「完全自動」と「半自動」によるニューラル機械翻訳のエラー修正手法 ～翻訳者目線での修正作業を効率化するツールの紹介～

新田順也
エヌ・アイ・ティー株式会社

1. はじめに

筆者の会社は、翻訳支援ツールと産業翻訳を提供しており、翻訳者のためのニューラル機械翻訳支援ツール「GreenT（グリーンティー）[1]」を開発している。GreenT は Google Translate API を利用して Word ファイルの文章を翻訳する Word アドインである。ニューラル機械翻訳（NMT）のエラーを自動修正し翻訳者が編集しやすい訳文を出力できる（**図 1**）。このツールに弊社の用語集の作成技術[2]や原文と訳文の用語や数字の比較技術[3]を用いている。

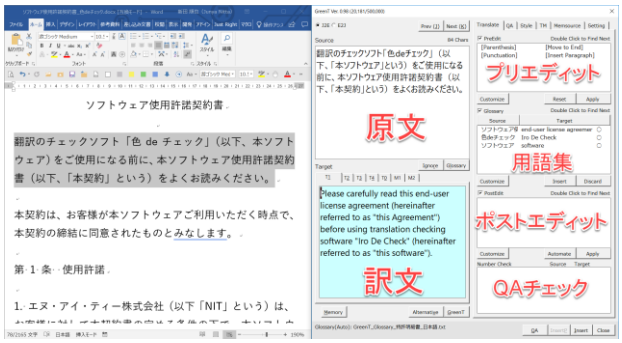

図 1　GreenT のユーザーインターフェイス

筆者は、2018 年 10 月の日本翻訳連盟（JTF）翻訳祭[1]や JTF ジャーナル[2]にて特許翻訳における NMT のエラーを指摘し、原文の修正（プリエディット）や訳文の修正（ポストエディット）によりエラーを解消する GreenT を紹介してきた。

本稿では、GreenT を用いて NMT を使う翻訳者の心理的な負担を減らしつつ翻訳品質を向上させる手法を紹介する。この手法には以下の特徴がある。

(1) 自前の用語集を用い 1 文単位で訳文を出力する
(2) プリエディットやポストエディットを自動化し修正作業を効率化する
(3) 訳文の用語や数字のエラーを自動検出する
(4) 自動処理を「完全自動」と「半自動」に区分し翻訳者が訳文作成に関与しやすくする

2. NMT のエラー修正の現状と課題

現在の NMT にはエラーがあるため、訳文の用途によっては人手による修正が欠かせない。NMT の出力結果を編集するポストエディット（MTPE）の活用方法が検証されてきた。NMT には特有のエラーが存在し、エラーの類型とともにプリエディットやポストエディットの具体例が紹介されている[3]。また、ポストエディットの自動化により品質の向上とともに効率化をはかることが提案されている[4]。一方で、特許翻訳において修正作業が多いと人件費がかかり低価格での翻訳提供には採算に合わないとする報告もある[5]。また「MT の流暢さが向上したことによって適切さの問題を検出しづらくなっている」[6]ことや、MTPE の単純作業による翻訳者への心理的負担が指摘されている[7]。

3. GreenT の翻訳手順

筆者はかねてより、翻訳における小さな繰り返し作業の自動化の重要性を説明しツールを提供してきた[8]。この考えに基づき、プリエディットやポストエディットに伴う作業を極力自動化し、翻訳者が訳文作成に注力できるよう GreenT を設計した。GreenT の翻訳手順を表 1 に示す。

1 GreenT　https://www.wordvbalab.com/word-addin/greent/
2 頻度のヒント https://www.wordvbalab.com/word-addin/hindo-hint/
3 色 de チェック https://www.wordvbalab.com/word-addin/iro-de-check/

表 1　GreenT の翻訳手順

ステップ	① 用語集の作成	② 原文の修正（プリエディット）	③ 訳文の出力	④ 訳文の修正（ポストエディット）	⑤ 訳文のチェック
作業目的	文書全体で用語集を作成	翻訳対象となる 1 文を訳し易く修正	用語集を適用し 1 文ずつ翻訳	訳文の誤訳箇所や表記を修正	数字/用語のチェック
完全自動	・用語の抽出	・否定語の検出	・表記統一	・用語の統一 ・文末処理	・用語/数字/否定語 ・用語の既出/初出
半自動	・訳語の決定 ・表記揺れ統一	・表現の修正提案 ・文字列の移動提案 ・文章の分割提案 ・用語の追記支援	・エンジン選択	・修正提案 ・用語の修正支援	

NMT において訳語の揺れや誤訳があるため用語統一の手法が研究されている。GreenT でも用語集に基づく訳文を出力できる。GreenT は翻訳者が使うことを前提としているため、用語集作りから始める手順としている。ただし、NMT に用語集を適用すると、訳語が正確に反映されなかったり文章が破綻したりすることがある[9]。また、数字も誤訳されることがある。このため、出力された訳文を最後にチェックする手順とした。

