卒業論文

BARTに基づく事後編集器による機械翻訳の自然性向上

指導教官　　村上 陽平 教授

立命館大学 情報理工学部
先端社会デザインコース 4回生
2600210248-9

富田 悠斗

2024年度（秋学期）卒業研究3（CH）
令和7年1月31日

BART に基づく事後編集器による機械翻訳の自然性向上

富田 悠斗

内容梗概

機械翻訳は, ニューラル機械翻訳の登場によって翻訳精度が大幅に向上した. その結果, 異なる言語を話す人々において, ニューラル機械翻訳は多言語コミュニケーションの拡大に大きく貢献している.

しかしながら, 機械翻訳は適切さと流ちょうさのみで評価されているため, 翻訳元言語と翻訳先言語の特性の違いにより, 翻訳先言語の母語話者にとって意味は通るものの自然ではない翻訳結果になることがある. 日英翻訳の場合，生成される英語が入力の日本語の影響を強く受け，副詞句を伴う非母語話者英語を出力する傾向にある．例えば,「強風で窓が割れた.」は「Windows were broken by strong winds.」と出力される．この翻訳結果は文法的に問題なく意味も正確に伝わるが，英語母語話者は「The strong wind broke the window.」と翻訳し，窓が割れたことと強風の因果関係をより明確に表現する．

そこで，本研究では，日英翻訳を対象として，機械翻訳の翻訳結果を翻訳先言語のスタイルを反映した自然な翻訳に変換する事後編集器を構築する．具体的には，非母語話者英語と母語話者英語の対訳コーパスにより，事前学習済みの Sequence-to-Sequence モデルをファインチューニングして事後編集モデルを構築する．本手法の実現にあたり，取り組むべき課題は以下の 2 点である.

非母語話者英語と母語話者英語の対訳コーパスの構築

非母語話者英語を母語話者英語に変換する事後編集器を構築するには，非母語話者英語とそれに対応する母語話者英語のペアが必要である．つまり，日本人が翻訳した英語とそれを英語母語話者が事後編集したデータを取得しなければならない．また，翻訳コストを抑えるためには，できる限り非母語話者英語と推定されるものに原文段階で限定する必要がある.

事後編集モデルのファインチューニング

非母語話者英語と母語話者英語のペアを用いて事後編集器を構築する必要がある．このような変換モデルを学習するには，非母語話者英語と母語話者英語のコーパスを用いて，事前学習済みモデルをファインチューニングしなければならない．また， Google 翻訳や DeepL 翻訳

などの多様な機械翻訳に対して母語話者英語を生成できるように，汎用的な事後編集モデルを構築する必要がある．

一つ目の課題に対しては，非母語話者による英語への翻訳と英語母語話者による事後編集の両方を行うのはコストが高いため，英語母語話者が作成した英語原文と日本語訳の対訳コーパスと機械翻訳を用いて，非母語話者英語と母語話者英語の対訳コーパスを作成する．具体的には，母語話者英語分類器によって母語話者英語と推定された対訳ペアのみを抽出する．次に，そのペアの日本語訳文を機械翻訳で英訳し，母語話者英語分類器を用いて非母語話者英語と推定された対訳ペアを抽出し，日本語訳を介して，母語話者英語と非母語話者英語のペアを取得する．この時，多様な機械翻訳に対して汎用的な事後編集モデルを構築するために，Google 翻訳，DeepL 翻訳，みらい翻訳の 3 種類の機械翻訳を用いる．

二つ目の課題に対しては，事前訓練された英語版の BART モデルを用いる．一つ目の課題で構築した対訳コーパスを用いて，それぞれ BART をファインチューニングし，非母語話者英語から母語話者英語への事後編集モデルを構築する．自然性を母語話者英語分類器の母語話者英語の予測確率と定義し，事後編集前後の英文の自然性を評価し，提案手法の有効性を検証する本研究の貢献は以下の通りである．

非母語話者英語と母語話者英語のペアの構築

NICT が提供する日英コーパスから約 11 万件のペアを取得した．このペアを分類器によって母語話者英語自然性が 0.5 以上のものでフィルタリングし，約 62,000 件の日本語と母語話者英語のペアを作成した．さらに，このペアから日本語を英訳して，分類器によって自然性が 0.5 以下のものでフィルタリングを行い，非母語話者英語と母語話者英語のペアを約 35,000 件取得した．

非母語話者英語から母語話者英語への変換器の構築

作成した対訳コーパスを用いて BART モデルにファインチューニングし，非母語話者英語から母語話者英語への事後編集モデルを構築した．そのモデルを事後編集前後における英語に母語話者英語分類器を用いて評価を行い，21%の自然性の向上が確認された．

Improving the Naturalness of Machine Translation Using a BART-Based Post-Editor

Yuto Tomida

Abstract

With the advent of neural machine translation, the accuracy of machine translation quality has significantly improved. As a result, neural machine translation has greatly contributed to the expansion of multilingual communication among speakers of different languages.

However, because machine translation is evaluated only on appropriateness and fluency, linguistic differences between the source and target languages may result in translation outputs that, although meaningful to native speakers of the target language, sound unnatural. In the case of Japanese-English translation, the generated English was strongly influenced by the input Japanese and often produced non-native English, particularly with excessive adverbial phrases.

Therefore, in this study, we built a post-editor for Japanese-English translation that converted the results of machine translation into more natural translations that align with the natural style of the target language. Specifically, we built a post-editing model by fine-tuning a pre-trained sequence-to-sequence model using a bilingual corpus of non-native English and native English. There were two challenges to be addressed in realizing this method:

Building a Parallel Corpus of Non-Native English and Native English

To develop a post-editor that refined non-native English into native-like English, we needed pairs of non-native English and their corresponding native English versions. In other words, we needed to collect translations produced by Japanese speakers and revised by native English speakers. Since translating into English by non-native speakers and post-editing by native speakers was costly, we created a bilingual corpus of non-native English and native English using a bilingual corpus of English original texts and Japanese translations created by native English speakers and machine translation. Here, we used three types of machine translation: Google Translate, DeepL Translate, and Mirai Translate, to build a general-purpose post-editing model for a

variety of machine translations.

Fine-tuning the post-editing model

A post-editor needed to be built using pairs of non-native and native English. To train such a conversion model, a pre-trained model needed to be fine-tuned using corpora of non-native and native English. Additionally, a versatile post-editing model should be developed so that it could generate native English for various machine translations, such as Google Translate, DeepL Translate and Mirai Translate.

For the second task, we used a pre-trained English BART model. Using the bilingual corpus constructed in the first task, we fine-tuned BART and constructed a post-editing model from non-native English to native English. We used a native English classifier to evaluate the English sentences before and after post-editing and verified the effectiveness of the proposed method.

Constructing a Parallel Corpus of Non-Native and Native English

Approximately 110,000 pairs were obtained from the Japanese-English corpus provided by NICT. These pairs were filtered using a classifier to find those with a predicted probability of 0.5 or higher in native English, yielding approximately 62,000 pairs of Japanese and native English. Furthermore, these pairs were translated into Japanese and filtered to find those with a predicted probability of 0.5 or lower using a classifier, resulting in approximately 35,000 pairs of non-native English and native English.

Developing a Non-Native to Native English Translator

We fine-tuned the BART model using the bilingual corpus we created and constructed a post-editing model for converting non-native English to native English. We evaluated the model using a native English classifier on English before and after post-editing and confirmed a 21

BART に基づく事後編集器による機械翻訳の自然性向上

目次

第1章　はじめに

機械翻訳は, ニューラル機械翻訳の登場によって翻訳精度が大幅に向上した. その結果, 異なる言語を話す人々において, ニューラル機械翻訳は多言語コミュニケーションの拡大に大きく貢献している. 特に, 英語は主要な国際共通語 [1] として側面を持っており, コミュニケーションの場において英語の重要性がさらに増している.

