それでも学生はポストエディターになれるのか？
ニューラル機械翻訳(Google NMT)を用いたポストエディットの検証

山田優　　大西菜奈美

関西大学

1 はじめに

機械翻訳＋ポストエディット(以下 PE)は，産業翻訳において一定のニーズや特定の目的に対する翻訳サービス提供のための手法として確立してきたと言える．日本語と英語の組合せによる PE でさえも，ニューラル機械翻訳(NMT)の実用化以降，その普及に拍車がかかっているようである．また，PE が業界的に認知されてきたという実情は，ポストエディットの国際基準 ISO18587 が発行された事実にも見受けられる．

2017 年時点での ISO18587 が求めるポストエディターの資格基準は，翻訳サービスの ISO17100 とほぼ同じ内容になっている．それに加え，ポストエディターに対して「専門的知識および能力(Professionalism)」に関する要求事項が追加されている [1]．その一つ目に「機械翻訳技術および機械翻訳エンジンが出力する典型的なエラーに対する一般的知識」を有することとある．

専門サービス提供者としてこのような知識を持つことは好ましいことであるが，一方で，ニューラル技術の進歩の速度を考えると「機械翻訳エンジンが出力する典型的なエラー」は知り得るものだろうか，という疑念も生まれる．しかし，産業や大学教育機関で ISO18587 に準拠したポストエディター教育を行なうのであれば，NMT のエラーの傾向を把握しておくことは，実務者・研究者・教育者として必要なことであろう．

このようにポストエディターの需要が高まる状況を想定し，Yamada[2]では，誰がポストエディターになるのかというリサーチクエスチョンを，具体的には，大学生のようなアマチュア翻訳者がポストエディターになれるのかという問いを，当時の統計ベースの Google 翻訳(GSMT)を使って検証した．しかし，結果は学生の PE の成績はプロ翻訳者の PE 品質には満たなかった．

2016 年秋にローンチした Google NMT(GNMT)は，日英間の翻訳品質が飛躍的に向上したと言われるが，はたして GNMT がポストエディットに適しているのかどうかは，まだ検証されていない．そこで本研究は，GNMT がポストエディットに向いているのかどうかを検証する．検証方法は，Yamada[2]と同様に大学生を対象に PE を実施し，GSMT と GNMT の PE への適合性の違いを，PE の品質結果，体感認知負荷，修正量に加え，エラーの種類の観点から検証する．

2 リサーチクエスチョン

GNMT と GSMT とでは，どちらが PE 適正が高いのか，を検証する．検証指標は，以下の通り．基本的には，すべて Yamada [2]の調査で用いたものである．これにより，2012 年の GSMT の PE との比較が可能になる．

1. 体感作業負荷(主観調査)
2. 修正量(GTM)
3. 翻訳品質(MNT-TT 校閲カテゴリ)

1 の体感作業負荷は，ポストエディットの体感作業負荷を数値化した指標である．人手翻訳(HT)の体感負荷を 100(基準)として，ポストエディットの体感作業負荷はどうだったのかを回答してもらう．例えば，ポストエディットが人手翻訳よりも 2 割程度の負荷軽減を実現したと感じたならば 80 と記入(100 マイナス 20)，逆に 2 割増えたと感じたら 120(10 プラス 20)という具合である．

2 の修正量は，MT 訳と最終翻訳物のテクスト類似度で，General Text Matcher(GTM)で測定した．修正量が少なければ，機械翻訳の訳が役に立ったということの証拠になるが，体感負荷と修正量が厳密に相関するわけではないので，あくまで目安としての指標である．

C の翻訳品質は，人手の評価者によってチェックした．みんなの翻訳実習(MNH-TT)[3]で提供されている校閲カテゴリ[4][5]を用いて評価を行った．

3 実験デザインと参加者

調査は，某大学外国語学部の3回生 28 名を対象に行った．翻訳演習という授業の履修者で，英日・日英の翻訳の基礎演習を中心に学んだ．学期中に数回ほどポストエディットの演習を実施したが，学生たちがポストエディットの方法を習得したとはいえない．これらの条件は，Yamada [2]と同じである．また実験デザインとして，テクストは，英語の Wikipedia から「Steve Jobs」の項目の抜粋を使用，これを GNMT で和文に翻訳し，MS Word の罫線フォーマ

