

機械翻訳向けプリエディットの有効性と多様性の調査

宮田 玲 (名古屋大学) 藤田 篤 (情報通信研究機構)

Abstract

Machine translation (MT) systems are not able to always produce translations of sufficient quality. As a practical means of utilizing such MT systems, we investigated the potential of pre-editing strategy, i.e., rewriting source texts into more machine translatable ones. We first formalized a human-in-the-loop protocol, in which human editors incrementally rewrite source sentences aiming at better MT results. Implementing this protocol with a Japanese-to-English MT system, we then collected a total of 12687 pre-edit instances for 400 source sentences on four different datasets and obtained promising results; more than 85% of source sentences turned out to be accurately translated by the MT system. We also revealed that the pre-edited Japanese source sentences were better translated into Chinese and Korean, confirming the usefulness of pre-editing strategy in a multilingual translation setting. Through analyzing the collected pre-edit instances, we built a typology of edit operations comprising 53 types, which unveils the subjects for further research.

1 はじめに

近年、機械翻訳 (Machine Translation: MT) の実用化が急速に進んでいる。MT の性能向上と商用・無料システムの普及を背景として、多言語サイトの閲覧など個人レベルでの MT の利用が促進されるだけでなく、産業翻訳においても MT が導入されつつある (Jocelyne et al., 2017)。しかし、翻訳の目的、文書タイプ、言語方向によっては、MT は必ずしも実用レベルには達していない。また、ブラックボックス化している MT システムの中身を利用者が独自に調整することも容易ではない。MT を実社会で効果的に運用するためには、MT システムそのものの改良のみならず、文書作成・翻訳工程における MT の活用に関する実践的な手法の開発が必要である。

MT 訳の品質向上のためには様々なアプローチがある。例えば、対訳辞書登録といった MT システム自体のカスタマイズのほか、プリエディット (pre-edit) とポストエディット (post-edit) を挙げることができる¹。プリエディットは、所与の MT システムを用いて翻訳を行う前に起点テキスト (ST) を翻訳しやすいように書き換える作業である。他方、ポストエディットは、MT システムの出力を人手で修正することで、目標テキスト (TT) の品質を必要なレベルにまで高める作業である。近年の産業翻訳のワークフローでは、MT とポストエディットが合わせて導入されることが多い。またポストエディットについては、2017 年に国際標準

MIYATA Rei and FUJITA Atsushi, "Investigating the Effectiveness of Pre-Editing Strategy and the Diversity of Pre-Edit Operations for Better Use of Machine Translation," *Invitation to Interpreting and Translation Studies*, No. 18, 2017, pages 53-72. ©by the Japan Association for Interpreting and Translation Studies

化機構 (ISO, 2017) で規格化されるなど、要件定義が進んでいる。一方、リエディットは、上流工程で ST を制御することによって全体の生産性を高めることができる点で、とりわけ多言語展開の場面でのメリットが大きいとされる (ibid.)。

しかしながら、ポストエディットに比べると、リエディットの翻訳フローへの導入は遅れている。この理由としては次の 2 つが考えられる。1 つ目は、リエディットの有効性に関する実証的な研究が十分に進んでおらず、費用対効果、すなわち導入コストと翻訳フローの効率化の関係が明確になっていないことである。不完全な MT 訳を修正するポストエディットが最終的に翻訳品質を担保するために必須の工程であるのに対し、ST を編集するリエディットは、必ずしも MT 訳の品質が十分に高くなることを保証するものではない。2 つ目は、リエディットの全体像、すなわち言語表現の編集操作の多様性、および実用上の具体的な手続きが明確化されていないことである。「長文を避ける」「曖昧な表現を避ける」といった一般的に有用とされるリエディット用の書き換えルールは存在するが、より広範かつ個別の言語表現を対象とした書き換えルールに関する体系的な調査はなされていない。また後述するように、使用する MT システム、言語方向、文書ドメインによって有効な書き換えルールの集合は異なるという点も、リエディットの範疇を分かりにくくする一因であろう。

これらの課題に対して、我々は、具体的に次の 2 つの観点から調査を行った。

- 有効性: リエディットと MT を組み合わせることで、どの程度の翻訳品質を実現できるか。
- 多様性: リエディットの作業において対象となる言語現象は、どの程度の広がりを持つか。

なお、本論文は、MT に有効なリエディットの具体的な方法 (書き換えルール) を提案するものではない。むしろリエディットの全体像、言語表現の編集操作を広範に観察するための方法論の提案、および実データを用いた基礎調査を目的とする。このような分析結果に基づいて、リエディット用の書き換えルールの作成や自動リエディット手法の開発が可能となる。

2 関連研究

2.1 リエディットとは

上述のように、リエディットとは ST を機械翻訳しやすい形に書き換えることである。「機械翻訳しやすさ (Machine Translatability)」とは、MT がよい結果を出せるような ST の性質である (Bernth and Gdaniec, 2001)。ここで注意すべきは、MT の挙動は完全には予測できないため、機械翻訳しやすさはリエディットの段階では確定できず、翻訳後に TT の品質を確認してはじめて把握できる、ということである。すなわち、一定の仮説やルールに基づいて ST を機械翻訳しやすく書き換えたつもりでも、必ずしも MT が期待通りの訳文を出力する保証はない。また、O'Brien (2006) や Miyata et al. (2015) が指摘するように、ST を書き換えることの有効性や書き換えの多様性は、翻訳言語方向・使用 MT システム・文書ドメインによって異なりうる。言語方向に関しては、ある目標言語 (TL) に翻訳しやすいからといって、他の TL にも翻訳しやすいとは限らない。リエディットで ST を制御することにより多言語の TT の品質向上が期待できるとされているが、実証的な評価はなされていない。使用する MT システムに関しては、例えば、Hartley et al. (2012) や Miyata et al. (2015) は複数の MT システムを用いて、ST 中の表現を制御することの効果を検証し、特定の MT システムに対してのみ有効な書き換えルールが少なくないことを明らかにしている。さらに、対象とする文書ドメインも、リエディットの範囲を見定める上で考慮すべきポイントである。特定の文

書ドメインで頻出する MT 訳の品質に関わる表現が、他の文書ドメインではほとんど出現しないこともある。プリエディットについて議論する際は、これらの条件を考慮する必要がある。

プリエディットと類似した概念として、ローカリゼーションの枠組みで主要な役割を持つ「国際化 (internationalization)」についても言及しておく。国際化とは、ソフトウェアなどの製品を様々な言語圏に展開する前に、あるいはそのような製品の設計段階で、言語的・文化的な慣習になるべく依存しない形に製品を整えるプロセスを指す (Pym, 2014)。これにより、特に製品を多言語で同時に展開する際の時間や翻訳コストを削減することができるとされる。国際化には、日付の表示形式や文字エンコーディングを含む様々なレベルの処理が含まれる。言語表現レベルでは、ST をより分かりやすく (翻訳しやすく) 整えておくための制限言語 (Controlled Language)² も注目されている (山田 2007)。制限言語とは、人間にとっての理解しやすさおよび機械的な処理しやすさを向上させるために、自然言語の語彙・文法・スタイルに一定の制限を加えた言語のことである (Kittredge, 2003; Kuhn, 2014)。制限言語の枠組みは、プリエディットとも密接に関係しているため、2.2 節でも取り上げる。

2.2 プリエディットの方法と有効性

プリエディットの具体的な手法について、人手プリエディットと自動プリエディットに分けて説明する。

まず、人手プリエディットについては、ルールに基づく方法とアドホックな方法がある。ルールに基づく手法は、規定の執筆・書き換えルールにしたがって、システムチックに原文を書き換える手法である。様々な書き換えの種類を抽出・同定してルールを作成し、テキストに適用するアプローチは、主に制限言語の枠組みで検討されてきた。制限言語の包括的な整理は本論文の範囲を超える。ここでは、「1 文をなるべく短くする」「省略表現を避ける」といった執筆ルールに関する制限言語に注目する。英語の制限言語の研究開発は、マニュアル等の技術文書を多様な読み手に提供するニーズに応じて、1980 年代頃から特に欧米圏の企業の主導で進められており、MT 向けの制限言語ルールも数多く存在する。これまで制限言語ルールを用いたプリエディットの有効性も様々な条件下で検証されており (Pym, 1990; Aikawa et al., 2007; Thicke, 2011)、人間の書き換え作業を機械的に支援する研究もある (Mitamura et al., 2003; Seretan et al., 2014)。近年は、日本語を対象とした制限言語・プリエディットの研究報告も増えている (Matsui and Magnusson, 2011; Hartley et al., 2012; Miyata et al., 2015)。また産業界でも、日本特許情報機構による「特許ライティングマニュアル」³ や各種の商用ベースの執筆ガイドライン⁴ が開発されてきた。

