。この機能を蓄積翻訳機能という。

この機能を使うと、翻訳ボタンを押すことなく、すぐに翻訳結果を表示することができる。例えば、要求言語が日本語のユーザの場合を考える。もともと日本語のページは、そのまま日本語で表示され、英語のページでキャッシュされたページもすぐに日本語で表示されるとすると、まさに自国語にリアルタイム翻訳された世界に入ることができる。

また、本システムは複数の WWW ブラウザ (ユーザ) が接続できるため、複数のユーザで翻訳結果を共有することができる。同じサーバを使用しているグループで誰か一人でもある URL を見れば、そのページはその時点で翻訳され、他の人はすぐに翻訳結果を得ることができる。同じページをよく見る人が集まっているグループなら、そのヒット率は高くなるため、蓄積翻訳機能の効果は大きい。

3.3.3 翻訳と表示の非同期処理

事前翻訳機能と蓄積翻訳機能を用いると、ほとんどの場合でリアルタイム翻訳を行なうことができる。しかし、初めて要求される URL で (翻訳結果がキャッシュされていない)、なおかつ、ユーザがすぐに翻訳結果の要求を行なった場合や、その文書が大きい場合に問題が生じる。つまり、翻訳が終了する前に翻訳結果の要求が行なわれた場合である。この場合に、翻訳が終了するのを待って、翻訳結果を WWW ブラウザに送るとすると、大きい文書の場合大変待たされる。

そこで本システムでは、翻訳と表示を非同期に行なうようにした。翻訳を始めるとその結果を次々とファイルに書きだす。WWW ブラウザから翻訳結果の要求があれば、その時点でのその翻訳結果のファイル内容をブラウザに送る。ブラウザは翻訳が終了するのを待つことなく、翻訳結果を得ることができる。

このようにすれば、文書の大きさによらず、翻訳要求があった時点ではある程度の翻訳が行なわれているため、WWW ブラウザの 1 画面分は翻訳が終っており、多くの場合、リアルタイム翻訳が実現される。もし、翻訳要求が 1 画面分翻訳が終る前になされても、ユーザの得たい情報 (たとえば、他の文書へのリンクなどのメニュー) が文書の先頭に書かれておれば、ユーザの要求が満たされるので翻訳の続きを待つ必要はない。

3.4 タグやレイアウトを保存した翻訳

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 [5, 6, 7]。従来の研究は、スペースとタグの混在についてあまり考慮されていない。

本システムでは、タグは翻訳せずに、それ以外のところを翻訳する。さらに、タグとレイアウト (空白やインデントや改行など) を両方保存しながら翻訳するタグレイアウト保存翻訳機能を持っている。本来、タグつき文書に含まれる空白や改行は意味を持たないので、無視しても構わないが、HTML の場合は、<PRE> タグのようにそのタグ内ではフォーマットを保存しなければならないものがある。したがって、本機能で翻訳すると、タグつき文書でもプレーンテキストでも翻訳する前と同じレイアウトを持った翻訳結果が得られる。これは、後で人が見て参考にしたり修正する場合にも非常に有効である。

また、タグによってはその範囲内を訳したくないもの (例えば、<ADDRESS> など) があるため、タグ名とその範囲を解析し、翻訳するかしないかも指定できるようにした。また、翻訳の際に不必要になる改行タグ (
 など) もヒューリスティックにより削除する。

4 WWW 用機械翻訳システム: W3-PENSÉE

我々は、機械翻訳システム PENSÉE [7] と前章で述べた機能とを組み合わせ、Sun[8] ワークステーション上に WWW 用機械翻訳システム: W3-PENSÉE を開発した。

構成を図 1に、使用画面を図 2に示す。

図 2(1) のように、通常の表示の最上部に W3-PENSÉE の翻訳ボタンを追加している。そのボタンをユーザは通常の操作と同様クリックすることにより、図 2(2) のような翻訳結果を得ることができる。

