Vol.2016-NL-228 No.4
2016/9/29

# ウェブページ・インターフェース翻訳の技術課題

松山宏樹[†1]　岡田勇[†2]　宮﨑瑞之[†3]　Dawn L.Miyazaki[†4]　横山晶一[†5]
江原暉将[†6]　宮澤信一郎[†7]

**概要**：機械翻訳技術の向上に伴い，よりシームレスな自動翻訳システムに対するニーズが高まってきた．異なる言語を用いてもスムーズにコミュニケーションするためには機械翻訳技術だけでなく，コミュニケーションを行う主体間のインターフェースにも適切な翻訳技術が用いられる必要がある．しかし，このようなインターフェース翻訳はこれまであまり注目されていない．我々はウェブページ・インターフェースを取り上げ，克服すべき技術課題の探索を行った．そのためウェブページで標準的に使用されている文法を中心に検討し，標準テスト項目を策定した．これを用いて商業的に機械翻訳を提供している主要 9 つのサイトに対し人手評価を行った．この結果，技術的課題を 5 カテゴリに集約させることができた．ウェブページ・インターフェース翻訳の主たる技術課題としては，異なるフォーマットとして記録されているデータの中に埋め込まれた文字情報の復元化が挙げられ，また技術的に可能な属性についてサイト別に評価し，翻訳再現率といった指標を提示することが有効であることが分かった．

# Technical Issues Concerning the Translation of Webpage Interfaces

HIROKI MATSUYAMA[†1]　ISAMU OKADA[†2]　MITSUYUKI MIYAZAKI[†3]
DAWN L.MIYAZAKI[†4]　SHOICHI YOKOYAMA[†5]　TERUMASA EHARA[†6]
SHINICHIRO MIYAZAWA[†7]

***Abstract***: The advancement of automatic translation systems has increased the demands for more seamless translations. Translation technique is not the only requirement for successful communication; equally important is the adequate translation technique of interfaces. Few studies, however, have examined issues associated with interface translation. This paper explored technical issues to be solved concerning the translation of webpage interfaces through the following procedures. First, standard test items were defined through the consolidation of HTML grammar frequently used on webpages. Next, nine of the leading commercial translation sites were manually evaluated using the standard tests items. Finally, the results of these tests were categorized into five classes. It has become evident that the primary technical issues of the interface translation are centered around the reproducibility of informational text and data embedded in various formats. Moreover, it has proven effective to evaluate technically possible attributes on every site and announce the indicator of the performance ratios of translation reproducibility.

## 1. はじめに

この数十年で機械翻訳技術は格段の進歩を達成している．その核となる翻訳アルゴリズムは，単語や句ベースを用いたもの[1]，構文木を用いるルールベース翻訳[2]や，用例ベース[3,4]，あるいは，統計翻訳[5]などが提案されそれぞれ一定の成果を上げている．これらの機械翻訳技術をどのように評価するかについても人手評価[6,7]のみならず，様々な評価指標を用いた自動評価[8,9]が提案されているが，実務翻訳者に対するアンケート分析では，多くの課題を見ることができる[10]．また，このような機械翻訳技術を用いたシステムは，通常，使用者にある決まった形式に従って作成された文の入力を要求し，システムが作成した翻訳文を特定の形式で受理できることが求められる．つまり，システム使用者にとってシステムが機能できるような入出力の形式は拘束される．

一方，機械翻訳システムの性能が向上し，その利用場面も多角化されるにつれ，システムと使用者とのインターフェースを考慮する必要が生じてくる．例えば，異なる言語の話者が同じウェブページを閲覧する際に，システムによって各言語に自動翻訳されたページを閲覧できるには，単に書かれている情報が翻訳されているのみならず，フォント情報やレイアウトなども保持されているべきである．また画像情報に文字が含まれている場合，その文字情報をそれとみなし翻訳されていることが望ましい．また例えば，異なる言語の話者が対面での対話を試みるとき，機械翻訳システムに音声認識システム[11]を連動させたインターフェースが要請される．

†1, †3, †7 秀明大学
Shumei Unicersity.
†2 創価大学
Soka University
†4 早稲田大学
Waseda University
†5 山形大学
Yamagata University
†6 江原自然言語処理研究室
Ehara Natural Language Processing Research Laboratory

