「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

AAMT年次大会
2021年12月9日

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

2 医薬特化型エンジンについて

- 医学・医薬分野の翻訳の特徴
  - 高度な専門性を必要とする
    - 語彙の特殊性、構文の複雑さ
  - 要求品質の高さ
    - 人類の資産となりうる研究
    - 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

- 分野特化性の追求
  - 「医学・医療分野（メディカル分野）」という括りでは不十分
  - より細分化された分野・用途に応じた翻訳モデルの開発が必要

3 特化度を重視したMTモデルの開発

- 特化型MTモデルの評価
  - 特化のレベルが異なる2種類のMTモデルを構築し、比較
  - 訓練にはドメインアダプテーションを使用（基盤のモデルは同一）
  - 細分化された教師データを用意し、特化のレベル差を表現
    - 特定の文書種に特化した特化度の深いMT
    - 医学・医薬分野から広く教師データを収集した特化度の浅いMT
  - 教師データには公開済みデータを翻訳したものを使用
    - 教師データのクオリティの差は無視できる

評価対象のMTモデル

- 英→日
  - 文書特化型エンジン
    - 医学論文のテキスト13万文で訓練

  VS

  - 医薬汎用型エンジン
    - 多様な医学・医薬分野のテキスト40万分で訓練

- 日→英
  - 文書特化型エンジン
    - 医薬品臨床試験の計画書から抜粋したテキスト19万文で訓練

  VS

  - 医薬汎用型エンジン
    - 多様な医学・医薬分野のテキスト40万分で訓練（英日と同一モデル）

5 出力サンプル

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

研究の目的

- カスタムMT＋フルPEの実力はいかに？

- 翻訳現場での使用時の「実力」
- プロ翻訳者によるフルPEの実施
- 効率性、満足度などから評価

7 実験デザイン

- プロ翻訳者（３名 x ２）、英日、日英で検証
- フルポストエディット
- MT1 （文書特化型）vs. MT2（医薬汎用型）
- BLEUスコア
  - 英日： 51.11 vs. 47.56
  - 日英： 44.50 vs. 40.82
- 評価用テキスト (reference付き)
  - 日英：医学論文 約50文
  - 英日：治験文書 約50文

実験手順と評価指標

- プロ翻訳者（３名 x ２）、英日、日英で検証
- MT1 or M2を選択
- 選択したMTをフルPEする
- 修正箇所をエラーアノテーションする（JTF品質評価ガイドライン）
- 修正理由を書く

- タスク終了後にアンケート実施
- 作業効率、作業負荷・疲労、満足度等を調査

9 実験作業環境例

Title
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

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
○ MT1	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、ランダム化、二重盲検、プラセボ対照、並行群間比較試験
◉ MT2	糖尿病性及び非糖尿病性慢性腎臓病患者を対象に、経口 XXX-12345 を 14 週間にわたり単独で及びエンパグリフロジンとの併用で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間試験
選択理由	MT2 のほうが修正が単純であると判断した（文章の構造を変える必要が無い）。

カテゴリー一覧	使用する？
正確さ	
正確さ – 誤訳	○
正確さ – 抜けと余分	○
正確さ – 未翻訳	○
正確さ – その他	○
流暢さ	
流暢さ – 誤入力	
流暢さ – 誤字	○
流暢さ – 同音異義語誤り	
流暢さ – 文法誤り	○
流暢さ – 誤用	
流暢さ – 待遇表現誤り	○
流暢さ – 不統一	
流暢さ – 読解不能	
流暢さ – その他	○
用語	
用語 – 指定用語違反	
用語 – 特定分野用語違反	○
用語 – 用語不統一	○
用語 – その他	○
スタイル	
スタイル – 指定スタイル違反	
スタイル – 特定分野スタイル違反	○
スタイル – スタイル不統一	○
スタイル – その他	○

