事例・研究

「カスタム MT＋フル PE」現状と見通し：医療翻訳分野の事例研究

山田 優[1]　早川 威士[2]

[1]立教大学　[2]株式会社アスカコーポレーション

1. はじめに

医学・医薬分野の翻訳の特徴は、高度な専門性がその 1 つと考えられる。例えば、語彙の特殊性や構文の複雑さなどがあげられる。また、医学・医薬分野は、人類医療の資産となりうる研究や、法規制に基づく新薬の承認申請、患者の生命を左右する情報を扱うことから、正確性の要求品質が高い。機械翻訳（MT）のドメインアダプテーションないしカスタム MT という技術を使うことで、専門領域に特化した MT エンジンを作ることができる。その際に、「医学・医薬分野（メディカル分野）」という括りでは専門性を絞りきれているとはいえない。より細分化された分野・用途に応じた文書特化型 MT エンジンの構築が必要となる。

さて、このようにして作ったカスタム MT を産業翻訳、とくに医学・医薬翻訳の分野で、ポストエディットに用いるとした場合に、プロ翻訳者は、それをどのように見るか。作業効率は向上するのか。品質はどうか。MT の産業への実装という観点から興味深い問いである。

このような背景から、本研究では医学・医薬分野の中でも詳細分野に特化した MT エンジンを作成し、そのカスタム MT を活用した実務翻訳における MT＋フルポストエディット（以下、フル PE）の実用性を評価する。

2. カスタム MT の構築

本研究では、医薬の文書特化型エンジンと医薬汎用型エンジンの 2 種類の MT モデルを構築した。基盤となる MT モデルには情報通信研究機構（NICT）の汎用 NT を使用した。まず、医学・医薬分野から広く収集した教師データ 40 万文で訓練を行い、医薬汎用型エンジンを構築した。続いて、英日方向については、医学論文のテクスト 13 万文を追加して訓練（学習の重みを強めるため data augmentation）を行い、医学論文に特化した MT を構築した（英日の文書特化型エンジン）。日英方向については、医薬品臨床試験の計画書から抜粋したテクスト 19 万文を追加して訓練を行い、日英の文書特化型エンジンを構築した。いずれも教師データには公開済みデータを原文として収集し、それを翻訳したものを使用した。

表 1 は、ある原文に対する英日の医薬文書特化型エンジンと医薬汎用型エンジンの出力の例である。例えば、原文の “chronic kidney disease” は、文書特化型エンジンでは「慢性腎疾患」、汎用型エンジンでは「CKD」と出力されているなど、用語に違いがみられる。

原文	Randomised, double-blind, placebo-controlled and parallel dose group trial to investigate efficacy and safety of multiple doses of oral XXX-12345 over 14 weeks, alone and in combination with empagliflozin, in patients with diabetic and non-diabetic chronic kidney disease
文書特化型エンジン	糖尿病性及び非糖尿病性慢性腎疾患患者を対象として、XXX-12345の経口反復投与(単独投与及びエンパグリフロジンとの併用投与)の効果及び安全を14週間にわたり検討する、無作為化、二重盲検、プラセボ対照、並行群間比較試験
医薬汎用型	糖尿病及び非糖尿病性CKD患者を対象に、経口XXX-12345を単独又はエンパグリフロジンと併用で14週間にわたり反復投与したときの有効性及び安全性を検討する無作為化二重盲検プラセボ対照並行群試験

表 1 カスタム MT の出力例

BLEU スコアは、英日の場合、文書特化型エンジンで 51.11、医薬汎用型エンジンで 47.56 であった。日英の場合、文書特化型エンジンで 44.50、医薬汎用型エンジンで 40.82 であった。つまり、MT を医学・医薬分野の中の特定の文書種に特化させることで

Custom MT + PE in medical domain: A case study.
Masaru Yamada[1] Takeshi Hayakawa[2]
[1]Rikkyo University, [2]ASCA Corporation
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International Public License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