本原稿は、オリジナルの発表原稿に、表４の内容を変えずに体裁を修正し、その表の引用・参照元を明記した訂正版である。(2022年9月7日)

ットに貼り付け，そのまま上書きで和文を修正（PE）できる形で，授業内の課題の一部として課した．課題終了後には，Web アンケートで，体感負荷や PE の感想などを回答してもらった．いずれも，前回と同じ手順・方法であり，異なるのは使用した MT エンジンが，GSMT から GNMT に変わったという点のみである．

4 実験結果

4.1. 1: 体感負荷

まず，GNMT＋PE の作業が，普通の HT と比較して「楽」になったかという質問に対して，28 名中 20 名が（71%）が「はい」と回答した．GSMT+PE の時は，74%だったので若干ポイント減である．HT を 100 とした GNMT+PE の体感負荷は 79.1，約 21%楽になったが，SMT+PE の時は 75.1 だったので，こちらも若干だが体感負荷は増になったことになる．ただし統計的有意差は，GNMT+PE と GSMT+PE の間には無い(p-value = 0.8008)．結果として，NMT の品質が向上したとはいえ，PE での体感負荷には変化がなかったという結果になった．

4.2. 修正量

修正量は GTM で，MT 出力と PE 後のテクスト類似度をみている．GTM は，0 - 1 のスケールで，1.00 が完全一致を意味するので，数値は多いほど一致率が高い（すなわち，修正量は多い）ことを示す．GNMT+PE vs. GSMT+PE の平均比較では，0.790 vs. 0.753 となった．若干であるが，GNMT のほうが高くなっている（つまり GNMT+PE では，修正量は減っている）．統計的にも，有意差が確認された(p-value = 0.002825)．

上2つの結果から推測されるのは，GNMT+PE は，修正量が少ないのに体感負荷が減らないため，Koppone [6]が言うように，「高い認知負荷がかかるようなエラー」が GNMT 出力に含まれているのかもしれないということである．

	GSMT+PE	GNMT+PE
体感負荷	75.1	79.2
修正量（GTM）	0.753	0.790

表 1: 体感負荷と修正量

4.3. 品質

品質は，幾つかの側面から検証しているが，まず最初に PE を行なう前の MT 出力（Raw MT output）に対して，MNH-TT 校閲カテゴリでアノテーション付けを行った結果から示す．便宜的に重要度（エラーの重にづけ：メジャー／マイナー）も付加して，下記の表にまとめた．

この表からわかることは，PE を実施する前の段階で MT の出力に含まれるエラー数が，GSMT 43，GNMT 27 と，GSMT のほうが 1.5 倍程度エラーが多いということである．つまり，一般的に言われるように，同じ原文の翻訳における GMNT の品質は向上したということである．

	GSMT	GNMT
メジャー・エラー数	31	10
マイナー・エラー数	12	17
合計	43	27

図2：GSMT と SNMT の訳出に含まれるエラー数

これに対して，PE 後のテクストについても同様にエラーのタグ付けをした．その表が下である．GNMT+PE では平均 3.2 個のエラーしか残らなかったので（平均 6.8 個のエラーが修正された），見かけ上は，GSMT+PE よりもエラーの絶対数は半減したと言える．しかし，エラー修正率でみると GSMT+PE のほうが 77.7%と高く，GNMT+PE は 68%と低かった．プロの PE であれば，エラー修正率 85%を達成できる[7]．

学生による修正率のバラツキは，GSMT+PE も GNMT+PE もほとんど同じであった．このことから，学生による GNMT+PE の成績は，むしろ GSMT+PE よりも悪化していると考えることもできる．むろん，繰り返しになるが，もともとの GNMT 出力に含まれるエラー数は少ないので，最終 PE プロダクトに残されるエラーの絶対数は，見かけ上は減少して，PE 品質が向上したと言えないことはない．

	GSMT+PE	GNMT+PE
未修正エラー数（修正エラー数）	6.9 (24.1)	3.20 (6.8)
エラー修正率	77.7%	68%
エラー修正率のバラツキ	41-93%	40-90%