コメントの追加 [A1]: 流暢さ - その他

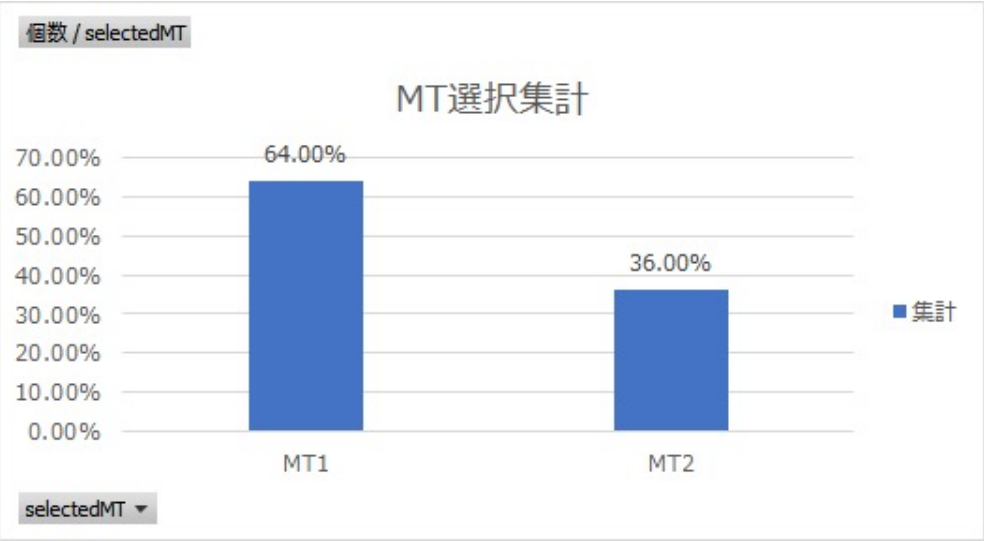
あいまい

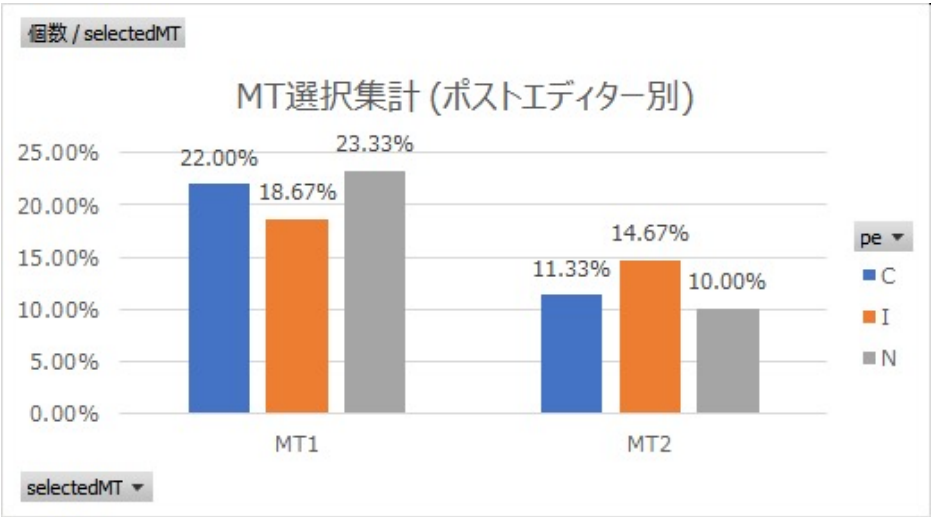
コメントの追加 [A2]: 用語 - 用語不統一

原文は chronic kidney disease だが、MT では CKD と略語に変換されている。

削除: CKD

MTの選好性（英日）

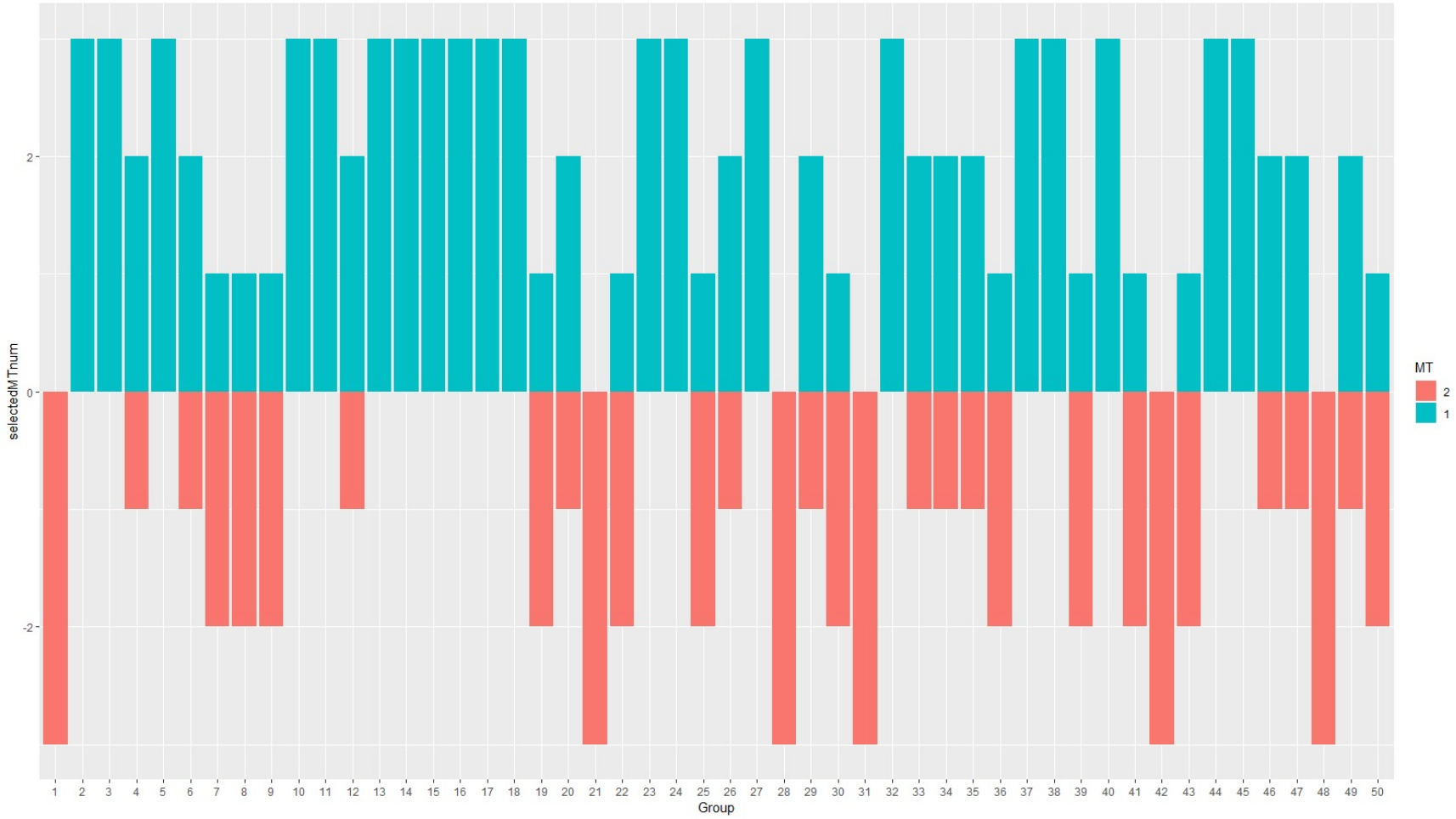