一方、アドホックな方法とは、翻訳エラーを引き起こすと思われる ST 中の表現を別の表現に書き換える試行錯誤的な方法である。例えば、逆方向の MT による折り返し翻訳文 (ある ST を訳した結果を、起点言語 (SL) に訳し戻したもの) を利用する方法が考えられる。Miyabe et al. (2009) は、折り返し翻訳文が ST と同じ意味になるように、ST を繰り返し書き換えるという方法を採用している。また、Uchimoto et al. (2006) は折り返し翻訳文を用いて、ST 中の機械翻訳しにくい箇所を検出する手法を提案し、人間の作業者が検出箇所を重点的に書き換えることができるようにしている。折り返し翻訳文を使わない手法としては、例えば、Resnik et al. (2010) は、MT 訳のエラーを TL 話者の作業者が検出し、それに対応する ST 中の表現箇所を自動で同定し、SL 側の作業者に書き換えを要求するというプロトコルを提案している。

自動プリエディットの手法としては、ルールに基づく手法と統計的な手法が提案されてきた。ルールに基づく手法では、上述のように、規定のルールにしたがって ST を書き換えるものである。基本的な言語

解析処理 (形態素解析・構文解析など) の結果を用いながら、書き換えるべき表現のパターンを検出し、規定の表現に置き換える様々な手法が開発されてきた (Shirai et al., 1998; 山口ら 1998; 吉見ら 2000; Yoshimi, 2001; 乾・藤田 2004)。語句の置換や文構造の軽微な変換であれば、高い精度で実現できるが、主語の補完や文構造の大幅な変更は自動的には難しい。他方、統計的な手法は、大量に集めた原文とその書き換えの対を学習データとして、同一言語間の書き換えモデルを学習する手法である (Sun et al., 2010, 南條ら 2012)。しかし、学習データとして必要な、人手によるプリエディットの書き換え事例は、通常の翻訳プロセスにおいて自然に蓄積されるものではなく、また意図的に収集することもされていない。このため、統計的な自動プリエディットは、翻訳から自動生成できる擬似的な書き換え文を用いるに留まっている。Mirkin et al. (2013) は、書き換え候補文の生成にテキスト平易化技術 (Specia, 2010) を用いているが、必ずしも MT 向けのプリエディットに適しているわけではない。統計的なプリエディット手法の実用化のためには、やはり人手による高品質なプリエディット事例の収集が不可欠である。

上述の先行研究を含め、これまでの研究のほとんどは英語を SL としており、日本語を SL とした場合のプリエディットの有効性はまだ十分に検証されていない。また、TL として複数言語を扱った研究もなく、多言語翻訳場面でのプリエディットの有効性は不明である。さらにこれまでの研究は、いずれの書き換え手法を採るにせよ、部分的な ST の書き換えに留まり、プリエディットの有効性の上限を調査するものではない。挙動の読めない MT に依存するプリエディットでは「どれだけ書き換えてもいい結果が出るとは限らない」という躊躇ゆえとも考えられる。これに対して我々は、今後の研究・実用の発展可能性を考えるためにも、プリエディット単体でどこまで MT 訳の品質を改善できるかを見極めることには意義があると考え。

2.3 プリエディットの全体像

プリエディットで求められる編集のレベルと扱われる言語現象について整理する。プリエディットの作業内容を具体的に整理した文献はないため、Mosso (2014) による一般的なエディタの業務分類を確認する。

- コピーエディティング (copy editing): 文法やスペル、用字・用語など、規定のルールにしたがってテキストを修正する作業
- スタイルエディティング (stylistic editing): 読み手に応じて語彙や文構造を調整したり、テキストを読みやすくしたりと、修正というよりもテキストを改善する作業
- 構造エディティング (structural editing): 内容をよりよく提示し、メッセージのパーツ同士の関係を明示化するためにテキストを再構成する作業
- 内容エディティング (content editing): トピックがカバーしている範囲に対して、追加・削除すべき内容を提案する作業

ISO (2017) によると、プリエディットは、広義には制限言語による入力、すなわちあらかじめ定められた表現ルールにしたがって ST を作成する作業と重なりとされる。狭義には、スペルチェック・フォーマットチェック・タグ処理など、ST を整える作業を指すとされる。このようにプリエディットは、比較的微細な言語表現レベルの書き換えが中心である。これらは、上記の分類によれば、主にコピーエディティングとスタイルエディティングの範疇に入る。また文を超えた構造レベルの書き換えや内容に踏み込んだ書き換えは、プリエディットというよりは、より上流の文書作成で扱われるべきものだろう。

それでは、言語表現レベルの書き換えとは具体的にはどのようなものか。1節で述べたように、プリエディットで観察される言語現象の体系的な調査を目的とした研究はこれまでほとんどない。プリエディットの多様性の調査が主たる目的ではないが、O'Brien (2003) は、既存の8つの制限言語ルールセットを比較して、共通するルールを調査した結果、全てに共通するルールは、文の長さ制限に関するもの1つのみであったと報告している。換言すれば、各ルールセットが対象とする言語表現の範囲は互いに大きく異なっている。各ルールセットは、合意のとれた網羅的なルールセットから、用途に合わせて適切な部分集合を選んで作られたわけではなく、作成者・作成機関独自の経験と判断で作成されたものであろう。まずは、プリエディットで対象となりうる言語現象の範囲を広く調査し、さらなる研究のための基礎を提供することが必要である。具体的な調査の方法としては、O'Brien (ibid.) にならい、既存の制限言語ルールセットを多数収集し、ルールの和集合をとることで、ある程度網羅的な書き換え体系を作ることが可能かもしれない。しかし、これまでに提案されてきたルールセットは、ほとんどの場合、その作成方法・過程が開示されておらず、そこで収録されているルール集合の位置づけを客観的に評価することが難しい。また特に日本語の場合は、既存のルールセット自体が少ないため、まずは実際のプリエディットの事例を探索的に調査する方法が有効であると考えられる。事例分析に基づく手法として、Miyata et al. (2015) は、「オリジナルのSTと、人間が書き換えたより機械翻訳しやすいSTを比較することで、MTに適した原文作成の知見が得られる」という仮説のもと、次の一連の手続きを示している。

1. MT訳が十分な品質に達するまで人間が原文の書き換えを繰り返す。
 2. 書き換え箇所を分析し、プリエディットに有効な書き換えの種類を同定する。
- 3種類の日英MTシステムを用いて、自治体ドメイン (特に税金・医療・教育など生活に関する情報) の日本語文100文を対象に、1名の作業者が書き換えを行い、最終的に選定した38の言語的な特徴を規制する形で制限言語ルールを作成している。しかし、ここで示される書き換えの手続きは、試行錯誤の過程を細かく記録はしていないため、MT訳の品質改善に寄与しうる書き換えの知見を取りこぼしている可能性がある。また、提示された38のルールは、あくまで特定の言語方向 (日英)、MTシステム (3種類)、ドメイン (自治体文書)、テキスト量 (100文)、作業者 (1名) のもとで抽出・作成されたものであり、いずれかを変えた時にどのような書き換えが見られるかは不明である。

以上のように、プリエディットの技術やルールは、経験的に蓄積されているものの、プリエディットの対象となる言語現象の全体像は十分に明らかになっていないだけでなく、全体像を明らかにするための方法論も確立されていないのが現状である。

3 プリエディット事例の収集

本論文は、人間による実際のプリエディット作業の分析を通じて、プリエディットの有効性と多様性を検証することを目的とする。このためにまず、Miyata et al. (ibid.) の手法を改良して、作業用のツールを構築した (3.1 節)。そして、複数の文書ドメインのSTに対するプリエディットを行い、事例を収集した (3.2 節および 3.3 節)。

3.1 事例収集プロトコル

Miyata et al. (ibid.) の手法と同様に、人間の作業者が、MT 訳の品質向上を目標として徐々に ST を書

き換える方法を採用する。Miyata et al. (ibid.) は、最終的な書き換え結果を原文と比較しているが、我々は、プリエディットの様相をなるべく広い範囲で観察するために、新たに以下の拡張を加えた。