4. 完全自動処理と半自動処理の紹介

プリエディットやポストエディットの処理を完全に自動化できる場合とできない場合とがある。翻訳者の判断を要せず実行できるものを「完全自動処理」とし、翻訳者の判断が必要な処理は「半自動処理」とした。この「半自動処理」では、修正箇所と修正内容を文字情報や文字書式、色を用いて翻訳者に提案する。以下、表 1 の①～⑤の各ステップにおける「完全自動処理」と「半自動処理」を紹介する。

4-1 ステップ① 用語集の作成

（完全自動処理）

用語集に用いる名詞句を原文ファイルでの使用頻度に応じて抽出する。なお、日本語・英語ともに複合語も抽出できる。使用頻度が低くても重要な用語はあるため（例：契約書における会社名）、翻訳者が下線を引いて特定した語句は使用頻度が少なくても抽出できる。

（半自動処理）

抽出した用語に対して、手持ちの用語集とニューラル機械翻訳を用い訳語を出力する（図 2）。GreenT が訳語の表記揺れや誤訳の可能性を指摘し、翻訳者が内容を確認して修正する。

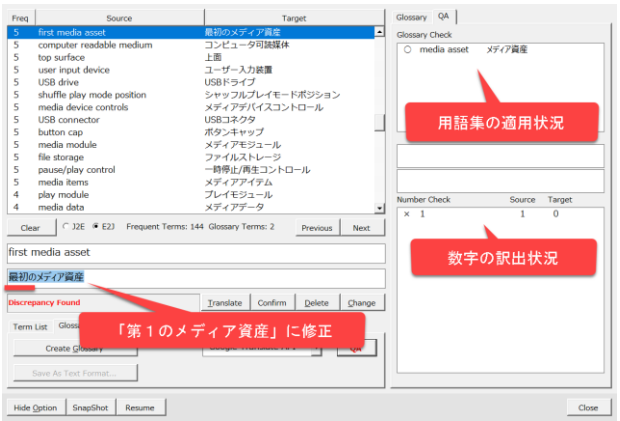

図 2 用語集のチェック結果と訳語の修正例

この際の調べものを効率化するため、Word からウェブ検索を実行するツール[4]とも GreenT は連動している。

この工程を経て作成した用語集を「③訳文の出力」と「⑤訳文のチェック」で使用する。このように、手持ちの用語集を活用し翻訳者自身の判断で選んだ訳語を NMT の訳文の出力に反映できるので、修正の負荷を低減できる。

4-2. ステップ② 原文の修正

GreenT には、原文を修正し訳文の出力結果を向上させる仕組みがある。ただし、機械翻訳に不向きな原文の場合は、原文の修正に時間をかけるよりも後述の用語や数字の入力支援機能を用いて、翻訳者が訳文を直接入力したほうが効率がよい。

（完全自動処理）

原文に否定語が含まれる場合、原文の文字列が斜体になる。NMT では否定語が正確に訳出されないことがある。注意喚起として原文中の否定語の存在を可視化する。

（半自動処理）

GreenT では、表現の修正案が表示される。NMT で誤訳を生じやすい表現がその修正案とともにプリエディット欄に表示される。たとえば、英日翻訳では省略文字があると誤訳や文章の破綻が生じやすい[2]。GreenT には化学分野でよく用いられる略語が登録されており、省略前の表記に戻す提案をする（図 3）。翻訳者が修正案を選択し、クリック操作で修正が完了する。

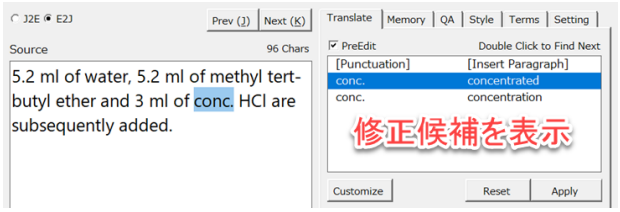

図 3 プリエディットの例

日英翻訳においても、誤訳を生じやすい表現と修正案が表示される。たとえば、「に記載の」を「に記載された」にする提案がなされる。様々な修正例が「特許ライティングマニュアル」に示されている[10]。NG ワードの登録もできる。「これ」のような、機械翻訳をするうえで誤訳が生まれやすい表現を登録して注意喚起のために使ってもよい。

NMT では、特定の文字列（which 節や括弧で囲まれた文字列など）を文章の末尾に移動して元の文章から分離すると、文章構造が明確になり出力結果が向上する場合がある。また、複雑な文章を改行し