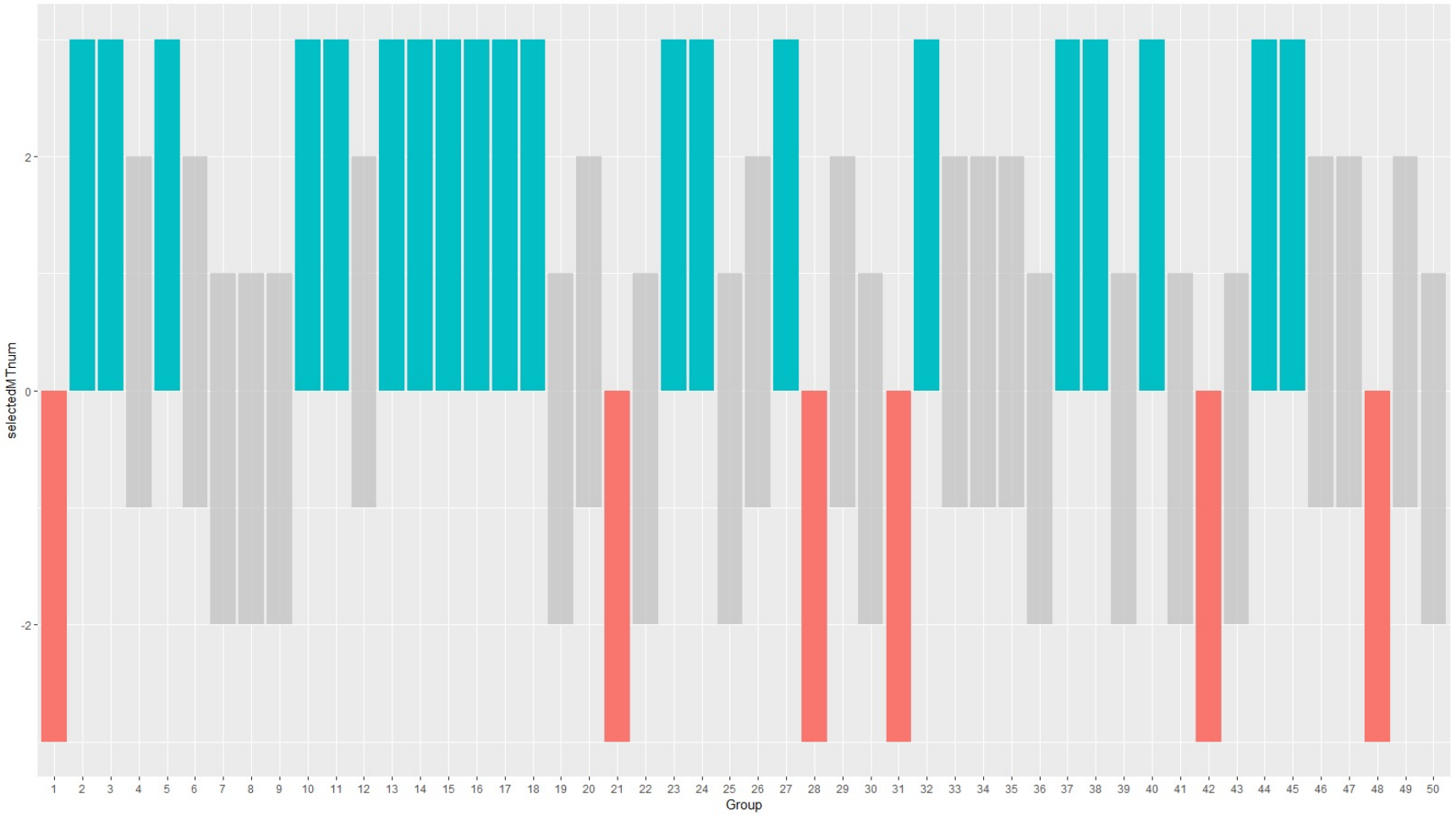
- MT1
- MT2

P<0.0001 (カイ二乗検定)

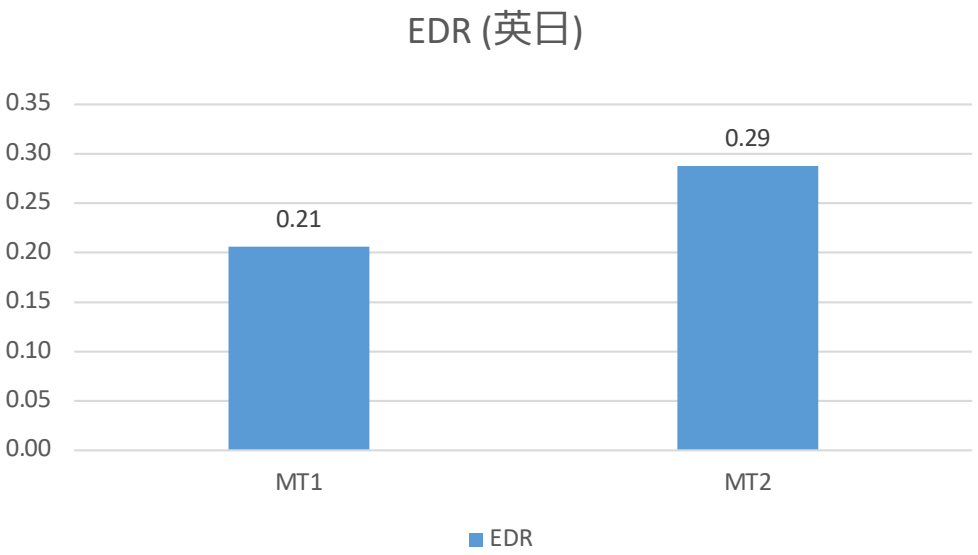
MTの選好性（英日）

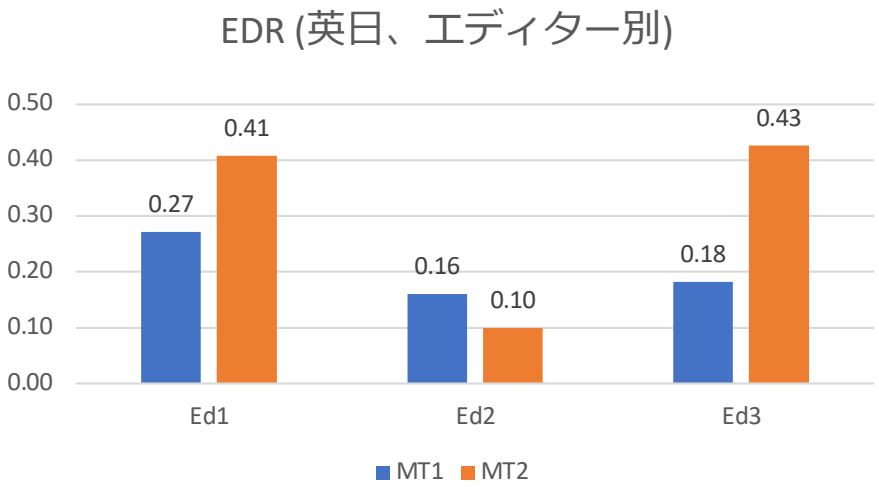

selectedMTnum
2
0
-2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Group
MT
2
1