翻訳モデルの訓練には, 膨大なテキストデータが必要であり, これはインターネット上で収集されたデータを学習させることで実現している. 現在, 機械翻訳は適切さと流ちょうさでのみ評価されており, 翻訳のスタイルは重視されていない [2]. これは, インターネット上から収集された英語は, 英語を主言語としない人によって生成された非母語話者英語も含むためである. その結果, 機械翻訳が出力した英語が, 英語母語話者にとって自然ではない場合がある. 日英翻訳の場合，生成される英語が入力の日本語の影響を強く受け，副詞句を伴う非母語話者英語を出力する傾向にある. 例えば,「この薬を飲めば風邪はすぐに治ります.」という文を Google 翻訳で英訳すると「If you take this medicine, your cold will be cured quickly.」と出力される. この翻訳結果は文法的に問題なく意味も正確に伝わるが，英語母語話者は「This medicine will cure your cold quickly.」と翻訳し，窓が割れたことと強風の因果関係をより明確に表現する.

そこで，本研究では，日英翻訳を対象として，機械翻訳の翻訳結果を翻訳先言語のスタイルを反映した自然な翻訳に変換する事後編集手法を提案する. 具体的には, 日英の対訳コーパスから非母語話者英語と母語話者英語のペアを抽出し, このデータを BART を用いてファインチューニングすることで, 母語話者英語から非母語話者英語への変換器を構築する. 本手法の実現にあたり，取り組むべき課題は以下の 2 点である.

非母語話者英語と母語話者英語の対訳コーパスの構築

非母語話者英語を母語話者英語に変換する事後編集器を構築するには，非母語話者英語とそれに対応する母語話者英語のペアが必要である. つまり，日本人が翻訳した英語とそれを英語母語話者が事後編集したデータを取得しなければならない. また，翻訳コストを抑えるためには，できる限り非母語話者英語と推定されるものに原文段階で限定する必要がある.

事後編集モデルのファインチューニング

非母語話者英語と母語話者英語のペアを用いて事後編集器を構築する必要がある．このような変換モデルを学習するには，非母語話者英語と母語話者英語のコーパスを用いて，事前学習済みモデルをファインチューニングしなければならない．また， Google 翻訳や DeepL 翻訳などの多様な機械翻訳に対して母語話者英語を生成できるように，汎用的な事後編集モデルを構築する必要がある．

以下, 本論文では, 第 2 章で関連研究と事後編集について述べる．次に, 第 3 章で母語話者英語と非母語話者英語の対訳ペアの抽出手法について述べ, 第 4 章でその対訳ペアを用いた事後編集モデルの構築について述べる．最後に, 第 5 章で構築した事後編集モデルの性能について述べ, 第 6 章で今後の展望と本研究の振り返りを述べる．

第２章　関連研究

現代社会において, 機械翻訳は非常に広く浸透しており, 多くの利用者が存在する. 日常生活や仕事の中で, 高品質な機械翻訳として Google 翻訳や DeepL 翻訳などがあり, その需要は急増している. 現在, 機械翻訳は適切さと流ちょうさで評価されており, 翻訳のスタイルは重視されていない. 機械翻訳は完全に正確かつ流ちょうな訳文を生成することは常に保証されているわけではないため, その翻訳結果を向上させるために, さまざまな研究が行われている. 翻訳結果の向上に対するアプローチとして, 大きく機械翻訳, 事前編集, 事後編集の３つが挙げられる. 機械翻訳に関しては, 翻訳エンジンそのものの精度向上を図るのではなく, 対訳辞書の整備により翻訳精度の向上に貢献するものである. 統計的機械翻訳は, 対訳辞書を用いることでデコーダ制約を実現し, 翻訳精度の改善が行われたが, ニューラル機械翻訳において対訳辞書を用いた自然さ改善手法は見られなかった. そこで, 竹林らは, 対訳辞書を単語報酬モデルによりニューラル機械翻訳に適用する手法を研究し, より少ない計算量での翻訳精度の改善を実現した [3]. 事前編集 (Pre-editing) ならびに事後編集 (Post-editing) は, 入力あるいは出力される文章を編集することで翻訳精度を向上させる手法である. 事前編集とは, 翻訳プロセスの最初に位置し原文を機械翻訳するのに適した形に前処理する手法である. 一方で, 事後編集とは, 翻訳プロセスの最終段階に位置し翻訳結果を英語母語話者スタイルの英語に変換する後処理である. 図 1 は, これらの手法が同一言語間で変換が行われていることを示している. プロの翻訳家を要するこれらの手法は, 翻訳精度や信頼性は非常に高いが, 翻訳コストや翻訳スピードが課題であり, 一般ユーザが簡単に使用できない傾向がある. 瓦は日英間のような語順が大きく異なる言語間において, RNN(Recursive Neural Network) や非自己回帰ニューラル機械翻訳などのさまざまなアプローチに対して, 事前編集が翻訳精度の向上に貢献することを示した [4]. この手法は, 人手による事前編集と同等の翻訳精度を達成している. 一方, 井佐原はプロの翻訳家ではなく, ボランティアの集合知による事後編集を紹介した [5]. これにより, 事後編集にかかる膨大なコストは緩和される.

翻訳結果を改善する手法とその基準はさまざまであるが, 本研究では出力される英語が母語話者にとって自然であるかを基準とし, ニューラルネットワークを用いて日英翻訳タスクにおける事後編集を行うことである. 対訳データを収

集ならびに生成し, BART モデルをファインチューニングして, 機械翻訳の翻訳結果から母語話者英語を生成する事後編集モデルを構築する. 訳文の品質を向上させる手法として, BART という新しいアプローチを提案する.

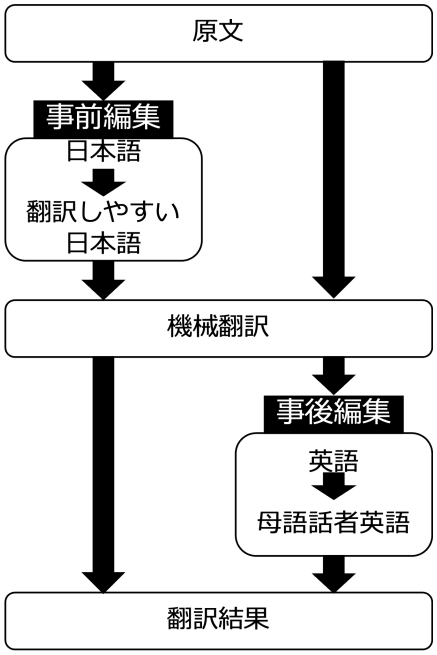

図 1: 事前編集ならびに事後編集の位置付け

第3章　非母語話者英語と母語話者英語の対訳コーパス

本章では, 事後編集モデルの構築にあたってファインチューニングを行うための対訳データの抽出方法を述べる. この対訳データは非母語話者英語と母語話者英語の対訳ペアにより構成され, これは事後編集モデルの性能向上において重要な役割を持つ. 具体的には, この対訳ペアは, 日本人が翻訳した英語とそれを英語母語話者が事後編集した英語のペアである. しかし, このような対訳コーパスを取得することは困難であるため, これは日英対訳のデータから生成する必要がある.

3.1　日英対訳コーパスの取得

BART にファインチューニングするための非母語話者英語と母語話者英語のペアの対訳コーパスを準備するには, まず日英対訳データを取得する必要がある. また, この時英語は英語母語話者によって書かれたものでなければならない. 情報通信機構 (NICT) が提供する「日英対訳文対応付けデータ [1)]」はこの条件を満たしている. これは, 著者が英語母語話者である英語とそれに対する日本語が対応付けされたテキストの集合であり, 英語と日本語が各 160 ファイル, および各 117,983 件の対訳ペアが内蔵する.

3.1.1　不適切な文の修正および削除

取得した日英対訳コーパスには, 不完全な文を含んでいる. 不完全な文とは, 最終的な対訳データを抽出するのに完全に不要な文や不適切な表現を含んだ文を意味する. データの品質を担保するためには, これらの不完全な文を修正あるいは削除する必要がある. 各ファイルの最後などに記述された脚注や注意書きは, データの品質を不安定にする原因として削除する. また, 不適切な表現として「ー-€$＄●□【】『』—｜/／※–＃＜＞」などの記号を含んだ文章や文字化けを起こした文章が挙げられる. これらを含む文章は, 機械翻訳への入力として一般的ではなく, ファインチューニングされた事後編集モデルに影響を与える可能性があるため, 修正する必要がある. 本研究ではこのようなデータのノイズを除去し, 各 117,983 件から各 108,742 件が抽出された. 以下ではこの抽出され

[1)] https://www2.nict.go.jp/astrec-att/member/mutiyama/align/index.html

た文を原文として扱う.