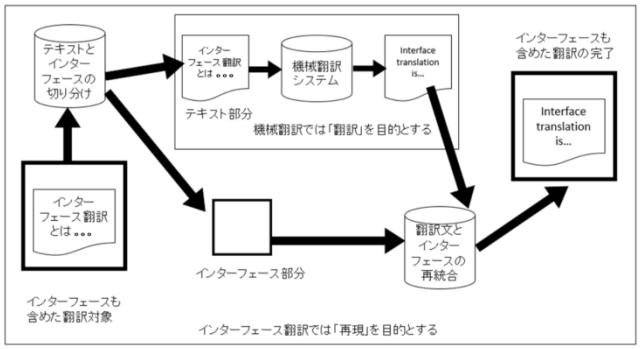


図 1：インターフェース翻訳の概念図

使用者がシームレスな自動翻訳システムを要求する場合，核となる機械翻訳システム以外に，そのシステムと使用者とのインターフェースも適切に再現できていなければならない．我々はこのような観点での翻訳をインターフェース翻訳と呼び（図 1），本稿でその研究の重要性を指摘する．

インターフェース翻訳は様々な場面で要請される[12-15]が，本稿ではウェブページに関するインターフェース翻訳の現状を分析し，技術課題を分類する．ウェブページは HTML 文法とその拡張を基本として開発されており[16]，タグ部分は異なる言語でも同じ表記になることから，ウェブページに関するインターフェース翻訳は容易と考えられるかもしれない．しかし，本稿で明らかになるように，主要なウェブページ翻訳システム（サイト）であっても少なくない属性が適切に再現されておらず，そのために翻訳できないウェブページが数多く存在する．

我々はウェブページ・インターフェース翻訳技術を開発するにあたり，現状のシステムではどのような属性が再現できており，どのような属性が再現できていないのかを分類する必要があると判断した．そのため，HTML 文法を用いて想定されるページ属性を列挙した標準テスト項目を定義し，それらを実装した標準テストサイトを開発した．これを用いて，商業的に機械翻訳を提供している主要 9 つのサイトに対し人手評価を行い，技術課題の分類を行った．

2 節では標準テスト項目の策定と標準テストサイトの開発について，3 節では主要サイトに対して人手評価を行った結果について検討する．最後に 4 節で議論をまとめる．

## 2. 標準テスト項目の策定と標準テストサイトの開発

ウェブページ・インターフェース翻訳の現状を探るために，HTML 文法規則[16]に基づき標準テスト項目を策定する．HTML 文法に関する標準化については，例えばソースコードを誤りなく記述するためのガイドラインを作成した HTML TIDY [17] や HTML 文書管理に関する標準的な手法の提案 [18], Web アプリケーションに対するテストの標準化法の提案 [19] などが行われている．

ここでは，HTML 文法規則 [16] をベースに，ほとんど使用されていない属性や，インターフェース翻訳においては考慮しなくてもよい属性を除いたテスト項目を 14 取り上げ，それぞれについて 1 つ以上の属性を抽出し，最終的に 33 の属性を列挙し，標準テスト項目として定義した（表 1）．

表 1：ウェブページ・インターフェース翻訳の標準テスト項目

| テスト項目 | 属性 | 簡単な説明 |
|---|---|---|
| ページ設定 | タイトルの表示 | <title>タグ内のタイトル |
| フォント設定 | フォントの再現 | イタリック体，アンダーライン，ボールド，ゴシック，フォントサイズ |
| | 訂正線の表示 | <s>タグ内の打ち消し線 |
| | タグを使ったフォントの再現 | ，，<sub>，<sup>，<blockquote>，<var>，<tt>，<basefont>，<big>，<small>，<font>タグによる強調 |
| イメージ設定 | クリッカブル画像中の文字 | 画像中の文字 |
| | PDF | 文章のみで書かれた PDF 形式の文書 |
| | Gif | GIF 形式の画像中の文字 |
| | メニュージャンプ | 画像をクリックすることで指定したリンク |
| レイアウト設定 | マーキー(Marquee)の再現 | <marquee>タグ内の文字が左右の移動可能性 |
| | 中央寄せ・改行・スペース | <align=”center”>,<br>,<p>, タグの指定 |
| | 一行空けての改行 | 一文の最後の<p>タグ |
| | <p>タグを使った文章の表示 | 段落としての機能 |
| | 水平線<hr>の表示 | 水平線の罫線 |
| | アスキーアートの表示 | レイアウトの再現 |