表3：ポストエディットの品質

5 ここまでのまとめ

GNMT+PE と GSMT+PE の比較では，体感負荷は HT 比で，両方共に約 75%と変わらない．しかし，修正量は，GNMT+PE のほうが少なくなっている．品質については，エラーの絶対数は，GNMT+PE で減少するので，最終プロダクトの品質は見かけは向上するが，エラー修正率でいうとむしろ GNMT+PE では減少している．

これらの結果から，GNMT+PE は，PE 修正量は少なくなるが，作業負荷やエラー検出率は横ばいか，むしろ悪化する．つまり GNMT+PE は GSMT+PE よりも高い作業負荷を強いられるわりに，あまりエラーを修正できないタスクになる．PE のしやすさという観点からいうと，適正が低下している可能性がある．以下のセクションでは，この原因について考察する．

6 NMT のエラーと人間の学習者のエラーの比較

本研究の狙いの1つに，翻訳学習者（大学生）がポストエディターになれるのか，という教育的側面から学生のパ

フォーマンスを考察することがある. そのため, 今回のエラー分析には, MNH-TT の校閲カテゴリを採用した. これは翻訳者教育に特化したエラーカテゴリを提供し, かつ英日間の翻訳でのエラーアノテーション用に最適化されている為である. MNH-TT 校閲カテゴリの詳細については, 先行研究を参考にされたい. エラーの一覧表は以下に記しておく.

校閲カテゴリ	
X1	原文内容の欠落
X2	原文にない要素の付加
X3	原文内容の歪曲
X4a	未翻訳
X4b	直訳調
X6	曖昧さ未解消
X7	用語の訳出誤り
X8	コロケーションの誤り
X9	その他の文法的・統語的な誤り
X10	前置詞や助詞の誤り
X11	活用の誤りや数・性などの不一致
X12	綴り誤り・誤変換
X13	句読法に関する誤り
X14	レジスタ違反
X15	表現のぎこちなさ
X16	結束性違反

表 4：MNH-TT 校閲カテゴリ（豊島ら[4], 山本ら[9]を参照）

6.1. 人手翻訳 HT のエラー分布

実験参加者には, PE タスクとは別に普段通りの HT の課題も課しておいた. これらの HT 訳出にエラーアノテーションを付与した. 学習者 28 名分の全エラー中, どのカテゴリのエラーを犯す頻度が高いのか／低いのかの分布を以下のグラフに示す. ポストエディットではなく, 翻訳の学習者が普段翻訳 HT をする際に, どのような翻訳エラーを起こすのかを知ることができる.

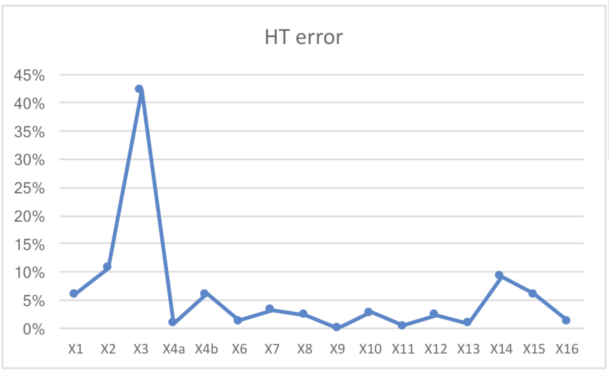

図 1：HT のエラー分布

このグラフが示すとおり, 学習者が起こしやすいエラーは, 全体の傾向として X3(原文内容の歪曲), 一般的にいうところの意味の「誤訳」が突出して多くなっている. これは, 豊島[4], 山本[5]ら先行研究結果とも合致する. その次に X14 や X4 が多くなる. いずれにしても, X3が突出している理由について, 大西ら[8]では, 学生に回顧法インタビューを実施し, 訳出結果だけでなくプロセス面から, 詳細原因にも迫っている. この研究を参照して, 今回の GNMT+PE の考察もこころみる.

6.2. GSMT と GNMT のエラー

学習者の HT のエラーと GSMT と GNMT の出力に含まれるエラーを比較して見る. エラーアノテーションを付与したのは PE を行う前の MT からの出力(raw MT output)である.