3.2 コーパス作成プロセス

取得した日英対訳の原文から非母語話者英語と母語話者英語の対訳コーパスを生成するためには, 大きく２つのステップを踏む必要がある. 図２は, 対訳コーパス作成のプロセス並びに２つのステップの位置付けを示す. まず, ステップ１では原文の英語に対して, 母語話者英語分類器を用いて母語話者英語とそれに対応する日本語を抽出する. 次に, ステップ２では抽出された日本語を機械翻訳を用いて英訳し, その英訳された文に対して分類器を用いて非母語話者英語とそれに対応する母語話者英語を抽出する. 原文は日本語と英語でそれぞれファイルが分かれており, それぞれファイルの行同士で対応し合っているが, ステップ１から互いの文の間に TAB を挿入し, 対訳データを同ファイルにおいて処理を行っていく.

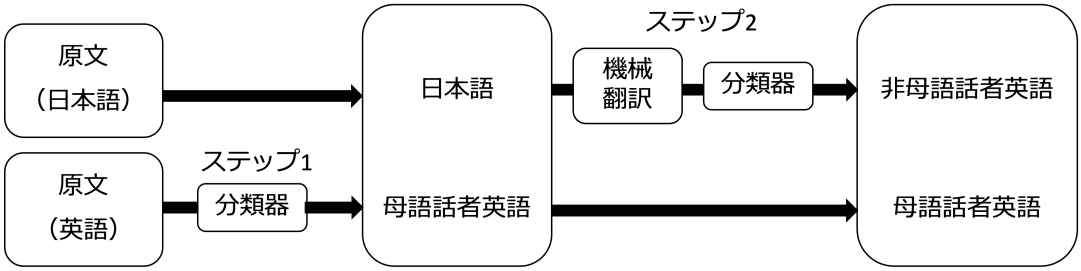

図 2: 対訳コーパス作成の手順

3.2.1 母語話者英語分類器の取得

非母語話者英語と母語話者英語の対訳コーパスを作成する上で, その英語が非母語話者英語であるのか, 母語話者英語であるのかを判定する必要がある. これは, このようなコーパスを作成する上で, 英語母語話者が述べた英語が一様に完全に自然な英語というわけではなく, また機械翻訳の翻訳結果は母語話者英語にとって自然であるかどうかは考慮されていないが, すべての出力が非母語話者英語であるわけではない. したがって, 品質の高い対訳コーパスを作成するために, その英文の自然性を図る母語話者英語分類器を構築する. この自然性とは, 英語母語話者にとって自然な翻訳スタイルであるかの程度を指し, これは

BERT により算出される母語話者英語予測確率をその文章が持つ自然性の値と定義する.

母語話者英語分類器は BERT を用いた分類器モデルである. BERT は, Transformers ベースの事前学習済みモデルであり, これにより文頭・文末の双方向から学習され, 文の理解だけでなく文脈を把握する能力も付与される [6]. これは, Mask Language Modeling(MLM) と Next Sentence Prediction(NSP) の２つのタスクにより実現される. MLM は BERT がマスク化された箇所の単語を予測する訓練タスクであり, NSP は BERT が２つの文の相互関連性を把握するための訓練タスクである. この事前学習済みモデルにファインチューニングを行うことで, 非母語話者英語と母語話者英語の分類器が構築される. ファインチューニングとは, ディープラーニングにおいて一般的な手法であり, 事前に訓練されたモデルをさらに訓練して調整を施すことである. これにより, 適切な訓練データをそのモデルにファインチューニングすることで, 特定のタスクにより適したモデルに調整することができる. 学習データとして, 「Wikipedia 日英京都関連文書対訳コーパス」を用いて, 英語母語話者が書いた英語をラベル 0, 日本人が書いた英語をラベル１を付け, データコーパスを構築した. この分類器では, 非母語話者英語の自然性並びに母語話者英語の自然性が数値化され, この２値は足して１になるように構成されている. テストセットでの精度検証の結果を図３に示されている. この図の横軸は, 構成される文章の単語数の範囲であり, 図が示すように, F1 スコアは共に 90%以上に達しており, これにより非母語話者英語と母語話者英語の分類が比較的正確に行えることが示された.

3.2.2　母語話者英語の抽出

求める対訳コーパスを作成するために, ステップ１では, 原文の英語に対して, 構築した母語話者英語分類器を用いて母語話者英語とそれに対応する日本語を抽出する. 英語の原文と日本語の原文は, 別々のテキストファイルであり, これらは行ごとで対応しあっているため, 双方を同時に読み込み, 英語が母語話者英語であるかを分類器を用いて判別し, 抽出する. この時, 分類器によって母語話者英語自然性が算出される. 本研究では, 母語話者英語の定義はこの自然性の閾値を 0.5 として, この値より大きい英文を母語話者英語と定義する. この処理により, 英語の原文 108,742 件において 38,864 件まで抽出され, 原文の 36%が母語話者英語であることが示された. また, 母語話者英語の定義上, 母語話者英語と判定された文の自然性は, 1.0 から 0.5 までであり, 幅がかなりある. したがって,

1.0 から 0.5 までの幅を 0.05 ずつに刻み, 自然性が 1.0 から 0.95 の間であり, 自然性が極めて高い英語から, 自然性が 0.55 から 0.5 の間であり, 自然性が比較的高くない英語まで 10 個にカテゴライズした (図 4). 図 4 の結果, 著者が英語母語話者であることから, 自然性が 1.0 から 0.95 で自然性が極めて高い英語が多いことが示された. このカテゴライズは, ファインチューニングするのに用いる訓練データとして母語話者英語であると判定されたより自然性の高いデータのみを扱う手法, 母語話者英語であると判定されたデータ全てを扱う手法の 2 通りのアプローチを比較する.

3.2.3 非母語話者英語の抽出

ステップ 1 によって日本語と母語話者英語の対訳コーパスを作成した. 次に, ステップ 2 では非母語話者英語と母語話者英語の対訳コーパスを作成するために, 日本語を機械翻訳を用いて英訳し, その英訳された文に対して母語話者英語分類器を用いて非母語話者英語を抽出し, 非母語話者英語と母語話者英語の対訳コーパスを作成する. 日本語を英訳する機械翻訳として Google 翻訳, DeepL 翻訳, みらい翻訳を用い, 作成される対訳コーパスは 3 通りとなる. 表 1 はそれぞれ Google 翻訳を用いたステップ 2, DeepL 翻訳を用いたステップ 2, みらい翻訳を用いたステップ 2 処理であり, これらはステップ 1 処理により取得した対訳コーパスに対して, それぞれ独立して処理が行われる. 表 1 における「1.0 - 0.95」は, ステップ 1 処理において抽出された英語の母語話者英語自然性が 1.0