- プリエディットの試行錯誤の過程を漸次的に細かく記録する。
- 過去の履歴に適宜戻ることを許し柔軟な書き換えを促進する。

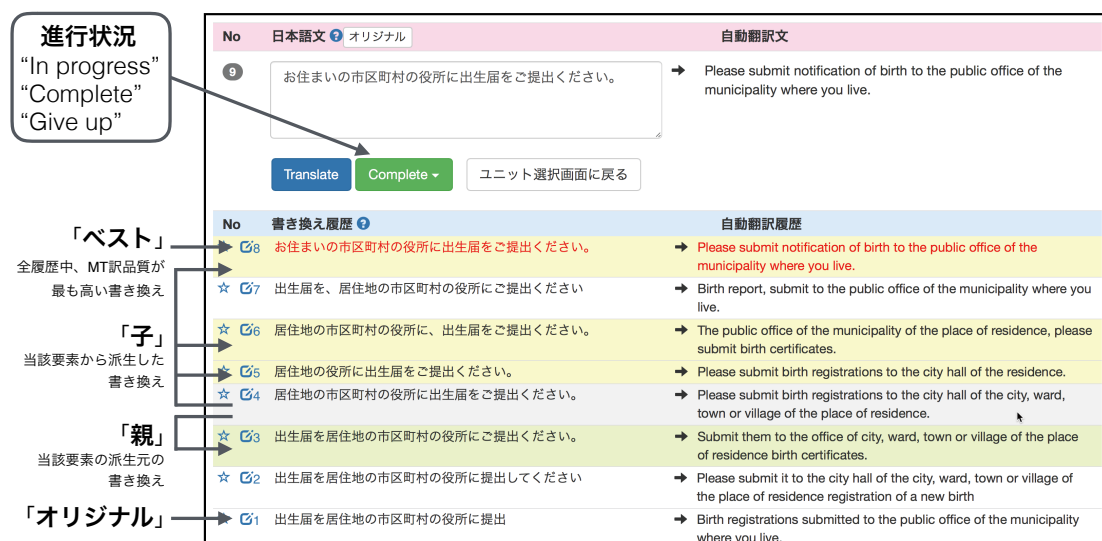


図1. 書き換え作業用ツール

具体的なプリエディット事例収集のプロトコルについて説明する。データの収集と作業の円滑化のために、図1に示す書き換え作業用ツールを作成した。この図に示すように、個々の日本語原文(「オリジナル」のST)について、その書き換え文および各々に対するMT訳を構造化したものを、「ユニット」と呼ぶ。作業者はユニットごとに次の手順でプリエディット作業を行う。

- Step 1.** 入力ボックスの文に対するMT訳の品質を評価する。十分な品質であれば進行状況を「Complete」として Step 4 に進む。十分な品質に達していなければ Step 2 に進む。
- Step 2.** 過去の書き換え履歴から、MT 訳も参照しながら、書き換える対象の ST を1つ選択し、Step 3 に進む。過去のどの書き換え履歴も MT 訳が十分な品質に達しておらず、かついずれを書き換えても MT 訳の品質改善が困難であると判断した場合は進行状況を「Give up」として Step 4 に進む。
- Step 3.** 原文の意味を保ちながら、最小単位の書き換えを行い、「Translate」ボタンを押す。その時点の ST と MT 訳の対が書き換え履歴の一番上の行に追加される。Step 1 に戻る。
- Step 4.** これまでの全ての書き換え履歴の中で MT 訳の品質が最も高いものを指定し(「ベスト」と呼ぶ)、このユニットの作業を終了する。

Step 1 および2におけるMT訳の「十分な品質」とは、表1の5段階の品質評価基準における「5. Perfect」もしくは「4. Good」とする。これまでに様々な翻訳評価方法・手段が提案されているが、本基準は比較的簡便に必要な最低限の評価をすることを目的として設計されたものであり、MT システムの研究開発の場面でよく使われている (Goto et al., 2013)。

作業には、Step 3 において、できるだけ最小単位の書き換え (藤田ら 2015) を行うよう指示した。表2に、「最小単位の書き換え」の説明に用いた例を示す。(a) の文は、語順変更により (b) に書き換えら

れ、さらに (b) の文は受動態への変更により (c) に書き換えることができる。例えば、(a) から (c) へ直接書き換えることは、最小単位ではない。

さらに Step 3 では、記録済の ST と同一の ST を二重に記録できないように制限をかけた。これにより、各ユニットにおいて書き換えの過程で蓄積される ST とそれに対応する MT 訳の対は、全て異なるものとなり、かつそのような対の集合は木構造を形成する (図 2 参照)。各ノードに付与された番号は、書き換え履歴の時系列を示し、オリジナル (「1」番) を除く全てのノードは必ず 1 つの親ノードから派生する。また全ノードのうち、1 つだけベストに指定される (図 2 では、「8」番がベスト)。N 個のノードがある時、直接的な関係にある書き換え前後のノードに着目すると、合計 (N-1) 対 (図 2 においては、1-2, 1-7, 2-3, 3-4, 4-5, 4-6, 4-8 の 7 対) を取り出すことができる。この数は書き換え試行総数を表す。このうち、ベストの ST はユニット中で最も MT 訳の品質が高いことから、オリジナルからベストに至る最短経路上の少なくとも 1 つ以上の書き換えが MT 訳の品質改善に寄与していることが保証される。以下、オリジナルからベストに至るパスを「ベストパス」と呼ぶ。例えば、図 2 においては灰色部分、1-2, 2-3, 3-4, 4-8 の 4 対がベストパスを構成する。

5. Perfect	原文の情報が完全に翻訳されている。訳文は文法誤りを含まない。母語話者から見ても、語句の選択が自然である。
4. Good	語句の選択がやや不自然であるが、原文の情報が完全に翻訳され、訳文は文法誤りを含まない。
3. Fair	原文のあまり重要でない情報の翻訳に些細な誤りがあるが、原文の内容は容易に理解できる。
2. Acceptable	重要な情報の一部が欠落したり、誤訳されたりしているが、原文の中核となる情報は何とか理解できる。
1. Incorrect/nonsense	原文の意味が全く理解できない。

表 1: MT 訳の品質評価基準

(a) 来院しなくても <u>十日前後で登録のクレジットカードから引き落としを行います。</u>
(b) 来院しなくても登録のクレジットカードから <u>十日前後で引き落としを行います。</u>
(c) 来院しなくても登録のクレジットカードから十日前後で引き落としが <u>行われます。</u>

表 2. 最小単位の書き換えの例

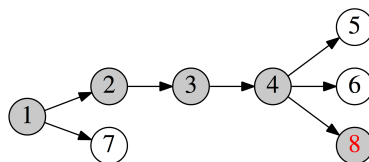


図 2. 木構造で示した書き換え履歴のパス (図 1 のデータに基づくサンプル)

3.2 事例収集

2.1節で述べたように、有効なブリエディットの範囲は、翻訳の言語方向、扱う文書ドメイン、使用する MT システムによって異なりうる。さらに、本手法では実際の人間の作業者が試行錯誤的に書き換えを行

うため、収集できる事例には作業者の能力や適性も反映されることに注意する必要がある。上記のプロトコルを用いて、プリエディット事例を収集するにあたり、今回の調査では以下の対象を設定した。

3.2.1 対象言語

翻訳の言語方向は日英とする。TLを英語とする理由は次の3つである。

- 翻訳ニーズ: 日本語から英語への翻訳市場は、他の言語への翻訳と比べて総じて大きい。
- 展開可能性: 英語は他の言語へのピボット言語となることが多いため、英語に最適化することで、派生的に他の言語への翻訳効率の改善が期待できる。
- 手法の妥当性: 言語構造が大きく異なる日英間のMTは難易度が高いとされ、本手法の適用可能性を検証する上で有用な基準となる。

3.2.2 MTシステム

使用するMTシステムの種類は、プリエディットの具体的な方法に関わる。今回の調査では、情報通信研究機構が開発、公開している統計的MT (SMT) である「みんなの自動翻訳@TexTra」⁵の「汎用」システム (2015年5月から2016年2月まで使用可能であった単一のシステムおよびモデル、以下、TexTra) をデフォルトの設定で使用した。SMTは、翻訳の対訳コーパスから翻訳モデル・言語モデルを学習する方式で、20年以上の研究開発の歴史があり、産業場面での導入実績も多数存在する。

その他のMTの方式としては、ルールベースMT (RBMT) やニューラルMT (NMT) がある。RBMTは、構文変換ルールと語彙変換ルールを用いてSTをTTに移し替える方式である。現時点では精度の観点からも、多言語展開の実現可能性の観点からも、他の手法に比べ遅れをとっているため、本論文では扱わない。NMTは、人手翻訳の対訳コーパスを用いてSTからTTまでの end-to-end のネットワークモデルを学習する方式である。近年研究・開発が急速に進み、特に翻訳の流暢さでSMTを凌駕しているが、挙動の制御しにくさや固有表現 (低頻度語) の処理の不安定さなどの課題があり、SMTに比べてSTの書き換えによってMT訳が変化しやすいことも知られている (高橋ら 2017)。また今後NMTの手法そのものや特性が大きく変化する可能性もあるため、今回の調査では比較的技術の安定しているSMTを採用した。