MTの選好性（英日）


selectedMTnum
2
0
-2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Group

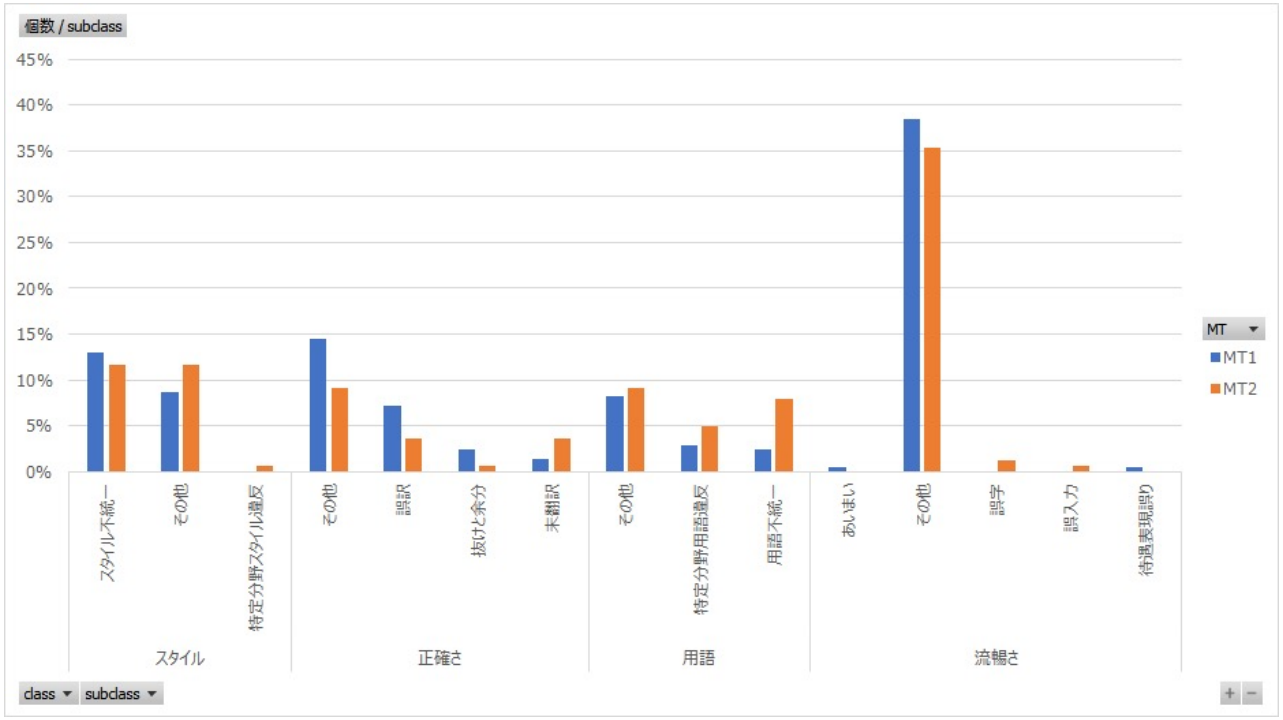
Edit Distance Rate





MTエラー分類（英日）

行ラベル	MT1	MT2	総計
スタイル	**22%**	**24%**	**23%**
スタイル不統一	13%	12%	12%
その他	9%	12%	10%
特定分野スタイル違反	0%	1%	0%
正確さ	**25%**	**17%**	**22%**
その他	14%	9%	12%
誤訳	7%	4%	6%
抜けと余分	2%	1%	2%
未翻訳	1%	4%	2%
用語	**13%**	**22%**	**17%**
その他	8%	9%	9%
特定分野用語違反	3%	5%	4%
用語不統一	2%	8%	5%
流暢さ	**39%**	**37%**	**38%**
あいまい	0%	0%	0%
その他	38%	35%	37%
誤字	0%	1%	1%
誤入力	0%	1%	0%
待遇表現誤り	0%	0%	0%
総計	**100%**	**100%**	**100%**



英日PE 主観品質評価

			MT1	MT2	
正確性	a	TRUE	100	80	MT1のほうが専門用語が訳出できていた気がします。
	b	TRUE	100	95	ほぼ同じ。
	c	TRUE	100	85	MT2の方が未翻訳と誤訳が多い印象でした。
流暢性	a	FALSE	80	100	MT2のほうは、やや長文の原文について日本語が自然なものが少し多かった気がしました。
	b	FALSE	95	100	両者に大差ない。特に流暢さは一見、MT2が優れているように思われるが、時に日本語の流暢さが原文の英語のリズムを失う。
	c	TRUE	100	80	語順や流れを理由にMT1を選んだことが多かったため。
用語	a	TRUE	100	80	MT1のほうが、正しい医薬用語の割合が多かったように感じました。
	b	TRUE	100	70	用語はむしろMT1の方が良かったように思います。
	c	TRUE	100	85	MT2は試験名の勝手翻訳が見られたため。
スタイル	a	TRUE	100	80	MT2は一般的ではない略語が出てくると、直訳気味になったり訳抜けが発生したように思います。
	b	TRUE	100	95	両者に大差ない。
	c	TRUE	100	95	どちらも（）の全角でないなどの問題があり、差はない。
		Mean	97.9	87.1	
		Wilcox.test. V = 66.5, p-value < 0.05			