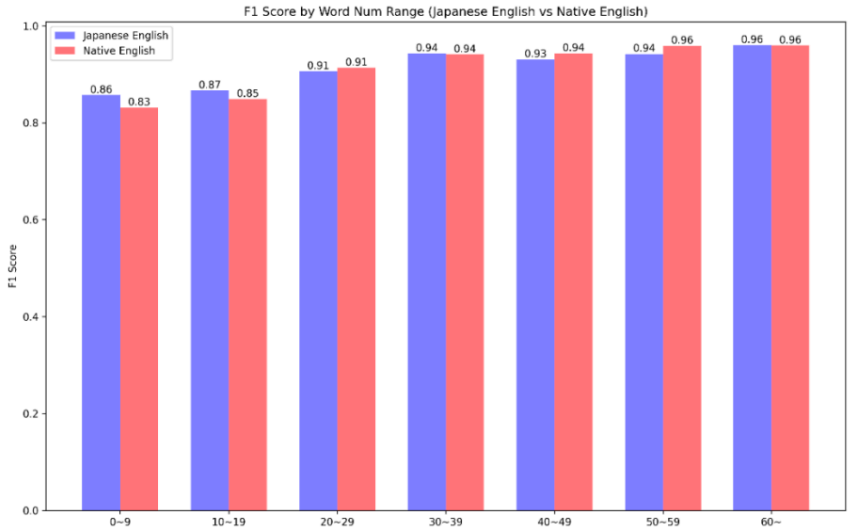

図 3: 分類器の F1 スコア

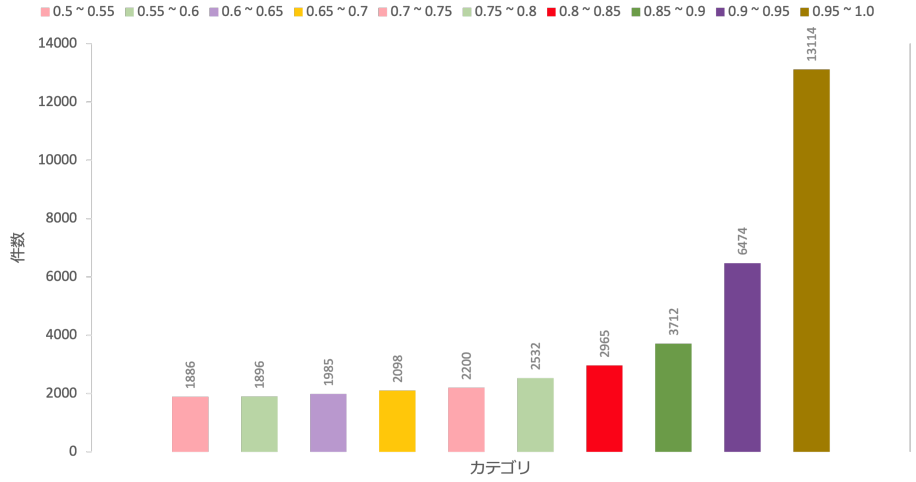

図 4: カテゴリ分けされた件数

から 0.95 の間の値を取ることを指している. それぞれ取得された対訳コーパスは処理前後において, Google 翻訳を用いた場合 21%, DeepL 翻訳を用いた場合 33%, みらい翻訳を用いた場合 35%を抽出した.

表 1: 抽出した対訳コーパス

		ステップ 2 処理に用いた機械翻訳:		
英語原文の自然性	処理前 (件)	Google 翻訳 (件)	DeepL 翻訳 (件)	みらい翻訳 (件)
1.0 - 0.95	13,114	1,337	2,971	2,936
0.95 - 0.9	6,474	1,305	2,353	2,364
0.9 - 0.85	3,712	703	1,129	1,212
0.85 - 0.8	2,965	648	1,013	1,121
0.8 - 0.75	2,532	616	919	1,026
0.75 - 0.7	2,200	603	866	953
0.7 - 0.65	2,098	661	890	950
0.65 - 0.6	1,985	682	884	974
0.6 - 0.55	1,896	722	850	968
0.55 - 0.5	1,886	818	929	1,037
合計	38,862	8,095	12,804	13,541

3.2.4 重複処理

非母語話者英語と母語話者英語の対訳コーパスを作成したが, それぞれのコーパスの中には完全に同じ文章が存在している. これは対訳データのサイズを大きくしているのみで, 完全に一致する複数文をファインチューニングのための訓練データとして用いることは, 不要である. 完全に一致する文が再度, 訓練データとして扱われないように削除する重複処理を行い, 3 通りの対訳コーパスにおいてより品質の高いコーパスを作成する. Google 翻訳を用いて作成された対訳コーパス, DeepL 翻訳を用いて作成された対訳コーパス, みらい翻訳を用いて作成された対訳コーパスをそれぞれ「Google 翻訳対訳コーパス」「DeepL 翻訳対訳コーパス」「みらい翻訳対訳コーパス」とする. 表 2 の結果からそれぞれの対訳コーパスにおいて, 100 件程度の重複処理が行われたことが確認された. この処理により, ファインチューニングにおいてより品質の高い対訳コーパスを取得した.

表 2: 重複処理済みの対訳コーパス

	Google 翻訳対訳コーパス		DeepL 翻訳対訳コーパス		みらい翻訳対訳コーパス	
英語原文の自然性	処理前 (件)	処理後 (件)	処理前 (件)	処理後 (件)	処理前 (件)	処理後 (件)
1.0 - 0.95	1,337	1,324	2,971	2,925	2,936	2,903
0.95 - 0.9	1,305	1,294	2,353	2,325	2,364	2,335
0.9 - 0.85	703	696	1,129	1,121	1,212	1,198
0.85 - 0.8	648	641	1,013	998	1,121	1,103
0.8 - 0.75	616	609	919	914	1,026	1,016
0.75 - 0.7	603	587	866	846	953	933
0.7 - 0.65	661	655	890	880	950	936
0.65 - 0.6	682	672	884	873	974	956
0.6 - 0.55	722	716	850	844	968	959
0.55 - 0.5	818	802	929	907	1,037	1,020
合計	8,095	7,996	12,804	12,633	13,541	13,359

第4章　母語話者英語変換器

本章では, 前章で作成した対訳コーパスを用いて, BART モデルのファインチューニングを行い, 非母語話者英語から母語話者英語への変換器を作成する. 本研究の目的は, この変換器を用いて機械翻訳の翻訳結果を事後編集し, 変換器によりその翻訳結果の英語母語話者にとっての自然さを改善されることである.

BART[7] は, 大規模なテキストデータで事前訓練された Transformer アーキテクチャに基づく Sequence-to-Sequence モデルであり, BERT の双方向エンコーダと GPT の自己回帰デコーダから構成されている. エンコーダは BERT のように全体的な文脈を把握し, 高品質なテキスト生成を可能し, デコーダは GPT のように各単語を生成する際に, 以前に生成された単語を考慮しテキストの一貫性と整合性を保証する. これにより, テキスト生成, 要約生成, 質問応答システム, 機械翻訳タスクなど多くの自然言語処理タスクにおいて優れた性能を発揮する. BERT と同様に, BART も特定のタスクに応じてファインチューニングすることで, その特定のタスクに特化したモデルの構築が行える.

4.1　母語話者英語への変換

本節では, 母語話者英語への変換を行う方法を紹介する. この変換には事後編集手法を用いられ, BART を用いた事後編集モデルにより事後編集を自動的に行う母語話者英語変換器を構築する. 事後編集のプロセスは, 主要な 3 つのステップがあり, 図 5 のように構成される. 入力された日本語から母語話者英語を出力する過程において, まず入力された日本語を機械翻訳を用いて英訳する. 使用する機械翻訳は, Google 翻訳, DeepL 翻訳, みらい翻訳の 3 通りである. 次に, 機械翻訳の翻訳結果を第 3 章で構築した母語話者英語分類器を用いて, その英語が母語話者英語であるかを母語話者英語自然性により判定する. これは, 機械翻訳の出力が完全に非母語話者英語を生成するとは限らないためである. 分類器による判定が母語話者英語である場合, その英語の自然性を改善する必要はないため, 翻訳のスピードと翻訳におけるコストがより高水準な翻訳が行われる. 本研究では, 閾値を 0.5 とし, 自然性が 0.5 以上の英語を母語話者英語, 0.5 未満の英語を非母語話者英語と定義する. 最後の処理として, 分類器により非母語話者英語と判定された英語を変換器を用いて, 母語話者英語に変換する. この処理は事後編集であり, BART を用いた事後編集モデルにより母語話者英語を

生成する. 次節では, 母語話者英語変換器の構築方法を述べる.

