機械翻訳向け前編集の事例収集と類型化

宮田 玲†‡　藤田 篤†　内山 将夫†　隅田 英一郎†
†情報通信研究機構　‡東京大学大学院教育学研究科

1 はじめに

近年、機械翻訳（MT）の性能は著しく向上しているが、日本語と英語など言語構造の大きく異なる言語間の翻訳は難しく、MT を実社会で効果的に運用していくためには、MT システムそのものを改良するだけでなく、文書作成・翻訳工程における MT の活用に関するより実践的な検討が必要である。

我々は現在、翻訳の前処理段階に着目し、起点言語テキストの言語表現を翻訳しやすい形に統制する手法について研究している。これまでの研究では、特定の言語方向、MT システム、ドメインにおいて有効な書き換えの種類も全体像も十分明らかではなく、まずは MT 向けの前編集においていかなる書き換えが実際に観察されるのか、という現象の理解が重要だと考える。そこで本稿では、Miyata et al. [1] が提案した、(1) MT 訳の品質向上を目指して人間が原文の書き換えを試行錯誤的に繰り返す、(2) その書き換え履歴を分析する、という手続きをベースに、より詳細に書き換え事例を収集・分析する手法を提示する。そして、本手法を病院内会話、自治体生活情報、新聞記事ドメインから抽出した日本語文の英訳タスクに適用して書き換え事例を収集し、その一部を類型化した結果を報告する。

2 先行研究

MT 向け前編集の研究としては、統計的な手法により直接学習データから前編集器を自動構築する研究 [2]、原文中の MT が困難な箇所を自動あるいは手動で検出しながら人間による書き換えをガイドする研究 [3, 4] がある。ただし、これらはあくまで目の前のテキストを MT 向けに書き換えることを目的としており、書き換えの全体像を捉えようとするものではない。

一方、様々な書き換えの種類を抽出・同定し、ルールを作成した上で、テキストに適用するアプローチも検討されてきた。人手による文章執筆を統制する観点からは、Japio による「特許ライティングマニュアル」[1] や各種の商用ベースの執筆ガイドライン [2] が開発されてきた。また書き換えルールの自動適用の観点からは、白井ら [6]、山口ら [7] の研究がある。しかし、いずれも書き換えルールを作成するために用いた事例と手法が十分明示されておらず、提示されているルールの集合がどこまで網羅的なのか、また個別のルールがどの程度有効なのかを検証することが困難である。

Miyata et al. [1] は、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し、後から書き換え箇所を分析することで、前編集に有効な書き換えの種類の候補を広く抽出する一連の手続きを示している。3 種類の日英 MT システムを用いて、自治体生活情報ドメインの日本語文 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、試行錯誤の過程を綿密に記録していないため、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。

3 書き換え事例の収集

3.1 手法と作業支援ツール

ある日本語原文について、その書き換え文および MT 訳の履歴をひとまとまりにしたものを、「ユニット」と呼ぶ。ユニットごとに完結した書き換え作業用ツールを準備し（図 1 ）、以下の手順で履歴を収集する。

Step 1 入力ボックスの文に対する MT 訳が十分な品質[3]であれば進行状況を「Complete」にして Step 5 へ

Step 2 必要に応じて書き換え元の文を過去の書き換え履歴から選択する

Step 3 最小単位の書き換えを行い、「Translate」ボタンを押す（その時点の日本語文と MT 訳の対が書き換え履歴の一番上の行に追加される）[4]

Step 4 (a) MT 訳が十分な品質に達しておらず、かつこれ以上書き換えても 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」にして Step 5 へ (b) そうでない場合は Step 1 へ

Step 5 これまでの全ての書き換え履歴中で MT 訳の品質が最も高いものを指定する（「ベスト」と呼ぶ）

[1] http://japio-tjp.org/wmanual.html
[2] 例えば、日本語を対象としたものだと Acrolinx による STE ルールセット [5] などがある。
[3] 本研究では、「十分な品質の訳文」を「多少流暢さに欠けても、情報の過不足がなく、文法的にも正しい訳文」と定める。
[4] 登録済の文は受け付けない。

図 1: 書き換え作業用ツール

ドメイン	原文の平均文長（文字数）	書き換え数 合計	平均	中央値	最大値	Complete/Give up（ユニット数）	原文=Complete（ユニット数）
病院	20.2	1199	12.0	3	105	97/3	40
自治体	34.8	2119	21.2	14.5	89	97/3	3
新聞（BCCWJ）	44.4	3823	38.2	26.5	209	86/14	0
新聞（Reuters）	62.0	3637	56.0	46	211	62/3	0

表 1: 収集データの基本情報

文献 [1] の手法とは、(1) 試行錯誤の過程を細かく記録する点、(2) 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点が異なる。