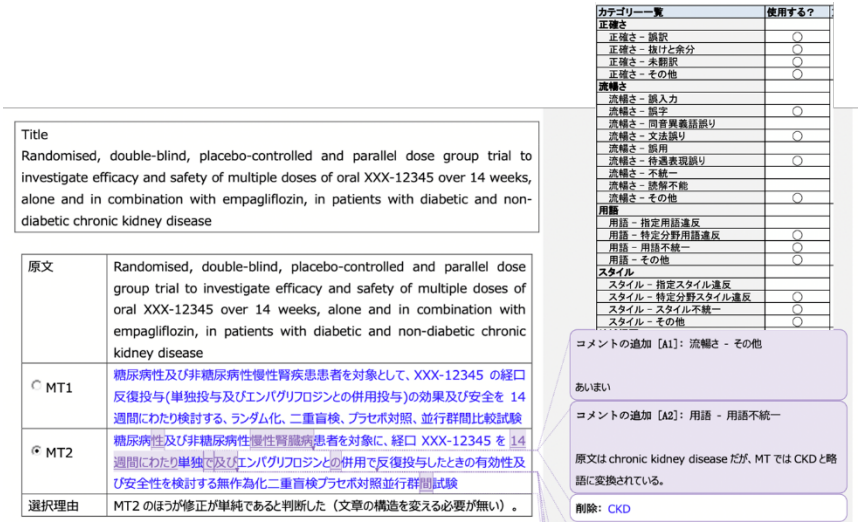

図 1 実験作業環境の例

BLEU スコアは両言語方向において向上した。

本研究の目的は、このようにして構築した特化性の異なる 2 種類の MT をプロ翻訳者にフル PE してもらい、エラー・主観評価・修正量などの観点から、カスタム MT＋フル PE の「実力」を評価することである。

3. 実験デザイン

実験参加者は英日、日英それぞれでプロ翻訳者 3 名（計 6 名）であった。翻訳経験年数は最低 2 年以上、最高で 16 年以上のベテランである。ポストエディットの経験年数は 2 年未満がほとんどだった。

評価用テクストは、英日では医学文書 50 文、日英では治験文書 50 文を用意した。これらを文書特化型と医薬汎用型の 2 種類の MT にかけ、参加者には MT の選択とフル PE を行ってもらった。実験に先立って、参加者には ISO 18587（ISO, 2017）[1]に基づきフル PE の要求事項を説明した。

実験手順は以下の通りである。

1. 2 種類の MT 出力からいずれの出力が PE しやすいかを判断し、MT を選択する。（評価はセグメント（おおむね 1 文）毎に行った。どちらの MT が特化型かは参加者には伏せた状態）。
2. 選択理由を書く。
3. 選んだ MT をフル PE する。
4. 修正箇所にエラーカテゴリーの付与をする（エラーカテゴリーは JTF 品質評価ガイドライン（日本翻訳連盟, 2018）[2]の一部を使用）。
5. 修正箇所について修正理由を書く。

全てのテクストのフル PE と上記の作業が終了した後に、主観品質評価・効率性・認知負荷・満足度などの質問項目を含むアンケートを実施した。

図 1 は実験作業環境の例である。作業はワードマクロを組み込んだ MS Word 上で行った。上部に 1 つの原文テクスト全体が表示され、その下にはセグメントごとに原文、2 種類の MT 出力と MT の選択ボタン、そして MT の選択理由を記入するボックスが表示される。各修正箇所にはエラーカテゴリーのオプションが表示され、カテゴリーを選択すると修正理由をコメントできる。

4. 結果

4.1 MT の選好性

英日方向の実験において、文書特化型エンジン（MT1）と医薬汎用型エンジン（MT2）のうち選ばれたセグメントの割合を比べたところ、MT1 は 64%、MT2 は 36%であった。χ^2 検定の結果、MT1 と MT2