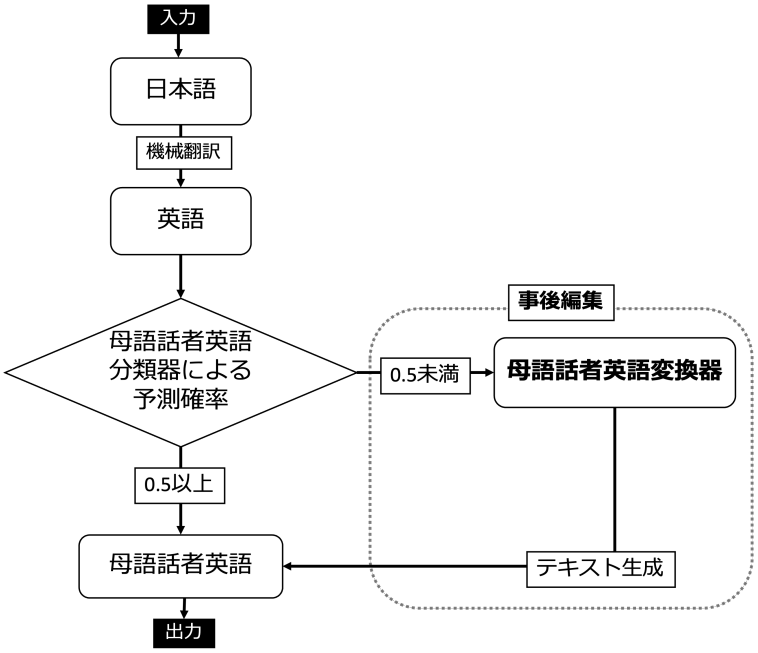

図 5: 事後編集の流れ

4.2 BART を用いた母語話者英語変換器

母語話者英語変換器では, 機械翻訳が出力した非母語話者英語から母語話者英語への変換, および事後編集を行う. この変換器は BART を用いた事後編集モデルであり, 入力された英語の自然さを改善した英語の生成し, 出力する. 入力された英語の自然さを改善することに重視した BART モデルを構築するために, 有効な訓練データによりファインチューニングを行う必要がある. 有効な訓練データとは, 適切な非母語話者英語と母語話者英語のペアデータであり, これによって訓練された事後編集モデルが生成した英語がより英語母語話者にとって自然であることを期待する. 本小節では, 訓練データの取得方法とファインチューニングの技術的詳細を説明する.

4.2.1 訓練データとテストデータ

高品質な事後編集モデルを構築するためには, 非母語話者英語と母語話者英語のペアで高品質な訓練データを準備しファインチューニングを行わなければな

らない. そのため, BART モデルのファインチィーニングに用いる訓練データを, 前章で作成した対訳コーパスから取得する. 作成した対訳コーパスは, Google 翻訳対訳コーパス, DeepL 翻訳対訳コーパス, みらい翻訳対訳コーパスの３通りを準備している. これらの対訳コーパスは, それぞれ総件数が 7,996 件, 12,633 件, 13,359 件であり, これらは対訳ペアの抽出におけるステップ１処理に対して, それぞれ独立してステップ２処理が行われており, 各対訳コーパスは「機械翻訳が出力した非母語話者英語 + TAB + 原文の母語話者英語」の形式において格納されている. それぞれステップ２処理が独立して行われているため, ある原文の母語話者英語に対して, 各機械翻訳が非母語話者英語を出力している場合がある. 例えば, Google 翻訳対訳コーパスと DeepL 翻訳対訳コーパスにおいて, それぞれ「(い)+ TAB +(あ)」「(う)+ TAB +(あ)」という対訳ペアが存在し, (あ) は原文の英語は母語話者英語であり, Google 翻訳または DeepL 翻訳による翻訳結果が (い), (う) のようにどちらも非母語話者英語である場合があり, 母語話者英語の重複が観察された. 図６は, このような対訳コーパス間における重複の関係を示す. 全ての対訳コーパスで重複しているデータ７はテストデータとして扱い, これは次章で詳細を述べる. したがって, データ１からデータ６までの組み合わせによって, ファインチューニングのための訓練データを取得する.

本研究では, 用いる機械翻訳に依存しない, 汎用性の高い事後編集モデル構築を目指すため, データ１からデータ６までのデータ７以外全ての対訳コーパスを訓練データとして用い, これを統合訓練データと呼ぶ. ここで, データ４は Google 翻訳対訳コーパスと DeepL 翻訳対訳コーパスの重複データであり, それぞれ「(い)+ TAB +(あ)」「(う)+ TAB +(あ)」のような対訳ペアが存在し, (あ) に対しての対訳が (い), (う) の２通り存在する. 重複データを双方とも訓練データとして扱う必要性はないので, random モジュールを使用し, (あ) の対訳ペアとして (い)(う) をランダムで選定する. これをデータ５, データ６にも同様に行うことで統合訓練データを取得した. また, 機械翻訳の依存性の有無による比較を行うために, Google 翻訳訓練データ, DeepL 翻訳訓練データ, みらい翻訳訓練データの３つの訓練データも取得する. Google 翻訳訓練データはデータ１, データ４, データ６の和集合, DeepL 翻訳訓練データはデータ２, データ４, データ５の和集合, みらい翻訳訓練データはデータ３, データ５, データ６の和集合により構成される. これにより, 統合訓練データ, Google 翻訳訓練データ, DeepL 翻訳訓練データ, みらい翻訳訓練データの４通りの訓練データを取得し

た. これらの訓練データは, ステップ２処理が母語話者英語自然性のカテゴリごとに取得され, 表３から取得した総件数はそれぞれ 13,662 件, 3,137 件, 7,774 件, 8,500 件である.

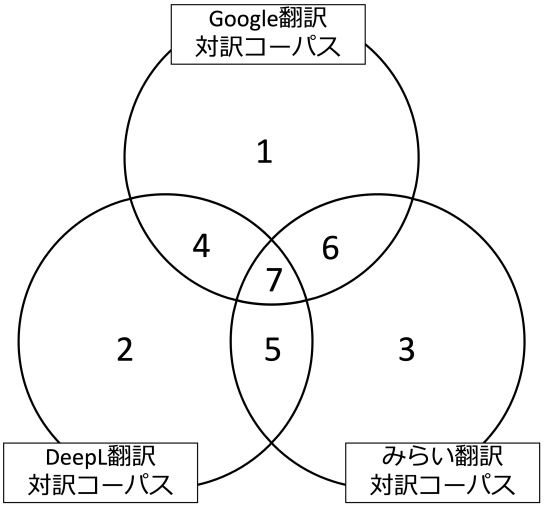

図 6: 各対訳コーパス間の重複関係

表 3: 取得した訓練データの分布

	訓練データ			
英語原文の自然性	統合 (件)	Google 翻訳 (件)	DeepL 翻訳 (件)	みらい翻訳 (件)
1.0 - 0.95	3,431	515	2,116	2,094
0.95 - 0.9	2,356	427	1,458	1,468
0.9 - 0.85	1,339	308	733	810
0.85 - 0.8	1,160	271	628	733
0.8 - 0.75	1,070	270	575	677
0.75 - 0.7	925	242	501	588
0.7 - 0.65	878	255	480	536
0.65 - 0.6	865	264	465	548
0.6 - 0.55	801	274	402	517
0.55 - 0.5	837	311	416	529
合計	13,662	3,137	7,774	8,500

4.2.2 BART のファインチューニング

前小節では, 統合訓練データ, Google 翻訳訓練データ, DeepL 翻訳訓練データ, みらい翻訳訓練データの４通りの訓練データを取得し, 本小節ではこれらの訓練データを用いて BART モデルをファインチューニングし, 事後編集モデルを構築する. このモデルの入力であるため, 事前学習済みモデルとして bart-base モデルを用い, これにより訓練に必要な時間と計算リソースを効率化する. これの大規模モデルとして bart-large があるが, bart-base モデルのパラメータが 139M であるのに対し, 約３倍の 406M であり, メモリ使用量の増加に伴い訓練時間が大幅に増加する問題があるため, 本研究ではこのモデルは用いない. バッチサイズを 16, エポック数 10 とし, 取得した４通りの訓練データを用いてそれぞれ BART をファインチューニングした. これにより構築された事後編集モデルをそれぞれ all モデル, google モデル, deepl モデル, mirai モデルと呼ぶ. また, 母語話者英語自然性が 1.0 から 0.95 に該当する訓練データのみでファインチューニングを行う手法と, 自然性が 1.0 から 0.5 までの全ての訓練データを用いてファインチューニングを行う手法の２つのアプローチを検証する. これにより, より自然性が優れた高品質な訓練データのみで訓練することが適切であるのか, 品質を問わず大量の訓練データで訓練することが適切であるのかの比較検証を行う. したがって, 自然性が 1.0 から 0.95 に該当する訓練データのみを用いてファインチューニングした事後編集モデルをそれぞれ all(1.0-0.95) モデル, google(1.0-0.95) モデル, deepl(1.0-0.95) モデル, mirai(1.0-0.95) モデルと呼ぶ. 一方で, 自然性が 1.0 から 0.5 までの全ての訓練データを用いてファインチューニングした事後編集モデルをそれぞれ all(1.0-0.5) モデル, google(1.0-0.5) モデル, deepl(1.0-0.5) モデル, mirai(1.0-0.5) モデルと呼び, 構築した事後編集モデルは８通りである.