3.2.3 文書ドメイン

対象の文書ドメインは、病院内会話⁶、自治体生活情報⁷、新聞記事の3ドメインとした。新聞記事については、BCCWJ⁸の当該セクションとReuters⁹のコーパスを用いた。BCCWJは、もともと日本語で書かれた新聞記事のデータである。一方Reutersは、もともと英語で書かれた新聞記事を日本語に翻訳したものである。病院内会話は話し言葉で、その他は書き言葉である。平均文長もデータ間で大きく異なる。このように多様な文書ドメインを扱うことで、本プロトコルの適用可能性を評価することができる。今回の調査では、これら4つのデータセット (以下、病院、自治体、BCCWJ、Reuters と表記する) からそれぞれ、TexTraの翻訳モデルの構築に使用されている文を除外した後、ランダムに100ずつ日本語文を抽出した。

3.2.4 プリエディット作業者

書き換え作業は、日本語を母語とし、かつ英語の翻訳文の品質を適切に評価できる者に依頼した。全データセットに対する書き換え作業は、これまでMT訳の評価（対象言語は英語）の経験が豊富にあり、MTの挙動や癖にもある程度理解がある作業者（作業者Aとする）に依頼した。なお書き換え作業そのものは、個人の力量・経験のみに頼るため、作業者によって書き換え方法や終了の判断にばらつきが出ることが予想された。そこで、作業者間の相違の有無を把握するため、予備的にReutersのデータセットに対しては、追加の作業者（作業者Bとする）を割り当てた。作業者Bは、言語学的な知識は一定程度有しており、英語の運用能力もある程度高いが、MTの評価については未経験であった。

両作業者ともに、作業概要と作業用ツールの操作方法を確認し、サンプルの文で書き換えの練習を行った上で、本番の作業に取り組んだ。なお作業遂行上の課題として、作業期間中、作業者Aとは対面で頻繁にコミュニケーションをとりながら作業を進めることができた一方で、作業者Bとの作業のやりとりはオンラインテキスト上でのみ行われたことを付記しておく。

ドメイン/ データセット	作業者	ST 平均文字数 (標準偏差)	書き換え数				ユニット数	
			合計	平均	中央値	最大値	Complete/Give up	オリジナル=ベスト
病院	A	20.2 (4.5)	1199	12.0	3	105	97/3	40
自治体	A	34.8 (12.0)	2119	21.2	14	89	97/3	3
新聞 (BCCWJ)	A	44.4 (16.0)	3823	38.2	26	209	86/14	0
新聞 (Reuters)	A	62.0 (12.6)	5546	55.5	45	258	93/7	4
新聞 (Reuters)	B	62.0 (12.6)	2503	25.0	25	46	13/87	0

表 3. 収集した書き換え事例の記述統計

3.3 収集した書き換え事例の統計

日英 TexTra を用いて、提案プロトコルを3ドメイン（4データセット）に適用した結果の概要を表3に示す。作業者Aから12687の書き換え事例（400ユニット）を、作業者Bから2503の書き換え事例（100ユニット）を収集できた。個々のユニットにおける書き換え数は、Reutersに対する作業者Aの結果において、最大258回にも達している。原文に対して、同内容の異なる書き換え事例を大量に収集できたことは、プリエディットの多様性の解明を目的とした本プロトコルの有効性を示すものであろう。

作業者Aが作成した複数のデータセットの結果を比較すると、全体的には、原文の平均文長が長いほど、平均書き換え数が多い傾向が観察された。例えば、平均文長が20.2文字と最も短い病院データセットでは、平均して12.0回の書き換えが行われた。また書き換えをせずともMT訳が十分な品質に達しているユニットが4割あった。このくらい短い文であれば、MTの出力結果は比較的十分な品質に達しうること示している。一方、平均文長が62.0文字（病院の約3倍）の長さであるReutersでは、平均55.5回（病院の4倍以上）の書き換えが行われた。また平均文長が34.8文字の自治体データでも、プリエディットなしのSTに対するMT訳が十分な品質に達しているものはほとんどない（3/100件のみ）。これは長い文ほど、MT訳の品質が低くなりがちで、また書き換えるべき箇所も多いためであろう。

今回の調査では、Reutersのデータセットのみ、作業者Aと作業者Bの2人の作業者から書き換え事例を収集した。その結果、多様な書き換えを収集する上で、作業者の能力の影響が大きいことが明らか

になった。例えば、「Give up」の割合に大きな差があった。作業者 A は多くの試行錯誤を行い、ほとんどのユニットを「Give up」にしている。一方作業者 B は、作業者 A の半分程度の試行錯誤しか行っておらず、比較的早期に書き換えを断念する傾向にあった。しかも、オンラインでの断続的な教示のみでは、試行錯誤を強く促すことが難しかった。この点は、本調査を通じて明らかになった作業遂行上の課題であり、作業環境・指示方法を改善する必要がある。

さらに書き換え結果の分析から、作業者 A の方が多様な種類の書き換え方略を試していることが分かった。作業者 A の詳細は次節で述べるが、作業者 B は、書き換えの方略に乏しく、ほとんどの書き換えは「語順変更」「読点の追加・削除」「助詞の追加・削除・変更」のいずれかに分類できるものであった。そのため書き換えが比較的早く行き詰まる傾向にあったといえる。以上をふまえ、試行錯誤的なリエディット作業に必要なスキルや経験として、大きく次の 2 つがあると考えられる。

- 編集の能力・経験：リエディットで要求される言語操作能力は幅広く（詳しくは 5 節でまとめる）、文章作成・編集経験や言語現象に対する理解が求められる。普段の編集経験を作業者間で直接比較することは難しいが、作業者 A は普段から日本語や英語の文章作成・編集作業に従事しており、最初から編集能力が高かったと考えられる。
- MT の理解：本作業は MT の挙動を予想しながら ST を書き換えるものなので、MT についてある程度の理解があると有利である。作業者 A は MT の挙動に関してもともと理解があり、多様な書き換え方略を思いつくことができたと考えられる。

4 リエディットの有効性の検証

本節では、日英を対象としたリエディット作業の結果から、MT 訳の品質がどの程度向上したかを確認する（4.1 節）。また、原文とリエディット後の ST を、MT を用いて中国語・韓国語に訳し、その品質を評価することにより、多言語翻訳におけるリエディットの有効性を検証する（4.2 節）。なお以下の節では、分析対象として、作業者 A の結果のみを用いる。

4.1 英訳品質

英訳品質の観点から、リエディットの有効性を検証する。表 3 の作業者 A の「Complete/Give up」の列に示すとおり、どのデータセットにおいても、85%以上（平均すると 90%以上）の原文が、最終的に「Complete」となっていた。BCCWJ のデータに関しては、「Complete」となった割合が他のデータセットと比べてやや低い。全体的には原文をうまく書き換えることで、意味内容を保持したまま MT 訳の品質を十分に高くできることが示された。もちろん現実のリエディット作業としては、3.1 節のプロトコルに従って数十回以上の原文書き換えを行うことは時間・コストの制約上難しいが、本結果は、リエディットと MT の組み合わせが潜在的にどの程度有効であるかを示唆するものである。表 4 に、十分な訳文品質を達成した、実際の書き換え結果（原文およびベスト ST）の例を示す。Reuters のデータから抜粋したこのユニットの全体の書き換え回数は 194 回であり、ベストパスの長さ（原文からベスト ST に至る書き換え回数）は 25 であった。「語順変更」「語彙の言い換え」「名詞句・動詞句の交替」などの様々な種類の書き換えを経て、最終的に MT 訳が十分な品質に達していた。

一方、多数の試行錯誤によっても十分な品質に達しなかった 27 件のユニットについて見ると、とりわけ

固有表現の誤訳・語彙の選択誤り・不要語の湧き出しといった MT エラーが解消できていなかった。例えば、表 5 のベスト ST には、「4 日」「5 日」「6 日」という表現があり、そのまま単純に訳せば「4th」「5th」「6th」となるべきところが、不要な「June」が挿入されている。これは SMT システムの学習段階におけるフレーズ間の不適切な対応付け¹⁰に起因すると考えられる。このようなエラーはプリエディットで解消することは極めて難しく、翻訳モデルの学習あるいはポストエディットの段階で解消すべきものである。

以上のように、提案したプロトコルを用いることにより、プリエディットと MT の高い潜在能力を示すことができた。またプリエディットでは本質的に解決が難しい MT エラーの一部を明らかにすることができた。

	ST	MT 訳
オリジナル	同国は、前年の過剰輸出と、今年の減産によって、穀物不足に直面しており、大量の小麦輸入の計画を表明している。	Excess exports in the previous year, and reduced production this year, is facing a shortage of grain, a large amount of wheat imports plan.
ベスト	当年の減産と前年の過剰輸出による穀物の不足をふまえ、この国は小麦を大量に輸入する計画を表明している。	Based on the shortage of grain due to production cuts in the current year and excessive exports last year, this country has announced plans to import a large amount of wheat.
参照訳	The country, currently battling an acute grain shortage due to excessive exports last year, faces a poor harvest this year and intends to import large quantities of wheat.	

