インターネット翻訳ソフトの現状と将来展望

宮澤信一郎

▶1　インターネット翻訳ソフト

インターネットの普及に伴い国際コミュニケーションの重要性はますます増大している。その中で機械翻訳（Machine Translation，以降MTと略）に対する需要と期待が増している。インターネット翻訳ソフトとはWWW翻訳，電子メール翻訳，チャット翻訳などのインターネットに関わるMT用ソフトのことである。この中ではWWW翻訳が中心であり，製品数も他のインターネット翻訳ソフトに比べて圧倒的に多い。そのため本論ではWWW機械翻訳を中心に述べる。

WWW機械翻訳は通常のテキストを対象としたMTとは異なり，次のような点が出来なければならない。

・HTMLのタグとテキストの区別
・HTMLのタグの意味を理解し，タグに応じた翻訳
・JAVAやJAVA Script，XMLなどの各種のプロトコルに対応
・次々に生まれる新しいWWW技術への対応
・ネットサーフィンの性格上，次から次へとネットサーフィンを行うための迅速な翻訳

日本で販売されている翻訳ソフトの数は次表のようになる。これは同じメーカの同じシリーズのソフトでもスタンダードやフルパックなどの違いで製品価格が異なる場合は１個と数えている。大多数が英日・日英であるが，イタリア語，韓国語，スペイン語，中国語，ドイツ語，フランス語，ポルトガル語，ロシア語用などもある。

	Windows版	Mac版	Linux版
英日翻訳ソフト	77	17	3
日英翻訳ソフト	45	14	1

（http://www.bekkoame.ne.jp/~oto3/から集計）

▶ 2　インターネット機械翻訳の機能と評価

2-1　機能項目

市販の製品を分析してみると，MTソフトには次のような機能があることが判明した。

⑴　翻訳機能

・翻訳方法：　訳文修正，自動巡回翻訳，範囲指定翻訳，オフライン翻訳

「自動巡回翻訳」は，あらかじめURLで指定された複数のページを自動的にキャッシュに取得して翻訳を行う機能である。インターネットの接続コストを削減できるとともにブラウザへの描画が早くなる。「オフライン翻訳」とは，インターネットへアクセスして取得したデータをキャッシュデータとしてディスクに保存させておき，オフラインの状態で翻訳を行なうものである。

・翻訳対象形式による翻訳機能：　HTML，XML，PDF，ロータスノーツ，ニュース，メール，テキスト

このうち「PDF翻訳」はアドビのアクロバットのPDFファイルを翻訳する機能である。「ロータスノーツ翻訳」はロータスノーツで作成された英文メールを翻訳する機能である。

⑵　辞書機能

・辞書ツール：　辞書編集，辞書追加，辞書削除，辞書マージ
・単語関係：　単語登録，品詞登録，品詞種類，訳語活用形設定，格助詞設定
・辞書種類：　専門用語辞書，共通辞書

「共通辞書」はアジア太平洋機械翻訳協会が推奨する，異なった機械翻訳ソフトのユーザ辞書を交換するための，共通フォーマットに対応した辞書である。

⑶　その他の機能：　英文メール作成支援，OCR，ホットキー，音声出力，音声認識，アドイン，訳振り

このうち「英文メール作成機能」は，英文のメールを作成時に項目を選んだり必要な情報を入力していくことで自然なメールが作成される機能である。日本語の質問に答えていくだけで，定型の英文メールを手軽に作成，メーラにワンタッチで転送できる。「ホットキー機能」は，自らキーを設定しキーボード上からワンタッチで実行できる機能である。「アドイン機能」はワープロソフトに組み込むことで，そのワープロソフトに翻訳機能を組み込むことができる。「訳振り」は全体の翻訳は行なわずに，原文中の主要な単語のみに訳語を付与する（原文の単語の下部が多い）機能である。辞書引きを人間に代わって行なう機能といえる。

2-2　機能評価

MTソフトの性能はまだ期待されているレベルには至っていない。MTの発展のためにはその評価が重要である。筆者らはインターネット用MTソフトの機能評価を行った[1]。対象にしたMTソフトは英日の13製品であり，1999年４月現在市販されているインターネット用とうたっているMTソフトについて，ほぼ全製品を網羅している。