4.2.3 母語話者英語変換器の構築

母語話者英語変換器は入力される日本語を機械翻訳を用いて英訳し, その翻訳結果である英語が非母語話者英語であれば, 構築した事後編集モデルを用いてより自然さを改善した英語を生成する. 用いる機械翻訳は, Google 翻訳, DeepL 翻訳, みらい翻訳の３通りである. また, 構築した事後編集モデルは all(1.0-0.95) モデル, google(1.0-0.95) モデル, deepl(1.0-0.95) モデル, mirai(1.0-0.95) モデル, all(1.0-0.5) モデル, google(1.0-0.5) モデル, deepl(1.0-0.5) モデル, mirai(1.0-0.5) モデルの８通りである. したがって, 母語話者英語変換器は機械翻訳と事後編集

モデルの組み合わせにより２４通りに構築される. 次章では, この構築したそれぞれの変換器の性能を評価し, 機械翻訳に依存しない事後編集モデルを明らかにする.

第5章　モデル評価

この章では, 構築した２４通りの母語話者英語変換器の性能を評価する. 具体的には, 出力された英語の自然性を検証し, 構築した各変換器の比較を行う. これは, 事後編集の有無による比較検証であり, 自動化された事後編集手法の有用性の検証である. 本章では, 性能評価に用いるテストデータと評価指標を紹介し, その指標に基づいた評価結果を述べる.

5.1　テストデータ

構築した変換器の評価および事後編集の有無による比較検証は, 前章で取得したテストデータを変換器に用いることで行う. このテストデータは, Google 翻訳対訳コーパス, DeepL 翻訳対訳コーパス, みらい翻訳対訳コーパスにおいて母語話者英語が重複しているデータの集合である. これは, 日本語とそれに対応する母語話者英語に対して, どの機械翻訳を用いてその日本語を英訳しても非母語話者英語が出力されることを意味する. 本研究において, 自然さが改善された英語を生成し, より汎用性の高い変換器の構築を目標としている. どの機械翻訳を用いて英訳しても非母語話者英語が出力される日本語文は, 構築した変換器を評価するためのテストデータとして最適なデータである. したがって, 英訳して得られた非母語話者英語文に対応する翻訳元の日本語文をテストデータとして使用する. 用いて使用する. テストデータは 4,861 件であり, 取得したテストデータはすべて訓練データには含まれておらず, テストデータの公平性と信頼性を確保している. また, テストデータは使用されるどの機械翻訳に用いても非母語話者英語を出力する日本語であるため, 全文が全ての構築した変換器において事後編集される.

5.2　評価指標

本研究の評価指標は, 英語の自然性である. つまり, その英語文が英語母語話者にとってどの程度自然であるかにおいて評価を行う. 具体的には, 日本語を機械翻訳により英訳して得られた英語と, 母語話者英語変換器により生成された英語に対して, 自然性を評価する. 本研究において, この自然性は母語話者英語予測確率の値あるいは単語の豊富性により求められる. 事後編集を行っていない英語と事後編集を行うことで得られた英語の自然性を比較することで, 構築

した事後編集モデルの評価が行われる.

5.2.1 母語話者英語分類器の自然性の向上率

事後編集を行っていない英語と事後編集を行うことで得られた英語の自然性を比較するために, 第３章で構築した母語話者英語分類器を用いる. 具体的に, 作成したテストデータを３通りの中から選択された機械翻訳により英訳し, その英語から８通りの中から選択された事後編集モデルにより事後編集済み英語を作成する. この２つの英語をそれぞれ事後編集前英語, 事後編集後英語とし, 分類器を用いて母語話者英語自然性をそれぞれ算出し比較評価する. テストデータの各文において事後編集前英語より事後編集後英語の自然性が上昇した件数を集計する. また, 各文の自然性の平均値をデータセット全体の自然性とし, 事後編集前英語と事後編集後英語の自然性の上昇率を算出し, 事後編集モデルによる自然さ改善の割合を評価する.

5.2.2 単語の豊富性

Markus らは, 英語母語話者にとってより自然な英語は豊富な語彙が用いられると述べている [2]. したがって, 構築した事後編集モデルが生成した英語の単語の豊富性を検証し, そのモデルに自然さの改善が観察できるかを評価する. 具体的に, 単語の豊富性は英語を構成する単語の延べ語数と異なり語数を算出し, 事後編集前英語と事後編集後英語において単語の豊富性の比較を行う. しかしながら, これらの値は文章の長さに依存する傾向にあるが, 事後編集前後で出力されるそれぞれの文章の長さは等しい限りではない. これでは, 単に長い文章を生成した方が自然性が高いと評価されかねない. そこで本研究では TTR[8] 手法を用いて得られた値を単語の豊富性とする. 延べ語数を N, 異なり語数を V とすると, TTR の値 t は

$$t = \frac{V}{N}$$

と定義され, この値 t が高いほどの単語の豊富性が高い. これを各行ずつ算出し, その平均をその事後編集モデルの持つ単語の豊富性とする.

5.3 変換器の性能評価

テストデータ 4,861 件を用いて, まず３通りの機械翻訳により英訳を行い, 事後編集前英語を取得した. ここで, それぞれ取得した英語の全体の自然性の分布は図７, 各行における t の値の分布は図８に示した. 加えて, これらの指標によ

る各機械翻訳の評価は表４で示される．表４の結果より Google 翻訳は自然性の平均と TTR が示す評価値 t が共に最大値であった．本研究の評価指標において，３通りの中で英語母語話者にとってより自然な翻訳を行った機械翻訳は Google 翻訳であったが，事後編集を用いて更なる自然さの改善を目指す．次小節では，各機械翻訳の翻訳結果に対して８通りの事後編集モデルを用いた検証を行う．

事後編集モデルの評価にあたり，事後編集前後において自然性が上昇した件数および総件数における上昇した件数の割合，事後編集により生成された英語の自然性の平均値と事後編集前後におけるその平均値の上昇率をそれぞれ上昇件数，上昇件数割合，平均値，自然性向上率と呼ぶ．また，事後編集前後における t の上昇率を単語上昇率と呼ぶ．

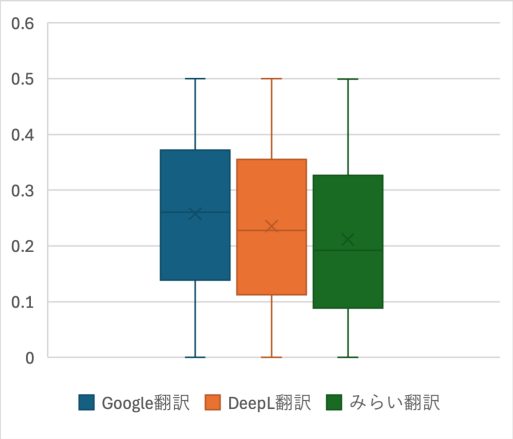

図 7: 自然性の箱ひげ図

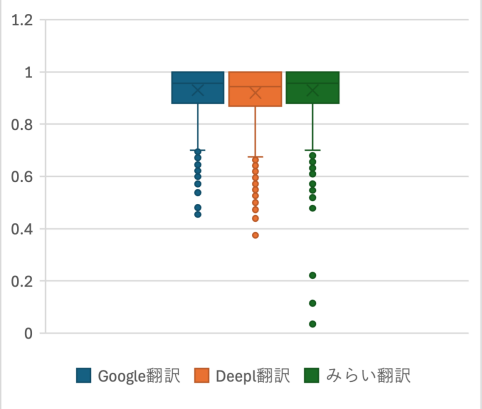

図 8: t の値の箱ひげ図

表 4: 各機械翻訳の評価

	Google 翻訳	DeepL 翻訳	みらい翻訳
自然性の平均	0.26	0.23	0.21
t	0.91	0.90	0.91

5.3.1 Google 翻訳が生成した英語の事後編集

Google 翻訳の翻訳結果に対して，８通りの構築した事後編集モデルの検証を行った．この時，自然性向上率は平均値から Google 翻訳を用いて得られた自然

性の平均 0.26 を用いて, 単語上昇率は t の値から Google 翻訳を用いて得られた t の値 0.91 をを用いて算出した値である. 検証結果は表 5 で示され, all(1.0-0.95) モデルは平均値並びに自然性向上率の結果より, 自然性による評価において, 自然性が 79.7%向上し, 最も優れた性能が示された. 一方, t の値および単語上昇率より, deepl(1.0-0.95) モデルは豊富性において 0.51%の向上が観察され, 最も優れていると示された.