表 4. 十分な品質に達した事例 (オリジナルとベストのみ抜粋)

	ST	MT 訳
オリジナル	WSCによると、4日には弱い複数の降雨の可能性があるものの、5-6日には全般に乾燥した天候が戻る見通し。	WSC, although the possibility of weak more rainfall within 4 days, the weather in general dry return to 5-6 days.
ベスト	WSCによると、4日には弱い降雨の可能性が存在する一方で、5日から6日にかけては、乾燥した天候が全般に戻ってくる見込み、とのこと。	WSC said, while the possibility of a weak rain exists on June 4, from June 5 to 6, the dry weather comes back, in general.
参照訳	WSC said the outlook was for a chance of a few light showers on 4th, and generally dry conditions on 5th and 6th.	

表 5. 十分な品質に達しなかった事例 (オリジナルとベストのみ抜粋)

4.2 多言語翻訳品質

多言語翻訳におけるプリエディットの有効性も検証した。ここでは特に、ある TL (今回の調査では英語) で十分な品質の MT 訳を出せるように書き換えられた ST は、他の TL (今回の調査では中国語と韓国語) への翻訳においても良い MT 訳を得られるか、という問いに対する解を示す。まず、4 つのデータセットの全ユニット中の原文および作業者 A のベスト ST (のべ 800 文) を、TexTra の日中、日韓の「汎用」システムを用いて中国語と韓国語に翻訳した (2016 年 3 月に実施)。そして、各 TL について、MT 評価経験のある 1 名の評価者を割り当て、800 件の MT 訳を表 1 の 5 段階評価基準にしたがって評価させた。

表 6 に示すように、全データセットに関して、ベスト ST の MT 訳は平均的に、原文の MT 訳よりも品質が高かった。英語に翻訳しやすい日本語 ST は、中国語や韓国語にも翻訳しやすい傾向にあることを示している。とりわけ病院と BCCWJ の 2 つのデータセットに関しては、ベスト ST の MT 訳が原文の MT 訳よりも有意に高い品質となった。今回の調査では、あくまで英語を TL としてプリエディットの事例を集めており、他の言語への翻訳しやすさは考慮していないにもかかわらず、他の言語を TL とする同じ方式の

MT システムを用いた場合のプリエディットの有効性が定量的に示された。英語に翻訳しやすい ST の表現特徴と中国語・韓国語に翻訳しやすい ST の表現特徴が、部分的に重なることが示唆される。

中国語	平均値		ユニット数 (O vs B)			韓国語	平均値		ユニット数 (O vs B)		
	O	B	>	=	<		O	B	>	=	<
病院	2.73	2.93**	7	70	23	病院	3.32	3.56**	12	57	31
自治体	2.84	2.89	32	31	37	自治体	3.58	3.67	32	29	39
BCCWJ	2.39	2.75**	13	42	45	BCCW	3.37	3.60*	18	47	35
Reuters	2.61	2.77	22	45	33	Reuters	3.31	3.36	24	47	29

表 6. MT 訳の品質の人手評価結果: * と ** は、ベスト (B) の MT 訳の品質がオリジナル (O) の MT 訳の品質よりも有意に高いことを表す (ウィルコクソンの符号順位検定で各々 $p<0.05$, $p<0.01$)

5 プリエディットの多様性の調査

本節では、3 節で収集した人間によるプリエディット事例の一部を、分解・類型化することで、プリエディットの作業において対象となる言語現象の広がりを調査する。

5.1 分析対象と分析手順

ベスト ST は原文よりも高い品質の MT 訳を出すことから、3 節で述べたように、ベストパス中の 1 つ以上の書き換えが MT 訳の品質改善に寄与している¹¹。そこで、作業員 A の 4 データセットからランダムに選んだ各々 10 ユニット、合計 40 ユニットの対象として、ベストパスに含まれる書き換えを詳細に分析した。具体的には、ベストパス中の直接的な書き換え対を、以下の手続きで分析した。

Step 1. 最小単位の書き換えへの分解¹²

Step 2. テキスト表層上の差異の抽出

Step 3. 書き換えの種類の類型化

例えば、表 2 の (a) と (c) のように、「十日前後で登録のクレジットカードから引き落としを行います。」と「登録のクレジットカードから十日前後で引き落としが行われます。」という書き換え対があったとすると、まず「登録のクレジットカードから十日前後で引き落としを行います。」を中間の書き換えとするように分解する (Step 1)。その上で、「～で～から～」→「～から～で～」と「～を行います」→「～が行われます」という表層上の差異を取り出し (Step 2)、それぞれ「語順変更」と「態の変更」として種類を定める (Step 3)。

表 7 に書き換え対の統計情報を示す。Step 1 で 568 の書き換え対を分解した結果、979 の最小単位の書き換えを得た。この作業を通じてベストパス上の書き換え対の数が約 1.4～2.0 倍に増加したことからも、書き換えを最小単位で行うことは一般の作業員には難しい作業であると分かる¹³。理由としては、我々研究者と作業員の間で書き換えの最小単位に対する認識のずれがあったことや、実際の書き換えの自然なリズムと最小単位の書き換えは必ずしも一致せず、指示の遵守が難しかったことが考えられる。

データセット	ベストパス中の書き換え対の数		
	(a) 分解前	(b) 分解後	(b)/(a)
病院	97	185	1.91
自治体	106	186	1.75
BCCWJ	174	340	1.95
Reuters	191	268	1.40

表 7. サンプルした 10 ユニット中のベストパスにおける分解前後の書き換え対の数

5.2 事例の類型化と頻度

本論文末尾の付録に、上述のボトムアップな類型化を通じて抽出・同定した書き換え種類の一覧を、各データセットにおける事例の頻度および書き換えの例とともに示す。ID の先頭の英数字はそれぞれ S (構造)、C (語彙: 内容語)、F (語彙: 機能語)、T (語彙: ターミノロジー)、O (表記)、I (情報)、E (その他: エラー) という 7 種類の上位カテゴリを示す。書き換えの種類は 53 種類となり、979 の書き換え対はそのいずれかに分類された。この類型には、S07「読点の削除・追加」や S05「語順変更」といったテキスト表層レベルの書き換えだけでなく、構文的に複雑な書き換えも含まれる。

まず頻度の高い種類について述べる。C01「特定表現の使用」は、「一度」を「一回」に、「習得する」を「学ぶ」に書き換えるなど、内容語の変更にに関するものであり、どのデータセットでも最頻出の種類である。この種の書き換えは、特定の MT に最適化された語彙資源を準備することによって、自動化することができると考えられる。一方、I01「内容の変更」は自明な要素を推論して補完したり、副詞を加えたりする書き換えであり、機械的な自動化が難しい高度な書き換えも含まれる。この他、S05「語順変更」、S07「読点の削除・追加」、S21「接続表現」、F01「敬語化・非敬語化」は、どのデータセットでも比較的頻出し、かつ MT 訳の品質向上に貢献する事例も多く、ここからブリエディットに有効な知見を導くことが期待できる。

特定のドメイン (データセット) に特徴的な書き換えの種類も同定できる。例えば、S15「体言止めの回避・使用」や S20「接尾辞の使用・解除」は特に新聞ドメイン (BCCWJ や Reuters) で多く観察された。新聞では、「タイ農民銀行が売買代金で1位を獲得し、2バツ高の141バツ。」のような簡略的な書き方がよく使われる。これが MT 訳の品質に関わると判断され、「2バツ高の」→「2バツ上がり」(S20「接尾辞の使用・解除」)、「141バツ。」→「141バツとなった。」(S16「体言止めの回避・使用」) などの書き換えが試みられた。