処理の流れは以下の通りである。

[6] 同時に元文書が更新されていないかもチェックする。

[7] PENSÉE (パンセ) は、沖電気工業株式会社、大阪ガス株式会社及び株式会社オージス総研の登録商標

[8] Sun は、 Sun Microsystems, Inc. の登録商標

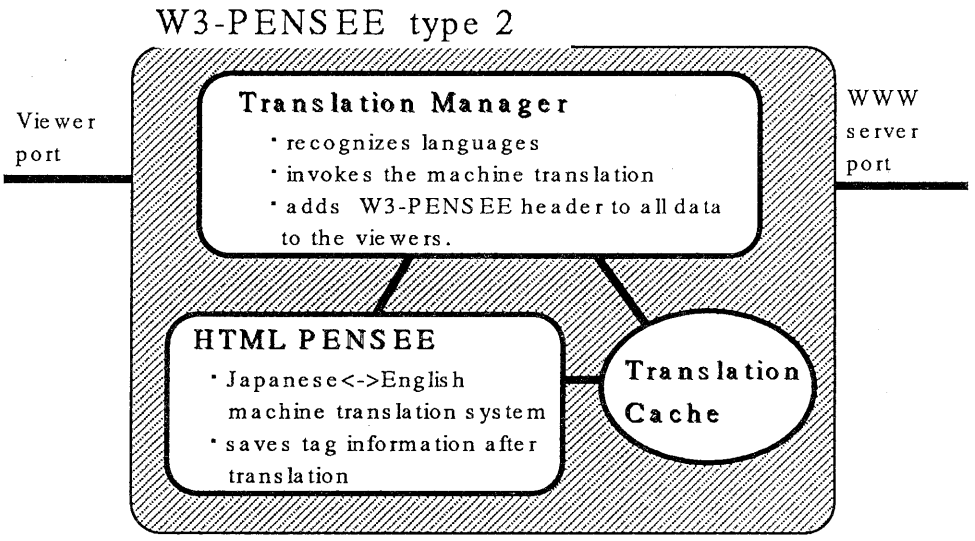

図 1: W3-PENSÉE の構成

(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 2: W3-PENSÉE の使用画面

1. ユーザが WWW ブラウザを用いて、W3-PENSÉE に対して URL を要求する。
2. W3-PENSÉE は、要求されたデータがディスクにあるかどうかをチェックし有効であれば、その内容をブラウザに送る (蓄積翻訳機能)。
3. なければ、W3-PENSÉE は要求されたデータを HTTP プロトコルを用いて WWW サーバから獲得する。
4. 獲得したデータの言語を判別 (言語自動判別機能) し、必要ならば翻訳機能を実行し (事前翻訳機能)、翻訳結果をディスク (図 1の Translation Cache) に格納する (蓄積翻訳機能)。
5. 3 と並行して、獲得したデータに、翻訳ボタンを付与し、ブラウザに転送する。
6. 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、翻訳結果の要求の場合は、ディスクから対応する翻訳結果の現時点での内容をブラウザに転送する (翻訳と表示の非同期性)。

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 3)。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 WWW 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。

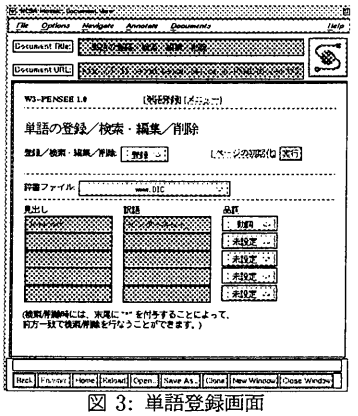

図 3: 単語登録画面

5 まとめ

本論文では、World Wide Web(WWW) の情報の翻訳を対象に、ユーザモデルを検討し、そこで必要となる機能を示し、それらを満たす WWW 用機械翻訳システムについて述べた。

本システムの特徴は以下の通りである。

- ユーザが使いなれた WWW ブラウザやサーバをそのまま使えるように、通信路上で翻訳を行なう。
- 翻訳のための操作は、翻訳ボタンを押すだけである。また、言語自動判別機能と蓄積翻訳機能により、翻訳結果がキャッシュされている場合は翻訳ボタンも押す必要がなく、翻訳の操作をほとんど意識する必要がない。
- 事前翻訳機能、蓄積翻訳機能、翻訳と表示の非同期性により、リアルタイム翻訳を実現している。
- タグレイアウト保存翻訳機能により、全く同じレイアウトの翻訳結果を得ることができ、翻訳結果から次の文書へリンクも行なえる。

本システムは、WWW 用機械翻訳に特化したもので、従来の方式では引き出せないリアルタイム翻訳を実現することができた。

本システムの開発によって、適用場面により機械翻訳をどのように使用すればよいかという研究が非常に重要であることがわかった。また、評価方法についても応用事例に応じた方法が必要であることがわかった。例えば、訳質に関して WWW 文書の場合は、短い文が不得意で長い文が得意であるよりは、短い文が得意で長い文が不得意の方が高い評価を与える必要があるなど、翻訳対象文書によってそれぞれ評価基準を設ける必要がある。

今後の課題は、上記のことを考慮した本システムの評価である。

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] 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).

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

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

[4] 村田稔樹, 山本秀樹, 永田淳次 : 快適なインターネットサーフィングを支援する WWW 用機械翻訳システム, 電子情報通信学会技術報告 (OFS), Vol. 95, No. 6, pp. 31 – 36 (1995).

[5] 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).

[6] 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).

[7] 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, pp. 23 – 26 (1995).