[4] 右クリックで Google！ https://www.wordvbalab.com/word-addin/rg/

て分割すると編集しやすい訳文や訳語を取得できる場合がある[11]。GreenT では文字列の移動や句読点の選択、改行の挿入等の作業をクリック操作だけで実行できる。

NMT では、原文に主語を補うことで正確な英訳が出力される場合がある[3][10]。GreenT は、原文修正の際に翻訳対象の文章の前後に記載される名詞句や数字を表示して文字の入力支援をする。

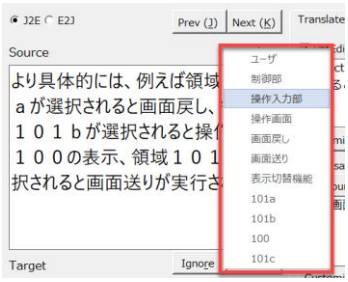

図 4　原文への用語の追記支援例

4-3. ステップ③　訳文の出力

（完全自動処理）

GreenT では用語や数字、日付の表記を統一できる。NMT では数字の誤訳が出ることがあるが、GreenT では既知のパターンの誤訳を解消できる。日付や図面番号については表記を指定できる。

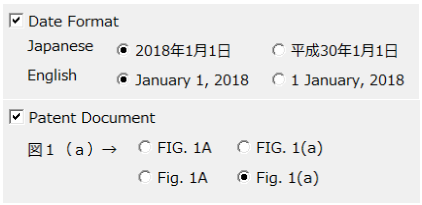

図 5　選択可能な表記例

また、日英特許翻訳における参照符号を正確に訳せる。NMT では「エンジン１００ａ」が「engine 100 a」となり、参照符号の 100 と a の間に半角スペースが挿入されてしまうことがあるが、GreenT では半角スペースの誤挿入を防止している。

そのほかに、日英翻訳の NMT 出力では、①や②などの特殊記号が全角文字のまま英文中に出力されてしまうことがあるが、GreenT ではこの全角文字の出力もできる限り回避している。英日翻訳においては、英数字記号を全角文字で出力することも可能である。

（半自動処理）

GreenT では 4 種類のエンジンを選択できる（表 2）。すべての種類において Google Translate API を利用している。他社製の翻訳エンジンも利用できるように現在準備中である。

表 2　利用可能な翻訳エンジン

	内容
T1	ニューラル機械翻訳（用語集を利用） 訳語が適用されないことがあるが、流暢な訳文を出力する
T2	ニューラル機械翻訳（用語集を利用） 訳語が確実に適用されるが、不自然な訳文になることがある
T3	ニューラル機械翻訳（用語集を利用しない）
T4	フレーズベース機械翻訳（用語集を利用） 訳もれは少ないが、不自然な訳文になることがある

それぞれ特徴が異なるため、同じ原文から別の訳文を出力できる。翻訳者は出力結果をそれぞれ比較して、編集しやすい訳文を採用する。

4-4. ステップ④　訳文の修正

（完全自動処理）

用語集を適用して訳文を出力した場合でも、訳揺れが時々発生する。既知の誤訳であれば自動で修正できる。たとえば、「media device」の訳語を「メディアデバイス」と定義した場合でも、「メディア・デバイス」や「メディアデバイド」になることがある。これらの訳語が出力されるたびに「メディアデバイス」に自動で変換するよう設定できる。

文末処理も自動化している。英訳文では文末のスペース数を 1 つまたは 2 つに指定できる。和訳文では、常体と敬体を指定できる。指定した文末処理の訳文となるように、既定のルールやユーザー定義のルールに基づいて自動で修正する。

さらに、原文に含まれる否定語の誤訳を防止するため、訳文中に否定語が含まれる場合には、訳文の文字列を斜体で表示し可視化する。

（半自動処理）

GreenT では誤記と思われる箇所を指摘できる。NMT では繰り返し表現が出力されることがある[3]。GreenT では、任意の 3 文字以上の繰り返し表現（例：メディアメディア）が検出され修正方法が提案される。修正提案の採択は翻訳者が判断をし、クリック操作だけで修正作業が完了する（図 6）。

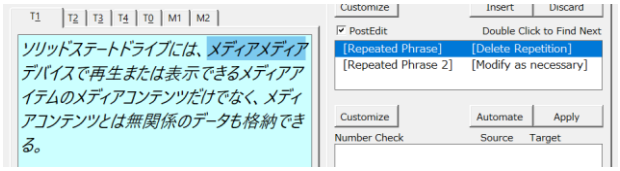

図 6　繰り返し表現の検出例

NG ワードを登録して、検出と修正案の提示がな

されるようにカスタマイズできる。また、好ましくない表現（冗長表現の「～することができる」など）を自身への注意喚起として登録してもよい。英文にて文章間のスペースを 2 つに統一する場合、完全自動処理では文末の判断ができないことがある。この場合、翻訳者がクリック操作だけで修正できる。