総合	a	TRUE	向上率	60	はい、効率化になります
	b	TRUE	向上率	80	はい、効率化になります
	c	TRUE	向上率	80	はい、効率化になります
			Mean	73.3	

効率性、認知負荷、満足度

	時間的な効率性 （良かったほうのMTを使った時のPE作業と従来のHT/PEと比較）		認知負荷 （認知負荷とは、心的負担、作業的疲労の度合い）		満足度 （報酬は変わらない）	
a	はい、効率化になります	60%	はい、認知負荷は軽減されます	60%	変わりません	0%
b	はい、効率化になります	80%	はい、認知負荷は軽減されます	80%	はい、満足度は高くなります	80%
c	はい、効率化になります	80%	いいえ、むしろ認知負荷は増大します	-30%	はい、満足度は高くなります	30%
		73%		37%		37%

	翻訳者もしくはチェッカーとしての経験年数	PEの経験年数	翻訳メモリ（CATツール）用いた翻訳経験年数
a	10～15年	2年未満	2～5年
b	2～5年	2～5年	２年未満
c	16年以上	2年未満	10～15年

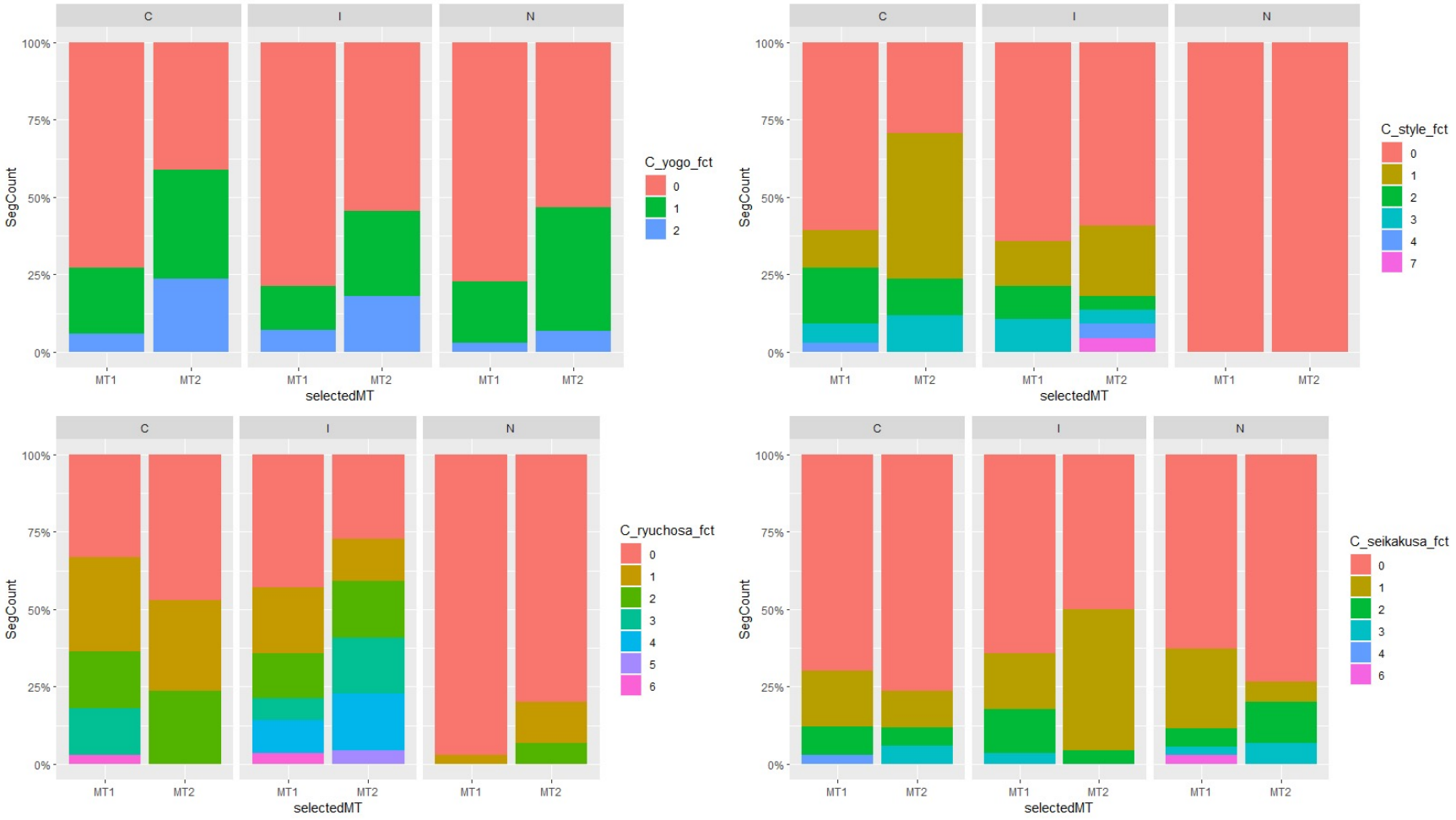
エラー分類とMTの選択の関係

GLMM= 一般化線形混合効果モデルで検証

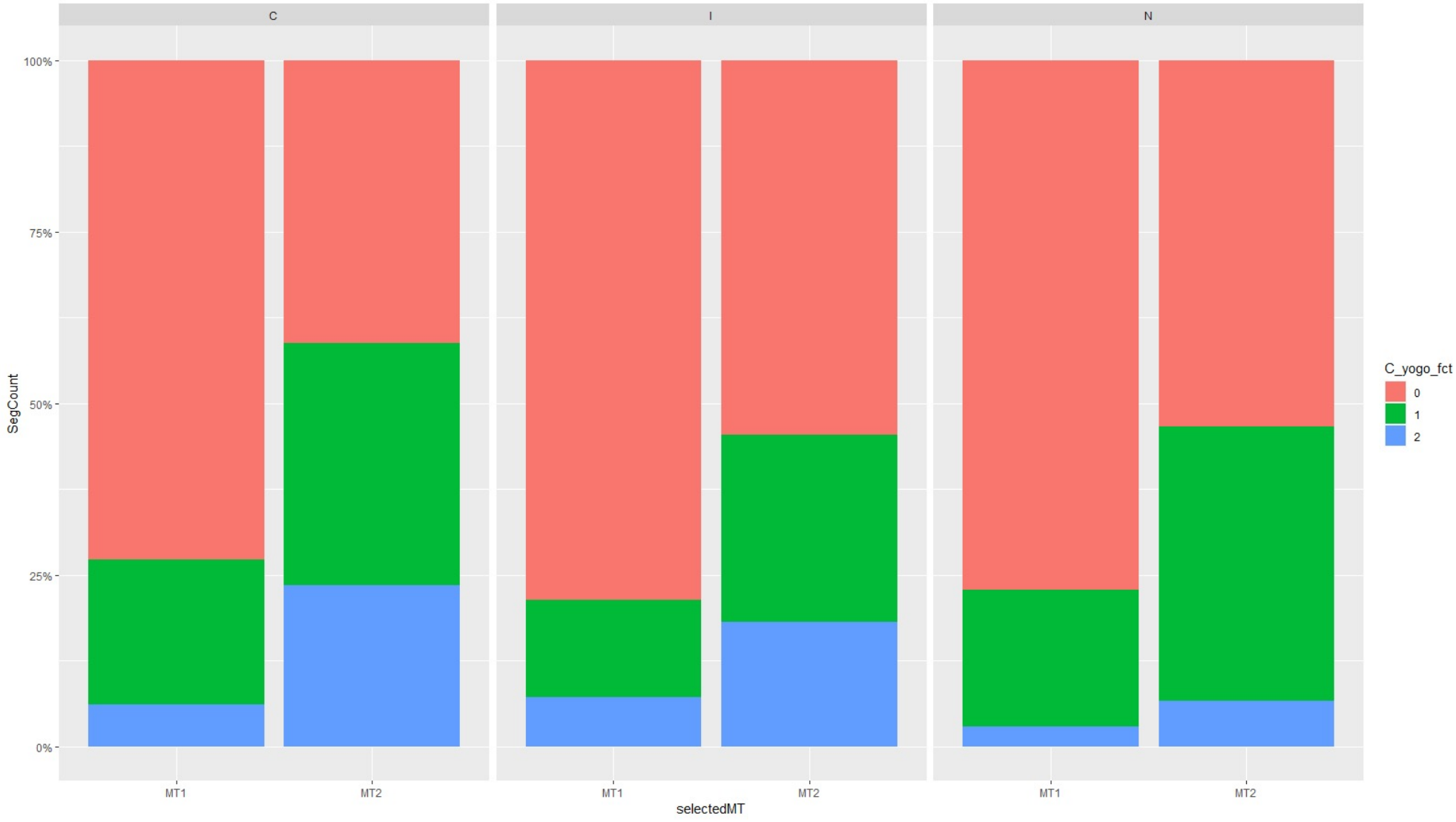
非説明変数： MTの選択（MT1 or MT1）の選択
説明変数　：品質エラー分類
変量効果：参加者