試行錯誤的な書き換えを細かく記録した結果、多様な書き換えが収集できた。例えば、「ガッドゥム副総裁」という表現に対して、「副総裁ガッドゥム」「副総裁であるガッドゥム」「副総裁のガッドゥム」といった複数のバリエーションが得られた (S23「同格表現」)。この他、「懸念を強め」→「強い懸念を抱き」(S13「動詞句の主辞交替」)や「トイレトペーパーをカートに山積みにした客」→「トイレトペーパーで山積みのカートを押す客」(S10「視点変更」) といった、日本語のブリエディット・制限言語の先行研究にも例がなく、また比較的網羅的な言い換え分類表¹⁴にも含まれない、新しい書き換えも観察された。

6 考察

本節では、3～5 節の結果をふまえ、将来的なブリエディット研究や実用方法について議論する。

6.1 機械翻訳+リエディットの可能性

作業者 A のリエディットにより、特徴の異なる 4 つのデータセット全てについて 85% 以上という高い割合で、十分な品質の MT 訳を達成できた (4.1 節)。このことは、優れた編集能力を有し、かつ MT システムの挙動にも慣れている作業者が労力をかければ、大部分の ST は十分に機械翻訳しやすい形に書き換えられることを示しており、リエディットの潜在的な有効性を強く示唆する。しかし、この結果は、あくまで有効性の上限というべきものであり、特定の作業者や予算の制約下で行われる現実のリエディットで達成することは難しい。実際の翻訳フローでは、リエディットと MT で完結するわけではなく、ポストエディットも含めて全体の生産性向上を目指すことが肝要である。例えば、リエディットの多言語翻訳への有効性については確認できた (4.2 節)。今後は、翻訳フロー全体の生産性を見据え、リエディット導入のコストと多言語翻訳におけるポストエディットを組み合わせた場合の効率改善度合いを検証すべきである。さらに、リエディットではどうしても解消できない用語の誤訳の問題は、対訳用語資源の構築・管理・活用の枠組みから別途取り組む必要がある。

また MT の進歩に応じて、リエディットの有効性は都度検証しなければならない。本論文で提案した事例収集プロトコルおよび事例の分解・分析の方法論は、どの MT システムを扱う場合でも適用可能であるが、我々が実際に調査したのは 1 種類の SMT システムのみである。近年は、NMT が目覚ましい勢いで発展しつつある。これにより、リエディットの負担が減じるというメリットが想定される一方、これまで有効だったリエディットの方法が効かなくなる、という状況が生まれつつある。特に、NMT は挙動の予測や説明が困難である (高橋ら 2017) ため、そもそもリエディットのアプローチがどの程度有効か、改めて調査する必要があるだろう。

6.2 リエディット作業者に必要な要件

試行錯誤的なリエディットは、作業者の能力・経験に大きく依存することが明らかになった (3 節)。とりわけ、本論文で提案したプロトコルでは、TT を評価する能力に加えて、高いレベルの ST 編集能力・MT システムの挙動の把握が必要であるといえる。一方、実務でのリエディット場面では、試行錯誤的な方法は現実的ではなく、むしろ規定のルールに基づき機械翻訳しにくい表現を書き換える作業が中心となるだろう。具体的には、5 節で類型化したような多様な言語現象を把握し、文章中の書き換え箇所を的確に同定し、ルールにしたがって書き換えることが求められる。また、リエディットの導入形態にもよるが、一般的には TT の評価は必須ではなく、SL における編集能力が十分にあればよい。また、この能力も一定の体系的な訓練で身につけることができると考えられる。

専門職としての要件定義が進んでいるポストエディットと比較する。まず、ポストエディットでは文法的誤りや情報欠落を含んだ MT の出力結果の修正が主な作業となるが、リエディットでは人間の作成した文法的に正しい ST をさらに書き換える作業である。また、ポストエディットでは TT と ST の内容を的確に理解する能力が必要であるが¹⁵、リエディットでは必ずしも TL の能力は必要とされない。以上のように、リエディットに必要な能力は、ポストエディットや翻訳の能力よりも、技術情報の正確かつ効率的な執筆技術として発展してきたテクニカルライティングの能力に近いといえる。

今回の調査は既存の ST のリエディットという設定であったが、理想的には執筆段階で後工程の MT を見据えて制限言語にしたがった ST を作成できれば、リエディットが担う範囲は、言語依存・システム

依存の表現書き換えのみに縮減できる。また MT の進歩により、プリエディットで解決しなければならない問題は次第に減るだろう。つまり長期的には、プリエディットそのものの貢献の程度は小さくなると考えられる。したがって、プリエディタという固有の専門職を新たに作り出すよりも、テクニカルライターの業務内容にプリエディット (あるいは、機械翻訳向け制限言語) を組み込む方が望ましい。

6.3 自動化・書き換え支援に向けて

現実の翻訳フローにプリエディットを導入する上で、その一部の自動化は重要な課題である。2.2 節で見たように、プリエディットの自動化には課題が多く、実用事例もほとんどない。

今回の調査で収集したプリエディットの事例および類型は、自動化手法の開発に大きく貢献する。まず、人手で作成した 10000 件以上のプリエディット事例 (書き換え前後の ST の対) は、統計的な自動化手法の学習データとして活用可能である。つまり、「機械翻訳しにくい方の ST」と「機械翻訳しやすい方の ST」の対から、機械翻訳しやすく書き換える変換器を構築できると考えられる。ただし今回収集した試行錯誤的なデータには、書き換え前後で機械翻訳しやすさが向上している事例だけでなく、低下している事例も含まれる。そのため、学習データを準備する際は、各事例がどちらのタイプであるかを判別することが課題となる。また、作成した書き換え類型は、プリエディットルールとしてまとめることができる。そしてルールによっては、書き換えるべき表現のパターンを明確に定義し、自動変換ルールを構築することができる。単純な例を挙げれば、S31「になる・となる表現」は表層文字列の照合によって自動的に検出し、削除することが可能である。特に、今回頻出した S05「語順変更」、S07「読点の削除・追加」、S21「接続表現」、C01「特定表現の使用」、F01「敬語化・非敬語化」などは重点的に取り組むべき書き換え種類といえる。

なお完全な自動化が難しい書き換え種類もあるため、人間のプリエディット作業を支援する技術の開発も重要である。これまでに、ルールに違反した箇所を同定する制限言語チェッカーが各種開発されてきた (Mitamura et al., 2003)。例えば、S19「複合名詞の使用・展開」については、複合名詞の検出までは機械的に可能であるが、具体的な書き換え作業は自動化が難しい。全自動で可能か、人間の作業支援が現実的か、完全に人手でまかなうべきか、を判断し、それぞれに適した方法を開発することが重要である。5 節の分析を経て得た類型は、網羅性の観点で十分であると断定はできないが、このような判断の道標となりうる。

7 おわりに

本論文では、MT 訳の品質改善を目指して人間が試行錯誤的に ST を書き換える手法による、プリエディット事例の収集の方法論、およびそれに基づいて収集した事例の分析について報告した。プリエディットの有効性の観点からは、(1) 人間による適切なプリエディットにより、平均して 90%以上の MT 訳を誤りのない高い品質にまで引き上げることができ、(2) さらに MT を用いて英語に翻訳しやすい ST は、中国語・韓国語にも翻訳しやすい傾向にあることが分かった。これは、プリエディットと MT の組み合わせの高い潜在能力を示すもので、実用場面でのプリエディット導入に関する判断材料となるだろう。また多様性の観点からは、プリエディット事例を手作業で類型化することで、7 カテゴリー 53 種類の書き換えを抽出・同定し、ドメインごとの傾向を一部明らかにした。ここで構築した書き換えに関する類型は、プリエディット操作に関する全事象を網羅するものではないが、今後の研究に資する基礎を提供するものである。

今後は、主に以下の3つの研究に取り組む。第1に、本論文で提案したプロトコルを用いて、NMTに対するプリエディットの有効性を調査する。第2に、MTに有効な具体的なプリエディットルールを作成する。今回は基礎調査を目的としており、どの書き換え種類がMT訳の品質向上に寄与するのかまでは、明らかにしていない。Miyata et al. (2015) の定量的な評価手法を参考にしながら、効果のある書き換え種類を同定し、プリエディットルールとしてまとめたい。第3に、プリエディットの自動化手法と作業支援手法を開発する。ルールに基づく手法と統計的な手法の比較・融合を含め、より効果的な自動化手法を模索すると同時に、人間による書き換え作業をサポートするツールの開発も進めていきたい。

【謝辞】

本研究の一部は、科研費若手研究 (B) (課題番号:25730139)、科研費基盤研究 (A) (課題番号:25240051) の支援を受けた。データの一部は、総務省の情報通信技術の研究開発「グローバルコミュニケーション計画の推進 -多言語音声翻訳技術の研究開発及び社会実証-I. 多言語音声翻訳技術の研究開発」から提供を受けた。