これら2つのエンジンのエラー分布を下記に示す. グラフの形状から一目瞭然であるが, GSMT と GNMT エラー分布は全く異なる. 両者とも, 学習者のエラー分布と同じように X3 が多くなっている. GNMT のほうは, X3 と同じくらい X7(不適切な訳語)の割合が高い. これは NMT のエラーの特徴とも言えるだろう. しかし, X7 を除けば, GNMT のエラー分布は学生の HT のエラー分布に類似する. それに対して, GSMT のエラーには, X3 の他に X4b, X7 X9, X10 など, 様々な種類のエラーが多く含まれる.

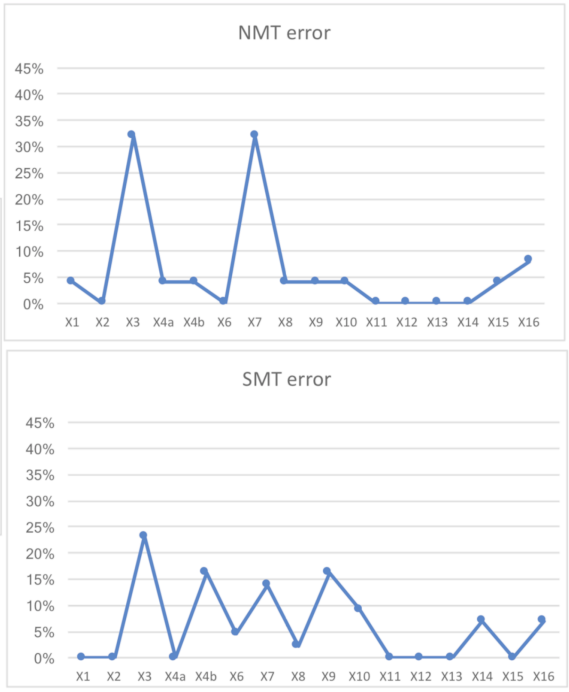

図 2：GNMT と GSMT のエラー分布

6.3. Human-Computer Interaction と補完関係

PE とは, 人間とコンピューターのインターアクション(Human-Computer Interaction = HCI)であるから, お互いがお互いの欠点を補完し合えるような関係(complementary)にあるのが望ましい. この観点からいうと, GSMT のエラー分布は, 人間(学生)翻訳者のエラーとは補完的な関係にあるといえる. つまり, GSMT のエラーは, 学生翻訳者が犯さないようなエラーの種類を多く含んでいるので, 学生にとっては意識しやすく気づきやすい

エラー，すなわち PE し易いエラーが多いと考えることができるのだ．

これに対して，GNMT のエラー分布は，X7 以外は学生翻訳者のエラー分布に近い．X3 は，そもそも学生翻訳者が多く犯すエラーであるので，このエラーを PE で修正するのは，難しく，認知負荷が高くなってしまう可能性がある．このことから，GNMT は GSMT よりも，学生の PE には不向きであると考えられるのである．

7 X3 エラーの謎

学習者の翻訳には X3のエラーが突出して多く含まれることから，大西ら[8]では，回顧法インタビューを実施して，学習者がエラーを犯してしまった箇所で何を考えていたのか（いなかったのか）を調査し，X3 エラーが起こるメカニズムの詳細を分析した．下図に示すように，同じ X3 エラーでも，発生メカニズムは様々で，大きく分類すると，問題の箇所に対して人間の翻訳者が「考慮」ないし「注意」を向けていたかどうかが，分岐点になる．エラーの箇所に注意が向けられていない場合は（図の×），(1)不注意による誤り（原文を読み違えている等）や，(2)思い込み起因する誤り（単語の意味を間違って記憶している）ものなどに分けられる．問題箇所に対して「注意」が向けられている場合でも（図○），(3)原文の解釈に関して迷いがあったり（解決策を見つけられない），(4)意味を取り違えていたり（間違った解決策にたどり着く）する．また(5)の「より適切な訳をと求めて」は，原文の意味を正しく理解したにもかかわらず，翻訳の修正の過程で不適切な訳文に改悪してしまうケースもある．