英日翻訳において、「用語（用語不統一）」のみに有意差を確認

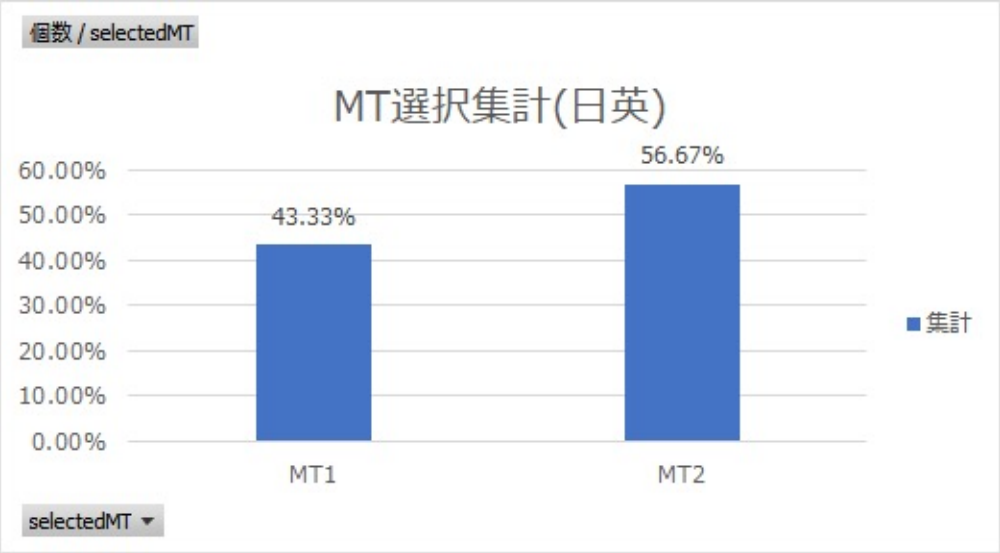
エラー分類、参加者、MT選択

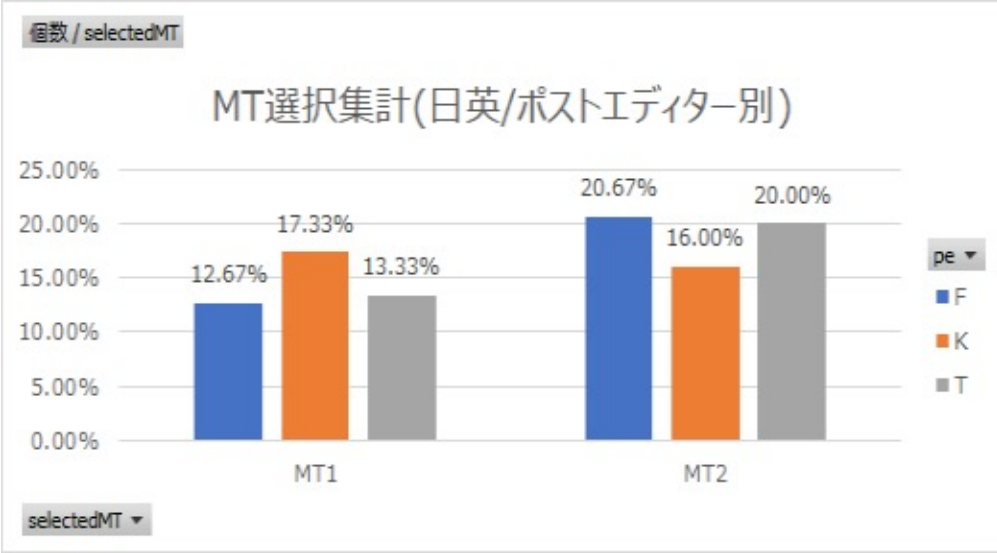


用語（不統一）、参加者、MT選択



MTの選好性 (日英)

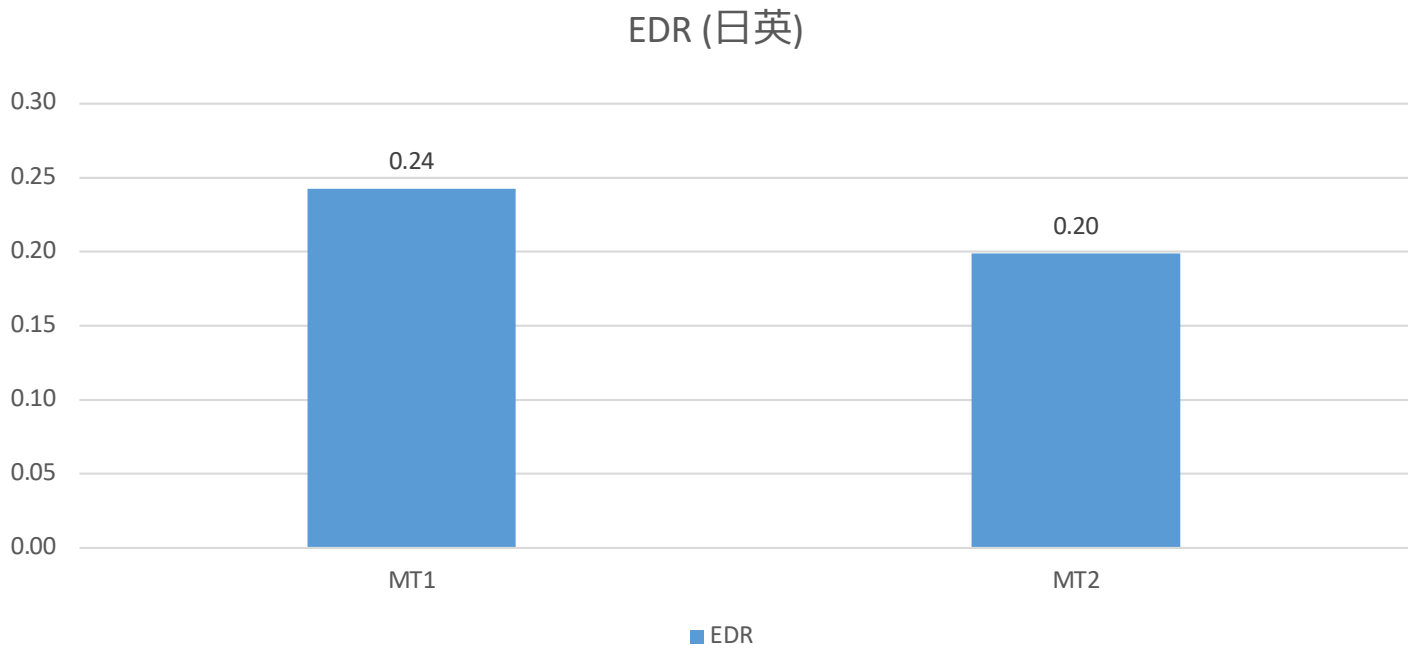




- MT1 = 文書特化型アダプテーションMT
- MT2 = 医薬分野MT

P<0.0001 (カイ二乗検定)

