様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月　 ６ 日現在

機関番号：１３９０１
研究種目：若手研究
研究期間：2019～2021
課題番号：１９Ｋ２０６２８

研究課題名（和文）機械翻訳活用のための前編集手法の体系化と段階的編集支援ツールの開発

研究課題名（英文）Systematising pre-editing methods and developing a staged editing support system for better use of machine translation

研究代表者
宮田　玲（Miyata, Rei）

名古屋大学・工学研究科・助教

研究者番号：７０８０４３００

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：人手による原文書き換え事例の収集と分析を通じて、機械翻訳出力の品質向上のための日本語前編集（プリエディット）ルールを作成した。特に、原文に直接書かれていない内容をテキスト中に明示する方法（明示化方略）が有効であることが示唆された。人間の編集者を支援するツールのプロトタイプを開発し、一部のルールを実装した。本ツールは、編集プロセスの段階に応じて、ルールに違反する箇所の検出、書き換え候補の提示、書き換え候補のランキング、選択した候補に基づくテキストの自動書き換えの各機能を提供する。

研究成果の学術的意義や社会的意義
近年、機械翻訳の性能が向上し、各種の情報発信場面での利用が増えている。原文を事前に翻訳しやすい表現に書き換えることの有効性は指摘されながらも、その具体的な手法は十分明らかになっていなかった。本研究は、何をどう書き換えればよいかに関する具体的な指針を与えるだけでなく、ツールの提供を通じて、そのような書き換えに不慣れな書き手を支援するものであり、様々な利用者による機械翻訳の活用に貢献する。

研究成果の概要（英文）：We created Japanese pre-editing rules for improving the quality of machine translation outputs by collecting and analysing manual pre-editing instances. Our analysis suggested that explicitation strategies, which are used to explicitly indicate the content of source text, are effective to improve machine translation outputs. We also developed a prototype editing support system and implemented some of the created rules. This tool provides the following functions according to editing processes: detection of rule violations, suggestion of candidates for rewriting, ranking of the candidates, and automatic rewriting of the source text based on the selected candidate.

研究分野：図書館情報学

キーワード：前編集（プリエディット）　機械翻訳　明示化方略　執筆支援システム　制限言語　翻訳しにくい表現　翻訳品質

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、機械翻訳の性能が大幅に向上し、様々な情報流通場面での利用が期待されている。これまで、機械翻訳の活用手法の一つとして、入力となる原文を適切に書き換えることで翻訳結果を改善する前編集（プリエディット）と呼ばれる手法が注目されてきた。しかし、翻訳実務における前編集の導入は必ずしも十分進んでおらず、特に以下の課題が残されている。

(1) 最新の機械翻訳方式における前編集手法の有効性と全体像が明らかではないこと。

(2) 人間の書き換え作業を支援する前編集用のツールが整備されていないこと。

(1)について、既存研究ではルールに基づく機械翻訳や統計的機械翻訳を対象とした前編集手法の開発や調査が進められてきたが（Miyata et al. 2015; 宮田・藤田 2017; Seretan et al. 2014)、最新の深層学習に基づくニューラル機械翻訳を対象とした研究は十分なされていない。入力文と出力文を単一のネットワークモデルで学習するニューラル機械翻訳は、利用者のみならず研究者にとってもブラックボックス化しており、その制御可能性については十分解明されていない。

(2)について、人間の前編集プロセスを支援する制限言語チェッカー（Bernth & Gdaniec 2001; Mitamura et al. 2003; Miyata et al. 2016）が開発されてきたが、ニューラル機械翻訳を想定したものはあまりなく、また日本語を対象としたツールも不足している。

２．研究の目的

本研究の目的は大きく、(1)機械翻訳向け前編集手法を言語表現の書き換えルールとして体系化することと、(2)書き換えルールごとに部品化された段階的な前編集支援ツールを開発することである。特に翻訳元の言語は日本語を対象とする。また、特定の機械翻訳システムに特化したルールではなく、ある程度汎用に利用可能なルールの開発を目指す。