| | | |
|---|---|---|
| ハイパーリンク | リンクの表示 | <a href="URL">タグ内のリンク |
| リスト | 見出し・連番などの再現 | 見出しタグの大きさと番号 |
| | リストのマークの有無 | リストタグの● |
| | リストのタイトルの有無 | <li>タグの先頭に表示されている disc 記号 |
| テーブル | 体言止めの可・不可 | テーブルの外枠およびセル内の体言止め |
| | 複数行のテキストボックス | 入力欄の縦横サイズとリストタイトルおよび色を指定したセル |
| フォーム | ドロップダウン型選択リスト | <select>メニュー |
| | ラジオボタン | <input type=”radio”>タグ |
| | ボタン | <input type=”button”>タグ |
| フレーム | 各フレームの翻訳 | <iframe>タグによるフレーム |
| タグミス | タグの省略 | の非存在 |
| | タグのスペルミス | 開始タグと終了タグの不一致 |
| Java | 最終更新日の表示 | Java Script による最終更新日そのもの |
| | Java テキストボックス内の文字 | Java Script を用いたスクロール文字 |
| XML | XML 形式のテーブル | 独自タグで作成された XML 形式のテーブル |
| 効果 | alt アンカーの翻訳 | リンクのアンカー機能 |
| | アドレスの翻訳と再現 | <address>タグのアドレス |
| 複合型 | 重複タグ | <big>と<font color>ならびに |
| テストページ | の表示 | <i>と<u>タグ |
| | 異関係の混在タグ | テーブル内に記載されたタグ |

次に標準テスト項目を実装するため標準テストサイトを開発した．このサイトでは表 1 で定義した標準テスト項目をそれぞれテストできるようにページが作成されている．各テストページは対象テスト項目の再現性のみを評価できるような単純な構造になっている．

## 3. 主要 9 サイトに対する評価

ウェブページ・インターフェース翻訳の現状を評価し，技術課題を探るために，商業的に機械翻訳を提供している主要 9 つのサイト（表 2）を取り上げ，開発した標準テストサイトを用いて評価した．評価は 2014 年 12 月から翌年 1 月にかけて行われ，それぞれ異なる複数の評価者の評価結果に基づいて著者の一人が評価を確定させた．属性別に各サイトに関する評価を「再現できている」「部分的に再現できている」「再現できていない」の 3 段階とし，それぞれ 2,1,0 の値で数値化した．また部分的に再現できている場合は，再現できていない属性について指摘した．

表 2：評価対象となった機械翻訳サイト

| サイト名 | URL |
|---|---|
| nifty | http://honyaku.nifty.com/ |
| Bing | http://www.bing.com/translator/Default.aspx?MKT=ja-JP |
| Excite | http://www.excite.co.jp/world/ |
| Google | https://translate.google.co.jp/?hl=ja |
| Infoseek | http://translation.infoseek.ne.jp/web.html |
| So-net | http://www.so-net.ne.jp/translation/ |
| SYSTRAN | http://www.systranet.com/ja/web |
| Yahoo | http://honyaku.yahoo.co.jp/url/ |
| worldlingo | http://www.worldlingo.com/ja/websites/url_translator.html |

この結果，いかなる翻訳サイトでも再現できていない属性や，ほとんどすべての翻訳サイトでも再現できている属性，または再現可能と不可能で分かれている属性などが混在することが分かった．これらから属性ごとにウェブページ・インターフェース翻訳の再現達成状況を表 3 の基準に従って 5 つのクラスに分類した．

表 3：ウェブページ・インターフェース翻訳の実装レベルのクラス基準

| クラス | 実装の実現 | 選択基準 |
|---|---|---|

| 番号 | レベル | |
|---|---|---|
| クラス 1 | 実装が極めて困難 | Max = 0 |
| クラス 2 | 実装がかなり困難 | Max = 1 |
| クラス 3 | 実装可能だが未実施サイトが多い | Max = 2 かつ Sum < 9 |
| クラス 4 | 実装可能だが一部のサイト未実施 | Max = 2 かつ 9≦Sum < 16 |
| クラス 5 | ほぼ実装済 | Max = 2 かつ Sum ≧ 16 |