機能項目につき13製品の評価を行った。その結果，各製品で機能にかなりの開きがあ

ることが判明した。また本研究で様々な知見が得られたが一部を述べると次のとおりである。

⑴ 翻訳機能

① 翻訳方法

翻訳方法はいくつかのパターンに分かれそれぞれ長短がある。またインターネットには様々な形式の翻訳対象が増えているが十分対応できていない。

・訳文修正・範囲指定翻訳・オフライン翻訳

全ソフトが可能であるが，実際の操作面では様々なバリエーションがあり，それぞれ長短がある。

・自動巡回翻訳

半数ほどのソフトが出来ないが，今後，自動的に取り込んで自動的に翻訳するこの機能は，国際コミュニケーションの増加に伴って重要性を増すと思われる。

② 翻訳対象形式

インターネットには様々な形式の翻訳対象が増えているが，十分対応できていない。特に次の翻訳対象が重要である。

・PDF…WWW上に多くある。

・チャット翻訳…直接のコミュニケーションが増えると思われる。

・メール翻訳…一々コピー＆ペーストしなくてもメーラー上で翻訳できる機能である。

⑵ 辞書機能

辞書への品詞や訳語の登録に関しては，分かりやすく工夫する必要があるソフトが見られる。翻訳精度をあげるために専門辞書の充実も重要である。またユーザ辞書共通フォーマットに対応した製品が少ない。MTソフトはユーザが辞書を充実させていくことにより翻訳精度が上がっていくので，メーカ間の辞書の互換性は重要である。

⑶ その他の機能

英文メール作成支援機能は，例文群があり，ヒットしたキーワードの英文が出力されるので，必要に応じて修正して使用する機能である。英文はネーティブの正しい英語であるので，この方向での機能の充実は有望である。しかし例文の数が問題である。専用のソフトで最大のものは５万文例ほどであり，これでも少ない。50万文例～500万文例位あると相当実用的と思われる。場合によっては機械翻訳ソフトより有用になる可能性がある。

▶ 3　インターネット翻訳ソフトの性能評価

テキストに関する機械翻訳の評価の研究は数多く行なわれている。しかし需要が多いWWW用機械翻訳ソフトの全体的な評価の研究は，その発展のために重要であるにもかかわらず，筆者等の知る限りなされていない。今後，国際社会のコミュニケーションの中心的な存在になり，ますます重要性を増しているWWWの機械翻訳の研究と評価は，重要な問題である。WWW機械翻訳評価の標準化が強く望まれる。

3-1　WWW機械翻訳の評価

テキスト以外のWWWに固有な部分を中心に評価を行なった[2]。この評価方法は特定の言語に依存しない汎用性のある評価方法を提供するものである。これらのWWW特有の機能を持った11のホームページ（HP）をインターネット上で探しテストページとした。

翻訳に使用した機械翻訳ソフトは６製品であり，総て英語から日本語への翻訳である。頁数の制約があるので，翻訳結果に対する分析と評価の一部のみを以下に簡単に示す。

⑴　ページ設定関係

翻訳結果においてホームページのタイトルが英語のままであったＭＴソフトは３つであった。他の３個のソフトは翻訳出来ていた。HTMLのタグからタイトルを把握することは可能であるので，翻訳すべきである。

⑵　フォント関係

２つのソフトで翻訳結果にアンダーラインが再現されなかった。アンダーラインが入っている部分は把握できるはずである。

英語のフォントが変更されている箇所で，翻訳結果のフォントが対応していない場合が多い。フォントの変更は意味があって行なっているのであるから，翻訳結果でもフォントを変更した方が良い。言語が異なるので，英語と同様のフォントが無い場合は，その言語に合った何か一般的な別のフォントに変えるだけでも効果的である。

文中でフォントサイズが変化した時に，翻訳結果もサイズが変化したのは３つのソフトだけである。文字サイズの変更はタグを見れば分かるのであるから，変更した方が良い。

フォント関係全体についての提言であるが，フォント関係の情報は翻訳結果に是非反映してほしい。ただし，英語で１つのまとまった語句が翻訳結果では２つ以上に分離してしまうことがあるので，機械翻訳ソフトの翻訳エンジン内部の改良が必要になるかもしれない。

⑶　イメージ関係

全ソフトとも文字画像部分の翻訳を行なっていない。文字画像などのイメージ形式の文字はホームページには非常に多いので，これが翻訳できないと翻訳不能部分が多くなってしまう。イメージ形式の文字の翻訳は，WWW機械翻訳において本質的な問題である。OCR機能を組み込んで認識させることも一つの方法である。またホームページ作成者の方で，記述するソースにイメージ形式文字に相当するテキストのコメントを書くようにすれば，翻訳が可能になる。またホームページ作成者の手間を省くために，自動的にこれらコメントが記述されるようなWWWプロトコルの研究も必要である。