３．研究の方法

(1) 人手前編集事例の分析に基づく書き換えルールの構築

原文の内容を保持したまま、十分な品質の機械翻訳結果が得られるまで試行錯誤的に書き換えを繰り返す手法を用いて、多様な条件（言語方向、テキスト分野、機械翻訳システム）で、書き換え事例を収集する。収集した書き換え事例を分析し、機械翻訳品質の向上に有効と考えられる書き換え操作をルール化する。

(2) 書き換え支援ツールの開発

(1)で作成したルールを実装し、人間の前編集プロセスを支援する Web ベースのアプリケーションを作成する。書き換えの段階に応じて、書き換え対象の検出、書き換え候補の提示、書き換え候補のランキングの各ステップからなる支援を行う。

４．研究成果

(1) 人手前編集事例の分析に基づく書き換えルールの構築

① 機械翻訳向け前編集事例の収集と分析

3 つの言語方向(日英、日中、日韓)、4 つのテキストドメイン(病院内会話、自治体文書、BCCWJ、ニュース記事)、2 つの機械翻訳システム(Google 翻訳、TexTra)の全組み合わせ 24 条件について、25 文ずつ（計 600 文）を対象に、作業者に書き換え作業を依頼した。作業を通じて、合計 6652 件の書き換え事例を収集した。1 つの原文に対する書き換えの履歴と機械翻訳結果をまとめたものを 1 ユニットとする。600 ユニット中 571 ユニット（約 95%）は、書き換えによって、情報の過不足や文法的な誤りのない翻訳文を出力できた。このことからも、収集した書き換え事例中には、機械翻訳品質の向上に寄与する書き換え方法が含まれることが分かった。

② 収集事例の分析

上記の 24 条件について 10 ユニットずつ、計 240 ユニットを対象に、書き換え事例 935 件の人手による類型化を進めた。具体的には、最小単位の書き換え前後のテキスト差分を抽出・分析しながら、書き換え方法のタイプを類型化していく方法をとった。その結果、6 カテゴリ 39 タイプの言語学的な書き換え方法を同定した。また、明示化・暗示化・情報保持の情報方略の観点から事例を分析し、特に明示化方略（原文に直接書かれていない内容をテキスト中に明示する方法)が、機械翻訳品質の向上に寄与することが示唆された。

③ 翻訳品質に寄与する書き換え方法の類型化

明示化方略に注目し、明示化の対象となる表現と明示化方略の体系化を行った。具体的には、262 件の最小単位の書き換え事例を対象に、どのような表現に対して(対象)、どのような書き換

えがなされ(操作)、どのような情報が明示されたか(明示内容)を分析し、3 カテゴリ・34 サブカテゴリからなる明示化対象表現(対象)、6 カテゴリ・26 サブカテゴリからなる明示化方略体系(操作・明示内容)を構築した。これらは、機械翻訳向け前編集ルールとして用いることができる。

(2) 書き換え支援ツールの開発

① インタフェースの開発

人間の執筆・編集プロセスを段階的に支援するためのツールのインターフェースを設計し、プロトタイプを構築した。ルールに違反する箇所をハイライトし、書き換え候補を適切な順番で提示する機能を持つ。また書き換え候補をクリックすることで、その表現に自動で書き換えることができる。

② 前編集ルールの実装

一部のルールを実装した。ルールに応じて、規則に基づく手法や機械学習手法を適宜組み合わせた。例えば、文の分割は、形態素解析・構文解析結果を用いた言語変換規則により実現した。また、語彙的な変換には、事前学習済みのマスク言語モデルの活用を試みた。書き換え支援ツールに実装したルールの性能は、テストセットを用いて評価し、人間の作業支援の観点からその有用性を確認した。