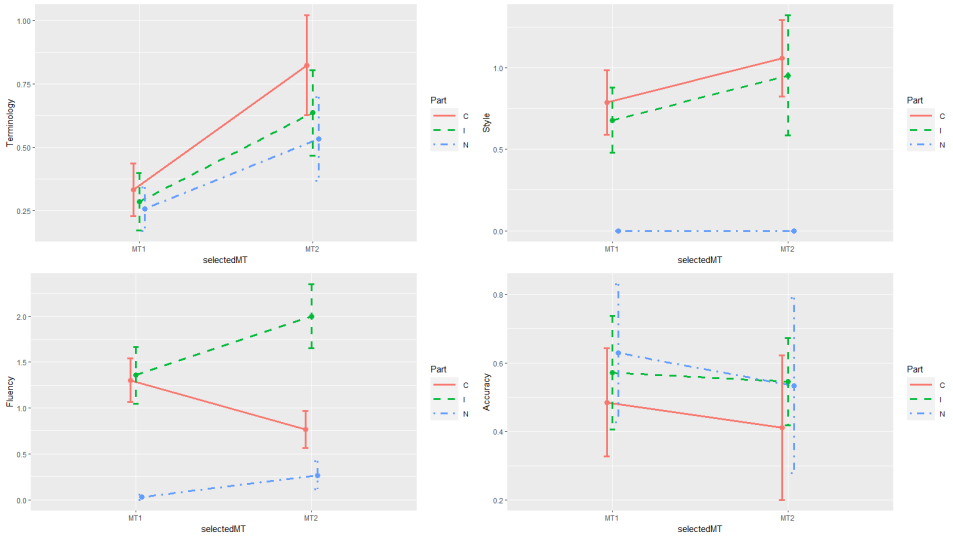

図 2 MT 間のエラー数の違い（英日）

の間には統計的に有意な違いがあった（χ^2(1)=11.76、p<.001、ϕ=.28）。

同様に日英方向の実験において、文書特化型エンジン（MT1）と医薬汎用型エンジン（MT2）の選ばれた割合を比べたところ、MT1 は 43%、MT2 は 57% であった。しかし、χ^2 検定の結果、MT1 と MT2 の間に統計的に有意な違いはなかった。

4.2 MT の選択とエラーの関係

まず英日方向のフル PE についての分析結果を報告する。図 2 は、文書特化型エンジン（MT1）と医薬汎用型エンジン（MT2）でのエラー数の違いを示す。参加者ごとに線の種類を変え、エラータイプ（用語・スタイル・流暢さ・正確さ）ごとにグラフを分けた。丸い点はエラー数をセグメント数で割った平均値で、エラーバーは標準誤差を表す。

図 2 で、例えば、左上の用語エラーについては、MT1 のほうが MT2 よりも平均エラー数が少ない傾向がはっきりとみえる。具体的には、用語エラー（terminology）において、MT1 では 1 セグメント当たりのエラー数は、実験参加者 N において、平均 0.25 個である。しかし、MT2 では、0.50 強と約 2 倍に増えているのが観察できる。これに対して、スタイル、流暢さ、正確さエラーについては参加者全員に共通してみられるような傾向はない。

これらのグラフの傾向をみて、どの種類のエラーの増減が MT1 か MT2 を選択するのかの決定に寄与しているのかを調べるためのモデルを立てた。被説明変数を MT1 と MT2 の選択、説明変数を各セグメントにおける各エラーカテゴリー（用語・スタイル・流暢さ・正確さの 4 種類）のエラーの数、そして変量効果を 3 名の参加者とした一般化線形混合モデル（generalized linear mixed effect model）を立てた。

その結果、表 2 に示すように用語エラーのみに統計的有意な差が見られた。つまり、MT1 と MT2 の選択に、用語エラー（の数）が深く関わっている。しかしスタイル・流暢さ・正確さエラーはあまり大きな影響を与えていないことが明らかとなった。またこれに加えて、被説明変数と変量効果は同じで、説明変数を各セグメントにおける Edit Distance Rate（EDR≒修正量）としたモデルも検討したが、有意にはならなかった。

日英方向のデータについても同様の分析を行った。しかし、参加者全員に共通する傾向はどのエラーカテゴリーにもみられなかった。また、一般化線形混合モデルも全てのエラーカテゴリーで有意な結果にはならなかった。つまり、MT1 と MT2 の違いに深く関連するエラーカテゴリーを特定することはできなかっ