【著者紹介】

宮田 玲 (MIYATA Rei) 名古屋大学大学院工学研究科・助教。2012年東京大学教育学部卒、2014年同大学院教育学研究科修士課程修了、2017年同博士課程修了。博士 (教育学)。2015年度に情報通信研究機構にて機械翻訳応用に関する研究に従事。2017年より名古屋大学工学部・大学院工学研究科 (情報・通信工学専攻) 助教。専門は図書館情報学・自然言語処理。

藤田 篤 (FUJITA Atsushi) 情報通信研究機構・主任研究員。2000年九州工業大学情報工学部卒、2002年同大学院情報工学研究科博士前期課程修了、2005年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。博士 (工学)。京都大学大学院情報学研究科産学官連携研究員、名古屋大学大学院工学研究科助手、同助教、公立はこだて未来大学システム情報科学部准教授を経て、2014年より情報通信研究機構ユニバーサルコミュニケーション研究所主任研究員、同先進的音声翻訳研究開発推進センター主任研究員。専門は自然言語処理。主に言い換え表現の生成と認識、機械翻訳の研究に従事。

【註】

¹ 用語としては「pre-edit / post-edit」を直接カタカナで表記した「プリエディット／ポストエディット」のほか、日本語に置き換えた「前編集／後編集」もあるが、本論文では、現在の翻訳研究分野で比較的広く用いられている前者を用いる (例えば、鳥飼 (2013: 72-73) 参照)。ポストエディットは「PE」と略記されることも多いが、本論文では同じくPEと略記されるプリエディットとの混同を避けるため、略記は使用しない。

² Controlled Natural Language という用語も使われている。本論文では、MTの応用分野でこれまで使われることが多かった Controlled Language に準拠する。日本語においても「制御言語」など他の訳もあるが、本論文では、現在主流である「制限言語」を用いる。

³ <http://japio-tjp.org/wmanual.html>

⁴ 例えば、日本語を対象としたものには Acrolinx による STE ルールセット (小倉ら 2010) などがある。

⁵ <https://mt-auto-minhon-mlt.ucri.jgn-x.jp>

⁶ NICTと東大病院による音声翻訳の模擬被験者実験で使用された会話文の対訳コーパスのうち、日本語部分を用いた。

⁷ 自治体国際化協会、新宿区、浜松市の3つのウェブサイトから抽出した日本語文を用いた。

- ⁸ 現代日本語書き言葉均衡コーパス, http://pj.ninjal.ac.jp/corpus_center/bccwj/
- ⁹ <http://www2.nict.go.jp/astrec-att/member/mutiyama/jea/reuters/index.html>
- ¹⁰ 例えば、「4 日」の翻訳として「June 4」というフレーズが学習されてしまった。
- ¹¹ 当然これらベストパス中の全ての書き換えが MT 訳の品質改善に寄与している保証はなく、またベストパスに含まれない書き換えにも MT 訳の品質改善に貢献しているものもある。これらの調査は、今後の課題である。
- ¹² 作業者による書き換えは、必ずしも十分に最小単位でなされていなかった。
- ¹³ 新聞 (Reuters) については、他の 3 データセットよりも細かく書き換えするように、改めて作業者に指示してあった。
- ¹⁴ 言い換えのあれこれ, <http://paraphrasing.org/paraphrase.html>
- ¹⁵ 近年は、TL のモノリンガル作業者によるポストエディットの研究も進められているが (Schwartz, 2014)、基本的には SL と TL の両言語の能力が要件とされる (ISO, 2017)。

【参考文献】

- Aikawa, T., Schwartz, L., King, R., Corston-Oliver, M. and Lozano, C. (2007). Impact of Controlled Language on Translation Quality and Post-Editing in a Statistical Machine Translation Environment. In *Proceedings of the Machine Translation Summit XI*, 1-7.
- Bernth, A. and Gdaniec, C. (2001). MTranslatibility. *Machine Translation*, 16 (3): 175-218.
- Goto, I., Chow, K. P., Lu, B., Sumita, E. and Tsou, B. K. (2013). Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *Proceedings of the 10th NTCIR Conference*, 260-286.
- Hartley, A., Tatsumi, M., Isahara, H., Kageura, K. and Miyata, R. (2012). Readability and Translatability Judgments for ‘Controlled Japanese’. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*, 237-244.
- ISO. (2017). *ISO 18587:2017 Translation services – Post-editing of machine translation output – Requirements*.
- Jocelyne, A., Samiotou, A., van der Meer, J. and Ruopp, A. (2017). *TAUS Machine Translation Market Report*. De Rijp: TAUS.
- Kittredge, R. (2003). Sublanguages and Controlled Languages. In Mitkov, R. (Ed.), *Oxford Handbook of Computational Linguistics* (pp. 430-437). Oxford: Oxford University Press.
- Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40 (1): 121-170.
- Matsui, J. and Magnusson, D. (2011). Six Pre-edit Techniques for Enhancing Japanese to English Machine Translations. *Interpreting and Translation Studies*. 11: 173-184.
- Miyabe, M., Yoshino, T. and Shigenobu, T. (2009). Effects of Undertaking Translation Repair Using Back Translation. In *Proceedings of the International Workshop on Intercultural Collaboration (IWIC)*, 33-40.
- Mirkin, S., Venkatapathy, S., Dymetman, M. and Calapodescu, I. (2013). SORT: An Interactive Source-Rewriting Tool for Improved Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, 85-90.
- Mitamura, T., Baker K., Nyberg, E. and Svoboda D. (2003). Diagnostics for Interactive Controlled Language Checking. In *Proceedings of the 4th Controlled Language Applications Workshop (EAMT/CLAW)*, 237-244.

- Miyata, R., Hartley, A., Paris, C., Tatsumi, M. and Kageura, K. (2015). Japanese Controlled Language Rules to Improve Machine Translatability of Municipal Documents. In *Proceedings of the Machine Translation Summit XV*, 90-103.
- Mossop, B. (2014). *Revising and Editing for Translators (3rd ed.)*. Oxon & New York: Routledge.
- O'Brien, S. (2003). Controlling Controlled English: An Analysis of Several Controlled Language Rule Sets. In *Proceedings of the 4th Controlled Language Applications Workshop (EAMT/CLAW)*, 105-114.
- O'Brien, S. (2006). Controlled Language and Post-Editing. *Multilingual*, October/November: 17-19.
- Pym, A. (2014). *Exploring Translation Theories (2nd ed.)*. New York: Routledge.
- Pym, P. (1990). Pre-Editing and the Use of Simplified Writing for MT. In Mayorcas, P. (Ed.), *Translating and the Computer 10: The Translation Environment 10 Years on* (pp. 80-95). London: Aslib.
- Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A. and Bederson, B. (2010). Improving Translation via Targeted Paraphrasing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 127-137.
- Schwartz, L. (2014). Monolingual Post-Editing by a Domain Expert is Highly Effective for Translation Triage. In *Proceedings of the 3rd Workshop on Post-editing Technology and Practice (WPTP)*, 34-44.
- Seretan, V., Roturier, J., Silva, D. and Bouillon, P. (2014). The ACCEPT Portal: An Online Framework for the Pre-editing and Post-editing of User-Generated Content. In *Proceedings of the Workshop on Humans and Computer-Assisted Translation (HaCaT)*, 66-71.
- Shirai, S., Ikehara, S., Yokoo, A. and Ooyama, Y. (1998). Automatic Rewriting Method for Internal Expressions in Japanese to English MT and Its Effects. In *Proceedings of the 2nd International Workshop on Controlled Language Applications (CLAW)*, 62-75.
- Specia, L. (2010). Translating from Complex to Simplified Sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 30-39.
- Sun, Y., O'Brien, S., O'Hagan, M. and Hollowood, F. (2010). A Novel Statistical Pre-Processing Model for Rule-Based Machine Translation System. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT)*, 8 pages.
- Thicke, L. (2011). Improving MT results: A study. *Multilingual*, January/February: 37-40.
- Uchimoto, K., Hayashida, N., Ishida, T. and Isahara, H. (2006). Automatic Detection and Semi-Automatic Revision of Non-Machine-Translatable Parts of a Sentence. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, 703-708.
- Yoshimi, T. (2001). Improvement of Translation Quality of English Newspaper Headlines by Automatic Pre-Editing. *Machine Translation*, 16 (4): 233-250.
- 藤田篤, 柴田知秀, 松吉俊, 渡邊陽太郎, 梶原智之 (2015). 「言い換え認識技術の評価に適した言い換えコーパスの構築指針」『言語処理学会第 21 回年次大会ワークショップ 自然言語処理におけるエラー分析 発表論文集』 11 pages.
- 乾健太郎, 藤田篤 (2004). 「言い換え技術に関する研究動向」『自然言語処理』 11 (5): 151-198.