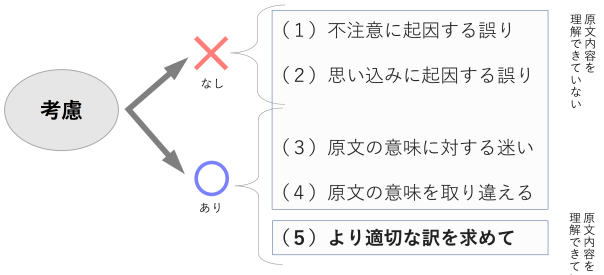

図３：[X3 原文内容の歪曲]の原因とその分類

PE で発生する X3 エラーも，「注意」が向けられていなかったケースや，「注意」は向けられていたにもかかわらず，修正に失敗しているケースなどがあることが分かった．注意が向けられていない X3 の場合は，PE では，認知負荷が低い（low effort or effortless）になり，注意が向けられているのに修正が失敗した X3 は，認知負荷が高いエラー（high effort）になる．

8 ポストエディター教育とは

これからのポストエディター教育を考えた場合，X3 エラーは，学習者にとって，そして翻訳という根本的なタスクにおいて，非常に根の深い問題であるので，これを PE するという行為は，もはや PE だけに限定した対策やスキルで解決できるものではない．従来から HT（人手翻訳）に求められてきたような翻訳のためのスキルセットやコンピテンスが，人間翻訳者に求められているのだ．つまり NMT+PE の時代のポストエディター教育を行なうには，翻訳者専門教育における翻訳コンピテンスと翻訳者コンピテンスを合せて考えていかなければならないだろう．

謝辞

本研究の一部は，日本学術振興会科学研究費補助金基盤(A)「翻訳知識アーカイヴ化を利用した協調・学習促進型翻訳支援プラットフォームの構築」（研究課題番号:25240051)の支援を受けて行われた．

参考文献

[1] 森口功造 (2017).ISO17100 と ISO/DIS 18587.2 の要求事項の比較とポストエディット現場への影響．*言語処理学会 第 23 回年次大会 発表論文集*, pp. 1157-1159.

[2] Yamada, M. (2014) Can College Students be Post-Editors? An Investigation into Employing Language Learners in Machine Translation plus Post-Editing, *Machine Translation*, 29(1), 49–67.

[3] Babych, B., Hartley, A., Kageura, K., Thomas, M., & Utiyama, M. (2012). "MNH-TT: a collaborative platform for translator training." In the proceedings of *Translation and the Computer 31* (November 29-30, 2012), 1–18.

[4] 豊島知穂，藤田篤，田辺希久子，影浦峡，Anthony Hartley. (2016). 校閲カテゴリ体系に基づく翻訳学習者の誤り傾向の分析.*通訳翻訳研究への招待*, Vol. 16, pp. 47-65,

[5] 山本真佑花，田辺希久子，藤田篤．(2016). 翻訳学習者の学習過程におけるエラーの傾向の変化．*言語処理学会第 22 回年次大会発表論文集*, E5-3, pp. 865-868.

[6] Koponen, M. (2012). “Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations.” In C. Callison-Burch et al. (eds), *7th Workshop on Statistical Machine Translation* . Proceedings of the workshop. Stroudsburg, PA: Association for Computational Linguistics, 181–190.

[7] Comparin, L. & Mendes, S. (2017). Using Error Annotation to Evaluate Machine Translation and Human Post-Editing in a Business Environment. In the proceedings of *EAMT 2017*.

[8] 大西菜奈美，山田優，藤田篤，影浦峡．(2017). 翻訳学習者が誤訳をする理由: MNT-TT の校閲カテゴリ「X3」から見る学習者の訳出プロセスと学習効果．通訳翻訳研究への招待, Vol. 18, pp. 88-16.

[9] 山本真佑花・田辺希久子・藤田篤（2015）エラーカテゴリーに基づく翻訳学習者の学習過程における習熟度の分析『日本通訳翻訳学会第16回年次大会予稿集』, 29.
（2022年9月1日http://paraphrasing.org/~fujita/publications/coauthor/yamamoto-JAITS2015-slides.pdf から取得）