(*) Max とは，全サイトでの評価得点の最大値，Sum は全サイトでの評価得点の合計を意味する．

この基準に基づいて現状のウェブページ・インターフェース翻訳の属性ごとのクラス分けを行った．表 4 は日英翻訳を，表 5 は英日翻訳の状況をまとめる．

表 4：日英翻訳におけるウェブページ・インターフェース翻訳の技術困難性の分類

| クラス | 数 | 属性 |
|---|---|---|
| 1 | 3 | ①クリッカブル画像中の文字，②gif，③XML 形式のテーブル |
| 2 | 2 | ①最終更新日の表示，②Java テキストボックス内の文字 |
| 3 | 4 | ①タイトルの表示，②PDF，③見出し・連番などの再現，④各フレームの翻訳 |
| 4 | 20 | ①フォントの再現，②訂正線の表示，③タグを使ったフォントの再現，④メニュージャンプ，⑤マーキー(Marquee)の再現，⑥中央寄せ・改行・スペース，⑦一行空けての改行，⑧アスキーアートの表示，⑨リンクの表示，⑩リストのマークの有無，⑪リストのタイトルの有無，⑫複数行のテキストボックス，⑬ドロップダウン型選択リスト，⑭ラジオボタン，⑮ボタン，⑯タグの省略，⑰タグのスペルミス，⑱アドレスの翻訳と再現，⑲重複タグの表示，⑳異関係の混在タグ |
| 5 | 4 | ①<p>タグを使った文章の表示，②水平線<hr>の表示，③体言止めの可・不可，④alt アンカーの翻訳 |

表 5：英日翻訳におけるウェブページ・インターフェース翻訳の技術困難性の分類

| クラス | 数 | 属性 |
|---|---|---|
| 1 | 5 | ①タイトルの表示，②クリッカブル画像中の文字，③gif，④Java テキストボックス内の文字，⑤XML 形式のテーブル |
| 2 | 2 | ①アスキーアートの表示，②見出し・連番などの再現 |
| 3 | 5 | ①フォントの再現，②PDF，③最終更新日の表示，④alt アンカーの翻訳，⑤アドレスの翻訳と再現 |
| 4 | 6 | ①メニュージャンプ，②リンクの表示，③体言止めの可・不可，④複数行のテキストボックス，⑤ボタン，⑥各フレームの翻訳 |
| 5 | 15 | ①訂正線の表示，②タグを使ったフォントの再現，③マーキー(Marquee)の再現，④中央寄せ・改行・スペース，⑤一行空けての改行，⑥<p>タグを使った文章の表示，⑦水平線<hr>の表示，⑧リストのマークの有無，⑨リストのタイトルの有無，⑩ドロップダウン型選択リスト，⑪ラジオボタン，⑫タグの省略，⑬タグのスペルミス，⑭重複タグの表示，⑮異関係の混在タグ |


図 2：クリッカブル画像中の文字の日英翻訳

クラスごとに検討する．クラス 1 に属している属性は，現状の技術レベルでは実現が極めて困難であると判断される．例として，クリッカブル画像中の文字の日英翻訳を挙げる．図 2 はクリッカブル画像中の文字を含むウェブページを nifty で日英翻訳した結果であるが，画像中の文字は英語に翻訳されず日本語のままである．この例のように画像に含まれる文字情報や，XML といった異なるフォーマットとして記録されるデータの中に埋め込まれた文字情報に対しては，翻訳技術の開発が困難であることが示された．しかし，

画像になっている文字データは文字認識システムの高性能化で文字データとして復元できる余地があり，また，XML ではテキストデータをテキストして抽出可能である．これらのことからクラス 1 に属しているといっても，技術開発が完全に困難であるというわけではないと予測される．本研究によって，最困難属性が明らかになることで，研究開発が促進されることが期待される．また，表 4 と表 5 のクラス 1 の数を比較すると，表 4 は 3 であり表 5 は 5 である．このことから，技術的に困難なウェブページ・インターフェース翻訳の対応状況においては，日英翻訳よりも英日翻訳の方が遅れていることが確認できる．

クラス 2 に属している属性は，一部の機能が再現できているものである．しかし，表 4 と表 5 を比較すると，両者でクラス 2 以下である属性は前述した画像や XML を除いて一つしかない．この場合，言語依存性は低いと思われるので，結局，両方向の翻訳を通して，一部しか翻訳できていない属性は「Java テキストボックス内の文字」のみということになる．その主原因は文字化けであったことから，クラス 1 で明らかになった課題と同様に，異なるフォーマットの中に埋め込まれた文字情報の復元が課題であることが分かる．

クラス 3 以後は少なくとも一つのサイトで実現できていることから，技術的には可能であると判断される．これはインターフェース翻訳の重要性が高まるにつれて，改善できると予想されることである．例えば，クラス 3 以後に属している属性に限ってサイトごとに評価し，ウェブページ・インターフェース翻訳再現率といった指標を翻訳市場などに提示することは，技術を整備する環境条件になりやすい．