ボタン名は全製品とも英語のままである。しかしHTMLのソースの中にはボタンの名前が記述されている。ソースの中の"VALUE="の部分を把握できれば，翻訳できる可能性がある。

PDF中の文字の部分は全ソフトとも英語のままである。しかしPDFの中には文字部分のテキストがあるから，機械翻訳ソフトの方でPDFのプロトコルに対応できるようにすれば翻訳できるようになる。

⑷　ハイパーリンク関係

すべてのソフトが次のようなリンク名を各々のリンク名毎に訳すことが出来なかった。

Yellow Pages Maps Farefinder Reservations by Preview Travel Related Books

現象としては次のようなことが起こっている。①翻訳されていないリンクがある。②リンクの順番が変わる。③１個のリンクが２つになっている。④逆に２つのリンクが１つになっている。⑤複数のリンクが１つの文になっている。これらは，<A> </A> で囲まれた句の連続を誤って１文として解釈したものである。リンクのタグ（<A HREF> と</A>）をうまく認識してその範囲内で翻訳することが出来るはずで，機械翻訳ソフト側の工夫が必要である。HTMLの問題点として，"Yellow Pages"，"Maps"，"Farefinder" 等が，独立した項目として(箇条書きのように)並んでいるということを HTMLタグで 論理的に記述できないという点があげられる。その結果，実際に１文である場合と区別できないという現象が起こりうる。

⑸ リスト関係

リストの翻訳結果では，頭文字が大文字の単語が訳されていないために，リストの半分位を訳せないソフトがある。これは頭文字が大文字のものは固有名詞とみなして訳していないものと思われる。このような単語はリストでは多用されるので，半分ほども訳さないのでは，機械翻訳ソフトの効果が半減する。もしもユーザがおかしいと思えば，原文に当たれば良いのであるから，訳した方が良いと思われる。また固有名詞を訳して失敗している例もある。総て大文字のものは訳さない方が良いかもしれない。
（原　文）UH Admissions　（UHはUniversity of Hawaiiの略称である）
（翻訳例）あの，承認　（"あの"は日本語で指示代名詞である）

リストが体言止めされていない例もある。リストは一般的に体言止めされた方がよい。

１つのソフトを除く５ソフトにおいて原文にあるような行頭記号 "・" が無くなっている。これはHTMLのタグ(<UL><LI>・・・・・・</UL>) に機械翻訳ソフトが対応できていないことを意味する。

⑹ テーブル関係

表中の文は体言止めが一般的である。表の大部分は体言止めで翻訳されているが，一部で体言止めが行なわれていない。

⑺ フォーム関係

・フィールド型選択リストで，選択リストが１つのソフトは英語のままであり，他のソフトは翻訳を行なっている。各項目が体言止めに翻訳されているのに，１項目のみ体言止めになっていないソフトが２つあった。同一の選択リスト内のことなので，体言止めにするかしないか統一した方が良い。

⑻ その他のWWW特有の問題点

フレームは４つのソフトが再現できたが，２ソフトではフレームがそれぞれ別ページに印字されてしまった。その内，１ソフトではフレーム中の地図が再現されなかった。

アクセスカウンターは全ソフトが再現出来たが，カウンターの数値は原HPの数値に関わりなく４ソフトがゼロを示している。１ソフトについてはアクセスカウンターが１になっている。元のページは，０や１でない数字が示されている。

本評価を通じて，WWW機械翻訳ソフトについて今後，改良していかなければならな

い問題点が明確になってきた。総てのソフトが出来ない項目については，今後工夫をしていかなければならない。しかし１つのメーカでも可能になっている項目については，他のメーカでも十分，取り入れることが可能なはずである。

▶４　インターネット翻訳ソフトの将来展望

今年開催された沖縄サミットでのコミュニケ「G8コミュニケ・沖縄2000」中の「文化の多様性」の項や，「グローバルな情報社会に関する沖縄憲章」で大きな頁を割いている文化の多様性を守るには，文化が言語と深く結びついていることを考えると，機械翻訳の発展と活用が欠かせない。またそれらの中で強調されているデジタル・デバイドやデジタル・オポチュニティの問題でも，世界中の諸国民のうち僅かな人しか自由に英語を使えない現状や，使えるにしてもマスターするのにネイティブに比べて膨大なエネルギーを浪費せざるをえない状況では，インターネットを代表とする情報の獲得と流通で機械翻訳が重要な意味をもってくる。幸い日本は異なった語族間の機械翻訳システムの経験が深いので，機械翻訳研究の分野で世界をリードしている。