表 5: Google 翻訳における事後編集モデルの評価

	(1.0-0.95)				(1.0-0.5)			
事後編集モデル	all	google	deepl	mirai	all	google	deepl	mirai
上昇件数 (件)	**3,442**	1,647	3,319	3,130	3,402	3,108	3,287	3,270
上昇件数割合 (%)	**70.8**	33.9	68.3	64.4	70.0	63.9	67.1	68.4
自然性の平均	**0.47**	0.30	0.46	0.42	0.45	0.40	0.44	0.42
自然性向上率 (%)	**79.7**	16.9	75.2	61.9	71.6	53.9	68.2	61.3
t	0.91	0.91	**0.92**	0.91	0.91	85.4	0.91	0.91
単語上昇率 (%)	0.39	0.35	**0.51**	0.23	0.18	0.07	0.43	0.20

5.3.2 DeepL 翻訳が生成した英語の事後編集

DeepL 翻訳の翻訳結果に対して, 8 通りの構築した事後編集モデルの検証を行った. この時, 自然性向上率は平均値から Google 翻訳を用いて得られた自然性の平均 0.23 を用いて, 単語上昇率は t の値から DeepL 翻訳を用いて得られた t の値 0.90 を用いて算出した値である. 検証結果は表 6 で示され, all(1.0-0.95) モデルは平均値並びに自然性向上率の結果より, 自然性による評価において, 自然性が 94,.4%向上し, 最も優れた性能が示された. 一方, t の値および単語上昇率より, all(1.0-0.95) モデル並びに deepl(1.0-0.95) モデルは豊富性において 1.31% の向上が観察され, 最も優れていると示された.

表 6: Deepl 翻訳における事後編集モデルの評価

	(1.0-0.95)				(1.0-0.5)			
事後編集モデル	all	google	deepl	mirai	all	google	deepl	mirai
上昇件数 (件)	3,584	1,943	3,438	3,297	**3,626**	3,323	3,527	3,433
上昇件数割合 (%)	73.7	38.1	70.7	67.8	**74.6**	68.4	72.6	70.6
自然性の平均	**0.45**	0.28	0.44	0.41	0.44	0.38	0.43	0.41
自然性向上率 (%)	**94.4**	21.3	86.5	75.0	89.1	61.8	82.3	75.5
t	**0.91**	0.91	**0.91**	0.90	0.91	0.90	0.91	0.90
単語上昇率 (%)	**1.31**	0.98	**1.31**	0.70	1.29	0.78	1.30	0.84

5.3.3　みらい翻訳が生成した英語の事後編集

みらい翻訳の翻訳結果に対して，８通りの構築した事後編集モデルの検証を行った．この時, 自然性向上率は平均値からみらい翻訳を用いて得られた自然性の平均 0.21 を用いて, 単語上昇率は t の値から Google 翻訳を用いて得られた t の値 0.91 を用いて算出した値である．検証結果は表７で示され, all(1.0-0.95) モデルは平均値並びに自然性向上率の結果より, 自然性による評価において, 自然性が 106.1%向上し, 最も優れた性能が示された．一方, t の値および単語上昇率より, all(1.0-0.95) モデルは豊富性において 0.47%の向上が観察され, 最も優れていると示された．

表 7: みらい翻訳における事後編集モデルの評価

	(1.0-0.95)				(1.0-0.5)			
事後編集モデル	all	google	deepl	mirai	all	google	deepl	mirai
上昇件数 (件)	**3,559**	1,851	3,405	3,278	3,548	3,135	3,405	3,419
上昇件数割合 (%)	**73.2**	38.1	70.0	67.4	73.0	64.5	70.0	70.3
自然性の平均	**0.44**	0.27	0.43	0.40	0.41	0.36	0.40	0.39
自然性向上率 (%)	**106.1**	25.7	103.2	86.8	95.7	67.7	90.6	84.0
t	**0.91**	0.91	0.91	0.91	0.91	0.91	0.91	0.91
単語上昇率 (%)	**0.47**	0.36	0.36	0.34	0.45	0.28	0.60	0.40

第6章　考察

本章では, 前章で得られた検証結果に基づき, 構築した事後編集モデルの性能を考察する. 事後編集モデルの有効性を示した上で, 事後編集モデルの課題や改善点を述べる.

6.1　変換器の有効性

構築した母語話者英語変換器２４通りにおいて, 事後編集前後の自然性および事後編集前に生成された英語の全体の母語話者英語自然性の平均と事後編集後に生成された英語の全体の自然性の差, 事後編集前後の t の値の差は共に全ての変換器で正の値が得られた. また, 事後編集前後において自然性が上昇した件数とテストデータとして用いた件数の割合は, google(1.0-0.95) モデルを除いた全ての事後編集モデルで正の値が得られた.

どの機械翻訳に用いても同等の性能であり, 汎用性の高い事後編集モデルとして, all(1.0-0.95) モデル並びに all(1.0-0.5) モデルを構築し, その比較として google(1.0-0.95) モデル, google(1.0-0.5) モデル, deepl(1.0-0.95) モデル, deepl(1.0-0.5) モデル, mirai(1.0-0.95) モデル, mirai(1.0-0.5) モデルを構築した. しかしながら, どの事後編集モデルも用いた機械翻訳に依存しない結果が得られ, deepl(1.0-0.95) モデルは all(1.0-0.5) モデルの性能を凌ぐほどの性能評価が得られた. (1.0-0.95) と (1.0-0.5) は, より自然性が優れた高品質な訓練データのみで訓練することが適切であるのか, 品質を問わず大量の訓練データで訓練することが適切であるのかの比較のために, 作成した訓練データから母語話者英語が 1.0 から 0.95 のデータを抽出してファインチューニングしたモデルと, 母語話者英語が 1.0 から 0.5 として作成された全訓練データを用いてファインチューニングしたモデルを意味する. それぞれの訓練データにより all モデル, google モデル, deepl モデル, mirai モデルを構築して評価を行ったが, どのモデルも (1.0-0.5) のモデルより (1.0-0.95) のモデルの方がより自然さ改善において優れた性能を有することが観察された.

特に, all(1.0-0.95) モデルは, どの機械翻訳の翻訳結果に対しても 70%以上の確率で事後編集前後における自然性の向上が行われた. また, 事後編集前後における全文の自然性の平均および自然性の上昇率は, どの機械翻訳の結果に対しても 79%以上の向上が見られ, 事後編集後の自然性は 0.45 前後であり, 構築し

た事後編集モデルの中で最も優れたこのモデルは, 確実に自然さの改善に貢献している.

また, 機械翻訳の翻訳結果は長い文章で構成されるのに対して, その結果から構築した全ての事後編集モデルにより生成された英語は比較的短い文章で構成されていた. しかしながら, t の値は事後編集モデルにより生成した英語の方が高い値を示した. つまり, より短い文章かつより様々な種類の単語を用いて文章を構成していることが示され, 事後編集モデルを用いることで, 単語の豊富性が向上した生成が行われ, これはどの事後編集モデルも似た性能を持つことが観察された.

以上のことから, 構築した事後編集モデルは翻訳のスタイルを, 英語母語話者にとって自然な文章に改善することに, 概ね貢献している. 例えば, all(1.0-0.95) モデルを用いて, 自然さが改善された例を表 8 に示した. これは「そうやって５２日間で完成したのです.」という日本語を用いて得られたそれぞれの英語とその英語の自然性である. 事後編集前後で自然性が大幅に改善されており, 事後編集後の英語は完成したこととそれに５２日間かかったことの因果関係が明確に表現されている. したがって, 提案した手法による翻訳の自然性の改善は, 有効であることが示された.