3.2 対象

対象言語方向は日英とし、NICT が開発している「みんなの自動翻訳」[5] の汎用 MT（以下 GPMT）をデフォルトの設定で使用した。

対象文書ドメインは、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（以下それぞれ、病院、自治体、新聞）。新聞については、BCCWJ[8] と Reuters[9] のコーパスを用いた。これらからそれぞれ、GPMT の訓練データに使用されているものを除外した後、ランダムに日本語文を抽出した。病院、自治体、新聞（BCCWJ）については 100 文ずつ、新聞（Reuters）については 65 文を用いた。

[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コーパス（日本語）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]http://pj.ninjal.ac.jp/corpus_center/bccwj/

[9]http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/jea/reuters/index.html

3.3 書き換え作業の実施

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 1 名に依頼した。作業概要とシステムの操作方法を伝えた上で、(1) できるだけ最小単位の書き換えを行うこと、(2) 書き換えの過程で原文の意味内容を保持することを指示した。

3.4 収集できた書き換え事例の統計

日英 GPMT を用いて、本手法を 3 ドメイン（4 データセット）に適用した結果の概要を表 1 に示す。原文の平均文長が長いほど、平均書き換え数が多い傾向にある。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。逆に、原文の平均文長が 20 語程度の病院ドメインでは、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。

また病院と自治体の 97/100 ユニット、新聞（Reuter）の 62/65 ユニットが Complete となった点は特筆すべきであろう。特定のドメインでは、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、すなわち GPMT のポテンシャルを示している。

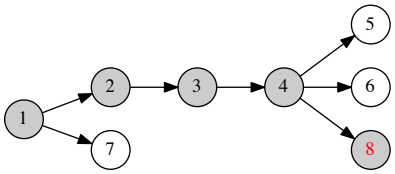

図 2: 書き換え履歴のパス

4 書き換え事例の類型化

4.1 得られるデータとその解釈

収集した書き換え履歴は、図 2 のようなツリー構造で表示することができる。各ノードは 1 つの書き換え履歴（日本語文及びその MT 訳）であり、ノードに付与された数字は履歴を保存した順番を示す [10]。

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 $(n-1)$ ペア（図 2 だと、1-2, 1-7, 2-3, 3-4, 4-5, 4-6, 4-8 の 7 ペア）を取り出すことができる。これらの書き換えはいずれも、MT 訳の品質向上を目指してなされたものではあるが、当然ながら実際に品質が向上しているとは限らない。そこで我々は、「ベスト」の履歴はユニット中で最も MT 訳の品質が高い、ということに注目し、原文から「ベスト」に至るパス（本稿では「ベストパス」と呼ぶ）を構成する書き換えペアを分析した（図 2 だと灰色部分、1-2, 2-3, 3-4, 4-8 の 4 ペア）。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わけではないが、大局的には「ベスト」は原文よりも MT 訳の品質が高いことは分かっているため、その過程にあるいずれかの書き換えは MT 訳の品質改善に寄与しているとみなすことができる。

4.2 ベストパスの分析

我々はベストパス中の直接的な書き換えペアを、以下の手続きで分析した。

(1) プリミティブな書き換えへの分解 [11]
(2) テキスト表層上の差異の抽出
(3) 書き換えの種類の類型化

例えば、「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。」という書き換えペアに対し、(1)「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。(2) 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 2 つの表層上の差異を取り出し、(3) それぞれ**「語順変更」**と**「態の変更」**として種類を定める。

[10]頂上に位置する 1 番のノードはオリジナルの文である。赤色で表示された数字は、作業者が「ベスト」と指定した履歴を示す。

[11]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。

ドメイン	全ペア数	ベストパス上のペア数	
		分解前	分解後
病院	131	97	185
自治体	217	106	186
新聞（BCCWJ）	484	174	340
新聞（Reuters）	483	191	268

表 2: 各ドメイン 10 ユニットの書き換えペア数

4.3 事例の類型化と頻度

4 データセット中の各々10 ユニット、合計 40 ユニットを分析した。表 2 は書き換えペア数の統計情報を示す。プリミティブな書き換えに分解することで、ベストパス上の書き換えペア数は約 1.4〜2.0 倍に増加した [12]。

表 3 に抽出・同定した書き換え種類の一覧と、各データセットにおける事例の頻度情報を示す。ID 列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）の上位カテゴリーを示す。書き換えの種類の総数は 53 である。

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。また **I01「内容の変更」**は自明な要素を推論して補完したり、副詞を加えてニュアンスを変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、ここから前編集に有効な知見を導くことが期待できる。

一方、新聞ドメインにおける **S15「体言止めの回避／使用」**や **S20「接尾辞の使用／解除」**のように、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ある。例えば、「タイ農民銀行が売買代金で１位を獲得し、２バーツ高の１４１バーツ。」のような新聞特有の簡略的な書き方があり、これが MT 訳の品質に関わると判断され、「２バーツ高の」→「２バーツ上がり」、「１４１バーツ。」→「１４１バーツとなった。」といった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」