このように、誤訳対応については、「完全自動処理」と「半自動処理」とで使い分け、修正作業を容易にするためにクリック操作で完了するように工夫がなされている。上記例のように「media device」が「メディアデバイド」と訳される場合には、「メディアデバイス」の修正候補を表示してもよい（図 7）。

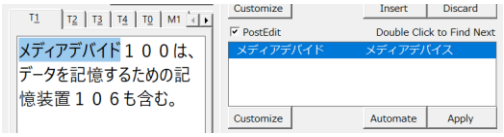

図 7　ポストエディットの修正提案例

原文の編集と同様に、文章修正の際に入力候補として用語と数字を表示できる。用語や数字の入力時間の短縮のみならず、タイプミスを回避できる。

4-5. ステップ⑤　訳文のチェック

上記のように、用語適用をしても訳語が正確に反映されるとは限らない。そのため、原文と訳文を比較して用語や数字、単位のチェックを行う。漢数字や英語表記の数字や、和暦と西暦との比較もできる。

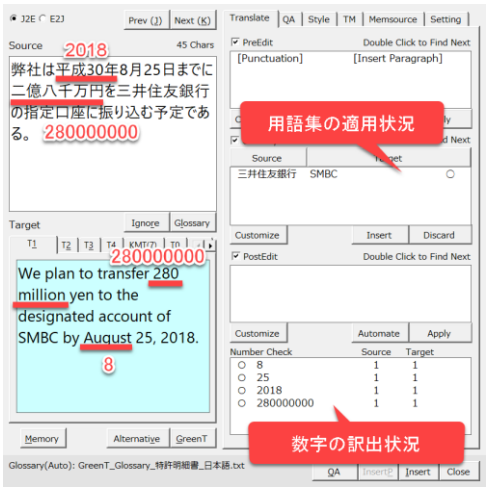

図 8　チェック結果の表示

5. まとめ

GreenT では、自分で決めた用語を用い訳文を出力できる。また、原文をクリック操作で修正すれば編集しやすい訳文も出力できる。細かい修正作業が自動化されるので、翻訳者は文章の流れや論理の確認、訳語・訳文の検証に時間を割けるようになる。さらに用語や数字のエラーを検出できるため、翻訳者が疑心暗鬼になって訳文と向き合う必要がなくなる。このようなことから、これまでの MTPE と比較して翻訳者の作業の負荷や心理的な負担が少なくなると期待できる。GreenT が翻訳の生産性と品質の向上に役立てば幸いである。

［参考文献］

[1] 新田順也（2018）．みんなのワードマクロ ブログ．https://www.wordvbalab.com/seminar/8888/

[2] 新田順也（2018）．ニューラル機械翻訳の弱点を補う支援ツール．JTF ジャーナル 2018 年 11/12 月号 #298，pp. 16-17.

[3] 森口功造，中安裕志，鈴木光（2018）．日英機械翻訳（NMT）のエラーの特徴ー機械翻訳の限界と人手による補完のポイントー．テクニカルコミュニケーションシンポジウム 2018 論文集，pp. 54-60.

[4] 徳田愛，澤田祐理子（2018）．10 社の事例から学ぶ！機械翻訳導入の課題と解決策．テクニカルコミュニケーションシンポジウム 2018 論文集，pp. 29-36.

[5] 梶木正紀（2018）．機械と人間との協働 LSP perspective ～学習型機械翻訳から翻訳プラットフォーム～．言語処理学会 第 24 回年次大会 発表論文集（2018 年 3 月），pp. 750-752.

[6] 藤田篤，山田優（2017）．翻訳の品質と効率:実社会におけるニーズと工学的実現可能性．言語処理学会 第 23 回年次大会 発表論文集（2017 年 3 月），pp. 915-918.

[7] 阪本章子（2018）．翻訳テクノロジー論考 第 3 回 翻訳の社会学テクノロジーとの共存を目指して．JTF ジャーナル 2018 年 11/12 月号 #298，pp. 24-25.

[8] 新田順也（2012）．Word マクロ×アイディア！×カンチガイ？＝なんだかすごいこと！！．JTF ジャーナル 2012 年 5/6 月号 #259，pp. 24-29.

[9] 飯田頌平，龍梓，木村龍一郎，宇津呂武仁，三橋朋晴，山本幹雄（2018）．ニューラル機械翻訳における大規模語彙および訳抜けへの対応の併用．言語処理学会 第 24 回年次大会 発表論文集（2018 年 3 月），pp. 877-880.

[10] 一般財団法人日本特許情報機構 特許情報研究所（2018）．特許ライティングマニュアル（第 2 版）「産業日本語」．

[11] 湯浅豊裕（2018）．機械翻訳と CAT ツールをいっしょにつかう．JTF ジャーナル 2018 年 11/12 月号 #298，pp. 18-19.