グローバルコミュニケーションを担う機械翻訳技術

# 2．翻訳ソフトウェアを取り巻く技術

　翻訳ソフトウェアが日本で発売されたのは，今からおよそ15年前。そのころは産業翻訳にしか利用されませんでしたが，近年のインターネットの普及とパソコンの高性能化により，一般家庭でも使用されるようになってきました。価格も当初の数百万円から1万円程度へと身近になり，今や翻訳ソフトウェアは，ネットサーフィンには欠かせないツールになりました。

　ここでは，インターネット翻訳を中心に機械翻訳を使用したソフトウェアを作るための技術を紹介します。


## インターネット翻訳

　翻訳ソフトウェアをもっとも身近に感じるのは，ネットサーフィンで英語のページを日本語で読むときでしょう。英文のページそのままのレイアウトで日本語で読むことができるので，内容が理解しやすく，ネットサーフィンの楽しさが増してきます。

　このように，レイアウトを保ったまま日本語で英文ページを読むことができるために必要な技術を紹介しましょう。

## プロキシ方式と直接方式

　インターネット翻訳ソフトウェアの仕組みは大きく二つに分けられます（図１）。

　第一の方式は，“ プロキシ方式 ”です。これは，インターネットからホームページを受信する際，ブラウザに表示する前に翻訳ソフトウェアが割り込み，データを翻訳してからブラウザに表示する方法です。ブラウザに表示する前に日本語にしてしまうので英語のページであることを意識せずに読むことができます。この方法では，ページの取得と翻訳を並行して行うため翻訳時間が気にならない，プロキシが使用できるブラウザであればどんなブラウザとでもいっしょに使用できるという利点があります。その反面，ページ全体の翻訳になり，欲しい部分だけの翻訳ができないという欠点もあります。

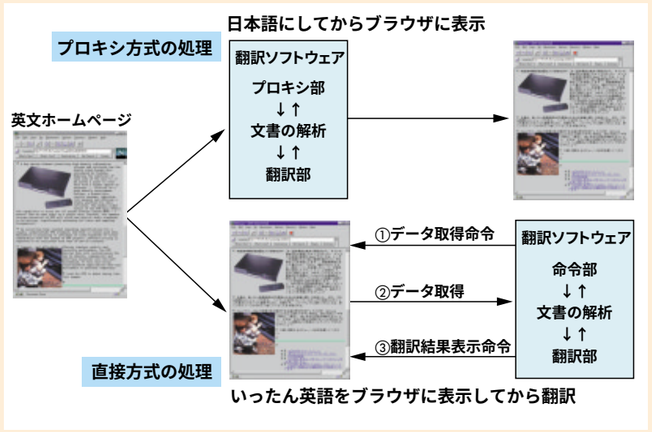



図１．プロキシ方式と直接方式　　インターネット翻訳には二つの方式があり，それぞれ特徴があります。

　もう一つの方式は，“ 直接方式 ”です。この方式は，ブラウザにいったん表示したデータを翻訳するものです。ブラウザにメッセージと呼ばれる命令を直接出して，データを取得して翻訳し，結果もブラウザに直接命令を出して表示するために，この名前があります。この方式では，翻訳したいページやページ中の翻訳したい部分だけを選んで翻訳できるという利点があります。更に，この方式では直接ブラウザに命令するために様々な処理が可能であるという特長があります。したがって，ブラウザに翻訳ボタンを埋め込んだり，ブラウザ上に文書を翻訳する領域を作成したりすることができるなど，インタフェースが優れた翻訳ソフトウェアとすることができます(図２)。しかし，この方法では，ブラウザごとに命令が異なるため，限られたブラウザでしか使用できないという欠点があります。

　The翻訳$_{TM}$シリーズでは，それぞれの長所を生かすため両方の方式を使用できるようにしています。

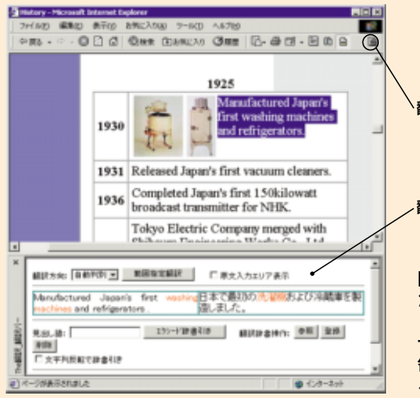
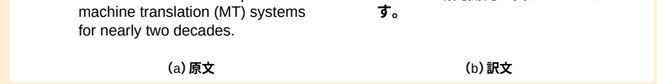



図２．ブラウザ上に組み込まれた翻訳ボタンと翻訳領域
　直接方式を使用すると，ブラウザ上にボタンを埋め込んだり翻訳領域（画面下部）を作成して翻訳することができます。

## 文書の解析と復元

　上記のいずれの方法を採用した場合においても問題となるのが，レイアウトの解析と復元です。インターネットの情報は，HTML（Hyper Text Markup Language）形式で記載されています。この形式は，文書のレイアウトをタグで指定しています（図３）。したがって，翻訳すべき“ 文 ”と翻訳しない“ タグ ”の情報を分離し，翻訳後に復元する必要があります。原文と訳文では語の順序が異なる，一つの語句が複数の語句になるなどの問題があります。この問題の解決のもっとも簡単な方法は，タグまでの部分をひと固まりとして翻訳することですが，The翻訳$_{TM}$では，更にタグに囲まれた部分が訳文のどの部分に対応しているかを調べることによって，文がこま切れにならないようにしています。



| | |
|---|---|
| <HTML><br><HEAD><br><TITLE><br>The HON-YAKU HOME PAGE<br></TITLE><br></HEAD><br><BODY><br><H1>The Hon-Yaku</H1><br><A HREF="www.toshiba.co.jp"><br>Toshiba</A> has been involved in the research and development of machine translation (MT) systems for nearly two decades. | <HTML><br><HEAD><br><TITLE><br>The翻訳ホームページ<br></TITLE><br></HEAD><br><BODY><br><H1>The 翻訳</H1><br><A HREF="www.toshiba.co.jp"><br>東芝</A>は，約20年間機械翻訳システムの研究開発に関わっています。 |
| (a) 原文 | (b) 訳文 |

図３．HTMLデータの例　　ホームページはHTMLで記載され，翻訳後もHTMLのタグを復元しなくてはなりません。

## WordやPDFへの応用

　The翻訳$_{TM}$では，ホームページの他にWordの文書やPDF（Portable Document Format）データもレイアウトを保ったまま翻訳することができます。これらの文書の場合も直接方式と同様に，WordやAcrobat[注1]に対して文書取得／復元の命令を出すことにより翻訳を実現しています。更に，この直接方式の応用は広く，マウスで指した文を翻訳する“ クイック翻訳 ”機能も同様の方法で実現しています。

（注１）　Acrobatは，アドビシステムズ社の商標。

伊藤　悦雄
e-ソリューション社　府中e-ソリューション工場
コンピュータソフトウェア部主務