## 4. まとめ

本研究は，よりシームレスに自動翻訳されるためには機械翻訳システムだけでなく，それを包含するようなインターフェース翻訳が必要であるとの問題認識から，ウェブページのインターフェース翻訳技術の現状について評価することで技術課題や対応策について検討した．そのため，HTML 文法規則に基づき，14 項目，33 属性からなる標準テスト項目を策定した．これを用いて商業的に機械翻訳を提供している主要 9 つのサイトに対し人手評価を行った．この結果，どのサイトでも再現できていない属性や，一部のサイトでは再現できている属性，ほぼすべてのサイトで再現できている属性など，5 つのカテゴリに分類した表を，日英翻訳，英日翻訳ともに作成した．これらの分類を検討する中で，ウェブページ・インターフェース翻訳の主要な技術課題が，異なるフォーマットとして記録されているデータの中に埋め込まれた文字情報の復元化であることが明らかとなった．また，サイトによって再現できていたり，できていなかったりする属性が多いことが分かった．これは，現状のサイトが機械翻訳技術に焦点があった技術開発をしているためであると予測されるので，インターフェース翻訳の重要性を認識させるため，技術的に可能な属性についてサイト別に評価し，翻訳再現率といった指標を公開することで技術開発環境の整備を促進できると考えられる．

## 謝辞

今回の調査研究の機会を与えていただいた Asia-Pacific Association for Machine Translation（AAMT）に感謝申し上げます．

## 参考文献

[1] Andreas Stolcke: SRILM - an Extensible Language Modeling Toolkit, 7th International Conference on Spoken Language Processing, pp.901-904 (2002).

[2] Lonsdale, Deryle; Mitamura, Teruko; Nyberg, Eric (1995). "Acquisition of Large Lexicons for Practical Knowledge-Based MT" (PDF). Machine Translation 9: 251-283. Kluwer Academic Publishers.

[3] Medin, D. L. and M. M. Schaffer: 1978, Context Theory of Classification Learning, Psychological Review 85, 207-238.

[4] Harold Somers. Review Article: Example-based Machine Translation. Machine Translation 14: 113-157, 1999.

[5] P.Koehn, F.J.Och, and D.Marcu: Statistical Phrase-Based Translation, Proc. of HLTNAACL 2003, pp. 127-133, 2003

[6] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, Josh Schroeder. Meta-Evaluation of Machine Translation, Proc Second Workshop on Statistical Machine Translation, pp. 136-158. 2007.

[7] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, Eiichiro Sumita. Overview of the 1st Workshop on Asian Translation. Proc 1st Workshop on Asian Translation. 2014.

[8] Papineni, Kishore, et al. "BLEU: a Method for Automatic Evaluation of Machine Translation." Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002.

[9] H. Isozaki, T. Hirao, K. Duh, K. Sudoh and H. Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp.944-952.

[10] 長瀬友樹，小谷克則，工藤竜広，佐久間みゆき，秋葉泰弘: 実務翻訳における機械翻訳の利用に関する調査報告，言語処理学会第 20 回年次大会発表論文集, pp.610-613 (2014)

[11] Ketkar, S. S., and M. Mukherjee. "Speech Recognition System." Proceedings of the International Conference & Workshop on Emerging Trends in Technology. ACM, 2011.

[12] Miyazawa, S., Yokoyama, S., Matsudaira, M., Kumano, A., Kodama, S., kashioka, H., Shirokizawa Y. and Nakajima Y.: Study on Evaluation of WWW MT Systems, Proceedings of MACHINE TRANSLATION SUMMIT Ⅶ '99 "MT in the Great Translation Era"（1999）

[13] 宮澤信一郎，林紘一郎: 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(2000)

[14] Miyazawa, S., Okada, I., Shimizu, N., Yokoyama, S. and Ohta, T.: Study on Evaluating Machine Translation in Cyber Commons, PROCEEDINGS World Multiconference on Systemics, Cybernetics and Informatics (2003)

[15] 宮澤信一郎: 機械翻訳の評価･比較，社団法人　情報科学技術協会『情報の科学と技術』Vol. 55, No.8，特集＝「機械翻訳」, p.339-344 (2005)

[16] HTML Working Group, Transitional Document Type Definition, HTML 4.01, http://www.w3.org/TR/html4/sgml/loosedtd.html, W3C Recommendation 24 December 1999.

[17] D. Raggett, Clean up your Web pages with HTML TIDY, http://www.w3.org/People/Raggett/tidy/, 1998.

[18] W. Chisholm, G. Vanderheiden, I. Jacobs, Techniques for Web Content Accessibility Guidelines 1.0, http://www.w3.org/TR/WAI-WEBCONTENT-TECHS/, W3C Note 6 November 2000.

[19] F. Ricca, P. Tonella, Web Site Analysis: Structure and Evolution, Proceedings of the International Conference on Software Maintenance, San Jose, CA (2000) 76-86.