- 南條浩輝, 山本祐司, 吉見毅彦 (2012). 「機械翻訳の品質向上のための対訳コーパスからの統計的ブリエディットシステムの自動構築」『情報処理学会論文誌』 53 (6): 1644-1653.
- 小倉英里, 工藤真代, 柳英夫 (2010). 「シンプリファイド・テクニカル・ジャパニーズ: 英訳を視野に入れて日本語を作る」『情報処理学会研究報告』 2010-DD-78 (5): 1-8.
- 高橋寛治, 竹野峻輔, 山本和英 (2017). 「機械翻訳システムの出力安定性の評価」『人工知能学会論文誌』 32 (5), D-H33: 1-4.
- 鳥飼久美子 (編) (2013). 『よくわかる翻訳通訳学』 ミネルヴァ書房.
- 山田優 (2007). 「ローカリゼーションにおける翻訳と翻訳理論研究」『翻訳研究への招待』 1: 57-68.
- 山口昌也, 乾伸雄, 小谷善行, 西村恕彦 (1998). 「前編集結果を利用した前編集自動化規則の獲得」『情報処理学会論文誌』 39 (1): 17-28.
- 吉見毅彦, 佐田いち子, 福持陽士 (2000). 「頑健な英日機械翻訳システム実現のための原文自動前編集」『自然言語処理』 7 (4): 99-117.

付録. 書き換え事例の類型化と頻度 (H: 病院, M: 自治体, B: BCCWJ, R: Reuters)

ID	種類	事例の頻度				例 (書き換え事例から部分的に抜粋)
		H	M	B	R	
S01	文分割・統合	4	1	7	2	1981年に「筆とエンピツ」を刊行してから、2001年4月に→1981年に「筆とエンピツ」を刊行した。それ以来、2001年4月に
S02	レイアウトの変更	0	3	0	0	下水道がつながっていない状態で、汲み取りタイプの便所をご利用の場合 業者名、申込方法および料金について、市町村の役所にお問い合わせください。 →下水道がつながっていない状態で、汲み取りタイプの便所をご利用の場合は、業者名、申込方法および料金について、市町村の役所にお問い合わせください。
S03	複文・重文の変更	0	0	0	1	消費者物価指数が前月比で下落したことが、相場を支援する材料として働いた。 → 消費者物価指数が前月比で下落し、そのことが、相場を支援する材料として働いた。
S04	句の分割	0	0	1	0	「筆とエンピツ」を刊行した小木さんは → 「筆とエンピツ」を刊行した後、小木さんは
S05	語順変更	24	6	22	13	暴風雨は、米国北東部で3日間にわたって続き → 米国北東部で3日間にわたって暴風雨は続き
S06	主語追加・削除	0	2	2	2	WTOの政府調達協定への加盟には合意するが → 我々はWTOの政府調達協定への加盟には合意するが
S07	読点の削除・追加	24	5	27	27	体重を中央に戻して両足で長めに爪先立ちを一回してください。 → 体重を中央に戻して、両足で長めに爪先立ちを一回してください。
S08	主題のスコープ変更	0	0	1	1	どの程度まで市場を開放するかについては結論が出ていない → どの程度まで市場を開放するかについての結論は出ていない
S09	ガ格と主題への変更	0	1	3	2	担当者は事情聴取を行った→担当者が事情聴取を行った
S10	視点変更	0	2	11	0	署員を対象に、事情聴取が行われた。 → 署員が、事情聴取を受けた。
S11	態の変更	3	1	13	3	講座を開催します。→講座が開催されます。
S12	修飾の仕方の変更	2	0	12	13	就労を支援するために“職親”制度を → 就労を支援するための“職親”制度を

S13	動詞句の主辞交替	0	0	0	3	寄り付きの売りを受けて懸念を強め → 寄り付きの売りを受けて強い懸念を抱き
S14	条件節の明示	2	7	2	0	日曜・祝日でも→ たとえ日曜・祝日でも
S15	体言止めの回避・使用	0	1	3	5	2 パーツ高の 141 パーツ。→ 2 パーツ高の 141 パーツとなった。
S16	名詞句の主辞交替	0	0	1	0	欧米で普及が進んでいる路面電車システム「LRT」のようなシステム→ 欧米で普及が進んでいる「LRT」のような路面電車システム
S17	名詞句・動詞句の交替	3	4	9	0	針付きのままで大丈夫です。→ 針は付いたままで大丈夫です。
S18	複合動詞の使用・展開	2	0	2	0	体重を中心にシフトしなおし→ もう一度体重を中心にシフトし
S19	複合名詞の使用・展開	2	7	5	8	富久小学校内学童クラブ→ 富久小学校内の学童クラブ
S20	接尾辞の使用・解除	2	1	10	5	2 パーツ高の 141 パーツに。→ 2 パーツ上がり 141 パーツに。
S21	接続表現	6	16	12	13	台湾の休日に当たるため、休場します。 → 台湾の休日に当たるので、休場します。
S22	並列表現	2	3	1	0	予備選・党員集会が行われた。→ 予備選と党員集会が行われた。
S23	同格表現	0	0	0	5	独連銀の副総裁、ガッドゥム氏 → 独連銀の副総裁であるガッドゥム氏
S24	限定表現	0	0	0	3	22 万 3000 トンに限られそうである → 22 万 3000 トンのみになりそうである
S25	場所の限定表現	0	0	0	2	ドイツ・ノルトライン・ウェストファーレン州 → ドイツのノルトライン・ウェストファーレン州
S26	伝聞表現	0	0	0	4	同氏によると→ 同氏曰く
S27	間接疑問表現	0	0	0	1	どのくらい市場が開放されるかについて → 市場開放の程度について
S28	サ変名詞表現の変更	1	2	7	4	勝利した→ 勝利を収めた
S29	形式名詞を使った表現	0	1	3	5	説明するという意向を伝えた。→ 説明する意向を伝えた。
S30	存在動詞表現	1	0	0	1	採尿室は右側にあるテレビの前にあります。 → 採尿室は右側にあるテレビの前です。
S31	になる・となる表現	0	0	0	11	それが相場を支援する材料になった。 → それが相場を支援する材料として働いた。
C01	特定表現の使用	29	36	69	33	整理券を無くしたり→整理券を紛失したり
C02	具体化	5	3	2	1	ご来院頂かなくても結構です。 → 病院にお越し頂かなくても結構です。
C03	端的な表現の使用	0	5	0	0	外国の人→ 外国人
C04	参照表現	0	0	0	1	連銀→ 同銀
C05	冗長化	0	1	0	1	12 日は台湾の休日です→ 12 日の日は台湾の休日です
F01	敬語化・非敬語化	19	11	14	4	来院して頂かなくても結構です。→ 来院しなくても結構です。
F02	時制の変更	0	3	1	2	出ていたとみられていた→ 出ていたとみられている
F03	並列語句のつながりの表現	4	4	0	1	手袋とインフューザーポンプ → 手袋およびインフューザーポンプ
F04	助動詞の変更	1	0	0	0	再度、体重を中央に移動させ→ 再度、体重を中央に移動し
F05	助詞の追加・削除・変更	4	9	24	9	外国人が生活するために必要な→ 外国人が生活するのに必要な
F06	助詞の使用・回避	4	3	3	10	登録のクレジットカード→ 登録されているクレジットカード
F07	複合助詞	0	1	1	5	12 日につきましては→ 12 日は
T01	固有表現	0	0	3	6	バーデン・ビュルテンベルク州→ バーデンビュルテンベルク州
O01	表記の変更	1	7	7	4	八十一年→ 81 年
O02	文末処理	0	1	2	0	もたらしたかどうか。→ もたらしたかどうか？
O03	記号の追加・削除・置換	0	6	0	0	◇たとえ日曜・祝日でも→ たとえ日曜・祝日でも
O04	省略の補完	0	0	3	2	81 年→ 1981 年
O05	チャンキングの変更	0	5	3	1	“職親制度”→ 「職親制度」
I01	内容の変更	18	20	27	16	ビニール袋に入れてください。 → ビニール袋の中に入れてください。
I02	ニュアンスの変更	0	7	17	6	折り曲げをしないでください。→ 折り曲げにご注意ください。
E01	表記ミス、文法ミス	3	1	4	6	針は付きのままでいいです。→ 針付きのままでいいです。
E02	必要要素の削除・復元、 原文にない情報の追加	19	0	6	6	両足で長めに爪先立ちを一回してください。 → 両足で爪先立ちを一回してください。