表 8: 事後編集の例

	生成した英語	自然性
事後編集前	That is how it was completed in 52 days.	0.05
事後編集後	It took 52 days to complete.	0.76

6.2 変換器の課題

構築した変換器および事後編集モデルは有効であることが示されたが, これらの課題, 改善点も残っている. まず, google モデルを構築するのに十分な訓練データを収集できなかったことである. 訓練データの件数が構築した事後編集モデルで異なっており, google(1.0-0.95) モデルに用いられた訓練データの量が他のモデルの訓練データの量と比べて, 極端に少なくなってしまった. また, 全てのモデルにおいて, 全テストデータの自然性の改善には至らなかった. 事後編集モデルを用いても, 事後編集前後で文章が修正されずそのまま出力された場合が観察された. また, 単に自然性が減少した場合も存在しており, 翻訳生成を自然性に重視した結果, 翻訳の適切性による評価は, 事後編集前後で劣っている場合が観察された.

第7章　おわりに

本研究では, 日英翻訳において, 英語母語話者の英語を生成するための機械翻訳の事後編集手法を提案した. 事後編集とは, 機械翻訳で翻訳した後にその翻訳結果を英語母語話者にとって自然なスタイルの英語に修正することである. 本研究では, 日本語を既存の機械翻訳に通して得られた非母語話者英語を母語話者英語に変換することを目的とする. 具体的には, 日英の対訳コーパスから非母語話者英語と母語話者英語のペアを抽出し, このデータを用いて BART 事前学習済みモデルをファインチューニングすることで, 非母語話者英語から母語話者英語への変換器を構築した.

本研究の貢献は以下の 2 点である.

非母語話者英語と母語話者英語の対訳コーパスの構築

非母語話者英語を母語話者英語に変換する事後編集器を構築するには, 非母語話者英語とそれに対応する母語話者英語のペアが必要である. つまり，日本人が翻訳した英語とそれを英語母語話者が事後編集したデータを取得しなければならない．また，翻訳コストを抑えるためには，できる限り非母語話者英語と推定されるものに原文段階で限定する必要がある. 本研究では, NICT が提供する日英対訳コーパスを用いて非母語話者英語と母語話者英語の対訳コーパスを作成した. Google 翻訳, DeepL 翻訳, みらい翻訳の 3 通りの機械翻訳を用いて, その対訳コーパスを約 35,000 件取得した.

事後編集モデルのファインチューニング

作成した非母語話者英語と母語話者英語のペアを用いて事後編集器を構築した. Google 翻訳や DeepL 翻訳などの多様な機械翻訳に対して母語話者英語を生成できるように，汎用的な事後編集モデルの構築を目指し, その事後編集モデルを用いて事後編集を行うことで, 機械翻訳の翻訳結果の英語母語話者にとっての自然性を向上させることが有効であることを示した. また, 構築したどのモデルも汎用的であるが, 特に all(1.0-0.95) モデルは常に最も優れた性能を示し, どの機械翻訳の出力に対しても約 21%の自然性の向上が見られた. さらに, 事後編集前後の結果を比較したところ, 事後編集モデルによって生成された英語は, より豊かな単語表現がされることが示された.

本研究では, 訓練データは量ではなく, より高品質な訓練データを用いたモデルほど優れた性能を示したことから, 将来的にはさらに多くの高品質なデータを収集することで, モデルの事後編集能力を一層向上させることが可能であると示唆された. また, 機械翻訳の自然性を向上させる方法は他にも存在する. 例えば, 機械翻訳で翻訳する前に日本語を, 母語話者英語を生成しやすい日本語に変換する事前編集手法がある. 今後は翻訳の適切性向上とともに, 翻訳の自然性の向上を目指し, これにより生成された英語が英語母語話者にとって本当に自然な表現であるのか評価を行う.

謝辞

本研究を行うにあたり, 熱心なご指導, ご助言を賜りました指導教官の村上 陽平教授と Mondheera Pituxcoosuvarn 助教, ZHANG Yuxuan 先輩に深謝申し上げます. また, この四年間普段からお世話になっていた社会知能研究室の皆さまに心より感謝申し上げます.

参考文献

[1] 鳥飼玖美子. ”国際共通語としての英語.” 東京: 講談社現代新 (2011).

[2] Markus Freitag, David Vilar, David Grangier, Colin Cherry and George Foster. ”A Natural Diet: Towards Improving Naturalness of Machine Translation Output.” Findings of the Association for Computational Linguistics: ACL 2022, 3340–3353 (2022).

[3] 竹林佑斗, 荒瀬由紀, 永田昌明. ”ニューラル機械翻訳における単語報酬モデルに基づく対訳辞書の利用.” 自然言語処理 26.4. 711-731 (2019).

[4] 瓦 祐希. ”事前並び替え手法による機械翻訳性能の向上.” (2021)

[5] 井佐原均. ”機械翻訳の実用的利用に向けた取り組み.” Japio year book 232-236 (2012).

[6] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” Proceedings of naacL-HLT. Vol. 1. No. 2. (2019).

[7] Lewis, Mike. ”Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.” arXiv preprint arXiv:1910.13461 (2019).

[8] Van Hout, Roeland, and Anne Vermeer. ”Comparing measures of lexical richness.” Modelling and assessing vocabulary knowledge: 93. 115 (2007).

付録

構築した all(1.0-0.95) モデルが生成した事例を DeepL 翻訳が生成した英語と比較して掲載する.

A.1　事後編集の成功事例

		自然性
例文１	基地内では、指定された喫煙所をご利用ください。	
事後編集前	Please use the designated smoking areas on the base.	0.03
事後編集後	Guests are encouraged to use designated smoking areas on the base.	0.84
例文２	8 つの商店街に約 5,000 軒の店舗が並び、市場で働く人の数は約 2 万人に及ぶ。	
事後編集前	About 5,000 stores line the eight shopping streets, and the number of people working in the market is approximately 20,000.	0.16
事後編集後	About 5,000 shops are located along eight shopping streets, employing approximately 20,000 people.	0.88
例文３	天気は、ひどく荒れ、きびしい寒さが続きました。 雪は、２メートルほども積もりました。そして至るところに氷が張って、牛や人間が乗っても割れないほど固く凍っていました	
事後編集前	The weather was severely stormy and bitterly cold. Snow piled up to two meters high. And there was ice everywhere, frozen so hard that cows and people could not break it even if they rode on it.	0.15

事後編集後	The weather was very stormy and bitterly cold, with snow as high as two meters high, and ice everywhere, so hard that the cows and men could not break it.	0.90
例文４	仕事の前や夕方に歩くことができます。	
事後編集前	You can walk before work or in the evening.	0.12
事後編集後	A walk may be taken before work or in the evening.	0.73
例文５	これ以外に農業復興の道はない。	
事後編集前	There is no other way to restore agriculture.	0.31
事後編集後	There is no other remedy for restoring agricultural productivity.	0.75

A.2　事後編集の失敗事例

		自然性
例文１	まさに活きる力がなかった。小枝１本つかもうとすることができず、まるで体から落ちてしまいそうな感じだった。	
事後編集前	I had exactly no power to live. I couldn't even try to grab a single twig; it was as if it were going to fall off my body.	0.46
事後編集後	I had no power to live; I could not reach a single twig; it was as if I were going to fall off my frame.	0.40
例文２	中島を歩き回ってできた道がある。すぐにたどり着いた。向こうに釣り人が座っていた。	
事後編集前	There is a path that was created by walking around Nakajima. I soon reached it. A fisherman was sitting over there.	0.09

事後編集後	A path was created by walking around Nakajima, and I soon reached it, where a fisherman was sitting on a bench.	0.69
例文３	今後は、国会で法改正を求めるしかありません。	
事後編集前	In the future, we have no choice but to ask the Diet to amend the law.	0.25
事後編集後	In the future, the Diet may decide to amend the law.	0.81
例文４	6月29日の午後、数十人の囚人が刑務所からの脱走を強引に実行した。	
事後編集前	On the afternoon of June 29, dozens of prisoners forced their way out of the prison.	0.36
事後編集後	On the afternoon of 30 June, hundreds of prisoners broke out of the prison.	0.95
例文５	自分のせいで遅くなったくせに、難癖をつけて相手のせいにする者がいる。 怠け者は得てしてそういうものだ。	
事後編集前	There are those who blame the other party for being late, even though it was their own fault. Lazy people are always like that.	0.07
事後編集後	There are those who blame the other party for being late, even though it was their own fault. Lazy people are always like that.	0.07

例文１: 自然性が減少した例

例文２, 例文３, 例文４: 自然性は上昇したが, 翻訳が不適切な例

例文５: 事後編集が行われず出力された例