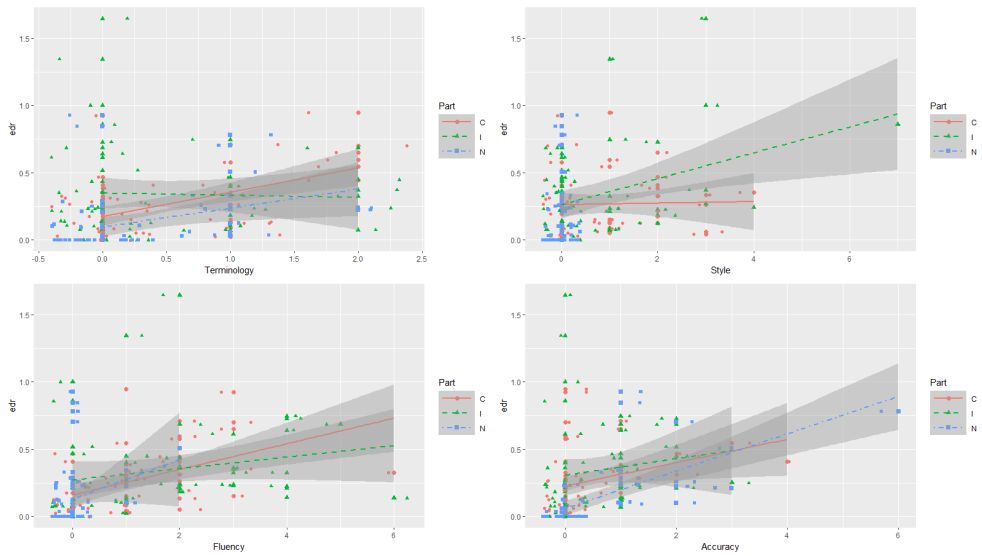

図 3 EDR とエラー数の関係（英日）

た。

	selectedMT			
Predictors	*Odds Ratios*	*std. Error*	*CI*	*p*
(Intercept)	0.38	0.08	0.25 – 0.58	**<0.001**
Terminology	2.36	0.63	1.40 – 3.98	**0.001**
Random Effects				
σ^2	3.29			
$\tau_{00\ Part}$	0.00			
N_{Part}	3			
Observations	150			
Marginal R^2 / Conditional R^2	0.089 / NA			

表 2 一般化線形混合モデル（英日）

4.3 修正量とエラーの関係

まず、英日方向の結果を示す。図 3 は EDR とエラー数の関係を示す。y 軸は EDR で、x 軸は 1 セグメントにおける各エラーカテゴリー（用語・スタイル・流暢さ・正確さ）のエラー数である。散布図の点は結果がわかりやすいように多少 x 軸方向にずらして表示している（そのためマイナスにみえるデータがある）が、回帰直線は散布図に依存しないように表示しているので解釈には影響しない。3 種類の線は参加者を表す。下 2 つの流暢さ、正確さエラーのグラフをみると、どの参加者でも MT が出すエラー数が多いセグメントほど EDR（フル PE での修正量）も大きくなる傾向がある。それに対して、上 2 つのスタイルエラーと用語エラーのグラフでは、参加者全員に共通するような傾向はみられない。

これらのグラフの結果にたいして、被説明変数を各セグメントの EDR、説明変数を各セグメントにおける各エラーカテゴリー（用語・スタイル・流暢さ・正確さ）のエラーの数、変量効果を 3 名の参加者とした線形混合モデル（linear mixed effect model）をたてた。まず、説明変数がひとつのみのモデルを 4 つのエラーカテゴリーそれぞれで検討した結果、4 つ全てで統計的有意な結果になった。その結果を踏まえ、説明変数を複数組み合わせたモデルをたて、ステップワイズ法でモデル選択をした結果、説明変数を「スタイル+流暢さ+正確さ」としたモデルが選ばれた（表 3）。この結果から、スタイル、流暢さ、正確さエラーが EDR に大きな影響を与えていることがわかった。

	edr			
Predictors	*Estimates*	*std. Error*	*CI*	*p*
(Intercept)	0.10	0.03	0.05 – 0.16	**<0.001**
Fluency	0.06	0.01	0.04 – 0.09	**<0.001**
Accuracy	0.09	0.02	0.05 – 0.13	**<0.001**
Style	0.07	0.02	0.03 – 0.10	**<0.001**
Random Effects				
σ^2	0.06			
$\tau_{00\ Part}$	0.00			
N_{Part}	3			
Observations	150			
Marginal R^2 / Conditional R^2	0.293 / NA			