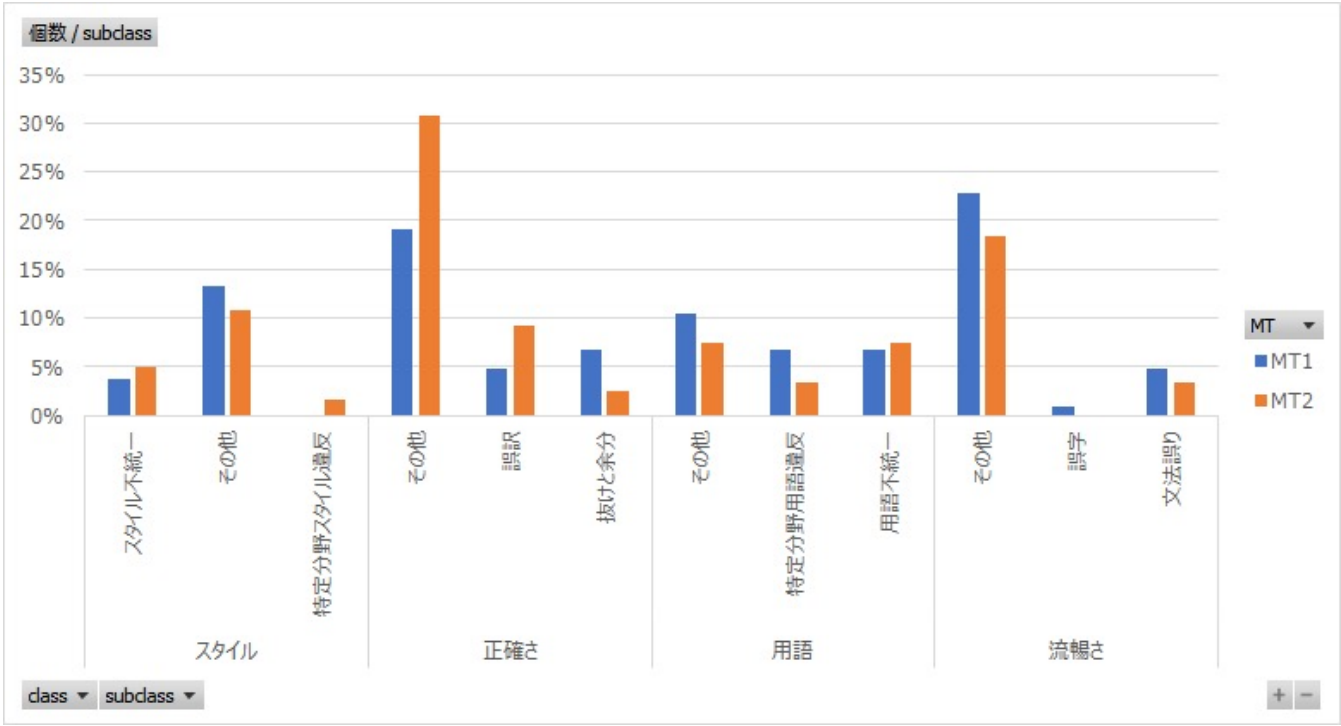
Edit Distance



- MT1 = 文書特化型アダプテーションMT
- MT2 = 医薬分野MT

MTのエラー分類 (日英)

行ラベル	MT1	MT2	総計
スタイル	**17%**	**18%**	**17%**
スタイル不統一	4%	5%	4%
その他	13%	11%	12%
特定分野スタイル違反	0%	2%	1%
正確さ	**30%**	**43%**	**37%**
その他	19%	31%	25%
誤訳	5%	9%	7%
抜けと余分	7%	3%	4%
用語	**24%**	**18%**	**21%**
その他	10%	8%	9%
特定分野用語違反	7%	3%	5%
用語不統一	7%	8%	7%
流暢さ	**29%**	**22%**	**25%**
その他	23%	18%	20%
誤字	1%	0%	0%
文法誤り	5%	3%	4%
総計	**100%**	**100%**	**100%**



日英PE　主観品質評価

正確性	e	FALSE	70	100	作業中にははっきりとした感想がなかったので、作業終了後に見直しました。 MT1とMT2のどちらかを選ぶ際に、40%以上でMT2。修正のしやすさ以外の理由で選択をした場合には、MT2を選ぶことがかなり多かったので、MT2の方が正確だったのではないかと思います。ただ、70点という数字にあまり根拠はありません。
	f	FALSE	70	100	正確さを理由に選んだ割合で出しました。
	g	FALSE	85	100	I found almost no difference in accuracy. Both contained isolated errors. MT1 seemed to have more sentences requiring no changes.
流暢性	e	TRUE	100	95	作業中に差は感じませんでした。修正のしやすさではMT1を選んだことの方が少し多かったので、語順などの点でMT1の方がやや自然だったのかもしれないと思ったので、MT2を95点としましたが、95という数字にあまり根拠はありません。
	f	FALSE	60	100	最終的に選択したMTの全体の割合により出しました。
	g	FALSE	85	100	MT1 seemed to have more sentences requiring no changes.
用語	e	TRUE	100	80	MT1のほうが、正しい医薬用語の割合が多かったように感じました。
	f	TRUE	100	70	用語はむしろMT1の方が良かったように思います。
	g	TRUE	100	85	MT1を選択しても修正するときにMT2を参考にしたこともあったため。MT1は試験名の勝手翻訳が見られたため。
スタイル	e	FALSE	95	100	これも、作業中に差は感じませんでしたが、見直すとMT2の方が少しよいと思われました。
	f	FALSE	60	100	最終的に選択したMTの全体の割合により出しました。
	g	FALSE	85	100	MT1 had more sentences requiring no changes, but I selected more MT2 sentences.
		Mean	84.2	94.2	
			V = 66.5, p-value > 0.05		

総合	e	FALSE	向上率	30	はい、効率化になります
	f	FALSE	向上率	0	変化なし
	g	TRUE	向上率	70	はい、効率化になります
			Mean	33.3	

効率性、認知負荷、満足度（日英）

		時間的な効率性 （良かったほうのMTを使った時のPE作業と従来のHT/PEと比較）		認知負荷 （認知負荷とは、心的負担、作業的疲労の度合い）		満足度 （報酬は変わらない）	
JE	e	はい、効率化になります	30%	変わりません	0%	変わりません	0%
	f	変化なし	0%	はい、認知負荷は軽減されます	30%	はい、満足度は高くなります	30%
	g	はい、効率化になります	70%	はい、認知負荷は軽減されます	25%	はい、満足度は高くなります	35%
			33%		18%		22%

まとめ

- カスタムMT（ドメインアダプテーション）は、翻訳品質の精度を上げる
- BLEUスコアで 5 点程度向上する（英日、日英）

- BLEUスコアの差は、英日のプロ翻訳者のPE作業によって確認された
- カスタムMTが選ばれる比率が優位に多かった
- 用語エラーの向上が寄与していることが判明した
- それに伴い、修正量（EDR）も少なくなった
- 品質面では、全般的に用語の正確性が向上する。他方で、流暢性やスタイルはあまり向上しないようである。
- 作業者の主観的な作業効率性の向上も認められた
- 作業に伴う認知負荷、満足度の向上も確認できた
- 次世代的な翻訳者は、より高い満足度を感じていることが推察された
- 今回のエンジンでは、日英のPEにおいては同様の結果が見られなかった

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。