筆者はアジア太平洋機械翻訳協会のネットワーク翻訳研究会の委員を兼ねて，インターネット機械翻訳技術の評価の研究を行なっている。この協会は日本，アジアの機械翻訳研究の中心であり，世界機械翻訳連盟に属している（日本が中心になって設立したものである）。この連盟のもとにアメリカ機械翻訳協会，欧州機械翻訳協会がある。せっかく日本が主導した今回のサミットであれば，日本がリードしている機械翻訳の振興と活用を取り上げるべきではなかったろうか。これは乏しい資金で研究開発を続けている機械翻訳の研究者やメーカも勇気づけることになる。本コミュニケによればディジタル・オポチュニティ作業部会（ドット・フォース）が設けられるとのことなので，その作業部会で日本が世界に貢献できる可能性の高い，機械翻訳技術の活用が大きなテーマになることを期待している。

WWW機械翻訳は，機械翻訳全般の中で大きな発展の可能性を有している。それはWWWがタグという構造を持っており，それを頼りに文章構造を把握し機械翻訳に活用できるという点である。この点で電子技術総合研究所が中心になって行なっているXMLによる言語のタグ付けの体系である大域文書修飾（Global Document Annotation，GDA）（http://www.etl.go.jp/etl/nl/gda/）のプロジェクトが注目される。

今後，インターネット機械翻訳システムに対する需要と期待はますます高まる。評価を行なうことにより，ソフトの進歩が促進されるので，評価技術の確立は非常に重要なことである。今回の評価を通じて，WWW機械翻訳はどの点に問題があるかが明確になってきたので，それらをテスト出来る汎用性のある標準評価テストサイトを作成中である。このサイトのホームページは英語版である。WWW用機械翻訳ソフトの性能をテストしたい場合は，このサイトにアクセスして機械翻訳することにより，性能を評価できる予定である。

以上の点を踏まえて，我々のグループでは今後のインターネット機械翻訳評価の課題として，次のような点を進めつつある。

① 上記，標準評価テストサイトの充実
② WWW機械翻訳品質の向上
- 機械翻訳ソフト側の工夫で出来ることについてメーカ側への提言
- HTML記述の工夫で可能になることついてページ作成者への提言
- 今の技術では非常に困難なことについて代替案を提言

・WWWの機能項目の出現頻度（重要性）の測定
・同じホームページの中で翻訳結果の用語の統一がとれているかの検証
・XML等を活用したタグ付け言語の活用とツールの開発

③　翻訳速度の評価
④　MTソフトが対応しているインターネットブラウザの調査
⑤　WWW以外のインターネット機械翻訳の評価

我々は今後，インターネット翻訳ソフトを更に発展させる責務がある。インターネット翻訳ソフトが真に発展と普及を遂げたときに，人類は多様性を保ちながら，初めて「バベルの塔」を克服できるのである。

●謝　辞

機能評価の研究に助成いただいた電気通信普及財団に感謝いたします。また性能評価に当たっては，アジア太平洋機械翻訳協会ネットワーク翻訳研究会の委員に，性能評価に当たっては秀明大学宮澤ゼミ生の助力を得たので，ここに記して謝意を表します。

●参 考 文 献

[1] 宮澤信一郎，林紘一郎（2000）インターネット機械翻訳の機能評価に関する研究　第17回情報通信学会大会
[2] S. Miyazawa, S. Yokoyama, M. Matudaira, A. Kumano, S. Kodama, H. Kashioka, Y. Shirokizawa and Y. Nakajima "Study on Evaluation of WWW MT Systems", Proceedings of Machine Translation Summit VII(Singapore), pp.290-298 (1999)
[3] S. Yokoyama, A. Kumano, M. Matudaira, Y. Shirokizawa, M. Kawagoe , S. Kodama, H. Kashioka, T. Ehara, S. Miyazawa and Y. Nakajima,"Quantitative Evaluation of Machine Translation using Two-way MT", Proceedings of Machine Translation Summit VII(Singapore) (1999)

（宮澤信一郎　秀明大学国際協力学部教授）