＜参考文献＞

- Bernth, A., Gdaniec, C. (2001). MTranslatability. *Machine Translation*, 16(3): 175-218.
- Mitamura, T., Baker, K. L., Nyberg, E., Svoboda, D. (2003). Diagnostics for Interactive Controlled Language Checking, *Proceedings of the Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (EAMT/CLAW)*, 237-244.
- Miyata, R., Hartley, A., Paris, C., Kageura, K. (2016). Evaluating and Implementing a Controlled Language Checker. *Proceedings of the 6th International Workshop on Controlled Language Applications (CLAW)*, 30-35.
- Miyata, R., Hartley, A., Paris, C., Midori T., Kageura, K. (2015). Japanese Controlled Language Rules to Improve Machine Translatability of Municipal Documents. *Proceedings of the Machine Translation Summit XV*, 90-103.
- 宮田玲, 藤田篤. (2017). 機械翻訳向けプリエディットの有効性と多様性の調査. 通訳翻訳研究への招待, No.18, pp.53-72.
- Seretan, V., Bouillon, P., Gerlach, J. (2014). A Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 1793-1799.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Miyata Rei	４．巻 －
２．論文標題 Formulating a terminology for source document profiling through a literature review: From functionalist to documentational approaches	５．発行年 2022年
３．雑誌名 Perspectives	６．最初と最後の頁 1～18
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/0907676X.2022.2049830	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 加藤汰一，宮田玲，佐藤理史	４．巻 62
２．論文標題 説明文を対象とした日本語文末述語の平易化	５．発行年 2021年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 1605～1619
掲載論文のDOI（デジタルオブジェクト識別子） 10.20729/00212765	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 宮田玲	４．巻 306
２．論文標題 翻訳テクノロジー論考 第9回 ～テクノロジーを論じ考えるために～	５．発行年 2020年
３．雑誌名 JTFジャーナル	６．最初と最後の頁 28-29
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 宮田玲	４．巻 301
２．論文標題 翻訳テクノロジー論考 第5回 ～「やさしい日本語」と翻訳テクノロジーその2～	５．発行年 2019年
３．雑誌名 JTFジャーナル	６．最初と最後の頁 24-25
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　3件）

1．発表者名
Rei Miyata, Hodai Sugino

2．発表標題
Building a Controlled Lexicon for Authoring Automotive Technical Documents

3．学会等名
XIX EURALEX Congress（国際学会）

4．発表年
2021年

1．発表者名
Rei Miyata, Atsushi Fujita

2．発表標題
Understanding Pre-Editing for Black-Box Neural Machine Translation

3．学会等名
16th conference of the European Chapter of the Association for Computational Linguistics（国際学会）

4．発表年
2021年

1．発表者名
Taichi Kato, Rei Miyata, Satoshi Sato

2．発表標題
BERT-Based Simplification of Japanese Sentence-Ending Predicates in Descriptive Text

3．学会等名
13th International Conference on Natural Language Generation（国際学会）

4．発表年
2020年

1．発表者名
杉野峰大，宮田玲，小川浩平，佐藤理史

2．発表標題
執筆・翻訳のための制限語彙の構築とその自動化の検討

3．学会等名
言語処理学会第27回年次大会

4．発表年
2021年

１．発表者名
加藤汰一，宮田玲，立見みどり，佐藤理史

２．発表標題
文化財説明文を対象とした平易化支援システムの設計と実装

３．学会等名
第34回人工知能学会全国大会

４．発表年
2020年

１．発表者名
宮田玲，宮内拓也，影浦峡

２．発表標題
翻訳のための起点文書分析：文献レビューの枠組み

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
永井利季，宮田玲，立見みどり，佐藤理史

２．発表標題
文化財関連の専門用語を対象とした平易な説明生成

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
永井利季，宮田玲，立見みどり，佐藤理史

２．発表標題
専門用語を平易に説明するプロセスとその支援技術：文化財説明文の平易化に向けて

３．学会等名
日本通訳翻訳学会第20回年次大会

４．発表年
2019年

〔図書〕　計1件

１．著者名 Rei Miyata	４．発行年 2020年
２．出版社 Routledge	５．総ページ数 236
３．書名 Controlled Document Authoring in a Machine Translation Age	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関