[12]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。

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「トイレットペーパーをカートに山積みにした客」→「トイレッ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、前編集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[13]においても新しい書き換えも見られた。

5 おわりに

本研究では、MT 訳の品質改善を目指して、試行錯誤的に人間が書き換えを繰り返す手法により、前編集事例の収集を試みた。さらに事例を手作業で類型化することで、53 の書き換えの種類を抽出・同定し、またドメインによる種類の傾向を一部明らかにした。現在、複数の作業者に依頼し、英日翻訳の書き換え事例を収集中である。引き続き、本手法により書き換えの全体像を明らかにし、その上で、MT 向け前編集手法を開発する予定である。

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。

参考文献

[1] Miyata, R., Hartley, A., Paris, C., Tatsumi, M. & Kageura, K. Japanese Controlled Language Rules to Improve Machine Translatability of Municipal Documents. *MT Summit XV*, 90–103, 2015.

[2] 南條浩輝, 山本祐司, 吉見毅彦. 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前編集システムの自動構築, 情報処理学会論文誌, 53(6): 1644–1653, 2012.

[3] Uchimoto, K., Hayashida, N., Ishida, T. & Isahara, H. Automatic Detection and Semi-Automatic Revision of Non-Machine-Translatable Parts of a Sentence. *LREC 2006*, 703–708, 2006.

[4] Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A. & Bederson, B. Improving Translation via Targeted Paraphrasing, *EMNLP 2010*, 127–137, 2010.

[5] 小倉英里, 工藤真代, 柳英夫. 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デジタルドキュメント, 2010(5): 1–8, 2010.

[6] Shirai, S., Ikehara, S., Yokoo, A. & Ooyama, Y. Automatic Rewriting Method for Internal Expressions in Japanese to English MT and Its Effects. *CLAW 1998*, 62–75, 1998.

[7] 山口昌也, 乾伸雄, 小谷善行, 西村恕彦. 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, 1998.

[13] 言い換えのあれこれ, http://paraphrasing.org/paraphrase.html

ID	書き換えの種類	事例の頻度			
		H	M	B	R
S01	文分割／統合	4	1	7	2
S02	レイアウトの変更	0	3	0	0
S03	複文／重文の変更	0	0	0	1
S04	句の分割	0	0	1	0
S05	語順変更	24	6	22	13
S06	主語追加／削除	0	2	2	2
S07	読点の削除／追加	24	5	27	27
S08	主題のスコープ変更	0	0	1	1
S09	ガ格と主題ハの変更	0	1	3	2
S10	視点変更	0	2	11	0
S11	態の変更	3	1	13	3
S12	修飾の仕方の変更	2	0	12	13
S13	動詞句の主辞交替	0	0	0	3
S14	条件節の明示	2	7	2	0
S15	体言止めの回避／使用	0	1	3	5
S16	名詞句の主辞交替	0	0	1	0
S17	名詞句・動詞句の交替	3	4	9	0
S18	複合動詞の使用／展開	2	0	2	0
S19	複合名詞の使用／展開	2	7	5	8
S20	接尾辞の使用／解除	2	1	10	5
S21	接続表現	6	16	12	13
S22	並列表現	2	3	1	0
S23	同格表現	0	0	0	5
S24	限定表現	0	0	0	3
S25	場所の限定表現	0	0	0	2
S26	伝聞表現	0	0	0	4
S27	間接疑問表現	0	0	0	1
S28	サ変名詞表現の変更	1	2	7	4
S29	形式名詞を使った表現	0	1	3	5
S30	存在動詞表現	1	0	0	1
S31	になる・となる表現	0	0	0	11
C01	特定表現の使用	29	36	69	33
C02	具体化	5	3	2	1
C03	端的な表現の使用	0	5	0	0
C04	参照表現	0	0	0	1
C05	冗長化	0	1	0	1
F01	敬語化／非敬語化	19	11	14	4
F02	時制の変更	0	3	1	2
F03	並列語句のつなぎの表現	4	4	0	1
F04	助動詞の変更	1	0	0	0
F05	助詞の追加／削除／変更	4	9	24	9
F06	助詞の使用／回避	4	3	3	10
F07	複合助詞	0	1	1	5
T01	固有表現	0	0	3	6
O01	表記の変更	1	7	7	4
O02	文末処理	0	1	2	0
O03	記号の追加／削除／置換	0	6	0	0
O04	省略の補完	0	0	3	2
O05	チャンキングの追加／削除	0	5	3	1
I01	内容の変更	18	20	27	16
I02	ニュアンスの変更	0	7	17	6
E01	表記ミス、文法ミス	3	1	4	6
E02	必要要素の削除／復元、原文にない情報の追加	19	0	6	6

表 3: 書き換え事例の類型化と頻度
（H: 病院、M: 自治体、B: BCCWJ、R: Reuters）