表 3 線形混合モデル（英日）

この結果は、すなわち、翻訳者が MT 出力をたくさんフル PE しなければならない時というのは、その MT 出力の中に、スタイルエラー、流暢さエラー、正確さエラーが多く含まれている時であることを表す。この中に、用語エラーが含まれていないということは、逆にいうと、プロ翻訳者が良い MT エンジン（MT1 か 2）を選ぶ理由は、用語が正しく訳出されているからであると推察される。

日英方向のデータについても同様の分析を行った。EDR と各エラーカテゴリーのエラー数の関係を観察したところ、3 名の参加者全員に共通して、用語・流暢さ・正確さエラーが多いセグメントほど EDR が高くなっている傾向がみられた。しかし、スタイルエラーについてはそのような傾向はないようであった。この結果について英日の場合と同様に、被説明変数を EDR、説明変数を各エラーカテゴリーのエラーの数、変量効果を 3 名の参加者とした線形混合モデルをたてて検証したところ、説明変数を「用語+流暢さ+正確さ」としたモデルが選ばれた（表 4）。したがって、日英においては、用語、流暢さ、正確さエラーが EDR に大きく影響していることが予想される。

	edr			
Predictors	*Estimates*	*std. Error*	*CI*	*p*
(Intercept)	0.11	0.04	0.02 – 0.20	**0.017**
Terminology	0.10	0.03	0.05 – 0.15	**<0.001**
Fluency	0.09	0.02	0.04 – 0.14	**<0.001**
Accuracy	0.07	0.02	0.04 – 0.10	**<0.001**
Random Effects				
σ^2	0.03			
$\tau_{00\ Part}$	0.00			
ICC	0.13			
N_{Part}	3			
Observations	150			
Marginal R^2 / Conditional R^2	0.244 / 0.341			

表 4 線形混合モデル（日英）

5. まとめ

本研究では、医学・医薬分野において特定の文書種に特化した MT を構築し、プロの翻訳者にフル PE を行ってもらうことで、エラー・主観評価・修正量の観点から MT の「実力」の分析を試みた。その結果、以下のような結論が得られた。

1. MT エンジンを医学・医薬分野の中でも細分化された文書種に特化させることで、BLEU スコアが英日・日英の両言語方向において向上した。
2. 英日方向においては、文書特化型エンジンが医薬汎用型エンジンよりも有意に多く選ばれた。
3. 英日方向においては、用語エラーが MT の違いに大きく関わっていることがわかった。
4. 英日方向においては、用語エラー以外が EDR の予測に大きく影響していた。これは 3 を裏付ける結果であると考えられる（ただし、今回は MT の違いに EDR 自体があまり影響を与えていないという結果であったことに注意が必要である）。
5. 日英方向においては、カスタム MT の品質の差が、フル PE 作業に影響していることが、はっきりとは確認できなかった。

謝辞

本研究の一部は、日本学術振興会科研費補助金基盤研究(S)「翻訳規範とコンピテンスの可操作化 を通した 翻訳プロセス・モデルと統合環境の構築」(研究課題番号:19H05660)の支援を受けて行われた。

参考文献

[1] International Standard Organization. (2017). *Translation services – Post-editing of machine translation output – Requirements* (ISO Standard No. 18587:2017). https://www.iso.org/standard/62970.html

[2] 日本翻訳連盟 (2018) 「JTF 翻訳品質評価ガイドライン」 [Online] https://www.jtf.jp/tq/pdf/jtf_translation_quality_guidelines_v1.pdf